Modeling and estimating income data in the presence of distinctive zero and heaped responses

Dissertation

zur Erlangung des akademischen Grades

eines Doktors der Sozial- und Wirtschaftswissenschaften
(Dr. rer. pol.)

an der Fakultät Sozial- und Wirtschaftswissenschaften
der Otto-Friedrich-Universität Bamberg

vorgelegt von
Ariane Würbach, Magistra Artium
geboren am 11. März 1981 in Erfurt

Bamberg, März 2016

Thesis Advisor:	Prof. Dr. Susanne Rässler University of Bamberg, Germany
Reviewers:	Prof. Dr. Ulrich Rendtel Freie Universität Berlin, Germany Prof. Dr. Guido Heineck University of Bamberg, Germany
Date of Submission:	March 8, 2016
Date of Defense:	September 20, 2016

Abstract

A major part of research data in the social sciences originates from survey interviews. Besides the issue of non-response, questions concerning the accuracy of self-reported data are important research objectives. The focus of this thesis is on heaping behavior in surveyed income data. Heaping, i.e. aberrant concentrations of response values at specific points of the range, is typical for retrospective data, when the respondent is either uncertain about the true value or hesitates to report. A theoretical framework and explanations for heaping are presented. Measurements for heaping and appropriate strategies to cope with it are discussed afterwards. Heaped data are linked with a loss of information and hence are found to deteriorate effects on the macro- and micro-level. Therefore, exploration of the relationships between heaping behavior and personal as well as context information is valuable. This work provides descriptive evidence for heaping behavior in the income data of the German National Educational Panel Study (NEPS). The data at hand strongly support the assumption that heaping behavior is not stochastic but deterministic, i.e. whether and to which degree heaping occurs is not random. Respective determinants influencing heaping behavior are the response value itself and common socio-economic characteristics. Male, higher educated, and older respondents have a higher propensity to heap their income. Because of that, there is a necessity of adequately addressing this issue, e.g. by a modeling strategy which explicitly takes the non-randomness of the heaping behavior into consideration. According to this, a heaping model is introduced enabling to account for different heaping behaviors. The model is a mixture of two components, the latent distribution and the model for the heaping behavior. A zero-inflated log-normal distribution with a piecewise constant heaping mechanism is defined as base model. The generality and flexibility of the established model is outlined by several modifications and extensions, with respect to the latent distribution, the heaping pattern as well as the heaping mechanism. In the application, all proposed models are explored concerning their fit to the NEPS income data. Posterior predictive checks are used to access the overall fit of the models. This thesis also includes a comparative analysis of different random-walk Metropolis (RWM) algorithms with respect to their estimation accuracy and efficiency. Besides the original RWM algorithm, blocking and adaptive strategies are inquired into. The results indicate that blocking can greatly improve mixing and convergence of the RWM algorithm, in contrast to the adaptive schemes considered. The performance of the models is fairly good, however, large differences in estimation exist with respect to runtime and efficiency. These differences are mainly attributable to the model assumed and the selected specification of the RWM algorithm.

Keywords: heaping, finite mixture model, random-walk Metropolis algorithm, block Metropolis-Hastings algorithm, adaptive MCMC, posterior predictive checks

Authorship and Publications

The first publication this thesis refers to is a journal article co-authored with Dr. Sabine Zinn, published in the *Journal of Applied Statistics* (Zinn & Würbach, 2015). A previous version of this paper was contributed as *NEPS Working Paper* (Zinn & Würbach, 2014). Main object of these papers is the introduction of the heaping model. Major differences between both versions relate to modifications in the model for the latent true distribution.

In the second publication, the author of this thesis introduces a Bayesian estimation procedure of the heaping model by means of some blocking strategies of the random-walk Metropolis algorithm (Würbach, 2015). This work was presented at the $30th$ International Workshop on Statistical Modelling in Linz (Austria).

Additionally, this thesis borrows to a small extent from the article published in *Sociological Methods and Research* by Aßmann, Würbach, Goßmann, Geissler, and Biedermann (2015) in which a multiple imputation technique for the NEPS income data is unfolded taking the peculiarities of the data structure into account.

All publications stem from joint work of members of the LIfBi working group "Methods, Weighting and Imputation" of Department 2 – Data Center and Method Development.

Acknowledgements

This thesis uses data from the German National Educational Panel Study (NEPS): Starting Cohort 6 – Adults, `doi:10.5157/NEPS:SC6:1.0.0`. From 2008 to 2013, NEPS data were collected as part of the Framework Programme for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, the NEPS survey is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

Contents

List of Figures

List of Tables

List of Abbreviations

ACF	autocorrelation function
adj.	adjusted
AIC	Akaike information criterion
ALLBUS	German General Social Survey
AM	adaptive Metropolis
AP	adaptive proposal density
approx.	approximately, approximated
AR	acceptance rate
asym.	asymmetric, asymmetrically
BF	Bayes factor(s)
BFGS	Broyden-Fletcher-Goldfarb-Shanno algorithm
BIC	Bayesian information criterion
BNM	constrained and simple bounded Nelder-Mead algorithm
CAPI	computer-assisted personal interview
CAR	coarsened at random
CART	classification and regression trees
CATI	computer-assisted telephone interview
cdf	(theoretical) cumulative distribution function
CHINTEX	Change from Input Harmonisation to Ex-post Harmonisation in National Samples of the European Community Household Panel
CI	confidence interval(s)
CLT	central limit theorem
coeff	coefficient
Cov	covariance
COV	coverage
cp.	compare
CPS	Current Population Survey
CPU	central processing unit
CV	cross-validation
Dag	Dagum distribution
df	degrees of freedom
DGP	data generating process
diag	diagonal of a matrix
ecdf	empirical cumulative distribution function
ECHP	European Community Household Panel
edf	empirical density function
e.g.	[from latin "exempli gratia"] for example
ESS	effective sample size
etc.	[from latin "et cetera"] and so forth
EUR	Euro (European currency)
exp	exponential function

ext	external factor
f.	and following (singular)
ff.	and following (plural)
fmi	fraction of missing information
GB	Generalized Beta distribution
GC	Gini coefficient
HDR	highest (posterior) density region(s)
HM	heaping mechanism
HP	heaping point(s)
i.a.	[from latin "inter alia"] among other things
IAB	Institute for Employment Research
IAT	integrated autocorrelation time
ibid.	[from latin "ibidem"] in the same place (book, etc.)
ICS	initial convex sequence estimator
id.	[from latin "idem"] the same (man)
i.e.	[from latin "id est"] that is
IHS	inverse hyperbolic sine
iid	independent identically distributed
IMS	initial monotone sequence estimator
Ineff	inefficiency factor
int	internal factor
IPS	initial positive sequence estimator
I-RWM	MB-RWM algorithm, blocks separated by interval
KDE	kernel density estimation
KS	Kolmogorov-Smirnov test
LFS	Labour Force Survey
LIfBi	Leibniz Institute for Educational Trajectories
log	logarithm function
LRI	latent optimal rounding intensity
LVM	latent variable modeling
MBI	Myers' blended index
MB-RWM	multiple-block random-walk Metropolis algorithm
MC	Monte Carlo simulation
MCMC	Markov Chain Monte Carlo (algorithm)
MH	Metropolis-Hastings algorithm
MICE	multivariate imputation by chained equations
min	minimum function
ML	Maximum Likelihood
mod	modulo function
MPSRF	multivariate potential scale reduction factor
M-RWM	MB-RWM algorithm, blocks separated by modulo
MSE	mean squared error
NA	not available

NEPS	National Educational Panel Study
NEPS-HH	NEPS net household income
NEPS-Ind	NEPS net individual income
NSE	numerical standard error
OECD	Organization for Economic Cooperation and Development
OMC	Ordinary Monte Carlo
Par	parameter(s)
pbsm	piecewise bell-shape model
pcm	piecewise constant model
pdf	probability density function
Perc	percentage, percentages
p.m.	per month
PPC	posterior predictive checks
ppp	posterior predictive p-value
$PSRF$	potential scale reduction factor(s)
RAM	robust adaptive Metropolis algorithm
$RAMA$	regional adaptive Metropolis algorithm
repeatc	(average) number of repeats for sampling a candidate value
repeatp	(average) number of repeats for sampling a starting value
RI	Rounding Indicator
RMB-RWM	randomized multiple-block random-walk Metropolis algorithm
RNE	relative numerical efficiency
RQ	Rounding Quotient
RSM	Rounding Strain Measure
RWM	random-walk Metropolis algorithm
SC	starting cohort
SC	Schwarz's criterion
$SCAM$	single-component adaptive Metropolis algorithm
SD	standard deviation
SE	standard error
SHP	The Swiss Household Panel
sic.	[from latin "sic erat scriptum"] thus it was written
sig.	significance, significant, significantly
SOEP	Socio-Economic Panel
S-RWM	simple random-walk Metropolis algorithm
std.	standard
SUF	scientific use-file
sym.	symmetric, symmetrically
Var	variance
VC	covariance matrix
vs.	[from latin "versus"] against
WI	Whipple's index

List of Symbols

$\mathcal{D}$	digit that is to be explored
$\mathbb{E}$	expectation value
$\mathcal{I}_x$	identity matrix of dimension x
$\mathbb{I}$	indicator function
∞	infinity
Ω	correlation matrix
Σ	covariance matrix
$\mathbb{N}$	natural numbers
$\mathbb{R}_0^+$	real positive numbers including zero
N	population size
n	sample size
$_i$	index for individuals
$_p$	index for processing step
$_g$	index for repeated measurements or trials
Y_i	value of a construct for the i-th individual in the population, with $i = 1, \ldots, N$
y_i	true value of the measurement for the i-th individual in the sample, with $i = 1, \ldots, n$
z_i	reported value of the measurement for the i-th individual in the sample, with $i = 1, \ldots, n$
z_{ip}	reported value of the measurement for the i-th individual in the sample after editing and other processing steps
z_{ig}	reported value of the measurement for the i-th individual in the g-th measurement or trial
ε_i	measurement error, difference between reported and true value of the measurement of an individual i
ξ	half interval width
ι	variance of random term in (log-)linear model
X	covariates determining Z
β	parameters for X, the covariates determining Z
W	covariates determining the heaping mechanism
γ	parameters for W, the covariates determining the heaping mechanism
$\mathcal{H}$	set of heaping points
$_b$	index for heaping points
S	number of heaping points considered
h_b	heaping point $h_b \in \mathbb{N}$, with $b = 1, \ldots, S$
I_b	heaping intervals for heaping point h_b
l_b	lower bound of the heaping interval I_b
u_b	upper bound of the heaping interval I_b
θ	model parameters, with $\theta = [\phi, \psi]$
$_d$	index for model parameter

D	number of parameters in vector θ
θ_d	component of θ, specific model parameter, with $d = 1, \ldots, D$
Θ	parameter space
$\Re$	parameter region
$\widehat{\theta}$	estimates of the model parameters
$\widehat{\theta}_{ML}$	maximum likelihood estimates
θ'	posterior mean
$h(\theta)$	quantity of interest, usually a specified scalar estimand
$\bar{h}_T$	approximation to the quantity of interest
ϕ	parameters of the heaping mechanism
$v_b(y)$	heaping probability function
ρ_b	heaping probabilities in the piecewise-constant heaping function, with $b = 1, \ldots, S$
$y_{i,b,(0.5)}$	median value of y_i for heaping point h_b
$\widehat{y}_{i,b,(0.5)}$	approximation to the median value of y_i for heaping point h_b
η_b	heaping probabilities in the piecewise bell-shaped heaping function, with $b = 1, \ldots, S$
$f(y)$	true underlying probability distribution function
$F(y)$	true underlying cumulative distribution function
ψ	parameters of the underlying true distribution
μ	shape parameter of the log-normal distribution
σ	scale parameter of the log-normal distribution
$\Phi(\cdot)$	standard normal distribution function
ρ_Z	inflation parameter
a	first shape parameter of the Dagum distribution
b	scale parameter of the Dagum distribution
p	second shape parameter of the Dagum distribution
$B(\cdot)$	Beta function
$\Gamma(\cdot)$	Gamma function
$g_1(z_i\|\psi,\phi)$	density of observing z_i if the true value y_i is not heaped, i.e. $z_i = y_i$
$g_2(z_i\|\psi,\phi)$	density of z_i falling on a heaping point h_b, i.e. $z_i \neq y_i$
$\mathcal{L}$	(approx.) likelihood function
ℓ	(approx.) logarithmized likelihood function
$\mathcal{C}_\psi$	linear restrictions on the parameters of the underlying distribution
$\mathcal{C}_\Sigma$	usual positivity and positive definiteness constraints on matrices
$\mathcal{C}_\Theta$	constraints imposed on the model parameters
$\mathbf{y}$	data
$p(\mathbf{y}\|\theta)$	likelihood function
$p(\theta)$	prior distribution function
$p(\theta\|\mathbf{y})$	posterior distribution function
C	normalizing constant
$g(\theta\|z_i)$	posterior distribution proportional up to the normalizing constant
$q(\cdot)$	proposal density/jumping distribution

n_0	number of iterations considered for burn-in
T	Markov chain sample size
t	iteration number in Markov chain
$\{\theta^{(t)}\}_{t=1}^{T}$	Markov chain, a sequence of random elements, with $t = 1, \ldots, T$
T^*	effective sample size
T_A	fraction of early iterations
T_B	fraction of later iterations
ϑ	part of θ, for either $t = 1, \ldots, T_A$ or $t = T - T_B + 1, \ldots, T$
$\mathcal{S}_h$	spectral density for time series h at point zero
$\theta^{(0)}$	starting value for RWM
θ^*	candidate draw
ϵ	stochastically independent random perturbation
δ	random perturbation dependent on $\theta^{(t-1)}$
λ	scale factor for proposal density
υ	mean vector for parameters of the latent distribution ψ
Υ	covariance matrix for the parameters of the latent distribution ψ
Σ_q	covariance matrix of the proposal distribution
α	probability of move
π	rejection probability
$\mathcal{K}$	transition kernel for the Hastings update
$\mathcal{U}$	uniform distribution function
u	random number from uniform distribution function
$\mathcal{N}_x$	(multivariate) normal distribution function of dimension x
$_k$	index for blocks of θ
K	number of blocks θ is divided into
$\bar{K}$	average number of blocks from multiple runs of the RMB-RWM algorithm
θ_k	specific block of θ, with $k = 1, \ldots, K$
Ψ_{k-1}	parameter blocks below k
Ψ^{k+1}	parameter blocks beyond k
m	thinning interval
τ	inefficiency factor
$_l$	index for a lag in the autocorrelation function (ACF)
L	lag at which the ACF tapers off
κ_l	ACF at lag l, with $l = 1, \ldots, L$
γ_l	the l-th autocovariance of the sequence h_t
$\mathcal{B}$	between-chain variation
$\mathcal{W}$	within-chain variation
$\widehat{R}$	squared potential scale reduction factor ($PSRF$)
$_o$	index for models
O	number of models
$\mathcal{M}_o$	model to be tested, with $o = 1, \ldots, O$
$\mathcal{A}$	set of independent $MCMC$

$\mathcal{A}_o$	set of 20 independent Markov chains of a specific RWM algorithm
r	repetition number, with $r = 1, \ldots, 100$
$z^{(r)}$	replicated data for $\mathbf{z}$, with $r = 1, \ldots, 100$
$\mathcal{R}$	set of replicated data
$m(\mathbf{y})$	marginal likelihood
j	draw from proposal density, with $j = 1, \ldots, J$
J	number of draws from proposal density, usually $J = T$
$p(\theta_k \mid \mathbf{y}, \theta_{-k})$	full conditional density of the block posterior
$p(\theta_{\tilde{k}} \mid \mathbf{y}, \theta_{-\tilde{k}})$	reduced set of the block posterior
c	constant
s_D	scaling factor of dimension D
ν	index for parameter estimate already being sampled
R	whole history of draws for θ
$\{\theta^{(\nu)}\}_{\nu=1}^{R}$	whole history of parameter estimates already being sampled
$\tilde{\nu}$	index for parameter estimate already being sampled and accepted
$\tilde{R}$	history of accepted draws for θ
$\mathbb{C}_R$	at R updated covariance matrix in adaptive Metropolis
M	a fixed integer that denotes the *memory* parameter
$\mathbb{M}$	matrix resulting from $M \times d$ used for calculation of $\mathbb{C}_R$
$\tilde{M}$	centered matrix resulting from $\mathbb{M} - \mathbb{E}[\mathbb{M}]$
t_0	length of initial period before adaption process starts
U	update frequency for adaptive $MCMC$

Chapter 1

Introduction and motivating example

A major part of research data in the social sciences originates from survey interviews, and a large body of literature on survey methodology focusses on data quality issues in particular. Beyond concerns about non-response also the manner in which responses are reported or recorded are crucial aspects for data quality, and the immense literature on measurement errors in surveys is still growing. Survey data scaled continuously can only be measured to a limited precision or are discretized otherwise. That is, data are either coarsened at reporting or recording, or grouped before further processing. Participation in survey studies is inherently connected to various response styles in self-reported data dependent on the respondents' characteristics but also on the issue in question. This often leads to different patterns of coarsening. On the contrary, coarsening before processing is mostly related to aggregation or tabulation, see Hanisch (2005a, p. 39). Such coarsened or grouped data are linked with a loss of information on structure, but they are also of important distributional information, see Howes (1996). Standard statistical problems might become complicated then, see e.g. Gastwirth and Glauberman (1976), Cowell (2000), and Pace, Salvan, and Ventura (2004).

One special artifact of coarsening in reported continuous or discrete numeric data is called heaping. Heaping means that a certain proportion of values falls on particular values, whereas all other values are reported at a reasonably high level of accuracy. Founding on the smoothness assumption for such data, deviations from this structure in form of spikes or heaps occur. To be concrete, at certain points of the distribution abnormal concentration of responses are striking (Torelli & Trivellato, 1993). The term *heaping* appears first in Myers (1940) with respect to age reportings and Eisenhart (1947) explores effects of rounding or grouping – both being special cases of heaping – for different sample sizes.

In principle, all numeric variables are susceptible for heaping, such as frequencies, amounts, fractions, scale measurements, but also time-related data, like starting and ending of episodes, or duration of episodes. Typical examples are age (Camarda, Eilers, & Gampe, 2007; Heitjan & Rubin, 1990; Myers, 1940;

Stockwell & Wicks, 1974), body weight (Camarda et al., 2007; Groß & Rendtel, 2015; Kroh, 2004; Rowland, 1990), number of cigarettes consumed (Harris & Zhao, 2007; Wang & Heitjan, 2008; Wang, Shiffman, Griffith, & Heitjan, 2012) or time of quitting cigarette consumption (Bar & Lillard, 2012; Lillard, Bar, & Wang, 2008), unemployment duration (M. Baker, 1992; Kraus & Steiner, 1995; Torelli & Trivellato, 1989, 1993), or other duration data (Augustin & Wolff, 2004; el Messlaki, Kuijvenhoven, & Moerbeck, 2010; Hobson, 1976; van der Laan & Kuijvenhoven, 2011; J. Wolff & Augustin, 2000, 2003). Many more examples are given in J. M. Roberts and Brewer (2001, p. 887f.), Camarda et al. (2007, p. 386), or Holbrook et al. (2014, p. 592).

Heaping is typical for retrospective data known to suffer from several recall errors (Torelli & Trivellato, 1989, 1993), when the respondent is either uncertain about the true value or hesitates to report. This indisposition leads to coarseness in convenient units, whereby the precision strongly depends on the data range (Torelli & Trivellato, 1993, p. 189). The preference for some set of numbers is, to a large extent, due to the feature of the quantity of interest. Huttenlocher, Hedges, and Bradburn (1990, p. 212) identify prototypes which can be either conventional calendar prototypes $(7, 10, 14, 21, 30, 60)$ or conventional arithmetic prototypes (multiples of 5 or 10). Of course, heaping has to be clearly demarcated from true observations, events, seasonal fluctuations, and other measurement errors (cp. J. M. Roberts & Brewer, 2001; Torelli & Trivellato, 1993).

A highly topical issue that arises from heaping – and coarsened data in general – is that it immediately affects the measurement scale and implies a loss of information about the true values (Hanisch & Rendtel, 2002, p. 2), distorts the distribution (Wang & Heitjan, 2008) but also influences results and yields biased inferences, variance deterioration, and inadequate interpretations. Besides attenuation on the macro-level, by hiding real effects, or on the contrary, exhibiting relationships not present in real data (Bound, Brown, & Mathiowetz, 2001; Schweitzer & Severance-Lossin, 1996; Torelli & Trivellato, 1993), also the micro-level is affected. Concretely, differences between respondents as well as individual changes over time (income mobility) can be obscured by the heaped values (Bound et al., 2001; Hanisch, 2003, 2006; Hanisch & Rendtel, 2002). An example for a macro-level effect is given in Schweitzer and Severance-Lossin (1996, p. 19). The authors find that subtle movements of the median from year to year could cause a larger shift than expected. On the opposite, subtle but meaningful changes in the true distribution might be eclipsed, because the point estimate of the median is still located at the mass point. Accordant effects can be excessive in particular when heaping does not occur at random. The prevalence and the pattern of heaping as well as the distribution structure of the data determine the performance of estimators, see Torelli and Trivellato (1993, p. 201).

A bunch of literature is dedicated to the problem of heaping and related problems, such as rounding or grouping of data. One part of the literature focusses on their evidences, with description of determinants and panel conditioning, see

e.g. Hanisch and Rendtel (2002) or Serfling (2006) with regard to income data. Others give an illustration of the effects. For example, Sheppard (1898) explores effects on moments and proposes a correction factor. The appropriateness of the so called Sheppard's correction factor is largely discussed by Dempster and Rubin (1983) and T. Liu, Zhang, Hu, and Bai (2007). Studies regarding the effects on parameter estimates are presented in Tricker (1992, 1995), J. Wolff and Augustin (2000) and Augustin and Wolff (2004). Effects on quantiles are given, e.g. in Schweitzer and Severance-Lossin (1996), Hanisch (2005a, 2006) or Drechsler and Kiesl (2012, 2014). Effects on measures of income inequality (e.g. Gini coefficient) are presented by Gastwirth and Glauberman (1976), Rendtel, Nordberg, Jäntti, Hanisch, and Basic (2004) and Daniels (2008). For example, Hanisch (2003) studies effects on poverty measures (e.g. headcount ratio). Hall (1982), DiNardo, Fortin, and Lemieux (1996), Schweitzer and Severance-Lossin (1996) as well as Hanisch (2006, pp. 40-52) examined the impact of rounding on nonparametric density estimation. Furthermore, studies exist that explore in which way test statistics are affected. For example, Pearson, D'Agostino, and Bowman (1977) study the influence on tests of normality, as the Shapiro-Wilk test. Preece (1982) explores effects on two-sample t-tests, and Rydén and Alm (2010) effects on the two-way ANOVA. Tricker published many different papers concerning changes in the significance level and statistical power of certain test statistics (Chi-squared test, one sample t-test and two sample t-test, F-test), either for normal data (Tricker, 1990b), or for non-normal data (Tricker, 1984, 1990a). Panel data analysis is also affected by rounding or heaping, as demonstrated by Pudney (2008) and Wang et al. (2012).

Literature attempting at explanations of heaping can be found in cognitive and social psychology. Early attempts address the satisficing theory, which was in general described by Simon (1955) and adapted to the theory of statistical survey satisficing by Krosnick (1991). Respondents are assumed to stop screening for further response options as soon as a sufficient outcome is achieved. The theory of Rosch (1975) attributes to *cognitive reference points*. Typical reference points in the decimal system are multiples of 10, for example. According to the theory of cognitive reference points, respondents have a strong tendency just to remember the magnitude of a value expressed by some leading digits and forget about the rest. During the retrieval process those terminal digits not being remembered are replaced with zeros producing a heaped value this way. If values are completely unknown, and the respondent is requested to take a guess, often a highly coarsened random number is produced, see also Hanisch (2005a, p. 40).

Another large part of the literature either discusses measurements and derives tests for heaping (Hanisch & Rendtel, 2002; Serfling, 2006) and related concepts, e.g. digit preference (Beaman & Grenier, 1998; J. M. Roberts & Brewer, 2001), or is concerned with appropriate strategies to cope with heaping, i.a. smoothing or modeling techniques (e.g. Camarda et al., 2007; Groß & Rendtel, 2015; Heitjan & Rubin, 1990; van der Laan & Kuijvenhoven, 2011; Wang & Heitjan, 2008).

Items for income information are one of the most important data in survey studies for political decisions-makers and economists. Income data often exhibit a substantial amount of heaping when being self-reported. To be concrete, in the majority of studies concerning heaping in income data interest is less in income rounded at a low level, i.e. rounding to the next integer, but on surveyed income data discretized at higher levels, i.e. multiples that fall on hundreds or thousands. Strong evidences exist that heaping in income data is related to the income level itself and further determinants, such as personal characteristics (e.g. Hanisch, 2005a, 2006; Hanisch & Rendtel, 2002; Schräpler, 1999; Serfling, 2006). Further exploration of the relationships between heaping behavior and personal as well as context information is valuable for a better understanding of the driving forces behind heaping. In summary, it can be stated that there is a necessity of addressing this issue adequately, e.g. by a modeling strategy which explicitly takes the non-randomness of the pattern into consideration.

Organization of this dissertation

This study has three main contributions to the existing literature. First, it provides descriptive evidence for heaping in the income data of the German National Educational Panel Study (NEPS). In particular, it is to be shown that heaping in survey data is not occurring at random. For this purpose, associations between heaping behavior and the true values as well as common socio-economic characteristics are explored. All findings speak against randomness with regard to certain predictors or the response value itself. The second research question deals with the introduction of a heaping model which is more general than other models proposed in the existing literature with respect to the distributional assumptions. In detail, a mixture model is established enabling to account for different heaping behaviors prevalent in self-reported income data. The proposed method assumes parametric models for the latent true distribution of the variable of interest and the heaping behavior. In doing so, the parameters of this mixture model can be estimated simultaneously. The generality of the proposed model is outlined by several modifications and extensions. The third main research objective is the comparative analysis of different estimation procedures. This work contains a concise comparison between a frequentist approach and Bayesian methods. Though, the complexity of the proposed model represents a very good opportunity to prove the efficiency of differing random-walk Metropolis (RWM) algorithms. Besides the original RWM scheme, the blocking strategy for sampling components of the proposal density and adaptive schemes with regular updates of the proposal covariance matrix are employed. In particular, different multiple-block schemes (Chib & Greenberg, 1995), the randomized-blocking strategy (Chib & Ramamurthy, 2010), the Adaptive Proposal algorithm (Haario, Saksman, & Tamminen, 1999), and the Adaptive Metropolis-Hastings algorithm (Haario, Saksman, & Tamminen, 2001) are compared to each other. To the best knowledge of the author of this thesis

no such comparison exists so far for mixture models[1] and work on adaptive or blocking Metropolis-Hastings (MH) algorithms is sparse to date.

The remainder of this doctoral thesis is organized as follows. Before addressing the three main research objectives listed above, this introduction proceeds with a clarification of central terms, their definitions and demarcations with respect to other related concepts (i.a. rounding and digit preference), and provides a theoretical framework on how to classify heaping and its corresponding structures by presenting some literature with distinct explanatory approaches. After that, the focus is on heaping behavior in income data in particular. The first part of the introduction is complemented by a brief summary of common literature on measurements and tests for rounding and heaping as well as differing strategies to cope with heaping and its consequences. The second introductory part of Chapter 1 explores the occurrence and relations of heaping in survey data of the German National Educational Panel Study (NEPS), with special focus on net individual income data. The findings from the NEPS income data are compared to findings from existing literature and supply further evidence for the fact that heaping does not occur at random. Heaping behavior largely depends on the true response value and several internal factors, i.e. factors attributed to the respondent. This endogenous and exogenous dependency of heaping provides a broad justification for the modeling approach suggested in this thesis. The proceeding in this section is purely exploratory and not hypothesis-driven. Furthermore, this illustration does not attempt to be representative. Owing to both facts, no general conclusion can be drawn, also comparisons with other studies have to be treated with caution.

Chapter 2 contains a thorough description of the heaping model, a mixture model allowing for simultaneous estimation of all model parameters. The model consists of two parts. One part models the latent true distribution and the other part constitutes the model for the heaping behavior, both parts being parametric. As latent true distribution a two-component model is assumed – the zero-inflated log-normal distribution. The log-normal distribution is considered to model income owing to its simplicity and because covariates can be easily included. Since the log-normal distribution does not support zero (or negative) values, a second component is included which additionally models distinctive zero responses. In the first place, a piecewise constant model is introduced as heaping mechanism assuming equiprobable heaping probabilities within predefined intervals for a priori fixed heaping points. Separating the whole range of income values into smaller parts enables flexible modeling of the heaping probabilities. The established model and the structures found in the NEPS data build the foundation for the data generating processes (DGP). Simulations are used to elicit the feasibility and effectiveness of the model. The work continues with a frequentist estimation approach by *Maximum Likelihood* (ML) using the *Nelder-Mead algorithm*.

[1] At the end of 2015, Herbst and Schorfheide published a book on Bayesian estimation of dynamic stochastic general equilibrium (DSGE) models in which block and adaptive $MCMC$ algorithms are explicitly compared to each other, see Herbst and Schorfheide (2015).

ML estimation can be problematic in models with finite mixture distribution and multi-modal likelihoods. Because of that, a Bayesian framework was set up using the random-walk Metropolis (RWM) algorithm as one method out of the pool of $MCMC$ methods. The author of this thesis refers to the RWM algorithm since no established distribution was found for the joint conditional distribution of all model parameters considered. Different specifications of the original RWM algorithm are explored to find a reasonable set-up for the RWM algorithm. The convergence behavior of $MCMC$ methods strongly depends on the specification of the algorithm. In this respect, the initial values as well as the definition of the proposal density – strictly speaking the covariance matrix of the proposal density – determine the exploration of the parameter space. In order to ascertain the impact of the proposal dispersion more precisely, variations of the RWM algorithm are tested. Such variations are the blocking strategy on the one hand and the adaptive $MCMC$ on the other hand. When considering blocking, the parameters of the model are summarized into clear-cut blocks arising either from the clustered structure of the model parameters – with respect to intervals or modulos – or the parameters are randomly assigned to blocks of varying lengths. In the updating schemes, the algorithm learns from the history of sampled draws and adapts the covariance matrix of the proposal density. Both methods attempt to yield a better mixing behavior with lower autocorrelations between consecutive iterations. The algorithms show a distinct behavior with respect to convergence and efficiency measures. The blocking schemes clearly outperform the adaptive schemes exhibiting high stability in estimation and very fast convergence.

In Chapter 3, the proposed heaping model is modified and extended in various ways illustrating the generality of the model. Suggestions for modification are (i) assuming the Dagum distribution as another latent true distribution, (ii) assuming wider heaping intervals to allow for more values being heaped, (iii) assuming asymmetric intervals due to underreporting of income data, (iv) assuming an alternative heaping mechanism that models higher probabilities for values in the proximity of a heaping point, and (v) modeling the heaping mechanism with less parameters by assuming constance of heaping probabilities in broader parts of the income range. Two extensions of the model are given afterwards by integrating internal factors, attributing to characteristics of the respondent, as covariates into the model. First, the covariates are used to determine the income level solely. The dependency of the heaping probabilities on the level of the true value (endogeneity) is already considered the heaping mechanism so far, but in the second extension personal characteristics are included as covariates to model individually different heaping behaviors (exogeneity). The performance of all models is fairly good, however, large differences exist with respect to runtime and efficiency.

In Chapter 4, the heaping model with its basic and all modified or extended versions is applied to the net individual income data of the National Educational Panel Study (NEPS). Posterior predictive checks and marginal likelihood estimates serve for comparative purposes. The overall fit of the models to the real

data is fairly good. Individual aspects of the distribution are also captured sufficiently. The results point to superiority of RMB-RWM estimation opposed to ML estimation.

An extension to external factors attributing to the interview situation is presumed in the conclusion. Limitations of this work as well as strategies for further research, either on the modeling strategy, or the Bayesian estimation technique, are also laid out in Chapter 5.

1.1 Heaping as measurement error

To put some structure in the discussion on where to place heaping in the series of survey errors, some recourses on survey methodology are necessary, in particular on quality issues. The whole survey process is prone to errors, beginning with the development of a research idea, verbally expressed as construct, up to a quantity of interest in form of a survey statistic. *Error* in this case means the deviation of what is desired from what is attained, and *measurement error* or *errors of observation* ($\varepsilon_i = z_i - y_i$) refer to deviations from responses given to a survey question and the true response value. Figure 1.1 is borrowed from Groves et al. (2004, pp. 48ff.) and depicts the errors typically found in survey data. Let Y_i denote the value of a construct, e.g. the true income for the i-th individual of the population ($i = 1, \ldots, N$), and y_i is the value of the measurement for the i-th sampled person ($i = 1, \ldots, n$). Although attempted to measure Y_i, the researcher is content with the imperfect indicator y_i. The difference between construct and measurement ($y_i = Y_i + \epsilon_i$) denotes the individual deviation from the true value. The response value evoked by application of the measurement is denoted by z_i. After all editing and processing steps one finally gets the edited response z_{ip} (ibid.).

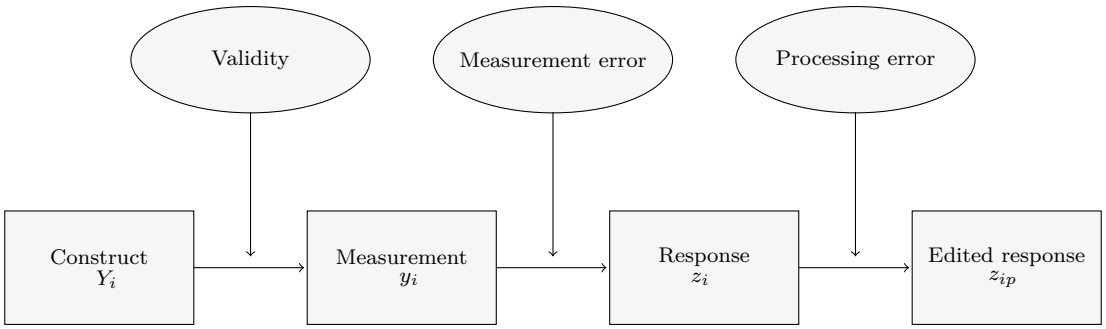


Figure 1.1: Errors in survey development, Source: Groves et al. (2004, p. 48).

Heaping is usually treated as measurement error (see i.a. Hanisch & Rendtel, 2002; Serfling, 2006; Torelli & Trivellato, 1993; Vardeman & Lee, 2003) along with the biases due to social desirability, acquiescence, central tendency, need for

social approval, and many more. The quantities y and ε are typically unknown. However, z is informative for the range of possible values for ε, see Hanisch and Rendtel (2002, p. 2). Besides this general categorization, it should be emphasized that heaping is also affected indirectly by validity and processing errors. Hence, it should be regarded as a result of the different survey errors intermingling.

Validity refers to the translation process when formulating a certain item (or more) as a representative of the theoretical construct. An item is ought to be valid if it actually measures for what it is intended. Several choices by the researcher concerning the instruments' design are important in this respect, see Daniels (2008, p. 2). For example, the order of questions and question wording play major roles, see Krosnick (1991, p. 213). Specific question wording can force certain responses and even slight variations can change the scope of interpretation and thus, the distribution of the outcome as well as its validity (ibid.). For example, the response to the income query might largely depend on the requested accuracy, if announced at all. The reporting period can further blur the reported data, i.e. annual reports are more prone to heaping, see Krosnick (1991, p. 221). Burton and Blair (1991, p. 77) discuss about the response formulation process and state that respondents do not always intuitively choose an optimal process. Exactly at this point, question formulation can help respondents to select a process that yields the desired outcome. For example, Becker and Diop-Sidibé (2003) find that a calendar-based query of events reduces heaping in duration data, and Huinink et al. (2011) show that a combination of Dependent Interviewing and a graphic event history calendar (DI-EHC) significantly reduces cognitive burden of the respondent when remembering life course data.

Finally, processing errors might result from range or consistency checks, outlier detection, or data aggregation, and can lead to biased outcomes. It should be further explored whether processing can exhibit a spillover effect on the response formulation process. That means, respondents might think that a given precise value will be aggregated or coarsened in some form anyway. Especially in connection with the absence of an instruction concerning the required accuracy of the response, the respondent might tend to an anticipatory coarsening or aggregation of the true precise value. This to prove is not within the scope of this thesis, but is left for further research.

Another classification of survey errors is provided by Henderson and Jarrett (2003). The authors differentiate between three categories: 1) *measurement error*, 2) *misreporting error*, and 3) *misclassification error*. The first category refers to an error in continuous data where the true value is erroneously reported as a more or less accurate value. The second error term concerns situations in which the true continuous value is reported as a discrete value, and the third error type results from reporting a true discrete value as another wrong discrete value. Following this scheme, heaping can be categorized as misreporting error, which represents a more distinctive description than the previous one. Moore, Stinson, and Welniak (2000) distinguish two further facets of misreporting errors: bias and

random error. Both facets – the systematic as well as the unsystematic one – are operating independently (id., p. 4).

1.1.1 Definitions and theoretical framework

Several terms exist to describe more or less the same phenomenon or overlap in meaning to a large part. This variety can be mainly explained by the respective discipline from which the focus is placed on. All terms have in common that they paraphrase surveyed data which are to some extent incomplete, i.e. only partial information about the true but unobservable data is available (Heitjan, 1989, p. 164f.). Though, the extent of inaccuracy can be highly variable. The differing extent of imprecision, the pattern of the outcome, and furthermore the underlying behavior that drives to the particular outcome can be utilized to distinguish these terms from each other.

The most general term in this respective is *coarse data* which implies rounded, grouped, interval, censored, but also aggregate data. Of grouped information – or aggregated data in general – individual data might be unavailable, because it is summarized into a small number of (equally-sized) groups prior to data provision, cp. Sheppard (1898), Heitjan (1989), Schneeweiß, Komlos, and Ahmad (2006). The reason herefor is often to preserve confidentiality. Another aim of supplying aggregate data can be to provide data that are easier to handle, e.g. by simple frequency tables, cp. Schneeweiß et al. (2006, p. 2). Dealing with aggregate data is straightforward, since the pattern that produced the outcome is known by the analyst, e.g. the procedure for aggregation, the intervals of the true values, and the time of censoring (for right censored data). A wide range of literature exists on how to cope with coarse data, e.g. by means of the coarsened data model according to Heitjan and Rubin (1991, p. 25f.). Moreover, Heitjan (1989, 1994, 1997) and Heitjan and Rubin (1990, 1991) established the foundations for inference from coarse data. This work was continued by Heeringa (1996) to further include point estimates – values being reported with accuracy – alongside the coarsened data. Daniels (2008) introduces, next to point estimates, also interval estimates and additionally missing data (uninformative intervals) into the model, and further implements a test for ignorability. He shows that the ignorability assumption does not hold in most cases owing to the structure of survey data. That is, interval data are often "not coarsened at random" (NCAR), see Heitjan and Rubin (1991) and Gill, van der Laan, and Robins (1997). J. Zhang and Heitjan (2006) propose an index of local sensitivity to nonignorability of the coarsening process by referring to the general coarse-data model of Heitjan and Rubin (1991). The index quantifies the extend to which inference changes and whether the coarsening can be ignored in analyses. If coarsening is ignorable one can revert to standard analysis, otherwise a nonignorable model of the coarsening process should be estimated. In a subsequent study by J. Zhang and Heitjan (2007), the index is adapted to Bayesian inference. Both studies show that the sensitivity to nonignorability increases as the percentage of coarsening increases.

A further general term besides coarsening is *heaping*. According to Torelli and Trivellato (1993, p. 188) and J. M. Roberts and Brewer (2001), heaping denotes an abnormal concentration of responses at certain values, durations, or dates. In this context, abnormality is relative to external validation data or prior expectations about the smoothness of the frequency distribution. Heaping is typically found in numeric data, either continuous or discrete. Rounding and heaping are often used synonymously as illustrated by the term *round-off error*, even though heaping denotes the more general case. When facing heaping, the points to which are heaped can be systematically different from typical rounded values. Not all points being spikes after rounding might also be preferred values for heaping. Likewise, a subset of preferred values might exhibit more probability mass than others. Furthermore, the heaping intervals can be of different widths or asymmetric. The propensity to heap, the preferred values, and the heaping intervals strongly depend on the object in question as well as on the specific range of values, see Torelli and Trivellato (1993, p. 189). Hence, these three figures can vary intra- and inter-individually, see Hanisch (2006). Respondents might relate to different interval widths yielding distinct observed values for the same true value.

Rounding refers to a special case of heaping and is most often reserved for continuous numeric data, see J. M. Roberts and Brewer (2001), or Wilrich (2005). Other disciplines, e.g. engineers, call this error quantization or digital resolution, see Vardeman and Lee (2003). Rounding means that quantities are usually measured or reported at a finite precision but does not only apply to decimal digits. Also precomma digits can be affected in that several final digits are replaced with zero, cp. Hanisch and Rendtel (2002, p. 2). Rounding represents some special kind of heaping since the intervals for rounding can be assumed to be symmetric, of equal width for the whole range of values, and the mechanism behind applies to all respondents. From this derives that rounding intervals do not overlap. Hanisch (2006, p. 27) distinguishes rounding, as a mathematical response type, from heaping, as an artificial response type. Overall, the error or degree of rounding is completely known and fix for all observations and can be corrected straightforwardly. Approaches for handling rounded data are given, e.g. in Qian (1996), Wright and Bray (2003), or Schneeweiß et al. (2006).

Another special kind of heaping is *digit preference* (Heitjan & Rubin, 1991), which is often synonymously called *number preference* (Beaman & Grenier, 1998), and is typical for discrete numeric or count data. In digit preference, the spikes correspond to values with terminal digit of a limited set, see J. M. Roberts and Brewer (2001, p. 888). The most common terminal digits are 0 and 5 (cp. Spoorenberg & Dutreuilh, 2007), but it also refers to preferred digits being multipliers of time units (e.g. 7, 14, 21), when respondents are asked for a specific duration at a given time unit, for example. First, the respondents compute the duration with respect to comfortable time units. Second, they multiply this rate by the number of time units occurring in the question's recall period (ibid.), see also Pickering (1992).

Sometimes the *birthday effect* is put into connection with heaping. The co-occurrence of death and birth on the same day is explained by Phillips and Feldman (1973) as the willingness of individuals to postpone their deaths, either to experience the own birthday or to participate in public ceremonies. The authors argue that the stronger an individual's feeling of affiliation and appreciation is, the greater is the motivation to participate. Abel and Kruger (2006-2007, p. 64) raise the concern that the birthday effect might be more likely to be due to *death heaping*. Data recording agencies often explicitly recommend to note 1 or 15 for unknown day of birth as well as day of death. This leads automatically to the above-mentioned co-occurrence. Strictly speaking, this phenomenon is artificial but does not originate from surveyed data.

Regarding the different constructs to describe heaping, Figure 1.2 systemizes all central terms used throughout this thesis. At the root stands the *heaping behavior* comprising the heaping pattern and the heaping mechanism. The first concept encompasses the heaping points and the associated intervals, and the latter refers to the assumed heaping function and the accordant heaping probabilities.

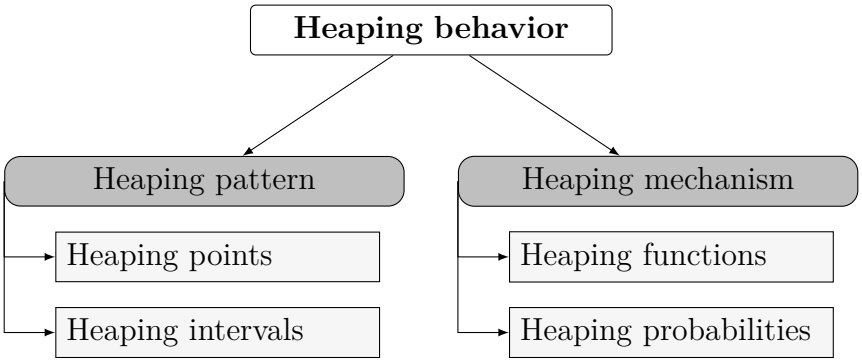


Figure 1.2: Systematization of central terms for the description of heaping.

The *heaping pattern* describes the structures on which the heaping mechanism bases and which are fixed in advance.[2] On the one hand, these structures are *heaping points* – also called spikes or heaps – which mark an abnormal concentration of frequencies in a continuous distribution that would be smooth otherwise, see Torelli and Trivellato (1993). Heaping points are often culturally defined, see Huttenlocher et al. (1990), e.g. by the common European metric system with typical multiples of 5, 10, 50, 100, 500, 1000, or the imperial and US customary systems of measurement on the contrary. Furthermore, the quantity of interest might evoke differing prototypes. While space dimensions or reports of frequencies are assumed to follow the metric decimal system, timescales are usually reported

[2]With regard to the definitions of the heaping pattern and function the author of this thesis does not follow Kraus and Steiner (1995).

as multiples of the duodecimal system (2, 3, 6, 12) for months, or the sexagesimal system (1, 5, 10, 15, 30, 60) for minutes, see Groß and Rendtel (2015, p. 4). The preference of certain values strongly depends on the scale of the variable and the magnitude of the response value. Whereas count data refer to small scales (often exhibiting digit preference), in variables with a wide range of possible values mostly multiples of 50, 100, 500, and 1000 occur owing to scale dependency. Furthermore, higher response values of at least five digits are more prone to fall on multiples of 1000 than on multiples of 10 because of level dependency, see e.g. Torelli and Trivellato (1993).

On the other hand, the *heaping intervals* need to be specified before estimation. The heaping interval defines the range of values contributing to a specific heaping point – also called *catchment area*. The true value y_i lies with full confidence (per assumption) in the interval $P(z_i - \xi \leq y_i < z_i + \xi) = 1$, with ξ specifying the parts of the interval below and above the heaping point. The heaping intervals can vary in their width and symmetry, but the width of the heaping intervals is largely determined by the multiple it covers. By assumption, and as far as not stated otherwise, ξ is considered as half of the respective multiple. For example, the heaping point 2000 falls on a multiple of 1000, thus $\xi = 500$. However, the intervals can be assumed to be even wider, e.g. ξ is equal to the multiple, or to be asymmetric, with one third of the respective multiple below the heaping point and two third above. The object in question can determine both, the width as well as the symmetry of the heaping interval. By definition, the heaping intervals for different heaping points may overlap (J. M. Roberts & Brewer, 2001). There are various items with high potential for coarsening and/or underreporting due to sensitivity of the issue or uncertainty about the topic. To sum up, the heaping pattern is expected to vary according to the object of interest and the magnitude of the response value (scale and level dependency).

The *heaping mechanism* describes the process governing the heaping function parameterized by heaping probabilities, for example. Here, the *heaping probabilities* are defined as the respondents' propensity to heap the true value y_i to a possible heaping point. The distribution of the heaping probabilities is expressed in form of a *heaping function*. Typical functional forms are simple piecewise constant models or more complex functions like a truncated bell-shaped function. Thus, the heaping probabilities are either considered to be equiprobably distributed within the heaping interval, or the heaping probabilities are assumed to increase with proximity to a particular heaping point. Throughout this thesis, the respective heaping function assumed applies to all heaped responses in a given setting. The heaping probabilities are also expected to show level dependency, see J. M. Roberts and Brewer (2001) and Torelli and Trivellato (1993), but might be further influenced by various exogenous factors, e.g. individual characteristics as well as conditions of the interview situation.

When referring to the general heaping model, the heaping behavior is meant together with the latent distribution assumed to model the true response values.

Attempts at explanation

Literature from cognitive and social psychology distinguishes four steps in the process of answering to survey questions, see Tourangeau, Rips, and Rasinski (2000), and cp. Figure 1.1 (from y_i to z_i). First, the respondent has to comprehend the question including interpretation of the question meaning. The second step refers to retrieval of the required information from memory. In the following judgement process, the response is evaluated while assessing accuracy and completeness of the eligible responses. Furthermore, the correspondence between the desired and retrieved information is assessed. All eligible answers are judged with respect to possible positive or negative sanctions, due to disclosure of privacy, social desirability and incentives, for example. In the final response step, the selected answer is edited and communicated. The editing can be attributed to the categories given in which the response has to be fit into.

Voluntary participation in surveys is crucial and affects the third and the fourth step in the answering process. Various attempts for explanation of this relationship exist. In the early literature, behavior in surveys is associated with the *economic man*, see Simon (1955). The respondent is assumed to act rational, to have full knowledge and control, and to optimize decisions. Due to internal as well as external constraints, e.g. limited ability, knowledge or time, substantial simplifications in decision making might be expected. In this concept of *bounded rationality*, respondents act as satisficers who straightaway watch out for a satisfactory solution instead of striving for an optimal one (Simon, 1955).

Two general terms are raised when respondents compensate for their limited resources. These are *satisficing*, widely used in economics and political science, and *heuristics*, a common concept in psychology and sociology. Rather than relying on an exaggerated rule of optimization, respondents might refer to heuristics or mental shortcuts when deliberation costs are high and other concurrent obligations are pending (Simon, 1955). *Heuristics* in general are techniques to speed up the process of problem solving, definitely not guaranteed to be optimal or even perfect but sufficient for the task ahead. Much of the research on heuristics is done by Tversky and Kahneman (1974). However, the concept was originally introduced by Simon (1955) as well as the term *satisficing*, which is a combination of the verbs "satisfy" and "suffice". Examples of such simple and efficient rules – also well theorized – are *anchoring* and *adjustment*, see Tversky and Kahneman (1974) and Hurd (1999), but also the *availability heuristic*, see Tversky and Kahneman (1974). *Satisficing* is a cognitive strategy in which a decision-maker screens through several eligible options but immediately stops searching when the first option is retrieved that meets the requirements, i.e. when sufficiency of the outcome is achieved. Foundation for this assumptions is that decision-makers are regarded as being incapable of evaluating all outcomes with comprehensive precision (Simon, 1955).

Krosnick (1991) adapts this theory and proposes a theory of statistical survey satisficing. Because optimal question answering involves high cognitive effort,

some people might tend to shortcut the same. Respondents might become fatigued and less motivated. Hence they are, at least to some extent, expected not to select the optimal but the first more or less acceptable answer coming into their mind, see also Tourangeau et al. (2000) and Schaeffer and Presser (2003). Satisficing for reduction of cognitive burden is found in two forms, either in a weak, or in a strong form (Krosnick, 1991, p. 215). In its weak form, respondents are assumed to execute all cognitive steps involved in optimizing but less elaborated. Hence, the answers are incomplete or inaccurate and possibly biased. In the strong form, respondents are assumed to offer seemingly reasonable responses but without employing any memory search or judgement. The overall probability of the satisficing strategy is related to the respondents' ability and motivation but also to the task difficulty. Typical artifacts resulting from satisficing are, e.g. selecting randomly from offered response options, choosing the "not applicable" or "don't know" option (strong satisficing), choosing socially desirable responses in sensitive questions, preferring middle and neutral answers (weak satisficing), but also skipping of items to abandon the survey more quickly, see Krosnick (1991) and Weisberg, Krosnick, and Bowen (1989, 1996).

During answering all survey questions, respondents who satisfice are likely to employ distinct patterns, dependent on the question, the respondents' capability to deal with the task given, the respondents' interest and enthusiasm, or a combination thereof, see Krosnick (1991, pp. 221ff.). When respondents take mental shortcuts this might entail suboptimal or even detrimental outcomes, e.g. being biased to some extent or lack information, instead of helping to increase the effectiveness or accuracy. Satisficing is typical in retro- and prospective questions, questions of estimation, in sum all questions requiring intuitive decisions under uncertainty (Holbrook et al., 2014).

The theory of satisficing was extensively explored by Holbrook et al. (2014) and to the largest extent refuted. Holbrook et al. (2014) analyze several studies to systematically examine the prevalence and predictors of heaping across a variety of question types. Strictly speaking, the authors checked for digit preference (values ending with 0 and 5). Satisficing is measured as shorter response latencies, less accuracy and lowered predictive validity. In their final judgement, Holbrook et al. (2014, p. 617) state that the processes leading to heaping are very different with respect to the type of question: objective constructs vs. subjective phenomena. Hence, the prevalence of heaping varies systematically across the question types. Burton and Blair (1991) explain this finding by the different processes respondents rely on when answering. If an estimation process is applied during response, as opposed to a counting process, frequencies of behaviors are more likely to be heaped. This in particular holds for high response values.

Other studies also point to the fact that heaping is not kind of a deliberate decision. Heaping behavior rather reflects the attempts of respondents to give the best estimate of the true value which they believe to be true, see Beaman and Grenier (1998). According to Huttenlocher et al. (1990), two factors impact

the reporting process and are assumed to fully account for the observed bias from inexact information. These factors are the imposition of an upper bound to constitute a reasonable answer on the one hand, and the overuse of prototypic values on the other hand (id., p. 208). Information is represented in terms of specific units which can have several possible forms of various sizes. If values are preserved at one level of detail all higher levels are preserved as well, and the stored information can be easily translated into. For example, when duration data are stored in weeks, the respondent can quickly answer in months. On the contrary, when the unit in question is more precise than the stored information, the latter one is translated into a heaped (prototypical) value (id., p. 198). For example, Huttenlocher et al. (1990, p. 212) quote conventional calendar prototypes (7, 10, 14, 21, 30, or 60) or conventional arithmetic prototypes (multiples of 5 or 10). Respondents refer to these prototypic values with increased inexactness of the information in memory. Thereby, two patterns are found in the use of prototypic values: the proportion of prototypic values relative to the remaining values and the size of the prototypic values increases with uncertainty (id., p. 199).

In this respect, heaping should be regarded as an option along the sequence of alternative respondent behaviors occurring during interviews. In the majority of survey studies, participation is voluntary. Therefore, respondents can behave in multiple ways. First, they are free to opt for participation. Second, they are free to decide whether to answer certain questions of the complete questionnaire. Third, they are free to decide on how to answer these questions, exactly or imprecisely. Of course, the last optionality also holds for mandatory surveys. Furthermore, in panel studies, the respondent is free to opt for panel consent. The following Figure 1.3 is borrowed and adapted from Schräpler (2002, p. 2) for illustration:

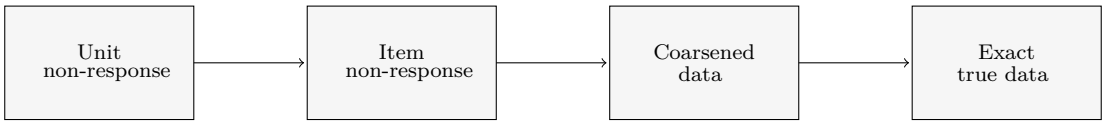


Figure 1.3: Alternative respondent behaviors occurring during interviews.

The order in Figure 1.3 originates from the information content being available after surveying. From the left to the right the information content decreases. In the worst case, the respondent completely refuses implicating total loss of information. At the best, all responses are reported truthfully and exactly. Along the sequence, two response behaviors are located entailing partial loss of information. The information loss is greater when the respondent refuses to answer a certain question than providing a coarsened value instead. These types of respondent behavior are not exhaustive and not distinct from each other. Once opted for participation, the respondent is still free to decide from question to question whether to answer and when, to which degree of accuracy. In line with this course, large

inter- as well as intra-individual differences exist emerging from varying motives, see Schräpler (2002, p. 3).

Inter-individual differences can be largely explained by basic personal characteristics, such as anxiety, need for social approval, or lack of motivation, but also with respect to the respondents' ability including comprehension and numeracy. Some respondents are a priori known to have greater knowledge than others. Hurd (1999, p. 114) demonstrates how biases vary with the level of respondents' uncertainty. Crayen and Baten (2008) use heaping as proxy for non-numeracy whereby numeracy composes of numerical ability and numerical discipline. The authors find significant age heaping at least until the turn of the 20th century. These improvements in age reports relate widely to the increased spread of education. Földvári, van Leeuwen, and van Leeuwen-Li (2012) find that age heaping is more spread among women than men. The authors attribute this to less numeracy or less human capital among women, as opposed to men.

Intra-individual differences can be attributed to the object in question. There are various items with high potential for coarsening due to sensitivity of the issue or uncertainty about the topic (J. M. Roberts & Brewer, 2001). An example for topic dependency is that a housewife might report information on household income heavily coarsened due to high uncertainty (Y. C. Zhang & Schwarz, 2012). On the contrary, she reports hours of child care precisely. The uncertainty about the requested household income can be caused by unawareness owing to item difficulty (complicated formulations), owing to lack of exposure to financial concerns (no self-earner), or owing to volatility (performance-based salaries, compensation according to the number of hours worked). The second object related factor is item sensitivity. Tourangeau et al. (2000) describe three aspects, each of them might cause response artifacts. These are invasion of privacy owing to inappropriate intrusive questions, risk of disclosure of answers to third parties (dissemination of data), and social desirability of the answers, i.e. expected social costs from admitting to deviate from a public or social standard. Item sensitivity is a strong motive for refusal and many studies have shown that response rates of sensitive items are in general lower, cp. Frick and Grabka (2007) and Krumpal (2013). Though, unwillingness to reveal sensitive information does not automatically mean that respondents completely skip the question. This reluctance or hesitation in reporting might express in form of coarsening as a moderate form of refusal, at least temporarily. Hanisch (2003, p. 16) demonstrates this close connection of heaped or coarsened data to item non-response in his transition overview. In his study, response types are regarded over consecutive waves. According to the author, the probability of switching between heaping and item non-response (and vice versa) is much more likely than the probability of switching between point estimates and interval estimates. Serfling (2006, p. 4) also point to the fact that heaping might be a precursor of subsequent non-response. Though, their hypothesis heaping is caused by a lack of motivation (id., p. 89) is to rigid and not sustainable as outlined later on by the author himself.

1.1.2 Heaping in income data

The focus of this thesis is on heaped income data, but before referring to the literature, it is crucial to distinguish the specific income types being in question first. Household income is of particular interest since it constitutes the *OECD equivalence income* which serves as a relevant measure to access relative wealth of all household members used to measure economic inequality. For this purpose, the total disposable household income is divided by the OECD equivalence weights. Additionally, earned income at the individual level is a preferred variable for labor market studies, see Rendtel et al. (2004, p. 4). Whereas register-based estimates can be used with clear conscience as proxies for true income (Rendtel et al., 2004, p. 6), information from survey data is most often prone to errors. However, register data are restricted by their availability. Only few countries dispose over register data in the required extent or provide full access to them, as e.g. Finland (cp. Hanisch & Rendtel, 2002). Quite the contrary, register data are seldom comprehensive. For example, register data from Germany concerning individual income are available only for employees (cp. Antoni, Vicari, & Bela, 2015).

An abundance of literature has emerged on evidences of heaping in income data in survey studies. Early evidences are given by Miller and Paley (1958) and Maynes (1968). Miller and Paley (1958) inquire into the accuracy of the reported 1950 census income data by a matching study of the post-enumeration survey sample with data from tax forms. According to the authors (id, p. 200), the variability of income reports is remarkable but at random. Albeit the authors do not find any systematic downward bias, Pechman comments that reports on tax returns often suffer from underreporting and can not be regarded as truth for validation (id., p. 204). Maynes (1968) documents response errors in financial data (assets and depts) and finds evidences on heaping as well as underreporting. As stated by Maynes, both effects are clearly visible, although the respondents were informed about the purpose of the study to test for accuracy of financial reportings. Both studies, Miller and Paley as well as Maynes, report an overrepresentation of respondents with higher incomes in the matching sample. Maynes (1968, p. 216) assumes that the respondents with low incomes might felt to be unimportant and therefore refuse respondence.

In the nineties, Schweitzer and Severance-Lossin (1996) and Schräpler (1999) provided evidences on heaping in gross income data. The prevalence of heaping in the Annual Demographic Supplements to the Current Population Survey (CPS) of March 1995 was 71% of all full-time earnings whereby the prevalence is estimated as the number of reported values being a multiple of 1000, see Schweitzer and Severance-Lossin (1996, p. 4). The authors find heaping to be highly systematic and correlated with the respondents' earnings level (id., p. 2). Schräpler (1999) investigates heaping behavior in the gross income data of respondents in the first 12 waves of the German Socio-Economic Panel (GSOEP). In his study, 67-77% of the reported income values fall on multiples of 100, 500, or 1000. He finds stability of heaping behavior over subsequent waves and many relationships to covariates

in a multinomial logit estimation. For example, heaping differs by gender, age, interview mode and length, and income level of the respondent. Concretely, he finds men to be less precise, and elderly respondents being more precise (ibid.).

Hanisch and Rendtel (2002), Rendtel et al. (2004) and Hanisch (2005a, 2006) inquire into several aspects on data quality within the CHINTEX project frame.[3] Among others, the authors explore heaping in the empirical distribution of different panel surveys, i.e. the German Socio-Economic Panel (SOEP) and the German as well as the Finnish subsample of the European Community Household Panel (ECHP). They provide descriptives on the prevalence of heaping but also comparisons of surveyed data with registered data for the Fin-ECHP as well as analyses on the transition of individual heaping behavior over consecutive panel waves. The authors focus on net household income but also on individual gross wage and earnings. In this respect, they find that heaping is less prevalent in individual data than in household data, see Hanisch and Rendtel (2002, p. 4). In the Fin-ECHP (1996, 2000) approx. 80% of gross wage and earnings and approx. 95% of net disposable household income are reported with one or two significant leading digits, see Hanisch (2005a, p. 43). Another major finding is the correlation of the intense of heaping to the level of earnings or income on the one hand, and systematic variations across different types of respondents on the other hand. The impact of related factors can be divided into personal and context-related factors. In the scope of context-related factors the authors regard the interview mode, interview duration, panel conditioning as well as the income type. Among the personal or household characteristics, age, gender, and job type are explored according to their relationship with heaping behavior. A measure for descriptives of the intensity of heaping is proposed – the Rounding Indicator. The possible influence of the factors is checked by an ordered probit analysis with the Rounding Indicator as dependent variable, see Hanisch (2005a, 2006). Selected findings of their analysis are: older respondents report with higher accuracy (but this was only a tendency). In the Fin-ECHP males reported more precisely than female respondents. This finding could not be confirmed in German and Luxembourgian data of the ECHP. Some general findings for the context-related factors are: higher accuracy of responses in (computer-assisted) personal interviews than in telephone interviews. Higher interview duration was correlated with higher precision of responses, strengthening the assumption of more deeper processing.

Serfling (2006) explores heaping in income data of the Swiss Household Panel (SHP), waves 1 to 5. In a stepwise augmented ordered probit model respondent, interviewer and respondent-interviewer-interaction effects as well as panel duration effects on a proposed relative measure for the rounding intensity are estimated. As stated by the author, gender, age, and health status are influential with respect to the intensity of heaping. Positive correlations are found for male gender, elderly people, low educational level, good health status, but also higher income values.

[3]Change from Input Harmonisation to Ex-post Harmonisation in National Samples of the European Community Household Panel (CHINTEX).

Against expectations, the authors found no continuous negative influence of panel duration on the rounding intensity. The assumption that respondent experience can positively affect respondents' willingness to cooperate is therefore neglected.

Drechsler and Kiesl (2012, 2014) and Drechsler, Kiesl, and Speidel (2015) study heaping in the German panel study "Labor Market and Social Security (PASS)" (2008/2009). About 62% of the observed values fall on multiples of 100, 500, or 1000. Relationships with external or internal factors are not explored, though being considered as covariates in the model for heaping.

Non-response and underreporting of income data

Questions regarding income are those with a high potential for item non-response and heaping due to sensitivity or uncertainty. For example, Frick and Grabka (2007, p. 5) report an overall item non-response in gross labor income in the SOEP (1992-2004) of 14%. About 8% of the data in the monthly net household income are missing. The German General Social Survey (ALLBUS) is also affected by high non-response rates of household net income. Between 21% and 26% of the values are missing in the surveys 1990, 2000, and 2006 (Krumpal, 2013).

Another typical pattern found in self-reported income data is *underreporting* or *downsizing*. This is the tendency for heaping downwards to the next multiple below the true value. Such an underreporting is also known for self-reported body weight (e.g. Kroh, 2004; Krul, Daanen, Hein A. M., & Choi, 2010; Qian, 1996; Rowland, 1990) and for cigarette consumption (e.g. Pérez-Stable, Marín, Marín, Brody, & Benowitz, 1990; Warner, 1978). In contrast, *upsizing* or *overreporting* is typically found in self-reported body height (Rowland, 1990). In general, people know about desirable attributes, e.g. avoiding overweight, non-smoking behavior, and have a strong propensity to match (more closely) to those ideals, of course, not in their real behavior, but in the perception by others (*social desirability*). Thus, people who do not fit to the norms "adjust" the real data to the idealiter expected ones. For example, Miller and Paley (1958, p. 204), David (1962), Greenberg, Moffitt, and Friedmann (1981), Rendtel et al. (2004) or Hurst, Li, and Pugsle (2014) examine underreporting of income to tax authorities but also in household surveys.[4] Moore et al. (2000, p. 4) find this general tendency toward underreporting especially for income sources whose magnitude is highly variable across different income types. For this, several causes can be driving, see Rendtel et al. (2004, p. 8). Respondents might ignore certain components of income if they are rarely received, e.g. special payments. People might forget about small incomes because they seem unnecessary or are less salient (Moore et al., 2000, p. 20), e.g. grants. Furthermore, when respondents are highly uncertain about the true

[4] A lot of nonscientific literature also covers underreporting of income data, see e.g. `http://money.howstuffworks.com/personal-finance/personal-income-taxes/tax-evasion1.htm`, `http://smallbusiness.chron.com/happens-dont-report-self-employment-income-16135.html`, or `http://www.investopedia.com/terms/u/underreporting.asp` and so on.

value they might tend to report a rather conservative estimate, see Rendtel et al. (2004, p. 8).

In a recent study, Antoni et al. (2015) inquired into the accuracy of reported individual gross income and underreporting of employees in the data of the adult cohort of the National Educational Panel Study (NEPS) by linking the survey data to administrative data of the Institute for Employment Research (IAB). With regard to personal and interviewer characteristics on accuracy, the authors find female respondents to report more precisely with lower deviation. Highly educated respondents show highest deviation but report more precisely on average. Among the interviewer characteristics, the experience exhibits a weak effect on accuracy, but no hints are given on any significant influences of the interviewers' gender. Clear evidence on underreporting emerges from comparing the median of registered and self-reported income data which reveals an underestimation of about 200 EUR on average.

1.1.3 Diagnostic tools for rounding and heaping

This section presents common measures and tests for digit preference and heaping, according to the differences between expected and observed frequencies of certain values. The first step in the detection of particular response patterns simply is to inspect the frequency distributions – either tabulated or plotted –, but to quantify the prevalence and extent of heaping, measures are highly desirable.

Myers' blended index (1940)

Myers' blended index (MBI) focusses on the terminal digit, concretely, the last digit of a reported or measured value. The MBI is calculated by summing up the absolute deviations of the observed percentages from the expected 10% when being reported randomly for each digit. The index relies on the assumption of a uniform distribution of all possible terminal digits. Since it regards all terminal digits together, the MBI constitutes an overall accuracy score, see Myers (1940). It is typically used for age data in demographics.

The MBI is not able to effectively measure heaping on multiples of any number aside from 2, 5 or 10, see J. M. Roberts and Brewer (2001, p. 888), which constitutes a major weakness and downgrades the MBI to a measure for digit preference only and not a general measurement of heaping.

The Whipple's index (Shryock and Siegel, 1976)

The most widely applied measure for detection of the preference or avoidance of particular terminal digits is the Whipple's index (WI). As the MBI, the WI is usually applied to measure the quality of age reports. The WI is conceptually similar to the MBI. However, it considers only the preference of a) 0 and 5, or only b) 0. In case of a), all values ending with digits 0 and 5 are counted and

divided by the sum of all frequencies times 1/5. If the WI is calculated for case b) the factor is 1/10. The basis for this calculation is the rectangular distribution assumption for all terminal digits, see Shryock and Siegel (1976, pp. 115-119) as cited by Spoorenberg and Dutreuilh (2007, p. 735). Disadvantageously, opposite effects of digit preference – avoidance of particular digits – can potentially affect the WI. Because the WI is a ratio of two sums, positive and negative deviations can cancel each other out, see Spoorenberg and Dutreuilh (2007, p. 730).

Spoorenberg and Dutreuilh (2007) introduce an extended version of the WI – the total modified WI – which considers all terminal digits. This results in an overall summary index. The total modified WI is the sum of the absolute differences between the digit-specific WI and 1. As a normalized overall accuracy measure, it eases comparisons through time and across countries (id., pp. 730ff.). Comparisons of the original WI and the modified WI reveal that the original WI clearly underestimates improvements in reporting quality (id., p. 736).

Benford's Law (Newcomb, 1881; Benford, 1938)

Benford's Law – or correctly *Newcomb-Benford's Law* – inspects leading digits according to their deviation from natural frequencies. It was perceived for the first time in science after Frank Benford publicized an article about it (Benford, 1938), but actually it can be attributed to Simon Newcomb who already quote on this phenomenological law in the late nineteenth century (Newcomb, 1881). A leading digit is the first digit unequal to zero or the first "significant" digit in a real number, e.g. 1 in 1256, and 2 in 0.0289. Usually, the frequency of all digits in any sequence is expected to be equally distributed. However, by investigating logarithmic tables, Benford finds smaller digits to be more frequent compared to higher digits. These natural frequencies could be detected elsewhere, e.g. the distribution of digits in the Bible or on car licence plates, see Humenberger (2008).

Of course, this is not a measure for the quality of reported data, but it is useful for detection of artificial – maybe fraudulent – scientific data, see Diekmann (2007).

Severity measure for rounding by Pace et al. (2004)

Pace et al. (2004, p. 39) introduce the severity measure in the scope of robustness of classical likelihood procedures for testing with respect to rounding. Referring to Tricker (1984, 1990a, 1990b, 1992, 1995), the authors point to the fact that the number of classes after rounding strongly determines the effect size. Here, number of classes simply means the number of cells of a frequency table. The severity measure is defined as the ratio between the length of the rounding interval and the standard deviation (SD). A decreasing number of classes leads to an increased severity measure. However, this measure is only applicable to data where all values are rounded to the same multiple, e.g. to hundreds or thousands.

Tests and measures for digit preference by Beaman and Grenier (1998)

A Chi-squared goodness-of-fit test which compares heaping points to the expected values of a smooth function is introduced by Beaman and Grenier (1998). As smooth function, a linear spline to adjacent values is employed. The knots of the spline are determined as the average of two values below and two values above the heaping point (id., p. 46). Disadvantageously, such a spline can produce poor estimates, in particular for cases where the adjacent values are zero. The authors alleviate this problem by adjusting the likelihood accordingly. Despite this fact, any heaping point surrounded by zeros and exhibiting a frequency of 5 and more will be significant at the 1%-level in this test (id., p. 49). In this test, the set of heaping points has to be specified explicitly. A multinomial distribution is considered for the frequency function of all spikes, though, other distributional assumptions as the binomial or Poisson distribution may be applied as well. By concept, the test also allows for testing single spikes and those not ending with 0 and 5. Beaman and Grenier (1998, p. 48) further introduce a measure for the influence on the mean as well as an estimate for the overall proportion of heaped values, which can be used for comparison of different data sets.

The major critique point of this test is that it is quite sensitive and often becomes hastily significant, especially when comparing heaping points in a long-tailed distribution where almost all values in the higher ranges do not have adjacent values above 0. Then it might be helpful to consider the k-th neighborhood, with $k > 10$ for example, as proposed i.a. by J. M. Roberts and Brewer (2001).

Formal measurement and test of heaping by Roberts and Brewer (2001)

J. M. Roberts and Brewer (2001) present a formal measurement for heaping that enables comparison of different data sets through time and across countries. Preliminary, a set of hypothetical values has to be identified as heaping points. This can simply be done by inspecting the data. Alternatively, the set of values can be derived from previous studies. Then, for the remaining values, it is determined whether or not the response is aberrant compared to the neighboring response values.[5] Regarded is either the difference between the response and the average response of the (two) adjacent neighbors, or if the response represents a local mode with respect to the neighbors considered. The test power is influenced by sample size and the number of values hypothesized as heaping points but also by the sampling variability (id., pp. 891ff.). Possible extensions raised by the authors are: the consideration of more than two adjacent neighbors or adaption to "reverse heaping", i.e. the avoidance of particular digits (id., p. 894).

[5] In the concrete implementation, the tails of the given distribution are truncated to avoid a large number of responses without any nearest neighbor.

Rounding Indicator (Hanisch, 2002)

For measuring the incidence and intensity of heaping, Hanisch and Rendtel (2002) introduce the Rounding Indicator (RI). The RI is a discrete measure of the precision of a given income statement. It quantifies the number of significant leading digits followed by zeros. Concretely, the RI is determined by the function $b(z)$ of the observed and possibly rounded value z, see Hanisch (2005a). The index can be calculated immediately for the data at hand, e.g. $b(5) = b(400) = b(6000) = 1$, $b(23) = b(180) = b(5100) = 2$, $b(169) = b(1540) = b(43200) = 3$, $b(3497) = b(40150) = b(614700) = 4$, and so on. As asserted by Hanisch (2005a, p. 41), the RI is an indicator that is better suited for comparisons (e.g. across countries or time periods) than the approach of simply counting multiples of integer values. For this purpose, the total frequencies of $b = 1, 2, 3, 4$ are counted and compared.

The author extends the RI in the accompanying presentation to his paper, see Hanisch (2005b). Here, he further considers the length of the observed value, i.e. the number of all digits of which a particular value consists. To get the relative measure, b is divided by the length of the value. The relative RI is a much more meaningful measurement than its predecessor as it corrects for the level of the response and its inherent ordinality. This facilitates a refined interpretation as the intensity can differ for the same RI dependent on the length (e.g. for $b = 3$ then $b/4 = 0.75$ and $b/5 = 0.6$). See also the doctoral thesis of Hanisch (2006, p. 28) for more details and examples of the RI in different survey studies.

Rounding Intensity (Serfling, 2006)

Serfling proposes the latent optimal rounding intensity (LRI) in his dissertation to characterize the unobservable heaping behavior. This quantity is based on a random utility maximization hypothesis. The underlying utility function can be regarded as a function of costs and benefits. The costs and benefits can be attributed to personal characteristics of the respondent and the interviewer but also to factors of the interview situation, see Serfling (2006, p. 87). In line with the LRI, several rounding measures are explored. Only those are assumed to be appropriate which are positively correlated to the LRI, thus full-filling the ordinality assumption. Furthermore, his focus is on relative measures which also consider the total number of digits because larger figures are found to be more severely heaped (ibid.). First, the percentage of rounding error is calculated as a simple measure. The width of the rounding interval is divided by the observed value. This measure is appropriate to detect the relative disturbance owing to heaping. Second, the Rounding Quotient (RQ) is introduced. The RQ is simply the ratio of the number of zero digits at the rear of an observed value – those following significant or non-zero digits – to the total number of digits minus 1. For example, the RQ of 4350 EUR is $1/(4-1) = 1/3$. The RQ ranges between $[0, 1]$, but some of the outcomes are standing alone due to unique combinations of the ratio. That is, a RQ of $1/6$ is only observed for 7-digit values rounded at tens.

The RQ is always zero for values reported accurately independent of the length of the value. This artifact induces the problem of indistinguishability (id., p. 88). Therefore, a third relative measure is considered, the Rounding Strain Measure (RSM). The RSM is calculated as difference between the number of zero digits at the rear and the significant leading digits minus 1. The outcomes can range from -6 to 6 for values up to 7 digits. The accordant RSM of the hypothetical values 4350 EUR is $1 - (3 - 1) = -1$. The RSM outcomes are aggregated to range between $[1, 5]$ for a better handling and to avoid the small cells problem. To be concrete, the values -6 to -4 are summarized into category 1, -3 and -2 form category 2, -1 to 1 form the third category, and categories 4 and 5 are built by the values 2 to 6, respectively (id., p. 89). Serfling (2006, p. 101) proves the ordinality of the RSM outcome and finds the observed categories of the RSM to be monotonically connected to the LRI. The outcomes of the measurements are usually reported in percentages.

Identification of heaping points (Zinn and Würbach, 2015)

For modeling heaping behavior according to the approach presented later on in Chapter 2, the set of heaping points (HP) must be fixed in advance. This can be achieved by either defining them ex ante, or by identifying them from the respective data. Zinn and Würbach (2015, Section 2.2) employ a heuristical procedure to identify HP in a reliable way from given income data of the Adult Cohort of the German National Educational Panel Study (NEPS SC6), see Section 1.2 in this thesis for more information concerning the data. Basically, a set of hypothetical HP is defined and checked for being real HP with respect to the given real data.

Let the vector $\mathbf{z} = (z_1, \ldots, z_n)$ comprise the reported values for the variable of interest. It is $z_i \in \mathbb{R}_0^+$ with $i = 1, \ldots, n$ for n given observations. The vector $\mathbf{y} = (y_1, \ldots, y_n)$ denotes the true values corresponding to $\mathbf{z}$, with $y_i \in \mathbb{R}_0^+, i = 1, \ldots, n$. Note that $\mathbf{y}$ is not directly observed. At first, the empirical cumulative distribution function (*ecdf*) $\widehat{F}(z)$ is estimated from the observed, and possibly heaped, data $\mathbf{z}$. Then, $\widehat{F}(z)$ is compared with the cumulative distribution function of a hypothetical income distribution $F^h(y)$ that roughly resembles the degree of smoothness of the real (unobserved) cumulative income function $F(y)$. In the present case, $F^h(y)$ is assumed to follow a log-normal distribution. For other issues than income data, different distributions can be assumed, of course. The accordant parameters of the log-normal distribution are derived from fitting the self-reported income values $\mathbf{z}$ without zero values to the log-normal distribution. $F^h(y)$ exhibits then the intended shape and degree of smoothness of the unknown $F(y)$. A sample of size n with hypothetical income values $i^h = (i_1^h, \ldots, i_n^h)$ is simulated and used for computation of the *ecdf* $\widehat{F}^e(y)$.

Based on this, a hypothetical HP h_b^0 is considered as real HP if the increment of $\widehat{F}(h_b^0)$ from the observed data exceeds the median of all increments of $\widehat{F}^e(y)$ from the simulated data as well as the corresponding increment of $\widehat{F}^e(h_b^0)$. In Figure 1.4,

the processing of the heuristic is illustrated using the NEPS SC6 income data (id., p. 5). Please note that this heuristic also allows identifying HP that are not common ones or so called prototypes, i.e. HP which are not typical multiples of 100, 500, or 1000. On the contrary, it might be the case that a particular multiple is not identified as HP since it represents no spike or an abnormal concentration of values. Note, a heuristic is a "rule of thumb". Some of the identified values might be true ones, e.g. 400 EUR.[6] Integer multiples of 100, 500, or 1000 are commonly used for reporting. In the subsequent, mod(100) denotes multiples of 100 but not of 500, mod(500) denotes multiples of 500 but not of 1000, and mod(1000) denotes multiples of 1000. As HP considered are multiples of 100 up to 5000 and multiples of 500 or 1000 up to 10,000. All mod(500) and mod(1000) are identified as real HP in the range of 0 up to 10,000, which was expected from inspection of Figure 1.5. Up to 5500 all mod(100) are identified and above only 5700, 5800, 5900, 6300, 7300, and 8800 are found to be spikes.

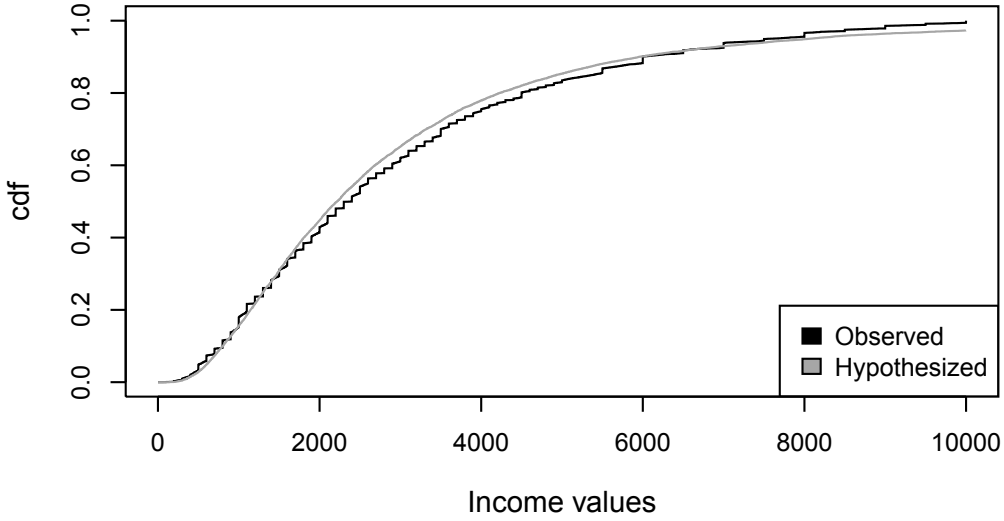
1.1.4 Coping with rounding and heaping

Rounded or heaped values and accordant analyses are treated in several possible ways. These approaches range from rather simple techniques as, for example, ignoring to more sophisticated ones, like imputation and/or modeling:

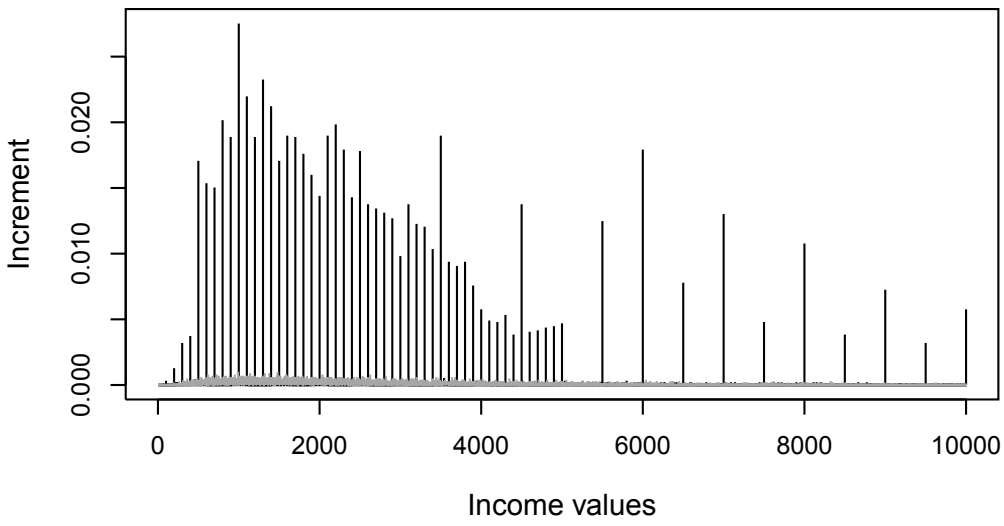
Ignoring

Rounded or heaped values could simply be ignored but it has to be tested first, whether rounding or heaping is ignorable, or whether it has to be regarded in analyses. Ignorability is a property permitting the researcher to neglect the rounding or heaping process. The construct of ignorability was first introduced as a condition for missing data by Rubin (1976, 1987), see Daniels (2008, p. 3). Practically, ignorability for interval data can be accomplished by imposing the uniform distribution across each interval. This relies on the assumption that observations within the interval are equiprobable for being observed, see Heeringa (1996) and Daniels (2008, p. 14). This might hold for rounded data, but heaped data are much more complex. Among others, Torelli and Trivellato (1993), Wright and Bray (2003), Augustin and Wolff (2004) and Daniels (2008) demonstrate the danger of such an approach. Furthermore, Pudney (2008, p. 10) raises the problem that aggregated measures derived from single heaped variables are also affected. This aggregation might, if at all, smooth out the effect of rounding or heaping to a small degree. This picture is substantiated by Marcus, Siegers, and Grabka (2013, p. 13). It calls into question whether even small errors might mount up when aggregating and hence exacerbate the problem of heaping and its effects.

[6] Since April 1, 2003, the threshold value for income values without obligation for paying social security contributions is 400 EUR in Germany.



(a) Empirical cumulative distribution functions (*ecdf*).



(b) Increments of the *ecdf*.

Figure 1.4: The upper graph shows the empirical cumulative distribution function (*ecdf*) estimated from observed net income (given in black), as opposed to the *cdf* estimated from values simulated from the respective hypothetical income distribution (given in grey). The lower graph gives the increments of the *ecdf* of the observed income values as well as the corresponding increments of the *cdf* of the hypothesized income values.

Comparing with external sources

Other studies deal with heaping by comparing survey-based estimates to independent benchmark estimates (Moore et al., 2000, p. 1f.). Usually, administrative records serve as benchmark, since register data are known to be unaffected by heaping or underreporting. Also data collected from diaries can serve as proxy. These figures are then regarded as true, although small deviations from the aggregated population total might exist, see Rendtel et al. (2004, p. 3). Matching to external (benchmark) sources requires the same population and a common set of conditioning variables (Battistin, Miniaci, & Weber, 2003, p. 358). Battistin et al. (2003) present an accordant fully non-parametric matching technique. Following the approach of Heitjan and Rubin (1990), imputations for household expenditures are generated based on the estimates of an inverse Engel curve fit to an external data set. Kraus and Steiner (1995) use register data of the Federal Labour Office to deal with heaped unemployment duration data in the GSOEP. The authors extend the model of Torelli and Trivellato (1993) by further regarding the beginning and ending of episode data.

Bound et al. (2001) provide a review on validation studies to assess more general information on the nature and magnitude of measurement errors across a wide range of survey data, i.a. monthly and annual income, unemployment reports, and health-related variables. The authors explore the perpetuation of the classical independency assumption of measurement errors from the true value and other determinants, on which most of the standard methods rely. This review shows that this assumption does not hold for many studies. To correct for the bias, Bound et al. (2001, p. 3729) suggest using instrumental variables.

So far, the focus was on macro-level comparison of external population-based parameters or estimates of aggregates derived from survey data. Much better suited is the comparison of records at an individual level, e.g. using external data of the individuals included in the survey (Bound et al., 2001, p. 3741). Hanisch and Rendtel (2002, p. 5) give an example for matching retrospectively reported survey data of the Finnish subsample of the European Community Household Panel (ECHP) to registered data from Statistics Finland. One problem that has to be addressed carefully when matching survey data to register data (e.g. for income) is the different reference period. Administered data usually encompass annual records of income covering a calender year, whereas survey data most often refer to income data on a monthly basis, i.e. the income of the preceding month. A formula for computation of monthly reference values based on annual register data is given by the authors as well. These reference values are then used for direct comparison. Hanisch (2006) inquires into the differences between reference values and survey values. He studies distinct measures of income mobility and finds the accordant values to be quite similar for registered and surveyed data. However, income mobility in register data exhibits a higher variance and extreme income changes are more frequent than in survey data. These findings strengthen the assumption that small changes of the true value might be masked by heaping.

Unfortunately, external benchmark sources typically refer to different populations, might cover different observation periods, and might base on different question wording. This clearly complicates comparisons, see Marcus et al. (2013, p. 13). A direct comparison of reported income data with administrative income data seems desirable, but even if linkage is possible[7], related benchmark data are, e.g. in Germany, only available for respondents with episodes of dependent employment and censored at the social security contributing ceiling.

Recently, Antoni et al. (2015) presented a study on the accuracy and underreporting of reported individual gross income data of the National Educational Panel Study (NEPS). The authors link survey data from the NEPS to register data of the Institute for Employment Research (IAB). Informed consent to linkage was available from about 90% of the respondents. Of those, 91% are successfully linked. Before record linkage, only those episodes from full-time employment ongoing or ended shortly before the time of interview are selected further reducing the sample for comparison.

Referring to longitudinal information

Individual specific information might also stem from internal sources, e.g. from longitudinal information. Parameters or estimates of the current wave are compared to parameters or estimates of previous and following waves, respectively. For example, Pudney (2008) enriches reported consumption data by referring to data from previous waves. For this purpose, he uses a multinomial logit model with first-order autoregressive random effects and unobservable individual effects to analyze expenditure behavior in a longitudinal framework (id., p. 17). The model generalizes an approach of Heitjan and Rubin (1990) and allows to specify stability of the responses over time, see Pudney (2008, p. 14).

Ad-hoc correction

One crude ad-hoc correction for heaping would be *omitting* heaped responses. For example, Barreca, Lindo, and Waddell (2001, p. 14f.) remove the heaped data from regression-discontinuity designs (called donut-RD). Discarding heaped entries involves a loss of information, decreases sample sizes and test power, and yields less precise estimates, especially when being not at random. This applies in particular to questions exhibiting large amounts of heaped observations.

Further ad-hoc methods consider the *inclusion of factors* to correct for estimates and their variances. For example, Sheppard's correction on variance estimators is a simple heuristic which adds a correction term to get an unbiased estimator of the variance, see Sheppard (1898). Dempster and Rubin (1983) and T. Liu et al. (2007) propose an ad-hoc adjustment of the parameters in a linear regression model. Dempster and Rubin (1983) use adjustments to the diagonal of the

[7]For data linkage of surveyed to registered data, the respondents have to give their informed consent.

covariance matrix of the variables considered, based on Sheppard's correction factor. T. Liu et al. (2007) propose a two-stage method to estimate the unknown parameters in linear models. Augustin and Wolff (2004, p. 216) derive a closed expression for bias correction which can directly be utilized to construct consistent estimators. Their bias correction approach assumes the heaping probabilities to be dependent only on the marginal distribution which is either known or can be derived from external or validation data, see Augustin and Wolff (2004, p. 215f.). Schneeweiß et al. (2006) develop an estimate for the rounding error based on a Taylor series.

Other suggested ad-hoc methods refer to the inclusion of a *dummy variable* or *random effects* into the model of interest. The performance of a dummy variable for heaped and non-heaped values was explored by Torelli and Trivellato (1993, p. 205) and Barreca et al. (2001, p. 14f.). According to Torelli and Trivellato (1993) this approach yields doubtful results. They stated that even ignoring is better than a dummy, and dummies are only appropriate for modeling spikes owing to true behavior. Another ad-hoc approach used by Barreca et al. (2001, p. 14f.) is a pooled model allowing for separate intercepts and slopes for the heaped data.

Smoothing

The fourth coping strategy *smoothing* requires to specify the heaping points in advance. The excess of observed values at these particular points is then re-allocated to neighboring values. This approach, however, is accompanied by a bunch of sensitive questions, e.g. which distributional assumption to pursue for the approximation of the empirical distribution, which interval width to consider as neighboring area (Marcus et al., 2013, p. 14), and whether to rely on external information or to revert to non-parametric smoothing techniques.

Usually, smoothing procedures are applied prior to inference or comparisons. Already in 1976, Hobson used smoothing functions (expressed in terms of matrix multiplication) to model time budget. In his study, he also investigates the preservation of properties when smoothing heaped data. Sider (1985) suggests assigning a certain proportion of the heaped values to an adjacent interval, as cited by M. Baker (1992, p. 118). Here, the proportion as well as the interval width for re-allocation is fixed in advance. M. Baker (1992, p. 117) uses functional smoothing. An exponential function of a polynomial of higher degree is fit piecewise for each spike and interval separately, allowing the degree of smoothing to vary. According to M. Baker (1992, p. 119), the resulting statistics are quite sensitive to the choice of the corrective. Torelli and Trivellato (1993, p. 205) study a technique to smooth the frequency distribution of spell durations around the spikes. Again, they find doubtful results.

Other approaches model the empirical frequency distribution as a whole. For example, Qian (1996, p. 448) applies a theoretical frequency distribution and employs three smoothing approaches. At start, a weighted average is considered. This weighted average is utilized for a) fitting a probability distribution, b) fitting

local regression models, and c) shrinking wavelet coefficients. The distribution in the heaping interval might not always be symmetric and could follow an asymmetric pattern. Therefore, the distributional assumptions have to be made with particular caution. Allie (2002, p. 94) extends Qian's approach and calls this *inversion procedure*. In the inversion procedure, information from fiscal data is used to derive the distribution in a heaping interval. Allie (2002) shows that his approach produces a better fit than Qian's approach. However, the procedure is marginally less efficient than simple re-allocation (ibid.).

In a recent study, Marcus et al. (2013, p. 14) propose a data-driven approach which automatically tests distinct continuous distributions concerning their fit to empirical income data and uses the best suitable to predict five new values. They find the Generalized Beta distribution II (GB2) to fit best to the data at hand. Given all heaping points, the observed values exceeding the expected frequency according to the GB2 are randomly assigned to the neighboring area. In the end, the re-allocated values follow a GB2 distribution, (id., p. 22). It remains unclear how this re-allocation process is governed and how the neighboring area is defined. Crucial is, that in long-tailed distributions only few values are available in the upper tail hampering the fit. For this reason, Marcus et al. (2013, p. 24) discarded all values above a certain threshold value. This constitutes a further weakness of the approach. Marcus et al. (2013) extend the approach and combine smoothing with multiple imputation. At first, five imputed data sets are created implementing regression-based multivariate imputation by chained equations (MICE). Afterwards, each imputed data is smoothed separately and, once again, using the GB2 to predict new values. This procedure is not state of the art.

Several studies exist with regard to non-parametric smoothing procedures. For example, Hall (1982) examines the influence of rounding errors on non-parametric densities and proposes the mean of neighboring empirical values as estimator. Pickering (1992) models digit preference in gestational age. He estimates a survival model by means of a polynomial regression. In this estimation procedure, he uses varying smoothing parameters and introduces a misclassification model for correct and incorrect reporting. This misclassification model determines probability to observe a misreported value. It constitutes a composite link function which is embedded into the survival model. The big advantage of this method is that it is a non-parametric technique. Non-parametric methods, e.g. kernel density estimation (KDE), do not require any prior knowledge about the nature of the true unobserved distribution, see Minoiu and Reddy (2007, p. 3). A further advantage is its convenience for aggregated data, hence it is very useful for international comparisons. Disadvantageously, there might be a high persistence of local modes even at considerably large bandwidths, see Qian (1996, p. 448), Schweitzer and Severance-Lossin (1996, p. 19) and DiNardo et al. (1996). Furthermore, the results are quite sensitive to the choice of smoothing parameter. Because of that, density estimates from different years or countries should be compared with caution, see Schweitzer and Severance-Lossin (1996) and Mair and Wilcox (2015).

Minoiu and Reddy (2007) study KDE methods on quantile means of income. According to them, these methods cause substantial distortions on grouped data, in particular in the tails of the distribution (id., p. 11). The biases vary with KDE parameter (bandwidth, kernel) but also with the number of analyzed data points. That is, KDE estimators are prohibitively difficult or impossible in small samples (id., p. 3). A further non-parametric smoothing technique combines a penalized likelihood function with a composite link model. Camarda et al. (2007, p. 386) rely on the assumption that the observed values are linear compositions of a latent true distribution. They use a Poisson model for the unobserved count data, and the non-parametric model for digit preference is estimated using B-splines. All non-parametric density estimation procedures rely on the presupposition that the estimate converges to the true density as the sample size rises to infinity and the bandwidth goes to zero. However, with rounding or heaping, convergence is not assured, see Schweitzer and Severance-Lossin (1996, p. 19). The more values are heaped, the less the observed values follow a continuous distribution function which is a prerequisite for convergence (id., p. 19). Groß and Rendtel (2015) suggest to model the kernel density of the latent distribution and the parameters of the heaping process simultaneously. They use a partially Bayesian Stochastic Expectation Maximization (SEM) algorithm for estimation. At start, the intervals are defined to be symmetric and the probabilities of the heaping process are assumed to be equal for all respondents and independent from the unobserved true value. As an extension, these suppositions are relaxed by introducing a parameter for the heaping direction bias and by imposing an ordered probit model to allow for non-constant heaping probabilities.

Imputation

Another approach to correct for heaping during data processing is using multiple imputation, see i.a. Heitjan and Rubin (1990), van der Laan and Kuijvenhoven (2011), or Drechsler and Kiesl (2012, 2014) and Drechsler et al. (2015). As well as smoothing, imputation requires specification of a model. The accordant parameter estimates are then used to generate new values as proxies for the true unobserved data. Imputation is performed multiply to account for the uncertainty in the imputation process (Rubin, 1976, 1987). Compared to methods which estimate the parameters from the underlying model directly, multiple imputation is robust against misspecification of the model parameters as well as against misspecification of the heaping intervals, as shown by el Messlaki et al. (2010), for example.

An early imputation approach proposed by Heitjan and Rubin (1990) assumes a naïve model on the one hand and a complex model with covariates on the other hand. In the naïve model, the set of imputes is drawn independently and uniformly from the heaping intervals. The authors study the impact of varying interval widths on diverse inferences. They find that widening the intervals results in increased standard errors owing to the larger between-imputation variance. In the complex model a linear regression for prediction of true age as well as for

prediction of the heaping type (full-year, half-year heapers vs. exact respondents) is used. A more recent approach by van der Laan and Kuijvenhoven (2011) employs a multinomial model to explain the heaping probabilities which are assumed to be constant within predefined intervals. By means of their approach, van der Laan and Kuijvenhoven (2011) model duration data. For this purpose, they use a discrete time model (piecewise constant hazard model) to model the latent true distribution. After estimating the surplus values on heaping points, new values are multiply imputed for randomly selected ones from the corresponding intervals. They find that the method works well with small intervals and a completely covered range.

Drechsler and Kiesl (2012) estimate the posterior distribution of the heaping probabilities given the observed data and than re-impute the partially heaped data. The threshold values of the ordinal probit model represent the unknown model parameters. Multiple plausible candidates for the unknown true values are generated by a simple rejection sampling approach. Those candidates being inconsistent with the observed heaped value and the drawn indicator for heaping are immediately rejected, as proposed by Heitjan and Rubin (1990). In a following study, Drechsler and Kiesl (2014) propose two correction methods: one for values with known intervals (grouped data) and one for values with unknown, possibly varying, intervals. The latter method allows estimating the degree of heaping as well. In addition, covariates are included in the modified model. In a further paper, Drechsler et al. (2015) likewise address the non-response problem. The authors use partial information from bracketed income questions and additionally adjust for non-response. The proposed approach is implemented in form of a sequential regression multivariate imputation procedure.

Modeling

Obviously, also smoothing and imputation require modeling. However, there are approaches which consider only modeling of heaping or rounding. Different authors borrow from the missing data techniques to set up models for coarsened data, see i.a. Heitjan and Rubin (1990, 1991), Gill et al. (1997), and Kim and Hong (2012). Heitjan and Rubin (1990) employ non-ignorable and ignorable versions of their model in which the probability of heaping either depends on the outcome, or does not. Concretely, Heitjan and Rubin (1990) use a bivariate normal distribution conditional on covariates to model true age and the heaping indicator. Kim and Hong (2012) adapt a Monte Carlo Expectation Maximization algorithm for missing data (Wei & Tanner, 1990) to handle coarse data. Their method is readily applicable to model a deterministic coarsening mechanism but also to model an extension for a more general, stochastic coarsening mechanism.

Huttenlocher et al. (1990) propose a simple model for digit preference in temporal reports (duration data). The distribution of responses is assumed to be a convolution of three distributions. First, values taken from memory are described by a normal distribution. Second, the feasible value ranges are specified

by a doubly truncated normal distribution. Third, a mixture model is introduced to differentiate three types of response behavior: reporting of unaltered values, values according to calendar prototypes, or values according to arithmetic prototypes (the concept of prototypes is explained in the introduction). To allow for reduction of the number of model parameters, several functional forms – exponential, linear, or fixing to one particular value – are employed to model the response behaviors. Ridout, Martin S., and Morgan (1991) model digit preference in fecundability studies. According to the authors, the inclusion of additional parameters for modeling misreporting can lead to substantial improvements in model fit. In their study, however, they find only minor changes to the estimated parameter values of the underlying beta-geometric distribution. A quite similar model propose Torelli and Trivellato (1993) to model event data. Though, Ridout et al. (1991, p. 1432) further include covariates into their misreporting model in form of a generalized linear interactive modeling approach. The approach yields reliable results according to the authors.

Wright and Bray (2003) inquire into the merits of a mixture model for rounded foetal measurements obtained from ultrasound images. A two-component mixture model is employed, incorporating a latent indicator variable to model the uncertainty owing to the undetermined grouping interval and a simple linear regression model of the assumed true values. To capture the heavy-tailed behavior of the data, the model for the true data is refined using a Student t error distribution. The refined model shows a better fit to the data at hand. Lin and Tsai (2013) use a mixture of a multinomial logistic and a Poisson distribution for modeling health survey data. The first component is used to account for excessive zero and heaped responses and the second one is used to model the count data.

Techniques to model heaping or rounding in event data are given i.a. in Torelli and Trivellato (1993), Petoussis, Gill, and Zeelenberg (2004), or Bar and Lillard (2012). The approach of Torelli and Trivellato (1993) combines a measurement model for heaping with a duration model. As heaping function, an exponential cumulative distribution function is used and the model of interest is a fully parametric proportional hazard model (Weibull, log-logistic, exponential), both without covariates. Petoussis et al. (2004) amend this method by including covariates and by modeling heaping at the beginning as well as at the ending of events (instead of modeling heaping of durations directly). The authors apply a Cox's Proportional Hazards Model. Assumptions made by the authors are: heaping is assumed to be dependent only on the true value, the distribution of beginning dates is assumed to be uniform, and heaping on ending dates is assumed to be dependent on censoring (while seasonal effects are neglected). The model of Bar and Lillard (2012) considers multiple heaping rules thereby allowing for distinction between true heaping and actual behavior. Bar and Lillard (2012) accommodate a wide range of distributions – also mixture distributions. The heaping probabilities are assumed to follow a functional form. In the absence of any particular properties assumed, this functional form can be as simple as a Bernoulli distribution. A

complex functional form facilitates modeling subject-specific heaping probabilities by including covariates. However, the authors leave the addition of covariates for future work. Also the multiple heaping rules, each of which can be modeled with a different probability distribution function, are left open for future research.

Heaped count data are modeled by Harris and Zhao (2007), Wang and Heitjan (2008) and Wang et al. (2012). In Wang and Heitjan (2008), an underlying precise count variable (Poisson or negative binomial distribution) and a heaping behavior variable are used for modeling. The authors compare models with and without zero-inflation regarding the sensitivity of the inference. A refinement of the model to consider the degree of heaping, as proposed by Heitjan and Rubin (1990), is implemented. Four separate logistic regressions for each cut point with common slopes and varying intercepts are estimated for this specific application. Constraints are imposed on the model in that the intercepts are assumed to have an ordered structure. Harris and Zhao (2007) extend the ordered probit model of Wang and Heitjan (2008) to a zero-inflated ordered probit model. Concretely speaking, the authors propose a combination of a split probit model and an ordered probit model to separate tobacco users into two latent groups of users and nonusers. In a further study, Wang et al. (2012) compare retrospective recall data with instantaneously recorded data. In a two component model, fixed and random effects are included to model the two-stage process of misremembering and heaping. At first, unobserved true data are predicted from recorded, precise data. Second, the heaping behavior is predicted by means of an ordinal logistic regression to model also the degree of inaccuracy.

Summary

This enlisting of strategies to cope with digit preference, rounding, heaping, or coarsened data is not exhaustive. The author of this thesis identifies two major drawbacks of existing approaches. First, all presented approaches apply to specific cases (since they consider certain variables and settings). Second, the non-parametric techniques presented usually need a lot of observations and can especially be problematic for long-tailed distributions with values in the upper tail that do not have immediate neighbors. In response to these obstacles, a parametric technique will be presented to deal with heaping in a more general and flexible way with respect to definitions of the latent distribution but also regarding the heaping behavior. It will be shown that the proposed method yields feasible results in plain settings but also when assuming much wider intervals and a long-tailed distribution in which the number of adjacent neighbors is low per se but gets even lower with increasing value, as is the case e.g. in self-reported income data.

Even though this thesis is motivated by heaping behavior found in self-reported income data of the National Educational Panel Study (NEPS) the established model is not restricted to income data solely. The specifications of the proposed method are admittedly adapted to heaping behavior in income data and the overall fit of the model is tested against NEPS income data in the application. However,

as will be demonstrated in Chapter 4, several modifications of the model can be taken into consideration. This enables to account for heaping in diverse settings. For example, to model heaping behavior in duration data, adjustments according to the heaping pattern – the sets of heaping points and corresponding intervals – as well as the latent distribution are necessary. With regard to duration data, piecewise constant hazard rates might be used as latent distribution, and according to Huttenlocher et al. (1990, p. 212) the calendar prototypes 7, 10, 14, 21, 30, or 60 are advisable as heaping points.

1.2 Motivating example

Net income data from the first wave of Starting Cohort 6 (adults) of the German National Educational Panel Study (NEPS) is used for illustration of heaping behavior and its associated structures (`doi:10.5157/NEPS:SC6:1.0.0`).[8] The NEPS is a large-scaled infrastructure project in Germany carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg. Longitudinal data covering the development of competencies, educational processes, educational decisions, and returns to education from six independent cohorts is collected, thereby covering the lifespan between early childhood and adulthood. Besides formal environmental influences, non-formal, and furthermore, informal contexts are regarded.[9] The NEPS provides access to these data in the form of scientific use-files (SUF). Access is admissible for registered users from the scientific community. The SUF 1.0.0 of SC6 consists of data from the first NEPS wave ($n = 5154$) and data of re-interviewed respondents ($n = 6495$) from the previous study ALWA "Working and Learning in a Changing World" conducted by the Institute for Employment Research (IAB, Nuremberg). That is, the NEPS sample integrates the ALWA sample, a refreshment sample, and an enhancement sample, altogether covering individuals born between 1944 and 1986. Both, the ALWA and the refreshment sample cover the birth cohorts from 1956 to 1986. The SUF is provided in a modular form consisting of 22 modules. For selected areas of life, a separate module is provided. Information on net household income data (NEPS-HH) is available in the target panel module (`pTarget`). The individual income data (NEPS-Ind) focused here is available in the employment history module (`spEmp`).[10] Overall, 47,368 employment episodes are available in the module and remain after correction by means of the biography module. From these spells, the number per each respondent is reduced to one, see Aßmann, Würbach, Goßmann, Geissler, and Biedermann (2014). Concretely, only one episode of main activity is used per each respondent, either finished or ongoing.

[8] See Blossfeld, Roßbach, and von Maurice (2011) for a general discussion of the study design and Leopold, Raab, and Skopek (2011) for a general documentation of the scientific use file.

[9] See `https://www.neps-data.de/en-us/home.aspx` for more detailed information.

[10] See Appendix A.1 for the specific question wording.

To keep it simple, the stratified multistage sampling design of SC6 is ignored and the observed income information is treated as if it was obtained from a simple random sampling. For this reason, no general statements can be made on the target population. Furthermore, cases with missing income values are omitted in a first place. This chapter starts with a short description of the data. After an initial inspection, marginal frequencies of some selected covariates are explored. The dependencies between the independent variables gender, educational level, and age and the dependent variable income are further traced by a regression tree and a log-linear model for logarithmized net income disregarding any heaping behavior at first. Third, some personal traits are further regarded for explanation of the occurrence of heaping. A classification tree is grown for the detection of dependencies between the covariates and heaping as binary outcome – whether a reported value falls on a heaping point or not –, and a probit regression is run to identify and quantify the dependencies.

Finally, these results from real data description are used for the data generating process to simulate data with plausible distributional characteristics. The estimates from the log-linear model are used for an alternative multivariate modeling of income data. The estimates from the probit regression are used for simulation of an accordant heaping mechanism. Simulated data will be used to test the performance of the model introduced in Chapter 2.

1.2.1 Description of the NEPS income data

Both types of income data, the net individual income and even more pronounced the net household income, reported in the NEPS adult cohort sample of wave 2009/2010 are evidently exposed to heaping behavior. A total of $n = 8685$ individuals gave reasonable information about their net individual income and $n = 10{,}012$ on net household income. Figure 1.5 shows the respective frequency distribution of the net individual income, and the corresponding figure for the net household income is given in Figure A.2 in the Appendix. The mean of the net individual income is located at 1881 EUR and the median at 1700 EUR, both indicating a distribution that is skewed to the right. The values 331 EUR and 4200 EUR mark the 5th and 95th percentile, hence 90% of the probability mass lies between them. The percentage of reported zeros is 1.69. Abnormal concentrations of reported values can be found at thousands, five-hundreds and even at hundreds (Figure 1.5). Spikes at values ending with 500 and 1000 are particularly striking above the mode of 2000 EUR (Zinn & Würbach, 2015, Section 5). In Table 1.1 the percentages for the integer multiples of 5 up to 1000 are presented for both types of income. Integer multiples of 100, 500, or 1000 are commonly used for reporting. The relative frequency of values at mod(500) is about 10% and of values at mod(1000) is about 14% in net individual income data. Most of the values are rounded off to mod(100), about 45%.

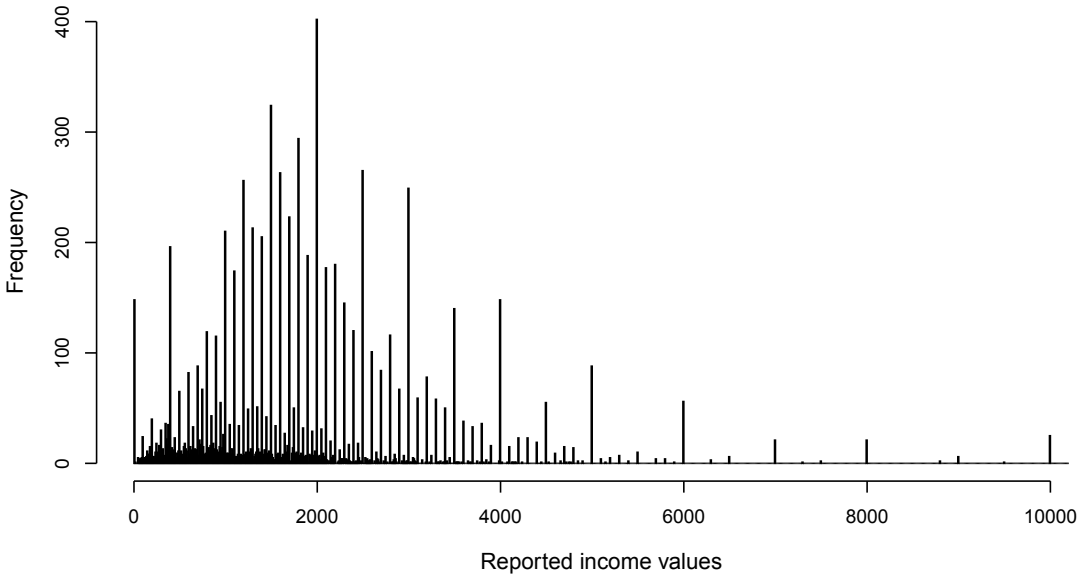


Figure 1.5: Self-reported net individual income data from the Adult Cohort in the NEPS wave 2009/2010, $n = 8685$ ($\leq$ 10,000 EUR).

Table 1.1: Percentages of heaped values per modulo in NEPS income data.

Source	mod(5)	mod(10)	mod(50)	mod(100)	mod(500)	mod(1000)
NEPS-Ind[1]	2.23	10.35	8.93	45.33	9.90	14.02
NEPS-HH[2]	0.59	3.22	3.62	40.98	19.42	29.60

[1]Net individual income data in the NEPS SC6 wave 2009/2010, $n = 8685$.
[2]Net household income data in the NEPS SC6 wave 2009/2010, $n = 10{,}012$.

As expected, the prevalence of heaping is higher in the reported household income values (90.0% vs. 69.3% with regard to multiples at 100, 500, and 1000). Additionally, reports on household income are much more inaccurate with more values falling on mod(500), about 20%, and mod(1000), 29.6%. These patterns are already shown, i.a. in Hanisch (2005a, p. 42f.). The author argues that the individual income is more familiar than the household income thus the first one being reported at a higher degree of accuracy, as opposed to the latter one.

Table 1.2 gives the proportions of the spikes (heaping points) with regard to specified intervals for the individual income in particular. In sum 69.26% of the reported individual income values fall on a modulo, which is a high prevalence. The concentration of heaped values is highest in the range of 1000 and 3000 (13-17% vs. 2-7.5%). This fact is not surprising because the intervals covering this range comprise the majority of income values, in sum 63.2% (Zinn & Würbach, 2015).

Table 1.2: Percentage of values located at the modulos in the NEPS net individual income data.

Interval	mod(100)	mod(500)	mod(1000)	Total
$[0, 500]$	2.99	0.74	–	3.73
$(500, 1000]$	4.38	–	2.35	6.73
$(1000, 1500]$	9.59	3.68	–	13.27
$(1500, 2000]$	11.07	–	4.61	15.68
$(2000, 3000]$	11.36	3.04	2.87	17.27
$(3000, 4000]$	4.19	1.59	1.70	7.48
$(4000, 5000]$	1.37	0.63	1.01	3.01
$(5000, 10{,}000]$	0.38	0.22	1.49	2.09
Total	45.33	9.90	14.03	69.26

Counting of integer multiples of 100, 500, or 1000 is practicable, though, only a crude measure of the degree of heaping. Great disadvantage of this measure is that it does not take the total length of the reported value into account. Because of this, the relative RI according to Hanisch (2005a) and the RSM according to Serfling (2006) are reported for the data at hand (cp. Section 1.1.3). High accuracy of a reported value yields a high relative RI but low RSM. Otherwise, a low relative RI and high RSM indicates an inflated degree of heaping. The relative RI from NEPS individual and household income is presented in Table 1.3 and the RSM in Table 1.4. Category 1 of the RSM is not assigned for the NEPS income data which means none of the values with total number of digits of five or more are reported straight with significant digits.[11] Since the proportion of reported values is overall quite low for those in the higher ranges, even when being heaped, category 5 is underrepresented.

Table 1.3: Proportionate frequencies of the relative Rounding Indicator (RI) in NEPS income data.

Source	0.25	0.33	0.40	0.50	0.60	0.67	0.75	0.80	1.00
NEPS-Ind[1]	13.62	8.05	0.62	47.03	–	8.87	10.33	–	11.48
NEPS-HH[2]	28.88	2.60	1.17	57.38	0.06	2.28	4.48	0.01	3.14

[1]Net individual income data in the NEPS SC6 wave 2009/2010, $n = 8685$.

[2]Net household income data in the NEPS SC6 wave 2009/2010, $n = 10{,}012$.

[11]Serfling (2006, p. 106) omits the observations at category 1 for further analyses due to the small cells problem and to assure ordinality of the measure.

Table 1.4: Proportionate frequencies of the Rounding Strain Measure (RSM) in NEPS income data.

Source	1	2	3	4	5
NEPS-Ind[1]	–	9.61	68.09	21.95	0.35
NEPS-HH[2]	–	2.91	64.43	32.06	0.59

[1]Net individual income data in the NEPS SC6 wave 2009/2010, $n = 8685$.
[2]Net household income data in the NEPS SC6 wave 2009/2010, $n = 10{,}012$.

1.2.2 Multivariate consideration of income data and heaping behavior

Besides the univariate description of reported income and the incidence of heaping, the relationships between some candidate covariates and the dependent variables net income and heaping behavior are to be explored. Proofs that heaping is strongly related to certain individual characteristics are given in Schräpler (1999), Hanisch (2005a, 2006), and Serfling (2006), cp. Section 1.1.2. The next two sections describe the analyses of the dependencies between selected individual characteristics (internal factors) with respect to the income level as well as their influence on the occurrence and degree of heaping. Gender, age, and educational level are taken into consideration. Elderly people, males, and people with higher educational level are expected to have higher income values. With respect to heaping, older people, females, and people with higher educational level are expected to report their income more accurately than younger people, males, and people with lower or middle educational level, according to the authors.

About 52% of the respondents are males. This surplus of males is due to the fact that women work less often. More than a half (52.6%) of the respondents with a net income have a medium level of education, less than 20% have a lower educational level, and about 28% have a higher educational level.[12] The average age is 46.58 years with minimum at 23.08 and maximum at 68.83.

In the following descriptions and models, missing values are imputed multiply using the full conditional multiple imputation approach of Aßmann et al. (2014, 2015). The authors adapted a non-parametric tree-based sequential regression approach that combines the binary recursive partition algorithm CART

[12]The following CASMIN groups are summarized to get three ordered categories: "[1a] No school leaving qualification", "[1b] General elementary education", and "[1c] Basic vocational training above and beyond compulsory schooling" for lower educational level; "[2b] Intermediate general education", "[2a] Intermediate vocational qualification, or secondary programmes in which general intermediate schooling is combined by vocational training", "[2c gen] General maturity: Full maturity certificates (e.g. the Abitur, A-levels)", and "[2c voc] Vocational maturity: Full maturity certificates including vocationally specific schooling or training" for medium level of education; "[3a] Lower tertiary education: Lower level tertiary degrees, generally of shorter duration and with a vocational orientation", "Higher tertiary education: The completion of a traditional, academically oriented university eduction" for higher educational level.

(classification and regression trees), see Burgette and Reiter (2010) and Breiman, Friedman, Olshen, and Stone (1984), and the imputation technique MICE (multivariate imputation by chained equations), see van Buuren (2007, 2012) and van Buuren and Groothuis-Oudshoorn (2011). That is, non-parametric characterizations of the full conditional distributions for the missing values are used and a set of identified donor observations operationalize the full conditional distributions. For this, CART splits up the observations into different groups where the intra-group homogeneity and inter-group heterogeneity are highest with respect to the relevant variable. Using the CART approach of Burgette and Reiter (2010), to find these optimal partitions, obviates the explicit specification of the conditional models for the considered variables, which is especially burdensome for categorical variables. The chained equations approach manages that all variables in the data set with missing data are imputed in sequence. To be concrete, for each variable with missing values an individual imputation model is specified, see van Buuren and Groothuis-Oudshoorn (2011). The conditional models are then sequentially chained by using all variables previously imputed for explanation in the following model. All missing information in all enlisted variables is imputed at once. The approach is particularly suitable to preserve also nonlinear relationships among the variables. After initialization, the algorithm iteratively runs 10 times through all imputation models to mitigate the effect of initialization and the following 10 sequences[13] are stored, yielding 10 completely imputed data sets ready for analyses, see van Buuren (2007).

The multiply imputed data sets are analyzed using Rubin's Combining Rules, see Rubin (1987, p. 76) and Barnard and Rubin (1999). For this, standard complete data analyses are performed on each imputed data set and the estimates and the corresponding standard errors are then combined. The pooled estimate is purely the average of all single estimates, whereas the combined standard errors result from the average within-imputation variance and the between-imputation variance. The corrected confidence intervals, fraction of missing information (fmi), and the proportion of the variance attributable to the missing data ($lambda$) will be given as well in the following analyses, see Schafer (2001, p. 15) and van Buuren (2012, p. 41f.). The quantities $lambda$ and fmi are measures of the severity of the missing data problem. Both can be interpreted similarly, though, the fmi is adjusted for the number of imputations (id., p. 41).[14]

[13]The author of this thesis set this value arbitrarily. No general recommendation exists, but van Buuren (2012, p. 50) points to the fact that a higher number of stored imputed data sets, i.e. more than 5, is better to obtain sufficient statistical power.

[14]For more details regarding the fmi and $lambda$ see also Rubin (1987, p. 77f.). An illustration for analyzing multiply imputed data as well as the complete R code is given in Würbach, Hammon, Geissler, and Goßmann (2014) for some examples with reference to the imputed data of NEPS Starting Cohort 6.

Personality traits and net income level

First of all, the combined mean statistics of net income with respect to gender and educational level are given in Table 1.5. The fmi and *lambda* are both well below 0.2 indicating that only few variation is caused by the missing data (van Buuren, 2012) and the influence of the imputation model on the final result is innocuous. The main effects of gender and educational level on income level are discernible. Female respondents have an average net income of 1340.46 EUR and their male counterparts have an average net income which is more than 1000 EUR higher. With respect to the educational level, only minor differences between lower educated people and respondents with medium educational level exist. Only those respondents with higher educational level earn on average 1000 EUR more. This pattern is even clearer when separating the respondents according to their gender and educational level. The boxplots given in Figure A.1 in the Appendix accentuate the differences between the subgroups. The effect of age becomes pronounced when separating the respondents according to their gender, see Figure 1.6. For men, the income gain with increasing age is much more evident than for women.

Table 1.5: Combined mean statistics for net income, divided by subgroups.

Group	Estimate	SE	CI	fmi	*lambda*
All	1927.19	20.833	[1886.34, 1968.04]	0.0432	0.0426
Female	1340.46	17.324	[1306.48, 1374.44]	0.0691	0.0679
Male	2461.18	34.595	[2393.36, 2529.01]	0.0369	0.0364
Lower edu	1616.95	25.958	[1566.05, 1667.84]	0.0376	0.0371
Middle edu	1646.97	29.223	[1589.68, 1704.25]	0.0276	0.0272
Higher edu	2666.14	42.466	[2582.88, 2749.40]	0.0385	0.0379
Female lower edu	995.56	29.111	[938.43, 1052.69]	0.0944	0.0925
Female middle edu	1195.02	19.301	[1157.16, 1232.88]	0.0708	0.0695
Female higher edu	1884.93	43.245	[1800.13, 1969.74]	0.0573	0.0564
Male lower edu	1989.10	32.782	[1924.82, 2053.37]	0.0408	0.0403
Male middle edu	2171.79	56.872	[2060.30, 2283.28]	0.0195	0.0192
Male higher edu	3261.08	62.249	[3139.05, 3383.12]	0.0270	0.0266

To get more insight into the joint influence of all three internal factors, a regression tree with net income as dependent variable is built based on the pooled data sets. Trees are especially helpful to visualize multivariate dependencies with three or more variables. They provide a good summary of the relationship between all factors considered as well as their importance in the context studied, see Breiman et al. (1984). Split nodes are denoted by ellipses and rectangles indicate terminal nodes. The different shades of grey indicate an increase in the income level. Splitting of the tree is supposed to stop when the minimum number of observations in any terminal node is reached, which is set to 100, or when the decrease in the overall lack of fit does not exceed the complexity parameter fixed at 0.01.

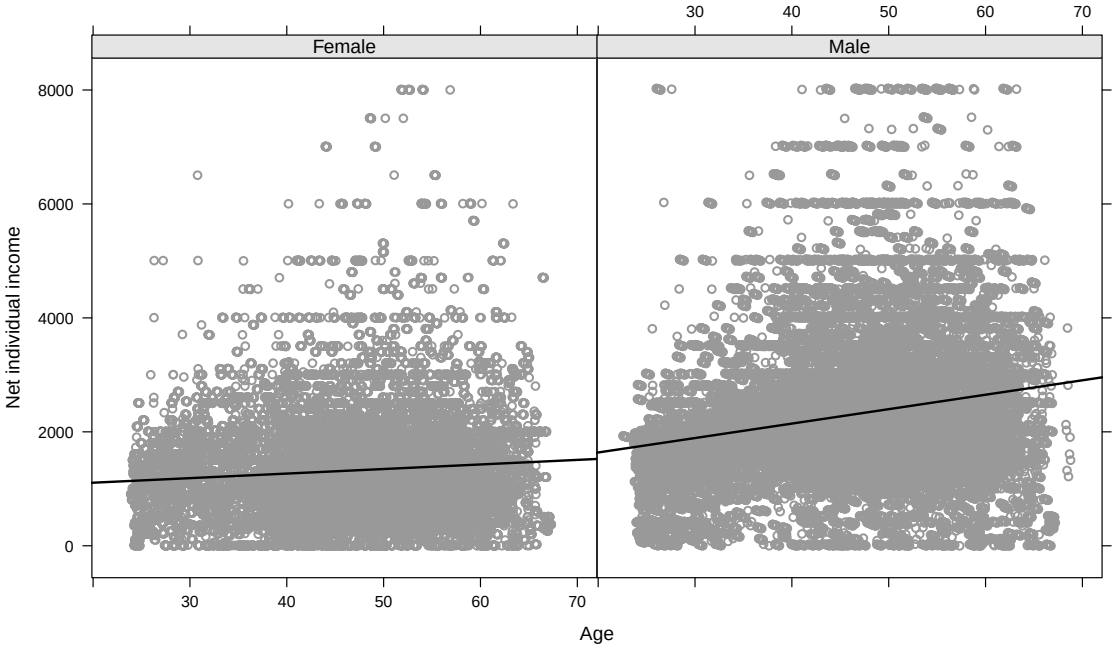


Figure 1.6: Net individual income of females and males by age.

The root node of the regression tree in Figure 1.7 starts with all 8714 (partially imputed) observations having an overall mean in net income of 1927 EUR, cp. also Table 1.5. The first split is made according to gender. Female respondents are separated in the left branch and males in the right branch. All 4152 females (47.6% of the respondents) are then split according to their educational level. Females with lower and middle educational level (35.5% of the respondents) form the terminal node with the lowest average net income (1154 EUR). The male respondents are also split by educational level. For those males with lower and middle educational level (36.4%), a further split is made by age at the cut-off point 34.1 years, whereas the higher educated men are separated at the cut-off point 38.8 years. The group of lower and middle educated men younger than 34.1 years form the terminal node with the second lowest average income (1401 EUR, 6.1%). The two groups with the highest average net income are formed by higher educated men, either younger than 38.8 years (2265 EUR, 3.0%), or older equal 38.8 years (3493 EUR, 12.9%).[15] The order of the relevance of all three individual characteristics can be extracted directly from inspecting the regression tree, but in the model summary also the relative variable importance[16] for each of the consid-

[15]The logarithmic or inverse hyperbolic sine (IHS) transformation of income before building the tree yields only minor changes regarding splitting variables and splitting points (cp. Figure A.4 and Figure A.5 in the Appendix).

[16]The variable importance summarizes the appearance of a variable as a primary and a surrogate splitting variable by means of the goodness of split measures. The variable importance is scaled to sum up to 100 and rounded to integers.

ered explanatory variables is given. Gender is considered as the most important variable here (54%), then educational level (34%), and age has the lowest relative importance (13%).

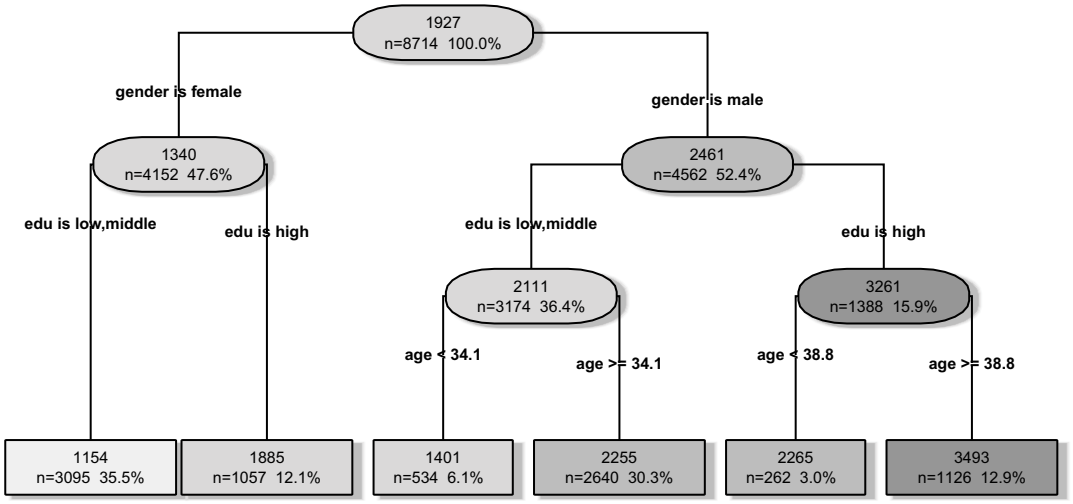


Figure 1.7: Regression tree for net individual income.

A log-linear model is estimated in order to quantify the effects of the previously described covariates on net income.[17] The results from Table 1.6 show that all of the variables considered have a highly significant influence on the net income level. About 27% of the variance in net income can be explained by the log-linear model ($AIC = 16{,}568.6$). The fmi and $lambda$ are both below 0.2 indicating that the influence of the imputation model on the final result is modest (van Buuren, 2012, p. 42).

Table 1.6: Results from combined log-linear regression for net income.

Predictor	Estimate	SE	CI	df	t-ratio	p-value	fmi	$lambda$
(Intercept)	6.296	0.042	[6.213,6.379]	4013.2	148.46	<0.001	0.0344	0.0340
Male	0.609	0.015	[0.580,0.638]	1475.0	41.30	<0.001	0.0717	0.0705
Age	0.010	0.001	[0.009,0.012]	2467.9	13.10	<0.001	0.0511	0.0504
Middle edu	0.137	0.020	[0.098,0.175]	899.2	6.98	<0.001	0.0961	0.0941
High edu	0.565	0.022	[0.522,0.607]	481.8	26.20	<0.001	0.1357	0.1322

[17]The log-linear model is estimated since net income does not follow a normal distribution, KS-test p-value <0.001, see Figure A.3 in the Appendix. The excess in zeros is omitted before estimation.

Personality traits and the tendency and degree of heaping

For detection of possible dependencies between the individual characteristics and the heaping behavior, two variables are created that represent the tendency to and the degree of heaping. First, a binary variable separates all values being located at modulos from those being not. Second, in an ordered variable reported values are further separated into those being located at hundreds (including five-hundreds), and those being located at thousands.

The overall percentage of heaped values is approx. 69%. The propensity for heaping grows with increasing level of income (level dependency). While less than half of the values below 1500 EUR are heaped, 80% of the values between 1500 and 3000 EUR, and almost 94% of the values above 3000 EUR fall on a modulo, see Table 1.7. Female respondents have indeed a lower percentage of heaped values opposed to male respondents (58.0% vs. 79.3%). The difference is remarkable with more than 20%. Educational level is considered as proxy for ability (Narayan & Krosnick, 1996) since schooling is by far the strongest determinant for numeracy (Crayen & Baten, 2008). Albeit people with high ability are assumed to be less likely to be satisficing, and hence to heap, higher educated people have a 10% increased propensity to heap values than lower or middle educated respondents in the data at hand. This also contradicts the findings from Antoni et al. (2015) with regard to the deviation from administered data but is in line with the findings of, e.g. Serfling (2006, p. 115).

Table 1.7: Combined percentages for observing heaping, divided by subgroups.

Group	Heaping	No heaping
All	69.16	30.84
Income less than 1500 EUR	48.37	51.63
Income 1500 up to 3000 EUR	80.09	19.91
Income more than 3000 EUR	93.94	6.06
Female	58.02	41.98
Male	79.30	20.70
Lower edu	67.83	32.17
Middle edu	65.34	34.66
Higher edu	77.22	22.78
Female lower edu	49.70	50.30
Female middle edu	55.90	44.10
Female higher edu	67.90	32.10
Male lower edu	78.69	21.31
Male middle edu	76.31	23.69
Male higher edu	84.33	15.67

When inspecting the differentiated degrees of heaping in Table 1.8, most of the values are heaped at hundreds, almost 55%. The higher percentages of heaped values with increasing income are mostly attributable to an increase in heaped values that fall on mod(1000) (level dependency). See Figure A.6 in the Appendix for more details regarding the quartiles. Hanisch and Rendtel (2002, p. 4) also find both, the frequency of heaped values and the level of heaping, being increased with income level. According to the authors, this association is not simply proportional. Hanisch and Rendtel (2002, p. 7) argue that this relation depends on the degree of heaping expressed by the relative Rounding Indicator (RI). For male respondents, both, heaping at hundreds as well as heaping at thousands, are increased by 10%. With regard to the educational level, no differences exist according to the proportion of values that fall on hundreds, but values that fall on thousands are increased by 10% for those with higher educational level. Separating the respondents by gender and educational level corroborates these findings more clearly. The increase in heaping is almost linearly among the subgroups.

Table 1.8: Combined percentages for different degrees of heaping, divided by subgroups.

Group	No heaping	Heaping at hundreds	Heaping at thousands
All	30.84	54.91	14.25
Income less than 1500 EUR	51.63	42.49	5.87
Income 1500 up to 3000 EUR	19.91	69.21	10.88
Income more than 3000 EUR	6.06	49.94	44.00
Female	41.98	49.08	8.94
Male	20.70	60.21	19.08
Lower edu	32.17	55.68	12.15
Middle edu	34.66	54.13	11.21
Higher edu	22.78	55.82	21.40
Female lower edu	50.30	44.53	5.17
Female middle edu	44.10	48.63	7.28
Female higher edu	32.10	52.83	15.07
Male lower edu	21.31	62.35	16.33
Male middle edu	23.69	60.53	15.78
Male higher edu	15.67	58.10	26.22

With regard to the joint influence of gender, age, and educational level, the author of this thesis now grows a classification tree that has the binary variable for occurrence of heaping as its dependent variable. Again, splitting is supposed to stop when the minimum number of observations in any terminal node is reached (100) or when the decrease in the overall lack of fit does not exceed 0.0015.

The proportions for heaping (0.69) or no heaping (0.31) for all 8714 observations are given in the root node of the classification tree (Figure 1.8). One more time, the first split is made according to gender, i.e. all male respondents are separated in the left branch and females in the right branch. No further split seems necessary for male respondents. They form the terminal node that exhibits the highest proportion of heaped values (0.79, 52%). The female respondents are split according to their educational level. Females with higher educational level (12% of the respondents) form the terminal node that has the second highest proportion of heaped values (0.68). A further split is made to separate the lower from the middle educated female respondents. For those females with lower educational level (7%), further splits are made by age at the cut-off points 51 and 57 years. The groups of lower educated women younger than 51 years and older equal 57 years form the terminal nodes in which the proportion of non-heaped values exceed those being heaped (0.53 and 0.51 vs. 0.47 and 0.49). The relative variable importance in the presented classification tree for each of the variables considered are 85%, 14% and 1%. That is, gender again is considered as being most influential, then the educational level followed by age. In Figure A.7 also the income level is included.

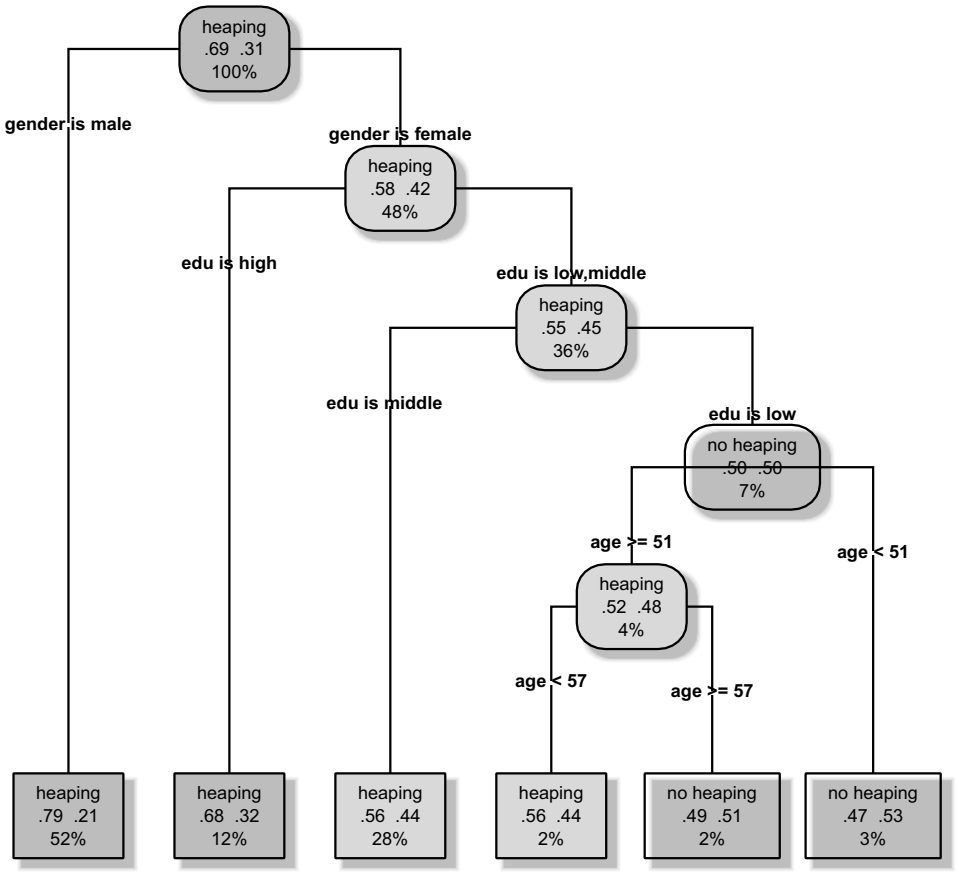


Figure 1.8: Classification tree for observing heaping.

To quantify the effects of the internal factors on the binary outcome heaping or no heaping, a probit model is estimated.[18] As expected, the effects of gender, age, and higher educational level are highly significant, see Table 1.9. Nagelkerke's pseudo-R^2 is 0.096 and the AIC[19] equals 9818.3. The quantities fmi and $lambda$ are both below 0.2 indicating a modest influence of the imputation model on the final result (van Buuren, 2012, p. 42). The results presented here replicate the findings from Hanisch (2005a, p. 46) that older people have the tendency to report more precisely. Whereas males report more accurately in the Finnish ECHP, this could not be confirmed here. However, this is in concordance with the analyzes of Hanisch (2005a) on the German ECHP data. The difference between respondents with lower educational level and those of middle level is not significant, but the difference to higher educated respondents is highly significant (p-value < 0.001). This points to an increased propensity for heaping when being higher educated, as opposed to having a lower educational level.

Table 1.9: Results from combined probit regression for the tendency to heap.

Predictor	Estimate	SE	CI	df	t-ratio	p-value	fmi	$lambda$
(Intercept)	0.410	0.088	[0.238,0.582]	1080.7	4.67	<0.001	0.0864	0.0847
Male	0.611	0.032	[0.549,0.673]	642.7	19.37	<0.001	0.1160	0.1132
Age	0.011	0.002	[0.008,0.014]	683.7	6.67	<0.001	0.1121	0.1095
Middle edu	0.084	0.041	[0.004,0.164]	715.4	2.05	0.040	0.1093	0.1068
Higher edu	0.385	0.045	[0.297,0.473]	3508.5	8.56	<0.001	0.0389	0.0383

Notes: Nagelkerke's pseudo-$R^2 = 0.096$ for each single probit regression, $AIC = 9818.3$.

Two relative measures have been introduced in Section 1.1.3 which give more insight into the degree of heaping. While the simple binary outcome captures only whether the reported value falls on a heaping point or not, the relative RI and RSM further regard the multiple on which a particular true value falls and its magnitude, hence measuring the intensity. Relying on the assumption of ordinality, both measures are tested against the predictors in an ordered probit model. The combined estimates of both models are not so different from the results of the binary model (see Table A.1 and Table A.2 in the Appendix). The marginal effects of both models are plotted in Figure 1.9 and Figure 1.10. Increasing the predictor (e.g. age) by one unit, increases (or decreases) the probability of selecting an alternative category. The marginal effect is usually expressed as percentage to simplify interpretation. In the RSM model, the marginal effects of category 5 are somewhat outstanding from the remaining impression. This might be largely due to the small proportion of observations in this particular category. When excluding this category for interpretation, the monotone trends in the marginal effects become obvious for all predictors. The trends in the relative RI are almost linear.

[18] The excess in zeros is omitted before estimation again.

[19] For calculation of the AIC in regression with multiply imputed data, see Chaurasia and Harel (2012, p. 5).

Though, be aware that the value range of the relative RI is still ordinal which leads to a disproportional X-axis in Figure 1.9. Only for male or higher educated respondents a small peak is discernible at the 0.50 category. All marginal effects are highly significant (p-value < 0.01) with except of those for middle educated respondents.

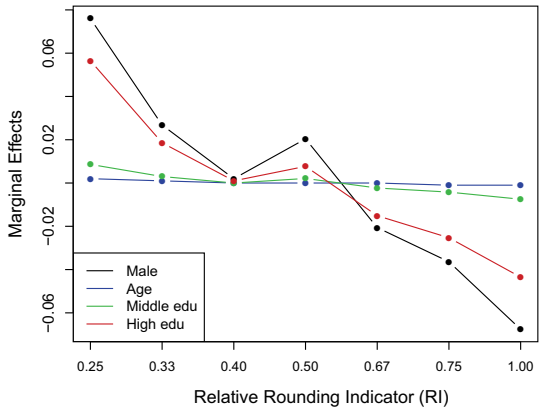


Figure 1.9: Marginal effects from ordered probit regression for the relative RI.

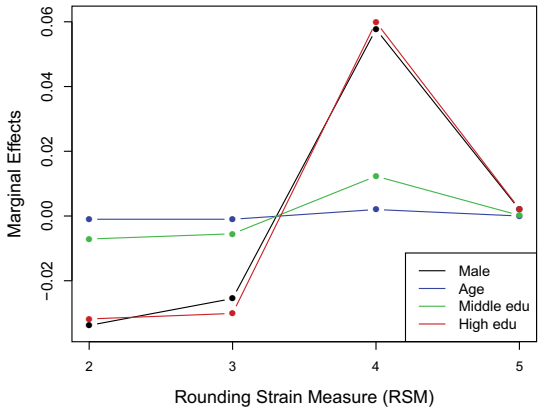


Figure 1.10: Marginal effects from ordered probit regression for the RSM.

Summary

In this section, further empirical evidence on heaping behavior of respondents in survey studies was provided. Analyses of the NEPS SC6 net income data strongly support the assumption that heaping behavior is not stochastic but deterministic. This means that whether and to which degree a true value is heaped is not random and the true values do not fall randomly on eligible heaping points. In contrast, the probabilities for being heaped and to fall on a particular heaping point are influenced by the true value itself (its range and magnitude) as well as internal factors. Significant relationships exist between heaping behavior and common socio-economic characteristics of the respondent. Male, higher educated, and older respondents are more likely to heap their income. When considering an ordered measure for the heaping intensity, a monotone, almost linear pattern was found for all three characteristics considered. The intensity of heaping is mainly determined by the magnitude of the true response value (level dependency).

These findings are important in three different ways. First, they illustrate the necessity to correct for heaping behavior. Second, the modeling of heaping behavior inevitably should be flexible for different subgroups. Third, the knowledge from the regression models is transferred to the data generating process. Strictly speaking, the estimates from the log-linear model are used for an alternative multivariate modeling of income data. The estimates from the probit regression are used for simulation of an accordant heaping mechanism. The simulated data will be used to elicit the performance of the model introduced in the following chapter.

Chapter 2

Modeling heaped income data

This chapter is dedicated to the description of the general heaping model and mainly follows Zinn and Würbach (2014, 2015). The authors establish a finite mixture model to account for different heaping behavior prevalent in self-reported income data. The method is an adaption of the procedure of van der Laan and Kuijvenhoven (2011) but allows for a more general and flexible likelihood. The proposed model is composed of two parts that can be estimated simultaneously. One part concerns the latent distribution of the true values and the other specifies the heaping mechanism operating on the data. Concretely, the latent model describes the distribution function of the non-heaped variable of interest – the true income being reported correctly –, whereas the heaping mechanism works on top of this by inducing shifts of the true income values to specific heaping points. Formalization of such a heaping behavior requires a priori specification of the heaping pattern – the set of heaping points and the corresponding heaping intervals – as well as the heaping mechanism – a function which quantifies the probabilities to heap to the heaping points. Both parts of the heaping model are described in the following subsections. The established model together with the findings from real data are then utilized to set up the data generating process.

2.1 Latent distribution of true income values

Since heaping can occur in discrete as well as continuous numerical data, various theoretical distributional assumptions can be taken into consideration for modeling the latent variable of interest. Recommendations for which distribution to choose can be found in the literature or can be deduced from the data at hand. The eligible theoretical distribution should describe the variable of interest well with respect to the shape and scale of the empirical distribution. The complexity of the model (e.g. the number of parameters) should also be considered carefully. Distributions of high complexity can complicate the estimation process substantially. Usually, it is advisable to fit diverse distributions and to select this one which fits best to the real data.

In the literature, several distributional assumptions exist for modeling income data in particular. The most important 2-parametric distributions are the exponential, the log-normal, and the Pareto distribution (according to Vilfredo Pareto), see Fahrmeir, Künstler, Pigeot, and Tutz (2007, p. 301) and Kleiber and Kotz (2003a). Among the 3-parametric distributions, the Dagum distribution (Dagum, 1977) and the Singh-Maddala's distribution (Singh & Maddala, 1976) are the most common ones, see McDonald (1984). The Generalized Beta distributions I and II (GB1, GB2) as well as the double-Pareto-log-normal are examples of 4-parametric distributions. Especially the last one is asserted to be favorable according to its superior fit to real data, see McDonald, Sorensen, and Turley (2013, p. 361). Besides the plain distributions, also composite distributions are regarded to model income. That is, the exponential or the 2-parameter log-normal distribution are considered for the lower and middle income ranges, whereas the Pareto functional form is used to model higher income ranges, usually the upper quantile, see Harrison (1981, p. 628) or Clementi and Gallegati (2005). Cowell (2000, p. 146) further extends this modeling strategy by using the log-normal distribution to approximate income in the lower tail, the gamma distribution approximates the major body in the middle of the distribution, and the Pareto distribution the upper tail. Many more distributional assumptions can be found i.a. in McDonald (1984). How to model a particular quantity is not a finite decision. On the contrary, the distributions enumerated only reflect expert opinions or conventions.

All of these modeling suggestions have in common that income data are modeled as continuous variable with a typical right-skewed shape. Due to its simple structure, the small number of parameters, and because of the fact that covariates can be easily incorporated, the log-normal distribution is taken into further consideration.[1] The log-normal distribution is commonly used in the economic literature, see e.g. Aitchison and Brown (1957), Johnson, Kotz, and Balakrishnan (1994) and Kleiber and Kotz (2003b). Clementi and Gallegati (2005) also show that the log-normal distribution is particularly well suited to describe the (net) income distribution in various countries. The probability density function (*pdf*) of the log-normal distribution is given by

$$f(y) = \frac{1}{y\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(\log y - \mu)^2\right\} \mathbb{I}(y > 0).$$

It is specified only for $y > 0$ as denoted by the indicator function. For further information on the log-normal distribution, see Appendix A.2.1.

The NEPS income data expose a non-negligible proportion of reported values at zero, cp. Section 1.2.1. To adjust for clumping at zero, income data are modeled as a semicontinuous variable with some probability mass at zero. In Zinn and Würbach (2014), the excess of zeros is considered as part of the heaping mechanism. This strategy is not followed here. Aligned with Zinn and Würbach (2015), the zeros are separately counted and included as an inflation parameter

[1] In Section 4.1.1 the Dagum distribution is applied alternatively.

into the latent model. Thus, as latent true distribution, a zero-inflated log-normal distribution is assumed. Such a zero-inflated distribution is a two-part model in which the zeros are modeled separately from the positive values of the continuous distribution (Lin & Tsai, 2013). It can be viewed as type of a latent class model with the two class probabilities, ρ_Z and $1 - \rho_Z$. The accordant density is

$$f(y|\psi) = (1-\rho_Z)\mathbb{I}(y=0) + \rho_Z \frac{1}{y\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(\log y - \mu)^2\right\} \mathbb{I}(y>0), \quad (2.1)$$

with $\psi = [\mu, \sigma, \rho_Z]$. The parameter space of ψ is $\mu, \sigma \in \mathbb{R}^+$ and $\rho_Z \in [0,1]$. The class parameter ρ_Z is one minus the inflation probability (Zinn & Würbach, 2015).

As short reminder (cp. Section 1.1.3 on page 24), $\mathbf{z} = (z_1, \ldots, z_n)$ comprises the observed values with $z_i \in \mathbb{R}_0^+, i = 1, \ldots, n$ for n given observations. The true (and possibly unknown) values corresponding to $\mathbf{z}$ are denoted by $\mathbf{y} = (y_1, \ldots, y_n)$ with $y_i \in \mathbb{R}_0^+, i = 1, \ldots, n$. Finally, $f(y|\psi)$ and $F(y|\psi)$ denote the assumed underlying probability distribution function (*pdf*) of the variable of interest and the corresponding cumulative distribution function (*cdf*).

2.2 Heaping mechanism

The overall heaping behavior can be described by the heaping pattern and the heaping mechanism, cp. Figure 1.2 on page 11. First of all, in this approach (cp. Zinn & Würbach, 2014, 2015), the heaping pattern has to be fixed. For this purpose, the set of numbers preferred when reporting values, the so-called heaping points (HP), has to be known in advance or needs to be identified. The set of HP is described by $\mathcal{H} = \{h_b : h_1, \ldots, h_S\}$, for $h_b \in \mathbb{N}$. Several multiples could be considered as HP. Typical prototypes are, e.g. $10, 50, 100, 500, 1000$, see Huttenlocher et al. (1990, p. 212). Second, the catchment areas for each of the HP, the so-called heaping intervals, have to be determined. Each HP has a certain catchment area from which values can fall on h_b. Hence, a HP cannot pull values from outside its catchment area. This assumption is made because it is implausible to state that all HP attract all possible (true) income values to the same extent. Based on this attraction assumption, the heaping intervals should be narrower for mod(100) than for mod(500), and the intervals of mod(500) should be narrower than those of mod(1000). Besides this differences, the heaping intervals are assumed to be identical for each modulo on the range of income values, e.g. the heaping interval for HP 1100 is considered to equal this of 5700 and so on. The catchment area for h_b is denoted by $I_b = [l_b, u_b]$, where l_b describes the lower and u_b the upper bound of the respective interval.

In addition, the functional form of the heaping mechanism and the corresponding parameters (ϕ) have to be determined. The model captures a respondents' propensity to heap his true income y to heaping point h_b. At this point, it becomes relevant what heaping distinguishes from simple (mathematical) rounding,

cp. Section 1.1.1. Heaping does not occur systematically according to one fixed rule, because it differs by object, by magnitude, and also by covariates. Whereas rounding is systematic – there is one fixed rule that applies across all observations, independent of covariates or the quantity itself. On this account, the heaping mechanism has to allow for different probabilities within the income range but in particular with regard to the modulos, as seen earlier in Section 1.2.1 in Table 1.2. A functional form is necessary that enables a high level of flexibility in modeling the heaping probabilities. According to van der Laan and Kuijvenhoven (2011), these probabilities are assumed to be symmetric within the predefined intervals (I_b). This assumption was adapted for the sake of simplicity. Further possible functions could be of triangle or bell-shaped form, where the propensity to heap is assumed not only to be dependent on the magnitude of income but additionally on the proximity of a true value to a HP, see Zinn and Würbach (2014, 2015).[2]

Zinn and Würbach (2014) call this model with equiprobable probabilities "piecewise constant heaping probabilities" (*pcm*). The model is quantified by the probability function $v_b(y)$. In general, the function $v_b(y)$ resembles a multinomial distribution. Figure 2.1 illustrates the heaping probability function described, and the *pcm* has the following form:

$$v_b(y) = \begin{cases} \rho_b, & \text{if } y \in I_b, \text{ for } y \neq h_b \\ 0, & otherwise. \end{cases} \tag{2.2}$$

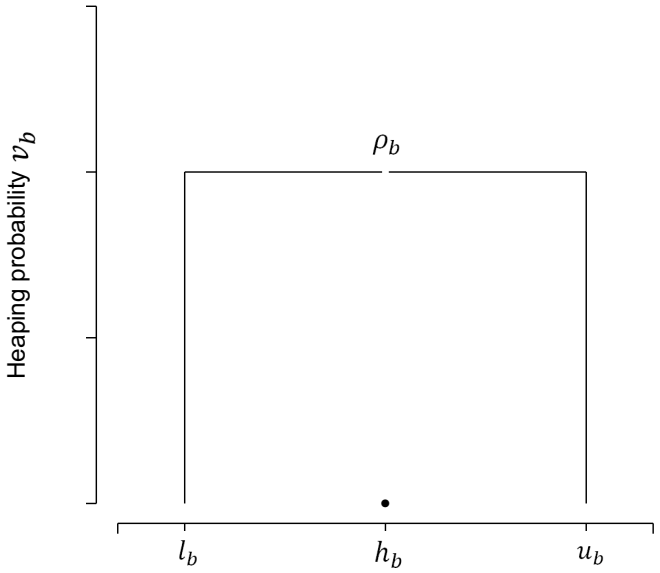


Figure 2.1: Illustration of the piecewise constant heaping mechanism with equal probabilities for heaping.

[2]In Section 4.1.3 the heaping mechanism is adapted accordingly.

Here, ρ_b denotes the constant heaping probability contributing to h_b. The probability of heaping a value y located on heaping point h_b to precisely that HP is zero because there is nothing to heap. Please note that in the given specifications the catchment areas are allowed to overlap, i.e. values can be heaped to more than one heaping point. Concretely, true income values have a positive probability for all possible intervals I_b into which y can fall and zero otherwise. For example, value 1263.14 EUR might be heaped to 1300, 1500, or 1000.

2.3 Constraint system

For estimation purposes, the constraints imposed on the parameter vectors ϕ and ψ have to be taken into account. It must be ensured that the heaping probabilities ρ_b, their sums, and the inflation parameter ρ_Z range between zero and one. Let $\mathcal{C}_\phi$ denote the constraint system for the parameters of the heaping mechanism, ϕ, imposed by the underlying theoretical model (2.2). The following requisites need to be met, cp. Zinn and Würbach (2014):

(i) ρ_Z and $\rho_b \in [0,1]$, for all $b = 1, \ldots, S$,

(ii) $\sum_{b:z_i \in I_b} \rho_b \in [0,1]$ for all $z_1, \ldots, z_n$.

For simplicity, all constraints on the model parameters are summarized with $\mathcal{C}_\Theta$. The constraint system $\mathcal{C}_\Theta$ can be specified in the form of inequality equations which are linear in the parameters of ψ and ϕ. Thus, the optimization problem at hand is a classical linear optimization problem.[3]

2.4 Log-Likelihood

First of all, the likelihood function of observing z_i needs to be constructed in order to estimate the unknown parameter vectors ψ and ϕ of the latent true distribution and the heaping mechanism, see Zinn and Würbach (2014). If the true value y_i is not heaped ($z_i = y_i$), the density of observing z_i is

$$g_1(z_i|\psi,\phi) = [1 - v_b(z_i|\phi)]\, f(z_i|\psi). \tag{2.3}$$

Please note that this definition also accounts for the fact that values located at heaping points might be reported correctly. Otherwise, if z_i falls on a heaping point h_b with interval $I_b = [l_b, u_b]$ and $z_i \neq y_i$, the corresponding density is

$$g_2(z_i|\psi,\phi) = v_b(z_i|\phi)\,[F(u_b|\psi) - F(l_b|\psi)]\,. \tag{2.4}$$

[3] An alternative re-parametrization would also be conceivable to reassure the parameters to lie within the defined parameter space, but this approach was not pursued further here. For example, van der Laan and Kuijvenhoven (2011) used the parametrization $\exp(-\exp(\rho))$, but this particular form is not suitable for the approach presented here owing to the different intervals a true income value can fall into.

In words, the probability of observing a value z_i, which falls on h_b, is determined by the difference between the *cdf* at the upper bound of I_b and the *cdf* at the lower bound of I_b multiplied by the density of heaping its unobserved correspondent y_i to h_b ($y_i \in I_b \setminus h_b$). In the considered case of constant heaping probabilities, $v_b(y_i)$ equals $v_b(z_i)$, and can be simply replaced by ρ_b. Combining the functions g_1 and g_2 yields the (approximated) likelihood function $\mathcal{L}$ of observing z_i:

$$\mathcal{L}(z_i|\psi,\phi) = g(z_i|\psi,\phi) = g_1(z_i|\psi,\phi)\mathbb{I}(z_i \in \mathbb{R}_0^+) + g_2(z_i|\psi,\phi)\mathbb{I}(z_i \in \mathcal{H}), \tag{2.5}$$

where

$$\mathbb{I}(z_i \in \mathbb{R}_0^+) = \begin{cases} 1, & \text{if } z_i \in \mathbb{R}_0^+, \\ 0, & \text{otherwise.} \end{cases}$$

and

$$\mathbb{I}(z_i \in \mathcal{H}) = \begin{cases} 1, & \text{if } z_i \in \mathcal{H}, \\ 0, & \text{otherwise.} \end{cases}$$

In a second step, the logarithmic density of one observation is defined. Let θ now comprise both parameter vectors ψ and ϕ, then

$$\ell(z_i|\theta) = \underbrace{\left(\left[1 - \sum_{b=1}^{S} \rho_b\right] f(z_i|\psi)dz_i\right)\mathbb{I}(z_i \in \mathbb{R}_0^+)}_{\text{non-heaped income values}}$$

$$+ \underbrace{\left(\sum_{b=1}^{S} \rho_b \left[F(u_b|\psi) - F(l_b|\psi)\right]\right)\mathbb{I}(z_i \in \mathcal{H}),}_{\text{heaped income values}} \tag{2.6}$$

with $\ell(\mathbf{z}|\theta) = \sum_{i=1}^{N} \ln g(z_i|\psi,\phi)$ being the log-likelihood function of the complete data. The first part of the log-likelihood function – restricting $z_i \in \mathbb{R}_0^+$ – describes the likelihood for non-heaped values, and the second part – restricting z_i to take values from a discrete indicator in $\mathcal{H} = \{h_b : h_1, \ldots, h_S\}$, for $h_b \in \mathbb{N}$ $\text{mod}(100, 500, 1000)$ – describes the likelihood for heaped values. Maximizing ℓ yields estimates $\widehat{\theta} = (\widehat{\psi}, \widehat{\phi})$ for the parameter vector $\theta = (\psi, \phi)$:

$$\widehat{\theta} = \text{argmax}_\theta \; \ell(\mathbf{z}|\theta).$$

2.5 Specification of the heaping model and data generating process

Simulations are performed to test the validity of the proposed heaping model. The key figures found in Section 1.2.1 are used as orientation to simulate data that closely resemble the real data previously described. First of all, a data set disregarding any covariates is simulated. Income data are generated with a zero-inflated log-normal distribution as latent distribution, and a heaping mechanism is applied assuming piecewise constant heaping probabilities.

Specification of the heaping pattern

Regarding the heaping pattern, the following assumptions are made to specify the HP and the associated I_b. As HP considered are: multiples of 100 up to 5000 and multiples of 500 or 1000 up to 10,000, which yields 60 HP in sum, see Figure 2.2. As far as not stated otherwise, the catchment areas I_b are assumed to be symmetric around the respective HP, while the widths depend on the modulo. One half of the modulo is located below and the other half is located above the HP, $I_b = [h_b - \frac{1}{2}\text{mod}, h_b + \frac{1}{2}\text{mod}]$. To be concrete, catchment areas associated with HP that are e.g. multiples of 100, but not of 500, have width 50. All $h_b \in$ mod(500) feature the interval $I_b = h_b \pm 250$ and lastly, all $h_b \in$ mod(1000) have an interval width of 500, i.e. $I_b = h_b \pm 500$.

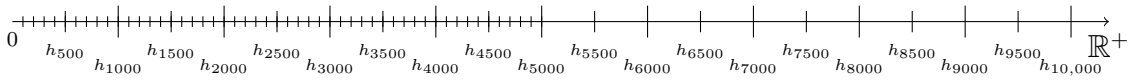


Figure 2.2: Heaping points within the considered income range.

Each of the 60 HP has an associated heaping probability. Defining the heaping probabilities this way leads to a large number of model parameters, 63 in total. These are 60 parameters for the heaping probability function plus three parameters for the underlying true distribution. This is likely to hamper the success and the efficiency of the estimation procedure. To alleviate this problem, Zinn and Würbach (2015) suggest a further restrain in the parameter space. They assume that some components of ϕ are equal. From inspection of the real data (Table 1.8), it can be seen that the propensity to heap strongly depends on the level of income. This issue concerns two points. First, the level of income determines if the true value is going to be heaped at all, i.e. higher income values are more often heaped (94% for income values above 3000 EUR vs. 48% for income values below 1500 EUR). Second, it can be seen that the proportion of heaped values falling on thousands increases remarkably in higher income ranges, whereas the probability to heap values at hundreds increases in the interval $(1500, 3000]$ but decreases again in ranges above 3000 EUR. The overall probabilities differ for each

modulo (cp. Table 1.1). In summary, this means that in the considered case the probabilities to fall on mod(100) are assumed to be higher in the lower to middle range intervals, whereas the probabilities to fall on mod(1000) are assumed to be smaller, and vice versa. Last but not least, the probabilities to fall on mod(500) should be the lowest overall. When taking these findings into account, it is highly plausible to assume a congenial heaping behavior for each of the three modulos within certain intervals of the whole income range to set equality constraints on the parameters of the heaping mechanism. Of course, the division into intervals is a trade-off between complexity reduction and closeness to reality. It seems appropriate to break down intervals in the lower income ranges more finely, whereas the higher income ranges should be more roughly subdivided. Such processing is advisable, because in such long-tailed distributions most of the distributional mass lies in the lower to middle income ranges.

According to this, the range of values is divided into eight intervals $[0, 500]$, $(500, 1000]$, $(1000, 1500]$, $(1500, 2000]$, $(2000, 3000]$, $(3000, 4000]$, $(4000, 5000]$, and $(5000, 10{,}000]$ within which the probabilities of heaping to a multiple of 100, 500, and 1000 are assumed to be identical (cp. Zinn & Würbach, 2015). Table 2.1 gives the respective sets that result from grouping the heaping probabilities that way.

Table 2.1: Sets of heaping probabilities for mod(100), mod(500), and for mod(1000).

Interval	mod(100)	mod(500)	mod(1000)
$[0, 500]$	Set 1	Set 8	–
$(500, 1000]$	Set 2	–	Set 14
$(1000, 1500]$	Set 3	Set 9	–
$(1500, 2000]$	Set 4	–	Set 15
$(2000, 3000]$	Set 5	Set 10	Set 16
$(3000, 4000]$	Set 6	Set 11	Set 17
$(4000, 5000]$	Set 7	Set 12	Set 18
$(5000, 10{,}000]$	–	Set 13	Set 19

It should be noted that in some intervals for particular multiples no heaping points exist by definition. For example, in the interval $(500, 1000]$ no multiple of 500 exists that is not a multiple of 1000 at the same time. Introducing equality constraints on the heaping probabilities as described reduces the number of parameters remarkably. Consequently, only 19 estimates are necessary to determine the heaping probabilities. The number of the set is used as identifier for the corresponding heaping probability, e.g. ρ_2 denotes the probability to heap to multiples of 100 but not of 500 within the interval $(500, 1000]$. Three further parameters have to be estimated to fully identify the zero-inflated log-normal distribution. This sums up to an overall number of model parameters of 22.

Univariate consideration of income data

From the zero-inflated log-normal distribution, a sample with $N = 10{,}000$ *iid* random values is drawn, parameterized with $\mu = 7.72$, $\sigma^2 = 0.85^2$, and $\rho_Z = 0.987$. The heaping mechanism working on top of the latent distribution assumes equiprobable heaping probabilities within predefined intervals. In order to obtain a data set which roughly resembles the structure of the NEPS income data (cp. Figure 1.5), the true values are shifted to the HP according to the heaping probabilities given in Table 2.2. To ease estimation in a first place, a lower proportion of heaped values is assumed – about 50% compared to approx. 69% in the NEPS income data. Figure 2.3 depicts the income distribution of the data example of simulation model one (Model I). In sum, 49.82% of the values are heaped and 1.33% of the values fall on zero. The mean of the heaped distribution is 2713.16 EUR ($SD = 2045.97$ EUR) and the median is 2072.68 EUR. Table 2.3 shows how many values are heaped to multiples of 100, 500, and 1000.

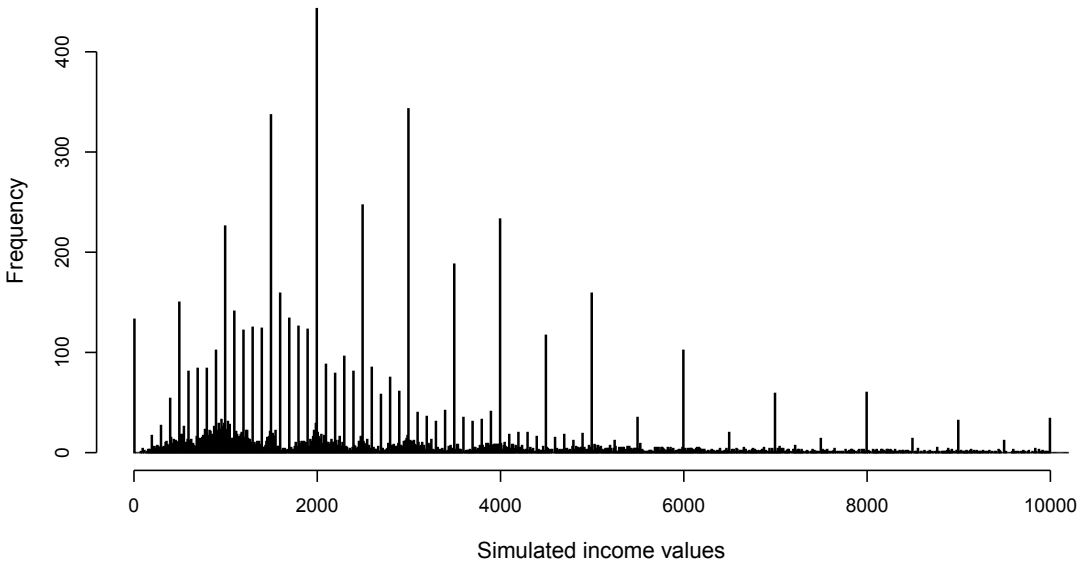


Figure 2.3: Data example of simulation model one (Model I).

Table 2.2: Heaping probabilities in Model I.

Interval	mod(100)	mod(500)	mod(1000)
$[0, 500]$	0.25	0.17	–
$(500, 1000]$	0.26	–	0.08
$(1000, 1500]$	0.39	0.24	–
$(1500, 2000]$	0.46	–	0.18
$(2000, 3000]$	0.37	0.27	0.22
$(3000, 4000]$	0.28	0.32	0.24
$(4000, 5000]$	0.19	0.30	0.25
$(5000, 10{,}000]$	–	0.17	0.26

Table 2.3: Percentages of heaped values in the data example of Model I.

Interval	mod(100)	mod(500)	mod(1000)	Total
$[0, 500]$	0.90	1.34	–	2.24
$(500, 1000]$	2.70	–	1.98	4.68
$(1000, 1500]$	4.63	3.23	–	7.86
$(1500, 2000]$	5.22	–	4.26	9.48
$(2000, 3000]$	5.68	2.36	3.30	11.34
$(3000, 4000]$	2.63	1.85	2.25	6.73
$(4000, 5000]$	1.08	1.17	1.54	3.79
$(5000, 10{,}000]$	–	0.91	2.79	3.70
Total	22.84	10.86	16.12	49.82

Chapter 3 elicits the feasibility and effectiveness of the proposed model and the accordant specifications made in this chapter. If the model is capable to accurately describe the simulated data, it is applied to real data. The data example of Model I is estimated by a maximum likelihood (ML) approach as well as a Bayesian approach. This dual strategy is employed since the likelihood of the given heaping model is assumed to suffer from multi-modality. Because of the peculiarities of the heaping model, e.g. multiple heaping probabilities and possibly overlapping heaping intervals, the assumption of multi-modality seems advisable. In case of multi-modality, ML approaches are known to weaken, since they are at risk of sticking at local modes. Whereas estimation by means of the random-walk Metropolis (RWM) algorithm is more likely to explore the whole parameter space. By design, the RWM algorithm makes random steps at each iteration and is therefore less likely to remain at a certain point (or mode). That is, in the presence of multi-modality, Bayesian estimation via RWM algorithm is expected to outperform the ML approach. Furthermore, the efficiency of the RWM algorithm will be checked. For this purpose, various specifications of the RWM algorithm are examined and the results of those settings turning out to be the best with regard to their mixing behavior and accuracy are directly compared to the results of the ML approach.

Chapter 3

Estimation of the heaping model

The parameters of the proposed mixture model for heaped income data are estimated using a frequentist approach and different Bayesian methods. For this purpose, a maximum likelihood (ML) technique is employed on the one hand, and on the other hand different random-walk Metropolis (RWM) algorithms are run. Besides the original RWM algorithm, blocking and updating strategies for sampling the proposal density are compared. The RWM schemes that perform best with respect to estimation accuracy and numerical efficiency will be considered for direct comparison with the ML approach.

3.1 Frequentist estimation of the heaping model

First, the heaping model is estimated by a maximum likelihood (ML) approach referring to the data example of simulation model one described thoroughly in the previous section (cp. Section 2.5).

3.1.1 Maximum Likelihood with constraints

The constraint system introduced in Section 2.3 has to be considered for estimation of the model parameters, i.e. the heaping probabilities as well as the parameters of the underlying zero-inflated log-normal distribution. Zinn and Würbach (2015) suggest to solve this linear optimization problem with inequality constraints by using the *Nelder-Mead algorithm*, see S. Wolff (2004). The *constrained and simple bounded Nelder-Mead* (BNM) is a method for direct optimization and is particularly suitable for solving nonlinear and discontinuous local optimization problems. The BNM bases on the Nelder-Mead, or downhill simplex, method which is a commonly used nonlinear optimization technique already proposed by Nelder and Mead (1965). It is an intuitive and well-defined numerical method for problems for which derivatives may not be known. It is relatively stable and approaches the optimum in great steps at the beginning of the heuristic search.

The *maxNM* function, which is part of the `maxLik` package (Henningsen & Toomet, 2011) of the statistical software `R`, implements the approach.[1] Along with the maximum likelihood estimates $\widehat{\theta}_{ML} = (\widehat{\psi}_{ML}, \widehat{\phi}_{ML})$, robust standard errors (SE) derived by means of the Huber-White sandwich estimator[2] (Kleiber & Zeileis, 2008, p. 136f.), and the respective 95% confidence intervals (CI) are calculated, cp. Zinn and Würbach (2015).

3.1.2 Specification and results of ML estimation

The first step requires definition of the initial values for triggering ML estimation. Initial values for the parameters of the latent distribution are gathered by fitting the values of the observed data to an ordinary log-normal distribution, cp. Zinn and Würbach (2015). When deleting zeros before fitting the model and disregarding any heaping behavior, the estimates are $\widehat{\mu} = 7.714$ and $\widehat{\sigma}^2 = 0.839^2$ with the corresponding SE 0.008 and 0.006. The initial values of the heaping probabilities are arbitrarily set to 0.2 and the initial value of the inflation parameter is set to 0.99. Other sets of initial values have been tested and yield similar results.

In a second step, the model parameters are estimated by ML with the BNM method (see corresponding `R` Code in Appendix B.1). On a desktop workstation equipped with Intel(R) Core(TM) i5-4570, CPU 3.20GHz, 8GB RAM, under Windows 7 using a 64bit system, model fitting takes approx. 45 minutes. Table 3.1 and Figure 3.1 show the ML estimates, their (robust) SE, and the respective 95% CI. As can be seen, the parameter estimates, SE, and CI are all in all very reasonable. The estimates of the parameters of the zero-inflated log-normal distribution are very precise and the estimates of the heaping probabilities are reasonably close to the true ones. Overestimation and the large SE of the estimate of heaping probability ρ_1 (cp. Figure 3.1) result from only few observations on the related heaping points. The estimate of ρ_{10} differs significantly from the true heaping probability albeit being not so far from the true one. However, in line with Zinn and Würbach (2015), the outcome of this simulation example can be assessed as being very good and as a solid evidence for the functionality of the method considering the high percentage of heaped values and the fact that heaping intervals are allowed to overlap.

In a last step, the model that accounts for heaping is tested against the model that does not account for. The *Akaike information criterion* (AIC) is used to check the fit of the models to the data. The accordant values of the AIC are 176,760.3 for disregarding heaping and 140,311.4 when accounting for heaping. The lower AIC of the model that regards heaping indicates a better fit to the data. Thus, simply ignoring heaping patterns when analyzing the data might lead to erroneous results. Therefore, an approach that explicitly accounts for heaping behavior can strongly be advocated.

[1] In Appendix B.3 all `R` packages being used in the course of this thesis are listed.

[2] The Huber-White sandwich estimator allows for deriving heteroscedasticity-consistent SE.

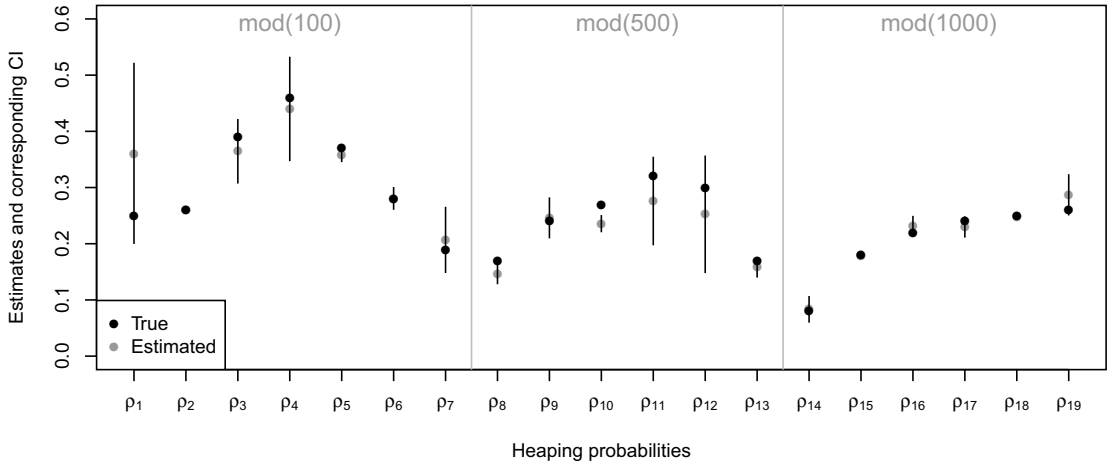


Figure 3.1: ML estimates with 95% confidence intervals for the data example of Model I.

Table 3.1: ML estimates and measures of uncertainty (standard errors SE and 95% confidence intervals CI) for the data example of Model I.

Par	θ	$\widehat{\theta}_{ML}$	SE	CI lower	CI upper
ρ_1	0.250	0.361	0.007	0.200	0.522
ρ_2	0.260	0.260	<0.001	0.259	0.260
ρ_3	0.390	0.365	0.001	0.307	0.422
ρ_4	0.460	0.440	0.002	0.347	0.533
ρ_5	0.370	0.359	<0.001	0.345	0.373
ρ_6	0.280	0.281	<0.001	0.260	0.301
ρ_7	0.190	0.207	0.001	0.148	0.266
ρ_8	0.170	0.147	<0.001	0.128	0.166
ρ_9	0.240	0.246	<0.001	0.210	0.282
ρ_{10}	0.270	0.236	<0.001	0.220	0.251
ρ_{11}	0.320	0.276	0.002	0.197	0.355
ρ_{12}	0.300	0.252	0.003	0.148	0.357
ρ_{13}	0.170	0.158	<0.001	0.140	0.176
ρ_{14}	0.080	0.083	<0.001	0.060	0.107
ρ_{15}	0.180	0.178	<0.001	0.177	0.179
ρ_{16}	0.220	0.232	<0.001	0.214	0.250
ρ_{17}	0.240	0.230	<0.001	0.211	0.249
ρ_{18}	0.250	0.248	<0.001	0.246	0.251
ρ_{19}	0.260	0.287	<0.001	0.250	0.324
μ	7.720	7.726	<0.001	7.697	7.756
σ	0.850	0.844	<0.001	0.835	0.854
ρ_Z	0.987	0.986	<0.001	0.983	0.988

3.2 Bayesian estimation of the heaping model

In this chapter, some Markov Chain Monte Carlo ($MCMC$) algorithms are suggested for Bayesian estimation of the heaping model (Model I). Ahead, some general words on Bayesian estimation are given. The main appeal of a Bayesian estimation scheme is the relative ease of implementation once the likelihood is built. Another great benefit of a Bayesian estimation approach is that it enables researchers to incorporate substantive information about the parameters through the prior distribution. This is important when parameters seem to be or can be shown to be ill-determined – or even unreasonable – when fit by maximum likelihood. A further advantage is the applicability to small data sets utilized for fitting. These facts and the growing availability of powerful $MCMC$ simulation methods, which provide the technology for sampling the posterior distribution of the parameters, and the corresponding improvement of computational power enhance the attractiveness of Bayesian methods remarkably, see Chib and Greenberg (1995), Chib and Jeliazkov (2001), and Chib (2009). Especially complex problems like in the present case – models with finite mixture distribution, non-standard and multi-modal likelihoods – can be decomposed into a sequence of smaller problems which are easier to solve, see Robert and Casella (2010, p. 168) and Hastings (1970, p. 97).

In the frequentist framework, data are considered as random and the model parameters are fixed, whereas in the Bayesian context model parameters are treated as random variables conditioned on the observed data. Thus, information about the model parameters is obtained from their *posterior distribution* $p(\theta|\mathbf{y})$, where the posterior distribution is calculated as the product of the likelihood function $p(\mathbf{y}|\theta)$ and the *prior distribution* $p(\theta)$ up to a normalizing constant

$$p(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta)p(\theta).$$

Both terms are proportional to each other as functions of θ. For equality of both terms, the posterior distribution is further divided by some constant of proportionality that could be computed as $p(\mathbf{y}) = \int_\Omega p(\mathbf{y}|\theta)p(\theta)d\theta$, see Hoff (2009, p. 32). Hence, the term on the right side has some normalizing constant, $C = 1/p(\mathbf{y})$.

Let *target distribution* denote the true (latent) distribution of the model parameters. Then, the posterior distribution without normalizing constant can be regarded as a close approximation to the target distribution. Hence, summaries of the draws from the posterior distribution (e.g. the posterior mean) are simulation-consistent estimates of the model parameters, see Jackman (2009, p. 134). In the illustrated case with an assumed finite mixture distribution as likelihood, the posterior summaries can not be obtained in a straightforward way. At this point, $MCMC$ methods are suggested to draw sample variates from the posterior distribution in order to find the posterior mean. In generating a Markov chain, whose invariant density is the specified target density, sample draws represent a (correlated) sample from the posterior density.

In the following section, some introductory words on $MCMC$ methods are given. Afterwards, the random-walk Metropolis (RWM) algorithm is explained followed by the concretization when applied to the heaping model considered. It is shown how this approach can be operationalized, e.g. by splitting the components of high-dimensional targets into k sub-blocks, or by adaptive schemes, respectively.

3.2.1 Introduction to $MCMC$ samplers

Let t denote specific iterations and T the Markov chain sample size. A Markov chain is a stochastic process, a sequence $\{\theta^{(t)}\}_{t=1}^{T}$ of random elements, whose future state is only dependent on the current state

$$p(\theta^{(t)}|\theta^{(t-1)}, \theta^{(t-2)}, \ldots, \theta^{(0)}) = p(\theta^{(t)}|\theta^{(t-1)}).$$

This conditional probability distribution of $\theta^{(t)}$ given $\theta^{(t-1)}$ is called a *transition kernel*, $\mathcal{K}$. As can be seen, Markov chains are independent of the past. This property is called the Markov property. Monte Carlo (MC) as well as Markov Chain Monte Carlo ($MCMC$) are two ways of sampling from the target probability distribution.

In MC simulation, the sampled sequence is a representative of the target distribution, see Hoff (2009, pp. 98ff.). This independence sampler, or Ordinary Monte Carlo (OMC), is the "gold standard" of MC simulation but requires full knowledge about the posterior density (Geyer, 2011, p. 6). When only minor knowledge of the posterior is available, $MCMC$ methods are preferred. $MCMC$ simulations following the scheme of Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller (1953) are called *Metropolis* algorithms, whereas $MCMC$ simulations referring to the generalized algorithm of Hastings (1970) are called *Metropolis-Hastings* algorithms. Special cases of the *Metropolis-Hastings* algorithm are the *Gibbs sampler* developed by Geman and Geman (1984) and the related *data augmentation* approach developed by Tanner and Wong (1987).

The Metropolis-Hastings algorithm (MH) is an important class of $MCMC$ methods, see i.a. Smith, A. F. M. and Roberts (1993) and Gilks, Richardson, and Spiegelhalter (1996). It is in particular suitable for problems where the target density is not known up to a normalizing constant (multi-modality), or does not look like any familiar distribution (no conjugacy), the full conditionals (some or all) do not look like any known distributions (no Gibbs sampling), cp. Hoff (2009, p. 171), or the target consists of more than two parameters (grid approximations are intractable, see Ritter and Tanner (1992) and Lam (2008)). All these aspects apply to the heaping model considered. That is, the MH does not require a closed form of the posterior distribution and no derivations. Thus, it is mathematical less demanding and therefore attractive for non-normal likelihoods with several minor (and possibly equal) modes commonly found in mixture models, see Sherlock (2005) and Sherlock, Fearnhead, and Roberts (2010).

MH algorithms generate *correlated* samples from a Markov chain, as opposed to *independent and identically distributed (iid)* variates from e.g. importance sampling. The existence of a *stationary probability distribution* and the other properties of *MCMC* settings (Robert & Casella, 2010, p. 169) ensure the sequence $\{\theta^{(t)}\}$ to reach its equilibrium,[3] no matter which starting value $\theta^{(0)}$ is selected. Even if the theoretical convergence of Metropolis-Hastings algorithms or Gibbs samplers is almost always guaranteed,[4] practical issues may imply very large convergence times. Even worse, one is lead to belief that convergence is achieved while important aspects of the target distribution are left unexplored (*pseudo-convergence*). In case of pseudo-convergence, only one local maximum is represented in the posterior distribution, although multiple maxima are existing, see Robert and Casella (2010, p. 170) and Geyer (2011, p. 18). Consequently, careful selection and thoughtful adjustment of the tuning parameters in *MCMC* settings is vital to construct an efficient, well-mixing *MCMC* sampler that covers the whole parameter space and does not produce highly autocorrelated draws.

3.2.2 The Metropolis-Hastings algorithm in general and the random-walk Metropolis algorithm in specific

The basic principle of *MH* algorithms is as follows, see e.g. Jackman (2009, pp. 202ff.) and Robert and Casella (2010, p. 170): given a target density $p(\theta|\mathbf{y})$ – the posterior density $p(\theta|\mathbf{y})$ which is proportional to $p(\mathbf{y}|\theta)p(\theta)$ – a transition kernel $\mathcal{K}$ is built that has a stationary distribution $p(\cdot)$. This kernel can be constructed from a defined set of *jumping rules* (proposal densities) that generate a Markov chain on the support of $p(\theta|\mathbf{y})$ so that the limiting distribution of $\{\theta^{(t)}\}$ is $p(\cdot)$. Strictly speaking, because of continuity of the state space, $\theta \in \mathbb{R}^p$, $\{\theta^{(t)}\}$ is a Markov process not a Markov chain. From this follows that the probability of a move to any specific value $\theta \in \Theta$ is 0. Without loosing any relevant properties, a discretized state space is considered for *MCMC*. Thus, the probability of a transition from $\theta^{(t-1)}$ to $\theta^{(t)}$ equals the probability of jumping from one region $\Re \in \Theta$ to another one, see Jackman (2009, p. 173). A proposal density (jumping distribution) q is needed such that the algorithm can jump to any region in Θ where p is positive. Here, q can be almost arbitrary, but it needs to be assured that the ratio $p(\theta)/p(\theta|\mathbf{y})$ is known up to a constant C independent of the *data*, $\mathbf{y}$. The corresponding transition kernel for the probability to move from $\theta^{(t-1)}$ to $\theta^{(t)}$ is:[5]

$$\mathcal{K}(\theta^{(t)}|\theta^{(t-1)}) = q(\theta^{(t)}|\theta^{(t-1)}) \min\left\{1, \frac{p(\theta^{(t)}|\mathbf{y})q(\theta^{(t-1)}|\theta^{(t)})}{p(\theta^{(t-1)}|\mathbf{y})q(\theta^{(t)}|\theta^{(t-1)})}\right\}.$$

[3] The terms stationarity and equilibrium are synonymous, see Geyer (2011, p. 3).

[4] Cases exist that never converge, but those are not discussed in this thesis. See G. O. Roberts and Rosenthal (2001) for more information on improper posterior distributions.

[5] As $\mathcal{K}$ does not cover all possible transitions there is also a probability of rejecting a move. The corresponding probability that the process remains at $\theta^{(t-1)}$ (in the continuous case) is: $p(\theta^{(t-1)}) = 1 - \int_\Theta q(\theta^{(t-1)}, \theta^{(t)})\alpha(\theta^{(t-1)}, \theta^{(t)})d\theta^{(t)}$ (Gilks et al., 1996, p. 54).

The algorithm of the Metropolis-Hastings is given in the following, cp. Jackman (2009, p. 204).[6]

Algorithm 1

Step 1 Initialize $\theta^{(0)} \in \Theta$ and fix the burn-in period (n_0) as well as the Markov chain sample size (T).

Step 2 Sample θ^* from a proposal density $q(\theta^*|\theta^{(t-1)})$.

Step 3 Calculate the Hastings ratio as $\alpha = \min\left\{1, \frac{p(\theta^*|\mathbf{y})q(\theta^{(t-1)}|\theta^*)}{p(\theta^{(t-1)}|\mathbf{y})q(\theta^*|\theta^{(t-1)})}\right\}$.

Step 4 Sample a random number u from $\mathcal{U}(0,1)$.

Step 5 Rejection decision: accept the candidate draw θ^* as current draw $\theta^{(t)}$, if $u \leq \alpha$, otherwise, keep $\theta^{(t-1)}$ as the current draw $\theta^{(t)}$.

Step 6 Repeat steps 2–5 $n_0 + T$ times, discard the draws from the burn-in phase (the first n_0 iterations), and store the following T draws $[\theta^{(n_0+1)}, \ldots, \theta^{(n_0+T)}]$.

In a first step, initial values for θ are set and the number of iterations is fixed, including those iterations discarded before analysis (Step 1). Following Step 2, proposal values for θ are drawn. Let θ^* denote the proposed draw (*candidate*) at iteration t and let α be the acceptance ratio (*Hastings ratio*), i.e. the plausibility of the candidate point θ^* as new draw (*current value*). The candidate draw is always accepted when the ratio is larger than one. Then, the chain is closer to a local mode and the probability of θ^* is higher than the probability of the current draw and α will equal 1. When the probability of θ^* is lower, α is compared with a standard uniform distributed random number. If u exceeds α, the proposed values are rejected and the current values are retained as the new values $\theta^{(t)}$. The so called Hastings update (Step 3) keeps the Markov chain in the main posterior mass most of the time, cp. Sherlock et al. (2010, p. 2). Finally, Steps 2–5 are repeated $n_0 + T$ times where n_0 is the number of iterations considered as burn-in and T is the Markov chain sample size.

In the Hastings ratio, the evaluations of the draws are weighted at the posterior densities, cp. Hastings (1970, p. 98) and Hoff (2009, p. 183). However, the ratio simplifies to the *Metropolis ratio* when the jumping distribution is symmetric: $\min\left\{1, \alpha = \frac{p(\theta^*|\mathbf{y})}{p(\theta^{(t-1)}|\mathbf{y})}\right\}$. Since $q(\theta^*|\theta^{(t-1)}) = q(\theta^{(t-1)}|\theta^*)$, q will cancel out, indicating independence of the acceptance probability from q, see Hastings (1970, p. 98). The symmetry property can be ensured by proposals of the general form $\theta^{(t)} = \theta^{(t-1)} + \epsilon$, where ϵ is a random perturbation stochastically independent of $\theta^{(t-1)}$. This can be either an uniform distribution, the so called *uniform-in-each-direction* proposal density, see Hoff (2009, p. 175) and Jackman (2009, p. 204), or

[6] For a more general notation of the algorithm see Gilks et al. (1996, p. 54).

a normal distribution.[7] Both are popular proposal densities referred to as *random-walk* proposals. If the construction of the proposal is complicated, it is helpful to gather information about the target stepwise by a local exploration of the neighborhood of the current value of the Markov chain. This can be achieved when using the current value either as the center of the uniform distribution, or as the mean in the multivariate normal distribution with Σ_q being the covariance matrix of the proposal distribution, see Robert and Casella (2010, p. 182) or Hoff (2009, p. 175), respectively:

$$q\left(\theta^*|\theta^{(t-1)}\right) \sim \mathcal{U}\left(\theta^{(t-1)} - \epsilon, \theta^{(t-1)} + \epsilon\right)$$

$$q\left(\theta^*|\theta^{(t-1)}\right) \sim \mathcal{N}\left(\theta^{(t-1)}, \Sigma_q\right).$$

The general idea of $MCMC$ methods, and random-walk Metropolis (RWM) algorithms in particular, is easy to grasp and straightforward to implement. Nevertheless, designing a sampler that mixes well and immediately converges to the invariant distribution is often demanding. The setting of the starting values and the choice of the variance of the increment in the RWM proposal distribution are crucial points for the performance of the $MCMC$ sampler, see Robert and Casella (2010, p. 175) and Jackman (2009, p. 205). This especially holds in higher-dimensional problems with many components of θ, see G. O. Roberts, Gelman, and Gilks (1997) and G. O. Roberts and Rosenthal (2001).

Poor starting values can prevent the algorithm to explore the parameter space entirely. Too small variances (dispersion parameters of the proposal density) can slow down the search process. Because small increments require a greater sample size and may hamper the exploration of the parameter space. However, candidate values from a proposal density with small increments are more likely to be accepted. On the contrary, large values of ϵ and on the diagonals of Σ_q might ensure good mixing but, at the same time, might entail many rejections of the candidate values. This is especially problematic in cases with constrained domains, see Robert and Casella (2010, p. 185). Then, a considerable proportion of drawn values does not change over many iterations, leading to high serial correlations and slow convergence to the target distribution, see Chib and Ramamurthy (2010, p. 21) and Robert and Casella (2010, p. 175).

There is a plenty of literature on the appropriate scaling of the proposal distribution for efficient mixing, see e.g. G. O. Roberts et al. (1997, p. 110) and G. O. Roberts and Rosenthal (2001). Gelman, Roberts, and Gilks (1996, p. 605) recommend an under-dispersed proposal distribution in higher dimensional sampling by using a smaller dispersion compared to the target density. An alternative option is to initially specify an identity matrix with small-scale diagonals (e.g. 0.001) and then to update the covariance matrix of the proposal density in a

[7] Alternatively, for each component of θ ($\theta_d : d = 1, \ldots, D$) different distributions can be assumed and also other distributions are conceivable, e.g. a finite mixture distribution $f(\epsilon)$, see Jackman (2009, p. 204).

manner of an *adaptive MCMC*, see Sherlock et al. (2010, p. 12), S. Brooks, Gelman, Jones, and Meng (2011, pp. 93ff.) as well as Haario et al. (1999) and Haario et al. (2001). Done repeatedly while processing, an adaptive $MCMC$ can result in a more desirable proposal covariance matrix, because it considers the dependence structure of the parameters appropriately.

Output of the Metropolis-Hastings algorithm

$MCMC$ samplers generate a dependent sequence $[\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(T)}]$ whose empirical distribution converges under mild regularity conditions[8] for $T \to \infty$ to the posterior distribution $p(\theta|\mathbf{y})$, see Hoff (2009, p. 177). Thus, the expected values of any measurable function h of θ can be approximated using the empirical distribution of $\{\theta^{(t)}\}_{t=1}^{T}$,

$$\bar{h}_T = \frac{1}{T}\sum_{t=1}^{T} h(\theta^{(t)}) \to \mathbb{E}[h(\theta)],$$

see Robert and Casella (2010, p. 169). This holds for any quantity of the posterior distribution one might be interested in, e.g. means or quantiles. Likewise, the variance of $\bar{h}_T$ can be approximated by the square of its standard deviation (Hoff, 2009, p. 54):

$$\frac{1}{T-1}\sum_{t=1}^{T}(h(\theta^{(t)}) - \bar{h}_T)^2 \to Var[h(\theta)].$$

As $\theta^{(t)} \in \Re^D$, with D equalling the dimension of the target (i.e. the number of parameters in the model), the output of a $MCMC$ sampler can be summarized by a matrix $\theta_{T\times D}$. After running the Markov chain, averages and SE of $\theta_{T\times D}$ can be built, see Jackman (2009, p. 192).

Two kinds of adjustments might be applied to the output of $MCMC$ samplers. The first one is to consider some *burn-in* period to mitigate the effect of initialization.[9] The algorithm is run until iteration number n_0 is reached for which it looks like the Markov chain has achieved stationarity. Afterwards, the algorithm runs T more times, generating the sequence $[\theta^{(n_0+1)}, \ldots, \theta^{(n_0+T)}]$. The initial draws $[\theta^{(1)}, \ldots, \theta^{(n_0)}]$ are discarded and the empirical distribution of $[\theta^{(n_0+1)}, \ldots, \theta^{(n_0+T)}]$ is used to approximate $p(\theta|\mathbf{y})$, see Geyer (1992).

[8] Properties of Markov chains are summarized in the Ergodic Theorem. Besides the existence of a stationary distribution (stationarity), four conditions need to be met for $MCMC$ samplers: irreducibility, recurrence, reversibility, and aperiodicity, cp. Jackman (2009, pp. 176ff.). Irreducibility assures that a Markov chain can get from any current state to any region $\Re \in \Theta$ with positive probability. When a Markov chain is said to be recurrent then the chain is allowed to visit a previous state infinitely many times. If the Markov chain possesses detailed balance it is said to be reversible. The last requirement aperiodicity bases on the irreducibility condition. Given irreducibility, a Markov chain is aperiodic if each state in Θ can be visited at each iteration.

[9] When the Markov chain is started near the center of the equilibrium distribution, e.g. at the mode found by optimization, no burn-in period is needed (Geyer, 2011, p. 19f.).

Second, *thinning* can be regarded. In case of slow-mixing, when the chain sticks to current values over many iterations, thinning reduces the output by using only every m-th sample (for $m \geq 1$), while discarding all other draws. This reduces autocorrelation but necessitates large $MCMC$ sample sizes, see Geyer (1992).

3.2.3 Tuning of the original RWM algorithm

For analytically tractable target distributions, e.g. of a conjugate Gibbs Sampler, the optimal parametrization (tuning) can be determined in advance. However, this does not apply here, because the posterior distribution of the heaping model does not follow any known distribution. Tuning parameters refer to the size and shape of the proposal distribution, to different starting values for the RWM chain, and to different prior specifications. Common choices for competing shapes of the proposal distribution are the uniform and multivariate normal distribution. The size is varied by means of different variance values, and thus, different widths for exploration of the parameter space. A possible way to find acceptable tuning parameters are preliminary experiments, cp. Gilks, Roberts, and Sahu (1998, p. 1045). The specifications and parametrization of a RWM algorithm can exhibit remarkable effects on its outcome. Besides basic specifications, such as the definition of the chain size, a proper burn-in period, or thinning, a number of further adjustments is available for tuning of RWM algorithms. A brief sensitivity analysis allows assessing the performance and effectiveness of a RWM algorithm with respect to these parameters. The performance of different parameter settings can be evaluated by means of convergence diagnostics as well as by the efficiency of the $MCMC$ sampler. The accordant quantities are elaborated in Section 3.2.6.

3.2.4 Different blocking strategies in the RWM algorithm

Usually, model parameters are sampled at once in each iteration. This is the so called "simple" or single-block RWM algorithm which will be abbreviated by S-RWM. However, in complex models with many model parameters, sampling in one block might constitute a slow-mixing sampler that yields highly autocorrelated draws, see Chib and Greenberg (1995) and Chib and Ramamurthy (2010). Furthermore, the parameters of a specific model possibly show a clear clustered structure. At this point, it is highly plausible to categorize the parameters into blocks. Regarding the heaping model, the parameters are separated into clear-cut blocks following a natural blocking strategy, either corresponding to the modulo (M-RWM), or the interval (I-RWM), respectively. This blocking of model parameters is expected to increase performance and efficiency of the RWM algorithm remarkably. Besides clustering of the model parameters into fixed blocks, in this thesis a further strategy is employed where clustering is performed randomly. At each iteration, the blocks are newly constructed with varying sizes and composition. The last scheme is henceforth referred to as RMB-RWM. The objective is, whether the performance and efficiency of the RWM algorithm can be additionally

increased that way. At least, the randomized block scheme facilitates the substantive considerations about how to block, making sophisticated a priori choices of blocks obsolete.

All schemes are described thoroughly in the following paragraphs. It should be noted that the parameter estimates of the different RWM schemes are all equal by theory, since the posterior distribution remains the same. Hence, the sampled draws of all schemes are expected to reach posterior summaries that are (approx.) similar.

Simple random-walk Metropolis algorithm (S-RWM)

The S-RWM algorithm draws random variates from the posterior density in one step for all the parameters using one proposal density for all parameter values given. This is achieved by sampling iteratively values on the basis of a proposal density in one single-block. A candidate value is always accepted as the current value, $\theta^* = \theta^{(t)}$, when α is 1. For $\alpha < 1$, the candidate draw is accepted when the accordant acceptance rate exceeds a random variate from the standard uniform distribution. Otherwise, the current draw is maintained, $\theta^{(t-1)} = \theta^{(t)}$. Here, the acceptance rate is computed as the ratio of the posterior density at the candidate values and the posterior density at the current values. This step refers to the Metropolis update for symmetric proposal densities. According to Chib and Ramamurthy (2010, p. 20), substantial pre-run tuning effort is often inevitable for single-block schemes, especially in higher-dimensional problems. Otherwise, large $MCMC$ sample sizes or judiciously selected starting values are necessary to ensure that the chain starts near the center of the target distribution.

Multiple-block random-walk Metropolis algorithm (MB-RWM)

In multiple-block RWM schemes (MB-RWM), the parameters of the considered model are grouped into several distinct blocks. Each block of parameters is updated in sequence by a Metropolis step conditioned on the current value of the parameters in the remaining blocks, cp. Chib and Ramamurthy (2010, p. 20). Chib and Greenberg (1995) suggest multiple-block RWM schemes to overcome some of the shortcomings of the single-block RWM scheme (S-RWM). Compared to the one-big-model-for-everything approach, the modular strategy is more flexible, reliable, and efficient, see Chib and Greenberg (1995). One explanation for the better performance of multiple-block vs. single-block algorithms is the better mixing of the Markov chain. Chib and Ramamurthy (2010) demonstrate the higher efficiency of multiple-block proposals, as opposed to single-block proposals, for DSGE (dynamic stochastic general equilibrium) models. For many known problems, the S-RWM algorithm is often unable to explore the whole parameter space, see e.g. the applications of Haario et al. (1999, 2001) with a non-linear, curved posterior distribution, or multidimensional parameter identification problems, especially with correlated components. This is in particular the case when using

constrained proposal densities, see Chib and Greenberg (1995, p. 329). The MB-RWM schemes suffer from this restriction as well but to a lower extent, since they explore the posterior distribution in a more efficient manner, see Chib and Ramamurthy (2010). A special variant of the multiple-block approach is the *variable-at-a-time* Metropolis-Hastings algorithm, where each component of θ is treated separately in one cycle of the algorithm, see Geyer (2011, p. 25), which is also called *RWM-within-Gibbs* algorithm in Sherlock et al. (2010, p. 3) or G. O. Roberts and Rosenthal (2001, 2007, 2009) and Bai, Roberts, and Rosenthal (2011).

A very important issue in the construction of multiple blocks is their number and composition. A key prerequisite is to form groups in such a way that parameters within a block are more correlated as compared to parameters from other blocks, cp. Chib and Ramamurthy (2010, p. 21). In this thesis, three different multiple-block variants are to be explored. The first two variants arise from the inherent cluster structure of the model parameters. The parameters of the heaping model can be clustered according to modulo or interval. Concretely, blocks arise either on the basis of the number of different types of modulos considered (M-RWM), or on the basis of an interval composition considered appropriate for the range of values at hand (I-RWM). The third variant of the MB-RWM randomizes the number of blocks and the respective assignment of the parameters (RMB-RWM).

In order to apply the multiple-block scheme, the components of θ are split into vector blocks: $[\theta_k, \theta_{-k}]$, where k is the current block to be updated, and $-k$ summarizes all components left out. The number of blocks considered is denoted by K. The proposal density of the MB-RWM can be split into two parts: $q_k\left(\theta_k, \theta_k^*|\theta_{-k}\right)$ and $q_{-k}\left(\theta_{-k}, \theta_{-k}^*|\theta_k\right)$, where the first parts gives the proposal of the current block to be updated, and the second part describes the proposal of the blocks remaining untouched. The respective acceptance rate α is:

$$\alpha\left(\theta_k, \theta_k^*|\theta_{-k}\right) = \min\left\{1, \frac{p(\theta_k^*|\mathbf{y}, \theta_{-k}) q_k(\theta_k^*, \theta_k|\theta_{-k})}{p(\theta_k|\mathbf{y}, \theta_{-k}) q_k(\theta_k, \theta_k^*|\theta_{-k})}\right\}$$

Similarly to the single-block approach, q cancels out when the proposal is symmetric, see e.g. Sherlock et al. (2010, p. 3). The algorithm of the multiple-block scheme is given in the following.

Algorithm 2: MB-RWM algorithm

Step 1 Specify initial values $\theta^{(0)} \in \Theta$ and fix the burn-in period (n_0) as well as the Markov chain sample size (T).

Step 2 Repeat for $t = 1, 2, \ldots, n_0 + T$:

Substep Repeat for $k = 1, \ldots, K$

I Propose a vector of values for the k-th block conditioned on the previous values $\theta_k^{(t-1)}$ and the current values of the other block θ_{-k}:
$\theta_k^* \sim q_k\left(\theta_k^{(t-1)}, \theta_{-k}\right)$.

II Calculate the acceptance rate:
$\alpha_k = \min\left\{1, \frac{p(\theta_k^*|\mathbf{y},\theta_{-k})}{p(\theta_k^{(t-1)}|\mathbf{y},\theta_{-k})}\right\}$.

III Update the k-th block as:
$\theta_k^{(t)} = \begin{cases} \theta_k^* & \text{with probability } \alpha_k \\ \theta_k^{(t-1)} & \text{with probability } 1-\alpha_k. \end{cases}$

Step 3 Store the values $[\theta^{(n_0+1)}, \theta^{(n_0+2)}, \ldots, \theta^{(n_0+T)}]$.

Note that the number of blocks can vary depending on the MB-RWM used. When assigning the parameters of the heaping mechanism according to the modulos (M-RWM), K equals 4. That is, the parameters of the heaping mechanism form three blocks plus an additional block for the parameters of the underlying distribution. When grouping the parameters according to the number of intervals along the income range (I-RWM), K equals 9 (8 intervals plus one block for ψ). The computational time increases by factor K accordingly.

Randomized multiple-block random-walk Metropolis algorithm (RMB-RWM)

The last blocking algorithm of the random-walk Metropolis presented here, the RMB-RWM, follows a Markov Chain Monte Carlo scheme illustrated by Chib and Ramamurthy (2010, p. 22). At each iteration, the parameters of the heaping model are randomly clustered into an arbitrary number of blocks. Within each iteration, each block is sequentially updated through a Metropolis-Hastings step, see Chib and Ramamurthy (2010, p. 23). A great advantage of this blocking strategy is that it overcomes the drawbacks resulting from a poor choice of a priori blocks, since the RMB-RWM scheme allows for a variable grouping of the model parameters. Chib (2009, p. 10) postulate that randomized blocking gives a flexibility at hand that allows to accommodate for different blocking needs owing to model specification. This is accomplished since random blocks by design facilitate capturing different scales and sizes of the parameters that would otherwise require special reasoning when using fixed blocking schemes, see Chib and Ramamurthy (2010, p. 21). For D model parameters, the number of possible blocks (K) can vary between one and D, when either all parameters are grouped into one single block, or each parameter is drawn separately, respectively. The randomized blocking scheme can also be organized in a way allowing for some components of θ to form a fixed block, while the other parameters in θ form random blocks, see Chib (2009, p. 10f).

Algorithm 3: RMB-RWM algorithm

Step 1 Initialize $\theta^{(0)} \in \Theta$ and fix the burn-in phase (n_0) and the Markov chain sample size (T).

Step 2 At each iteration t, $t = 1, \ldots, n_0 + T$ randomly generate K_t blocks $\left(\theta_1^{(t)}, \theta_2^{(t)}, \ldots, \theta_K^{(t)}\right)$.

Step 3 Within each iteration, sample sequentially candidate draws $\theta_{t,k}^*$ for each block $\theta_{t,k}, k = 1, \ldots, K_t$ from proposal density $q_{t,k}$ by a Metropolis step.

Step 4 Repeat steps 2–3 $n_0 + T$ times, discard the draws from the burn-in phase (the first n_0 iterations), and save the subsequent T draws $[\theta^{(n_0+1)}, \ldots, \theta^{(n_0+T)}]$.

Step 2 of Algorithm 3 implies the random construction of parameter blocks of different lengths and composition, $K \sim \mathcal{U}(1, D)$. Cycling through the K_t randomly constructed blocks $(\theta_1^{(t)}, \theta_2^{(t)}, \ldots, \theta_K^{(t)})$, for each block a Metropolis step is performed. The remaining steps equal those of the MB-RWM scheme. At the end of the $(t-1)$st iteration, K_t blocks have been updated.

MCMC samplers with multiple-block proposal distributions show preferable acceptance rates with minimum effort. However, even in case of sophisticated tuning, multiple-block schemes can exhibit slow convergence when the components of $\theta^{(t)}$ are highly correlated, see Haario et al. (2001, p. 232). In such situations, adaptive Gaussian proposal distributions have been suggested to be more promising (Haario et al., 1999, 2001). The next section is dedicated to their description and exploration.

3.2.5 Adaptive *MCMC* for a Gaussian proposal density

Updating strategies, so called adaptive *MCMC* schemes, learn from the output of early iterations. That way, an efficient exploration of the parameter space is aimed at and thus a fast convergence to the target distribution. Especially in situations where the a priori knowledge of the target distribution is quite limited, e.g. with respect to the correlation structure between the parameters, those methods develop their virtues. Generally, adaptive approaches are intended for Gaussian proposals with specification $\theta^* \sim \mathcal{N}(\theta^{(t-1)}, s_D^2 \mathcal{I}_D)$, where s_D^2 is a scale parameter for the covariance matrix of the proposal distribution and $\mathcal{I}_D$ denotes an identity matrix of dimension D.

In 1999, Haario et al. describe an *Adaptive Proposal* (*AP*) algorithm where the proposal distribution is tuned according to the covariance matrix calculated from a fixed number of previous draws. Let R denote the whole history $[\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(t-1)}]$ already being sampled. Conditioned on R, the next candidate value θ^* is sampled from a proposal distribution $q_R(\cdot|\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(t-1)})$ with mean $\frac{1}{t-1} \sum_{\nu=1}^{t-1} \theta^{(\nu)}$ and the covariance matrix that depends on the empirical covariance matrix of R.

The difficulty when applying the AP algorithm, is to specify how and to which part the proposal distribution depends on the history R. Here, Haario et al. (1999) suggest the usage of a *memory* parameter M to determine the update of the covariance matrix. Let at time $t-1$ at least M points being sampled $[\theta^{(1)}, \ldots, \theta^{(t-M)}, \ldots, \theta^{(t-1)}]$ and the M points form a subset of the whole history R. Then, according to Haario et al. (1999, p. 6), the proposal distribution $q_M\left(\theta^{(t)}|\theta^{(t-M)}, \ldots, \theta^{(t-1)}\right)$ is defined as

$$\mathcal{N}\left(\theta^{(t-1)}, s_D^2 \mathbb{C}_M\right),$$

where $\mathbb{C}_M$ denotes the empirical covariance matrix determined by the M points and the scaling factor s_D depends on the dimension of θ, D. According to Gelman, Roberts, and Gilks (1996, p. 603), in the subsequent the scaling factor is set to $s_D = 2.38/\sqrt{D}$. Defining the scaling factor this way is supposed to lead to good mixing properties of the $MCMC$ algorithm, see Haario et al. (1999, p. 7). The matrix $\mathbb{C}_M$ is calculated by means of the $M \times D$-dimensional matrix $\mathbb{M}$ whose rows are made up by the sampled points. Denoting by $\tilde{M}$ the centered matrix $\tilde{M} = \mathbb{M} - \mathbb{E}[\mathbb{M}]$, $\mathbb{C}_M$ is

$$\mathbb{C}_M = \frac{1}{M-1} \tilde{M}' \mathbb{M}.$$

According to Haario et al. (1999, p. 6), the proposal distribution q_M can thus be written as

$$\theta^{(t-1)} + \frac{s_D}{\sqrt{M-1}} \tilde{M}' \mathcal{N}(0, \mathcal{I}_M),$$

with $\mathcal{N}(0, \mathcal{I}_M)$ being the normalized M-dimensional Gaussian distribution. For reasons of performance, the matrix $\mathbb{C}_M$ is only updated at each U-th iteration based on the last M iterations and kept fixed in between. Hence, the parameter U denotes the *update frequency*, see Haario et al. (1999, p. 7). Haario et al. (1999, p. 15) have shown in a sensitivity analysis that the choice of M and U is rather less influential, as opposed to the choice of s_D. Since there is no way to determine M and U properly, both parameters are set to arbitrarily values ($M = U = 1000$). This way, the author of this thesis follows the suggestions of Haario et al. (1999, p. 17f.) that in non-linear problems larger values should be chosen for M and U.

Tests of Haario et al. (1999) show that in their settings the mixing properties of the AP algorithm are superior compared to the non-adaptive RWM algorithms. However, Haario et al. (1999) also identified problems concerning the ergodicity of the chain produced by the AP algorithm, see G. O. Roberts and Rosenthal (2007, 2009) and Bai et al. (2011). Haario et al. (1999) find a biased simulation of the posterior distribution and thus biased $MCMC$ estimates. In favor for the AP algorithm, the authors posit that the differences between the estimated posterior distribution and the true target distribution are negligible when the chain of the AP algorithm explores the parameter space in a meaningful and comprehensive manner, see Haario et al. (1999, p. 18). Furthermore, the authors postulate that the overall accuracy of the AP algorithm is commensurable to the accuracy of a well-tuned, non-adaptive RWM scheme (ibid.).

The *AP* algorithm is presented in the following. In Algorithm 4, a *greedy start* procedure is integrated, cp. Haario et al. (1999, p. 7). This means, the proposal covariance matrix is updated after a short initial period using only the accepted draws for calculating the empirical covariance matrix multiplied by the scaling factor. Cumulating information this way, right at the beginning of the simulation, ensures a rapid start of the adaption process. Thus, already at an early stage of the simulation, the exploration of the parameter space becomes more efficient (cp. Haario et al., 2001, p. 226). The iteration at which the greedy start occurs depends on the number of draws accepted so far. Hence, it can vary from chain to chain. The number of accepted draws used for the greedy start is usually arbitrarily set. However, the number is strongly determined by the overall number of accepted draws, when the chain is run without any adaption. Concretely, it is implausible to assume a greedy start after 100 iterations when the chain mixes quite slowly. In such cases, the adaption process would start at the very end of the chain which would be the opposite of what has been intended.

Algorithm 4: AP-RWM algorithm

Step 1 Initialize $\theta^{(0)} \in \Theta$ and fix the burn-in phase (n_0) as well as the Markov chain sample size (T).

Step 2 At each iteration t, sample sequentially candidate draws θ^* from the proposal density q by a Metropolis step.

Step 3 After 10 accepted draws, update $\mathbb{C}_M$ by a greedy start.

Step 4 Update $\mathbb{C}_M$ at each U-th iteration, based on $[\theta^{(t-M)}, \ldots, \theta^{(t-1)}]$.

Step 5 Repeat steps 2–4 $n_0 + T$ times, discard the draws from the burn-in phase (the first n_0 iterations), and store the following T draws $[\theta^{(n_0+1)}, \ldots, \theta^{(n_0+T)}]$.

In a subsequent article from 2001, Haario et al. propose an alternative *Adaptive Metropolis* (*AM*) algorithm which constructs a chain that has the correct ergodic properties thus providing a correct simulation of the target distribution.[10] In contrast to the *AP* algorithm where the covariance matrix of the proposal distribution is updated on the basis of a fixed number of previous points, in the *AM* algorithm the covariance matrix is calculated using *all* of the previous points $R = [\theta^{(1)}, \ldots, \theta^{(t-1)}]$. The proposal distribution q_R is thus

$$\mathcal{N}(\theta^{(t-1)}, s_D^2 \mathbb{C}_R + s_D^2 c \mathcal{I}_D).$$

[10]The proof for ergodicity in the *AM* algorithm is given in detail in Haario et al. (2001, pp. 225ff.).

For $R+1$ the covariance matrix can be computed using the recursion formula

$$\mathbb{C}_{R+1} = \frac{t-1}{t}\mathbb{C}_R + \frac{s_D^2}{t}\left(t\overline{X}_{t-1}\overline{X}'_{t-1} - (t+1)\overline{X}_t\overline{X}'_t + X_tX'_t + c\mathcal{I}_D\right).$$

Again, s_D is the scaling factor depending only on dimension D of θ. The constant $c > 0$ is chosen to be very small (e.g. 0.001) to ensure that $\mathbb{C}_R$ does not become singular.[11] To minimize perturbation due to initial parameter settings, up to $t_0 > 0$, the initially defined covariance matrix of the proposal distribution is used, and as off $t_0 + 1$, the adaption process is triggered (cp. Haario et al., 2001, p. 225f.).

Algorithm 5: AM-RWM algorithm

Step 1 Initialize $\theta^{(0)} \in \Theta$ and fix the burn-in phase (n_0), the initial period (t_0) as well as the Markov chain sample size (T).

Step 2 At each iteration t, sample sequentially candidate draws θ^* from proposal density q by a Metropolis step.

Step 3 After 10 accepted draws, update $\mathbb{C}_M$ by a greedy start.

Step 4 Update $\mathbb{C}_R$ at each iteration $t \geq t_0$, based on $[\theta^{(1)}, \ldots, \theta^{(t-1)}]$.

Step 5 Repeat steps 2–4 $n_0 + T$ times, discard the draws from the burn-in phase (the first n_0 iterations), and store the following T draws $[\theta^{(n_0+1)}, \ldots, \theta^{(n_0+T)}]$.

Both, the *AP* and the *AM* are widely used and established approaches which are easy to implement, fast and viable. Even G. O. Roberts and Rosenthal (2007, p. 473f.) and Vihola (2011, p. 47) claim that the *AM* approach is the most natural and useful adaption scheme to date. For further variants of adaptive schemes, see also Gilks et al. (1998), G. O. Roberts and Rosenthal (2001, 2007, 2009), Brockwell and Kadane (2005), Levine, Yu, Hanley, and Nitao (2005), Andrieu and Thoms (2008), Bai et al. (2011) and Vihola (2011, 2012). However, Haario et al. (2001, p. 232) show that the results from adaptive algorithms are superior as compared to the results of non-adaptive schemes relying on a systematic tuning of the acceptance rate (id., remark 6).

Please note that adaptive algorithms have to be handled with caution. The major point of criticism is that they typically rely too much on the past, cp. Robert and Casella (2010, p. 263). A further drawback refers to the specification of the greedy start procedure. Here, G. O. Roberts and Rosenthal (2009, p. 365) point to the fact that a sampler being too "greedy" adapts to closely to initial information and possibly sticks at a specific mode. Gilks et al. (1998) and G. O. Roberts and Rosenthal (2009) describe some ways to overcome this problem. Among other things, these findings will be re-examined in the context of the heaping model.

[11] Vihola (2011) states that such a lower bound on the adapted covariance matrix can deteriorate the efficiency of the sampler thus proposes an unconstrained *AM* algorithm.

3.2.6 Tools for comparison of different RWM algorithms

Simulation efficiency, convergence, and estimation accuracy are the most common means to evaluate how well model parameters have been estimated. In conjunction therewith, acceptance rates, inefficiency factors, graphics for visual exploration, convergence diagnostics as well as the computational time are documented.

Acceptance rate

The empirical frequency of acceptance in a $MCMC$ run serves as a performance indicator for optimization (tuning) purposes and comparison of different algorithms. G. O. Roberts et al. (1997) stated an acceptance of about 25% for models of higher dimension and roughly about 50% for models of dimension one and two to be sufficient when using the *uniform-in-each-direction* proposal density, see also Gelman, Roberts, and Gilks (1996, p. 600). However, Robert and Casella (2010, p. 193) emphatically refer to being careful with the acceptance rate in RWM algorithms. High rates might indicate poor convergence patterns as the moves on the support of the target are more limited. Low rates can result when the sampler moves quickly on the surface of the target, often reaching the "borders" of its support (parameter space). This behavior is typical for normally distributed proposal densities with a wide spread. Scaling down the covariance matrix of the jumping distribution is usually a promising means when trying to achieve viable acceptance rates. Acceptance rate tuning is only a suggestion, though, found to be robust in practice, see Sherlock et al. (2010, p. 9f.). In multiple-block settings, it might be rational to construct suitable scale parameters for each block, since the components in high-dimensional settings are split into different sub-blocks.

Inefficiency factor

The RWM algorithms can also be compared by their *inefficiency factors* (*Ineff*), so named in Chib and Ramamurthy (2010, p. 24) but also known as the *integrated autocorrelation time* (IAT), i.e. the computational inefficiency of a $MCMC$ sampler. The *Ineff* or IAT summarizes the serial correlations among the sampled draws. The higher variance of $MCMC$ methods compared to the MC variance can be expressed in terms of *simulation inefficiency*. Concretely, it measures how the dependency of the draws degrades the precision of the quantity of interest. It can be approximated by the ratio of the variance of the posterior mean derived from the $MCMC$ sample relative to that one derived from hypothetical *iid* draws, see i.a. Jackman (2009, p. 192f.) and Hoff (2009, p. 102f.). For a given sequence of draws $\{\theta_t\}_{t=1}^{T}$, the inefficiency factor can be computed as

$$\tau = 1 + 2\sum_{l=1}^{L} \kappa_l,$$

where κ_l denotes the autocorrelation function (ACF) at lag l and L is set to 5000, according to Chib and Ramamurthy (2010, p. 25). Following Geyer (1992, p. 477), the sum of the ACF is truncated when the sum of adjacent sample ACF values is negative to obtain a consistent estimator, also called *initial positive sequence (IPS) estimator*. A well-mixing sampler exhibits low autocorrelations, i.e. it decays to zero within a few lags, hence leading to small *Ineff*, see Chib and Ramamurthy (2010, p. 24f). The simulation inefficiency in turn is used in order to estimate the *effective sample size* (ESS) of the $MCMC$ sampler applied. Due to the dependence of successive draws in a Markov chain, the parameter space Θ is explored slower, as compared to the independence sampler. Hence, more iterations of the Markov chain are required to generate a random sample. The ESS is an estimate of the equivalent number of independent iterations that the chain represents and is related to the *Ineff* in the following way

$$T^* = \frac{T}{\tau},$$

with T^* being the effective sample size and T being the Markov chain sample size (Jackman, 2009, p. 193). See Appendix A.2.3 for more details on the inefficiency factor. For the purpose of comparability, the averaged *Ineff* across all model parameters will be reported as well. It represents a good single measure for simulation efficiency of a $MCMC$ sampler.

Graphical inspection

To visualize the results of the different RWM algorithms, in this thesis, four types of plots are used. First, the point estimates (posterior means) with their corresponding SD (or CI for the 2.5th and 97.5th quantiles) are shown to give an overview of the accuracy of the estimates as compared to the true parameter values from simulation. Evaluating the fit of a $MCMC$ sampler with respect to the SD and CI is usually quite conservative. For this reason, *highest posterior density regions* (HDR) are often inspected. The second plot depicts the marginal posterior densities of a selected set of scalar estimates $\widehat{\theta}_d$ with a kernel smoothed histogram and shaded HDR for inspection, see e.g. Hyndman (1996) and Sherlock et al. (2010). The HDR is an interval where all points in a HDR have a higher posterior density than points outside, see e.g. Hoff (2009, p. 42f.). This is opposed to quantile-based intervals where the posterior quantiles of $\widehat{\theta}_d$ are used for construction of the $100(1-\alpha)\%$ confidence interval (CI), probably leading to values of $\widehat{\theta}_d$ that lie outside the CI but have a higher probability (density) than values inside. The HDR allows evaluating the plausibility of the probability distribution of the scalar estimate with respect to all sampled draws. That way, e.g. multi-modality can be easily detected, see Jackman (2009, p. 28). Third, *trajectory plots* (trace plots) are given to visualize the effect of the starting values. Trace plots graph the iterative history of a $MCMC$ sampler and are a simple heuristic tool to examine slow mixing and convergence. Concretely, trace plots can reveal how quickly the

Markov chain converges to some equilibrium or target density.[12] Finally, for each model parameter, plots of the *autocorrelation functions* (ACF) up to lag 100 are depicted to give an insight into the sampler performance. A well-mixing sampler has autocorrelations that decay to zero within a few lags.

Convergence diagnostics

Besides graphical inspection of the output of an $MCMC$ sampler, formal diagnostics are almost inevitable to assess its convergence behavior. Under mild regularity conditions, convergence of the sampler to the target density is almost always assured. The conditions that need to be met are irreducibility and aperiodicity, which are part of the Ergodic Theorem, cp. Chib and Greenberg (1995, p. 329). This means that the chain can move from any $\theta^{(t-1)}$ to the current $\theta^{(t)}$ in a finite number of iterations with non-zero probability. This condition is satisfied when the proposal density has a positive density on the support of target density, $q(\theta^{(t)}|\theta^{(t-1)}) > 0$. When relaxing this condition – i.e. $\theta^{(t-1)}$ reaches $\theta^{(t)}$ in a finite number of iterations –, it is also satisfied by proposal densities with restricted support, see Chib and Greenberg (1995, p. 329). An accordant example is the standard uniform proposal density with a finite dispersion.

Although convergence can be assumed theoretically and diagnosed quickly from graphical inspection of the trace plots, the rate of convergence has to be monitored. Comprehensive reviews of corresponding convergence diagnostics can be found, e.g. in Cowles and Carlin (1996). Here, the focus lies on two of the most widely used diagnostics: Geweke's test of non-stationarity (1992) for single chains and S. P. Brooks and Gelman's criterion (1998) for multiple independent chains.

Geweke's test of non-stationarity can be applied to any $MCMC$ approach, see Geweke (1992). It requires only a single chain and is essentially univariate (cp. Cowles & Carlin, 1996, p. 866). In detail, the $MCMC$ outcome for a given component of θ is decomposed into two parts. The first part consists of a number of early iterations and runs from $t = 1$ to $t = 0.1T$ when considering the first 10% of the chain. The second part usually contains the last 50% of the iterations in the chain and runs from $t = 0.5T$ to $t = T$. Then, for a given component θ_d of θ, the values $\theta_d^{(t)}$ of the respective parts are averaged and compared to each other. If the resulting averages are statistically different, non-stationarity is postulated.[13] The test is applied with a suitable burn-in period ($n_0 = 1000$) for all model parameters. Scores well within two SD give no indication for lack of convergence, while deviations exceeding ± 1.96 suggest that additional samples are required to achieve convergence.

[12] At this point, the modeler has to be aware of *pseudo-convergence* in which a Markov chain appears to have converged to its equilibrium distribution but has only explored parts of the target distribution. Pseudo-convergence often occurs in cases where parts of the state space are poorly connected by the Markov chain thus taking many iterations to get from one part to another. This phenomenon often occurs in case of *multi-modality*, see Geyer (2011, p. 18).

[13] See Appendix A.2.4 for more details on Geweke's convergence diagnostic.

S. P. Brooks and Gelman's convergence criterion generalizes the convergence diagnostic of Gelman and Rubin (1992). Both diagnostics are particularly useful to uncover multi-modality. In case of multi-modality, several chains with different starting values might run to different modes and stuck there. The criterion of S. P. Brooks and Gelman (1998) is aimed at comparing the output of multiple independent chains according to the between-chain and within-chain variation for each scalar component of θ. High variances between the chains are penalized and high variances within the chains are rewarded. In other words, the estimated values of each iteration should be similar and each chain should cover the parameter space well. To express this behavior, Gelman and Rubin (1992) introduce the quantity $\sqrt{\widehat{R}}$, which is interpreted as the potential scale reduction factor ($PSRF$) or shrink factor. The $PSRF$ declines to 1 as T increases. For slowly mixing samplers, the variance between the means from the different chains exceeds the average of the within-chain variances. This leads to values substantially above 1. On the contrary, in fast-mixing samplers, the effect of the initial values mitigates quickly. According to Gelman, Carlin, John, B., Stern, and Rubin (2004, p. 297), values of $\sqrt{\widehat{R}}$ below 1.1 are considered to be acceptable.[14]

S. P. Brooks and Gelman (1998) extend the $PSRF$ i.a. to multivariate summaries. The multivariate version of the $PSRF$ ($MPSRF$) is introduced as an upper bound for the univariate $PSRF$s corresponding to the D scalar estimates of $\widehat{\theta}$, see S. P. Brooks and Gelman (1998, p. 447). Again, a suitable burn-in period should be considered and terminated for calculation.[15]

Marginal likelihood

To compare alternative Bayesian models, for each model specification a single summary number is needed (cp. Gelfand & Dey, 1994, p. 502). One appropriate aggregate for model selection is the *marginal likelihood*. The marginal likelihood indicates what $\mathbf{z}$ should look like before the data have been observed. It can be calculated as

$$p(\mathbf{z}) = \int_{\Theta} p(\mathbf{z}|\theta)p(\theta)d\theta.$$

The marginal likelihood is the often omitted denominator in the calculation of the posterior distribution, which serves as a constant of proportionality. Unfortunately, the marginal likelihood is notoriously difficult to estimate. Several methods of approximation are available, see e.g. Frühwirth-Schnatter (2006, Chapter 5) or Han and Carlin (2001) for an overview. These include i.a. trans-dimensional methods, like the product space $MCMC$ (Carlin & Chib, 1995) and the reversible jump $MCMC$ (Green, 1995). Those methods are across-model strategies considering the joint posterior distribution of model indicators and model parameters,

[14] In Jackman (2009, p. 255), a threshold value of 1.2 is reported as being acceptable.

[15] See Appendix A.2.5 for more details on S. P. Brooks and Gelman's convergence criterion.

$p(\theta_o, \mathcal{M}_o|\mathbf{z})$. Let $\mathcal{M}_o$ index models under consideration to describe $\mathbf{z}$. In contrast, within-model strategies examine posterior distributions separately for each model, $p(\theta_o|\mathbf{z}, \mathcal{M}_o)$. Those include simulation-based approaches, e.g. the bridge sampling technique (Frühwirth-Schnatter, 2004, 2006), the Gelfand-Dey estimator (Gelfand & Dey, 1994), approximations to the marginal likelihood based on density ratios as the Chib's estimator (Chib, 1995; Chib & Jeliazkov, 2001) as well as marginal likelihood estimation via power posteriors (Friel & Pettitt, 2008).

In this thesis, the marginal likelihood is calculated according to Chib and Jeliazkov (2001). The authors extend and complete the method of Chib (1995) for Gibbs sampler frameworks by adapting it to Metropolis-Hastings outputs and to multiple-block sampling with fixed blocks. Chib's estimators are more accurate yet demanding for programming and computation. That is why it is necessary to run $MCMC$ samplers for each block of parameters as compared to the Gelfand-Dey method for example, see C. Liu and Liu (2012). The estimation of the marginal likelihood according to Chib and Jeliazkov (2001) is described in Appendix A.2.6.

By taking the logarithms, the marginal likelihood can be estimated for a given model $\mathcal{M}_o$, with $o = 1, \ldots, O$, from the following identity

$$\log m(\mathbf{z}|\mathcal{M}_o) = \log p(\mathbf{z}|\mathcal{M}_o, \theta'_o) + \log p(\theta'_o|\mathcal{M}_o) - \log p(\theta'_o|\mathbf{z}, \mathcal{M}_o), \tag{3.1}$$

see Chib and Jeliazkov (2001, p. 270). This identity only requires the evaluation of the log-likelihood function, the prior, and an estimate of the posterior ordinate. The parameter vector θ' is approx. by an appropriate high-density point in the support of the posterior. In general, the posterior mean is used as approx. high density point. The first as well as the second term on the right hand side can be calculated as soon as the $MCMC$ sampling is completed. Moreover, it needs an estimate of the third term, the posterior ordinate, $p(\theta'_o|\mathbf{z}, \mathcal{M}_o)$. For simplicity, the notation that indicates specific models is subsequently omitted.

Let $\{\theta^{(t)}\}_{t=1}^{T}$ denote the sampled draws from the posterior distribution and $\{\theta^{(j)}\}_{j=1}^{J}$ the sampled draws from the proposal distribution, both given a fixed value θ'. Then, a simulation-consistent estimate of the posterior ordinate is available as

$$\hat{p}(\theta'|\mathbf{z}) = \frac{T^{-1} \sum_{t=1}^{T} \alpha(\theta^{(t)}, \theta'|\mathbf{z}) q(\theta'|\theta^{(t)})}{J^{-1} \sum_{j=1}^{J} \alpha(\theta', \theta^{(j)}|\mathbf{z})} \tag{3.2}$$

for a single-block sampling approach, see Chib and Jeliazkov (2001, p. 270). With regard to multiple parameter blocks, the posterior ordinate at θ' is denoted as $p(\theta'_1, \ldots, \theta'_K|\mathbf{z})$ for the K blocks of model parameters considered. Let $\Psi_{k-1} = (\theta_1, \ldots, \theta_{k-1})$ and $\Psi^{k+1} = (\theta_{k+1}, \ldots, \theta_K)$ be the parameter blocks below and beyond k. The simulation-consistent estimate of the posterior ordinate for

each fixed blocks is now available as

$$\hat{p}(\theta'_k|\mathbf{z}, \theta'_1, \ldots, \theta'_{k-1}) = \frac{T^{-1}\sum_{t=1}^{T} \alpha\left(\theta_k^{(t)}, \theta'_k|\mathbf{z}, \Psi'_{k-1}, \Psi^{k+1,(t)}\right) q\left(\theta_k^{(t)}, \theta'_k|\Psi'_{k-1}, \Psi^{k+1,(t)}\right)}{J^{-1}\sum_{j=1}^{J} \alpha\left(\theta'_k, \theta_k^{(j)}|\mathbf{z}, \Psi'_{k-1}, \Psi^{k+1,(j)}\right)}. \tag{3.3}$$

The accordant marginal likelihood estimate (at a logarithmic scale) is

$$\log \hat{m}(\mathbf{z}) = \log p(\mathbf{z}|\theta') + \log p(\theta') - \sum_{k=1}^{K} \log \hat{p}(\theta'_k|\mathbf{z}, \theta'_1, \ldots, \theta'_{k-1}). \tag{3.4}$$

To ease computation, Chib and Jeliazkov (2001) recommend to fix an appropriate set of parameters for reduced runs. In this thesis, the parameters of the underlying distribution $\psi = [\mu, \sigma, \rho_Z]$ are fixed to the posterior means. The model constraints are implemented as in the RWM runs by a simple rejection sampling method.

Chib and Ramamurthy (2010, p. 29) further extend the framework of Chib and Jeliazkov (2001) to accommodate the randomized-block sampling strategy. First, the number of blocks is fixed to the average number of blocks $\bar{K}$ realized in a RMB-RWM run for estimation of the posterior ordinate. Second, $\bar{K}$ parameter blocks are constructed by randomly assigning components of θ. Estimation of the log-marginal likelihood proceeds in a similar manner as for the fixed blocks.

Besides the estimates of the log-marginal likelihood, the corresponding SD will be reported. The SD can be derived from multiple independent runs and is a good approximation to the numerical standard error (NSE). A clear correspondence to the inefficiency factor exists, because a scheme that is efficient for sampling the posterior distribution is also efficient for estimating the log-marginal likelihood. As the *Ineff* decreases, the NSE of the log-marginal likelihood gets smaller, see Chib and Jeliazkov (2001, p. 274).

Major critique points on marginal likelihood estimation are raised by Gelfand and Dey (1994, p. 504), Han and Carlin (2001, p. 1132), and Frühwirth-Schnatter (1995, p. 240). The authors consistently point out that the marginal likelihood is not interpretable when improper priors are used. Furthermore, most estimators of the marginal likelihood are prone to be biased on draws of a poor mixing sampler, see Frühwirth-Schnatter (2006, p. 145). To overcome these obstacles, alternative Bayesian model selection methods exist. An overview is given in Carlin and Louis (2009). All of these methods for estimation of the marginal likelihood require substantial time and effort. There is a tradeoff between investment and benefit of the single-number summary of relative model worth (Han & Carlin, 2001, p. 1132).[16]

[16]Reichl (2015) proposes a new estimator of the marginal likelihood that can be implemented generically in almost any sampling scheme and is supposed to decrease the computational burdens associated with marginal likelihood estimation significantly, yet, yielding stable results – also when the number of model parameters is large. This could serve as a starting point for further research.

Once the marginal likelihood is calculated, the *Bayes factor* (BF) of competing models $\mathcal{M}_O$ can be calculated, cp. Chib and Jeliazkov (2001). If all models have the same prior probability $p(\mathcal{M}_o)$, this model is chosen that has the highest log-marginal likelihood among all models considered, cp. Frühwirth-Schnatter (2006, p. 121). The BF is the ratio of the marginal likelihood in favor of one of the models. It can be calculated as

$$\widehat{BF}_{12} = \exp\{\log \widehat{m}(\mathbf{z}|\mathcal{M}_1) - \log \widehat{m}(\mathbf{z}|\mathcal{M}_2)\}.$$

That is, the BF provides a measure of whether the data $\mathbf{z}$ support $\mathcal{M}_1$ relative to $\mathcal{M}_2$, or vice versa. If BF is greater than one, the odds on $\mathcal{M}_1$ are greater than those for $\mathcal{M}_2$, in words, $\mathcal{M}_1$ is more plausible than $\mathcal{M}_2$ with respect to $\mathbf{z}$. The opposite is true, i.e. $\mathcal{M}_2$ is more plausible than $\mathcal{M}_1$, if the BF is smaller than one, see Frühwirth-Schnatter (2006, p. 119). Interpretative ranges for the BF are provided by Jeffrey's scale cited in Jackman (2009, p. 38) or in Kass and Rafferty (1995, p. 777), both referring to Jeffreys (1961).

The BF implicitly penalizes model complexity if two nested models provide a comparable fit to $\mathbf{z}$. This also constitutes a relationship of Bayesian model selection by BF and frequentist model selection based on criteria such as AIC, see Section 3.1.2, cp. Frühwirth-Schnatter (2006, p. 120f.). In the AIC, the first term measures the goodness of fit, and the second term adds a quantity which penalizes model complexity (id., p. 116). As Gelfand and Dey (1994, p. 508) point out, *Schwarz's criterion* (SC) can be derived as an asymptotic approximation to the marginal likelihood.

3.2.7 Specification and results of different RWM algorithms

The RWM schemes presented in the previous sections are now utilized to estimate the heaping model described in Chapter 2 (Model I). To explore the variety of ad-hoc tuning methods for the original single-block RWM algorithm, different set-ups are chosen and compared with respect to their performance. All RWM set-ups and schemes are enlisted in the Tables 3.2 to 3.4. The Markov chain sample size is fixed to $T = 10{,}000$ following a burn-in of $n_0 = 1000$ iterations.

In the specified heaping model, the parameter vector θ comprises the parameters of the zero-inflated log-normal distribution, $\psi = [\mu, \sigma, \rho_Z]$, and the heaping probabilities, $\phi = [\rho_b]_{b=1}^{S}$. As suitable prior distribution, a multivariate normal distribution is chosen assuming that the parameters are a priori independent. The RWM algorithms are also initialized that way (Step 1 of Algorithm 1), by sampling from

$$\phi^{(0)} \sim \mathcal{N}_{19}(\varphi, \Sigma)$$

with $\varphi = [0.2]_{b=1}^{19}$ as far as not stated otherwise. The corresponding covariance matrix is $\Sigma = \frac{1}{\lambda}\mathcal{I}_{19}$, with λ equalling 100, and $\mathcal{I}$ being a matrix of ones (identity matrix) of dimension S. The prior of the components of ψ (also used for initialization) is as follows:

$$\psi^{(0)} \sim \mathcal{N}_3(\upsilon, \Upsilon),$$

with $\upsilon = (7.714, 0.839, 0.990)'$ and covariance matrix $\Upsilon = \text{diag}(0.1, 0.001, 0.0001)\mathcal{I}_3$. In doing so, φ and υ are equal to the initial values already used for ML estimation in Section 3.1.2 . By definition, Σ and Υ satisfy the usual positivity and positive definiteness constraints on matrices. These constraints are denoted by $\mathcal{C}_\Sigma$. $\mathcal{C}_\psi$ denotes the linear restrictions on the parameters of the underlying distribution (2.1) ensuring positivity of all components of ψ. Several constraints imposed on the model parameters are summarized by $\mathcal{C}_\Theta$. Hence, the log-posterior is (up to a constant of proportionality):

$$\begin{aligned}
&\log p(\theta|z_i) \propto \log p(z_i|\theta) + \log p(\theta) \\
&\log p(\theta|z_i) \propto \Bigg(\left[1 - \sum_{b=1}^{S} \rho_b \right] f(z_i|\psi) dz_i \mathbb{I}(z_i \in \mathbb{R}_0^+) + \sum_{b=1}^{S} \rho_b \left[F(u_b|\psi) - F(l_b|\psi) \right] \mathbb{I}(z_i \in \mathcal{H}) \\
&\qquad - \frac{\lambda}{2}(\phi - \varphi)^2 - \frac{1}{2}(\psi - \upsilon)' \Upsilon^{-1} (\psi - \upsilon) \Bigg) \mathbb{I}(\theta \in \mathcal{C}_\Theta).
\end{aligned}$$

This is the target distribution that is to be explored. Be aware that the density $p(\theta|z_i)$ is not known exactly. However, it can be calculated up to the normalizing constant, $p(\theta|z_i) = Cg(\theta|z_i)$, where $g(\theta|z_i)$ is either known or easy to compute by taking the likelihood function times the prior distribution. The positive scalar C corresponds to $1/p(\mathbf{z})$ and is usually either unknown, or hard to compute.

The posterior density has a restricted parameter space, given by $\mathcal{C}_\Theta$. The implementation of these constraints is as follows: in the initialization step the prior density and in the Metropolis step the proposal density are subject to constraints. To be concrete, if a draw violates any of the constraints it is rejected immediately just before the random-walk process starts.[17] The algorithm is allowed to draw up to 100 starting or proposal values before giving an error message and stopping the procedure. Since the requirement of multiple redraws indicates an inappropriate specification of the prior and proposal densities with respect to the parameter space, the number of redraws is documented at the relevant points, cp. Tables 3.2 to 3.4. Draws fitting to the constraints are taken as starting or candidate values, respectively.

The parameters of the uniform proposal density (Step 2 of Algorithm 1) are set to $\epsilon_\rho = [0.01]_{b=1}^{19}$ and $\epsilon_\psi = [0.1, 100, 0.01]$ at start. When using the multivariate normal proposal density, Σ_q is defined as an identity matrix times diag$(0.001, \ldots, 0.001, 0.01, 0.0001, 0.00001)$. The length of the diagonal is determined by D.

All $MCMC$ samplers are coded in `R 3.1.2` and executed on a Windows 7 64-bit machine with a 3.20 GHz Intel Core i5-4570 architecture CPU. Implementation of the RWM algorithm (cp. Appendix B.2) borrows partially from the function `MCMCmetrop1R` in the `R` package *MCMCpack*, see Martin, Quinn, and Park (2011).

[17] This approach is called simple rejection sampling method, see e.g. Jackman (2009, pp. 159ff.).

Table 3.2: Tuning set-ups for the original RWM algorithm.

Trial	**Par**	**Prior**	**VC**	**Posterior** $\propto \ell$	**Proposal**	**Repeats & Acceptance rates**
1	$\phi = [\rho_b]_{b=1}^{S}$	$\phi \sim \mathcal{N}_{19}(\varphi, \Sigma)$ $\varphi = [0.2]_{b=1}^{19}$, $\Sigma = VC$	$VC = \frac{1}{\lambda}\mathcal{I}_{19}$ $\lambda = 100$	$-\frac{\lambda}{2}(\phi - \varphi)^2$	$\theta^* \sim \mathcal{U}(\theta^{(t)} - \epsilon, \theta^{(t)} + \epsilon)$ $\epsilon_\phi = [0.01]_{b=1}^{19}$	repeatp: 0 repeatc: 0 avineff: 74.3 runtime: 2.0 AR: 0.234
2	$\phi = [\rho_b]_{b=1}^{S}$	$\phi \sim \mathcal{N}_{19}(\varphi, \Sigma)$ $\varphi = [0.2]_{b=1}^{19}$, $\Sigma = VC$	$VC = \frac{1}{\lambda}\mathcal{I}_{19}$ $\lambda = 1000$ [1]	$-\frac{\lambda}{2}(\phi - \varphi)^2$	$\theta^* \sim \mathcal{U}(\theta^{(t)} - \epsilon, \theta^{(t)} + \epsilon)$ $\epsilon_\phi = [0.01]_{b=1}^{19}$	repeatp: 0 repeatc: 0 avineff: 57.0 runtime: 2.1 AR: 0.244
3	$\phi = [\rho_b]_{b=1}^{S}$	$\phi \sim \mathcal{N}_{19}(\varphi, \Sigma)$ $\varphi = [0.2]_{b=1}^{19}$, $\Sigma = VC$	$VC = \frac{1}{\lambda}\mathcal{I}_{19}$ $\lambda = 1000$	$-\frac{\lambda}{2}(\phi - \varphi)^2$	$\theta^* \sim \mathcal{U}(\theta^{(t)} - \epsilon, \theta^{(t)} + \epsilon)$ $\epsilon_\phi = [0.05]_{b=1}^{19}$ [2]	repeatp: 0 repeatc: 0 avineff: 523.5 runtime: 2.1 AR: 0.003
4	$\phi = [\rho_b]_{b=1}^{S}$	$\phi \sim \mathcal{N}_{19}(\varphi, \Sigma)$ $\varphi = [0.1]_{b=1}^{19}$ [3], $\Sigma = VC$	$VC = \frac{1}{\lambda}\mathcal{I}_{19}$ $\lambda = 1000$	$-\frac{\lambda}{2}(\phi - \varphi)^2$	$\theta^* \sim \mathcal{U}(\theta^{(t)} - \epsilon, \theta^{(t)} + \epsilon)$ $\epsilon_\phi = [0.01]_{b=1}^{19}$	repeatp: 0 repeatc: 0 avineff: 93.3 runtime: 2.1 AR: 0.247
5	$\phi = [\rho_b]_{b=1}^{S}$	$\phi \sim \mathcal{N}_{19}(\varphi, \Sigma)$ $\varphi = [0.3]_{b=1}^{19}$ [3], $\Sigma = VC$	$VC = \frac{1}{\lambda}\mathcal{I}_{19}$ $\lambda = 1000$	$-\frac{\lambda}{2}(\phi - \varphi)^2$	$\theta^* \sim \mathcal{U}(\theta^{(t)} - \epsilon, \theta^{(t)} + \epsilon)$ $\epsilon_\phi = [0.01]_{b=1}^{19}$	repeatp: 0 repeatc: 0 avineff: 56.0 runtime: 2.1 AR: 0.236

Notes: $MCMC$ samples sizes are $T = 10{,}000$ following $n_0 = 1000$ iterations considered as burn-in. Runtimes are given in hours.
repeatp: repeats necessary for starting values, repeatc: repeats necessary for candidate values, avineff: averaged inefficiency coefficient, AR: acceptance rate.
[1]Increase λ to get smaller variances for the prior densities, i.e. less informative priors.
[2]One possibility to reduce autocorrelation is to increase the proposal variance, see Hoff (2009, p. 181).
[3]Vary starting points of the $MCMC$ to control for convergence.

Results for tuning of the original RWM algorithm

The tuning of the original RWM algorithm in a complex setting – like the one considered – is not straightforward. A brief sensitivity analysis is therefore used to illustrate the influence of certain tuning parameters. The specification of the prior and the proposal densities are varied as well as different starting values are tested, see Table 3.2. To focus in particular on justifiable specifications of ϕ, the parameters of the underlying true distribution are fixed to true parameter values from simulation at start.

The average runtime for estimation of the tuning set-up is about 2 hours. The term *runtime* is used to refer to the elapsed time taken by the entire system to complete the RWM algorithm task. Trial 1 corresponds to the specification of the RWM algorithm mentioned above and yields the following results: all ρ, with except of ρ_1, ρ_4, ρ_9, ρ_{10}, ρ_{13} and ρ_{14} are well estimated,[18] as can be seen in Table 3.5. The acceptance rate is as high as expected for a 19-dimensional setting (23.4%).[19]

First of all, in trial 2, the dispersion of the prior distribution is decreased by enhancing λ from 100 to 1000. As can be seen, the estimates for ρ_4, and ρ_{10} are again outside one standard deviation (SD), whereas ρ_1, ρ_9, ρ_{13}, and ρ_{14} are now within. In contrast, heaping probabilities ρ_3, ρ_5, ρ_{11}, ρ_{12}, and ρ_{15} are outside in trial 2. The high acceptance rate of 24.4% indicates a sufficient exploration of the parameter space (as in trial 1). However, with regard to the averaged *Ineff*, choosing a less informative prior seems to be more efficient (74.3 vs. 57.0). Thus, for the following specifications, a prior dispersion with λ equalling 1000 is assumed.

In trial 3, the dispersion of the standard uniform proposal density is increased to enable an even better exploration of the parameter space in each Metropolis step. The increase of ϵ leads to a markedly decrease in the acceptance rate (0.3%) and a higher averaged *Ineff* (523.5). This finding strongly confirms that the performance of the sampler can be severely affected by a poor choice of the variance of the proposal density. Additionally, most of the estimates are far beyond one SD. Since in trial 2 almost all estimates are within one SD and owing to the higher simulation efficiency, ϵ will be fixed at 0.01 for the heaping probabilities in the following.

In trials 4 and 5, the starting values for the $MCMC$ sampler are varied. Trial 3 starts with $\varphi = [0.1]_{b=1}^{19}$ and trial 4 with $\varphi = [0.3]_{b=1}^{19}$, see Table 3.5. No improvement in estimation of the model parameters can be found. In trial 5, the averaged *Ineff* is markedly smaller than in trial 4 (56.0 vs. 93.3).The acceptance rates are comparable to those of trials 1 and 2 (24.7% in trial 4 and 23.6% in trial 5).

[18] In the following, estimated parameter values (given as the posterior mean) equalling the true parameter values from the DGP, or being located in the range of one *SD* below or above the posterior mean, are considered as being well estimated. The *SD* enables one to give a range of the estimated parameter values where the true parameter values are expected to lie within.

[19] This is exactly the acceptance rate Gelman, Roberts, and Gilks (1996) recommend for multi-parameter settings. Although, this is not a theoretical value for optimality.

In summary, the results of the sensitivity analysis conducted indicate that the chains fully navigate the posterior distribution no matter how they have been initialized. Thus, there is no need to increase the Markov chain sample size. Since the specifications of trial 2 are superior to the others considered, those will be assumed for the following variations of the RWM algorithm.

Results for the blocking strategies

To overcome the pitfalls of a single-block RWM scheme, different blocking strategies have been proposed and are now evaluated on the basis of their efficiency in estimation. First, the parameters for ψ, that have been kept fix in the foregoing set-ups, are now included for estimation. In trial 6, the parameters of the heaping mechanism (ρ_b) are estimated in one block together with the parameters of the underlying zero-inflated log-normal distribution. When adding $\psi = [\mu, \sigma, \rho_Z]$, the estimates noticeably get worse. The acceptance rate shrinks to 4.7% and the averaged *Ineff* increases to 212.7. Estimating the parameters of the heaping mechanism together with the parameters of the underlying model in one single block yields unsatisfactory estimates and figures, see Table 3.6.

In further trials, the model parameters are grouped into different blocks. The M-RWM (trial 7) encompasses the heaping probabilities according to the modulo and the I-RWM (trial 8) corresponding to the intervals defined in advance. The RMB-RWM (trials 9 and 10) constitutes random blocks. A separate block for the parameters of the underlying true distribution was assumed in trials 7–9. In trial 10, all model parameters are randomly grouped into different blocks, see Table 3.3. The posterior means and their corresponding 95% CI for trials 6–8 and 10 are given in Figure 3.2 and Figure 3.3. The estimates are slightly closer to the true ones (in trial 7). However, the efficiency gain compared to the S-RWM scheme is considerable: the averaged *Ineff* decreases to 31.0 and the acceptance rates increase ($AR_1 = 53.4\%$, $AR_2 = 56.3\%$, $AR_3 = 40.2\%$, and 10.5% for ψ). Albeit the estimates are approx. equal to those in trial 7, a further efficiency gain is obvious due to the increased number of blocks considered in trial 8 ($K = 9$ vs. $K = 4$). Furthermore, the even higher acceptance rates (Table 3.3) indicate superiority of specifying 9 blocks instead of 4. The same applies to the estimates from trials 9 and 10 where the parameters are grouped randomly.

The most efficient set-up exhibits trial 10, having the by far lowest averaged *Ineff* (19.6). However, ρ_3–ρ_6 as well as ρ_{10}–ρ_{12} are outside one SD. When considering the 95% confidence intervals (CI) instead (Figure 3.2), only ρ_3, ρ_5, and ρ_{10} lie outside the CI. The estimates of the parameters μ and ρ_Z of the underlying distribution are both within one SD, and $\widehat{\sigma}$ is within two SD in trials 6–10, see Table 3.6. The respective 95% CI for the parameter vector ψ of trials 6–8 and 10 are graphically shown in Figure 3.3. As can be seen, the true value of σ is well within the 95% CI.

The results for blocking highlight the adequate sampler length and well-mixing behavior, by chains that sufficiently explore the range of the posterior distribution.

Table 3.3: Blocking set-ups for the multiple-block strategy.

Trial	**Par**	**Prior**	**VC**	**Posterior** $\propto \ell$	**Proposal**	**Repeats & Acceptance rates**
6	$\phi = [\rho_b]_{b=1}^{S}$ $\psi = [\mu, \sigma, \rho_Z]$	$\phi \sim \mathcal{N}_{19}(\varphi, \Sigma)$ $\varphi = [0.2]_{b=1}^{19}$, $\Sigma = VC$ $\psi \sim \mathcal{N}_3(\upsilon, \Upsilon)$ $\upsilon = (7.714, 0.839, 0.990)'$	$VC = \frac{1}{\lambda}\mathcal{I}_{19}$ $\lambda = 1000$ $\Upsilon = \text{diag}(0.1, 0.01, 0.001)\mathcal{I}_3$	$-\frac{\lambda}{2}(\phi - \varphi)^2$ $-\frac{1}{2}(\psi - \upsilon)'\Upsilon^{-1}(\psi - \upsilon)$	$\theta^* \sim \mathcal{U}(\theta^{(t)} - \epsilon, \theta^{(t)} + \epsilon)$ $\epsilon_\phi = [0.01]_{b=1}^{19}$ $\epsilon_\psi = [0.1, 0.01, 0.001]$	repeatp: 0 repeatc: 0 avineff: 212.7 runtime: 2.1 AR_S: 0.047
7	$\phi = [\rho_b]_{b=1}^{S}$ $\psi = [\mu, \sigma, \rho_Z]$	$\phi \sim \mathcal{N}_{19}(\varphi, \Sigma)$ $\varphi = [0.2]_{b=1}^{19}$, $\Sigma = VC$ $\psi \sim \mathcal{N}_3(\upsilon, \Upsilon)$ $\upsilon = (7.714, 0.839, 0.990)'$	$VC = \frac{1}{\lambda}\mathcal{I}_{19}$ $\lambda = 1000$ $\Upsilon = \text{diag}(0.1, 0.01, 0.001)\mathcal{I}_3$	$-\frac{\lambda}{2}(\phi - \varphi)^2$ $-\frac{1}{2}(\psi - \upsilon)'\Upsilon^{-1}(\psi - \upsilon)$	$\theta^* \sim \mathcal{U}(\theta^{(t)} - \epsilon, \theta^{(t)} + \epsilon)$ $\epsilon_\phi = [0.01]_{b=1}^{19}$ $\epsilon_\psi = [0.1, 0.01, 0.001]$ 4-Block proposal[1]	repeatp: 0 repeatc: 0 avineff: 31.0 runtime: 8.4 AR_ψ: 0.105 AR_M: 0.534, 0.563, 0.402
8	$\phi = [\rho_b]_{b=1}^{S}$ $\psi = [\mu, \sigma, \rho_Z]$	$\phi \sim \mathcal{N}_{19}(\varphi, \Sigma)$ $\varphi = [0.2]_{b=1}^{19}$, $\Sigma = VC$ $\psi \sim \mathcal{N}_3(\upsilon, \Upsilon)$ $\upsilon = (7.714, 0.839, 0.990)'$	$VC = \frac{1}{\lambda}\mathcal{I}_{19}$ $\lambda = 1000$ $\Upsilon = \text{diag}(0.1, 0.01, 0.001)\mathcal{I}_3$	$-\frac{\lambda}{2}(\phi - \varphi)^2$ $-\frac{1}{2}(\psi - \upsilon)'\Upsilon^{-1}(\psi - \upsilon)$	$\theta^* \sim \mathcal{U}(\theta^{(t)} - \epsilon, \theta^{(t)} + \epsilon)$ $\epsilon_\phi = [0.01]_{b=1}^{19}$ $\epsilon_\psi = [0.1, 0.01, 0.001]$ 9-Block proposal[2]	repeatp: 0 repeatc: 0 avineff: 22.2 runtime: 18.6 AR_ψ: 0.103 AR_I: 0.794, 0.592, 0.715, 0.656, 0.625, 0.701, 0.745, 0.794
9	$\phi = [\rho_b]_{b=1}^{S}$ $\psi = [\mu, \sigma, \rho_Z]$	$\phi \sim \mathcal{N}_{19}(\varphi, \Sigma)$ $\varphi = [0.2]_{b=1}^{19}$, $\Sigma = VC$ $\psi \sim \mathcal{N}_3(\upsilon, \Upsilon)$ $\upsilon = (7.714, 0.839, 0.990)'$	$VC = \frac{1}{\lambda}\mathcal{I}_{19}$ $\lambda = 1000$ $\Upsilon = \text{diag}(0.1, 0.01, 0.001)\mathcal{I}_3$	$-\frac{\lambda}{2}(\phi - \varphi)^2$ $-\frac{1}{2}(\psi - \upsilon)'\Upsilon^{-1}(\psi - \upsilon)$	$\theta^* \sim \mathcal{U}(\theta^{(t)} - \epsilon, \theta^{(t)} + \epsilon)$ $\epsilon_\phi = [0.01]_{b=1}^{19}$ $\epsilon_\psi = [0.1, 0.01, 0.001]$ RMB+1-Block proposal[3]	repeatp: 0 repeatc: 0 avineff: 23.6 runtime: 18.4 AR_ψ: 0.106 AR_{RMB}: 0.671, 0.646, 0.638, 0.634, 0.623, 0.639, 0.650, 0.640, 0.622, 0.631, 0.654, 0.671, 0.655, 0.524, 0.576, 0.621 0.636, 0.654, 0.643
10	$\phi = [\rho_b]_{b=1}^{S}$ $\psi = [\mu, \sigma, \rho_Z]$	$\phi \sim \mathcal{N}_{19}(\varphi, \Sigma)$ $\varphi = [0.2]_{b=1}^{19}$, $\Sigma = VC$ $\psi \sim \mathcal{N}_3(\upsilon, \Upsilon)$ $\upsilon = (7.714, 0.839, 0.990)'$	$VC = \frac{1}{\lambda}\mathcal{I}_{19}$ $\lambda = 1000$ $\Upsilon = \text{diag}(0.1, 0.01, 0.001)\mathcal{I}_3$	$-\frac{\lambda}{2}(\phi - \varphi)^2$ $-\frac{1}{2}(\psi - \upsilon)'\Upsilon^{-1}(\psi - \upsilon)$	$\theta^* \sim \mathcal{U}(\theta^{(t)} - \epsilon, \theta^{(t)} + \epsilon)$ $\epsilon_\phi = [0.01]_{b=1}^{19}$ $\epsilon_\psi = [0.1, 0.01, 0.001]$ RMB-Block proposal[4]	repeatp: 0 repeatc: 0 avineff: 19.6 runtime: 18.2 AR_ψ: 0.104, 0.486, 0.566 AR_R: 0.604, 0.579, 0.572, 0.562, 0.543, 0.562, 0.574, 0.577, 0.549, 0.554, 0.585, 0.596, 0.581, 0.467, 0.515, 0.543, 0.573, 0.596, 0.573

Notes: $MCMC$ samples sizes are $T = 10{,}000$ following $n_0 = 1000$ iterations considered as burn-in. Runtimes are given in hours.
repeatp: repeats necessary for starting values, repeatc: repeats necessary for candidate values, avineff: averaged inefficiency coefficient, AR: acceptance rate.
[1]Assume a multiple-block proposal. The blocks for ϕ are summarized by modulo (M-RWM) with an additional block for ψ.
[2]Assume a multiple-block proposal. The blocks for ϕ are summarized by interval (I-RWM) with an additional block for ψ.
[3]Assume a multiple-block proposal. The blocks for ϕ are built randomly with an additional block for ψ.
[4]Assume a multiple-block proposal. The blocks for ϕ together with ψ are built randomly (RMB-RWM).

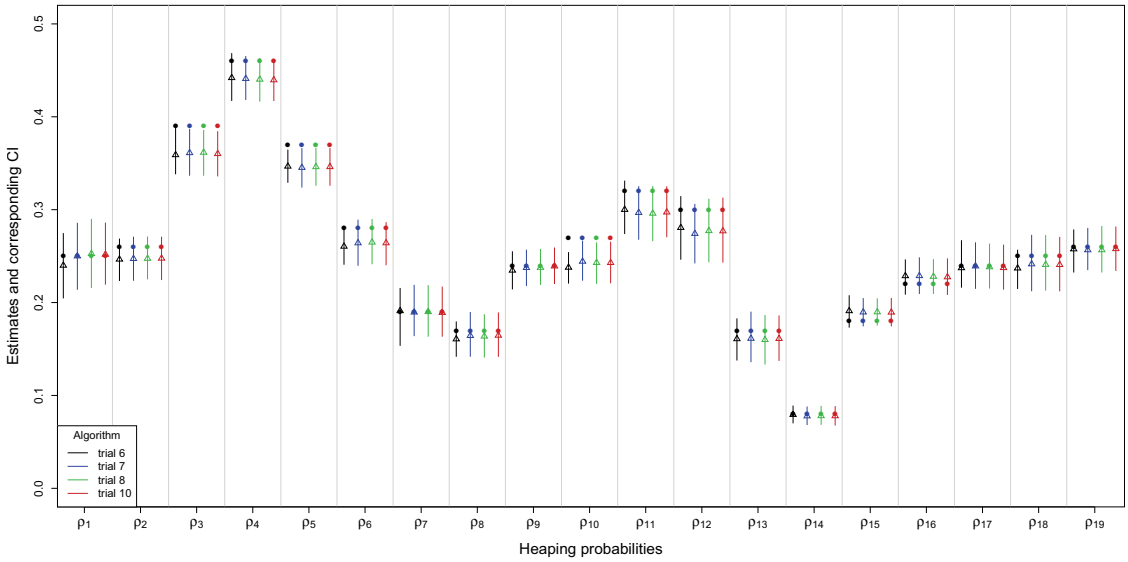


Figure 3.2: Posterior means with 95% confidence intervals (CI) of four multiple-block random-walk MH algorithms for the heaping probabilities. The true parameter values are marked by a dot, parameter estimates are marked by a triangle.

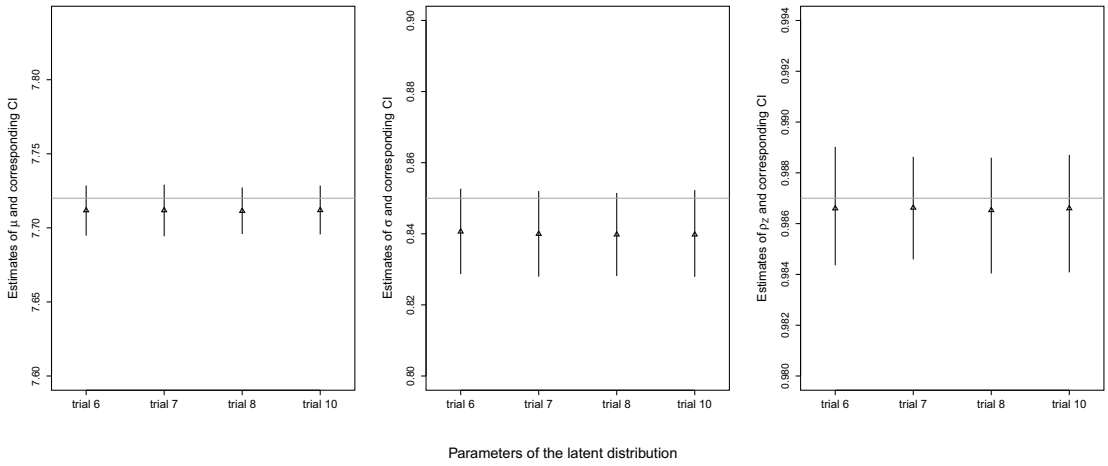


Figure 3.3: Posterior means with 95% confidence intervals (CI) of four multiple-block random-walk MH algorithms for the parameters of the underlying true distribution. The true parameter values are marked by a dashed line, parameter estimates are marked by a triangle.

Results for the adaptive $MCMC$ schemes

Finally, adaptive $MCMC$ schemes on a Gaussian proposal density are compared with the previous results from a standard uniform proposal density, either in a single-block, or in a multiple-block setting, see Table 3.4. At the beginning of this part (trial 11), only the shape of the proposal density is changed from uniform to a multivariate normal. No update is scheduled. The proposal covariance matrix is specified as $\Sigma_q = \text{diag}(0.001, \ldots, 0.001, 0.01, 0.001, 0.0001)$ times the identity matrix indicating independency between the parameters and taking the dimensionality of the parameters into account. When using a multivariate normal distribution as proposal density, it is quite likely that proposed values more often collide with the constraint system given. This can be explained by the wider range of possible values to sample from in the multivariate normal distribution owing to the tails, as compared to the standard uniform distribution with its clear-cut ranges. The counter *repeatc* (cp. Table 3.4) indicates the number of repeats (on average) that were necessary because a sampled candidate draw did not fit the constraint system. The acceptance rate shrinks to 0.1% and the averaged *Ineff* rises to 1043.4. The SD as well as the CI are much larger in case of a multivariate normal proposal distribution. In trial 12, the proposal covariance matrix is assumed to be less disperse, i.e Σ_q is now specified as $\text{diag}(0.0001, \ldots, 0.0001, 0.001, 0.0001, 0.00001)\mathcal{I}_{22}$. Now, the estimates are closer to the true parameter values. The averaged *Ineff* is decreased remarkably to 217.0, and the acceptance rate is increased to 2.0%. The counter *repeatc* falls below 0.001. Hence, the reduced dispersion in the covariance matrix of the proposal density is used in the following trials with updating scheme.

The results improved only modest when updating the proposal covariance matrix based on information from preceding iterations (cp. Table 3.4). Updating is applied to capture possible dependencies among the parameters (trials 13–15). In trial 13, an AP-proposal is used. Here, Σ_q is updated every 1000th iteration (U) based on the last 1000 iterations (M). In trial 14, the AM-proposal is employed with constant c set to 0.0001 and t_0 fixed to 1000. The averaged *Ineff* of trial 13 is slightly decreased to 138.9 and decreases even more in trial 14 (127.5). The corresponding acceptance rates are 49.8% and 48.5%. The covariance matrices $\Sigma_{q_{AP}}$ and $\Sigma_{q_{AM}}$ of the respective last update in trials 13 and 14 are given in the Appendix in Equations (A.3) and (A.4).[20] Both matrices differ to a large extent which can be explained by the constant c additionally used in the AM-Proposal to keep the proposal covariance matrix explicitly small (cp. Haario et al., 2001).

On estimation accuracy, great differences are visible compared to the blocking schemes. The parameter values estimated by updating schemes are farther away from the true ones. The overall deficiency might be mainly attributable to the failed convergence of the adaptive schemes. As Haario et al. (1999) and Bai et al. (2011) point out, adaptive proposals typically fail in settings with many parameters and suggest *RWM-within-Gibbs* schemes (e.g. Single Component Adaptive Metropolis ($SCAM$) algorithm, see G. O. Roberts & Rosenthal, 2009, p. 355).

[20]The correlation matrices $\Omega_{q_{AP}}$ and $\Omega_{q_{AM}}$ are given as well (cp. Equations (A.1) and (A.2)).

Table 3.4: Updating set-ups for the adaptive $MCMC$ schemes.

Trial	Par	Prior	VC	Posterior $\propto \ell$	Proposal	Repeats & Acceptance rates
11	$\phi = [\rho_b]_{b=1}^{S}$ $\psi = [\mu, \sigma, \rho_Z]$	$\phi \sim \mathcal{N}_{19}(\varphi, \Sigma)$ $\varphi = [0.2]_{b=1}^{19}$, $\Sigma = VC$ $\psi \sim \mathcal{N}_3(\upsilon, \Upsilon)$ $\upsilon = (7.714, 0.839, 0.990)'$	$VC = \frac{1}{\lambda}\mathcal{I}_{19}$ $\lambda = 1000$ $\Upsilon = \text{diag}(0.1, 0.01, 0.001)\mathcal{I}_3$	$-\frac{\lambda}{2}(\phi - \varphi)^2$ $-\frac{1}{2}(\psi - \upsilon)'\Upsilon^{-1}(\psi - \upsilon)$	$\theta^* \sim \mathcal{N}_{S+3}(\theta^{(t)}, \Omega)$ $\Omega = \text{diag}(0.001, \ldots, 0.001,$ $0.01, 0.001, 0.0001)\mathcal{I}_{22}$	repeatp: <0.001 avineff: 1043.4 AR: 0.001 repeatc: 0.09 runtime: 2.0
12	$\phi = [\rho_b]_{b=1}^{S}$ $\psi = [\mu, \sigma, \rho_Z]$	$\phi \sim \mathcal{N}_{19}(\varphi, \Sigma)$ $\varphi = [0.2]_{b=1}^{19}$, $\Sigma = VC$ $\psi \sim \mathcal{N}_3(\upsilon, \Upsilon)$ $\upsilon = (7.714, 0.839, 0.990)'$	$VC = \frac{1}{\lambda}\mathcal{I}_{19}$ $\lambda = 1000$ $\Upsilon = \text{diag}(0.1, 0.01, 0.001)\mathcal{I}_3$	$-\frac{\lambda}{2}(\phi - \varphi)^2$ $-\frac{1}{2}(\psi - \upsilon)'\Upsilon^{-1}(\psi - \upsilon)$	$\theta^* \sim \mathcal{N}_{S+3}(\theta^{(t)}, \Omega)$ $\Omega = \text{diag}(0.0001, \ldots, 0.0001,$ $0.001, 0.0001, 0.00001)\mathcal{I}_{22}$	repeatp: <0.001 avineff: 217.0 AR: 0.020 repeatc: <0.001 runtime: 2.0
13	$\phi = [\rho_b]_{b=1}^{S}$ $\psi = [\mu, \sigma, \rho_Z]$	$\phi \sim \mathcal{N}_{19}(\varphi, \Sigma)$ $\varphi = [0.2]_{b=1}^{19}$, $\Sigma = VC$ $\psi \sim \mathcal{N}_3(\upsilon, \Upsilon)$ $\upsilon = (7.714, 0.839, 0.990)'$	$VC = \frac{1}{\lambda}\mathcal{I}_{19}$ $\lambda = 1000$ $\Upsilon = \text{diag}(0.1, 0.01, 0.001)\mathcal{I}_3$	$-\frac{\lambda}{2}(\phi - \varphi)^2$ $-\frac{1}{2}(\psi - \upsilon)'\Upsilon^{-1}(\psi - \upsilon)$	$\theta^* \sim \mathcal{N}_{S+3}(\theta^{(t)}, \Omega)$ $\Omega = \text{diag}(0.0001, \ldots, 0.0001,$ $0.001, 0.0001, 0.00001)\mathcal{I}_{22}$ AP[1]	repeatp: <0.001 avineff: 138.9 AR: 0.489 repeatc: 1.617 runtime: 3.0
14	$\phi = [\rho_b]_{b=1}^{S}$ $\psi = [\mu, \sigma, \rho_Z]$	$\phi \sim \mathcal{N}_{19}(\varphi, \Sigma)$ $\varphi = [0.2]_{b=1}^{19}$, $\Sigma = VC$ $\psi \sim \mathcal{N}_3(\upsilon, \Upsilon)$ $\upsilon = (7.714, 0.839, 0.990)'$	$VC = \frac{1}{\lambda}\mathcal{I}_{19}$ $\lambda = 1000$ $\Upsilon = \text{diag}(0.1, 0.01, 0.001)\mathcal{I}_3$	$-\frac{\lambda}{2}(\phi - \varphi)^2$ $-\frac{1}{2}(\psi - \upsilon)'\Upsilon^{-1}(\psi - \upsilon)$	$\theta^* \sim \mathcal{N}_{S+3}(\theta^{(t)}, \Omega)$ $\Omega = \text{diag}(0.0001, \ldots, 0.0001,$ $0.001, 0.0001, 0.00001)\mathcal{I}_{22}$ AM[2]	repeatp: <0.001 avineff: 127.5 AR: 0.485 repeatc: 1.446 runtime: 3.0
15	$\phi = [\rho_b]_{b=1}^{S}$ $\psi = [\mu, \sigma, \rho_Z]$	$\phi \sim \mathcal{N}_{19}(\varphi, \Sigma)$ $\varphi = [0.2]_{b=1}^{19}$, $\Sigma = VC$ $\psi \sim \mathcal{N}_3(\upsilon, \Upsilon)$ $\upsilon = (7.714, 0.839, 0.990)'$	$VC = \frac{1}{\lambda}\mathcal{I}_{19}$ $\lambda = 1000$ $\Upsilon = \text{diag}(0.1, 0.01, 0.001)\mathcal{I}_3$	$-\frac{\lambda}{2}(\phi - \varphi)^2$ $-\frac{1}{2}(\psi - \upsilon)'\Upsilon^{-1}(\psi - \upsilon)$	$\theta^* \sim \mathcal{N}_{S+3}(\theta^{(t)}, \Omega)$ $\Omega = \text{diag}(0.0001, \ldots, 0.0001,$ $0.001, 0.0001, 0.00001)\mathcal{I}_{22}$ 2-Block AM proposal[3]	repeatp: <0.001 avineff: 44.8 $AR_{\rho_1 - \rho_{11}}$: 0.523 $AR_{\rho_{12} - \psi}$: 0.494 repeatc: 0.647 runtime: 5.0

Notes: $MCMC$ samples sizes are $T = 10{,}000$ following $n_0 = 1000$ iterations considered as burn-in. Runtimes are given in hours.
repeatp: repeats necessary for starting values, repeatc: repeats necessary for candidate values, avineff: averaged inefficiency coefficient, AR: acceptance rate.
[1]Consider Adaptive Proposal (AP) with greedy start based on 10 accepted draws and covariance matrix update at each 1000th run based on the last 1000 iterations.
[2]Consider Adaptive Metropolis (AM) with greedy start based on 10 accepted draws and covariance matrix update after 1000 runs at each iteration based on all preceding iterations.
[3]Consider a 2-Block AM with greedy start based on 10 accepted draws and covariance matrix update after 1000 runs at each iteration based on all preceding iterations. The two blocks are fixed as $\theta_1 = [\rho_1, \ldots, \rho_{11}]$ and $\theta_2 = [\rho_{12}, \ldots, \psi]$.

In trial 15, it is to be checked whether updating can be improved by additional blocking. For this purpose, the model parameters are arbitrarily split into two blocks according to their sequential order (ρ_{1-11} and $\rho_{12-19,\psi}$). The mean vector of the Gaussian proposal density is split into $\theta_k^{(t-1)}$ and $\theta_{-k}^{(t-1)}$, and the complete covariance matrix Σ_q is split into the smaller matrices Σ_k and Σ_{-k}. For example, the mean vector and the covariance matrix of the first block proposal ($k = 1, \ldots, 11$) are now defined as

$$\theta_k^{(t-1)} = \theta_1^{(t-1)}, \ldots, \theta_{11}^{(t-1)}; \qquad \Sigma_k = \left\{ \begin{array}{cccc} \Sigma_{1,1} & \Sigma_{1,2} & \ldots & \Sigma_{1,11} \\ \Sigma_{2,1} & \Sigma_{2,2} & \ldots & \Sigma_{2,11} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{11,1} & \Sigma_{11,2} & \ldots & \Sigma_{11,11} \end{array} \right\}.$$

The remaining specifications of trial 15 refer to those of trial 14. That is, an *AM*-Proposal is used with $c = 0.0001$ and $t_0 = 1000$. Estimates derived from trial 15 are slightly better compared to the foregoing ones, see Table 3.7. The very wide CI bands are a little bit smaller than those of trials 13 and 14 (cp. Figures 3.4 and 3.5). The acceptance rate increases to 52.3% for ρ_{1-11} and 49.4% for $\rho_{12-19,\psi}$ (cp. Table 3.4). The averaged *Ineff* is the smallest (44.8) compared to the previous updating set-ups. These improvements are attributable to the grouping of the parameters into two blocks.

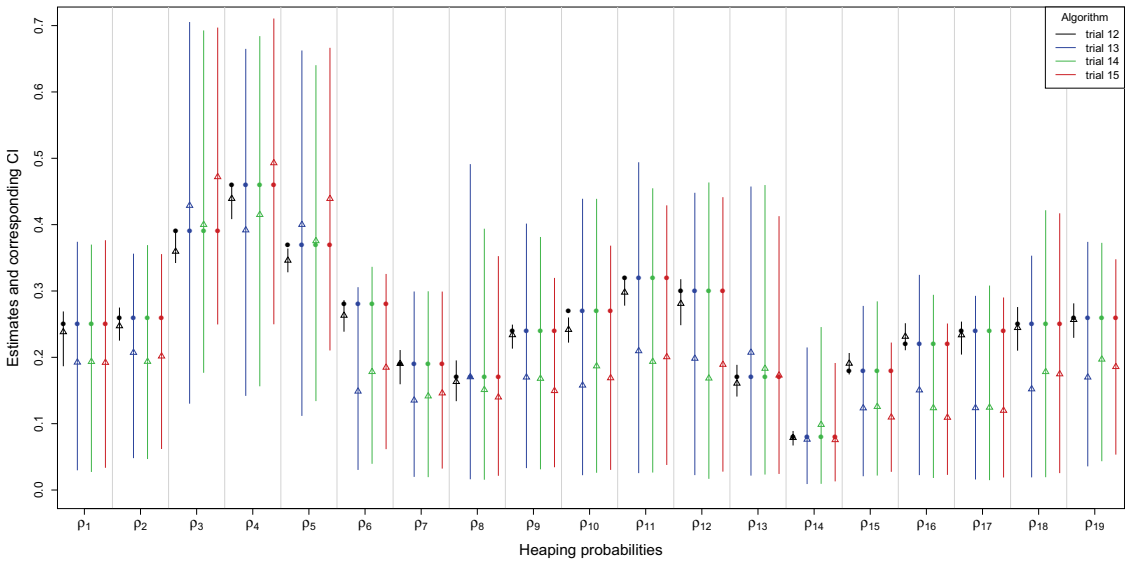


Figure 3.4: Posterior means with 95% confidence intervals (CI) of four adaptive random-walk MH algorithms for the heaping probabilities. The true parameter values are marked by a dot, parameter estimates are marked by a triangle.

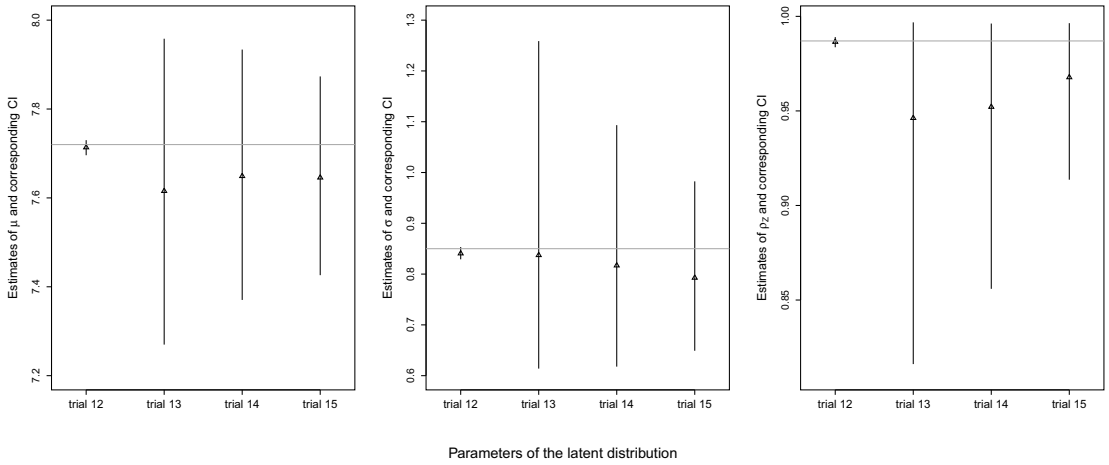


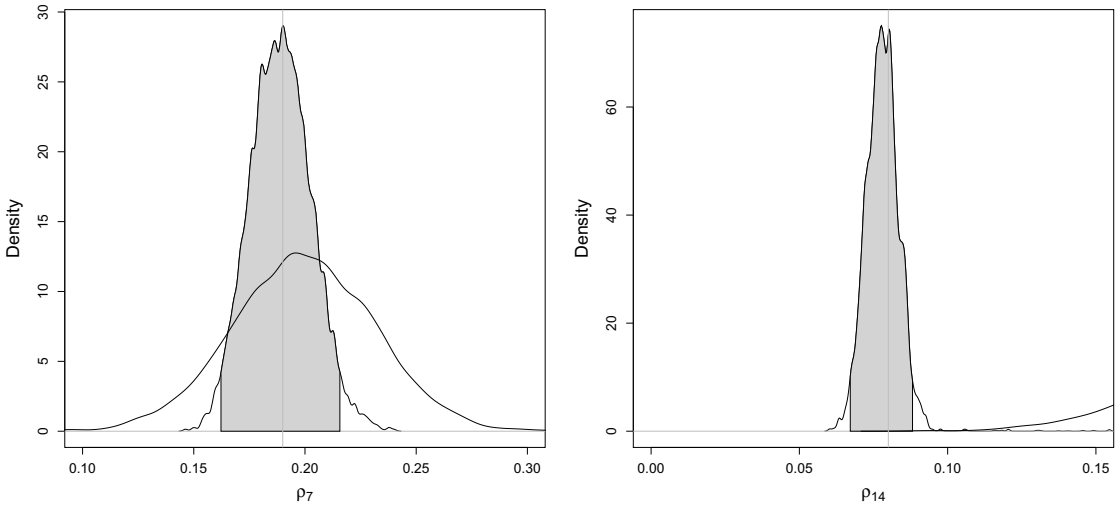
Figure 3.5: Posterior means with 95% confidence intervals (CI) of four adaptive random-walk MH algorithms for the parameters of the underlying distribution. The true parameter values are marked by a dashed line, parameter estimates are marked by a triangle.

Results of graphical inspection and convergence

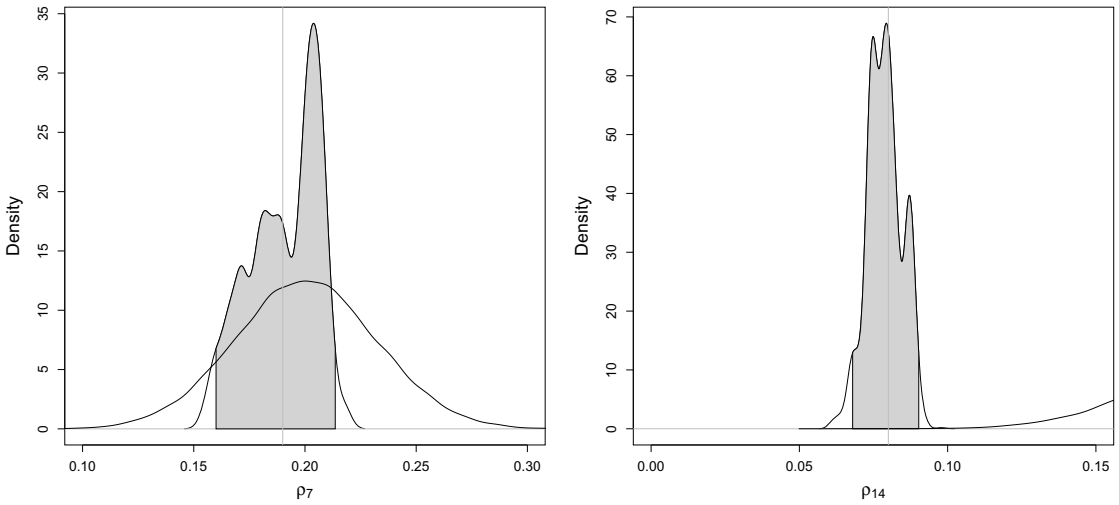
Exemplarily, HDR plots of selected parameters are given to receive more insight into the behavior of the $MCMC$ samplers employed. Figure 3.6 shows the kernel smoothed histogram of the marginal posterior density for ρ_7 and ρ_{14} along with the corresponding prior distribution derived from trials 10 and 12. The marginal posterior densities for both parameters are precisely centered around the true parameter values from the DGP indicating high estimation accuracy of these trials with respect to both parameters.

In contrast, Figure 3.7 shows the HDR plots for ρ_4 and ρ_{10}, two parameters that are unsatisfactory estimated (again stemming from trials 10 and 12). These estimates seem to be a little bit problematic, because all considered schemes fail to include the true parameter into the range of posterior mean $\pm$ one SD. As can be seen, the true parameter value of ρ_4 is at least within the 95% HDR, whereas the true parameter value of ρ_{10} clearly lies outside.

The corresponding trace plots and ACF plots of all trials are given in Figures A.8 to A.37 in the Appendix. The effect of the starting values wears off within less than 1000 iterations in trials 1–10 (cp. Figures A.8 to A.17), but not in the trials 11–15 with multivariate normal proposal density. This corresponds to the very low acceptance rates of trials 11 and 12. In the trace plots of trials 13–15, it clearly can be seen when the update process triggers and that the amplitudes of the chain get much wider, which in turn indicates that the increments in the proposal covariance matrix are getting larger. Similar figures are found, e.g. by Mathew et al. (2012, p. 241).

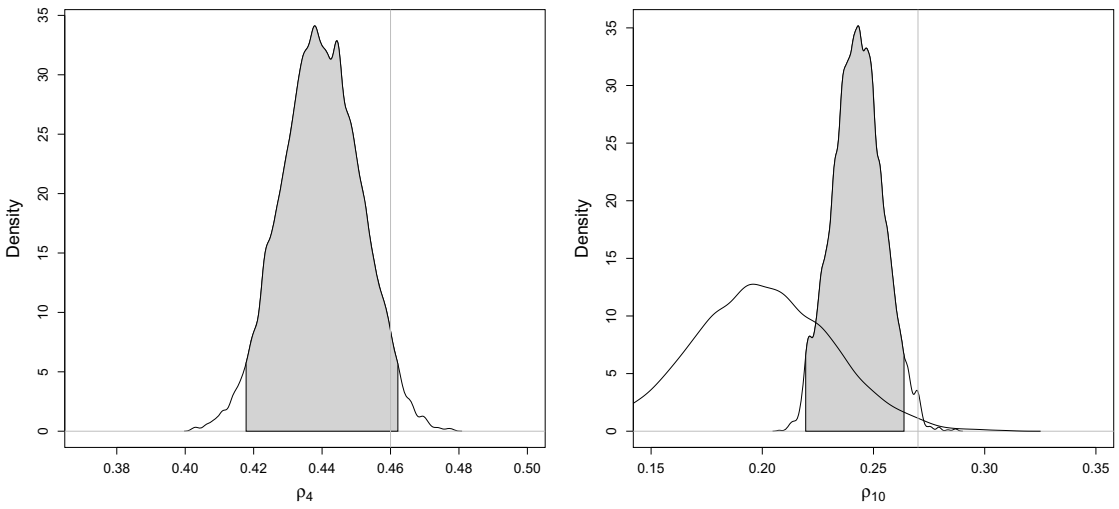


(a) Parameter values of trial 10.

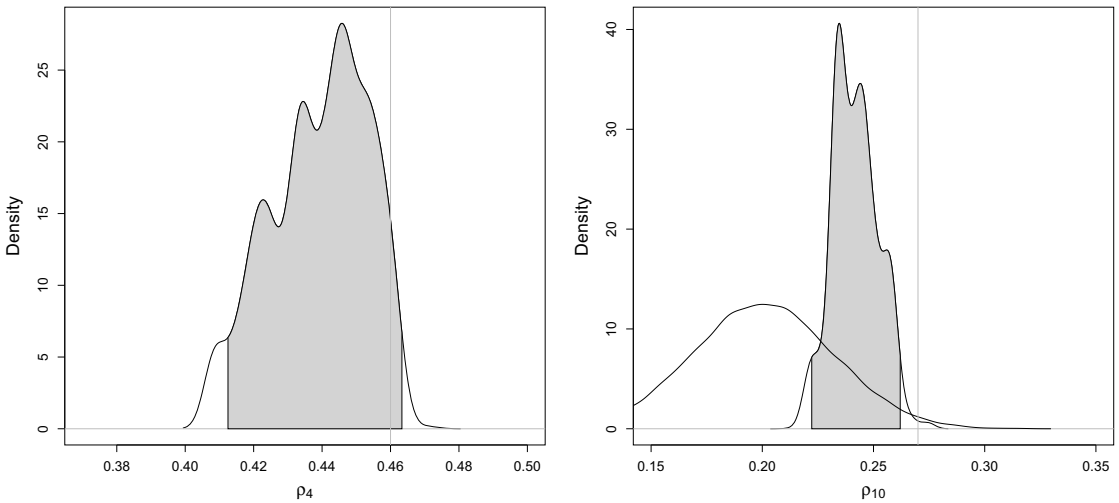


(b) Parameter values of trial 12.

Figure 3.6: Marginal prior-posterior plots for two well estimated parameter values of trial 10 (a) and trial 12 (b). Solid lines indicate posterior distributions, whereas the dashed lines indicate the prior distributions. The gray vertical lines indicate the position of the true parameter values. The 95% highest posterior density region (HDR) is shaded in gray.



(a) Parameter values of trial 10.



(b) Parameter values of trial 12.

Figure 3.7: Marginal prior-posterior plots for two unsatisfactory estimated parameter values of trial 10 (a) and trial 12 (b). Solid lines indicate posterior distributions, whereas the dashed lines indicate the prior distributions. The gray vertical lines indicate the position of the true parameter values. The 95% highest posterior density region (HDR) is shaded in gray.

The ACF plots in Figures A.23 to A.37 in the Appendix reveal that the performance of the RWM schemes is affected by the dimension of the posterior distribution and its complexity, especially when sampling from a multivariate normal proposal density. The sample ACF always decays to zero within 100 lags for most of the parameters in trials 2, 7–10, 14, and 15. Hence, these samplers seem to be efficient. Adding the parameters of the underlying true distribution when drawing samples for the posterior results in significant autocorrelation even at lags 200 for most of the parameters in trials 6 and 11. This effect alleviates when applying the blocking strategy. The sampled draws from trials 3–5, 12, and 13 are highly persistent owing to a slowly mixing sampler (trials 3–5 and 12) and the large amplitudes (trial 13). This fact underlines the assumption that the S-RWM algorithm does not explore the high density regions of the target distribution as efficient as in the MB-RWM algorithms, especially when using the multivariate normal proposal density, as already demonstrated in Chib and Ramamurthy (2010, p. 19).

The behavior of the $MCMC$ samplers of trials 13 and 14 is highly unsatisfactory. To check whether convergence can be achieved over a longer run, both trials are further run with $n_0 + T = 105{,}000$ iterations. The acceptance rates are quite the same for trial 13 (48.7%) and trial 14 (48.5%). Comparison of the empirical correlation and covariance matrices indicates now dependencies among at least half of the model parameters, see Equations (A.5) to (A.8) in the Appendix. All correlations exceeding a value of 0.06 show the same direction and almost the same magnitude, cp. Equations (A.5) and (A.6). All other correlations seem to be noise. The correlations get more pronounced when thinning the output ($m = 5$) to reduce the dependencies between consecutive draws, cp. the correlation matrices in Equations (A.9) and (A.10).

The ACF is reduced considerably when thinning the $MCMC$ output, as expected, see Figures A.38 to A.39 in the Appendix. The ACF for the thinned AP_5 decays to zero within 50 lags for most of the parameters, and for the thinned AM_5 within 20 lags. However, even with increased $MCMC$ sample size, the estimated parameter values do not indicate an improvement in estimation accuracy for the adaptive RWM schemes. The main criticism against adaptive Gaussian proposals owing to model complexity and high dimensionality, see e.g. Gilks et al. (1998, p. 1052), might apply to these results. To check this supposition, an example with lower dimensionality and complexity is introduced. The findings of the Excursion support the assumption that adaptive schemes work well for models with a small number of model parameters.

Excursion: Downsized heaping model, simulation and estimation

To illustrate the effectiveness of adaptive schemes for Gaussian proposals in the RWM algorithm, a downsized version of the heaping model has been considered. The Dagum distribution has been assumed as latent true distribution with the following parameter specification: $a = 3.6$, $b = 2416$, and $p = 0.43$. The sample size is $n = 10{,}000$ and the heaping mechanism corresponds to that of Equation (2.2) in Section 2.2, but with two ρ_b to describe heaping to mod(1000) in the range of $[1000, 10{,}000)$. In the interval $[1000, 5000)$, the probability to heap is $\rho_1 = 0.23$ and in the second interval $[5000, 10{,}000)$, ρ_2 is set to 0.34. The intervals do not overlap. Thus, no constraint is needed that restricts the probabilities to not exceed one in sum. The other constraints remain the same, see Section 2.3. The parameters of the underlying true distribution are kept fix. Hence, the example is still a finite mixture model but small in the meaning of low dimensionality.

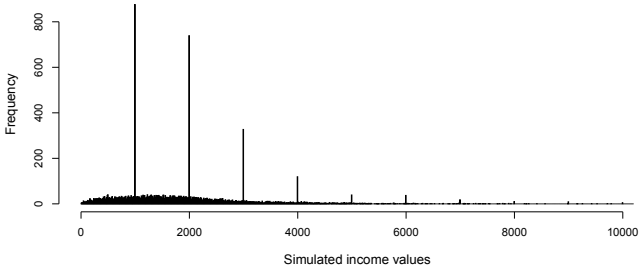


Figure 3.8: Data example of the downsized heaping model.

The percentages of heaped values in the small simulation example are 20.39% in the interval $[1000, 5000)$ and 0.77% in the interval $[5000, 10{,}000)$, yielding a proportion of heaped values of 21.16% in total, see Figure 3.8. A Gaussian proposal scheme without adaption and with adaption and a greedy start procedure after 100 accepted draws has been employed. Runtime for the small example has been 59 minutes for one scheme. When comparing the corresponding trace plots of both settings, see Figure 3.9, it is obvious that the behavior of the chain changes immediately at the iteration where the covariance matrix of the proposal distribution has been updated. The updated $MCMC$ sampler explores the parameter space more efficiently, as also indicated by the higher acceptance rate (2.6% vs. 29.7%) and decreased *Ineff* (33.5 vs. 17.6).

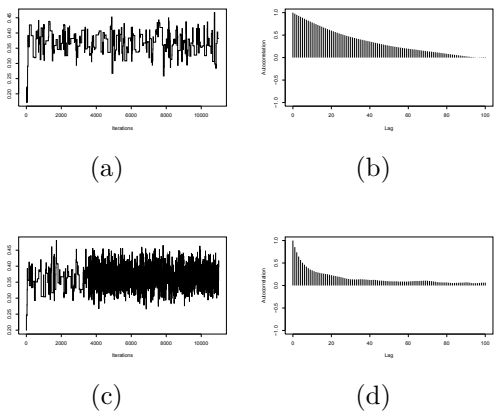


Figure 3.9: Trace plots (a, c) and ACF plots (b, d) for ρ_2 in the RWM algorithm with multivariate normal proposal density without (a, b) and with update (c, d) of the covariance matrix of the proposal density. $MCMC$ sample size is $n_0 + T = 11{,}000$, and the acceptance rates are 2.6% vs. 29.7%, with corresponding inefficiency coefficients 33.5 vs. 17.6.

Table 3.5: Posterior summaries for different tunings of the original RWM algorithm.

		Trial 1			**Trial 2**			**Trial 3**			**Trial 4**			**Trial 5**		
Par	θ	$\widehat{\theta}$	*SD*	*Ineff*	$\widehat{\theta}$	*SD*	*Ineff*	$\widehat{\theta}$	*SD*	*Ineff*	$\widehat{\theta}$	*SD*	*Ineff*	$\widehat{\theta}$	*SD*	*Ineff*
ρ_1	0.250	0.283	0.021	98.4	0.254	0.018	153.4	0.202	0.015	980.7	0.223	0.018	114.9	0.286	0.019	82.0
ρ_2	0.260	0.259	0.013	35.9	0.254	0.013	53.7	0.259	0.019	825.7	0.254	0.013	101.8	0.255	0.013	34.3
ρ_3	0.390	0.391	0.013	106.8	0.362	0.012	66.4	0.364	0.007	117.0	0.348	0.013	106.5	0.373	0.013	61.9
ρ_4	0.460	0.482	0.012	91.7	0.439	0.012	57.3	0.436	0.003	78.4	0.430	0.014	128.9	0.451	0.013	53.4
ρ_5	0.370	0.363	0.012	53.8	0.345	0.011	67.5	0.344	0.017	1027.0	0.340	0.011	136.6	0.350	0.011	56.9
ρ_6	0.280	0.267	0.014	56.5	0.265	0.014	46.1	0.281	0.012	440.2	0.258	0.012	47.5	0.274	0.012	36.5
ρ_7	0.190	0.176	0.015	132.8	0.187	0.014	49.7	0.219	0.013	330.4	0.175	0.014	50.8	0.201	0.015	68.0
ρ_8	0.170	0.157	0.012	47.8	0.164	0.012	44.2	0.170	0.010	872.5	0.152	0.010	27.4	0.175	0.014	64.1
ρ_9	0.240	0.228	0.010	66.3	0.239	0.010	23.2	0.230	0.012	723.4	0.235	0.010	87.6	0.241	0.010	27.9
ρ_{10}	0.270	0.245	0.013	49.1	0.242	0.012	39.5	0.242	0.014	591.4	0.235	0.011	129.3	0.253	0.012	62.1
ρ_{11}	0.320	0.318	0.017	77.6	0.293	0.015	57.5	0.316	0.014	518.4	0.282	0.014	133.5	0.308	0.015	64.0
ρ_{12}	0.300	0.309	0.021	210.1	0.277	0.017	89.6	0.245	0.013	694.3	0.256	0.017	155.6	0.298	0.018	102.6
ρ_{13}	0.170	0.150	0.013	35.9	0.153	0.016	68.5	0.181	0.011	533.9	0.154	0.015	114.9	0.152	0.015	63.1
ρ_{14}	0.080	0.073	0.005	63.8	0.078	0.006	43.0	0.079	0.008	387.2	0.077	0.005	22.5	0.079	0.005	34.4
ρ_{15}	0.180	0.178	0.007	22.5	0.190	0.007	39.8	0.192	0.006	411.6	0.189	0.008	48.5	0.190	0.007	43.9
ρ_{16}	0.220	0.222	0.009	24.6	0.229	0.010	33.2	0.233	0.004	286.5	0.224	0.010	65.3	0.231	0.010	39.2
ρ_{17}	0.240	0.238	0.012	85.3	0.240	0.012	38.9	0.234	0.013	992.2	0.232	0.012	95.0	0.245	0.012	48.6
ρ_{18}	0.250	0.248	0.017	84.5	0.242	0.015	69.3	0.272	0.007	99.0	0.225	0.016	93.8	0.255	0.014	72.4
ρ_{19}	0.260	0.267	0.013	68.5	0.256	0.012	41.4	0.244	0.002	37.6	0.242	0.011	112.9	0.273	0.013	49.3

Notes: The parameters of the latent distribution are kept fix to the true ones, $\psi = [\mu = 7.72, \sigma = 0.85, \rho_Z = 0.987]$.
Posterior means are based on $T = 10{,}000$ iterations following a burn-in period of $n_0 = 1000$ iterations. Inefficiency factors (*Ineff*) are based on 5000 lags.

Table 3.6: Posterior summaries for the blocking strategy.

		Trial 6			**Trial 7**			**Trial 8**			**Trial 9**			**Trial 10**		
Par	θ	$\widehat{\theta}$	SD	*Ineff*	$\widehat{\theta}$	SD	*Ineff*	$\widehat{\theta}$	SD	*Ineff*	$\widehat{\theta}$	SD	*Ineff*	$\widehat{\theta}$	SD	*Ineff*
ρ_1	0.250	0.240	0.020	326.8	0.250	0.018	49.3	0.253	0.019	43.1	0.252	0.018	38.4	0.252	0.017	27.7
ρ_2	0.260	0.247	0.011	219.7	0.247	0.012	38.0	0.247	0.012	14.4	0.248	0.012	21.5	0.248	0.012	16.3
ρ_3	0.390	0.359	0.013	274.4	0.361	0.013	65.8	0.362	0.012	19.8	0.360	0.012	19.2	0.360	0.012	29.8
ρ_4	0.460	0.442	0.013	352.4	0.441	0.012	53.7	0.440	0.012	34.7	0.441	0.012	31.6	0.440	0.012	34.1
ρ_5	0.370	0.347	0.010	196.4	0.345	0.011	37.0	0.346	0.011	15.2	0.346	0.011	19.4	0.346	0.010	24.0
ρ_6	0.280	0.260	0.010	125.4	0.264	0.013	29.7	0.265	0.012	15.7	0.265	0.012	13.4	0.264	0.012	14.0
ρ_7	0.190	0.191	0.013	201.1	0.190	0.014	21.0	0.190	0.014	18.9	0.190	0.014	21.7	0.189	0.014	18.1
ρ_8	0.170	0.161	0.011	307.3	0.164	0.012	14.6	0.164	0.012	8.2	0.163	0.011	16.1	0.165	0.012	13.8
ρ_9	0.240	0.235	0.010	146.8	0.238	0.010	20.5	0.238	0.010	15.2	0.238	0.010	17.8	0.239	0.010	12.3
ρ_{10}	0.270	0.238	0.009	123.9	0.244	0.011	16.6	0.243	0.011	19.9	0.243	0.013	20.3	0.243	0.011	15.2
ρ_{11}	0.320	0.300	0.014	229.8	0.297	0.015	25.4	0.296	0.015	23.0	0.296	0.014	19.8	0.297	0.014	23.1
ρ_{12}	0.300	0.281	0.019	444.9	0.274	0.016	34.1	0.277	0.017	25.4	0.278	0.017	44.6	0.277	0.018	24.7
ρ_{13}	0.170	0.161	0.012	264.9	0.161	0.014	19.4	0.160	0.014	16.9	0.162	0.014	18.9	0.161	0.013	15.4
ρ_{14}	0.080	0.079	0.005	83.4	0.078	0.005	19.6	0.078	0.005	10.8	0.078	0.005	7.6	0.078	0.005	16.0
ρ_{15}	0.180	0.191	0.009	221.1	0.190	0.008	25.7	0.190	0.008	8.3	0.189	0.008	7.6	0.189	0.008	10.4
ρ_{16}	0.220	0.229	0.010	180.6	0.229	0.010	15.8	0.228	0.010	10.5	0.227	0.009	8.9	0.227	0.010	16.6
ρ_{17}	0.240	0.237	0.012	149.0	0.239	0.013	21.6	0.238	0.012	15.0	0.238	0.012	13.7	0.237	0.012	18.1
ρ_{18}	0.250	0.237	0.011	311.6	0.241	0.015	46.6	0.241	0.015	17.8	0.239	0.015	20.9	0.241	0.015	20.5
ρ_{19}	0.260	0.258	0.012	238.2	0.257	0.012	24.7	0.257	0.013	11.2	0.258	0.012	12.9	0.258	0.012	23.0
μ	7.720	7.712	0.009	17.3	7.712	0.009	6.2	7.711	0.008	6.8	7.711	0.008	5.6	7.712	0.009	6.5
σ	0.850	0.841	0.006	53.2	0.840	0.006	21.4	0.840	0.006	19.7	0.841	0.006	46.7	0.840	0.006	7.5
ρ_Z	0.987	0.987	0.001	211.3	0.987	0.001	76.0	0.987	0.001	117.8	0.986	0.001	92.2	0.987	0.001	44.9

Notes: Posterior means are based on $T = 10{,}000$ iterations following a burn-in period of $n_0 = 1000$ iterations. Inefficiency factors (*Ineff*) are based on 5000 lags.

Table 3.7: Posterior summaries for the adaptive $MCMC$ schemes.

		Trial 11			**Trial 12**			**Trial 13**			**Trial 14**			**Trial 15**		
Par	θ	$\widehat{\theta}$	SD	*Ineff*	$\widehat{\theta}$	SD	*Ineff*	$\widehat{\theta}$	SD	*Ineff*	$\widehat{\theta}$	SD	*Ineff*	$\widehat{\theta}$	SD	*Ineff*
ρ_1	0.250	0.261	0.012	613.3	0.238	0.020	390.9	0.193	0.092	54.9	0.193	0.094	45.4	0.192	0.093	29.5
ρ_2	0.260	0.267	0.037	1799.0	0.247	0.013	241.9	0.207	0.081	272.0	0.194	0.086	23.0	0.202	0.076	66.5
ρ_3	0.390	0.365	0.011	1486.5	0.360	0.011	128.2	0.429	0.141	59.6	0.400	0.131	310.7	0.472	0.117	65.2
ρ_4	0.460	0.404	0.037	1760.0	0.439	0.014	248.9	0.392	0.133	51.7	0.415	0.140	321.2	0.493	0.117	31.2
ρ_5	0.370	0.351	0.018	1307.6	0.346	0.010	232.6	0.400	0.139	117.4	0.376	0.129	36.6	0.439	0.116	28.4
ρ_6	0.280	0.279	0.028	1445.3	0.263	0.012	237.4	0.149	0.070	88.6	0.178	0.078	330.2	0.185	0.069	60.1
ρ_7	0.190	0.191	0.043	1830.2	0.191	0.015	377.2	0.135	0.076	47.2	0.142	0.075	67.7	0.146	0.070	33.5
ρ_8	0.170	0.179	0.011	936.0	0.163	0.013	250.4	0.170	0.126	214.6	0.151	0.099	34.1	0.140	0.088	27.5
ρ_9	0.240	0.237	0.026	1856.7	0.234	0.010	114.3	0.170	0.094	86.3	0.168	0.090	23.4	0.150	0.076	146.1
ρ_{10}	0.270	0.208	0.017	265.4	0.242	0.010	212.0	0.158	0.107	480.4	0.187	0.111	23.7	0.169	0.088	32.3
ρ_{11}	0.320	0.283	0.019	1245.4	0.298	0.014	236.5	0.210	0.127	77.1	0.194	0.114	62.2	0.201	0.103	24.3
ρ_{12}	0.300	0.251	0.021	1048.3	0.281	0.019	220.8	0.198	0.109	58.1	0.168	0.121	246.4	0.189	0.109	28.6
ρ_{13}	0.170	0.149	0.009	181.5	0.161	0.014	173.9	0.207	0.119	174.5	0.183	0.115	42.4	0.173	0.102	90.0
ρ_{14}	0.080	0.090	0.009	319.4	0.079	0.006	57.4	0.076	0.053	75.1	0.099	0.064	361.9	0.076	0.046	31.8
ρ_{15}	0.180	0.206	0.020	1650.0	0.191	0.008	212.6	0.124	0.068	91.3	0.126	0.069	311.3	0.110	0.052	31.0
ρ_{16}	0.220	0.220	0.010	560.1	0.231	0.011	608.3	0.151	0.080	71.0	0.124	0.072	34.1	0.109	0.059	26.2
ρ_{17}	0.240	0.207	0.012	474.8	0.234	0.010	125.9	0.124	0.072	45.5	0.124	0.077	30.2	0.120	0.071	47.9
ρ_{18}	0.250	0.313	0.018	539.7	0.245	0.014	256.8	0.152	0.087	44.0	0.178	0.108	105.7	0.175	0.102	31.8
ρ_{19}	0.260	0.250	0.024	1576.7	0.257	0.015	287.2	0.170	0.089	403.5	0.197	0.089	44.3	0.186	0.075	26.8
μ	7.720	7.709	0.010	1300.6	7.714	0.008	26.8	7.616	0.167	33.2	7.649	0.144	69.8	7.646	0.115	18.9
σ	0.850	0.841	0.009	408.2	0.841	0.006	84.0	0.837	0.151	65.2	0.817	0.121	60.8	0.793	0.086	69.1
ρ_Z	0.987	0.985	0.001	351.1	0.987	0.001	49.3	0.946	0.048	444.6	0.952	0.039	220.7	0.968	0.022	39.5

Notes: Posterior means are based on $T = 10{,}000$ iterations following a burn-in period of $n_0 = 1000$ iterations. Inefficiency factors (*Ineff*) are based on 5000 lags.

Finally, lets have a look at the convergence diagnostic of Geweke (1992), see Table 3.8. According to this diagnostic, trials 3 and 11 do not converge for most of the parameter estimates. Trials 2, 5, 8, 9, and 12–15 seem to be problematic as well since here – for one or more parameter estimates – the corresponding $MCMC$ sampler does not converge at a 99%-level (three $SD = \pm 2.58$). In contrast, trials 1, 4, 6, and 10 converge for most of the parameter estimates at a 95%-level, and for two or three parameter estimates at least at level 99%. The best convergence behavior is found for trial 7. Here, the posterior distribution converges for 21 out of 22 parameter estimates within a significance level of 90%, and only for one parameter estimate at a 95%-significance-level. These findings are not surprising, because several trials showing lack of convergence have already been shown to be problematic with respect to efficiency and acceptance rates.

The criterion of S. P. Brooks and Gelman (1998) for determining convergence based on multiple independent chains is to be explored in Section 3.3. Due to the high computational burden, because runs have to be performed multiple times, it is not pursued further here.

Summary

The aim of this sections was the exploration of different tuning parameters and specifications of the RWM algorithm and their effect on the outcome. For this purpose, a substantial number of trials was generated with distinct specifications. Their performance was measured and the best ones are selected for further analyses outspread in the next sections. Selection of the best settings ensues from the predictive accuracy of their estimates, their acceptance rates, convergence behavior, and efficiency. According to these factors, trial 10 outperforms all other trials. The samplers of trials 7–9 appear to be efficient as well, because they make large moves across the support of the target density while generating reasonable acceptance rates at the same time. The results reinforce the comments given by Chib and Ramamurthy (2010) that the problem of high-dimensional proposal densities can be overcome by dividing the parameters of interest into separate blocks and by running dependent cycles for each block during one iteration.

The trace plots indicate that the sample draws all in all quickly move to the target distribution. In other words, the Markov chain sample size of $T = 10{,}000$ with a burn-in of $n_0 = 1000$ iterations seems to be sufficient to explore the target distribution globally in the considered schemes. Overall, this indicates the likelihood function to carry substantial information about the parameter values. With respect to the ACF plots, trial 10 shows the lowest dependencies between the $MCMC$ draws. The previous results are supported by the corresponding convergence behaviors of the different trials. No indication for lack of convergence is obvious for trials 1, 4, 6, 7, and 10, because most of the parameter estimates converge within a significance level of 95%-level and two or three parameter estimates converge at least within a significance level of 99%.

Table 3.8: Geweke's convergence criterion for all $MCMC$ samplers.

Par	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	Trial 6	Trial 7	Trial 8	Trial 9	Trial 10	Trial 11	Trial 12	Trial 13	Trial 14	Trial 15
ρ_1	0.924	3.814	5.741	0.393	1.291	-1.050	0.387	2.166	1.367	-1.465	3.373	-0.121	0.868	1.496	0.039
ρ_2	-0.006	0.385	6.978	0.998	-0.379	-2.146	-0.566	0.332	0.986	-0.283	13.068	-0.927	7.101	0.031	2.203
ρ_3	-0.582	0.288	-7.650	-1.928	1.149	-0.147	0.250	0.768	0.218	-0.847	-38.645	0.083	3.023	-5.577	-2.601
ρ_4	0.647	0.016	-2.574	0.097	1.263	-0.007	0.310	0.038	1.381	0.411	-7.477	4.686	-0.734	-3.014	1.993
ρ_5	-0.391	-0.843	31.634	-1.975	0.042	1.077	-0.733	1.354	-1.288	-0.336	5.048	1.558	3.029	-1.086	0.173
ρ_6	-1.239	-0.025	4.190	1.397	-0.450	-0.823	-0.864	-2.191	0.301	-0.312	4.152	-0.161	-0.063	5.729	1.848
ρ_7	-2.056	0.475	0.954	0.313	0.069	-0.658	0.802	-0.184	1.992	0.909	-11.255	-2.443	-2.870	3.671	-0.804
ρ_8	-0.654	-1.044	-4.467	1.065	-1.330	1.207	0.302	0.133	-1.134	-1.607	2.208	-0.927	-4.832	-2.581	-2.510
ρ_9	0.661	-0.912	14.315	0.834	-1.200	0.545	-0.222	-1.428	-0.955	0.046	13.128	-0.818	-0.325	1.175	1.928
ρ_{10}	0.713	0.127	-3.706	2.479	0.921	-1.186	0.661	-2.759	0.830	1.404	-0.660	0.971	-5.011	1.876	0.929
ρ_{11}	1.853	0.672	2.074	0.875	-0.070	0.297	0.677	0.789	-0.782	0.449	-16.445	0.123	-1.135	-1.904	-1.588
ρ_{12}	1.771	1.200	1.858	-0.132	0.695	1.116	0.357	0.078	-3.010	-0.447	3.855	2.514	2.166	-7.667	1.031
ρ_{13}	-1.304	1.286	0.051	0.309	2.705	0.092	0.025	0.163	-0.081	2.577	-1.909	0.728	3.760	-2.435	-2.350
ρ_{14}	0.622	0.433	-1.046	-1.039	0.803	0.370	-0.370	-0.160	0.468	0.232	-2.619	0.377	-0.551	2.210	-1.528
ρ_{15}	-2.492	0.407	-49.004	-0.465	-1.184	0.108	-0.163	1.695	0.868	-0.097	9.942	-1.949	2.814	4.400	0.143
ρ_{16}	1.338	0.168	5.181	-0.884	-0.998	2.081	-0.280	0.710	-1.730	-2.415	-1.943	-4.370	1.946	-1.500	-0.747
ρ_{17}	-1.166	0.484	-56.101	-0.817	-0.626	-1.998	0.751	1.167	1.601	-0.799	2.067	-0.657	0.007	0.249	-1.876
ρ_{18}	1.092	-1.779	-4.109	0.445	1.763	1.393	-1.809	1.305	0.439	-0.894	-1.145	-0.701	-1.577	2.075	-0.347
ρ_{19}	0.625	-1.102	-1.566	0.981	0.012	-0.885	-1.196	-0.462	-0.199	-1.667	-7.309	1.340	-8.046	0.061	0.533
μ	–	–	–	–	–	-1.245	-1.196	-0.762	-0.684	-0.204	20.608	1.148	0.940	1.947	0.855
σ	–	–	–	–	–	0.838	-0.066	0.656	2.223	0.450	-2.231	-2.259	-2.054	-1.852	-3.334
ρ_Z	–	–	–	–	–	0.162	0.351	0.933	-0.559	-1.088	1.936	-0.324	8.124	3.793	3.973

Notes: Geweke's convergence diagnostics are based on $T = 10{,}000$ iterations following a burn-in period of $n_0 = 1000$ iterations. Window 1 is set to 25% and window 2 to 50%.

The uniform proposal density yields very good approximations to the true parameters. Concretely, the true parameters lie within one SD of the posterior mean for most of the parameters. Additionally, the averaged *Ineff* are the smallest and it exhibits good convergence behavior (cp. Chib & Greenberg, 1995, p. 329). However, the uniform proposal density has a strong theoretical drawback. It does not ensure that the whole support of the target distribution is covered at each iteration, especially not the tails. This in particular means that the irreducibility condition is not met, and the log-marginal likelihood becomes incalculable in the consequence.

Choosing the multivariate normal distribution as proposal density ensures irreducibility, but appears to be problematic in higher dimensional problems, even when considering adaption of the proposal covariance matrix. Only when introducing blocking, the efficiency gain is remarkable. This fact strongly supports usage of multiple-block proposals over adaptive settings. The results obtained clearly show that some schemes of the RWM algorithm perform exceptionally well for estimation of a posterior distribution with a finite mixture likelihood.

Overall, the posterior means and corresponding SD or CI from the different RWM schemes considered capture the true parameters well. Only few model parameters are underestimated (ρ_3–ρ_6 and ρ_{10}–ρ_{12} as well as σ). Worth recognizing is that the overall accuracy of estimation is high for all RWM schemes of the tuning and blocking set-ups. Remarkably are also simulation efficiency and convergence. Only the adaptive schemes fall behind with respect to accuracy in parameter estimation. Even consideration of longer runs gave no improvement. This might be, to a large extent, due to the scaling constant c in the algorithm. Vihola (2011, p. 48) states that the potential collapse of the covariance matrix to singularity does not tend to happen in general thus making c unnecessary. Furthermore, the author argues that such a predefined lower bound on the adapted covariance matrix can deteriorate efficiency of the sampler. The author hence proposes an unconstrained AM algorithm. A more deeply investigation of the effect of the scaling constant c can be found in Herbst and Schorfheide (2015). According to the authors, c follows an U-shape function regarding the accuracy of the posterior mean approximation. Highest precision (the smallest value of c) is attained for 0.5, with respect to the DSGE models considered. In the updating set-ups in this thesis, the problem was not singularity but noticing the adapted covariance matrix to become diffuse. Because of that, the scaling constant c was set to 0.001 or 0.0001, respectively. Additionally, it could be checked in further studies whether the scaling factor $s_D = 2.38/\sqrt{D}$ according to Gelman, Roberts, and Gilks (1996) is appropriate for estimation of the proposed heaping model.

With regard to the inefficiency factors, it can be said that in trials 7–10 and 15 the *Ineff* most of the parameters look favorable, i.e. are smaller than 50, indicating high efficiency. That is, the different algorithms produce virtually *iid* draws for the model parameters. Considering trials 1, 2, 4, and 5, the *Ineff* is smaller than 100 (or 200 in most cases), also indicating acceptable efficiency. The remaining

trials, 3 and 12–14 have remarkably high *Ineff* indicating slow-mixing samplers or too large amplitudes. The fact that the *Ineff* of trial 12 are lower compared to those of trial 11 indicates that smaller step sizes in the proposal density seem to be preferable in higher dimensional problems. This is also confirmed by the higher acceptance rate. By far the best (averaged) *Ineff* was found for trial 10, all below a value of 40. Hence, tuning of trial 10 – and the other blocking schemes as well – leads to well-mixing samplers. Trials 10 and 15 have the best simulation efficiency and seem to be preferable to all other trials.

Herbst and Schorfheide (2015, p. 120) suggest an overall efficiency measure that regards the numerical and computational efficiency at the same time. The overall efficiency measure is calculated as $\frac{T}{\text{runtime (sec)}} \cdot \frac{1}{Ineff}$. This measure determines the number of *iid*-equivalent draws produced per unit of runtime. According to this, the M-RWM (trial 7) performs best. The M-RWM produces the most *iid*-equivalent draws per second (0.0055). This result is perpetuated when using the *Ineff* of the slowest mixing model parameter instead of the averaged *Ineff*, as proposed by Turek, Valpine, Paciorek, and Anderson-Bergman (2015, p. 13).

3.3 A comparison of ML and RWM estimation of the heaping model

Guided by the assumption that ML and $MCMC$ approaches behave different in estimation of models with finite mixture distribution, simulation is now repeated several times to compare statistical accuracy of ML and some selected RWM settings that proved to be efficient in the last section. The RWM schemes following a blocking strategy clearly outperformed the adaptive $MCMC$ schemes. Thus, the M-RWM, the I-RWM and the RMB-RWM with uniform and multivariate proposal densities are compared to the ML approach. The corresponding S-RWM schemes are used as benchmark.

Although the parameter estimates equal the true parameter values closely, there might be a remarkable error due to $MCMC$ simulation. Besides a convergence diagnostic and the bias, additionally two performance criteria are calculated to assess statistical accuracy: the mean squared error and coverage rates. For this purpose, 100 repetitions are run with foregoing simulation of the data. Each simulated data set is then estimated by means of ML and the RWM settings of trials 6–8, 10, and 12. Trial 12 was also considered as blocked proposal with M-RWM, I-RWM, and RMB-RWM. This sequence of settings is denoted by $\mathcal{A}_1, \ldots, \mathcal{A}_8$.

3.3.1 Convergence of multiple independent chains

The last chapter concluded with the convergence diagnostic according to Geweke (1992). It remained open to check the convergence rate with Brooks and Gelman's (1998) convergence criterion over multiple independent runs. For this purpose, all 100 repetitions of trials $\mathcal{A}_1, \ldots, \mathcal{A}_8$ are used as multiple independent chains.

As can be seen in Table 3.9, the $MCMC$ samplers with blocking strategy converge for all parameters of the heaping mechanism (the $PSRF$ are well below the threshold of 1.2, or even 1.1). This holds for the S-RWM with multivariate normal proposal density as well. In the S-RWM with uniform proposal density, ρ_1 and ρ_{12} exceed the threshold of 1.2 to some extent. The $PSRF$ of the $MCMC$ samplers with multivariate proposal density look more favorable opposed to those of the $MCMC$ samplers with uniform proposal density, indicating a better convergence behavior. When inspecting the parameter vector ψ, the posterior distribution of μ fails to converge in the RMB-RWM settings. The posterior distribution of σ converges hardly in the fixed block settings and even worse in the RMB-RWM schemes. These results point to the fact that the likelihood is not very sensitive to small variations in σ making it more difficult to estimate. However, the posterior distribution of the inflation parameter ρ_Z converges in all schemes.

Table 3.9: Potential scale reduction factors ($PSRF$) and multivariate $PSRF$ ($MPSRF$) at 95% confidence level for selected RWM settings.

Par	$PSRF$ $\mathcal{A}_1$	$PSRF$ $\mathcal{A}_2$	$PSRF$ $\mathcal{A}_3$	$PSRF$ $\mathcal{A}_4$	$PSRF$ $\mathcal{A}_5$	$PSRF$ $\mathcal{A}_6$	$PSRF$ $\mathcal{A}_7$	$PSRF$ $\mathcal{A}_8$
ρ_1	1.26	1.08	1.10	1.12	1.08	1.05	1.07	1.09
ρ_2	1.11	1.09	1.11	1.17	1.06	1.06	1.07	1.11
ρ_3	1.12	1.09	1.09	1.16	1.03	1.05	1.06	1.10
ρ_4	1.11	1.08	1.07	1.14	1.06	1.04	1.05	1.10
ρ_5	1.08	1.08	1.08	1.16	1.04	1.06	1.06	1.10
ρ_6	1.12	1.09	1.08	1.16	1.05	1.06	1.06	1.12
ρ_7	1.14	1.08	1.09	1.17	1.06	1.06	1.05	1.12
ρ_8	1.10	1.07	1.13	1.14	1.04	1.05	1.07	1.09
ρ_9	1.08	1.07	1.08	1.14	1.05	1.04	1.05	1.09
ρ_{10}	1.11	1.09	1.08	1.17	1.05	1.05	1.05	1.10
ρ_{11}	1.13	1.07	1.06	1.12	1.05	1.05	1.04	1.09
ρ_{12}	1.21	1.09	1.08	1.16	1.10	1.06	1.05	1.09
ρ_{13}	1.13	1.07	1.06	1.15	1.04	1.06	1.05	1.11
ρ_{14}	1.04	1.05	1.07	1.15	1.02	1.05	1.05	1.11
ρ_{15}	1.05	1.06	1.08	1.16	1.04	1.04	1.05	1.10
ρ_{16}	1.09	1.06	1.07	1.15	1.04	1.05	1.05	1.10
ρ_{17}	1.15	1.07	1.07	1.16	1.05	1.05	1.05	1.10
ρ_{18}	1.14	1.09	1.10	1.16	1.06	1.05	1.06	1.10
ρ_{19}	1.11	1.06	1.06	1.15	1.04	1.06	1.05	1.11
μ	1.08	1.10	1.10	1.35	1.10	1.20	1.20	1.49
σ	1.20	1.27	1.27	2.15	1.19	1.36	1.38	1.85
ρ_Z	1.13	1.10	1.10	1.18	1.02	1.04	1.04	1.10
$MPSRF$	1.29	1.35	1.35	2.33	1.27	1.51	1.53	2.13

Notes: Brooks and Gelman's (1998) convergence criterion based on 100 multiple independent runs, $n_0 = 1000$.
$\mathcal{A}_1$: S-RWM, $\mathcal{A}_2$: M-RWM, $\mathcal{A}_3$: I-RWM, $\mathcal{A}_4$: RMB-RWM, all with uniform proposal density.
$\mathcal{A}_5$: S-RWM, $\mathcal{A}_6$: M-RWM, $\mathcal{A}_7$: I-RWM, $\mathcal{A}_8$: RMB-RWM, all with multivariate normal proposal density.

3.3.2 Performance assessment

To assess the statistical accuracy of ML estimation vs. selected RWM algorithms, the estimated parameter values are compared to the true parameter values over all 100 runs for all nine estimation methods. The following criteria are used for performance assessment, cp. Haario et al. (1999, p. 11) and Hoff (2009, p. 81, 103):

- empirical $Bias[\widehat{\theta}]$ = mean distance of the parameter estimates $\widehat{\theta}$ from the true parameter values θ calculated over all 100 repetitions.

  $\overline{Bias}[\widehat{\theta}] = \frac{1}{100}\sum_{r=1}^{100}\left(\theta - \widehat{\theta}^{(r)}\right)$

- empirical $MSE[\widehat{\theta}]$ = mean squared error of an estimator, i.e. the variance of the parameter estimates $\widehat{\theta}$. MSE for ML estimates bases on the robust standard errors (SE). For $MCMC$ samples, the variances of the estimates are divided by their effective sample size, plus the squared bias. The MSE is also calculated over all 100 repetitions.

  $\overline{MSE}[\widehat{\theta}] = \frac{1}{100}\sum_{r=1}^{100}\left(Var[\widehat{\theta}^{(r)}]/T^* + Bias[\widehat{\theta}]^2\right)$

- $COV = \mathcal{I}(\theta \in CI)$ = coverage, i.e. proportion of the sampled values hitting the 95% confidence interval or highest posterior density region, respectively.

  $COV_{ML} = \frac{1}{100}\sum_{r=1}^{100}\theta \in \left(\widehat{\theta}^{(r)} \pm 1.96\sqrt{MSE}\right)$

  $COV_{\mathcal{A}_s} = \frac{1}{100}\sum_{r=1}^{100}\theta \in HDR(\widehat{\theta}^{(r)})$

The results produced by the 100 repetitions are averaged and given separately for all model parameters, see Tables 3.10 to 3.12. The results show small biases and very small MSE for all parameters and all estimation procedures. The absolute averaged biases are all smaller than 0.03 in all estimation settings with except of those for ρ_3 and ρ_4 in the RWM schemes. In direct comparison, the RWM schemes only surpass the ML estimation procedure with smaller biases for five parameters: ρ_1 has a smaller averaged bias only in the S-RWM schemes, whereas ρ_{13} and ρ_{14} have smaller averaged biases only in the MB-RWM schemes. Only the biases for ρ_7 and ρ_8 are smaller in all RWM schemes, as opposed to the ML biases. The averaged MSE from ML estimation closely resemble those from RWM estimation.

The coverage rates for RWM estimation indicate how well the chains actually cover the range of the target distribution. In Table 3.13, the coverage rates for all estimation procedures considered are given. The coverage rates are higher for RWM estimation which is mostly due to the fact that the 95% HDR cover more parameter space, and the target distribution in particular, as compared to the 95% CI. In summary, all results indicate good statistical precision.

Table 3.10: Averaged parameter estimates ($\overline{\widehat{\theta}}$), averaged biases ($\overline{Bias}[\widehat{\theta}]$) and averaged mean squared errors ($\overline{MSE}[\widehat{\theta}]$) for ML estimation.

Par	θ	$\overline{\widehat{\theta}}_{ML}$	$\overline{Bias}[\widehat{\theta}_{ML}]$	$\overline{MSE}[\widehat{\theta}_{ML}]$
ρ_1	0.2500	0.2666	0.0166	0.0022
ρ_2	0.2600	0.2538	-0.0062	0.0004
ρ_3	0.3900	0.3601	-0.0299	0.0009
ρ_4	0.4600	0.4385	-0.0215	0.0008
ρ_5	0.3700	0.3542	-0.0158	0.0004
ρ_6	0.2800	0.2772	-0.0028	0.0003
ρ_7	0.1900	0.2011	0.0111	0.0009
ρ_8	0.1700	0.1814	0.0114	0.0006
ρ_9	0.2400	0.2407	0.0007	0.0001
ρ_{10}	0.2700	0.2694	-0.0006	0.0003
ρ_{11}	0.3200	0.3084	-0.0116	0.0005
ρ_{12}	0.3000	0.2906	-0.0094	0.0009
ρ_{13}	0.1700	0.1849	0.0149	0.0009
ρ_{14}	0.0800	0.0889	0.0089	0.0003
ρ_{15}	0.1800	0.1824	0.0024	0.0001
ρ_{16}	0.2200	0.2203	0.0003	0.0002
ρ_{17}	0.2400	0.2409	0.0009	0.0002
ρ_{18}	0.2500	0.2543	0.0043	0.0004
ρ_{19}	0.2600	0.2612	0.0012	0.0002
μ	7.7200	7.7139	-0.0061	0.0001
σ	0.8500	0.8494	-0.0006	0.0001
ρ_Z	0.9870	0.9864	-0.0006	<0.0001

Note: Estimates are based on 100 repetitions, $n_0 = 1000$.

Table 3.11: Averaged parameter estimates ($\overline{\widehat{\theta}}$), averaged biases ($\overline{Bias}[\widehat{\theta}]$) and averaged mean squared errors ($\overline{MSE}[\widehat{\theta}]$) for RWM estimation with uniform proposal density and different blocking strategies.

Par	θ	$\overline{\widehat{\theta}}_{\mathcal{A}_1}$	$\overline{Bias}[\widehat{\theta}_{\mathcal{A}_1}]$	$\overline{MSE}[\widehat{\theta}_{\mathcal{A}_1}]$	$\overline{\widehat{\theta}}_{\mathcal{A}_2}$	$\overline{Bias}[\widehat{\theta}_{\mathcal{A}_2}]$	$\overline{MSE}[\widehat{\theta}_{\mathcal{A}_2}]$	$\overline{\widehat{\theta}}_{\mathcal{A}_3}$	$\overline{Bias}[\widehat{\theta}_{\mathcal{A}_3}]$	$\overline{MSE}[\widehat{\theta}_{\mathcal{A}_3}]$	$\overline{\widehat{\theta}}_{\mathcal{A}_4}$	$\overline{Bias}[\widehat{\theta}_{\mathcal{A}_4}]$	$\overline{MSE}[\widehat{\theta}_{\mathcal{A}_4}]$
ρ_1	0.2500	0.2421	-0.0079	0.0018	0.2287	-0.0213	0.0007	0.2279	-0.0221	0.0007	0.2278	-0.0222	0.0007
ρ_2	0.2600	0.2492	-0.0108	0.0005	0.2453	-0.0147	0.0004	0.2459	-0.0141	0.0004	0.2453	-0.0147	0.0004
ρ_3	0.3900	0.3525	-0.0375	0.0017	0.3532	-0.0368	0.0015	0.3537	-0.0363	0.0015	0.3537	-0.0363	0.0015
ρ_4	0.4600	0.4049	-0.0551	0.0034	0.4133	-0.0467	0.0023	0.4148	-0.0452	0.0022	0.4151	-0.0449	0.0021
ρ_5	0.3700	0.3423	-0.0277	0.0010	0.3461	-0.0239	0.0007	0.3473	-0.0227	0.0006	0.3475	-0.0225	0.0006
ρ_6	0.2800	0.2672	-0.0128	0.0007	0.2701	-0.0099	0.0003	0.2700	-0.0100	0.0003	0.2713	-0.0087	0.0002
ρ_7	0.1900	0.1983	0.0083	0.0007	0.1929	0.0029	0.0002	0.1950	0.0050	0.0002	0.1933	0.0033	0.0002
ρ_8	0.1700	0.1786	0.0086	0.0004	0.1747	0.0047	0.0001	0.1734	0.0034	0.0001	0.1734	0.0034	0.0001
ρ_9	0.2400	0.2457	0.0057	0.0002	0.2478	0.0078	0.0002	0.2481	0.0081	0.0002	0.2486	0.0086	0.0002
ρ_{10}	0.2700	0.2618	-0.0082	0.0005	0.2635	-0.0065	0.0002	0.2638	-0.0062	0.0002	0.2649	-0.0051	0.0001
ρ_{11}	0.3200	0.2924	-0.0276	0.0013	0.2975	-0.0225	0.0007	0.2980	-0.0220	0.0007	0.2986	-0.0214	0.0006
ρ_{12}	0.3000	0.2685	-0.0315	0.0021	0.2721	-0.0279	0.0011	0.2710	-0.0290	0.0012	0.2725	-0.0275	0.0010
ρ_{13}	0.1700	0.1881	0.0181	0.0008	0.1802	0.0102	0.0002	0.1813	0.0113	0.0003	0.1785	0.0085	0.0002
ρ_{14}	0.0800	0.0894	0.0094	0.0001	0.0863	0.0063	0.0001	0.0852	0.0052	0.0001	0.0851	0.0051	0.0001
ρ_{15}	0.1800	0.1919	0.0119	0.0002	0.1910	0.0110	0.0002	0.1903	0.0103	0.0002	0.1903	0.0103	0.0002
ρ_{16}	0.2200	0.2236	0.0036	0.0002	0.2252	0.0052	0.0001	0.2245	0.0045	0.0001	0.2246	0.0046	0.0001
ρ_{17}	0.2400	0.2400	>-0.0001	0.0005	0.2419	0.0019	0.0002	0.2414	0.0014	0.0001	0.2412	0.0012	0.0001
ρ_{18}	0.2500	0.2435	-0.0065	0.0007	0.2450	-0.0050	0.0003	0.2440	-0.0060	0.0003	0.2422	-0.0078	0.0002
ρ_{19}	0.2600	0.2542	-0.0058	0.0004	0.2521	-0.0079	0.0002	0.2522	-0.0078	0.0002	0.2507	-0.0093	0.0002
μ	7.7200	7.7134	-0.0066	0.0005	7.7135	-0.0065	0.0005	7.7134	-0.0066	0.0005	7.7135	-0.0065	0.0005
σ	0.8500	0.8472	-0.0028	0.0004	0.8454	-0.0046	0.0004	0.8458	-0.0042	0.0004	0.8445	-0.0055	0.0004
ρ_Z	0.9870	0.9839	-0.0031	<0.0001	0.9850	-0.0020	<0.0001	0.9852	-0.0018	<0.0001	0.9867	-0.0003	<0.0001

Notes: Estimates are based on 100 repetitions. $MCMC$ samples are of length $T = 10{,}000$ following a burn-in period of $n_0 = 1000$ iterations.
$\mathcal{A}_1$: S-RWM, $\mathcal{A}_2$: M-RWM, $\mathcal{A}_3$: I-RWM, $\mathcal{A}_4$: RMB-RWM, all with uniform proposal density.

Table 3.12: Averaged parameter estimates ($\overline{\widehat{\theta}}$), averaged biases ($\overline{Bias}[\widehat{\theta}]$) and averaged mean squared errors ($\overline{MSE}[\widehat{\theta}]$) for RWM estimation with multivariate normal proposal density and different blocking strategies.

Par	θ	$\overline{\widehat{\theta}}_{\mathcal{A}_5}$	$\overline{Bias}[\widehat{\theta}_{\mathcal{A}_5}]$	$\overline{MSE}[\widehat{\theta}_{\mathcal{A}_5}]$	$\overline{\widehat{\theta}}_{\mathcal{A}_6}$	$\overline{Bias}[\widehat{\theta}_{\mathcal{A}_6}]$	$\overline{MSE}[\widehat{\theta}_{\mathcal{A}_6}]$	$\overline{\widehat{\theta}}_{\mathcal{A}_7}$	$\overline{Bias}[\widehat{\theta}_{\mathcal{A}_7}]$	$\overline{MSE}[\widehat{\theta}_{\mathcal{A}_7}]$	$\overline{\widehat{\theta}}_{\mathcal{A}_8}$	$\overline{Bias}[\widehat{\theta}_{\mathcal{A}_8}]$	$\overline{MSE}[\widehat{\theta}_{\mathcal{A}_8}]$
ρ_1	0.2500	0.2396	-0.0104	0.0008	0.2306	-0.0194	0.0006	0.2271	-0.0229	0.0007	0.2271	-0.0229	0.0007
ρ_2	0.2600	0.2488	-0.0112	0.0003	0.2464	-0.0136	0.0003	0.2462	-0.0138	0.0004	0.2449	-0.0151	0.0004
ρ_3	0.3900	0.3530	-0.0370	0.0016	0.3527	-0.0373	0.0015	0.3534	-0.0366	0.0015	0.3536	-0.0364	0.0015
ρ_4	0.4600	0.4062	-0.0538	0.0031	0.4136	-0.0464	0.0023	0.4136	-0.0464	0.0023	0.4139	-0.0461	0.0023
ρ_5	0.3700	0.3416	-0.0284	0.0010	0.3462	-0.0238	0.0007	0.3467	-0.0233	0.0007	0.3480	-0.0220	0.0006
ρ_6	0.2800	0.2679	-0.0121	0.0004	0.2696	-0.0104	0.0003	0.2711	-0.0089	0.0002	0.2715	-0.0085	0.0002
ρ_7	0.1900	0.1964	0.0064	0.0004	0.1944	0.0044	0.0002	0.1939	0.0039	0.0002	0.1949	0.0049	0.0002
ρ_8	0.1700	0.1792	0.0092	0.0003	0.1748	0.0048	0.0001	0.1736	0.0036	0.0001	0.1738	0.0038	0.0001
ρ_9	0.2400	0.2452	0.0052	0.0002	0.2474	0.0074	0.0001	0.2476	0.0076	0.0001	0.2481	0.0081	0.0001
ρ_{10}	0.2700	0.2612	-0.0088	0.0003	0.2635	-0.0065	0.0002	0.2635	-0.0065	0.0002	0.2635	-0.0065	0.0002
ρ_{11}	0.3200	0.2900	-0.0300	0.0012	0.2965	-0.0235	0.0007	0.2958	-0.0242	0.0008	0.2986	-0.0214	0.0006
ρ_{12}	0.3000	0.2715	-0.0285	0.0015	0.2715	-0.0285	0.0011	0.2726	-0.0274	0.0010	0.2726	-0.0274	0.0010
ρ_{13}	0.1700	0.1909	0.0209	0.0007	0.1813	0.0113	0.0003	0.1819	0.0119	0.0003	0.1787	0.0087	0.0002
ρ_{14}	0.0800	0.0894	0.0094	0.0001	0.0868	0.0068	0.0001	0.0862	0.0062	0.0001	0.0854	0.0054	0.0001
ρ_{15}	0.1800	0.1911	0.0111	0.0002	0.1901	0.0101	0.0002	0.1900	0.0100	0.0002	0.1902	0.0102	0.0002
ρ_{16}	0.2200	0.2244	0.0044	0.0002	0.2249	0.0049	0.0001	0.2247	0.0047	0.0001	0.2252	0.0052	0.0001
ρ_{17}	0.2400	0.2397	-0.0003	0.0002	0.2417	0.0017	0.0001	0.2406	0.0006	0.0001	0.2406	0.0006	0.0001
ρ_{18}	0.2500	0.2448	-0.0052	0.0004	0.2434	-0.0066	0.0002	0.2431	-0.0069	0.0002	0.2410	-0.0090	0.0002
ρ_{19}	0.2600	0.2545	-0.0055	0.0002	0.2528	-0.0072	0.0002	0.2525	-0.0075	0.0002	0.2507	-0.0093	0.0002
μ	7.7200	7.7135	-0.0065	0.0005	7.7133	-0.0067	0.0005	7.7133	-0.0067	0.0005	7.7137	-0.0063	0.0005
σ	0.8500	0.8474	-0.0026	0.0004	0.8452	-0.0048	0.0004	0.8454	-0.0046	0.0004	0.8450	-0.0050	0.0004
ρ_Z	0.9870	0.9845	-0.0025	<0.0001	0.9859	-0.0011	<0.0001	0.9859	-0.0011	<0.0001	0.9866	-0.0004	<0.0001

Notes: Estimates are based on 100 repetitions. $MCMC$ samples are of length $T = 10{,}000$ following a burn-in period of $n_0 = 1000$ iterations.
$\mathcal{A}_5$: S-RWM, $\mathcal{A}_6$: M-RWM, $\mathcal{A}_7$: I-RWM, $\mathcal{A}_8$: RMB-RWM, all with multivariate normal proposal density.

3.3.3 Model selection by marginal likelihood

So far, no clear preference for one particular estimation procedure could be found. The log-marginal likelihoods are estimated for further model selection among the RWM schemes (cp. Section 3.2.6).

Table 3.14: Averaged log-posterior densities and averaged log-marginal likelihoods.

Quantity	$\mathcal{A}_1$	$\mathcal{A}_2$	$\mathcal{A}_3$	$\mathcal{A}_4$	$\mathcal{A}_5$	$\mathcal{A}_6$	$\mathcal{A}_7$	$\mathcal{A}_8$
$\log p(\mathbf{z}\|\theta')$	-65,582.1	-65,557.6	-65,556.7	-65,552.8	-65,571.5	-65,600.5	-65,555.7	-65,557.2
$\log p(\theta')$	-18.952	-19.291	-19.964	-20.074	-16.884	-18.387	-19.155	-19.615
$\log p(\theta'\|\mathbf{z})$	-65,601.1	-65,576.9	-65,576.7	-65,572.8	-65,588.4	-65,618.9	-65,574.9	-65,576.8
$\log \widehat{p}(\theta'\|\mathbf{z})$	–	–	–	–	-9.692	47.869	51.622	63.627
$\log m(\mathbf{z})$	–	–	–	–	-65,573.2	-65,632.2	-65,615.6	-65,624.7
(NSE)	–	–	–	–	(28.95)	(50.14)	(30.15)	(31.70)

$\mathcal{A}_1$ S-RWM and uniform proposal density, $\mathcal{A}_2$ M-RWM and uniform proposal density,
$\mathcal{A}_3$ I-RWM and uniform proposal density, $\mathcal{A}_4$ RMB-RWM and uniform proposal density,
$\mathcal{A}_5$ S-RWM and multivariate normal proposal density, $\mathcal{A}_6$ M-RWM and multivariate normal proposal density,
$\mathcal{A}_7$ I-RWM and multivariate normal proposal density, $\mathcal{A}_8$ RMB-RWM and multivariate normal proposal density.

Results of marginal likelihood estimation are given in Table 3.14. Additionally, the numerical standard error (NSE) of the log-marginal likelihood estimate is approx. by the SD of 100 repetitions. The highest log-marginal likelihood estimate can be found for the S-RWM with multivariate normal proposal density. This scheme has also the smallest NSE. To some extent, this finding contradicts the results so far which strongly supported the blocking schemes. As can be seen, the log-posterior ordinate is increased for the multiple-block settings which might be interpreted as a penalty term in this respect. Using *Schwarz's Criterion* ($SC = -\frac{1}{2}BIC$, see Frühwirth-Schnatter, 2006, p. 116) additionally allows to compare the Bayesian models with ML estimation because of their clear correspondence (cp. Gelman, Meng, & Stern, 1996, p. 786). The SC can be regarded as an asymptotic approximation to the log-marginal likelihood, see Frühwirth-Schnatter (2006, p. 120) and Gelfand and Dey (1994, p. 508). The log-marginal likelihoods are lower than the SC supporting the RWM schemes against the ML approach.

Comparable results of Würbach (2015), for the heaping model with underlying Dagum distribution and the excess of zeros modeled within the heaping mechanism as separate modulo, show the log-marginal likelihood of the I-RWM algorithm to be the greatest (-66,237.1), as compared to the log-marginal likelihood of the S-RWM algorithm (-66,363.1), the M-RWM algorithm (-66,285.7), and the RMB-RWM algorithm (-66,283.6). The increase in the log-marginal likelihood in the current setting averages 680. This increase indicates an improvement of the model, when including alternatively an inflation parameter for the excess of zeros instead of modeling the proportion of zero values as part of the heaping mechanism.

Table 3.13: Coverage rates for ML and RWM estimation.

Par	COV ML	COV $\mathcal{A}_1$	COV $\mathcal{A}_2$	COV $\mathcal{A}_3$	COV $\mathcal{A}_4$	COV $\mathcal{A}_5$	COV $\mathcal{A}_6$	COV $\mathcal{A}_7$	COV $\mathcal{A}_8$
ρ_1	71	99	100	100	100	100	100	100	100
ρ_2	66	100	99	100	96	100	100	100	100
ρ_3	67	100	96	95	56	100	100	100	83
ρ_4	81	97	85	88	36	100	100	99	59
ρ_5	65	100	100	100	90	100	100	100	100
ρ_6	58	100	100	100	100	100	100	100	100
ρ_7	74	100	100	100	100	100	100	100	100
ρ_8	75	100	100	100	100	100	100	100	100
ρ_9	60	100	100	100	100	100	100	100	100
ρ_{10}	64	100	100	100	100	100	100	100	100
ρ_{11}	69	100	100	100	98	100	100	100	99
ρ_{12}	76	100	100	100	96	100	100	100	100
ρ_{13}	72	100	100	100	100	100	100	100	100
ρ_{14}	76	100	100	100	100	100	100	100	100
ρ_{15}	72	100	100	100	99	100	100	100	100
ρ_{16}	64	100	100	100	100	100	100	100	100
ρ_{17}	55	100	100	100	100	100	100	100	100
ρ_{18}	63	100	100	100	99	100	100	100	100
ρ_{19}	65	100	100	100	100	100	100	100	100
μ	61	100	100	100	92	100	93	93	92
σ	62	93	93	93	92	94	92	92	92
ρ_Z	72	100	100	100	100	100	100	100	100

Notes: Coverage rates are based on 100 repetitions.
$\mathcal{A}_1$ S-RWM and uniform proposal density,
$\mathcal{A}_2$ M-RWM and uniform proposal density,
$\mathcal{A}_3$ I-RWM and uniform proposal density,
$\mathcal{A}_4$ RMB-RWM and uniform proposal density,
$\mathcal{A}_5$ S-RWM and multivariate normal proposal density,
$\mathcal{A}_6$ M-RWM and multivariate normal proposal density,
$\mathcal{A}_7$ I-RWM and multivariate normal proposal density,
$\mathcal{A}_8$ RMB-RWM and multivariate normal proposal density.

Summary

In general, the RWM schemes need considerably more runtime. The advantage is, however, that the RWM schemes immediately return the error of the estimates (SD and quantiles) due to the availability of the complete posterior distribution. The ML approach needs either a bootstrap, which is time consuming, or the derivation of the hessian, which is computational laboriously. No clear statement can be given relating to differences in the performance of ML and RWM estimation, pointing to no superiority of one the approaches considered in the simulation.

With regard to the alternative RWM schemes worked through in this thesis, several remarks can be given. First, no significant differences exist between the schemes according to $PSRF$, averaged bias, averaged mean squared error, and averaged numerical standard error. Second, in the considered case, the multiple-block schemes seem to be more efficient as indicated by the lower *Ineff*, which is confirmed by the lower SD of the log-marginal likelihood. Third, this efficiency gain comes at the price of an increased runtime (k-fold, due to the number of blocks considered). This cost is expressed by the penalty term for multiple blocks in the log-marginal likelihood, lowering the favor for multiple-block schemes to some extent.

The conclusions drawn here only hold for the heaping model described so far. In the following chapter several modifications and also extensions of the heaping model are explored. These modifications of the heaping model increase model complexity. It is advisable to employ the multiple-block schemes for RWM estimation to reach sufficient efficiency, but the increase in model complexity complicates the decision about the appropriate number and composition of the blocks. According to Chib and Ramamurthy (2010), this is a reliable reason to continue to use the randomized multiple-block schemes for RWM estimation.

Chapter 4

Modifications and extensions of the heaping model

In the previous chapter, different estimation procedures have been compared with regard to a particular specification of the heaping model, referred to as Model I. For this, data were generated that accommodate the specific modeling strategy as described in Chapter 2. In its general form, the heaping model is far more flexible. Different specifications concerning the latent distribution, the heaping pattern or the heaping mechanism can be considered and tested against each other. In this chapter, five alternative modeling approaches are unfolded and estimated on the basis of simulated data. The chapter concludes with an extension of the heaping model to multivariate contexts considering additionally information from the respondent. In contrast to the univariate modeling, the extended models are assumed to give a better reflection of reality thus enhancing estimation accuracy.

4.1 Modifications of the heaping model

In a first place, Model I is modified by assuming an alternative latent true distribution to model income (Model II). The more complex 3-parametric Dagum distribution is used instead of the 2-parametric log-normal distribution. The second modification assumes wider intervals for the heaping points to enable a higher proportion of heaped values in total (Model III). The rationale behind this is the supposition of higher uncertainty or reluctance on the part of the respondent, both causing more coarseness. Then, asymmetric intervals are considered to account for possible downsizing (Model IV). A phenomenon typically found for self-reported income in which values are underreported (cp. Section 1.1.2). The next two modifications are with respect to the heaping mechanism. An alternative function is employed to describe the heaping probabilities. In Model V, the heaping probabilities are considered to increase steadily with proximity to a heaping point. And finally, a reduced set of heaping probabilities is explored (Model VI) by enlarging the equality constraints. Concretely, the ranges of the sets are spread for each

modulo in which the heaping probabilities are assumed to be equal. The heaping mechanism is then described by less parameters in total, which is expected to ease estimation. An accordant data example is simulated and used for estimation of each of the competing models.

An overview with respect to the modifications is given in Table 4.1. Furthermore, descriptive statistics from the simulated data of each model are presented. Mean, median and SD for Model II are lower compared to the other models. This fact is attributable to the shape of the Dagum distribution which has much more mass in the peak than in the tails, as opposed to the log-normal distribution. All data examples closely resemble each other when inspecting the corresponding histograms. The five modifications of the univariate heaping model are estimated by ML and the RMB-RWM algorithm, respectively. Runtimes for estimation differ remarkably for Model II and even more for Model VI, compared to the others. Using the Dagum distribution is computational more expensive, while reducing the number of model parameters decreases computational effort noticably.

Table 4.1: Descriptive statistics of the data examples for all modeling strategies.

Descriptive	Model I	Model II	Model III	Model IV	Model V	Model VI
Latent distr.	log-norm	Dagum	log-norm	log-norm	log-norm	log-norm
Perc of heaping	∼50	∼50	∼70	∼50	∼50	∼50
Heaping intervals	sym.	sym.	sym.	asym.	sym.	sym.
Heaping function	*pcm*	*pcm*	*pcm*	*pcm*	*pbsm*	*pcm*
Components of ϕ	19	19	19	19	19	9
Perc of zeros	1.33	1.16	1.26	1.36	1.20	1.24
Mean	2713.16	1866.54	2719.11	2689.99	2733.03	2708.40
Median	2072.68	1600.00	2064.44	2036.61	2020.29	2042.37
SD	2045.97	1276.85	2050.10	2052.31	2077.65	2038.58
Run.* ML	0.39	0.91	0.43	0.48	0.50	0.09
Run.* RMB-RWM	13.26	33.19	13.89	13.29	34.34	8.48

*Runtime is given in hours.

4.1.1 Modification with respect to the latent distribution

The first modification concerns an alternative modeling of the latent true distribution (Model II). Whereas 2-parameter models do not allow for intersecting Lorenz curves, 3-parameter models do due to the increased flexibility, see McDonald et al. (2013, p. 361). This is an important factor when dealing with income distributions. Among the 3-parameter models, the Dagum is shown to perform best (id., p. 373). It is also proved to be suitable for temporal and spatial comparisons, see Bandourian, McDonald, and Turley (2002). Thus, the Dagum distribution (Dagum, 1977) might help to model income data more appropriately. The Dagum distribution is a right-skewed distribution, in the statistical literature known as Burr

Type 3 distribution (Burr, 1942) and belongs to the family of Generalized Beta distributions, see Kleiber and Kotz (2003c, Chapter 6.3). Hence, its *pdf* can be expressed by the *pdf* of the 4-parametric Generalized Beta II distribution (GB2) in which one of the scale parameters is set to one, i.e. $q = 1$:

$$f(y) = apy^{ap-1}b^{-ap}\left[1 + \left(\frac{y}{b}\right)^{a}\right]^{-p-1} \mathbb{I}(y > 0),$$

where $a, b, p \in \mathbb{R}^+$. Parameters a and p are for shape and b is a scale parameter (id., p. 184). Corresponding moments and other properties are given in the Appendix A.2.2. Both, the Dagum and the log-normal distribution are fitted to the non-zero values from the NEPS data.[1] Accordant data are generated ($N = 100{,}000$) based on these estimates. Figure 4.1 shows the fitted densities in direct comparison. As can be seen, the Dagum distribution has a higher peak and a flatter tail, as opposed to the log-normal distribution.

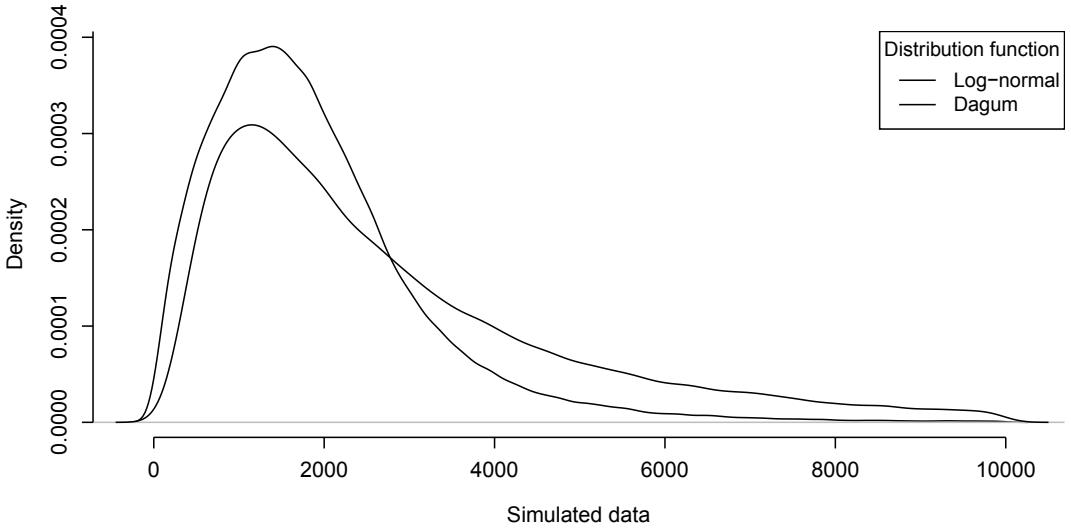


Figure 4.1: Densities of the log-normal and Dagum distribution. Parameters of the log-normal distribution are $\mu = 7.72$, $\sigma^2 = 0.85^2$, and parameters of the Dagum distribution are $a = 3.6$, $b = 2416$, and $p = 0.43$.

Alike the log-normal, the Dagum distribution is not defined at $y = 0$. To overcome this problem, an additional parameter for inflation is introduced (cp. Section 2.1) yielding the following density:

$$f(y|\psi) = (1 - \rho_Z)\mathbb{I}(y = 0) + \rho_Z apy^{ap-1}b^{-ap}\left[1 + \left(\frac{y}{b}\right)^{a}\right]^{-p-1} \mathbb{I}(y > 0),$$

where ψ now comprises a, b, p, ρ_Z, and ρ_Z ranges between 0 and 1 again. The specification of the heaping behavior is as in Model I. Based on this Model II, an accordant data example is simulated (cp. Figure 4.2).

[1] Accordant functionality for fitting the Dagum distribution is given by the `R` package `VGAM`.

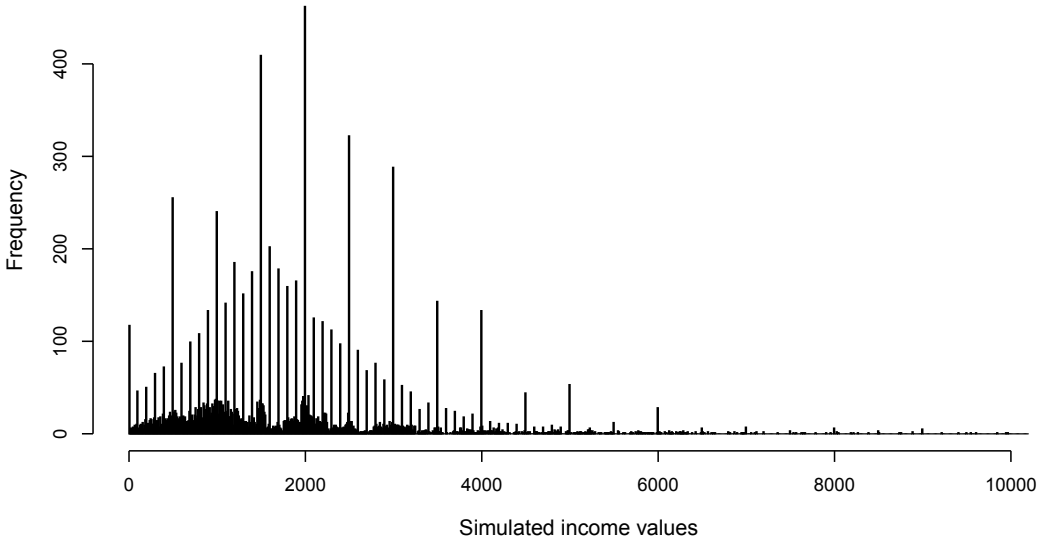


Figure 4.2: Data example of Model II with Dagum distribution.

The probabilities of true values to fall on a modulo in a given set are shown in Table 4.2, and the corresponding percentages of heaped values in the data example are given in Table 4.3. In sum, 50.29% of the simulated values are heaped.

Table 4.2: Heaping probabilities in Model II.

Interval	mod(100)	mod(500)	mod(1000)
$[0, 500]$	0.25	0.17	–
$(500, 1000]$	0.26	–	0.06
$(1000, 1500]$	0.39	0.22	–
$(1500, 2000]$	0.46	–	0.14
$(2000, 3000]$	0.39	0.27	0.20
$(3000, 4000]$	0.31	0.32	0.21
$(4000, 5000]$	0.24	0.30	0.25
$(5000, 10{,}000]$	–	0.17	0.26

Table 4.3: Percentages of heaped values in the data example of Model II.

Interval	mod(100)	mod(500)	mod(1000)	Total
$[0, 500]$	1.95	2.33	–	4.28
$(500, 1000]$	3.32	–	1.96	5.28
$(1000, 1500]$	5.83	3.92	–	9.75
$(1500, 2000]$	6.68	–	4.25	10.93
$(2000, 3000]$	6.97	3.03	2.81	12.81
$(3000, 4000]$	2.26	1.41	1.25	4.92
$(4000, 5000]$	0.69	0.43	0.49	1.61
$(5000, 10{,}000]$	–	0.25	0.46	0.71
Total	27.70	11.37	11.22	50.29

The initial values for estimation of the ρ_b's are set to $\varphi = [0.2]_{b=1}^{19}$. The initial values for estimation of ψ are set to $\upsilon = (3.60, 2000.00, 0.43, 0.99)'$ which result from fitting the non-zero values of the simulated data to the Dagum distribution. These values are considered also as prior means for RWM estimation with covariance matrix $\Sigma = 0.001\mathcal{I}_{19}$ for the parameter vector ϕ and covariance matrix $\Upsilon = \text{diag}(0.1, 100, 0.01, 0.001)\mathcal{I}_4$ for the parameter vector ψ. The covariance matrix of the proposal density is assumed to equal $\text{diag}(0.0001, \ldots, 0.0001, 0.1, 10, 0.0001, 0.00001)\mathcal{I}_{23}$.

Estimation of Model II lasts less than one hour when using the ML approach. Bayesian estimation via RMB-RWM algorithm takes roughly 33 hours (with 11,000 $MCMC$ draws sampled). This relatively long runtime is caused by the required evaluation of the integrals of the log-likelihood at each sampled draw. In Model II, each evaluation takes approx. 1.6 seconds as compared to approx. 0.4 seconds when considering the standard log-normal distribution. Multiple evaluations of the log-likelihood – as in case of the randomized multiple-blocking strategy – are therefore computational expensive.

Results from estimation are presented in Table 4.4. The parameter estimates for ϕ are close to the true ones, as opposed to the parameter estimates for ψ. The parameters a and b of the Dagum distribution are both significantly underestimated in both estimation procedures, while the scale parameter p is significantly overestimated. In an overall assessment, both estimation techniques (ML and RWM) clearly fail in correct estimation of the parameters of the underlying assumed true model. Comparing the squared bias averaged over all model parameters yields 2006.61 for ML, and 6144.14 for RWM. These particularly high deviations from the true parameters are caused by the clear underestimation of the parameters a and b (3.288 and 3.060 vs. 3.600 for a, and 2201.2 and 2040.1 vs. 2416.0 for b). This might be, to a large extent, attributable to the less restrictive 3-parametric distribution. Three parameters might not discriminate sufficiently. To be concrete, different parameterizations might result in densities that closely resemble each other. Since 2-parameter models are evidently limited in the variety of shapes, multi-parameter distributions have a higher flexibility and can describe the shape of the income distribution presumably better. However, this advantage comes at the price of insensitivity to small changes in their specification. The increment of parameters in the model specification also imposes the considerably greater burden in terms of computation, cp. Cowell (2000, p. 146). When comparing only the parameter vector ϕ, the accordant deviations are 0.00023 for ML and 0.00044 for RWM. This means that the ML estimates are marginally closer to the true parameters than the $MCMC$ estimates.

4.1.2 Modifications of the heaping pattern

Assume an extreme high proportion of heaping

Second, a data set with an extreme high proportion of heaped values is simulated (Model III). To resemble the proportion of heaping found in the NEPS data more closely, about 70% of the values from the latent assumed true distribution are assigned to heaping points. One way to reach such a high proportion but taking into account the constraint system at the same time, requires the heaping intervals to be specified markedly wider, i.e. not half of the modulo on each side of the heaping point yielding a width equalling to the modulo, but the widths being doubled. To be concrete, heaping points at modulo 100 (h_b mod(100)) have now a catchment area of $h_b \pm 100$ instead of $h_b \pm 50$. In general, the heaping intervals are

Table 4.4: Parameter estimates and 95% confidence intervals (CI) or 95% highest density region (HDR) for the data example of Model II.

Par	θ	$\widehat{\theta}_{ML}[CI]$	$\widehat{\theta}_{\mathcal{A}_8}[HDR]$
ρ_1	0.250	0.260[0.248,0.272]	0.255[0.201,0.310]
ρ_2	0.260	0.251[0.250,0.252]	0.244[0.201,0.286]
ρ_3	0.390	0.383[0.382,0.384]	0.358[0.312,0.403]
ρ_4	0.460	0.472[0.465,0.480]	0.439[0.396,0.481]
ρ_5	0.390	0.397[0.381,0.413]	0.369[0.329,0.409]
ρ_6	0.310	0.304[0.297,0.312]	0.294[0.242,0.347]
ρ_7	0.240	0.249[0.247,0.252]	0.241[0.164,0.318]
ρ_8	0.170	0.179[0.170,0.188]	0.178[0.140,0.216]
ρ_9	0.220	0.218[0.212,0.223]	0.222[0.187,0.258]
ρ_{10}	0.270	0.275[0.263,0.287]	0.280[0.236,0.323]
ρ_{11}	0.320	0.334[0.326,0.343]	0.303[0.246,0.363]
ρ_{12}	0.300	0.301[0.289,0.312]	0.256[0.175,0.343]
ρ_{13}	0.170	0.195[0.184,0.207]	0.201[0.113,0.285]
ρ_{14}	0.060	0.056[0.055,0.057]	0.060[0.042,0.078]
ρ_{15}	0.140	0.131[0.130,0.131]	0.139[0.114,0.164]
ρ_{16}	0.200	0.204[0.203,0.206]	0.210[0.171,0.249]
ρ_{17}	0.210	0.246[0.242,0.250]	0.247[0.189,0.307]
ρ_{18}	0.250	0.215[0.198,0.233]	0.226[0.145,0.308]
ρ_{19}	0.260	0.249[0.239,0.258]	0.238[0.160,0.318]
a	3.600	3.288[3.277,3.300]	3.060[2.905,3.214]
b	2416.0	2201.2[2201.1,2201.2]	2040.1[2009.9,2072.3]
p	0.430	0.516[0.513,0.519]	0.594[0.558,0.630]
ρ_Z	0.987	0.989[0.988,0.989]	0.988[0.983,0.993]

now $[h_b - \text{mod}, h_b + \text{mod}]$. To give an example, the lower bound of $h_b = 1600$ is now fixed at $l_b = 1500$ and the upper bound is fixed at $u_b = 1700$. For interpretation of real heaping behavior, this specification assumes higher uncertainty with respect to the total amount of income or reluctance due to sensitivity of the topic, each expressed by increased coarseness in the responses.

Table 4.5 shows the probabilities of true values to fall on a modulo in a given set used for simulation of Model II. The corresponding percentages of heaped values in the accordant data example for each modulo in each set are given in Table 4.6. In sum, 69.91% of the simulated values are heaped in this data example.

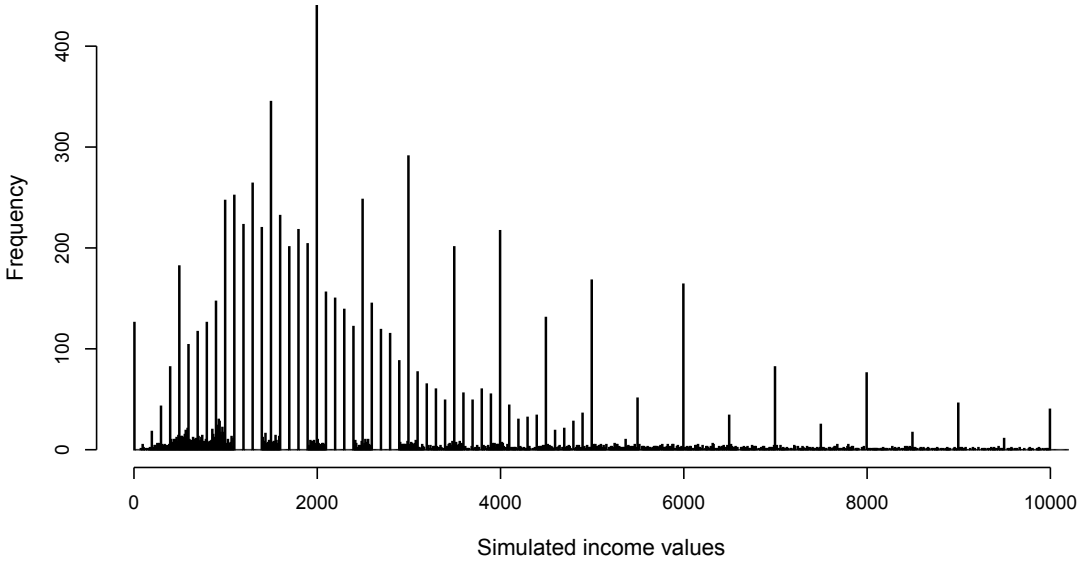


Figure 4.3: Data example of Model III with extreme heaping.

Table 4.5: Heaping probabilities in Model III.

Interval	mod(100)	mod(500)	mod(1000)
$[0, 500]$	0.20	0.11	–
$(500, 1000]$	0.23	–	0.05
$(1000, 1500]$	0.39	0.13	–
$(1500, 2000]$	0.43	–	0.13
$(2000, 3000]$	0.33	0.14	0.09
$(3000, 4000]$	0.24	0.17	0.12
$(4000, 5000]$	0.19	0.15	0.14
$(5000, 10{,}000]$	–	0.10	0.18

Table 4.6: Percentages of heaped values in the data example of Model III.

Interval	mod(100)	mod(500)	mod(1000)	Total
$[0, 500]$	1.33	1.66	–	2.99
$(500, 1000]$	4.52	–	2.25	6.77
$(1000, 1500]$	9.46	3.37	–	12.83
$(1500, 2000]$	8.46	–	4.32	12.78
$(2000, 3000]$	10.19	2.40	2.85	15.44
$(3000, 4000]$	4.46	1.96	2.13	8.55
$(4000, 5000]$	2.28	1.27	1.66	5.21
$(5000, 10{,}000]$	–	1.29	4.05	5.34
Total	40.70	11.95	17.26	69.91

The parameters of the heaping probability function are initiated with $\varphi = [0.1]_{b=1}^{19}$. The initial values for estimation of the parameters of the zero-inflated log-normal distribution result from fitting the non-zero values of the observed data to a log-normal distribution, hence equalling those values in Section 3.1.2 with $\upsilon = (7.714, 0.839, 0.990)'$.

Results from ML estimation and Bayesian estimation via RMB-RWM algorithm are presented in Table 4.7. As can be seen, the parameter estimates are less close to the true ones and CI and HDR, as measures of uncertainty, are much greater, as opposed to those in Model I (cp. Table 3.1 in Section 3.2.7). This can be explained by the wider catchment areas as well as the higher overall proportion

of heaping that has been assumed in Model II. First, the wider heaping intervals that are per se allowed to overlap enhance uncertainty in estimation. Second, the greater proportion of heaped values further extends this uncertainty. Comparing the squared bias averaged over all model parameters yields 0.00639 for ML and 0.00344 RWM. That is, the $MCMC$ estimates are closer to the true ones than the ML estimates. However, the parameter estimates ρ_3, ρ_4, and ρ_5 are significantly overestimated in both cases but even stronger when using ML. Neither CI nor HDR cover the true parameter values. These three parameters together cover the range $(1000, 3000]$ for mod(100). This is the income range with the highest probability mass and the modulo with the highest heaping probabilities overall. Unfortunately, the estimation fails by putting even more weight on the contribution of these parameters.

Table 4.7: Parameter estimates and 95% confidence intervals (CI) or 95% highest density region (HDR) for the data example of Model III.

Par	θ	$\widehat{\theta}_{ML}[CI]$	$\widehat{\theta}_{\mathcal{A}_8}[HDR]$
ρ_1	0.200	0.222[0.219,0.225]	0.200[0.139,0.260]
ρ_2	0.230	0.224[0.222,0.225]	0.215[0.177,0.256]
ρ_3	0.390	0.624[0.609,0.639]	0.554[0.505,0.603]
ρ_4	0.430	0.619[0.596,0.643]	0.553[0.500,0.609]
ρ_5	0.330	0.546[0.517,0.576]	0.506[0.458,0.556]
ρ_6	0.240	0.237[0.232,0.243]	0.229[0.191,0.267]
ρ_7	0.190	0.199[0.192,0.205]	0.185[0.140,0.229]
ρ_8	0.110	0.100[0.099,0.100]	0.102[0.064,0.141]
ρ_9	0.130	0.113[0.105,0.122]	0.141[0.107,0.177]
ρ_{10}	0.140	0.132[0.131,0.133]	0.144[0.103,0.186]
ρ_{11}	0.170	0.152[0.127,0.178]	0.162[0.115,0.211]
ρ_{12}	0.150	0.134[0.075,0.194]	0.171[0.110,0.231]
ρ_{13}	0.100	0.110[0.089,0.130]	0.106[0.063,0.152]
ρ_{14}	0.050	0.055[0.044,0.066]	0.058[0.038,0.078]
ρ_{15}	0.130	0.097[0.088,0.106]	0.109[0.084,0.134]
ρ_{16}	0.090	0.096[0.094,0.097]	0.104[0.075,0.133]
ρ_{17}	0.120	0.118[0.105,0.130]	0.114[0.076,0.156]
ρ_{18}	0.140	0.132[0.117,0.147]	0.138[0.091,0.185]
ρ_{19}	0.180	0.182[0.179,0.185]	0.179[0.140,0.218]
μ	7.720	7.703[7.698,7.709]	7.705[7.657,7.754]
σ	0.850	0.841[0.832,0.850]	0.837[0.803,0.870]
ρ_Z	0.987	0.988[0.987,0.988]	0.987[0.980,0.993]

Assume asymmetric heaping intervals

Besides increased uncertainty and coarseness, people might tend to downsize their income. *Downsizing* means the preference for more heaping downwards, as opposed to heaping upwards, i.e. intervals for heaping are shifted to the right. Such an *underreporting* of income is found in many surveys, cp. Section 1.1.2. Antoni et al. (2015) give a strong evidence for this phenomenon in the NEPS data when comparing self-reported gross income with registered data. Thus, it stands to reason that net income is also underreported but to an unknown extent. To model downsizing, the interval bounds l_b and u_b are assumed to be $h_b - \frac{1}{3}\text{mod}$ and $h_b + \frac{2}{3}\text{mod}$ for *pcm* (Model IV), see Figure 4.4. The catchment areas are assumed to be of the same widths as in Model I, e.g. heaping points at modulo 100 (h_b mod(100)) have an interval width of 100 and so on. For example, l_b of $h_b = 1600$ is now fixed at 1567 and u_b is fixed at 1667. Table 4.8 gives the probabilities of a true value to fall on a modulo and Table 4.9 shows the corresponding percentages of heaped values in the simulated data for each modulo. In sum, 50.23% of the simulated values are heaped.

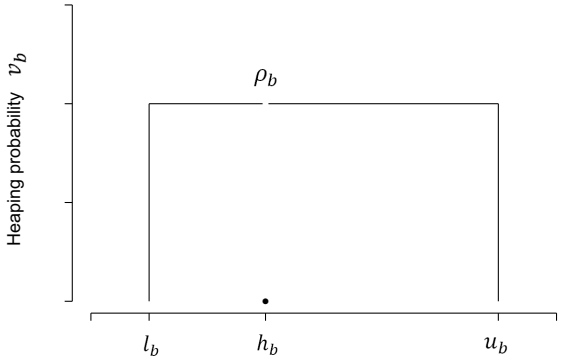


Figure 4.4: Illustration of the *pcm* with asymmetric intervals for the heaping probabilities.

Table 4.8: Heaping probabilities in Model IV.

Interval	mod(100)	mod(500)	mod(1000)
$[0, 500]$	0.26	0.16	–
$(500, 1000]$	0.27	–	0.07
$(1000, 1500]$	0.45	0.23	–
$(1500, 2000]$	0.50	–	0.19
$(2000, 3000]$	0.43	0.26	0.21
$(3000, 4000]$	0.33	0.31	0.23
$(4000, 5000]$	0.22	0.29	0.24
$(5000, 10{,}000]$	–	0.16	0.25

Table 4.9: Percentages of heaped values in the data example of Model IV.

Interval	mod(100)	mod(500)	mod(1000)	Total
$[0, 500]$	0.78	1.75	–	2.53
$(500, 1000]$	2.91	–	1.95	4.86
$(1000, 1500]$	5.30	2.88	–	8.18
$(1500, 2000]$	5.00	–	4.03	9.03
$(2000, 3000]$	6.42	2.44	2.71	11.57
$(3000, 4000]$	3.42	1.47	1.83	6.72
$(4000, 5000]$	1.42	0.98	1.39	3.79
$(5000, 10{,}000]$	–	0.99	2.56	3.55
Total	25.25	10.51	14.47	50.23

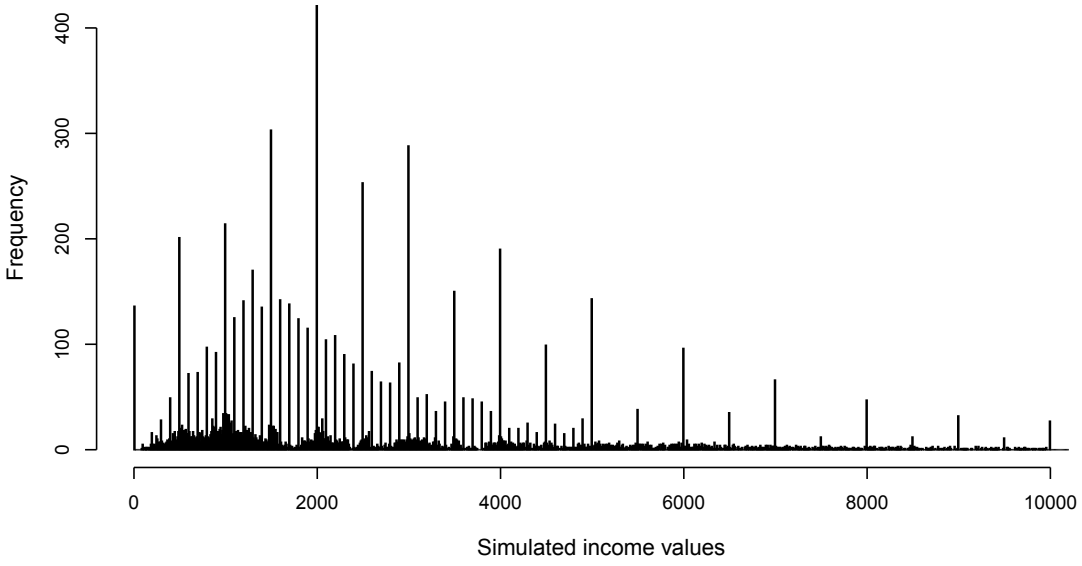


Figure 4.5: Data example of Model IV with asymmetric intervals.

Initial values for estimation are set to $\varphi = [0.2]_{b=1}^{19}$ and $\upsilon = (7.714, 0.839, 0.990)'$. Results from ML and RMB-RWM estimation are presented in Table 4.10. The parameter estimates are much closer to the true ones than in the previous models. None of the parameters is remarkably over- or underestimated. Comparing the squared bias averaged over all model parameters shows that the $MCMC$ estimates are slightly closer to the true ones (0.00053) than the ML estimates (0.00070).

4.1.3 Modifications of the heaping mechanism

Assume an alternative heaping mechanism

Zinn and Würbach (2014) employ two distinct models for the heaping mechanism. A piecewise constant model (*pcm*) which assumes uniform heaping behavior within all catchment areas I_b on the one hand (cp. Section 2.2), and on other hand heaping probabilities are considered to increase steadily with proximity to a heaping point h_b. The *pcm* assumption might be too simple to explain real heaping behavior. In case of "piecewise bell-shaped heaping probabilities" (*pbsm*), the model accounts for the fact that people's propensity to heap might by more likely to increase with proximity to a HP (Model V). The name derives from the bell-shaped curve of the normal density on which the function is based. As opposed to equiprobable distributed probabilities within the intervals, the steadily increasing/decreasing heaping probabilities indicate – because of the curvature within each interval – that falling on a HP is not equally likely to occur for each value in the respective

Table 4.10: Parameter estimates and 95% confidence intervals (CI) or 95% highest density region (HDR) for the data example of Model IV.

Par	θ	$\widehat{\theta}_{ML}[CI]$	$\widehat{\theta}_{\mathcal{A}_8}[HDR]$
ρ_1	0.260	0.251[0.200,0.301]	0.219[0.147,0.292]
ρ_2	0.270	0.295[0.272,0.318]	0.266[0.219,0.316]
ρ_3	0.450	0.423[0.368,0.477]	0.405[0.357,0.455]
ρ_4	0.500	0.485[0.441,0.529]	0.451[0.400,0.502]
ρ_5	0.430	0.394[0.330,0.458]	0.397[0.351,0.443]
ρ_6	0.330	0.390[0.369,0.412]	0.347[0.297,0.396]
ρ_7	0.220	0.248[0.236,0.260]	0.233[0.183,0.286]
ρ_8	0.160	0.182[0.148,0.216]	0.172[0.131,0.214]
ρ_9	0.230	0.198[0.155,0.241]	0.235[0.197,0.274]
ρ_{10}	0.260	0.283[0.255,0.311]	0.268[0.221,0.317]
ρ_{11}	0.310	0.287[0.284,0.290]	0.273[0.213,0.333]
ρ_{12}	0.290	0.235[0.193,0.278]	0.248[0.180,0.317]
ρ_{13}	0.160	0.183[0.107,0.259]	0.161[0.110,0.213]
ρ_{14}	0.070	0.082[0.055,0.109]	0.078[0.056,0.101]
ρ_{15}	0.190	0.202[0.174,0.229]	0.202[0.169,0.235]
ρ_{16}	0.210	0.191[0.187,0.196]	0.201[0.162,0.239]
ρ_{17}	0.230	0.244[0.191,0.296]	0.217[0.167,0.266]
ρ_{18}	0.240	0.264[0.233,0.294]	0.237[0.179,0.297]
ρ_{19}	0.250	0.232[0.206,0.257]	0.235[0.182,0.290]
μ	7.720	7.705[7.690,7.720]	7.713[7.672,7.753]
σ	0.850	0.841[0.823,0.859]	0.853[0.827,0.879]
ρ_Z	0.987	0.987[0.986,0.989]	0.986[0.980,0.991]

interval. The function for the *pbsm* has the following form, cp. Zinn and Würbach (2014, p. 7):

$$v_b(y) = \begin{cases} \eta_b C_b(z_i), & \text{if } y \in I_b, \text{ for } y \neq h_b \\ 0, & \text{otherwise,} \end{cases} \tag{4.1}$$

where $\eta_b \in [0, 1]$. Let $C_b(z_i)$ be defined as follows:

$$C_b(z_i) = \begin{cases} \exp\left\{ -2\xi_b^{-2}(z_i - h_b)^2 \right\}, & \text{if } z_i \in I_b, \text{ for } z_i \neq h_b, \\ 0, & \text{otherwise,} \end{cases}$$

with ξ_b equalling $0.5(u_b - l_b)$.

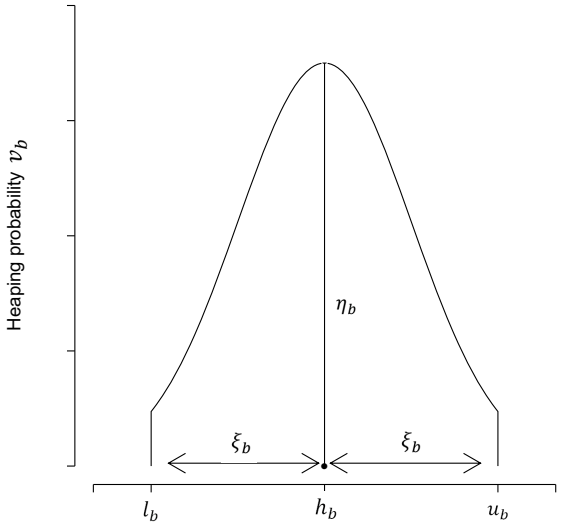


Figure 4.6: Illustration of the piecewise bell-shaped heaping mechanism with steadily increasing/decreasing probabilities for heaping.

Figure 4.6 illustrates the piecewise bell-shaped heaping probability function. As well as in the *pcm*, the catchment areas are by definition allowed to overlap, the functional form resembles a multinomial distribution and the probability for y not being heaped is $1 - v_b(y)$. The vector ϕ is now defined to comprise the alternative parameters of the heaping mechanism, $\phi = (\eta_b)_{b\in\{1,\dots,S\}}$.

Along estimation, $v_b(y_i)$ from Equation (4.1) has to be computed. Whereas in the *pcm* $v_b(y_i)$ is simply p_b, cp. Equations (2.2) to (2.4) in Section 2.4 on pages 52 to 53, in case of *pbsm*, $v_b(y_i)$ cannot be derived as easily. Furthermore, depending on the distinct probabilities caused by the curvature within each interval, $v_b(y_i)$ does not equal $v_b(z_i)$. Zinn and Würbach (2014, 2015) suggest a way to determine g_2 by using the heaping probability $v_b[\widehat{y}_{i,b,(0.5)}]$ of an approximation $\widehat{y}_{i,b,(0.5)}$ of the median value $y_{i,b,(0.5)}$ of y_i within the interval I_b of heaping point h_b, instead of $v_b(y_i)$. That is, during model estimation at every iteration step an approximation $\widehat{y}_{i,b,(0.5)}$ of $y_{i,b,(0.5)}$ is computed by means of the current guess of the assumed latent distribution. Accordingly, the following representation can be given for function g_2:

$$g_2^*(z_i|\psi,\phi) = v_b[\widehat{y}_{i,b,(0.5)}]\,[F(u_b|\psi) - F(l_b|\psi)]\,.$$

The parameters of the heaping mechanism η_b and their sums have to range between zero and one. Hence, constraint (ii) of the constraint system $\mathcal{C}_\Theta$ (cp. Section 2.3) now changes to:

(i) $\eta_b \in [0,1]$ for all $b = 1,\dots,S$,

(ii) $\sum_{b:z_i\in I_b} \eta_b C_b(z_i) \in [0,1]$ for all $z_1,\dots,z_n$.

The corresponding log-likelihood function is defined as (cp. Zinn & Würbach, 2015):

$$\ell(\theta|z_i) = \left(\left[1 - \sum_{b=1}^{S} \eta_b C_b(z_i) \right] f(z_i|\psi) dz_i \right) \mathbb{I}(z_i \in \mathbb{R}_0^+) + \left(\sum_{b=1}^{S} \eta_b C_b(z_i) \left[F(u_b|\psi) - F(l_b|\psi) \right] \right) \mathbb{I}(z_i \in \mathcal{H}). \tag{4.2}$$

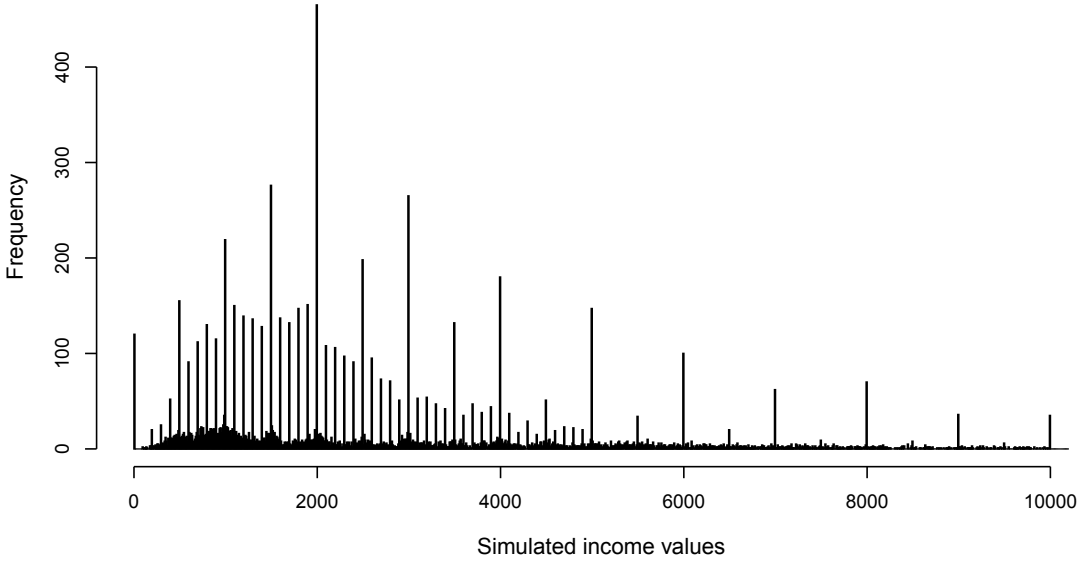


Figure 4.7: Data example of Model V with steadily increasing/decreasing probabilities for heaping.

The probabilities of true values to fall on a modulo in a given set are shown in Table 4.11. The corresponding percentages of heaped values in the data example of Model V are given in Table 4.12. In sum, 50.02% of the simulated values are heaped in the accordant data example.

Table 4.11: Heaping probabilities in Model V.

Interval	mod(100)	mod(500)	mod(1000)
$[0, 500]$	0.26	0.14	–
$(500, 1000]$	0.29	–	0.07
$(1000, 1500]$	0.39	0.16	–
$(1500, 2000]$	0.42	–	0.15
$(2000, 3000]$	0.37	0.19	0.14
$(3000, 4000]$	0.31	0.18	0.17
$(4000, 5000]$	0.22	0.10	0.20
$(5000, 10{,}000]$	–	0.08	0.21

Table 4.12: Percentages of heaped values in the data example of Model V.

Interval	mod(100)	mod(500)	mod(1000)	Total
$[0, 500]$	0.88	1.43	–	2.31
$(500, 1000]$	4.00	–	2.02	6.02
$(1000, 1500]$	5.02	2.52	–	7.54
$(1500, 2000]$	5.23	–	4.48	9.71
$(2000, 3000]$	6.52	1.87	2.50	10.89
$(3000, 4000]$	3.34	1.31	1.74	6.39
$(4000, 5000]$	1.56	0.50	1.43	3.49
$(5000, 10{,}000]$	–	0.72	2.95	3.67
Total	26.55	8.35	15.12	50.02

Initial values for estimation of the model parameters are set to the same values as in Model III, i.e. $\varphi = [0.1]_{b=1}^{19}$ and $\upsilon = (7.714, 0.839, 0.990)'$. With approx. 30 minutes, the ML procedure needs more time to fit the model to the data than in Models I and III (23 and 26 minutes). The increment is considerable for estimation via RMB-RWM algorithm. The according runtime increases by factor 2.5 which is mainly attributable to the increased effort needed to compute the likelihood function of a model with *pbsm*. For values that fall on a heaping point, the corresponding true value has to be approximated in order to compute the related heaping probability. Additionally, the related heaping probability function has to be evaluated at each iteration. In contrast, in the *pcm*, all values in the catchment of a heaping point have the same probability to fall on it.

Alike in Models III and IV, the $MCMC$ estimates are closer to the true values than the ML estimates, see Table 4.13. The corresponding averaged squared biases are 0.00061 for the RMB-RWM and 0.00870 for ML. As in Model III, the parameter estimates ρ_3, ρ_4, and ρ_5 are conspicuous. All three parameter values are now significantly underestimated, even stronger by ML estimation.

Assume a heaping mechanism with less parameters

Finally, a model is assumed whose piecewise constant heaping mechanism is described only by nine heaping probabilities (Model VI), three heaping probabilities for each of the modulos. To be concrete, the equality constraints are adapted in such a way that the range of income values is now divided into three parts for each modulo yielding nine heaping probabilities. Table 4.14 shows the corresponding probabilities of true values to fall on a modulo used for simulation. Reducing the number of components of ϕ, and hence the number of model parameters, is expected to ease estimation. The percentages of heaped values in the data example of Model VI for each modulo in each set are given in Table 4.15. In sum, 49.16% of the simulated values are heaped. Figure 4.8 shows the corresponding distribution.

Table 4.13: Parameter estimates and 95% confidence intervals (CI) or 95% highest density region (HDR) for the data example of Model V.

Par	θ	$\widehat{\theta}_{ML}[CI]$	$\widehat{\theta}_{\mathcal{A}_8}[HDR]$
η_1	0.260	0.127[0.081,0.173]	0.205[0.145,0.272]
η_2	0.290	0.168[0.151,0.185]	0.271[0.225,0.318]
η_3	0.390	0.192[0.157,0.227]	0.332[0.288,0.376]
η_4	0.420	0.242[0.224,0.261]	0.368[0.325,0.412]
η_5	0.370	0.233[0.228,0.238]	0.353[0.310,0.395]
η_6	0.310	0.175[0.159,0.191]	0.280[0.234,0.327]
η_7	0.220	0.136[0.130,0.141]	0.210[0.154,0.266]
η_8	0.140	0.089[0.075,0.102]	0.135[0.092,0.178]
η_9	0.160	0.094[0.088,0.100]	0.161[0.125,0.196]
η_{10}	0.190	0.096[0.082,0.110]	0.172[0.129,0.218]
η_{11}	0.180	0.107[0.090,0.124]	0.182[0.127,0.238]
η_{12}	0.100	0.064[0.056,0.073]	0.122[0.066,0.181]
η_{13}	0.080	0.049[0.042,0.056]	0.090[0.050,0.130]
η_{14}	0.070	0.052[0.035,0.069]	0.071[0.049,0.092]
η_{15}	0.150	0.091[0.078,0.104]	0.172[0.141,0.204]
η_{16}	0.140	0.077[0.062,0.092]	0.152[0.114,0.191]
η_{17}	0.170	0.112[0.070,0.154]	0.157[0.110,0.205]
η_{18}	0.200	0.129[0.100,0.159]	0.180[0.132,0.230]
η_{19}	0.210	0.143[0.113,0.173]	0.203[0.159,0.248]
μ	7.720	7.717[7.710,7.725]	7.722[7.678,7.766]
σ	0.850	0.860[0.848,0.871]	0.872[0.844,0.899]
ρ_Z	0.987	0.988[0.987,0.988]	0.988[0.982,0.992]

Table 4.14: Heaping probabilities in Model VI.

Interval	mod(100)	mod(500)	mod(1000)
$[0, 1500]$	0.27	0.19	0.08
$(1500, 4000]$	0.52	0.20	0.17
$(4000, 10{,}000]$	0.38	0.13	0.23

Table 4.15: Percentages of heaped values in the data example of Model VI.

Interval	mod(100)	mod(500)	mod(1000)	Total
$[0, 1500]$	6.70	4.48	2.22	13.40
$(1500, 4000]$	17.42	2.90	8.06	28.38
$(4000, 10{,}000]$	2.33	1.25	3.80	7.38
Total	26.45	8.63	14.08	49.16

The initial values for estimation of the model parameters are set to $\varphi = [0.2]_{b=1}^{9}$ and $\upsilon = (7.714, 0.839, 0.990)'$. Results from ML and RMB-RWM estimation are presented in Table 4.16. The estimates are closer to the true parameter values and the CI and the HDR are smaller than in the previous models. This result is not surprising since the uncertainty in estimation decreases because of the reduced set of model parameters. The squared bias averaged over all model parameters is only 0.00005 for ML estimation and 0.00028 for estimation via RWM algorithm.

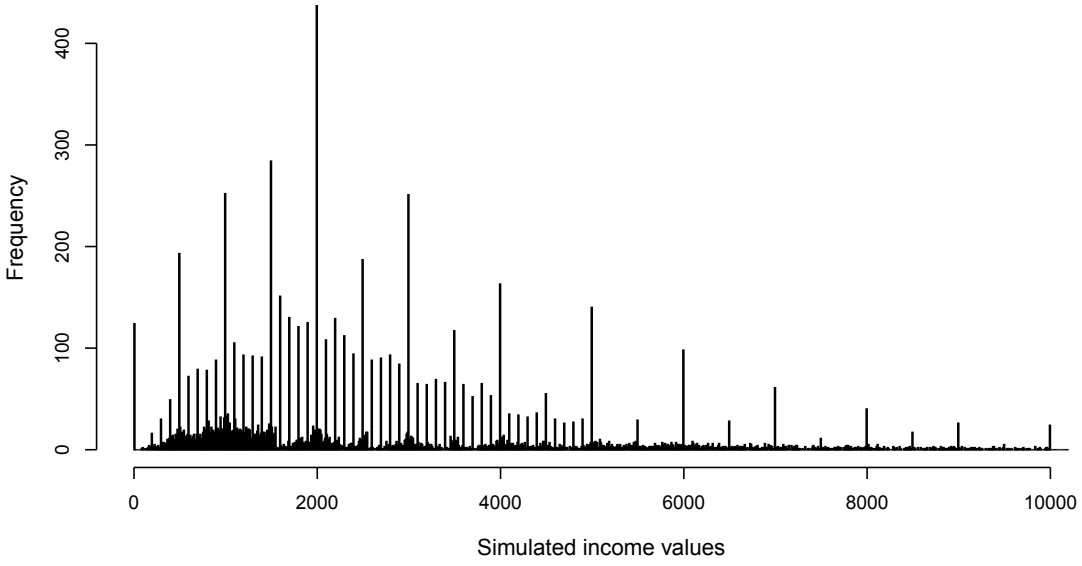


Figure 4.8: Data example of Model VI with less heaping probabilities.

The ML estimates are closer to the true parameter values than the $MCMC$ estimates. However, differences at such decimal places are negligible. The parameter estimates ρ_1 and ρ_2 are slightly underestimated in both cases, but the corresponding HDR cover the true parameter values. A further efficiency gain concerns the runtimes which are much shorter compared to the Models I-V (less than 6 minutes for ML estimation). It will be tested in the application (Chapter 5) whether this gain in efficiency and accuracy comes at the price of an impaired fit to real income data.

Table 4.16: Parameter estimates and 95% confidence intervals (CI) or 95% highest density region (HDR) for the data example of Model VI.

Par	θ	$\widehat{\theta}_{ML}[CI]$	$\widehat{\theta}_{\mathcal{A}_8}[HDR]$
ρ_1	0.270	0.250[0.236,0.264]	0.254[0.214,0.295]
ρ_2	0.520	0.514[0.512,0.516]	0.491[0.452,0.531]
ρ_3	0.380	0.371[0.339,0.404]	0.335[0.269,0.401]
ρ_4	0.190	0.195[0.185,0.204]	0.193[0.158,0.227]
ρ_5	0.200	0.204[0.193,0.214]	0.205[0.165,0.246]
ρ_6	0.130	0.130[0.117,0.143]	0.138[0.090,0.186]
ρ_7	0.080	0.080[0.077,0.084]	0.083[0.057,0.110]
ρ_8	0.170	0.170[0.166,0.173]	0.179[0.149,0.208]
ρ_9	0.230	0.231[0.222,0.240]	0.226[0.179,0.271]
μ	7.720	7.712[7.707,7.717]	7.715[7.669,7.760]
σ	0.850	0.848[0.847,0.850]	0.851[0.820,0.882]
ρ_Z	0.987	0.988[0.987,0.988]	0.987[0.981,0.993]

Summary

Modifying the first heaping model with respect to the latent distribution, the heaping pattern, or the heaping mechanism yields two major findings. First, ML estimation approaches have in general problems with multi-modality which usually occurs in case of mixture models. Using $MCMC$ estimation via RWM instead allows tackling the problem of finding all local modes or a flat global mode within multiple modes, respectively, when mixing well.

Second, the heaping probabilities ρ_3, ρ_4 but also ρ_5 are estimated with relatively high uncertainty in all models, as expressed by the strong tendencies for either under- or overestimation and the wider CI and HDR. These probabilities capture the range $(1000, 3000]$ for mod(100). This range comprises the majority of income values and has the highest concentration of heaped values. In Model VI, the corresponding parameters are ρ_1 and ρ_2, also being estimated with less accuracy. The overall high percentage of heaped values in this range seems to impair estimation more than the lower number of observations in the upper tail of the distribution.

The generality and flexibility of the established model was outlined by several modifications, whereby the modified models were tested for their capability by means of simulations so far. All models proposed are applied to the net individual income data of the Adult Cohort of the German National Educational Panel Study (NEPS SC6) in Chapter 5.

4.2 Extension of the heaping model to a multivariate context

Instead of modeling income data purely by means of a univariate distribution, one might consider a multivariate model that predicts income on the basis of some personal characteristics. However, these covariates might also have an effect on the heaping behavior. Figure 4.9 gives an overview with regard to the embedding of internal factors into the previous concept. The analyses presented in Section 1.2.2 revealed that gender, age, and educational level have a significant influence on the income level and the tendency to heap in the NEPS data. In the following model specification, they are used as independent variables in a log-linear response model to model income data. Furthermore, factors with respect to gender and educational level are included into the heaping mechanism altering the heaping probabilities for females and respondents with low and middle educational level.

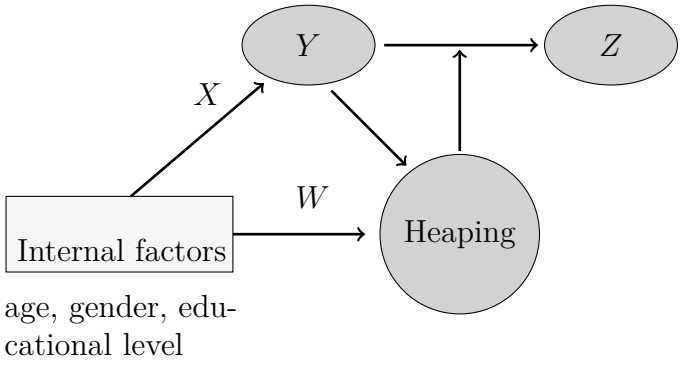


Figure 4.9: Relationships between heaping and internal factors.

Let X denote a matrix of covariates that determine the logarithmized income Y, and W the matrix of covariates that determine the heaping mechanism. The corresponding parameters are β for X and γ for W. The covariate matrix X is composed of a vector of ones for the intercept, a binary variable (gender), an ordered variable with three categories (educational level - low, middle, high), and a continuous variable (age). The matrix W is a subset of X excluding age.

4.2.1 Adding covariates to model income level

In the first extension of the heaping model, the effects of covariates on income level are included into the latent distribution (Model VII). The specifications of the heaping behavior correspond to those of Model VI. The following log-linear response model is used to generate the outcome data (cp. Zinn, 2014):

$$\log(Y) = \beta_0 + \beta_1 \mathbb{I}(X_1 = \text{male}) + \beta_{21} \mathbb{I}(X_2 = \text{middle}) + \beta_{22} \mathbb{I}(X_2 = \text{high}) + \beta_3 X_3 + \epsilon \tag{4.3}$$

The likelihood function in Equation (2.5) on page 54 is adapted by replacing the parameters of the latent distribution ψ by the log-linear model $X\beta$ as follows:

$$g(z_i|X\beta,\phi) = g_1(z_i|X\beta,\phi)\mathbb{I}(z_i \in \mathbb{R}_0^+) + g_2(z_i|X\beta,\phi)\mathbb{I}(z_i \in \mathcal{H}). \tag{4.4}$$

Equation (2.3) and Equation (2.4) are now

$$g_1(z_i|X\beta,\phi) = [1 - v(z_i|\phi)]\, f(z_i|X\beta), \text{ and} \tag{4.5}$$

$$g_2(z_i|X\beta,\phi) = v(z_i|\phi)\,[F(u_b|X\beta) - F(l_b|X\beta)]\,. \tag{4.6}$$

Hence, the log-likelihood function for all observations with integrated effects of covariates on income level is

$$\ell(\mathbf{z}|X\beta,\phi) = \sum_{i=1}^{N} \ln g(z_i|X\beta,\phi). \tag{4.7}$$

Specification and estimation of Model VII

The DGP of Model VII differs from the DGP of Model VI only in the following: the true income values are generated by letting

gender following a binomial distribution $X_1 \sim Bin(0.52)$,

educational level following a multinomial distribution $X_2 \sim \mathcal{M}(0.2, 0.5, 0.3)$,

age following a truncated-normal distribution $X_3 \sim t\mathcal{N}(0, \text{Inf}, 47, 6^2)$, and

the error term $\epsilon \sim \mathcal{N}(0, \iota^2)$.

The coefficients of the model are set to the following values to mirror the picture of the real data and the percentage of explained variance by the model, cp. Table 1.6 on page 43:

$$\beta_0 = 6.3,\ \beta_1 = 0.6,\ \beta_{21} = 0.14,\ \beta_{22} = 0.56,\ \beta_3 = 0.01,\ \iota = 0.62.$$

Because of the increased number of parameters for the latent distribution, the reduced set of heaping probabilities (as in Model VI) was selected in order to keep computational burden comparatively small. The heaping mechanism is therefore described by nine heaping probabilities (cp. Table 4.14). Table 4.17 gives the percentages of heaped values in the accordant simulated data example.

Table 4.17: Percentages of heaped values in the data example of Model VII.

Interval	mod(100)	mod(500)	mod(1000)	Total
$[0, 1500]$	11.13	6.34	3.49	20.96
$(1500, 4000]$	16.71	2.39	7.31	26.41
$(4000, 10{,}000]$	1.27	0.70	1.24	3.21
Total	29.11	9.43	12.04	50.58

In Model VII, 50.58% of the simulated values are heaped in sum and 146 zeros have been simulated. The mean of the heaped distribution is 1932.03 EUR (SD = 1610.04 EUR) and the median is 1500 EUR, see Table 4.18. Table 4.19 gives the mean income values per subgroup. The mean income values separated by gender closely correspond to the observed values in the NEPS data, cp. Table 1.5. The mean income values for lower and middle educated individuals differ between observed and simulated data. However, the values for gender and educational level again closely mirror the findings from the NEPS data.

Table 4.18: Descriptives of the data examples for the extended modeling strategies.

Descriptive	Model VII	Model VIII
Latent distribution	log-norm	log-norm
Perc of heaping	∼50	∼43
Heaping intervals	sym.	sym.
Heaping function	*pcm*	*pcm*
Components of ϕ	9	9
Covariates	on income	on income and HM
Perc of zeros	1.46	1.46
Mean	1885.62	1884.88
Median	1500.00	1500.00
SD	1440.21	1440.41
Run.* RMB-RWM	14.39	16.14

*Runtime is given in hours.

Table 4.19: Mean statistics of the data examples for the extended modeling strategies, divided by subgroups.

Group	Model VII	Model VIII
Female	1347.16	1345.01
Male	2462.25	2462.79
Lower edu	1495.72	1496.12
Middle edu	1707.92	1706.88
Higher edu	2620.58	2619.59
Female lower edu	1028.34	1028.24
Female middle edu	1208.11	1205.15
Female higher edu	1813.07	1810.96
Male lower edu	1906.90	1907.74
Male middle edu	2172.91	2173.65
Male higher edu	3334.72	3334.72

The initial values for estimation by RMB-RWM algorithm are set to $\varphi = [0.2]_{b=1}^{9}$ for the ρ_b's. The initial values for the parameters of the assumed latent model are derived from fitting a log-linear model to the non-zero simulated data, i.e. $\upsilon = (\beta_0, \beta_1, \beta_{21}, \beta_{22}, \beta_3)' = (6.30, 0.59, 0.15, 0.55, 0.01)'$. The inflation parameter for the zeros and the variance for the random perturbation are initiated with $\rho_Z = 0.99$ and $\iota = 0.5$, respectively. Furthermore, these values are considered as prior means with covariance matrix $\Sigma = 0.001\mathcal{I}_9$ for the parameter vector ϕ and covariance matrix $\Upsilon = \text{diag}(0.001, 0.1, 0.01, 0.01, 0.01, 0.001, 0.01)\mathcal{I}_7$ for the parameter vector ψ. The current draws are used as mean vector for the proposal density and the covariance matrix is fixed to $\Omega = \text{diag}(0.0001, \ldots, 0.0001, 0.00001, 0.001, 0.0001, 0.0001, 0.0001, 0.00001, 0.0001)\mathcal{I}_{16}$.

Table 4.20: Parameter estimates and 95% confidence intervals (CI) or 95% highest density region (HDR) for the data examples of Model VII and Model VIII.

Par	θ	$\widehat{\theta}_{\mathcal{A}_8}[HDR]$ Model VII	$\widehat{\theta}_{\mathcal{A}_8}[HDR]$ Model VIII
ρ_1	0.270	0.269[0.206,0.334]	0.265[0.192,0.337]
ρ_2	0.520	0.492[0.412,0.569]	0.487[0.392,0.578]
ρ_3	0.380	0.286[0.143,0.436]	0.264[0.119,0.444]
ρ_4	0.190	0.187[0.130,0.242]	0.189[0.116,0.263]
ρ_5	0.200	0.201[0.122,0.282]	0.201[0.110,0.293]
ρ_6	0.130	0.197[0.048,0.347]	0.160[0.053,0.274]
ρ_7	0.080	0.085[0.046,0.128]	0.092[0.048,0.140]
ρ_8	0.170	0.178[0.118,0.236]	0.186[0.123,0.250]
ρ_9	0.230	0.191[0.085,0.301]	0.231[0.097,0.356]
ρ_Z	0.987	0.983[0.970,0.996]	0.984[0.971,0.995]
β_0	6.300	6.108[4.313,7.977]	7.389[6.650,7.938]
β_1	0.600	0.597[0.498,0.691]	0.600[0.494,0.713]
β_{21}	0.140	0.106[-0.023,0.231]	0.147[-0.026,0.295]
β_{22}	0.560	0.526[0.384,0.677]	0.567[0.373,0.718]
β_3	0.010	0.014[-0.026,0.054]	-0.014[-0.025,0.002]
ι	0.620	0.599[0.535,0.659]	0.613[0.567,0.658]
γ_1	0.800	–	0.794[0.692,0.903]
γ_2	0.900	–	0.894[0.803,0.993]

Runtime of the extended heaping model lasts 14.39 hours for RMB-RWM estimation. This increase results from evaluation of the log-likelihood, which is now approx. 0.7 seconds for each sampled draw owing to the regression step instead of the simple univariate evaluation. Results from estimation are presented in the third column of Table 4.20. The parameter estimates are all within the HDR. Most of the estimates are close to the true values, with except of ρ_3, ρ_6, and β_0. Both, ρ_3 and ρ_6 cover the interval $(4000, 10{,}000]$, either for mod(100),

or mod(500). The higher inaccuracy might be explained by the low number of observations falling on both modulos in this range. The wider HDR of these estimates strengthen this assumption. The parameter ρ_9, also covering the range $(4000, 10{,}000]$, is estimated quite accurately. However, the remarkably wider HDR of the estimate of ρ_9 indicates an increased uncertainty also with respect to mod(1000) in the highest range. The overall assessment gives a squared bias averaged over all model parameters in the amount of 0.0035, and 0.0017 when only considering the heaping probabilities.

4.2.2 Adding covariates to model income level and the heaping mechanism

The covariate model described in the previous section is extended to further regard influences of gender and educational level on the heaping mechanism (Model VIII). Thus, in the likelihood function, the parameters of the heaping mechanism ϕ are additionally extended by covariate specific factors $W\gamma$ that act multiplicative on the propensity to heap. The likelihood function $\mathcal{L}$ of observing one specific income value is now

$$g(z_i|X\beta, W\gamma) = g_1(z_i|X\beta, W\gamma)\mathbb{I}(z_i \in \mathbb{R}_0^+) + g_2(z_i|X\beta, W\gamma)\mathbb{I}(z_i \in \mathcal{H}). \tag{4.8}$$

with

$$g_1(z_i|X\beta, W\gamma) = [1 - v(z_i|W\gamma)]\, f(z_i|X\beta), \text{ and} \tag{4.9}$$

$$g_2(z_i|X\beta, W\gamma) = v(z_i|W\gamma)\, [F(u_b|X\beta) - F(l_b|X\beta)]\,. \tag{4.10}$$

The related log-likelihood function for all observations is

$$\ell(\mathbf{z}|X\beta, W\gamma) = \sum_{i=1}^{N} \ln g(z_i|X\beta, W\gamma). \tag{4.11}$$

Specification and estimation of Model VIII

In the accordant DGP of Model VIII, the heaping probabilities of female individuals are decreased by 20% compared to those of male individuals. For individuals with low or middle educational level, the probabilities for heaping are decreased by 10% compared to those of higher educated individuals. The implementation is demonstrated in Zinn (2014) for one skill factor determining the heaping mechanism.

Table 4.21 gives the percentages of heaped values in a corresponding simulated data example. In sum, 42.78% of the simulated values fall on modulos. The decline of heaped values compared to the preceding simulation model (Model VII) is due to the decreased heaping probabilities for females and lower or middle educated individuals. The remaining descriptives considered are almost similar to

the previous data example with mean at 1931.28 EUR ($SD = 1610.24$ EUR) and median at 1500 EUR, see Table 4.18 and Table 4.19.

For illustration on how the factor for gender alters the heaping probabilities, Figure 4.10 depicts the income values by gender. In this figure, also the gender effect on income is clearly visible. Table 4.22 gives the percentages of heaped and non-heaped values by subgroup. A lower percentage of heaped values results for females, as opposed to male individuals (35.4% vs. 49.4%). The proportion of heaped values is increased by 7% for individuals with higher educational level as compared to individuals with low or middle educational level. The differences are less pronounced than in the NEPS data (cp. Table 1.7) which is largely attributable to the fact that the overall percentage of heaped income values is smaller in the data example than in the observed data (42.8% vs. 69.2%).

Table 4.21: Percentages of heaped values in the data example of Model VIII.

Interval	mod(100)	mod(500)	mod(1000)	Total
$[0, 1500]$	8.61	4.99	3.00	16.60
$(1500, 4000]$	14.48	2.05	6.69	23.22
$(4000, 10{,}000]$	1.14	0.60	1.22	2.96
Total	24.23	7.64	10.91	42.78

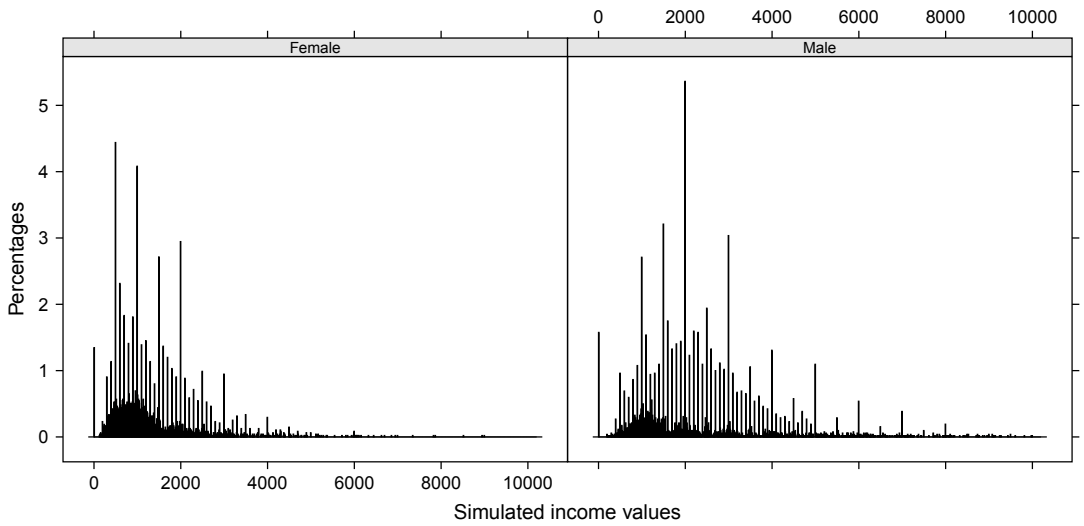


Figure 4.10: Data example of Model VIII with reduced heaping probabilities for female individuals.

Table 4.22: Percentages for observing heaping in Model VIII by gender or educational level.

Group	Heaping	No heaping
Female	35.44	64.56
Male	49.44	50.56
Lower edu	39.60	60.40
Middle edu	41.04	58.96
Higher edu	47.99	52.01

As compared to the foregoing model (Model VII), the specifications for γ have to be added in the estimation model. The starting values and the prior means for γ are set to 1. This simply means that a priori both effects on heaping (due to gender or educational level) are assumed to be without any influence. Variances for γ are fixed to 0.01 each. The covariance matrix of the proposal density is adjusted to $\Omega = \text{diag}(0.0001, \ldots, 0.0001, 0.00001, 0.001, 0.0001, 0.0001, 0.0001, 0.00001, 0.0001, 0.0001, 0.0001)\mathcal{I}_{18}$.

Estimation of Model VIII increases to 17.56 hours. The accordant results are given in the fourth column of Table 4.20. With except of β_0 and β_3, all model parameters are within the HDR. As a whole, the HDR are not much wider in Model VIII, opposed to those of Model VII, indicating that the inclusion of the factors for gender and educational level determining the HM does not yet increase the uncertainty of estimation. The squared bias averaged over all model parameters is 0.0668 in Model VIII (0.0018 when only considering ϕ). To a large part, the deviations owe to the imprecisely estimated β_0.

Summary

Multivariate modeling of income data in form of a log-linear response model provides feasible results. Even when including further two parameters to affect the HM the estimation of the parameters of the HM is not hampered. Because of the fact that the accuracy in estimation of the model parameters is still high, it may be concluded that an increase in the number of parameters – and thereby determinants for income level and/or the HM – is still possible until the overall fit gets deteriorated.

The better performance of the RMB- vs. the S-RWM algorithm becomes striking in the extended examples. Both models are additionally estimated via S-RWM algorithm. The respective estimates are given in Table A.5 in the Appendix for reasons of comparability. The parameters are poorly estimated by S-RWM and the averaged squared biases are increased (Model VII: 0.0048, Model VIII: 0.0598). Thus, referring to RMB-RWM estimation is beneficial in models with increased complexity.

Chapter 5

Application of the heaping model to NEPS data

All modeling strategies introduced in the previous chapters – Models I-VI as well as the extended models (Models VII and VIII) – are employed and compared to each other with respect to their fit to the NEPS income data. Formal measures of the relative quality of the given models are calculated for model comparison. Furthermore, posterior predictive checks are used to evaluate replicated data from $MCMC$ estimates with respect to descriptive statistics.

5.1 Apply alternative models to real data

The different models described in Chapter 2, Section 4.1, and Section 4.2 are now applied to real income data of the NEPS Adult Cohort Wave 1 (cp. Section 1.2.1). To avoid the problem of small cells, observations that exceed 10,000 EUR are not used during estimation processes inducing a loss of 26 observations. This proceeding was suggested and implemented by J. M. Roberts and Brewer (2001, p. 888), Serfling (2006, p. 102), and Marcus et al. (2013, p. 24). In a first place, all six heaping models without covariates are considered and their outcomes will be compared. Model estimation is performed using the ML approach described in Section 3.1 and further by the RMB-RWM algorithm described in Section 3.2.4. Afterwards, both models with covariates are applied to the NEPS data. To account for the variability of $MCMC$ estimates, 20 runs with differing starting values are performed for each model and the averaged estimates are displayed. The tools for model comparison are: AIC, BIC and SC with respect to the ML estimates, and averaged log-marginal likelihood for RWM model comparison. Preferable are those models with the lowest AIC and BIC as well as the highest SC and averaged log-marginal likelihood, accordingly. Models VII and VIII are solely estimated by the RMB-RWM algorithm and compared by their log-marginal likelihoods.

Table 5.1: Application to real data, Models I to IV.

Feature	Model I	Model II	Model III	Model IV
Latent distribution	log-norm	Dagum	log-norm	log-norm
Interval widths	HP$\pm\frac{1}{2}$mod	HP$\pm\frac{1}{2}$mod	HP$\pm$mod	HP$\pm\frac{1}{2}$mod
Heaping intervals	sym.	sym.	sym.	asym.
Heaping function	*pcm*	*pcm*	*pcm*	*pcm*
Components of ϕ	19	19	19	19
Covariates	–	–	–	–
AIC ML	97,835.7	96,712.4	98,240.7	99,648.2
BIC ML	97,991.3	96,875.0	98,396.2	99,803.8
SC ML	−48,995.6	−48,437.5	−49,198.1	−49,901.9
$\overline{\log m(\mathbf{z})}$	−49,309.7	−48,872.9	−49,342.6	−50,012.9

Table 5.2: Application to real data, Models V to VIII.

Feature	Model V	Model VI	Model VII	Model VIII
Latent distribution	log-norm	log-norm	log-norm	log-norm
Interval widths	HP$\pm\frac{1}{2}$mod	HP$\pm\frac{1}{2}$mod	HP$\pm\frac{1}{2}$mod	HP$\pm\frac{1}{2}$mod
Heaping intervals	sym.	sym.	sym.	sym.
Heaping function	*pbsm*	*pcm*	*pcm*	*pcm*
Components of ϕ	19	9	9	9
Covariates	–	–	on income	income, HM
AIC ML	98,240.7	98,061.3	–	–
BIC ML	98,396.2	98,146.1	–	–
SC ML	−49,198.1	−49,073.1	–	–
$\overline{\log m(\mathbf{z})}$	−47,060.7	−49,277.4	−43,892.8	−43,915.3

Tables 5.1 and 5.2 present the goodness of fit measures for the considered models with respect to their fit to the NEPS data. The least suitable model is Model IV which assumes asymmetric heaping intervals. The supposition that people underreport their income does not hold for the observed data. Models III and V are also negligible. Both models have the same AIC, BIC and SC, up to the first decimal point. Models I and II show the best goodness of fit measures. That is, assuming symmetric intervals of widths equalling the modulo and a heaping function based on piecewise constant heaping probabilities seems to be preferable over all other settings considered. Using the more complex 3-parametric Dagum distribution instead of the log-normal distribution further improves modeling. Disadvantageously, it slows down the estimation process. Models V, VII, and VIII are highly preferable with respect to their log-marginal likelihoods. No indication is given that the efficiency and accuracy gain in Model VI, due to the reduced number of components of ϕ, comes at the price of an impaired fit to the real income data.

Table 5.3: Parameter estimates and 95% confidence intervals (CI) or 95% highest density region (HDR) for Models I to IV in the application.

	Model I		**Model II**		**Model III**		**Model IV**	
Par	$\widehat{\theta}_{ML}[CI]$	$\overline{\widehat{\theta}}_{RMB}[HDR]$	$\widehat{\theta}_{ML}[CI]$	$\overline{\widehat{\theta}}_{RMB}[HDR]$	$\widehat{\theta}_{ML}[CI]$	$\overline{\widehat{\theta}}_{RMB}[HDR]$	$\widehat{\theta}_{ML}[CI]$	$\overline{\widehat{\theta}}_{RMB}[HDR]$
ϕ_1	0.344[0.202,0.485]	0.358[0.287,0.428]	0.475[0.372,0.579]	0.358[0.295,0.422]	0.159[0.109,0.210]	0.181[0.130,0.232]	0.443[0.301,0.585]	0.330[0.264,0.396]
ϕ_2	0.363[0.250,0.477]	0.371[0.310,0.432]	0.459[0.387,0.531]	0.372[0.316,0.428]	0.202[0.170,0.234]	0.181[0.138,0.224]	0.412[0.297,0.526]	0.323[0.268,0.378]
ϕ_3	0.593[0.529,0.657]	0.559[0.507,0.611]	0.604[0.562,0.646]	0.558[0.511,0.605]	0.273[0.268,0.277]	0.266[0.230,0.301]	0.472[0.301,0.643]	0.505[0.455,0.555]
ϕ_4	0.516[0.322,0.710]	0.551[0.505,0.598]	0.574[0.503,0.645]	0.552[0.510,0.595]	0.289[0.260,0.317]	0.293[0.259,0.325]	0.463[0.283,0.644]	0.504[0.458,0.549]
ϕ_5	0.580[0.514,0.646]	0.544[0.496,0.592]	0.561[0.491,0.631]	0.545[0.502,0.588]	0.308[0.283,0.333]	0.285[0.251,0.318]	0.462[0.282,0.643]	0.508[0.463,0.553]
ϕ_6	0.450[0.295,0.606]	0.448[0.384,0.514]	0.542[0.502,0.582]	0.448[0.390,0.506]	0.270[0.252,0.287]	0.247[0.201,0.292]	0.327[0.062,0.593]	0.430[0.369,0.492]
ϕ_7	0.324[-0.004,0.652]	0.351[0.279,0.423]	0.400[0.198,0.603]	0.356[0.290,0.421]	0.197[0.166,0.229]	0.174[0.120,0.228]	0.168[-0.082,0.418]	0.348[0.279,0.416]
ϕ_8	0.086[0.010,0.162]	0.073[0.037,0.110]	0.043[0.022,0.064]	0.072[0.040,0.106]	0.030[0.026,0.034]	0.037[0.013,0.064]	0.050[0.042,0.058]	0.066[0.034,0.100]
ϕ_9	0.154[0.132,0.177]	0.161[0.128,0.195]	0.148[0.132,0.164]	0.161[0.131,0.192]	0.097[0.086,0.109]	0.095[0.066,0.126]	0.184[0.095,0.273]	0.158[0.125,0.191]
ϕ_{10}	0.144[0.078,0.210]	0.204[0.161,0.248]	0.199[0.166,0.232]	0.204[0.165,0.244]	0.114[0.070,0.158]	0.144[0.098,0.191]	0.178[0.136,0.221]	0.209[0.167,0.252]
ϕ_{11}	0.177[0.105,0.249]	0.242[0.177,0.307]	0.191[0.160,0.223]	0.243[0.187,0.299]	0.161[0.122,0.200]	0.181[0.113,0.250]	0.291[0.157,0.425]	0.238[0.180,0.299]
ϕ_{12}	0.177[0.065,0.289]	0.263[0.180,0.345]	0.303[0.034,0.572]	0.260[0.188,0.333]	0.207[0.169,0.245]	0.198[0.110,0.287]	0.257[0.250,0.264]	0.254[0.178,0.329]
ϕ_{13}	0.228[-0.263,0.719]	0.196[0.109,0.282]	0.088[-0.094,0.271]	0.197[0.117,0.276]	0.184[-0.108,0.476]	0.111[0.029,0.193]	0.272[-0.303,0.847]	0.196[0.110,0.280]
ϕ_{14}	0.090[0.044,0.136]	0.082[0.057,0.107]	0.066[0.058,0.075]	0.082[0.060,0.105]	0.037[0.037,0.038]	0.039[0.023,0.057]	0.088[0.045,0.132]	0.078[0.055,0.102]
ϕ_{15}	0.218[0.054,0.382]	0.167[0.136,0.199]	0.161[0.137,0.185]	0.166[0.138,0.194]	0.071[0.063,0.078]	0.077[0.055,0.101]	0.213[0.081,0.346]	0.168[0.138,0.200]
ϕ_{16}	0.251[0.120,0.383]	0.217[0.178,0.258]	0.210[0.186,0.234]	0.217[0.181,0.253]	0.093[0.087,0.098]	0.100[0.066,0.135]	0.312[0.151,0.474]	0.244[0.204,0.284]
ϕ_{17}	0.364[0.144,0.584]	0.286[0.228,0.345]	0.260[0.251,0.269]	0.288[0.235,0.339]	0.159[0.139,0.178]	0.157[0.098,0.218]	0.340[0.315,0.366]	0.296[0.239,0.351]
ϕ_{18}	0.412[0.409,0.416]	0.323[0.245,0.400]	0.265[0.159,0.371]	0.324[0.251,0.397]	0.179[0.131,0.226]	0.194[0.112,0.278]	0.560[0.294,0.826]	0.334[0.256,0.411]
ϕ_{19}	0.449[-0.444,1.341]	0.412[0.325,0.497]	0.485[-0.487,1.456]	0.411[0.332,0.489]	0.237[-0.069,0.543]	0.302[0.224,0.377]	0.552[-0.552,1.657]	0.414[0.334,0.494]
μ	7.317[7.306,7.328]	7.323[7.282,7.364]	–	–	7.317[7.311,7.323]	7.314[7.260,7.368]	7.346[7.342,7.350]	7.344[7.304,7.385]
σ	0.740[0.731,0.749]	0.737[0.710,0.764]	–	–	0.721[0.705,0.737]	0.731[0.696,0.766]	0.786[0.665,0.907]	0.735[0.709,0.762]
ρ_Z	0.984[0.982,0.986]	0.983[0.976,0.989]	0.980[0.975,0.986]	0.983[0.976,0.988]	0.983[0.983,0.984]	0.982[0.973,0.990]	0.987[0.981,0.993]	0.983[0.976,0.989]
a	–	–	3.687[3.215,4.160]	3.178[3.018,3.337]	–	–	–	–
b	–	–	2199.3[2198.9,2199.8]	2013.2[1982.3,2047.3]	–	–	–	–
q	–	–	0.485[0.422,0.549]	0.612[0.574,0.651]	–	–	–	–

Table 5.4: Parameter estimates and 95% confidence intervals (CI) or 95% highest density region (HDR) for Models V to VIII in the application.

	Model V		**Model VI**		**Model VII**	**Model VIII**
Par	$\widehat{\theta}_{ML}[CI]$	$\widehat{\overline{\theta}}_{RMB}[HDR]$	$\widehat{\theta}_{ML}[CI]$	$\widehat{\overline{\theta}}_{RMB}[HDR]$	$\widehat{\theta}_{RMB}[HDR]$	$\widehat{\overline{\theta}}_{RMB}[HDR]$
ϕ_1	0.159[0.109,0.210]	0.274[0.220,0.330]	0.519[0.496,0.543]	0.480[0.433,0.527]	0.476[0.384,0.566]	0.509[0.409,0.613]
ϕ_2	0.201[0.172,0.230]	0.278[0.230,0.326]	0.560[0.484,0.636]	0.568[0.527,0.609]	0.556[0.475,0.637]	0.572[0.494,0.650]
ϕ_3	0.272[0.267,0.278]	0.425[0.383,0.467]	0.242[0.169,0.316]	0.302[0.215,0.387]	0.284[0.119,0.452]	0.282[0.125,0.444]
ϕ_4	0.289[0.260,0.317]	0.456[0.417,0.496]	0.103[0.082,0.124]	0.122[0.092,0.153]	0.123[0.067,0.182]	0.134[0.076,0.193]
ϕ_5	0.308[0.283,0.333]	0.451[0.413,0.490]	0.182[0.160,0.204]	0.203[0.166,0.240]	0.204[0.130,0.278]	0.198[0.129,0.266]
ϕ_6	0.269[0.253,0.285]	0.390[0.338,0.441]	0.172[0.152,0.193]	0.234[0.147,0.321]	0.243[0.078,0.418]	0.238[0.074,0.406]
ϕ_7	0.196[0.166,0.227]	0.329[0.265,0.392]	0.090[0.081,0.098]	0.093[0.062,0.126]	0.101[0.043,0.164]	0.109[0.047,0.173]
ϕ_8	0.030[0.026,0.034]	0.058[0.029,0.089]	0.217[0.131,0.303]	0.188[0.155,0.221]	0.189[0.126,0.253]	0.194[0.133,0.255]
ϕ_9	0.097[0.086,0.109]	0.155[0.121,0.189]	0.574[0.478,0.669]	0.423[0.335,0.514]	0.384[0.215,0.548]	0.400[0.228,0.572]
ϕ_{10}	0.114[0.069,0.159]	0.221[0.174,0.268]	–	–	–	–
ϕ_{11}	0.161[0.121,0.201]	0.248[0.186,0.309]	–	–	–	–
ϕ_{12}	0.206[0.164,0.248]	0.235[0.154,0.317]	–	–	–	–
ϕ_{13}	0.185[-0.106,0.476]	0.179[0.091,0.270]	–	–	–	–
ϕ_{14}	0.037[0.037,0.038]	0.066[0.046,0.087]	–	–	–	–
ϕ_{15}	0.070[0.062,0.079]	0.130[0.103,0.157]	–	–	–	–
ϕ_{16}	0.093[0.087,0.098]	0.188[0.146,0.230]	–	–	–	–
ϕ_{17}	0.159[0.139,0.180]	0.259[0.196,0.321]	–	–	–	–
ϕ_{18}	0.179[0.133,0.225]	0.292[0.212,0.371]	–	–	–	–
ϕ_{19}	0.237[-0.070,0.544]	0.387[0.310,0.462]	–	–	–	–
μ	7.318[7.310,7.326]	7.338[7.298,7.379]	7.331[7.314,7.348]	7.323[7.273,7.373]	–	–
σ	0.722[0.707,0.736]	0.759[0.733,0.786]	0.739[0.732,0.745]	0.737[0.704,0.770]	–	–
ρ_Z	0.983[0.983,0.984]	0.983[0.976,0.989]	0.982[0.981,0.984]	0.982[0.974,0.990]	0.979[0.962,0.994]	0.979[0.963,0.994]
β_0	–	–	–	–	6.449[4.248,8.929]	6.127[3.525,8.084]
β_1	–	–	–	–	0.599[0.439,0.755]	0.605[0.455,0.753]
β_{21}	–	–	–	–	0.129[-0.151,0.397]	0.143[-0.103,0.407]
β_{22}	–	–	–	–	0.555[0.325,0.791]	0.562[0.338,0.790]
β_3	–	–	–	–	0.007[-0.043,0.051]	0.014[-0.027,0.066]
ι	–	–	–	–	0.586[0.408,0.711]	0.572[0.327,0.717]
γ_1	–	–	–	–	–	0.940[0.878,0.998]
γ_2	–	–	–	–	–	0.971[0.927,1.001]

The parameter estimates from application of all eight models are presented in Tables 5.3 and 5.4. ML and RWM estimates are given for Models I-VI, but only for Model III the estimated parameter values of both techniques closely resemble each other. In most cases, large differences exist between ML and RWM estimates for individual parameters or even blocks of parameters. To give some examples, the ML estimates for ρ_{15-19} in Models I and IV are significantly larger than the RWM estimates. This parameter block corresponds to mod(1000) and the findings suggest a stronger tendency to heap values to thousands according to the ML estimates. With respect to Model II, the ML estimates are larger than the RWM estimates in the parameter block for mod(100) (ρ_{1-7}), suggesting a stronger tendency to heap values to hundreds compared to the RWM estimates. Regarding Model V, the ML estimates for all η (with except of η_{13}) are considerably smaller than the RWM estimates. This indicates variable behavior between both estimation techniques, in particular for the *pbsm* function. Further checks are needed to evaluate whether the ML or the RWM estimates capture the real structure of the NEPS data better. Finally, the negative lower CI bound for ρ_7 in Model I and Model IV as well as for ρ_{13} and ρ_{19} in Models I-V is striking. The accordant probabilities correspond to the respective highest range in each modulo considered. Concretely, ρ_7 refers to the range $(4000, 5000]$ for mod(100), and ρ_{13} as well as ρ_{19} refer to the range $(5000, 10{,}000]$ for mod(500) or mod(1000), respectively. The large widths of the corresponding CI and HDR underline the uncertainty in parameter estimation using ML or RMB-RWM.

The following findings relate exclusively to the RWM estimates. First, the parameter estimates of the underlying log-normal distribution are of the same magnitude across all models with zero-inflated log-normal distribution or those two with log-linear response model. In contrast, the parameter estimates of the heaping mechanism often differ largely across the models owing to the different interpretations of the model parameters in the considered settings. Only Models I and II (to some extent also Model IV), or Models VI-VIII respectively, show quite similar estimates for ϕ among each other. Third, the estimated parameter values of Model III are expectedly lower which is attributable to the wider intervals assumed in Model III. More values are likely for being heaped to a certain HP lowering the heaping probabilities for that particular interval. Fourth, the estimates of ϕ are also smaller in Model V but not as low as in Model III. This is caused by the wider intervals as well, but the fact that the η_b are in general greater than the ρ_b corresponds to the functional form. Strictly speaking, the parameters of the heaping mechanism in Model V are not to be interpreted as probabilities alike in the *pcm*, since they represent the height of the *pbsm* function. Lastly, comparison of the models with or without covariates reveals that the HDR for the estimates of ρ_b of Models VII and VIII are wider compared to the corresponding HDR of Model VI without covariates. This indicates an increase of uncertainty in estimation of the models with regression-based modeling of income.

The estimated parameter values are exemplarily interpreted for Model II, the model with the best goodness of fit measures. The by far highest heaping probabilities (ρ_{3-5}) indicate that most of the values fall on HP of mod(100) within the range $(1500, 3000]$. These probabilities are followed by ρ_1, ρ_2, ρ_6, ρ_7, ρ_{18}, and ρ_{19}. The parameters ρ_1, ρ_2, ρ_6, and ρ_7 complete the set for mod(100), i.e. the probabilities to heap values at hundreds. The parameters ρ_{18} and ρ_{19} refer to mod(1000) and correspond to the upper tail of the distribution. The results corroborate the findings from Section 1.2.1 that the tendency for heaping at hundreds is the largest, compared to heaping at five-hundreds or thousands. Furthermore, the propensity for heaping at thousands increases with income level due to level dependency (cp. Table 1.8 on page 45).

Further analyses are taken into consideration for a better interpretation of the parameter estimates. Concretely, posterior predictive checks are used but also a graphical inspection by quantile-quantile plots (QQ-plots) is considered.

5.2 Posterior predictive checks

Posterior predictive checks (PPC) according to Gelman, Meng, and Stern (1996) and Rubin (1984, p. 1165) are used to assess the quality of the estimated heaping models. For PPC, posterior predictive reference sets – also called replicated data sets – are generated such that those reflect a repetition of the data collection at a later time but with the same model $\mathcal{M}$ and the same, but unknown, model parameters that produced the current data (cp. Gelman, Meng, & Stern, 1996, p. 738). To generate the replicated data sets $\mathcal{R}$, draws from the posterior distribution of the model parameters estimated owing to model $\mathcal{M}_o$ are used for simulation under the respective model.[1]

PPC can be used for model evaluation in general but are especially advisable when comparing multiple models with respect to their fit to the real data. The true parameter values are not known in advance and other graphical diagnostics like residual plots are not applicable in the considered case. With PPC one is able to assess how well the true DGP can be replicated on the basis of the parameter estimates received. Furthermore, large occasional differences can be detected by resorting to the examination of repeated draws from the posterior distributions. In cases when certain statistics (e.g. mean) derived from each replicate consistently deviate from real data in one direction – indicated by a small two-sided *posterior predictive p-value* (*ppp*) – one can conclude that the estimated model is not able to replicate the observed data. For example, Wang et al. (2012, p. 1695), Mathew et al. (2012) and Wright and Bray (2003) use this kind of diagnostic check to assess the ability of their model(s) for preservation of the inherent data structure. However, the explicit objective of this approach is to learn about the nature of the departures rather than to reject a specific model (Wright & Bray, 2003, p. 9).

[1] There is a clear correspondence to Bootstrap tests, if only one parameter set is used for simulation of $\mathcal{R}$ (as is the case for ML estimates).

In the present case, the PPC proceeds as follows: for each of the 20 independent $MCMC$ runs of all Models (I-VIII) – denoted by $\mathcal{A}_o$, the set of independent $MCMC$ under model o – a corresponding set of replicated data $\mathcal{R}_o$ is generated containing the replicated data $\mathbf{z}^{(r)}$, with $r = 1, \ldots, 100$. As samples from the posterior distribution of the model parameters serve the posterior means of each independent run for each model o. After generation, each $\mathbf{z}^{(r)}$ is then compared to the observed data $\mathbf{z}$ with respect to their mean, quartiles, and proportions of modulos.

The results from PPC are given in Tables 5.5 and 5.6. Highly significant differences (*ppp*-value < 0.01) are bolded. The replicated data from all eight models have a good fit to the real data with respect to the mean, but also deviations of the 1. and 3. quartile are small (below 100) despite being partially significant. The relatively large deviations in median are striking. Almost all deviations amount to approx. 200 and all are significant. Inspection of the ranges and their signs reveals sometimes large, and in particular systematic, differences between the replicated and the observed data for several statistics used for analysis. Especially the median and the 1. quartile are significantly underestimated, whereas in most cases, the 3. quartile is overestimated. Models VII and VIII show the lowest deviations supporting the first impression from the log-marginal likelihood according to their superior model fit. Regarding these statistics (point estimates), the models seem to systematically distort the original data structure in part, they should have captured ideally. On this account, the following analysis evaluates whether the models can replicate the distribution on the whole.

Table 5.5: Averaged absolute differences of descriptive statistics and their ranges between real and replicated data for Models I to IV.

Descriptive	Model I	Model II	Model III	Model IV
Mean	24.98	44.41	23.68	**29.32**
	[-6.86,58.66]	[-73.17,2.53]	[-6.47,66.65]	[0.85,61.36]
Median	**200.05**	**144.51**	**200.00**	**200.08**
	[-203.83,-200.00]	[-194.76,-100.00]	[-200.00,-200.00]	[-208.17,-200.00]
1. Quartile	**90.56**	0.06	**99.90**	**85.59**
	[-100.00,-54.38]	[-5.59,0.00]	[-101.78,-93.84]	[-100.00,-45.21]
3. Quartile	79.07	**83.53**	**95.16**	85.35
	[-2.00,98.00]	[-102.00,-2.00]	[38.36,98.00]	[-2.00,98.00]

Note: significant differences (*ppp*-value < 0.01) are bolded.

Table 5.6: Averaged absolute differences of descriptive statistics and their ranges between real and replicated data for Models V to VIII.

Descriptive	Model V	Model VI	Model VII	Model VIII
Mean	**77.78**	26.90	13.67	14.65
	[42.71,106.69]	[-12.14,63.73]	[-51.57,29.29]	[-47.18,26.03]
Median	**199.78**	**200.82**	**200.42**	**200.16**
	[-200.00,-184.20]	[-223.73,-200.00]	[-217.57,-200.00]	[-213.42,-200.00]
1. Quartile	**97.69**	**88.69**	**47.91**	**37.14**
	[-100.00,-69.88]	[-100.00,-46.97]	[-100.00,-18.97]	[-76.22,-11.17]
3. Quartile	**100.44**	80.77	20.61	13.80
	[98.00,149.82]	[-2.00,98.00]	[-102.00,98.00]	[-102.00,98.00]

Note: significant differences (*ppp*-value < 0.01) are bolded.

The graphical inspection via QQ-plots confirms the findings from PPC, see Figure 5.1. The distributions of the replicated data from each model are compared to the distribution of the real data from NEPS SC6. Distorting effects become obvious by the divergence between the distributions expressed as deviations from the diagonal. Overall, the models estimate the real data fairly well in the lower tail and middle part, whereas the higher income ranges are only reasonably good approximated. Again, Model II has the best fit to the real data. This holds in particular for the lower tail of the distribution up to 7500 EUR. The best fit for values above 7500 EUR was found for Model VIII. Here, differences between the real data and replicates from Model VII become insignificant. The fact that the model with the Dagum distribution as underlying assumed true distribution is more suitable to model real income data corroborates the conjectures made earlier. In general, to approx. the first half of the distribution a univariate distribution seems to be sufficient in modeling true income data. Only with respect to the upper tails of the distribution further factors become relevant for prediction. The superiority of the heaping models with covariates was already alluded by the higher log-marginal likelihoods, see Table 5.2. The corresponding QQ-plot for the ML estimates of Models I-VI is depicted below in Figure 5.2. As can be seen there, the distributions of the replicated data based on ML estimates are not closer to the observed data as compared to those of Figure 5.1, the opposite is true.

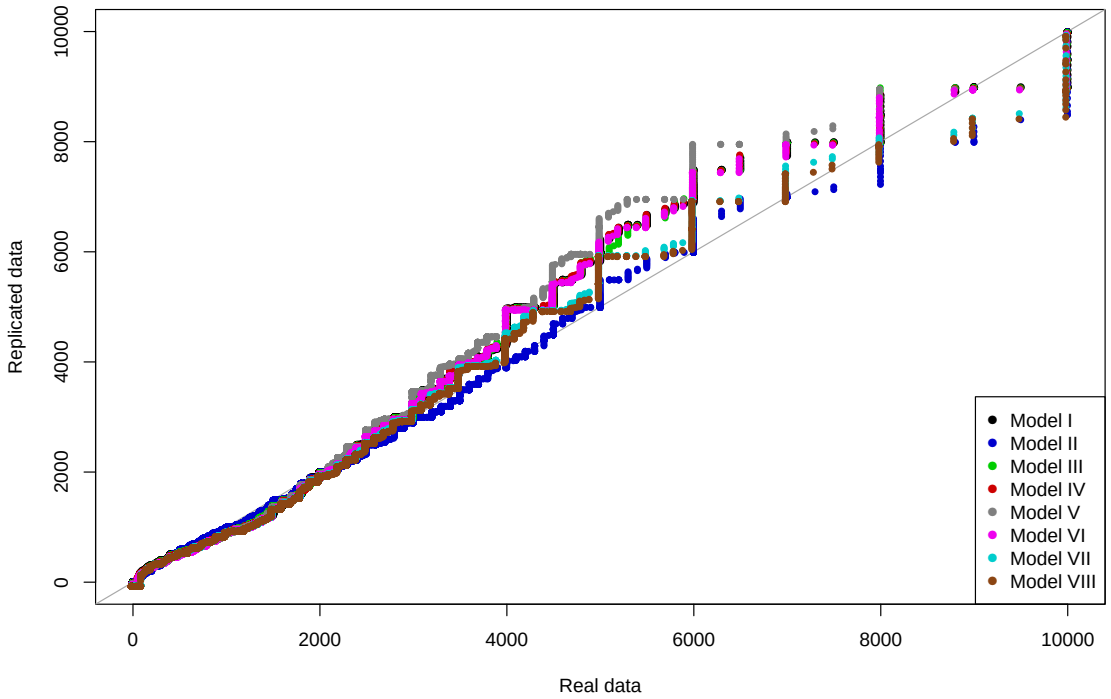


Figure 5.1: Quantile-quantile plot for individual net income data from the Adult Cohort of the NEPS and replicated data from RWM estimates of each model.

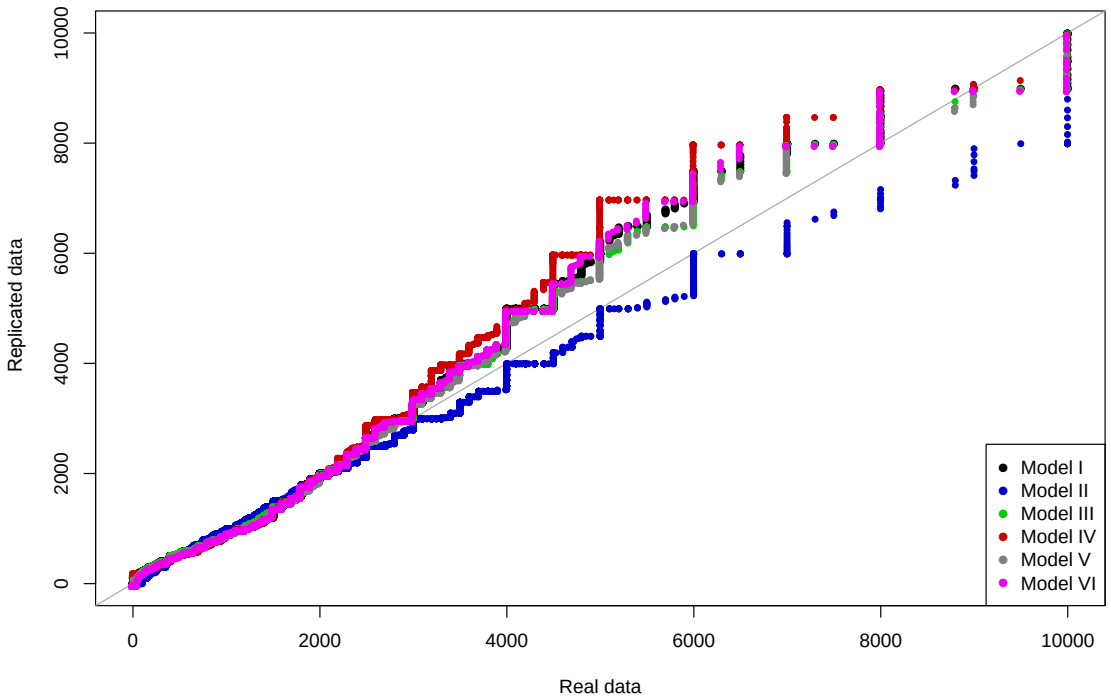


Figure 5.2: Quantile-quantile plot for individual net income data from the Adult Cohort of the NEPS and replicated data from ML estimates of each model.

Drechsler and Kiesl (2014) and Drechsler et al. (2015) compare the outcomes of their models by listing the proportions of values falling on modulos from replicated data against the observed data. Table 5.7 gives the respective summary of all percentages. The total amount of heaped values is underestimated in all models, by approx. 10%. In each model, deviations are rather small for each modulo. The frequency of zeros is significantly overestimated in all models. Model II has the closest estimates for zero values and the parameters of the HM corresponding to mod(1000). The frequency of values falling on mod(100) is best replicated by Model VI and the frequency of values falling on mod(500) by Model III. Overall, the results indicate a good fit of the heaping models to the observed data.

Table 5.7: Percentage of values located at the modulos in the observed and replicated income data from RWM estimates with 95% CI given in parentheses.

Interval	0	mod(100)	mod(500)	mod(1000)	Total
NEPS data	1.69	45.33	9.90	14.03	70.95
Model I	1.77	35.51	7.71	14.27	59.26
	[1.75, 1.79]	[35.41, 35.60]	[7.65, 7.77]	[14.20, 14.34]	
Model II	1.72	37.43	7.91	14.01	61.07
	[1.70, 1.75]	[37.33, 37.53]	[7.86, 7.96]	[13.94, 14.07]	
Model III	1.79	36.08	9.68	14.31	61.87
	[1.77, 1.82]	[35.97, 36.18]	[9.63, 9.74]	[14.25, 14.38]	
Model IV	1.73	32.53	7.40	13.57	55.23
	[1.71, 1.76]	[32.43, 32.62]	[7.35, 7.46]	[13.50, 13.64]	
Model V	1.77	33.80	8.86	14.43	58.86
	[1.74, 1.80]	[33.71, 33.89]	[8.79, 8.92]	[14.37, 14.50]	
Model VI	1.76	38.00	7.67	14.68	62.11
	[1.74, 1.79]	[37.90, 38.09]	[7.62, 7.72]	[14.62, 14.74]	
Model VII	2.06	37.65	7.65	14.79	62.15
	[2.03, 2.09]	[37.56, 37.74]	[7.59, 7.70]	[14.52, 14.65]	
Model VIII	2.07	37.47	7.51	14.80	61.85
	[2.04, 2.10]	[37.39, 37.56]	[7.45, 7.56]	[14.74, 14.86]	

In the replicated data from ML estimates (Table 5.8) only two percentages are closer to the percentages from true data, as compared to the RWM replicates (1.67% zeros in Model V and 41.77% of values falling on hundreds in Model II), but on the whole even larger differences exist. Especially the percentages of values falling on mod(1000) are more distant, as opposed to those of the RWM replicates. Alike in the RWM replicates, the total amount of heaped values is underestimated in all models considered. Furthermore, the larger ML estimates for ρ_{15-19} found in Models I and IV (Table 5.4) yield higher percentages for values falling on thousands compared to the percentages in the NEPS data. The consistently smaller ML

estimates for Model V indicate underestimation as shown by the lower percentages in the ML replicates compared to the real data. Both findings support higher accuracy of the RWM estimates, as opposed to the ML estimates.

Table 5.8: Percentage of values located at the modulos in the observed and replicated income data from ML estimates with 95% CI given in parentheses.

Interval	0	mod(100)	mod(500)	mod(1000)	Total
NEPS data	1.69	45.33	9.90	14.03	70.95
Model I	1.59	36.02	6.92	17.05	61.59
	[1.57, 1.61]	[35.93, 36.12]	[6.88, 6.97]	[16.98, 17.12]	
Model II	1.97	41.77	6.87	12.72	63.33
	[1.95, 2.00]	[41.66, 41.88]	[6.82, 6.91]	[12.65, 12.80]	
Model III	1.66	38.09	9.23	13.06	62.03
	[1.63, 1.68]	[37.99, 38.18]	[9.18, 9.28]	[13.01, 13.12]	
Model IV	1.29	31.24	7.76	17.25	57.54
	[1.27, 1.32]	[31.15, 31.34]	[7.72, 7.81]	[17.17, 17.32]	
Model V	1.67	22.71	5.51	7.90	37.79
	[1.65, 1.70]	[22.62, 22.79]	[5.46, 5.56]	[7.85, 7.96]	
Model VI	1.74	38.96	6.50	16.88	64.08
	[1.72, 1.77]	[38.87, 39.04]	[6.44, 6.55]	[16.81, 16.96]	

The model fit is also evident when inspecting the histogram from observed and replicated data (cp. Würbach, 2015). The histogram in Figure 5.3 shows the frequencies of the replicated data from RWM estimates of Model II and the observed frequencies of individual net income. The frequencies overlap to a great part as indicated by the red bars. Orange and blue parts indicate surpluses, where either more values have been observed, or have been replicated. Data simulated on the basis of RWM estimates of Model II yields a frequency distribution that is more skewed to the right than the empirical one. Here, income values below the mean seem to be overestimated, whereas income values above the mean are underestimated. This assumption is confirmed when comparing with the ranges of differences in Tables 5.5 and 5.6 for the 1. and the 3. quartile from PPC. Negative ranges for estimates of the 1. quartile indicate a systematic underestimation, whereas positive ranges for estimates of the 3. quartile point to overestimation.

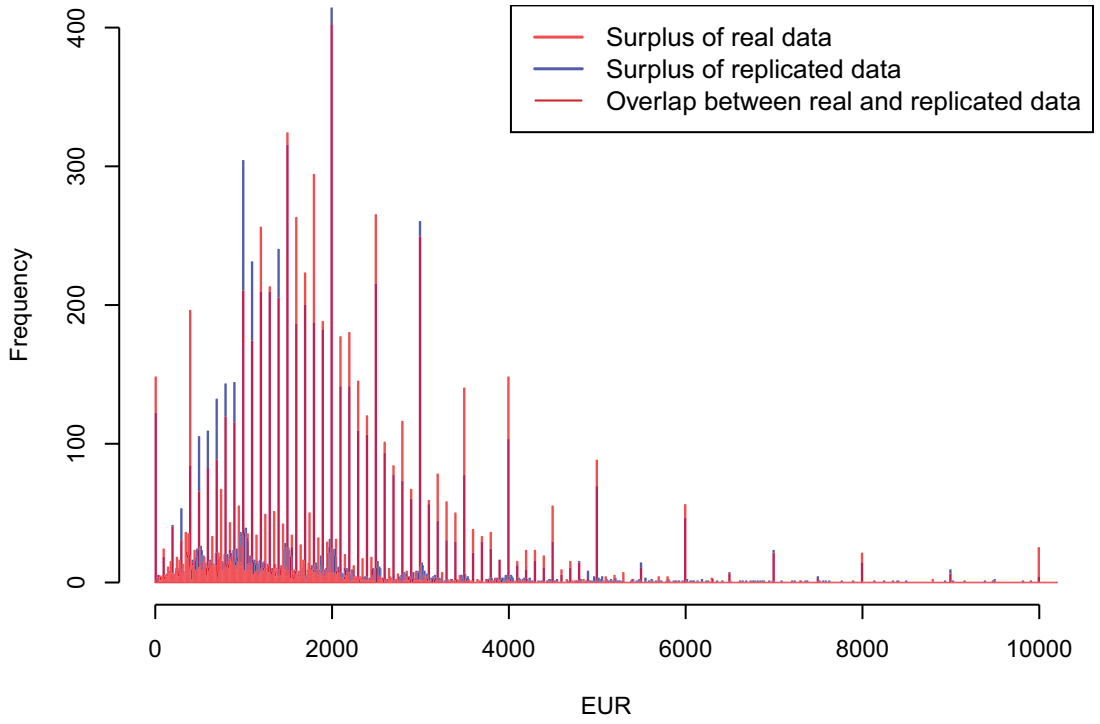


Figure 5.3: Net individual income data from the Adult Cohort of the NEPS (orange) and replicated data from RWM estimates of Model II (blue). Overlapping proportions are colored in red.

Chapter 6

Conclusion

In this thesis, three main issues were to be addressed. First, this work provided further evidence on heaping behavior of respondents in survey studies, exemplarily shown by the income data of the Adult Cohort (SC6) of the German National Educational Panel Study (NEPS) wave 2009/2010. Questions concerning income data are typically known to be refused to be answered which might be attributable to the high uncertainty and sensitivity of the topic. Besides non-response, respondents might also disclose a form of misreporting due to insecurity about the true value or due to hesitation in reporting. For example, they enhance the amount of inaccuracy by sticking to (higher) multiples. As already shown in other studies, the data at hand strongly support the assumption that such a heaping behavior is not stochastic but deterministic, i.e. it is not at random whether and to which degree a true value is heaped. The level of income and common socio-economic characteristics of the respondent were found to have significant effects on the respondents' propensity to heap. Concretely, male and higher educated respondents are more likely to heap their income. The effect due to age is rather small, although older respondents have significantly more heaped responses. With respect to the heaping intensity, e.g. measured by the relative Rounding Indicator (RI), a monotone almost linear pattern was found for the marginal effects of all three characteristics considered. The magnitude of the true response value largely determines the intensity of heaping, a pattern introduced as level dependency by Torelli and Trivellato (1993). The results presented in this work are in concordance with findings from the existing literature and clearly show that heaping is not random. Thus, standard assumptions on misreporting errors and inferences are violated. This motivates the construction of a model for heaped data which considers the true response value and the heaping behavior at once.

The second issue considered in this thesis therefore was to establish a flexible mixture model which allows to model different heaping behaviors prevalent in self-reported income data. The proposed method assumes a parametric model for the latent true distribution of the variable of interest and a parametric model for the heaping behavior. As a starting point, the latent distribution was assumed to follow a log-normal distribution enriched by an inflation parameter to account for observed zero values. Heaping behavior was modeled on the basis of heaping prob-

abilities. At start, these probabilities are assumed to be piecewise constant (*pcm*). This modeling technique regards the probability for heaping to be equiprobably distributed within predefined intervals for each modulo taken into consideration (here: 100, 500, 1000). The generality and flexibility of the proposed approach was outlined by five modifications and two extensions. The modifications relate either to the assumed latent distribution, the heaping pattern or the heaping mechanism. To vary the latent model, the Dagum distribution was considered because of its wide popularity for modeling income data. The heaping pattern was relaxed in two respects: first, by allowing wider heaping intervals widths to reflect higher uncertainty or hesitation in reporting; second, by assuming asymmetric heaping intervals instead of symmetric ones shifted to the right to model underreporting. The heaping mechanism can be specified alternatively by a bell-shaped probability function (*pbsm*) on the one hand and a model for the heaping mechanism with a reduced number of components on the other hand. The *pbsm* considers values in the proximity of a heaping point to be more likely to fall on this. In the model with less parameters of the heaping mechanism a distinct number of equality constraints concerning the heaping probabilities is determined for the *pcm*. That is, equiprobable heaping behavior across larger parts of the income range is assumed leading to a smaller number of parameters to estimate. The above-mentioned heaping models are all without any covariates. However, this is truly short-sighted. Relationships have been found between income level, heaping behavior, and observable personal characteristics. Therefore, implementation of an extended heaping model is required which introduces these personal characteristics as covariates. The covariates are specified to affect both, the income value and additionally the heaping mechanism allowing for individually different heaping behaviors. All model specifications have been applied to analyze the heaped income data of the NEPS SC6. Measures of relative quality and several posterior predictive checks (PPC) were used to assess how well the models fit to the observed data. Among the univariate models without covariates, this one assuming the Dagum distribution outperformed the remaining models. Though, the models with covariates show a preferable fit in particular for the upper tail of the distribution (above 7500 EUR). According to this, it seems advisable to specify distinct latent models across the income range to adequately and flexibly resemble the empirical distribution.

The third research question referred to the comparative analysis of different estimation procedures. It was ascertained that a Bayesian estimation technique could be more valuable for the estimation purpose at hand, because ML estimation techniques in general have problems with multi-modality of the likelihood function. $MCMC$ techniques are often used in the scope of integration or optimization problems, when being confronted with large dimensional spaces and multiple local maxima. Since no established distribution was found for the joint conditional distribution of all model parameters considered, the random-walk Metropolis (RWM) algorithm from the family of Metropolis-Hastings algorithms (MH) was employed.

MH algorithms are constructed in such a way that the chain spends more time in the most important regions of the parameter space. Due to that fact the draws mimic samples from the unknown target distribution. No matter what initial distribution, the chain will stabilise at the target distribution, which means the chain will converge in the majority of cases. The definition of the proposal density is crucial for generating draws. These draws can determine success or failure of the algorithm. A too small or too large dispersion of the proposal causes a loss of the ergodicity property, see Andrieu and Thoms (2008, p. 355). The aim is to find a specification of the RWM algorithm making it well-mixing, i.e. ensuring all modes are visited while the acceptance rate is still high. Desired properties of the algorithm are therefore *approximating ability* (the proposal density fits the target distribution well) and *exploring ability* (the proposal density searches mainly in the high density regions), cp. Giordani and Kohn (2010, p. 11). Furthermore, computational requirements ought not be forgotten when comparing algorithms and specifications along with their efficiency. Preferable exploration of the parameter space and accelerated convergence while having extreme computational costs is clearly suboptimal (Turek et al., 2015, p. 3f.). Objective functions for evaluation and comparison are: the convergence rate, the precision of the estimates, algorithmic efficiency, and computational efficiency.

Different RWM algorithms are inquired into. Besides the original RWM scheme, which samples all model parameters at once with a fixed proposal distribution, blocking strategies for sampling of the proposal density and adaptive schemes with regularly updates of the proposal covariance matrix, are taken into consideration. In cases when the parameter values are highly correlated or when having a hierarchical structure it is strongly recommended to split the multivariate state vector into non-overlapping blocks updated sequentially (cp. Chib & Greenberg, 1995). Other literature (cp. Haario et al., 1999, 2001) proposes to automate the process of choosing a suitable proposal distribution by using the information from past samples for consecutive draws to obtain a distribution which is closer to the target distribution. In this thesis, both techniques, blocking and updating of the proposal density, are applied and compared to each other.

The empirical evidences presented here suggest that blocking can greatly improve mixing and convergence of the RWM algorithm. All blocking algorithms are very stable for the given specifications and show very good and fast convergence, in contrast to the adaptive schemes considered. The adaptive schemes work provably well in a downsized setting of the established model. It was found that any gain in algorithmic efficiency is moderated by the higher computational effort required for blocking. This means that there is a trade-off between the size of the blocks and the structure of the model parameters. In other words, considered blocks should be as large as possible but capturing the inherent structure at the same time. The overall efficiency measure proposed by Herbst and Schorfheide (2015) points to superiority of the M-RWM, the multiple-block scheme in which the parameters of the heaping mechanism are separated with respect to the modulos considered.

Besides high algorithmic efficiency, this algorithm is also computational more efficient relative to the other blocking schemes. To overcome the necessity of prior knowledge on the correlation structure of the model parameters and appropriate block sizes, the randomized blocking strategy was applied for estimation of the modifications, extensions, and in the application to NEPS income data.

Estimation accuracy was slightly higher for RMB-RWM estimation compared to the ML approach. Based on simulated data examples of all model specifications without covariates, estimated parameter values from RMB-RWM were closer to the true parameter values from the DGP compared to the ML estimates in four of six models (Models I-IV, the averaged squared biases of Model I are 0.00095 for ML, and 0.00021 for RMB-RWM estimation). In particular the PPC in the application to NEPS income data indicated superiority of RWM estimation, because replicated data from ML estimates was not able to mirror point estimates of the real data or the empirical distribution on the whole in certain cases. The performance of RMB-RWM estimation for all eight models considered is fairly good, however, large differences exist with respect to runtime and efficiency. This fact is largely attributable to the assumed latent distribution and the related evaluations of the likelihood. Estimation of the presented modified or extended heaping models by RMB-RWM is widely robust against misspecifications. Though not presented here for all models, varying priors or starting values yield comparable results (cp. Section 3.2.3 for a brief sensitivity analysis of the basic model).

6.1 Limitations

The findings of this study are limited by several points. Starting with limitations of the heaping model, some remarks regarding the latent model and the model for the heaping behavior are outlined. With respect to the latent model, it might be argued that heaping points could result from truly reported values and the model should be adapted accordingly. However, when assuming that the occurrences of such values are distributed equiprobable across all heaping points this should not constrain the generality of the results, cp. Serfling (2006, p. 122). Another potential criticism of the introduced approach is the parametric model for the latent distribution. A non-parametric kernel density estimation approach, as proposed by Groß and Rendtel (2015), could be used instead. Such an attempt was not pursued any further here and is left for further research. Objections relating the heaping mechanism could concern the restricted number of multiples and covariates regarded. Considering only the modulos 100, 500, and 1000 in the model for the heaping mechanism could be expected to hamper the validity of the heaping model. The results suggest that the proposed models fit at least as well as models additionally considering the modulos 5, 10, and 50 (e.g. Drechsler & Kiesl, 2014). More covariates than the introduced ones are conceivable to determine the heaping mechanism. It is to be shown how estimation behaves and to which extent it might be possible to include even more parameters until estimation blurs out.

Finally, some remarks regarding the estimation via RWM algorithm have to be given before describing further approaches in the next section. The total chain lengths are set to 11,000 with T fixed at 10,000 following a burn-in of $n_0 = 1000$ iterations. Other studies use by far more iterations, e.g. Yeung and Wilkinson (2001) and Levine et al. (2005) use $T = 60{,}000$ and $n_0 = 10{,}000$. Andrieu and Thoms (2008) use as much as 100,000 or 200,000 iterations. In the sensitivity analysis conducted and for the blocking schemes it was shown that convergence was achieved at early iterations for most of the model parameters thus making larger $MCMC$ sample sizes unnecessary. On a trial basis, the adaptive schemes are run with an increased $MCMC$ sample sizes of $T = 100{,}000$ following a burn-in period of $n_0 = 5000$. In sum, the behavior of the samplers did not show a better mixing behavior over the longer run. Another finding is that the advantage of higher estimation accuracy of the MB-RWM algorithms comes at the price of computational efforts. It is widely known that the computational problems encountered can be prohibitive for very complex models, see e.g. Wilkinson and Yeung (2002) and Turek et al. (2015). In line with this, $MCMC$ runtimes often constitute a limiting factor for the range of models considered and the diagnostics used for comparison. The work presented here is limited to the blocking strategies proposed by Chib and Greenberg (1995) and Chib and Ramamurthy (2010) and the adaptive schemes introduced by Haario et al. (1999, 2001). Several other possibilities exist for efficient MH sampling listed in the following section.

6.2 Further research

Notations on possible avenues for future research are also given regarding the heaping model and the estimation procedure. As mentioned earlier, further covariates could be included into the model. Besides internal factors (observed personal characteristics), also external factors relating to the interview context might be of interest. The influence of interview quality indicators on reporting accuracy is demonstrated by Battistin et al. (2003, p. 370). This concerns interview duration and interviewers' assessment on how well the respondent understood the question, for example. Also Hanisch (2005a, 2006) explores how interview-related factors influence response behavior. The author highlights the importance of the interview mode on the propensity to heap. Personal interviews (CAPI – computer-assisted personal interview) have an overall lower proportion of heaped values than interviews from proxy or telephone interviews (CATI – computer-assisted telephone interview). Hanisch argues that respondents are likely to be more reluctant to disclose income figures to a stranger on the telephone, as opposed to an interviewer the respondent is faced with and probably can identify with. To be concrete, the spatial distance between interviewer and interviewee is assumed to reduce the commitment of the respondent. Again, interview duration was found to be associated with response accuracy. Longer durations indicate a more careful evaluation of the questions and eligible answers, whereas shorter interviews are shown to

suffer from satisficing, i.e. shortcuts in reporting accuracy. These patterns are also supported by Schräpler (1999). According to this reasoning, external factors are additionally taken into account for explanation of the heaping behavior in a further model, see Figure 6.1. The context specific attributes assumed to influence the tendency and intensity to heap are the mode and duration of the interview as well as the incentive height. The amount of the incentive is supposed to increase the extrinsic motivation of individuals to be more inclined towards giving correct responses and hence to reduce satisficing.

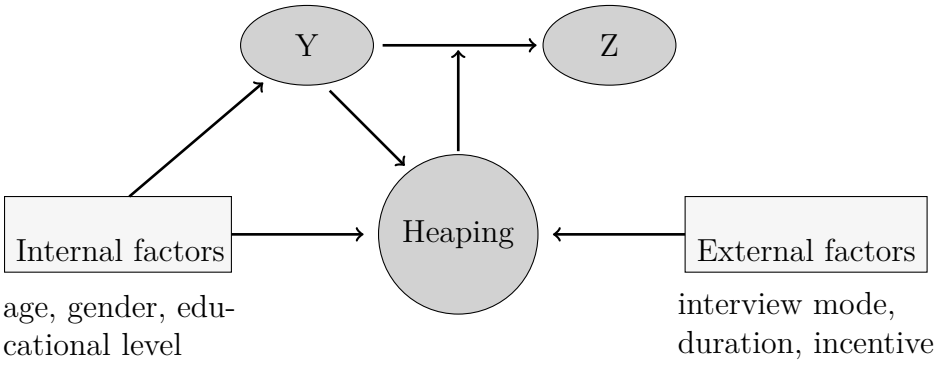


Figure 6.1: Relationships between heaping, internal and external factors

Multivariate setting with personal characteristics and interview effects

To study a multivariate setting with added context information (Figure 6.1), at first, marginal frequencies according to the external factors are explored. Most of the respondents (about 85.1%) are interviewed by CATI and only 14.9% by CAPI (Table 6.1). This finding is not surprising since CATI by design was the preferred interview mode. The average interview duration[1] is 44.93 minutes with a minimum of 10.76 minutes and a maximum of 210.90 minutes. An incentive of 10 EUR was held in prospect for more than 80% of the respondents. The other 20% received 50 EUR after interview completion.

Second, a further classification tree[2] is grown for joint consideration of all internal and external factors as explanatory variables for the occurrence of heaping, see Figure 6.2. The difference compared to the tree with only internal factors in Figure 1.8 on page 46 is in the fourth split. Instead of the separation of low and middle educated female respondents, groups are formed according to interview duration (cut-off point 39 minutes). The terminal node of those with less than

[1]Interview durations from first-time respondents are halved to account for the large discrepancies to the panel respondents. Data on retrospective life-course information has to be reported completely in the first interview. Because of that fact the interviews of first-time respondents are on average twice as long as those from panel respondents (88.8 vs. 44.3 minutes).

[2]Splits are supposed to stop when the minimum number of observations in any terminal node would fall below 100. The complexity parameter is now fixed at 0.0035.

39 minutes of interview duration have a proportion of 0.59 of heaped values. The group of respondents with interview durations longer equal 39 minutes are further split into lower and middle educated females. Female lower educated respondents with interviews of 39 minutes or longer form the terminal node with a higher proportion of non-heaped values (0.56 vs. 0.44). The relative variable importance of gender is still very high with 82%, 14% are attributable to the educational level and 4% to interview duration.

Table 6.1: Combined percentages for observing heaping, divided by subgroups according to selected context factors.

Group	Heaping	No heaping
All	69.16	30.84
CAPI	69.51	30.49
CATI	69.10	30.90
Interview duration shorter than 39 minutes	73.29	26.71
Interview duration 39 up to 49 minutes	68.77	31.23
Interview duration longer equal 49 minutes	64.91	35.09
Low incentive	68.72	31.28
High incentive	70.96	29.04

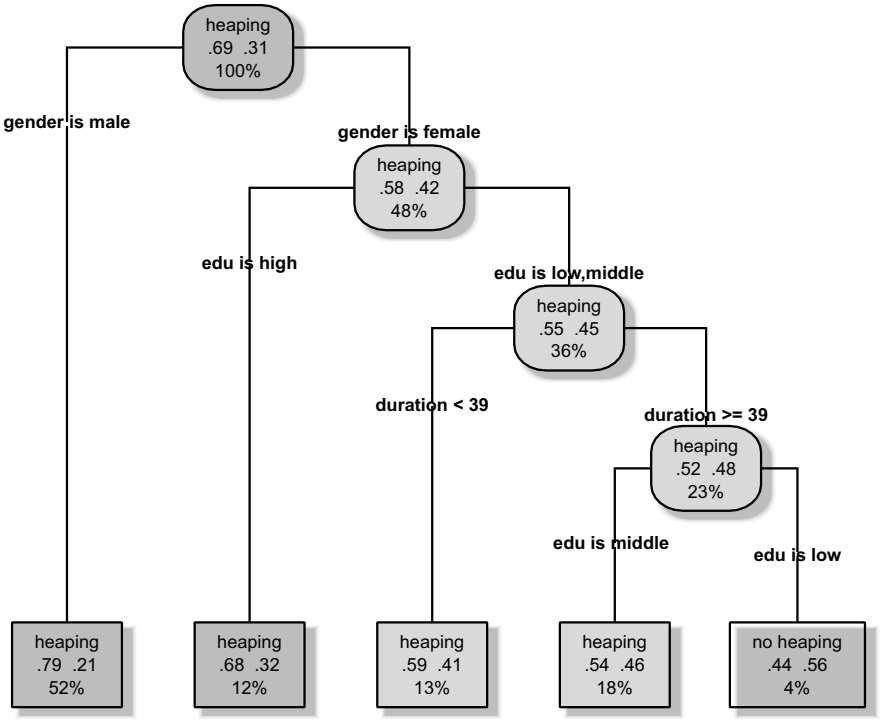


Figure 6.2: Classification tree for observing heaping with external factors.

Third, probit regressions are estimated to identify and quantify the joint effects of the individual characteristics and the context related covariates on the binary outcome for heaping (value located at a modulo or not), see Table 6.2.[3] The quantities fmi and $lambda$ are both below 0.2 indicating a modest influence of the imputation model on the final result. The estimates of the individual characteristics do not change remarkably and remain as significant as before (cp. Table 1.6 on page 43). Neither the interview mode nor the height of the incentive have significant effects on the propensity to heap. Only interview duration exhibits a significant relationship with the tendency to heap. With increasing duration decline the percentages for observing heaped values. This finding strengthens the assumption that interview duration is an indicator for the deepness of evaluation, and that respondents with longer interview durations less often resort to shortcuts in answering. Nagelkerke's pseudo-R^2 in the augmented model is 0.096 and the AIC decreases slightly to 9767.4 (compared to 9818.3, see Chaurasia & Harel, 2012, p. 5).

Table 6.2: Results from combined probit regression for the tendency to heap with external factors.

Predictor	Estimate	SE	CI	df	t-ratio	p-value	fmi	$lambda$
(Intercept)	0.039	0.109	[-0.174,0.252]	3639.1	0.36	0.717	0.0377	0.0371
Male	0.609	0.032	[0.547,0.671]	644.6	19.24	<0.001	0.1158	0.1130
Age	0.011	0.002	[0.008,0.015]	761.6	7.03	<0.001	0.1055	0.1032
Middle edu	0.090	0.041	[0.009,0.170]	672.2	2.19	0.029	0.1131	0.1105
High edu	0.412	0.045	[0.323,0.500]	3779.6	9.07	<0.001	0.0364	0.0359
CATI	0.049	0.045	[-0.040,0.138]	746.2	1.08	0.279	0.1067	0.1043
Duration	-0.008	0.001	[-0.011,-0.006]	1392.5	-6.78	<0.001	0.0743	0.0730
High incentive	-0.086	0.040	[-0.164,-0.008]	1265.8	-2.16	0.031	0.0787	0.0773

Notes: Nagelkerke's pseudo-$R^2 = 0.096$ for each single probit regression, $AIC = 9767.4$.

Augmented ordered probit regressions computed for the relative RI and the RSM give similar results, although differing in sign.[4] The corresponding marginal effects are given in Figure 6.3 and Figure 6.4, for each model separately. Remember, a decreased relative RI and an increased RSM indicate higher heaping intensity. Again, the categories 0.40 in the relative RI model and 5 in the RSM model are conspicuous. The latter can be left out for interpretation owing to the low number of observations in this category. However, the monotone trends in the marginal effects are still obvious for all predictors in both models. The trends in the relative RI are, once again, almost linear. The picture in Tables A.3 and A.4 remains the same as in the models without external factors, cp. Tables A.1 and A.2. To be concrete, all personal characteristics except of having a middle educational level are still highly significant (at a 1%-level). CATI as interview mode

[3]The excess in zeros is omitted before estimation.

[4]The combined estimates of both models are given in Tables A.3 and A.4 in the Appendix.

is insignificant in both models, whereas interview duration is. A high incentive being held in prospect is significant only in the extended ordered probit model for the relative RI (at a 5%-level). Model improvement is highly significant (p-value < 0.001) when including the external factors. This holds for both measures, the relative RI as well as the RSM.

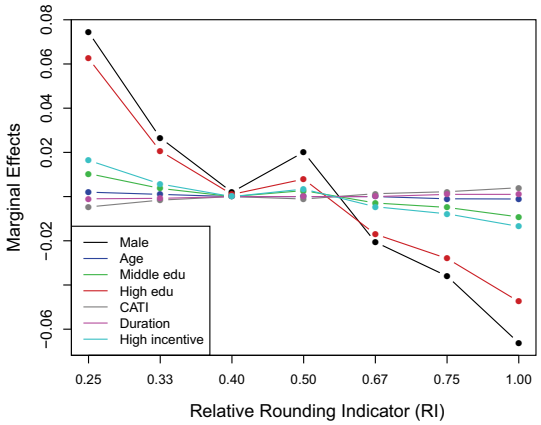


Figure 6.3: Marginal effects from ordered probit regression for the relative RI with additional external factors.

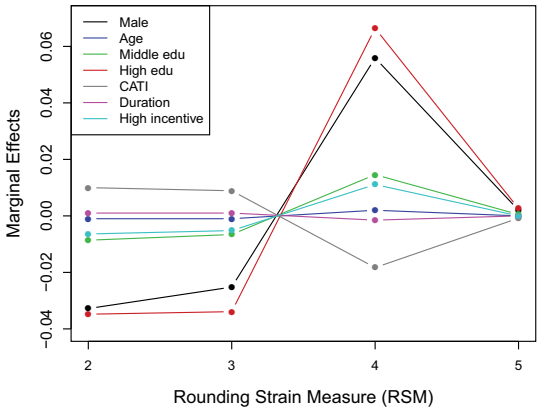


Figure 6.4: Marginal effects from ordered probit regression for the RSM with additional external factors.

The results underpin the suggestions to include these external factors into the model. Of course, even more covariates could be considered. In the scope of internal factors, the occupational status of a respondent can be of interest. For example, Miller and Paley (1958, p. 199) find that occupation groups with stable incomes (wage or salary workers) show the lowest variation in response. The regularity of income – as opposed to highly variable incomes of side jobs – strongly determines the uncertainty with respect to the true income value, see Hanisch (2005a, p. 41). Regarding further external factors, the experience of the interviewer could be included. Serfling (2006, p. 115) finds a small quadratic interviewer experience effect. Social proximity could be another factor. Serfling (2006, p. 91) argues that social distance between respondent and interviewer negatively influences the cost-benefit calculation in the judgement and response step. Social proximity builds the foundation for a more trustful interview situation in which the respondent feels more comfortable. However, Serfling (2006, p. 116) rejected this hypothesis based on his empirical findings. Question comprehension (Battistin et al., 2003, p. 370) is also an interesting issue to be studied. In the NEPS, interviewer information concerning the reliability of responses is available. While it seems desirable to include as many variables as possible into the model, increasing the number of covariates might impose identification problems or raises the possibility of numerical instability, and is likely to increase the runtime of RMB-RWM estimation.

Further crucial aspects for extension of the proposed model are the implementation of bracketed information and panel conditioning. Bracketed income queries constitute an active strategy to minimize item non-response. In the NEPS, after an initial question regarding the exact value of the actual income a follow-up question is given to the respondents who refused to answer the first one. In the following prompt, the respondent is asked to point to the correct interval into which the true income falls. The resulting data imply coarsening of the true income distribution and a loss of important information (Howes, 1996). This complicates the computation of some standard statistical problems (Cowell, 2000, p. 142). Fortunately, the level of discretization is specified by the question designer (exogenous). This avoids the statistical complexities entailed by endogenous heaping with unknown intervals, which is a strong argument in favour of bracketed questions, see Pudney (2008, p. 27f.).

In this thesis, only effects of heaping on cross-sectional reports have been regarded. However, panel conditioning might be of importance and, if available, longitudinal information might be included. Respondents might become more and more acquainted with the survey and successive interview contacts might increase trust in the interviewer thus encourage to report more carefully, see Schräpler (2002, p. 7). Also at side of the interviewer a learning and experience gain might be expected (id.). In an analysis of data from the German Socio-Economic Panel (G-SOEP), Rendtel et al. (2004, pp. 56ff.) do not find evidence for the hypothesis that panel conditioning has a positive effect on the precision of responses to income question, see also Schräpler (1999), and regarding the SHP, see Serfling (2006). Hanisch (2005a, p. 43) also finds no evidence for a panel conditioning effect with regard to the Fin-ECHP data (1996-2000). Hanisch (2003, p. 16) compares transitions in consecutive waves with respect to the response type and finds stability of response behavior over subsequent waves, see also Hanisch and Rendtel (2002).

Regarding the model of the latent distribution, non-parametric techniques can be evaluated with respect to their applicability. These can be alternatively kernel density estimation approaches (cp. Groß & Rendtel, 2015) or models based on p-splines and wavelets. Usually, non-parametric methods are known to weaken in settings with long-tailed distributions. In the higher ranges, more and more values do not have any adjacent neighbor which is not a heaping point itself making them inapplicable for estimation. Three-parameter models have an increased flexibility and allow for intersecting Lorenz curves which might have caused the superiority of the Dagum distribution over the log-normal distribution. Going one step further, a 4-parameter model could be considered to approximate the empirical distribution more closely, e.g. the 4-parametric double-Pareto-log-normal distribution (cp. Section 2.1) or combinations of distinct densities. An option would also be to refer to latent variable modeling (LVM), see Frühwirth-Schnatter (2006), e.g. the latent class model used by Vermunt, van Ginkel, Joost R., van der Ark, L. Andries, and Sijtsma (2008). In this respect, mixed membership models would be preferable over finite mixture models because of the possibly overlapping

intervals, see Airoldi, Blei, Erosheva, and Fienberg (2015).

With regard to the estimation procedure, the author of this thesis is currently faced with stability and computing speed issues. Further steps are taken into the direction of exploration of some of the other blocking or adaption algorithms available. Albeit the literature on blocking and adaptive MH algorithms is sparse, several innovative approaches exist that could be tested for their suitability. Early attempts are made by Besag, Green, Higdon, and Mengersen (1995) and G. O. Roberts and Sahu (1997), mostly attributed to the Gibbs sampler. G. O. Roberts and Sahu (1997) propose random and non-random updating strategies for Gaussian models. Here, updating means the visiting sequence within each iteration. The authors show that deterministic updating schemes are preferable – with respect to convergence – for hierarchically structured problems or densities with positive partial correlations, see G. O. Roberts and Sahu (1997, p. 293) and Besag et al. (cp. also 1995). Similar effects of dimensionality and correlation structure on convergence hold for the Gibbs sampler as well. High dimensionality, involving a large number of model parameters or latent variables, leads to inefficiency of the Gibbs sampler, see Giordani and Kohn (2010). However, this interplay between correlation and dimension is highly complicated (G. O. Roberts & Sahu, 1997, p. 308). Examples of deterministic updating schemes are the deterministic sweep strategy, i.e. updating the components in their natural order (G. O. Roberts & Sahu, 1997, p. 295), a reversible version, i.e. forward and backward updating alternates, or a checker-board type, i.e. updating first all odd-numbered components then all even-numbered components (G. O. Roberts & Sahu, 1997, p. 300). A stochastic updating scheme is the random sweep strategy originally introduced by Geman and Geman (1984) and can be performed with or without replacement. That means, the visited components are distinct or not thus allowing blocks to overlap, see G. O. Roberts and Sahu (1997). Mathew et al. (2012) propose a mixed technique that uses a hybrid Gibbs sampler – a combination of scalar and blocked updates – in the first stage to learn about the covariance structure. This knowledge is used in the second stage for formulation of an efficient proposal density for the MH algorithm.

Alternative blocking strategies arise from the work of Giordani and Kohn (2010) and Turek et al. (2015). Giordani and Kohn (2010) propose a *k-harmonic means* clustering of the model parameters and Turek et al. (2015) suggest a tree-based clustering of the components for automated blocking. Both methods attempt to overcome the disadvantage of lacking prior knowledge with respect to the correlation structure of the components and do not require explicit computation of the covariance matrix. In the self-tuning procedure of Turek et al. (2015), model parameters are iteratively clustered based upon the empirical posterior correlations which are transformed into distances and used for a *hierarchical clustering tree*. Other blocking techniques can be found in Andrieu and Thoms (2008). The authors quote block sampling algorithms by principal directions and principal component analysis.

Promising adaptive MH algorithms are provided, e.g. by Yeung and Wilkinson (2001), Levine et al. (2005), G. O. Roberts and Rosenthal (2009), and M. J. Baker (2014). Andrieu and Thoms (2008) provide an overview of the theory and practice of selected adaptive $MCMC$ algorithms. Some of the procedures perform adaption only at a preliminary period, and afterwards the algorithms run with the current proposal. For example, Browne and Draper (2000) adapt the proposal until a previously defined acceptance rate is achieved. Then, adaption stops and the chain continues with the fixed proposal. On the contrary, other approaches explicitly allow the proposal to change (all the time) during sampling. The work of Gilks et al. (1998) is similar to this of Browne and Draper (2000), but the authors allow for adaption at regeneration times (cp. also Brockwell & Kadane, 2005). Gelfand and Sahu (1994) monitor the eigenvalues of transition kernels as stochastic matrices for finding better proposals. Tierney and Mira (1999) also allow for adaption by the delayed rejection technique (cp. Andrieu, Freitas, Doucet, & Jordan, 2003). In Yeung and Wilkinson (2001, pp. 3ff.), adaption proceeds by means of a tuning parameter specified alternatively by a stochastic search algorithm, a quadratic response surface algorithm, or a hybrid of both. For linear Gaussian Directed Acyclic Graph models, an adaptive MH scheme can be more efficient than a Gibbs sampler according to Yeung and Wilkinson (2001), in particular in highly correlated settings or hierarchical models. In Haario, Saksman, and Tamminen (2005) a single-component variant of the AM algorithm is proposed: $SCAM$. $SCAM$ works well in high dimensional settings but requires the unknown target distribution to be uni-modal. In the absence of correlation among the components no further adjustments are necessary. Otherwise, to increase efficiency, the proposal distribution is rotated so that the sampling directions of the algorithm are in accordance with the directions of the principal vectors. Based on the covariance matrix of the chain retrieved so far the principal vectors are computed and used as sampling directions. The direction of the proposal distribution is fixed for the further proceeding and the sampling proceeds by only updating the covariance matrix of the proposal distribution. However, the $SCAM$ algorithm raises the runtime by factor $K = D$, with D being the number of parameters in the model. Since the resulting efficiency gain does not charge with the immense increased computational burden, the $SCAM$ algorithm is not recommendable in general. In Andrieu and Thoms (2008), the technique proposed by Haario et al. (1999, 2001) is extended in several ways: to a Rao-Blackwellized AM algorithm, an AM algorithm with global adaptive scaling (adaption of the scaling parameter to fit to the acceptance rate globally), a componentwise AM (originally from Haario et al. (2005)) with componentwise adaptive scaling, and a global AM with componentwise adaptive scaling. G. O. Roberts and Rosenthal (2009) adapt the AM and the $SCAM$ algorithm and propose an algorithm adapting the proposal distribution locally, called the regional adaptive Metropolis ($RAMA$) algorithm. Vihola (2012) proposes a robust adaptive Metropolis (RAM) algorithm that attains a given acceptance probability during estimation of the shape of the target distribu-

tion. The RAM avoids the usage of the empirical covariance matrix from previous iterations by relying on a single matrix update formula (id., p. 998). Levine et al. (2005) suggest an adaptive MH that generates a random variate to choose the optimal sweep strategy and another variate to find the optimal proposal distribution on the fly. Referring to the random proposal distribution of Besag et al. (1995) assuming component specific proposal distributions, the selection probabilities in the componentwise MH sampler is assumed to follow a functional form also being updated at each iteration (Levine et al., 2005). According to Besag et al. (1995), the random proposal distributions are adaptively updated but not restricted to Gaussian proposal distributions as in Haario et al. (1999, 2001). M. J. Baker (2014) augment the proposal distribution with a tuning parameter, which is also updated at the preliminary state. This adaptive $MCMC$ algorithm with normal proposal and vanishing adaption tunes towards a predefined acceptance rate.

Further projects could include implementation of the proposed heaping model into multivariate imputation by chained equations (MICE). Currently, popular imputation techniques as hot decking or non-parametric imputation duplicate the heaping behavior at approximately the same frequency as to which the heaped observations occur (Schweitzer & Severance-Lossin, 1996, p. 4). Zinn (2014) shows how such an implementation could look like when adjusting for non-response and misreporting patterns simultaneously (cp. van der Laan & Kuijvenhoven, 2011).

In this thesis, the main focus was on heaping in income data. Applications of the proposed model to duration data or discrete numerical data are conceivable, e.g. by using a Weibull distribution or a discrete hazard model as latent distribution for duration data, and a Poisson distribution for count data, respectively. The heaping pattern can be adapted accordingly, e.g. by referring to calendar prototypes 7, 10, 14, 21, 30, or 60 as heaping points.

Bibliography

Abel, E. L., & Kruger, M. L. (2006-2007). Heaping in anniversary reaction studies: a cautionary note. *Omega*, *54*(1), 59–65. doi: 10.2190/V752-6773-1KMW-3334

Airoldi, E. M., Blei, D., Erosheva, E. A., & Fienberg, S. E. (Eds.). (2015). *Handbook of Mixed Membership Models and Their Applications*. Boca Raton, FL: Chapman & Hall/CRC Press.

Aitchison, J., & Brown, J. A. C. (1957). *The Lognormal Distribution with special reference to its uses in economics* (Vol. 5). Cambridge, UK: Cambridge University Press.

Allie, É. (2002). Inversion procedures for systematically randomly rounded income data. *Proceedings of the Survey Methods Section, SSC Annual Meeting*, 93–97. Retrieved 10.11.2015, from `http://www.ssc.ca/survey/documents/SSC2002_E_Allie.pdf`

Andrieu, C., Freitas, N. d., Doucet, A., & Jordan, M. I. (2003). An Introduction to MCMC for Machine Learning. *Machine Learning*, *50*(1), 5–43. doi: 10.1023/A:1020281327116

Andrieu, C., & Thoms, J. (2008). A tutorial on adaptive MCMC. *Statistics and Computing*, *18*(4), 343–373. doi: 10.1007/s11222-008-9110-y

Antoni, M., Vicari, B., & Bela, D. (2015). *Interviewers' influence on bias in reporting income, Presentation at the 4th Baltic-Nordic Conference on Survey Statistics (BaNoCoSS).* Helsinki, Finland.

Aßmann, C., Würbach, A., Goßmann, S., Geissler, F., & Biedermann, A. (2014). *A nonparametric multiple imputation approach for multilevel filtered questionnaires* (No. 36). Bamberg, Germany. Retrieved 21.08.2015, from `https://www.neps-data.de/Portals/0/Working%20Papers/WP_XXXVI.pdf`

Aßmann, C., Würbach, A., Goßmann, S., Geissler, F., & Biedermann, A. (2015). Nonparametric multiple imputation for questionnaires with individual skip patterns and constraints: The case of income imputation in the National Educational Panel Study. *Sociological Methods and Research*. doi: 10.1177/0049124115610346

Augustin, T., & Wolff, J. (2004). A bias analysis of Weibull models under heaped data. *Statistical Papers*, *45*, 211–229. doi: 10.1007/BF02777224

Azzalini, A., & Genz, A. (2014). *The R package mnormt: The multivariate normal and t distributions (version 1.5-1).* Retrieved 13.09.2015, from

http://azzalini.stat.unipd.it/SW/Pkg-mnormt

Bai, Y., Roberts, G. O., & Rosenthal, J. S. (2011). On the Containment Condition for Adaptive Markov Chain Monte Carlo Algorithms. *Advances and Applications in Statistics*, *21*(1), 1–54. Retrieved 10.02.2015, from http://www.pphmj.com/abstract/5793.htm

Baker, M. (1992). Digit preference in CPS unemployment data. *Economics Letters*, *39*(1), 117–121. doi: 10.1016/0165-1765(92)90112-C

Baker, M. J. (2014). Adaptive Markov chain Monte Carlo sampling and estimation in Mata. *The Stata Journal*, *14*(3), 623–661. Retrieved 22.12.2015, from http://www.stata-journal.com/article.html?article=st0354

Bandourian, R., McDonald, J. B., & Turley, R. S. (2002). *A comparison of parametric models of income distribution across countries and over time.* Retrieved 23.08.2012, from http://www.lisproject.org/publications/liswps/305.pdf

Bar, H. Y., & Lillard, D. R. (2012). Accounting for heaping in retrospectively reported event data – a mixture-model approach. *Statistics in Medicine*, *31*(27), 3347–3365. doi: 10.1002/sim.5419

Barnard, J., & Rubin, D. B. (1999). Small-Sample Degrees of Freedom with Multiple Imputation. *Biometrika*, *86*(4), 948–955. Retrieved 14.11.2013, from http://www.jstor.org/stable/2673599

Barreca, A. I., Lindo, J. M., & Waddell, G. R. (2001). *Heaping-induced bias in regression-discontinuity designs* (No. 17408). Retrieved 12.04.2012, from http://www.nber.org/papers/w17408

Bates, D., Maechler, M., Bolker, B. M., & Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-7.* Retrieved 13.09.2015, from http://arxiv.org/abs/1406.5823

Battistin, E., Miniaci, R., & Weber, G. (2003). What do we learn from recall consumption data? *Journal of Human Resources*, *38*(2), 354–385. doi: 10.2307/1558748

Beaman, J., & Grenier, M. (1998). Statistical tests and measures for the presence and influence of digit preference. In H. G. Vogelsong (Ed.), *Proceedings of the 1997 Northeastern Recreation Research Symposium* (pp. 44–50). Retrieved 01.04.2012, from http://www.treesearch.fs.fed.us/pubs/17075

Becker, S., & Diop-Sidibé, N. (2003). Does Use of the Calendar in Surveys Reduce Heaping? *Studies in Family Planning*, *34*(2), 127–132. Retrieved 07.02.2014, from http://www.jstor.org/stable/3181184

Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society*, *78*(4), 551–572. Retrieved from http://www.jstor.org/stable/984802

Besag, J., Green, P., Higdon, D., & Mengersen, K. (1995). Bayesian Computation and Stochastic Systems. *Statistical Science*, *10*(1), 3–41. Retrieved 02.01.2016, from http://projecteuclid.org/download/pdf_1/euclid.ss/1177010123

Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (Eds.). (2011). *Education as a Lifelong Process: The German National Educational Panel Study (NEPS)* (No. 14). Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.

Bound, J., Brown, C., & Mathiowetz, N. (2001). Measurement Error in Survey Data: Chapter 59. In J. J. Heckman & E. E. Leamer (Eds.), *Handbook of Econometrics* (Vol. 5, pp. 3705–3843). Amsterdam, North-Holland: Elsevier. Retrieved 20.02.2014, from `http://www.sciencedirect.com/science/article/pii/S1573441201050127`

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (Eds.). (1984). *Classification and regression trees*. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Brockwell, A. E., & Kadane, J. B. (2005). Identification of Regeneration Times in MCMC Simulation, with Application to Adaptive Schemes. *Journal of Computational and Graphical Statistics*, *14*(2), 436–458. Retrieved 26.02.2015, from `http://www.jstor.org/stable/27594123`

Brooks, S., Gelman, A., Jones, G., & Meng, X.-L. (Eds.). (2011). *Handbook of Markov Chain Monte Carlo*. Boca Raton, FL: Chapman & Hall/CRC Press/Taylor & Francis.

Brooks, S. P., & Gelman, A. (1998). General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics*, *7*(4), 434–455. Retrieved 13.01.2015, from `http://www.jstor.org/stable/1390675`

Browne, W. J., & Draper, D. (2000). Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational Statistics*, *15*(3), 391–420. doi: 10.1007/s001800000041

Burgette, L. F., & Reiter, J. P. (2010). Multiple Imputation for Missing Data via Sequential Regression Trees. *American Journal of Epidemiology*, *172*(9), 1070–1076. doi: 10.1093/aje/kwq260

Burr, I. W. (1942). Cumulative Frequency Functions. *The Annals of Mathematical Statistics*, *13*(2), 215–232. doi: 10.1214/aoms/1177731607

Burton, S., & Blair, E. (1991). Task Conditions, Response Formulation Processes, and Response Accuracy for Behavioral Frequency Questions in Surveys. *Public Opinion Quarterly*, *55*(1), 50. doi: 10.1086/269241

Camarda, C. G., Eilers, P. H., & Gampe, J. (2007). Modelling general patterns of digit preference. In J. del Castillo, A. Espinal, & P. Puig (Eds.), *Proceedings of the 22nd International Workshop on Statistical Modelling, Barcelona* (pp. 148–153). Retrieved 01.04.2012, from `http://www.demogr.mpg.de/de/projekte_publikationen/publikationen_1904/beitraege_in_einem_sammelband/modelling_general_patterns_of_digit_preference_2692.htm`

Canty, A., & Ripley, B. (2014). *boot: Bootstrap R (S-Plus) Functions. R package version 1.3-13.* Retrieved 13.09.2015, from `http://CRAN.R-project.org/package=boot`

Carlin, B. P., & Chib, S. (1995). Bayesian Model Choice via Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(3), 473–484. Retrieved 07.01.2015, from http://www.jstor.org/stable/2346151

Carlin, B. P., & Louis, T. A. (2009). *Bayesian Methods for Data Analysis* (3rd ed.). Boca Raton, FL: Chapman & Hall/CRC Press.

Changyou, S. (2015). *erer: Empirical Research in Economics with R. R package version 2.4.* Retrieved 24.11.2015, from http://CRAN.R-project.org/package=erer

Chaurasia, A., & Harel, O. (2012). Using AIC in Multiple Linear Regression framework with Multiply Imputed Data. *Health services & outcomes research methodology*, *12*(2-3), 219–233. doi: 10.1007/s10742-012-0088-8

Chib, S. (1995). Marginal Likelihood from the Gibbs Output. *Journal of the American Statistical Association*, *90*(432), 1313–1321. doi: 10.2307/2291521

Chib, S. (Ed.). (2009). *Tailored Multiple-block MCMC Methods for Analysis of DSGE models.* Washington University. Retrieved 06.10.2014, from http://apps.olin.wustl.edu/faculty/conferences/sbies2009/uploads/Ramamurthy_Srikanth.pdf

Chib, S., & Greenberg, E. (1995). Understanding the Metropolis-Hastings Algorithm. *The American Statistician*, *49*(4), 327–335. doi: 10.1080/00031305.1995.10476177

Chib, S., & Jeliazkov, I. (2001). Marginal Likelihood from the Metropolis-Hastings Output. *Journal of the American Statistical Association*, *96*(453), 270–281. doi: 10.1198/016214501750332848

Chib, S., & Jeliazkov, I. (2005). Accept–reject Metropolis–Hastings sampling and marginal likelihood estimation. *Statistica Neerlandica*, *59*(1), 30–44. doi: 10.1111/j.1467-9574.2005.00277.x

Chib, S., & Ramamurthy, S. (2010). Tailored randomized block MCMC methods with application to DSGE models. *Journal of Econometrics*, *155*(1), 19–38. doi: 10.1016/j.jeconom.2009.08.003

Clementi, F., & Gallegati, M. (2005). Pareto's Law of Income Distribution: Evidence for Germany, the United Kingdom, and the United States. In A. Chatterjee, S. Yarlagadda, & B. K. Chakrabarti (Eds.), *Econophysics of wealth distributions* (pp. 3–14). Milan, Italy and Berlin, Germany and Heidelberg, Germany and New York, NJ: Springer. Retrieved 18.02.2015, from http://arxiv.org/pdf/physics/0504217.pdf

Cowell, F. A. (2000). Measurement of inequality. In A. B. Atkinson & F. Bourguignon (Eds.), *Handbook of income distribution* (Vol. 16, pp. 87–166). Amsterdam, Holland and New York, NJ: Elsevier. Retrieved 27.10.2015, from http://www.sciencedirect.com/science/handbooks/15740056/1

Cowles, M. K., & Carlin, B. P. (1996). Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. *Journal of the American Statistical Association*, *91*(434), 883–904. doi: 10.2307/2291683

Crayen, D., & Baten, J. (2008). *Global Trends in Numeracy 1820-1940 and its Implications for Long-Run Growth* (No. 2218). Retrieved 10.11.2015, from http://ssrn.com/abstract=1092396

Croissant, Y. (2013). *mlogit: multinomial logit model. R package version 0.2-4.* Retrieved 13.09.2015, from http://CRAN.R-project.org/package=mlogit

Curtis, S. M. (2015). *mcmcplots: Create Plots from MCMC Output. R package version 0.4.2.* Retrieved 13.09.2015, from http://CRAN.R-project.org/package=mcmcplots

Dagum, C. (1977). A New Model of Personal Income Distribution: Specification and Estimation. *Economie Appliquée*, *30*, 413–437. Retrieved 05.11.2015, from http://link.springer.com/chapter/10.1007/978-0-387-72796-7_1?null

Dahl, D. B. (2014). *xtable: Export tables to LaTeX or HTML. R package version 1.7-4.* Retrieved 13.09.2015, from http://CRAN.R-project.org/package=xtable

Daniels, R. C. (2008). *The income distribution with coarse data* (No. 82). Retrieved 27.10.2015, from http://www.econrsa.org/papers/w_papers/wp82.pdf

David, M. (1962). The Validity of Income Reported by a Sample of Families who Received Welfare Assistance During 1959. *Journal of the American Statistical Association*, *57*(299), 680–685. doi: 10.2307/2282405

Dempster, A. P., & Rubin, D. B. (1983). Rounding Error in Regression: The Appropriateness of Sheppard's Corrections. *Journal of the Royal Statistical Society, Series B (Methodological)*, *45*(1), 51–59. Retrieved 28.08.2013, from http://www.jstor.org/stable/2345623

Diekmann, A. (2007). Not the First Digit! Using Benford's Law to Detect Fraudulent Scientific Data. *Journal of Applied Statistics*, *34*(3), 321–329. doi: 10.1080/02664760601004940

DiNardo, J., Fortin, N. M., & Lemieux, T. (1996). Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach. *Econometrica*, *64*(5), 1001–1044. doi: 10.2307/2171954

Drechsler, J., & Kiesl, H. (2012). *MI double feature: multiple imputation to address nonresponse and rounding errors in income questions simultaneously.* Nuremberg and Regensburg. Retrieved from http://www.fcsm.gov/12papers/Drechsler_2012FCSM_VIII-A.pdf

Drechsler, J., & Kiesl, H. (2014). *Beat the Heap - An Imputation Strategy for Valid Inference from Rounded Income Data* (No. 2). Nuremberg and Regensburg. Retrieved 27.02.2014, from http://doku.iab.de/discussionpapers/2014/dp0214.pdf

Drechsler, J., Kiesl, H., & Speidel, M. (2015). MI Double Feature: Multiple Imputation to Address Nonresponse and Rounding Errors in Income Questions. *Austrian Journal of Statistics*, *44*(2), 59–71. doi: 10.17713/ajs.v44i2.77

Eisenhart, C. (1947). Effects of Rounding or Grouping Data: Chap. 4. In C. Eisenhart, M. W. Hastay, & W. A. Wallis (Eds.), *Techniques of Statistical Analysis*

(pp. 185–224). New York, NJ and London, UK: McGraw-Hill Book Company, Inc.

el Messlaki, F., Kuijvenhoven, L., & Moerbeck, M. (2010). *Making Use of Multiple Imputation to Analyze Heaped Data* (Master Thesis). Utrecht University, The Netherlands.

Elff, M. (2015). *memisc: Tools for Management of Survey Data, Graphics, Programming, Statistics, and Simulation. R package version 0.97.* Retrieved 13.09.2015, from `http://CRAN.R-project.org/package=memisc`

Fahrmeir, L., Künstler, R., Pigeot, I., & Tutz, G. (2007). *Statistik: Der Weg zur Datenanalyse* (6th ed.). Berlin and Heidelberg, Germany: Springer.

Földvári, P., van Leeuwen, B., & van Leeuwen-Li, J. (2012). How did women count? A note on gender-specific age heaping differences in the sixteenth to nineteenth centuries. *Economic History Review*, *65*(1), 304–313. doi: 10.1111/j.1468-0289.2010.00582.x

Frick, J. R., & Grabka, M. M. (2007). *Item Non-response and Imputation of Annual Labor Income in Panel Surveys from a Cross-National Perspective* (No. 49). Berlin. Retrieved 14.09.2015, from `http://www.diw.de/documents/publikationen/73/diw_01.c.65965.de/diw_sp0049.pdf`

Friel, N., & Pettitt, A. N. (2008). Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society, Series B (Methodological)*, *70*(3), 589–607. doi: 10.1111/j.1467-9868.2007.00650.x

Frühwirth-Schnatter, S. (1995). Bayesian Model Discrimination and Bayes Factors for Linear Gaussian State Space Models. *Journal of the Royal Statistical Society, Series B (Methodological)*, *57*(1), 237–246. Retrieved 01.04.2015, from `http://www.jstor.org/stable/2346097`

Frühwirth-Schnatter, S. (2004). Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques. *Econometrics Journal*, *7*(1), 143–167. doi: 10.1111/j.1368-423X.2004.00125.x

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models.* New York, NJ: Springer.

Gastwirth, J. L., & Glauberman, M. (1976). The Interpolation of the Lorenz Curve and Gini Index from Grouped Data. *Econometrica*, *44*(3), 479–483. doi: 10.2307/1913977

Gelfand, A. E., & Dey, D. K. (1994). Bayesian Model Choice: Asymptotics and Exact Calculations. *Journal of the Royal Statistical Society. Series B (Methodological)*, *56*(3), 501–514. Retrieved 10.03.2015, from `http://www.jstor.org/stable/2346123`

Gelfand, A. E., & Sahu, S. K. (1994). On Markov Chain Monte Carlo Acceleration. *Journal of Computational and Graphical Statistics*, *3*(3), 261–276. doi: 10.2307/1390911

Gelman, A., Carlin, John, B., Stern, H. S., & Rubin, D. B. (Eds.). (2004). *Bayesian Data Analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC Press.

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of

model fitness via realized discrepancies: (with discussion). *Statistica Sinica*, *6*, 733–807. Retrieved 07.01.2015, from http://www3.stat.sinica.edu.tw/statistica/oldpdf/A6n41.pdf

Gelman, A., Roberts, G. O., & Gilks, W. R. (1996). Efficient Metropolis Jumping Rules. *Bayesian Statistics*, *5*, 599–607. Retrieved 10.06.2014, from http://www.stat.columbia.edu/~gelman/research/published/baystat5.pdf

Gelman, A., & Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, *7*(4), 457–472. Retrieved 07.01.2015, from http://www.stat.columbia.edu/~gelman/research/published/itsim.pdf

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distribution, and the Bayesian Restoration of Images. *IEEE Trans. Pat. Anal. Mach. Intel.*, *6*(721-741). Retrieved 26.02.2015, from http://www.stat.cmu.edu/~acthomas/724/Geman.pdf

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., & Scheipl, F. (2014). *mvtnorm: Multivariate Normal and t Distributions. R package version 1.0-2.* Retrieved 13.09.2015, from http://CRAN.R-project.org/package=mvtnorm

Geweke, J. (1992). Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments. *Bayesian Statistics*, *4*, 169–193. Retrieved 15.01.2015, from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.27.2952

Geyer, C. J. (1992). Practical Markov Chain Monte Carlo. *Statistical Science*, *7*(4), 473–483. Retrieved 19.11.2014, from http://www.jstor.org/stable/2246094

Geyer, C. J. (2011). Introduction to Markov Chain Monte Carlo. In S. Brooks, A. Gelman, G. Jones, & X.-L. Meng (Eds.), *Handbook of Markov Chain Monte Carlo* (pp. 3–48). Boca Raton, FL: Chapman & Hall/CRC Press/Taylor & Francis. Retrieved 19.11.2014, from http://www.mcmchandbook.net/HandbookChapter1.pdf

Gibrat, R. (1931). *Les Inégalités économiques: Applications: aux inégalités des richesses, à la concentration des entreprises, aux populations des villes, aux statistiques des familles, etc., d'une loi nouvelle, la loi de l'effect proportionnel.* Paris, France: Recueil Sirey.

Gilbert, P., & Varadhan, R. (2012). *numDeriv: Accurate Numerical Derivatives. R package version 2012.9-1.* Retrieved 13.09.2015, from http://CRAN.R-project.org/package=numDeriv

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (Eds.). (1996). *Markov chain Monte Carlo in practice.* London, UK: Chapman & Hall/CRC Press.

Gilks, W. R., Roberts, G. O., & Sahu, S. K. (1998). Adaptive Markov Chain Monte Carlo through Regeneration. *Journal of the American Statistical Association*, *93*(443), 1045–1054. doi: 10.2307/2669848

Gill, R. D., van der Laan, M. J., & Robins, J. M. (1997). Coarsening at Random: Characterisations, Conjectures, Counter-Examples. In D. Y. Lin &

T. R. Fleming (Eds.), *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis* (pp. 255–294). New York, NJ: Springer-Verlag.

Giordani, P., & Kohn, R. (2010). Adaptive Independent Metropolis–Hastings by Fast Estimation of Mixtures of Normals. *Journal of Computational and Graphical Statistics*, *19*(2), 243–259. doi: 10.1198/jcgs.2009.07174

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, *82*(4), 711–732. doi: 10.1093/biomet/82.4.711

Greenberg, D., Moffitt, R., & Friedmann, J. (1981). Underreporting and Experimental Effects on Work Effort: Evidence from the Gary Income Maintenance Experiment. *The Review of Economics and Statistics*, *63*(4), 581–589. doi: 10.2307/1935854

Groß, M., & Rendtel, U. (2015). *Kernel Density Estimation for Heaped Data* (No. 2015/27). Berlin, Germany. Retrieved 10.09.2015, from http://edocs.fu-berlin.de/docs/servlets/MCRFileNodeServlet/FUDOCS_derivate_000000005368/discpaper2015_27.pdf;jsessionid=18DF6FFF8B27B06F8570F1FEEC1D0BAC?hosts=

Groves, R. M., Fowler, F. J., JR., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (Eds.). (2004). *Survey methodology.* Hoboken, NJ: Wiley-Interscience.

Haario, H., Saksman, E., & Tamminen, J. (1999). Adaptive proposal distribution for random walk Metropolis algorithm. *Computational Statistics*, *14*(3), 375–395. Retrieved 30.01.2015, from http://link.springer.com/article/10.1007/s001800050022

Haario, H., Saksman, E., & Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli*, *7*(2), 223–242. doi: 10.1007/s001800050022

Haario, H., Saksman, E., & Tamminen, J. (2005). Componentwise adaptation for high dimensional MCMC. *Computational Statistics*, *20*(2), 265–273. doi: 10.1007/BF02789703

Hall, P. (1982). The Influence of Rounding Errors on Some Nonparametric Estimators of a Density and its Derivatives. *SIAM Journal on Applied Mathematics*, *42*(2), 390–399. doi: 10.1137/0142030

Han, C., & Carlin, B. P. (2001). Markov Chain Monte Carlo Methods for Computing Bayes Factors: A Comparative Review. *Journal of the American Statistical Association*, *96*(455), 1122–1132. doi: 10.1198/016214501753208780

Hanisch, J. U. (2003). *Data Quality in Panel Surveys: The Impact of Panel Effect on the Precision of Income Data, Presentation at May, 26th.* Retrieved from https://www.destatis.de/DE/Methoden/Methodenpapiere/Chintex/ResearchResults/FinalConference/Downloads/Hanisch.pdf

Hanisch, J. U. (2005a). Rounded responses to income questions. *Allgemeines Statistisches Archiv*, *89*(1), 39–48. Retrieved 19.11.2012, from http://link.springer.com/article/10.1007/s101820500190

Hanisch, J. U. (2005b). *Rounding Simulation: Hanisch (2005): Rounded responses*

to income questions, AStA 89, 39-48.

Hanisch, J. U. (2006). *Rounding of income data: an empirical analysis of the quality of income data with respect to rounded values and income brackets with data from the European community household panel* (Vol. 9). Frankfurt am Main [u.a.], Germany: Peter Lang. Retrieved from `http://www.gbv.de/dms/bsz/toc/bsz260393487inh.pdf`

Hanisch, J. U., & Rendtel, U. (2002). *Quality of Income Data from Panel Surveys with Respect to Rounding* (No. 6). Retrieved 25.10.2012, from `https://www.destatis.de/DE/Methoden/Methodenpapiere/Chintex/Projekt/Workpackage5.html`

Harris, M. N., & Zhao, X. (2007). A zero-inflated ordered probit model, with an application to modelling tobacco consumption. *Journal of Econometrics*, *141*(2), 1073–1099. doi: 10.1016/j.jeconom.2007.01.002

Harrison, A. (1981). Earnings by Size: A Tale of Two Distributions. *The Review of Economic Studies*, *48*(4), 621–631. Retrieved 27.10.2015, from `http://www.jstor.org/stable/2297201`

Hastings, W. K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, *57*(1), 97–109. doi: 10.2307/2334940

Heeringa, S. G. (1996). Application of Generalized Iterative Bayesian Simulation Methods to Estimation and Inference for Coarsened Household Income and Asset Data. In Survery Research Methods Section (Ed.), *JSM Proceedings* (pp. 42–51). Alexandria, VA: American Statistical Association. Retrieved 27.10.2015, from `SurveyResearchMethodsSection`

Heitjan, D. F. (1989). Inference from Grouped Continuous Data: A Review. *Statistical Science*, *4*(2), 164–183. Retrieved 29.08.2013, from `http://www.jstor.org/stable/2245350`

Heitjan, D. F. (1994). Ignorability in general incomplete-data models. *Biometrika*, *81*(4), 701–708. doi: 10.1093/biomet/81.4.701

Heitjan, D. F. (1997). Ignorability, sufficiency and ancillarity. *Journal of the Royal Statistical Society. Series B (Methodological)*, *59*(2), 375–381. doi: 10.1111/1467-9868.00073

Heitjan, D. F., & Rubin, D. B. (1990). Inference from Coarse Data Via Multiple Imputation with Application to Age Heaping. *Journal of the American Statistical Association*, *85*(410), 304–314. doi: 10.2307/2289765

Heitjan, D. F., & Rubin, D. B. (1991). Ignorability and Coarse Data. *The Annals of Statistics*, *19*(4), 2244–2253. Retrieved from `http://www.jstor.org/stable/10.2307/2241929`

Henderson, B., & Jarrett, R. (2003). Models with Errors due to Misreported Measurements. *Australian & New Zealand Journal of Statistics*, *45*(4), 431–444. doi: 10.1111/1467-842X.00296

Henningsen, A., & Toomet, O. (2011). maxLik: A package for maximum likelihood estimation in R. *Computational Statistics*, *26*(3), 443–458. doi: 10.1007/s00180-010-0217-1

Herbst, E. P., & Schorfheide, F. (Eds.). (2015). *Bayesian Estimation of DSGE Models: (The Econometric and Tinbergen Institutes Lectures)*. Princeton, NJ: Princeton University Press Group Ltd.

Hlavac, M. (2014). *stargazer: LaTeX/HTML code and ASCII text for well-formatted regression and summary statistics tables. R package version 5.1.* Cambridge, MA. Retrieved 13.09.2015, from http://CRAN.R-project.org/package=stargazer

Hobson, R. (1976). Properties Preserved by Some Smoothing Functions. *Journal of the American Statistical Association*, *71*(355), 763–766. doi: 10.1080/01621459.1976.10481564

Hoff, P. D. (2009). *A first course in Bayesian statistical methods*. New York, NJ and London, UK: Springer.

Holbrook, A. L., Anand, S., Johnson, Timothy, P., Cho, Y. I., Shavitt, S., Chávez, N., & Weiner, S. (2014). Response Heaping in Interviewer-administered Surveys: Is it really a form of Satisficing? *Public Opinion Quarterly*, *78*(3), 591–633. doi: 10.1093/poq/nfu017

Hope, R. M. (2013). *Rmisc: Rmisc: Ryan Miscellaneous. R package version 1.5.* Retrieved 13.09.2015, from http://CRAN.R-project.org/package=Rmisc

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651–674. Retrieved 13.09.2015, from http://www.jstor.org/stable/27594202

Howes, S. (1996). The Influence of Aggregation on the Ordering of Distributions. *Economica*, *63*(250), 253–272. doi: 10.2307/2554762

Huinink, J., Brüderl, J., Nauck, B., Walper, S., Castiglioni, L., & Feldhaus, M. (2011). Panel analysis of intimate relationships and family dynamics (pairfam): conceptual framework and design. *Zeitschrift für Familienforschung*, *23*(1), 77–101. Retrieved 05.08.2016, from http://nbn-resolving.de/urn:nbn:de:0168-ssoar-376463

Humenberger, H. (2008). Eine elementarmathematische Begründung des Benford-Gesetzes. *Der Mathematikunterricht*, *54*(1), 24–34. Retrieved 30.08.2012, from http://www.oemg.ac.at/DK/Didaktikhefte/2008%20Band%2041/VortragHumenberger.pdf

Hurd, M. D. (1999). Anchoring and acquiescence bias in measuring assets in household surveys. *Journal of Risk and Uncertainty*, *19*(1-3), 111–136. Retrieved 01.04.2012, from http://journals.kluweronline.com/issn/0895-5646/contents

Hurst, E., Li, G., & Pugsle, B. (2014). Are household surveys like tax forms? Evidence from income underreporting of the self-employed. *The Review of Economics and Statistics*, *96*(1), 19–33. doi: 10.1162/REST_a_00363

Huttenlocher, J., Hedges, L. V., & Bradburn, N. M. (1990). Reports of Elapsed Time: Bounding and Rounding Processes in Estimation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(2), 196–213. doi:

10.1037/0278-7393.16.2.196

Hyndman, R. J. (1996). Computing and Graphing Highest Density Regions. *The American Statistician*, *50*(2), 120–126. doi: 10.2307/2684423

Hyndman, R. J., Einbeck, J., & Wand, M. (2013). *hdrcde: Highest density regions and conditional density estimation. R package version 3.1.* Retrieved 13.09.2015, from `http://CRAN.R-project.org/package=hdrcde`

Iniesta, R., & Moreno, V. (2013). *BayHap: Bayesian analysis of haplotype association using Markov Chain Monte Carlo. R package version 1.0.1.* Retrieved 13.09.2015, from `http://CRAN.R-project.org/package=BayHap`

Jackman, S. (2009). *Bayesian Analysis for the Social Sciences.* Chichester, UK: John Wiley & Sons.

Jeffreys, H. (1961). *Theory of probability.* Oxford and New York, NJ: Clarendon Press and Oxford University Press.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1994). *Continuous univariate distributions Vol. 1* (2nd ed.). New York, NJ: John Wiley & Sons.

Kass, R. E., & Rafterty, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773–795. doi: 10.1080/01621459.1995.10476572

Kim, J. K., & Hong, M. (2012). Imputation for statistical inference with coarse data. *The Canadian Journal of Statistics*, *40*(3), 604–618. doi: 10.1002/cjs.11142

Kleiber, C., & Kotz, S. (2003a). Beta-type Size Distributions. In *Statistical Size Distributions in Economics and Actuarial Sciences* (pp. 183–234). Hoboken, NJ: Wiley-Interscience.

Kleiber, C., & Kotz, S. (2003b). Lognormal Distributions. In *Statistical Size Distributions in Economics and Actuarial Sciences* (pp. 107–145). Hoboken, NJ: Wiley-Interscience.

Kleiber, C., & Kotz, S. (2003c). *Statistical Size Distributions in Economics and Actuarial Sciences.* Hoboken, NJ: Wiley-Interscience.

Kleiber, C., & Zeileis, A. (2008). *Applied econometrics with R.* New York, NJ: Springer.

Knaus, J. (2013). *snowfall: Easier cluster computing (based on snow). R package version 1.84-6.* Retrieved 13.09.2015, from `http://CRAN.R-project.org/package=snowfall`

Komsta, L., & Novomestky, F. (2012). *moments: Moments, cumulants, skewness, kurtosis and related tests. R package version 0.13.* Retrieved 13.09.2015, from `http://CRAN.R-project.org/package=moments`

Kraus, F., & Steiner, V. (1995). *Modelling heaping effects in unemployment duration models - With an application to retrospective event data in the German Socio-Economic Panel* (No. 95-09). Mannheim, Germany. Retrieved 01.04.2012, from `http://www.zew.de/en/publikationen/publikation.php3?action=detail&nr=142`

Kroh, M. (2004). *Interviewеffekte bei der Erhebung des Körpergewichts: die Qua-*

lität von umfragebasierten Gewichtsangaben (No. 439). Berlin, Germany. Retrieved 14.11.2015, from http://www.diw.de/sixcms/detail.php?id=diw_02.c.230208.de

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213–236. doi: 10.1002/acp.2350050305

Krul, A. J., Daanen, Hein A. M., & Choi, H. (2010). Self-reported and measured weight, height and body mass index (BMI) in Italy, the Netherlands and North America. *European Journal of Public Health*, *21*(4), 414–419. doi: 10.1093/eurpub/ckp228

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & Quantity*, *47*, 2025–2047. doi: 10.1007/s11135-011-9640-9

Lam, P. (2008). *Gov 2002 Section 6: The Metropolis-Hastings Algorithm and Convergence Diagnostics, November 3rd.* Retrieved from http://www.people.fas.harvard.edu/~plam/teaching/gov2002/section6/metropolis.pdf

Lander, J. P. (2013). *coefplot: Plots Coefficients from Fitted Models. R package version 1.2.0.* Retrieved 13.09.2015, from http://CRAN.R-project.org/package=coefplot

Lemon, J., & et al. (2006). Plotrix: a package in the red light district of R. *R-News*, *6*(4), 8–12. Retrieved 13.09.2015, from http://CRAN.R-project.org/package=plotrix

Leopold, T., Raab, M., & Skopek, J. (2011). *Data Manual: Starting Cohort 6 - Adult Education and Lifelong Learning.* Bamberg, Germany. Retrieved 12.11.2015, from https://www.neps-data.de/Portals/0/Neps/Datenzentrum/Forschungsdaten/SC6/1-0-0/SC6_1-0-0_DataManual_EN.pdf

Levine, R. A., Yu, Z., Hanley, W. G., & Nitao, J. J. (2005). Implementing componentwise Hastings algorithms. *Computational Statistics & Data Analysis*, *48*(2), 363–389. doi: 10.1016/j.csda.2004.02.002

Lillard, D. R., Bar, H., & Wang, H. (2008). *A Heap of Trouble? Accounting for Mismatch Bias in Retrospectively Reported Data: (with application to smoking cessation and (non) employment).*

Lin, T. H., & Tsai, M.-H. (2013). Modeling health survey data with excessive zero and K responses. *Statistics in Medicine*, *32*(9), 1572–1583. doi: 10.1002/sim.5650

Liu, C., & Liu, Q. (2012). Marginal likelihood calculation for the Gelfand–Dey and Chib methods. *Economics Letters*, *115*(2), 200–203. doi: 10.1016/j.econlet.2011.12.034

Liu, T., Zhang, B., Hu, G., & Bai, Z. (2007). *Revisit of Sheppard Corrections in Linear Regression* (No. 07/06). China. Retrieved 29.08.2013, from http://www.rmi.nus.edu.sg/_files/rmiworkingpp/wp0706.pdf

Mair, P., Schönbrodt, F. D., & Wilcox, R. R. (2015). *WRS2: Wilcox robust estimation and testing. R package version 0.4-0.* Retrieved 23.10.2015, from http://CRAN.R-project.org/package=WRS2

Mair, P., & Wilcox, R. R. (2015). *Robust Statistical Methods: The R Package WRS2.* Retrieved 29.10.2015, from https://cran.r-project.org/web/packages/WRS2/vignettes/WRS2.pdf

Marcus, J., Siegers, R., & Grabka, M. M. (2013). *Preparation of Data from the New SOEP Consumption Module: Editing, Imputation, and Smoothing* (No. 70). Berlin. Retrieved 07.02.2014, from http://www.diw.de/sixcms/detail.php?id=diw_01.c.428032.de

Martin, A. D., Quinn, K. M., & Park, J. H. (2011). MCMCpack: Markov Chain Monte Carlo in R. *Journal of Statistical Software*, *42*(9), 1–21. doi: 10.18637/jss.v042.i09

Mathew, B., Bauer, A. M., Koistinen, P., Reetz, T. C., Léon, J., & Sillanpää, M. J. (2012). Bayesian adaptive Markov chain Monte Carlo estimation of genetic parameters. *Heredity*, *109*(4), 235–245. doi: 10.1038/hdy.2012.35

Maynes, E. S. (1968). Minimizing Response Errors in Financial Data: The Possibilities. *Journal of the American Statistical Association*, *63*(321), 214–227. doi: 10.1080/01621459.1968.11009236

McDonald, J. B. (1984). Some generalized functions for the size distribution of income. *Econometrica*, *52*(3), 647–663. doi: 10.1007/978-0-387-72796-7_3

McDonald, J. B., Sorensen, J., & Turley, P. A. (2013). Skewness and kurtosis properties of income distribution models. *Review of Income and Wealth*, *59*(2), 360–374. doi: 10.1111/j.1475-4991.2011.00478.x

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, *21*(6), 1087–1092. doi: 10.1007/s002149900053

Milborrow, S. (2014). *rpart.plot: Plot rpart Models. An Enhanced Version of plot.rpart. R package version 1.5.0.* Retrieved 13.09.2015, from http://CRAN.R-project.org/package=rpart.plot

Miller, H. P., & Paley, L. R. (1958). Income Reported in the 1950 Census and on Income Tax Returns. In Conference on Research in Income and Wealth (Ed.), *An Appraisal of the 1950 Census Income Data* (pp. 177–204). Princeton, NJ: Princeton University Press. Retrieved from http://www.nber.org/chapters/c1053

Minoiu, C., & Reddy, S. G. (2007). *Kernel Density Estimation based on Grouped Data: The Case of Poverty Assessment.* Retrieved 02.04.2014, from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=991503 doi: 10.2139/ssrn.991503

Moore, J. C., Stinson, L. L., & Welniak, E. J., JR. (2000). *Income Measurement Error in Surveys: A Review.* Retrieved 22.08.2012, from http://www.census.gov.edgekey.net/srd/papers/pdf/sm97-05.pdf

Myers, R. J. (1940). Errors and bias in the reporting of ages in census data.

Transactions of the Actuarial Society of America, *41*(104), 395–415. doi: 10.2307/2281542

Narayan, S., & Krosnick, J. A. (1996). Education Moderates Some Response Effects in Attitude Measurement. *Public Opinion Quarterly*, *60*(1), 58. doi: 10.1086/297739

Navarro, D. (2014). *Learning statistics with R: A tutorial for psychology students and other beginners. R package version 0.4.* Adelaide, Australia. Retrieved 13.09.2015, from `http://CRAN.R-project.org/package=adaptMCMC`

Neal, R. M. (1998). *Erroneous Results in "Marginal Likelihood from the Gibbs Output".* Toronto and Ontario, Canada. Retrieved 31.05.2015, from `http://www.cs.utoronto.ca/~radford/chib-letter.html`

Nelder, J. A., & Mead, R. (1965). A Simplex Method for Function Minimization. *The Computer Journal*, *7*(4), 308–313. doi: 10.1093/comjnl/7.4.308

Newcomb, S. (1881). Note on the Frequency of the Use of different Digits in Natural Numbers. *American Journal of Mathematics*, *4*(1), 39–40. doi: 10.2307/2369148

Pace, L., Salvan, A., & Ventura, L. (2004). The Effects of Rounding on Likelihood Procedures. *Journal of Applied Statistics*, *31*(1), 29–48. doi: 10.1080/0266476032000148939

Pearson, E. S., D'Agostino, R. B., & Bowman, K. O. (1977). Tests for departure from normality: Comparison of powers. *Biometrika*, *64*(2), 231–246. doi: 10.1093/biomet/64.2.231

Pérez-Stable, E. J., Marín, B. V., Marín, G., Brody, D. J., & Benowitz, N. L. (1990). Apparent underreporting of cigarette consumption among Mexican American smokers. *American Journal of Public Health*, *80*(9), 1057–1061. Retrieved 08.06.2015, from `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1404846/`

Petoussis, K., Gill, R. D., & Zeelenberg, C. (2004). *Statistical Analysis of Heaped Duration Data.* Retrieved 12.04.2012, from `http://igitur-archive.library.uu.nl/math/2001-0712-164541/UUindex.html`

Phillips, D. P., & Feldman, K. A. (1973). A Dip in Deaths Before Ceremonial Occasions: Some New Relationships Between Social Integration and Mortality. *American Sociological Review*, *38*(6), 678–696. Retrieved 10.11.2015, from `http://www.jstor.org/stable/2094131`

Pickering, R. M. (1992). Digit preference in estimated gestational age. *Statistics in Medicine*, *11*(9), 1225–1238. doi: 10.1002/sim.4780110908

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News*, *6*(1), 7–11. Retrieved 13.09.2015, from `http://CRAN.R-project.org/doc/Rnews/Rnews_2006-1.pdf`

Preece, D. A. (1982). T is trouble (and textbooks): A critique of some examples of the paired-samples t-test. *Statistician*, *31*, 169–195. doi: 10.2307/2987888

Pudney, S. (2008). *Heaping and leaping: Survey response behaviour*

and the dynamics of self-reported consumption expenditure (No. 2008-09). Retrieved 07.11.2015, from https://www.iser.essex.ac.uk/files/iser_working_papers/2008-09.pdf

Qian, J. (1996). Restoration of Data with Rounding and Bounding Errors. In Survery Research Methods Section (Ed.), *JSM Proceedings* (Vol. 1, pp. 446–451). Alexandria, VA: American Statistical Association. Retrieved 10.11.2015, from http://www.amstat.org/sections/srms/Proceedings/

R Core Team. (2014a). *foreign: Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, Weka, dBase, ... R package version 0.8-61.* Retrieved 13.09.2015, from http://CRAN.R-project.org/package=foreign

R Core Team. (2014b). *R: A Language and Environment for Statistical Computing.* Vienna, Austria. Retrieved 13.09.2015, from http://www.R-project.org/

Reichl, J. (2015). Estimating marginal likelihoods from the posterior draws through a geometric identity. In H. Friedl & H. Wagner (Eds.), *Proceedings of the 30th International Workshop on Statistical Modelling* (Vol. 1, pp. 324–329). Linz, Austria.

Rendtel, U., Nordberg, L., Jäntti, M., Hanisch, J. U., & Basic, E. (2004). *Report on quality of income data* (No. 21). Retrieved 25.10.2012, from https://www.destatis.de/DE/Methoden/Methodenpapiere/Chintex/Projekt/Workpackage5.html

Revelle, W. (2015). *psych: Procedures for Psychological, Psychometric, and Personality Research. R package version 1.5.6.* Evanston, IL. Retrieved 13.09.2015, from http://CRAN.R-project.org/package=psych

Revolution Analytics, & Weston, S. (2014a). *doParallel: Foreach parallel adaptor for the parallel package. R package version 1.0.8.* Retrieved 13.09.2015, from http://CRAN.R-project.org/package=doParallel

Revolution Analytics, & Weston, S. (2014b). *foreach: Foreach looping construct for R. R package version 1.4.2.* Retrieved 13.09.2015, from http://CRAN.R-project.org/package=foreach

Ridout, Martin S., & Morgan, B. J. T. (1991). Modelling digit preference in fecundability studies. *Biometrics*, *47*(4), 1423–1433. doi: 10.2307/2532396

Ritter, C., & Tanner, M. A. (1992). Facilitating the Gibbs sampler: the Gibbs stopper and the griddy-Gibbs sampler. *Journal of the American Statistical Association*, *87*(419), 861–868. doi: 10.2307/2290225

Robert, C. P., & Casella, G. (Eds.). (2010). *Introducing Monte Carlo methods with R.* New York, NJ: Springer.

Roberts, G. O., Gelman, A., & Gilks, W. R. (1997). Weak convergence and optimal scaling of Random Walk Metropolis algorithms. *The Annals of Applied Probability*, *7*(1), 110–120. Retrieved 06.10.2014, from http://www.jstor.org/stable/2245134

Roberts, G. O., & Rosenthal, J. S. (2001). Optimal Scaling for Various Metropolis-Hastings Algorithms. *Statistical Science*, *16*, 351–367. Retrieved 19.12.2014, from http://www.jstor.org/stable/3182776

Roberts, G. O., & Rosenthal, J. S. (2007). Coupling and Ergodicity of Adaptive Markov Chain Monte Carlo Algorithms. *Journal of Applied Probability*, *44*(2), 458–475. Retrieved 10.02.2015, from http://www.jstor.org/stable/27595854

Roberts, G. O., & Rosenthal, J. S. (2009). Examples of Adaptive MCMC. *Journal of Computational and Graphical Statistics*, *18*(2), 349–367. doi: 10.1198/jcgs.2009.06134

Roberts, G. O., & Sahu, S. K. (1997). Updating Schemes, Correlation Structure, Blocking and Parameterization for the Gibbs Sampler. *Journal of the Royal Statistical Society. Series B (Methodological)*, *59*(2), 291–317. doi: 10.1111/1467-9868.00070

Roberts, J. M., & Brewer, D. D. (2001). Measures and tests of heaping in discrete quantitative distributions. *Journal of Applied Statistics*, *28*(7), 887–896. doi: 10.1080/02664760120074960

Rosch, E. (1975). Cognitive Reference Points. *Cognitive Psychology*, *7*, 532–547. doi: 10.1016/0010-0285(75)90021-3

Rowland, M. L. (1990). Self-reported weight and height. *American Journal of Clinical Nutrition*, *52*(6), 1125–1133. Retrieved 08.06.2015, from http://ajcn.nutrition.org/content/52/6/1125

RStudio Team. (2015). *RStudio: Integrated Development for R.* Boston, MA. Retrieved 14.09.2015, from http://www.rstudio.com/

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. doi: 10.1093/biomet/63.3.581

Rubin, D. B. (1984). Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician. *The Annals of Statistics*, *12*(4), 1151–1172. Retrieved 27.01.2015, from http://projecteuclid.org/euclid.aos/1176346785

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York, NJ: John Wiley & Sons.

Rydén, J., & Alm, S. E. (2010). The effect of interaction and rounding error in two-way ANOVA: example of impact on testing for normality. *Journal of Applied Statistics*, *37*(10), 1695–1701. doi: 10.1080/02664760903143925

Sarkar, D. (2008). *Lattice: Multivariate Data Visualization with R.* New York, NJ: Springer. Retrieved 13.09.2015, from http://lmdvr.r-forge.r-project.org

Schaeffer, N. C., & Presser, S. (2003). The Science of Asking Questions. *Annual Review of Sociology*, *29*, 65–88. doi: 10.1146/annurev.soc.29.110702.110112

Schafer, J. L. (2001). *Analyzing the NHANES III Multiply Imputed Data Set: Methods and Examples.* Hyattsville Maryland. Retrieved 04.09.2015, from ftp://ftp.cdc.gov/pub/health_statistics/nchs/nhanes/nhanes3/7a/doc/analyzing.pdf

Scheidegger, A. (2012). *adaptMCMC: Implementation of a generic adaptive Monte Carlo Markov Chain sampler. R package version 1.1.* Retrieved 13.09.2015,

from http://CRAN.R-project.org/package=adaptMCMC

Schneeweiß, H., Komlos, J., & Ahmad, A. (2006). *Symmetric and Asymmetric Rounding* (No. 479). Ludwig-Maximilians-University Munich, Germany. Retrieved 17.08.2012, from http://epub.ub.uni-muenchen.de/1847/1/paper_479.pdf

Schräpler, J.-P. (1999). *Das Befragtenverhalten im Sozio-oekonomischen Panel: Analysen an ausgewählten Beispielen* (PhD thesis, Ruhr-Universität Bochum, Germany). Retrieved 15.11.2015, from http://astro.stat.rub.de/JPS/

Schräpler, J.-P. (2002). *Respondent Behavior in Panel Studies -A Case Study for Income-Nonresponse by means of the German Socio-Economic Panel (GSOEP)* (No. 299). Berlin, Germany. Retrieved 14.11.2015, from https://www.diw.de/documents/publikationen/73/diw_01.c.38541.de/dp299.pdf

Schweitzer, M. E., & Severance-Lossin, E. K. (1996). *Rounding in Earnings Data* (No. 12). Retrieved 27.10.2015, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.362.1622&rep=rep1&type=pdf

Serfling, O. (2006). *Respondent Behavior and Data Quality Aspects in Panel Surveys: Four Empirical Contributions* (Dissertation, University of Basel). Retrieved 01.08.2012, from http://edoc.unibas.ch/509/1/DissB_7684.pdf

Sheppard, W. F. (1898). On the Calculation of the most Probable Values of Frequency-Constants, for Data arranged according to Equidistant Division of a Scale. *Proceedings of the London Mathematical Society*, *29*, 353–380. doi: 10.1112/plms/s1-29.1.353

Sherlock, C. (2005). Discussion on the Paper by Kou, Xie and Liu 'Bayesian analysis of single-molecule experimental data'. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *54*(3), 500. doi: 10.1111/j.1467-9876.2005.00509.x

Sherlock, C., Fearnhead, P., & Roberts, G. O. (2010). The Random Walk Metropolis: Linking Theory and Practice Through a Case Study. *Statistical Science*, *25*(2), 172–190. doi: 10.1214/10-STS327

Shryock, H. S., & Siegel, J. S. (Eds.). (1976). *The methods and materials of demography*. New York, NJ: Academic Press.

Sider, H. (1985). Unemployment Duration and Incidence: 1968-82. *The American Economic Review*, *75*(3), 461–472. Retrieved 08.11.2015, from http://www.jstor.org/stable/1814811

Simon, H. A. (1955). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, *69*(1), 99–118. doi: 10.2307/1884852

Singh, S. K., & Maddala, G. S. (1976). A Function for Size Distribution of Incomes. *Econometrica*, *44*(5), 963–970. doi: 10.2307/1911538

Smith, A. F. M., & Roberts, G. O. (1993). Bayesian Computation Via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, *55*(1), 3–23. Retrieved 06.10.2014, from http://www.jstor.org/stable/2346063

Spoorenberg, T., & Dutreuilh, C. (2007). Quality of Age Reporting: Extension and Application of the Modified Whipple's Index. *Population (English Edition)*, *62*(4), 729–741. Retrieved 04.07.2015, from `http://www.jstor.org/stable/27645330`

Stockwell, E. G., & Wicks, J. W. (1974). Age heaping in recent national censuses. *Social Biology*, *21*(2), 163–167. doi: 10.1080/19485565.1974.9988102

Tanner, M. A., & Wong, W. H. (1987). The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association*, *82*(398), 528–540. doi: 10.2307/2289457

Templ, M., Alfons, A., Kowarik, A., & Prantner, B. (2014). *VIM: Visualization and Imputation of Missing Values. R package version 4.1.0.* Retrieved 13.09.2015, from `http://CRAN.R-project.org/package=VIM`

Therneau, T., Atkinson, B., & Ripley, B. D. (2015). *rpart: Recursive Partitioning and Regression Trees. R package version 4.1-10.* Retrieved 13.09.2015, from `http://CRAN.R-project.org/package=rpart`

Tierney, L., & Mira, A. (1999). Some adaptive Monte Carlo methods for Bayesian inference. *Statistics in Medicine*, *18*(17-18), 2507–2515. doi: 10.1002/(SICI)1097-0258(19990915/30)18:17/18<2507::AID-SIM272>3.0.CO;2-J

Tierney, L., Rossini, A. J., Li, N., & Sevcikova, H. (2013). *snow: Simple Network of Workstations. R package version 0.3-13.* Retrieved 13.09.2015, from `http://CRAN.R-project.org/package=snow`

Torelli, N., & Trivellato, U. (1989). Youth unemployment duration from the Italian labour force survey: Accuracy issues and modelling attempts. *European Economic Review*, *33*(2-3), 407–415. doi: 10.1016/0014-2921(89)90118-9

Torelli, N., & Trivellato, U. (1993). Modelling inaccuracies in job-search duration data. *Journal of Econometrics*, *59*(1-2), 187–211. doi: 10.1016/0304-4076(93)90045-7

Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). *The Psychology of Survey Response*. Cambridge, UK and New York, NJ: Cambridge University Press.

Trautmann, H., Steuer, D., Mersmann, O., & Bornkamp, B. (2014). *truncnorm: Truncated normal distribution. R package version 1.0-7.* Retrieved 13.09.2015, from `http://CRAN.R-project.org/package=truncnorm`

Tricker, A. R. (1984). Effects of Rounding Data Sampled from the Exponential Distribution. *Journal of Applied Statistics*, *11*(1), 54–87. doi: 10.1080/02664768400000007

Tricker, A. R. (1990a). The effect of rounding on the significance level and power of certain test statistics for non-normal data. *Journal of Applied Statistics*, *17*(3), 329–340. doi: 10.1080/02664769000000005

Tricker, A. R. (1990b). The effect of rounding on the significance level of certain normal test statistics. *Journal of Applied Statistics*, *17*(1), 31–38. doi: 10.1080/757582644

Tricker, A. R. (1992). Estimation of parameters for rounded data from non-normal distributions. *Journal of Applied Statistics*, *19*(4), 465–471. doi:

10.1080/02664769200000041

Tricker, A. R. (1995). Effect on Bayesian inference of the degree of precision of recorded data. *Journal of Applied Statistics*, *22*(2), 235–244. doi: 10.1080/757584617

Turek, D., Valpine, P. d., Paciorek, C. J., & Anderson-Bergman, C. (2015). *Automated Parameter Blocking for Efficient Markov-Chain Monte Carlo Sampling.* Retrieved 22.12.2015, from `http://arxiv.org/pdf/1503.05621.pdf`

Tversky, A., & Kahneman, D. (1974). Judgement under Uncertainty: Heuristics and Biases: Biases in judgements reveal some heuristics of thinking under uncertainty. *Science*, *185*(4157), 1124–1131. Retrieved from `http://people.hss.caltech.edu/~camerer/Ec101/JudgementUncertainty.pdf`

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully condititional specification. *Statistical Methods in Medical Research*, *16*(3), 219–242. doi: 10.1177/0962280206074463

van Buuren, S. (2012). *Flexible imputation of missing data.* Boca Raton, FL: Chapman & Hall/CRC Press.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(3), 1–67. doi: 10.18637/jss.v045.i03

van der Laan, J., & Kuijvenhoven, L. (2011). *Imputation of rounded data* (No. 201108). The Hague/Heerlen, The Netherlands. Retrieved from `http://www.cbs.nl/NR/rdonlyres/3CBACED8-5B4C-415E-8E8F-F9C2752C8E23/0/2011x1008.pdf`

Vardeman, S. B., & Lee, C.-S. (2003). *Likelihood-Based Statistical Estimation From Quantized Data* (No. 38/03). Retrieved 15.09.2015, from `https://www.statistik.uni-dortmund.de/fileadmin/user_upload/Lehrstuehle/MSind/SFB_475/2003/tr39-03.pdf`

Venables, W. N., & Ripley, B. D. (Eds.). (2002). *Modern Applied Statistics with S* (4th ed.). New York, NJ: Springer. Retrieved 13.09.2015, from `http://www.stats.ox.ac.uk/pub/MASS4`

Vermunt, J. K., van Ginkel, Joost R., van der Ark, L. Andries, & Sijtsma, K. (2008). Multiple Imputation of Incomplete Categorical Data Using Latent Class Analysis. *Sociological Methodology*, *38*(1), 369–397. doi: 10.1111/j.1467-9531.2008.00202.x

Vihola, M. (2011). Can the adaptive Metropolis algorithm collapse without the covariance lower bound? *Electronic Journal of Probability*, *16*(2), 45–75. Retrieved 10.01.2016, from `http://ejp.ejpecp.org/article/viewFile/840/1060`

Vihola, M. (2012). Robust adaptive Metropolis algorithm with coerced acceptance rate. *Statistics and Computing*, *22*(5), 997–1008. doi: 10.1007/s11222-011-9269-5

Wang, H., & Heitjan, D. F. (2008). Modeling heaping in self-reported cigarette counts. *Statistics in Medicine*, *27*, 3789–3804. doi: 10.1002/sim.3281

Wang, H., Shiffman, S., Griffith, S. D., & Heitjan, D. F. (2012). Truth and memory: linking instantaneous and retrospective self-reported cigarette consumption. *The Annals of Applied Statistics*, *6*(4), 1689–1706. doi: 10.1214/12-AOAS557

Warner, K. E. (1978). Possible Increases in the Underreporting of Cigarette Consumption. *Journal of the American Statistical Association*, *73*(362), 314–318. doi: 10.2307/2286658

Wei, G. C. G., & Tanner, M. A. (1990). A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms. *Journal of the American Statistical Association*, *85*(411), 699–704. doi: 10.2307/2290005

Weisberg, H. F., Krosnick, J. A., & Bowen, B. D. (Eds.). (1989). *An introduction to survey research and data analysis* (2nd ed.). Glenview, IL: Scott, Foresman.

Weisberg, H. F., Krosnick, J. A., & Bowen, B. D. (Eds.). (1996). *An introduction to survey research, polling, and data analysis* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Dordrecht, The Netherlands and New York, NJ: Springer.

Wickham, H. (2011). The Split-Apply-Combine Strategy for Data Analysis. *Journal of Statistical Software*, *40*(1), 1–29. doi: 10.18637/jss.v040.i01

Wilcox, R. R., & Schönbrodt, F. D. (2014). *The WRS package for robust statistics in R* (No. R package version 0.24). Retrieved 23.10.2015, from `http://r-forge.r-project.org/projects/wrs/`

Wilkinson, D. J., & Yeung, S. K. H. (2002). Conditional simulation from highly structured Gaussian systems, with application to blocking-MCMC for the Bayesian analysis of very large linear models. *Statistics and Computing*, *12*(3), 287–300. doi: 10.1023/A:1020711129064

Williams, G. J. (2011). *Data Mining with Rattle and R: The art of excavating data for knowledge discovery*. New York, NJ: Springer. Retrieved 13.09.2015, from `http://CRAN.R-project.org/package=rattle`

Wilrich, P.-T. (2005). Rounding of measurement values or derived values. *Measurement*, *37*(1), 21–30. doi: 10.1016/j.measurement.2004.08.005

Wolff, J., & Augustin, T. (2000). *Heaping and its Consequences for Duration Analysis* (No. 203). Retrieved 12.04.2012, from `http://epub.ub.uni-muenchen.de/1593/1/paper_203.pdf`

Wolff, J., & Augustin, T. (2003). Heaping and its consequences for duration analysis: A simulation study. *Allgemeines Statistisches Archiv*, *87*(1), 59–86.

Wolff, S. (2004). *A local and globalized, constrained and simple bounded Nelder-Mead method*. Istanbul, Turkey. Retrieved 03.03.2015, from `http://home.ku.edu.tr/~daksen/2004-Nelder-Mead-Method-Wolff.pdf`

Wright, D. E., & Bray, I. (2003). A Mixture Model for Rounded Data. *Journal of the Royal Statistical Society, Series D (The Statistican)*, *52*(1), 3–13. doi:

10.1111/1467-9884.00338

Würbach, A. (2015). Estimation of a general heaping model via random-walk Metropolis algorithms. In H. Friedl & H. Wagner (Eds.), *Proceedings of the 30th International Workshop on Statistical Modelling* (Vol. 1, pp. 392–397). Linz, Austria.

Würbach, A., Hammon, A., Geissler, F., & Goßmann, S. (2014). *Data Documentation: Imputed Data File of Starting Cohort 6.* Bamberg, Germany. Retrieved from `https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC6/1-0-0/SC6_DataDoc_Imp_1-0-0.pdf`

Yee, T. W. (2014). *VGAM: Vector Generalized Linear and Additive Models. R package version 0.9-5.* Retrieved 13.09.2015, from `http://CRAN.R-project.org/package=VGAM`

Yeung, S. K. H., & Wilkinson, D. J. (2001). *Adaptive Metropolis-Hastings samplers for the Bayesian analysis of large linear Gaussian systems.* Retrieved 22.12.2015, from `http://www.interfacesymposia.org/I01/I2001Proceedings/SYeung/SYeung.pdf`

Zhang, J., & Heitjan, D. F. (2006). A Simple Local Sensitivity Analysis Tool for Nonignorable Coarsening: Application to Dependent Censoring. *Biometrics*, *62*(4), 1260–1268. doi: 10.1111/j.1541-0420.2006.00580.x

Zhang, J., & Heitjan, D. F. (2007). Impact of nonignorable coarsening on Bayesian inference. *Biostatistics*, *8*(4), 722–743. Retrieved 27.08.2013, from `http://biostatistics.oxfordjournals.org/content/8/4/722.full`

Zhang, Y. C., & Schwarz, N. (2012). How and Why 1 Year Differs from 365 Days: A Conversational Logic Analysis of Inferences from the Granularity of Quantitative Expressions. *Journal of Consumer Research*, *39*(2), 248–259. doi: 10.1086/662612

Zinn, S. (2014). *A multiple imputation approach to address the problem of nonignorable nonresponse and misreporting patterns in income data, Presentation at the XVIII ISA World Congress of Sociology.* Yokohama, Japan.

Zinn, S., & Würbach, A. (2014). *A Statistical Approach to Account for Heaping Patterns: An Application to Self-Reported Income Data* (No. 40). Bamberg, Germany. Retrieved 26.02.2015, from `https://www.neps-data.de/Portals/0/Working%20Papers/WP_XXXX.pdf`

Zinn, S., & Würbach, A. (2015). A Statistical Approach to Address the Problem of Heaping in Self-Reported Income Data. *Journal of Applied Statistics*. doi: 10.1080/02664763.2015.1077372

Appendix A

Additional material

A.1 Supplemental tables and figures

The specific question wording for an exact estimate of the net individual income is:

> "Wie hoch war im letzten Monat Ihr Netto-Arbeitsverdienst für Ihre Tätigkeit als ‹Berufsbezeichnung (KldB 1988)›? Bitte geben Sie die Summe an, die Sie !!nach!! Abzug der Steuern und Sozialversicherungsbeiträge erhalten haben. Wenn Sie im letzten Monat Sonderzahlungen hatten, z.B. Urlaubsgeld oder Nachzahlungen, rechnen Sie diese bitte nicht mit. Entgelt für Überstunden rechnen Sie dagegen mit. Bitte schätzen Sie Ihren derzeitigen monatlichen Gewinn nach Steuer für Ihre Tätigkeit als ‹Berufsbezeichnung (KldB 1988)›." «Falls nicht genau bekannt: monatlichen Betrag schätzen lassen.»

The specific question wording for an exact estimate of the net household income is:

> "Jetzt geht es um alle Einkünfte ihres Haushalts: Wie hoch ist Ihr monatliches Haushaltseinkommen aller Haushaltsmitglieder heute? Bitte geben Sie den Netto-Betrag an. Regelmäßige Zahlungen wie Renten, Wohngeld, Kindergeld, BAFöG, Unterhaltszahlungen, Arbeitslosengeld usw. rechnen Sie bitte dazu!" «Falls nicht genau bekannt: monatlichen Betrag schätzen lassen. Hinweis auf Anonymität geben. Bei Unklarheit bzgl. Nettoeinkommen: Bitte geben Sie die Summe an, die Sie nach Abzug der Steuern und Sozialabgaben erhalten haben.»

See the codebook (Supplement A) of Starting Cohort 6 SUF 1.0.0 in Leopold et al. (2011).

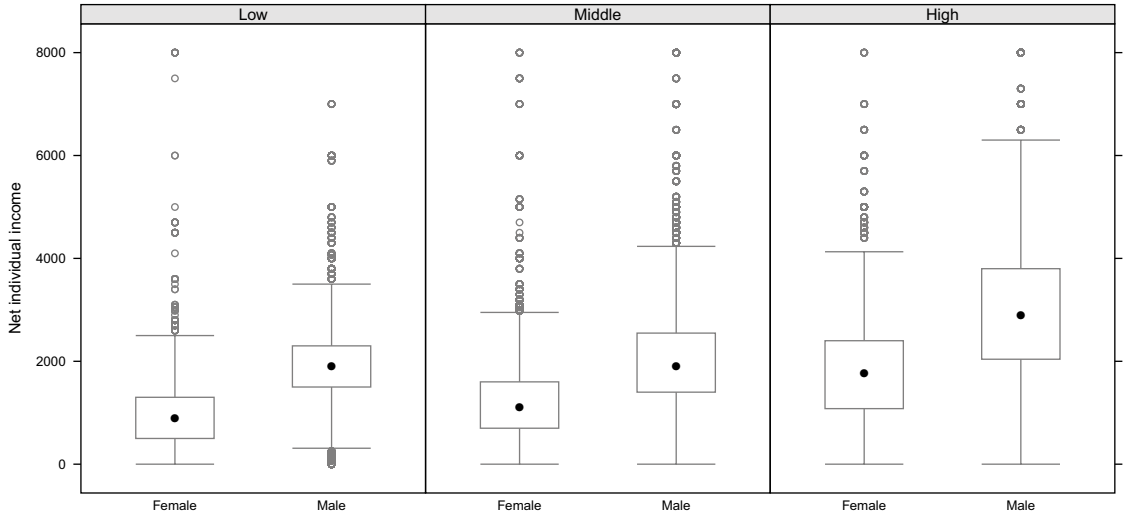


Figure A.1: Self-reported net individual income of females and males separated by educational level. The distribution is truncated at 8000 EUR for a better visualization.

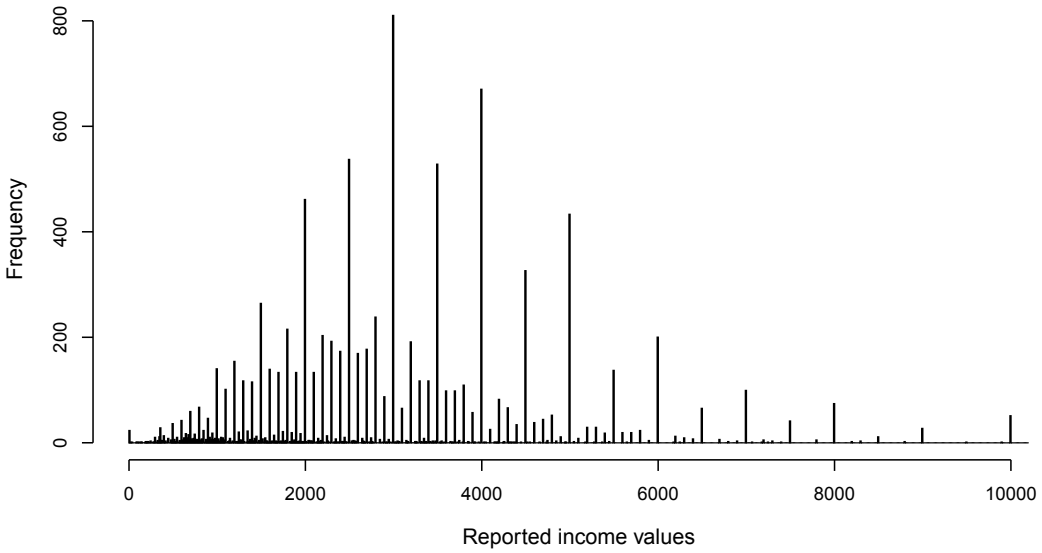


Figure A.2: Self-reported net household income data from the Adult Cohort in the NEPS wave 2009/2010, $n = 10{,}012$ ($\leq$ 10,000 EUR).

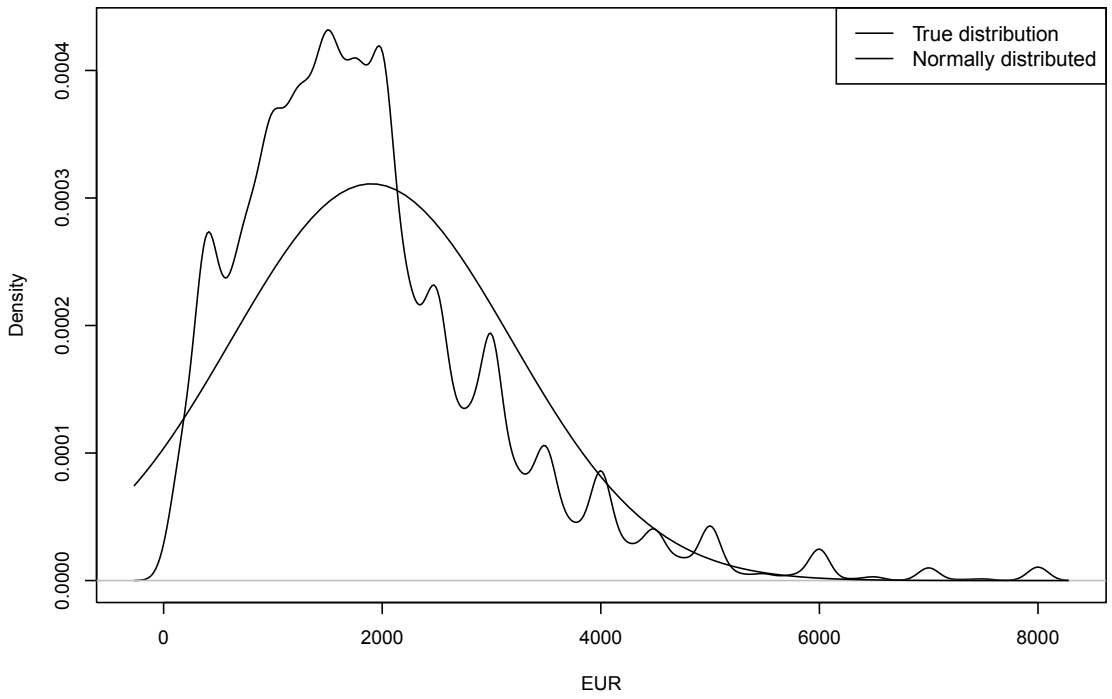


Figure A.3: Kolmogorov-Smirnov test of net income against the normal distribution function, p-value < 0.001.

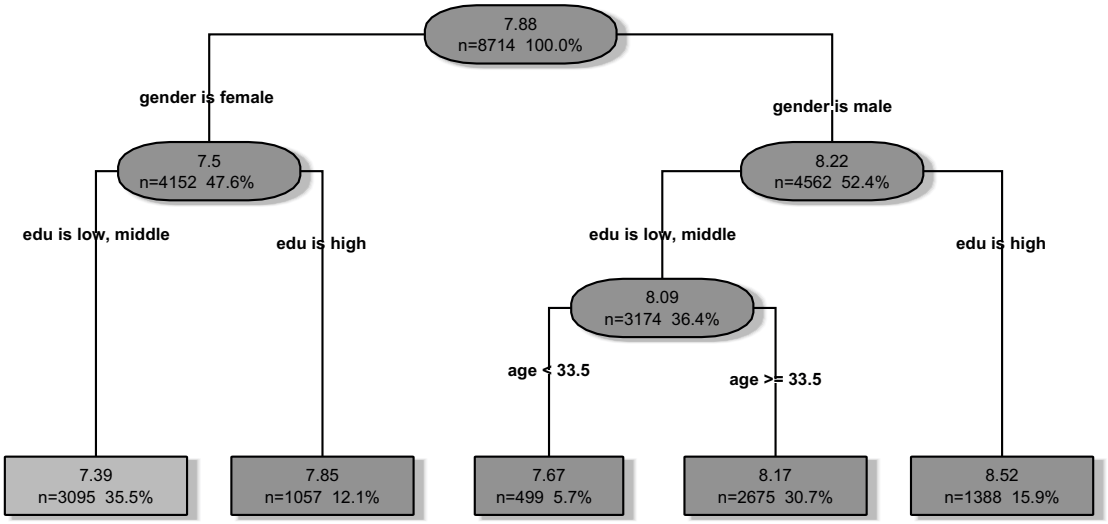


Figure A.4: Regression tree for net individual income with IHS-transformation.

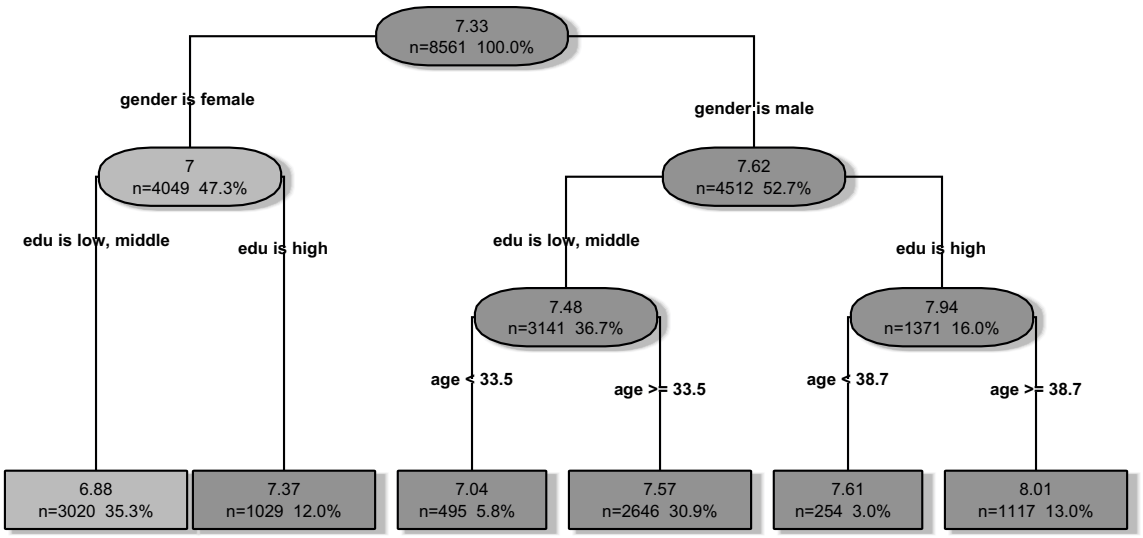


Figure A.5: Regression tree for logarithmized net individual income (zeros excluded).

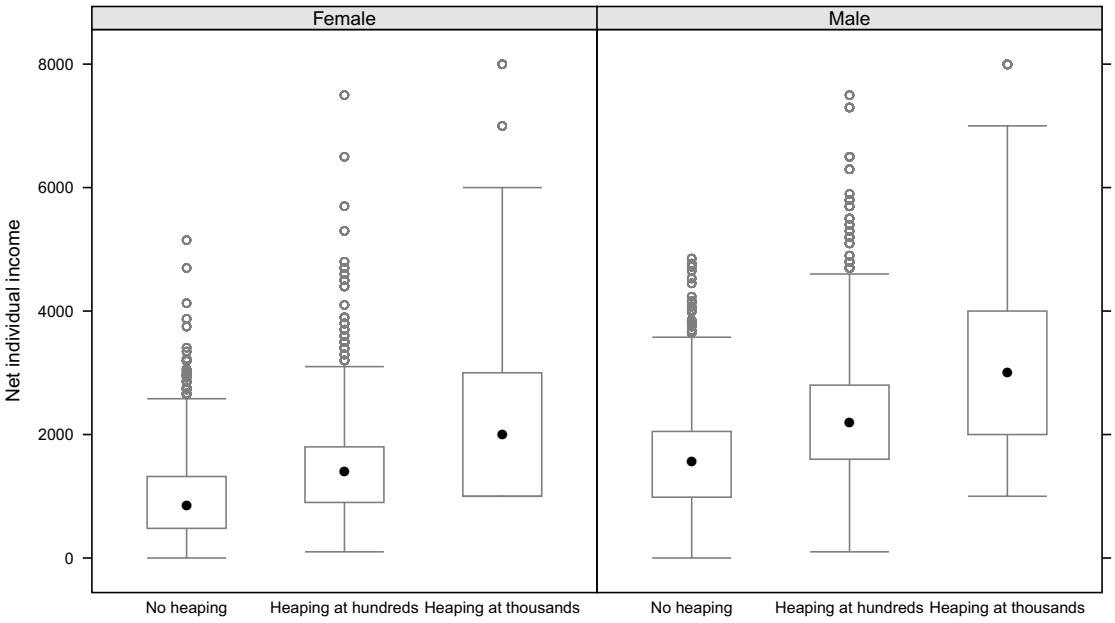


Figure A.6: Self-reported net individual income of females and males separated by degree of heaping. The distribution truncated at 8000 EUR for a better visualization.

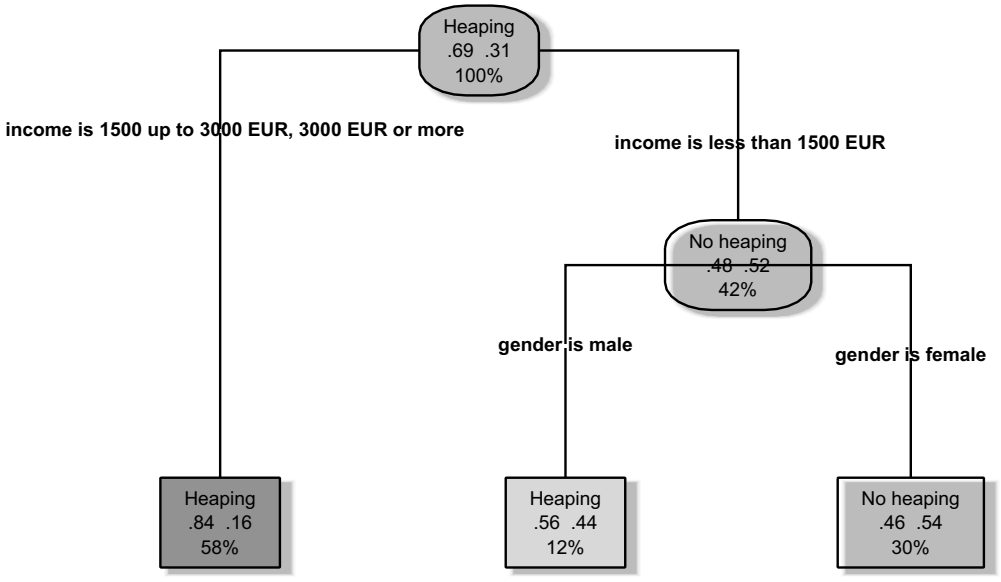


Figure A.7: Classification tree for observing heaping, with income level. (Variable importance: income 72, gender 24, age 4.)

Table A.1: Results from combined ordered probit regression for the relative RI.

Predictor	Estimate	*SE*	*CI*	*df*	*t*-ratio	*p*-value	*fmi*	*lambda*
Male	-0.357	0.025	[-0.405,-0.309]	605.9	-14.53	<0.001	0.1199	0.1170
Age	-0.008	0.001	[-0.010,-0.005]	1288.9	-6.26	<0.001	0.0780	0.0766
Middle edu	-0.040	0.032	[-0.103,0.024]	921.8	-1.24	0.217	0.0948	0.0929
Higher edu	-0.248	0.035	[-0.318,-0.179]	1069.2	-7.01	<0.001	0.0871	0.0854
0.25|0.33	-1.745	0.068	[-1.879,-1.612]	2977.0	-25.57	<0.001	0.0447	0.0441
0.33|0.40	-1.428	0.068	[-1.561,-1.295]	2574.0	-21.04	<0.001	0.0498	0.0491
0.40|0.50	-1.408	0.068	[-1.541,-1.275]	2741.6	-20.77	<0.001	0.0476	0.0469
0.50|0.67	-0.113	0.066	[-0.243,0.018]	3241.5	-1.70	0.090	0.0418	0.0412
0.67|0.75	0.170	0.066	[0.039,0.300]	3480.9	2.56	0.011	0.0394	0.0389
0.75|1.00	0.601	0.067	[0.471,0.732]	4550.2	9.03	<0.001	0.0306	0.0302

Notes: $AIC = 26{,}804.1$. The intercepts of 0.33|0.40 and 0.40|0.50 do not differ to a large extent indicating that these categories could be summarized into one category. The quantities fmi and *lambda* are both below 0.2 indicating a modest influence of the imputation model on the final result.

Table A.2: Results from combined ordered probit regression for the RSM.

Predictor	Estimate	*SE*	*CI*	*df*	*t*-ratio	*p*-value	*fmi*	*lambda*
Male	0.201	0.027	[0.148,0.255]	486.1	7.38	<0.001	0.1352	0.1316
Age	0.006	0.001	[0.004,0.009]	2463.5	4.93	<0.001	0.0514	0.0506
Middle edu	0.043	0.036	[-0.028,0.114]	651.3	1.18	0.239	0.1152	0.1125
Higher edu	0.203	0.041	[0.122,0.284]	277.9	4.94	<0.001	0.1823	0.1764
2|3	-0.833	0.073	[-0.976,-0.690]	4961.4	-11.41	<0.001	0.0278	0.0274
3|4	1.256	0.074	[1.110,1.402]	2745.9	16.92	<0.001	0.0475	0.0468
4|5	3.216	0.097	[3.026,3.405]	5466.3	33.32	<0.001	0.0246	0.0242

Notes: $AIC = 14{,}511.9$. All intercepts differ to a large extent indicating that the categories should not be summarized anyway. The quantities fmi and *lambda* are both below 0.2 indicating a modest influence of the imputation model on the final result.

Table A.3: Results from combined ordered probit regression for the relative RI with additional external factors.

Predictor	Estimate	*SE*	*CI*	*df*	*t*-ratio	*p*-value	*fmi*	*lambda*
Male	-0.352	0.025	[-0.400,-0.304]	600.5	-14.31	<0.001	0.1205	0.1175
Age	-0.008	0.001	[-0.010,-0.005]	1655.0	-6.45	<0.001	0.0669	0.0658
Middle edu	-0.049	0.032	[-0.113,0.014]	1006.6	-1.52	0.129	0.0902	0.0884
Higher edu	-0.274	0.036	[-0.344,-0.204]	1156.8	-7.70	<0.001	0.0832	0.0816
CATI	0.022	0.036	[-0.050,0.093]	232.4	0.60	0.550	0.2004	0.1935
Duration	0.007	0.001	[0.005,0.009]	2667.2	7.28	<0.001	0.0485	0.0478
High incentive	-0.076	0.030	[-0.134,-0.017]	2967.3	-2.52	0.012	0.0448	0.0442
0.25|0.33	-1.462	0.085	[-1.628,-1.296]	2701.4	-17.23	<0.001	0.0481	0.0474
0.33|0.40	-1.143	0.084	[-1.309,-0.978]	2651.8	-13.53	<0.001	0.0487	0.0480
0.40|0.50	-1.123	0.084	[-1.288,-0.957]	2770.5	-13.30	<0.001	0.0472	0.0465
0.50|0.67	0.177	0.084	[0.013,0.341]	3186.0	2.12	0.034	0.0424	0.0418
0.67|0.75	0.461	0.084	[0.297,0.625]	3128.4	5.50	<0.001	0.0430	0.0424
0.75|1.00	0.894	0.084	[0.730,1.058]	4427.8	10.68	<0.001	0.0315	0.0311

Notes: $AIC = 26{,}745.8$. The intercepts of 0.33|0.40 and 0.40|0.50 do not differ to a large extent indicating that these categories could be summarized into one category. The quantities fmi and *lambda* are both below 0.2 indicating a modest influence of the imputation model on the final result.

Table A.4: Results from combined ordered probit regression for the RSM with additional external factors.

Predictor	Estimate	*SE*	*CI*	*df*	*t*-ratio	*p*-value	*fmi*	*lambda*
Male	0.196	0.027	[0.143,0.250]	496.0	7.20	<0.001	0.1337	0.1302
Age	0.007	0.001	[0.004,0.009]	3418.3	4.97	<0.001	0.0400	0.0394
Middle edu	0.051	0.036	[-0.020,0.122]	717.3	1.42	0.155	0.1092	0.1067
Higher edu	0.224	0.041	[0.143,0.305]	311.3	5.46	<0.001	0.1716	0.1663
CATI	-0.063	0.042	[-0.146,0.020]	124.1	-1.50	0.137	0.2782	0.2667
Duration	-0.005	0.001	[-0.007,-0.003]	3603.5	-5.11	<0.001	0.0383	0.0377
High incentive	0.038	0.033	[-0.026,0.103]	3978.1	1.17	0.242	0.0350	0.0345
2|3	-1.107	0.094	[-1.291,-0.922]	1485.5	-11.75	<0.001	0.0716	0.0703
3|4	0.988	0.094	[0.802,1.173]	1174.6	10.46	<0.001	0.0824	0.0809
4|5	2.950	0.113	[2.729,3.170]	2227.6	26.21	<0.001	0.0551	0.0543

Notes: $AIC = 14{,}489.6$. All intercepts differ to a large extent indicating that the categories should not be summarized anyway. The quantities fmi and *lambda* are both below 0.3 indicating a moderately large influence of the imputation model on the final result.

$$
\Omega_{q_{AP}} = \begin{pmatrix}
1.000 & 0.130 & -0.182 & 0.362 & 0.200 & 0.305 & 0.228 & -0.313 & 0.032 & 0.186 & -0.183 & 0.064 & 0.035 & -0.043 & -0.184 & -0.520 & 0.206 & -0.142 & -0.160 & -0.205 & -0.026 & 0.018 \\
0.130 & 1.000 & 0.026 & 0.325 & 0.154 & 0.185 & < 0.001 & -0.047 & 0.040 & 0.010 & -0.124 & -0.059 & 0.215 & -0.449 & -0.020 & -0.300 & 0.260 & -0.062 & -0.103 & -0.002 & -0.018 & 0.068 \\
-0.182 & 0.026 & 1.000 & 0.213 & 0.355 & 0.250 & -0.326 & 0.204 & -0.333 & -0.062 & -0.272 & 0.428 & -0.015 & -0.224 & -0.238 & 0.002 & -0.044 & < 0.001 & -0.254 & -0.067 & 0.041 & 0.025 \\
0.362 & 0.325 & 0.213 & 1.000 & 0.137 & 0.406 & 0.170 & -0.126 & -0.203 & 0.152 & -0.367 & 0.231 & -0.010 & -0.345 & -0.253 & -0.301 & 0.167 & -0.158 & -0.134 & -0.013 & -0.032 & 0.072 \\
0.200 & 0.154 & 0.355 & 0.137 & 1.000 & 0.145 & -0.004 & -0.036 & -0.203 & -0.447 & -0.032 & 0.039 & 0.157 & -0.022 & -0.262 & -0.403 & 0.085 & 0.114 & 0.056 & -0.054 & 0.031 & -0.169 \\
0.305 & 0.185 & 0.250 & 0.406 & 0.145 & 1.000 & 0.089 & -0.037 & -0.241 & 0.397 & -0.680 & 0.388 & -0.074 & -0.068 & -0.084 & -0.516 & 0.020 & -0.252 & -0.267 & -0.054 & 0.119 & -0.032 \\
0.228 & < 0.001 & -0.326 & 0.170 & -0.004 & 0.089 & 1.000 & -0.407 & < 0.001 & -0.121 & -0.038 & -0.420 & 0.131 & 0.089 & -0.068 & -0.048 & -0.175 & -0.233 & 0.038 & 0.239 & -0.153 & -0.164 \\
-0.313 & -0.047 & 0.204 & -0.126 & -0.036 & -0.037 & -0.407 & 1.000 & 0.082 & -0.083 & -0.153 & -0.030 & -0.061 & -0.033 & -0.010 & 0.118 & 0.122 & 0.117 & 0.063 & -0.374 & 0.076 & 0.117 \\
0.032 & 0.040 & -0.333 & -0.203 & -0.203 & -0.241 & < 0.001 & 0.082 & 1.000 & 0.050 & -0.005 & -0.103 & 0.336 & -0.040 & -0.195 & 0.151 & 0.152 & 0.186 & -0.181 & 0.008 & -0.146 & 0.103 \\
0.186 & 0.010 & -0.062 & 0.152 & -0.447 & 0.397 & -0.121 & -0.083 & 0.050 & 1.000 & -0.306 & 0.387 & -0.219 & 0.106 & -0.094 & -0.258 & 0.144 & -0.260 & -0.206 & -0.010 & 0.186 & 0.227 \\
-0.183 & -0.124 & -0.272 & -0.367 & -0.032 & -0.680 & -0.038 & -0.153 & -0.005 & -0.306 & 1.000 & -0.294 & < 0.001 & 0.023 & 0.324 & 0.141 & -0.197 & 0.067 & 0.234 & 0.098 & 0.044 & < 0.001 \\
0.064 & -0.059 & 0.428 & 0.231 & 0.039 & 0.388 & -0.420 & -0.030 & -0.103 & 0.387 & -0.294 & 1.000 & -0.159 & -0.079 & 0.016 & -0.112 & -0.034 & -0.024 & -0.417 & -0.085 & 0.022 & -0.114 \\
0.035 & 0.215 & -0.015 & -0.010 & 0.157 & -0.074 & 0.131 & -0.061 & 0.336 & -0.219 & < 0.001 & -0.159 & 1.000 & -0.229 & -0.169 & -0.004 & 0.149 & 0.305 & -0.375 & 0.153 & -0.080 & 0.141 \\
-0.043 & -0.449 & -0.224 & -0.345 & -0.022 & -0.068 & 0.089 & -0.033 & -0.040 & 0.106 & 0.023 & -0.079 & -0.229 & 1.000 & -0.081 & 0.110 & -0.073 & 0.087 & 0.072 & 0.022 & -0.134 & -0.135 \\
-0.184 & -0.020 & -0.238 & -0.253 & -0.262 & -0.084 & -0.068 & -0.010 & -0.195 & -0.094 & 0.324 & 0.016 & -0.169 & -0.081 & 1.000 & -0.017 & -0.241 & -0.158 & 0.209 & 0.143 & 0.023 & -0.146 \\
-0.520 & -0.300 & 0.002 & -0.301 & -0.403 & -0.516 & -0.048 & 0.118 & 0.151 & -0.258 & 0.141 & -0.112 & -0.004 & 0.110 & -0.017 & 1.000 & -0.360 & 0.282 & 0.010 & -0.038 & -0.147 & -0.093 \\
0.206 & 0.260 & -0.044 & 0.167 & 0.085 & 0.020 & -0.175 & 0.122 & 0.152 & 0.144 & -0.197 & -0.034 & 0.149 & -0.073 & -0.241 & -0.360 & 1.000 & -0.014 & 0.006 & -0.040 & 0.028 & 0.139 \\
-0.142 & -0.062 & < 0.001 & -0.158 & 0.114 & -0.252 & -0.233 & 0.117 & 0.186 & -0.260 & 0.067 & -0.024 & 0.305 & 0.087 & -0.158 & 0.282 & -0.014 & 1.000 & -0.050 & -0.073 & -0.117 & -0.044 \\
-0.160 & -0.103 & -0.254 & -0.134 & 0.056 & -0.267 & 0.038 & 0.063 & -0.181 & -0.206 & 0.234 & -0.417 & -0.375 & 0.072 & 0.209 & 0.010 & 0.006 & -0.050 & 1.000 & -0.191 & 0.119 & -0.009 \\
-0.205 & -0.002 & -0.067 & -0.013 & -0.054 & -0.054 & 0.239 & -0.374 & 0.008 & -0.010 & 0.098 & -0.085 & 0.153 & 0.022 & 0.143 & -0.038 & -0.040 & -0.073 & -0.191 & 1.000 & 0.009 & 0.036 \\
-0.026 & -0.018 & 0.041 & -0.032 & 0.031 & 0.119 & -0.153 & 0.076 & -0.146 & 0.186 & 0.044 & 0.022 & -0.080 & -0.134 & 0.023 & -0.147 & 0.028 & -0.117 & 0.119 & 0.009 & 1.000 & 0.088 \\
0.018 & 0.068 & 0.025 & 0.072 & -0.169 & -0.032 & -0.164 & 0.117 & 0.103 & 0.227 & < 0.001 & -0.114 & 0.141 & -0.135 & -0.146 & -0.093 & 0.139 & -0.044 & -0.009 & 0.036 & 0.088 & 1.000
\end{pmatrix} \tag{A.1}
$$

$$
\Omega_{q_{AM}} = \begin{pmatrix}
1.000 & -0.192 & -0.530 & -0.361 & -0.196 & -0.358 & -0.358 & 0.362 & 0.565 & -0.035 & -0.229 & 0.438 & 0.129 & 0.158 & -0.258 & -0.073 & 0.362 & 0.224 & 0.047 & -0.061 & 0.221 & 0.200 \\
-0.192 & 1.000 & 0.248 & 0.026 & -0.271 & -0.023 & < 0.001 & -0.339 & -0.014 & 0.587 & 0.110 & -0.060 & -0.082 & -0.471 & 0.110 & -0.282 & -0.031 & -0.286 & 0.030 & -0.170 & -0.211 & 0.064 \\
-0.530 & 0.248 & 1.000 & 0.568 & 0.577 & 0.481 & 0.253 & -0.767 & -0.719 & -0.100 & 0.263 & -0.709 & -0.213 & -0.112 & 0.669 & -0.104 & -0.511 & -0.316 & -0.198 & 0.435 & -0.104 & -0.141 \\
-0.361 & 0.026 & 0.568 & 1.000 & 0.419 & 0.546 & 0.287 & -0.407 & -0.635 & -0.085 & -0.115 & -0.503 & -0.311 & -0.042 & 0.622 & -0.059 & -0.232 & -0.215 & -0.150 & 0.181 & 0.207 & -0.161 \\
-0.196 & -0.271 & 0.577 & 0.419 & 1.000 & 0.210 & 0.204 & -0.519 & -0.516 & -0.392 & 0.430 & -0.674 & -0.225 & 0.206 & 0.449 & 0.168 & -0.629 & 0.175 & -0.091 & 0.414 & 0.144 & -0.196 \\
-0.358 & -0.023 & 0.481 & 0.546 & 0.210 & 1.000 & 0.416 & -0.086 & -0.437 & -0.228 & 0.041 & -0.404 & -0.062 & -0.302 & 0.308 & 0.184 & -0.380 & -0.128 & 0.167 & 0.220 & -0.132 & -0.358 \\
-0.358 & < 0.001 & 0.253 & 0.287 & 0.204 & 0.416 & 1.000 & -0.194 & -0.303 & -0.095 & 0.154 & -0.588 & -0.462 & -0.094 & < 0.001 & -0.164 & -0.324 & 0.200 & 0.208 & 0.018 & 0.132 & -0.447 \\
0.362 & -0.339 & -0.767 & -0.407 & -0.519 & -0.086 & -0.194 & 1.000 & 0.621 & -0.119 & -0.235 & 0.649 & 0.373 & -0.127 & -0.581 & 0.479 & 0.279 & 0.280 & 0.382 & -0.346 & -0.142 & 0.048 \\
0.565 & -0.014 & -0.719 & -0.635 & -0.516 & -0.437 & -0.303 & 0.621 & 1.000 & 0.080 & -0.207 & 0.608 & 0.253 & 0.048 & -0.661 & 0.066 & 0.458 & 0.242 & 0.147 & -0.204 & 0.083 & 0.136 \\
-0.035 & 0.587 & -0.100 & -0.085 & -0.392 & -0.228 & -0.095 & -0.119 & 0.080 & 1.000 & 0.081 & 0.243 & -0.335 & -0.149 & -0.152 & -0.312 & 0.051 & 0.063 & 0.165 & -0.145 & 0.052 & < 0.001 \\
-0.229 & 0.110 & 0.263 & -0.115 & 0.430 & 0.041 & 0.154 & -0.235 & -0.207 & 0.081 & 1.000 & -0.341 & < 0.001 & -0.241 & -0.099 & 0.358 & -0.759 & 0.410 & 0.427 & -0.070 & -0.168 & -0.286 \\
0.438 & -0.060 & -0.709 & -0.503 & -0.674 & -0.404 & -0.588 & 0.649 & 0.608 & 0.243 & -0.341 & 1.000 & 0.431 & 0.099 & -0.404 & 0.115 & 0.574 & 0.028 & 0.058 & -0.205 & 0.023 & 0.313 \\
0.129 & -0.082 & -0.213 & -0.311 & -0.225 & -0.062 & -0.462 & 0.373 & 0.253 & -0.335 & < 0.001 & 0.431 & 1.000 & -0.181 & -0.074 & 0.380 & 0.228 & -0.231 & -0.040 & -0.141 & -0.190 & 0.258 \\
0.158 & -0.471 & -0.112 & -0.042 & 0.206 & -0.302 & -0.094 & 0.048 & -0.149 & -0.241 & -0.099 & -0.404 & -0.181 & 1.000 & 0.091 & -0.271 & 0.254 & 0.141 & -0.344 & 0.346 & 0.542 & 0.105 \\
-0.258 & 0.110 & 0.669 & 0.622 & 0.449 & 0.308 & < 0.001 & -0.184 & -0.661 & -0.152 & -0.099 & -0.074 & -0.074 & 0.091 & 1.000 & -0.158 & -0.145 & -0.462 & -0.361 & 0.282 & 0.032 & 0.086 \\
-0.073 & -0.282 & -0.104 & -0.059 & 0.168 & 0.184 & -0.164 & 0.066 & -0.312 & 0.358 & 0.115 & 0.380 & -0.271 & -0.158 & 1.000 & -0.425 & 0.197 & 0.343 & -0.015 & -0.378 & -0.073 \\
0.362 & -0.031 & -0.511 & -0.232 & -0.629 & -0.380 & -0.324 & 0.279 & 0.458 & 0.051 & -0.759 & 0.574 & 0.228 & 0.254 & -0.145 & -0.425 & 1.000 & -0.367 & -0.450 & -0.178 & 0.248 & 0.386 \\
0.224 & -0.286 & -0.316 & -0.215 & 0.175 & -0.128 & 0.200 & 0.280 & 0.242 & 0.063 & 0.410 & 0.028 & -0.231 & 0.141 & -0.462 & 0.197 & -0.367 & 1.000 & 0.542 & -0.092 & 0.263 & -0.268 \\
0.047 & 0.030 & -0.198 & -0.150 & -0.091 & 0.167 & 0.208 & 0.382 & 0.147 & 0.165 & 0.427 & 0.058 & -0.040 & -0.344 & -0.361 & 0.343 & -0.450 & 0.542 & 1.000 & -0.268 & -0.103 & -0.280 \\
-0.061 & -0.170 & 0.435 & 0.181 & 0.414 & 0.220 & 0.018 & -0.346 & -0.204 & -0.145 & -0.070 & -0.205 & -0.141 & 0.346 & 0.282 & -0.015 & -0.178 & -0.092 & -0.268 & 1.000 & 0.211 & -0.164 \\
0.221 & -0.211 & -0.104 & 0.207 & 0.144 & -0.132 & 0.132 & -0.142 & 0.083 & 0.052 & -0.168 & 0.023 & -0.190 & 0.542 & 0.032 & -0.378 & 0.248 & 0.263 & -0.103 & 0.211 & 1.000 & -0.073 \\
0.200 & 0.064 & -0.141 & -0.161 & -0.196 & -0.358 & -0.447 & 0.048 & 0.136 & < 0.001 & -0.286 & 0.313 & 0.258 & 0.105 & 0.086 & -0.073 & 0.386 & -0.268 & -0.280 & -0.164 & -0.073 & 1.000
\end{pmatrix} \tag{A.2}
$$

$$\Sigma_{q,\text{AP}} = \begin{Bmatrix}
0.00300 & 0.00031 & -0.00078 & 0.00190 & 0.00077 & 0.00077 & 0.00049 & -0.00101 & 0.00005 & 0.00071 & -0.00067 & 0.00026 & 0.00010 & -0.00006 & -0.00034 & -0.00151 & 0.00036 & -0.00035 & -0.00047 & -0.00140 & -0.00011 & 0.00002 \\
0.00031 & 0.00188 & 0.00009 & 0.00135 & 0.00047 & 0.00037 & < 0.00001 & -0.00012 & 0.00005 & 0.00003 & -0.00036 & -0.00019 & 0.00049 & -0.00050 & -0.00003 & -0.00069 & 0.00036 & -0.00012 & -0.00024 & -0.00001 & -0.00006 & 0.00006 \\
-0.00078 & 0.00009 & 0.00613 & 0.00160 & 0.00195 & 0.00090 & -0.00100 & 0.00094 & -0.00075 & -0.00034 & -0.00142 & 0.00248 & -0.00006 & -0.00045 & -0.00063 & 0.00001 & -0.00011 & > -0.00001 & -0.00107 & -0.00065 & 0.00025 & 0.00004 \\
0.00190 & 0.00135 & 0.00160 & 0.00919 & 0.00092 & 0.00179 & 0.00064 & -0.00071 & -0.00056 & 0.00101 & -0.00235 & 0.00164 & -0.00005 & -0.00085 & -0.00082 & -0.00153 & 0.00051 & -0.00068 & -0.00069 & -0.00016 & -0.00024 & 0.00014 \\
0.00077 & 0.00047 & 0.00195 & 0.00092 & 0.00492 & 0.00047 & -0.00001 & -0.00015 & -0.00041 & -0.00218 & -0.00015 & 0.00020 & 0.00058 & -0.00004 & -0.00062 & -0.00150 & 0.00019 & 0.00036 & 0.00021 & -0.00047 & 0.00017 & -0.00024 \\
0.00077 & 0.00037 & 0.00090 & 0.00179 & 0.00047 & 0.00212 & 0.00016 & -0.00010 & -0.00032 & 0.00127 & -0.00209 & 0.00132 & -0.00018 & -0.00008 & -0.00013 & -0.00126 & 0.00003 & -0.00052 & -0.00066 & -0.00031 & 0.00043 & -0.00003 \\
0.00049 & < 0.00001 & -0.00100 & 0.00064 & -0.00001 & 0.00016 & 0.00154 & -0.00094 & > -0.00001 & -0.00033 & -0.00010 & -0.00122 & 0.00027 & 0.00009 & -0.00009 & -0.00010 & -0.00022 & -0.00041 & 0.00008 & 0.00117 & -0.00047 & -0.00013 \\
-0.00101 & -0.00012 & 0.00094 & -0.00071 & -0.00015 & -0.00010 & -0.00094 & 0.00347 & 0.00014 & -0.00034 & -0.00060 & -0.00013 & -0.00019 & -0.00005 & -0.00002 & 0.00037 & 0.00023 & 0.00031 & 0.00020 & -0.00274 & 0.00035 & 0.00014 \\
0.00005 & 0.00005 & -0.00075 & -0.00056 & -0.00041 & -0.00032 & > -0.00001 & 0.00014 & 0.00083 & 0.00010 & -0.00001 & -0.00022 & 0.00051 & -0.00003 & -0.00019 & 0.00023 & 0.00014 & 0.00024 & -0.00028 & 0.00003 & -0.00033 & 0.00006 \\
0.00071 & 0.00003 & -0.00034 & 0.00101 & -0.00218 & 0.00127 & -0.00033 & -0.00034 & 0.00010 & 0.00483 & -0.00142 & 0.00199 & -0.00080 & 0.00019 & -0.00022 & -0.00095 & 0.00032 & -0.00081 & -0.00077 & -0.00009 & 0.00101 & 0.00032 \\
-0.00067 & -0.00036 & -0.00142 & -0.00235 & -0.00015 & -0.00209 & -0.00010 & -0.00060 & -0.00001 & -0.00142 & 0.00445 & -0.00145 & < 0.00001 & 0.00004 & 0.00073 & 0.00050 & -0.00042 & 0.00020 & 0.00084 & 0.00081 & 0.00023 & > -0.00001 \\
0.00026 & -0.00019 & 0.00248 & 0.00164 & 0.00020 & 0.00132 & -0.00122 & -0.00013 & -0.00022 & 0.00199 & -0.00145 & 0.00547 & -0.00062 & -0.00015 & 0.00004 & -0.00044 & -0.00008 & -0.00008 & -0.00166 & -0.00078 & 0.00013 & -0.00017 \\
0.00010 & 0.00049 & -0.00006 & -0.00005 & -0.00018 & 0.00058 & 0.00027 & -0.00019 & 0.00051 & -0.00080 & < 0.00001 & -0.00062 & 0.00277 & -0.00031 & -0.00030 & -0.00001 & 0.00025 & 0.00072 & -0.00106 & 0.00100 & -0.00033 & 0.00015 \\
-0.00006 & -0.00050 & -0.00045 & -0.00085 & -0.00004 & -0.00008 & 0.00009 & -0.00005 & -0.00003 & 0.00019 & 0.00004 & -0.00015 & -0.00031 & 0.00066 & -0.00007 & 0.00015 & -0.00006 & 0.00010 & 0.00010 & 0.00007 & -0.00027 & -0.00007 \\
-0.00034 & -0.00003 & -0.00063 & -0.00082 & -0.00062 & -0.00013 & -0.00009 & -0.00002 & -0.00019 & -0.00022 & 0.00073 & 0.00004 & -0.00030 & -0.00007 & 0.00114 & -0.00003 & -0.00026 & -0.00024 & 0.00038 & 0.00060 & 0.00006 & -0.00010 \\
-0.00151 & -0.00069 & 0.00001 & -0.00153 & -0.00126 & -0.00010 & 0.00037 & 0.00023 & -0.00095 & 0.00023 & 0.00050 & -0.00044 & -0.00001 & 0.00015 & -0.00003 & 0.00281 & -0.00061 & 0.00067 & 0.00003 & -0.00025 & -0.00061 & -0.00010 \\
0.00036 & 0.00036 & -0.00011 & 0.00051 & 0.00019 & -0.00022 & 0.00023 & 0.00014 & 0.00032 & -0.00042 & -0.00008 & 0.00025 & -0.00006 & -0.00026 & -0.00061 & 0.00102 & -0.00002 & 0.00001 & -0.00016 & 0.00007 & 0.00009 \\
-0.00035 & -0.00012 & > -0.00001 & -0.00068 & 0.00036 & -0.00052 & -0.00041 & 0.00031 & 0.00024 & -0.00081 & 0.00072 & 0.00010 & 0.00024 & 0.00067 & -0.00002 & 0.00201 & -0.00012 & -0.00041 & -0.00041 & -0.00004 \\
-0.00047 & -0.00024 & -0.00107 & -0.00069 & 0.00021 & -0.00066 & 0.00008 & 0.00020 & -0.00028 & -0.00077 & 0.00084 & -0.00166 & -0.00106 & 0.00010 & 0.00038 & 0.00003 & 0.00001 & -0.00012 & 0.00289 & -0.00128 & 0.00050 & -0.00001 \\
-0.00140 & -0.00001 & -0.00065 & -0.00016 & -0.00047 & -0.00031 & 0.00117 & -0.00274 & 0.00003 & -0.00009 & 0.00081 & -0.00078 & 0.00100 & 0.00007 & 0.00060 & -0.00025 & -0.00016 & -0.00041 & -0.00128 & 0.01550 & 0.00009 & 0.00014 \\
-0.00011 & -0.00006 & 0.00025 & -0.00024 & 0.00017 & 0.00043 & -0.00047 & 0.00035 & -0.00033 & 0.00101 & 0.00023 & 0.00013 & -0.00033 & -0.00027 & 0.00006 & -0.00061 & 0.00007 & -0.00041 & 0.00050 & 0.00009 & 0.00612 & 0.00014 \\
0.00002 & 0.00006 & 0.00004 & 0.00014 & -0.00024 & -0.00003 & -0.00013 & 0.00014 & 0.00006 & 0.00032 & > -0.00001 & -0.00017 & 0.00015 & -0.00007 & -0.00010 & -0.00010 & 0.00009 & -0.00004 & -0.00001 & 0.00009 & 0.00014 & 0.00041
\end{Bmatrix} \tag{A.3}$$

$$\Sigma_{q,\text{AM}} = \begin{Bmatrix}
0.00020 & -0.00006 & -0.00015 & -0.00009 & -0.00006 & -0.00010 & -0.00008 & 0.00015 & 0.00025 & -0.00001 & -0.00008 & 0.00014 & 0.00003 & 0.00003 & -0.00006 & -0.00002 & 0.00015 & 0.00005 & 0.00001 & -0.00003 & 0.00006 & 0.00002 \\
-0.00006 & 0.00049 & 0.00011 & 0.00001 & -0.00013 & -0.00001 & > -0.00001 & -0.00022 & -0.00001 & 0.00026 & 0.00006 & -0.00003 & -0.00003 & -0.00014 & 0.00004 & -0.00012 & -0.00002 & -0.00010 & 0.00001 & -0.00013 & -0.00009 & 0.00001 \\
-0.00015 & 0.00011 & 0.00040 & 0.00020 & 0.00025 & 0.00019 & 0.00008 & -0.00045 & -0.00045 & -0.00004 & 0.00013 & -0.00032 & -0.00007 & -0.00003 & 0.00022 & -0.00004 & -0.00030 & -0.00010 & -0.00006 & 0.00030 & -0.00004 & -0.00002 \\
-0.00009 & 0.00001 & 0.00020 & 0.00031 & 0.00016 & 0.00019 & 0.00008 & -0.00021 & -0.00035 & -0.00003 & -0.00005 & -0.00020 & -0.00009 & -0.00001 & 0.00018 & -0.00002 & -0.00012 & -0.00006 & -0.00004 & 0.00011 & 0.00007 & -0.00002 \\
-0.00006 & 0.00025 & 0.00025 & 0.00016 & 0.00047 & 0.00039 & 0.00007 & -0.00033 & -0.00035 & -0.00017 & 0.00023 & -0.00008 & 0.00006 & -0.00008 & 0.00016 & 0.00007 & -0.00040 & 0.00006 & 0.00031 & 0.00006 & -0.00003 \\
-0.00010 & -0.00001 & 0.00019 & 0.00019 & 0.00009 & 0.00039 & 0.00013 & -0.00005 & -0.00027 & -0.00009 & 0.00002 & -0.00018 & -0.00002 & -0.00008 & 0.00010 & 0.00007 & -0.00022 & -0.00004 & 0.00005 & 0.00015 & -0.00005 & -0.00005 \\
> -0.00008 & -0.00001 & 0.00008 & 0.00007 & 0.00013 & 0.00013 & 0.00025 & -0.00009 & -0.00015 & -0.00003 & 0.00006 & -0.00021 & -0.00012 & -0.00002 & > -0.00001 & -0.00005 & -0.00015 & 0.00005 & 0.00005 & 0.00001 & 0.00004 & -0.00005 \\
0.00015 & -0.00022 & -0.00045 & -0.00021 & -0.00033 & -0.00005 & -0.00009 & 0.00086 & 0.00057 & -0.00007 & -0.00017 & 0.00043 & 0.00018 & -0.00005 & -0.00028 & 0.00027 & 0.00024 & 0.00013 & 0.00017 & -0.00035 & -0.00008 & 0.00001 \\
0.00025 & -0.00001 & -0.00045 & -0.00035 & -0.00035 & -0.00027 & -0.00015 & 0.00057 & 0.00098 & 0.00005 & -0.00016 & 0.00043 & 0.00013 & 0.00002 & -0.00034 & 0.00004 & 0.00042 & 0.00012 & 0.00007 & -0.00022 & 0.00005 & 0.00003 \\
-0.00001 & 0.00026 & -0.00004 & -0.00003 & -0.00017 & -0.00009 & -0.00003 & -0.00007 & 0.00005 & 0.00040 & 0.00004 & 0.00011 & -0.00011 & -0.00004 & -0.00005 & -0.00012 & 0.00003 & 0.00002 & 0.00005 & -0.00010 & 0.00002 & < 0.00001 \\
-0.00008 & 0.00006 & 0.00013 & -0.00005 & 0.00023 & 0.00002 & 0.00006 & -0.00017 & -0.00016 & 0.00004 & 0.00061 & -0.00019 & < 0.00001 & -0.00008 & -0.00004 & 0.00017 & -0.00055 & 0.00016 & 0.00016 & -0.00006 & -0.00008 & -0.00005 \\
0.00014 & -0.00003 & -0.00032 & -0.00020 & -0.00033 & -0.00018 & -0.00021 & 0.00043 & 0.00043 & 0.00011 & -0.00019 & 0.00051 & 0.00016 & 0.00003 & -0.00015 & 0.00005 & 0.00038 & 0.00001 & 0.00002 & -0.00016 & 0.00001 & 0.00005 \\
0.00003 & -0.00003 & -0.00007 & -0.00009 & -0.00008 & -0.00002 & -0.00012 & 0.00018 & 0.00013 & -0.00011 & < 0.00001 & 0.00016 & 0.00027 & -0.00004 & -0.00002 & 0.00012 & 0.00011 & -0.00006 & -0.00001 & -0.00008 & -0.00006 & 0.00003 \\
0.00003 & -0.00014 & -0.00003 & -0.00001 & 0.00006 & -0.00008 & -0.00002 & -0.00005 & 0.00002 & -0.00004 & -0.00008 & 0.00003 & -0.00004 & 0.00018 & 0.00002 & -0.00007 & 0.00010 & 0.00003 & -0.00007 & 0.00016 & 0.00014 & 0.00001 \\
-0.00006 & 0.00004 & 0.00022 & 0.00018 & 0.00016 & 0.00010 & > -0.00001 & -0.00028 & -0.00034 & -0.00005 & -0.00004 & -0.00015 & -0.00002 & 0.00002 & 0.00027 & -0.00005 & -0.00007 & -0.00012 & -0.00009 & 0.00016 & 0.00001 & 0.00001 \\
-0.00002 & -0.00012 & -0.00004 & -0.00002 & 0.00007 & 0.00007 & -0.00005 & 0.00027 & 0.00004 & -0.00012 & 0.00017 & 0.00005 & 0.00012 & -0.00007 & -0.00005 & 0.00037 & -0.00024 & 0.00006 & 0.00010 & -0.00001 & -0.00014 & -0.00001 \\
0.00015 & -0.00002 & -0.00030 & -0.00012 & -0.00040 & -0.00022 & -0.00015 & 0.00024 & 0.00042 & 0.00003 & -0.00055 & 0.00038 & 0.00011 & 0.00010 & -0.00007 & -0.00024 & 0.00086 & -0.00017 & -0.00020 & -0.00018 & 0.00014 & 0.00008 \\
0.00005 & -0.00010 & -0.00010 & -0.00006 & 0.00006 & -0.00004 & 0.00005 & 0.00013 & 0.00012 & 0.00002 & 0.00016 & 0.00001 & -0.00006 & 0.00003 & -0.00012 & 0.00006 & -0.00017 & 0.00025 & 0.00013 & -0.00005 & 0.00008 & -0.00003 \\
0.00001 & 0.00001 & -0.00006 & -0.00004 & -0.00003 & -0.00005 & 0.00007 & 0.00017 & 0.00005 & 0.00016 & 0.00002 & 0.00016 & 0.00007 & -0.00001 & -0.00009 & 0.00010 & -0.00020 & 0.00013 & 0.00023 & -0.00014 & -0.00003 & 0.00003 \\
-0.00003 & 0.00013 & -0.00001 & 0.00004 & 0.00031 & 0.00015 & 0.00001 & -0.00035 & -0.00022 & -0.00010 & -0.00006 & -0.00016 & -0.00008 & 0.00016 & 0.00016 & -0.00001 & -0.00018 & -0.00005 & -0.00014 & 0.00119 & 0.00014 & -0.00004 \\
0.00006 & -0.00009 & -0.00004 & 0.00007 & 0.00006 & -0.00005 & 0.00004 & -0.00008 & 0.00005 & -0.00002 & -0.00008 & 0.00001 & -0.00006 & 0.00014 & 0.00001 & -0.00014 & 0.00014 & 0.00008 & -0.00003 & 0.00014 & 0.00037 & -0.00001 \\
0.00002 & 0.00001 & -0.00002 & -0.00002 & -0.00003 & -0.00005 & -0.00005 & 0.00001 & 0.00003 & < 0.00001 & -0.00005 & 0.00005 & 0.00003 & 0.00001 & 0.00001 & -0.00001 & 0.00008 & -0.00003 & -0.00004 & -0.00001 & 0.00005
\end{Bmatrix} \tag{A.4}$$

$$\Omega_{q_{AP_{100}}} = \begin{pmatrix}
1.000 & 0.114 & 0.033 & 0.029 & -0.004 & -0.026 & -0.018 & -0.287 & 0.023 & 0.008 & -0.004 & 0.028 & 0.004 & -0.142 & 0.015 & -0.025 & 0.019 & 0.002 & 0.008 & -0.016 & 0.033 & 0.007 \\
0.114 & 1.000 & 0.036 & 0.014 & 0.011 & -0.019 & -0.035 & -0.311 & 0.044 & -0.002 & 0.016 & -0.009 & 0.004 & -0.164 & 0.011 & -0.023 & -0.008 & < 0.001 & -0.027 & -0.007 & 0.002 & 0.011 \\
0.033 & 0.036 & 1.000 & 0.264 & 0.051 & -0.009 & -0.002 & 0.010 & -0.352 & 0.012 & -0.008 & 0.013 & 0.016 & -0.163 & -0.213 & 0.028 & 0.015 & -0.013 & < 0.001 & 0.004 & -0.010 & 0.032 \\
0.029 & 0.014 & 0.264 & 1.000 & 0.037 & 0.025 & 0.045 & -0.004 & -0.354 & 0.043 & -0.003 & -0.003 & 0.016 & -0.185 & -0.208 & 0.018 & -0.038 & -0.024 & 0.010 & 0.018 & 0.019 & -0.020 \\
-0.004 & 0.011 & 0.051 & 0.037 & 1.000 & 0.047 & 0.020 & -0.021 & 0.032 & -0.449 & 0.022 & 0.001 & -0.006 & 0.024 & -0.200 & -0.267 & 0.032 & -0.015 & -0.012 & 0.017 & 0.011 & 0.028 \\
-0.026 & -0.019 & -0.009 & 0.025 & 0.047 & 1.000 & 0.052 & 0.011 & -0.026 & 0.034 & -0.385 & 0.022 & -0.044 & 0.015 & 0.037 & -0.193 & -0.213 & 0.058 & 0.019 & 0.022 & 0.013 & 0.017 \\
-0.018 & -0.035 & -0.002 & 0.045 & 0.020 & 0.052 & 1.000 & 0.024 & -0.016 & -0.041 & 0.003 & -0.323 & 0.016 & 0.002 & -0.021 & 0.049 & -0.174 & -0.250 & 0.021 & 0.009 & -0.010 & 0.002 \\
-0.287 & -0.311 & 0.010 & -0.004 & -0.021 & 0.011 & 0.024 & 1.000 & 0.018 & -0.006 & -0.005 & 0.010 & 0.029 & -0.097 & 0.041 & 0.020 & 0.005 & -0.009 & 0.023 & -0.008 & -0.024 & -0.034 \\
0.023 & 0.044 & -0.352 & -0.354 & 0.032 & -0.026 & -0.016 & 0.018 & 1.000 & 0.006 & 0.007 & -0.010 & -0.002 & -0.134 & -0.152 & 0.011 & 0.005 & 0.021 & 0.002 & < 0.001 & -0.003 & 0.013 \\
0.008 & -0.002 & 0.012 & 0.043 & -0.449 & 0.034 & -0.041 & -0.006 & 0.006 & 1.000 & 0.033 & -0.006 & -0.002 & -0.002 & -0.119 & -0.186 & 0.023 & 0.010 & 0.005 & -0.024 & 0.009 & -0.017 \\
-0.004 & 0.016 & -0.008 & -0.003 & 0.022 & -0.385 & 0.003 & -0.005 & 0.007 & 0.033 & 1.000 & 0.044 & 0.011 & -0.024 & 0.020 & -0.177 & -0.189 & 0.052 & -0.024 & -0.033 & -0.019 & -0.029 \\
0.028 & -0.009 & 0.013 & -0.003 & 0.001 & 0.022 & -0.323 & 0.010 & -0.010 & -0.006 & 0.044 & 1.000 & 0.018 & -0.030 & 0.038 & 0.015 & -0.190 & -0.208 & -0.001 & -0.034 & -0.003 & 0.023 \\
0.004 & 0.004 & 0.016 & 0.016 & -0.006 & -0.044 & 0.016 & 0.029 & -0.002 & -0.002 & 0.011 & 0.018 & 1.000 & -0.026 & -0.024 & 0.036 & 0.033 & -0.149 & -0.368 & -0.005 & -0.005 & -0.034 \\
-0.142 & -0.164 & -0.163 & -0.185 & 0.024 & 0.015 & 0.002 & -0.097 & -0.134 & -0.002 & -0.024 & -0.030 & -0.026 & 1.000 & -0.071 & 0.014 & 0.021 & 0.027 & -0.018 & 0.006 & -0.018 & -0.023 \\
0.015 & 0.011 & -0.213 & -0.208 & -0.200 & 0.037 & -0.021 & 0.041 & -0.152 & -0.119 & 0.020 & 0.038 & -0.024 & -0.071 & 1.000 & -0.081 & -0.020 & -0.005 & -0.014 & -0.013 & 0.005 & 0.007 \\
-0.025 & -0.023 & 0.028 & 0.018 & -0.267 & -0.193 & 0.049 & 0.020 & 0.011 & -0.186 & -0.177 & 0.015 & 0.036 & 0.014 & -0.081 & 1.000 & -0.088 & -0.032 & -0.001 & -0.020 & -0.010 & -0.005 \\
0.019 & -0.008 & 0.015 & -0.038 & 0.032 & -0.213 & -0.174 & 0.005 & 0.005 & 0.023 & -0.189 & -0.190 & 0.033 & 0.021 & -0.020 & -0.088 & 1.000 & -0.122 & -0.022 & 0.011 & -0.013 & 0.004 \\
0.002 & < 0.001 & -0.013 & -0.024 & -0.015 & 0.058 & -0.250 & -0.009 & 0.021 & 0.010 & 0.052 & -0.208 & -0.149 & 0.027 & -0.005 & -0.032 & -0.122 & 1.000 & -0.029 & 0.011 & -0.013 & 0.009 \\
0.008 & -0.027 & < 0.001 & 0.010 & -0.012 & 0.019 & 0.021 & 0.023 & 0.002 & 0.005 & -0.024 & -0.001 & -0.368 & -0.018 & -0.014 & -0.001 & -0.022 & -0.029 & 1.000 & -0.013 & 0.011 & 0.032 \\
-0.016 & -0.007 & 0.004 & 0.018 & 0.017 & 0.022 & 0.009 & -0.008 & < 0.001 & -0.024 & -0.033 & -0.034 & -0.005 & 0.006 & -0.013 & -0.020 & 0.011 & 0.011 & -0.013 & 1.000 & -0.038 & 0.019 \\
0.033 & 0.002 & -0.010 & 0.019 & 0.011 & 0.013 & -0.010 & -0.024 & -0.003 & 0.009 & -0.019 & -0.003 & -0.005 & -0.018 & 0.005 & -0.010 & -0.013 & -0.013 & 0.011 & -0.038 & 1.000 & 0.037 \\
0.007 & 0.011 & 0.032 & -0.020 & 0.028 & 0.017 & 0.002 & -0.034 & 0.013 & -0.017 & -0.029 & 0.023 & -0.034 & -0.023 & 0.007 & -0.005 & 0.004 & 0.009 & 0.032 & 0.019 & 0.037 & 1.000
\end{pmatrix} \tag{A.5}$$

$$\Omega_{q_{AM_{100}}} = \begin{pmatrix}
1.000 & 0.116 & 0.042 & 0.029 & 0.008 & 0.007 & 0.006 & -0.284 & -0.006 & -0.017 & -0.041 & 0.002 & -0.015 & -0.125 & 0.027 & -0.007 & -0.007 & 0.005 & -0.006 & 0.021 & -0.020 & 0.023 \\
0.116 & 1.000 & 0.041 & 0.058 & -0.018 & -0.004 & 0.005 & -0.317 & -0.021 & 0.015 & 0.003 & -0.013 & 0.009 & -0.132 & 0.007 & -0.011 & 0.009 & -0.006 & -0.011 & -0.008 & -0.007 & -0.010 \\
0.042 & 0.041 & 1.000 & 0.248 & 0.082 & -0.004 & 0.003 & 0.004 & -0.342 & 0.022 & 0.005 & -0.014 & < 0.001 & -0.208 & -0.237 & 0.002 & -0.009 & 0.002 & 0.005 & -0.015 & 0.011 & -0.005 \\
0.029 & 0.058 & 0.248 & 1.000 & 0.081 & 0.019 & 0.006 & -0.001 & -0.334 & 0.001 & -0.016 & -0.007 & -0.017 & -0.183 & -0.207 & 0.005 & -0.010 & 0.004 & -0.010 & 0.017 & 0.016 & 0.013 \\
0.008 & -0.018 & 0.082 & 0.081 & 1.000 & 0.040 & -0.023 & 0.014 & 0.010 & -0.432 & 0.032 & 0.010 & -0.029 & -0.014 & -0.203 & -0.245 & 0.016 & 0.022 & 0.010 & -0.004 & 0.012 & 0.010 \\
0.007 & -0.004 & -0.004 & 0.019 & 0.040 & 1.000 & 0.036 & < 0.001 & -0.021 & 0.026 & -0.400 & 0.035 & -0.008 & -0.007 & 0.003 & -0.198 & -0.225 & 0.038 & -0.005 & 0.002 & -0.020 & 0.010 \\
0.006 & 0.005 & 0.003 & 0.006 & -0.023 & 0.036 & 1.000 & 0.019 & -0.023 & < 0.001 & 0.032 & -0.364 & 0.035 & 0.005 & 0.003 & 0.043 & -0.207 & -0.250 & -0.013 & -0.028 & 0.013 & -0.011 \\
-0.284 & -0.317 & 0.004 & -0.001 & 0.014 & < 0.001 & 0.019 & 1.000 & 0.026 & -0.006 & 0.002 & -0.005 & 0.027 & -0.083 & 0.022 & -0.022 & -0.031 & -0.010 & 0.020 & -0.004 & -0.014 & -0.039 \\
-0.006 & -0.021 & -0.342 & -0.334 & 0.010 & -0.021 & -0.023 & 0.026 & 1.000 & 0.038 & -0.006 & 0.016 & -0.012 & -0.110 & -0.137 & 0.026 & -0.023 & 0.006 & -0.010 & -0.022 & -0.013 & 0.032 \\
-0.017 & 0.015 & 0.022 & 0.001 & -0.432 & 0.026 & < 0.001 & -0.006 & 0.038 & 1.000 & 0.030 & 0.004 & 0.050 & -0.002 & -0.128 & -0.151 & 0.015 & -0.019 & -0.014 & -0.006 & 0.029 & -0.017 \\
-0.041 & 0.003 & 0.005 & -0.016 & 0.032 & -0.400 & 0.032 & 0.002 & -0.006 & 0.030 & 1.000 & 0.030 & -0.013 & -0.005 & 0.011 & -0.169 & -0.152 & 0.019 & -0.021 & -0.007 & 0.006 & 0.013 \\
0.002 & -0.013 & -0.014 & -0.007 & 0.010 & 0.035 & -0.364 & -0.005 & 0.016 & 0.004 & 0.030 & 1.000 & 0.015 & -0.016 & -0.008 & 0.006 & -0.147 & -0.223 & 0.008 & 0.007 & 0.004 & 0.022 \\
-0.015 & 0.009 & < 0.001 & -0.017 & -0.029 & -0.008 & 0.035 & 0.027 & -0.012 & 0.050 & -0.013 & 0.015 & 1.000 & 0.022 & -0.035 & 0.020 & 0.013 & -0.136 & -0.352 & 0.004 & -0.005 & -0.023 \\
-0.125 & -0.132 & -0.208 & -0.183 & -0.014 & -0.007 & 0.005 & -0.083 & -0.110 & -0.002 & -0.005 & -0.016 & 0.022 & 1.000 & -0.058 & 0.040 & 0.044 & -0.009 & > -0.001 & -0.007 & 0.013 & -0.006 \\
0.027 & 0.007 & -0.237 & -0.203 & 0.003 & 0.003 & 0.022 & -0.137 & -0.128 & 0.011 & -0.008 & -0.035 & -0.058 & 1.000 & -0.101 & 0.024 & 0.001 & 0.037 & 0.016 & -0.023 & -0.022 \\
-0.007 & -0.011 & 0.002 & -0.245 & -0.198 & 0.043 & -0.022 & 0.026 & -0.151 & -0.169 & 0.006 & 0.020 & 0.040 & -0.101 & 1.000 & -0.084 & -0.001 & -0.002 & -0.035 & 0.003 & 0.012 \\
-0.007 & 0.009 & -0.009 & -0.010 & 0.016 & -0.225 & -0.207 & -0.031 & -0.023 & 0.015 & -0.152 & -0.147 & 0.013 & 0.044 & 0.024 & -0.084 & 1.000 & -0.141 & 0.029 & 0.014 & 0.021 & -0.012 \\
0.005 & -0.006 & 0.002 & 0.004 & 0.022 & 0.038 & -0.250 & -0.010 & 0.006 & -0.019 & 0.019 & -0.223 & -0.136 & -0.009 & 0.001 & -0.001 & -0.141 & 1.000 & 0.003 & 0.005 & -0.021 & 0.012 \\
-0.006 & -0.011 & 0.005 & -0.010 & 0.010 & -0.005 & -0.013 & 0.020 & -0.010 & -0.014 & -0.021 & 0.008 & -0.352 & > -0.001 & 0.037 & -0.002 & 0.029 & 0.003 & 1.000 & -0.007 & 0.020 & -0.002 \\
0.021 & -0.008 & -0.015 & 0.017 & -0.004 & 0.002 & -0.028 & -0.004 & -0.022 & -0.006 & -0.007 & 0.007 & 0.004 & -0.007 & 0.016 & -0.035 & 0.014 & 0.005 & -0.007 & 1.000 & 0.017 & -0.024 \\
-0.020 & -0.007 & 0.011 & 0.016 & 0.012 & -0.020 & 0.013 & -0.014 & -0.013 & 0.029 & 0.006 & 0.004 & -0.005 & 0.013 & -0.023 & 0.003 & 0.021 & -0.021 & 0.020 & 0.017 & 1.000 & 0.012 \\
0.023 & -0.010 & -0.005 & 0.013 & 0.010 & 0.010 & -0.011 & -0.039 & 0.032 & -0.017 & 0.013 & 0.022 & -0.023 & -0.006 & -0.022 & 0.012 & -0.012 & 0.012 & -0.002 & -0.024 & 0.012 & 1.000
\end{pmatrix} \tag{A.6}$$

$$\Sigma_{q_{AP_{100}}} = \begin{Bmatrix}
0.0101 & 0.0011 & 0.0005 & 0.0005 & -0.0001 & -0.0002 & -0.0001 & -0.0035 & 0.0002 & 0.0001 & -0.0001 & 0.0004 & 0.0001 & -0.0009 & 0.0001 & -0.0002 & 0.0002 & < 0.0001 & 0.0001 & -0.0004 & 0.0006 & < 0.0001 \\
0.0011 & 0.0085 & 0.0005 & 0.0002 & 0.0001 & -0.0001 & -0.0002 & -0.0035 & 0.0004 & > -0.0001 & 0.0002 & -0.0001 & 0.0001 & -0.0010 & 0.0001 & -0.0002 & -0.0001 & < 0.0001 & 0.0001 & -0.0002 & < 0.0001 & 0.0001 \\
0.0005 & 0.0005 & 0.0243 & 0.0066 & 0.0012 & -0.0001 & > -0.0001 & 0.0002 & -0.0059 & 0.0002 & -0.0002 & 0.0003 & 0.0004 & -0.0016 & -0.0024 & 0.0004 & 0.0002 & -0.0002 & < 0.0001 & 0.0001 & -0.0003 & 0.0003 \\
0.0005 & 0.0002 & 0.0066 & 0.0253 & 0.0009 & 0.0003 & 0.0005 & -0.0001 & -0.0060 & 0.0009 & -0.0001 & -0.0001 & 0.0004 & -0.0019 & -0.0023 & 0.0003 & -0.0006 & -0.0004 & 0.0001 & 0.0006 & 0.0005 & -0.0002 \\
-0.0001 & 0.0001 & 0.0012 & 0.0009 & 0.0225 & 0.0005 & 0.0002 & -0.0004 & 0.0005 & -0.0087 & 0.0005 & < 0.0001 & -0.0001 & 0.0002 & -0.0021 & -0.0036 & 0.0004 & -0.0003 & -0.0002 & 0.0006 & 0.0003 & 0.0002 \\
-0.0002 & -0.0001 & -0.0001 & 0.0003 & 0.0005 & 0.0058 & 0.0003 & 0.0001 & -0.0002 & 0.0003 & -0.0041 & 0.0002 & -0.0005 & 0.0001 & 0.0002 & -0.0013 & -0.0015 & 0.0005 & 0.0001 & 0.0004 & 0.0002 & 0.0001 \\
-0.0001 & -0.0002 & > -0.0001 & 0.0005 & 0.0002 & 0.0003 & 0.0057 & 0.0002 & -0.0001 & -0.0004 & < 0.0001 & -0.0034 & 0.0002 & < 0.0001 & -0.0001 & 0.0003 & -0.0012 & -0.0022 & 0.0002 & 0.0002 & -0.0001 & < 0.0001 \\
-0.0035 & -0.0035 & 0.0002 & -0.0001 & -0.0004 & 0.0001 & 0.0002 & 0.0147 & 0.0002 & -0.0001 & -0.0001 & 0.0002 & 0.0005 & -0.0008 & 0.0004 & 0.0002 & 0.0001 & -0.0001 & 0.0003 & -0.0002 & -0.0005 & -0.0002 \\
0.0002 & 0.0004 & -0.0059 & -0.0060 & 0.0005 & -0.0002 & -0.0001 & 0.0002 & 0.0115 & 0.0001 & 0.0001 & -0.0001 & > -0.0001 & -0.0009 & -0.0012 & 0.0001 & < 0.0001 & 0.0003 & < 0.0001 & < 0.0001 & > -0.0001 & 0.0001 \\
0.0001 & > -0.0001 & 0.0002 & 0.0009 & -0.0087 & 0.0003 & -0.0004 & -0.0001 & 0.0001 & 0.0165 & 0.0006 & -0.0001 & > -0.0001 & > -0.0001 & -0.0011 & -0.0022 & -0.0024 & 0.0001 & 0.0001 & -0.0007 & 0.0002 & -0.0001 \\
-0.0001 & 0.0002 & -0.0002 & -0.0001 & 0.0005 & -0.0041 & < 0.0001 & -0.0001 & 0.0001 & 0.0006 & 0.0193 & 0.0009 & 0.0002 & -0.0002 & 0.0002 & -0.0022 & -0.0024 & 0.0008 & -0.0003 & -0.0010 & -0.0004 & -0.0002 \\
0.0004 & -0.0001 & 0.0003 & -0.0001 & < 0.0001 & 0.0002 & -0.0034 & 0.0002 & -0.0001 & -0.0001 & 0.0009 & 0.0192 & 0.0004 & -0.0003 & 0.0004 & 0.0002 & -0.0024 & -0.0033 & > -0.0001 & -0.0010 & -0.0001 & 0.0002 \\
0.0001 & 0.0001 & 0.0004 & 0.0004 & -0.0001 & -0.0005 & 0.0002 & 0.0005 & > -0.0001 & > -0.0001 & 0.0002 & 0.0004 & 0.0227 & -0.0002 & -0.0003 & 0.0005 & 0.0004 & -0.0026 & -0.0052 & -0.0002 & -0.0001 & -0.0003 \\
-0.0009 & -0.0010 & -0.0016 & -0.0019 & 0.0002 & 0.0001 & < 0.0001 & -0.0008 & > -0.0009 & > -0.0001 & -0.0002 & -0.0003 & -0.0002 & 0.0041 & -0.0003 & 0.0001 & 0.0001 & 0.0002 & -0.0001 & 0.0001 & -0.0002 & -0.0001 \\
0.0001 & 0.0001 & -0.0024 & -0.0023 & -0.0021 & 0.0002 & -0.0001 & 0.0004 & -0.0012 & -0.0011 & 0.0002 & 0.0004 & -0.0003 & -0.0003 & 0.0050 & -0.0005 & -0.0001 & > -0.0001 & -0.0001 & -0.0002 & 0.0001 & < 0.0001 \\
-0.0002 & -0.0002 & 0.0004 & 0.0003 & -0.0036 & -0.0013 & 0.0003 & 0.0002 & 0.0001 & -0.0022 & -0.0022 & 0.0002 & 0.0005 & 0.0001 & -0.0005 & 0.0081 & -0.0007 & -0.0003 & > -0.0001 & -0.0004 & -0.0002 & > -0.0001 \\
0.0002 & -0.0001 & 0.0002 & -0.0006 & 0.0004 & -0.0015 & -0.0012 & -0.0001 & < 0.0001 & 0.0003 & -0.0024 & 0.0004 & 0.0001 & -0.0001 & -0.0001 & -0.0007 & 0.0084 & -0.0013 & -0.0002 & 0.0002 & -0.0002 & < 0.0001 \\
< 0.0001 & < 0.0001 & -0.0002 & -0.0004 & -0.0003 & 0.0005 & -0.0022 & -0.0001 & 0.0003 & 0.0001 & 0.0008 & -0.0033 & > -0.0001 & -0.0026 & 0.0004 & 0.0002 & -0.0001 & -0.0003 & 0.0013 & -0.0003 & 0.0003 & 0.0001 \\
0.0001 & < 0.0001 & 0.0001 & 0.0001 & -0.0002 & 0.0001 & 0.0002 & 0.0003 & < 0.0001 & 0.0001 & -0.0003 & -0.0001 & > -0.0001 & -0.0002 & -0.0001 & -0.0001 & > -0.0001 & -0.0002 & -0.0002 & -0.0003 & 0.0089 & -0.0003 & 0.0002 \\
-0.0004 & -0.0002 & 0.0001 & 0.0006 & 0.0004 & 0.0002 & -0.0001 & -0.0005 & < 0.0001 & -0.0007 & -0.0010 & -0.0002 & 0.0001 & -0.0002 & 0.0004 & -0.0002 & -0.0002 & 0.0003 & -0.0003 & 0.0002 & 0.0497 & -0.0015 & 0.0002 \\
0.0006 & < 0.0001 & -0.0003 & 0.0005 & 0.0003 & 0.0002 & -0.0001 & -0.0005 & > -0.0001 & -0.0004 & -0.0001 & -0.0001 & -0.0002 & 0.0001 & -0.0002 & -0.0002 & -0.0002 & -0.0003 & 0.0002 & -0.0015 & 0.0298 & 0.0003 \\
< 0.0001 & 0.0001 & 0.0003 & -0.0002 & 0.0002 & 0.0001 & < 0.0001 & -0.0002 & 0.0001 & -0.0001 & -0.0002 & 0.0002 & -0.0003 & -0.0001 & < 0.0001 & > -0.0001 & < 0.0001 & 0.0001 & 0.0002 & 0.0002 & 0.0003 & 0.0029
\end{Bmatrix} \tag{A.7}$$

$$\Sigma_{q_{AM_{100}}} = \begin{Bmatrix}
0.0098 & 0.0010 & 0.0007 & 0.0005 & 0.0001 & 0.0001 & < 0.0001 & -0.0035 & -0.0001 & -0.0002 & -0.0006 & < 0.0001 & -0.0002 & -0.0008 & 0.0002 & -0.0001 & -0.0001 & 0.0001 & -0.0001 & 0.0005 & -0.0004 & 0.0001 \\
0.0010 & 0.0083 & 0.0006 & 0.0008 & -0.0002 & > -0.0001 & < 0.0001 & -0.0036 & -0.0002 & 0.0002 & < 0.0001 & -0.0002 & 0.0001 & -0.0008 & < 0.0001 & -0.0001 & 0.0001 & -0.0001 & -0.0002 & -0.0001 & -0.0001 \\
0.0007 & 0.0006 & 0.0250 & 0.0062 & 0.0019 & -0.0001 & < 0.0001 & 0.0001 & -0.0054 & 0.0005 & 0.0001 & -0.0003 & < 0.0001 & -0.0021 & -0.0027 & < 0.0001 & -0.0001 & < 0.0001 & -0.0001 & -0.0005 & 0.0003 & > -0.0001 \\
0.0005 & 0.0008 & 0.0062 & 0.0248 & 0.0019 & 0.0002 & 0.0001 & > -0.0001 & -0.0053 & < 0.0001 & -0.0004 & -0.0002 & -0.0004 & -0.0018 & -0.0024 & 0.0001 & -0.0001 & 0.0001 & -0.0001 & 0.0006 & 0.0005 & 0.0001 \\
0.0001 & -0.0002 & 0.0019 & 0.0019 & 0.0227 & 0.0005 & -0.0003 & 0.0003 & 0.0002 & -0.0085 & 0.0007 & 0.0002 & -0.0006 & -0.0001 & -0.0022 & -0.0031 & 0.0002 & 0.0004 & > -0.0001 & -0.0001 & -0.0003 & < 0.0001 \\
> -0.0001 & -0.0001 & 0.0002 & 0.0005 & 0.0064 & 0.0002 & < 0.0001 & -0.0002 & 0.0003 & -0.0044 & 0.0004 & -0.0001 & > -0.0001 & < 0.0001 & -0.0001 & -0.0013 & -0.0016 & 0.0004 & > -0.0001 & < 0.0001 & -0.0003 & < 0.0001 \\
< 0.0001 & < 0.0001 & 0.0001 & -0.0003 & 0.0002 & 0.0059 & 0.0002 & -0.0002 & < 0.0001 & 0.0003 & -0.0039 & 0.0004 & < 0.0001 & < 0.0001 & 0.0003 & -0.0014 & -0.0022 & -0.0001 & > -0.0005 & 0.0002 & > -0.0001 \\
-0.0035 & -0.0036 & 0.0001 & > -0.0001 & 0.0003 & < 0.0001 & 0.0002 & 0.0153 & 0.0003 & -0.0001 & < 0.0001 & -0.0001 & 0.0005 & -0.0006 & 0.0002 & -0.0002 & -0.0003 & -0.0001 & 0.0002 & -0.0001 & -0.0003 & -0.0003 \\
-0.0001 & -0.0002 & -0.0054 & -0.0053 & 0.0002 & -0.0002 & -0.0002 & 0.0003 & 0.0101 & 0.0005 & -0.0001 & 0.0002 & -0.0002 & -0.0007 & -0.0010 & 0.0002 & -0.0002 & 0.0001 & -0.0001 & -0.0005 & -0.0002 & 0.0002 \\
-0.0002 & 0.0002 & 0.0005 & < 0.0001 & -0.0085 & 0.0003 & < 0.0001 & -0.0001 & 0.0005 & 0.0169 & 0.0005 & 0.0001 & 0.0009 & > -0.0001 & -0.0012 & -0.0017 & 0.0002 & -0.0003 & -0.0002 & -0.0002 & 0.0007 & -0.0001 \\
-0.0006 & < 0.0001 & 0.0001 & -0.0004 & 0.0007 & -0.0044 & 0.0003 & < 0.0001 & -0.0001 & 0.0005 & 0.0193 & 0.0006 & -0.0003 & > -0.0001 & 0.0001 & -0.0020 & -0.0019 & 0.0003 & -0.0003 & -0.0002 & 0.0001 & 0.0001 \\
< 0.0001 & -0.0002 & -0.0003 & -0.0002 & 0.0002 & 0.0004 & -0.0039 & -0.0001 & 0.0002 & 0.0001 & 0.0006 & 0.0194 & 0.0003 & -0.0001 & -0.0001 & 0.0001 & -0.0018 & -0.0036 & 0.0001 & 0.0002 & 0.0001 & 0.0002 \\
-0.0002 & 0.0001 & < 0.0001 & -0.0004 & -0.0006 & -0.0001 & 0.0004 & 0.0005 & -0.0002 & 0.0009 & -0.0003 & 0.0003 & 0.0202 & 0.0002 & -0.0004 & 0.0002 & 0.0002 & -0.0023 & -0.0047 & 0.0001 & -0.0001 & -0.0002 \\
-0.0008 & -0.0008 & -0.0021 & -0.0018 & -0.0001 & > -0.0001 & < 0.0001 & -0.0006 & -0.0007 & > -0.0001 & > -0.0001 & -0.0001 & 0.0002 & 0.0040 & -0.0003 & 0.0002 & 0.0003 & -0.0001 & > -0.0001 & -0.0001 & 0.0001 & > -0.0001 \\
0.0002 & < 0.0001 & -0.0027 & -0.0024 & -0.0022 & < 0.0001 & < 0.0001 & 0.0002 & -0.0010 & -0.0012 & 0.0001 & -0.0001 & -0.0004 & -0.0003 & 0.0052 & -0.0006 & 0.0002 & < 0.0001 & 0.0003 & 0.0003 & -0.0003 & -0.0001 \\
-0.0001 & -0.0001 & < 0.0001 & 0.0001 & -0.0031 & -0.0013 & 0.0003 & -0.0002 & 0.0002 & -0.0017 & -0.0020 & 0.0001 & 0.0002 & 0.0002 & -0.0006 & 0.0072 & -0.0006 & > -0.0001 & < 0.0001 & -0.0007 & 0.0001 & 0.0001 \\
-0.0001 & -0.0001 & -0.0001 & -0.0001 & 0.0002 & -0.0016 & -0.0014 & -0.0003 & -0.0002 & 0.0002 & -0.0019 & -0.0018 & 0.0002 & 0.0003 & 0.0002 & -0.0006 & 0.0081 & -0.0015 & 0.0002 & 0.0003 & 0.0003 & -0.0001 \\
0.0001 & -0.0001 & < 0.0001 & 0.0001 & 0.0004 & 0.0004 & -0.0022 & -0.0001 & 0.0001 & -0.0002 & -0.0003 & 0.0003 & 0.0003 & 0.0001 & -0.0002 & -0.0001 & > -0.0001 & -0.0015 & < 0.0001 & 0.0001 & -0.0005 & 0.0001 \\
-0.0001 & -0.0001 & -0.0001 & 0.0001 & 0.0001 & > -0.0001 & 0.0002 & -0.0001 & -0.0005 & -0.0002 & -0.0002 & 0.0001 & -0.0002 & -0.0001 & -0.0047 & > -0.0001 & -0.0001 & 0.0002 & < 0.0001 & 0.0087 & -0.0001 & > -0.0001 \\
-0.0005 & -0.0002 & -0.0005 & 0.0006 & -0.0001 & < 0.0001 & -0.0005 & -0.0001 & -0.0005 & -0.0002 & -0.0002 & 0.0002 & 0.0001 & -0.0001 & 0.0003 & -0.0007 & 0.0003 & 0.0001 & -0.0001 & 0.0512 & 0.0007 & -0.0003 \\
-0.0004 & -0.0001 & 0.0003 & 0.0005 & 0.0003 & -0.0003 & 0.0002 & -0.0003 & -0.0002 & 0.0007 & 0.0001 & 0.0001 & -0.0001 & 0.0001 & -0.0003 & < 0.0001 & 0.0003 & -0.0005 & 0.0003 & 0.0003 & 0.0336 & 0.0001 \\
0.0001 & -0.0001 & > -0.0001 & 0.0001 & 0.0001 & < 0.0001 & > -0.0001 & -0.0003 & 0.0002 & -0.0001 & 0.0001 & 0.0002 & > -0.0002 & > -0.0001 & -0.0001 & 0.0001 & -0.0001 & > -0.0001 & -0.0003 & 0.0001 & 0.0031
\end{Bmatrix} \tag{A.8}$$

$$\Omega_{m=5,AF_{100}} = \begin{pmatrix}
1.000 & 0.118 & 0.033 & 0.029 & -0.007 & -0.027 & -0.022 & -0.288 & 0.019 & 0.009 & -0.003 & 0.027 & 0.005 & -0.140 & 0.019 & -0.026 & 0.018 & 0.004 & 0.008 & -0.015 & 0.036 & 0.010 \\
0.118 & 1.000 & 0.036 & 0.014 & 0.009 & -0.019 & -0.035 & -0.312 & 0.043 & -0.003 & 0.016 & -0.009 & 0.004 & -0.169 & 0.014 & -0.021 & -0.012 & -0.001 & -0.027 & -0.007 & 0.005 & 0.014 \\
0.033 & 0.036 & 1.000 & 0.266 & 0.049 & -0.006 & -0.001 & 0.007 & -0.355 & 0.011 & -0.007 & 0.015 & 0.014 & -0.160 & -0.213 & 0.028 & 0.012 & -0.014 & 0.003 & 0.004 & -0.009 & 0.033 \\
0.029 & 0.014 & 0.266 & 1.000 & 0.034 & 0.026 & 0.044 & -0.005 & -0.354 & 0.041 & -0.003 & -0.004 & 0.015 & -0.185 & -0.208 & 0.025 & -0.038 & -0.024 & 0.010 & 0.016 & 0.017 & -0.024 \\
-0.007 & 0.009 & 0.049 & 0.034 & 1.000 & 0.049 & 0.022 & -0.020 & 0.031 & -0.451 & 0.023 & 0.003 & -0.009 & 0.024 & -0.198 & -0.268 & 0.029 & -0.012 & -0.011 & 0.018 & 0.017 & 0.027 \\
-0.027 & -0.019 & -0.006 & 0.026 & 0.049 & 1.000 & 0.054 & 0.013 & -0.027 & 0.033 & -0.386 & 0.020 & -0.045 & 0.016 & 0.039 & -0.194 & -0.206 & 0.056 & 0.017 & 0.021 & 0.017 & 0.019 \\
-0.022 & -0.035 & -0.001 & 0.044 & 0.022 & 0.054 & 1.000 & 0.028 & -0.015 & -0.040 & 0.003 & -0.324 & 0.016 & < 0.001 & -0.021 & 0.048 & -0.175 & -0.247 & 0.023 & 0.007 & -0.010 & 0.003 \\
-0.288 & -0.312 & 0.007 & -0.005 & -0.020 & 0.013 & 0.028 & 1.000 & 0.021 & -0.006 & -0.007 & 0.009 & 0.029 & -0.096 & 0.039 & 0.022 & 0.005 & -0.009 & 0.025 & -0.009 & -0.024 & -0.036 \\
0.019 & 0.043 & -0.355 & -0.354 & 0.031 & -0.027 & -0.015 & 0.021 & 1.000 & 0.006 & 0.009 & -0.010 & -0.003 & -0.134 & -0.148 & 0.010 & 0.004 & 0.024 & > -0.001 & < 0.001 & < 0.001 & 0.013 \\
0.009 & -0.003 & 0.011 & 0.041 & -0.451 & 0.033 & -0.040 & -0.006 & 0.006 & 1.000 & 0.034 & -0.007 & > -0.001 & > -0.001 & -0.118 & -0.184 & 0.025 & 0.008 & 0.002 & -0.027 & 0.009 & -0.017 \\
-0.003 & 0.016 & -0.007 & -0.003 & 0.023 & -0.386 & 0.003 & -0.007 & 0.009 & 0.034 & 1.000 & 0.046 & 0.012 & -0.023 & 0.017 & -0.175 & -0.192 & 0.051 & -0.024 & -0.035 & -0.017 & -0.030 \\
0.027 & -0.009 & 0.015 & -0.004 & 0.003 & 0.020 & -0.324 & 0.009 & -0.010 & -0.007 & 0.046 & 1.000 & 0.016 & -0.030 & 0.037 & 0.016 & -0.189 & -0.206 & -0.003 & -0.036 & 0.001 & 0.018 \\
0.005 & 0.004 & 0.014 & 0.015 & -0.009 & -0.045 & 0.016 & 0.029 & -0.003 & > -0.001 & 0.012 & 0.016 & 1.000 & -0.021 & -0.022 & 0.037 & 0.033 & -0.150 & -0.367 & -0.007 & -0.005 & -0.032 \\
-0.140 & -0.169 & -0.160 & -0.185 & 0.024 & 0.016 & < 0.001 & -0.096 & -0.134 & > -0.001 & -0.023 & -0.030 & -0.021 & 1.000 & -0.074 & 0.011 & 0.022 & 0.028 & -0.018 & 0.008 & -0.020 & -0.024 \\
0.019 & 0.014 & -0.213 & -0.208 & -0.198 & 0.039 & -0.021 & 0.039 & -0.148 & -0.118 & 0.017 & 0.037 & -0.022 & -0.074 & 1.000 & -0.083 & -0.018 & -0.006 & -0.015 & -0.011 & 0.004 & 0.011 \\
-0.026 & -0.021 & 0.028 & 0.025 & -0.268 & -0.194 & 0.048 & 0.022 & 0.010 & -0.184 & -0.175 & 0.016 & 0.037 & 0.011 & -0.083 & 1.000 & -0.091 & -0.030 & -0.002 & -0.020 & -0.014 & -0.009 \\
0.018 & -0.012 & 0.012 & -0.038 & 0.029 & -0.206 & -0.175 & 0.005 & 0.004 & 0.025 & -0.192 & -0.189 & 0.033 & 0.022 & -0.018 & -0.091 & 1.000 & -0.123 & -0.021 & 0.011 & -0.017 & 0.006 \\
0.004 & -0.001 & -0.014 & -0.024 & -0.012 & 0.056 & -0.247 & -0.009 & 0.024 & 0.008 & 0.051 & -0.206 & -0.150 & 0.028 & -0.006 & -0.030 & -0.123 & 1.000 & -0.027 & 0.017 & -0.012 & 0.009 \\
0.008 & -0.027 & 0.003 & 0.010 & -0.011 & 0.017 & 0.023 & 0.025 & > -0.001 & 0.002 & -0.024 & -0.003 & -0.367 & -0.018 & -0.015 & -0.002 & -0.021 & -0.027 & 1.000 & -0.012 & 0.012 & 0.032 \\
-0.015 & -0.007 & 0.004 & 0.016 & 0.018 & 0.021 & 0.007 & -0.009 & < 0.001 & -0.027 & -0.035 & -0.036 & -0.007 & 0.008 & -0.011 & -0.020 & 0.011 & 0.017 & -0.012 & 1.000 & -0.036 & 0.020 \\
0.036 & 0.005 & -0.009 & 0.017 & 0.017 & 0.017 & -0.010 & -0.024 & < 0.001 & 0.009 & -0.017 & 0.001 & -0.005 & -0.020 & 0.004 & -0.014 & -0.017 & -0.012 & 0.012 & -0.036 & 1.000 & 0.039 \\
0.010 & 0.014 & 0.033 & -0.024 & 0.027 & 0.019 & 0.003 & -0.036 & 0.013 & -0.017 & -0.030 & 0.018 & -0.032 & -0.024 & 0.011 & -0.009 & 0.006 & 0.009 & 0.032 & 0.020 & 0.039 & 1.000
\end{pmatrix} \tag{A.9}$$

$$\Omega_{m=5,AM_{100}} = \begin{pmatrix}
1.000 & 0.114 & 0.044 & 0.030 & 0.008 & 0.007 & 0.008 & −0.281 & −0.005 & −0.017 & −0.043 & 0.001 & −0.013 & −0.126 & 0.025 & −0.006 & −0.006 & 0.004 & −0.007 & 0.021 & −0.022 & 0.023 \\
0.114 & 1.000 & 0.042 & 0.058 & −0.016 & 0.001 & 0.006 & −0.318 & −0.021 & 0.015 & 0.002 & −0.014 & 0.007 & −0.130 & 0.005 & −0.010 & 0.006 & −0.008 & −0.007 & −0.008 & −0.007 & −0.011 \\
0.044 & 0.042 & 1.000 & 0.250 & 0.082 & −0.002 & 0.002 & 0.003 & −0.345 & 0.022 & 0.005 & −0.012 & −0.002 & −0.206 & −0.236 & > −0.001 & −0.010 & < 0.001 & 0.007 & −0.016 & 0.009 & −0.008 \\
0.030 & 0.058 & 0.250 & 1.000 & 0.078 & 0.021 & 0.005 & > −0.001 & −0.335 & 0.003 & −0.016 & −0.008 & −0.018 & −0.182 & −0.208 & 0.004 & −0.006 & 0.004 & −0.009 & 0.015 & 0.016 & 0.014 \\
0.008 & −0.016 & 0.082 & 0.078 & 1.000 & 0.042 & −0.022 & 0.015 & 0.012 & −0.432 & 0.031 & 0.011 & −0.026 & −0.015 & −0.206 & −0.245 & 0.016 & 0.019 & 0.008 & −0.006 & 0.011 & 0.012 \\
0.007 & 0.001 & −0.002 & 0.021 & 0.042 & 1.000 & 0.036 & −0.001 & −0.023 & 0.025 & −0.399 & 0.036 & −0.010 & −0.007 & 0.001 & −0.198 & −0.226 & 0.035 & −0.005 & 0.002 & −0.020 & 0.011 \\
0.008 & 0.006 & 0.002 & 0.005 & −0.022 & 0.036 & 1.000 & 0.018 & −0.023 & < 0.001 & 0.032 & −0.364 & 0.037 & 0.003 & 0.005 & 0.040 & −0.207 & −0.251 & −0.015 & −0.025 & 0.011 & −0.009 \\
−0.281 & −0.318 & 0.003 & > −0.001 & 0.015 & −0.001 & 0.018 & 1.000 & 0.025 & −0.007 & 0.002 & −0.003 & 0.027 & −0.081 & 0.021 & −0.024 & −0.030 & −0.009 & 0.020 & −0.002 & −0.014 & −0.038 \\
−0.005 & −0.021 & −0.345 & −0.335 & 0.012 & −0.023 & −0.023 & 0.025 & 1.000 & 0.022 & 0.005 & −0.012 & −0.002 & −0.206 & −0.236 & > −0.001 & −0.010 & < 0.001 & 0.007 & −0.016 & 0.009 & −0.008 \\
−0.017 & 0.015 & 0.022 & 0.003 & −0.432 & 0.025 & < 0.001 & −0.007 & 0.038 & 1.000 & 0.027 & 0.002 & 0.047 & 0.001 & −0.126 & −0.151 & 0.016 & −0.017 & −0.017 & −0.004 & 0.026 & −0.018 \\
−0.043 & 0.002 & 0.005 & −0.016 & 0.031 & −0.399 & 0.032 & 0.002 & −0.006 & 0.027 & 1.000 & 0.030 & −0.017 & −0.005 & 0.014 & −0.170 & −0.151 & 0.019 & −0.019 & −0.008 & 0.004 & 0.011 \\
0.001 & −0.014 & −0.012 & −0.008 & 0.011 & 0.036 & −0.364 & −0.003 & 0.017 & 0.002 & 0.030 & 1.000 & 0.013 & −0.016 & −0.007 & 0.008 & −0.147 & −0.221 & 0.008 & 0.005 & 0.005 & 0.020 \\
−0.013 & 0.007 & −0.002 & −0.018 & −0.026 & −0.010 & 0.037 & 0.027 & −0.011 & 0.047 & −0.017 & 0.013 & 1.000 & 0.024 & −0.035 & 0.019 & 0.013 & −0.133 & −0.354 & 0.004 & −0.004 & −0.025 \\
−0.126 & −0.130 & −0.206 & −0.182 & −0.015 & −0.007 & 0.003 & −0.081 & −0.110 & 0.001 & −0.005 & −0.016 & 0.024 & 1.000 & −0.058 & 0.042 & 0.043 & −0.007 & −0.003 & −0.009 & 0.015 & −0.008 \\
0.025 & 0.005 & −0.236 & −0.208 & −0.206 & 0.001 & 0.005 & 0.021 & −0.135 & −0.126 & 0.014 & −0.007 & −0.035 & −0.058 & 1.000 & −0.099 & 0.023 & 0.003 & 0.037 & 0.018 & −0.022 & −0.021 \\
−0.006 & −0.010 & > −0.001 & 0.004 & −0.245 & −0.198 & 0.040 & −0.024 & 0.026 & −0.151 & −0.170 & 0.008 & 0.019 & 0.042 & −0.099 & 1.000 & −0.083 & −0.002 & > −0.001 & −0.036 & 0.005 & 0.011 \\
−0.006 & 0.006 & −0.010 & −0.006 & 0.016 & −0.226 & −0.207 & −0.030 & −0.022 & 0.016 & −0.151 & −0.147 & 0.013 & 0.043 & 0.023 & −0.083 & 1.000 & −0.141 & 0.031 & 0.013 & 0.023 & −0.012 \\
0.004 & −0.008 & < 0.001 & 0.004 & 0.019 & 0.035 & −0.251 & −0.009 & 0.007 & −0.017 & 0.019 & −0.221 & −0.133 & −0.007 & 0.003 & −0.002 & −0.141 & 1.000 & 0.005 & 0.004 & −0.022 & 0.012 \\
−0.007 & −0.007 & 0.007 & −0.009 & 0.008 & −0.005 & −0.015 & 0.020 & −0.010 & −0.017 & −0.019 & 0.008 & −0.354 & −0.003 & 0.037 & > −0.001 & 0.031 & 0.005 & 1.000 & −0.006 & 0.023 & 0.003 \\
0.021 & −0.008 & −0.016 & 0.015 & −0.006 & 0.002 & −0.025 & −0.002 & −0.022 & −0.004 & −0.008 & 0.005 & 0.004 & −0.009 & 0.018 & −0.036 & 0.013 & 0.004 & −0.006 & 1.000 & 0.017 & −0.023 \\
−0.022 & −0.007 & 0.009 & 0.016 & 0.011 & −0.020 & 0.011 & −0.014 & −0.016 & 0.026 & 0.004 & 0.005 & −0.004 & 0.015 & −0.022 & 0.005 & 0.023 & −0.022 & 0.023 & 0.017 & 1.000 & 0.015 \\
0.023 & −0.011 & −0.008 & 0.014 & 0.012 & 0.011 & −0.009 & −0.038 & 0.033 & −0.018 & 0.011 & 0.020 & −0.025 & −0.008 & −0.021 & 0.011 & −0.012 & 0.012 & 0.003 & −0.023 & 0.015 & 1.000
\end{pmatrix} \tag{A.10}$$

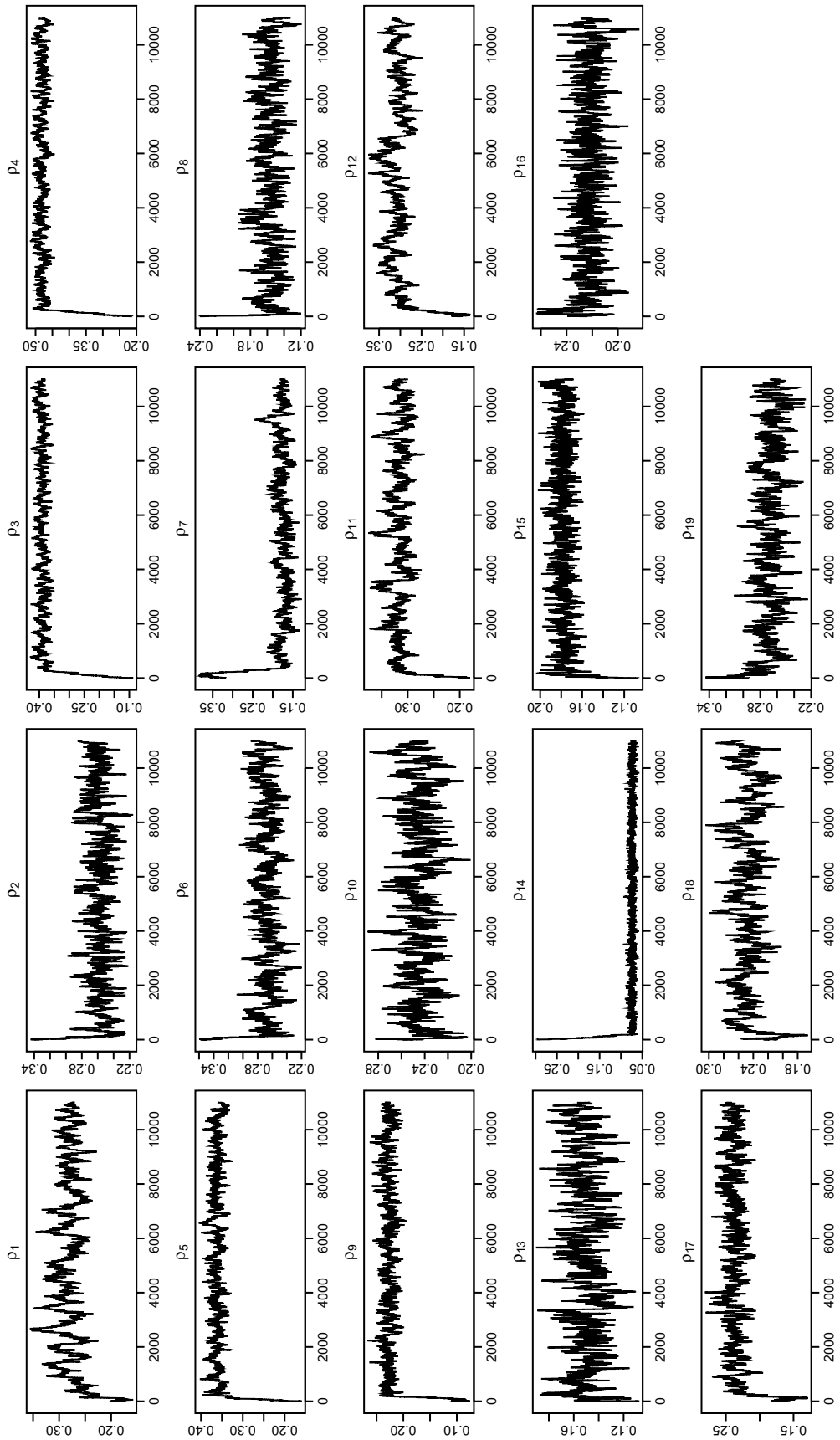


Figure A.8: Traceplots of the $MCMC$ estimates of trial 1.

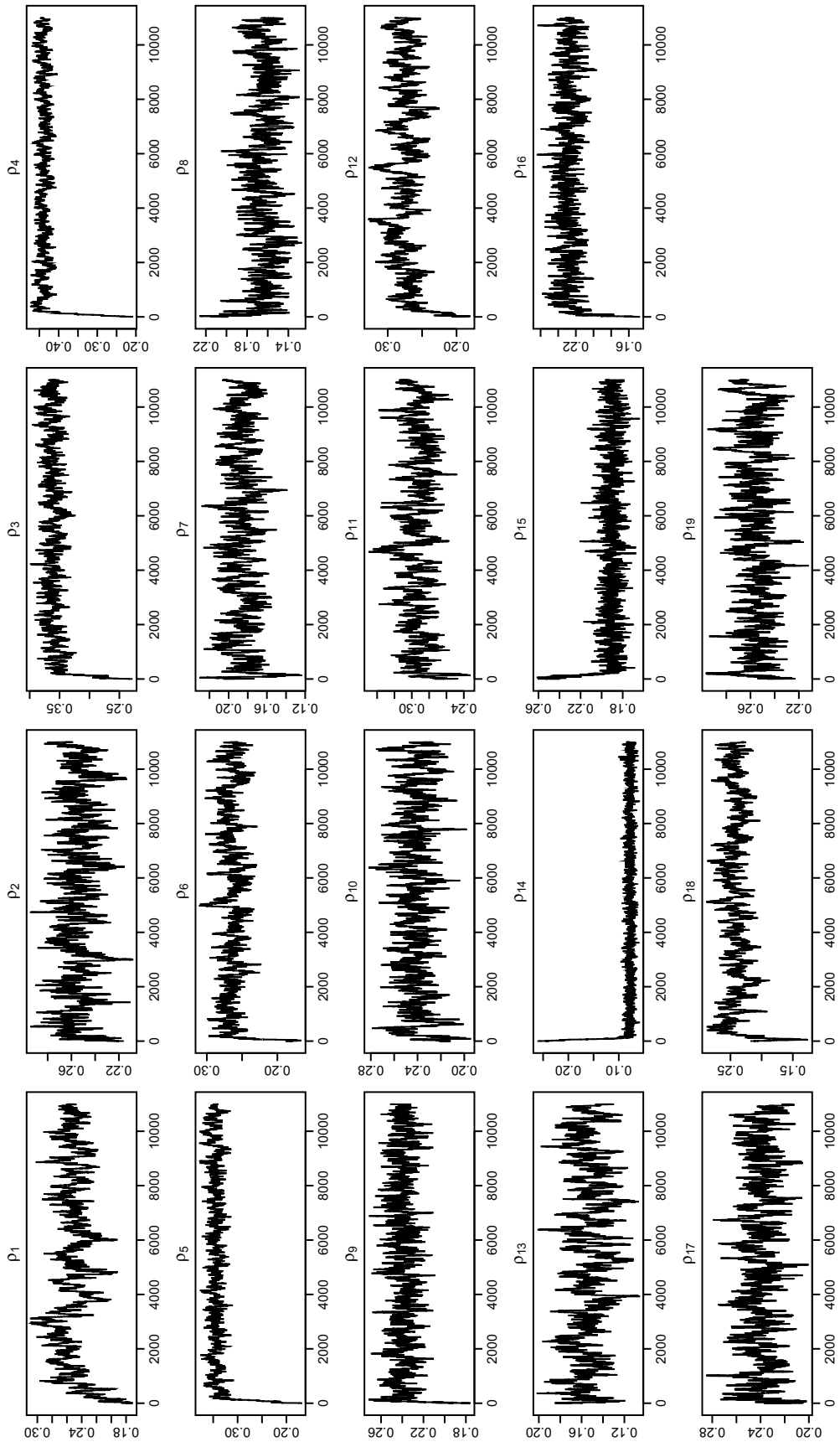


Figure A.9: Traceplots of the $MCMC$ estimates of trial 2.

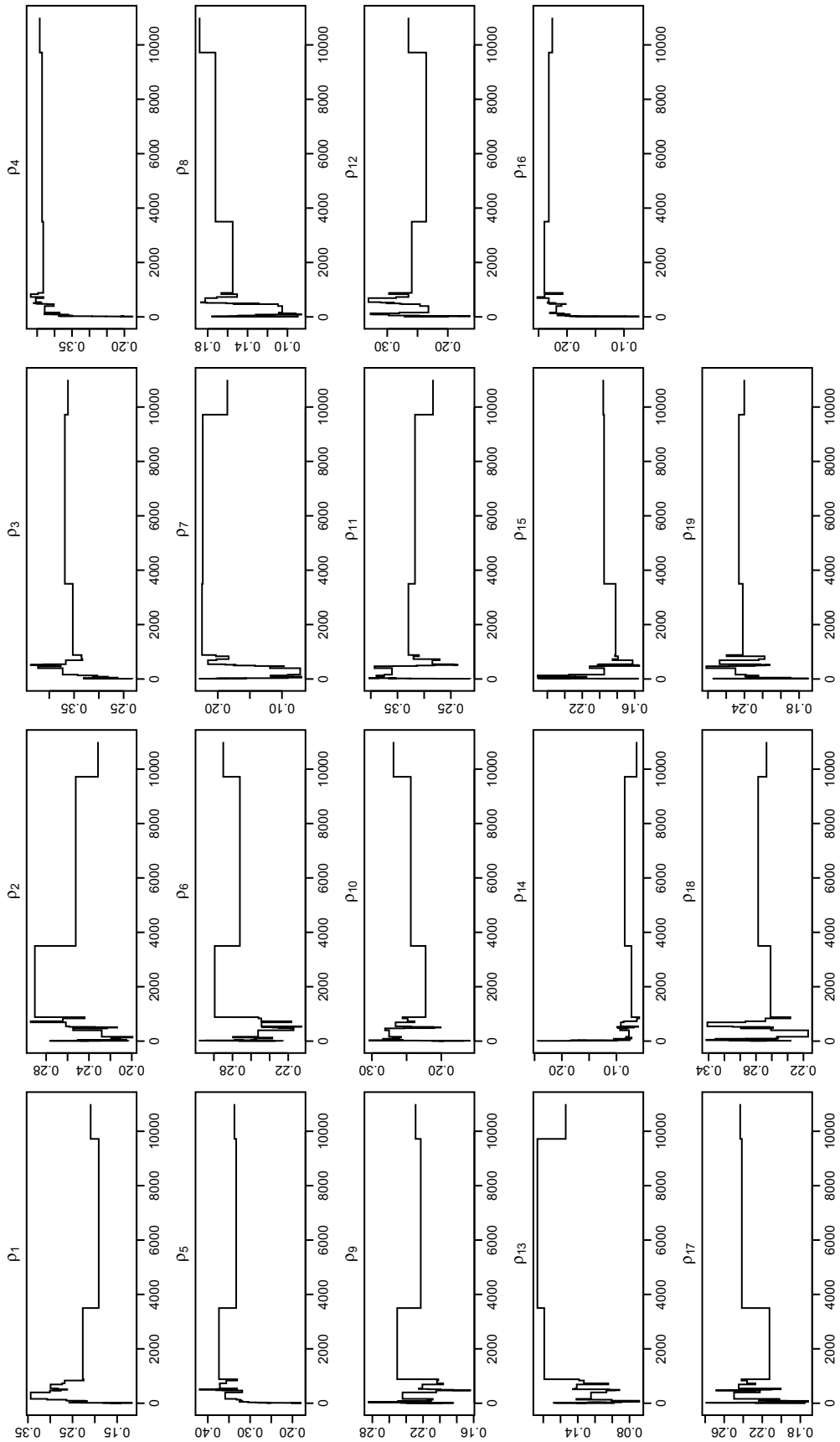


Figure A.10: Traceplots of the $MCMC$ estimates of trial 3.

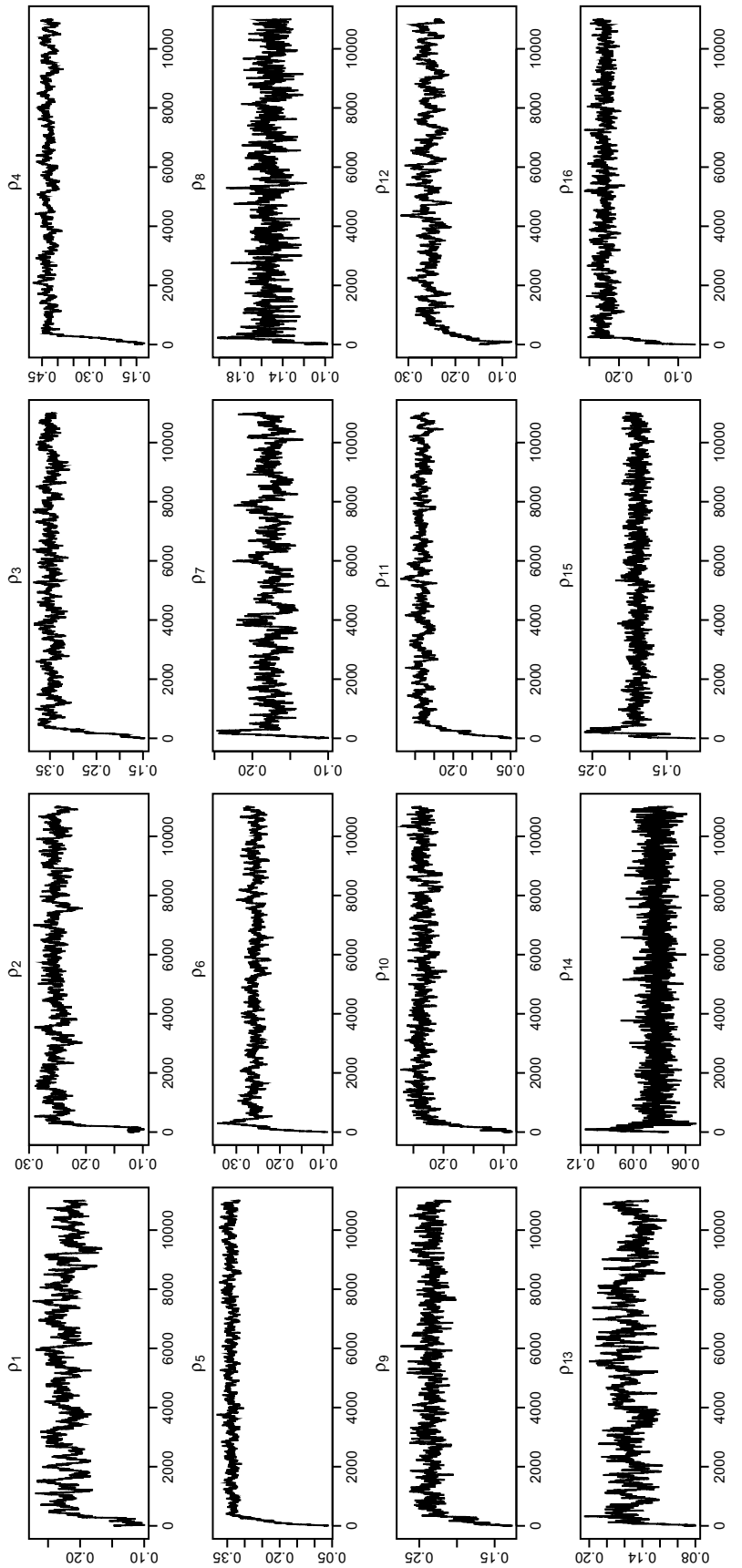


Figure A.11: Traceplots of the $MCMC$ estimates of trial 4.

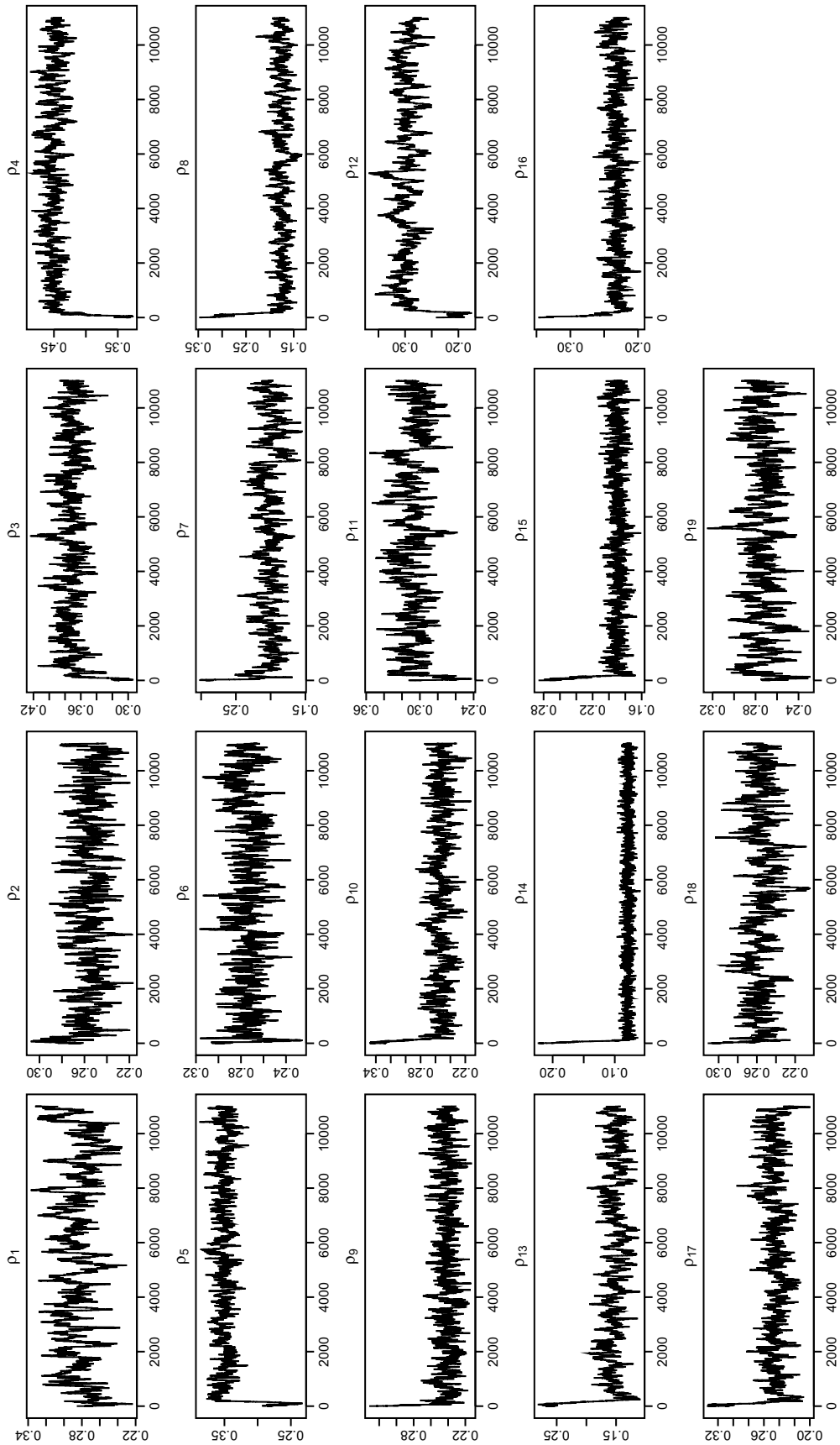


Figure A.12: Traceplots of the $MCMC$ estimates of trial 5.

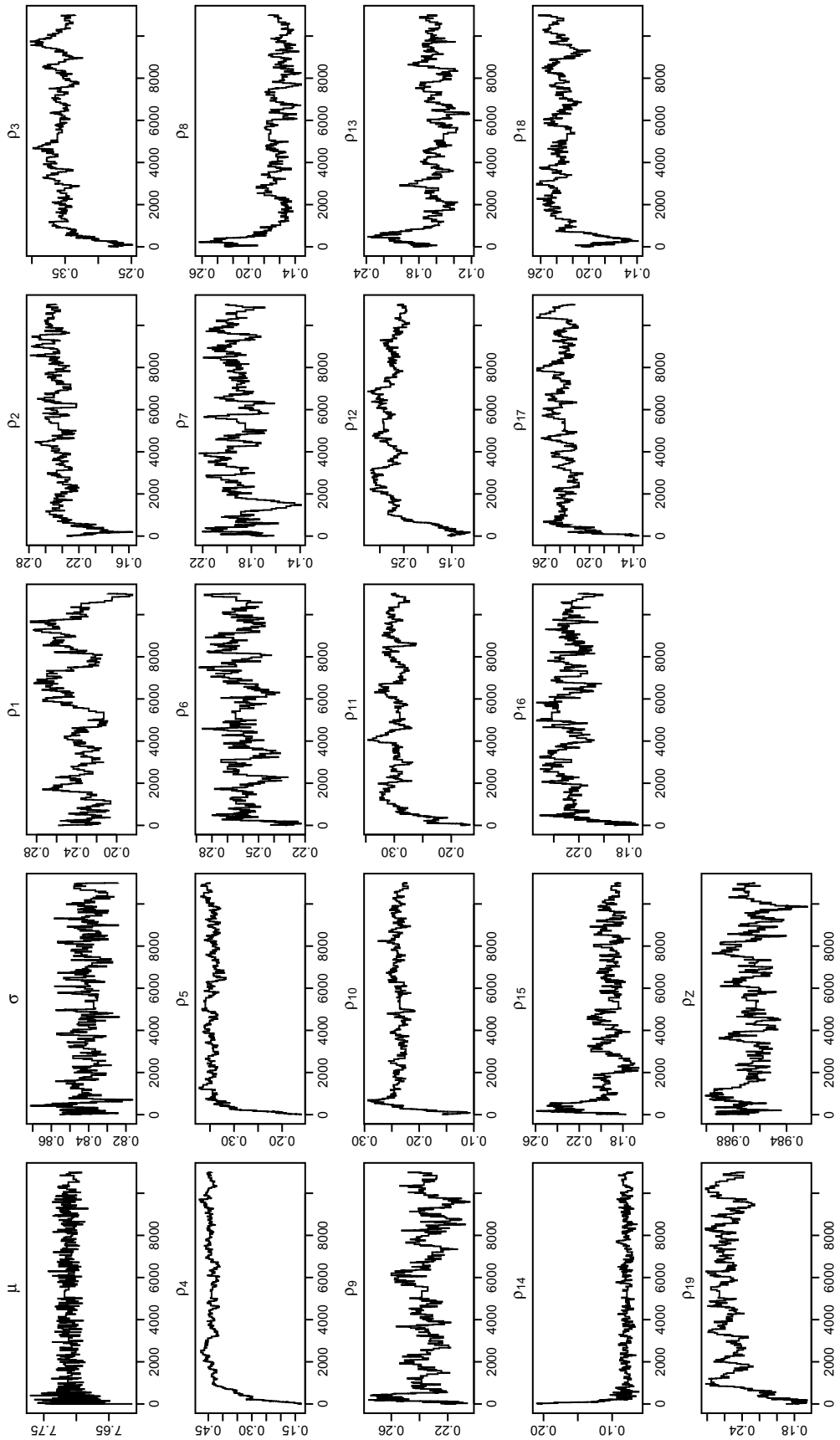


Figure A.13: Traceplots of the $MCMC$ estimates of trial 6.

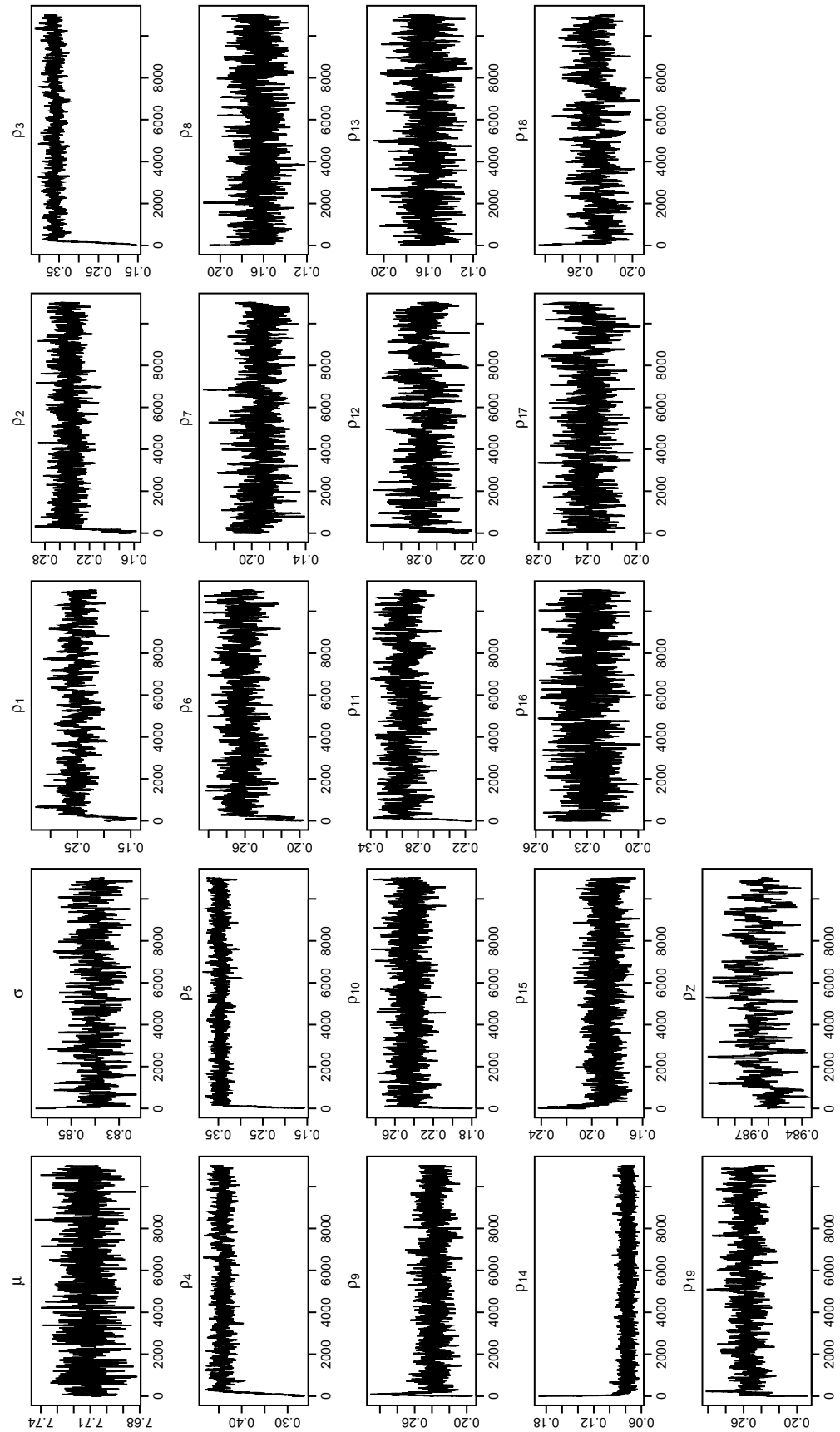


Figure A.14: Traceplots of the $MCMC$ estimates of trial 7.

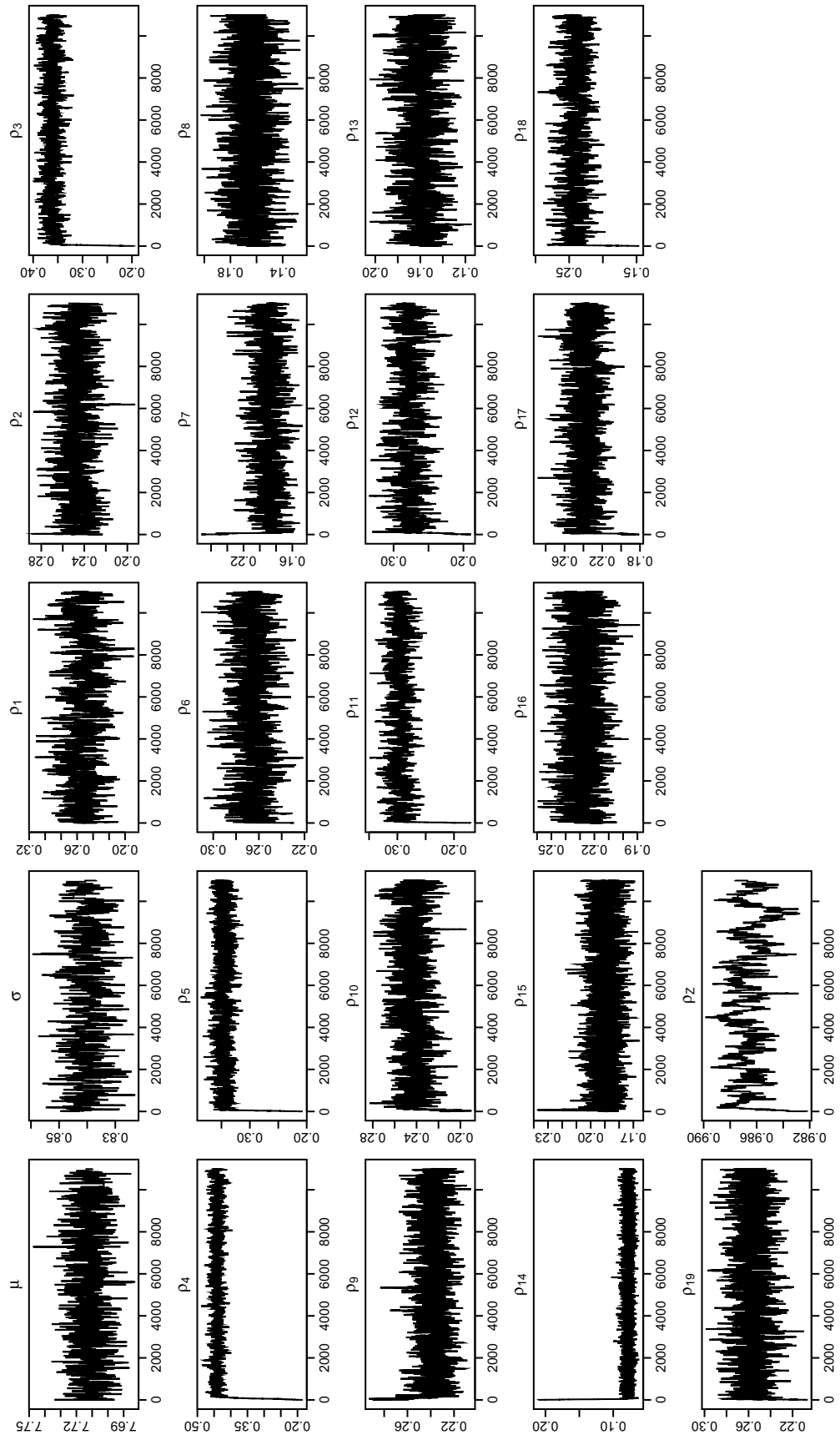


Figure A.15: Traceplots of the $MCMC$ estimates of trial 8.

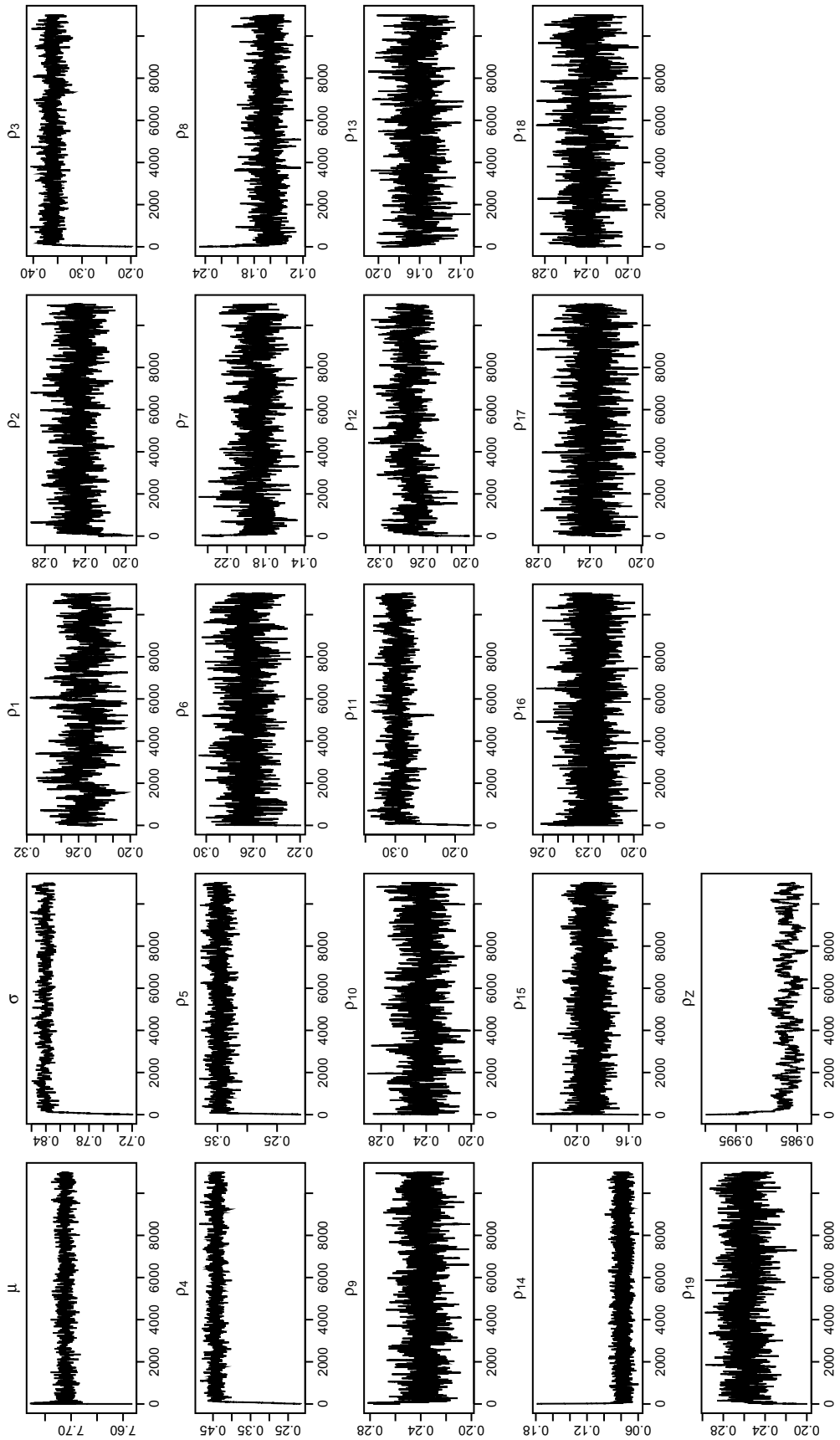


Figure A.16: Traceplots of the $MCMC$ estimates of trial 9.

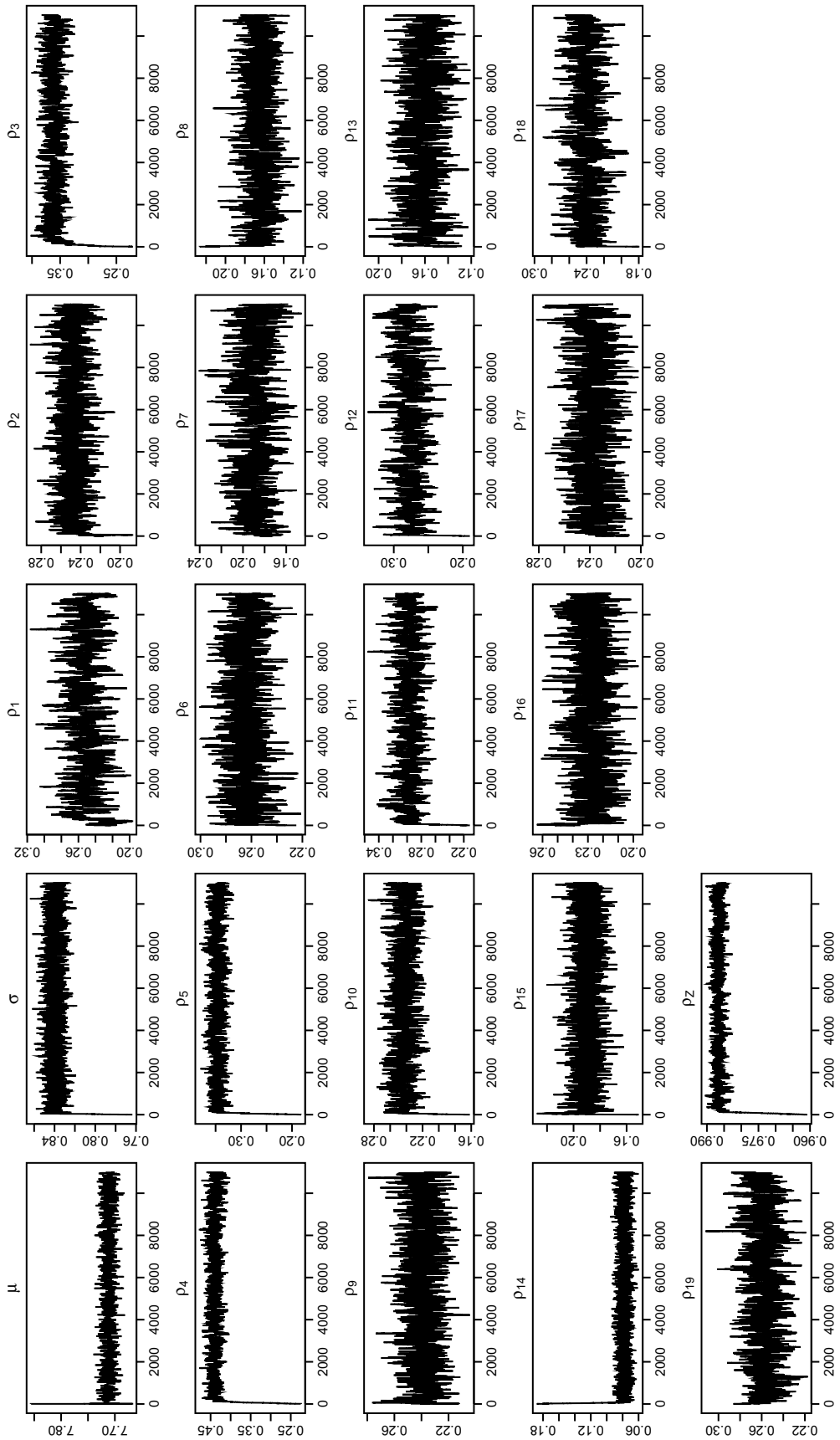


Figure A.17: Traceplots of the $MCMC$ estimates of trial 10.

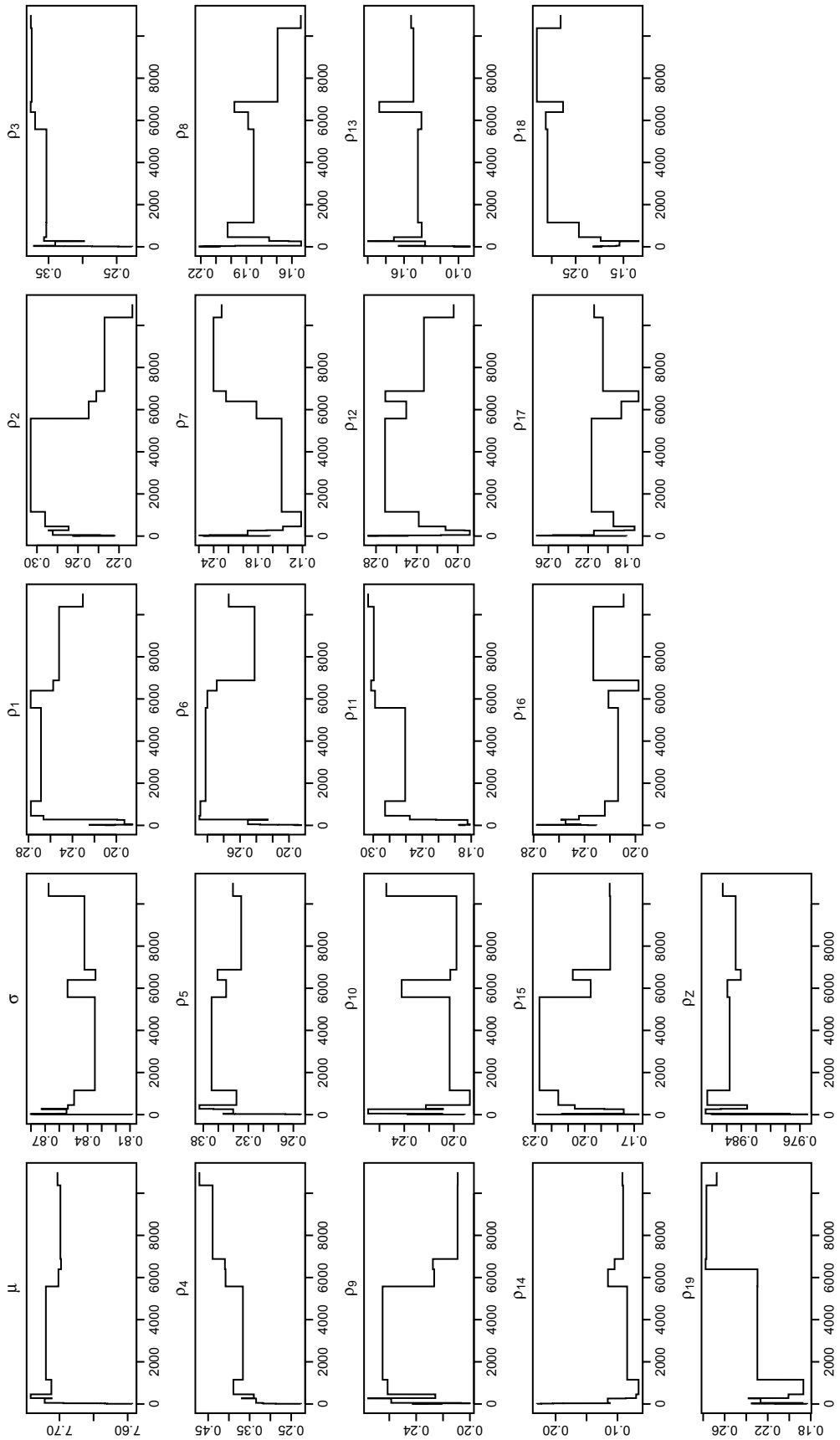


Figure A.18: Traceplots of the $MCMC$ estimates of trial 11.

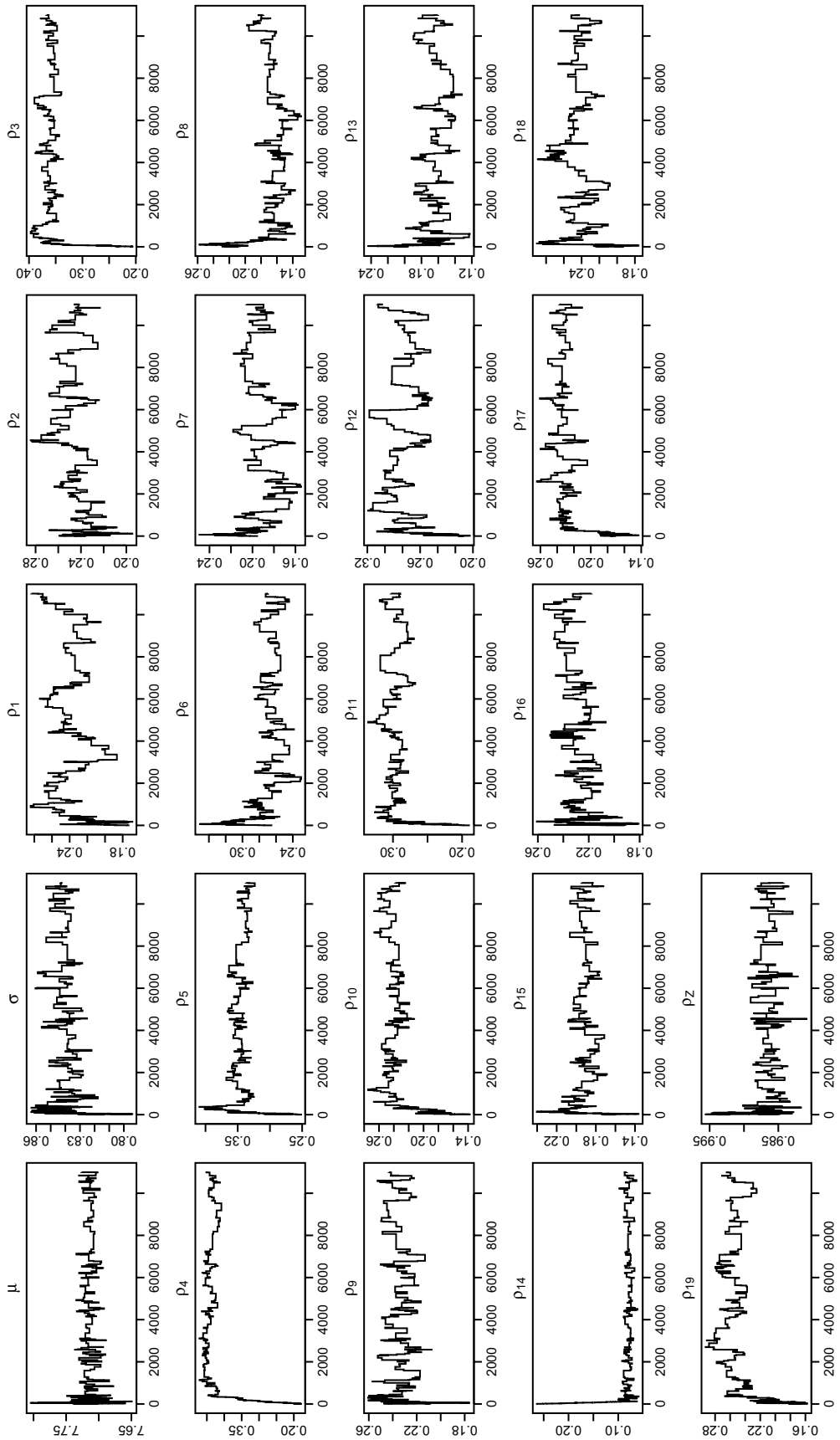


Figure A.19: Traceplots of the $MCMC$ estimates of trial 12.

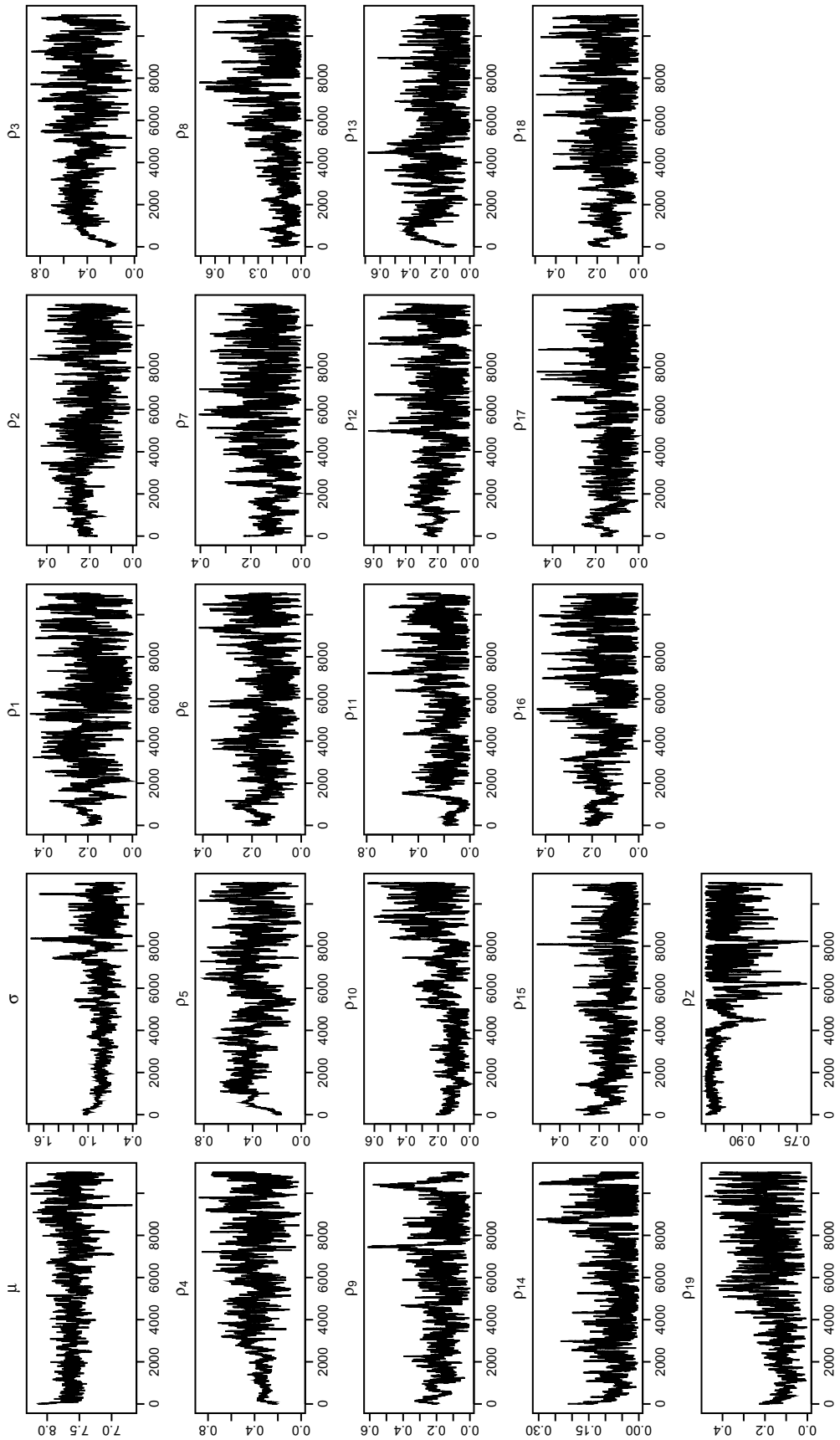


Figure A.20: Traceplots of the $MCMC$ estimates of trial 13.

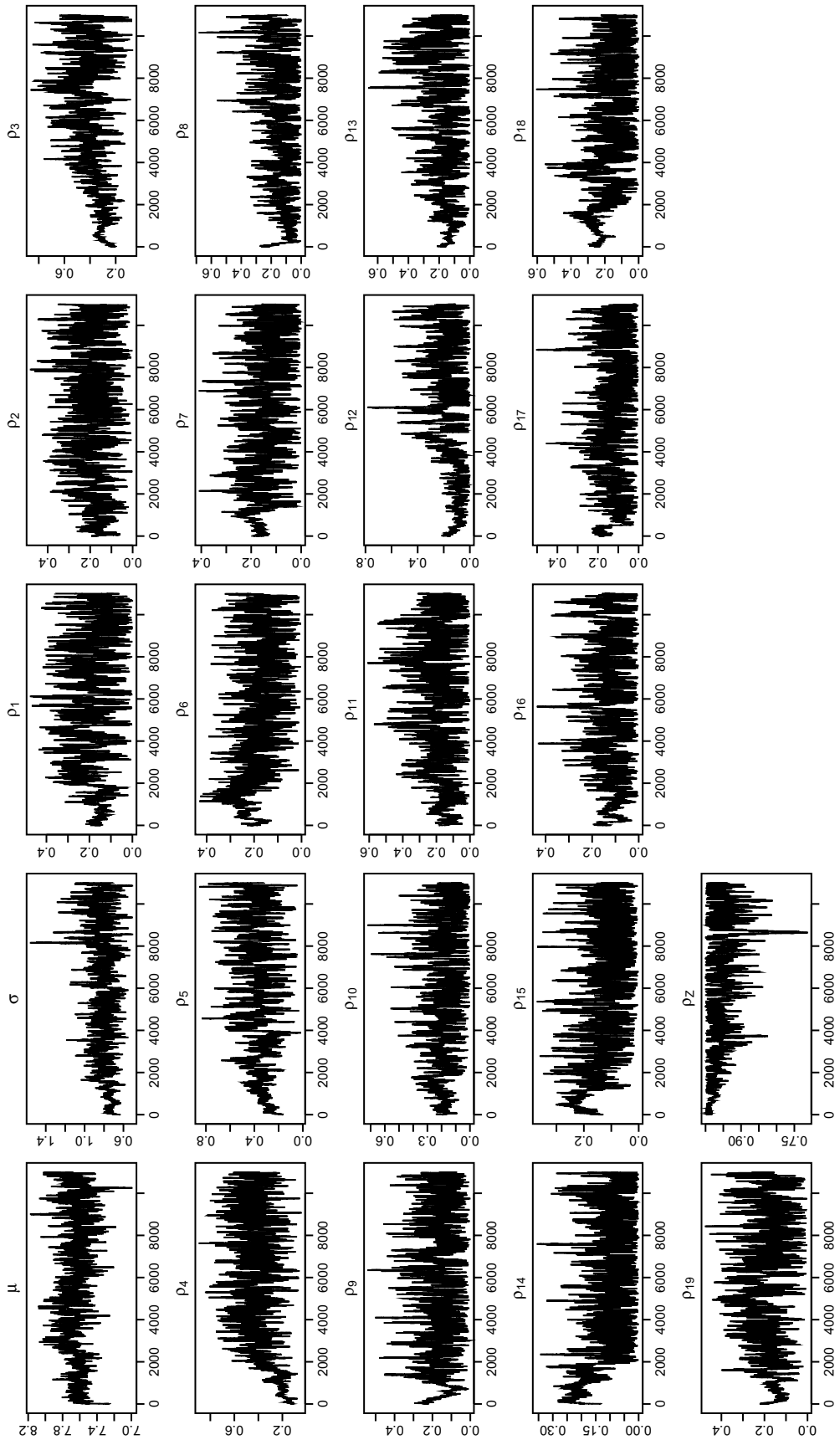


Figure A.21: Traceplots of the $MCMC$ estimates of trial 14.

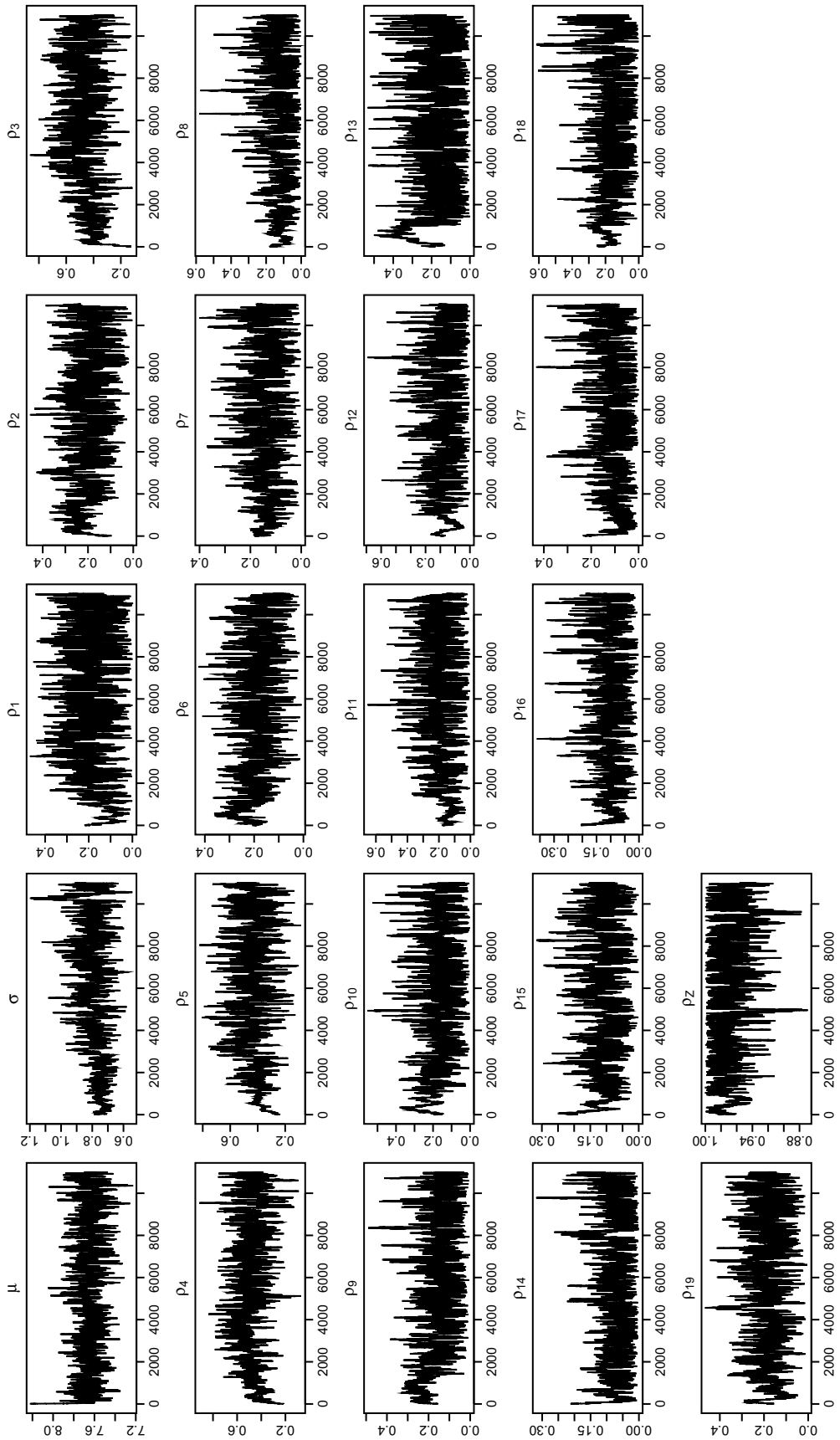


Figure A.22: Traceplots of the $MCMC$ estimates of trial 15.

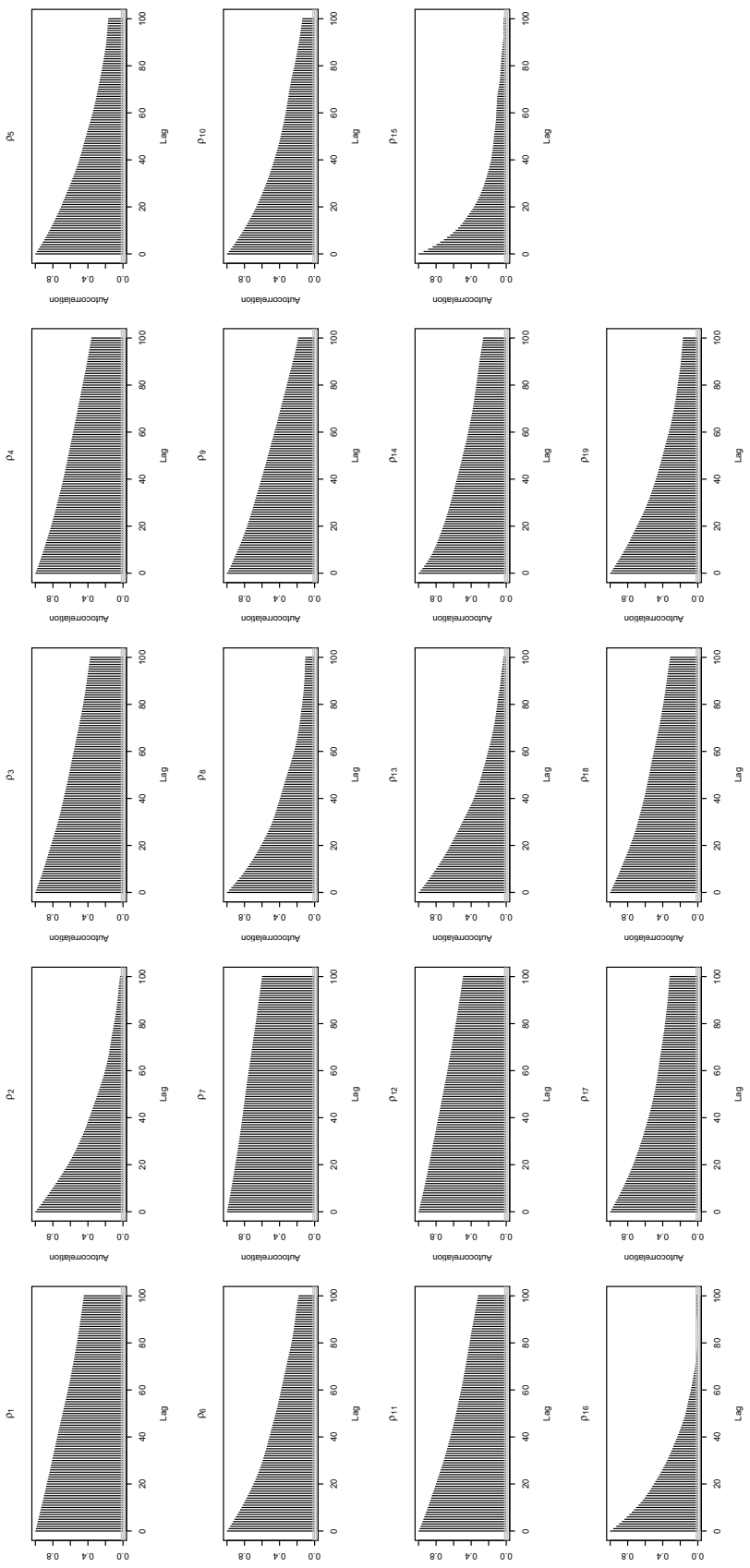


Figure A.23: ACF plots of the $MCMC$ estimates of trial 1.

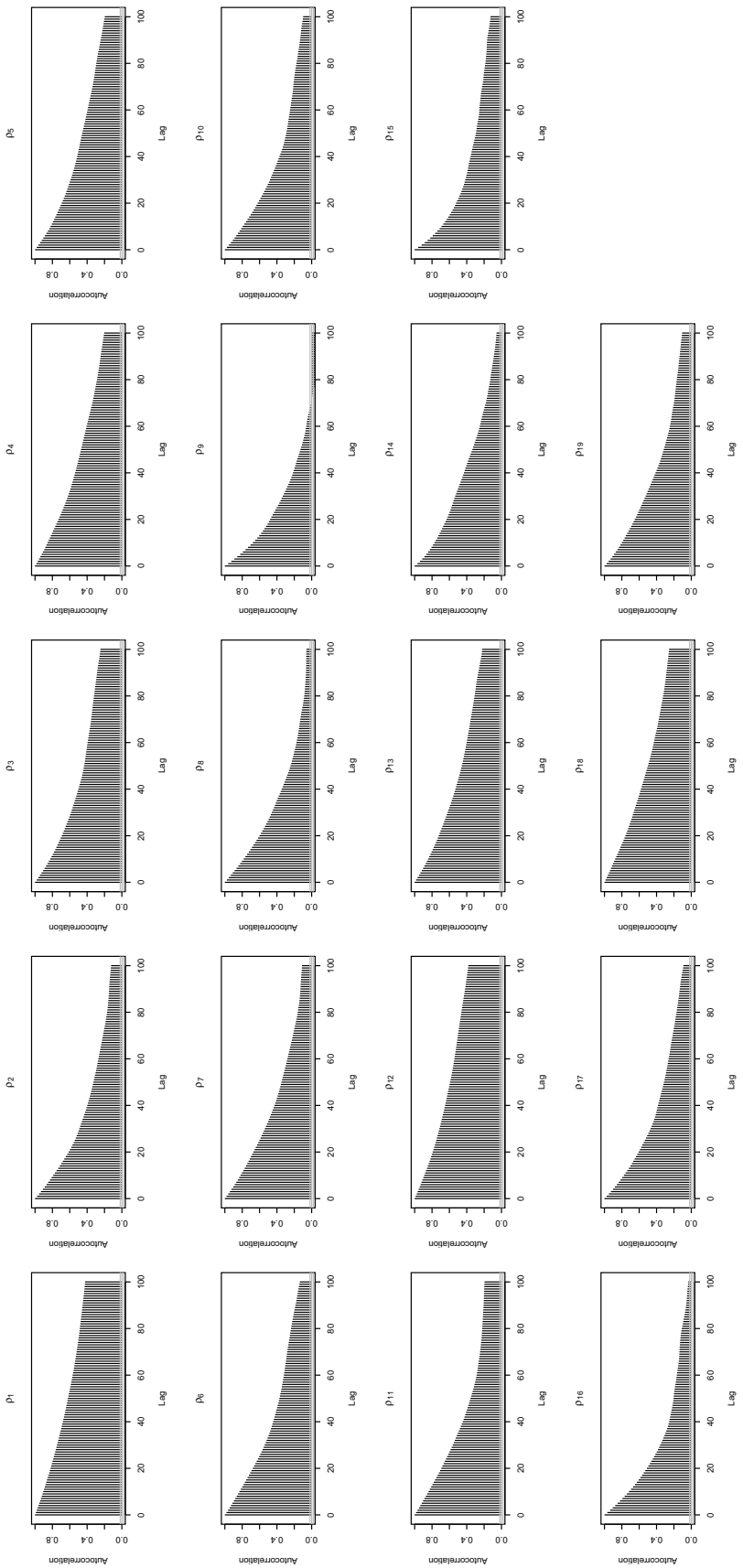


Figure A.24: ACF plots of the $MCMC$ estimates of trial 2.

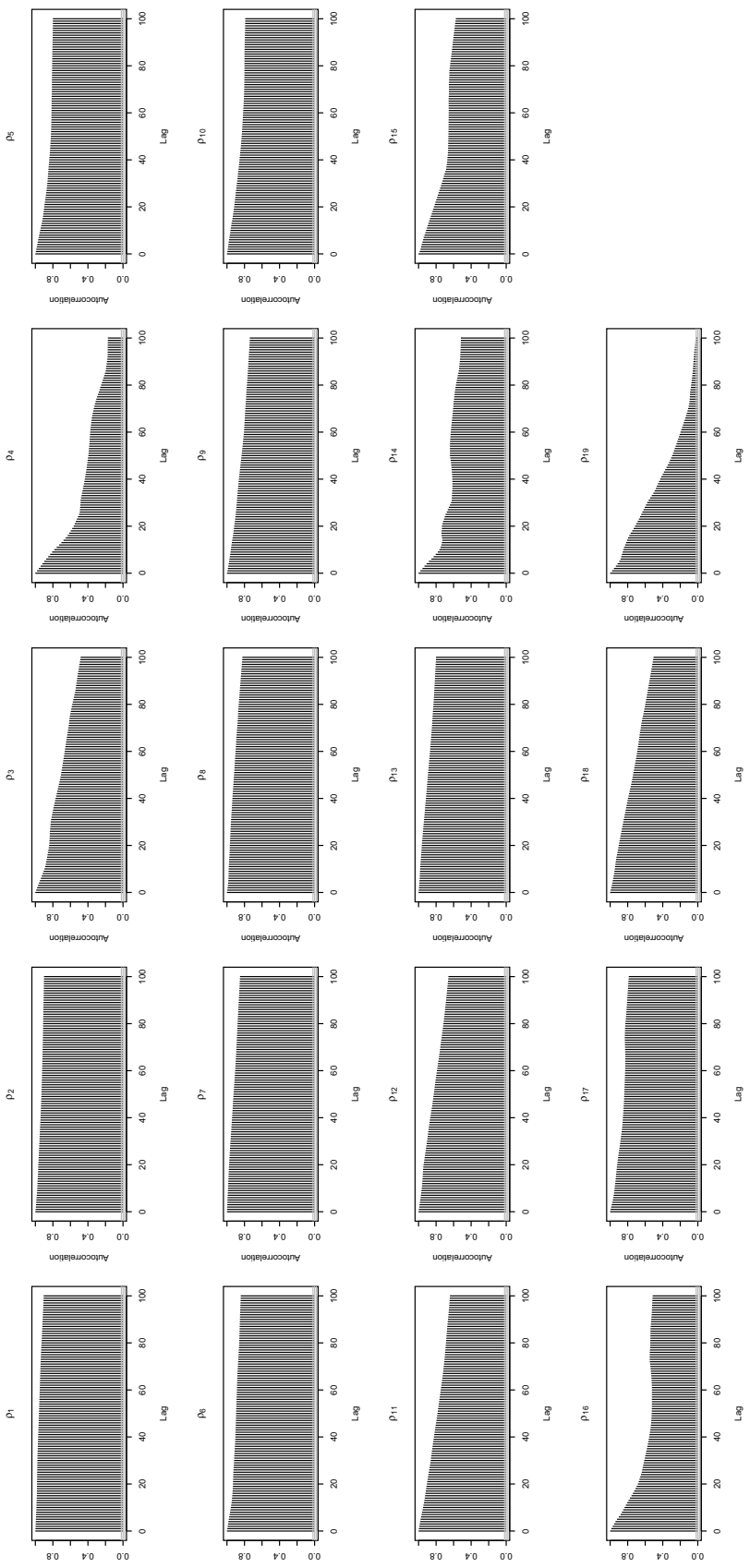


Figure A.25: ACF plots of the $MCMC$ estimates of trial 3.

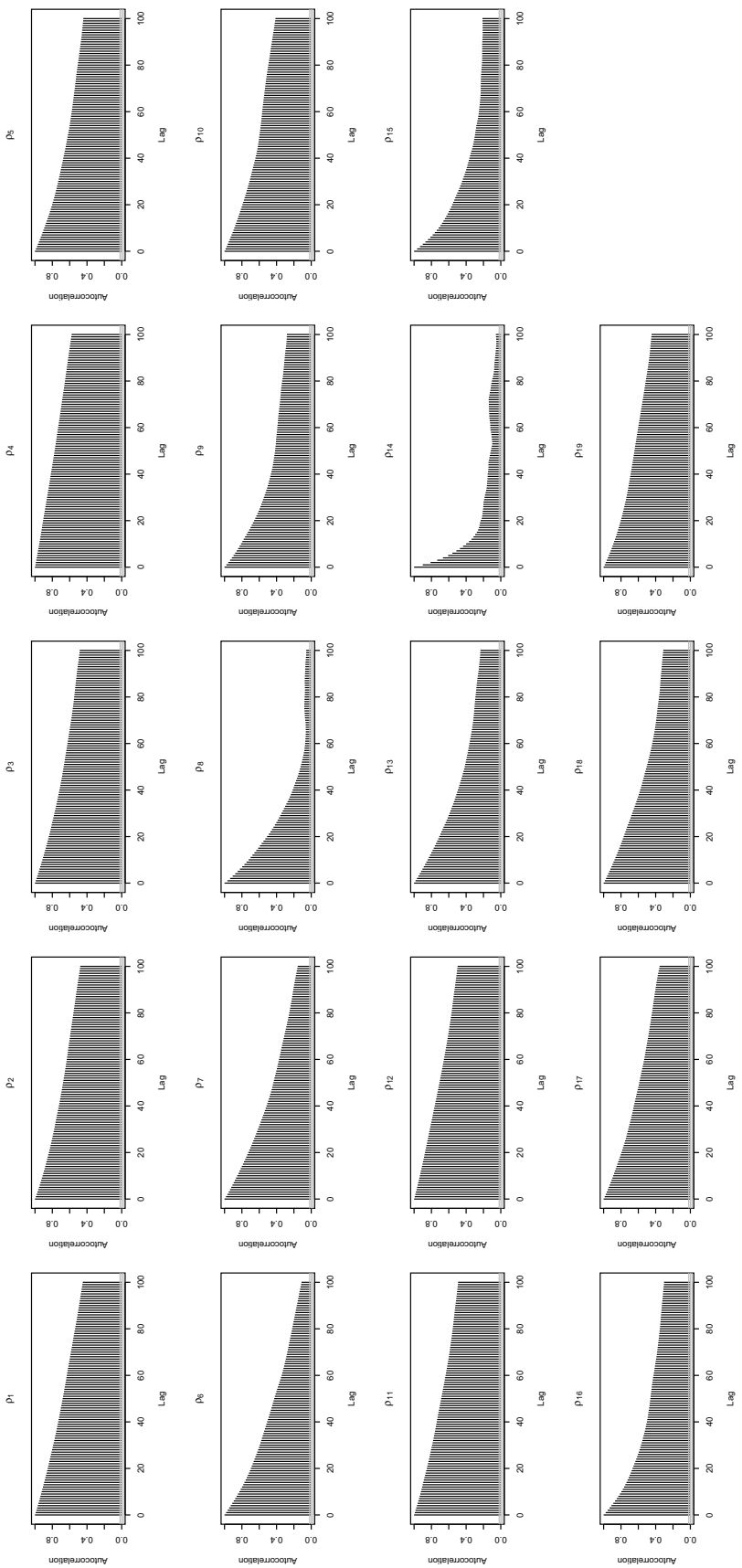


Figure A.26: ACF plots of the $MCMC$ estimates of trial 4.

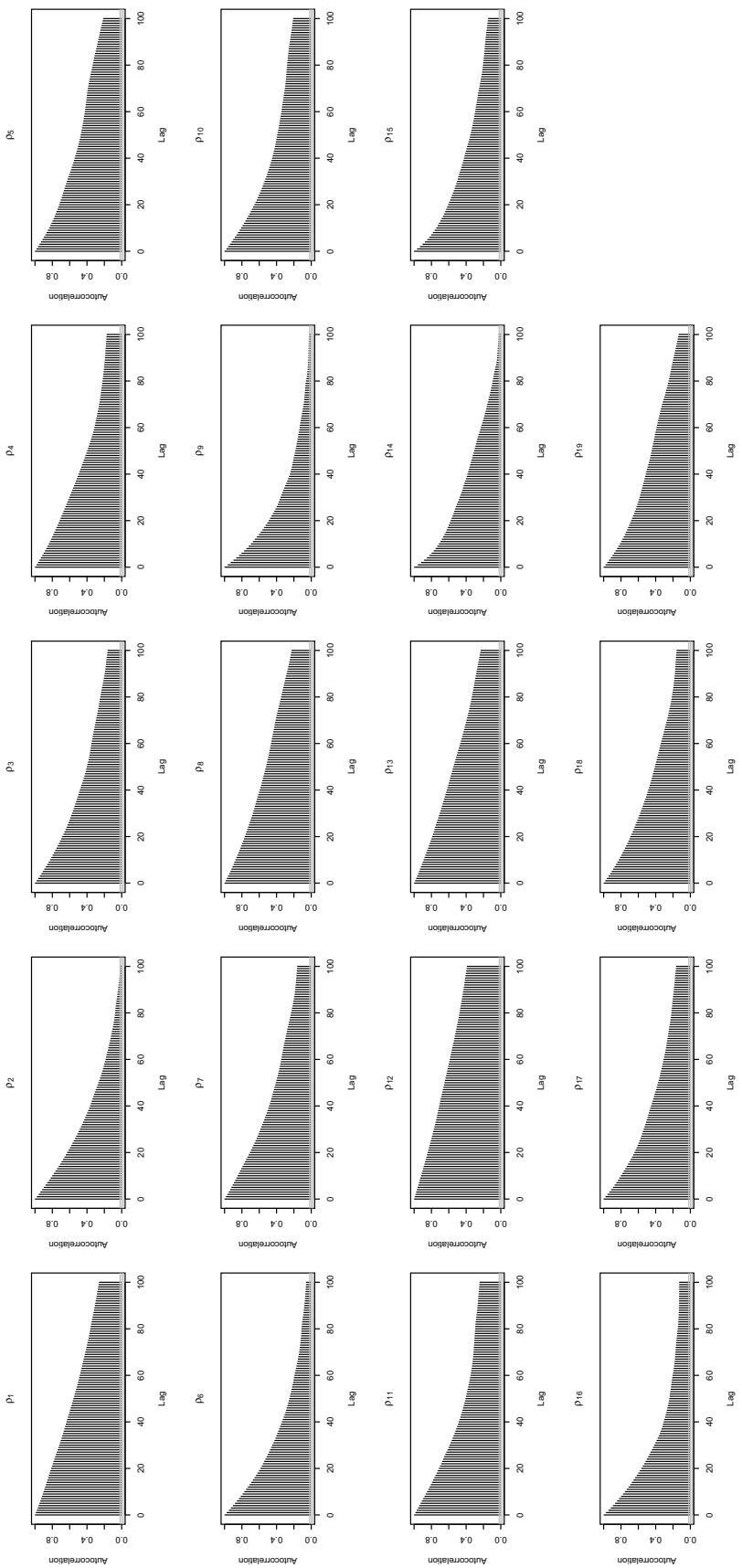


Figure A.27: ACF plots of the $MCMC$ estimates of trial 5.

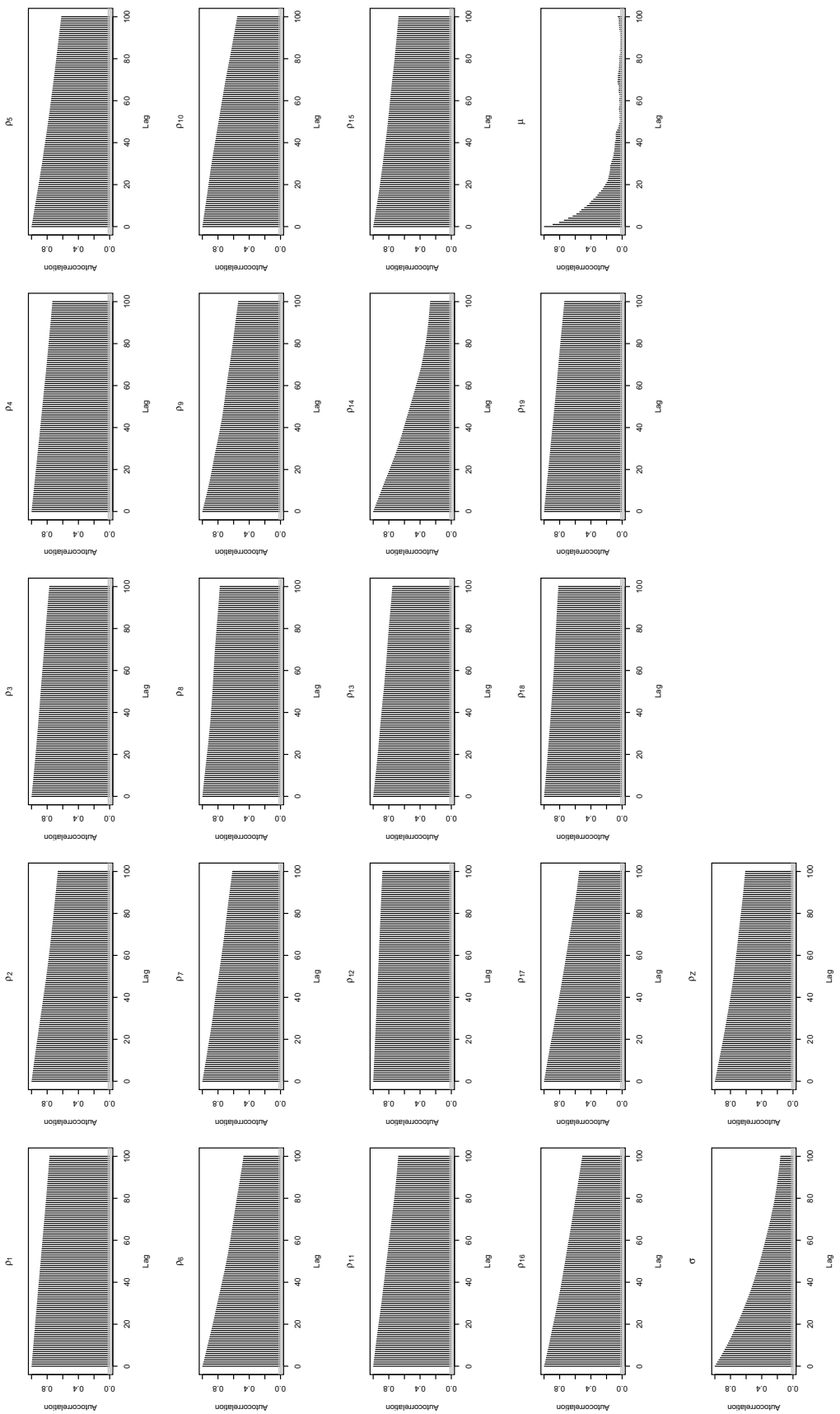


Figure A.28: ACF plots of the $MCMC$ estimates of trial 6.

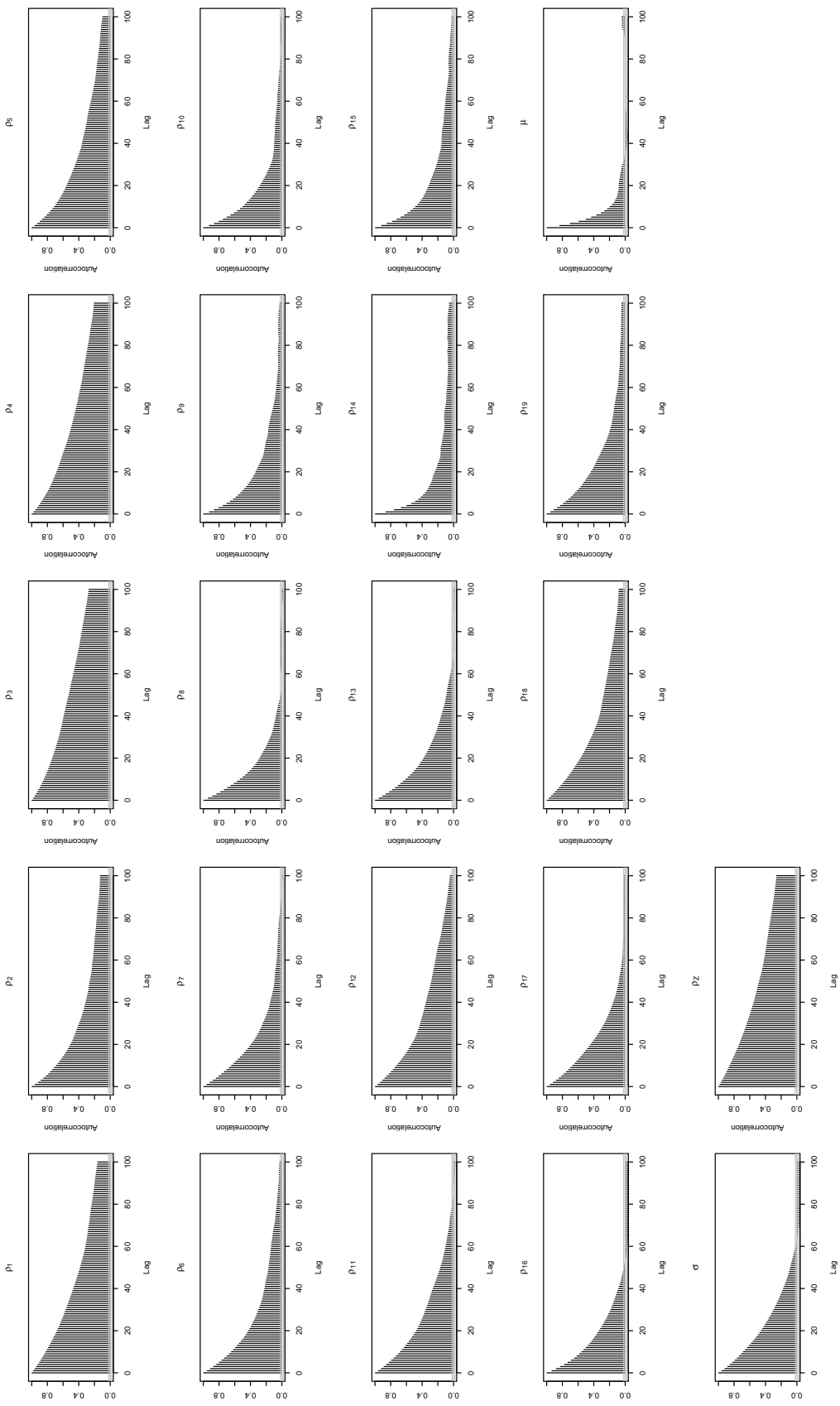


Figure A.29: ACF plots of the $MCMC$ estimates of trial 7.

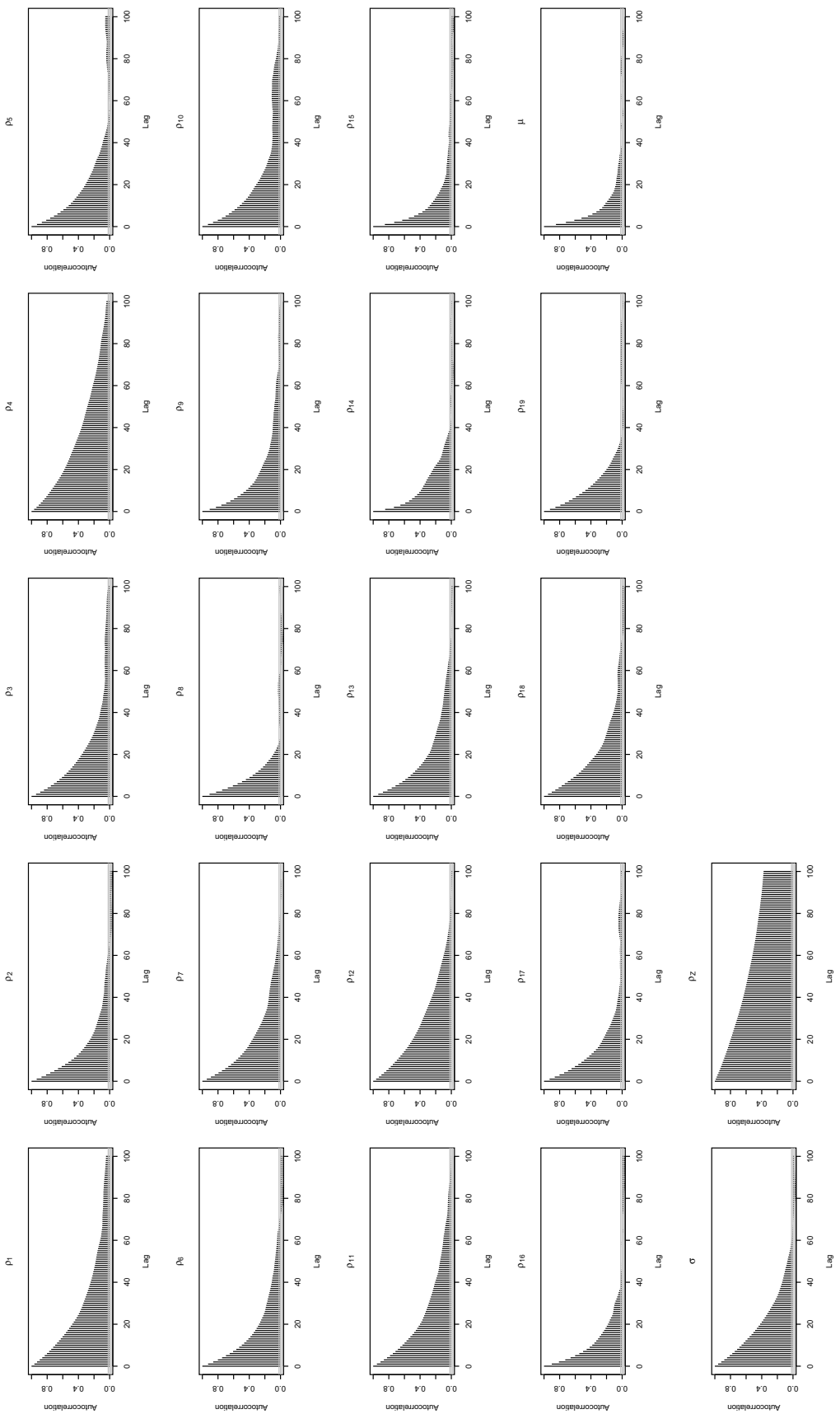


Figure A.30: ACF plots of the $MCMC$ estimates of trial 8.

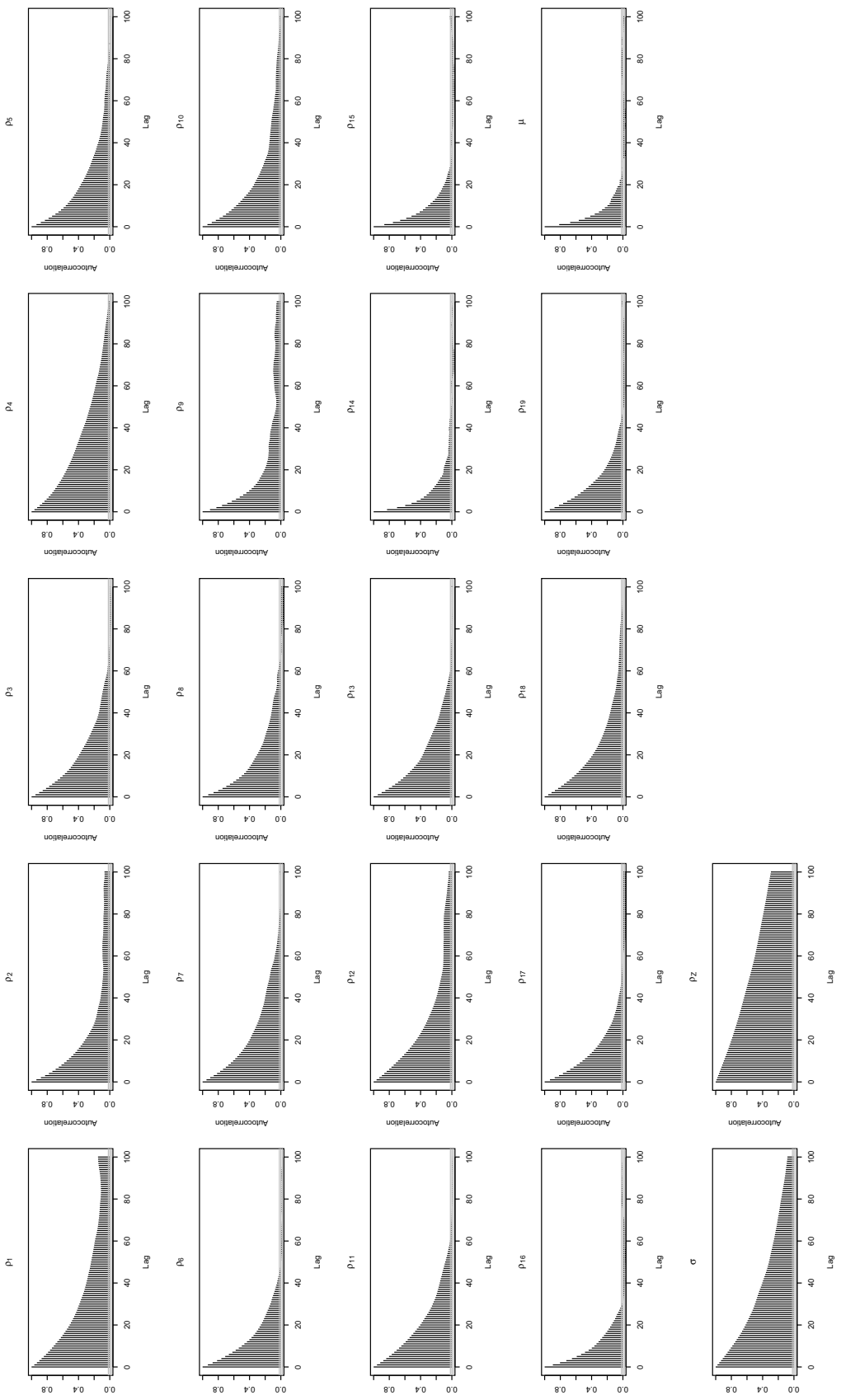


Figure A.31: ACF plots of the $MCMC$ estimates of trial 9.

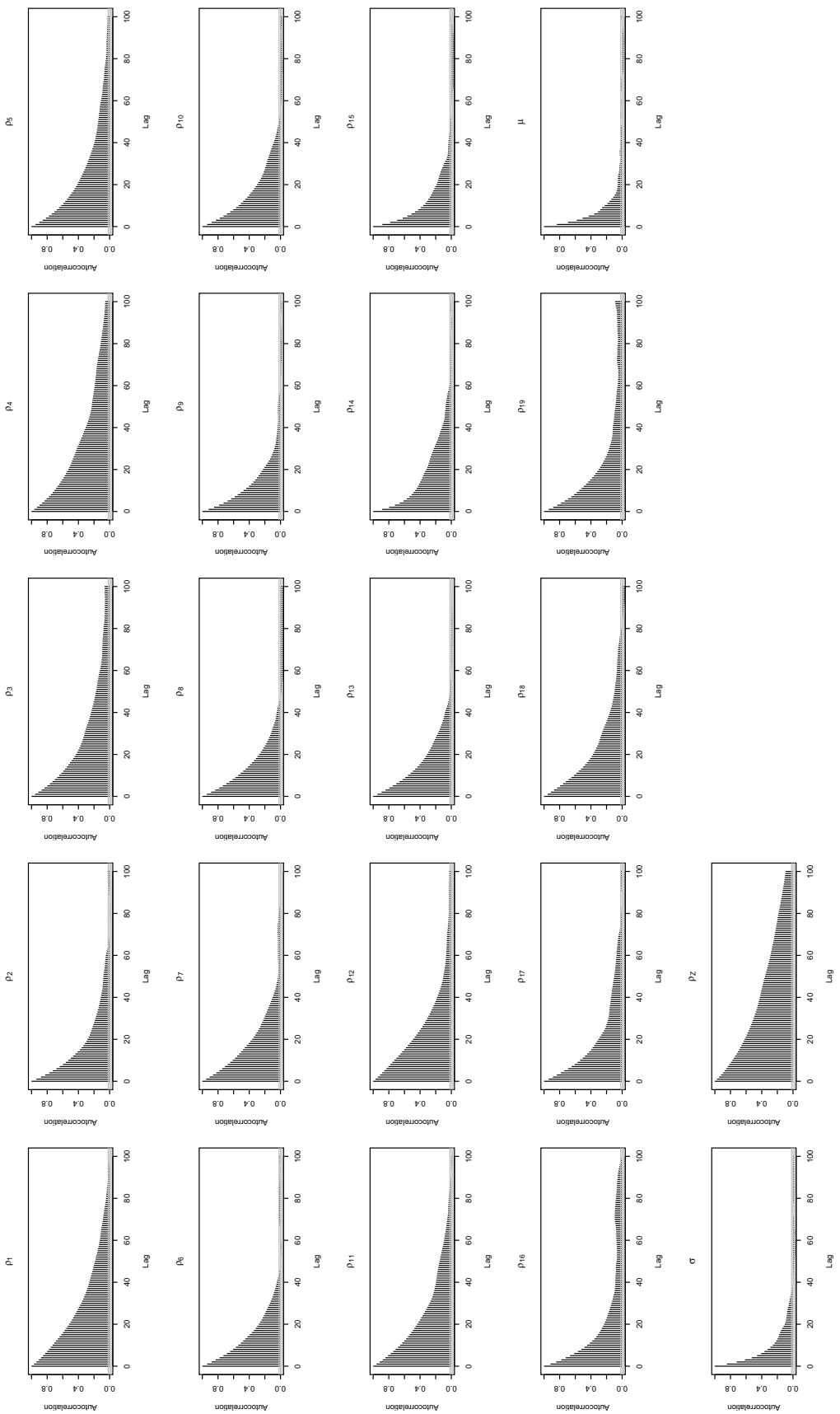


Figure A.32: ACF plots of the $MCMC$ estimates of trial 10.

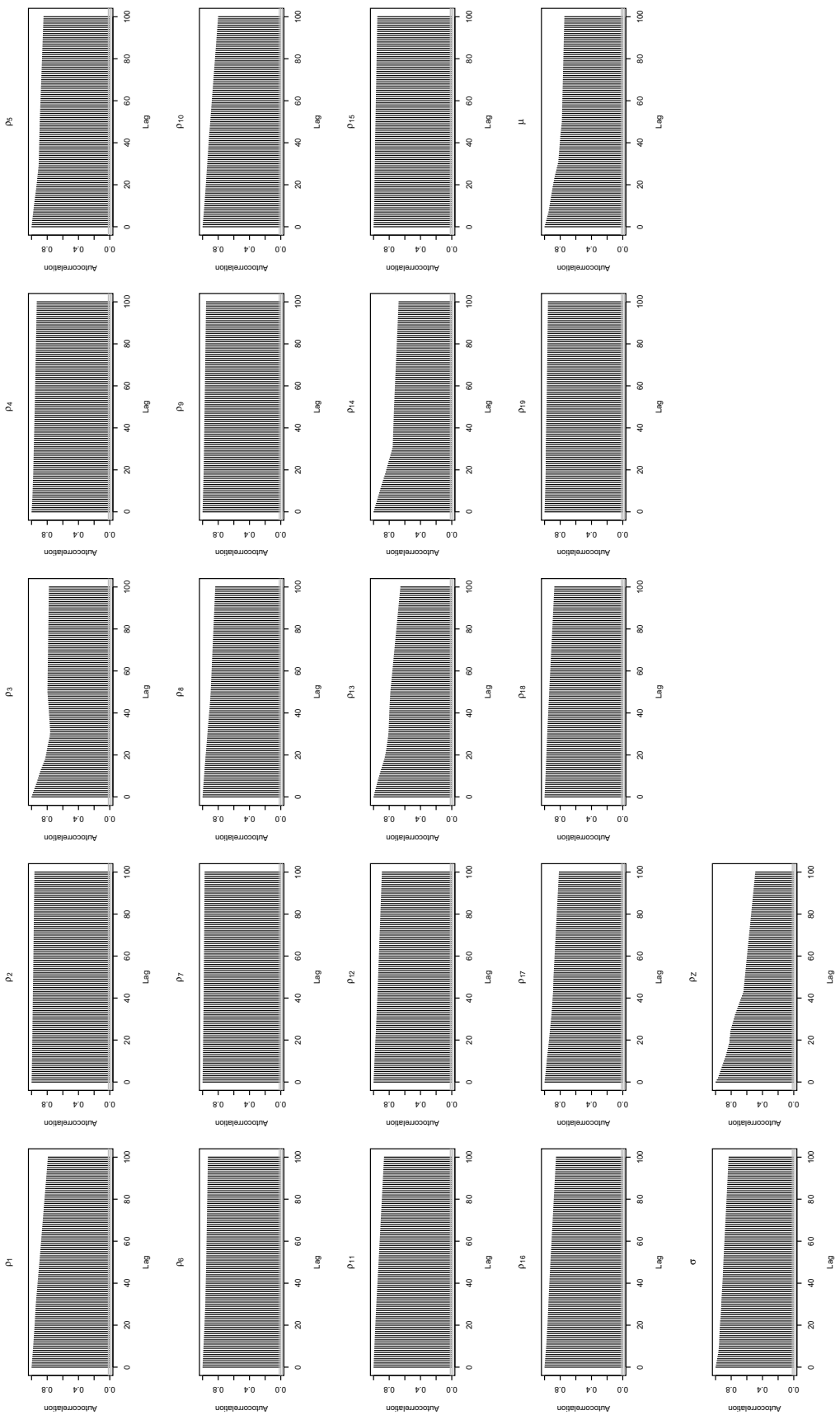


Figure A.33: ACF plots of the $MCMC$ estimates of trial 11.

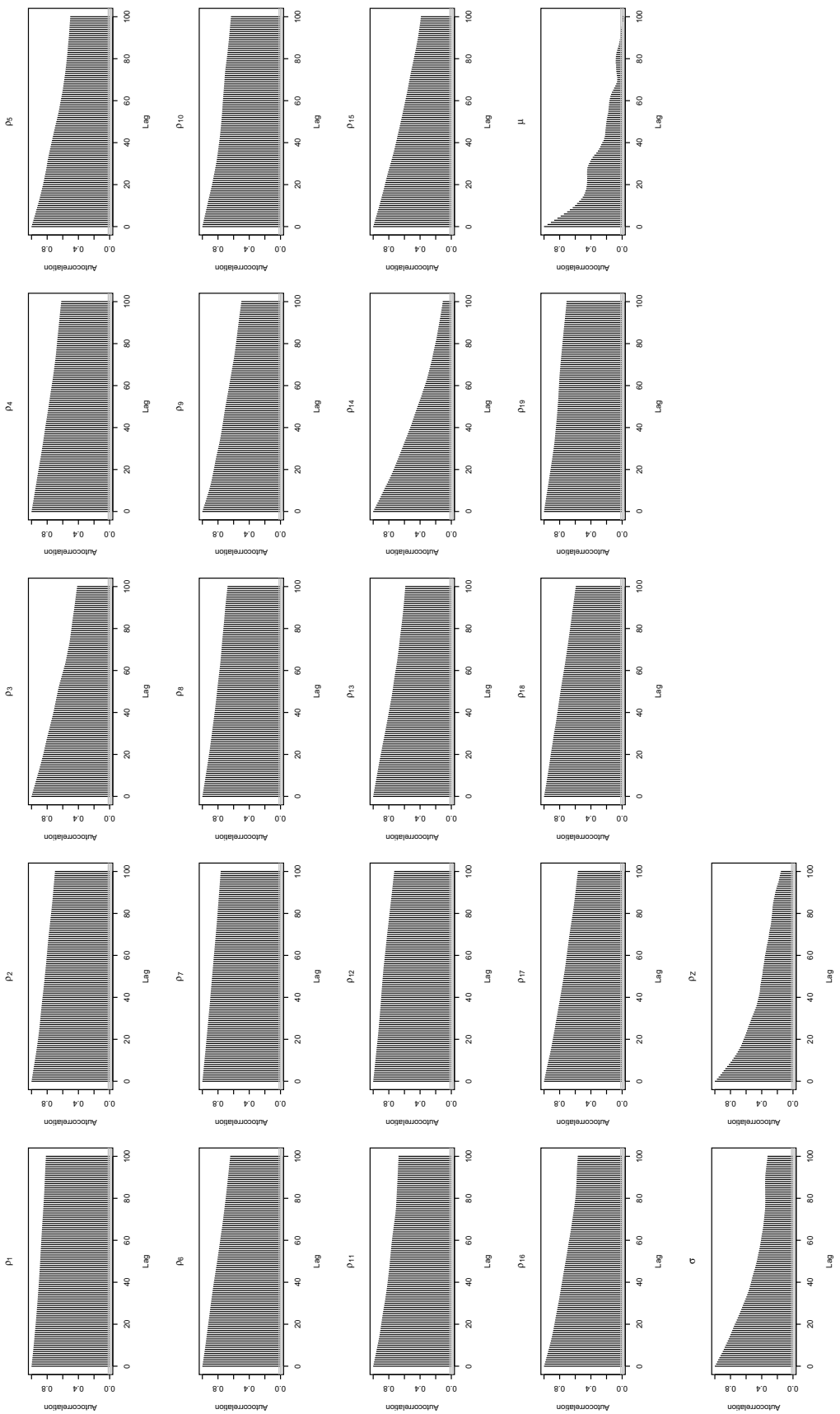


Figure A.34: ACF plots of the $MCMC$ estimates of trial 12.

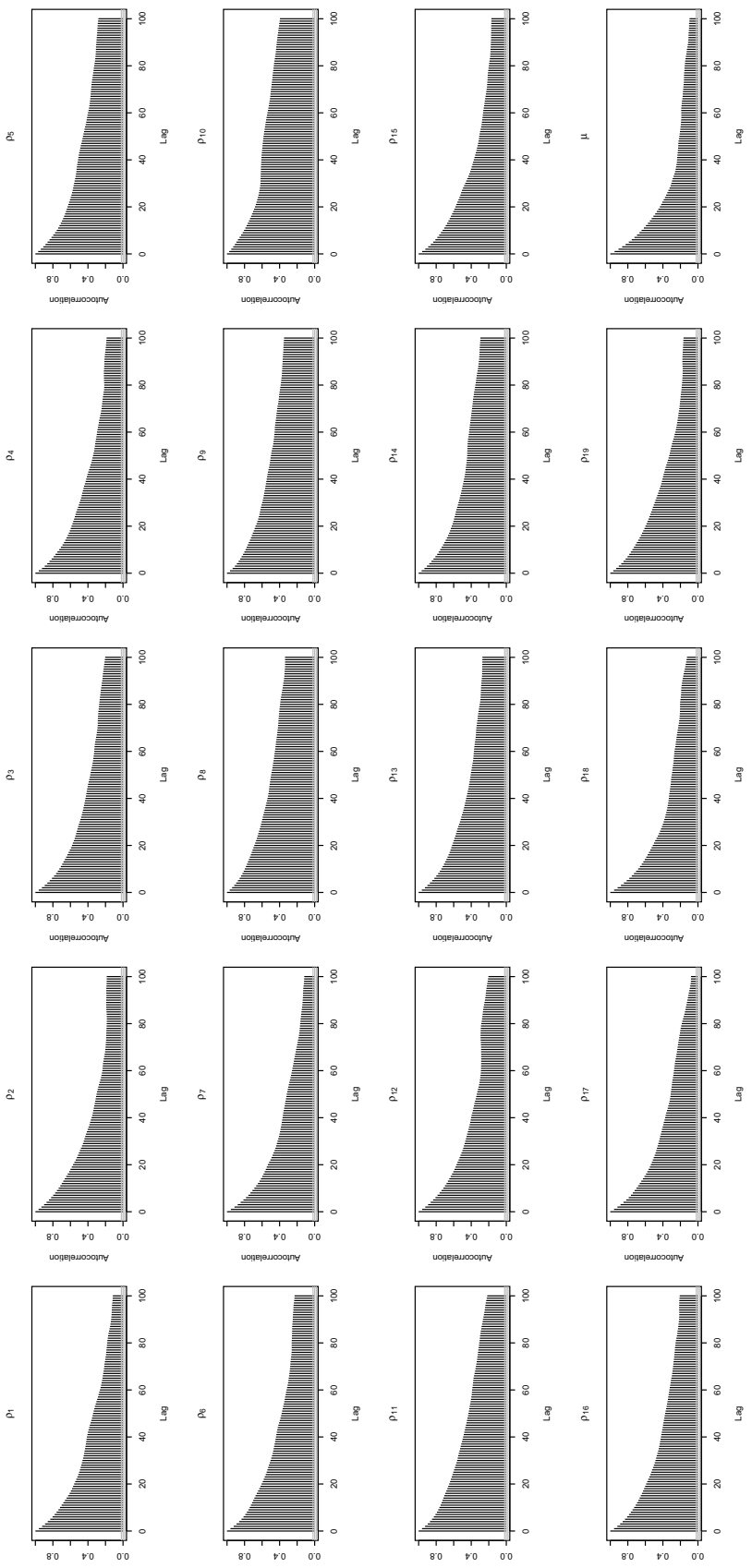


Figure A.35: ACF plots of the $MCMC$ estimates of trial 13.

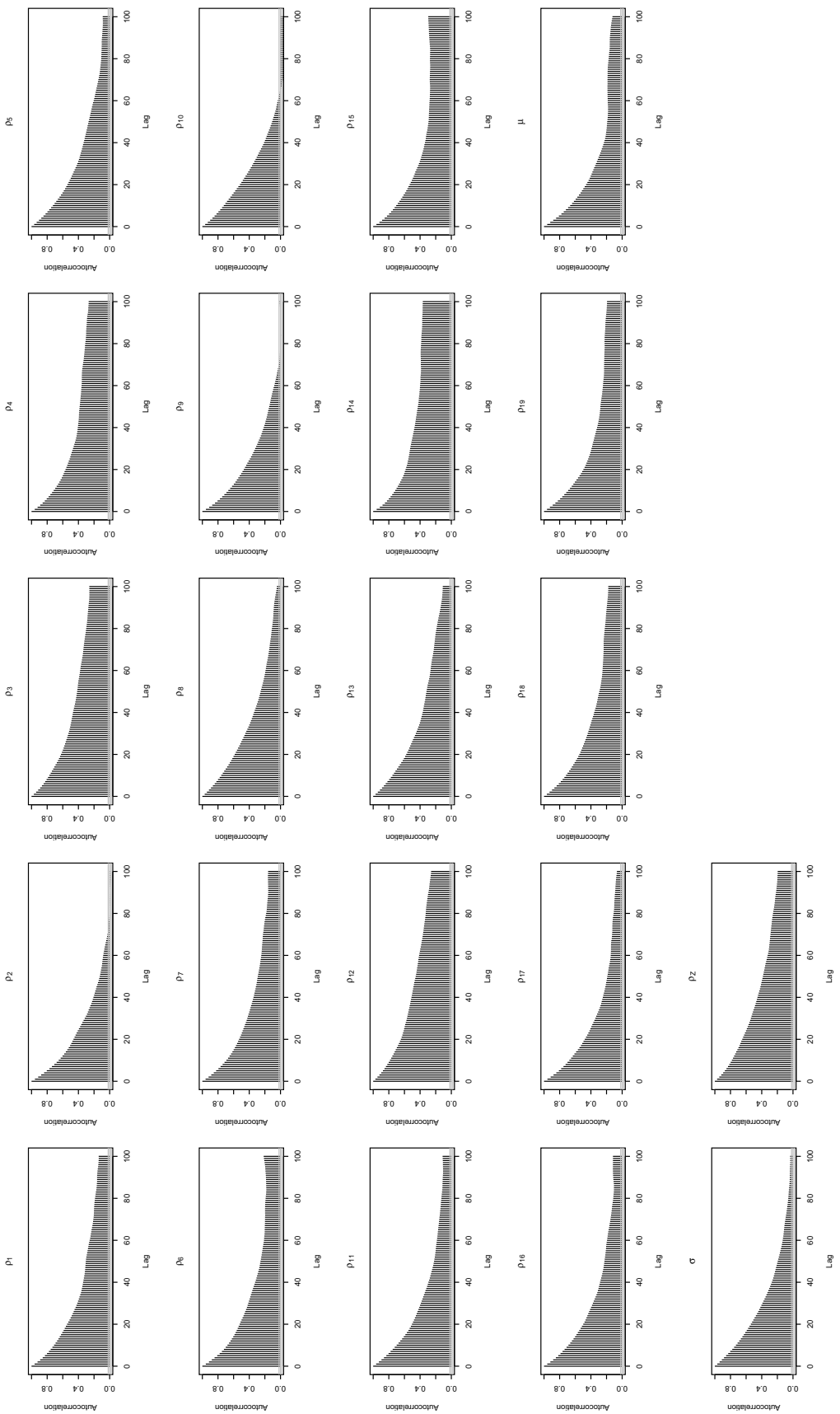


Figure A.36: ACF plots of the $MCMC$ estimates of trial 14.

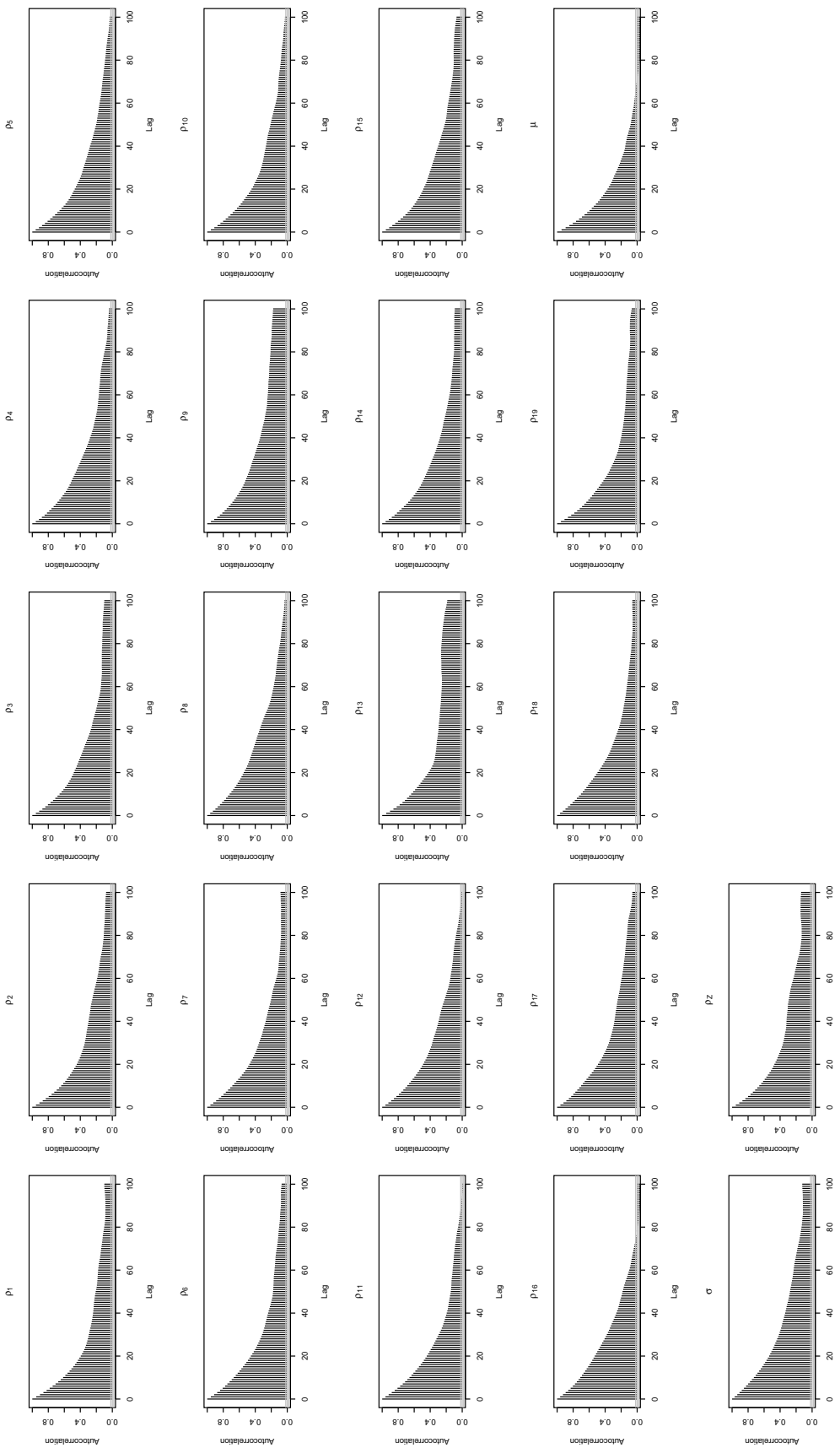


Figure A.37: ACF plots of the $MCMC$ estimates of trial 15.

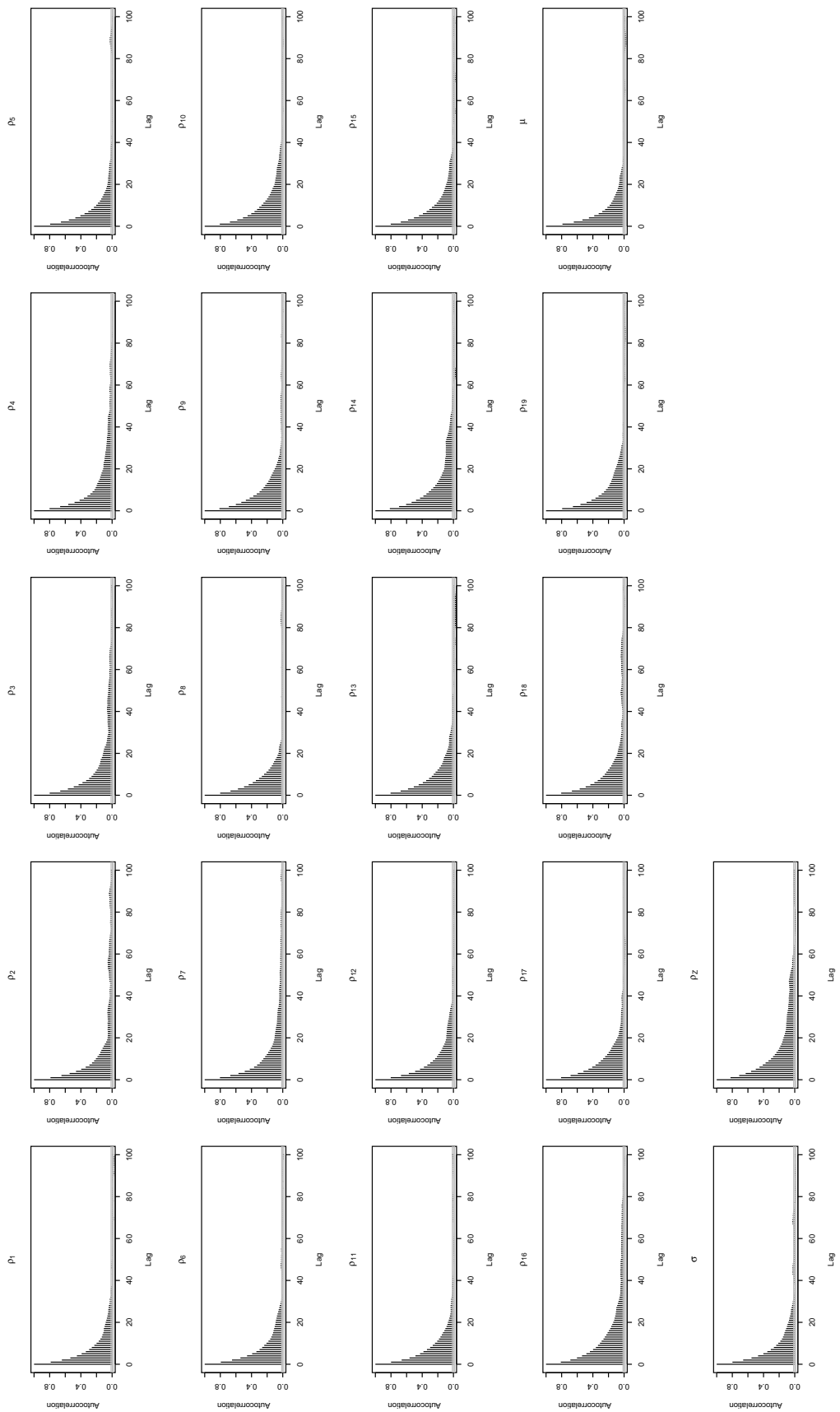


Figure A.38: *ACF* plots of the *MCMC* estimates of trial 13 for $T = 100{,}000$ following a burn-in of $n_0 = 5000$, with thinning interval $m = 5$.

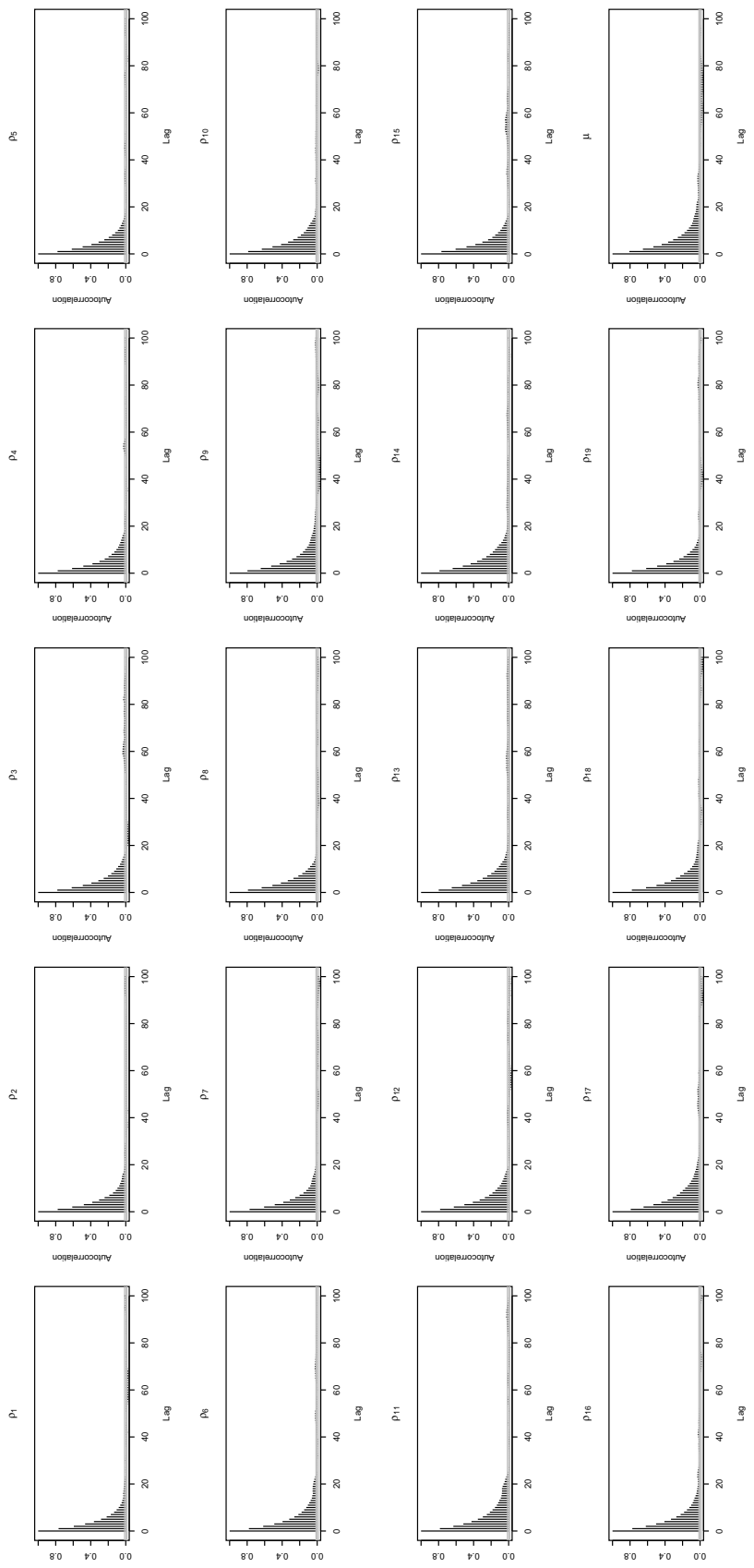


Figure A.39: *ACF* plots of the *MCMC* estimates of trial 14 for $T = 100{,}000$ following a burn-in of $n_0 = 5000$, with thinning interval $m = 5$.

Table A.5: S-RWM parameter estimates and 95% highest density region (HDR) for the data examples of Model VII and Model VIII.

Par	θ	$\widehat{\theta}_{\mathcal{A}_8}[HDR]$ Model VII	$\widehat{\theta}_{\mathcal{A}_8}[HDR]$ Model VIII
ρ_1	0.270	0.299[0.110,0.461]	0.305[0.102,0.509]
ρ_2	0.520	0.451[0.298,0.577]	0.418[0.149,0.721]
ρ_3	0.380	0.266[0.056,0.431]	0.175[-0.019,0.379]
ρ_4	0.190	0.185[0.050,0.305]	0.207[0.065,0.360]
ρ_5	0.200	0.175[0.041,0.315]	0.171[0.017,0.338]
ρ_6	0.130	0.307[0.038,0.531]	0.216[0.019,0.424]
ρ_7	0.080	0.094[0.007,0.195]	0.156[0.011,0.340]
ρ_8	0.170	0.172[0.076,0.271]	0.187[0.041,0.345]
ρ_9	0.230	0.178[0.016,0.340]	0.397[0.243,0.578]
ρ_Z	0.987	0.976[0.943,1.002]	0.965[0.927,0.998]
β_0	6.300	6.442[4.327,8.888]	7.255[5.970,7.816]
β_1	0.600	0.601[0.327,0.899]	0.575[0.309,0.807]
β_{21}	0.140	0.202[-0.110,0.526]	0.153[-0.380,0.438]
β_{22}	0.560	0.551[0.171,1.011]	0.747[0.466,1.068]
β_3	0.010	0.006[-0.044,0.049]	-0.012[-0.022,0.021]
ι	0.620	0.627[0.476,0.809]	0.655[0.544,0.896]
γ_1	0.800	–	0.706[0.544,0.896]
γ_2	0.900	–	0.751[0.569,0.973]

Runtimes are 2.16 hours for Model VII, and 2.17 hours for Model VIII. Averaged squared biases are 0.0048 for Model VII, and 0.0598 for Model VIII.

A.2 Derivations and mathematics

A.2.1 Moments of the log-normal distribution

The initial use of the log-normal distribution traces back to 1931 – in Gibrat's approach concerning individual income, see Gibrat (1931) cited by Kleiber and Kotz (2003b, p. 108). Accordant moments and other basic properties follow directly from the close relationship to the normal distribution. Some selected moments of the log-normal distribution are as follows (Kleiber & Kotz, 2003b, pp. 112ff.):

$$y_{mode} = \exp\{\mu - \sigma^2\},$$

$$y_{(0.5)} = \exp\{\mu\},$$

$$\mathbb{E}[Y] = \exp\left\{\mu + \frac{\sigma^2}{2}\right\},$$

$$Var[Y] = \exp\{2\mu + \sigma^2\}(\exp\{\sigma^2\} - 1),$$

$$y_{skew} = (\exp\{\sigma^2\} + 2)\sqrt{\exp\{\sigma^2\} - 1},$$

$$y_{kurt} = \exp\{\sigma^2\}^4 + 2\exp\{\sigma^2\}^3 + 3\exp\{\sigma^2\}^2 - 3.$$

The log-normal distribution satisfies the mean – median – mode inequality: $\mathbb{E}(Y) > y_{(0.5)} > y_{mode}$. And the Gini coefficient as measure of inequality is given by:

$$GC = 2\Phi\left(\frac{\sigma}{\sqrt{2}}\right) - 1.$$

A.2.2 Moments of the Dagum distribution

Since the Dagum distribution is a special form of the Generalized Beta II distribution (GB2), with one of the scale parameters being set to one ($q = 1$), moments as well as other properties can be directly derived (Kleiber & Kotz, 2003b, pp. 213ff.):

$$y_{mode} = b\left(\frac{ap-1}{a+1}\right)^{\frac{1}{a}}, \quad \text{if } ap > 1,$$

$$y_{(0.5)} = b\left(2^{\frac{1}{p}} - 1\right)^{-\frac{1}{a}},$$

$$\mathbb{E}[Y] = \frac{b(B(p+1/a, 1-1/a))}{B(p,1)} = \frac{b\Gamma(p+1/a)\Gamma(1-1/a)}{\Gamma(p)},$$

$$Var[Y] = \frac{b^2\left[\Gamma(p)\Gamma(p+2/a)\Gamma(1-2/a) - \Gamma^2(p+1/a)\Gamma^2(1-1/a)\right]}{\Gamma^2(p)},$$

$$y_{skew} = \frac{\Gamma^2(p)\lambda_3 - 3\Gamma(p)\lambda_2\lambda_1 + 2\lambda_1^3}{\left[\Gamma(p)\lambda_2 - \lambda_1^2\right]^{\frac{3}{2}}},$$

$$y_{kurt} = \frac{\Gamma^3(p)\lambda_4 - 4\Gamma^2(p)\lambda_3\lambda_1 + 6\Gamma(p)\lambda_2\lambda_1^2 - 3\lambda_1^4}{\left[\Gamma(p)\lambda_2 - \lambda_1^2\right]^2},$$

where $\lambda_i = \Gamma(1 - i/a)\Gamma(p + i/a)$ for $i = 1, 2, 3, 4$. The Dagum distribution also satisfies the mean – median – mode inequality: $\mathbb{E}(Y) > y_{(0.5)} > y_{mode}$. The corresponding Gini coefficient is given by:

$$GC = \frac{\Gamma(p)\Gamma(2p+1/a)}{\Gamma(2p)\Gamma(p+1/a)} - 1.$$

A.2.3 Derivation of the inefficiency factor

Let $\{\theta^{(t)}\}_{t=1}^{T}$ be a sequence of $MCMC$ draws with stationary distribution. The mean of the posterior density is considered as quantity of interest, $h(\theta)$. A simulation consistent estimator of $\mathbb{E}[h(\theta|\mathbf{y})]$ is the mean of the $MCMC$ sample:

$$\bar{h}_T = \frac{1}{T}\sum_{t=1}^{T} h(\theta^{(t)}).$$

Due to the law of large numbers for stationary time series it can be shown that the sample mean converges to the expectation of $h(\theta)$:

$$\bar{h}_T \to \mathbb{E}[h(\theta|\mathbf{y})].$$

An approximation of this convergence rate is needed which is also equivalent to the convergence of the Monte Carlo standard error ($MCSE$) to zero:

$$MCSE(\bar{h}_T) \to 0.$$

The $MCSE$ measures how much error is in the estimate due to the fact that $MCMC$ is used instead of an independence sampler. Because of the slower exploration of the parameter space Θ, more steps of the Markov chain are required to characterize the features of the posterior density sufficiently, as compared to *iid* draws. For an independence sampler, the rate of convergence can be simply expressed by $\sqrt{T}$. Not so for a Markov chain. The stochastic dependence in the Markov chain changes the standard error by adding the autocovariance of consecutive or l-lagged draws.

The *inefficiency factor (Ineff)* , see Chib and Ramamurthy (2010), or *initial (positive) sequence* (IPS) estimator, see Geyer (1992), is calculated as

$$\tau = 1 + 2\sum_{l=1}^{L} \kappa_l$$

where κ_l is the lag-l autocorrelation function (ACF) of the sequence $\{\theta^{(t)}\}_{t=1}^{T}$, for convenience written in the short form h_t. L is the magnitude at which to stop summation. For the IPS estimator, truncation occurs when the sum of adjacent sample ACF values is negative. The factor is derived from the integrated autocorrelation time of h_t:

$$\tau_{int} = \frac{1}{2} + \sum_{l=1}^{\infty} \kappa_l.$$

Let σ^2 denote the variance of h_t, and γ_l is the l-th autocovariance of the sequence h_t. Then

$$\begin{aligned} TVar[\bar{h}_T] &= TVar\left[\frac{1}{T}\sum_{t=1}^{T} h_t\right] \\ &= \frac{1}{T}\left(\sum_{t=1}^{T} Var[h_t] + 2\sum_{i<l} Cov(h_i, h_l)\right) \\ &= \frac{1}{T}\left(T\sigma^2 + 2\sum_{l=1}^{T}(T-l)\gamma_l\right) \\ &= \frac{1}{T}\left(T\sigma^2 + 2T\sigma^2\sum_{l=1}^{T}\frac{T-l}{T}\frac{\gamma_l}{\sigma^2}\right) \\ &= \sigma^2\left(1 + 2\sum_{l=1}^{T}\frac{T-l}{T}\kappa_l\right). \end{aligned}$$

When T is sufficiently large, $\frac{T-l}{T}$ converges to 1 and cancels out yielding an approximative estimate

$$TVar[\bar{h}_T] \approx \sigma^2\left(1 + 2\sum_{l=1}^{\infty}\kappa_l\right).$$

The asymptotic variance of the sample mean can thus be written as follows:

$$Var[\bar{h}_T] \approx \frac{\sigma^2}{T}\tau.$$

When τ is large, autocorrelations slowly decay to zero. In case of *iid* draws, all autocorrelations are zero, hence τ becomes zero. Since in samples produced by Markov chains, $\theta^{(t+1)}$ is more or less dependent on $\theta^{(t)}$, more iterations are needed. The *effective sample size* (ESS) of the chain h_t can be calculated as

$$T^* = \frac{T}{\tau},$$

see Jackman (2009, p. 192f.) and Chib and Ramamurthy (2010, p. 24).

Geyer (1992, p. 476) also refers to a windowing procedure for downweighting large-lag terms as well as *initial monotone sequence* (IMS) estimators and *initial convex sequence* (ICS) estimators, but these approaches are not pursued in this thesis.

A.2.4 Geweke's (1992) test of non-stationarity

Background for monitoring convergence is the failure of $MCMC$ methods to produce iid sequences. Assuming that $\{\theta^{(t)}\}_{t=1}^{T}$ converges in distribution to the limiting density $p(\theta)$ facilitates the assessment of numerical accuracy and convergence. In Geweke's test of non-stationarity (1992) the averages across two stages, say $T_A = [1, \ldots, 0.1T]$ and $T_B = [0.5T, \ldots, T]$, of a $MCMC$ output are compared for a given part of θ, ϑ. Non-stationarity is assumed if these averages are statistically different from each other. Comparison is done with respect to the mean of a specified scalar estimand $h(\vartheta)$, thereby relying on the fact that the spectral density of the time series $\{h(\vartheta^{(t)})\}$ can be used to estimate the asymptotic variance of an estimate of the average of the time series. The estimators of $\mathbb{E}[h(\vartheta)]$ based on T_A and T_B are

$$\bar{h}_A = \frac{1}{T_A}\sum_{t=1}^{T_A} h(\vartheta^{(t)}), \quad \text{and} \quad \bar{h}_B = \frac{1}{T_B}\sum_{t=T-T_B+1}^{T} h(\vartheta^{(t)}),$$

and the asymptotic variances are

$$\mathcal{S}_{h_A}(0)/T_A, \quad \text{and} \quad \mathcal{S}_{h_B}(0)/T_B,$$

with $\mathcal{S}_h$ being the spectral density for this time series at point zero. The square root of this asymptotic variance provides an estimate of the standard error of the mean, also called the *numeric standard error* (NSE), see Geweke (1992) and Cowles and Carlin (1996, p. 866). The difference of the estimates $\bar{h}(\theta)_A$ and $\bar{h}(\theta)_B$ divided by the asymptotic standard error of the difference tends to a standard normal distribution as $T \to \infty$ by the central limit theorem (CLT). For this to hold, two conditions need to be met: the ratios T_A/T and T_B/T are fixed and $T_A + T_B < T$.

Geweke's test criterion requires only a single chain, can be applied to any $MCMC$ method, and is essentially univariate (Cowles & Carlin, 1996, p. 866). The question regarding the optimal sizes of T_A and T_B, for which Geweke only gives suggestions, remains open.

A.2.5 Brooks and Gelman's (1998) convergence criterion

The convergence criterion recommended by S. P. Brooks and Gelman (1998) generalizes the convergence diagnostic of Gelman and Rubin (1992). Gelman and Rubin (1992) advise to use different starting values from an overdispersed proposal for the generation of several chains. *Overdispersion* means that the variance of the proposal is attempted to be greater than the target distribution, though, not spread too widely (Gelman & Rubin, 1992, p. 458f.). This approach is particularly expected to uncover multi-modality, because single chain methods run the risk of sticking at one of the modes, hence neglecting further exploration of the posterior distribution. The output of such a multiple $MCMC$ is compared according to the between-chain ($\mathcal{B}$) and within-chain ($\mathcal{W}$) variation for each scalar component of θ, denoted as ϑ. Conditioned on the observed data the marginal posterior variance of ϑ, for any finite chain length, can be estimated by

$$\widehat{Var}[\vartheta|\mathbf{y}] = \frac{T-1}{T}\mathcal{W} + \frac{1}{T}\mathcal{B},$$

As the $MCMC$ sample size T increases and goes to infinity, the contribution of $\mathcal{B}$ gets smaller and the contribution of $\mathcal{W}$ gets greater, since $(T-1)/T \to 1$ and $1/T \to \infty$. The proposed convergence diagnostic is expressed as a variance ratio

$$\widehat{R} = \frac{\widehat{Var}[\vartheta|\mathbf{y}]}{\mathcal{W}}.$$

The quantity $\sqrt{\widehat{R}}$ can be interpreted as the *potential scale reduction factor (PSRF)* or *shrink factor* which declines to 1 as $T \to \infty$. Large values of $\widehat{R}$ suggest that the marginal posterior variance can be further deceased by more simulations, whereas values of the *PSRF* close to 1 indicate that the r different chains of length T are close to the target distribution.

In the generalization of S. P. Brooks and Gelman (1998, p. 435), graphical inspection methods are added, the scale reduction factor is corrected as well as generalized, and the diagnostic is extended to multivariate summaries. First, the authors recommend an iterated graphical approach (id., pp. 438ff.), but this will not be discussed further here, because it becomes impractical given a large number of parameters (id., p. 440). Second, S. P. Brooks and Gelman (1998, p. 437) present a new factor to correctly account for sampling variability in the variance estimates. The correction factor $(df+3)/(df+1)$ should be used instead of $df/(df-2)$, as recommended by Gelman and Rubin (1992, p. 465), where df is the estimated degrees of freedom, the key parameter for adjustment. Assuming a Student-t-distribution for the posterior inference, df can be estimated my the method of moments: $df \approx 2\widehat{Var}(\vartheta|\mathbf{y})/\widehat{Var}[\widehat{Var}[\vartheta|\mathbf{y}]]$. The corrected *PSRF* is hence defined as

$$\widehat{R}_{corr} = \frac{df+3}{df+1}\frac{\widehat{Var}[\vartheta|\mathbf{y}]}{\mathcal{W}}.$$

In addition, an alternative interpretation of the $\widehat{R}$ diagnostic is introduced that bases on the (squared) ratio of interval lengths opposed to the variance ratio, see S. P. Brooks and Gelman (1998, p. 441). This method avoids the assumption of normality of the marginal distribution of each ϑ and does not require second moments making it appealing for usage. Third, and most important, S. P. Brooks and Gelman (1998, pp. 445ff.) provide multivariate extensions of the approach. Both, $\mathcal{W}$ and $\mathcal{B}$ are now d-dimensional within and between-sequence covariance matrix estimates of the d-variate functional ϑ. And the distance between $\widehat{Var}$ and $\mathcal{W}$ is summarized with a scalar measure which approaches 1 as convergence is achieved. The *multivariate PSRF (MPSRF)* is expressed by means of the maximum root statistic as

$$\widehat{R}^d = \max_a \frac{a'\widehat{Var}a}{a'\mathcal{W}a}.$$

The *MPSRF* $\widehat{R}^d$ is used as an approximate upper bound to the largest of the univariate *PSRF* $\widehat{R}$ statistics over all d variables (id., p. 447). With regard to the *MPSRF*, convergence might be diagnosed later than by the *PSRF*, but this is due to the lack of convergence attributed to the interaction between the single scalar estimands (id., p. 448).

In summary, S. P. Brooks and Gelman's convergence test requires multiple chains which makes it computational more expensive, but it can be applied to any *MCMC* method. Cowles and Carlin (1996, p. 885) further criticise the reliance on normal approximation for the scalar of interest.

A.2.6 Marginal likelihood estimation according to Chib and Jeliazkov (2001)

The marginal likelihood is also called the integrated likelihood or the prior predictive distribution of the data $\mathbf{z}$

$$p(\mathbf{z}) = \int_{\Theta} p(\mathbf{z}|\theta)p(\theta)d\theta.$$

Efforts in calculation of the marginal likelihood can be summarized by the words of (Frühwirth-Schnatter, 2006, p. 125): "The computation of the marginal likelihood for a finite mixture model is quite a challenge." Multiple approaches exist, but here, the methods of Chib and Jeliazkov (2001) and Chib and Ramamurthy (2010) are described in more detail. Both methods are based on Chib's estimator for Gibbs sampler outputs, see Chib (1995). Chib and Jeliazkov (2001) adapted it to Metropolis-Hastings outputs and to multiple parameter blocks. Using Metropolis-Hastings steps instead of full conditionals overcomes the main criticism on Chib (1995) that the posterior modes are not fully explored. Neal (1998) and Frühwirth-Schnatter (2006, p. 161f.) point to the fact that the marginal likelihood estimator is biased because, owing to the full conditionals specified, the posterior (almost) never visits all modes. The more general solution by Chib and Jeliazkov (2001) is based on a representation of the marginal likelihood enabling direct calculation from $MCMC$ output and is described tightly in the following, cp. Chib and Jeliazkov (2001, pp. 271ff.), but see also Chib and Jeliazkov (2005, pp. 32ff.).

By rearrangement of the Bayes theorem, the marginal likelihood can be expressed in the form

$$m(\mathbf{z}|\mathcal{M}_o) = \frac{p(\mathbf{z}|\mathcal{M}_o, \theta)p(\theta|\mathcal{M}_o)}{p(\theta|\mathbf{z}, \mathcal{M}_o)},$$

for a given model $\mathcal{M}_o$, with $o = 1, \ldots, O$. Thus, the basic marginal likelihood identity is the normalizing constant $1/C$ of the posterior density. Since this form is an identity in θ it needs to be evaluated at value θ' to obtain the marginal likelihood. For this to do, it only requires the evaluation of the log-likelihood function, the prior and an estimate of the posterior ordinate $p(\theta'_o|\mathbf{z}, \mathcal{M}_o)$. At a logarithmic scale the marginal likelihood can be estimated from the following identity

$$\log m(\mathbf{z}|\mathcal{M}_o) = \log p(\mathbf{z}|\mathcal{M}_o, \theta'_o) + \log p(\theta'_o|\mathcal{M}_o) - \log p(\theta'_o|\mathbf{z}, \mathcal{M}_o). \tag{A.11}$$

For simplicity, the notation indicating specific models is omitted subsequently. To evaluate – with respect to some θ' – an appropriate high-density point in the support of the posterior, e.g. the posterior mode or posterior mean can be selected. In the presence of multi-modality, multiple local posterior mode may exist and it might be possible that no global mode can be found. To avoid this, the posterior

mean is used as appropriate point. The posterior mean has two advantages: it is unique and the error of the mean is known in advance.

The objective is estimating the posterior ordinate $p(\theta'|\mathbf{z})$ given the posterior sample $[\theta^{(1)}, \ldots, \theta^{(T)}]$. The proposal density in the MH step is denoted as $q(\theta, \theta')$, the probability of getting to the posterior mean from any sampled θ. For this, the irreducibility and aperiodicity conditions from the Ergodic Theorem need to be fulfilled. That is, an $MCMC$ is irreducible and aperiodic if it is possible to move from any state to any other state in one step. To ensure convergence it is not necessarily in one step. The multivariate normal distribution usually fits to the requirements. When using such a symmetric proposal density, the MH step simplifies to a Metropolis step, i.e. $q(\cdot)$ cancels out in the probability of move. With respect to the posterior ordinate, $q(\cdot)$ has to be taken into account thus yielding to the following formula when considering a single-block sampling approach:

$$p(\theta'|\mathbf{z}) = \frac{\int \alpha\left(\theta, \theta'|\mathbf{z}\right) q(\theta, \theta') p(\theta|\mathbf{z}) d\theta}{\int \alpha(\theta', \theta|\mathbf{z}) q(\theta', \theta) d\theta} \tag{A.12}$$

for any point θ'. Let $\{\theta^{(t)}\}_{t=1}^{T}$ denote the sampled draws from the posterior distribution and $\{\theta^{(j)}\}_{j=1}^{J}$ the sampled draws from the proposal distribution, both given the fixed value θ'. A simulation-consistent estimate of the posterior ordinate is then available as

$$\widehat{p}(\theta'|\mathbf{z}) = \frac{T^{-1} \sum_{t=1}^{T} \alpha(\theta^{(t)}, \theta'|\mathbf{z}) q(\theta^{(t)}, \theta')}{J^{-1} \sum_{j=1}^{J} \alpha(\theta', \theta^{(j)}|\mathbf{z})} \tag{A.13}$$

which can be calculated as soon as the $MCMC$ sampling is finished. Only the sampling of the $\theta^{(j)}$ by reduced runs from $q(\theta', \theta)$ for the summands in the denominator requires some additional amount of time.

In Equation (A.13) it can be seen that proposal densities which do not fit to the Ergodicity conditions might hamper estimation of the posterior ordinate. For example, when using the uniform distribution it might be the case that it is not possible to reach θ' from any θ in one step at each iteration due to the fixed boundaries $(-\delta, +\delta) = (\theta^{(t)} - \epsilon, \theta^{(t)} + \epsilon)$. Especially in multi-parameter settings it is highly possible that at least one of the components does not have a positive probability. Hence, $q(\theta, \theta')$ becomes zero most of the time. In the consequence, the numerator will become zero at all. The denominator (in part) can become zero as well if some values $\theta^{(j)}$ do not lie in the support of the target Θ. Here, however, it does not matter, because those are included as $\alpha(\theta', \theta^{(j)}|\mathbf{z}) = 0$ in the average of the denominator.

Multiple parameter blocks might be more convenient and efficient in higher-dimensional problems. With respect to multiple-block settings the posterior ordinate at θ' is denoted as $p(\theta'_1, \ldots, \theta'_K|\mathbf{z})$ according to the K fixed blocks of the parameter values $\theta = (\theta_1, \ldots, \theta_K)$, with $\theta_k \in \Theta_k$. Given $\theta_{-k} \equiv (\Psi_{k-1}, \Psi^{k+1})$, with $\Psi_{k-1} = (\theta_1, \ldots, \theta_{k-1})$ and $\Psi^{k+1} = (\theta_{k+1}, \ldots, \theta_K)$ being the parameter blocks below and beyond k, the posterior ordinate can be decomposed by the law of total

probability into

$$\begin{aligned} p(\theta'|\mathbf{z}) = p(\theta'_1,\ldots,\theta'_K|\mathbf{z}) &= p(\theta'_1|\mathbf{z})p(\theta'_2|\mathbf{z},\theta'_1)\ldots p(\theta'_K|\mathbf{z},\theta'_1,\ldots,\theta'_{K-1}) \\ &= \Pi_{k=1}^{K} p(\theta'_k|\mathbf{z},\theta'_1\ldots,\theta'_{k-1}) \\ &= \Pi_{k=1}^{K} p(\theta'_k|\mathbf{z},\Psi'_{k-1}), \end{aligned} \tag{A.14}$$

with the typical reduced ordinate being $p(\theta'_k|\mathbf{z},\theta'_1\ldots,\theta'_{k-1}) = p(\theta'_k|\mathbf{z},\Psi'_{k-1})$. At this point consider the estimation of the reduced ordinate instead of the product.

Suppose that the block posterior $p(\theta_k|\mathbf{z},\theta_{-k})$ is approx. the complete posterior $p(\theta|\mathbf{z})$ and the proposal density for each block is given as $q(\theta'_k,\theta_k|\Psi_{k-1},\Psi^{k+1})$, then the probability of move, i.e. the transition from θ_k to θ^*_k, for each fixed block, is

$$\begin{aligned} \alpha(\theta_k,\theta^*_k|\Psi_{k-1},\Psi^{k+1}) = \min\Bigg\{1, &\frac{p(\mathbf{z}|\theta^*_k,\Psi_{k-1},\Psi^{k+1})p(\theta^*_k,\theta_{-k})}{p(\mathbf{z}|\theta_k,\Psi_{k-1},\Psi^{k+1})p(\theta_k,\theta_{-k})} \\ &\cdot\frac{q(\theta^*_k,\theta_k|\Psi_{k-1},\Psi^{k+1})}{q(\theta_k,\theta^*_k|\Psi_{k-1},\Psi^{k+1})}\Bigg\} \end{aligned}$$

Again assuming a Metropolis step, then the ratio of the proposal densities cancels out due to symmetry. The simulation-consistent estimate of the posterior ordinate for each fixed block is now

$$\begin{aligned} &\widehat{p}(\theta'_k|\mathbf{z},\theta'_1,\ldots,\theta'_{k-1}) = \widehat{p}(\theta'_k|\mathbf{z},\Psi'_{k-1}) = \\ &\frac{T^{-1}\sum_{t=1}^{T}\alpha\left(\theta_k^{(t)},\theta'_k|\mathbf{z},\Psi'_{k-1},\Psi^{k+1,(t)}\right)q\left(\theta_k^{(t)},\theta'_k|\Psi'_{k-1},\Psi^{k+1,(t)}\right)}{J^{-1}\sum_{j=1}^{J}\alpha\left(\theta'_k,\theta_k^{(j)}|\mathbf{z},\Psi'_{k-1},\Psi^{k+1,(j)}\right)}. \end{aligned} \tag{A.15}$$

Chib and Jeliazkov (2001, p. 273) formulate four steps for calculation of the marginal likelihood, see Algorithm 6. As Chib and Jeliazkov (2001) recommend, the reduced runs are obtained by fixing an appropriate set of parameters. In this thesis, the parameters of the underlying distribution $\psi = [\mu,\sigma,\rho_Z]$ are fixed to the true parameter values. The constraints are implemented as in the RWM runs by a simple rejection sampling method. This means that samples first are checked if they fit to the constraint system and either kept and used for the Metropolis step, or rejected immediately.

Algorithm 6: Marginal likelihood estimation from multiple-block proposal density

Step 1 Set $\Psi_{k-1} = \Psi'_{k-1}$ and sample $p(\theta_k|\mathbf{z},\theta_{-k})$ for each $k = k,\ldots,K$ which yields the generated draws $\{\theta_k^{(t)},\ldots,\theta_K^{(t)}\}_{t=1}^{T}$.

Step 2 Include θ'_k in the conditioning set and let $\Psi'_k = (\Psi'_{k-1},\theta'_k)$. Then remove the full conditional distribution of θ_k from Step 1 and sample the remaining distributions $p(\theta_k|\mathbf{z},\theta_{-k})$, $k = k+1,\ldots,K$. Now yielding the generated draws $\{\theta_{k+1}^{(j)},\ldots,\theta_K^{(j)}\}_{j=1}^{J}$. Also draw $\theta_k^{(j)}$ from $q\left(\theta'_k,\theta_k|\Psi'_{k-1},\Psi^{k+1,(j)}\right)$ at each step of sampling.

Step 3 Estimate the reduced ordinate by Equation (A.15).

Step 4 The marginal likelihood estimate on the log-scale is then

$$\log \widehat{m}(\mathbf{z}) = \log p(\mathbf{z}|\theta') + \log p(\theta') - \sum_{k=1}^{K} \log \widehat{p}(\theta'_k|\mathbf{z}, \theta'_1, \ldots, \theta'_{k-1}). \qquad \text{(A.16)}$$

The numerical standard error (NSE) of the marginal likelihood estimate gives the expected variation for repeated simulations and can be calculated according to Chib and Jeliazkov (2001, p. 274). The NSE of the log-marginal likelihood estimate will be derived from the SD resulting from 100 repetitions that have been performed. As already mentioned by Chib and Jeliazkov (2001, p. 274), the estimate of the NSE closely resembles the SD of the log-marginal likelihood estimates stemming from such a frequency analysis.

Chib and Ramamurthy (2010, p. 29) made a further extension of the framework of Chib and Jeliazkov (2001) to accommodate the randomized block sampling strategy. Modifications include: fixing the number of blocks to the average number $\bar{K}$ realized in a RMB-RWM run for estimation of the posterior ordinate, and construction $\bar{K}$ parameter blocks by randomly assigning components of θ. As in the approach for the fixed multiple parameter blocks the posterior ordinate can be decomposed, see Equation (A.14), and the typical ordinate then can be estimated according to Chib and Jeliazkov (2001), see Algorithm 7.

Algorithm 7: Marginal likelihood estimation from randomized multiple-block proposal density

Step 1 Construct $\bar{K}$ blocks with randomly assigned components of θ.

Step 2 Repeat for $k = 1, \ldots, \bar{K}$:

Substep Repeat for $t = 1, \ldots, T$:

I Use $p(\theta_k, \Psi^{k+1}|\mathbf{z}, \Psi'_{k-1})$ to generate the RMB-RWM draws $\theta_k^{(t)}, \Psi^{k+1,(t)}$. The parameters from the preceding block Ψ_{k-1} are held fixed at Ψ'_{k-1}. Thus, randomizing is only over the parameters in Ψ^{k+1}. Then calculate the k-th stage numerator summand in Equation (A.15).

II Calculate the $(k-1)$-th stage denominator summand in Equation (A.15) by supplementing the preceding draw with a draw $\theta_{k-1}^{(t)}$ from $q(\theta'_{k-1}, \theta_{k-1}|\mathbf{z}, \Psi'_{k-2}, \Psi^{k,(t)})$.

III Store the values.

Step 3 Draw T values $\{\theta_{\bar{K}}\}$ from $q(\theta'_{\bar{K}}, \theta_{\bar{K}}|\mathbf{z}, \Psi'_{\bar{K}-1})$.

Step 4 The marginal likelihood estimate on the log-scale is then A.16.

Appendix B

Sources

B.1 R Code for ML estimation

```
##############################################################################
## ESTIMATE HEAPING MODEL WITH PIECEWISE CONSTANT HEAPING PROBABILITIES      ##
##                                                                          ##
## Add-on material to paper:                                                ##
## "A Statistical Approach to Address the Problem of Heaping in             ##
##  Self-Reported Income Data", CJAS, doi:10.1080/02664763.2015.1077372     ##
## Code by Zinn, S. (February 2015), Adapted by Wuerbach, A. (March 2015)   ##
##                                                                          ##
##############################################################################

# load libraries
library(numDeriv)
library(maxLik)

# log-likelihood for ML (sequence of parameters changed) ----------------------

### for log-normal
G_LN <- function(z_i,param) {

fval <- dlnormZ(z_i, pZ=param[3], logmu=param[1], logsd=param[2])
intAll <- matrix(HPs[z_i > HPs[,2] & z_i < HPs[,3],],
                 ncol=ncol(HPs), byrow=FALSE)
G1 <- fval*ifelse(nrow(intAll)==0,1,1-sum(param[-c(1,2,3)][intAll[,4]]))
if(z_i %in% HPs[,1]){
   intH <- HPs[HPs[,1] %in% z_i,]
   G2 <- param[-c(1,2,3)][intH[4]]*param[3]*
               (plnorm(intH[3],meanlog=param[1],sdlog=param[2])-
                plnorm(intH[2],meanlog=param[1],sdlog=param[2]))
  } else {
    return(G1)
  }
return(G1+G2)
}
```

```
LLIKE_LN <- function(pars) {
   E <- sum(log(unlist(apply(matrix(simdata, ncol=1), 1, G_LN, param=pars))))
   return(E)
}

### for Dagum
G_Dag <- function(z_i,param) {

fval <- dDagumZ(z_i, para=param[1], parb=param[2], parp=param[3], pZ=param[4])
intAll <- matrix(HPs[z_i > HPs[,2] & z_i < HPs[,3],],
                 ncol=ncol(HPs), byrow=FALSE)
G1 <- fval*ifelse(nrow(intAll)==0,1,1-sum(param[-c(1:4)][intAll[,4]]))
if(z_i %in% HPs[,1]){
   intH <- HPs[HPs[,1] %in% z_i,]
   G2 <- param[-c(1:4)][intH[4]]*param[4]*
        (pdagum(intH[3], shape1.a=param[1], scale=param[2], shape2.p=param[3])-
         pdagum(intH[2], shape1.a=param[1], scale=param[2], shape2.p=param[3]))
  } else {
    return(G1)
  }
  return(G1+G2)
}

LLIKE_Dag <- function(pars) {
  E <- sum(log(unlist(apply(matrix(simdata, ncol=1), 1, G_Dag, param=pars))))
  return(E)
}

# specify constraint system ----------------------------------------------------

determineConsSys_pcm <- function(parz=3) {

  findProbIndS <- function(z_i) {
    return(HPs[z_i > HPs[,2] & z_i < HPs[,3],4])
  }
  isAlreadyIn <- function(eL, vec) {
  if(length(vec)==0)
     return(FALSE)
    for(i in 1: length(vec)) {
      c0 <- as.numeric(unlist(strsplit(vec[i],split="-")))
      if(sum(c0 %in% eL)==length(eL))
      return(TRUE)
      }
    return(FALSE)
  }
  giveConstr <- function(dataIN) {
    liS <- lapply(dataIN,findProbIndS)
    erS <- unlist(lapply(liS, length))
    li_cS <- liS[which(erS>1)]
    li_sS <- lapply(li_cS,sort)
    li_uS <- unique(lapply(li_sS,paste,collapse="-"))
    constr0 <- c()
```

```
    for(j in 1:length(li_uS)){
        ell <- unlist(li_uS[j])
        isIN <- isAlreadyIn(ell,constr0)
        if(!isIN)
           constr0 <- c(constr0, unlist(lapply(li_uS[j],paste,collapse="-")))
    }
    # counter check
    constr <- c()
    for(j in 1:length(constr0)){
        ell <- as.numeric(unlist(strsplit(unlist(constr0[j]),split="-")))
        isIN <- isAlreadyIn(ell,constr0[-j])
        if(!isIN)
           constr <- c(constr, paste(ell,collapse="-"))
    }
    CoND <- NULL
    for(k in 1:length(constr)){
        p1 <- sort(as.numeric(unlist(strsplit(constr[k],split="-"))))
        p2 <- setdiff(1:nuPr,p1)
        p3 <- c(-1*as.numeric(table(p1)), rep(0, nuPr-length(unique(p1))))
        p4 <- rbind(c(as.numeric(names(table(p1))), p2) ,p3)
        p5 <- p4[,order(p4[1,])]
        CoND <- rbind(CoND, p5[2,])
    }
    return(CoND)
  }
  nuPr <- 19
  constrA <- giveConstr(simdata)
  part1 <- rbind(diag(1,parz), c(rep(0,parz-1),-1),
                 matrix(0, ncol=parz, nrow=nrow(constrA)+nuPr))
  part2 <- matrix(0, nrow=parz+1, ncol=nuPr)
  part3 <- rbind(diag(1,nuPr),constrA)
  part4 <- cbind(matrix(0,ncol=parz, nrow=nuPr),diag(-1,nuPr))
  Amat <- rbind(cbind(part1, rbind(part2, part3)),part4)
  bvec <- c(rep(0,parz), 1, rep(0,nuPr),rep(1,nrow(constrA)),rep(1,nuPr))
  return(list(Amat=Amat, bvec=bvec))
}

# find maximum of log-likelihood function ------------------------------------

MLestFun <- function(inits=c(c(7.714,0.839,0.99),rep(0.2,19)),distr="inflLN") {

if(distr=="inflLN") {
sol_V <- maxNM(LLIKE_LN, start=inits, constraints=list(ineqA=Amat,ineqB=bvec),
               iterlim = 10000, tol = 1e-06)
} else if(distr=="inflDagum") {
sol_V <- maxNM(LLIKE_Dag, start=inits, constraints=list(ineqA=Amat,ineqB=bvec),
               iterlim = 10000, tol = 1e-06)
}
FisherInf <- sol_V$gradient %o% sol_V$gradient
invH <- qr.solve(-sol_V$hessian)
vars <- round(diag(invH%*%FisherInf%*%invH),15)
# Huber-White standard errors
```

```
stds <- sqrt(vars)
alpha <- 0.05
ci_l <- sol_V$estimate - qt(1-alpha/2,length(simdata))*stds
ci_u <- sol_V$estimate + qt(1-alpha/2,length(simdata))*stds
MLr <- cbind(sol_V$estimate, ci_l, ci_u, stds)
colnames(MLr) <- c("estim", "lower CI", "upper CI", "stds")
return(list(ML=sol_V$maximum, MLresults=MLr))
}
```

B.2 R Code for RWM estimation

B.2.1 Log-Likelihood

```
##############################################################################
# Evaluation of the log-likelihood                                           ##
# with zero-inflated log-normal distribution                                 ##
##############################################################################

library(MASS)
library(mnormt)
options(scipen=10)

# zero-inflated log-normal distribution ------------------------------------

dlnormZ <- function(x, pZ, logmu, logsd) {
  return(ifelse(x==0, (1-pZ), pZ*dlnorm(x, meanlog=logmu, sdlog=logsd)))
}

# likelihood for ONE observation and underlying log-normal distribution -------

g <- function(z_i,param) {
  # z_i = for which value to be evaluated
  # param = input as vector -> parameters of 'g' (p_j/logmu,logsd,pZ)
  l <- length(param)-3

  fval <- dlnormZ(z_i, pZ=param[l+3], logmu=param[l+1], logsd=param[l+2])
  intAll <- matrix(HPs[z_i>HPs[,2] & z_i<HPs[,3],], ncol=4, byrow=FALSE)
  ud <- c(l+1,l+2,l+3)
  g1 <- fval*ifelse(nrow(intAll)==0,1,1-sum(param[-ud][intAll[,4]]))
  if(z_i %in% HPs[,1]) {
    intH <- HPs[HPs[,1] %in% z_i,]
    g2 <- param[-ud][intH[4]]*param[l+3]*
                (plnorm(intH[3],meanlog=param[l+1],sdlog=param[l+2])-
                 plnorm(intH[2],meanlog=param[l+1],sdlog=param[l+2]))
  } else {
    return(g1)
  }
  return(g1 + g2)
}
```

```
# calculate log-likelihood ------------------------------------------------

llike <- function(pars) {
  E <- sum(log(apply(matrix(simdata, ncol=1), 1, g, param=pars)))
  return(E)
}
```

B.2.2 Constraints

```
##############################################################################
# Function to check the probabilities according to their sum (29.12.2014)   ##
# Extended for bell-shaped heaping function (15.06.2015)                     ##
##############################################################################

getIntSum <- function(val, HP, pss) {
  # val = point of evaluation
  # HP = corresponding to val
  # pss = probabilities to check
  set <- HP[which(val > HP[,2] & val < HP[,3]),, drop=FALSE]
  pset <- set[,4]
  return(sum(pss[pset]))
}

probnorFun <- function(val, HP, pss) {
  intv <- matrix(HP[val > HP[,2] & val < HP[,3],], ncol=4, byrow=FALSE)
  if(dim(intv)[1]==0)   probNoR <- 1
  else {
    probNoR <- 1
    for(i in 1:dim(intv)[1]) {
      l <- intv[i,2]
      u <- intv[i,3]
      ps <- intv[i,4]
      pr <- pss[ps]
      probNoR <- round(probNoR - pr,6)
    }
  }
  return(probNoR)
}

getEtaSum <- function(val, HP, etas) {
  # val = point of evaluation
  # HP = corresponding to val
  # etas = probabilities to check
  set <- HP[which(val > HP[,2] & val < HP[,3]),, drop=FALSE]
  etab <- rep(NA,nrow(set))
  if(nrow(set) > 0) {
    for(i in 1:nrow(set)) {
      conset <- set[i,]
      hb <- conset[1]
```

```
      xi <- conset[5]
      posM <- conset[4]
      etab[i] <- exp(-2*((val-hb)^2)/(xi^2))
    }
    sume <- sum(etab*etas[set[,4]])
  } else sume <- 0
  return(sume)
}

# give sum and/or break ---------------------------------------------------

### for piecewise constant heaping probabilities
check_p_Fun <- function(HP=HPs, simdata=simdata, pss=pss) {
  checkp <- apply(matrix(simdata,ncol=1), 1, getIntSum, HP=HPs, pss=pss)
  if(any(checkp > 1)) {
     cat("\n --- Sum of rhos exceeds one. --- \n")
     return(TRUE)
  } else return(FALSE)
}

### for piecewise bell-shaped heaping probabilities
check_eta_Fun <- function(HP=HPs, simdata=simdata, etas=etas) {
  checketa <- apply(matrix(simdata,ncol=1), 1, getEtaSum, HP=HPs, etas=etas)
  if(any(checketa > 1)) {
    cat("\n --- Sum of etas exceeds one. --- \n")
    return(TRUE)
  } else return(FALSE)
}
```

B.2.3 RWM algorithm

```
################ Random-walk Metropolis-Hastings algorithms ###################
## 1) define starting values for theta                                       ##
## 2) draw theta* from jumping distribution J_t(theta*|theta_t-1) at         ##
##    iteration t, with multi-block proposal for varying sets of p's         ##
## 3) compute acceptance ratios (probability) r for p's                      ##
## 4) accept theta* as theta(t) definitely if the candidate has higher       ##
##    probability than our current draw, otherwise accepted according to     ##
##    probability ratio r in the contrary case stick to theta(t-1)           ##
## 5) repeat T times                                                         ##
## 6) consider burn-in and/or thinning                                       ##
###############################################################################

# functions ---------------------------------------------------------------

unifdrawFun <- function(x) {
  if (x[1] != x[2]) runif(1,x[1],x[2])
  else cat("Problem: del is zero, no interval to draw from.")
}
```

```
psFun <- function(x) paste("p",x,sep="")
lp <- 0
lps <- NULL
GMprop <- NULL
repeatc <- 0

blockingFun <- function(p,ud=0) {
  x <- seq(1:p)
  n <- sample(seq(1:p),1)
  blocks <- split(x, factor(sample(n,p,replace=TRUE)))
  if(ud==0) BLOCK <- lapply(blocks,psFun)
  if(ud==1) BLOCK <- c(lapply(blocks,psFun),list(UD=c("p20","p21","p22")))
  if(ud==3) BLOCK <- c(lapply(blocks,psFun),UD=c("p20","p21","p22"))
  return(BLOCK)
}

# define posterior density ------------------------------------------------

logpost <- function(BLOCK,thetax,init,GM) {
  LLike <- l.like(BLOCK,thetax)

  if(!is.null(LLike)) {
  # posterior = likelihood times prior
    LLike + dmnorm(thetax, mean=init, varcov=GM, log=TRUE)
    }
  }

# build the random walk metropolis ----------------------------------------

RWMA <- function(nit=10000,burnin=0,thin=1,meth="unif",eps=0.01,update=NULL,
                 lambda=100,omega=NULL,dtrue=NULL,delta=0,psin=rep(0.2,19),
                 udin=NULL,BLOCK=NULL,FUN=NULL,HFun="pcm") {
  # nit = number of iterations (in total)
  # burnin = number of iterations to discard
  # thin = sequence for thinning
  # meth = density for proposal
  # eps = epsilon = boundary widths (for meth="unif")
  # update = whether to update variance-covariance matrix (for meth="mvnorm")
  # lambda = scale factor for unity matrix
  # omega = prior covariance for parameters of the underlying distribution
  # delta = specification for proposal covariance
  # init = initialized theta (ps~ for p and ud~ for underlying distribution)
  if(burnin >= nit) stop("Iterations for Burn-in greater than total.")
  lep <- length(psin)
  lepu <- length(c(psin,udin))
  # starting values
  VC <- 1/lambda*diag(lep)
  # check dimensions of theta and variance-covariance-matrix
  if(nrow(VC) != ncol(VC) | nrow(VC)!= lep)
     stop("VC is not of appropriate dimension! Check init and VC.")
  # check if variance-covariance-matrix is positive definite
  cd <- NULL
```

```
try(cd <- chol(VC), silent=TRUE)
if(is.null(cd)) stop("VC is not positive definite!")
Ups <- omega*diag(length(udin))
GMdiag <- c(rep(1/lambda,lep),omega)
GM <- GMdiag*diag(lepu)
GMprop <<- delta*diag(lepu)
eps <- c(rep(eps,lep),omega)

# set auxilarily matrices and other objects
td <- matrix(NA,nrow=nit+burnin,ncol=lepu)
repeatp <- 0
repeatm <- rep(NA,sum(nit,burnin))
acr <- matrix(NA,nrow=nit+burnin,ncol=lepu)
a <- matrix(NA,nrow=nit+burnin,ncol=lepu)
theta_cur <- rep(NA,lepu)
if(is.null(BLOCK)) { lplist <- matrix(NA,nrow=nit+burnin,ncol=1)
} else lplist <- matrix(NA,nrow=nit+burnin,ncol=length(BLOCK))

# priors for theta --> p
repeat{ # be aware that the priors fit the restrictions!
theta_cur[1:lep] <- rmnorm(1,mean=psin,varcov=VC)
if(any(theta_cur[1:lep] <= 0 |
       theta_cur[1:lep] >= 1) == FALSE) {
   if(HFun=="pcm") {
      if(check_p_Fun(HPs,simdata,
                pss=theta_cur[1:lep]) == FALSE) { break
      } else repeatp <<- repeatp+1
   } else if(HFun=="pbsm") {
      if(check_eta_Fun(HPs,simdata,
                etas=theta_cur[1:lep]) == FALSE) { break
      } else repeatp <<- repeatp+1
   }
} else repeatp <<- repeatp+1
if(repeatp==100) {
   print(theta_cur)
   stop("Problem: repeatp reached limit 100.\n")
   }
}

# priors for theta --> psi=(mu,sd,pZ)
if(!is.null(udin)) { # sample
theta_cur[(lep+1):lepu] <- abs(rmnorm(1,mean=udin,varcov=Ups))
   if(theta_cur[lepu] > 1) {
      theta_cur[lepu] <- 1-diff(c(1,theta_cur[lepu]))
      }
} else {theta_cur[lep+1] <- dtrue[1]
        theta_cur[lep+2] <- dtrue[2]
        theta_cur[lep+3] <- dtrue[3]} # otherwise fix psi

# evaluate likelihood for the start values
lp <<- logpost(paste("p", seq(1:lepu), sep=""),
               theta_cur,init=c(psin,udin),GM)
```

```
# define when to update theta
theta_update <- function(thetacur,lp,psin,udin=NULL,meth,eps,
                         GM,GMprop,BLOCK) {
   #*****************************************************************
   # for each block separately
   for (j in 1:length(BLOCK)) {

        pos <- as.numeric(substr(BLOCK[[j]], 2, 3))
        neg <- which(!(seq(1:length(thetacur)) %in% pos))
        thetacan <- thetacur
        logpost_can <- Inf

   while(is.null(logpost_can) | abs(logpost_can)==Inf) {
     repeatw <- 0

   # 2) draw theta* from jumping distribution J_t
   #*****************************************************************
   # J_t as "uniform-in-each-direction" proposal
   #-----------------------------------------------------------------
   if(meth=="unif") {
      del <- matrix(NA,nrow=length(thetacur[pos]),ncol=2)
      del[,1] <- thetacur[pos] - eps[pos]
      del[,2] <- thetacur[pos] + eps[pos]

      repeat{ # be aware that the candidates fit the restrictions!
         thetacan_b <- apply(del,1,unifdrawFun)
         thetacan[pos] <- thetacan_b
         if(any(thetacan[1:lep] <= 0 |
                thetacan[1:lep] >= 1) == FALSE &
            any(thetacan[(lep+1):length(thetacan)]
                <= 0) == FALSE &
            thetacan[length(thetacan)] <= 1) {
            if(HFun=="pcm") {
               if(check_p_Fun(HPs,simdata,
                        pss=thetacan[1:lep]) == FALSE) { break
               } else repeatc <<- repeatc+1
            } else if(HFun=="pbsm") {
               if(check_eta_Fun(HPs,simdata,
                        etas=thetacan[1:lep]) == FALSE) { break
               } else repeatc <<- repeatc+1
            }
         } else repeatc <<- repeatc+1
         if(repeatc==100) {
            print(thetacan)
            stop("Problem: repeatc reached limit 100.\n")
           }
         }

   # J_t as multivariate normal proposal
   #-----------------------------------------------------------------
   } else if(meth=="mvnorm") {
```

```
        repeat{ # be aware that the candidates fit the restrictions!
           thetacan_b <- rmnorm(1,mean=as.vector(thetacur[pos]),
                                varcov=GMprop[pos,pos])
           thetacan[pos] <- thetacan_b
           if(any(thetacan[1:lep] <= 0 |
                     thetacan[1:lep] >= 1) == FALSE &
              any(thetacan[(lep+1):length(thetacan)] <= 0) == FALSE &
              thetacan[length(thetacan)] <= 1) {
              if(HFun=="pcm") {
                 if(check_p_Fun(HPs,simdata,
                           pss=thetacan[1:lep]) == FALSE) { break
                 } else repeatc <<- repeatc+1
              } else if(HFun=="pbsm") {
                 if(check_eta_Fun(HPs,simdata,
                           etas=thetacan[1:lep]) == FALSE) { break
                 } else repeatc <<- repeatc+1
              }
           } else repeatc <<- repeatc+1
           if(repeatc==1000) {
              print(thetacan)
              stop("Problem: repeatc reached limit 1000.\n")
              }
           }
     }

     # 3) compute acceptance ratio (probability) r
     #*****************************************************************
     logpost_can <- logpost(BLOCK[[j]],thetacan,init=c(psin,udin),GM)

     repeatw <- repeatw+1
     if(repeatw==10) stop("Problem: while loop reached limit 10.\n")
     } # end while-loop for logpost_can

     lacceptprob <- logpost_can - lp

     # 4) accept or reject theta* as theta(t)
     #*****************************************************************
     if(log(runif(1)) < lacceptprob) {
        thetacur <- thetacan
        lp <<- logpost_can
        }
     lps <<- c(lp,logpost_can)
     } # end for-loop for blocks

     return(list(thetacur,lps))
     } # end update function

#5) repeat T times
#*****************************************************************
for(i in 1:(nit+burnin))
  {
```

```
repeatc <<- 0
options(warn=-1)

# decision for draws
if(!is.null(FUN)) BLOCK <- eval(FUN)
   theta <- theta_update(thetacur=theta_cur,lp,psin,udin,meth,eps,
                         GM,GMprop,BLOCK)
   options(warn=0)
   if(!is.null(udin)) {
      td[i,] <- theta[[1]]
   } else { hp <- theta[[1]]
            td[i,] <- hp[-((length(hp)-2):length(hp))] }
   # document acceptance ratio
   if(!is.null(udin)) {
      a[i,] <- as.vector(ifelse(theta_cur != theta[[1]], 1, 0))
   } else { ar <- as.vector(ifelse(theta_cur != theta[[1]], 1, 0))
            a[i,] <- ar[-((length(ar)-2):length(ar))]
   }
     if(i==1) { acr[i,] <- a[i,]
        } else  acr[i,] <- colSums(a[1:i,], na.rm=TRUE)/i

   # document repeatings
   repeatm[i] <- repeatc
   theta_cur <- theta[[1]]
   lplist[i,] <- theta[[2]][-1]
   lastlp <- length(theta[[2]])
   lp <<- theta[[2]][lastlp]

   # update VarCov for proposal
   if(!is.null(update)) {
      if(i>=10) {
         if(update=="GS") { # greedy start (crude)
            if(min(colSums(a[1:i,])) %in% seq(50, 500, 10)) {
               accTD <- td[1:i,]
               pos <- which(a[,1]==1)
               GMprop <<- cov(accTD[pos,])
            }
         } else if(update!="thetamean") {
            cf <- 2.38^2/ncol(a)
            if(min(colSums(a[1:i,]))==50) { # greedy start procedure
            up <- matrix(NA,nrow=min(colSums(a[1:i,])),ncol=ncol(a))
            accTD <- td[1:i,]
            pos <- which(a[,1]==1)
            GMprop <<- cf*cov(accTD[pos,])
            }
            if(update=="AP") {
               if(i%%1000==0) {
                  M <- td[c(i-999):i,]
                  tM <- scale(M,center=colMeans(M),scale=FALSE)
                  Rt <- 1/(999)*t(tM)%*%tM
                  GMprop <<- cf*Rt
               }
```

```
                } else if(update=="AM") {
                   M <- td[1:i,]
                   if(i==1000) {
                      GMprop <<- cf*cov(M)+cf*0.0001*diag(ncol(a))
                   } else if(i>1000) {
                      GMprop <<- (i-1)/i*GMprop +
                        cf/i*( i*colMeans(td[1:i-1,])%*%t(colMeans(td[1:i-1,])) -
                               (i+1)*colMeans(M)%*%t(colMeans(M)) +
                                td[i,]%*%t(td[i,]) + 0.0001*diag(ncol(a)) )
                   }
                }
             }
          }
       }
    }

 # output
 #****************************************************************
 # discard burn-in period
 selection <- seq((burnin+1), (nit+burnin), 1)
 # and consider thinning
 sel <- selection[seq(1, length(selection), thin)]
 # give back
 return(list(posterior_draws=data.frame(td[sel,]),
             acceptance_ratio=data.frame(acr[sel,]),
             repeatm=repeatm[sel],repeatp=repeatp,
             logposteriorlist=lplist))
}
```

B.2.4 Call RWM settings

```
# load sources ------------------------------------------------------------

library(snowfall)
path <- paste0("Z:\\02_Abteilung_2\\02_Methoden\\01_MA-Restricted\\AW\\Heaping",
               "\\GIT_R_Codes_Heaping\\GIT_Bayes_est\\")
source(paste0(path,"RWMA_Heaping_estpInflLN_MB_final.R"))
source(paste0(path,"likelihood_estInflLN_final.R"))
source(paste0(path,"checkInt_final.r"))
source(paste0(path,"mcmc_parallel.R"))

# load (simulated) data
pathdata <- paste0("Z:\\02_Abteilung_2\\02_Methoden\\01_MA-Restricted\\AW\\",
                   "Heaping\\Simulierte_Daten\\")
simdata <- unlist(read.table(paste0(pathdata,"simdata_inflLN_pwc_p19_70perc_",
                                    "sym.txt"),sep="\t"))
simdata <- simdata[simdata <= 10000]
HPs <- matrix(unlist(read.table(paste0(pathdata,"HPointssim_inflLN_pwc_p19_70",
                                       "perc_sym.txt"),sep="\t")), ncol=4)
```

```
# call RWMA - all trials ------------------------------------------------------

convlist <- mcmcparallel2(mccores=3, seeds=c(123,65,786),
                          list(
             FUN1=quote(RWMA(nit=11000,burnin=0,thin=1,meth="unif",
                        eps=0.01,lambda=100,
                        dtrue=c(7.72,0.85,0.987),psin=rep(0.2,19),
                        BLOCK=list(SB=paste("p", seq(1:19), sep="")))),

             FUN2=quote(RWMA(nit=11000,burnin=0,thin=1,meth="unif",
                        eps=0.01,lambda=1000,omega=c(0.1,0.01,0.001),
                        dtrue=c(7.72,0.85,0.987),psin=rep(0.2,19),
                        BLOCK=list(SB=paste("p", seq(1:19), sep="")))),

             FUN3=quote(RWMA(nit=11000,burnin=0,thin=1,meth="unif",
                        eps=0.05,lambda=1000,omega=c(0.1,0.01,0.001),
                        dtrue=c(7.72,0.85,0.987),psin=rep(0.2,19),
                        BLOCK=list(SB=paste("p", seq(1:19), sep="")))),

             FUN4=quote(RWMA(nit=11000,burnin=0,thin=1,meth="unif",
                        eps=0.01,lambda=1000,omega=c(0.1,0.01,0.001),
                        dtrue=c(7.72,0.85,0.987),psin=rep(0.1,19),
                        BLOCK=list(SB=paste("p", seq(1:19), sep="")))),

             FUN5=quote(RWMA(nit=11000,burnin=0,thin=1,meth="unif",
                        eps=0.01,lambda=1000,omega=c(0.1,0.01,0.001),
                        dtrue=c(7.72,0.85,0.987),psin=rep(0.3,19),
                        BLOCK=list(SB=paste("p", seq(1:19), sep="")))),

             FUN6=quote(RWMA(nit=11000,burnin=0,thin=1,meth="unif",
                        eps=0.01,lambda=1000,omega=c(0.1,0.01,0.001),
                        psin=rep(0.2,19),udin=c(7.714,0.839,0.99),
                        BLOCK=list(SB=paste("p", seq(1:22), sep="")))),

             FUN7=quote(RWMA(nit=11000,burnin=0,thin=1,meth="unif",
                        eps=0.01,lambda=1000,omega=c(0.1,0.01,0.001),
                        psin=rep(0.2,19),udin=c(7.714,0.839,0.99),
                        BLOCK=list(M1=paste("p", seq(1,7,1), sep=""),
                                   M2=paste("p", seq(8,13,1), sep=""),
                                   M3=paste("p", seq(14,19,1), sep=""),
                                   UD=paste("p", seq(20,22,1), sep="")))),

             FUN8=quote(RWMA(nit=11000,burnin=0,thin=1,meth="unif",
                        eps=0.01,lambda=1000,omega=c(0.1,0.01,0.001),
                        psin=rep(0.2,19),udin=c(7.714,0.839,0.99),
                        BLOCK=list(B1=c("p1","p8"), B2=c("p2","p14"),
                                   B3=c("p3","p9"), B4=c("p4","p15"),
                                   B5=c("p5","p10","p16"),
                                   B6=c("p6","p11","p17"),
                                   B7=c("p7","p12","p18"),
                                   B8=c("p13","p19"),
                                   UD=paste("p", seq(20,22,1), sep="")))),
```

```
            FUN9=quote(RWMA(nit=11000,burnin=0,thin=1,meth="unif",
                       eps=0.01,lambda=1000,omega=c(0.1,0.01,0.001),
                       psin=rep(0.2,19),udin=c(7.714,0.839,0.99),
                       FUN=call("blockingFun",19,1))),

            FUN10=quote(RWMA(nit=11000,burnin=0,thin=1,meth="unif",
                        eps=0.01,lambda=1000,omega=c(0.1,0.01,0.001),
                        psin=rep(0.2,19),udin=c(7.714,0.839,0.99),
                        FUN=call("blockingFun",22))),

            FUN11=quote(RWMA(nit=11000,burnin=0,thin=1,meth="mvnorm",
                        eps=0.01,lambda=1000,omega=c(0.1,0.01,0.001),
                        delta=c(rep(0.001,19),0.01,0.001,0.0001),
                        psin=rep(0.2,19),udin=c(7.714,0.839,0.99),
                        BLOCK=list(SB=paste("p", seq(1:22), sep="")))),

            FUN12=quote(RWMA(nit=11000,burnin=0,thin=1,meth="mvnorm",
                        eps=0.01,lambda=1000,omega=c(0.1,0.01,0.001),
                        delta=c(rep(0.0001,19),0.001,0.0001,0.00001),
                        psin=rep(0.2,19),udin=c(7.714,0.839,0.99),
                        BLOCK=list(SB=paste("p", seq(1:22), sep="")))),

            FUN13=quote(RWMA(nit=11000,burnin=0,thin=1,meth="mvnorm",eps=0.01,
                        update="AP",lambda=1000,omega=c(0.1,0.01,0.001),
                        delta=c(rep(0.0001,19),0.001,0.0001,0.00001),
                        psin=rep(0.2,19),udin=c(7.714,0.839,0.99),
                        BLOCK=list(SB=paste("p", seq(1:22), sep="")))),

            FUN14=quote(RWMA(nit=11000,burnin=0,thin=1,meth="mvnorm",eps=0.01,
                        update="AM",lambda=1000,omega=c(0.1,0.01,0.001),
                        delta=c(rep(0.0001,19),0.001,0.0001,0.00001),
                        psin=rep(0.2,19),udin=c(7.714,0.839,0.99),
                        BLOCK=list(SB=paste("p", seq(1:22), sep="")))),

            FUN15=quote(RWMA(nit=11000,burnin=0,thin=1,meth="mvnorm",eps=0.01,
                        update="AM",lambda=1000,omega=c(0.1,0.01,0.001),
                        delta=c(rep(0.0001,19),0.001,0.0001,0.00001),
                        psin=rep(0.2,19),udin=c(7.714,0.839,0.99),
                        BLOCK=list(B1=paste("p", seq(1:11), sep=""),
                                   B2=paste("p", seq(12,22,1), sep="")))
            ))
```

Written R code for simulations and analyses is available upon request.

B.3 R session information

To a large extent, the analyses presented in this thesis are programmed and performed using `RStudio` (RStudio Team, 2015) and R version 3.1.2 (2014-10-31) – "Pumpkin Helmet", (R Core Team, 2014b) (`x86_64-w64-mingw32/x64`) with the following Base packages: base, graphics, grDevices, methods, stats, tools, utils.

Further packages that have been used are: `adaptMCMC 1.1` (Scheidegger, 2012), `BayHap 1.0.1` (Iniesta & Moreno, 2013), `boot 1.3-13` (Canty & Ripley, 2014), `coda 0.17-1` (Plummer, Best, Cowles, & Vines, 2006), `coefplot 1.2.0` (Lander, 2013), `doParallel 1.0.8` (Revolution Analytics & Weston, 2014a), `erer 2.4` (Changyou, 2015), `foreach 1.4.2` (Revolution Analytics & Weston, 2014b), `foreign 0.8-61` (R Core Team, 2014a), `ggplot2 1.0.1` (Wickham, 2009), `hdrcde 3.1` (Hyndman, Einbeck, & Wand, 2013), `lattice 0.20-31` (Sarkar, 2008), `lme4 1.1-7` (Bates, Maechler, Bolker, & Walker, 2014), `lsr 0.4` (Navarro, 2014), `MASS 7.3-40` (Venables & Ripley, 2002), `maxLik 1.2-4` (Henningsen & Toomet, 2011), `mcmcplots 0.4.2` (Curtis, 2015), `memisc 0.97` (Elff, 2015), `mlogit 0.2-4` (Croissant, 2013), `mnormt 1.5-1` (Azzalini & Genz, 2014), `moments 0.13` (Komsta & Novomestky, 2012), `mvtnorm 1.0-2` (Genz et al., 2014), `numDeriv 2012.9-1` (Gilbert & Varadhan, 2012), `party 1.0-20` (Hothorn, Hornik, & Zeileis, 2006), `plotrix 3.5-12` (Lemon & et al., 2006), `plyr 1.8.2` (Wickham, 2011), `psych 1.5.6` (Revelle, 2015), `rattle 3.5.0` (Williams, 2011), `Rmisc 1.5` (Hope, 2013), `rpart 4.1-10` (Therneau, Atkinson, & Ripley, 2015), `rpart.plot 1.5.0` (Milborrow, 2014), `snow 0.3-13` (Tierney, Rossini, Li, & Sevcikova, 2013), `snowfall 1.84-6` (Knaus, 2013), `stargazer 5.1` (Hlavac, 2014), `truncnorm 1.0-7` (Trautmann, Steuer, Mersmann, & Bornkamp, 2014), `VGAM 0.9-5` (Yee, 2014), `VIM 4.1.0` (Templ, Alfons, Kowarik, & Prantner, 2014), `WRS 0.24` (Wilcox & Schönbrodt, 2014), `WRS2 0.4-0` (Mair, Schönbrodt, & Wilcox, 2015), `xtable 1.7-4` (Dahl, 2014).