DOI: https://doi.org/10.20378/irb-40127

The validity of data fusion

Hans Kiesl, Susanne Rässler
Institute for Employment Research, Federal Employment Services, Regensburger Straße 104, 9047Nürnberg, Germany
University of Bamberg, Feldkirchenstraße 21, 96045 Bamberg, Germany
e-mail: hans.kiesl@iab.de, susanne.raessler@uni-bamberg.de

1. Introduction

Statistical matching techniques typically aim to achieve a complete data file from different sources that do not contain the same units. On the contrary, if samples are exactly matched using identifiers such as social security numbers or name and address, this is called record linkage. Traditionally, statistical matching is done on the basis of variables common to all files. Statistical twins, i.e., donor and recipient units that are similar according to their common variables, are usually found by means of nearest neighbor or hot deck procedures. The specific variables of a donor unit which are observed only in one file are added to the record of the recipient unit to finally create the matched sample. We like to note that in our sense statistical matching is not restricted to the case of merging different samples without overlap. Also one single file may contain some records with observations on more variables than others, then, these records can be matched with those containing less information based on the variables common to all units.

In this paper we refer to the situation of data fusion which means there are groups of variables that are *never jointly observed*, say *X* and *Y*. In all other cases of statistical matching we assume that, at least, every pair of variables has been jointly observed in one or the other data set. The fusion of data sets with the aim of analyzing the unobserved relationship between X and Y and addressing quality of data fusion is done, e.g., by National Statistical Institutes such as Statistics Canada or the Italian National Institute of Statistics, see, e.g., Liu and Kovacevic (1997) or D'Orazio et al. (2003). The focus often is on analyzing consumers' expenditures and income, which are in detail only available from different surveys. In the U.S., e.g., data fusion is used for microsimulation modeling, where "what if" analyses of alternative policy options are carried out using matched data sets, see Moriarity and Scheuren (2001, 2003). Especially in Europe and among marketing research companies, data fusion has become a powerful tool for media planning, see, e.g., Wendt (1986). Often surveys concerning the purchasing behavior of individuals or households are matched to those containing valuable information about print, radio and television consumption.

2. Data Fusion and its Identification Problem

2.1 Traditional Fusion Algorithms

The general benefit of data fusion is the creation of one complete data source containing information about all variables. Without loss of generality, let the (X,Z) sample be the

recipient sample B of size n_B and the (Y,Z) sample the donor sample A of size n_A. The traditional matching procedures determine for every unit i, $i = 1,\ldots, n_B$, of the recipient sample with the observations (x_i,z_i) a value y from the observations of the donor sample. Thus, a composite data set $(x_1, \tilde{y}_1, z_1),\ldots,(x_{n_B}, \tilde{y}_{n_B}, z_{n_B})$ with n_B elements of the recipient sample is constructed. The main idea is to search for a statistical match, i.e., for a donor unit j with $(y_j, z_j) \in \{(y_1, z_1),\ldots,(y_{n_A}, z_{n_A})\}$ whose observed data values of the common variables z_j are identical to those z_i of the recipient unit i for $i = 1,\ldots,n_B$. Notice that $\tilde{y}_i$ is not the true y-value of the i-th recipient unit but the y-value of the matched statistical twin. In the following, all density functions (joint, marginal, or conditional) and their parameters produced by the fusion algorithm are marked by the symbol $\sim$. Notice that $\tilde{Y}$ is called fusion or imputed variable herein.

A typical matching algorithm chooses randomly among all possible statistical matches for each recipient unit i (i.e. among all (y_j, z_j) with $z_j = z_i$); we shall call this the ideal case thereafter. In reality, not every recipient allows for an exact match in the common variables; therefore some nearest neighbor rules are usually imposed. There are very sophisticated fusion techniques in practice; for an overview see Rässler (2002).

In order to judge the quality of any data fusion procedure, it is essential to study how the true (only partially known) distribution $f(x,y,z)$ and the fusion distribution $\tilde{f}(x,y,z)$ are related. In the ideal case, it can be shown that the joint distributions of X and Z and of Y and Z are unaltered by the matching algorithm. The overall joint distribution satisfies

$$\tilde{f}_{X,Y,Z}(x,y,z) = f_{X,Z}(x,z) \cdot f_{Y|Z}(y \mid z);$$

see Rässler (2002) for technical details. Obviously, the fusion distribution equals the true distribution if and only if $f_{Y|X,Z} = f_{Y|Z}$, i.e., if Y and X are conditionally independent given Z. This implicit assumption of traditional algorithms was first pointed out by Sims (1972); see also Rodgers (1984) for an enlightening discussion.

Rässler and Fleischer (1998) show that in the ideal case, the fusion covariance between X and Y is given by

$$\tilde{\mathrm{cov}}(X,Y) = \mathrm{cov}(\mathrm{E}(X \mid Z), \mathrm{E}(Y \mid Z)).$$

Because in general,

$$\mathrm{cov}(X,Y) = \mathrm{E}(\mathrm{cov}(X,Y \mid Z)) + \mathrm{cov}(\mathrm{E}(X \mid Z), \mathrm{E}(Y \mid Z))$$

holds, the fusion covariance $\tilde{\mathrm{cov}}(X,Y)$ equals the true covariance, if and only if $\mathrm{E}(\mathrm{cov}(X,Y \mid Z)) = 0$, i.e., if X and Y are on the average conditionally uncorrelated given Z. Notice that variables which are conditionally independent are also conditionally uncorrelated and, of course, on the average conditionally uncorrelated, but not vice versa in general. If f is multinormally distributed, however, these concepts coincide,

since in this case the conditional covariance $\mathrm{cov}(X, Y \mid Z = z)$ is given by $\mathrm{cov}(X, Y) - \mathrm{cov}(X, Z)\,\mathrm{var}(Z)^{-1}\,\mathrm{cov}(Z, Y)$, which is independent of z.

With small sample sizes, the ideal case is seldom observed. However, simulation studies have shown that these derivations are even approximately valid, if nearest neighbour algorithms are applied (see Rässler 2002).

Summing it up: Traditional algorithms produce fusion data sets which reflect the true joint distribution only in the case of conditional independence of X and Y given Z. The true covariance structure is retained in the fused file only in the case of X and Y being on the average conditionally uncorrelated given Z. The question that naturally arises is: can we learn from the data, whether these assumptions are met?

2.2 The Identification Problem of Data Fusion

2.2.1. Joint Distributions

Data fusion initially is connected to an identification problem concerning the joint distribution and the association of the specific variables that are never jointly observed. For every pair of specific variables (X_i, Y_j), the marginal joint cumulative distribution function $F_{X_i, Y_j}(x, y)$ is bounded by the Fréchet-Hoeffding inequality, although it is usually not very informative:

$$\max\{F_{X_i}(x) + F_{Y_j}(y) - 1,\ 0\} \le F_{X_i, Y_j}(x, y) \le \min\{F_{X_i}(x), F_{Y_j}(y)\}. \qquad (1)$$

With common variables Z these bounds can be slightly improved, since the same inequalities are valid for the conditional distributions either (Ridder and Moffitt 2006):

$$\begin{aligned}\max\{F_{X_i \mid Z=z}(x \mid Z = z) + F_{Y_j \mid Z=z}(y \mid Z = z) - 1,\ 0\} \quad &\le \quad F_{X_i, Y_j \mid Z=z}(x, y \mid Z = z) \\ \le \min\{F_{X_i \mid Z=z}(x \mid Z = z), F_{Y_j \mid Z=z}(y \mid Z = z)\}.&\end{aligned}$$

Taking expectations over Z, we have

$$\begin{aligned}\mathrm{E}\Big(\max\{F_{X_i \mid Z=z}(x \mid Z = z) + F_{Y_j \mid Z=z}(y \mid Z = z) - 1,\ 0\}\Big) \quad &\le \quad F_{X_i, Y_j}(x, y) \\ \le \mathrm{E}\Big(\min\{F_{X_i \mid Z}(x \mid Z), F_{Y_j \mid Z}(y \mid Z)\}\Big)&\end{aligned} \qquad (2)$$

While F_{X_i} and F_{Y_j} might be estimated with sufficient accuracy from the samples, this is probably not always true for the expectations in (2), especially in the case of continuous Z. Thus, in practice the unconditional bounds might be the more reliable choice, although the lower and upper bounds are usually quite far apart and therefore rather useless in reality. The lesson to be learned is, by means of the observed data we are not able to decide which joint distribution (given that it lies within the Fréchet-Hoeffding bounds) could have generated the data.

2.2.2 Correlation Structure

Consider, for example, a univariate common variable Z determining another variable X which is only observed in one file. Suppose first that X and Z be linearly dependent, i.e., let the correlation $\rho_{ZX} = 1$, and thus $X = a+bZ$ for some real-valued a and b ($b \neq 0$). The correlation between this common variable Z and a variable Y in a second file may be $\rho_{ZY} = 0.8$. It is easy to see that the unconditional correlation of the two variables X and Y which are not jointly observed is determined by Z with $\rho_{XY} = \rho_{a+bZ\,Y} = \rho_{ZY} = 0.8$. If the correlation between X and Z is less than one, say 0.9, we can easily calculate the possible range of the unconditional association between X and Y by means of the determinant of the covariance matrix which has to be positive semidefinite; i.e., the determinant of the covariance matrix cov(Z,Y,X) must be positive or at least zero, see, e.g., Cox and Wermuth (1996).

Given the above values and setting the variances to one without loss of generality, the covariance matrix of (Z,Y,X) is

$$\operatorname{cov}(Z,Y,X) = \begin{pmatrix} 1 & 0.9 & 0.8 \\ 0.9 & 1 & \operatorname{cov}(X,Y) \\ 0.8 & \operatorname{cov}(X,Y) & 1 \end{pmatrix}$$

with

$$\det(\operatorname{cov}(Z,Y,X)) = -\operatorname{cov}(X,Y)^2 + 2 \cdot 0.72 \operatorname{cov}(X,Y) - 0.45 .$$

Calculating the roots of det(cov(Z,Y,X)) = 0, we get the two solutions cov(X,Y) = $0.72 \pm \sqrt{0.0684}$. Hence we find the correlation bounded between [0.4585, 0.9815]; i.e., every value of the unknown covariance cov(X,Y) greater than 0.4585 and less than 0.9815 leads to a valid and thus feasible covariance structure for (Z,Y,X). By means of the observed data we are not able to decide which covariance matrix could have generated the data, provided that it is positive semidefinite.

Bearing these identification problems in mind, note that traditional data fusion algorithms make specific implicit assumptions (conditional independence or at least conditional uncorrelatedness on average) about the data. The need for alternative approaches that overcome these assumptions is obvious, although little research has been done in the literature so far.

Only few approaches, basically three different procedures, have been published to assess the effect of alternative assumptions about the inestimable correlation structure. One approach is due to Kadane (2001; reprinted from 1978), generalized by Moriarity and Scheuren (2001). The next approach dates back to Rubin and Thayer (1978), it is used to address data fusion explicitly by Rubin (1986), and generalizations are presented by Moriarity and Scheuren (2003). Both approaches use regression based procedures to produce synthetic data sets under various assumptions on this unknown association.

Finally, a full Bayesian regression approach using multiple imputations is first given by Rubin (1987, p. 188), and then generalized by Rässler (2002).

3. Calculation of Feasible Correlations

To ease notation, we again set all variances equal to 1. Consider again the correlation matrix $\Sigma := \mathrm{cov}(Z, Y, X)$ of all observed variables. Recall that Z is the vector of variables observed in both samples; Y and X are the vectors of variables which are only observed in sample A and B, respectively. The matrix Σ and its inverse can be partitioned corresponding to the partition of the complete data vector (Z,Y,X), to give

$$\Sigma = \begin{pmatrix} \Sigma_{ZZ} & \Sigma_{ZY} & \Sigma_{ZX} \\ \Sigma_{YZ} & \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{XZ} & \Sigma_{XY} & \Sigma_{XX} \end{pmatrix} \qquad \Sigma^{-1} = \begin{pmatrix} \Sigma^{ZZ} & \Sigma^{ZY} & \Sigma^{ZX} \\ \Sigma^{YZ} & \Sigma^{YY} & \Sigma^{YX} \\ \Sigma^{XZ} & \Sigma^{XY} & \Sigma^{XX} \end{pmatrix}$$

In the case of data fusion, Σ_{YX} consists of the correlations between variables that are never jointly observed and may therefore not be directly estimated from the data. However, as we will discuss below, there is information in the data about their feasible values.

Correlation matrices have to be positive semidefinite; apart from the case of exact linear dependence they are positive definite. We will ignore this distinction and assume positive definiteness, since an exact linear relationship never occurs in sample data (or can be easily detected and removed).

All other submatrices of Σ apart from Σ_{YX} can be estimated from the two samples. Therefore, Σ is only partially determined; since we know that it has to be positive definite, Σ is called a partial positive definite matrix. Finding the set of feasible correlation matrices in this case is a special application of what is called matrix completion problems in matrix theory; we are interested in positive definite completions of Σ.

Due to the special structure of Σ, a positive definite completion of Σ always exists. Moreover, there is a unique positive definite completion, whose determinant is maximal, and this matrix is the unique one whose inverses has zeros in those positions corresponding to the unspecified entries in Σ, i.e. $\Sigma^{YX} = 0$ (see Grone et al. 1984). Consider now the matrix $\Sigma^*_{YX|Z}$ of partial covariances of X and Y given Z, i.e. the covariance matrix of the residuals of linear least squares regression of every component of X and Y on all components of Z. (Notice that partial covariances and conditional covariances are different concepts. In case of multivariate normality these matrices coincide, whereas in general the two concepts produce different results.)

$\Sigma^*_{YX|Z}$ can be easily derived from the simple correlation matrix as the Schur complement of Σ_{ZZ} in Σ (see e.g. Whittaker 1990, p.135):

$$\Sigma^{*}_{YX|Z} = \begin{pmatrix} \Sigma_{YY|Z} & \Sigma_{YX|Z} \\ \Sigma_{XY|Z} & \Sigma_{XX|Z} \end{pmatrix} = \begin{pmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{XX} \end{pmatrix} - \begin{pmatrix} \Sigma_{YZ} \\ \Sigma_{XZ} \end{pmatrix} \Sigma_{ZZ}^{-1} \begin{pmatrix} \Sigma_{ZY} & \Sigma_{ZX} \end{pmatrix} \tag{3}$$

There is an interesting relationship between the partitioned inverse of Σ and the partial covariance matrix: The term $\Sigma^{YX} = 0$ if and only if the partial correlations between X and Y given Z vanish, i.e. $\Sigma_{YX|Z} = 0$ (Whittaker 1990, p. 144). Hence zero partial correlations given Z maximize the determinant of Σ among all feasible correlation matrices; the corresponding simple correlations being $\Sigma_{YX} = \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX}$. Notice that in case of normality, this is the correlation matrix of the fused data set that traditional algorithms create.

Positive definiteness places restrictions on the feasible correlations between X and Y. In general it is a difficult task to describe the set of feasible values in closed form. Kadane (2001) and Moriarity and Scheuren (2001) provide formulae for univariate X and univariate Y with multivariate Z. For multivariate X or multivariate Y, no closed form yet exists in the literature. One way to numerically tackle this problem is via grid search over all possible completions of Σ and deciding for every value if the completion is positive definite; see Rässler (2002) for an example of this approach.

In the following, we show that even in case of either multivariate X or multivariate Y (though not both), one can derive the range of all feasible solutions analytically.

Let (without loss of generality) X be univariate, i.e. $\Sigma_{XX} = 1$, so that Σ_{ZX} and Σ_{YX} are column vectors. Since all leading principal submatrices of Σ are fully specified and (by assumption of consistency) positive definite, the positive definiteness of Σ is equivalent to the determinant of Σ being positive, i.e. $\det(\Sigma) > 0$. Partitioning Σ and using a standard argument on the determinant of a partitioned matrix leads to the following condition:

$$\begin{pmatrix} \Sigma'_{ZX} & \Sigma'_{YX} \end{pmatrix} \begin{pmatrix} \Sigma_{ZZ} & \Sigma_{ZY} \\ \Sigma'_{ZY} & \Sigma_{YY} \end{pmatrix}^{-1} \begin{pmatrix} \Sigma_{ZX} \\ \Sigma_{YX} \end{pmatrix} < 1 \tag{4}$$

The inverse can be written in closed form:

$$\begin{pmatrix} \Sigma_{ZZ} & \Sigma_{ZY} \\ \Sigma'_{ZY} & \Sigma_{YY} \end{pmatrix}^{-1} = \begin{pmatrix} \Sigma_{ZZ}^{-1}(\mathrm{I} + \Sigma_{ZY} C \Sigma'_{ZY} \Sigma_{ZZ}^{-1}) & -\Sigma_{ZZ}^{-1}\Sigma_{ZY} C \\ -C\Sigma'_{ZY}\Sigma_{ZZ}^{-1} & C \end{pmatrix} =: \begin{pmatrix} A & B \\ B' & C \end{pmatrix}$$

with $C := (\Sigma_{YY} - \Sigma'_{ZY}\Sigma_{ZZ}^{-1}\Sigma_{ZY})^{-1}$.

After straightforward calculation (4) evolves into

$$\Sigma'_{YX} C \Sigma_{YX} + 2\Sigma'_{ZX} B \Sigma_{YX} + \Sigma'_{ZX} A \Sigma_{ZX} < 1. \tag{5}$$

From this inequality, the geometric shape of the set of feasible correlations can be determined. Since C is positive definite, the set of possible vectors Σ_{YX} satisfying (5) is

the interior of an n-dimensional ellipsoid (n being the dimension of vector Y). Transforming (5) into the normal form of an ellipsoid in order to be able to calculate its centre and axes, we get

$$(\Sigma_{YX} + C^{-1}B'\Sigma_{ZX})' \cdot \widetilde{C} \cdot (\Sigma_{YX} + C^{-1}B'\Sigma_{ZX}) < 1$$

with $\widetilde{C} := (1 + \Sigma'_{ZX}(BC^{-1}B' - A)\Sigma_{ZX})^{-1}C$.

Thus, the centre of the ellipsoid is $-C^{-1}B'\Sigma_{ZX}$. Plugging in the formulae for B and C yields

$$-C^{-1}B'\Sigma_{ZX} = \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX};$$

from this it can be seen that the correlation vector providing zero partial correlation (which maximizes the determinant) is the centre of the ellipsoid.

Final calculations give $1 + \Sigma'_{ZX}(BC^{-1}B' - A)\Sigma_{ZX} = 1 - \Sigma_{XZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX}$, from which $\widetilde{C}$ can be computed:

$$\widetilde{C} = (1 - \Sigma_{XZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX})^{-1} \cdot (\Sigma_{YY} - \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZY})^{-1}.$$

The semi-axes of the ellipsoid are in the direction of the eigenvectors of $\widetilde{C}$ (or C), the lengths of the semi-axes are given by $1/\sqrt{\lambda_i}$, where λ_i is the i-th eigenvalue of $\widetilde{C}$ ($i = 1,\ldots, n$).

The volume of the ellipsoid of feasible correlations (which is proportional to the product of the lengths of its semi-axes) might be considered as a new quality index for a data fusion process: the less volume the ellipsoid has, the greater is the explanatory power of the common variables and the less uncertainty remains for creating the fused data set.

In some cases, the marginal distributions might restrict the set of feasible correlation matrices even further. To see this, consider again the Fréchet-Hoeffding inequality (1). The upper and lower bounds are valid bivariate distributions, whose correlation coefficients are upper and lower bounds of possible correlations given the marginals (Tchen 1980). Thus, for every pair (X_i, Y_j) of specific variables, this inequality might place an additional restriction to the feasible correlations (in case of normality every correlation can be achieved with any marginal distributions, therefore no further restriction can be imposed).

If there are lots of ordinal variables in the samples, it is appropriate not to consider Bravais-Pearson correlation coefficients but to use association measures based on ranks. Frequently Spearman's ρ or Kendall's τ are measures of interest, even in metric settings. Since correlation matrices based on these measures also have to be positive definite (note that they can be expressed as Bravais-Pearson correlations for recoded variables), the results of this section remain valid, if consideration is upon matrices of Spearman or Kendall correlations rather than upon Bravais-Pearson correlation coefficients.

4. Summary and Outlook

In this paper we derived bounds for the correlations between variables not jointly observed, provided that one of the vectors of specific variables is univariate, and suggest a new quality index of data fusion which is built upon these bounds. Using our results, multiply imputed datasets can be produced according to different admissible correlation structures between X and Y by using appropriate algorithms (e.g. NIBAS, see Rässler 2002; notice that since data fusion can be viewed as a problem of missing data, multiple imputation procedures are applicable in general). Analyzing the different fused data sets can then reveal sensitivity to the different assumptions about the correlation structure between the variables that have never been jointly observed.

References

Box G.E.P., Tiao G.C. (1992) *Bayesian Inference in Statistical Analysis*, New York, Wiley.

Cox D.R., Wermuth N. (1996) *Multivariate Dependencies*, London, Chapman and Hall.

Grone R., Johnson C.R., Sá E.M., Wolkowicz H. (1984) Positive Definite Completions of Partial Hermitian Matrices, *Linear Algebra and its Applications*, 58, 109-124.

Kadane J.B. (2001) Some Statistical Problems in Merging Data Files, *Journal of Official Statistics*, 17, 423-433.

Liu T.P., Kovacevic M.S. (1997) An Empirical Study on Categorically Constrained Matching, *Proceedings of the Survey Methods Section, Statistical Society of Canada*, 167-178.

Moriarity C., Scheuren F. (2001) Statistical Matching: A Paradigm for Assessing the Uncertainty in the Procedure, *Journal of Official Statistics*, 17, 407-422.

Moriarity C., Scheuren F. (2003) A Note on Rubin's Statistical Matching Using File Concatenation With Adjusted Weights and Multiple Imputations, *Journal of Business and Educational Studies*, 21, 65-73.

D'Orazio M., Di Zio M., Scanu M. (2004) Statistical matching and the likelihood principle: uncertainty and logical constraints, *ISTAT Technical Report* 1/2004.

Rässler S. (2002) *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*, Lecture Notes in Statistics, 168, New York, Springer.

Rässler S., Fleischer K. (1998), Aspects Concerning Data Fusion Techniques, *ZUMA Nachrichten Spezial*, 4, 317-333.

Ridder G., Moffitt R. (2006) The Econometrics of Data Combination, in Heckman, J.J., and E.E. Leamer (eds.), *Handbook of Econometrics Volume 6*, Amsterdam: North Holland (to appear).

Rodgers W.L. (1984) An Evaluation of Statistical Matching, *Journal of Business and Economic Statistics*, 2, 91-102.

Rubin D.B. (1986) Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations, *Journal of Business and Economic Statistics*, 4, 87-95.

Rubin D.B. (1987) *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.

Rubin D.B., Thayer D. (1978) Relating Tests Given to Different Samples, *Psychometrika*, 43, 3-10.

Sims C.A. (1972) Comments, *Annals of Economic and Social Measurement*, 1, 343-345.

Tchen A.H. (1980) Inequalities for Distributions with Given Marginals, *Annals of Probability*, 8, 814-827.

Wendt F. (1986) Einige Gedanken zur Fusion, in Arbeitsgemeinschaft Media-Analyse e.V. (eds.), *Auf dem Wege zum Partnerschaftsmodell*, Frankfurt, Media-Micro-Census GmbH, 109-140. [In German]

Whittaker J. (1990) *Graphical Models in Applied Multivariate Statistics*, Chichester, Wiley.