IAB DiscussionPaper

Beiträge zum wissenschaftlichen Dialog aus dem Institut für Arbeitsmarkt- und Berufsforschung

No. 15/2006

How Valid Can Data Fusion Be?

Hans Kiesl, Susanne Rässler

Bundesagentur für Arbeit

How Valid Can Data Fusion Be?

Hans Kiesl and Susanne Rässler (IAB)

How Valid Can Data Fusion Be?

Hans Kiesl and Susanne Rässler[1]
Institute for Employment Research of the Federal Employment Services
Competence Centre Empirical Methods
Regensburger Straße 104, 90478 Nürnberg, Germany
email: hans.kiesl@iab.de susanne.raessler@iab.de

Abstract

Data fusion techniques typically aim to achieve a complete data file from different sources which do not contain the same units. Traditionally, this is done on the basis of variables common to all files. It is well known that those approaches establish conditional independence of the specific variables given the common variables, although they may be conditionally dependent in reality. We discuss the objectives of data fusion in the light of their feasibility and distinguish four levels of validity that a fusion technique may achieve. For a rather general situation, we derive the feasible set of correlation matrices for the variables not jointly observed and suggest a new quality index for data fusion. Finally, we present a suitable and efficient multiple imputation procedure to make use of auxiliary information and to overcome the conditional independence assumption.

Key words: Correlation matrix, data fusion, multiple imputation, missing data, missing by design, observed-data posterior, statistical matching.

JEL classification: C11, C15, C81.

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

[1]Acknowledgment: The authors want to thank Friedrich Wendt, who was one of the first and most engaged persons to develop data fusion techniques in Europe. Moreover, we are grateful to Donald B. Rubin and Fritz Scheuren for giving us lots if insights and stimulating discussions.

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. Basically, there are a couple of different situations, when statistical matching can be applied. Figure 1 gives an overview of these occasions. The white boxes represent the missing variables.

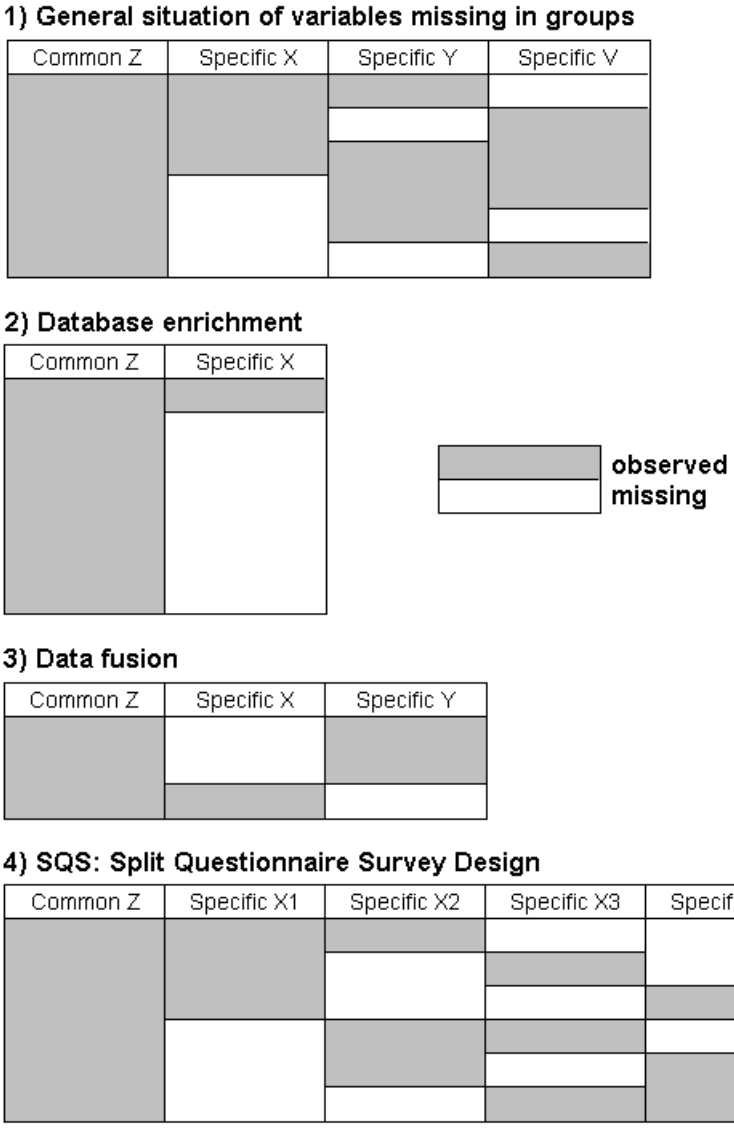

Figure 1: Different situations for statistical matching

In this paper we refer to the situation of picture no. 3 in Figure 1 which we call data fusion. This figure illustrates that only in the case of data fusion there are groups of variables that are *never jointly observed*, say X and Y. In all other cases 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.

Our article is organized as follows. Section 2 reviews the crucial identification problem inherent in data fusion. With this in mind, we define in Section 3 four different levels of validity a data fusion can achieve. Investigating further the most promising of these levels in Section 4, we present a new result on the calculation of feasible correlations between variables not jointly observed. In Section 5 a multiple imputation algorithm for assessing the impact of different correlation structures is developed, which is validated by a simulation study in Section 6.

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, $i = 1, 2, \ldots, n_B$, 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 $(x_1, \widetilde{y}_1, z_1), \ldots, (x_{n_B}, \widetilde{y}_{n_B}, z_{n_B})$ 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 $(y_j, z_j) \in \{(y_1, z_1), (y_2, z_2), \ldots, (y_{n_A}, z_{n_A})\}$ whose observed data values of the common variables z_j are identical to those z_i of the recipient unit i for $i = 1, 2, \ldots, n_B$. Notice that $\widetilde{y}_i$ 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 $\widetilde{\ }$. The variable $\widetilde{Y}$ is called fusion or imputed variable.

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 $\widetilde{f}(x, y, z)$ 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

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

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

$$\widetilde{\mathrm{cov}}(X, Y) = \mathrm{cov}(\mathrm{E}(X|Z), \mathrm{E}(Y|Z)).$$

Because in general,

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

holds, the fusion covariance $\widetilde{\mathrm{cov}}(X, Y)$ equals the true covariance, if and only if $\mathrm{E}(\mathrm{cov}(X, Y|Z)) = 0$, 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 $\mathrm{cov}(X, Y|Z = z)$ 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 neighbor 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:

$$\max\left\{F_{X_i}(x) + F_{Y_j}(y) - 1, 0\right\} \leq F_{X_i,Y_j}(x, y) \leq \min\left\{F_{X_i}(x), F_{Y_j}(y)\right\} \tag{1}$$

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):

$$\begin{aligned}\max\left\{F_{X_i|Z=z}(x|Z=z) + F_{Y_j|Z=z}(y|Z=z) - 1, 0\right\} \\ \leq F_{X_i,Y_j|Z=z}(x, y|Z=z) \leq \min\left\{F_{X_i|Z=z}(x|Z=z), F_{Y_j|Z=z}(y|Z=z)\right\}.\end{aligned}$$

Taking expectations over Z, we have

$$\begin{aligned}\mathrm{E}\left(\max\left\{F_{X_i|Z}(x|Z) + F_{Y_j|Z}(y|Z) - 1, 0\right\}\right) \\ \leq F_{X_i,Y_j}(x, y) \leq \mathrm{E}\left(\min\left\{F_{X_i|Z}(x|Z), F_{Y_j|Z}(y|Z)\right\}\right).\end{aligned} \tag{2}$$

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 $a, b \in \mathbb{R}^2, b \neq 0$. 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 $\mathrm{cov}(Z, Y, X)$ 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

$$\begin{aligned} \mathrm{cov}(Z, Y, X) &= \begin{pmatrix} 1 & 0.9 & 0.8 \\ 0.9 & 1 & \mathrm{cov}(X, Y) \\ 0.8 & \mathrm{cov}(X, Y) & 1 \end{pmatrix} \quad \text{with} \\ \det(\mathrm{cov}(Z, Y, X)) &= -\mathrm{cov}(X, Y)^2 + 2 \cdot 0.72\,\mathrm{cov}(X, Y) - 0.45. \end{aligned}$$

Calculating the roots of $\det(\mathrm{cov}(Z, Y, X)) = 0$, we get the two solutions $\mathrm{cov}(X, Y) = 0.72 \pm \sqrt{0.0684}$. Hence we find the correlation bounded between [0.4585, 0.9815]; i. e., every value of the unknown covariance $\mathrm{cov}(X, Y)$ 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 Validity Levels of Data Fusion

With the need for alternative data fusion algorithms, we have to make clear how to judge the quality of a data fusion procedure, i. e. we have to focus on the validity of the fusion process. We suggest to distinguish four levels of validity a fusion procedure may achieve. The term validity rather than efficiency will be used, because efficiency usually refers to a minimum mean squared error criterion as it is common, for example, in survey sampling theory and not to different levels of reproduction and preservation of the original associations and distributions.

3.1 First Level: Preserving Marginal Distributions

We shall say that a data fusion procedure attains the first (and lowest) level of validity, if the marginal and joint distributions of the variables in the donor sample are preserved in the fused file. Then $\widetilde{f}_Y = f_Y$ and $\widetilde{f}_{Y,Z} = f_{Y,Z}$ are expected to hold, if Y is imputed in the (X, Z) sample.

In the ideal case as described in Section 2, this level of validity is always attained. However, with small sample sizes or different sampling designs in the two samples, this need not be the case. In practice, the preservation of the distributions observed in the separate samples is usually required. Analysis concerning the marginal distributions based on the fused file should provide the same valid inference when based on the separate samples. Therefore, the empirical distributions of the common variables Z as well as the imputed variables Y in the resulting fused file are compared with their empirical distributions in the donor sample to evaluate the similarity of both samples. The empirical distributions $\widehat{\widetilde{f}}_Y$ and $\widehat{\widetilde{f}}_{Y,Z}$ should not differ from $\widehat{f}_Y$ and $\widehat{f}_{Y,Z}$ more than two random samples drawn from the true underlying population. Notice that this implies the different samples being drawn according to the same sampling design.

In the typical fusion situation (with traditional algorithms) only this level can be controlled for. Therefore, often data fusion is said to be successful, if the marginal and joint empirical distributions of Z and Y, as they are observed in the donor sample, are "nearly" the same in the fused file.

In common approaches, first of all, averages for all common variables Z between the donor and the recipient sample are compared. Then the average values between the imputed variables $\widetilde{Y}$ and the corresponding variables Y in the donor sample are compared. Often the preservation of the relation between variables is measured by means of correlations. Therefore, for each common variable Z, the correlation with every original variable Y and imputed variable $\widetilde{Y}$ is computed, both for the fused data set and the donor sample. The

mean difference between common-fusion correlations in the donor versus the fused data set are calculated and empirically evaluated, see, e.g., van der Putten et al. (2002).

The German association for media analysis[2], for example, still postulates the following data controls after a match has been performed.

- First the empirical distributions of the common variables Z in the recipient and the donor sample are compared to evaluate whether their marginal distributions are the same in both samples.
- Next the empirical distributions of the imputed variables $\widetilde{Y}$ in the recipient and Y in the donor sample are compared.
- Finally the joint distribution $f_{Z,Y}$ as observed in the donor sample is compared to the joint distribution $\widetilde{f}_{Z,Y}$ as observed in the fused file.

All these comparisons are done using different tests such as χ^2-tests or t-tests to compare empirical distributions or their moments. A successful match should lead to similar relationships between common and specific variables in the donor and the fused file; discrepancies should not be larger than expected between two independent random samples from the same underlying population. In particular, often each pair of variables Y and Z in the donor sample is tested at a significance level α for positive or negative association by, for example, a χ^2-test or a t-test (depending on the scale of the variables). Then the same test of association between $\widetilde{Y}$ and Z is performed for each pair in the fused file. If the results of the tests only differ in about α percent of the possible (Y, Z) combinations, then the fusion procedure is regarded as successful, although this means accepting the Null hypotheses rather than discarding them. Among others, nonparametric tests and multiple regression models may be used in the same manner.[3]

3.2 Second Level: Preserving Correlation Structures

If additionally the correlation structure is preserved after data fusion, i. e. $\widetilde{\text{cov}}(X, Y, Z) = \text{cov}(X, Y, Z)$, the second level of validity is achieved. In that case, the fusion data set

[2] "Media Analysis Association" called in German Arbeitsgemeinschaft Media Analyse, for short, AG.MA. The AG.MA is a media association, i. e., publishing houses, radio and TV stations, and many advertising agencies, as well as a certain number of advertisers.

[3] If the samples have different structures, e.g., due to oversampling in one survey or differing sampling designs, weights can be applied accounting for differing selection probabilities of the units in the separate samples. Also, samples drawn according to different sampling designs could be made "equal" by using propensity scores according to an idea by Rubin (2002) before performing the final match. However, this is beyond the scope of this article.

might not reflect the true joint distribution of all variables, but it could be considered as randomly generated from an artificial population which has, at least, the same moments and correlation structure as the actual population of interest. Thus any analysis which is based on covariances or correlations only, will produce reliable outcomes.

Traditional fusion algorithms achieve this level only, if the specific variables are on average conditionally uncorrelated, an assumption that cannot be validated with the given data. To overcome this assumption, two steps of research are needed: We first have to determine, which correlation structure the original data set might have. After that, we must design algorithms that are able to create fused data sets with prescribed correlation structures, so that we can assess the quality of a data fusion process by comparing analyses based on different fused files with different feasible correlation structures.

3.3 Third Level: Preserving Joint Distributions

If the overall joint distribution is preserved after data fusion, the true joint distribution of all variables is reflected in the fused file, i. e. $\widetilde{f}_{X,Y,Z} = f_{X,Y,Z}$. We will call this the third level of validity.

We usually assume that the units of both samples are drawn independently within and between the two samples and the fused file can be regarded as a random sample from the underlying fusion distribution $\widetilde{f}_{X,Y,Z}$. The most important objective of data fusion is the generation of a complete sample that can be used as a single-source sample drawn from the underlying distribution $f_{X,Y,Z}$. It is less the reconstruction of individual values but the possibility of making valid statistical inference based on the fused file.

With traditional algorithms, this level of validity is only achieved, if the specific variables Y and X are conditionally independent given the common variables Z, an assumption that cannot be tested with the given data in the case of data fusion. Unless the common variables have extremely high explanatory power (resulting in tight Fréchet-Hoeffding bounds), it is unrealistic to expect that a data fusion process might attain this level of validity.

3.4 Forth Level: Preserving Individual Values

The individual values are preserved when the true but unknown values of the (multivariate) Y variable of the recipient units are reproduced; i. e., $\widetilde{y}_i = y_i$ for $i = 1, 2, \ldots, n_B$. If all individual values were preserved, that would be the highest level of validity a fusion algorithm can achieve. But obviously it is totally out of reach to reproduce all true values

with certainty. We might call the preservation of an individual value a "hit" for any unit in the recipient sample and may calculate some kind of "hit rate". However, you should keep in mind that this hit rate is not as useful as it seems at first sight.

Within continuous distributions the probability of drawing a certain value y is zero; counting the hits is meaningless then. In the case of discrete or classified variables Y a hit rate may be calculated for the purpose of demonstration, counting a hit for the imputation of a p-dimensional variable Y when the whole imputed vector equates the original vector; i. e.,

$$(\widetilde{y}_{1i}, \widetilde{y}_{2i}, \ldots, \widetilde{y}_{pi}) = (y_{1i}, y_{2i}, \ldots, y_{pi})$$

for $i = 1, 2, \ldots, n_B$. Notice that the calculation of a single hit rate for each variable may mislead the interpretation because it does not ensure that the joint distributions are well preserved. One should always remember that imputations are not meant to exactly reflect the real values and that their microdata interpretation is usually meaningless; emphasis should be placed on marginal or joint distributions and correlation structures.

Any discussion of validity of a data fusion technique can now be based on these four levels. Besides so-called split half or simulation studies, all tests actually applied in practice only indicate the first-level validity. While the third and forth level are out of reach or even potentially misleading, it seems promising to further explore the predictive power of the common variables for bounding the set of valid correlation structures.

4 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.[4] 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 $\Sigma^{\star}_{YX|Z}$ 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.)

$\Sigma^{\star}_{YX|Z}$ 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):

$$\Sigma^{\star}_{YX|Z} = \begin{pmatrix} \Sigma_{YY|Z} & \Sigma_{YX|Z} \\ \Sigma_{XY|Z} & \Sigma_{XX|Z} \end{pmatrix} = \begin{pmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{XX} \end{pmatrix} - \begin{pmatrix} \Sigma_{YZ} \\ \Sigma_{XZ} \end{pmatrix} \Sigma_{ZZ}^{-1} \begin{pmatrix} \Sigma_{ZY} & \Sigma_{ZX} \end{pmatrix}. \qquad (3)$$

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

[4] It should be noted, that since the correlations are not known but estimated from different samples, the estimates might be inconsistent in the sense that no positive definite completion of Σ exists. This problem will disappear in large samples; in general however, this condition has to be checked. If it turns out that Σ ist not partial positive definite, one should look for the nearest partial positive definite approximation (w.r.t. some matrix norm); see Higham (2002) for details.

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 (w.l.o.g.) 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:

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

The inverse can be written in closed form:

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

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

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

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

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

$$-C^{-1}B'\Sigma_{ZX} = \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX}; \tag{6}$$

from this it can be seen that the correlation vector providing zero partial correlation (which maximizes the determinant) is the center 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:

$$\widetilde{C} = \left(1 - \Sigma_{XZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX}\right)^{-1} \cdot \left(\Sigma_{YY} - \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZY}\right)^{-1}. \tag{7}$$

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, 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.

5 A Multiple Imputation Algorithm

In the cases pictured in Figure 1 (at least in nos. 2 to 4), it is assumed that the data are missing completely at random or, at least, missing at random because the missingness is induced by design. Thus, the fusion task can be viewed as a typical imputation problem. In the presence of missing data, the theory of multiple imputation, initially introduced by Rubin (1978) and extensively described in Rubin (1987), provides very flexible procedures

for imputation with good statistical properties from a Bayesian as well as a frequentist view. We follow this approach and suggest a non-iterative Bayesian multiple imputation procedure, called NIBAS, especially suited for data fusion.

Let us assume a multivariate normal data model for $(X, Y|Z = z) = (X_1, X_2, \ldots, X_q, Y_1, Y_2, \ldots, Y_p|Z = z)$ with expectation $\mu_{XY|Z}$ and covariance matrix $\Sigma_{XY|Z}$ is denoted by

$$\Sigma_{XY|Z} = \begin{pmatrix} \Sigma_{XX|Z} & \Sigma_{XY|Z} \\ \Sigma_{YX|Z} & \Sigma_{YY|Z} \end{pmatrix}.$$

(Note that due to the assumption of normality, no distinction between conditional and partial covariance matrices is necessary.)

Moreover, the general linear model for both data sets is applied with

$$\begin{aligned} \text{(file } A\text{)} \quad Y &= Z_A\beta_{YZ} + U_A, \qquad U_A \sim N_{pn_A}(0, \Sigma_{YY|Z} \otimes I_{n_A}), \\ \text{(file } B\text{)} \quad X &= Z_B\beta_{XZ} + U_B, \qquad U_B \sim N_{qn_B}(0, \Sigma_{XX|Z} \otimes I_{n_B}), \end{aligned}$$

with Z_A and Z_B denoting the corresponding parts of the common derivative matrix Z. This data model assumes that the units can be observed independently for $i = 1, 2, \ldots, n$. The correlation structure refers to the variables X_{1i}, X_{2i}, ..., X_{qi}, Y_{1i}, Y_{2i}, ..., Y_{pi} for each unit $i = 1, 2, \ldots, n$. For abbreviation we use the Kronecker product $\otimes$ denoting that the variables X_i and Y_i of each unit i, $i = 1, 2, \ldots, n$, are correlated but no correlation of the variables is assumed between the units.

As a suitable noninformative prior we assume prior independence between β and Σ choosing

$$f_{\beta_{YZ},\beta_{XZ},\Sigma_{XX|Z},\Sigma_{YY|Z},R_{XY|Z}} \propto \Sigma_{XX|Z}^{-(\frac{q+1}{2})}\Sigma_{YY|Z}^{-(\frac{p+1}{2})} f_{R_{XY|Z}}.$$

The joint posterior distribution for the fusion case can be factored into the prior and likelihood derived by file A and file B, respectively, see Rässler (2002). Then the joint posterior distribution can be written with

$$\begin{aligned} f_{\beta_{XZ},\beta_{YZ},\Sigma_{XX|Z},\Sigma_{YY|Z},R_{XY|Z}}\big|_{X,Y} &= c_X^{-1} L(\beta_{XZ}, \Sigma_{XX|Z}; x) f_{\Sigma_{XX|Z}|R_{XY|Z}} \\ & \quad c_Y^{-1} L(\beta_{YZ}, \Sigma_{YY|Z}; y) f_{\Sigma_{YY|Z}|R_{XY|Z}} f_{R_{XY|Z}}. \end{aligned}$$

Thus, our problem of specifying the posterior distributions reduces to standard derivation tasks described, for example, by Box and Tiao (1992, p. 439). $\Sigma_{XX|Z}$ and $\Sigma_{YY|Z}$ given the observed data each are following an inverted-Wishart distribution. The conditional posterior distribution of β_{XZ} (β_{YZ}) given $\Sigma_{XX|Z}$ ($\Sigma_{YY|Z}$) and the observed data is

a multivariate normal distribution. The posterior distribution of $R_{XY|Z}$ equals its prior distribution. Having thus obtained the observed-data posteriors and the conditional predictive distributions a multiple imputation procedure for multivariate variables X and Y can be proposed with the following algorithm:

Algorithm NIBAS

- Compute the ordinary least squares estimates $\widehat{\beta}_{YZ} = (Z_A'Z_A)^{-1}Z_A'Y$ and $\widehat{\beta}_{XZ} = (Z_B'Z_B)^{-1}Z_B'X$ from the regression of each data set. Note that $\widehat{\beta}_{YZ}$ is a $k \times p$ matrix and $\widehat{\beta}_{XZ}$ is a $k \times q$ matrix of the OLS or ML estimates of the general linear model.

- Calculate the following matrices proportional to the sample covariances for each regression with

$$\begin{aligned} S_Y &= (Y - Z_A\widehat{\beta}_{YZ})'(Y - Z_A\widehat{\beta}_{YZ}), \\ S_X &= (X - Z_B\widehat{\beta}_{XZ})'(X - Z_B\widehat{\beta}_{XZ})\,. \end{aligned}$$

- Choose a value for the correlation matrix $R_{XY|Z}$ or each $\rho_{X_iY_j|Z}$ for $i = 1, 2, \ldots, q, j = 1, 2, \ldots, p$

  (a) from its prior according to some distributional assumptions, i. e., uniform over the set of feasible values, or

  (b) several arbitrary levels, or

  (c) estimate a value from a small but completely observed data set.

  The latter might be the most realistic case in many practical situations.

- Perform random draws for the parameters from their observed data posterior distribution according to the following scheme.

$$\begin{aligned} \text{Step 1:}\quad \Sigma_{YY|Z}|y &\sim W_p^{-1}(v_A, S_Y^{-1})\; v_A = n_A - (k+p) + 1 \\ \Sigma_{XX|Z}|x &\sim W_q^{-1}(v_B, S_X^{-1})\; v_B = n_B - (k+q) + 1 \\ \text{Step 2:}\quad \beta_{YZ}|\Sigma_{YY|Z}, y &\sim N_{pk}(\widehat{\beta}_{YZ}, \Sigma_{YY|Z} \otimes (Z_A'Z_A)^{-1}), \\ \beta_{XZ}|\Sigma_{XX|Z}, y &\sim N_{qk}(\widehat{\beta}_{XZ}, \Sigma_{XX|Z} \otimes (Z_B'Z_B)^{-1}), \\ \text{Step 3:}\quad \text{Set } \Sigma_{XY|Z} &= \{\sigma_{X_iY_j|Z}\} \text{ with } \sigma_{X_iY_j|Z} \\ &= \rho_{X_iY_j|Z}\sqrt{\sigma^2_{X_i|Z}\sigma^2_{Y_j|Z}} \\ &\quad \text{with } \sigma^2_{X_i|Z}, \sigma^2_{Y_j|Z} \text{ derived by Step 1} \\ &\quad \text{for } i = 1, 2, \ldots, q, j = 1, 2, \ldots, p\,. \end{aligned}$$

$$
\begin{aligned}
\text{Step 4:} \quad X|y,\beta,\Sigma &\sim N_{qn_A}\Big(Z_A\beta_{XZ} + (Y - Z_A\beta_{YZ})\Sigma^{-1}_{YY|Z}\Sigma_{YX|Z}; \\
&\qquad (\Sigma_{XX|Z} - \Sigma_{XY|Z}\Sigma^{-1}_{YY|Z}\Sigma_{YX|Z}) \otimes I_{n_A}\Big) \\
Y|x,\beta,\Sigma &\sim N_{pn_B}\Big(Z_B\beta_{YZ} + (X - Z_B\beta_{XZ})\Sigma^{-1}_{XX|Z}\Sigma_{XY|Z}; \\
&\qquad (\Sigma_{YY|Z} - \Sigma_{YX|Z}\Sigma^{-1}_{XX|Z}\Sigma_{XY|Z}) \otimes I_{n_B}\Big).
\end{aligned}
$$

The predictive power of the common variables Z especially affects the last step.

6 Simulation Study

6.1 Data Model

Let (Z_1, Z_2, Y_1, Y_2, X) each be univariate standard normally distributed variables with their joint distribution

$$(Z_1, Z_2, Y_1, Y_2, X) \sim N_5(0, \Sigma) \tag{8}$$

and

$$
\Sigma = \left(\begin{array}{cc|cc|c}
1.0 & 0.2 & 0.8 & 0.5 & 0.5 \\
0.2 & 1.0 & 0.6 & 0.6 & 0.5 \\
\hline
0.8 & 0.6 & 1.0 & 0.4 & \rho_{Y_1X} \\
0.5 & 0.6 & 0.4 & 1.0 & \rho_{Y_2X} \\
\hline
0.5 & 0.5 & \rho_{Y_1X} & \rho_{Y_2X} & 1.0
\end{array}\right) = \left(\begin{array}{ccc}
\Sigma_{ZZ} & \Sigma_{ZY} & \Sigma_{ZX} \\
\Sigma_{YZ} & \Sigma_{YY} & \Sigma_{YX} \\
\Sigma_{XZ} & \Sigma_{XY} & \Sigma_{XX}
\end{array}\right).
$$

Assume that file A contains (Z_1, Z_2, Y_1, Y_2) and file B (Z_1, Z_2, X), thus X and $Y = (Y_1, Y_2)'$ are never jointly observed. Thus, the partial correlations of X and Y_1 and X and Y_2, respectively, cannot be estimated from the observed data. Also the simple covariance matrix Σ_{XY} does not have a unique estimate, however, there is information in the data about their admissible values, as was shown in Section 4.

6.2 Calculation of the feasible correlations

Since X is univariate in the example above, we can calculate the admissible correlations via the formulae in Section 4. According to (6) the center of the corresponding ellipse is given by

$$
\begin{pmatrix} \rho_{Y_1X}{}^{\text{center}} \\ \rho_{Y_2X}{}^{\text{center}} \end{pmatrix} = \Sigma_{YZ}\Sigma^{-1}_{ZZ}\Sigma_{ZX} = \begin{pmatrix} 0.8 & 0.6 \\ 0.5 & 0.6 \end{pmatrix} \cdot \begin{pmatrix} 1.0 & 0.2 \\ 0.2 & 1.0 \end{pmatrix}^{-1} \cdot \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix} = \begin{pmatrix} 0.5833 \\ 0.4583 \end{pmatrix}.
$$

This is the vector of unconditional correlations which is the result of traditional matching techniques in the case of normality (zero conditional correlations given Z).

Next we compute matrix $\widetilde{C}$ according to (7):

$$\begin{aligned}
\tilde{C} &= \left(1 - \Sigma_{XZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX}\right)^{-1} \cdot \left(\Sigma_{YY} - \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZY}\right)^{-1} \\
&= \left(1 - \begin{pmatrix} 0.5 & 0.5 \end{pmatrix} \cdot \begin{pmatrix} 1.0 & 0.2 \\ 0.2 & 1.0 \end{pmatrix}^{-1} \cdot \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}\right)^{-1} \cdot \\
&\quad \cdot \left(\begin{pmatrix} 1.0 & 0.4 \\ 0.4 & 1.0 \end{pmatrix} - \begin{pmatrix} 0.8 & 0.6 \\ 0.5 & 0.6 \end{pmatrix} \cdot \begin{pmatrix} 1.0 & 0.2 \\ 0.2 & 1.0 \end{pmatrix}^{-1} \cdot \begin{pmatrix} 0.8 & 0.5 \\ 0.6 & 0.6 \end{pmatrix}\right)^{-1} \\
&= \begin{pmatrix} 33.571 & 15.714 \\ 15.714 & 10.857 \end{pmatrix}.
\end{aligned}$$

The eigenvalues of $\widetilde{C}$ are 41.603 and 2.826, thus the lengths of the semi-axes are 0.155 and 0.595, respectively. Recall that the product of these lengths might be seen as a quality index for data fusion.

The first (normed) eigenvector of $\widetilde{C}$ is (0.890 0.455), the other being orthogonal. With these values in mind, the ellipse representing the set of feasible correlations ρ_{Y_1X} and ρ_{Y_2X} can be displayed. Since the relationship between the unconditional and the partial covariances is linear according to (3), the feasible set of partial correlations can easily be derived; obviously it has an analogous elliptical form. Both sets of correlations are shown in Figure 2.

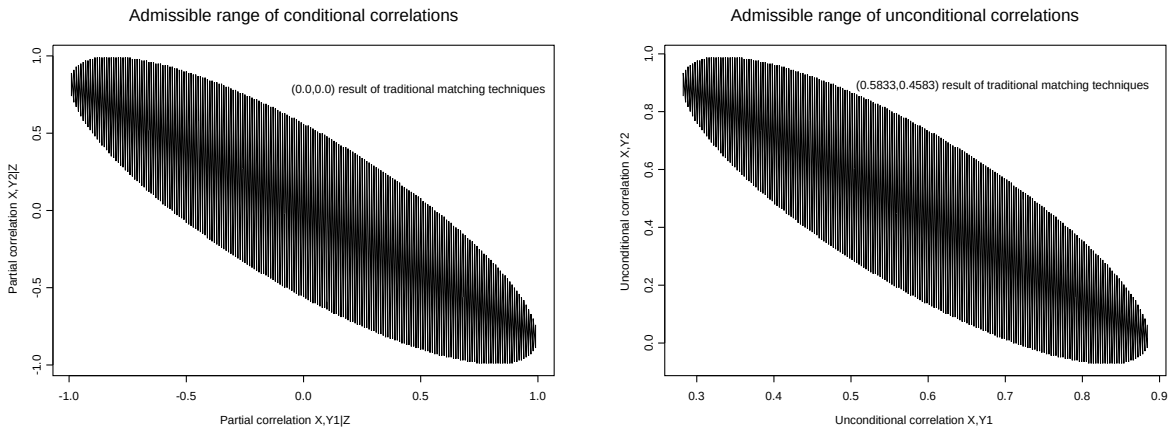

Figure 2: Admissible combinations of conditional/unconditional correlations

6.3 Simulation setup

From the data model of (8) we simulate $k = 200$ complete data sets of size $n_A + n_B = 5000$ and part them into two separate files A with $n_A = 3000$ and $n_B = 2000$ observations. Then the Y values of file A are matched or imputed in file B according to the following algorithms

- NN; i. e., a nearest neighbor match (always assuming conditional independence),
- RI; i. e., a regression imputation under different conditional correlations,
- RIEPS; i. e., a regression imputation with stochastic residual under different conditional correlations, and
- NIBAS; i. e., the proposed MI algorithm assuming different prior conditional correlations.

For details of the formulae used in RI and RIEPS see Rässler (2002). Notice that NN and RI are single imputation procedures whereas RIEPS and NIBAS create more than one imputed data set. However, imputations produced by RIEPS are expected to underestimate variability because they lack from additional random draws of the parameters. Finally, small 1% and 5% complete auxiliary files are created according to the data model and used with the multiple imputation algorithm NORM (standalone software NORM 2.03) that is provided by Schafer (1997). With NORM it is not possible to use a real informative prior for the unknown correlations, therefore, NORM is applied herein for the data fusion situation when some auxiliary data are available containing information about all variables X, Y, and Z.

This procedure of creating the data, dividing and matching them is carried out 200 times. Relevant point and interval estimates are stored and tabulated. For the MI procedures $m = 5$ imputations are used. The MI estimates are calculated according to $\widehat{\theta}_{MI} = \frac{1}{m}\sum_{t=1}^{m}\widehat{\theta}^{(t)}$, as well as the within-imputation variance $W = \frac{1}{m}\sum_{t=1}^{m}\widehat{var}(\widehat{\theta}^{(t)})$, and the between-imputation variance $B = \frac{1}{m-1}\sum_{t=1}^{m}(\widehat{\theta}^{(t)} - \widehat{\theta}_{MI})^2$. The 95% MI interval estimates are calculated with $\widehat{\theta}_{MI} \pm \sqrt{T}\, t_{0.975,\nu}$, $T = W + (1 + m^{-1})B$, and degrees of freedom $v = (m-1)\left(1 + \frac{W}{(1+m^{-1})B}\right)^2$. According to the MI principle we assume that based on the complete data the point estimates $\widehat{\theta}$ are approximately normal with mean θ and variance $\widehat{var}(\widehat{\theta})$.[5] Therefore, some estimates should be transformed to a scale for which the normal approximation works well. For example, the sampling distribution of Pearson's

[5]Notice that Barnard and Rubin (1999) relax this assumption of a normal reference distribution to allow a t-distribution for the complete-data interval estimates and tests.

correlation coefficient $\widehat{\rho}$ is known to be skewed, especially if the corresponding correlation coefficient of the population is large. Thus, usually the multiple imputation point and interval estimates of a correlation ρ are calculated by means of the Fisher z-transformation $z(\widehat{\rho}) = 0.5 \ln(\frac{1+\widehat{\rho}}{1-\widehat{\rho}})$, which makes $z(\widehat{\rho})$ approximately normally distributed with mean $z(\rho)$ and constant variance $1/(n-3)$, see, e.g., Schafer (1997, p. 216). By back transforming the corresponding MI point and interval estimates of $z(\rho)$ via the inverse Fisher transformation the final estimates and confidence intervals for ρ are achieved.

6.4 Results

The following tables show the estimated expectations of some point estimates. In addition, the tables give the simulated actual coverage, i. e., the number of times out of 200 that cover the true parameter value. To ease the reading we display the percentage. Also the average length of the confidence intervals is reported (ALCI). The following Tables 1, 2, and 3 concentrate on the most important results.

Table 1 shows the preservation of the prior values of the conditional correlation between X and Y_1 or Y_2, respectively. As it was to be expected, the nearest neighbor match always establishes conditional independence. Thus, this matching procedure only works, when the conditional independence assumption is satisfied. Even with slight derivations from it, see block 4 in Table 1, the simulated actual coverage is far beyond its true nominal value. Also the single regression imputation does not reflect the correct coverage and typically leads to a strong overestimation of the true population correlation. The regression imputation with random residual performs quite well as long as the true unconditional correlation is not too high. Best in all cases is the new procedure NIBAS. In every setting it preserves the prior correlation with a higher nominal coverage than expected.[6]

Moreover, its average confidence intervals are only a little bit larger than those produced by RIEPS. Also Table 1 demonstrates that the multiple imputation procedure NORM very efficiently allows to use auxiliary data. With an additional file of size 5%, i. e., a file of only 250 observations completely observed in X, Y, and Z, the simulated actual coverage in most of the cases is higher than its nominal value. For NIBAS and RIEPS we could also use auxiliary information to estimate the potential prior correlations therefrom, but other simulations have shown that NORM is more powerful here, see Rässler (2002). With NORM the confidence intervals are typically much larger than with NIBAS. Thus, when prior information has to be used, NIBAS is the best choice at hand.

[6]Notice that according to classical and current formal definition of confidence intervals such conservative intervals are valid.

	$\widehat{E}(\widehat{\rho}_{XY_1})$	ALCI	Cvg.	$\widehat{E}(\widehat{\rho}_{XY_2})$	ALCI	Cvg.
Procedure	$\rho_{XY_1\|Z} = -0.6032, \rho_{XY_1} = 0.4$			$\rho_{XY_2\|Z} = 0.6393, \rho_{XY_2} = 0.8$		
NN	0.5810	0.0581	0.000	0.4566	0.0694	0.000
RI	0.4204	0.0722	0.805	0.9480	0.0089	0.000
RIEPS	0.4054	0.0771	0.960	0.8496	0.0328	0.000
NIBAS	0.3984	0.0819	0.985	0.7989	0.0467	1.000
NORM 1%	0.4201	0.1256	0.890	0.7824	0.1330	0.785
NORM 5%	0.3948	0.0985	0.960	0.8010	0.1372	1.000
Procedure	$\rho_{XY_1\|Z} = -0.2742, \rho_{XY_1} = 0.5$			$\rho_{XY_2\|Z} = 0.2651, \rho_{XY_2} = 0.6$		
NN	0.5828	0.0579	0.000	0.4584	0.0692	0.000
RI	0.5415	0.0620	0.265	0.8125	0.0298	0.000
RIEPS	0.5029	0.0730	0.960	0.6108	0.0764	0.980
NIBAS	0.5003	0.0753	0.995	0.6000	0.0815	1.000
NORM 1%	0.5331	0.1626	0.865	0.5604	0.2704	0.820
NORM 5%	0.4931	0.1084	0.960	0.6102	0.1963	0.970
Procedure	$\rho_{XY_1\|Z} = 0, \rho_{XY_1} = 0.5833$			$\rho_{XY_2\|Z} = 0, \rho_{XY_2} = 0.4583$		
NN	0.5817	0.0580	0.970	0.4579	0.0693	0.940
RI	0.6354	0.0523	0.025	0.6410	0.0517	0.000
RIEPS	0.5828	0.0664	0.995	0.4589	0.0941	1.000
NIBAS	0.5830	0.0664	0.995	0.4581	0.0993	1.000
NORM 1%	0.6018	0.1741	0.920	0.4310	0.3184	0.920
NORM 5%	0.5732	0.1033	0.955	0.4702	0.2033	0.950
Procedure	$\rho_{XY_1\|Z} = 0.0548, \rho_{XY_1} = 0.6$			$\rho_{XY_2\|Z} = 0.078, \rho_{XY_2} = 0.5$		
NN	0.5818	0.0580	0.765	0.4590	0.0692	0.345
RI	0.6540	0.0502	0.015	0.6981	0.0450	0.000
RIEPS	0.5999	0.0646	0.975	0.5014	0.0914	1.000
NIBAS	0.5998	0.0648	0.960	0.5006	0.0934	1.000
NORM 1%	0.6286	0.1891	0.940	0.4536	0.3275	0.905
NORM 5%	0.5921	0.0984	0.905	0.5063	0.1831	0.945
Procedure	$\rho_{XY_1\|Z} = 0.7129, \rho_{XY_1} = 0.8$			$\rho_{XY_2\|Z} = -0.6705, \rho_{XY_2} = 0.1$		
NN	0.5821	0.0580	0.000	0.4578	0.0693	0.000
RI	0.8331	0.0268	0.030	0.1168	0.0864	0.850
RIEPS	0.8156	0.0316	0.535	0.1055	0.0965	0.975
NIBAS	0.7999	0.0385	0.965	0.1008	0.1202	0.995
NORM 1%	0.7993	0.0888	0.945	0.1012	0.1986	0.970
NORM 5%	0.7972	0.0559	0.990	0.1063	0.1339	0.980

Table 1: Results for preserving the correlation structure

	$\widehat{E}(\widehat{\mu}_{Y_1})$	Cvg.	$\widehat{E}(\widehat{\mu}_{Y_2})$	Cvg.	$\widehat{E}(\widehat{\sigma}^2_{Y_1})$	Cvg.	$\widehat{E}(\widehat{\sigma}^2_{Y_2})$	Cvg.	$\widehat{E}(\widehat{\rho}_{Y_1Y_2})$	Cvg.
Procedure	$\rho_{XY_1\|Z} = -0.6032, \rho_{XY_1} = 0.4$ and $\rho_{XY_2\|Z} = 0.6393, \rho_{XY_2} = 0.8$									
NN	0.0032	0.93	-0.0025	0.880	0.9903	0.910	0.9972	0.855	0.3975	0.885
RI	0.0019	0.96	-0.0009	0.935	0.8986	0.090	0.7116	0.000	0.6532	0.000
RIEPS	0.0019	0.98	-0.0014	0.980	0.9668	0.820	0.8863	0.055	0.4994	0.000
NIBAS	0.0009	0.99	0.0001	1.000	0.9998	0.980	1.0025	0.985	0.3995	0.995
NORM 1%	0.0066	0.97	-0.0023	0.995	0.9978	0.985	1.0052	1.000	0.4065	0.990
NORM 5%	0.0022	0.99	-0.0002	1.000	0.9906	0.960	1.0187	0.995	0.3964	0.990
Procedure	$\rho_{XY_1\|Z} = -0.2742, \rho_{XY_1} = 0.5$ and $\rho_{XY_2\|Z} = 0.2651, \rho_{XY_2} = 0.6$									
NN	0.0006	0.920	-0.0016	0.890	0.9917	0.915	0.9956	0.875	0.3988	0.815
RI	0.0000	0.925	-0.0005	0.880	0.8540	0.000	0.5464	0.000	0.8925	0.000
RIEPS	0.0003	0.965	-0.0009	0.985	0.9890	0.965	0.9687	0.945	0.4288	0.825
NIBAS	-0.0001	0.955	-0.0006	1.000	1.0018	0.980	1.0006	1.000	0.3996	0.985
NORM 1%	0.0019	0.935	0.0040	0.980	0.9999	0.995	1.0014	1.000	0.4069	0.995
NORM 5%	-0.0008	0.990	0.0026	1.000	0.9929	0.985	1.0180	0.985	0.3953	0.980
Procedure	$\rho_{XY_1\|Z} = 0, \rho_{XY_1} = 0.5833$ and $\rho_{XY_2\|Z} = 0, \rho_{XY_2} = 0.4583$									
NN	0.0027	0.920	-0.0017	0.930	0.9897	0.915	0.9925	0.870	0.3995	0.895
RI	0.0015	0.955	0.0012	0.920	0.8428	0.000	0.5112	0.000	0.9600	0.000
RIEPS	0.0013	0.975	0.0010	0.995	1.0008	0.980	1.0003	0.990	0.4012	1.000
NIBAS	0.0013	0.970	0.0011	1.000	1.0006	0.995	1.0013	0.995	0.4009	0.995
NORM 1%	0.0014	0.975	0.0092	0.985	1.0017	1.000	0.9983	1.000	0.4061	0.995
NORM 5%	0.0005	0.995	0.0046	1.000	0.9931	0.990	1.0179	0.990	0.3964	0.990

Table 2: Results for preserving the moments of the fused/imputed variable (1)

	$\widehat{E}(\widehat{\mu}_{Y_1})$	Cvg.	$\widehat{E}(\widehat{\mu}_{Y_2})$	Cvg.	$\widehat{E}(\widehat{\sigma}^2_{Y_1})$	Cvg.	$\widehat{E}(\widehat{\sigma}^2_{Y_2})$	Cvg.	$\widehat{E}(\widehat{\rho}_{Y_1Y_2})$	Cvg.
Procedure	$\rho_{XY_1\|Z} = 0.0548, \rho_{XY_1} = 0.6$ and $\rho_{XY_2\|Z} = 0.078, \rho_{XY_2} = 0.5$									
NN	0.0028	0.910	0.0021	0.880	0.9914	0.935	0.9973	0.910	0.3969	0.885
RI	0.0017	0.925	0.0035	0.835	0.8414	0.000	0.5150	0.000	0.9582	0.000
RIEPS	0.0015	0.950	0.0041	0.985	0.9998	0.985	0.9994	0.990	0.3987	0.990
NIBAS	0.0021	0.970	0.0032	1.000	0.9990	0.995	1.0038	1.000	0.3996	0.995
NORM 1%	0.0014	0.950	0.0107	0.970	0.9994	0.995	1.0018	1.000	0.4058	1.000
NORM 5%	0.0007	0.980	0.0068	0.995	0.9916	0.985	1.0198	0.995	0.3948	0.990
Procedure	$\rho_{XY_1\|Z} = 0.7129, \rho_{XY_1} = 0.8$ and $\rho_{XY_2\|Z} = -0.6705, \rho_{XY_2} = 0.1$									
NN	-0.0033	0.935	0.0015	0.920	0.9891	0.880	0.9922	0.845	0.3970	0.890
RI	-0.0017	0.935	0.0005	0.910	0.9205	0.235	0.7303	0.000	0.6033	0.000
RIEPS	-0.0018	0.940	0.0009	0.965	0.9604	0.775	0.8963	0.115	0.4968	0.005
NIBAS	-0.0017	0.960	0.0005	0.980	0.9975	0.985	1.0019	0.995	0.3993	0.995
NORM 1%	-0.0014	0.955	0.0075	0.970	0.9944	0.990	1.0122	0.990	0.4025	1.000
NORM 5%	-0.0003	0.965	0.0000	0.985	0.9867	0.960	1.0108	1.000	0.4005	0.995

Table 3: Results for preserving the moments of the fused/imputed variable (2)

The preservation of the distributions of the matched or imputed variables Y_1 and Y_2 is displayed in Tables 2 and 3. Again the nearest neighbor match leads to similar results regardless of the true correlations between X and Y_1 or Y_2. Always the coverage is too low and the variances are underestimated as it is typical for single imputation approaches. Regression imputation also typically underestimates the variances even more, the coverage often is 0. As before, adding a random residual improves the regression imputation considerably but not in all cases. The best preservation again provides NIBAS, throughout the coverage is higher than its nominal value. Using auxiliary data works fine for NORM also if only 1% (i. e., 50 observations) are completely observed.

7 Summary

In this paper we structure the validity a data fusion procedure may achieve by four levels. It is shown that the forth level is meaningless, and only the first level typically is controlled for when traditional techniques of data fusion are applied. Provided that one of the vectors of specific variables is univariate, we derive bounds for the correlations between variables not jointly observed and suggest a new quality index of data fusion which is built upon these bounds. Then, the preservation of the joint distribution and the correlation structure of the variables not jointly observed can be evaluated by using the non-iterative multiple imputation procedure NIBAS. Since data fusion can be viewed as a problem of missing data, MI procedures are applicable in general. Auxiliary data can be easily and efficiently used by standard MI procedures such as NORM. In a simulation study, we find the multiple imputation approaches superior to the traditional matching techniques.

References

Barnard, J. and Rubin, D.B. (1999). Small-Sample Degrees of Freedom with Multiple Imputation, *Biometrika*, **86**, 948–955.

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

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

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

Higham, N.J. (2002). Computing the Nearest Correlation Matrix - a Problem from Finance, *IMA Journal of Numerical Analysis*, **22**, 329–343.

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

Liu, T.P. and 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. and Scheuren, F. (2001). Statistical Matching: A Paradigm for Assessing the Uncertainty in the Procedure, *Journal of Official Statistics*, **17**, 407–422.

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

D'Orazio, M., Di Zio, M., and 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, Springer, New York.

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

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

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

Rubin, D.B. (1978). Multiple Imputations in Sample Surveys - a Phenomenological Bayesian Approach to Nonresponse, *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 20–34.

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*. Wiley, New York.

Rubin, D.B. (2002). Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation, *Health Services & Outcomes Research Methodology*, **2**, 178–186.

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

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman and Hall, London.

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.

Van der Putten, P., Kok, J.N., and Gupta, A. (2002). Data Fusion Through Statistical Matching, *MIT Sloan School of Management*, Working Paper 4342-02.

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

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

Recently published

No.	Author(s)	Title	Date
1/2004	Bauer, T. K. Bender, S. Bonin, H.	Dismissal protection and worker flows in small establishments	7/04
2/2004	Achatz, J. Gartner, H. Glück, T.	Bonus oder Bias? : Mechanismen geschlechtsspezifischer Entlohnung published in: Kölner Zeitschrift für Soziologie und Sozialpsychologie 57 (2005), S. 466-493 (revised)	7/04
3/2004	Andrews, M. Schank, T. Upward, R.	Practical estimation methods for linked employer-employee data	8/04
4/2004	Brixy, U. Kohaut, S. Schnabel, C.	Do newly founded firms pay lower wages? First evidence from Germany	9/04
5/2004	Kölling, A. Rässler, S.	Editing and multiply imputing German establishment panel data to estimate stochastic production frontier models published in: Zeitschrift für ArbeitsmarktForschung 37 (2004), S. 306-318	10/04
6/2004	Stephan, G. Gerlach, K.	Collective contracts, wages and wage dispersion in a multilevel model	10/04
7/2004	Gartner, H. Stephan, G.	How collective contracts and works councils reduce the gender wage gap	12/04
1/2005	Blien, U. Suedekum, J.	Local economic structure and industry development in Germany, 1993-2001	1/05
2/2005	Brixy, U. Kohaut, S. Schnabel, C.	How fast do newly founded firms mature? : empirical analyses on job quality in start-ups published in: Michael Fritsch, Jürgen Schmude (Ed.): Entrepreneurship in the region, New York et al., 2006, S. 95-112	1/05
3/2005	Lechner, M. Miquel, R. Wunsch, C.	Long-run effects of public sector sponsored training in West Germany	1/05
4/2005	Hinz, T. Gartner, H.	Lohnunterschiede zwischen Frauen und Männern in Branchen, Berufen und Betrieben published in: Zeitschrift für Soziologie 34 (2005), S. 22-39, as: Geschlechtsspezifische Lohnunterschiede in Branchen, Berufen und Betrieben	2/05
5/2005	Gartner, H. Rässler, S.	Analyzing the changing gender wage gap based on multiply imputed right censored wages	2/05
6/2005	Alda, H. Bender, S. Gartner, H.	The linked employer-employee dataset of the IAB (LIAB) published in: Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften 125 (2005), S. 327-336, (shortened) as: The linked employer-employee dataset created from the IAB establishment panel and the process-produced data of the IAB (LIAB)	3/05
7/2005	Haas, A. Rothe, T.	Labour market dynamics from a regional perspective : the multi-account system	4/05
8/2005	Caliendo, M. Hujer, R. Thomsen, S. L.	Identifying effect heterogeneity to improve the efficiency of job creation schemes in Germany	4/05
9/2005	Gerlach, K. Stephan, G.	Wage distributions by wage-setting regime	4/05
10/2005	Gerlach, K.	Individual tenure and collective contracts	4/05

	Stephan, G.		
11/2005	Blien, U. Hirschenauer, F.	Formula allocation : the regional allocation of budgetary funds for measures of active labour market policy in Germany	4/05
12/2005	Alda, H. Allaart, P. Bellmann, L.	Churning and institutions : Dutch and German establishments compared with micro-level data	5/05
13/2005	Caliendo, M. Hujer, R. Thomsen, S. L.	Individual employment effects of job creation schemes in Germany with respect to sectoral heterogeneity	5/05
14/2005	Lechner, M. Miquel, R. Wunsch, C.	The curse and blessing of training the unemployed in a changing economy : the case of East Germany after unification	6/05
15/2005	Jensen, U. Rässler, S.	Where have all the data gone? : stochastic production frontiers with multiply imputed German establishment data	7/05
16/2005	Schnabel, C. Zagelmeyer, S. Kohaut, S.	Collective bargaining structure and its determinants : an empirical analysis with British and German establishment data published in: European Journal of Industrial Relations, Vol. 12, No. 2. S. 165-188	8/05
17/2005	Koch, S. Stephan, G. Walwei, U.	Workfare: Möglichkeiten und Grenzen published in: Zeitschrift für ArbeitsmarktForschung 38 (2005), S. 419-440	8/05
18/2005	Alda, H. Bellmann, L. Gartner, H.	Wage structure and labour mobility in the West German private sector 1993-2000	8/05
19/2005	Eichhorst, W. Konle-Seidl, R.	The interaction of labor market regulation and labor market policies in welfare state reform	9/05
20/2005	Gerlach, K. Stephan, G.	Tarifverträge und betriebliche Entlohnungsstrukturen	11/05
21/2005	Fitzenberger, B. Speckesser, S.	Employment effects of the provision of specific professional skills and techniques in Germany	11/05
22/2005	Ludsteck, J. Jacobebbinghaus, P.	Strike activity and centralisation in wage setting	12/05
1/2006	Gerlach, K. Levine, D. Stephan, G. Struck, O.	The acceptability of layoffs and pay cuts : comparing North America with Germany	1/06
2/2006	Ludsteck, J.	Employment effects of centralization in wage setting in a median voter model	2/06
3/2006	Gaggermeier, C.	Pension and children : Pareto improvement with heterogeneous preferences	2/06
4/2006	Binder, J. Schwengler, B.	Korrekturverfahren zur Berechnung der Einkommen über der Beitragsbemessungsgrenze	3/06
5/2006	Brixy, U. Grotz, R.	Regional patterns and determinants of new firm formation and survival in western Germany	4/06
6/2006	Blien, U. Sanner, H.	Structural change and regional employment dynamics	4/06
7/2006	Stephan, G. Rässler, S. Schewe, T.	Wirkungsanalyse in der Bundesagentur für Arbeit : Konzeption, Datenbasis und ausgewählte Befunde	4/06
8/2006	Gash, V. Mertens, A. Romeu Gordo, L.	Are fixed-term jobs bad for your health? : a comparison of West-Germany and Spain	5/06
9/2006	Romeu Gordo, L.	Compression of morbidity and the labor supply of older people	5/06

10/2006	Jahn, E. J. Wagner, T.	Base period, qualifying period and the equilibrium rate of unemployment	6/06
11/2006	Jensen, U. Gartner, H. Rässler, S.	Measuring overeducation with earnings frontiers and multiply imputed censored income data	6/06
12/2006	Meyer, B. Lutz, C. Schnur, P. Zika, G.	National economic policy simulations with global interdependencies : a sensitivity analysis for Germany	7/06
13/2006	Beblo, M. Bender, S. Wolf, E.	The wage effects of entering motherhood : a within-firm matching approach	8/06
14/2006	Niebuhr, A.	Migration and innovation : does cultural diversity matter for regional R&D activity?	8/06

Letzte Aktualisierung: 30.8.2006, 43 Einträge

Imprint

IABDiscussionPaper
No. 15 / 2006

Editorial address
Institut für Arbeitsmarkt- und Berufsforschung
der Bundesagentur für Arbeit
Weddigenstr. 20-22
D-90478 Nürnberg

Editorial staff
Regina Stoll, Jutta Palm-Nowak

Technical completion
Jutta Sebald

Download of this DiscussionPaper:
http://doku.iab.de/discussionpapers/2006/dp1506.pdf

Website
http://www.iab.de

For further inquiries contact the author:
Hans Kiesl, Tel. 0911/179-1358,
or e-mail: hans.kiesl@iab.de