Multiple Imputation in Data Fusion: Making Better Assumptions than Conditional Independence

Volker Bosch* Philipp Gaffert †

Abstract

The missing pattern of data fusion implies that the variables that are specific to the data sets are never jointly observed. When applying standard imputation techniques, independence conditioned on the common variables is implicitly assumed. In general, however, this assumption does not hold; consequently, the estimated correlations between the fused specific variables are usually biased toward zero. We argue that in the absence of further information, a correlation lying well within the bounds of the conditional independence assumption (CIA) and one specific measurement error model is a significantly more sensible assumption. This argument is derived from a simple trivariate model and empirically supported by data from various fields.

Key Words: Data Fusion, Conditional Independence Assumption, Measurement Error Model

1. Introduction

Data fusion can be regarded as a missing data problem (Little & Rubin, 2002, p. 5) and has been thoroughly treated in textbooks (Raessler (2002), D'Orazio et al. (2006)). Two data sets with specific variables Y measured in the first set and Z measured in the second set are being combined through common variables X. Specifically, Y and Z have not been jointly observed.

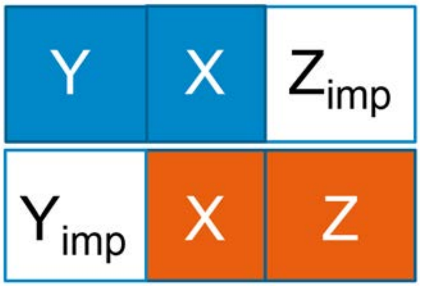

Figure 1: Fusion of two data sets with a common variable X and specific variables Y and Z never jointly observed

Consequently, the correlation ρ_{yz} is not identified. Nevertheless, the correlations with X, namely ρ_{xy} and ρ_{xz}, impose some boundaries. In the case of X, Y, and Z being univariate, the boundaries imposed by ρ_{xy} and ρ_{xz} on ρ_{yz} are set by the positive semi-definiteness of the correlation matrix Σ_{xyz}. Solving $det(\Sigma_{xyz}) = 0$ yields (Raessler, 2002, p. 10)

$$\rho_{yz} \in \left[\rho_{xy} \cdot \rho_{xz} \pm \sqrt{(1-\rho_{xy}^2)\cdot(1-\rho_{xz}^2)}\,\right]. \tag{1}$$

[*] Marketing & Data Sciences, GfK SE, Nordwestring 101, 90419 Nuremberg, Germany

[†] Marketing & Data Sciences, GfK SE, Nordwestring 101, 90419 Nuremberg, Germany

Consequently, the boundaries are relatively strict only when the absolute values of ρ_{xy} and ρ_{xz} are high. In cases where the correlations with X are only moderate, the range of possible values for ρ_{yz} is close to the maximal interval $[-1, 1]$.

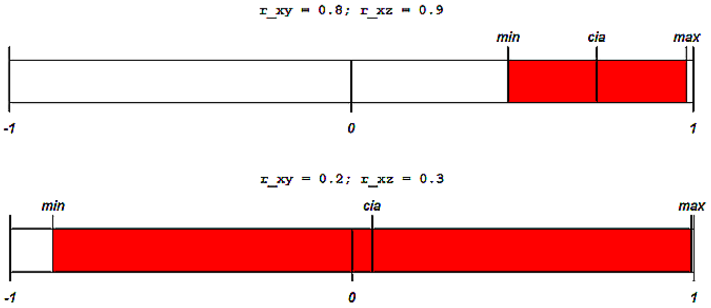

Figure 2: Boundaries on ρ_{yz} imposed by ρ_{xy} and ρ_{xz}

This is why, the imputer must essentially base the value of ρ_{yz} on assumptions. The default assumption for estimating ρ_{yz} in data fusion is the Conditional Independence Assumption (CIA), which implies that $E(\rho_{yz|x}) = 0$ and consequently $\rho_{yz}^{cia} = \rho_{xy} \cdot \rho_{xz}$. Adopting the CIA is convenient for two main reasons:

1. Standard imputation algorithms (as with distance matching (D'Orazio et al., 2006, pp. 13, 29, 41) and predictive mean matching (Rubin (1986), Little (1988)) implicitly produce results based on the CIA.

2. The resulting correlation ρ_{yz}^{cia} is the midpoint of the theoretically possible interval. Hence, in the absence of additional information, ρ_{yz}^{cia} seems to be a bias-minimizing guess.

In market research, a classical application is the fusion of a TV audience measurement panel (TAM) with a consumer panel (CP) (Wendt, 1986). The objective here is to estimate ad effectiveness, that is, the return on investment (ROI) for TV advertisements. Let Y measure ad contact, and let Z measure the purchase of the respective advertised product. Then, the correlation ρ_{yz} is a first indicator for the ROI. Obviously, it is extremely important for creative agencies to obtain estimates for ρ_{yz} with a low bias. This is especially relevant, as the observed correlations of Y or Z with X rarely exceed 0.5. In this case, the boundaries according to formula 1 are hardly restrictive. A typical range is $\rho_{yz} \in \Big[-0.7;\ 1.0\Big]$.

2. Is the CIA Biasing?

Sims (1972) already noted that the assumption of conditional independence is lacking theoretical backing. Indeed, there are several reasons to doubt the CIA's validity. For instance, when analyzing a correlation matrix

$$\Sigma = \begin{bmatrix} 1 & . & . \\ \rho_{12} & 1 & . \\ \rho_{13} & \rho_{23} & 1 \end{bmatrix}$$

and labeling any of the correlations with ρ_{yz}, an estimate based on the CIA will be the product of the remaining two correlations. Since $|\rho_{yz}^{cia}| \leq \min(|\rho_{xy}|, |\rho_{xz}|)$, the correlations estimated through the CIA $\rho_{12}^{cia}, \rho_{13}^{cia}$, and ρ_{23}^{cia} will on average contain strongly attenuated absolute values compared to the original. This indicates that the CIA biases the results toward zero because on average the absolute values should be maintained.

An alternative identifying assumption is given by the proposition that X is an unbiased but imperfect measure of Y, i.e., $X = Y + \epsilon$. In this case, a measurement error model (MEM) can be utilized (Raghunathan, 2015, p. 156), and $\rho_{yz}^{mem} = \rho_{xz}/\rho_{xy}$ follows. In contrast to the CIA, the MEM will on average always increase the absolute values of the estimated correlations. This is as implausible as adopting the CIA unless additional information justifies either of the two assumptions.

3. A Simple Trivariate Model

The above arguments suggest that both the CIA and the MEM tend to provide strongly biased estimates unless additional information about the data generating process indicate that either assumption is appropriate. In the absence of such information, what might be a low bias assumption for $\rho_{yz} \mid \rho_{xz}, \rho_{xy}$? To address this question, we propose a simple trivariate model as a framework. In this framework, all components X, Y, Z, and $\epsilon_{.}$ follow standard normal distributions. Moreover, all $\epsilon_{.}$ are orthogonal to each other. The observed variables X, Y, and Z are constructed by up to four constituents ϵ that are either unique for the observed variable $(\epsilon_x, \epsilon_y, \epsilon_z)$, shared by exactly two observables $(\epsilon_{xy}, \epsilon_{xz}, \epsilon_{yz})$ or shared by all three observables (ϵ_{xyz}). X is constructed as a linear combination of its constituents, i.e.,

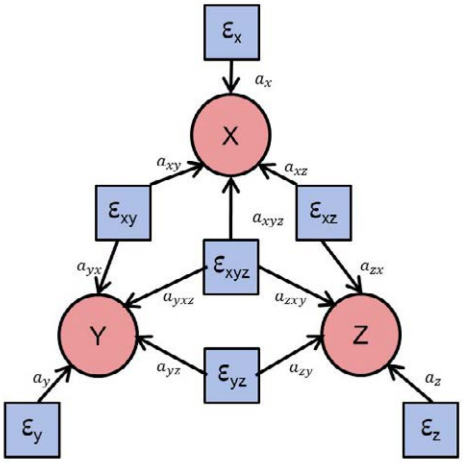

Figure 3: The framework

$$X = a_x\epsilon_x + a_{xy}\epsilon_{xy} + a_{xz}\epsilon_{xz} + a_{xyz}\epsilon_{xyz},$$

where the respective factors $a_{.}$ have to obey

$$a_x^2 + a_{xy}^2 + a_{xz}^2 + a_{xyz}^2 = 1$$

to ensure the unitary variance of X. This respectively holds for Y and Z as well. Correlations are determined by the respective factors of the mutually shared constituents of the model. For instance,

$$\rho_{xy} = a_{xy}a_{yx} + a_{xyz}a_{yxz}.$$

As there are twelve parameters and only six equations, this framework is clearly overspecified. Apparently, the model is easily identifiable when a highly simplified core model (see Figure 4) that is only extended when necessary is used. As a special case, we focus on the core model without extensions. This simple model is unable to describe the entire range of $\rho_{yz} \mid \rho_{xz}, \rho_{xy}$ (see equation 1) but rather is restricted to

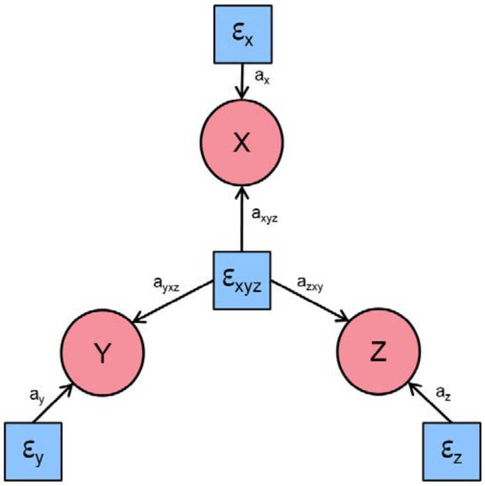

Figure 4: The Core Model

$$|\rho_{xz} \cdot \rho_{xy}| \leq |\rho_{yz}^{core}| \leq \min\left(|\frac{\rho_{xz}}{\rho_{xy}}|, |\frac{\rho_{xz}}{\rho_{xy}}|\right).$$

Specifically,

$$|\rho_{yz}^{cia}| \leq |\rho_{yz}^{core}| \leq |\rho_{yz}^{mem}|.$$

The CIA constitutes the lower bound and the MEM constitutes the upper bound of the core model's limited range. To extend beyond these boundaries, exactly one pairwise component, i.e., either $\epsilon_{xy}, \epsilon_{xz}$, or ϵ_{yz}, must be added. As ρ_{yz} is the unknown correlation, ϵ_{yz} seems to be a suitable candidate.

However, as any of the pairwise ϵ may be added to fill the entire range as given by equation 1, the core model can be regarded as constituting the 'center' of all the possible configurations. Specifically, although the CIA is at the center of the interval given by equation 1, it is an extreme case of the core model.

4. Alternatives to the CIA and the MEM

When we accept that the core model is indeed at the center of all possible data configurations, then an immediate candidate to replace the estimates based on the CIA or MEM is the mean of the core model's boundaries:

$$\rho_{yz}^{mean} = (\rho_{yz}^{cia} + \rho_{yz}^{mem})/2.$$

Another candidate results from minimizing the variance of the factors $a_{xyz}, a_{yxz}, a_{zxy}$ given the observed correlations ρ_{xy} and ρ_{xz} via a Lagrange approach, which gives

$$\rho_{yz}^{mva} = \frac{\rho_{xy} \cdot \rho_{xz}}{\sqrt{(\rho_{xy}^2 + \rho_{xz}^2)/2}}.$$

From Figure 5, it becomes apparent that the range of the possible values for ρ_{yz} (depicted in green) is considerably smaller when restricted to the core model. Moreover, both ρ_{yz}^{mean} and ρ_{yz}^{mva} are placed well within the core model's range. ρ_{yz}^{mva} is a more conservative estimate than ρ_{yz}^{mean} when the observed correlations are small. The opposite holds for high observed correlations ρ_{xy} and ρ_{xz}.

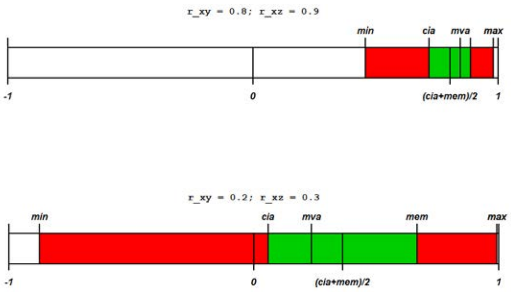

Figure 5: Values of ρ_{yz}^{cia}, ρ_{yz}^{mem} , ρ_{yz}^{mean}, and ρ_{yz}^{mva} as a function of ρ_{xy} and ρ_{xz}

5. Empirical Evidence

To assess the plausibility of ρ_{yz}^{cia} and ρ_{yz}^{mem} in a real data situation, a simple analysis was performed. Three variables were randomly selected out of correlation matrices from various sources, and one of the three correlations was assigned to ρ_{yz} (to not focus on noise, the absolute value of this correlation had to exceed 0.1). Next, the estimates ρ_{yz}^{cia} and ρ_{yz}^{mem} were computed based on ρ_{xy} and ρ_{xz}. Finally, the true correlation ρ_{yz} was positioned relative to the two estimates based on the CIA and the MEM.

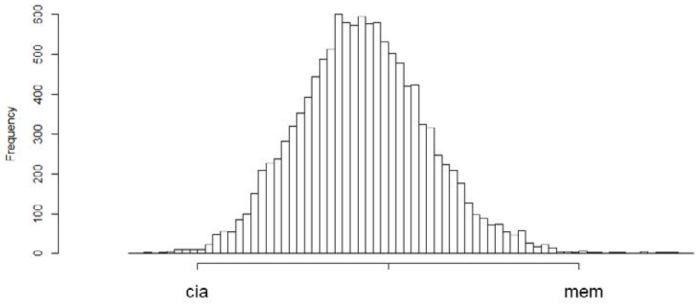

Figure 6: Values of ρ_{yz} relative to the estimates based on the CIA and the MEM. The underlying correlation matrix is derived from the Dow-Jones-30 constituent's closing-value log returns from 1988 from Yahoo Finance. The data are provided by the R-package 'rugarch', (Ghalanos, 2015), data set 'dji30_ret'.

Apparently, most of the correlations indeed stay within the range described by the core model (see Figure 6 based on financial data). In other correlation matrices of different sources, similar results emerged. Figure 7 is based on data from psychology (left panel) and education research (right panel).

6. Outlook

Because ρ_{yz} is unidentified in the data fusion problem, an assumption is required. It is still common to choose the Conditional Independence Assumption (CIA), which

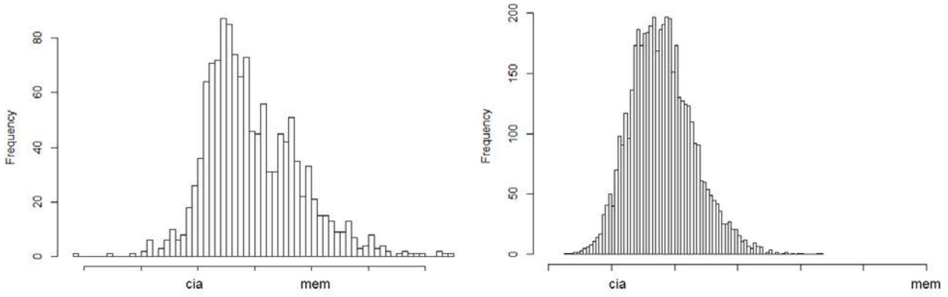

Figure 7: Values of ρ_{yz} relative to estimates based on the CIA and the MEM. Left Panel: Correlations between the Big Five measured on three occasions (Biesanz & West, 2004). Right Panel: Derived from items for a grade 12 science assessment test (SAT) measuring topics of chemistry, biology, and physics. Data from the R-package 'mirt', (Chalmers, 2012) data set 'SAT12'.

is most convenient for the statistician but likely to be inappropriate (Sims, 1972). We propose to learn from other complete data sets instead. Our above empirical examples show that the vast majority of observed correlations falls in an interval much smaller than theoretically possible. This interval is covered by the core model and bounded by the CIA and the measurement error model (MEM) denoted by $X = Y + \epsilon$. We thus argue that the CIA tends to underestimate ρ_{yz} and that alternative estimators like ρ_{yz}^{mean} or ρ_{yz}^{mva} are more suitable.

Clearly, external knowledge about the data generating process can provide a solution to the identification problem. It remains to be investigated, though, as to whether meta-data about the data set at hand contain any information about the non-identified correlation. Is the expected value of a non-identified correlation, for instance, larger if the data set contains financial data than if it were to contain psychological data?

Acknowledgements

The authors thank Susanne Rässler, Jakob Gräßel, Florian Meinfelder and Trivellore Raghunathan for their valuable inputs.

References

Biesanz, J. C. & West, S. G. (2004). Towards understanding assessments of the Big Five: Multitrait-multimethod analyses of convergent and discriminant validity across measurement occasion and type of observer. *Journal of Personality* **72**, 845–76.

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software* **48**, 1–29. R package version: 1.25.

D'Orazio, M., Di Zio, M. & Scanu, M. (2006). Statistical Matching: Theory and Practice. New York: Wiley.

Ghalanos, A. (2015). rugarch: Univariate GARCH models. R package version: 1.3-6.

Little, R. J. A. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics* **6**, 287–96.

Little, R. J. A. & Rubin, D. B. (2002). *Statistical Analysis with Missing Data.* Hoboken, NJ: Wiley, 2nd ed.

R Core Team (2017). *A Language and Environment for Statistical Computing: Version 3.4.1.* Vienna, Austria: R Foundation for Statistical Computing.

Raessler, S. (2002). *Statistical Matching: A Frequentist Theory. Practical Applications, and Alternative Bayesian Approaches.* New York: Springer.

RAGHUNATHAN, T. E. (2015). *Missing Data Analysis in Practice.* Boca Raton, FL: Chapman & Hall/CRC.

RUBIN, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics* **4**, 87–94.

SIMS, C. A. (1972). Comments and Rejoinder. *Annals of Economic and Social Measurement* **1**, 343–5 and 355–7.

WENDT, F. (1986). Einige Gedanken zur Fusion. *Auf dem Weg zum Partnerschaftsmodell*, 109–40. Frankfurt, Germany: Arbeitsgemeinschaft Media-Analyse e.V., Media-Micro-Census GmbH.