Congenial Multiple Imputation and Matched Pair Models for Square Tables

An Example of Patients' Self-management

Shakir Hussain[1], Mohamed A Mohammed[2] & Ghazi Shukur[3]

[1&3] Department of Economics and Statistics Linnaeus University, Sweden

[2] Primary care and general practice, School of Medicine, University of Birmingham, UK

[3] Department of Economics, Finance and Statistics, Jönköping International Business School, Sweden

Correspondence: Dr. Shakir Hussain, Department of Economics and Statistics Linnaeus University, Sweden. E-mail: Shakir.hussain@lnu.se

Received: December 1, 2012 Accepted: December 25, 2012 Online Published: January 7, 2013

doi:10.5430/jbar.v2n1p1 URL: http://dx.doi.org/10.5430/jbar.v2n1p1

Abstract

Experimental studies often measure an individual's quality of life before and after an intervention, with the data organized into a square table and analyzed using matched pair modeling. However, it is not unusual to find missing data in either round (i.e., before and/or after) of such studies and the use of multiple imputations with matched-pair modeling remains relatively unreported in the applied statistics literature. In this paper we introduce an approach which maintains dependency of responses over time and makes a match between the imputer and the analyst. We use 'before' and 'after' quality-of-life data from a randomized controlled trial to demonstrate how multiple imputation and matched-pair modeling can be congenially combined, avoiding a possible mismatch of imputation and analyses, and to derive a properly consolidated analysis of the quality-of-life data. We illustrate this strategy with a real-life example of one item from a quality-of-life study that evaluates the effectiveness of patients' self-management of anticoagulation versus standard care as part of a randomized controlled trial.

Keywords: Multiple imputation, Hierarchical log-linear models, Square table, Matched-pair model

1. Introduction

Matched-pair data is frequently seen in practice and is exemplified by studies that measure the patients' health status before and after some healthcare intervention. When the survey variables are polytomous and have the same categories, we can arrange the before and after data in a two-dimensional square table. Interest then lies in estimating of the marginal distribution of the row and column variables and the testing of whether or not these two distributions are identical. Analysis of matched-pair data can be undertaken using a Poisson generalized linear model for the counts in the square table as described by Agresti (2012).

As with other surveys, it is not unusual to find missing data in either round or square tables. Listwise deletions or 'complete case analysis' are in general to be avoided in preference to multiple imputation (MI), see Little and Rubin (2002). Single imputation, such as the mean or the regression imputation, ignores the uncertainty caused by the missing data. Multiple imputations, involving more than one set of imputation, allow valid assessment of the uncertainty that comes from missing data, see Rubin (1987). This may result in less biased but less precise estimates. MI maintains power by preserving the original sample size. It also reduces the risk of bias, especially when missingness is ignorable and occurs at random, i.e. missing at random (MAR), or if one uses a missing completely at random (MCAR) assumption for MI. Note that in the case when the data are MCAR, which is also an ignorable missing, complete case analysis does not bias the results when the rate of missing is not considerable. However, MI for square tables needs to address the dependency of paired data. As mentioned in Schafer (2003), this happens when we have data over an item for different subjects, but some portions are missing for reasons beyond our control. Moreover, if the item represents measures of the same variables on different occasions, and if the subjects who drop out do not return, then the MAR is easy to understand. In other words, it means that the subject's probability of responding out at specific occasion, given that it has not yet dropped out, may depend on previous responses, may

depend only on its own set of observed items but not on the present or future. In the presence of such correlation or dependency, which differs from one individual to another and cannot be generalized to the entire sample, we are forced to classify this kind missing as not at random (i.e. MNAR), and the MAR should not be assumed for this missing mechanism.

For MI under a variety of MNAR models and the example of MI for MNAR mechanisms, see Schafer (2003). See also Little (1985) for different paths of MNAR associations.

MI is undertaken to generate completed data sets where the imputed values should be close to the values that could have been observed. Moreover, marginal distribution should be the same as the true corresponding distribution and the MI procedure should lead to imputed values that are plausible. The most important conditions are that the estimation results of parameters are unbiased and inferences based on them should be efficient. These conditions may be achieved if imputation and analyses are undertaken congenially under the same model (or if the imputer and the analyst / data editor are integrated) (Schafer 1997), thereby avoiding a possible mismatch between imputation and analyses.

Nevertheless, several inconsistencies may appear between the imputer and the analyst. For example, the analyst may assume or impose more restrictions than the imputer. Schafer (1997) shows that in this case there is nothing wrong with the imputed data set and that inferences using MI are still valid, but interval estimates will be wider (i.e. the results are inefficient)(Note 1). In this case the analyst's model can be considered as a special case of the imputer's model. At this stage it is important to mention that this problem may be due to the analyst and not the data imputer (i.e. the analyst omits relevant information from the analysis which can lead to some kind of model mis-specification). Another type of inconsistency arises when the imputer assumes more (or puts more restrictions on the data) than the analyst. This happens when the analyst's model is more general than the imputer's model. Even in this case, the practical consequences of this kind of inconsistency will depend on whether or not the imputer's extra assumptions are valid.

Several researchers have investigated the case when these extra assumptions are valid, as described by Meng (1994) and Rubin (1987), and found that the point estimates were still unbiased, more efficient and had shorter intervals than observed data estimates derived completely from the analyst's model. Meng and Rubin explain that this happens because MI incorporates the imputer's superior knowledge about the state of nature. That is, the additional information supported by the imputer cannot invalidate the model; on the contrary, it can only help. Conversely, where the imputer's extra assumptions are not valid, the model will be erroneous and any MI data created under a mis-specified model may lead to misleading results and conclusions. So it seems that the most serious concern is when the imputer imposes invalid assumptions that will later be the subject of the analyst's inquiry.

In this paper we recommend a congenial working strategy that can be used to minimize possible inconsistencies between the imputer and the analyst (or the data editor) which might happen when the data include missing values in either round (before and/or after) of a square table. We illustrate this strategy with a real-life example of one item from a quality-of-life study which evaluates the effectiveness of patients' self-management (PSM) of anticoagulation versus standard care as part of a randomized controlled trial. The imputation and analyses are congenially combined by ensuring that they consider only plausible mechanisms for the pattern of missing data in the square table (e.g., round two missing cannot influence round one missing, but round one missing can influence round two missing). The calculations in this study are performed using the statistical program packages S+ version 8.0.1 and some libraries in R. The R and S+ codes used in this paper are available from the authors upon request.

In a special volume of the *Journal of Statistical Software*, Recai (2011) provides a brief history of multiple imputation and relevant software. Potential directions for the future of the software development are also provided.

The rest of the paper is organized as follows: Section 2 gives a brief presentation of the example data, while the methodology is discussed in Section 3. In Section 4 we present the estimated results. Finally, we give a short summary and conclusions in Section 5.

2. Data description

The data were collected in the West Midlands region of the UK for two groups of patients, one self-managing (PSM) anticoagulation and the other receiving standard care (control). After a period of 12 months the patients were asked to complete a Spielberger quality-of-life questionnaire. For more information about the study we refer the reader to Fitzmaurice et al. (2005). Note that, in our paper, we only consider the first item, 'I am worried', from this questionnaire for PSM patients to illustrate our strategy; the other items can be analyzed in a similar manner, see

Table 1. The Spielberger questionnaire has the following four response options: not at all; somewhat; moderately; and very much.

Table 1. Spielberger questionnaire

Item	Response			
	Not at all	Somewhat	Moderately	Very much
I am worried	1	2	3	4
I feel calm	1	2	3	4
I am tense	1	2	3	4
I feel upset	1	2	3	4
I feel relaxed	1	2	3	4
I feel content	1	2	3	4

Since there were many empty cells in the last option of the first analyzed item, we collapsed the total number of options to three (instead of four) by joining options three and four together.

Table 2 shows that 26% (46+18-9/213) of the data is missing: 18 observations in the first round and 46 observations in the second round, and 9 observations are missing in both rounds. If there were no missing data, then the quality-of-life responses could be organized into a 3x3 square table. Note that a listwise deletion of the missing data would end up with a square table, but this would be based on 158 observations from an original sample size of 213.

Table 2. Participant counts in two rounds of the quality-of-life study

I am worried...	Not at all	Somewhat	Moderately / very much"	Missing	Total
Not at all	95 (1,1)	9 (1,2)	10 (1,3)	22	136
Somewhat	13 (2,1)	13 (2,2)	2 (2,3)	12	40
Moderately / very much	6 (3,1)	7 (3,2)	3 (3,3)	3	19
Missing	4	4	1	9	18
Total	118	33	16	46	213

Rows are before-intervention and columns are after-intervention responses and () show row and column numbers for the 3x3 square table.

3. Investigating the mechanism for missingness

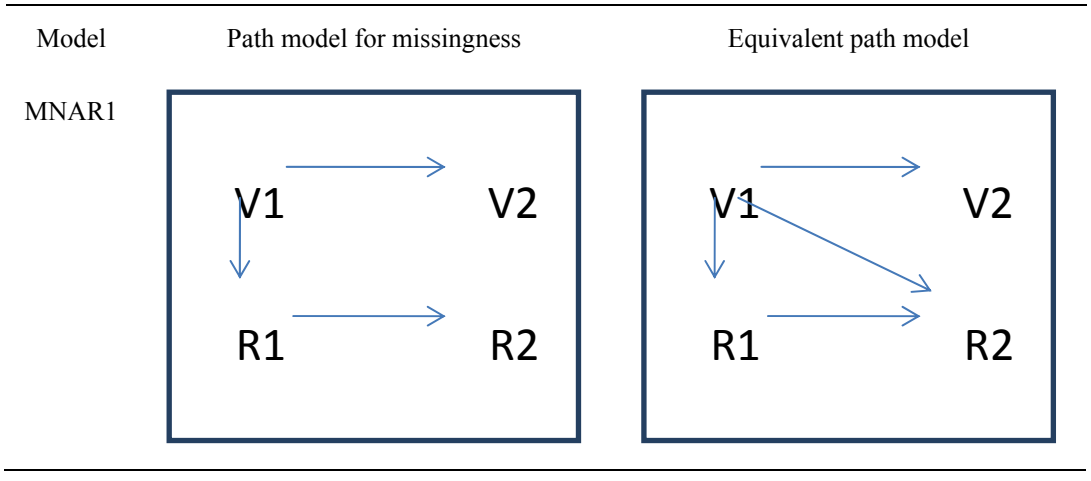
Since the quality-of-life variables are polytomous and have the same categories, we can arrange the quality-of-life data in a square table. Let V1 and V2 represent patient status at the beginning and at the end of the study respectively. Define R1 to take the value 1 if V1 is missing in a given cell in the table and 0 if V1 is not missing, and R2 takes the value 1 if V2 is missing and 0 if V2 is not missing (see Table 3). The responses R1 and R2 are generated by the imputer and are completely observed.

Table 3. Representation of missing data

V1	V2	R1	R2	Frequency
1	1	0	0	95
2	1	0	0	13
3	1	0	0	6
NA	1	1	0	4
1	2	0	0	9
2	2	0	0	13
3	2	0	0	7
NA	2	1	0	4
1	3	0	0	10
2	3	0	0	2
3	3	0	0	3
NA	3	1	0	1
1	NA	0	1	22
2	NA	0	1	12
3	NA	0	1	3
NA	NA	1	1	9

Since missing data can occur in the beginning and/or at the end of the study (i.e. V1 or V2), or at both time points (i.e. V1 and V2), we consider two plausible mechanisms (MNAR1 and MNAR2) for missingness in the path diagrams in Figure 1. Failing to restrict ourselves to plausible path models could result in overlooking the dependency structure of square-table data and yield potentially misleading results (Fay 1992).

Since the status V1 and V2 include missing data for some cases, the models that associate the responses R1 or R2 with the status in V1 or V2 are non-ignorable. The relationship as dependency paths is presented in Figure 1. Responses R1 and R2 may depend on V1 or V2 in different ways, see Little (1985). MNAR1 path model considers that missingness in R1 depends on the status in the first visit, V1, that is [V1V2, R1R2, R1V1] or equivalent [V1V2, R1R2, R1V1, R2V1]. In the MNAR2 path, the missingness in R2 depends on the status in the second visit, V2, that is [V1V2, R1R2, R2V2] or equivalent [V1V2, R1R2, R2V2, R1V2]. MNAR2 is implausible (since V2 cannot influence R1).



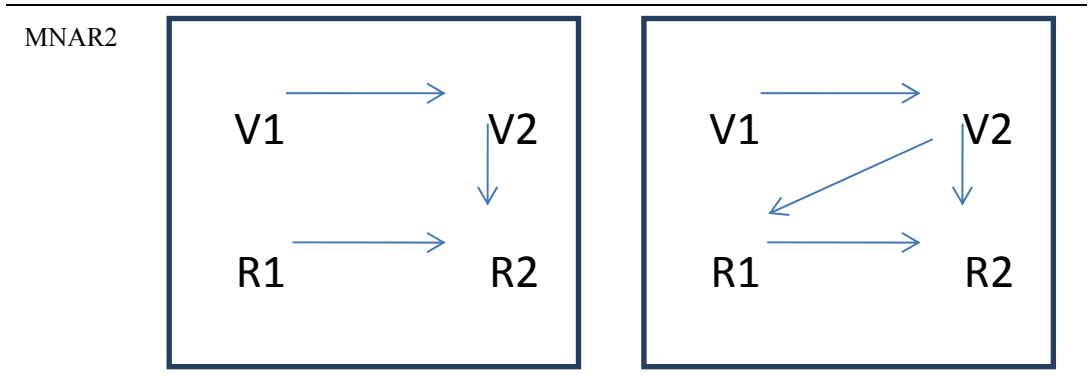


Figure 1. The different mechanisms for missing data:

top panel, MNAR1 before effect; lower panel, MNAR2 after effect.

For more paths of MNAR pattern, see Little (1985).

Following Schafer (2003) we investigate the null hypothesis of the missingness mechanism being MCAR versus the alternative hypothesis of MNAR. Using likelihood ratio tests, we can compare the MCAR model with MNAR1 (χ^2=3.00; 1df; P=0.08) and MNAR2 (χ^2=2.33; 1df; P=0.13). We reject MCAR in favor of MNAR1 but not MNAR2 (at the 10% level) and therefore proceed to the analysis with MNAR1.

3.1 Multiple imputations based on plausible mechanisms of missingness

To produce the imputations, some assumptions about the data and the mechanism producing missing data need to be made. The assumed data model should be plausible and should be somewhat related to the analyst's investigation. This model forms the basis to approximate the distribution in which the missing data are conditional upon observed data (i.e. predictive distribution of missing data), Recai (2011).

Using the MANR1 path model from the previous section we now generate multiply imputed data sets using data augmentation algorithms (as implemented in the S+ Missing Data library based on Schafer (1997) and Tanner and Wong (1987)). The data augmentation (DA) algorithm may be used to draw a sample of parameters from the posterior distribution from which further inferences are achieved.

Following Rubin (1996) we produced five multiply imputed data sets, using MNAR1 (with 1000 iterations), although increasing the number to ten did not make any material difference to our findings. Table 5 shows the cell probabilities (multiplied by 10 000) using a complete case analysis and using MNAR1.

Table 4. The 3x3 square-table parameter estimates of complete case and imputed data

Parameter	Complete cases sample size = 158	MNAR1 sample size = 213
π_{11*}	6011	4882
π_{21}	823	892
π_{31}	380	423
π_{12}	570	798
π_{22}	823	986
π_{32}	443	610
π_{13}	633	751
π_{23}	127	282
π_{33}	190	376
Total	10 000	10 000

* π_{11} denotes count estimates of the cell number (1, 1) in the 3x3 square table.

We now proceed to a consolidated analysis of imputed.

3.2 Consolidated analysis

Following Agresti (2012), we organise the quality-of-life data in a square table and analyze it using matched pair modeling. We start with the symmetry model and proceed to conditional symmetry and finally quasi-symmetry. The symmetry model implies, for example, that the number of patients who change their response from 1→2 is approximately equals to the number who change from 2→1, etc. The model has no interest in the number of patients who do not change their opinion between the two occasions, i.e., 1→1, 2→2 and 3→3. In the conditional-symmetry model there is an extra parameter, τ, for the off-diagonal elements to quantify the effect on the structure of agreement. The quasi-symmetry model permits marginal distributions to differ and allows for marginal heterogeneity. The symmetry model is a special case of the quasi-symmetry model where the marginal distributions for the row and column variables are the same or the score coefficients are equal to zero. In other words, symmetry has one parameter less than conditional symmetry and two parameters less than quasi symmetry. Looking at Table 5, the first column represents the symmetry with 6 coefficients, the second column represents conditional symmetry with 7 coefficients and finaly the third column Quasi with 8 coefficients.

For any patient, indexed by s, suppose that the log probabilities for the k response categories are:

$\lambda_{s1},...,\lambda_{sk}$ before treatment, and

$\lambda_{s1}+\tau_1,...,\lambda_{sk}+\tau_k$ after treatment

so that $(\tau_1,\ldots,\tau_k)$ is the treatment effect, assumed to be the same for all individual patients. To simplify the idea of modeling square tables, we introduce the main type of models we use in the analysis. The first is the symmetry model where we like to test the null hypothesis:

$$H_0 \quad : \qquad \pi_{ij} = \pi_{ji} .$$

In other words, we like to test if the cell probability on the one side of the main diagonal is a mirror image of that on the other side.

For expected frequency the logarithm value is:

$$\log \mu_{ij} = \lambda + \lambda_i + \lambda_j + \lambda_{ij} \tag{1}$$

Conditional symmetry models may be used when categories are ordered. This kind of model estimates an extra parameter, τ, for the off-diagonal elements on the structures of agreement. The main objective of this extra parameter is to quantify the effect on the structure of agreement. The symmetry model is a special case of the conditional-symmetry model when τ = 0. τ is equal to log (π_{ij} / π_{ji}). The following model represents a generalization that includes the condition when symmetry does not hold with ordered category:

$$\log \mu_{ij} = \lambda + \lambda_i + \lambda_j + \lambda_{ij} + \tau I(i < j), \quad I(\bullet) \text{ is the indicator function} \tag{2}$$

The quasi-symmetry model implies some association and allows the main effect terms to differ so that:

$$(\log\mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}) \neq (\log\mu_{ji} = \lambda + \lambda_j^X + \lambda_i^Y + \lambda_{ij}) \tag{3}$$

The main effect in (3) is different for rows (X) and columns (Y) and the resulting estimates are the differences in $\{\lambda_j^Y - \lambda_j^X\}$ for j = 1, 2,…, . Now, set $\{\lambda_j^Y - \lambda_j^X = \beta u_j\}$, thus the difference in the effect of a category from one occasion to the other follows a linear trend in the category score.

The quasi-symmetry model is:

$$\log\mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \beta u_j + \lambda_{ij} , \tag{4}$$

where u_j denotes order scores for the row and the column category. The greater the $|\beta|$, the greater the difference between the joint probabilities π_{ij} *and* π_{ji}, that is the difference between the marginal row and the column distributions.

The marginal homogeneity model compares marginal distribution in the square table, and one may test the hypothesis of marginal homogeneity by comparing the fit of the specific square table. It is known that a table which satisfies both quasi-symmetry and marginal homogeneity also satisfies symmetry, and one can test marginal homogeneity by comparing goodness-of-fit statistics for the symmetry and quasi-symmetry models, see Bishop et al. (1975).

4. Results

A symmetrical of before-after array of matched-pair data indicates that no association likely exists between before and after PSM patients' decisions with regard to treatment effects. When statistical evidence emerges of non-symmetry, two reasons exist for significant lack of symmetry: positive opinion change of PSM patients toward treatment effects and negative opinion change. These two sources of deviation for a symmetrical array are easily identified and indicate different dimensions of the association.

We present results in Table 5 for the matched-pair model of the square table with order categories for the MNAR1 mechanism as shown in Figure 1. These results are the product of ten consolidated estimates of the multiply imputed sets based on Rubin's rules. The results show that there is a significant difference between the PSM patients who changed their opinion from 1→2 against 2→1, but we are unable to reject the hypothesis of symmetry with regard to 1→3 against 3→1 and 2→3 against 3→2.

We follow Schafer (1997) procedures for combining hypothesis tests from multiple data sets. These procedures are based on Wald tests, likelihood ratio tests, or simple methods for combining chi-square statistics.

The consolidated F-statistics for our symmetric model is (2.218) and the p-value is (0.481), this shows an adequate model fit.

The estimated parameter of the conditional symmetry model and its uncertainty, $\tau = 0.066$ (SE 0.345), implies that in general the number of patients below the square table diagonal, who positively changed their opinion, is not significantly different from those above the diagonal, who negatively changed their opinion. The third column in Table 5 shows the results from the quasi-symmetry model; the score coefficients and their uncertainties are β_1=0.156 (SE 0.399), β_2=0.359 (SE 0.505) respectively, and the p-values of both models are (0.085 and 0.040 respectively). The conditional and quasi-symmetry models are rejected.

Table 5. Testing for symmetry, conditional symmetry and quasi-symmetry, using imputed data

Coefficients	Symmetry	Conditional symmetry	Quasi-symmetry
Intercept	1.445 (0.577)	1.744 (0.591)	1.386 (0.673)
Sym (3,2) ⇔ (2,3)	0.486 (0.653)	0.349 (0.747)	0.466 (0.681)
Sym (3,1) ⇔ (1,3)	0.895 (0.622)	0.831 (0.605)	1.020 (0.612)
Sym (2,2) ⇔ (3,3)	1.604 (0.644)	1.431 (0.676)	1.633 (0.734)
Sym (2,1) ⇔ (1,2)	1.372 (0.644)	0.885 (0.720)	1.197 (0.650)
Sym (1,1) ⇔ (3,3)	3.331 (0.585)	2.944 (0.619)	3.302 (0.677)
τ	-	0.066 (0.345)	
β_1 β_2			0.156 (0.399) 0.359 (0.505)
P-value	0.481	0.085	0.040

( ) indicate standard errors.

'(3,2) ⇔ (2,3)' denotes a comparison of the number of patients who change responses from 3 to 2 against 2 to 3 in a 3x3 square table.

5. Conclusion

In this paper we suggest a congenial strategy for integrating the performance of the imputer and the data analyst when considering data which can be organized in a square table and which have a clear dependency structure. We use before and after quality-of-life data from a randomized controlled trial to demonstrate how multiple imputation and matched-pair modeling can be properly combined, avoiding a possible mismatch between imputation and analyses, to derive a properly consolidated analysis of the quality-of-life data. We implement congenial conditions presented in Meng (1994), which are essential in integrating the performance of the imputer and the data analyst. With an illustrated example of the opinion study with different non-response mechanism, this paper shows how the performance of the imputer (testing for independence using hierarchical log-linear model and parameter estimate using MI) and the analyst (quantifying different effects using ML and consolidating the estimate under Rubin rules) to provide reliable results.

Acknowledgements

The authors are grateful to acknowledges Prof. D. Fitzmaurice and Dr. E. Murray (Primary care and General Practice, University of Birmingham, UK) for providing the data at the final stage of the study.

References

Agresti, A. (2012). *Categorical Data Analysis*. New York, NY: John Wiley and Sons.

Bishop, Y.M.M., Fienberg, S.E. & Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.

Fay, R.E. (1992). When are inferences from multiple imputation valid? *Proceedings of the Survey Research Methods Section*, *American Statistical Association*, 227-232.

Fitzmaurice, D., Murray, E., Hussain, S., McCahon, S., Holder, R., Raftery, J., … Hobbs, F. (2005, 15 October). SMART: Self management of oral anticoagulation: A randomised trial. *British Medical Journal*. http://dx.doi.org/10.1136/bmj.38618.580903.AE

Little, R.J. (1985). Nonresponse adjustments in longitudinal surveys: Models for categorical data. *Bulletin of the International Statistics Institute,* 15, 1-15.

Little, R.J. & Rubin, D.B. (2002). *Statistical Analysis with Missing Data* (2nd ed.). New York, NY: John Wiley and Sons.

Meng, X.L. (1994). Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statistical Science*, 10, 538-573.

Recai M. Yucel. (2011). State of the multiple imputation software. *Journal of Statistical Software*, 45(1), 1-7.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York, NY: John Wiley and Sons. http://dx.doi.org/10.1002/9780470316696

Rubin, D.B. (1996). Multiple imputations after 18+ years. *Journal of the American Statistical Association.* 91, 473-489. http://dx.doi.org/10.1080/01621459.1996.10476908

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall. http://dx.doi.org/10.1201/9781439821862

Schafer, J.L. (2003). Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica Neerlandica*, 57, nr. 1, 19-35. http://dx.doi.org/10.1111/1467-9574.00218

Tanner, M. A. & Wing Hung Wong. (1987). The calculation of posterior distribution by data augmentation. *Journal of the American Statistical Association*, 82, No.398, 528-540. http://dx.doi.org/10.1080/01621459.1987.10478458

Notes

Note 1. Note that this happens under the validity of the imposed assumption, otherwise imposing an invalid assumption can lead to biased results, rendering invalid any subsequent inferences and predictions.