Every missingness not at random model has a missingness at random counterpart with equal fit

Peer-reviewed author version

MOLENBERGHS, Geert; BEUNCKENS, Caroline; SOTTO, Cristina & Kenward, Michael G. (2008) Every missingness not at random model has a missingness at random counterpart with equal fit. In: JOURNAL OF THE ROYAL STATISTICAL SOCIETY SERIES B-STATISTICAL METHODOLOGY, 70. p. 371-388.

DOI: 10.1111/j.1467-9868.2007.00640.x
Handle: http://hdl.handle.net/1942/7996

Every Missing Not at Random Model Has Got a Missing at Random Counterpart with Equal Fit

Geert Molenberghs, Caroline Beunckens, Cristina Sotto

Center for Statistics, Hasselt University, B-3590 Diepenbeek, Belgium

geert.molenberghs@uhasselt.be caroline.beunckens@uhasselt.be cristina.sotto@uhasselt.be

and Michael G. Kenward

Medical Statistics Unit, London School of Hygiene and Tropical Medicine, London WC1E7HT, UK

Mike.Kenward@lshtm.ac.uk

Abstract

Over the last decade a variety of models to analyze incomplete multivariate and longitudinal data have been proposed, many of which allowing for the missingness to be not at random (MNAR), in the sense that the unobserved measurements influence the process governing missingness, in addition to influences coming from observed measurements and/or covariates. The fundamental problems implied by such models, to which we refer as sensitivity to unverifiable modelling assumptions, has, in turn, sparked off various strands of research in what is now termed *sensitivity analysis*. The nature of sensitivity originates from the fact that an MNAR model is not fully verifiable from the data, rendering the empirical distinction between MNAR and random missingness (MAR), where only covariates and observed outcomes influence missingness, hard or even impossible, unless one is prepared to accept the posited MNAR model in an unquestioning way. In this paper, we show that the empirical distinction between MAR and MNAR is not possible, in the sense that each MNAR model fit to a set of observed data can be reproduced exactly by an MAR counterpart. Of course, such a pair of models will produce different predictions of the unobserved outcomes, given the observed ones. Theoretical considerations are supplemented with an illustration based on the Slovenian Public Opinion survey, analyzed before in the context of sensitivity analysis.

Some Key Words: Contingency table; Ignorability; Missing completely at random; Pattern-mixture model; Selection model; Shared parameter model.

1 Introduction

Incomplete sets of data are common throughout all branches of empirical research. Incomplete data have always posed problems of imbalance in the data matrix, but more importantly incompleteness often destroys a trial's randomization justification or a survey's representativeness.

The extent to which this happens depends on the nature of the missing data mechanism. Rubin (1976) distinguished between *missing complete at random* (MCAR), where the outcomes are independent of the mechanism governing missingness, *missing at random* (MAR), where there is dependence between both, but only in the sense that missingness may depend on the observed, but not further on the unobserved measurements. Finally, when a *missing not at random* (MNAR) mechanism operates, missingness depends on the unobserved outcomes, perhaps in addition to the observed ones.

Traditionally, such simple methods as a complete case analysis or simple forms of imputation (e.g., last observation carried forward) have been in use. While they have the advantage of restoring balance and/or a rectangular data matrix, it is sufficiently documented that such analyses are prone to severe bias and/or losses of efficiency (Molenberghs *et al*, 2004; Jansen *et al*, 2006) and should be avoided. Since a likelihood-based or Bayesian analysis is valid when the missing data mechanism is MAR, as long as all observed data are included into the analysis, the so-called *ignorability* property, so-called direct likelihood analyses, their Bayesian counterparts, or multiple imputation (Rubin, 1987), are regarded by many as candidates for the primary analyses of a study. When semi-parametric inferences are desired, the methods proposed by Robins *et al* (1995) can be applied.

However, one can never exclude the possibility that MNAR models may be operating. Even though a variety of statistical models have been proposed for the MNAR situation (Diggle, and Kenward, 1994; Baker, 1995; Molenberghs *et al*, 1997; Troxel *et al*, 1998), and in spite of the dramatically increased computational power, such models are prone to considerable sensitivity. This was made clear by a variety of discussants to Diggle, and Kenward (1994), such as Laird (1994), Little (1994b), and Rubin (1994). Several authors have laid bare such sensitivities and proposed methods for informal and formal sensitivity analysis (Kenward, 1998; Robins *et al*, 1998; Molenberghs *et al*, 2001; Van Steen *et al*, 2001; Verbeke *et al*, 2001; Thijs *et al*, 2002; Jansen *et al*, 2003). Overviews are provided in Verbeke, and Molenberghs (2000) and Molenberghs, and Verbeke (2005).

One view is that testing the MAR null hypothesis against an MNAR alternative is of a conventional nature. While indeed Diggle, and Kenward (1994) have conducted such tests, it is very important to realize that they are conditional upon the alternative model holding.

The contribution of this paper is to show that, strictly speaking, the correctness of the alternative model can only be verified in as far as it fits the *observed* data. Thus, evidence for or against MNAR can only be provided within a particular, predefined parametric family, the plausibility of which cannot be verified in empirical terms alone. We show that an overall (omnibus) assessment of MAR *versus* MNAR is not possible, since every MNAR model can be doubled up with a uniquely defined MAR counterpart, producing exactly the same fit as the original MNAR model, in the sense that it produces exactly the same predictions to the observed data (e.g., fitted counts in an incomplete contingency table) as the original MNAR model, and depending on exactly the same parameter vector. We show that, while this so-called MAR counterpart generally does not belong to a conventional parametric family, its existence has important ramifications. While this broad issue is still open to debate and even confusion, it has been pointed out in the literature. For example, the issue has been referred to, in general terms, by Little, and Rubin (2002) and, in a non- and semi-parametric context, by Gill, van der Laan, and Robins (1997). An excellent

exposition, together with related references, can be found in Schafer and Graham (2002). Here, we focus on a general construction method for this counterpart, which we make explicit for the case of categorical data.

The rest of the paper is organized as follows. In Section 2 we outline the necessary concepts, terminology, and notation. Section 3 provides the general result and sketches the proof. In Section 4 the specific case of incomplete contingency tables is studied. In Section 5 we apply the ideas developed to data from the Slovenian Public Opinion Survey, analyzed before by Rubin *et al* (1995) and Molenberghs *et al* (2001).

2 Notation and Concepts

Let the random variable Y_{ij} denote the response of interest, for the ith study subject, designed to be measured at occasions t_{ij}, $i = 1, \ldots, N$, $j = 1, \ldots, n_i$. Independence across subjects is assumed. This setting covers both the longitudinal as well as the multivariate settings. In the latter case, $t_{ij} = t_j$ would merely be indicators for the various variables studied, and typically $n_i \equiv n$. The outcomes can conveniently be grouped into a vector $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$. In addition, define a vector of missingness indicators $\boldsymbol{R}_i = (R_{i1}, \ldots, R_{in_i})'$ with $R_{ij} = 1$ if Y_{ij} is observed and 0 otherwise. In the specific case of dropout, $\boldsymbol{R}_i$ can usefully be replaced by the dropout indicator

$$D_i = \sum_{j=1}^{n_i} R_{ij}.$$

Note that the concept of dropout refers to time-ordered variables, such as in longitudinal studies. For a complete sequence, $\boldsymbol{R}_i = \boldsymbol{1}$ and/or $D_i = n_i$. It is customary to split the vector $\boldsymbol{Y}_i$ into observed ($\boldsymbol{Y}_i^o$) and missing ($\boldsymbol{Y}_i^m$) components, respectively.

In principle, one would like to consider the density of the full data $f(\boldsymbol{y}_i, \boldsymbol{r}_i | \boldsymbol{\theta}, \boldsymbol{\psi})$, where the parameter vectors $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ describe the measurement and missingness processes, respectively. Covariates are assumed to be measured and grouped in a vector $\boldsymbol{x}_i$.

This full density function can be factorized in different ways, each leading to a different framework. The *selection model* (SeM) framework is based on the following factorization (Rubin, 1976; Little, and Rubin, 2002):

$$f(\boldsymbol{y}_i, \boldsymbol{r}_i | \boldsymbol{x}_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y}_i | \boldsymbol{x}_i, \boldsymbol{\theta}) f(\boldsymbol{r}_i | \boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{\psi}). \tag{1}$$

The first factor is the marginal density of the measurement process and the second one is the density of the missingness process, conditional on the outcomes. As an alternative, one can consider so-called *pattern-mixture models* (PMM; Little (1993, 1994a)) using the reversed factorization

$$f(\boldsymbol{y}_i, \boldsymbol{r}_i | \boldsymbol{x}_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y}_i | \boldsymbol{x}_i, \boldsymbol{r}_i, \boldsymbol{\theta}) f(\boldsymbol{x}_i, \boldsymbol{r}_i | \boldsymbol{\psi}).$$

This can be seen as a mixture density over different populations, each of which defined by the observed pattern of missingness.

Instead of using the selection modelling or pattern-mixture modelling frameworks, the measurement and the dropout process can be jointly modelled using a *shared-parameter model* (Wu, and Carroll, 1988; Wu, and Bailey, 1988, 1989; TenHave *et al*, 1998; Follmann, and Wu, 1995; Little, 1995). One then assumes there exists a vector of random effects $\boldsymbol{b}_i$, conditional upon which the measurement and dropout processes are independent. This *shared-parameter model* (SPM) is formulated by way of the following factorization

$$f(\boldsymbol{y}_i, \boldsymbol{r}_i|\boldsymbol{x}_i, \boldsymbol{b}_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{b}_i, \boldsymbol{\theta}) f(\boldsymbol{r}_i|\boldsymbol{x}_i, \boldsymbol{b}_i, \boldsymbol{\psi}). \tag{2}$$

Here, $\boldsymbol{b}_i$ are shared parameters, often considered to be random effects and following a specific parametric distribution.

The taxonomy of missing data mechanisms, introduced by Rubin (1976) and informally described in the introduction, can easily be formalized using the second factor on the right hand side of selection-model factorization (1). A mechanism is MCAR if

$$f(\boldsymbol{r}_i|\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{\psi}) = f(\boldsymbol{r}_i|\boldsymbol{x}_i, \boldsymbol{\psi}), \tag{3}$$

i.e., when the measurement and missingness processes are independent, perhaps conditional on covariates. For a given set of data, MAR holds when

$$f(\boldsymbol{r}_i|\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{\psi}) = f(\boldsymbol{r}_i|\boldsymbol{x}_i, \boldsymbol{y}_i^o, \boldsymbol{\psi}), \tag{4}$$

strictly weaker than the MCAR condition, but still a simplification of the MNAR case, where missingness depends on the unobserved outcomes $\boldsymbol{y}_i^m$, regardless of the observed outcomes and the covariates.

Note that MCAR is equally trivial in the pattern-mixture model framework, where $\boldsymbol{r}_i$ does not influence the mixture components, and in the shared-parameter model framework, where no random-effects are shared between the two factors in (2).

A final useful concept we need is *ignorability*. Note that the contribution to the likelihood of subject i, based on (1), equals

$$L_i = \int f(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta}) f(\boldsymbol{r}_i|\boldsymbol{x}_i, \boldsymbol{y}_i^o, \boldsymbol{y}_i^m, \boldsymbol{\psi})\, d\boldsymbol{y}_i^m. \tag{5}$$

In general, (5) does not simplify, but under MAR, we obtain:

$$L_i = f(\boldsymbol{y}_i^o|\boldsymbol{x}_i, \boldsymbol{\theta}) f(\boldsymbol{r}_i|\boldsymbol{x}_i, \boldsymbol{y}_i^o, \boldsymbol{\psi}). \tag{6}$$

Hence, likelihood and Bayesian inferences for the measurement model parameters $\boldsymbol{\theta}$ can be made without explicitly formulating the missing data mechanism, provided the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are distinct, meaning that their joint parameter space is the Cartesian product of the two component parameter spaces (Rubin, 1976). For Bayesian inferences, further the priors need to be independent (Little, and Rubin, 2002).

It is precisely this result which makes so-called direct likelihood analyses, valid under MAR, viable candidates for the status of primary analysis in clinical trials and a variety of other settings (Molenberghs *et al*, 2004). Since we will be concerned with expressions for MAR counterparts to MNAR models, we will explicitly describe the missing data mechanism. This implies we typically will not invoke the ignorability property.

3 General Result

In this section, we will show that for every MNAR model fitted to a set of data, there is an MAR counterpart providing exactly the same fit to the data. Here, the concept of model fit should be understood as measured using such conventional methods as deviance measures and, of course, in as far as the observed data are concerned. The following steps are involved: (1) fitting an MNAR model to the data; (2) reformulating the fitted model in PMM form; (3) replacing the density or distribution of the unobserved measurements given the observed ones and given a particular response pattern by its MAR counterpart; (4) establishing that such an MAR counterpart uniquely exists. Throughout this section, we will suppress covariates $\boldsymbol{x}_i$ from notation, but assume them to be present.

In the first step, we fit an MNAR model to the observed set of data. The observed data likelihood is:

$$L = \prod_i \int f(\boldsymbol{y_i}^o, \boldsymbol{y_i}^m, \boldsymbol{r}_i | \boldsymbol{\theta}, \boldsymbol{\psi}) d\boldsymbol{y_i}^m. \tag{7}$$

Upon denoting the obtained parameter estimates, e.g., obtained by likelihood-based or Bayesian methods, by $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\psi}}$ respectively, the fit to the hypothetical full data is

$$f(\boldsymbol{y_i}^o, \boldsymbol{y_i}^m, \boldsymbol{r}_i | \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}) = f(\boldsymbol{y_i}^o, \boldsymbol{y_i}^m | \widehat{\boldsymbol{\theta}}) f(\boldsymbol{r}_i | \boldsymbol{y_i}^o, \boldsymbol{y_i}^m, \widehat{\boldsymbol{\psi}}). \tag{8}$$

To undertake the second step, full density (8) can be re-expressed in PMM form as:

$$f(\boldsymbol{y_i}^o, \boldsymbol{y_i}^m | \boldsymbol{r}_i, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}) f(\boldsymbol{r}_i | \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}) = f(\boldsymbol{y_i}^o | \boldsymbol{r}_i, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}) f(\boldsymbol{r}_i | \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}) f(\boldsymbol{y_i}^m | \boldsymbol{y_i}^o, \boldsymbol{r}_i, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}). \tag{9}$$

A similar reformulation can be considered for an SPM. In a PMM, the model will have been expressed in this form to begin with.

Note that, in line with PMM theory, the final term on the right hand side of (9), $f(\boldsymbol{y_i}^m | \boldsymbol{y_i}^o, d_i, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}})$, is not identified from the observed data. In this case, it is determined solely from modelling assumptions. Within the PMM framework, identifying restrictions have to be considered (Little, 1994a; Molenberghs *et al*, 1998; Kenward *et al*, 2003).

The third step requires replacing this factor by the appropriate MAR counterpart. To this end, we need the following lemma, formulating MAR equivalently within the PMM framework.

Lemma 1 *In the PMM framework, the missing data mechanism is MAR if and only if*

$$f(\boldsymbol{y}_i^m | \boldsymbol{y}_i^o, \boldsymbol{r}_i, \boldsymbol{\theta}) = f(\boldsymbol{y}_i^m | \boldsymbol{y}_i^o, \boldsymbol{\theta}).$$

This means that, in a given pattern, the conditional distribution of the unobserved components given the observed ones equals the corresponding distribution marginalized over the patterns. The proof, which is rather straightforward and similar to what can be found in Molenberghs *et al* (1998), is deferred to the appendix. Note that, owing to this result, MAR can be formulated in terms of R given Y, but also in terms of Y given R.

Using Lemma 1, it is clear that $f(\boldsymbol{y_i}^m|\boldsymbol{y_i}^o, \boldsymbol{r}_i, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}})$ needs to be replaced with

$$h(\boldsymbol{y_i}^m|\boldsymbol{y_i}^o, \boldsymbol{r}_i) = h(\boldsymbol{y_i}^m|\boldsymbol{y_i}^o) = f(\boldsymbol{y_i}^m|\boldsymbol{y_i}^o, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}), \tag{10}$$

where the $h(\cdot)$ notation is used for shorthand purposes. Note that the density in (10) follows from the SeM-type marginal density of the complete data vector. Sometimes, therefore, it may be more convenient to replace the notation $\boldsymbol{y_i}^o$ and $\boldsymbol{y_i}^m$ by one that explicitly indicates which components are observed and missing in pattern $\boldsymbol{r}_i$ under consideration:

$$h(\boldsymbol{y_i}^m|\boldsymbol{y_i}^o, \boldsymbol{r}_i) = h(\boldsymbol{y_i}^m|\boldsymbol{y_i}^o) = f[(y_{ij})_{r_j=0}|(y_{ij})_{r_j=1}, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}]. \tag{11}$$

Thus, (11) provides a unique way of extending the model fit to the observed data, belonging to the MAR family. As stated before, the above construction does not lead to a member of a conventional parametric family. While this obviously implies limitations on its use, such is not dissimilar to the construction of some semi- and non-parametric estimators. Also, it helps to understand that an overall, definitive conclusion about the nature of the missing data mechanism is not possible, even though one can make progress if attention is confined to a given parametric family, in which one puts sufficiently strong prior belief. To show formally that the fit remains the same, we consider the observed-data likelihood based on (7) and (9):

$$\begin{aligned}
\widehat{L} &= \prod_i \int f(\boldsymbol{y_i}^o, \boldsymbol{y_i}^m|\widehat{\boldsymbol{\theta}}) f(\boldsymbol{r}_i|\boldsymbol{y_i}^o, \boldsymbol{y_i}^m, \widehat{\boldsymbol{\psi}}) d\boldsymbol{y_i}^m \\
&= \prod_i \int f(\boldsymbol{y_i}^o|\boldsymbol{r}_i, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}) f(\boldsymbol{r}_i|\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}) f(\boldsymbol{y_i}^m|\boldsymbol{y_i}^o, \boldsymbol{r}_i, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}) d\boldsymbol{y_i}^m \\
&= \prod_i f(\boldsymbol{y_i}^o|\boldsymbol{r}_i, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}) f(\boldsymbol{r}_i|\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}) \\
&= \prod_i \int f(\boldsymbol{y_i}^o|\boldsymbol{r}_i, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}) f(\boldsymbol{r}_i|\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}) h(\boldsymbol{y_i}^m|\boldsymbol{y_i}^o) d\boldsymbol{y_i}^m.
\end{aligned}$$

The above results justify the following theorem:

Theorem 1 *Every fit to the observed data, obtained from fitting an MNAR model to a set of incomplete data, is exactly reproducible from an MAR decomposition.*

The key computational consequence is the need to compute $h(\boldsymbol{y_i}^m|\boldsymbol{y_i}^o)$ in (10) or (11). This means, for each pattern, the conditional density of the unobserved measurements given the observed ones needs to be extracted from the marginal distribution of the complete set of measurements. Molenberghs *et al* (1998) have shown that, for the case of dropout, the so-called *available case missing value restrictions* (ACMV) provide a practical computational scheme. Precisely, ACMV states that

$$\forall t \geq 2, \forall s < t : f(y_{it}|y_{i1}, \cdots, y_{i,t-1}, d_i = s) = f(y_{it}|y_{i1}, \cdots, y_{i,t-1}, d_i \geq t). \tag{12}$$

In other words, the density of a missing measurement, conditional on the measurement history, is determined from the corresponding density over all patterns for which all of these measurements are observed. For example, the density of the third measurement in a sequence, given the first and second ones, in patterns with only 1 or 2 measurements taken, is determined from the

corresponding density over all patterns with 3 or more measurements. Thijs *et al* (2002) and Verbeke, and Molenberghs (2000)(p. 347) derived a practical computational method for the factors in (12):

$$f(y_{it}|y_{i1},\cdots,y_{i,t-1},d_i=s) = \frac{\sum_{d=s}^{n}\alpha_d f_d(y_{i1},\ldots,y_{is})}{\sum_{d=s}^{n}\alpha_d f_d(y_{i1},\ldots,y_{i,s-1})} \quad (13)$$

$$= \sum_{d=s}^{n}\left(\frac{\alpha_d f_d(y_{i1},\ldots,y_{i,s-1})}{\sum_{d=s}^{n_i}\alpha_d f_d(y_{i1},\ldots,y_{i,s-1})}\right) f_d(y_s|y_{i1},\ldots,y_{i,s-1}). \quad (14)$$

Here, α_d is the probability to belong to pattern d.

The above identifications for the monotone case are useful in case an MNAR pattern-mixture model has been fitted to begin with, since then the identifications under MAR can be calculated from the pattern-specific marginal distributions. When a selection model has been fitted in the initial step, $f(y_{i1},\ldots,y_{in_i}|\widehat{\boldsymbol{\theta}})$ has been estimated, from which all conditional distributions, needed in (11), can be derived. When the initial model is an MNAR PMM model and the missing data patterns are non-monotone, then it is necessary to first rewrite the PMM in SeM form, and derive the required conditional distributions from the so-obtained SeM measurement model. This essentially comes down to calculating a weighted average of the pattern-specific measurement models. In some cases, such as for contingency tables, this step can be done in an alternative way by fitting a saturated MAR selection model to the fit obtained from the PMM model.

We will illustrate and contrast the monotone and non-monotone cases using a bivariate and trivariate outcome with dropout on the one hand and a bivariate non-monotone outcome on the other hand. While the theorem applies to both the monotone and non-monotone settings, it is insightful to see that only for the former relatively simple and intuitively appealing expressions arise, while the latter setting involves the need for iterative computation. In the next section, the aforementioned general contingency table setting to which a PMM has been fitted, will be studied.

3.1 A Bivariate Outcome With Dropout

Here and in the following examples, we will present and equate the SeM and PMM decompositions, enabling us to derive expressions for the MAR counterparts. It is interesting and straightforward to derive results for the MCAR case, and hence these will be presented, too.

Dropping covariates, parameters, and the subject index i from notation, the SeM-PMM equivalence for the case of two outcomes, the first of which is always observed but the second one partially missing, is given by:

$$\begin{aligned} f(y_1,y_2)\widetilde{g}(d=2|y_1,y_2) &= f_2(y_1,y_2)\widetilde{\alpha}(d=2), \\ f(y_1,y_2)\widetilde{g}(d=1|y_1,y_2) &= f_1(y_1,y_2)\widetilde{\alpha}(d=1). \end{aligned}$$

Note that this is the setting considering by Glynn *et al* (1986). Here, $\widetilde{g}(\cdot)$ is used for the SeM dropout model, with $\widetilde{\alpha}(\cdot)$ denoting the PMM probabilities to belong to one of the patterns.

Since $\widetilde{\alpha}(d=1)+\widetilde{\alpha}(d=2)=1$ and a similar result holds for the $\widetilde{g}(\cdot)$ functions, it is convenient to write:

$$\begin{aligned} f(y_1,y_2)g(y_1,y_2) &= f_2(y_1,y_2)\alpha \quad (15)\\ f(y_1,y_2)[1-g(y_1,y_2)] &= f_1(y_1,y_2)[1-\alpha]. \quad (16)\end{aligned}$$

Assuming MCAR, it is clear that $\alpha = g(y_1,y_2)$, producing, without any difficulty:

$$f(y_1,y_2) = f_2(y_1,y_2) = f_1(y_1,y_2). \quad (17)$$

Under MAR, y_2 has to be removed from $g(\cdot)$ for incomplete observations, but since we assume a single parametric function for the missingness model, it follows that $g(y_1,y_2)=g(y_1)$ and hence (15) produces

$$f(y_1)f(y_2|y_1)g(y_1) = f_2(y_1)f_2(y_2|y_1)\alpha.$$

Upon reordering, we find:

$$\frac{f(y_1)g(y_1)}{f_2(y_1)\alpha} = \frac{f_2(y_2|y_1)}{f(y_2|y_1)}. \quad (18)$$

The same arguments can be applied to (16), from which we derive:

$$f(y_2|y_1) = f_2(y_2|y_1) = f_1(y_2|y_1). \quad (19)$$

Note that (19) is strictly weaker than (17). The last term in (19) is not identified by itself, and hence, we see it needs to be set equal to its counterpart from the completers which, in turn, is equal to the marginal distribution. This is in agreement with (11) as well as with the specific identifications applicable in the monotone and hence ACMV setting.

3.2 A Trivariate Outcome With Dropout

Note that identification (19) does not involve mixtures. This changes as soon as there are three or more outcomes. The equations corresponding to (15)–(16), specialized to the MAR case, are:

$$\begin{aligned} f(y_1,y_2,y_3)g_0 &= f_0(y_1,y_2,y_3)\alpha_0, \quad (20)\\ f(y_1,y_2,y_3)g_1(y_1) &= f_1(y_1,y_2,y_3)\alpha_1, \quad (21)\\ f(y_1,y_2,y_3)g_2(y_1,y_2) &= f_2(y_1,y_2,y_3)\alpha_2, \quad (22)\\ f(y_1,y_2,y_3)g_3(y_1,y_2) &= f_3(y_1,y_2,y_3)\alpha_3. \quad (23)\end{aligned}$$

We have chosen to include pattern 0, the one without follow-up measurements, as well, and will return to this one. We could write $g_3(\cdot)$ as a function of y_3 as well, but because the sum of the $g_d(\cdot)$ equals one, it is clear that $g_3(\cdot)$ ought to be independent of y_3. With arguments similar to the ones developed in the case of two measurements, we can rewrite (23) as:

$$\frac{f(y_1,y_2)}{f_3(y_1,y_2)}\cdot\frac{g_3(y_1,y_2)}{\alpha_3} = \frac{f_3(y_3|y_1,y_2)}{f(y_3|y_1,y_2)}.$$

Exactly the same consideration can be made based on (22), and hence

$$f_3(y_3|y_1,y_2) = f(y_3|y_1,y_2) = f_2(y_3|y_1,y_2). \quad (24)$$

The first factor identifies the second one, and hence also the third one. Starting from (21), we obtain:

$$f_1(y_2, y_3|y_1) = f(y_2, y_3|y_1),$$

which produces, in fact, two separate identities:

$$\begin{aligned} f_1(y_2|y_1) &= f(y_2|y_1), & (25)\\ f_1(y_3|y_1, y_2) &= f(y_3|y_1, y_2) = f_3(y_3|y_1, y_2) = f_2(y_3|y_1, y_2). & (26) \end{aligned}$$

For the latter one, identity (24) has been used as well. The density $f(y_2|y_1)$, needed in (25), is determined from the general ACMV result (14):

$$f(y_2|y_1) = \frac{\alpha_2 f_2(y_2|y_1) + \alpha_3 f_3(y_2|y_1)}{\alpha_2 + \alpha_3}.$$

Finally, turning attention to (20), it is clear that $g_0 = \alpha_0$ and hence also $f_0(y_1, y_2, y_3) = f(y_1, y_2, y_3)$. From the latter density, only $f(y_1)$ has not been determined yet, but this one follows again very easily from the general ACMV result:

$$f(y_1) = \frac{\alpha_1 f_1(y_1) + \alpha_2 f_2(y_1) + \alpha_3 f_3(y_1)}{\alpha_1 + \alpha_2 + \alpha_3}.$$

In summary, the necessary MAR identifications easily follow from both the PMM and the SeM formulations of the model.

3.3 A Bivariate Outcome With Non-Monotone Missingness

The counterparts to (15)–(16) and (20)–(23) for a bivariate outcome with non-monotone missingness are

$$\begin{aligned} f(y_1, y_2)g_{00}(y_1, y_2) &= f_{00}(y_1, y_2)\alpha_{00}, & (27)\\ f(y_1, y_2)g_{10}(y_1, y_2) &= f_{10}(y_1, y_2)\alpha_{10}, & (28)\\ f(y_1, y_2)g_{01}(y_1, y_2) &= f_{01}(y_1, y_2)\alpha_{01}, & (29)\\ f(y_1, y_2)g_{11}(y_1, y_2) &= f_{11}(y_1, y_2)\alpha_{11}. & (30) \end{aligned}$$

Clearly, under MCAR, the $g_{r_1r_2}(\cdot)$ functions do not depend on the outcomes and hence $f_{r_1r_2}(y_1, y_2) = f(y_1, y_2)$ for all four patterns. For the MAR case, (27)–(30) simplify to

$$\begin{aligned} f(y_1, y_2)g_{00} &= f_{00}(y_1, y_2)\alpha_{00}, & (31)\\ f(y_1, y_2)g_{10}(y_1) &= f_{10}(y_1, y_2)\alpha_{10}, & (32)\\ f(y_1, y_2)g_{01}(y_2) &= f_{01}(y_1, y_2)\alpha_{01}, & (33)\\ f(y_1, y_2)g_{11}(y_1, y_2) &= f_{11}(y_1, y_2)\alpha_{11}. & (34) \end{aligned}$$

Observe there are four identifications across the $g_{r_1r_2}(y_1, y_2)$ functions:

$$g_{00} + g_{10}(y_1) + g_{01}(y_2) + g_{11}(y_1, y_2) = 1,$$

for each (y_1, y_2). Also $\sum_{r_1,r_2} \alpha_{r_1,r_2} = 1$. Applying the usual algebra to (31)–(34), we obtain three identifications for the unobservable densities:

$$f_{00}(y_1, y_2) = f(y_1, y_2), \tag{35}$$
$$f_{10}(y_1|y_2) = f(y_1|y_2), \tag{36}$$
$$f_{01}(y_2|y_1) = f(y_2|y_1). \tag{37}$$

Using these in conjunction with the identifiable parts of the distributions yields the MAR counterpart.

4 The General Case of Incomplete Contingency Tables

In Sections 3.1–3.3 we have derived general identification schemes for an MAR extension of a fitted model to a binary or trivariate outcome with dropout, as well as to a bivariate outcome with non-monotone missingness. Whereas the monotone cases provide explicit expressions in terms of the pattern-specific densities, (35)–(37) provide an identification only in terms of the marginal probability. This in itself is not a problem, since the marginal density is always available, either directly when a SeM is fitted, or through marginalization when a PMM or an SPM is fitted.

In the specific case of contingency tables, further progress can be made. Indeed, we can show a saturated MAR model is always available, for any incomplete contingency table setting. This implies one can start from the fit of an MNAR model to the observed data, and then extend it, using this result, towards MAR. We will present the general result and then discuss its precise implications for practice.

Assume we have a $\prod_{k=1}^{n} c_k$ contingency table with supplemental margins, where k indexes the n dimensions in the table and c_k is the number of alternatives the kth categorical variable can take. The table of completers is indexed by $\boldsymbol{r} = \mathbf{1} = (1, \ldots, 1)$. A particular incomplete table is indexed by a $\boldsymbol{r} \neq \mathbf{1}$. The full set of tables can but does not have to be present. The number of cells is:

$$\#\text{cells} = \sum_{\boldsymbol{r}} \prod_{k=1}^{n} c_k^{r_k}. \tag{38}$$

Denote the measurement model probabilities by $p_{\boldsymbol{j}} = p_{j_1 \ldots j_n}$ for $j_k = 1, \ldots c_k$ and $k = 1, \ldots, n$. Clearly, these probabilities sum to one. The missingness probabilities, assuming MAR, are:

$$p(\boldsymbol{r}|\boldsymbol{j}) = \begin{cases} p(\boldsymbol{r}|j_k \text{ with } r_k = 1) & \text{if } \boldsymbol{r} \neq \mathbf{1}, \\ 1 - \sum_{\boldsymbol{r} \neq \mathbf{1}} p(\boldsymbol{r}|\boldsymbol{j}) & \text{if } \boldsymbol{r} = \mathbf{1}. \end{cases} \tag{39}$$

Summing over $\boldsymbol{r}$ implies summing over those patterns for which actual observations are available. The number of parameters in the saturated model is

$$\#\text{parameters} = \left(\prod_{k=1}^{n} c_k - 1 \right) + \sum_{\boldsymbol{r} \neq \mathbf{1}} \prod_{k=1}^{n} c_k^{r_k}. \tag{40}$$

The first term in (40) is for the measurement model, the second one is for the missingness model. Clearly, the number of parameters equals one less than the number of cells, establishing the claim. The situation where covariates are present is covered automatically, merely by considering one extra dimension in the contigency table, $j = 0$ say, with c_0 referring to the total number of covariate levels in the set of data.

We will now study the implications for the simple but important settings studied in Sections 3.1 and 3.3.

4.1 A Bivariate Contingency Table With Dropout

In Section 3.1 identifications have been derived for the bivariate case with monotone missingness. For contingency tables, these can be derived as well by further fitting the saturated MAR model, described in the previous section, to the fit obtained from the original MNAR model. Denote the counts obtained from the fit of the original model by $z_{2,jk}$ and $z_{1,j}$, for the completers and dropouts, respectively. Denote the measurement model probabilities by p_{jk} and the dropout probabilities by q_j. Then, due to ignorability, the likelihood factors into two components:

$$\ell_1 = \sum_{j,k} z_{2,jk} \ln p_{jk} + \sum_j z_{1,j} \ln p_{j+} - \lambda \left(\sum_{j,k} p_{jk} - 1 \right), \tag{41}$$

$$\ell_2 = \sum_{j,k} z_{2,jk} \ln q_j + \sum_j z_{1,j} \ln(1 - q_j). \tag{42}$$

We have used an undetermined Lagrange multiplier λ to incorporate the sum constraint on the marginal probabilities. Solving the score equations for (41) and (42) produces, with simple and well-known algebra:

$$\widehat{p_{jk}} = \frac{1}{n} z_{2,jk} \left(\frac{z_{2,j+} + z_{1,j}}{z_{2,j+}} \right), \tag{43}$$

$$\widehat{q_j} = \frac{z_{2,j+}}{z_{2,j+} + z_{1,j}}, \tag{44}$$

where n is the total sample size. Combining parameter estimates leads to the new, MAR-based, fitted counts:

$$\widehat{z_{2,jk}} = n\widehat{p_{jk}}\widehat{q_j} = z_{2,jk}, \tag{45}$$

$$\widehat{z_{1,jk}} = n\widehat{p_{jk}}(1 - \widehat{q_j}) = z_{1,j} \frac{z_{2,jk}}{z_{2,j+}}, \tag{46}$$

$$\widehat{z_{1,j+}} = z_{1,j+}. \tag{47}$$

From (45) and (47) it is clear that the fit in terms of the observed data has not changed. The expansion of the incomplete data into a complete one is described by (46). Equations (45) and (46) can be used to produce the MAR counterpart to the original model, without any additional calculations. This is not so simple for the non-monotone case, as we will show next.

4.2 A Bivariate Contingency Table With Non-Monotone Missingness

The counterparts to (41)–(42) for this case are:

$$\ell_1 = \sum_{j,k} z_{11,jk} \ln p_{jk} + \sum_j z_{10,j} \ln p_{j+} + \sum_k z_{01,k} \ln p_{+k} + z_{00} \ln p_{++} - \lambda \left(\sum_{j,k} p_{jk} - 1 \right), \quad (48)$$

$$\ell_2 = \sum_{j,k} z_{11,jk} \ln(1 - q_{10,j} - q_{01,k} - q_{00}) + \sum_j z_{10,j} \ln q_{10,j} + \sum_k z_{01,k} \ln q_{01,k} + z_{00} \ln g_{00}. \quad (49)$$

Notation has been modified in accordance with the design. The q quantities correspond to the $g(\cdot)$ model in Section 3.3.

While $p_{++} = 1$ and hence z_{00} does not contribute information to the measurement probabilities, it does add to the estimation of the missingness model.

Deriving the score equations from (48) and (49) is straightforward but, unlike in the previous section, no closed form exists. Chen, and Fienberg (1974) derived an iterative scheme for the probabilities p_{jk}, based on setting the expected sufficient statistics equal to their *complete-data* counterparts:

$$np_{jk} = z_{11,jk} + z_{10,j} \frac{p_{jk}}{p_{j+}} + z_{01,k} \frac{p_{jk}}{p_{+k}} + z_{00} \frac{p_{jk}}{p_{++}},$$

(with $p_{++} = 1$) and hence

$$(n - z_{00}) p_{jk} = z_{11,jk} + z_{10,j} \frac{p_{jk}}{p_{j+}} + z_{01,k} \frac{p_{jk}}{p_{+k}}. \quad (50)$$

The same equation is obtained from the first derivative of (48). Chen and Fienberg's iterative scheme results from initiating the process with a set of starting values for the p_{jk}, e.g., from the completers, and then evaluating the right hand side of (50). Equating it to the left hand side provides an update for the parameters. The process is repeated until convergence.

While there are no closed-form counterparts to (43) and (44), the expressions equivalent to (45)–(47) are

$$\widehat{z_{11,jk}} = z_{11,jk}, \quad (51)$$

$$\widehat{z_{10,jk}} = z_{10,j} \frac{p_{jk}}{p_{j+}}, \quad (52)$$

$$\widehat{z_{01,jk}} = z_{01,k} \frac{p_{jk}}{p_{+k}}, \quad (53)$$

$$\widehat{z_{00,jk}} = z_{00} p_{jk}. \quad (54)$$

However, there is an important difference between (45)–(47) on the one hand and (51)–(54) on the other hand. In the monotone case, the expressions on the right hand side are in terms of the counts z only, whereas here the marginal probabilities p_{jk} intervene, which have to be determined from a numerical fit.

The practical use of the results in this section are illustrated next on data from the Slovenian Public Opinion Survey.

Table 1: *Data from the Slovenian Public Opinion Survey. The* Don't Know *category is indicated by* $*$.

		Independence		
Secession	Attendance	Yes	No	$*$
Yes	Yes	1191	8	21
	No	8	0	4
	$*$	107	3	9
No	Yes	158	68	29
	No	7	14	3
	$*$	18	43	31
$*$	Yes	90	2	109
	No	1	2	25
	$*$	19	8	96

5 The Slovenian Public Opinion Survey

In 1991 Slovenians voted for independence from former Yugoslavia in a plebiscite. To prepare for this result, the Slovenian government collected data in the Slovenian Public Opinion Survey (SPO), a month prior to the plebiscite. Rubin *et al* (1995) studied the three fundamental questions added to the SPO and, in comparing it to the plebiscite's outcome, drew conclusions about the missing data process.

The three questions added were: (1) Are you in favour of Slovenian independence? (2) Are you in favour of Slovenia's secession from Yugoslavia? (3) Will you attend the plebiscite? In spite of their apparent equivalence, questions (1) and (2) are different since independence would have been possible in confederal form as well and therefore the secession question is added. Question (3) is highly relevant since the political decision was taken that not attending was treated as an effective NO to question (1). Thus, the primary estimand is the proportion θ of people that will be considered as voting YES, which is the fraction of people answering yes to both the attendance and independence question. The raw data are presented in Table 1. We will return to this question in Section 5.2.

Molenberghs *et al* (2001) reanalyzed these data and used them as motivation to introduce their so-called *intervals of ignorance*, a formal way of incorporating uncertainty stemming from incompleteness into the analysis of incomplete contingency tables. To this end, they used the convenient model family proposed by Baker *et al* (1992). We will now introduce the model family.

5.1 The BRD Models

Baker *et al* (1992) proposed a log-linear based family of models for the four-way classification of both outcomes, together with their respective missingness indicators: $\nu_{10,jk} = \nu_{11,jk}\beta_{jk}$, $\nu_{01,jk} = \nu_{11,jk}\alpha_{jk}$, and $\nu_{00,jk} = \nu_{11,jk}\alpha_{jk}\beta_{jk}\gamma$, with

$$\alpha_{jk} = \frac{\phi_{01|jk}}{\phi_{11|jk}}, \qquad \beta_{jk} = \frac{\phi_{10|jk}}{\phi_{11|jk}}, \qquad \gamma = \frac{\phi_{11|jk}\phi_{00|jk}}{\phi_{10|jk}\phi_{01|jk}}.$$

Furthermore $\nu_{r_1r_2,jk}$ is the model for the four cells, indexed by j and k, in pattern (r_1, r_2), where $(r_1, r_2) = (1, 1)$ corresponds to completers, etc.

The α (β) parameters describe missingness in the independence (attendance) question, and γ captures the interaction between both. The subscripts are missing from γ since Baker *et al* (1992) have shown that this quantity is independent of j and k in every identifiable model. These authors considered nine models, based on setting α_{jk} and β_{jk} constant in one or more indices, and enumerated using the 'BRD' abbreviation:

BRD1	:	(α, β)	BRD4	:	(α, β_k)	BRD7	:	(α_k, β_k)
BRD2	:	(α, β_j)	BRD5	:	(α_j, β)	BRD8	:	(α_j, β_k)
BRD3	:	(α_k, β)	BRD6	:	(α_j, β_j)	BRD9	:	(α_k, β_j).

Interpretation is straightforward, for example, BRD1 is MCAR, and in BRD4 missingness in the first variable is constant, while missingness in the second variable depends on its value. BRD6–BRD9 saturate the observed data degrees of freedom, while the lower numbered ones leave room for a non-trivial model fit to the observed data.

5.2 Analysis of the Slovenian Public Opinion Data

The ideas developed in this paper can be illustrated easily by means of 4 models from the BRD family, fitted to the independence and attendance outcomes, i.e., collapsing Table 1. We select models BRD1, BRD2, BRD7, and BRD9. Model BRD1 assumes missingness to be MCAR. All others are of the MNAR type. Model BRD2 has 7 free parameters, and hence does not saturate the observed data degrees of freedom, while models BRD7 and BRD9 saturate the 8 data degrees of freedom. The collapsed data, together with the model fits, are displayed in Table 2. Each of the four models is doubled up with its MAR counterpart.

Table 2 presents, apart from the raw data, for each of the models and its MAR counterpart, the fit to the observed and the hypothetical complete data. The fits of models BRD7, BRD9, and their MAR counterparts to the observed data, coincide with the observed data. As the theory states, every MNAR model and its MAR counterpart produce exactly the same fit to the observed data, which is therefore also seen for BRD1 and BRD2. However, while Models BRD1 and BRD1(MAR) coincide in their fit to the hypothetical complete data, this is not the case for the other three models. The reason is clear: since model BRD1 belongs to the MAR family from the start, its counterpart BRD1(MAR) will not produce any difference, but merely copies the fit of BRD1 to the unobserved data, given the observed ones. Finally, while BRD7 and BRD9 produce a different fit to the complete data, BRD7(MAR) and BRD9(MAR)

coincide. This is because the fits of BRD7 and BRD9 coincide with respect to their fit to the observed data, and indeed, due to their saturation, coincide with the observed data as such. This fit is the sole basis for the models' MAR extensions. It is noteworthy that, while BRD7, BRD9, and BRD7(MAR)≡BRD9(MAR) all saturate the observed data degrees of freedom, their complete-data fits are dramatically different.

Let us return to the implications of our results for the primary estimand θ, the proportion of people voting YES by simultaneously being in favor of independence and deciding to take part in the vote. Rubin *et al* (1995) considered, apart from simple models such as complete case analysis ($\widehat{\theta} = 0.928$) and available case analyses ($\widehat{\theta} = 0.929$), both ignorable models ($\widehat{\theta} = 0.892$ when based on the two main questions and $\widehat{\theta} = 0.883$ when using the secession question as an auxiliary variable) and a non-ignorable one ($\widehat{\theta} = 0.782$). Since the value of the plebiscite was $\theta_{\text{pleb}} = 0.885$, an important benchmark obtained four weeks after the SPO, they concluded the MAR was preferable. Molenberghs *et al* (2001) supplemented these analysis with a so-called pessimistic-optimistic interval, obtained from replacing the incomplete data with NO and YES, respectively, and obtained: $\theta \in [0.694, 0.904]$. Further, they considered all nine BRD models, producing a range for θ from 0.741 to 0.892. Ultimately, these authors devised a method to consider overspecified models, in which point estimates are replaced by interval estimates, so-called *intervals of ignorance*.

Let us consider the results obtained from fitting each of the nine BRD models. Molenberghs *et al* (2001) presented a summary table but unfortunately there was a small computational error that had to be corrected, for which reason the corrected results are reproduced here (Table 3). BRD1 produces $\widehat{\theta} = 0.892$, exactly the same estimate as the first MAR estimate obtained by Rubin *et al* (1995). This should not come as a surprise, since both BRD1 and Rubin's model assume MAR and use information from the two main questions. Before continuing with the models' interpretation, it is necessary to assess their fit. Conducting likelihood ratio tests for BRD1 versus the ones with 7 parameters, BRD2–BRD5, and then in turn for BRD2–BRD5 versus the saturated modes BRD6–BRD9, suggests the lower numbered models do not fit well, leaving us with BRD6–BRD9. The impression might be generated that the poor model fit of BRD1 might be seen as evidence for discarding the MAR-based value 0.892. However, studying the MAR values from each of the models BRD1(MAR)–BRD9(MAR), as displayed in the last column of Table 3, it is clear that this value is remarkably stable and hence a value of $\widehat{\theta} = 0.892$, based on the four counterparts BRD6(MAR)–BRD9(MAR), is a sensible choice after all. Thus, a main contribution resulting from considering the counterparts in this particular example, is the provision of a solid basis for the MAR-based estimate. Obviously, since Models BRD6(MAR)–BRD9(MAR) are exactly the same and exhibit a perfect fit, the corresponding probabilities $\widehat{\theta}_{\text{MAR}}$ are exactly equal too. In this particular case, even though BRD2(MAR)–BRD5(MAR) differ among each other, the probability of being in favor of independence and attending the plebiscite is constant across these four models. This is a mere coincidence, since all three other cell probabilities are different, but only slightly so. For example, the probability of being in favour of independence combined with not attending ranges over 0.066–0.0685 across these four models.

We have made the following two-stage use of Models BRD6(MAR)–BRD9(MAR). At the first stage, in a conventional way, the fully saturated model is selected as the only adequate description of the observed data. At the second stage, these models are transformed into their MAR

counterpart, from which inferences are drawn. As such, the MAR counterpart usefully supplements the original models BRD6–BRD9 and provide one further, important scenario to model the incomplete data. In principle, the same exercise can be conducted when the additional secession variable would be used.

6 Discussion

In this paper, we have shown that every MNAR model, fitted to a set of incomplete data, can be replaced by an MAR version which produces exactly the same fit to the observed data. There are several important implications of this. First, unless one puts strong *a priori* belief in the posited MNAR model, it is not possible to use the fit of an MNAR model for or against MAR, a message in line with Gill, van der Laan, and Robins (1997) and Schafer and Graham (2002). Second, it sometimes happens that the obvious parametric MAR model does not fit the observed data well. This is the case for BRD1, fitted to the Slovenian Public Opinion Survey. It is then appealing to fit a sufficiently versatile MNAR model, to ensure a good fit to the observed data, and then to use the MAR version. Various forms of use can be given to this MAR counterpart. To begin with, the MAR counterpart of a single, well-fitting MNAR model can be used as the sole basis for inference. More realistically, one can consider a variety of well-fitting MNAR models, such as BRD6–9, and then switch to the corresponding collection of MAR counterparts. For example, for the SPO data, BRD6(MAR)–BRD9(MAR) all provide the same answer, unlike the MNAR models they originate from. Finally, an MAR counterpart or several MAR counterparts, can be used as a component of a sensitivity analysis. In this respect, it is useful to recall that all MNAR models, saturating the observed data, produce the same counterpart.

Note further that the collection of counterparts provides a way of constructing an entire collection of MAR models. Especially with non-monotone missing data this is a less than trivial matter. Indeed, the determination of the MAR version of an MNAR model is straightforward in the case of dropout, since the ACMV restrictions, established by Molenberghs *et al* (1998) and translated in a computational scheme by Thijs *et al* (2002), provides a convenient algorithm. In the case of non-monotone missingness, the marginal density of the outcomes is needed. This is straightforward when the model fitted is of the SeM type. When a PMM is fitted, the marginal density follows from a weighted sum over the pattern-specific measurement models, for which no explicit construction exists.

It is also worth noting that the best possible MAR model can always be obtained through the construction proposed in this paper. One merely has to consider the best fitting model, perhaps a saturated model, and then construct its MAR counterpart, which by definition does not alter the fit.

The analyst might want to examine the differences between how the MNAR model on the one hand and its MAR counterpart on the other hand fit the hypothetical complete data, so as to better understand what individuals and/or which parts of the data make the missing data mechanism appear MNAR. For a thorough discussion, we refer to the sensitivity analysis part of Molenberghs and Kenward (2007). Also, Beunckens *et al* (2007) have considered this issue in the context of the Slovenian Public Opinion Survey.

Our re-analysis of the Slovenian Public Opinion Survey data has shown that, while a set of MNAR models produces a widely varying range of conclusions about the proportion of people who are jointly in favor of independence and plan to attend the plebiscite, the corresponding MAR models produce a very narrow range of estimates, which in addition all lie close to the outcome of the plebiscite. This provides evidence for the claim, also made in Rubin *et al* (1995), that choosing an MAR model as one's main route of analysis is a sensible one.

While the result of Theorem 1 is general, we have focused in the paper on SeM and PMM formulations. It is worth re-emphasizing that also the SPM is covered without any problem. In this case, the likelihood is expressed as

$$L = \prod_i \int f(\boldsymbol{y_i}^o, \boldsymbol{y_i}^m|\boldsymbol{\theta}, \boldsymbol{b_i}) f(\boldsymbol{r}_i|\psi, \boldsymbol{b_i}) d\boldsymbol{y_i}^m, \tag{55}$$

with $\boldsymbol{b_i}$ the shared parameter, often taking the form of random effects. To apply our result, $f(\boldsymbol{y_i}^o, \boldsymbol{y_i}^m|\widehat{\boldsymbol{\theta}}, \boldsymbol{b_i})$ needs to be integrated over the shared parameter. The model as a whole needs to be used to produce the fit to the observed data, and then (11) is used to extend the observed-data fit to complete-data MAR version.

A Proof of Lemma 1

Suppressing parameters and covariates from notation, the decomposition of the full data density, in both SeM and PMM fashion, whereby MAR is applied to the SeM version, produces:

$$f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m) f(\boldsymbol{r}_i|\boldsymbol{y}_i^o) = f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m|\boldsymbol{r}_i) f(\boldsymbol{r}_i). \tag{56}$$

Further factoring the right hand side and moving the second factor on the left to the right as well gives:

$$\begin{aligned} f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m) &= f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o, \boldsymbol{r}_i) \frac{f(\boldsymbol{y}_i^o|\boldsymbol{r}_i) f(\boldsymbol{r}_i)}{f(\boldsymbol{r}_i|\boldsymbol{y}_i^o)} \\ f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m) &= f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o, \boldsymbol{r}_i) \frac{f(\boldsymbol{y}_i^o, \boldsymbol{r}_i)}{f(\boldsymbol{r}_i|\boldsymbol{y}_i^o)} \\ f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o) f(\boldsymbol{y}_i^o) &= f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o, \boldsymbol{r}_i) f(\boldsymbol{y}_i^o), \end{aligned}$$

and hence

$$f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o) = f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o, \boldsymbol{r}_i).$$

Acknowledgment

Ivy Jansen and Geert Molenberghs gratefully acknowledge support from *Fonds Wetenschappelijk Onderzoek-Vlaanderen* Research Project G.0002.98 "Sensitivity Analysis for Incomplete and Coarse Data" and from IAP research Network P6/03 of the Belgian Government (Belgian Science Policy).

References

Baker, S.G. (1995). Marginal regression for repeated binary data with outcome subject to non-ignorable non-response. *Biometrics* **51**, 1042-52.

Baker, S.G., Rosenberger, W.F., and DerSimonian, R. (1992). Closed-form estimates for missing counts in two-way contingency tables. *Statistics in Medicine* **11**, 643-57.

Beunckens, C., Sotto, C., **Molenberghs, G.,**, and Verbeke, G. (2007). An integrated sensitivity analysis of the Slovenian Public Opinion Survey data. *Submitted for publication.*

Chen, T., and Fienberg, S.E. (1974). Two-dimensional contingency tables with both completely and partially cross-classified data. *Biometrics* **30**, 629-42.

Diggle, P.J., and Kenward, M.G. (1994). Informative drop-out in longitudinal data analysis (with discussion). *Applied Statistics* **43**, 49-93.

Follmann, D., and Wu, M. (1995). An approximate generalized linear model with random effects for informative missing data. *Biometrics* **51**, 151-68.

Gill RD, van der Laan MJ, Robins JM. (1997). Coarsening at random: characterizations, conjectures and counterexamples. In: *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis*, DY Lin and TR Fleming (eds), pp. 255–294. New York: Springer.
In: Proceedings of the First Seattle Symposium on Survival Analysis, pp. 255–294.

Glynn, R.J., Laird, N.M., and Rubin, D.B. (1986). Selection modelling versus mixture modelling with non-ignorable nonresponse. In: *Drawing Inferences from Self Selected Samples*, Ed. H. Wainer, pp. 115-142. New York: Springer-Verlag.

Jansen, I., Hens, N., Molenberghs, G., Aerts, M., Verbeke, G., and Kenward, M.G. (2006). The nature of sensitivity in missing not at random models. *Computational Statistics and Data Analysis* **50**, 830-58.

Jansen, I., Molenberghs, G., Aerts, M., Thijs, H., and Van Steen, K. (2003). A Local influence approach applied to binary data from a psychiatric study. *Biometrics* **59**, 410-9.

Kenward, M.G. (1998). Selection models for repeated measurements with non-random dropout: an illustration of sensitivity. *Statistics in Medicine* **17**, 2723-32.

Kenward, M.G., Molenberghs, G., and Thijs, H. (2003). Pattern-mixture models with proper time dependence. *Biometrika* **90**, 53-71.

Laird, N.M. (1994). Discussion to Diggle, P.J. and Kenward, M.G.: Informative dropout in longitudinal data analysis. *Applied Statistics*, **43**, 84.

Little, R.J.A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association* **88**, 125-34.

Little, R.J.A. (1994a). A class of pattern-mixture models for normal incomplete data. *Biometrika* **81**, 471-83.

Little, R.J.A. (1994b). Discussion to Diggle, P.J. and Kenward, M.G.: Informative dropout in longitudinal data analysis. *Applied Statistics*, **43**, 78.

Little, R.J.A. (1995). Modelling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association* **90**, 1112-21.

Little, R.J.A., and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. New York: Wiley.

Molenberghs, G., Kenward, M.G., and Goetghebeur, E. (2001). Sensitivity analysis for incomplete contingency tables: the Slovenian plebiscite case. *Applied Statistics* **50**, 15-29.

Molenberghs, G. and Kenward, M.G. (2007). *Missing Data in Clinical Studies*. Chichester: John Wiley.

Molenberghs, G., Kenward, M.G., and Lesaffre, E. (1997). The analysis of longitudinal ordinal data with non-random dropout. *Biometrika* **84**, 33-44.

Molenberghs, G., Michiels, B., Kenward, M.G., and Diggle, P.J. (1998). Monotone missing data and pattern-mixture models. *Statistica Neerlandica* **52**, 153-61.

Molenberghs, G., Thijs, H., Jansen, I., Beunckens, C., Kenward, M.G., Mallinckrodt, C., and Carroll, R.J. (2004). Analyzing incomplete longitudinal clinical trial data. *Biostatistics* **5**, 445-64.

Molenberghs, G., and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. New York: Springer-Verlag.

Robins, J.M., Rotnitzky, A., and Scharfstein, D.O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association* **93**, 1321-39.

Robins, J.M., Rotnitzky, A., and Zhao, L.P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* **90**, 106-21.

Rubin, D.B. (1976). Inference and missing data. *Biometrika* **63**, 581-92.

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

Rubin, D.B.. (1994). Discussion to Diggle, P.J. and Kenward, M.G.: Informative dropout in longitudinal data analysis. *Applied Statistics*, **43**, 80-82.

Rubin, D.B., Stern, H.S., and Vehovar, V. (1995). Handling "don't know" survey responses: the case of the Slovenian plebiscite. *Journal of the American Statistical Association* **90**, 822-8.

Schafer, J.L. and Graham, J.W. (2002). Missing data: our view of the state of the art. *Psychological Methods* **2**, 147–177.

TenHave, T.R., Kunselman, A.R., Pulkstenis, E.P., and Landis, J.R. (1998). Mixed effects logistic regression models for longitudinal binary response data with informative drop-out. *Biometrics* **54**, 367-83.

Thijs, H., Molenberghs, G., Michiels, B., Verbeke, G., and Curran, D. (2002). Strategies to fit pattern-mixture models. *Biostatistics* **3**, 245-65.

Troxel, A.B., Harrington, D.P., and Lipsitz, S.R. (1998). Analysis of longitudinal data with non-ignorable non-monotone missing values. *Appl. Statist.* **47**, 425-38.

Van Steen, K., Molenberghs, G., Verbeke, G., and Thijs, H. (2001). A local influence approach to sensitivity analysis of incomplete longitudinal ordinal data. *Statistical Modelling: An International Journal* **1**, 125-142.

Verbeke, G., and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer-Verlag.

Verbeke, G., Molenberghs, G., Thijs, H., Lesaffre, E., and Kenward, M.G. (2001). Sensitivity analysis for nonrandom dropout: a local influence approach. *Biometrics* **57**, 7-14.

Wu, M.C., and Bailey, K.R. (1988). Analysing changes in the presence of informative right censoring caused by death and withdrawal. *Statistics in Medicine* **7**, 337-46.

Wu, M.C., and Bailey, K.R. (1989). Estimation and comparison of changes in the presence of informative right censoring: conditional linear model. *Biometrics* **45**, 939-55.

Wu, M.C., and Carroll, R.J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modelling the censoring process. *Biometrics* **44**, 175-88.

Table 2: *Analysis of the Slovenian Public Opinion Survey, restricted to the independence and attendance questions. The observed data are shown, as well as the fit of models BRD1, BRD2, BRD7, and BRD9, and their MAR counterparts, to the observed data and to the hypothetical complete data. The contingency tables' rows (columns) correspond to 'yes'* vs. *'no' on the independence (attendance) question. The four tables in each row correspond to: (i) people responding to both questions; (ii) people responding to independence only; (iii) people responding to attendance only; (iv) people responding to neither question.*

Observed data &
fit of BRD7, BRD7(MAR), BRD9, and BRD9(MAR) to incomplete data

1439	78	159	144	54	136
16	16	32			

Fit of BRD1 and BRD1(MAR) to incomplete data

1381.6	101.7	182.9	179.7	18.3	136.0
24.2	41.4	8.1			

Fit of BRD2 and BRD2(MAR) to incomplete data

1402.2	108.9	159.0	181.2	16.8	136.0
15.6	22.3	32.0			

Fit of BRD1 and BRD1(MAR) to complete data

1381.6	101.7	170.4	12.5	176.6	13.0	121.3	9.0
24.2	41.4	3.0	5.1	3.1	5.3	2.1	3.6

Fit of BRD2 to complete data

1402.2	108.9	147.5	11.5	179.2	13.9	105.0	8.2
15.6	22.3	13.2	18.8	2.0	2.9	9.4	13.4

Fit of BRD2(MAR) to complete data

1402.2	108.9	147.7	11.3	177.9	12.5	121.2	9.3
15.6	22.3	13.3	18.7	3.3	4.3	2.3	3.2

Fit of BRD7 to complete data

1439	78	3.2	155.8	142.4	44.8	0.4	112.5
16	16	0.0	32.0	1.6	9.2	0.0	23.1

Fit of BRD9 to complete data

1439	78	150.8	8.2	142.4	44.8	66.8	21.0
16	16	16.0	16.0	1.6	9.2	7.1	41.1

Fit of BRD7(MAR) and BRD9(MAR) to complete data

1439	78	148.1	10.9	141.5	38.4	121.3	9.0
16	18	11.8	20.2	2.5	15.6	2.1	3.6

Table 3: *Analysis of the Slovenian Public Opinion Survey, restricted to the independence and attendance questions. Summaries on each of the Models BRD1–BRD9 are presented.*

Model	Structure	d.f.	loglik	$\widehat{\theta}$	C.I.	$\widehat{\theta}_{\text{MAR}}$
BRD1	(α, β)	6	-2495.29	0.892	[0.878;0.906]	0.8920
BRD2	(α, β_j)	7	-2467.43	0.884	[0.869;0.900]	0.8915
BRD3	(α_k, β)	7	-2463.10	0.881	[0.866;0.897]	0.8915
BRD4	(α, β_k)	7	-2467.43	0.765	[0.674;0.856]	0.8915
BRD5	(α_j, β)	7	-2463.10	0.844	[0.806;0.882]	0.8915
BRD6	(α_j, β_j)	8	-2431.06	0.819	[0.788;0.849]	0.8919
BRD7	(α_k, β_k)	8	-2431.06	0.764	[0.697;0.832]	0.8919
BRD8	(α_j, β_k)	8	-2431.06	0.741	[0.657;0.826]	0.8919
BRD9	(α_k, β_j)	8	-2431.06	0.867	[0.851;0.884]	0.8919