Bayesian methods for dealing with missing data problems

Article *in* Journal of the Korean Statistical Society · April 2018

DOI: 10.1016/j.jkss.2018.03.002

2 authors, including:

Zhihua Ma
Shenzhen University
19 PUBLICATIONS **128** CITATIONS

Journal of the Korean Statistical Society 47 (2018) 297–313

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

journal homepage: www.elsevier.com/locate/jkss

Review

Bayesian methods for dealing with missing data problems

Zhihua Ma *, Guanghui Chen

Department of Statistics, School of Economics, Jinan University, Guangzhou, China

ARTICLE INFO

Article history:
Received 5 September 2017
Accepted 9 March 2018
Available online 13 April 2018

AMS 2000 subject classifications:
primary 62-02
secondary 62D99

Keywords:
Missing data
Bayesian approach
Non-ignorable missing data mechanism
Missing data model

ABSTRACT

Missing data, a common but challenging issue in most studies, may lead to biased and inefficient inferences if handled inappropriately. As a natural and powerful way for dealing with missing data, Bayesian approach has received much attention in the literature. This paper reviews the recent developments and applications of Bayesian methods for dealing with ignorable and non-ignorable missing data. We firstly introduce missing data mechanisms and Bayesian framework for dealing with missing data, and then introduce missing data models under ignorable and non-ignorable missing data circumstances based on the literature. After that, important issues of Bayesian inference, including prior construction, posterior computation, model comparison and sensitivity analysis, are discussed. Finally, several future issues that deserve further research are summarized and concluded.

Contents

* Corresponding author.
E-mail addresses: mazh1993@stu2016.jnu.edu.cn (Z. Ma), tcghui@jnu.edu.cn (G. Chen).

https://doi.org/10.1016/j.jkss.2018.03.002
1226-3192/

1. Introduction

Missing data, a challenging problem that complicates the process of data analysis, is common and generally unavoidable in many studies, including longitudinal studies, clinical trials and sociological surveys. If missing data in these studies are ignored or handled inappropriately, inferences would be biased and inefficient (Mason, Best, Plewis, & Richardson, 2010). Various approaches have been proposed for dealing with missing data, including ad hoc methods like complete-case (CC) analysis and available-case analysis, as well as "statistical principled" methods including maximum likelihood (ML), multiple imputation (MI), and fully Bayesian (FB) approach. Although ad-hoc approaches have the advantage of simplicity, they are generally inappropriate as they lead to bias and loss of precision. By contrast, "statistical principled" methods are better alternatives since they take account of information from the observed data and the uncertainty introduced by the missing data through setting assumptions on missing data mechanisms (Mason et al., 2010).

As two commonly used "statistically principled" methods, FB and MI are similar in spirit and have tight connection. For MI, comprehensive reviews can be seen in Harel and Zhou (2007), Lee and Simpson (2014), Rubin (2008), and Zhang (2003). MI adopts a two-step procedure: (i) impute the missing values through imputation model and create a small number of datasets; (ii) fit analysis model on the imputed datasets and obtain the pooled estimates. However, in FB, this two steps are combined as a single step, which is the major difference compared to MI. By simultaneously fitting the imputation and analysis model, FB can jointly and directly obtain estimates from the posterior distributions of the parameters and missing variables. Besides, the uncertainty due to missing data is automatically taken into account (Erler, Rizopoulos, Rosmalen, et al., 2016). In this paper, we mainly focus on FB approach. However, since Bayesian approaches can be applied in the imputation step of MI, some frameworks and approaches we introduced, such as Markov chain Monte Carlo (MCMC) and Metropolis–Hasting (M–H) algorithms, can also be adapted to MI.

Bayesian approach provides a natural way to take the uncertainty from missing data into account when making inferences on incomplete data (Daniels & Hogan, 2008; Ibrahim, Chen, Lipsitz, & Herring, 2005). In Bayesian context, missing data are considered as random variables, whose posterior distributions can be obtained by specifying priors on the parameters and missing covariate distributions. The missing variables can be sampled from the corresponding conditional distributions through MCMC, and then inferences can be obtained from the posterior distributions (Ahmed, 2011). By estimating the unknown parameters and the missing data simultaneously, inferences are coherent (Mason, 2010). Also, by allowing informative priors and extra information, Bayesian approach can achieve better and more reliable results even under small sample size (Cai, Song, & Hser, 2010). In addition, models under Bayesian framework for dealing with missing data are constructed in a modular way. To be specific, these models consist of three units: response model, missing covariate distribution and missingness model, so analysts can adapt different units to various situations, and explore a range of assumptions about the missing data mechanism (Mason, 2010).

Recent advances in computation capacity and the rapid development of efficient algorithms have made Bayesian methods more feasible and popular in a wide array of missing data problems (Huang, Chen, & Ibrahim, 2005). There are several available software, such as the BUGS family of programs like WinBUGS (Lunn, Spiegelhalter, Thomas, et al., 2009), JAGS (Martyn, 2003), (Stan Development Team, 2012), and Proc MCMC (SAS/Stat, 2014). WinBUGS is quite powerful and can handle various types of missing data problems, but convergence would be slow with large and hierarchical structured datasets. JAGS, similar to WinBUGS, is an open-source implementation of BUGS model specification, and can be called without opening any IDEs, and have more flexibility to incorporate with other software like R and Python. Stan is another open-source software with similar functionality as WinBUGS but uses a more complicated simulation algorithm, which allows it to converge more quickly than WinBUGS, JAGS and Proc MCMC in complex model circumstances (Liu, Han, Zhao, & Lin, 2016).

Related reviews on Bayesian methods for dealing with missing data are mainly comparative reviews, which compare Bayesian methods with other common methods in missing data circumstances. Ibrahim et al. (2005) reviewed the performance of ML, MI, FB and weighted estimating equations (WEE) in dealing with missing covariate data under generalized linear models (GLMs). They explored the relationships between these methods as well as the properties of each methodology. Through simulated and real data examples, they pointed out that Bayesian methods are generally considered as more powerful in dealing with various missing data problems. These four methods were also discussed by Ibrahim, Chu, and Chen (2012) in the cox regression setting in longitudinal studies. Chen and Ibrahim (2014) examined the performance and relationships between MI, ML and FB under Missing at Random (MAR) assumption and they found a close connection between these three methods. With a large sample size, Bayesian methods with non-informative priors on all parameters will lead to ML estimates, and the imputation step in MI is based on sampling from a posterior predictive distribution. Liu et al. (2016) compared Bayesian approaches with frequentist methods through a clinical trial to show the properties, advantages and flexibility of Bayesian methods. These review papers mainly focus on discovering the relationships between

different methods and comparing their performances under specific settings. However, existing reviews do not summarize how Bayesian methods are employed in various settings. In the literature of applying Bayesian approach in missing data problems, researchers mainly focus on dealing with missing response or missing covariates under ignorable or non-ignorable missing mechanisms. In this paper, we will summarize the commonly used missing data models and some issues in Bayesian inference procedure based on the literature.

This article is a review of Bayesian methods for handling missing data problems. Different from other review papers about Bayesian approach in missing data, this article focus on the recent developments and applications of Bayesian methods for dealing with missing data. In Section 2, we will give some background knowledge about missing data mechanisms and Bayesian framework for dealing with missing data. Sections 3 and 4 introduce the commonly used missing data models under different missingness mechanism assumptions. Several crucial steps in Bayesian inference is discussed in Section 5. Section 6 gives a brief review of some other related topics. Conclusion and future issues are presented in Section 7.

2. Background knowledge

2.1. Missing data mechanism

Let $\{(Y_i, \boldsymbol{X}_i), i = 1, \ldots, N\}$ denote an N-component vector of the data from a study. Here Y_i is the response variable and $\boldsymbol{X}_i = (x_{i1}, \ldots, x_{ip})'$ is a vector of p covariates measured on the ith subject. Let $\boldsymbol{R} = (R_y, \boldsymbol{R_x})$ be the corresponding N-component indicator vector of observed response and covariate data, with $R_{yi} = 1$ if Y_i is observed and $R_{yi} = 0$ if Y_i is missing, $R_{xij} = 1$ if x_{ij} is observed and $R_{xij} = 0$ if x_{ij} is missing ($j = 1, \ldots, p$). Then $Y_{(1)} = (Y_i : r_{yi} = 1)$ and $\boldsymbol{X}_{(1)} = (\boldsymbol{X}_i : \boldsymbol{r}_{xi} = \{1\}^p)$ correspond to the observed response and covariate vectors, while $Y_{(0)} = (Y_i : r_{yi} = 0)$ and $\boldsymbol{X}_{(0)} = (\boldsymbol{X}_i : \boldsymbol{r}_{xi} = \{0\}^p)$ denote the missing response and covariate vectors.

For simplicity, here we assume that missing data only exist in the response variable. Missing data mechanism is the conditional distribution of R given Y and parameter $\boldsymbol{\phi}$. Let the probability that R_y takes the value $r_y = (r_{y1}, \ldots, r_{yN})$ given that Y takes the value y be $f(R_y = r_y|Y = y, \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{\phi})$. Let $\widetilde{r}_y$ and $\widetilde{y}_{(1)}$ denote a particular sample realization of R_y and $Y_{(1)}$, respectively. When dealing with missing data, it is helpful to distinguish between ignorable and non-ignorable missingness mechanisms. "Ignorable" means that inferences from a model for the data alone are equivalent to that from a joint model for the data and missingness mechanism, indicating that we can ignore the missingness model when analyzing (Seaman, Galati, Jackson, & Carlin, 2013). Conversely, "non-ignorable" missingness mechanism means that a joint model capturing the data and the missingness pattern should be constructed when modeling.

Within the Bayesian framework, the missingness mechanism is termed ignorable when the parameters governing the measurement and missingness process are distinct, and the missing data are Missing Completely at Random (MCAR) or Missing at Random (MAR). And non-ignorable missingness refers to the situation when missing data are Missing not at Random (MNAR) (Ibrahim & Molenberghs, 2009).

Missing data are MCAR if the missingness does not depend on any values of Y, either observed or unobserved data. Under MCAR, the observed data is just a random sample of the whole data. In this case, ad-hoc methods like CC analysis may lose efficiency, but the resulting estimator is unbiased (Ibrahim & Molenberghs, 2009). According to Mealli and Rubin (2015), the missing data are MCAR if

$$f(R_y = \widetilde{r}_y|Y = y, \boldsymbol{\phi}) = f(R_y = \widetilde{r}_y|\boldsymbol{\phi}), \forall y, \boldsymbol{\phi}. \tag{1}$$

Missing data are MAR if the missingness does not depend on the unobserved values of Y, given the observed ones. Under MAR, a CC analysis will be both inefficient and biased. As stated in Rubin (1976), the missing data are MAR if

$$f(R_y = \widetilde{r}_y|Y_{(1)} = \widetilde{y}_{(1)}, Y_{(0)} = y_{(0)}, \boldsymbol{\phi}) = f(R_y = \widetilde{r}_y|Y_{(1)} = \widetilde{y}_{(1)}, Y_{(0)} = y'_{(0)}, \boldsymbol{\phi}), \forall \boldsymbol{\phi}, y_{(0)}, y'_{(0)}. \tag{2}$$

When neither MCAR nor MAR holds, the missing data are MNAR. MNAR is the most general situation and is frequently encountered in reality, especially in longitudinal studies with repeated measures. Under MNAR, extra model for the missingness mechanism is required. According to Mealli and Rubin (2015), MNAR holds if for some $\boldsymbol{\phi}$ and some $y_{(0)} \neq y'_{(0)}$,

$$f(R_y = \widetilde{r}_y|Y_{(1)} = \widetilde{y}_{(1)}, Y_{(0)} = y_{(0)}, \boldsymbol{\phi}) \neq f(R_y = \widetilde{r}_y|Y_{(1)} = \widetilde{y}_{(1)}, Y_{(0)} = y'_{(0)}, \boldsymbol{\phi}). \tag{3}$$

2.2. Bayesian framework for dealing with missing data

In this section we introduce the Bayesian inference procedure for missing data, which involves four crucial parts (Fig. 1). The first part is constructing the missing data model, including a response model, a missing covariate distribution if needed, and a factorization framework if non-ignorable missing data exist. Suitable response models can be specified by considering the types of the responses, the relationship between the response variables and the covariates, and other factors. In addition to the response model, when missing covariates exist in the data, a covariate distribution is needed as well. These two issues will be discussed in Section 3.

When missingness is non-ignorable, then which analyzing framework to be applied should be determined. Different frameworks can be built according to different factorization forms. Selection model (SM), pattern mixture model (PMM) and

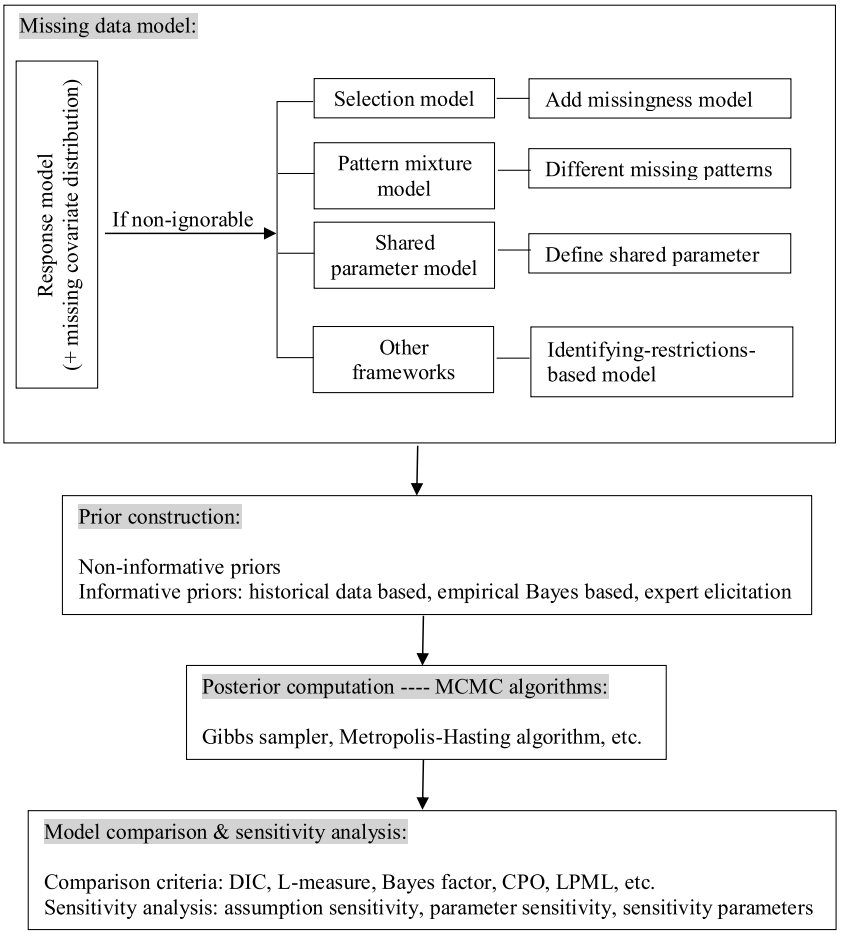

Fig. 1. Bayesian framework with missing data.

shared parameter model (SPM) are three common frameworks. In SM, a missingness model should be specified explicitly, while in PMM, the same model structure with distinct parameters are fitted according to different missing patterns. In SPM, a shared latent random effect should be defined. Other frameworks like identifying-restrictions-based model can also be considered.

The second part is prior construction. In Bayesian analysis, prior distributions should be assigned for the unknown parameters. Non-informative priors are usually used when no additional information can be imposed. However, when external information can be utilized, informative priors are more helpful, especially for the problem of identification. In this paper we introduce three common ways for constructing informative priors: historical data based priors, empirical Bayes based priors and expert elicitation priors.

After constructing the priors, a posterior distribution can be obtained through Bayes Theorem, and MCMC algorithms can be applied to make explicit inferences. After that, sensitivity analysis is necessary to test the sensitivity of the assumptions due to the inability to know the real data model and real missingness mechanism.

3. Missing data model for ignorable missing data

Following the notations in Section 2.1, the joint distribution of $(Y_i, \boldsymbol{X}_i)$ can be specified as the product of the conditional distribution of Y_i given $\boldsymbol{X}_i$, denoted by $[Y_i|\boldsymbol{X}_i, \boldsymbol{\beta}]$, and the marginal distribution of $\boldsymbol{X}_i$, denoted by $[\boldsymbol{X}_i|\boldsymbol{\alpha}]$. $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ parameterize the distribution of Y_i given $\boldsymbol{X}_i$ and the distribution of $\boldsymbol{X}_i$, respectively. Let $D = (N, Y, \boldsymbol{X})$ denote the complete data. The complete data likelihood for all subjects is given by

$$L(\boldsymbol{\beta}, \boldsymbol{\alpha}|D) = \prod_{i=1}^{N} f(Y_i|\boldsymbol{X}_i, \boldsymbol{\beta}) f(\boldsymbol{X}_i|\boldsymbol{\alpha}). \tag{4}$$

Our main interest lies in inferences of $\boldsymbol{\beta}$ based on the observed data. Let $D_{obs} = (N, Y_{(1)}, \boldsymbol{X}_{(1)})$ denote the observed data. When the missingness mechanism is ignorable, the joint posterior distribution of $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is given by:

$$f(\boldsymbol{\beta}, \boldsymbol{\alpha}|D_{obs}) \propto \left\{ \prod_{1}^{N} \left[\int_{Y_{(0)}} \int_{X_{(0)}} f(Y|\boldsymbol{X}, \boldsymbol{\beta}) f(\boldsymbol{X}_{(0)}|\boldsymbol{X}_{(1)}, \boldsymbol{\alpha}) d\boldsymbol{X}_{(0)} dY_{(0)} \right] \right\} \pi(\boldsymbol{\beta}, \boldsymbol{\alpha}), \tag{5}$$

where $f(Y|\boldsymbol{X}, \boldsymbol{\beta})$ denotes the full-data response model, and $\pi(\boldsymbol{\beta}, \boldsymbol{\alpha})$ denotes the joint prior distribution of $(\boldsymbol{\beta}, \boldsymbol{\alpha})$.

In general, the multi-dimensional integrals in (5) do not have a closed form. In addition, if there are several missing covariates, the formula would be of high dimension, leading to difficulty in direct computation from the posterior. In this case, MCMC methods can be used to sample from the posterior, which will be discussed in the following section.

The literature on Bayesian methods for dealing with ignorable missing data can be classified according to the missing components. In ignorable missing data settings, a response model and covariate distributions for missing covariates if needed should be constructed, but without a missingness model.

3.1. Missing response

In ignorable missing response settings, a suitable response model can be built according to the type of response as well as the relationship between response variable and the covariates. Here we introduce several popular response models in the literature of Bayesian methods for dealing with missing data.

Generalized linear model (GLM), which allows response variables to have non-normal error distributions, is the most popular response model. If data are organized at more than one levels, multilevel models and its extension, generalized linear mixed model (GLMM), are more suitable. When dealing with multiple responses, especially responses that are of mixed types, finite mixture models are much more welcomed. Growth mixture model (GMM), a combination of finite mixture model and latent growth curve models, is a flexible approach for analyzing longitudinal data with mixture distributions (Knott & Bartholomew, 1999). Structural equation modeling (SEM) is often used to assess unobservable latent constructions, which is a powerful multivariate regression technique when the variables are latent or unobserved (Das, Chen, Kim, & Warren, 2008; Kaplan, 2000). Its extensions, including mixture SEM (Zhu & Lee, 2001), nonlinear SEM (Lee & Zhu, 2000) are also employed in related researches. Quantile regression (QR) models have become increasingly popular due to its robust property since no assumptions are needed on the error distributions, and it provides a more complete picture of the covariate effects by assessing them at different quantiles of the response (Koenker, 2005). When capturing within-subject serial correlation in longitudinal studies, transition Markov model (TMM) is usually used to allow the expected response at a given time to depend on the previous responses (Kaciroti, Raghunathan, Schork, & Clark, 2008).

Theoretically, any statistical model is suitable as a response model, so analysts can choose the most suitable one according to necessity. For example, in order to analyze data on the incidence of the childhood diabetes in Finland, Moltchanova, Penttinen, and Karvonen (2005) assumed a multinomial model for the MAR missing count response with the probability following a hazard function or a survival function. In longitudinal circumstance, in order to take serial dependence structure into account, Su and Hogan (2008) developed a Bayesian TMM for MAR binary data and used nonparametric smooth functions. In multiple responses situation, the correlations between responses should be considered additionally. Hong, Chu, Zhang, and Carlin (2016) proposed a Bayesian hierarchical model for multiple responses in mixed treatment comparison settings. They introduced novel Bayesian approaches for multiple count or continuous responses simultaneously by incorporating missing data and correlation structure between responses through parameterizations. Deyoreo, Reiter, and Hillygus (2016) presented a Bayesian mixture model for mixed ordinal and nominal data under ignorable missingness assumption in an analysis of the 2012 American National Election Study.

3.2. Missing covariates

Missing covariate data occur frequently in various settings, including surveys, epidemiological studies, environmental studies and clinical trials (Tang & Zhao, 2014). In missing covariate cases, missing covariate distributions are needed in addition to a response model. The construction of covariate distributions is also related to the data types and the correlation between the missing components. When there are more than one missing covariates in the dataset, two ways are commonly used in the literature. The first one is modeling all of the missing covariates using multivariate distributions. For example, using a multivariate normal distribution for several continuous missing covariates, or a multivariate probit regression for correlated binary covariates instead. The second approach is factorizing the joint distribution as a product of a sequence of one-dimensional conditional distributions of each missing covariate (Ibrahim, Chen, & Lipsitz, 2002). More precisely, let $\boldsymbol{X}_i = (x_{i1}, \ldots, x_{is})'$ denote the s-dimensional missing covariates, the joint distribution can be written as a product of a series of one-dimensional conditional distributions, given by:

$$f(x_{i1}, \ldots, x_{is}|\boldsymbol{\alpha}) = f(x_{is}|x_{i1}, \ldots, x_{i,s-1}, \alpha_s) \cdots f(x_{i2}|x_{i1}, \alpha_2) f(x_{i1}|\alpha_1) \tag{6}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_s)'$ and α_k denotes the indexing parameters of the kth conditional distribution.

When missing covariates are of high dimension (i.e. s is large) or of mixed data types, the second approach is preferred. That is because many nuisance parameters from directly specifying a joint distribution will be unidentifiable. What is

more, Gibbs sampling will become computationally intensive and inefficient. Using (6) can reduce the number of nuisance parameters as well as the loss of efficiency of Gibbs sampler. In addition, (6) is more suitable for missing covariates of mixed types. In some situations that continuous and discrete covariates are both missing, it will be difficult to specify a joint distribution for these covariates directly but easier to specify conditional distributions for covariates of each type. Ibrahim et al. (2002) had also shown that (6) had other attractive advantages over the first approach, such as easing the prior elicitation for nuisance parameters.

It should be noted that the specification in the second approach is not invariant to the order of the conditioning, meaning that different orderings can lead to different joint distributions (Xu, Daniels, & Winterstein, 2016). For the order of the covariates, Chen and Ibrahim (2001) suggested to condition the categorical variables on the continuous variables. Erler et al. (2016) took the order according to the proportion of missing values and started with the variable with the least missing values. Based on both the data type and proportion of missingness, Xu et al. (2016) suggested a default ordering that specifying categorical variable firstly, followed by binary variables and then continuous variables. And within the same data type, variables with less missingness are specified before those with more missingness. With this order, the efficiency of the MCMC algorithm can be facilitated. To alleviate the issue of the order of the variables, Xu et al. (2016) used Bayesian additive regression trees (BART) for modeling the conditional mean function to flexibly impute continuous and binary missing covariates. However, it has been shown that sequential specifications used in Bayesian approach are quite robust against changes in the ordering, and as long as the models fit the data well enough, the results would be unbiased even if the order is misspecified. (Chen & Ibrahim, 2001; Zhu & Raghunathan, 2015).

In the literature of applying Bayesian methods for dealing with ignorable missing covariates, writing the joint covariate distribution as a product of piecewise conditional distribution is relatively more popular. Ibrahim et al. (2002) proposed Bayesian inference for GLMs with missing covariate data. They used a GLM with a logit link to fit binary response variable and wrote the missing covariate distribution as a product of one-dimensional conditional distribution. The same response model was used in Carrigan, Barnett, Dobson, and Mishra (2007) to fit longitudinal data with continuous missing covariates and took account of the longitudinal study design by introducing random effects in the model. Their paper gave a detailed instruction of how to construct the model in WinBUGS. Similarly, Chen, Ibrahim, and Lipsitz (2002) analyzed survival data with multiple continuous missing covariates using semiparametric survival model as the response model and the product of normal distributions as the joint covariate distribution.

SEMs were employed in Lee and Song (2004) and Das et al. (2008) for analyzing data with both missing responses and missing covariates. In Lee and Song (2004), missing continuous and ordinal categorical data were treated as latent quantities and were linked through a nonlinear SEM. Motivated by a multilevel survey, Das et al. (2008) also employed a SEM which involves a set of latent variables and random effects capturing dependence between responses and heterogeneity respectively.

In longitudinal studies, one important case is time-varying missing covariates. Pettitt, Tran, Haynes, and Hay (2006) employed a Bayesian hierarchical model to analyze categorical longitudinal data with time-varying missing covariates. A GLMM was built for binary response and a transition model taking the previous time points into account was built for time-varying missing covariates. Yu, Chen, Huang, and Anderson (2013) developed a generalized linear mixed probit regression model for the repeated binary responses and a joint model for time-dependent missing covariates.

Parametric models are usually used in specifying the joint distribution of the missing covariates, while nonparametric and semiparametric approaches are also considered. For example, Si and Reiter (2013) presented a nonparametric Bayesian joint modeling approach for multivariate categorical data based on Dirichlet process mixtures of multinomial distributions. This method was extended by Murray and Reiter (2016) to adapt for multivariate continuous and categorical variables. Two Dirichlet Process mixtures including a mixture of multinomial distributions for the categorical data, and a mixture of multivariate normal distributions for the continuous variables were employed.

4. Missing data model for non-ignorable missing data

When missing data mechanism is believed to be non-ignorable, a missingness model is needed additionally. Following the notations in the previous section, the missingness model can be denoted by $f(R|Y, \boldsymbol{X}, \boldsymbol{\phi})$, and the corresponding full-data model is denoted by $f(Y, \boldsymbol{X}, R|\boldsymbol{\theta})$ with $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\phi})$. In the missing data literature, both parametric and semiparametric approaches can be used for missingness model construction.

4.1. Missingness model

4.1.1. Parametric missingness model

For the binary missing indicator R_i, it is common to assume a Bernoulli distribution with missing probability τ_i as

$$R_i \sim Bernoulli(\tau_i). \tag{7}$$

When there is only one missing variable in the data, the relationship between the missing probability τ_i and the missing component can be modeled using a link function. The most popular link functions include a logit link $f(\tau_i) = [1 + \exp(\tau_i)]^{-1}$ or a probit link with $f(\tau_i)$ being the distribution function of the standard normal.

When there are more than one missing variables, the joint distribution of the missing indicators can be of the form of a multinomial model, or be represented as a product of one-dimensional conditional distributions similar to (6). To be specific, assuming that there are more than one missing covariates in the data, let $\boldsymbol{R}_{xi} = (r_{xi1}, \ldots, r_{xis})'$ be the corresponding vector of missing indicators. Then the joint distribution of $\boldsymbol{R}_{xi}$ for subject i can be represented as a product of one-dimensional conditional distributions given by

$$f(\boldsymbol{R}_{xi}|Y_i, \boldsymbol{X}_{(0)}, \boldsymbol{X}_{(1)}, \boldsymbol{\phi}) = f(r_{is}|r_{i1}, \ldots, r_{i,s-1}, Y_i, \boldsymbol{X}_{(0)}, \boldsymbol{X}_{(1)}, \phi_s) \cdots f(r_{i2}|r_{i1}, Y_i, \boldsymbol{X}_{(0)}, \boldsymbol{X}_{(1)}, \phi_2) f(r_{i1}|Y_i, \boldsymbol{X}_{(0)}, \boldsymbol{X}_{(1)}, \phi_1) \tag{8}$$

where $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_s)'$. For each one-dimensional conditional distribution of r_{ik}, a logit or probit regression can be built as above.

4.1.2. Semiparametric missingness model

In parametric missingness model, the relationship between the logit or probit form of τ_i and the missing component is assumed to be linear. However, this is quite a rigorous assumption since the relationship is actually unknown and might be nonlinear. Therefore, semiparametric approaches are presented to model the missing data mechanism.

One common semiparametric missingness model is generalized additive models (GAMs). GAM provides a flexible way to characterize the relationship between the missing components and binary missing indicators (Hastie & Tibshirani, 1987). Kalaylioglu and Ozturk (2013) applied GAM in non-ignorable missing covariates settings. In their research, r_{ik} was modeled in (7) with missing probability τ_{ik}, and the relationship between τ_{ik} and the missing components was modeled through a smooth function $m(\cdot)$ given by

$$h(\tau_{ik}) = \phi_{k0} + \boldsymbol{r}_{-ik}\phi_{k1} + \sum_{k=1}^{s} m(\boldsymbol{x}_{i(0)}) + \boldsymbol{x}_{i(1)}\boldsymbol{\phi}_{k2} + y_i\phi_{k3}, \tag{9}$$

where $h(\cdot)$ is a suitable chosen link function such as logit or probit. ϕ_{k0} and ϕ_{k3} are scalars, $\boldsymbol{\phi}_{k2}$ is parameter vectors associated with the observed covariates, and $\boldsymbol{r}_{-ik}$ is the set of missing indicators that excludes component k. $m(\cdot)$ is an unspecified smooth function. The advantage of this model is that it accommodates any possible nonlinear relationship between the missing indicator and missing covariates. In Kalaylioglu and Ozturk (2013), a low-rank thin-plate splines was employed to specify the smooth function. Certainly, other nonparametric approaches can also be used to specify the smooth function, like natural cubic splines, B-splines, truncated polynomials, etc.

4.2. Frameworks for non-ignorable missing data

Little and Rubin (2002) proposed three popular frameworks for dealing with non-ignorable missing data according to the factorization forms of the full-data model: selection model (SM), pattern-mixture model (PMM), and shared-parameter model (SPM). Apart from these three common frameworks, we also introduce other non-ignorable assumptions such as identifying-restrictions-based model.

4.2.1. Selection model

The SM approach is the most commonly used factorization in the literature. It factorizes the full-data model as:

$$f(Y, \boldsymbol{X}, R|\boldsymbol{\theta}) = f(Y|\boldsymbol{X}, \boldsymbol{\beta}) f(\boldsymbol{X}|\boldsymbol{\alpha}) f(R|Y, \boldsymbol{X}, \boldsymbol{\phi}), \tag{10}$$

so we should explicitly specify the response model, missing covariate distribution and missingness model.

One of the advantages of SM is that it specifies the response model $f(Y|\boldsymbol{X}, \boldsymbol{\beta})$ directly, which is usually the main interest of investigators. However, this approach is not advantageous in sensitivity analysis since parameters in SM cannot be easily partitioned as identified and non-identified parameters (Daniels & Hogan, 2008). Besides, identification problem in SM is still not explicitly specified. Constraints on the missingness mechanism should be set in order to ensure identifiability, but how these constraints can be translated into assumptions on the distributions of the missing components is still unclear (Ibrahim & Molenberghs, 2009). Model identifiability is more obscure in the SM approach, so in this case, one needs to characterize identifiability theoretically. For example, Wang, Shao, and Kim (2014) dealt with the problem of identifiability by utilizing a missing instrument, an auxiliary variable that is useful in predicting the study variable but is conditionally independent of the missing indicator given the study variable and other covariates.

On the basis of SM, Zhang and Wang (2012) proposed a simplified SM for regression analysis, assuming that missingness in response only related to itself and no auxiliary variables were used in the model. The advantage of simplified selection model lies in the avoidance of selecting auxiliary variables. By simulation study they showed that simplified SM can recover regression parameters under both correctly specified and misspecified situations.

Applications of Bayesian SM framework to deal with non-ignorable missing data problem is abundant in the literature. For studies with non-ignorable missing response variable only, a response model and a missingness model is needed. Mason et al. (2010) performed a Bayesian SM framework with linear regression for the response model and a logit model as the missingness model, and focused on discovering the effect of the addition of missingness model on the performance of parameter estimation. They found that the addition of missingness model could greatly improve the overall fit of the response

model and lead to better prediction, but skewness in the response would have negative effect on the estimation. Lee and Tang (2006) used nonlinear SEM as response model and a product of logit conditional models as missingness model. Kim and Choi (2014) proposed a Bayesian binomial mixture model for collaborative prediction with factors related to the missingness model following Bernoulli distributions.

When non-ignorable missing covariates also exist in data, a joint missing covariate distribution is required additionally. Cai et al. (2010) employed a mixture SEM to analyze latent variables and heterogeneous data, and logistic models were applied as missingness model. Poleto, Paulino, Singer, and Molenberghs (2015) constructed an analysis framework with a GLM for binary response as response model, a non-parametric model based on a Dirichlet process mixture for the continuous missing covariates as covariate distribution, and a logit link for the missingness mechanism. Similarly within the GLM framework, Kalaylioglu and Ozturk (2013) considered a regression spline based semiparametric approach to model the missingness mechanism of the missing covariates with each piecewise conditional density having the form of a GLM density. Tang and Zhao (2014) considered a nonlinear reproductive dispersion mixed models for longitudinal data and employed logit link for the missingness model.

When additionally take measurement error in covariates into account, a framework consisting of a QR-based mixed-effects model as response model, a measurement error model for missing covariates, and a logit link for missingness model was proposed in Huang (2016).

4.2.2. Pattern mixture model

Unlike SMs, PMMs partition the full-data model as:

$$f(Y, \boldsymbol{X}, R|\boldsymbol{\theta}) = f(Y|R, \boldsymbol{X}, \boldsymbol{\beta})f(\boldsymbol{X}|R, \boldsymbol{\alpha})f(R|\boldsymbol{\phi}). \tag{11}$$

PMMs stratify the data by different missing patterns and allow distinct model parameters for each stratum. In PMMs, response models are built with coefficients variant with different missing patterns. The response model here is a mixture model:

$$f(Y|\boldsymbol{X}, \boldsymbol{\beta}) = \sum_{R\in\Re} f(Y|R, \boldsymbol{X}, \boldsymbol{\beta})f(R|\boldsymbol{X}, \boldsymbol{\phi}). \tag{12}$$

The missingness model can be derived using Bayes' rule:

$$f(R|Y, \boldsymbol{X}, \boldsymbol{\beta}) = \frac{f(Y|R, \boldsymbol{X}, \boldsymbol{\beta})f(R|\boldsymbol{X}, \boldsymbol{\phi})}{f(Y|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\phi})}. \tag{13}$$

PMM approach is well suited in missing data problems as it does not require specific modeling of the missingness model, and it can be easily transformed into extrapolation factorization which makes sensitivity analysis more feasible (Tran, 2008). The extrapolation factorization is:

$$f(Y, \boldsymbol{X}, R|\boldsymbol{\theta}) = f(Y_{(0)}|Y_{(1)}, R, \boldsymbol{X}, \boldsymbol{\beta}_E)f(Y_{(1)}|R, \boldsymbol{X}, \boldsymbol{\beta}_O)f(R|\boldsymbol{X}, \boldsymbol{\phi})f(\boldsymbol{X}|\boldsymbol{\alpha}), \tag{14}$$

where $\boldsymbol{\beta}_E$ and $\boldsymbol{\beta}_O$ correspond to parameters indexing an extrapolation distribution and a model for observables. In general, $\boldsymbol{\beta}_E$ cannot be identified from data alone.

Unlike SM approach, problems of identifiability can be made explicitly in PMM framework since the responses are modeled separately for each missing pattern. Common ways include setting some restrictions and assigning informative prior distributions on the unidentified parameters. In Bayesian framework, imposing informative priors is always preferred. For example, Kaciroti, Raghunathan, Schork, Clark, and Gong (2006) and Little and Rubin (2002) constructed prior distributions on the parameters of missing patterns conditioning on parameters of the observed data to solve the problem of identification in PMM.

Application of incorporating Bayesian methods with PMM framework to deal with missing data is also rich in the literature, especially in longitudinal studies. Motivated by Metabolic Syndrome data, Kyoung and Lee (2015) constructed a GLMM for longitudinal binary response with random effects describing the effect of covariates on response, and a PMM was applied for dropout missingness. The follow-up time for dropout was constructed according to the missing indicator and was assumed to follow a multinomial model. For four missing patterns, different coefficients were produced in the same GLMM framework. For convenience of sensitivity analysis, the parameters were reparametrized in terms of sensitivity parameters and a component was defined to capture information about the missingness mechanism. Similarly, Kaciroti et al. (2008) analyzed longitudinal data with non-ignorable dropout using PMM framework. In their work, a TMM with random effects following Poisson distributions was used for count responses. For each missing data patterns, TMM was applied but allowing the parameters of the random effects to differ across patterns. Informative priors were used to solve the problem of identification. Linero and Daniels (2015) conducted a nonparametric Bayesian inference under non-ignorable monotone missingness using an extrapolation factorization with Dirichlet process mixtures, which enabled introducing sensitivity parameters to vary the untestable assumptions about the missing data mechanism. Their approach was extended by Linero (2017) to adapt for non-monotone missingness.

When additionally considering missing covariates in the analysis, covariate distributions were necessary in the framework. For example, Kaciroti et al. (2006) used a TMM with random effects to investigate changes in ordinal responses over time and PMM was employed to analyze missing response and time-varying covariates. A joint multivariate distribution for the missing time-varying covariates was used. Informative priors using cumulative odds were imposed to identify parameters. When the covariates are MNAR, a PMM with different parameters can be fitted similarly for these covariates as well.

4.2.3. Shared parameter model

Another approach for specifying the full-data model is SPM, where latent random effects are used to relate the response with missing indicators. The general form is given by:

$$f(Y, \boldsymbol{X}, R|\boldsymbol{\theta}) = \int f(Y, \boldsymbol{X}, R, \boldsymbol{b}|\boldsymbol{\theta})d\boldsymbol{b}, \tag{15}$$

where $\boldsymbol{b}$ denotes the latent random effects, and

$$f(Y, \boldsymbol{X}, R, \boldsymbol{b}|\boldsymbol{\theta}) = f(Y|\boldsymbol{X}, \boldsymbol{b}, \boldsymbol{\beta})f(R|Y, \boldsymbol{X}, \boldsymbol{b}, \boldsymbol{\phi})f(\boldsymbol{X}|\boldsymbol{\alpha}, \boldsymbol{b})f(\boldsymbol{b}|\boldsymbol{\varphi}). \tag{16}$$

One advantage of SPMs is that it simplifies the specification of response model and missingness model. Through random effects, SPMs are able to handle multilevel structured data or data with measurement error. However, SPM is difficult to understand and may not have a closed form since it requires integration over the random effects (Daniels & Hogan, 2008).

Yuan and Yin (2010) studied quantile regression for longitudinal responses with non-ignorable intermittent missing data and dropout. The informative missing data were assumed to be related to the longitudinal response process through the shared latent random effects. They firstly extended QR to longitudinal setting, and then introduced individual random effects into the model to link the missingness with longitudinal response process. They assumed the missing data process and the response process to share the same random effects and modeled the missing data process using transition probabilities. Lu, Zhang, and Lubke (2011) also used SPM to deal with non-ignorable missing data. An extended GMM with latent class membership indicator was employed to analyze longitudinal data. The missing indicator followed a Bernoulli distribution with the missing probability following a probit link function of the latent class membership and the covariates.

In order to illustrate the differences between these three common frameworks, a simple example is presented here. Consider a normal response Y and a covariate vector $\boldsymbol{X}$, with Y has missing data and $\boldsymbol{X}$ is completely observed. Define the missing indicator $R_i = 1$ if Y_i is observed. A SM framework factors the full-data distribution as

$$f(Y, \boldsymbol{X}, R|\boldsymbol{\theta}) = f(Y|\boldsymbol{X}, \boldsymbol{\beta})f(R|Y, \boldsymbol{X}, \boldsymbol{\phi}).$$

The response model can be specified as a normal density $N(\mu, \sigma^2)$, while the missingness model can be set as a Bernoulli distribution with a simple regression like

$$R_i \sim Bernoulli(\tau_i), g(\tau_i) = \phi_0 + \phi_1 y_i + \boldsymbol{\phi}_2 \boldsymbol{x}_i,$$

where $g(\cdot)$ is a link function and can take logit or probit.

For a PMM, the full-data model is factorized as

$$f(Y, \boldsymbol{X}, R|\boldsymbol{\theta}) = f(Y|R, \boldsymbol{X}, \boldsymbol{\beta})f(R|\boldsymbol{\phi}).$$

For the response model, we assume the normal response given missing indicator R follows a normal distribution with common variance as

$$Y|R = 1 \sim N(\mu^{(1)}, \sigma^2), Y|R = 0 \sim N(\mu^{(0)}, \sigma^2).$$

And the missingness model $f(R|\boldsymbol{\phi})$ is defined as

$$R \sim Bernoulli(\tau).$$

For the SPM, a latent random effect is considered. Here we assume that $\boldsymbol{W} \subseteq \boldsymbol{X}$ denotes the random effect covariates. The full-data model is given as

$$f(Y, \boldsymbol{X}, R|\boldsymbol{\theta}) = \int f(Y, \boldsymbol{X}, R, \boldsymbol{b}|\boldsymbol{\theta})d\boldsymbol{b},$$

where $f(Y, \boldsymbol{X}, R, \boldsymbol{b}|\boldsymbol{\theta}) = f(Y|\boldsymbol{X}, \boldsymbol{b}, \boldsymbol{\beta})f(R|\boldsymbol{b}, \boldsymbol{\phi})f(\boldsymbol{b}|\boldsymbol{\varphi})$. For the response model, we assume that a normal density $N(\boldsymbol{X\beta} + \boldsymbol{Wb}, \sigma^2)$ is specified. And a normal distribution $N(0, \Psi)$ is assumed for the random effects. For the missingness model, it is specified as

$$R_i|\boldsymbol{b} \sim Bernoulli(\tau_i), g(\tau_i) = \Phi \boldsymbol{b}_i.$$

Similarly, $g(\cdot)$ is a link function.

4.2.4. Identifying-restrictions-based model

Identifying-restrictions-based model (Thijs, Molenberghs, Michiels, et al., 2002) is a strategy for fitting PMMs. As we mentioned above, imposing restrictions can help solve the problem of identifiability in PMMs. Thijs et al. (2002) propose an alternative strategy to deal with this problem. In their work, attention is restricted to monotone patterns. For pattern $t = 1, \ldots, T, y_s = \{Y|r = s\}$, the model is

$$f_t(y_1, \ldots, y_T, r = t) = f_t(y_1, \ldots, y_t)f_t(y_{t+1}, \ldots, y_T|y_1, \ldots, y_t)f(r = t), \tag{17}$$

with identifying restrictions applying on the second component. For a given identified component y_s, a general expression is

$$f_t(y_s|y_1,\ldots,y_{s-1}) = \sum_{j=s}^{T} w_{sj} f_j(y_s|y_1,\ldots,y_{s-1}),\ s = t+1,\ldots,T, \tag{18}$$

and w_{sj} provides a valid identification scheme. Thijs et al. (2002) considered three important identification cases including CCMV, NCMV and ACMV.

The strategy above assumes that missingness depends on past measurement and on the present, but not on future ones. Kenward, Molenberghs, and Thijs (2003) made an extension to develop this so called non-future dependent missingness. The model is given by

$$f_t(y_1,\ldots,y_T, r = t) = f_t(y_1,\ldots,y_t) f_t(y_{t+1}|y_1,\ldots,y_t) f_t(y_{t+2},\ldots,y_T|y_1,\ldots,y_{t+1}) f(r = t), \tag{19}$$

with the first three components represent the distributions of past, present and future measurements, respectively. In (19), the second and third components are unidentifiable from the data. The non-future dependent missing value states that

$$f_t(y_s|y_1,\ldots,y_{s-1}) = f_{(\geq s-1)}(y_s|y_1,\ldots,y_{s-1}), \tag{20}$$

for $s = t+2,\ldots,T$. Using a general expression similar to (19), w_{sj} can be modeled to accommodate different cases. Kenward et al. (2003) considered the same cases as in Thijs et al. (2002) to illustrate the approach.

5. Bayesian inference

When missing data exist in the model framework, they should be integrated out from the likelihood function. In Bayesian approach, it is easy to achieve this goal without additional inferential procedures. Missing data in Bayesian frameworks are regarded as random variables that can be sampled from their corresponding conditional distributions (Tanner & Wong, 1987). In addition, more information can be extracted from the observed data to construct informative priors, which is helpful since there is insufficient information about parameters related to missingness mechanism from the likelihood alone.

In order to obtain the estimates of parameters of interest, posterior distribution $f(\boldsymbol{\beta}, \boldsymbol{\alpha}|D_{obs})$ should be firstly constructed using prior distributions, and then samples can be drawn from the joint posterior distribution through MCMC methods, such as Gibbs sampler (Geman & Geman, 1984) and M–H algorithm (Hastings, 1970; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953).

5.1. Prior construction

Prior distributions quantify the knowledge and information about the unknown parameters. Prior selection is an important issue since the posterior estimates may be sensitive to the choice of the priors of the hyperparameters. Also, some prior distributions may lead to improper posterior distributions and poor mixing as well as slow convergence in MCMC algorithm. Some commonly used non-informative priors and conjugate priors are introduced in Daniels and Hogan (2008). In missing data problems, the problem of identification is common and not enough information from likelihood alone can be used for inference (Ibrahim et al., 2002). In these situations, Bayesian framework with informative priors can be of great help.

5.1.1. Informative priors using historical data

One way of constructing informative priors is making use of historical data (Ibrahim et al., 2002), i.e. data from previous similar studies. Related application of historical informative priors can also be seen in Chen et al. (2002).

Let $D_0 = (N_0, Y_0, \boldsymbol{X}_0)$ denote the complete historical data, where N_0, Y_0 and $\boldsymbol{X}_0$ denote the sample size, response variable and covariates of the historical data, respectively. Similarly, we allow the historical data to have missing response or covariates. The joint prior for $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is of the form:

$$\pi(\boldsymbol{\beta}, \boldsymbol{\alpha}|D_{0,obs}) \propto \pi^*(\boldsymbol{\beta}, \boldsymbol{\alpha}|D_{0,obs})\pi_0(\boldsymbol{\beta}, \boldsymbol{\alpha}), \tag{21}$$

where $\pi_0(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is the initial prior of $(\boldsymbol{\beta}, \boldsymbol{\alpha})$, and

$$\begin{aligned}\pi^*(\boldsymbol{\beta}, \boldsymbol{\alpha}|a_0, D_{0,obs}) = \prod_{i=1}^{n} \int_{Y_{0(0)}} \int_{\boldsymbol{X}_{0(0)}} & f(Y_{0i}|\boldsymbol{X}_{0i}, \boldsymbol{\beta})^{a_0} f(\boldsymbol{X}_{0i1}|\boldsymbol{\beta}, \boldsymbol{\alpha}_1)^{a_{0i1}} \\ & \times \prod_{j=1}^{p-1} f(X_{0i,p-j+1}|X_{0i,1},\ldots,X_{0i,p-j}, \alpha_{p-j+1})^{a_{0i,p-j+1}} d\boldsymbol{X}_{0(0)} dY_{0(0)}\end{aligned} \tag{22}$$

where $D_{0,obs} = (N_0, Y_{0(1)}, \boldsymbol{X}_{0(1)})$ denotes the observed historical data. For simplicity, (22) can be written as

$$\pi^*(\boldsymbol{\beta}, \boldsymbol{\alpha}|a_0, D_{0,obs}) = \int_{Y_{(0)}} \int_{X_{(0)}} \{f(Y_0|\boldsymbol{X}_0, \boldsymbol{\beta}, \boldsymbol{\alpha})\}^{a_0} d\boldsymbol{X}_{(0)} \tag{23}$$

where $a_0 (0 \leq a_0 \leq 1)$ is a scalar prior parameter that weighs the complete data likelihood of the historical data relative to the current ones, which can be interpreted as a precision parameter that controls the heaviness of the tails of the joint prior for $(\boldsymbol{\beta}, \boldsymbol{\alpha})$. $a_0 = 0$ means no historical data is incorporated in the priors while $a_0 = 1$ means historical data are equally weighed with the likelihood of the current study.

Assume a vague prior for $\pi_0(\boldsymbol{\beta}, \boldsymbol{\alpha})$, that is:

$$\pi_0(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \pi_0(\boldsymbol{\beta}|c_0)\pi_0(\boldsymbol{\alpha}|d_0), \tag{24}$$

where c_0 and d_0 are prior parameters of $\pi_0(\boldsymbol{\beta}, \boldsymbol{\alpha})$. Consider a Beta prior for a_0, then the joint prior distribution for $(\boldsymbol{\beta}, \boldsymbol{\alpha}, a_0)$ is given by:

$$\pi(\boldsymbol{\beta}, \boldsymbol{\alpha}, a_0|D_{0,obs}) \propto \pi^*(\boldsymbol{\beta}, \boldsymbol{\alpha}|a_0, D_{0,obs})\pi_0(\boldsymbol{\beta}|c_0)\pi_0(\boldsymbol{\alpha}|d_0)a_0^{\delta_0-1}(1-a_0)^{\lambda_0-1}, \tag{25}$$

where (δ_0, λ_0) are prior parameters. Ibrahim et al. (2002) introduced several choices for the prior parameters (δ_0, λ_0).

5.1.2. Empirical Bayes based priors

Another way of constructing informative priors is empirical Bayes based priors (Huang et al., 2005; Kalaylioglu & Ozturk, 2013). Hyperparameters in empirical Bayes based priors are obtained from the observed data as well as the possible datasets that could be observed from the considered model. Carlin and Louis (1997) introduced the usefulness of empirical Bayes based priors.

The parameter space consists of $(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\phi})'$. Huang et al. (2005) had shown that under GLM, improper priors on $\boldsymbol{\beta}$ (the parameters in response model) and $\boldsymbol{\alpha}$ (the location parameters in missing covariate distribution) will not lead to improper joint posterior distribution as long as proper priors are given for $\boldsymbol{\phi}$ (the parameters in the missingness model). Besides, they also showed that empirical Bayes based priors for $\boldsymbol{\phi}$ will accelerate the convergence of MCMC algorithm. Therefore, improper uniform priors can be taken for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, while an empirical Bayes based prior is constructed for $\boldsymbol{\phi}$. For simplicity, we assume that response variables are completely observed. Let $\boldsymbol{\phi}$ denote the regression coefficients in the missingness model, then its prior distribution is set as:

$$\boldsymbol{\phi} \sim N(\hat{\boldsymbol{\phi}}, c_0\Phi_0), \tag{26}$$

where $\hat{\boldsymbol{\phi}}$ denotes an estimate of mean vector, Φ_0 is an estimate of variance–covariance matrix, and c_0 is a constant used to account for the variation introduced by estimating the prior parameters. $\hat{\boldsymbol{\phi}}$ and Φ_0 can be obtained using the imputation technique as follows:

(1) Obtain the maximum likelihood estimates of $\boldsymbol{\alpha}$, denoted by $\hat{\boldsymbol{\alpha}}_{MLE}$, based on the subjects with fully observed covariates;
(2) Generate K independent samples $\boldsymbol{X}_{(0)k} \sim f(\boldsymbol{X}_{(0)}|\boldsymbol{X}_{(1)}, \hat{\boldsymbol{\alpha}}_{MLE})$ to obtain the imputed missing covariates, and denote the imputed dataset as $D_{imp,k} = (Y, \boldsymbol{X}_{(0)k}, \boldsymbol{X}_{(1)})$ for $k = 1, \ldots, K$;
(3) Plug the imputed missing covariates in the missingness model and obtain $\hat{\boldsymbol{\phi}}_k$ and the information matrix $I_k(\hat{\boldsymbol{\phi}}_k)$ by maximizing the imputed likelihood function, where the information matrix can be calculated as $I(\boldsymbol{\phi}) = \frac{-\partial^2 \ln L(\boldsymbol{\phi})}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}'}$.

Then the hyperparameters of the empirical Bayes based prior, $\hat{\boldsymbol{\phi}}$ and Φ_0, can be calculated as:

$$\hat{\boldsymbol{\phi}} = \frac{1}{K}\sum_{k=1}^{K}\hat{\boldsymbol{\phi}}_k \text{ and } \Phi_0 = \frac{1}{K}\sum_{k=1}^{K}\left[I_k(\hat{\boldsymbol{\phi}}_k)\right]^{-1}. \tag{27}$$

5.1.3. Expert elicitation priors

Expert elicitation is usually used for specifying the priors for one or more unknown parameters of a statistical model. The expert's current knowledge of several aspects of the problem is translated into probabilistic form and then incorporated into the posterior inference through the Bayes' Theorem (Garthwaite, Kadane, & O'hagan, 2005). The elicitation process is divided into four parts: preparing for the elicitation, eliciting specific summaries of the experts' distributions for the unknown parameters, fitting a (joint) probability distribution to the summaries, and assessing the adequacy of the elicitation. Detailed discussions of these four issues can be seen in Garthwaite et al. (2005).

Mason (2010) gave a brief review of the application of expert elicitation in missing data problems and introduced several software packages for the elicitation process. Besides, he also gave a detailed example of expert elicitation using MSC income data. In the elicitation process, it is always difficult to construct a specific distribution from a finite number of statements of the expert. Oakley and O'hagan (2007) proposed a nonparametric approach to allow the expert's distribution to take any continuous form in order to overcome the deficiencies of the commonly used parametric approaches.

5.2. Posterior computation

A revolutionary approach in Bayesian computation to obtain exact inferences for complex model settings is MCMC. The crucial idea is to obtain a sample from the posterior distribution without explicitly evaluating normalizing constant of the posterior distribution by constructing a Markov chain, which has the posterior distribution of interest as its stationary

distribution. Then by doing Monte Carlo integration using the samples from the Markov chain, the marginal posteriors and the posteriors of functions of the parameters can be easily obtained (Daniels & Hogan, 2008). Gibbs sampler and M–H algorithm are two popular MCMC algorithms used in Bayesian inference. Here we give a brief introduction to these two algorithms.

(1) The Gibbs sampler.

The Gibbs sampler (Geman & Geman, 1984) is the most popular MCMC algorithm. It involves sampling a Markov chain which takes the product of the sequentially updated full conditional distributions of the parameters as the kernel and the posterior as the stationary distribution. Let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)'$ denote a q-dimensional parameter vector and let $\beta_j^{(k)}$ denote the sample of the jth component of $\boldsymbol{\beta}$ at iteration k, and then sample from the following distributions sequentially,

1. $\beta_1^{(k)} \sim f(\beta_1^{(k)}|\beta_2^{(k-1)}, \beta_3^{(k-1)}, \ldots, \beta_q^{(k-1)}, y)$;
2. $\beta_2^{(k)} \sim f(\beta_2^{(k)}|\beta_1^{(k)}, \beta_3^{(k-1)}, \ldots, \beta_q^{(k-1)}, y)$;

q. $\beta_q^{(k)} \sim f(\beta_q^{(k)}|\beta_1^{(k)}, \beta_2^{(k)}, \ldots, \beta_{q-1}^{(k)}, y)$.

(2) The Metropolis–Hastings algorithm.

M–H algorithm (Hastings, 1970) is a popular approach for sampling from non-standard full conditional distributions. The full conditional distribution of interest is given by

$$f(\beta_j^{(k)}|\{\beta_l^{(k)} : l < j\}, \{\beta_l^{(k-1)} : l > j\}, y).$$

For simplicity, we write the above conditional distribution as $f(\beta_j^{(k)})$. At the kth iteration, sampling $\beta_j^{(k)}$ from some candidate distributions $p(\beta_j^{(k)})$, and the sampled values are accepted with probability α^* given by:

$$\alpha^* = \min\left\{1, \frac{f(\beta_j^{(k)})p(\beta_j^{(k-1)}|\beta_j^{(k)})}{f(\beta_j^{(k-1)})p(\beta_j^{(k)}|\beta_j^{(k-1)})}\right\}.$$

The common choices of candidate distribution include normal distribution and an approximation to the full conditional distribution. More details can be seen in Daniels and Hogan (2008).

Combination of Gibbs sampler and M–H algorithm is also researched in Tang and Zhao (2014), they considered a hybrid sampling procedure combining the Gibbs sampler and M–H algorithm for Bayesian estimation.

Data augmentation (DA) (Tanner & Wong, 1987) is an alternative to M–H algorithm when the full conditional distributions are difficult to sample. DA introduces latent data $\boldsymbol{Z}$ which enables sampling $f(\boldsymbol{\beta}|\boldsymbol{z}, y)$ and $f(\boldsymbol{z}|\boldsymbol{\beta}, y)$ easily. This approach firstly samples $f(\boldsymbol{z}|\boldsymbol{\beta}, y)$, and then samples $f(\boldsymbol{\beta}|\boldsymbol{z}, y)$ using Gibbs sampler. DA is also a natural way for dealing with missing data. In incomplete data cases, latent data $\boldsymbol{Z}$ will denote the missing data, and sampling $f(\boldsymbol{\beta}|\boldsymbol{z}, y)$ and $f(\boldsymbol{z}|\boldsymbol{\beta}, y)$ corresponds to sampling from the complete data posterior and posterior predictive distribution of the missing data, respectively.

The above MCMC methods may give misleading answers before they converge to the stationary distribution. Therefore, assessing convergence is a key step in implementing the MCMC methods. Comprehensive introduction of the convergence diagnostics can be seen in Little and Rubin (2002) and Mengersen, Robert, and Guihenneuc (1999). Here we briefly introduce the most popular convergence statistics called potential scale reduction (PSR) statistic proposed by Gelman and Rubin (1992). For calculating the PSR statistic, we should firstly generate L parallel chains with starting points dispersed over the parameter space, and then calculate the variance between the sequence means B and the average of the within-sequence variances W for each quantity of interest. After that, by a weighted average of B and W, we can estimate the posterior variance $\hat{V}$, and finally, the PSR statistic can be calculated as

$$\sqrt{\hat{R}} = \sqrt{\hat{V}/W},$$

where $\hat{V} = B/K + W(K-1)/K$ is the posterior variance and W is the within-sequence variance. When the value of PSR is near 1.0 for all quantities of interest, we can say that the chains have already converged to the stationary distribution.

5.3. Model comparison

One issue in sensitivity analysis is how to select among several alternative models. Here we present some common criteria for model comparison, including deviance information criterion (DIC), L-measure, Bayes factor, the conditional predictive ordinate (CPO), and LPML.

The DIC proposed by Spiegelhalter, Best, Carlin, and Van Der Linde (2002) is the most popular criterion to compare various competing models in the Bayesian framework since it is easily obtained from the observations simulated by the MCMC algorithm (Cai et al., 2010; Das et al., 2008; Kyoung & Lee, 2015; Lu et al., 2011; Pettitt et al., 2006; Yu et al., 2013). It is a model-based criterion composed of a goodness of fit term and a penalty term. Let θ denote the parameter related to data y, then the overall fit of the model is defined as the deviance, a linear function of the log likelihood, given by

$$Dev(\theta) = -2\log L(\theta|y).$$

The penalty term is given by

$$p_D = E\{Dev(\theta)|y\} - Dev\{E(\theta|y)\}.$$

Then the DIC is defined as

$$DIC = Dev\{E(\theta|y)\} + 2p_D. \tag{28}$$

In the presence of non-ignorable missing data, DIC should be reconstructed due to taking account of the missingness mechanism. Daniels and Hogan (2008) discussed two different constructions for selection models that based on the observed data likelihood and the full-data likelihood respectively.

Ibrahim, Chen, and Sinha (2001) introduced L-measure criterion for measuring the adequacy of a given model. Let $\boldsymbol{z} = (z_1, \ldots, z_n)'$ denote a future response vector with the same sampling density as $f(y|\boldsymbol{\beta})$. L-measure is defined as:

$$L(y) = \sum_{i=1}^{n} Var(z_i|y_i) + v\sum_{i=1}^{n}(\mu_i - y_i)^2, \tag{29}$$

where $\mu_i = E(z_i|y_i)$ and $0 < v < 1$. v can be interpreted as a weight term in the formula, and it has an impact on the ordering of the models as well as characterizing the properties of L-measure. Ibrahim et al. (2002) proposed a calibration form of L-measure. Chen, Dey, and Ibrahim (2004) proposed the weighted L-measure and Huang et al. (2005) extended it to accommodate GLMs with missing covariates.

Following Berge (1985), Bayes factor for comparing two competing models M_0 and M_1 is given by:

$$B_{10} = \frac{f(Y_{(1)}, \boldsymbol{X}_{(1)}, \boldsymbol{R}|M_1)}{f(Y_{(1)}, \boldsymbol{X}_{(1)}, \boldsymbol{R}|M_0)}, \tag{30}$$

where $f(Y, \boldsymbol{X}, \boldsymbol{R}|M_i)$ is the margin probability distribution of model M_i based on the observed data for $i = 0, 1$. The path sampling idea can be used to evaluate $f(Y, \boldsymbol{X}, \boldsymbol{R}|M_i)$ as it is difficult to derive this density function directly. Details can be seen in Tang and Zhao (2014). Application of Bayes factor as model comparison criterion can also be seen in Lee and Song (2004) and Lee and Tang (2006).

The CPO statistic is a useful tool for model comparison (Chen et al., 2002; Ibrahim et al., 2001; Tang and Zhao, 2014). Let $D_{obs}^{(-i)}$ be the observed dataset $D_{obs} = (N, Y_{(1)}, \boldsymbol{X}_{(1)})$ with the ith observation deleted. For the ith observation, define the CPO statistic as $CPO_i = f(Y_{(1)i}|X_{(1)i}, D_{obs}^{(-i)})$. A larger value of CPO statistic indicates a better fit of the model. Chen et al. (2004) proposed a new definition of the CPO statistic in the presence of missing covariates.

A summary statistic for measuring the plausibility of a model is LPML, also named as pseudo-Bayes factor, which is given by:

$$LPML = \sum_{i=1}^{N} \log(CPO_i). \tag{31}$$

Similarly, a larger value of LPML means a better fit of the model.

Although all of these criteria can be used for model comparison, there is no final conclusion about which criterion is the best and the most reliable in practice. Efforts have been paid in discovering the relationship between these criteria. For example, Chen, Huang, Ibrahim, et al. (2008) explored the theoretical and computational connections between these criteria for model selection in GLM settings. They found that under conjugate priors, these criteria are quite similar in terms of model selection, especially under small values of the prior parameters. These criteria were also compared in Chen and Kim (2008) for selecting constrained ANOVA models. They discovered that the Bayes factor is extremely more sensitive to the specification of the prior distributions of model parameters than the other criteria since it suffers from the Bartlett's or Lindley's paradox, which means that the other criteria are more robust to the specification of prior distributions. As a result, other criteria do not require proper priors while Bayes factor does. DIC and LPML usually perform similarly and agree with each other, while Bayes factor may give different answers. In addition, L-measure indicates how well a model fits the data through the posterior predictive variance and bias, while the dimensional penalty term in DIC is regarded as a measure of model complexity. In computational point of view, L-measure, DIC, CPO and LPML are much easier to be computed than Bayes factor.

5.4. Sensitivity analysis

How a model fits to the observed data can be assessed, while its fit to the unobserved data given the observed data cannot be assessed (Mason, 2010). As a result, sensitivity analysis becomes a crucial part in Bayesian framework because of the inability to distinguish the real missing data mechanism (Molenberghs & Kenward, 2007). Draper (1995) pointed out that in the case of incomplete data, there are parametric and structural uncertainty in the models, so it is important to take the uncertainty into account.

Mason et al. (2010) concluded two types of sensitivity analysis, an assumption sensitivity and a parameter sensitivity. For the assumption sensitivity, several alternative models should be explored by changing the key assumptions. For example,

specifying different prior distributions, response models, error distributions, covariate distributions, or missingness models to fit the incomplete data and proceed model comparison (Das et al., 2008; Lee & Song, 2004; Yuan & Yin, 2010). For the parameter sensitivity, running the missingness model with parameters controlling the extent of departure from MAR fixed to values in a plausible range. The parameters controlling the extent of departure from MAR are usually called sensitivity parameters. The construction of sensitivity parameters can be seen in Daniels and Hogan (2008). Specifically for the PMM framework, Daniels and Hogan (2008) examined several global and local sensitivity methods in Bayesian analysis especially for the extrapolation factorization approach. Kaciroti and Raghunathan (2009) proposed a Bayesian parameterization using a PMM approach, which measured the difference between the distributions of the missing data from that of the observed data, for sensitivity analysis. This parameterization also allows for the translation between PMMs and SMs. More recently, Zhu, Ibrahim, and Tang (2011) developed a general framework of Bayesian analysis for assessing different perturbation schemes to the data, and Zhu, Ibrahim, and Tang (2014) developed a Bayesian perturbation manifold and performed sensitivity analysis using intrinsic influence measures.

The assumptions of response model and missingness model are the crucial parts to be checked in sensitivity analysis. However, nonparametric response model and missingness model that weaken model assumptions have become more popular in the literature. A comprehensive review of Bayesian nonparametric approaches for longitudinal data under non-ignorable missingness can be seen in Daniels and Linero (2015). In nonparametric modeling, we can find sensitivity parameters in the missingness mechanism and specify informative priors on them to make sure fitting models to the observed data will not be affected (Daniels & Hogan, 2008). For example, Scharfstein, Daniels, and Robins (2003) presented a fully Bayesian method by incorporating prior beliefs about non-identifiable selection bias parameters under a univariate continuous missing response circumstance. For longitudinal binary missing responses, Wang, Danies, Scharfstein, et al. (2010) proposed a Bayesian shrinkage approach to incorporate expert opinion about non-identifiable parameters. In addition, the Bayesian nonparametric framework presented in Linero and Daniels (2015) and Linero (2017) allowed the introduction of sensitivity parameters to vary the untestable assumptions about the missingness mechanism.

6. Other related topics

The above review of Bayesian methods for dealing with missing data is mainly about the application of different frameworks and models in various settings. Some other related topics include dealing with missing categorical data in contingence tables, incorporating information from similar studies, and improving the robustness of the estimators.

Contingence table is used for displaying the frequency distribution of the variables in a matrix form, usually used in survey researches. In contingence tables, unit nonresponse and item nonresponse may result in partial classification. ML estimates calculated from the data table may suffer from the problem of instability due to boundary solutions, so Bayesian approach is a good alternative for dealing with these problems. Green and Park (2003) developed a Bayesian hierarchical model with a log-linear model in the prior specification. Nandram, Han, and Choi (2002) proposed two hierarchical models, under ignorable and non-ignorable missingness respectively, to analyze count data from several areas in one-way tables, based on which Nandram, Cox, and Choi (2005) considered several two-way categorical tables and developed a method to study the association between the categorical variables. Nandram, Liu, Choi, and Cox (2005) extended to a general $r \times c$ categorical table with partial classification and proposed a Bayesian approach that allowed the missingness to be ignorable or non-ignorable, and a Bayes factor was used for model comparison. Other types of data, including binomial, ordinal and repeated measured data, can also be considered in the contingence table with missingness settings.

In Bayesian framework, data combination is natural and easy to implement through prior distributions. Information from other related sources can help improve the performance of the estimators, so it has become a popular area in the literature. For example, Jackson, Best, and Richardson (2006) combined aggregate-level and individual-level data from related sources by carrying out simultaneous regressions with common coefficients on data from two levels to improve inferences. Bayesian graphical model was employed by Molitor, Best, Jackson, and Richardson (2009) to analyze multiple data sources in which sub-models were linked by shared parameters, while Raghunathan et al. (2007) used a hierarchical Bayesian approach to consider multiple data sources and showed that the combined estimation procedure can help improve the performance of the estimates.

The robustness of the estimates is a common problem in missing data problems. Nonparametric modeling is a popular way to improve the robustness of the estimators as parametric assumptions are relaxed in nonparametric settings. In SM framework, Chen et al. (2002) employed splines in both the response model and the missingness model, while Daniels and Hogan (2008) and Su and Hogan (2008) used splines in response models within PMM framework. Rizopoulos and Ghosh (2011) used nonparametric random effects in SPM framework. Auxiliary information can also be used to improve robustness. For example, Daniels, Wang, and Marcus (2014) made inference of the marginal distributions of the auxiliary covariates using the saturated multinomial approach for ignorable missing data in Bayesian framework. However, although nonparametric approaches are more robust to model specification and estimation than a parametric model, sensitivity analysis and model assessment still should be thought highly of.

7. Conclusions and future issues

This paper is a review of recent developments and applications of Bayesian methods for dealing with missing data. We firstly give a brief introduction of ignorable and non-ignorable missing data mechanisms, as well as the Bayesian

framework for dealing with missing data. Then according to the inference procedure, missing data models under different missing data settings are reviewed. Following is some key issues of Bayesian inference, including prior construction, posterior computation, model comparison, and sensitivity analysis. We introduce how researchers use these structures and frameworks to analyze data of various types and under different missing data settings. We also briefly introduce several related topics. Summarizing from the existing researches, several future issues are concluded as follows.

The robustness of results and sensitivity analysis. The robustness of results when different parts of the missing data model are incorrectly specified is a common problem in missing data analysis, which also happens in Bayesian framework. In particular, when the error distribution of response model or the form of missingness model is misspecified, the performance of Bayesian methods will negatively affected (Mason, 2010). In addition, when the model is weakly identifiable, the inferences will be sensitive to the choices of the hyperparameters. As a result, sensitivity analysis about these key assumptions should be performed and informative priors elicited from external information or expert knowledge can be used to help solve this problem (Molenberghs, Fitzmaurice, Kenward, Tsiatis, & Verbeke, 2014). Nonparametric modeling and auxiliary information can also be developed to improve the robustness of the estimates. Incorporating Bayesian approach with MI or other methods can also be an alternative.

More complex data structures. When response variables or covariates are multiple and of mixed types, the complexity of the analysis increase rapidly due to the correlations between the variables. Different methods should be applied to deal with these complex data. For example, when dealing with mixed correlated ordinal and count data, factorization or latent variables can be used. When dealing with multiple responses in longitudinal studies, marginal modeling, random-effects models and Markov transition models can be employed (Samani & Ganjali, 2014). In addition, when measurement error, heterogeneity, skewness, censored or other features exist in missing variables, attention should be paid to fit more suitable parts of the models. For example, as in Huang (2016), a measurement error model was built as covariate distribution when the missing covariates suffer from measurement error. More efforts should be done to take different situations into account to achieve more reliable results. Also, special attention should be paid to deal with longitudinal data, spatial data, survival data and multilevel data, especially in the case of high-dimension.

Quantile regression with MNAR missingness in Bayesian framework. Most Bayesian approaches for missing data in the literature mainly focus on mean regression and few builds a model in the form of QR. With QR, the impact of covariates on quantiles can also be investigated and the results will be more robust since QR does not require the assumptions about the error distributions. The existing researches that employing QR for dealing with missing data in Bayesian framework do not allow for sensitivity parameters, which needs further developments (Molenberghs et al., 2014). Composite quantile regression (CQR) is a new extension of QR which can improve the efficiency and robustness of the estimates even for non-normal errors. Bayesian methods incorporating with CQR to analyze incomplete data also need further research (Samani & Ganjali, 2014).

Specialized MCMC sampling algorithms. Complex statistical models for large datasets do not run quickly or the MCMC may not converge easily in the existing available software like WinBUGS. The current capability of existing software limits the scope for easily implementing complex models that incorporate multiple correlated missing covariates of mixed types (Mason, 2010). As a result, specified MCMC sampling algorithms need further investigation to adapt for complex and large datasets.

Acknowledgment

This work is supported by Chinese National Program for Support of Top-notch Young Professionals [grant number 2015338].

References

Ahmed, M. R. (2011). *An investigation of methods for missing data in hierarchical models for discrete data*. (Ph.D. thesis), Canada: University of Waterloo.
Berger, J. O. (1985). Prior information and subjective probability. In *Statistical Decision Theory and Bayesian Analysis* (pp. 74–117). New York: Springer.
Cai, J. H., Song, X. Y., & Hser, Y. I. (2010). A Bayesian analysis of mixture structural equation models with non-ignorable missing responses and covariates. *Statistics in Medicine*, *29*, 1861–1874.
Carlin, B. P., & Louis, T. A. (1997). Bayes and empirical Bayes methods for data analysis. *Statistics and Computing*, *7*, 153–154.
Carrigan, G., Barnett, A. G., Dobson, A. J., & Mishra, G. (2007). Compensating for missing data from longitudinal studies using WinBUGS. *Journal of Statistical Software*, *19*, 1–17.
Chen, M. H., Dey, D. K., & Ibrahim, J. G. (2004). Bayesian criterion based model assessment for categorical data. *Biometrika*, *91*, 45–63.
Chen, M. H., Huang, L., Ibrahim, J. G., et al. (2008). Bayesian variable selection and computation for generalized linear models with conjugate priors. *Bayesian Analysis*, *3*, 585–614.
Chen, M. H., & Ibrahim, J. G. (2001). Maximum likelihood methods for cure rate models with missing covariates. *Bioemtrics*, *57*, 43–52.
Chen, M. H., Ibrahim, J. G., & Lipsitz, S. R. (2002). Bayesian methods for missing covariates in cure rate models. *Lifetime Data Analysis*, *8*, 117–146.
Chen, M. H., & Kim, S. (2008). The Bayes factor versus other model selection criteria for the selection of constrained models. *Statistics for Social & Behavioral Sciences*, *15*, 5–180.
Chen, Q., & Ibrahim, J. G. (2014). A note on the relationships between multiple imputation, maximum likelihood and fully Bayesian methods for missing responses in linear regression models. *Statistics and its Interface*, *6*, 315.
Daniels, M. J., & Hogan, J. W. (2008). *Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis*. New York: CRC Press.
Daniels, M. J., & Linero, A. R. (2015). Bayesian nonparametrics for missing data in longitudinal clinical trials. In *Nonparametric Bayesian inference in biostatistics* (pp. 423–446). Springer.

Daniels, M., Wang, C., & Marcus, B. (2014). Fully Bayesian inference under ignorable missingness in the presence of auxiliary covariates. *Bioemtrics*, *70*, 62–72.
Das, S., Chen, M.-H., Kim, S., & Warren, N. (2008). A Bayesian structural equations model for multilevel data with missing responses and missing covariates. *Bayesian Analysis*, *3*, 197–224.
Deyoreo, M., Reiter, J. P., & Hillygus, D. S. (2016). Bayesian mixture models with focused clustering for mixed ordinal and nominal data. *Bayesian Analysis TBA*, 1–25.
Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society. Series B. Methodology*, *4*, 5–97.
Erler, N. S., Rizopoulos, D., Rosmalen, J., et al. (2016). Dealing with missing covariates in epidemiologic studies: A comparison between multiple imputation and a full Bayesian approach. *Statistics in Medicine*, *35*, 2955–2974.
Garthwaite, P. H., Kadane, J. B., & O'hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, *100*, 680–701.
Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, *7*, 457–511.
Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *72*, 1–741.
Green, P. E., & Park, T. (2003). A bayesian hierarchical model for categorical data with non-ignorable nonresponse. *Bioemtrics*, *59*, 886–896.
Harel, O., & Zhou, X. H. (2007). Multiple imputation: review of theory, implementation and software. *Statistics in Medicine*, *26*, 3057–3077.
Hastie, T., & Tibshirani, R. (1987). Non-parametric logistic and proportional odds regression. *Applied Statatistics-Journal of the Royal Statistical Society*, 260–276.
Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*, 97–109.
Hong, H., Chu, H., Zhang, J., & Carlin, B. P. (2016). A Bayesian missing data framework for generalized multiple outcome mixed treatment comparisons. *Research Synthesis Methods*, *7*, 6–22.
Huang, L., Chen, M. H., & Ibrahim, J. G. (2005). Bayesian analysis for generalized linear models with nonignorably missing covariates. *Bioemtrics*, *61*, 767–780.
Huang, Y. (2016). Quantile regression-based bayesian semiparametric mixed-effects models for longitudinal data with non-normal, missing and mismeasured covariate. *Journal of Statistical Computation and Simulation*, *86*, 1183–1202.
Ibrahim, J. G., Chen, M. H., & Lipsitz, S. R. (2002). Bayesian methods for generalized linear models with covariates missing at random. *The Canadian Journal of Statistics. La Revue Canadienne de Statistique*, *30*, 55–78.
Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R., & Herring, A. H. (2005). Missing-data methods for generalized linear models: a comparative review. *Journal of the American Statistical Association*, *100*, 332–346.
Ibrahim, J. G., Chen, M.-H., & Sinha, D. (2001). Criterion-based methods for Bayesian model assessment. *Statistica Sinica*, 419–443.
Ibrahim, J. G., Chu, H., & Chen, M.-H. (2012). Missing data in clinical studies: issues and methods. *Journal of Clinical Oncology*, *30*, 3297–3303.
Ibrahim, J. G., & Molenberghs, G. (2009). Missing data methods in longitudinal studies: a review. *Test*, *18*, 1–43.
Jackson, C., Best, N., & Richardson, S. (2006). Improving ecological inference using individual-level data. *Statistics in Medicine*, *25*, 2136–2159.
Kaciroti, N., & Raghunathan, T. E. (2009). Bayesian sensitivity analysis of incomplete data using pattern-mixture and selection models through equivalent parameterization. *Ann Arbor*, *1001*, 48109.
Kaciroti, N. A., Raghunathan, T. E., Schork, M. Anthony., & Clark, N. M. (2008). A Bayesian model for longitudinal count data with non-ignorable dropout. *Applied Statatistics-Journal of the Royal Statistical Society*, *57*, 521–534.
Kaciroti, N. A., Raghunathan, T. E., Schork, M. A., Clark, N. M., & Gong, M. (2006). A Bayesian approach for clustered longitudinal ordinal outcome with non-ignorable missing data: Evaluation of an asthma education program. *Journal of the American Statistical Association*, *101*, 435–446.
Kalaylioglu, Z., & Ozturk, O. (2013). Bayesian semiparametric models for non-ignorable missing mechanisms in generalized linear models. *Journal of Applied Statistics*, *40*, 1746–1763.
Kaplan, D. E. (2000). *Structural equation modeling: foundations and extensions*. Sage Publications.
Kenward, M. G., Molenberghs, G., & Thijs, H. (2003). Pattern-mixture models with proper time dependence. *Biometrika*, *90*, 53–71.
Kim, Y. D., & Choi, S. (2014). Bayesian binomial mixture model for collaborative prediction with non-random missing data. In *Proceedings of the 8th ACM Conference on recommender systems*. ACM.
Knott, M., & Bartholomew, D. J. (1999). *Latent variable models and factor analysis*. Edward Arnold.
Koenker, R. (2005). *Quantile regression*. Cambridge University Press.
Kyoung, Y., & Lee, K. (2015). Bayesian pattern mixture model for longitudinal binary data with non-ignorable missingness. *Communications for Statistical Applications and Methods*, *22*, 589–598.
Lee, K. J., & Simpson, J. A. (2014). Introduction to multiple imputation for dealing with missing data. *Respirology*, *19*, 162–167.
Lee, S. Y., & Song, X. Y. (2004). Bayesian model comparison of nonlinear structural equation models with missing continuous and ordinal categorical data. *British Journal of Mathematical and Statistical Psychology*, *57*, 131–150.
Lee, S. Y., & Tang, N. S. (2006). Bayesian analysis of nonlinear structural equation models with non-ignorable missing data. *Psychometrika*, *71*, 541.
Lee, S. Y., & Zhu, H. T. (2000). Statistical analysis of nonlinear structural equation models with continuous and polytomous data. *British Journal of Mathematical and Statistical Psychology*, *53*, 209–232.
Linero, A. R. (2017). Bayesian nonparametric analysis of longitudinal studies in the presence of informative missingness. *Biometrika*, *104*, 327–341.
Linero, A. R., & Daniels, M. J. (2015). A flexible Bayesian approach to monotone missing data in longitudinal studies with informative dropout with application to a schizophrenia clinical trial. *Journal of the American Statistical Association*, *110*, 45–55.
Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data* (2nd ed.). New York: Wiley.
Liu, G. F., Han, B., Zhao, X., & Lin, Q. (2016). A comparison of frequentist and Bayesian model based approaches for missing data analysis: case study with a schizophrenia clinical trial. *Statistics in Biopharmaceutical Research*, *8*, 116–127.
Lu, Z. L., Zhang, Z., & Lubke, G. (2011). Bayesian inference for growth mixture models with latent class dependent missing data. *Multivariate Behavioral Research*, *46*, 567–597.
Lunn, D., Spiegelhalter, D., Thomas, A., et al. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, *28*, 3049–3067.
Martyn, P. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*.
Mason, A. J. (2010). Bayesian methods for modelling non-random missing data mechanisms in longitudinal studies. In *Technical report*. London: Imperial College.
Mason, A., Best, N., Plewis, I., & Richardson, S. (2010). Insights into the use of Bayesian models for informative missing data. In *Technical report*. London: Imperial College.
Mealli, F., & Rubin, D. B. (2015). Clarifying missing at random and related definitions, and implications when coupled with exchangeability. *Biometrika*, *102*, 995–1000.
Mengersen, K. L., Robert, C. P., & Guihenneuc, J. C. (1999). MCMC convergence diagnostics: a review. *Bayesian Statistics*, *6*, 415–440.
Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, *21*, 1087–1092.

Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A., & Verbeke, G. (2014). *Handbook of missing data methodology*. CRC Press.
Molenberghs, G., & Kenward, M. (2007). *Missing data in clinical studies*. John Wiley & Sons.
Molitor, N. T., Best, N., Jackson, C., & Richardson, S. (2009). Using Bayesian graphical models to model biases in observational studies and to combine multiple sources of data: application to low birth weight and water disinfection by-products. *Journal of the Royal Statistical Society. Series A. Satistics in Society*, *172*, 615–637.
Moltchanova, E., Penttinen, A., & Karvonen, M. (2005). A hierarchical Bayesian birth cohort analysis from incomplete registry data: evaluating the trends in the age of onset of insulin-dependent diabetes mellitus (T1DM). *Statistics in Medicine*, *24*, 2989–3004.
Murray, J. S., & Reiter, J. P. (2016). Multiple imputation of missing categorical and continuous values via Bayesian mixture models with local dependence. *Journal of the American Statistical Association*, *111*, 1466–1479.
Nandram, B., Cox, L. H., & Choi, J. W. (2005). Bayesian analysis of non-ignorable missing categorical data: an application to bone mineral density and family income. *Surv. Methodol.*, *31*, 213.
Nandram, B., Han, G., & Choi, J. (2002). A hierarchical Bayesian non-ignorable nonresponse model for multinomial data from small areas. *Surv. Methodol.*, *28*, 145–156.
Nandram, B., Liu, N., Choi, J. W., & Cox, L. (2005). Bayesian non-response models for categorical data from small areas: an application to BMD and age. *Statistics in Medicine*, *24*, 1047–1074.
Oakley, J. E., & O'hagan, A. (2007). Uncertainty in prior elicitations: a nonparametric approach. *Biometrika*, *94*, 427–441.
Pettitt, A., Tran, T., Haynes, M., & Hay, J. (2006). A Bayesian hierarchical model for categorical longitudinal data from a social survey of immigrants. *Journal of the Royal Statistical Society. Series A. Satistics in Society*, *169*, 97–114.
Poleto, F. Z., Paulino, C. D., Singer, J. M., & Molenberghs, G. (2015). Semi-parametric Bayesian analysis of binary responses with a continuous covariate subject to non-random missingness. *Statistical Modelling*, *15*, 1–23.
Raghunathan, T. E., Xie, D., Schenker, N., Parsons, V. L., Davis, W. W., Dodd, K. W., & Feuer, E. J. (2007). Combining information from two surveys to estimate county-level prevalence rates of cancer risk factors and screening. *Journal of the American Statistical Association*, *102*, 474–486.
Rizopoulos, D., & Ghosh, P. (2011). A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Statistics in Medicine*, *30*, 1366–1380.
Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581–592.
Rubin, D. B. (2008). *Multiple imputation for nonresponse in surveys*. New York: Wiley.
Samani, E. B., & Ganjali, M. (2014). Mixed correlated bivariate ordinal and negative binomial longitudinal responses with non-ignorable missing values. *Coommunications in Statistics - Theory and Methods*, *43*, 2659–2673.
SAS/STAT, 13.2. (2014). User's guide SAS Institute Inc., Cary, NC.
Scharfstein, D. O., Daniels, M. J., & Robins, J. M. (2003). Incorporating prior beliefs about selection bias in the analysis of randomized trials with missing outcomes. *Biostatistics*, *4*, 495.
Seaman, S., Galati, J., Jackson, D., & Carlin, J. (2013). What is meant by" missing at random"?. *Statistical Science*, *25*, 7–268.
Si, Y., & Reiter, J. P. (2013). Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *Journal of Educational and Behavorial Statistics*, *38*, 499–521.
Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *64*, 583–639.
Stan Development Team. (2012). A C++ library for probability and sampling, version 1.0. http://mc-stanorg/.
Su, L., & Hogan, J. W. (2008). Bayesian semiparametric regression for longitudinal binary processes with missing data. *Statistics in Medicine*, *27*, 3247–3268.
Tang, N.-S., & Zhao, H. (2014). Bayesian analysis of nonlinear reproductive dispersion mixed models for longitudinal data with non-ignorable missing covariates. *Communications in Statistics-Simulation and Computation*, *43*, 1265–1287.
Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*, 528–540.
Thijs, H., Molenberghs, G., Michiels, B., et al. (2002). Strategies to fit pattern-mixture models. *Biostatistics*, *3*, 245–265.
Tran, T. T. (2008). *Bayesian model estimation and comparison for longitudinal categorical data*. Queensland University of Technology.
Wang, C., Danies, M. J., Scharfstein, D. O., et al. (2010). A Bayesian shrinkage model for incomplete longitudinal binary data with application to the breast cancer prevention trial. *Journal of the American Statistical Association*, *105*, 1333–1346.
Wang, S., Shao, J., & Kim, J. K. (2014). An instrument variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica*, *24*, 1097–1116.
Xu, D., Daniels, M. J., & Winterstein, A. G. (2016). Sequential BART for imputation of missing covariates. *Biostatistics*, *17*, 589–602.
Yu, F., Chen, M.-H., Huang, L., & Anderson, G. J. (2013). *Hierarchical Bayesian analysis of repeated binary data with missing covariates*. New York: Springer.
Yuan, Y., & Yin, G. (2010). Bayesian quantile regression for longitudinal studies with non-ignorable missing data. *Bioemtrics*, *66*, 105–114.
Zhang, P. (2003). Multiple imputation: theory and method. *International Statistical Review*, *71*, 581–592.
Zhang, Z., & Wang, L. (2012). A note on the robustness of a full Bayesian method for non-ignorable missing data analysis. *Brazilian Journal of Probability and Statistics*, *26*, 244–264.
Zhu, H., Ibrahim, J. G., & Tang, N. (2011). Bayesian influence analysis: a geometric approach. *Biometrika*, *98*, 307–323.
Zhu, H., Ibrahim, J. G., & Tang, N. (2014). Bayesian sensitivity analysis of statistical models with missing data. *Statistica Sinica*, *24*, 871.
Zhu, H. T., & Lee, S.-Y. (2001). A Bayesian analysis of finite mixtures in the LISREL model. *Psychometrika*, *66*, 133–152.
Zhu, J., & Raghunathan, T. E. (2015). Convergence properties of a sequential regression multiple imputation algorithm. *Journal of the American Statistical Association*, *110*, 1112–1124.