Bayesian Nonparametric Analysis of Longitudinal Studies in the Presence of Informative Missingness

Antonio R. Linero

Department of Statistics, Florida State University, 214 Rogers Building, 117 N. Woodward Ave, Tallahassee, Florida, 32306-4330
arlinero@stat.fsu.edu

May 11, 2016

Abstract

In longitudinal clinical trials, one often encounters missingness which is thought to be non-ignorable. It is well-known that non-ignorable missingness introduces fundamental identifiability issues, resulting in intention-to-treat effects being unidentified; the best one can do is to conduct a sensitivity analysis to assess how much of the inference is being driven by missingness. We introduce a Bayesian nonparametric framework for conducting inference in the presence of non-ignorable, non-monotone missingness. This framework focuses on the specification of an auxiliary prior, which we refer to as a working prior, on the space of complete data generating mechanisms. This prior is not used to conduct inference, but instead is used to induce a prior on the observed data generating mechanism, which is then used in conjunction with an identifying restriction to conduct inference. Advantages of this approach include a flexible modeling framework, access to simple computational methods, strong theoretical support, straightforward sensitivity analysis, and its applicability to non-monotone missingness.
Keywords. Bayesian nonparametrics; non-monotone missingness; identifying restrictions; sensitivity analysis; mixture models; posterior consistency.

1 Introduction

In longitudinal clinical trials, the estimation of causal effects, such as intention-to-treat effects, is complicated by missingness in the response variable. Identification of causal effects requires the analyst to invoke untestable assumptions about the missing data. A common approach is to make an identifying assumption such as the missing at random assumption, and assume it holds with no further critique. Because such assumptions cannot be nonparametrically tested they must be justified on a case-by-case basis; hence, in the absence of strong subject-matter knowledge, this approach is inadvisable, and the best one can do is to conduct a sensitivity analysis to determine how much influence the missingness has on inferences.

We propose a Bayesian nonparametric framework for addressing non-ignorable missingness. We argue for the Bayesian nonparametric approach for two reasons. First, the additional flexibility is needed, as it is rare that parametric assumptions are plausible for longitudinal data. In this paper we find in a simulation study that, for the data we consider, parametric models drastically underestimate the uncertainty in estimates. Second, given the inherent identifiability concerns, incorporation of subject-matter expertise is often essential, and we feel that the most natural approach to incorporating this information in a principled manner is through the elicitation of informative priors.

We focus on the setting of non-monotone, or intermittent, missingness. By contrast, much of the existing literature has focused on the case of monotone missingness, where missingness is due to dropout; exceptions include Vansteelandt et al. (2007) and Harel & Schafer (2009). Bayesian nonparametric approaches when missingness is monotone are given by Linero & Daniels (2015) and Wang et al. (2010), with a review given by Daniels & Linero (2015). Non-
monotone missingness has received a lighter treatment in the literature and presents different challenges (National Research Council, 2010). An interesting aspect of this problem is that it has been argued that the popular missing at random assumption should not be expected to hold when missingness is non-monotone (Robins, 1997; Vansteelandt et al., 2007).

This paper makes several methodological contributions. Let the complete data consist of a longitudinal response vector $Y_i = (Y_{i1}, \ldots, Y_{iJ})$ and a binary indicator vector $R_i = (R_{i1}, \ldots, R_{iJ})$ for $i = 1, \ldots, n$ observations, and let $O_i$ denote the observed data for the $i^{th}$ subject. First, we introduce a generic approach to Bayesian analysis in the presence of missing data based on the notion of a working prior $\Pi^*$ on the space of distributions of the complete data $(Y_i, R_i)$. Rather than base inference directly on this prior, we instead use it to induce a prior $\Pi_{\text{obs}}$ on the space of observed data generating mechanisms for $O_i$.

The working prior framework possess several advantages. Because it respects the fact that the distribution of the missing data is unidentified, it forces the analyst to make their assumptions about the missing data precise, and allows for a simple sensitivity analysis. By contrast, specifying a prior $\Pi$ directly on the space of complete data generating mechanisms may inadvertently identify the joint of $(Y_i, R_i)$, confounding the assumptions made about the missingness with assumptions made about the model; this occurs, for example, when one uses parametric selection models (Heckman, 1979; Daniels & Hogan, 2008). The working prior $\Pi^*$ can additionally be used to facilitate computations so that inference is no more cumbersome than if $\Pi^*$ had been used directly. In particular, general purpose Markov-chain Monte Carlo software can be used.

We establish that, under regularity conditions on $\Pi^*$ and the true distribution of the complete data $p_0(y, r)$, posterior consistency is attained for the distribution of the observed data $p_{0,\text{obs}}$. Additionally, given a correctly-specified identifying restriction (Little, 1993; Kenward et al., 2003), we show that the posterior is consistent at the complete data generating mechanism $p_0(y, r)$, and, in particular, we can consistently estimate the intention-to-treat effect. A benefit of our results is that they are at the level of the complete data distribution.
$p_0$ rather than the observed data distribution $p_{0,\text{obs}}$, and as a result we can take advantage of a large number of existing results from the Bayesian nonparametric literature.

We propose a combination of a Bayesian bootstrap (Rubin, 1981) for the distribution of $R_i$ and a Bayesian nonparametric kernel mixture model for the distribution of $[Y_i \mid R_i = r]$ as a working prior. We advocate for this model on the basis of its flexibility and ease of implementation in standard software package such as JAGS (Plummer, 2003). Our framework is, however, broad enough to allow for models beyond the kernel mixture model we propose.

As a second contribution, we introduce a new family of identifying restrictions for continuous data $Y_i$ and argue that it provides a useful basis for a sensitivity analysis. This family assumes that, for each fixed $r$, there exists a transformation $T_r$ such that

$$[Y_{ij} \mid Y_{ir}, R = r] \equiv [T_r(Y_{ij}; Y_{ir}) \mid Y_{ir}, R = r^*_j],$$

where $r^*_j$ denotes the vector $r$ with the $j^{th}$ entry changed from a 0 to a 1 and $Y_{ir} = (Y_{ij} : r_j = 1)$. Here, $\equiv$ denotes equality in distribution. We refer to this identifying restriction as the transform observed data restriction, or $T$-OD. We argue that this family of identifying restrictions is well-suited to the setting where missingness is non-monotone, as it is both easy to explain to subject-matter experts and leads to simple computational algorithms.

We illustrate our methodology on data from a clinical trial designed to assess the impact of several treatments on insomnia, which we refer to as the sleep study. The clinical outcome of interest was wake-after-sleep-onset, defined to be the number of hours a subject was awake after first falling asleep. First, we compare our approach to a suite of multiple imputation techniques in a simulation designed to closely mirror the clinical trial. It is shown that intervals based on the multiple imputation methods have coverage far below the nominal level, while our approach attains close to the nominal coverage level. We then apply our methodology to the data and conduct a sensitivity analysis.

In Section 2, we give the formal framework we use for studying the problem and give
conditions for posterior consistency. In Section 3, we propose a kernel mixture model for use within our framework. In Section 4, we illustrate our methodology on the sleep study data. We conclude in Section 5 with a discussion. Proofs of all results are deferred to the supplementary material.

2 The general framework

2.1 Definitions and notation

Let $\mathcal{Y} \subset \mathbb{R}^J$ denote the response space and $\mathcal{R} = \{0, 1\}^J$ the collection of missing data patterns. We assume that the complete data on the $i^{th}$ subject, $(Y_i, R_i)$, is distributed according to a density $p_0(y, r)$ with respect to the product of Lebesgue measure on $\mathbb{R}^J$ and counting measure on $\mathcal{R}$. We will abuse notation and write, for example, $\int p_0(y, r) \, dy \, dr = 1$, where integration against $dr$ denotes summing over $r \in \mathcal{R}$. Ignoring measure-theoretic subtleties, we will consider priors $\Pi$ on the space of densities with respect to $dy \times dr$,

$$\mathcal{P} = \left\{ p : p \geq 0, \int p \, dy \, dr = 1 \right\}.$$

Define the coarsening

$$C(y, r) = y_j r_j - \infty (1 - r_j),$$

and let $O_i = C(Y_i, R_i), i = 1, \ldots, n$ denote the observed data on the $i^{th}$ subject, with $-\infty$ a placeholder for a missing value (and defining $0 \times \infty = 0$). If $(Y_i, R_i) \sim p$, then $O_i$ has density

$$p_{\text{obs}}(o) = \int p(y, r) \, dy_{\text{mis}},$$

where the integration over those components of $y$ such that $o_j = -\infty$, with the remaining components of $(y, r)$ fixed at the values implied by $o$. We also consider priors $\Pi_{\text{obs}}$ on the
$P_{\text{obs}} = \left\{ p_{\text{obs}} : p_{\text{obs}}(o) = \int p(y, r) \, dy_{\text{mis}}, \quad p \geq 0, \quad \int p \, dy \, dr = 1 \right\}.$

It is common in the literature to make the decomposition $(Y, R) = (Y_{\text{obs}}, Y_{\text{mis}}, R)$ where $Y_{\text{obs}}$ denotes the “observed data” and $Y_{\text{mis}}$ the “missing data.” As noted by Seaman et al. (2013), this notation invites confusion, particularly in the context of identifying restrictions. We give a more precise notation. If $r \in \{0, 1\}^J$ is a binary vector and $y \in \mathbb{R}^J$, we let $y_r = (y_j : r_j = 1)$. For example, the “everywhere” variant of missing at random holds if, for all $r$,

$$[Y_{1-r} \mid Y_r, R = r] \equiv [Y_{1-r} \mid Y_r],$$

where $1$ is a vector of 1’s and $\equiv$ denotes equality in distribution.

### 2.2 The working prior framework

We now introduce our principle tool for prior specification, the working prior. This development is abstract, with Section 3 and Section 4 focusing on concreteness.

**Definition 2.1** (Working prior). Let $\Pi_{\text{obs}}$ denote a prior on $P_{\text{obs}}$. A prior $\Pi^*$ on $P$ is said to be a working prior for $\Pi_{\text{obs}}$ if

$$\Pi_{\text{obs}}(A) = \Pi^*(p_{\text{obs}} \in A).$$

In words, $\Pi_{\text{obs}}$ is the marginal distribution of $p_{\text{obs}}$ under $\Pi^*$.

In practice, one specifies the working prior $\Pi^*$ to derive the prior $\Pi_{\text{obs}}$. Rather than using the posterior $\Pi^*(\cdot \mid O_1, \ldots, O_n)$ for inference, we use a combination of $\Pi_{\text{obs}}(\cdot \mid O_1, \ldots, O_n)$ and an identifying restriction.
Our approach can be compared to directly specifying a prior $\Pi_{\text{obs}}$ on $\mathcal{P}_{\text{obs}}$ or a prior $\Pi$ on $\mathcal{P}$. Indeed, one may wonder what the benefit of taking an indirect approach to prior specification is. Our approach possesses the following benefits. First, by passing to the level of priors on complete data distributions, we can reason about how the prior induces sharing of information across timepoints and missingness patterns. We have found it more difficult to construct $\Pi_{\text{obs}}$ directly in an intuitive manner, while specifying $\Pi$ directly may result in the analyst inadvertently identifying the distribution of $[Y_{1-r} | Y_r, R = r]$. Second, as shown in Section 2.4, the working prior perspective leads to automatic proofs of several desirable properties of the posterior. The working prior perspective can also be used to construct Markov chain Monte Carlo sampling schemes based on $\Pi^*$ via data augmentation. In particular, Markov chain Monte Carlo can be carried out as though $\Pi^*$ was our prior, giving draws of $p$ from the posterior based on $\Pi^*$, which are then reduced to draws $p_{\text{obs}}$ from the posterior based on $\Pi_{\text{obs}}$; an example is given in Section 3.

When reasoning about a working prior, the following interpretation is helpful: when compared to $\Pi_{\text{obs}}$, the working prior $\Pi^*$ tells an alternate story about how the observed data $O_i$ was generated. This interpretation is depicted schematically in Figure 1. The path in solid arrows from $\Pi^*$ to $O_i$ represents the use of $\Pi^*$ in our approach, while the path in dashed arrows represents a latent interpretation of our approach giving rise to a working model $p^*$ and complete data $(Y^*, R)$. We emphasize that we do not base inference on the posterior $\Pi^*(\cdot | O_1, \ldots, O_n)$, but on a combination of the posterior $\Pi_{\text{obs}}(\cdot | O_1, \ldots, O_n)$ and an identifying restriction.

### 2.3 Identifying restrictions

We use identifying restrictions to complete our prior specification.

**Definition 2.2.** An identifying restriction is a mapping

$$h : \mathcal{P}_{\text{obs}} \rightarrow \mathcal{P},$$
such that \( p_{\text{obs}} = h(p_{\text{obs}})_{\text{obs}} \); that is, \( h(p_{\text{obs}}) \) is compatible with \( p_{\text{obs}} \). We call \( h \) a partial identifying restriction if

\[
h : \mathcal{P}_{\text{obs}} \rightarrow \{ A : A \subseteq \mathcal{P} \}
\]

such that, for all \( p' \in h(p_{\text{obs}}) \), one has \( p'_{\text{obs}} = p_{\text{obs}} \). An identifying restriction is said to hold at \( p \) if \( p = h(p_{\text{obs}}) \). A partial identifying restriction is said to hold at \( p \) if \( p \in h(p_{\text{obs}}) \).

The most famous example of an identifying restriction is the missing at random assumption. Partial identifying restrictions are useful for expressing assumptions which do not identify the whole distribution \( p_0 \), but are sufficient to identify whatever effects the analyst is interested in. For example, if for some functional \( \psi(\cdot) \) we have \( \psi(p') = \psi(p_0) \) for all \( p' \in h(p_{0,\text{obs}}) \) then we expect to be able to consistently estimate \( \psi(p_0) \) despite the fact that \( p_0 \) itself is not identified.

A sensitivity analysis is performed by allowing \( h \) to vary. To be useful, \( h \) must vary through a low-dimensional space \( \{ h_\xi : \xi \in \Xi \} \) such that each \( h \) has an interpretation that can be easily communicated to subject matter experts. The parameter \( \xi \) is referred to as a sensitivity parameter.

A sensitivity analysis can be carried out through a tipping-point analysis (National Re-
search Council, 2010), which aims to identify the values of $\xi$ for which substantive conclusions change. Ideally, inferences are stable over the region of $\xi$’s which are clinically plausible. If a final inference is required, one can take a fully-Bayesian approach and elicit an informative prior on $\xi$ to average together the inferences under different $h$’s in a principled fashion; see Daniels & Hogan (2008), Wang et al. (2010), and Linero & Daniels (2015).

It has been argued that the missing at random assumption is not easy to interpret and may be generally inappropriate when missingness is non-monotone (Robins, 1997; Vansteelandt et al., 2007; National Research Council, 2010). For the sleep data, we use the baseline assumption:

\[
[Y_j \mid Y_r, R = r] \equiv [Y_j \mid Y_r, R = r_j^*],
\]  

(1)

where $r_j^*$ is identical to $r$, but with $j$th entry fixed at 1. This assumption is generalized to provide a full class of identifying restrictions by taking $h_\xi$ to correspond to the assumption that $p_0$ is such that the following holds:

\[
[Y_j \mid Y_r, R = r] \equiv [T_r(Y_j; Y_r, \xi) \mid Y_r, R = r_j^*].
\]  

(2)

This identifying restriction can be explained, in words, as follows:

Suppose $A$ and $B$ are two subjects who are equivalent at baseline. Let $r$ be the missingness pattern of $B$. Suppose we observe $A$ at the same times we observe $B$, with the exception of time $j$, where $A$ is observed but $B$ is not. Moreover, $A$ and $B$ have the same response values at the common times they are observed. Then, the response of subject $B$ at time $j$ is stochastically identical to the response of subject $A$ at time $j$, after applying the correction $T_r$ to the response of subject $A$.

We refer to this as the transform observed data assumption, or the $T$-OD assumption. $T$-OD
Figure 2: Graphical depiction of $\mathcal{T}$-OD when $\mathcal{T}_r$ is a location shift for all $r$. Subjects $A$ and $B$ have identical responses up-to time 4. The conditional distribution of $B$’s response at time 4 (red) is a location shift of the distribution of $A$’s response.

is a partial identifying restriction, as it is not sufficient to identify the joint distribution of $[Y_j, Y_k | Y_r, R = r]$; any copula linking $[Y_j | Y_r, R = r]$ to $[Y_k | Y_r, R = r]$ gives a joint distribution which satisfies the restriction. Figure 2 illustrates how $\mathcal{T}$-OD identifies the distribution of a response at time 4 with pattern $r = (1,1,1,0)$ using the distribution of an equivalent subject with pattern $r = (1,1,1,1)$. The transformation $\mathcal{T}_r$ should be chosen such that the interpretation above is easy to understand. In the simplest case, $\mathcal{T}_r$ is a simple location shift, i.e., $\mathcal{T}_r(Y_j; \xi) = Y_j + \xi$, essentially stating that the response of subject $B$ will be $\xi$ units larger on average than the response of subject $A$.

G-computation algorithms (Scharfstein et al., 2014; Robins, 1986) can be used to perform inference under the $\mathcal{T}$-OD assumption. Given a distribution $p_{\text{obs}}$ one can calculate causal effects of interest using Algorithm 1. This algorithm produces draws of $(Y, R)$ from a distribution with the correct marginal distribution for each $Y_j$; hence, it can be used to calculate causal effects which depend only on the marginal distributions of the $Y_j$’s. One can conduct inference by embedding Algorithm 1 inside a Markov chain Monte Carlo algorithm and, at
Algorithm 1 G-Computation algorithm for drawing \((Y, R)\) under the \(T\)-OD restriction.

1. Draw \((Y_R, R)\) by drawing \(O \sim p_{\text{obs}}\).

2. For each \(j\) such that \(R_j = 0\):
   
   (a) Draw \(Y_j \sim p(y_j \mid Y_R, R^*_j)\), where \(R^*_j\) is identical to \(R\), but with \(j^{\text{th}}\) entry 1.
   (b) Set \(Y_j \leftarrow T_R(Y_j; Y_R, \xi)\).

each iteration, calculating the effects of interest by Monte Carlo integration.

2.4 Posterior consistency

We develop some theoretical support for our approach by establishing posterior consistency results. With complete data, one hopes that \(\Pi(A \mid (Y_1, R_1), \ldots, (Y_n, R_n)) \rightarrow 1\) \(p_0\)-almost-surely for all neighborhoods \(A\) of \(p_0\) in some topology. Given the identifiability issues inherent in missing data problems, this type of result cannot be expected to hold in general. We instead focus on two related questions. First, when is it true that

\[
\Pi_{\text{obs}}(A \mid O_1, \ldots, O_n) \xrightarrow{n \to \infty} 1 \quad \text{(almost-surely)}
\]  

for all neighborhoods \(A\) of \(p_{0, \text{obs}}\)? That is, when is the prior \(\Pi_{\text{obs}}\) consistent at \(p_{0, \text{obs}}\)? Second, given that \(\Pi_{\text{obs}}\) is consistent at \(p_{0, \text{obs}}\), does it follow that

\[
\Pi_{\text{obs}}(h(p_{\text{obs}}) \in A \mid O_1, \ldots, O_n) \xrightarrow{n \to \infty} 1 \quad \text{(almost-surely)}
\]  

for all neighborhoods \(A\) of \(p_0\), provided that the identifying restriction \(h\) holds at \(p_0\)?

Within the working prior framework, these questions have satisfying answers. For \(p, p' \in \mathcal{P}\), let \(\|p - p'\|_1 = \int |p - p'| \, dy \, dr\) denote the \(L_1\) norm on \(\mathcal{P}\), let \(K(p\|p') = \int p \log(p/p') \, dy \, dr\) denote the Kullback-Leibler divergence from \(p\) to \(p'\), and let \(K_\epsilon(p) = \{p' : K(p\|p') < \epsilon\}\). Similarly define \(\|p_{\text{obs}} - p'_{\text{obs}}\|, \quad K(p_{\text{obs}}\|p'_{\text{obs}}), \quad K_\epsilon(p_{\text{obs}})\) for \(p_{\text{obs}}, p'_{\text{obs}} \in \mathcal{P}_{\text{obs}}\).
When the neighborhood $A$ in (3) is defined to be a set containing a set of the form

$$W_e(p_{0,\text{obs}}) = \{ p_{\text{obs}} : |E_{p_{0,\text{obs}}}(t(O_i)) - E_{p_{\text{obs}}}(t(O_i))| < \epsilon \},$$

for some bounded, continuous function $t$, then $\Pi_{\text{obs}}$ is said to be weakly consistent. Weak consistency guarantees that we can consistently estimate expectations of the form $E_{p_0}(t(O_i))$ provided that $t$ is a bounded continuous function. In particular, this includes the probability of a given pattern $E_{p_0}(I(R_i = r))$ and, if $Y_{ij} \leq M$ for some known $M$, the observed data means $E_{p_0}(Y_{ij} | R_{ij} = 1) = E_{p_0}(Y_{ij} R_{ij})/E_{p_0}(R_{ij})$.

The key condition for weak consistency to hold is the condition of Schwartz (1965), that $\Pi_{\text{obs}}(K_\epsilon(p_{0,\text{obs}})) > 0$ for every $\epsilon > 0$. The following theorem states that this property can be inferred directly from the working prior $\Pi^*$.

**Theorem 2.3.** Let $p_0 \in \mathcal{P}$ and suppose $(Y_1, R_1), (Y_2, R_2), \ldots$ are independently distributed according to $p_0$. Let $p_{0,\text{obs}}$ be the density of $O_i = C(Y_i, R_i)$. Let $\Pi_{\text{obs}}$ be a prior on $\mathcal{P}_{\text{obs}}$ with working prior $\Pi^*$ on $\mathcal{P}$. Suppose that $\Pi^*(K_\epsilon(p_0)) > 0$ for some $\epsilon > 0$. Then $\Pi_{\text{obs}}(K_\epsilon(p_{0,\text{obs}})) > 0$.

**Corollary 2.4.** Under the assumptions of Theorem 2.3, $\Pi_{\text{obs}}$ is weakly consistent at $p_{0,\text{obs}}$.

The value of this result is that it gives sufficient conditions for posterior consistency of the prior $\Pi_{\text{obs}}$ in terms of standard sufficient conditions for posterior consistency of $\Pi^*$. This immediately gives posterior consistency for a large library of existing priors.

While Corollary 2.4 is sufficient to obtain posterior consistency for many quantities of interest, it is too weak to give guarantees about expectations with respect to $p_0$. For example, it does not make any guarantees about the posterior of $E_p(Y_j)$. Due to technical issues concerning weak convergence of joint and conditional distributions, even if we assume $h$ is correctly specified it is inconvenient to proceed with weak convergence (Sethuraman, 1961). To sidestep the technical problems of weak convergence, we instead focus on strong consistency. The prior $\Pi_{\text{obs}}$ is said to be strongly consistent if the set $A$ in (3) contains a set
of the form
\[ S_\epsilon(p_{0,\text{obs}}) = \{ p_{\text{obs}} : \| p_{0,\text{obs}} - p_{\text{obs}} \| < \epsilon \} . \]

Similar to Corollary 2.4, the conditions for strong consistency can be expressed in terms of \( \Pi^* \) so that we can take advantage of a large number of existing results in the Bayesian nonparametric literature.

**Theorem 2.5.** Let \( p_0, p_{0,\text{obs}}, \Pi^*, \Pi_{\text{obs}}, Y_i, R_i, \text{and } O_i \) be as in Theorem 2.3. Suppose that \( \Pi^* \) and \( p_0 \) satisfy the regularity conditions A1–A3 in the supplementary material. Then \( \Pi_{\text{obs}} \) is strongly consistent at \( p_{0,\text{obs}} \).

The regularity conditions A1–A3 are typical conditions in the Bayesian nonparametric literature. Unlike weak consistency, we can prove the following result for strong consistency.

**Theorem 2.6.** Suppose that \( \Pi_{\text{obs}} \) is strongly consistent at \( p_{0,\text{obs}} \) and that the identifying restriction \( h(p_{\text{obs}}) \) is correctly specified. Suppose further that \( h \) is continuous at \( p_{0,\text{obs}} \) with respect to \( \| \cdot \|_1 \) as a mapping \( h : P_{\text{obs}} \to P \). Then, for every \( \epsilon > 0 \),
\[ \Pi_{\text{obs}} (p_{\text{obs}} : \| p_0 - h(p_{\text{obs}}) \|_1 < \epsilon \mid O_1, \ldots, O_n) \to 1 \]
p_{0,\text{obs}}-almost-surely as \( n \to \infty \). Consequently, for any bounded measurable function \( \psi(y, r) \),
\[ \Pi_{\text{obs}} \left( p_{\text{obs}} : \left| \int \psi(y, r) h(p_{\text{obs}})(y, r) \, dy \, dr - \int \psi(y, r) p_0(y, r) \, dy \, dr \right| > \epsilon \mid O_1, \ldots, O_n \right) \to 0. \]

The condition that \( h(p_{\text{obs}}) \) is continuous at \( p_{0,\text{obs}} \) is mild, and will be satisfied by any reasonable choice of \( h \), given suitable conditions on \( p_{0,\text{obs}} \). In the monotone setting, Linero (2015) shows that continuity holds for a wide class of non-future-dependent models (Kenward et al., 2003) and models \( p_{0,\text{obs}} \) including the missing at random assumption.

Theorem 2.5 does not cover partial identifying restrictions such as the \( T \)-OD assumption. For any reasonable partial identifying restriction \( h \) used in practice, one can expect that when
\(\Pi_{\text{obs}}\) is strongly consistent then all effects which are identified under \(h\) will be consistently estimable. The following theorem addresses \(T\)-OD in particular.

**Theorem 2.7.** Suppose that \(\Pi_{\text{obs}}\) is strongly consistent at \(p_{0,\text{obs}}\), and that \(p_0\) satisfies the \(T\)-OD assumption (2) with \(T_r\) continuously differentiable and monotonically increasing. Suppose that, under \(p_0\),

\[
[Y_r \mid R = r] \quad \text{and} \quad [Y_r \mid R = r_j^*]
\]

have the same support, and for all patterns \(r\), if \([R = r]\) has positive probability then \([R = r_j^*]\) does as well. Let \(q_j(p_{\text{obs}})\) denote the marginal distribution of \(Y_j\) under \(p_{\text{obs}}\) and the \(T\)-OD restriction. Then \(q_j(\cdot)\) is well-defined, and for every \(\epsilon > 0\),

\[
\Pi_{\text{obs}}(p_{\text{obs}} : \|q_j(p_{\text{obs}}) - p_{0j}\|_1 > \epsilon \mid O_1, \ldots, O_n) \xrightarrow{n \to \infty} 0,
\]

\(p_{0,\text{obs}}\)-almost-surely, where \(p_{0j}\) is the marginal of \(Y_j\) under \(p_0\). In particular, we can consistently estimate \(E_{p_0}(t(Y_j))\) for any bounded, measurable, \(t(\cdot)\).

### 3 Nonparametric mixture models

We propose a kernel mixture model as a working prior in the presence of non-monotone missingness. This model is computationally tractable and is shown to have theoretical support in Section 2.4. Let \(p(y, r) = \phi(r)g(y \mid r)\) be the pattern-mixture factorization of \(p \in \mathcal{P}\) (Daniels & Hogan, 2008). We model \(R\) with a Dirichlet process prior (Ferguson, 1973),

\[
\phi(r) \sim \mathcal{D}(cH_R).
\]

For the sleep data, we use the Bayesian bootstrap (Rubin, 1981), which is obtained from the limiting posterior as \(c \to 0\), to give a noninformative prior on \(\phi\).
Our working prior for \( g(y \mid r) \) places probability 1 on mixtures of Gaussian distributions

\[
g(y \mid r) = \sum_{k=1}^{K} \pi_k(r) \mathcal{N}(y \mid \mu_k, \Sigma).
\]  \hspace{1cm} (5)

with

\[
\bar{\pi}_k(r) = \gamma_k \prod_{j=1}^{J} \theta_{kj}^{r_j} (1 - \theta_{kj})^{1-r_j}, \quad \pi_k(r) = \frac{\bar{\pi}_k(r)}{\sum_{\ell=1}^{K} \bar{\pi}_\ell(r)}.
\]

To achieve a suitably flexible model, we choose \( K \) to be a slowly-growing function of \( n \), such as \( K = \lceil \sqrt{n} \rceil \). To be more nonparametric, one could use an infinite mixture (Pati et al., 2013) or place a prior on \( K \) (Norets & Pelenis, 2014), both resulting in excellent theoretical properties. Within mixture component, we set \( \mu_k \sim H_{\mu} \) and \( \Sigma \sim G \) independently. In our illustrations, we take \( H_{\mu} \) to be a Gaussian distribution with mean \( m \) and covariance \( \Omega \) where \((m, \Omega)\) is chosen to correspond to an AR-1 process,

\[
b_{k1} \sim \mathcal{N}(0, \sigma_b^2),
b_{kj} \sim \mathcal{N}(\rho^{t_j - t_{j-1}} b_{k(j-1)}, \sigma_b^2 \{1 - \rho^2(t_j - t_{j-1})\}) \quad j = 2, \ldots, J,
\]

\[
\mu_{kj} = m_j + b_{kj},
\]

where \( \rho \in [0, 1] \) and \( t_j \) denotes the time at which \( t_j \) is observed. The prior \( G \) is chosen so that \( \Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_J^2) \) with \( \sigma_j^{-2} \sim \text{Gam}(a_\sigma, b_\sigma) \). Lastly, we set \( \gamma \sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K) \) and \( \theta_{kj} \sim \text{Beta}(a_\theta, b_\theta) \).

Numerous modifications of this model are possible. For example, one might include a linear term \( r^T \beta \) to further adjust for \( R \), or allow a more flexible model for \( \Sigma \) which possibly varies across mixture components; see Quintana et al. (2016) for possibilities. For the sleep data, these modifications do not improve the performance of the model, and require more computationally intensive Markov chain Monte Carlo methods to fit.

The model features hyperparameters \((m, \sigma_b^2, \rho, a_\sigma, b_\sigma, a_\theta, b_\theta, \alpha)\). We standardize the data
so that the grand mean and variance of the observed data, over all time points, are 0 and 1, respectively. For the sleep data, we set $a_\theta, b_\theta \sim \text{Gam}(2, 1)$ and $\rho \sim \text{Uniform}(0, 1)$ and complete the specification by setting $\sigma_b \sim \text{Uniform}(0, 3)$, $m \sim \mathcal{N}(0, 10^2)$, and $\alpha^{-1} \sim \text{Gam}(1, 1)$.

With the exception of the G-computation step, the model can be fit easily using general purpose software; for example, JAGS is used to fit our model in all of our illustrations, and mixing for all quantities of interest was very good. The use of a finite mixture allows us to directly sample $p_{\text{obs}}$ at each iteration, so that we can implement the G-computation algorithm of Section 2.3. Given a sample of $\vartheta = (\phi, \gamma, \mu, \Sigma, \theta)$, we have

$$p_{\text{obs}}(o) = \phi(r) \times \sum_{k=1}^{K} \pi_k(r) \mathcal{N}(y_r | \mu_{kr}, \Sigma_{rr})$$

where $\mu_{kr} = (\mu_{kj} : r_j = 1)$ and $\Sigma_{rr}$ is the submatrix of $\Sigma$ from the indices with $r_j = 1$. To accomplish Step 1 of Algorithm 1, we sample from $p_{\text{obs}}$ by first sampling $(Y, R) \sim \phi(r) \times g(y \mid r)$ and discarding the components of $Y$ with $R_j = 0$. To accomplish Step 2, we sample from the normal mixture

$$g(y_j \mid Y_R, R^*_j) = \sum_{k=1}^{K} \omega_k(R^*_j, Y_r) \times \mathcal{N}(y_j \mid \mu_{j|R}, \Sigma_{j|R}),$$

where $\omega_k(r^*_j, y_r) \propto \pi_k(r^*_j) \mathcal{N}(y_r \mid \mu_{kr}, \Sigma_{rr})$ and $\mu_{k,j|R}$ and $\Sigma_{j|R}$ are the conditional mean and variance of $Y_j$ given $Y_r = y_r$ in mixture component $k$.

There exists a large literature giving conditions on the data generating mechanism $p_0$ and the prior $\Pi^*$ under which strong consistency holds for various kernel mixture models. Relevant for us are the results of Norets & Pelenis (2014) to handle the conditional specification of $g(y \mid r)$, with Shen et al. (2013) giving tools for handling flexible models for $\Sigma$. These results are immediately applicable to various extensions and modifications of the model of this section. For example, under suitable regularity conditions, we can attain consistency for a large class of Dirichlet process mixture models (Escobar & West, 1995); technical details are
given by Linero (2015). We note that, due to the form of \( \pi(r) \) chosen, these results require modification before they cover the model presented here; however, the minor modification \( \pi_k(r) \propto \gamma_k e^{-\|r - \theta_k\|^2} \) is covered by Norets & Pelenis (2014) and is just as effective.

4 Application to the sleep study

4.1 Simulation study and model fit

For the sleep study, we have \( n \approx 125 \) observations for each of 5 treatments for insomnia. We consider observations collected at baseline and 7, 14, 21, 28 and 34 days from baseline. The response was the wake-after-sleep-onset (WASO) score, defined to be the number of minutes a subject was awake after first falling asleep as reported by the subject the day after.

Due to its complexity, a major concern is whether our approach produces better inferences about the effects of interest. Even if simpler models do not fit the data well, inferences might be robust to deviations of the data generating mechanism from the assumed model. To assess whether simpler methods might be adequate for the sleep data, we conducted a simulation study. The goal is to estimate the expectation \( E(Y_1) \). To create a setting in which the ground-truth is known, we combined all treatments and removed the data for which \( R_1 = 0 \), conveniently resulting in \( n = 126 \) observations. We then generated simulated datasets by bootstrap sampling from this dataset, and introduced missingness into the datasets by setting \( R_1 = 0 \) with probability 0.7, giving the same rate of missingness at baseline as the actual data.

Because missingness for \( R_1 \) in the simulated datasets is known to be missing at random, we considered the model of Section 3 with dependence on \( R \) removed to give a model for the marginal distribution of \( Y \). This was compared to three multiple imputation approaches (Carpenter & Kenward, 2012) which make the missing at random assumption. First, we considered multiple imputation using the Gaussian model \( Y \sim \mathcal{N}(\mu, \Sigma) \). Second, we considered chained imputation (Azur et al., 2011) with imputations done by predictive mean matching.
This approach imputes missing data by generating linear predictions for the missing data and randomly choosing an imputation value from among the observed values with similar predicted means; see Little (1988). Lastly, we considered chained imputation using random forests with the method of Doove et al. (2014). For the chained imputation methods, 20 imputed datasets were generated with 10 Gibbs-sampling iterations per dataset; results are robust to these choices. Rubin’s rules (Rubin, 1987) were used to combine inferences across datasets.

We replicated this simulation 1000 times, and another 1000 times with Gaussian noise with standard deviation 0.3 added to each response. Results are reported in Table 1. All methods produced reasonable point estimates of the mean, with bias accounting for a negligible amount of the mean-squared error. The multiple imputation approaches, however, severely underestimated the uncertainty in their estimates, with coverage of each method’s confidence interval being well below the nominal coverage of 95%. Using predictive mean matching produced the most competitive results, while the methods using multivariate normality or random forests performed very poorly. The nonparametric Bayes method produced intervals which are near the nominal coverage level.

We conclude that, while the multiple imputation procedures can be expected to produce accurate point estimates, it is likely that they severely underestimate the uncertainty in the sleep data. The Bayesian nonparametric approach provides an attractive alternative, as it does not appear to severely underestimate the uncertainty in the parameters of interest.

### 4.2 Heterogeneity

A common feature of longitudinal studies is a marked heterogeneity among subjects in the underlying population. A benefit of the Bayesian nonparametric approach we take is that it also functions as a latent class model, and can be used to assess heterogeneity in the data. Assignments of observations to latent classes are produced by the Markov chain Monte Carlo algorithm used to fit the model. Figure 3 displays different latent classes found in treatment
<table>
<thead>
<tr>
<th>Method</th>
<th>RMSE</th>
<th>ŠE</th>
<th>SE</th>
<th>Coverage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No noise</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bayesian nonparametric</td>
<td>1.00</td>
<td>0.932</td>
<td>0.973</td>
<td>92.4</td>
</tr>
<tr>
<td>Random forest</td>
<td>0.96</td>
<td>0.684</td>
<td>0.946</td>
<td>84.2</td>
</tr>
<tr>
<td>Normal</td>
<td>0.99</td>
<td>0.645</td>
<td>0.984</td>
<td>77.9</td>
</tr>
<tr>
<td>Predictive mean matching</td>
<td>0.97</td>
<td>0.822</td>
<td>0.956</td>
<td>89.5</td>
</tr>
<tr>
<td>Gaussian noise</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bayesian Nonparametric</td>
<td>1.017</td>
<td>0.962</td>
<td>0.998</td>
<td>93.2</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.988</td>
<td>0.709</td>
<td>0.982</td>
<td>83.8</td>
</tr>
<tr>
<td>Normal</td>
<td>1.015</td>
<td>0.666</td>
<td>1.008</td>
<td>78.8</td>
</tr>
<tr>
<td>Predictive Mean Matching</td>
<td>1.004</td>
<td>0.864</td>
<td>0.997</td>
<td>90.4</td>
</tr>
</tbody>
</table>

Table 1: Results of simulation of Section 4.1. RMSE corresponds to root mean-squared error for estimating \( E(Y_1) \), ŠE corresponds to the average estimated standard error of the point-estimate of \( E(Y_1) \), SE corresponds to the true standard error, and Coverage corresponds to the coverage of the 95% interval constructed with each method. RMSE, ŠE, and SE are all normalized by the RMSE of the Bayesian nonparametric model with no added noise.

group 1. In group 1, most observations can be placed into two latent classes; the third class displayed in Figure 3 is the union of all classes besides the first two. Although not pursued here, this gives the opportunity for a more refined analysis in which treatments possibly have different effects depending on the subpopulation.

4.3 Inference and sensitivity analysis

The goal of the sleep study was to determine whether individuals assigned to one of a collection of candidate treatments have lower WASO scores on average than individuals assigned to a placebo. Let \( \psi_{zj} \) denote the intention-to-treat effect for individuals assigned to treatment \( z \). For simplicity, we will focus on the comparison of the placebo (treatment \( z = 1 \)) to subjects assigned to a high dose of a candidate drug (treatment \( z = 5 \)).

We fit the model of Section 3 independently to both treatment groups. Using JAGS, we obtained draws of \( p_{obs} \) from the posterior. We then used the G-computation algorithm to generate samples of \( \psi_{zj} \) under the identifying restriction (1).

We illustrate sensitivity analysis by conducting a tipping-point analysis. We apply the
Figure 3: Trajectories of two latent classes of individuals in the placebo arm of the trial, and
mean response over time, and a third class consisting of all other subjects. Time is measured
in days from baseline, and WASO is normalized to have observed mean 0 and variance 1
across all times.

\[ T_r(Y_j; \xi) = Y_j + \xi_j. \]  

(6)

Because of the simple form of (6), we can obtain \( \psi_{zj} \) for different values of \( \xi_j \) as

\[ \psi_{zj}(\xi) = \psi_{zj}^{(0)} + \xi_j \times \Pr(R_j = 0) = \psi_{zj}^{(0)} + \xi_j \times (R_{1j}, \ldots, R_{nj})^T \varpi, \]  

(7)

where \( \varpi \) is the weight assigned to each \( R_i \) from the Bayesian bootstrap. After standardizing
the data so that the observed mean and variance in the placebo group are 0 and 1 respectively,
we varied \( \xi \in (-0.8, 0.8) \); the value 0.8 was chosen to correspond roughly to 1.5 conditional
standard deviations of \( Y_{i6} \) given \( Y_{i1}, \ldots, Y_{i5} \) as estimated by a multivariate normal model. In
words, we allowed missingness to be associated with an average shift of up-to 1.5 standard
To display the results, we constructed contour plots giving the posterior probability $P = \Pr(\psi_{5j} > \psi_{1j})$ for different values $\xi_j$, allowing for different $\xi$’s for the two treatments. A heuristic invocation of the Bernstein–von Mises theorem (Bickel et al., 2012) justified by our simulation results suggests that we can treat $P$ as an approximate $P$-value of the null $\psi_{5j} = \psi_{1j}$ against the alternative $\psi_{5j} < \psi_{1j}$. Results are displayed in Figure 4. Contour lines corresponding to $P = (0.01, 0.05, 0.1, 0.2, 0.8, 0.9, 0.95, 0.99)$ are given.

The choice of $\xi$ for the two treatments exerts a large influence in our conclusions. Focusing on day 35, there is some weak evidence of an effect when $\xi = 0$ for both treatments, with smaller values of $\xi$ corresponding to more evidence of an effect. This suggests that, if those who are not recorded on a given day tend to sleep better, then non-ignorable missingness is masking the treatment effect. In reality, many of the missing responses are thought to correspond to WASO scores of 0, so $\xi < 0$ makes sense from a subject-matter perspective.
5 Discussion

We have introduced a general framework for missing data which achieves flexible modeling in the presence of non-monotone missingness while still allowing for a principled sensitivity to be easily conducted. The usefulness of this approach was illustrated on the sleep study and included a simulation study suggesting the need for flexible modeling. Theoretical support for our approach was established by giving conditions under which posterior consistency is attained.

For simplicity, we did not take into account any covariates and stratified the analysis by treatment. One can adjust for covariates by incorporating them into (5). For example, for continuous covariates $x \in \mathbb{R}^d$, we can set $\pi(x, r) \propto \gamma_k \left\{ \prod_{j=1}^{J} \theta_{kj}^{r_j}(1 - \theta_{kj})^{1-r_j} \right\} e^{-\|x-\mu_x\|^2/\tau_x}$, as in Norets & Pelenis (2014), or set $\mu_{kj}(x) = m_j + b_{kj}(t_j, x)$ with $b_k(\cdot)$ a Gaussian process and $t_j$ denoting the $j^{th}$ time. We also did not take into account any information on reason for missingness, as this information was unavailable. Auxiliary information on reason for dropout can be incorporated as in Hogan & Laird (1997) or Linero & Daniels (2015).

This paper focuses strictly on continuous responses, but the framework carries over unmodified to categorical or ordinal responses, with the kernel mixture model replaced by a model more suitable for discrete responses.

Supplementary Material

Supplementary material includes the proofs of all propositions and theorems.

References


Wiley & Sons.

and Hall/CRC.

longitudinal clinical trials. In *Nonparametric Bayesian Inference in Biostatistics*.

data imputation in the presence of interaction effects. *Computational Statistics & Data 


153–161.


25