Research Statement

Antonio Linero[*]

Department of Statistics, University of Florida

October 26, 2014

Introduction and Overview

My research is focused broadly on nonparametric Bayesian methods. One stream of work concerns the application of nonparametric Bayesian methods to inference in missing data problems when the missing data does not satisfy the *missing at random* assumption. The primary application here is inference in longitudinal clinical trials. A primary obstacle in this problem is that, due to identifiability issues, it is necessary to make unfalsifiable assumptions about the missing data in order to conduct inference. Here the Bayesian approach is convenient because it allows subject matter experts to quantify their uncertainty about these unfalsifiable assumptions through informative priors; an assessment of robustness of inferences to assumptions about the missing data is referred to as a *sensitivity analysis*. Additionally, simple parametric models can introduce substantial bias into inference which is potentially exacerbated as the model deviates from missing at random. This suggests the need for flexible models.

My research work presents, to my knowledge, the first application of nonparametric Bayes techniques in this area. In my work, I introduce a general framework I refer to as the *working model framework* for specifying flexible, computationally attractive priors which allow for a principled sensitivity analysis. In future work, I hope to further address the frequentist properties of this approach and make anticipated innovations to the methodology to improve efficiency properties. I also develop methods which address *non-monotone* or *intermittent* missingness and describe how to implement a sensible sensitivity analysis in this setting. To promote the methodology, I have an `R` package in development which implements these techniques.

A second area of my research develops model selection and empirical Bayes techniques in nonparametric models. This methodology is applied to hierarchical nonparametric random effects models, which have become increasingly popular in machine learning applications such as genomics and topic modeling. In cases where the nonparametric model is not a priori essential, one might wish to test whether a more parsimonious model would suffice.

[*] Email Address: `theodds@ufl.edu` Website: `antoniolinero.github.io`

In my work, I develop tools for calculating Bayes factors for nonparametric models that can be used to test these models against parametric or semiparametric alternatives, and develop EM-like algorithms for hyperparameter optimization in these models.

Nonignorable Missing Data and Sensitivity Analysis

In clinical trials it is common that one collects data on individuals at only a subset of J scheduled measurement times. Complete observations are assumed to be drawn iid from some probability distribution P. For concreteness, suppose that the goal is to estimate the mean response at termination of the study $\mu_J = \int y_J \; P(d\boldsymbol{y})$. The presence of missing data introduces two substantial problems into the data analysis. The first problem I label the *smoothing problem*; in order to conduct efficient inference, the missing data must, in some sense, be imputed. This requires tackling the problem of conditional distribution estimation, which is intrinsically difficult in light of the curse of dimensionality. The second and more foundational problem I label the *identifiability problem*; the functional μ_J is unidentified and cannot be consistently estimated without invoking statistically unfalsifiable assumptions about the missing data. A common default assumption is the *missing at random* (MAR) assumption, which removes the identifiability problem but must be justified on a substantive basis.

Because of the inherent dependence of inference on unfalsifiable assumptions, it is important to assess the degree to which inferences are influenced by these assumptions; this is strongly argued as being essential in a recent report from the National Research Council (2010). Proposed models are designed to be flexible enough to reasonably model complex longitudinal data while simultaneously allowing for the analyst to conduct a principled sensitivity analysis.

Monotone Missing Data

Missing data is referred to as *monotone* if missingness at time j implies missingness at time $j+1$ and non-monotone otherwise. In Linero and Daniels (2014) I consider the monotone setting and introduce the concept of a *working model* for the complete-data distribution (i.e. the distribution of the data if we actually observed the missing data). The essential idea introduced is that one can — in a computationally feasible manner — place a nonparametric prior $\pi_f^\star$ on the space of complete-data distributions to induce a prior π_o on the space of observed-data distributions. Rather than being used for inference, the prior $\pi_f^\star$ is discarded, with the posterior of P being determined by the posterior $\pi_{o|Y}$ of the observed-data distribution P_o and an *identifying restriction* I. An identifying restriction is a map from distributions on the observed data P_o to distributions on the missing data given the observed data $P_{m|o}$. This approach decomposes the problem into a completely identified component (estimating P_o) and a completely unidentified component (specifying I).

The working model framework is convenient for several reasons. First, the decomposition into identified and unidentified components above is useful. Second, for computational purposes, inference about P_o can be made as though $\pi_f^\star$ were the actual prior on P. This allows, for example, data augmentation schemes to be used with the working model rather than the actual model; as a result, inference about P_o may be carried out with standard Bayesian inference tools such as JAGS , with the sensitivity analysis implemented in post-processing. I have applied the working model framework to data analysis from clinical trials under monotone missingness where the working model is taken to be a *Dirichlet process mixture of missing at random (MAR) models*.

It is useful to interpret different choices of I as variations from a baseline assumption. We do this by considering a family of identifying restrictions $\{I_\xi : \xi \in \Xi\}$ such that $\xi = \mathbf{0}$ corresponds to the MAR assumption. We may then interpret deviations from $\xi = \mathbf{0}$ as deviations of our identifying restriction from the MAR assumption. It is critical to work in conjunction with clinical experts to elicit a plausible range of values for ξ, and to ensure that ξ has an interpretation which can be clearly communicated to a non-statistical audience. I introduce a general methodology for choosing the family of identifying restrictions by specifying an interpretable transformation of the observed data, with ξ indexing the transformation. It can then be determined whether inferences are robust to deviations from MAR by seeing whether substantive conclusions continue to hold within the range of plausible values of ξ. If a final decision is required, the Bayesian approach is also convenient as a prior may be elicited on ξ from experts and a formal decision-theoretic analysis undertaken.

Non-Monotone Missing Data

Compared to the monotone setting, the non-monotone setting is relatively unexplored, and the National Research Council (2010) points this out as an area requiring much further work. The problem in non-monotone settings is complicated by the fact that (1) it turns out to be much more difficult to choose I_ξ so that $I_{\mathbf{0}}$ corresponds to the MAR assumption, (2) even if this is possible, it has been argued that MAR often does not correspond to plausible data-generating mechanisms in the first place (Robins, 1997), and (3) it is more difficult to model the response and dropout mechanisms jointly in the non-monotone setting.

To address these issues, I propose alternatives to the MAR identifying restriction which are both computationally tractable and interpretable. In current work, I develop generalizations of the Latent Missing at Random (LMAR) assumption (Harel and Schafer, 2009). Methods based on this approach improve upon MAR-based methods in that they are very amenable to computation, allow for very flexible models, and correspond to plausible data-generating mechanisms. In addition, the relationship between the response and dropout mechanisms can be captured by the correlation between latent continuous processes, which is intuitively very natural.

Model Selection for Nonparametric Bayesian Problems

Nonparametric Bayesian techniques have been applied in a variety of hierarchical settings. Consider for example the following so-called *hierarchical Dirichlet process* due to Teh et al.

(2006):

$$\begin{array}{lll l}
\text{conditional on } \psi_{ij}, & \boldsymbol{Y}_{ij} \overset{\text{indep}}{\sim} f_{\psi_{ij}}, & i = 1, \ldots, n_j,\ j = 1, \ldots, J \\
\text{conditional on } \{G_j\}_{j=1}^{J}, & \psi_{ij} \overset{\text{iid}}{\sim} G_j, & i = 1, \ldots, n_j,\ j = 1, \ldots, J \\
\text{conditional on } G_0, & G_j \overset{\text{iid}}{\sim} \mathcal{D}(\alpha G_0), & \qquad\qquad j = 1, \ldots, J \\
\text{conditional on } H_\omega, & G_0 \sim \mathcal{D}(\gamma H_\omega), &
\end{array}$$

Here $\mathcal{D}(\alpha H)$ denotes the Dirichlet process prior with mean distribution H and concentration parameter α. This is a nonparametric model on the group-specific distributions $F_j = \int f_\psi(y)\ G_j(d\psi)$ of J groups, with pooling of information across groups by sharing atoms among the distributions G_j. Priors of this form are useful, for example, in topic modeling, genomics, and multilevel models.

This prior possesses tuning parameters (α, γ) which can greatly influence the quality of inference. In certain multilevel models, we may be interested in whether or not this extra complexity is supported by the data or if a more parsimonious model might suffice. This can be examined by looking at the limiting cases $\alpha \to 0, \infty$ and $\gamma \to 0, \infty$, which reduce the model to certain parametric and semiparametric submodels. The Bayesian solution to selecting between models $\mathcal{M}_0$ and $\mathcal{M}_1$ requires calculating the *Bayes factor* of $\mathcal{M}_0$ relative to $\mathcal{M}_1$, $\text{BF}(\mathcal{M}_0, \mathcal{M}_1) = m_0/m_1$ where m_i is the *marginal likelihood* of the observed data under model $\mathcal{M}_i$. We may also consider a continuum of models $\{\mathcal{M}_{\alpha,\gamma} : \alpha, \gamma > 0\}$ and instead aim for choosing the values of α and γ by optimizing the marginal likelihood $m_{\alpha,\gamma}$.

The *empirical Bayes* approach to model selection simply selects the model with the largest marginal likelihood, or equivalently the model which has the largest Bayes factor relative to a given model $\mathcal{M}_{\alpha^\star,\gamma^\star}$. In my dissertation work, I develop techniques for calculating families of Bayes factors in hierarchical nonparametric models like the one given above. The Dirichlet process is intimately connected to a distribution on partitions referred to as the *Chinese Restaurant Process* (CRP), and exploiting the fact that the CRP is an exponential family on the space of partitions I also develop EM and stochastic approximation algorithms for optimization with respect to the concentration parameters of a Dirichlet process.

Future Directions

I intend to further develop the nonparametric Bayesian methodology in applications to the analysis of missing data. An important open question is the extent to which the methodology can be adapted to achieve optimal frequentist inference. While the priors in Linero and Daniels (2014) work well in practical settings, it is likely that they are not fully efficient. It will be interesting to see whether this issue can be addressed in a manner which also allows for a similar sensitivity analysis to be carried out.

In addition to developing new methodology for missing data, it is important to also provide practitioners with practical tools for implementing the new methodology. There is currently a dearth of software for actually implementing this. I am currently developing an `R` package which implements the methodology developed in my dissertation which I hope to release soon.

I am also interested in extending the methodology developed to causal inference. The Rubin Causal Model effectively embeds the statistical analysis of causal models into inference in a missing data model, where *potential outcomes* under different treatment regimes are treated as missing data. Like the missing data problem in clinical trials, drawing causal conclusions from observational studies is not entirely a statistical problem, and one is required to make assumptions which are statistically unfalsifiable.

I have several interests outside my thesis work, primarily related to theory and applications of nonparametric Bayesian methods. One is analysis of Bayesian sum-of-tree models as proposed by Chipman et al. (2010). In unpublished work, I have shown that as the number of trees tends to infinity one obtains a particular Gaussian process prior. This limiting distribution is undesirable for several reasons; in particular, the Gaussian process prior does not favor models with sparse interactions between a small number of variables. Chipman et al. (2010) propose these models for variable selection, but, for essentially the reason outlined above, they are forced to use a small number of trees for variable selection. By modifying the prior in a suitable manner, we address these issues and an alternate limiting distribution as the number of trees tends to infinity derived.

I am fortunate to enjoy a collaboration with the Department of Psychiatry concerning repetitive behavior. Repetitive behaviors such as twitching or rocking in a chair are common symptoms in individuals with autism. The goal of this project is to understand how to elicit repetitive behavior in model species such as deer mice, how the tendency to exhibit repetitive behavior is influenced by genotype, and what changes in the brain are associated with the development of symptoms.

References

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298.

Harel, O. and Schafer, J. (2009). Partial and latent ignorability in missing-data problems. *Biometrika*, 96:37–50.

Linero, A. R. and Daniels, M. J. (2014). A flexible Bayesian approach to monotone missing data in longitudinal studies with informative dropout with application to a schizophrenia clinical trial. *to appear in Journal of the American Statistical Association.*

National Research Council (2010). *The Prevention and Treatment of Missing Data in Clinical Trials.* The National Academies Press.

Robins, J. (1997). Non-response models for the analysis of non-monotone non-ignorable missing data. *Statistics in Medicine*, 16:21–37.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American statistical Association*, 101(476).