A Bayesian Nonparametric Approach to Monotone Missing Data in Longitudinal Studies with Informative Missingness

A. Linero and M. Daniels

UF, UT-Austin

SRC 2014, Galveston, TX

Notation

- y: full data response
- y_{mis}: missing data
- y_{obs}: observed data
- r: missingness indicators
- S: number of observed responses
- $p_s(\cdot) = p(\cdot|S = s)$
- $\bar{Y}_s = (Y_1, \ldots, Y_s)$
- J observation times

Quick Review of missing data I

- nonignorable missingness: need to model $p(y, s)$
  - ignorable
    - missingness is MAR
    - distinct parameters for $p(y|\boldsymbol{\theta})$ and $p(S|y, \gamma)$
    - a priori independence
- of interest

$$p(y|\boldsymbol{\omega}) = \int p(y|s; \boldsymbol{\omega})dF(s; \boldsymbol{\omega})$$

Quick Review of missing data II

- extrapolation factorization:

$$p(y, s; \omega) = p(y_{mis} | y_{obs}, s; \omega) p(y_{obs}, s; \omega)$$

- unidentified parameters/sensitivity analysis/informative priors (NAS 2010 report)
  - not possible with parametric SM and SPM

General approach

$$p(y, s; \omega) = p(y_{mis} | y_{obs}, s; \omega) p(y_{obs}, s; \omega)$$

- specify a 'working model' for the joint distribution of the responses and the dropout process (S)
- extract the observed data model from this working model
- identify the extrapolation distribution using priors grounded off of identifying restrictions

Why this approach?

- avoids parametric assumptions on the observed data distribution
- will scale up in a reasonable way
- allows sensitivity parameters
- allows for fair characterization of uncertainty since within the Bayesian paradigm (which offers advantages over semiparametric doubly robust approaches)

What is a 'Working model'? I

We begin by specifying a *working model* for the joint distribution of the response and dropout processes.

Definition

For a model $p(y, s \mid \omega)$, a model $p^\star(y, s \mid \omega)$ is called a *working model* if for all $s \in \{1, 2, \ldots, J\}$,

$$p(y_{obs}, s \mid \omega) = \int p^\star(y_{obs}, y_{mis}, s \mid \omega) \; dy_{mis}. \tag{1}$$

A given specification of $p^\star(y, s \mid \omega)$ identifies $p(y, s \mid \omega)$ only up to $p(y_{obs}, s \mid \omega)$, leaving $p(y_{mis} | y_{obs}, s, \omega)$ unidentified.

What is a 'Working model'? II

For the purposes of likelihood based inference, it suffices to fit $p^\star(y, s|\omega)$ to the data.

Proposition

A model $p(y, s \mid \omega)$ and corresponding working model $p^\star(y, s \mid \omega)$ have the same observed data likelihood.

Implications of the working model

1. We can focus on specifying $p^\star$ to fit the data well without affecting the extrapolation distribution $p(y_{mis} | y_{obs}, s, \omega)$.
2. often easier conceptually to design $p^\star$ to induce desired sharing of information across dropout times without needing to take precautions in leaving the extrapolation distribution unidentified rather than specifying p directly.
3. also convenient because it allows us to specify a single model of dimension J (as opposed to for $(\bar{Y}_s, S)$)
4. For computational purposes, Y_{mis} may be imputed via data augmentation using $p^\star$ rather than p, which is substantially simpler.

Form of the 'Working model' I

- We take $p^\star$ to be a mixture of models as follows
  - $f(y|\boldsymbol{\theta}_1)$ model for the outcome process
  - $g(s|y, \boldsymbol{\theta}_2)$ model for the dropout processes conditional on the outcome process

$$p^\star(y, s|\boldsymbol{\omega}) = \int f(y|\boldsymbol{\theta}_1) g(s|y, \boldsymbol{\theta}_2)\ F(d\boldsymbol{\theta}).$$

- distribution of F is modeled as a Dirichlet process with (parameters) base distribution H and mass $\alpha > 0$
- specify $g(s|y, \boldsymbol{\theta}_2)$ as MAR so resulting working model is a mixture of MAR models

Form of the 'Working model' II

- specification on previous slide is equivalent to the "stick-breaking" construction (Sethuraman, 1994), which shows the Dirichlet process mixture is a prior on latent class models,

$$p^{\star}(y, s|\boldsymbol{\omega}) = \sum_{k=1}^{\infty} \beta_k f(y|\boldsymbol{\theta}_1^{(k)}) g(s|y, \boldsymbol{\theta}_2^{(k)}), \tag{2}$$

where $\beta_k = \beta_k' \prod_{j<k}(1-\beta_j')$, $\beta_j' \sim \text{Beta}(1, \alpha)$, and $\boldsymbol{\theta}_k \overset{iid}{\sim} H(d\boldsymbol{\theta})$.

Choice of the models in the 'Working model'

- $f(y|\boldsymbol{\theta}_1)$ a normal kernel (for continous data)
- $g(s|y,\boldsymbol{\theta}_2)$: sequence of regression on hazards
- or $g(s|y,\boldsymbol{\theta}_2) = g(s|\boldsymbol{\theta}_2)$, a categorical or ordinal distribution on $\{1, ..., J\}$.
  - A convenient choice of $g(s|\boldsymbol{\theta}_2)$ in this case is the ordinal probit model

'Nonparametric'?

- While the proposed method is "nonparametric" in the sense of having large support in the space of random probability measures, draws from the Dirichlet process mixture resemble finite mixture models.
- but in light of the curse of dimensionality, it is not feasible to estimate the distribution of longitudinal data for even moderate J in a fully nonparametric manner
- the Dirichlet process mixture combats this by shrinking towards latent class models with a small number of dominant components.

Extrapolation distribution I

We now need to specify the extrapolation distribution

$$p(y_{mis}|y_{obs}, s, \omega)$$

- Identifying restrictions, which express the extrapolation distribution as a function of the observed data distribution, provide a natural starting point.
  - available case missing value (ACMV) restriction sets

$$p_k(y_j|\bar{y}_{j-1}, \omega) = p_{\geq j}(y_j|\bar{y}_{j-1}, \omega),$$

for all $k < j$ and $2 \leq j < J$; equivalent to the MAR restriction under monotone missingness (Molenberghs et al., 1998),

$$P(S = s|Y, \omega) = P(S = s|\bar{Y}_s, \omega).$$

Extrapolation distribution II

- A subclass of identifying restrictions is generated by the non-future dependence assumption (NFD) (Kenward et al., 2003),
  - the probability of dropout at time s depends only $\bar{y}_{s+1}$,

$$P(S = s | Y, \omega) = P(S = s | \bar{Y}_{s+1}, \omega). \tag{3}$$

  - NFD holds if and only if

$$p_k(y_j | \bar{y}_{j-1}, s, \omega) = p_{\geq j-1}(y_j | \bar{y}_{j-1}, s, \omega), \tag{4}$$

    for $k < j - 1$ and $2 < j \leq J$, but places no restrictions on $p_{j-1}(y_j | \bar{y}_{j-1}, s, \omega)$.
  - MAR when $p_{j-1}(y_j | \bar{y}_{j-1}, s, \omega) = p_{\geq j}(y_j | \bar{y}_{j-1}, s, \omega)$.

Extrapolation distribution III

	$j = 1$	$j = 2$	$j = 3$	$j = 4$
$S = 1$	$p_1(y_1)$	?	$p_{\geq 2}(y_3 \mid \bar{y}_2)$	$p_{\geq 3}(y_4 \mid \bar{y}_3)$
$S = 2$	$p_2(y_1)$	$p_2(y_2 \mid y_1)$	?	$p_{\geq 3}(y_4 \mid \bar{y}_3)$
$S = 3$	$p_3(y_1)$	$p_3(y_2 \mid y_1)$	$p_3(y_3 \mid \bar{y}_2)$	?
$S = 4$	$p_4(y_1)$	$p_4(y_2 \mid y_1)$	$p_4(y_3 \mid \bar{y}_2)$	$p_4(y_4 \mid \bar{y}_3)$

Table : Schematic representation of NFD when $J = 4$. Distributions above the dividing line are not identified by the observed data.

Identification within NFD I

We consider two methods to identify the distribution $p_{j-1}(y_j|\bar{y}_{j-1}, s, \omega)$ under NFD.

1. consider the existence of a transformation $T_j(y_j|\bar{y}_{j-1}, \xi_j)$ such that

$$[Y_j|\bar{Y}_{j-1}, S = j-1, \omega] \stackrel{d}{=} [T_j(Y_j|\bar{Y}_{j-1}, \xi_j)|\bar{Y}_{j-1}, S \geq j, \omega],$$

   where $\stackrel{d}{=}$ denotes equality in distribution.
   - If T_j is chosen so that $T_j(Y_j|\bar{Y}_{j-1}, \mathbf{0}) = Y_j$ then deviations of ξ_j from $\mathbf{0}$ represent deviations of the assumed model from MAR.
   - Wang and Daniels (2011) implicitly take this approach,

Identification within NFD II

2 an exponential tilting assumption (Birmingham et al., 2003)

$$p_{j-1}(y_j|\bar{y}_{j-1}, s, \omega) = \frac{p_{\geq j}(y_j|\bar{y}_{j-1}, s, \omega)e^{q_j(\bar{y}_j)}}{\int p_{\geq j}(y_j|\bar{y}_{j-1}, s, \omega)e^{q_j(\bar{y}_j)}\, dy_j}. \quad (5)$$

- The function $q_j(\bar{y}_j)$ characterizes the influence of y_j on the scale of log-odds ratios of the probability of dropout at time $S = j - 1$ conditional on $S \geq j - 1$.
- When a normal kernel is used to model the outcome response, $q_j(\bar{y}_j) = \gamma_j y_j$ leads to tractable inference

Analysis of Schizophrenia trial I

- we analyze data from a clinical trial designed to determine the efficacy of a new drug for the treatment of acute schizophrenia.
- response was measured at baseline, Day 4, and Weeks 1, 2, 3, and 4 (so $J = 6$).
- three treatment (V) arms corresponding to the test drug (81 subjects), active control (45 subjects), and placebo (78 subjects)
- primary endpoint: mean change from baseline of the Positive and Negative Syndrome Scale (PANSS) after Week 4,

$$\eta_v = E(Y_6 - Y_1 | V = v, \omega)$$

Analysis of Schizophrenia trial II

- Moderate dropout: 33%, 19%, and 25% dropout for $V = 1, 2, 3$
- Many dropout patterns were sparse
  - for example, in the active control arm, only a single subject dropped out at each time $j = 1, 2, 3$.

Informative priors under MNAR I

- We compare inferences under MAR to inferences under NFD.
- To complete the NFD specification we first introduce sensitivity parameters ξ_j common across treatments such that

$$[Y_j | \bar{Y}_{j-1}, S = j-1, V = v, \omega] \stackrel{d}{=} [Y_j + \xi_j | \bar{Y}_{j-1}, S \geq j, V = v, \omega].$$

- MAR corresponds to $\xi_j \equiv 0$ for all j.
- We specify an informative prior $\xi_j \stackrel{iid}{\sim} \text{Uniform}(0, 8)$. So, deviations from MAR
  - are chosen to be at most roughly one residual standard deviation (8)
  - are restricted to be positive to reflect the fact that those who dropped out are a priori believed to have higher PANSS scores on *average*.

Informative priors under MNAR II

- we also compare to different missingness mechanisms across treatments.
  - fewer of the dropouts in the active arm dropped out due to lack of efficacy versus the placebo and test arms
  - as a result, we consider MAR for the active arm and the MNAR priors for the other two arms.

Informative priors under MNAR III

Model	$\eta_1 - \eta_3$	η_2 - η_3	DIC
	MAR Model		
DP	$-1.9(-8.9, 5.2)$	$-6.5(-14.1, 1.2)$	7762
	DP Sensitivity Analysis		
MNAR	$-1.7(-9.3, 5.9)$	$-7.4(-15.3, 0.6)$	7762
MNAR-2	$-1.7(-9.3, 5.9)$	$-8.3(-16.3, -0.3)$	7762

Table : Inferences under different models for the Schizophrenia data. MNAR refers to the model which takes $\xi_j \sim \text{Uniform}(0, 8)$ for all treatments; MNAR-2 assumes that $\xi_j \equiv 0$ for the active control arm only.

- 'signficant' CI under MNAR-2

Informative priors under MNAR IV

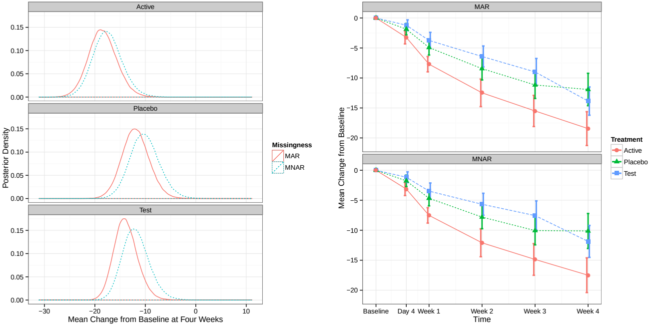

Figure : Density estimates for η_v (left) and posterior estimate of mean change from baseline over time under MAR and MNAR.

Conclusions I

- We have introduced a general methodology for conducting nonparametric Bayesian inference under nonignorable missingness
  - allows for a clear separation of the observed data distribution and the extrapolation distribution.
  - allows both flexible modeling of the observed data and flexible specification of the extrapolation distribution.
  - provides similar robustness to semiparametric AIPW (simulations not shown)

Conclusions II

- there is nothing particular about the Dirichlet process to our specification; in principle

$$p(y_{obs}, s|\omega) = \int p^{\star}(y_{obs}, y_{mis}, s|\omega)\, dy_{mis}.$$

could be applied to any joint distribution $p^{\star}(y, s|\omega)$ provided that inference is tractable.

Conclusions III

- Model complexity controlled both by our prior on α, the mass of the Dirichlet distribution, or using a hierarchical specification of the prior on H forcing the latent classes to be similar.
- R package available to implement these models
- intermittent missingness implicitly handled under assumption of partial ignorability

Extensions

- extend work to applications with covariates.
  - Often covariates are used to help with imputation of missing values, or to make the MAR assumption more plausible, but are not of primary interest (i.e., auxiliary covariates).
  - incorporation of covariates in a semiparametric fashion
- extend these methods to non-monotone missingness.
- extend these methods to more complex data (e.g., multivariate longitudinal data) - challenge is specification of the extrapolation distribution

Inference I

- We work with an approximation of the Dirichlet process mixture based on truncating the stick-breaking construction at a fixed K by setting $\beta'_K \equiv 1$ (Ishwaran and James [IJ], 2001)
- We break inference into two steps.
  1. Draw a sample of $(\boldsymbol{\theta}^{(1)}, \beta_1, \ldots, \boldsymbol{\theta}^{(K)}, \beta_K)$ from the posterior distribution given the observed data using the working model $p^{\star}(y, s|\boldsymbol{\omega})$.
  2. Calculate the posterior distribution of desired functionals of the true distribution $p(y|\boldsymbol{\omega})$.
- We use a data-augmentation scheme similar to IJ but which also includes augmenting the missing data for step 1.

Inference II

- Once we have a sample from the posterior distribution of $(\boldsymbol{\theta}^{(k)}, \beta_k)$ in step 1, scientific interest often lies in functionals of the form

$$E[t(Y)|\boldsymbol{\omega}] = \int t(y)p(y|\boldsymbol{\omega})\ dy.$$

Define $\phi_j \equiv P(S = j|\boldsymbol{\omega}) = \sum_{k=1}^{K} \beta_k g(j|\boldsymbol{\theta}_2^{(k)})$. Then,

$$E[t(Y)|\boldsymbol{\omega}] = \sum_{j=1}^{J} \phi_j E[t(Y)|S = j, \boldsymbol{\omega}].$$

  - ϕ_j is typically available in closed form given $\boldsymbol{\omega}$,
  - the expectation $E[t(Y)|S = j, \boldsymbol{\omega}]$ has a complicated form and depends on the missing data assumption

Inference III

- under MAR,

$$E[t(Y)|S=j,\omega] = \int t(y)\cdot p_j(\bar{y}_j|\omega)\cdot p_{\geq j+1}(y_{j+1}|\bar{y}_j,\omega) \cdot p_{\geq j+2}(y_{j+2}|\bar{y}_{j+1},\omega)\cdots p_J(y_J|\bar{y}_{J-1},\omega)\; dy.$$

- to calculate $E[t(Y)|\omega]$ we use Monte Carlo integration, sampling pseudo-data $Y_1^\star, \ldots, Y_{N^\star}^\star$, and forming the average $\frac{1}{N^\star}\sum_{i=1}^{N^\star} t(Y_i^\star)$ (easy for NFD in general).

Simulations I

- assess the performance of our method as an estimator of the population mean at the end of a clinical trial with $J = 3$ time points and $N = 100$.
- we generate data under the following conditions:
  - S1: $Y \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Missingness is MAR.
  - S2: Y is distributed as a 50-50 mixture of normal distributions; chosen to make the distributions of (Y_1, Y_2) and (Y_1, Y_3) highly non-linear while (Y_2, Y_3) is roughly linear. Missingness is MAR.

Simulations II

- we compare out method to
  - (a) modeling the data as normally distributed
  - (b) augmented inverse-probability weighting (AIPW); solves the estimating equation

$$\sum_{i=1}^{n} \left\{ \frac{I(S_i = J)}{P(S_i = J | Y_i)} \varphi(Y_i, \boldsymbol{\theta}) + \sum_{j=1}^{J-1} \frac{I(S_i = j) - \lambda_j(Y_i) I(S_i \geq j)}{P(S_i > j | Y_i)} E[\varphi(Y_i, \boldsymbol{\theta}) | \bar{Y}_{ij}] \right\} = 0,$$

    where $\sum_{i=1}^{n} \varphi(Y_i, \boldsymbol{\theta}) = \mathbf{0}$ is a complete data least-squares estimating equation for the regression of Y_1 on Y_2 and (Y_1, Y_2) on Y_3.
    - "doubly robust" in the sense that if either the dropout model or mean response model is correctly specified then the associated estimator is CAN

Simulations III

- Under (S1) the AIPW estimator used was constructed with the correct mean and dropout models
- Under (S2) the AIPW estimator was constructed assuming $E[Y_2|Y_1]$ to be quadratic in Y_1 and $E[Y_3|Y_1, Y_2]$ quadratic in Y_1 and linear in Y_2, with dropout modeled correctly (and so the estimator is consistent by double robustness).
- The expectation $E[\varphi(Y, \boldsymbol{\theta})|\bar{Y}_j]$ is taken under the assumption that the mean response is modeled correctly.

Simulations IV

	Bias	95% CI Width	95% CI Coverage Probability	Mean Squared Error
		Normal Model (S1)		
DP	−0.001(0.004)	0.493(0.001)	0.963(0.006)	0.01443(0.0006)
Normal	−0.005(0.004)	0.494(0.002)	0.944(0.007)	0.01524(0.0007)
AIPW	−0.001(0.004)	0.470(0.002)	0.943(0.007)	0.01530(0.0007)
		Mixture of Normal Models (S2)		
DP	−0.010(0.004)	0.542(0.001)	0.950(0.007)	0.0182(0.0008)
Normal	−0.039(0.005)	0.586(0.001)	0.949(0.007)	0.0220(0.0010)
AIPW	0.001(0.004)	0.523(0.001)	0.944(0.007)	0.0185(0.0008)
Mixture	−0.006(0.004)	0.536(0.001)	0.952(0.007)	0.0182(0.0008)

Table : Comparison of methods for estimating the population mean at time $J = 3$.