Biometrics 000, 000–000 DOI: 000
000 0000

Bayesian Model Selection For Incomplete Data using the Posterior Predictive Distribution

Michael J. Daniels[1,*] , Arkendu S. Chatterjee[1], and Chenguang Wang[2]

[1,*]Department of Statistics, University of Florida

[2]Division of Oncology Biostatistics and Bioinformatics, Johns Hopkins University

email: mdaniels@stat.ufl.edu

Summary: We explore the use of a posterior predictive loss criterion for model selection for incomplete longitudinal data. We begin by identifying a property that most model selection criteria for incomplete data should consider. We then show that a straightforward extension of the Gelfand and Ghosh (1998) criterion to incomplete data has two problems. First, it introduces an extra term (in addition to the goodness of fit and penalty terms) that compromises the criterion. Second, it does not satisfy the aforementioned property. We propose an alternative and explore its properties via simulations and on a real dataset and compare it to the deviance information criterion (DIC). In general, the DIC outperforms the posterior predictive criterion, but the latter criterion appears to work well overall and is very easy to compute unlike the DIC in certain classes of models for missing data.

Key words: DIC, Bayes Factor, MCMC, Model Selection.

1. Introduction

When several parametric models are under consideration, it is often of interest to determine which one fits the data the best. More specifically, choosing a probability model for the observed $\boldsymbol{Y}$, indexed by m, conditioned on a parameter vector $\theta^{(m)}$,

$$p(\boldsymbol{y}|m, \theta^{(m)}), m \in M, \theta^{(m)} \in \Theta^{(m)}$$

where M is the model space and $\Theta^{(m)}$ is the parameter space. We choose the model with the best value for the chosen criterion.

In the context of Bayesian inference, there have been many criteria proposed for model selection. We will briefly review three popular choices: Bayes Factors (BF), likelihood based penalized criteria, and posterior predictive distribution based criteria. We will then discuss issues in using these different criteria for incomplete longitudinal data.

1.1 *Bayes Factors*

The standard Bayesian approach to compare models is based on the ratio of marginal likelihoods, or the Bayes Factor (for an excellent review, see Kass and Raftery, 1995). The marginal likelihood for model m is defined as

$$p(\boldsymbol{y}|m) = \int p(\boldsymbol{y}|\theta^{(m)}, m)p(\theta^{(m)}|m)d\theta^{(m)}.$$

The main issues with Bayes Factors are related to computation (i.e., of the marginal likelihoods of the models under consideration) and the need to use proper priors for the parameters being 'compared' across models. However, an attractive feature of Bayes Factors is their connection to posterior model probabilities; among other things, this provides a good way to calibrate them.

Chib and colleagues (Chib, 1995; Chib and Jeliazkov, 2001 & 2005) in a series of papers have proposed computationally efficient ways to compute Bayes Factors using MCMC output. Recent work by Johnson and colleagues (2005, 2009) have proposed Bayes Factors based on test statistics. We will connect Johnson's work to our approach later.

1.2 *Likelihood based penalized criteria*

Given the popularity of sampling based approaches to compute posterior distributions, the most common likelihood based penalized criterion is the 'easy to compute' Deviance information criterion (DIC). Spiegelhalter et al (2002) proposed this criterion which is composed of two terms, a goodness of fit term and a complexity/penalty term. The goodness of fit term is the deviance evaluated at a summary of the posterior distribution of the parameters (often the posterior mean). The complexity penalty is defined as the posterior mean deviance ($\overline{D}$) minus the deviance evaluated at the posterior mean of the parameters; this is related to the idea of residual information. Two of the drawbacks of this criterion are the lack of invariance to the parameterization of the model and the choice of the likelihood in hierarchical/multilevel models. The seminal paper by Spiegelhalter et al. has been followed by numerous papers examining the DIC in more complex settings. Quite relevant for our setting is the work of Celeux et al (2006) who proposed several versions of DIC for settings with missing data. However, their recommendations were based on latent data, not responses that could be observed. We focus on the latter. Daniels and Hogan (2008) and Wang and Daniels (2011) recommended constructing the DIC based on the observed data likelihood for comparison of models based on incomplete data with the latter examining its performance with simulation studies. Treating the missing responses as 'latent' data and using the recommendations in Celeux et al. will result in criteria that do match satisfy desired properties, including the one to be introduced in Section 1.4.

1.3 *Posterior Predictive Distribution Based Criteria*

Numerous papers have proposed Bayesian criteria based on the posterior predictive distribution (Geisser and Eddy, 1979; Laud and Ibrahim, 1994; Ibrahim and Laud, 1995; Gelman, Meng and Stern, 1996; Gelfand and Ghosh, 1998; Ibrahim, Chen, and Sinha, 2001; Chen, Dey, and Ibrahim, 2004). The posterior predictive distribution for the replicated data $\boldsymbol{y}_{rep}$

under model m is given by

$$p(\boldsymbol{y}_{\text{rep}}|\boldsymbol{y}, m) = \int p(\boldsymbol{y}_{\text{rep}}|\theta^{(m)}, m)\pi(\theta^{(m)}|\boldsymbol{y}, m)d\theta^{(m)}.$$

In what follows, for clarity we drop dependence on the model m. Ibrahim and colleagues have proposed general Bayesian criteria from the posterior predictive distribution of the data. In general, good models should make predictions, $\boldsymbol{y}_{\text{rep}}$ close to what was observed, $\boldsymbol{y}$. Ibrahim and Laud (1994) defined their criterion as the expected squared Euclidean distance between $\boldsymbol{y}$ and $\boldsymbol{y}_{\text{rep}}$,

$$L = \text{E}\{(\boldsymbol{y}_{\text{rep}} - \boldsymbol{y})^{'}(\boldsymbol{y}_{\text{rep}} - \boldsymbol{y})\},$$

where the expectation was taken with respect to the posterior predictive distribution, $p(\boldsymbol{y}_{rep}|\boldsymbol{y})$. L can be re-expressed as

$$L = \sum_{i=1}^{n} \left[\text{Var}(y_{\text{rep},i}|\boldsymbol{y}) + \{\text{E}(y_{\text{rep},i}|\boldsymbol{y}) - y_i\}^2\right].$$

They call the proposed predictive criterion the L-measure. They examined the L-measure in detail for a variety of models. They also suggest approaches for calibration of the criterion and explore a variety of weighting strategies.

Gelfand and Ghosh (1998) proposed a more general loss function

$$\mathcal{L}(\boldsymbol{y}_{\text{rep}}, a; \boldsymbol{y}) = L(\boldsymbol{y}_{\text{rep}}, a) + kL(\boldsymbol{y}, a), k > 0.$$

For a model m they minimized $\text{E}\{\mathcal{L}(\boldsymbol{y}_{\text{rep}}, a; \boldsymbol{y})|\boldsymbol{y}\}$, the posterior predictive expectation of the loss with respect to an action, a. We provide some more details on this approach in Section 2 and use this as the starting point for our proposal. Chen et al. (2004) later used this loss function in the context of categorical regression models.

Model comparison is an important part of inferential statistics. We have briefly reviewed the most relevant literature on Bayesian methods for model comparison. We now discuss issues specific to incomplete data.

1.4 *Issues with Bayesian model selection with incomplete data*

For Bayesian inference with incomplete data, we often want to compare the fit of selection models (Heckman, 1976; Diggle and Kenward, 1994; Fitzmaurice, Molenberghs, and Lipsitz, 1995), shared parameter models (Wu and Carroll, 1988; Rizopoulos, Verbeke, and Molenberghs, 2008), and mixture models (Little, 1994; Daniels and Hogan, 2000; Kenward et al., 2003). For a good review of models, see texts by Molenberghs and Kenward (2007) and Daniels and Hogan (2008). Here we will focus on incomplete *longitudinal* data.

Model selection criteria for incomplete data should have a certain property in most situations; we identify situations when this is less important in the discussion. Before we introduce it, we first introduce some notation and review the extrapolation factorization (Daniels and Hogan, 2008). Let $\boldsymbol{R}$ be the vector of observed data indicators; i.e., $R_{ij} = \mathrm{I}(Y_{ij}$ is observed) and $\boldsymbol{Y}_{\text{obs}}$ as $\{Y_{ij} : r_{ij} = 1\}$. The full data is given as $(\boldsymbol{y}, \boldsymbol{r})$; the observed data as $(\boldsymbol{y}_{\text{obs}}, \boldsymbol{r})$. The extrapolation factorization is

$$p(\boldsymbol{y}, \boldsymbol{r}; \omega) \;\; = \;\; p(\boldsymbol{y}_{\text{mis}} | \boldsymbol{y}_{\text{obs}}, \boldsymbol{r}; \omega_E) p(\boldsymbol{y}_{\text{obs}}, \boldsymbol{r}; \omega_O),$$

where $p(\boldsymbol{y}_{\text{obs}}, \boldsymbol{r}; \omega_O)$ is the observed data model and $p(\boldsymbol{y}_{\text{mis}} | \boldsymbol{y}_{\text{obs}}, \boldsymbol{r}; \omega_E)$ is the (extrapolation) distribution of the missing data given the observed data. There is no information in the observed data about the extrapolation distribution.

Property I (Invariance to Extrapolation Distribution): Two models for the full data with the same model specification for the observed data, $p(\boldsymbol{y}_{\text{obs}}, \boldsymbol{r}; \omega_O)$ and same prior for $p(\omega_O)$ should give the same value of the Bayesian model selection criterion.

The deviance information criterion based on the observed data likelihood has this property (Daniels and Hogan, 2008 ; Wang and Daniels, 2011).

A main complication with criteria for incomplete data is computational. For example, both the DIC and Bayes Factors require computation of observed data likelihood which is very difficult for most selection models and shared parameter models. Approaches based on

the posterior predictive distribution based criteria in general do not need to use a closed form for the observed data likelihood. Our proposal will be simple and computationally attractive and will satisfy Property I. Our ultimate objective will be to choose the model under consideration that provides the best fit, and then to proceed with a sensitivity analysis (Daniels and Hogan, 2008).

In section 2, we review the Posterior Predictive Loss (PPL) model selection criterion proposed by Gelfand and Ghosh (and Ibrahim and Laud and colleagues) and propose a simple modification for complete longitudinal data. In section 3, we propose extensions for incomplete longitudinal data pointing out problems using the criterion based on a straightforward generalization. In section 4, we apply our criterion to incomplete longitudinal data from a recent clinical trial. Finally in section 5 we conduct some simulations to examine the operating characteristics of this criterion and compare its performance to the DIC. We offer conclusions and extensions in Section 6.

2. Posterior Predictive Loss: A quick review

Posterior Predictive Loss (PPL), is the model selection criterion proposed by Gelfand and Ghosh (1998). PPL quantifies the fit of the model by comparing features of the posterior predictive distribution, $p(\boldsymbol{y}_{\text{rep}}|\boldsymbol{y})$ to equivalent features of the observed data. The comparison is based on a loss function $\mathcal{L}(\boldsymbol{y}_{\text{rep}}, a; \boldsymbol{y}|\boldsymbol{y})$, where a is chosen to minimize the expectation of the loss with respect to the posterior predictive distribution $\text{E}\{\mathcal{L}(\boldsymbol{y}_{\text{rep}}, a; \boldsymbol{y}|\boldsymbol{y})\}$. Gelfand and Ghosh [GG] (among others) proposed the following loss function

$$\mathcal{L}(\boldsymbol{y}_{\text{rep}}, a; \boldsymbol{y}) = L(\boldsymbol{y}_{\text{rep}}, a) + kL(\boldsymbol{y}, a) \qquad k > 0.$$

When $L(\cdot)$ is chosen as squared error loss they showed that,

$$\begin{aligned} \min[\text{E}\{\mathcal{L}(\boldsymbol{y}_{\text{rep}}, a; \boldsymbol{y})|\boldsymbol{y}\}] &= \sum_{i=1}^{n} \text{Var}(y_{\text{rep},i}|\boldsymbol{y}) + \frac{k}{k+1} \sum_{i=1}^{n} \{\text{E}(y_{\text{rep},i}|\boldsymbol{y}) - y_i\}^2 \\ &= \text{Penalty Term} + \text{Goodness Of Fit Term} \end{aligned}$$

The expectation is with respect to the posterior predictive distribution associated with $\boldsymbol{y}_{\text{rep}}$. As the models become increasingly complex the Goodness of Fit term will decrease but the penalty term will begin to increase. Overfitting of model results in large predictive variances σ_i^2 and large values of the penalty function. The choice of k determines how much weight is placed on the goodness of fit term relative to the penalty term. As k goes to infinity, equal weight is placed on these two terms; and corresponds to the original $\mathcal{L}$ criterion in Ibrahim and Laud (1994) . The criterion is easy to calculate using samples from the posterior predictive distribution.

2.1 *A simple modification for (complete) longitudinal data*

Now let $\boldsymbol{y}_i$ be a $T \times 1$ vector of longitudinal responses observed at times $t_1, \ldots, t_T$. One issue in applying a PPL criterion to multivariate observations is the lack of independence of components of $\boldsymbol{y}_i$. Weighting each of the components of the $\boldsymbol{y}_i$ vector equally may not be a good choice. To address this, options include a multivariate loss function (e.g., deviance based loss or multivariate weighted squared error loss) or using a univariate summary. The multivariate loss alternative has complications including the intractability of the *observed* data likelihood and weighted multivariate normal loss type measures (Ibrahim and Laud, 1994 ; Chen et al., 2004) require knowing the weight matrix (i.e., the inverse of the covariance matrix). Here we propose replacing $\boldsymbol{y}$ in the criterion by a univariate summary of $\boldsymbol{y}$, $h(\boldsymbol{y})$, possibly of (inferential) interest. The resulting criterion can be shown to be,

$$C_k(h) = \sum_i^n \text{Var}\{h(\boldsymbol{y}_{\text{rep},i})|\boldsymbol{y}\} + \frac{k}{1+k}\sum_i^n [\text{E}\{h(\boldsymbol{y}_{\text{rep},i})|\boldsymbol{y}\} - h(\boldsymbol{y}_i)]^2 \tag{1}$$

A derivation can be found in Web Appendix A.

Choosing a summary measure as we do above, is similar, to some extent to the approach of Johnson who computes Bayes Factors based on a test statistic (Johnson, 2005; Hu and Johnson, 2009). However, using the statistic as he does creates several complications in our

setting. First, we will typically not be able to obtain closed forms for the Bayes factors based on the test statistics in the setting of models for incomplete data and the distributions of the test statistics will likely be complex. Second, most of the models we compare are not nested models and the likelihood is not available in closed form so the approach to model selection in Hu and Johnson (2009) can not be readily adapted to our setting.

3. PPL for incomplete longitudinal data

The obvious extension from the complete longitudinal data case is to just take expectations with respect to $p(\boldsymbol{y}_{\text{rep}}|\boldsymbol{y}_{\text{obs}}, \boldsymbol{r})$ (instead of $p(\boldsymbol{y}_{\text{rep}}|\boldsymbol{y})$). The criterion can then be shown to have the following form (see Web Appendix A for the derivation),

$$\begin{aligned} C_k(h) &= \sum_i^n \text{Var}\{h(\boldsymbol{y}_{\text{rep},i})|\boldsymbol{y}_{\text{obs}}, \boldsymbol{r}\} + k\sum_i^n \text{Var}\{h(\boldsymbol{y}_i)|\boldsymbol{y}_{\text{obs}}, \boldsymbol{r}\} \\ &\quad + \frac{k}{1+k}\sum_i^n [\text{E}\{h(\boldsymbol{y}_i)|\boldsymbol{y}_{\text{obs}}, \boldsymbol{r}\} - \text{E}\{h(\boldsymbol{y}_{\text{rep},i})|\boldsymbol{y}_{\text{obs}}, \boldsymbol{r}\}]^2. \end{aligned} \tag{2}$$

The resulting criterion has an extra term, $k\sum_{i=1}^n \text{Var}\{h(\boldsymbol{y}_i)|\boldsymbol{y}_{\text{obs}}, \boldsymbol{r}\}$. This is the conditional variance of $h(\boldsymbol{y})$ with respect to $p(\boldsymbol{y}|\boldsymbol{y}_{\text{obs}}, \boldsymbol{r})$; note that $\text{Var}(\boldsymbol{y}|\boldsymbol{y}_{\text{obs}}, \boldsymbol{r}) \equiv \text{Var}(\boldsymbol{y}_{\text{mis}}|\boldsymbol{y}_{\text{obs}}, \boldsymbol{r})$. This term is problematic for model selection criteria which we show in the following theorem. However, note that when there is no missingness, this term is zero and (2) simplifies to (1).

Theorem I: For two models with

(1) the same observed data model, $p(\boldsymbol{y}_{\text{obs}}, \boldsymbol{r}; \omega_O)$,

(2) the same prior, $p(\omega)$, and

(3) the same conditional expectation, $\text{E}(\boldsymbol{y}_{\text{mis}}|\boldsymbol{y}_{\text{obs}}, \boldsymbol{r}; \omega_E)$ for the extrapolation distribution,

the criterion in (2) (for $k > 0$) is minimized when the extrapolation distribution, $p(\boldsymbol{y}_{\text{mis}}|\boldsymbol{y}_{\text{obs}}, \boldsymbol{r}; \omega_E)$ is degenerate.

See Web Appendix A for a proof.

The theorem implies that this criterion will always pick a 'single imputation type' procedure

that gives the same values for $\mathrm{E}\{h(\boldsymbol{y}_{\mathrm{rep}})|\boldsymbol{y}_{\mathrm{obs}}, \boldsymbol{r}\}$ as a corresponding multiple imputation type procedure. Obviously this is bad practice and the criterion is flawed as it favors *not* allowing uncertainty about the 'filled-in' missing data (and *penalizes* extra uncertainty about it). In addition, the criterion does not satisfy Property I. So the form of the extrapolation distribution impacts the model selection criterion even though the data provide no information about it.

A way to avoid this problem would be to allow k to be unit-specific, i.e., k_i and set $k = 0$ if $h(\boldsymbol{y}_i)$ is *not* observed; GG suggest this as an option (top of p. 4). However, this alternative does not use all the data as part of $\boldsymbol{y}_i$ will be observed and this option 'throws' away the entire vector $\boldsymbol{y}_i$ if it is incomplete; in addition, it will likely introduce bias in model selection as it would be done on 'completers only'.

In the next section, we provide an alternative formulation that avoids the problems of (2).

3.1 *A re-formulation*

The complication with a direct extension of the PPL to incomplete longitudinal data arising from the fact that $h(\boldsymbol{y})$ is not always observed and this results in an extra term in the criterion. A straightforward and natural way to overcome this complication is to use a new univariate function of the data that is only a function of *observables*, i.e., $(\boldsymbol{r}, \boldsymbol{r} \circ \boldsymbol{y})$, where $(\boldsymbol{r} \circ \boldsymbol{y}) = (r_1 y_1, r_2 y_2, \ldots, r_T y_T)$. To derive the criterion here, we just replace $h(.)$ by $T(\boldsymbol{r}, \boldsymbol{r} \circ \boldsymbol{y})$ from the previous derivation and obtain

$$\begin{aligned} C_k(T) &= \sum_i^n \mathrm{Var}\{T(\boldsymbol{r}_{\mathrm{rep},i}, \boldsymbol{r}_{\mathrm{rep},i} \circ \boldsymbol{y}_{\mathrm{rep},i})|\boldsymbol{y}_{\mathrm{obs}}, \boldsymbol{r}\} \\ &\quad + \frac{k}{1+k} \sum_i^n [T(\boldsymbol{r}_i, \boldsymbol{r}_i \circ \boldsymbol{y}_i) - \mathrm{E}\{T(\boldsymbol{r}_{\mathrm{rep},i}, \boldsymbol{r}_{\mathrm{rep},i} \circ \boldsymbol{y}_{\mathrm{rep},i})|\boldsymbol{y}_{\mathrm{obs}}, \boldsymbol{r}\}]^2. \end{aligned}$$

This no longer has the problematic extra term. We discuss the choice of $T(\cdot)$ and some computational issues in the next two sections and then evaluate the criterion via simulations. Note that the criterion assesses *replicated observed* data here (as opposed to replicated full

(or complete) data). This version of the criterion satisfies Property I, i.e., it is invariant to the extrapolation distribution and will only give information about the fit of $p(\boldsymbol{y}_{\text{obs}}, \boldsymbol{r})$.

3.2 *Choices for* $T(\boldsymbol{r}, \boldsymbol{r} \circ \boldsymbol{y})$

We discuss some choices of the summary function $T(\cdot)$ in the following. Functions of $\boldsymbol{r}$ relate to how well we model the missingness. Functions of $\boldsymbol{r} \circ \boldsymbol{y}$ relate to how well we model the observed y's including how likely that y was observed under the model. Some possible choices for $T(\boldsymbol{r}, \boldsymbol{r} \circ \boldsymbol{y})$ follow.

- $T_1(\boldsymbol{r}, \boldsymbol{r} \circ \boldsymbol{y}) = r_T y_T - r_1 y_1$; difference in mean of observed at end of study and observed at beginning of study
- $T_2(\boldsymbol{r}, \boldsymbol{r} \circ \boldsymbol{y}) = r_T(r_T y_T - r_1 y_1)$; observed change from baseline
- $T_3(\boldsymbol{r}, \boldsymbol{r} \circ \boldsymbol{y}) = \sum_{j=1}^{T} r_j$; number of observed components of $\boldsymbol{y}$
- $T_4(\boldsymbol{r}, \boldsymbol{r} \circ \boldsymbol{y}) = \frac{\sum_{j=1}^{T} r_j y_j}{\sum_{j=1}^{T} r_j}$; the mean of the observed responses.
- $T_5(\boldsymbol{r}, \boldsymbol{r} \circ \boldsymbol{y}) = \frac{\sum_{j=1}^{T} t_j r_j y_j}{\sum_j r_j t_j}$; the observed least square slopes
- $T_6(\boldsymbol{r}, \boldsymbol{r} \circ \boldsymbol{y}) = \sum_{j=1}^{T} \{\mathrm{I}(r_j = 1, r_{j+1} = 0) r_j y_j\} - \mathrm{I}(r_2 = 1) r_1 y_1$; change from baseline to last observed response under monotone missingness.
- $T_7(\boldsymbol{r}, \boldsymbol{r} \circ \boldsymbol{y}) = \{r_T(r_T y_T - r_1 y_1)\}^2$; second moment of difference in mean of observed at end of study and observed at beginning of study.
- $T_8(\boldsymbol{r}, \boldsymbol{r} \circ \boldsymbol{y}) = \left[\sum_{j=1}^{T} \{\mathrm{I}(r_j = 1, r_{j+1} = 0) r_j y_j\} - \mathrm{I}(r_2 = 1) r_1 y_1\right]^2$; second moment of change from baseline to last observed response under monotone missingness.

In the data analysis and simulations, we focus on $T_1(\cdot)$, $T_2(\cdot)$, $T_6(\cdot)$ and $T_8(\cdot)$.

3.3 *Computations*

Assume the model is parameterized via a vector of parameters, ω. Computation of the PPL criterion here can be done more efficiently using output from an MCMC algorithm when the following expectations can be expressed in closed form, $\mathrm{E}\{T^p(\boldsymbol{r}_{\text{rep}}, \boldsymbol{r}_{\text{rep}} \circ \boldsymbol{y}_{\text{rep}}) | \omega\} : p = 1, 2$.

This expectation corresponds to the following integral,

$$\int\int T^p(\boldsymbol{r}_{\text{rep}}, \boldsymbol{r}_{\text{rep}} \circ \boldsymbol{y}_{\text{rep}}) p(\boldsymbol{r}_{\text{rep}}, \boldsymbol{y}_{\text{rep}}|\omega) d\boldsymbol{r}_{\text{rep}} d\boldsymbol{y}_{\text{rep}}. \tag{3}$$

The availability of the expectation in closed form depends on both the model and the choice of $T(\cdot)$.

4. Data Example

We use the PPL criterion in Section 3.1 to select among models for data from a randomized clinical trial conducted to examine the effects of recombinant human growth hormone therapy for building and maintaining muscle strength in the elderly. The study, which we will refer to as GH, enrolled 161 participants and randomized them to one of four treatments arms. The response of interest here was mean quadriceps strength, measured as the maximum foot-pounds of torque that can be exerted against resistance provided by a mechanical device, which was recorded at baseline, 6 months, and 12 months. We restrict our analyses to only two of the treatment groups, Exercise + Growth Hormone (EG) and Exercise + Placebo (EP). Of the 78 randomized to these two arms, only 53 had complete follow-up (and the missingness was monotone); see Table S.1 in Web Appendix B.

Define $\boldsymbol{Y} = (Y_1, Y_2, Y_3)^T$ to be quad strength measured at months 0, 6, and 12 with corresponding observed data indicators, $\boldsymbol{R} = (R_1, R_2, R_3)^T$. In this data, the baseline quad strength is always observed, so $P(R_1 = 1) = 1$. Given that the dropout is monotone, without any loss of information, in specifying our models we replace $\boldsymbol{R}$ with $S = \sum_{j=1}^{3} R_j$ (the number of quad strength measures observed).

4.1 *Models Considered*

We considered both pattern mixture models and selection models to jointly model the distribution of the full data, $(\boldsymbol{y}, \boldsymbol{r})$. The mixture model we consider for each treatment

is specified as

$$\begin{aligned} Y_1|S=k &\sim \mathrm{N}(\mu_1^{(k)}, \sigma_1^{(k)}) : k=1,2,3 \\ Y_2|Y_1,S=k &\sim \mathrm{N}(\alpha_2+\phi_{21}Y_1, \tau_2) : k=1,2,3 \\ Y_3|Y_1,Y_2,S=k &\sim \mathrm{N}(\alpha_3+\phi_{31}Y_1+\phi_{32}Y_2, \tau_3) : k=1,,2,3 \\ S &\sim \mathrm{Mult}(\eta). \end{aligned} \tag{4}$$

The multinomial parameter is $\eta = (\eta_1, \eta_2, \eta_3)$, where $\eta_s = \mathrm{P(S{=}s)}$ and $\sum_s \eta_s = 1$. Recall that the PPL is invariant to the extrapolation distribution, i.e., the distributions $p(y_2|y_1, S=1)$ and $p(y_3|y_1, y_2, S=1)$ and $p(y_3|y_1, y_2, S=2)$. In the above, without loss of generality, we have set the parameters of the extrapolation distribution to their values under MAR.

We also consider a more parsimonious versions of the mixture model, MM2 which allows some equality of parameters between treatments. MM2 assumes the conditional distributions $[Y_3|Y_1, Y_2, S=j]$ and $[Y_2|Y_1, S=j]$ are same over the both treatments (i.e., the parameters $(\alpha_3, \phi_{31}, \phi_{32}, \tau_3, \alpha_2, \phi_{21}, \tau_2)$).

For the selection model, for each treatment, the full data response model is specified as

$$\begin{aligned} \boldsymbol{Y} &\sim \mathrm{N}(\mu, \Sigma) \\ R_2|\boldsymbol{y} &\sim \mathrm{Ber}(\pi_2) \\ R_3|R_2=1, \boldsymbol{y} &\sim \mathrm{Ber}(\pi_3), \end{aligned} \tag{5}$$

where $\mathrm{logit}(\pi_2) = \psi_{02} + \psi_1 Y_1 + \psi_2 Y_2$ and $\mathrm{logit}(\pi_3) = \psi_{03} + \psi_1 Y_2 + \psi_2 Y_3$. In the missing data mechanism in the selection model above, we have implicitly assumed non-future dependence (Kenward et al, 2003) and first order Markov dependence (constant over time). The former corresponds to the missingness at month j depending on the past and the potential response at month j, but not responses after month j. The latter corresponds to the dependence only depending on the immediate past (the previous visit time).

For both the mixture and selection models, we use diffuse priors for most of the parameters.

In particular, for the mean/regression parameters (μ, α, ϕ) in the mixture models we use normal priors with variances, $(10^6/10^4)$. For the variances (σ, τ), we use uniform priors with upper bound of 100. For the selection model, the marginal mean μ has a normal prior with variance 10^6, Σ^{-1} has a Wishart prior, and the parameters in the logistic model (ψ) for missingness have diffuse normal priors specified as the prior for μ except for ψ_2 which was given a normal prior with mean 0 and variance 5 (note that inferences were not sensitive to choices of the variance between 1 and 10). We chose a somewhat informative prior for ψ_2 for stability.

4.2 *Results*

We ran the Gibbs sampling algorithm in `WinBUGS` for 100K iterations. Trace plots suggested good mixing (not shown). We computed the PPL criterion for the four choices of $T(\cdot)$: $T_1(\boldsymbol{r}, \boldsymbol{r} \circ \boldsymbol{y}) = r_3 y_3 - r_1 y_1$, $T_2(\boldsymbol{r}, \boldsymbol{r} \circ \boldsymbol{y}) = r_T(r_T y_T - r_1 y_1)$, $T_6(\boldsymbol{r}, \boldsymbol{r} \circ \boldsymbol{y}) = \sum_{j=1}^{T} \{\mathrm{I}(r_j = 1, r_{j+1} = 0) r_j y_j\} - \mathrm{I}(r_2 = 1) r_1 y_1$, and $T_8(\boldsymbol{r}, \boldsymbol{r} \circ \boldsymbol{y}) = \left[\sum_{j=1}^{T} \{\mathrm{I}(r_j = 1, r_{j+1} = 0) r_j y_j\} - \mathrm{I}(r_2 = 1) r_1 y_1\right]^2$. Note that in Web Appendix C, we derive explicit forms for (3) for the some of the choices of $T(\cdot)$ considered here in the context of the model given in (4). There are not closed forms available for the selection model in (5).

Table 1 gives the PPL criterion values for the three models fit to the GH data for each of the four choices of $T(\cdot)$. All favor the selection model over the two mixture models. The selection model also had the smallest complexity (penalty) and a similar fit to the most complex mixture model (MM1).

We also computed DIC based on the observed data likelihood (see (6) in Section 5) for the three models. The results are presented in Table S.2 in web appendix B. DIC based on the observed data likelihood also favors the selection model.

5. Simulations

To assess the ability of the PPL to select the best model, we conducted several simulations. We simulated 200 datasets based on the parameter values given in Table 2 (these values are partially based on the GH data). We fit three models to data simulated under these same three models with sample sizes per treatment of 50, 100, and 2000. The three true models were MM1 and MM2 from Section 4 and the selection model from (5) with $\psi_2 = 0$. We denote this final model as SM0. To compare the models we used the proposed PPL criteria with the four different choices for $T(\boldsymbol{r}, \boldsymbol{r} \circ \boldsymbol{y})$ considered in Section 4.

We also computed the DIC based on the observed data likelihood, $\boldsymbol{L(\theta|y_{\text{obs}}, r)}$ to compare to the proposed criterion. We expect the DIC to be more powerful since it uses the entire likelihood, but for many models, such as selection models, its computation is quite burdensome, which discourages its use. The observed data likelihood DIC is defined as

$$\text{DIC}_O = -4E_{\theta|y,r}\left\{\log L(\theta|\boldsymbol{y}_{\text{obs}}, r)\right\} + 2\log L\{E_{\theta|y,r}(\theta)|\boldsymbol{y}_{\text{obs}}, r\}. \tag{6}$$

We put the restriction that $\psi_2 = 0$ from the selection model so that the DIC would be available in closed form.

The percentages of times the PPL and DIC_o criterion choose the true model are presented in Table 3. The average PPL values of several scenarios are presented in Tables 4-6. The detailed PPL and DIC_o results are reported in Web Appendix D, Tables S.3-S.12.

When MM1 was the true model, all the choices of $T(\cdot)$ did well and as the sample size increased, the probability of choosing the correct model approached one, with the least power for $T_1(\cdot)$ and much higher powers for the other choices.

When MM2 was the true model, it was chosen with probability of around 50% for the small and medium samples for all choices of $T(\cdot)$ expect for $T_6(\cdot)$ for which it was chosen with probability around 60%. For the largest sample size ($n = 2000$), it was picked approximately 50/50 with MM1. For all the sample sizes, the criterion gave very similar values under both

mixture models (see Table 6 and Tables S.10-S.12 in Web Appendix D). Note that when MM2 is the true model, both are correct since MM2 is nested in MM1. We discuss this further in the next section.

When the selection model was true, it was selected with high probability in non-large samples ($n = 50, 100$ per treatment arm), with probabilities $> 80\%$ (Table 3) for all choices of $T(\cdot)$. $T_2(\cdot)$ appeared to be the best discriminator among models for this setting, picking SM0 with probability $> 80\%$ for all sample sizes.

The DIC based on the observed data likelihood does very well in all situations though for comparing MM2 to MM1 under true MM2, the probability of choosing MM2 does not appear to be approaching one. The overall behavior is not surprising as it uses the data in the most efficient way in terms of comparing full probability models. However, as stated earlier, it is often a computational burden to implement it given the need to evaluate the observed data likelihood.

5.1 *Simulation conclusions*

In non-large samples ($n = 50, 100$), the criterion does a very good job selecting the best model with the specific performance depending on the choice of $T(\cdot)$ (Table 3).

For larger sample sizes ($n = 2000$), in most cases, the probability of selecting the correct model approaches one with an appropriately chosen $T(\cdot)$. However, for nested models, the criterion takes the same value for larger samples. As such, in this case, one might choose the more parsimonious model for final inference. Under SM0, when the wrong model was chosen with high probability, the PPL values were very similar (see Table S.6).

We also note that certain choices of $T(\cdot)$ do considerably better here, e.g., $T_2(\cdot)$ for true SM0 or true MM1. In general, we recommend similar choices for comparing SM's and MM's.

We also point out that for certain choices of $T(\cdot)$, the wrong model is selected in the larger sample sizes. However, this is arguably of less importance if $T(\cdot)$ is chosen as a function of

interest and the 'wrong' model provides a better (or equivalent) 'fit' to this function, which is the case when this happens. In such cases in the simulations, the actual PPL values were (essentially) the same.

For small to medium size samples, the PPL does a good job in choosing the correct model. In larger sample sizes (e.g., n=2000 per treatment arm), the computationally intensive DIC might sometimes be a better choice. In all the simulations, as the sample size increased, the probability of the DIC choosing the correct model was approaching one (noting that when MM2 is the true model, both MM1 and MM2 are the correct model). Once the 'best' model is chosen, the user would then conduct a sensitivity analysis (Daniels and Hogan, 2008) using the chosen model.

6. Discussion

We have proposed a computationally convenient way to compare models for incomplete longitudinal data that satisfies the property of being invariant to the specification of the extrapolation distribution (Property I). Via simulations, the proposed criterion appears to work well, especially for typical sample sizes of 50 to 100 subjects per treatment arm. Nevertheless, the DIC based on the observed data likelihood performs best, and may be preferred whenever it can be calculated. In other situations, for example when comparing selection models and/or shared parameter models, the PPL offers a computationally attractive alternative. Clearly, the choice of the summary $T(\cdot)$ affects the power and discriminative ability of the criterion. Care should be taken in choosing an appropriate summary $T(\cdot)$ (ideally based on a feature of the data of interest); however, the ability to choose a feature of interest allows more focused and targeted model selection based on a specific quantity of interest for inference. In future work, we will be exploring in more detail the best choices for $T(\cdot)$ for comparing different types of model for incomplete data.

It is also possible to use a Deviance based loss (Chen et al., 2004); however, the problem in

our case is the intractability of the observed data likelihood for many models for incomplete data and the same computational problems would arise as with DIC. The criteria proposed here is in the spirit of Ibrahim and Laud in that it measures discrepacy from the observed data (which here is $(\boldsymbol{r}, \boldsymbol{r} \circ \boldsymbol{y})$).

One issue with our approach is aliasing, i.e., small values of y being similar to ry when $r = 0$. However, we typically do not expect this to be a major issue, especially for continuous responses. For binary responses, coding the response as -1 and 1 (and similarly for categorical data in general) will alleviate problems; in addition, weighted versions of these criteria could also help (Chen et al., 2004). Moreover, it would be of interest to explore summary statistics such $T(\boldsymbol{r}, \boldsymbol{r} \circ \boldsymbol{y}) = a_1 t_1(\boldsymbol{r}) + a_2 t_2(\boldsymbol{r}, \boldsymbol{r} \circ \boldsymbol{y})$. However, these would need to be appropriately calibrated to ensure one of the two terms does not inadvertently dominate the criterion.

A general issue with posterior predictive based criteria is calibration. Calibration requires additional *straightforward* computations (see, e.g., Chen et al., 2004) and requires proper (informative) priors. However, the strategy from Chen et al. could be implemented in our setting with an appropriate choice of priors. For the simulation scenario of comparing the two mixture models where the more parsimonious model is true, calibration could be used to choose the simpler one. However, as pointed out earlier, in this setting, for larger sample sizes, we obtain (essentially) the same value of the criterion. And we also recall that in these incomplete settings, the ultimate goal is to choose a model and then do sensitivity analysis on this model. So to some extent, picking a good model (in terms of providing a good 'fit' to the quantity of interest, $T(\cdot)$), but not necessarily the correct model, can be sufficient.

Proving consistency of posterior predictive based criteria is difficult and specific to the model setting; Ibrahim et al. (2001) prove some results for linear models. For an appropriate choice of $T(\cdot)$ the PPL criterion appears to pick the correct model with probability going

to one in certain cases. We are currently working on analytical results to verify and better understand the behavior seen here; however, such derivations are very complex except for the simplest model settings. In particular, exploring the large sample behavior of the penalty term in these situations would be of major interest. It would also be of interest to examine more formally the large sample behavior of the DIC, in particular for nested model settings.

A general issue in model selection for incomplete longitudinal data is comparing ignorable and non-ignorable models; for the former $p(\boldsymbol{r}|\boldsymbol{y}) = p(\boldsymbol{r}|\boldsymbol{y}_{\text{obs}})$ is not explicitly modeled. It is not clear that such model comparisons can be made based on a criterion that satisfies Property I. This is also related to posterior predictive checks based on replicated observed data versus replicated complete data the latter which was explored in Gelman et al. (2005). Dobson and Henderson (2003) proposed exploratory residuals for the response conditional on not dropping out. However, both of these approaches focus on graphical and exploratory model checking, not formal model comparison.

In Section 1, we describe how model selection criterion for incomplete data should satisfy Property I. However, there may be situations where external information is available about the distribution of the full data response such that this property might become less important.

Ibrahim et al. (2008) recently considered frequentist methods for the computation of model selection criteria in missing-data problems based on output of the EM algorithm in a frequentist setting. They developed a class of information criteria for missing-data problems. The general form satisfies the property of being invariant to the distribution of the missing data conditional on the observed data (more detail in Section 3). However, they need an analytic approximation to compute this (similar problem to not having closed form for the observed data likelihood). The simpler form they propose that does not require the approximation does not satisfy the frequentist version (no priors) of Property I.

Acknowledgments

This research was supported by NIH R01 CA85295. We thank Joe Hogan for helpful discussions over the years on this topic.

Supplementary Materials

Web Appendices A-D, referenced in Sections 2, 3, 4 and 5, are available with this paper at the Biometrics website on Wiley Online Library.

References

B. Carlin and S. Chib. Bayesian Model Choice via Markov Chain Monte Carlo Methods. *Journal Of Royal Statistical Society, Series B*, 57:473–484, 1995.

G. Celeux, F. Forbes, C. Robert, and M. Titterington. Deviance Information Criteria for Missing Data Models. *Bayesian Analysis*, 1:651–674, 2006.

M. Chen, D. Dey, and J. Ibrahim. Bayesian Criterion based Model Assessment for Categorical Data. *Biometrika*, 91:45–63, 2004.

S. Chib and I. Jeliazkov. Marginal Likelihood from the Metropolis-Hastings Output. *Journal Of the American Statistical Association*, 96:270–281, 2001.

S. Chib and I. Jeliazkov. Accept Reject Metropolis Hastings sampling and Marginal Likelihood Estimation. *Statistica Nederlandica*, 59:30–44, 2005.

M. Daniels and J. Hogan. Reparameterizing the Pattern Mixture Model for Sensitivity Analyses under Informative Dropout. *Biometrics*, 56:1241–1248, 2000.

M. Daniels and J. Hogan. *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*. Chapman & Hall, 2008.

P. Diggle and M. Kenward. Informative drop-out in Longitudinal Data Analysis. *Applied Statistics*, 43:49–93, 1994.

A. Dobson and R. Henderson. Diagnostics for Joint Longitudinal and Dropout Time Modeling. *Biometrics*, 59:741–751, 2003.

G. Fitzmaurice, G. Molenberghs, and S. Lipsitz. Regression Models for Longitudinal Binary Responses with Informative Drop-outs. *Journal Of Royal Statistical Society, Series B*, 57:691–704, 1995.

S. Geisser and W. Eddy. A Predictive approach to Model Selection. *Journal of the American Statistical Association*, 74:153–160, 1979.

A. Gelfand and S. Ghosh. Model Choice: A Minimum Posterior Predictive Loss Approach. *Biometrika*, 85:1–11, 1998.

A. Gelman, X. Meng, and H. Stern. Posterior Predictive Assessment of Model Fitness via Realized Discrepancies. *Statistical Sinica*, 6:733–807, 1996.

A. Gelman, I. Mechelen, G. Verbeke, D. Heitjan, and M. Meulders. Multiple Imputation for Model Checking: Completed-Data Plots with Missing and Latent Data. *Biometrics*, 61: 74–85, 2005.

J. Heckman. The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Independent Variables and a Simple Estimator for such Models. *Annals of Economic and Social Measurement*, 5:120–137, 1976.

J. Ibrahim and P. Laud. A Predictive Approach to the Analysis of Designed Expreiments. *Journal Of the American Statistical Association*, 89:309–319, 1994.

J. Ibrahim, M. Chen, and D. Sinha. Criterion-Based Methods for Bayesian Model Assessment. *Statistical Sinica*, 11:419–443, 2001.

J. Ibrahim, H. Zhu, and N. Tang. Model Selection Criteria for Missing-Data Problems using the EM Algorithm. *Journal Of the American Statistical Association*, 103:1648–1658, 2008.

V. Johnson. Bayes Factors based on Test Statistics. *Journal Of Royal Statistical Society,*

Series B, 67:689–701, 2005.

V. Johnson and J. Hu. Bayesian Model Selection using Test Statistics. *Journal Of Royal Statistical Society, Series B*, 71:143–158, 2009.

R. Kass and A. Raftery. Bayes Factors. *Journal Of the American Statistical Association*, 90:773–795, 1995.

M. Kenward, G. Molenberghs, and H. Thijs. Pattern-Mixture Models with Proper Time Dependence. *Biometrika*, 90:53–71, 2003.

P. Laud and J. Ibrahim. Predictive Model Selection. *Journal Of Royal Statistical Society, Series B*, 57:247–262, 1995.

R. Little. A Class of Pattern-Mixture Models for Normal Incomplete Data. *Biometrika*, 81: 471– 483, 1994.

G. Molenberghs and M. Kenward. *Missing Data in Clinical Trials*. Wiley, 2007.

D. Rizopoulos, G. Verbeke, and G. Molenberghs. Shared Parameter Models under Random Effects Misspecication. *Biometrika*, 95:63 – 74, 2008.

D. Spiegelhalter, N. Best, B. Carlin, and A. Van Der Linde. Bayesian Measures of Model Complexity and Fit. *Journal Of Royal Statistical Society, Series B*, 64:583–639, 2002.

C. Wang and M. Daniels. A Note on MAR, Identifying Restrictions, and Sensitivity Analysis in Pattern Mixture Models with and without Covariates for Incomplete Data. *Biometrics*, 67:810–818, 2011.

M. Wu and R. Carroll. Estimation and Comparison of Changes in the Presence of Informative Right Censoring by Modeling the Censoring Process. *Biometrics*, 44:175 – 188, 1988.

Table 1

PPL criterion for the three models fit to the growth hormone data: Selection model (SM), Mixture model 1 (MM1), and Mixture Model 2 (MM2) for four choices of $T(\boldsymbol{r}, \boldsymbol{r} \circ \boldsymbol{y})$. C_∞ is the criterion with $k = \infty$ and GOF is the goodness of fit component of the criterion.

Model	GOF	Complexity	C_∞
$T(\boldsymbol{r}, \boldsymbol{r} \circ \boldsymbol{y}) = r_T y_T - r_1 y_1$			
SM	2960.2	2907.6	5867.8
MM1	2961.7	3958.6	6920.3
MM2	3058.3	3498.5	6556.8
$T(\boldsymbol{r}, \boldsymbol{r} \circ \boldsymbol{y}) = r_T(r_T y_T - r_1 y_1)$			
SM	390.7	425.2	815.9
MM1	390.2	517.8	907.9
MM2	484.7	605.7	1090.3
$T(\boldsymbol{r}, \boldsymbol{r} \circ \boldsymbol{y}) = \sum_{j=1}^{T} \{\mathrm{I}(r_j = 1, r_{j+1} = 0) r_j y_j\} - \mathrm{I}(r_2 = 1) r_1 y_1$			
SM	1670.5	1759.7	3430.2
MM1	1670.0	2211.4	3881.4
MM2	1768.1	2606.3	4374.4
$T(\boldsymbol{r}, \boldsymbol{r} \circ \boldsymbol{y}) = \left[\sum_{j=1}^{T} \{\mathrm{I}(r_j = 1, r_{j+1} = 0) r_j y_j\} - \mathrm{I}(r_2 = 1) r_1 y_1\right]^2$			
SM	15563039	11655064	27218103
MM1	15712294	23472467	39184760
MM2	15760469	22043555	37804025

Table 2

Parameter settings of MAR Selection model (SM0), Mixture model 1 (MM1), and Mixture Model 2 (MM2) for Simulation Study in Section 5.

Arm	Parameter	Values
SM0		
1	μ_1, μ_2, μ_3	11,12,9
1	$\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_{12}, \sigma_{13}, \sigma_{23}$	7,7,5,4,3,4
1	$\phi_{02}, \phi_{03}, \phi_1$	0.9, 1.5, -0.25
2	μ_1, μ_2, μ_3	8,11,10
2	$\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_{12}, \sigma_{13}, \sigma_{23}$	7,13,13,7,8,12
2	$\phi_{02}, \phi_{03}, \phi_1$	0.3, 0.9, -0.25
MM1		
1	$P(S=1)$, $P(S=2)$, $P(S=3)$	0.15, 0.25, 0.6
1	$\mu_1^{(1)}$,$\mu_1^{(2)}$, $\mu_1^{(3)}$	20, 30, 27
1	$\sigma_1^{(1)}$,$\sigma_1^{(2)}$, $\sigma_1^{(3)}$	2, 1.5, 2
1	α_2, ϕ_{21}, α_3, ϕ_{31}, ϕ_{32}	2, 0.9, 3, 1, 1.1
1	τ_2, τ_3	2, 3
2	$P(S=1)$, $P(S=2)$, $P(S=3)$	0.15, 0.2, 0.6
2	$\mu_1^{(1)}$,$\mu_1^{(2)}$, $\mu_1^{(3)}$	22, 32, 28
2	$\sigma_1^{(1)}$,$\sigma_1^{(2)}$, $\sigma_1^{(3)}$	2, 1.5, 2
2	α_2, ϕ_{21}, α_3, ϕ_{31}, ϕ_{32}	4, 0.2, -5, 0.9, 1.3
2	τ_2, τ_3	2, 3
MM2		
1,2	parameters in treatment arm 1 of MM1	

Table 3

Number of times (out of 200) the PPL and DIC_o (observed data likelihood DIC) criterion choose the true model when fitting one of the following three models: MAR Selection model (SM0), Mixture model 1 (MM1), and Mixture Model 2 (MM2) for four choices of $T(\boldsymbol{r}, \boldsymbol{r} \circ \boldsymbol{y})$: $T_1(\boldsymbol{r}, \boldsymbol{r} \circ \boldsymbol{y}) = r_T y_T - r_1 y_1$, $T_2(\boldsymbol{r}, \boldsymbol{r} \circ \boldsymbol{y}) = r_T(r_T y_T - r_1 y_1)$, $T_6(\boldsymbol{r}, \boldsymbol{r} \circ \boldsymbol{y}) = \sum_{j=1}^{T}\{\mathrm{I}(r_j = 1, r_{j+1} = 0) r_j y_j\} - \mathrm{I}(r_2 = 1) r_1 y_1$, and $T_8(\boldsymbol{r}, \boldsymbol{r} \circ \boldsymbol{y}) = \left[\sum_{j=1}^{T}\{\mathrm{I}(r_j = 1, r_{j+1} = 0) r_j y_j\} - \mathrm{I}(r_2 = 1) r_1 y_1\right]^2$.

True Model	Size	Model	T_1	T_2	T_6	T_8	DIC_o
SM0	50	SM0	193	198	192	194	200
		MM1	7	0	2	3	0
		MM2	0	2	6	3	0
SM0	100	SM0	161	197	158	168	199
		MM1	39	3	2	20	1
		MM2	0	0	40	12	0
SM0	2000	SM0	16	165	47	113	200
		MM1	184	35	1	74	0
		MM2	0	0	152	13	0
MM1	50	SM0	117	4	21	21	0
		MM1	83	196	179	179	200
		MM2	0	0	0	0	0
MM1	100	SM0	111	0	2	0	0
		MM1	89	200	198	200	200
		MM2	0	0	0	0	0
MM1	2000	SM0	79	0	0	0	0
		MM1	121	200	200	200	200
		MM2	0	0	0	0	0
MM2	50	SM0	29	0	5	7	0
		MM1	91	98	98	78	40
		MM2	80	102	97	115	160
MM2	100	SM0	5	0	0	0	0
		MM1	87	90	100	72	46
		MM2	108	110	100	128	154
MM2	2000	SM0	0	0	0	0	0
		MM1	101	110	106	103	57
		MM2	99	90	94	97	143

Table 4

Simulating (true) model Mixture model 1 (MM1) and sample size 2000: average PPL criteria over 200 replications for four choices of $T(\boldsymbol{r}, \boldsymbol{r} \circ \boldsymbol{y})$ for models MAR Selection model (SM0), Mixture model 1 (MM1), and Mixture Model 2 (MM2). C_∞ is the PPL criterion with $k = \infty$ and GOF is the goodness of fit component of the criterion.

Model	GOF	Complexity	C_∞
$T(\boldsymbol{r}, \boldsymbol{r} \circ \boldsymbol{y}) = r_T y_T - r_1 y_1$			
SM0	1114.3	1114.1	2228.4
MM1	1113.2	1113.7	2226.8
MM2	1601.5	1644.6	3246.1
$T(\boldsymbol{r}, \boldsymbol{r} \circ \boldsymbol{y}) = r_T(r_T y_T - r_1 y_1)$			
SM0	270.3	291.6	561.9
MM1	270.0	270.2	540.2
MM2	758.6	873.7	1632.3
$T(\boldsymbol{r}, \boldsymbol{r} \circ \boldsymbol{y}) = \sum_{j=1}^{T} \{\mathrm{I}(r_j = 1, r_{j+1} = 0) r_j y_j\} - \mathrm{I}(r_2 = 1) r_1 y_1$			
SM0	418.9	445.6	864.5
MM1	417.5	417.7	835.2
MM2	1074.9	1354.4	2429.3
$T(\boldsymbol{r}, \boldsymbol{r} \circ \boldsymbol{y}) = \left[\sum_{j=1}^{T} \{\mathrm{I}(r_j = 1, r_{j+1} = 0) r_j y_j\} - \mathrm{I}(r_2 = 1) r_1 y_1\right]^2$			
SM0	299940	350146	650086
MM1	298273	298677	596950
MM2	1019002	2908665	3927667

Table 5

Simulating (true) model MAR Selection model (SM0) and sample size 100: average PPL criteria over 200 replications for four choices of $T(\boldsymbol{r}, \boldsymbol{r} \circ \boldsymbol{y})$ for models MAR Selection model (SM0), Mixture model 1 (MM1), and Mixture Model 2 (MM2). C_∞ is the PPL criterion with $k = \infty$ and GOF is the goodness of fit component of the criterion.

Model	GOF	Complexity	C_∞
$T(\boldsymbol{r}, \boldsymbol{r} \circ \boldsymbol{y}) = r_T y_T - r_1 y_1$			
SM0	41.4	41.7	83.2
MM1	41.5	43.2	84.7
MM2	41.9	47.6	89.5
$T(\boldsymbol{r}, \boldsymbol{r} \circ \boldsymbol{y}) = r_T(r_T y_T - r_1 y_1)$			
SM0	8.7	8.9	17.6
MM1	8.7	9.4	18.1
MM2	9.2	10.0	19.2
$T(\boldsymbol{r}, \boldsymbol{r} \circ \boldsymbol{y}) = \sum_{j=1}^{T} \{\mathrm{I}(r_j = 1, r_{j+1} = 0) r_j y_j\} - \mathrm{I}(r_2 = 1) r_1 y_1$			
SM0	30.5	30.8	61.4
MM1	30.5	33.1	63.6
MM2	31.1	31.7	62.8
$T(\boldsymbol{r}, \boldsymbol{r} \circ \boldsymbol{y}) = \left[\sum_{j=1}^{T} \{\mathrm{I}(r_j = 1, r_{j+1} = 0) r_j y_j\} - \mathrm{I}(r_2 = 1) r_1 y_1\right]^2$			
SM0	2122	2140	4263
MM1	2124	2566	4690
MM2	2127	2611	4739

Table 6

Simulating (true) model Mixture model 2 (MM2) and sample size 2000: average PPL criteria over 200 replications for four choices of $T(\boldsymbol{r}, \boldsymbol{r} \circ \boldsymbol{y})$ for models MAR Selection model (SM0), Mixture model 1 (MM1), and Mixture Model 2 (MM2). C_∞ is the PPL criterion with $k = \infty$ and GOF is the goodness of fit component of the criterion.

Model	GOF	Complexity	C_∞
$T(\boldsymbol{r}, \boldsymbol{r} \circ \boldsymbol{y}) = r_T y_T - r_1 y_1$			
SM0	1669.6	1699.4	3369.0
MM1	1668.4	1668.3	3336.7
MM2	1668.4	1668.5	3337.0
$T(\boldsymbol{r}, \boldsymbol{r} \circ \boldsymbol{y}) = r_T(r_T y_T - r_1 y_1)$			
SM0	511.4	552.0	1063.3
MM1	511.0	511.2	1022.3
MM2	511.1	511.2	1022.3
$T(\boldsymbol{r}, \boldsymbol{r} \circ \boldsymbol{y}) = \sum_{j=1}^{T} \{\mathrm{I}(r_j = 1, r_{j+1} = 0) r_j y_j\} - \mathrm{I}(r_2 = 1) r_1 y_1$			
SM0	409.3	462.0	871.3
MM1	409.1	409.1	818.3
MM2	409.1	409.3	818.4
$T(\boldsymbol{r}, \boldsymbol{r} \circ \boldsymbol{y}) = \left[\sum_{j=1}^{T} \{\mathrm{I}(r_j = 1, r_{j+1} = 0) r_j y_j\} - \mathrm{I}(r_2 = 1) r_1 y_1\right]^2$			
SM0	497319	580996	1078315
MM1	494065	494774	988839
MM2	494143	494568	988711