Published in final edited form as:

Comput Stat Data Anal. 2009 September 1; 53(11): 3765–3772. doi:10.1016/j.csda.2009.03.024.

Bayesian Model Checking for Multivariate Outcome Data

Catherine M. Crespi^{a,*} and W. John Boscardin^b

^aDepartment of Biostatistics, University of California, Los Angeles, CA, USA

^bDivisions of Geriatrics and Biostatistics, University of California, San Francisco, CA, USA

Abstract

Bayesian models are increasingly used to analyze complex multivariate outcome data. However, diagnostics for such models have not been well-developed. We present a diagnostic method of evaluating the fit of Bayesian models for multivariate data based on posterior predictive model checking (PPMC), a technique in which observed data are compared to replicated data generated from model predictions. Most previous work on PPMC has focused on the use of test quantities that are scalar summaries of the data and parameters. However, scalar summaries are unlikely to capture the rich features of multivariate data. We introduce the use of dissimilarity measures for checking Bayesian models for multivariate outcome data. This method has the advantage of checking the fit of the model to the complete data vectors or vector summaries with reduced dimension, providing a comprehensive picture of model fit. An application with longitudinal binary data illustrates the methods.

Keywords

Bayesian analysis; dissimilarity measures; longitudinal data; multivariate data; posterior predictive model checking

1 Introduction

It is common in medicine, social sciences and other fields to collect outcome data with a multivariate structure. Common models for multivariate outcome data include the multinomial model for categorical data (Agresti, 1990), the multivariate normal (Laird and Ware, 1982) and *t*-distribution (Pinheiro et al., 2001) models for continuous data, and a wide variety of models for longitudinal data (Diggle et al., 2002), time series (Hamilton, 1994), spatial data (Cressie, 1993) and microarray data (Speed, 2003), among others. Bayesian methods have been increasingly employed for inference with such data (Congdon, 2001; Gelman et al., 2004). However, model checking and diagnostics remain underdeveloped for such models.

Posterior predictive model checking (PPMC) is a broadly applicable diagnostic method for Bayesian models in which the observed data are compared to the posterior predictive distribution of replicated data under the posited model (Rubin, 1984; Gelman et al., 1996, 2004). If the posited model accurately represents the process that generated the data, then replicated data generated under the model should look similar to the observed data. PPMC is usually implemented by drawing simulated values from the posterior predictive distribution of replicated data and comparing these samples to the observed data using test quantities that characterize important features of the data. Differences between the test quantities for the observed and replicated data indicate model misfit.

*Corresponding author. *Email address:* ccrespi@ucla.edu (Catherine M. Crespi).

The intuitive appeal, flexibility and straightforward implementation of PPMC have made it a popular tool used in a wide variety of fields; see, for example, Duffull et al. (2005); Lewin et al. (2006); Sinharay et al. (2006). However, previous work has focused on test quantities that are scalar summaries of the data and parameters. While scalar summaries may be adequate for univariate outcome data, they are unlikely to capture the rich features of multivariate outcome data.

To address this need, we propose a method of checking Bayesian models for multivariate outcome data that directly evaluates the fit of the model to the data vectors themselves or to vector summaries of the data with reduced dimension. In our approach, vectors of observed data and replicated data are compared using dissimilarity measures that quantify the proximity of two vector quantities. In particular, pairwise dissimilarities between the observed and replicated data, and among the replicated data, are computed. Comparison of the distributions of the “between” and “among” dissimilarity measures provide insights about model fit, with divergence indicating lack of fit. Nonparametric test procedures are used to obtain p -values that can serve as numerical summaries of lack of fit and be used to diagnose the source of model misfit, leading to improved model specification. Other approaches to PPMC for multivariate outcome data have focused on either informative graphical displays of multivariate data summaries (Gelman et al., 2000) or posterior predictive p -values for scalar data summaries (Carlin et al., 2001).

We illustrate our methods using data collected as part of a clinical trial for treatment of a viral infection (Wald et al., 1996, 1997). In the trial, subjects infected with herpes simplex virus (HSV) collected mucosal secretion specimens daily for periods of several months. The specimens were assayed for the presence or absence of HSV DNA, resulting in longitudinal binary data showing a pattern of recurrent HSV shedding. Longitudinal binary data of this type are common in biomedical and epidemiological studies that involve the repeated administration of a diagnostic or screening procedure on the same subjects. We evaluate the fit of three different models for the data, including a semi-Markov model proposed for the data in Crespi et al. (2005). The models include hierarchical structure to accommodate subject-to-subject heterogeneity.

The paper is organized as follows. In Section 2, we review and present notation for posterior predictive model checking. Our motivating example is discussed in Section 3. The proposed methods are presented in Section 4, and illustrated on the application in Section 5. We close with a discussion.

2 Posterior Predictive Model Checking

Presentations of posterior predictive model checking can be found in Rubin (1984), Gelman et al. (1996) and Gelman et al. (2004). Let $p(\mathbf{y}|\psi)$ denote the likelihood for a model, where $\mathbf{y}$ denotes the data and ψ denotes all of the parameters in the model, including all hyperparameters if the model is hierarchical. The posterior distribution of ψ is $p(\psi|\mathbf{y}) \propto p(\mathbf{y}|\psi)p(\psi)$, where $p(\psi)$ denotes the prior distribution of the parameters. Posterior predictive model checking compares the observed data $\mathbf{y}$ to the posterior predictive distribution of replicated data $\mathbf{y}^{rep}$,

$$p(\mathbf{y}^{rep}|\mathbf{y}) = \int p(\mathbf{y}^{rep}|\psi)p(\psi|\mathbf{y})d\psi. \quad (1)$$

The replicated data $\mathbf{y}^{rep}$ may be regarded as the data we would expect to see if the model and values of ψ were correct; that is, $\mathbf{y}^{rep}$ is a replication that should look like $\mathbf{y}$ if the model is a good fit.

This leads to the motivation to compare $\mathbf{y}$ and $\mathbf{y}^{rep}$ to evaluate the fit of the model and perform diagnostics. In PPMC as it is commonly used, the observed and replicated data are compared using test quantities $T(\mathbf{y}, \psi)$, also called discrepancy measures, that are scalar summaries of the data and parameters (Gelman et al., 1996, 2004). The realized values of the test quantity for the observed data, $T(\mathbf{y}, \psi)$, are compared to the posterior predictive distribution of $T(\mathbf{y}^{rep}, \psi)$, with differences indicating lack of fit. Comparisons may be made using graphical plots, and the lack of fit of the data with respect to the posterior predictive distribution may be quantified by the tail-area probability

$$P(T(\mathbf{y}^{rep}, \psi) \geq T(\mathbf{y}, \psi) | \mathbf{y}) = \iint_{T(\mathbf{y}^{rep}, \psi) \geq T(\mathbf{y}, \psi)} p(\mathbf{y}^{rep} | \psi) p(\psi | \mathbf{y}) d\mathbf{y}^{rep} d\psi, \quad (2)$$

referred to as the posterior predictive p -value (Gelman et al., 1996). In some cases the sign of the test quantity may need to be changed in order to convert p -values close to one to p -values close to zero.

The posterior predictive distribution is usually computed using simulation (Rubin, 1984; Gelman et al., 1996, 2004). After fitting the model, one draws M simulations $\psi^1, \psi^2, \dots, \psi^M$ from the posterior distribution $p(\psi | \mathbf{y})$ and then simulates one replicated data set for each draw using the likelihood $p(\mathbf{y} | \psi^m)$, $m = 1, 2, \dots, M$. This process results in M draws, $\mathbf{y}^{rep,1}, \mathbf{y}^{rep,2}, \dots, \mathbf{y}^{rep,M}$, from the joint posterior distribution $p(\mathbf{y}^{rep}, \psi | \mathbf{y})$. One then computes and compares the realized test quantities, $T(\mathbf{y}, \psi^m)$, and the predictive test quantities, $T(\mathbf{y}^{rep,m}, \psi^m)$. The estimated posterior predictive p -value is the proportion of the M simulations for which $T(\mathbf{y}^{rep,m}, \psi^m) \geq T(\mathbf{y}, \psi^m)$, $m = 1, 2, \dots, M$. A model is suspect if the posterior predictive p -value is close to zero. Theoretical difficulties in the interpretation of these posterior predictive p -values has motivated a spirited discussion in the recent literature. We summarize some of this debate in the context of our approach in Section 6.

Test quantities may be chosen that depend on the data only, so that $T(\mathbf{y}, \psi) = T(\mathbf{y})$, in which case $T(\mathbf{y})$ is called a test statistic (Gelman et al., 2004). For example, one might choose as a test quantity the smallest observation in the data set. In this case, $T(\mathbf{y}, \psi^m) = T(\mathbf{y})$ is the smallest observation in the observed data (identical for all m) and the $T(\mathbf{y}^{rep,m}, \psi^m)$ are the $m = 1, 2, \dots, M$ smallest observations from each of the $m = 1, 2, \dots, M$ replicated data sets. In this case, the M draws of replicated data are from $p(\mathbf{y}^{rep} | \mathbf{y})$ and the p -value is computed using $T(\mathbf{y}^{rep,m}) \geq T(\mathbf{y})$.

This procedure for conducting PPMC requires specification of a test quantity, which ideally should be relevant to the inferences and scientific purposes of the model (Rubin, 1984; Gelman et al., 1996, 2004). One may evaluate a model with respect to several different test quantities that reflect different features of the data.

Next we use our motivating example to illustrate how this framework for model checking has shortcomings in the multivariate data context.

3 Motivating Example

Individuals infected with herpes simplex virus experience recurrent episodes of shedding of infectious virus from the mucosa (Corey and Wald, 1999). A focus of scientific inquiry is the frequency and duration of HSV shedding episodes, during which an individual is presumed to be infectious to others. Wald et al. (1996, 1997) report a crossover clinical trial in which longitudinal data on HSV shedding were collected from a sample of women with genital HSV infection. In the study, the subjects collected secretion specimens from the genital mucosa once

daily for periods of several months. The specimens were assayed for the presence (+) or absence (−) of HSV DNA, resulting in longitudinal binary data showing a pattern of recurrent HSV shedding. The assay results for Subject 21, for example, are

[illegible]

where the NA's represent missing data. We take our data from the placebo arm of the trial, during which the subjects were not taking any antiviral therapy and exhibited their natural pattern of shedding. Very little shedding was detected under treatment, making the treatment arm data uninteresting from a modeling perspective. Our data set consists of 21 subjects on placebo with daily specimens for 57–82 days.

These data show only the shedding state of a subject at discrete time points. The number and duration of a subject's shedding episodes cannot be determined directly from these data since the onset and resolution times of the episodes are not recorded, the data are censored at the beginning and end, and there is intermittent missing data. Crespi et al. (2005) proposed a semi-Markov model for the HSV shedding process that allows one to infer the frequency and duration of shedding episodes from the data. The model assumes that stimuli initiating a shedding state arrive according to a homogeneous Poisson process and each stimuli is associated with a period in the shedding state, with durations independently and identically distributed as exponential. Periods of shedding caused by independent stimuli may overlap, resulting in sojourns in the shedding state that depend on time-in-state. The result is a semi-Markov process which can be fit using a hidden Markov modeling approach (MacDonald and Zucchini, 1997).

This model proposes mechanisms that generate the observed longitudinal patterns of HSV shedding, and it is important to assess the extent to which the model accurately captures these patterns. Crespi et al. (2007) report posterior predictive model checking of the model using scalar summaries as test statistics, as well as further application to data analysis. The test statistics were percent time with mucosal shedding and the number and average length of HSV + runs. The PPMC found no evidence of model inadequacy, other than a slight tendency of the model to underpredict the number of HSV+ runs for individuals with very frequent runs. However, much information about the shedding patterns of subjects was lost in summarizing the data using these scalar quantities. For example, the data for Subject 21 (who was not in the data set for Crespi et al. 2007) would have been summarized as 44% time shedding, 7 runs and average length of 5.14 days. There are many possible patterns of HSV shedding that are consistent with these values. These scalars do not capture distinguishing features of the pattern of shedding such as the wide range of run lengths, from 1–16 days. It is possible that the model could fit these scalar summaries but provide a poor fit to the overall pattern of shedding.

A richer summary of the data that captures salient features of a shedding pattern is the entire distribution of positive run lengths. For example, the data for Subject 21 could be summarized as one run of length 1, three runs of length 2, etc, up to one run of length 16. Although some information is lost using the run length distribution as a summary of the data, it preserves much of the clinical information of interest on the frequency and duration of shedding episodes. Thus a more rigorous model checking procedure, and one that is more closely aligned with scientific objectives, would be to check the fit of the model to the complete distribution of run lengths, rather than to summary statistics such as the average run length.

Using the run length distribution involves summarizing the data as ordered vectors indicating the frequency of positive runs of various lengths. In the following, we show how model checking can be conducted using vector summaries of the data rather than scalars.

4 Methods

Consider data $\mathbf{y}_1, \dots, \mathbf{y}_n$ consisting of n random vectors of correlated random variables, which may be of differing lengths. Through dimension reduction or another summarizing operation, these vectors may be represented by the vectors $\mathbf{g}_1, \dots, \mathbf{g}_n$ of length p . For clarity in distinguishing the observed and replicated data, we will use the notation

$$(\mathbf{y}_1, \dots, \mathbf{y}_n) = (\mathbf{y}_1^{obs}, \dots, \mathbf{y}_n^{obs}) \text{ and } (\mathbf{g}_1, \dots, \mathbf{g}_n) = (\mathbf{g}_1^{obs}, \dots, \mathbf{g}_n^{obs}).$$

In the following, we apply the methods to $\mathbf{g}_1^{obs}, \dots, \mathbf{g}_n^{obs}$; however, the methods can be applied directly to $\mathbf{y}_1^{obs}, \dots, \mathbf{y}_n^{obs}$.

We are concerned with measuring the similarity or dissimilarity of vector quantities. The well-established methodology of cluster analysis (Everitt et al., 2001) provides a framework for doing so. We will work with distances between vectors. A real-valued function $d(\mathbf{g}, \mathbf{h})$, mapping $R^p \times R^p$ onto R^1 , is a distance function if it has the following properties (Mardia et al., 1979):

- (1) symmetry: $d(\mathbf{g}, \mathbf{h}) = d(\mathbf{h}, \mathbf{g})$;
- (2) non-negativity: $d(\mathbf{g}, \mathbf{h}) \geq 0$;
- (3) identification mark: $d(\mathbf{g}, \mathbf{g}) = 0$.

If d also satisfies two additional properties:

- (4) definiteness: $d(\mathbf{g}, \mathbf{h}) = 0$ if and only if $\mathbf{g} = \mathbf{h}$; and
- (5) triangle inequality: $d(\mathbf{g}, \mathbf{h}) \leq d(\mathbf{g}, \mathbf{k}) + d(\mathbf{k}, \mathbf{h})$ for all $\mathbf{g}, \mathbf{h}, \mathbf{k}$,

then it is called a metric. One would expect $d(\mathbf{g}, \mathbf{h})$ to increase as “dissimilarity” between $\mathbf{g}$ and $\mathbf{h}$ increases even when d does not satisfy the metric properties (4) and (5). In cluster analysis, $d(\mathbf{g}, \mathbf{h})$ with d satisfying only (1)–(3) is called a coefficient of dissimilarity or dissimilarity measure (Mardia et al., 1979). This specification allows distances functions that have $d(\mathbf{g}, \mathbf{h}) = 0$ for some $\mathbf{g} \neq \mathbf{h}$, violating property (4), which may be appropriate for some types of data. Regarding property (5), Sibson (1972) argues that in the comparison of distances, order relationships are more important than numerical values, and a monotonic increasing function of a dissimilarity measure might perform just as well as the original measure. Since a monotonic transformation of a metric need not satisfy the triangle inequality, e.g., the square of the Euclidean metric (Seber, 1984), the triangle inequality is not an important requirement for measuring dissimilarity.

Our approach will be to compare the observed and replicated data using a suitable distance function. There is a large literature on distance functions; see, e.g., Everitt et al. (2001); Mardia et al. (1979). One should be chosen as appropriate for the data. Familiar functions include

Euclidean distance, $d(\mathbf{g}, \mathbf{h}) = \sqrt{\sum_{i=1}^p (g_i - h_i)^2}$, and Manhattan distance,

$d(\mathbf{g}, \mathbf{h}) = \sum_{i=1}^p |g_i - h_i|$. In the case of multivariate normal data, one might choose to use

Mahalanobis distance, $d(\mathbf{g}, \mathbf{h}) = \sqrt{(\mathbf{g} - \mathbf{h})^T \Sigma^{-1} (\mathbf{g} - \mathbf{h})}$. One may also devise a new distance function to meet the objectives of the analysis.

Having specified the distance function, to proceed, we simulate from the posterior predictive distribution of replicated data as described in Section 2, drawing M simulations of parameters $\psi^1, \psi^2, \dots, \psi^M$ from the posterior distribution $p(\psi|y)$, and simulating one replicated data set for each draw, obtaining replicated data sets $(y_1^{rep,m}, y_2^{rep,m}, \dots, y_n^{rep,m}), m = 1, 2, \dots, M$. Next, we obtain summary vectors of each replicated data set, $(g_1^{rep,m}, g_2^{rep,m}, \dots, g_n^{rep,m}), m = 1, 2, \dots, M$. We then compute the dissimilarity measures. For each subject i in our data set, we consider the collection of $M+1$ vectors $(g_i^{rep,1}, g_i^{rep,2}, \dots, g_i^{rep,M}, g_i^{obs})$, and compute the pairwise dissimilarities among these vectors. There will be M values, $d(g_i^{obs}, g_i^{rep,m}), m = 1, 2, \dots, M$, of dissimilarities between the observed and replicated summary vectors, and $M(M-1)/2$ values of dissimilarities $d(g_i^{rep,r}, g_i^{rep,s}), r \neq s$, among the summary vectors of replicated data.

If the model is consistent with the data, we expect the distances between the vector summaries of observed and replicated data $\{d^{obs,m}\}$ and the distances among the replicated data vector summaries $\{d^{rep,m}\}$ to largely coincide. If the model is a poor fit, we expect $\{d^{obs,m}\}$ to be stochastically higher than $\{d^{rep,m}\}$. To formally compare the two sets, we can consider an appropriate hypothesis test. To choose a suitable test, it is important to consider that distance values are bounded below by zero. Thus the distribution of distances taking small values (as expected for $\{d^{rep,m}\}$) are likely to be right-skewed, and the distribution of distances taking larger values ($\{d^{obs,m}\}$ in the case of a poor-fitting model) are likely to be less skewed due to displacement away from zero. This is illustrated in Figure 1, which is discussed as part of the application in the following section. Based on this consideration and the fact that the dissimilarity measures may be ordinal in nature, it would appear sensible to use a test for a robust measure of central tendency such as the median, which leads to the Mann-Whitney-Wilcoxon procedure (see, for example, Gibbons 1997) as an appropriate choice. Since we expect a shift of location in one direction only (the distribution of $\{d^{obs,m}\}$ cannot be expected to be lower than that of $\{d^{rep,m}\}$), a one-sided alternative hypothesis is indicated.

The Mann-Whitney-Wilcoxon procedure can be applied to the sets of dissimilarity measures from each subject to obtain subject-specific p -values. These p -values are useful numerical summaries of model fit, and they can be used for model diagnostics by, for example, plotting them against salient covariates or the data vectors themselves in a search for characteristics of the data that are poorly reflected in the model specification. We illustrate these techniques in the application.

5 Application

We conduct PPMC using the new methods for three candidate models for the HSV longitudinal binary data, the semi-Markov model proposed in Crespi et al. (2005) and two simpler models. The three candidate models are increasingly complex and allow for increasing levels of autocorrelation in the positive test results. To accommodate subject heterogeneity, each model includes a hierarchical structure in which each subject's parameters are drawn from a population distribution.

Binomial Model

In this model, each test result is assumed to be independent, with probability π_i of a positive result for individual i . For the population distribution we use

$$\text{logit}(\pi_i) \sim N(\mu, \tau) \quad (4)$$

with independent prior distributions $\mu \sim N(b, c)$, where c is the variance, and $\tau \sim \text{Gamma}(u, v)$. This model would be plausible in cases in which the intervals between assays are long and

thus autocorrelation is not apparent. However, in our case, the data show obvious autocorrelation, and we use this model to illustrate the results when the model is a poor fit to the data. The model was fit in WinBUGS Version 1.4 (Spiegelhalter et al., 2003).

Markov Process

In this model, individuals are assumed to alternate between shedding and non-shedding states according to a two-state continuous-time Markov process (Ross, 2003). The longitudinal binary data are considered to arise from sampling the underlying states at discrete time points. Let μ_{1i} be the intensity of leaving the non-shedding state and μ_{2i} be the intensity of leaving the shedding state. For the population distribution we use

$$\begin{bmatrix} \log(\mu_{1i}) \\ \log(\mu_{2i}) \end{bmatrix} \sim N(\boldsymbol{\theta}, \boldsymbol{\tau}) \quad (5)$$

with independent prior distributions $\boldsymbol{\theta} \sim N(\mathbf{b}, \mathbf{C})$, where $\mathbf{C}$ is a covariance matrix, and $\boldsymbol{\tau} \sim \text{Wishart}(v, \mathbf{R})$, where v denotes the degrees of freedom and $\mathbf{R}$ is a scale matrix. In this model, the frequency of shedding episodes for patient i is estimated as $\mu_{1i}\mu_{2i}/(\mu_{1i} + \mu_{2i})$ and the mean duration is estimated as $1/\mu_{2i}$. This model was also fit in WinBUGS Version 1.4.

Semi-Markov Process

In the semi-Markov model (Crespi et al., 2005), stimuli creating a shedding state arrive according to a homogeneous Poisson process with intensity λ_{1i} for patient i and each stimuli is associated with a period in the shedding state, with durations independently and identically distributed as exponential with rate parameter λ_{2i} . This model is distinguished from the two-state Markov model in that the periods of shedding caused by independent stimuli may overlap, resulting in sojourns in the shedding state that depend on time-in-state. This model allows for more autocorrelation in the positive test results than the Markov model. For the population distribution we use

$$\begin{bmatrix} \log(\lambda_{1i}) \\ \log(\lambda_{2i}) \end{bmatrix} \sim N(\boldsymbol{\theta}, \boldsymbol{\tau}) \quad (6)$$

with independent prior distributions $\boldsymbol{\theta} \sim N(\mathbf{b}, \mathbf{C})$ and $\boldsymbol{\tau} \sim \text{Wishart}(v, \mathbf{R})$. Procedures for fitting this model, including a Gibbs sampler for sampling from the posterior distribution, are described in Crespi et al. (2005). In this model, the frequency of shedding episodes for patient i is estimated as $e^{\lambda_{1i}/\lambda_{2i}}/\lambda_{1i}$ and the mean duration is $(e^{\lambda_{1i}/\lambda_{2i}} - 1)/\lambda_{1i}$. The model was fit using a C language program.

To put all three models on an equal footing with regard to prior information, we used the same historical data set to specify the priors for each model. The historical data set was from a previous study with a similar protocol; from this previous study, we used data from the subset of individuals who shared common characteristics with our analysis data set (female, HSV-1 seronegative and HSV-2 seropositive). This historical data comprised 10 individuals, with sequences of data ranging from 49 to 63 daily assays.

To obtain the historical priors, our approach was to fit each of the models to the historical data using vague priors and find approximate analytical distributions for the resulting posteriors. After inflating the variance by a factor of 2, these analytical distributions served as the historical prior distributions for the main analysis. Thus for Model 1, we fit the density $N(\mu|b, c)$ to obtain the posterior means $\hat{b}$ and $\hat{c}$, from which we obtained $\hat{b}_{hist} = \hat{b}$ and $\hat{c}_{hist} = 2\hat{c}$. For Models 2 and

3, we used the same approach for the multivariate normal densities $N(\mathbf{b}, \mathbf{C})$. For Model 1, we obtained $\hat{u}_{hist}$ and $\hat{v}_{hist}$ by fitting $\text{Gamma}(\tau|\mu, \nu)$ then doubling the variance. For Models 2 and 3, a good prior guess of the covariance matrix was required for adequate performance of the Gibbs sampler. Thus we used the posterior mean of τ to obtain $\mathbf{R}_{hist}$ as $\hat{v}\hat{\tau}^{-1}$. When fitting the main data set using these historical priors, we used $\nu = 5$, which discounts the number of historical subjects by half. To fit each model, we ran the Gibbs sampler for 20,000 iterations and discarded the first 5,000 iterations as burn-in. Convergence was attained within 4,000 iterations or less for all three models.

For the posterior predictive model checking, we generated 500 sets of replicated data for each model. For each subject i , $i = 1, 2, \dots, 21$, we drew a set of subject-specific parameters from the posterior (π_i for Model 1, (μ_{1i}, μ_{2i}) for Model 2, and $(\lambda_{1i}, \lambda_{2i})$ for Model 3) and generated a replicate sequence of binary data using the specified likelihood. When simulating replicated data, we conditioned on the number and timing of each subject's assays, including missed time points, so that the number and timing replicate the experimental conditions. The number and timing of assays constitute auxiliary statistics, functions of the data that are held constant in replications (Gelman et al., 1996). This yielded 500 sets of replicated data,

$\mathbf{y}_i^{rep,1}, \mathbf{y}_i^{rep,2}, \dots, \mathbf{y}_i^{rep,500}$, for each subject.

Next we reduced the data to run length distributions, expressed as ordered vectors of frequencies of positive run lengths. This summarizing operation is aligned with our scientific focus on shedding episode frequency and duration. The data reduction operation maps a sequence of binary observations $\mathbf{y}$ to a vector $\mathbf{g}$ of length p , where the first element of $\mathbf{g}$, g_1 , is the number of runs of positive test results of length 1, the second element g_2 is the number of runs of length 2, etc., and the p th element g_p is the number of runs of length p , where p is the length of the longest HSV+ run in the data set $\{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n\}$. In our data set, we had $p = 16$, and the data for Subject 21 (presented in Equation 3), for example, are summarized as $\mathbf{g}_{21} = (1, 3, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1)$, where the first element of the vector indicates the number of runs of length 1, the second element indicates the number of runs of length 2, etc.

Thus from the replicated data $\mathbf{y}_i^{rep,1}, \mathbf{y}_i^{rep,2}, \dots, \mathbf{y}_i^{rep,500}$, we obtained the ordered vectors of run lengths $\mathbf{g}_i^{rep,1}, \mathbf{g}_i^{rep,2}, \dots, \mathbf{g}_i^{rep,500}$, and from the observed data for subject i , we obtained the ordered vector $\mathbf{g}_i^{obs}$, $i = 1, \dots, 21$.

The next step was to choose a function to measure the distance between vectors. For ordered vectors, suitable distance functions include cumulative Euclidean distance

$$d(\mathbf{g}, \mathbf{h}) = \sqrt{\sum_{i=1}^p \left(\sum_{j=1}^i g_j - \sum_{j=1}^i h_j \right)^2} \quad (7)$$

and landmover distance $d(\mathbf{g}, \mathbf{h}) = \sum_{i=1}^p \left| \sum_{j=1}^i g_j - \sum_{j=1}^i h_j \right|$. Cumulative Euclidean and landmover distance both measure the spatial concentration of the values in the vector, and the order of the values affects the distance. For ordered vectors, these distance functions are likely to provide better results than standard Euclidean and Manhattan distances (Kamarainen et al., 2003).

We used cumulative Euclidean distance to calculate pairwise dissimilarities between the ordered vectors. Distances between vectors for the observed and replicated data for each subject

i were calculated as $d_i^{obs,m} = d(\mathbf{g}_i^{obs}, \mathbf{g}_i^{rep,m})$, $m = 1, 2, \dots, 500$, and distances among vectors for replicated data were calculated as $d(\mathbf{g}_i^{rep,r}, \mathbf{g}_i^{rep,s})$, $r \neq s$.

Figure 1 presents the distributions of the dissimilarity measures for the three models for Subject 21. For all three models, the distances $\{d^{rep,m}\}$ among the replicated data summary vectors are clustered near the origin, reflecting the similarity among these vectors within each model. For the binomial model, the distances $\{d^{obs,m}\}$ between the observed and replicated data are large and well-separated from $\{d^{rep,m}\}$, indicating gross model misfit. In contrast, for Models 2 and 3, the two sets of distances largely coincide, indicating a lack of evidence of model misfit. The Mann-Whitney-Wilcoxon procedure yields p -values of $<.00001$, $.974$, and $.999$ for Models 1, 2 and 3, respectively.

Evaluation of the model fit on a subject-by-subject basis such as in Figure 1 can identify subjects for whom the model is a poor fit. Further diagnostics to find the characteristics of the data that are poorly reflected in the model specification can be performed by using the subject-specific p -values as numerical summaries of model fit. These p -values can be compared to the data vectors or vector summaries, or plotted against ordered covariate values, as a method of diagnosing the source of the model misfit. Figure 2 provides the p -values for the 21 subjects along with their vector summaries. The subjects are in order of increasing proportion of days HSV+, which ranges from zero for Subjects 1–3 to 0.44 for Subject 21. The figure shows that all three models have an acceptable fit for Subjects 1–3, who have no HSV+ runs, and Subjects 4–7, who have only single-day runs. For all other subjects, the binomial model is clearly inadequate, which we expected due to the autocorrelation in the data.

For both the Markov and semi-Markov models, the fit of the models is questionable for Subjects 12, 13 and 16. Inspection of the run length distributions for these subjects shows that they are bimodal, having only very short (1- and 2-day) and very long (6+-day) runs, and none of intermediate length. This suggests a model inadequacy that may merit a re-specification (e.g., a mixture model for shedding duration). There is no evidence of lack of fit of the semi-Markov model for the other 18 subjects, however, and some evidence for a better fit of the semi-Markov model over the Markov model for several subjects. There is no evidence of differential quality of fit based on proportion of days HSV+ for either the Markov or semi-Markov models, suggesting this aspect of the data is well-reflected in both models. Overall, the results suggest that the semi-Markov model provides a somewhat better fit to these data than the simpler Markov model, but the semi-Markov model might also be improved upon.

6 Discussion

We have presented a general method of checking the fit of a Bayesian model for multivariate data using posterior predictive model checking. The method uses distance functions to quantify the dissimilarity of the observed and replicated data and posterior predictive p -values as numerical summaries of model misfit. Other approaches to PPMC for multivariate outcomes data include Gelman et al. (2000) which focuses on direct graphical display of replicated residual plot-type summaries and Carlin et al. (2001) which examines a set of scalar discrepancy measures for longitudinal binary data.

Although our illustration of the technique compares three different models, PPMC is primarily a technique for checking the adequacy of a model, which can be distinguished from model selection, in which a number of candidate models are compared. Useful model selection tools for Bayesian models include the deviance information criterion (DIC) (Spiegelhalter et al., 2002) and partial Bayes factors (O'Hagan, 2019501). PPMC and model selection can be viewed as complementary: PPMC answers the question, "Is the model adequate?" while model selection addresses the question, "Is Model 1 better than Model 2?" Thus our method is not

designed to replace the use of formal model selection measures, but rather to be used as part of a larger model building and selection process that could include the use of DIC, partial Bayes factors or other techniques.

We have used our method to evaluate model fit on a subject-by-subject basis. Our approach is analogous to other subject-by-subject diagnostic techniques such as residuals analysis. Our data set of 21 subjects is relatively small; however, there is no conceptual impediment to using this method on large data sets, since the computed p -values for hundreds or thousands of subjects could be easily screened for extreme observations. On the other hand, simulation from the posterior predictive distribution could be computationally intensive for some models, which could make the technique time-consuming for large data sets. This is a limitation of PPMC in general rather than our method in particular, however.

In this paper we have not addressed model checking specifically for the hierarchical structure of the model. Several authors (for example, Dey et al. 1998; Sinharay and Stern 2003) have addressed this issue. Sinharay and Stern (2003) found that it is very difficult to detect violations of the assumptions made about the population distribution unless the extent of the violation is huge or the observed data have small standard errors.

There is debate as to the appropriateness of using the frequentist idea of a p -value in the context of Bayesian model checking; an informative view of this debate is given in O'Hagan and Forster (2004, Section 8.36). Arguments in favor are based on both pragmatism and a philosophical sense that model criticism falls outside a strictly Bayesian framework. Arguments against note that frequentist methods were not designed for Bayesian model criticism and their interpretation is unclear. In particular, it is not clear how small a p -value needs to be in order to warn of misspecification; conventional frequentist thresholds such as $p = .05$ do not have an obvious rationale in this context. Adding to the debate, posterior predictive checks have been criticized as being conservative in the sense that they are biased against detecting departures from the model (Dey et al., 1998). Robins et al. (2000) show that posterior predictive p -values are conservative in that their asymptotic distributions are more concentrated around $1/2$ than a uniform distribution, and suggest ways to modify the p -values to make their distributions asymptotically uniform. Draper and Krnjajic (2007) describe a highly computationally intensive simulation-based method for adjusting posterior predictive p -values. Gelman (2007) suggests that careful graphical display of replicated multivariate data summaries is more appropriate and informative than focusing on posterior predictive p -values. The method that we have presented takes a pragmatic view of the posterior predictive p -values as indicators of model misfit for particular subjects; this view is consistent with the applied use of regression diagnostics. Furthermore, we agree with the view in Gelman (2007) and have attempted to produce useful graphical displays of the PPMC results. Bayesian model diagnostics is a rapidly developing field, and future developments may lead to better resolution of the pragmatic issues, such as p -value calibration and interpretation, if not the philosophical issues. Despite these issues, PPMC is straightforward to conduct and gives practical insight into model misspecification, making it a useful tool for practitioners in an area for which diagnostic tools are lacking.

Acknowledgments

The authors thank the associate editor and two anonymous reviewers for thoughtful comments and suggestions on an earlier version of this article.

This work was supported by NIH grants T32 AI007370 and NCI CA016042 (CMC), and NIAID AI28697 (WJB).

References

- Agresti, A. *Categorical Data Analysis*. New York, New York: John Wiley and Sons; 1990.
- Carlin J, Brown H, Gelman A, Wolfe R. Alternative models for longitudinal binary outcome data: a case study on their choice, interpretation and checking. *Biostatistics* 2001;2:397–416. [PubMed: 12933632]
- Congdon, P. *Bayesian Statistical Modelling*. West Sussex, England: John Wiley and Sons; 2001.
- Corey, L.; Wald, A. Genital herpes. In: Holmes, KK., et al., editors. *Sexually Transmitted Diseases*. New York: McGraw-Hill; 1999. p. 285–312.
- Crespi CM, Cumberland WG, Blower S. A queueing model for chronic recurrent conditions under panel observation. *Biometrics* 2005;61:194–199.
- Crespi CM, Cumberland WG, Wald A, Corey L, Blower S. Longitudinal study of herpes simplex virus type 2 infection using viral dynamic modeling. *Sexually Transmitted Infections* 2007;83:359–364. [PubMed: 17475687]
- Cressie, NAC. *Statistics for Spatial Data*. New York, N. Y.: John Wiley and Sons; 1993.
- Dey DK, Gelfand AE, Swartz TB, Vlachos PK. A simulation-intensive approach for checking hierarchical models. *Test* 1998;7(2):323–344.
- Diggle, P.; Heagerty, P.; Liang, K-Y.; Zeger, S. *Analysis of Longitudinal Data*. 2nd Edition. Oxford, U.K.: Oxford University Press; 2002.
- Draper, D.; Krnjajic, M. *Bayesian Model Specification*. University of California, Santa Cruz: Tehnical Report ams2007–09, Department of Mathematics and Statistics; 2007. available at <http://ams.ucsc.edu/reports/trview.php>.
- Duffull SB, Kirkpatrick CMJ, Green B, Holford NHG. Analysis of population pharmacokinetic data using NONMEM and WinBUGS. *Journal of Biopharmaceutical Statistics* 2005;15(1):53–73. [PubMed: 15702605]
- Everitt, BS.; Landau, S.; Leese, M. *Cluster Analysis*. 4th Edition. Arnold Publishing; 2001.
- Gelman A. Comment on “Bayesian checking of the second level of hierarchical models: Cross-validated posterior predictive checks using discrepancy measures” by M. J. Bayarri and M. E. Castellanos. *Statistical Science* 2007;22:349–352.
- Gelman, A.; Carlin, JB.; Stern, HS.; Rubin, DB. *Bayesian Data Analysis*. Boca Raton: Chapman & Hall/CRC; 2004.
- Gelman A, Goegebeur Y, Tuerlinckx F, Mechelen IV. Diagnostic checks for discrete-data regression models using posterior predictive simulations. *Applied Statistics* 2000;49:247–268.
- Gelman A, Meng X-L, Stern H. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica* 1996;66:733–807.
- Gibbons, J. *Nonparametric Methods for Quantitative Analysis*. Columbus, Ohio: American Sciences Press, Inc.; 1997.
- Hamilton, JD. *Time Series Analysis*. Princeton, N. J.: Princeton University Press; 1994.
- Kamarainen J-K, Kyrki V, Ilonen J, Kalviainen H. Improving similarity measures of histograms using smoothing projections. *Pattern Recognition Letters* 2003;24:2009–2019.
- Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics* 1982;38:963–974. [PubMed: 7168798]
- Lewin A, Richardson S, Marshall C, Glazier A, Aitman T. Bayesian modeling of differential gene expression. *Biometrics* 2006;62(1):1–9. [PubMed: 16542224]
- MacDonald, IL.; Zucchini, W. *Hidden Markov and Other Models for Discrete-Valued Time Series*. New York: Chapman and Hall; 1997.
- Mardia, KV.; Kent, JT.; Bibby, JM. *Multivariate Analysis*. London: Academic Press; 1979.
- O’Hagan A. Fractional bayes factors for model comparison. *Journal of the Royal Statistical Society Series B* 57:99–138. 20199501.
- O’Hagan, A.; Forster, J., editors. *Bayesian Inference*. London: Arnold Publishers; 2004.
- Pinheiro JC, Liu CH, Wu YN. Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. *Journal of Computational and Graphical Statistics* 2001;10:249–276.

- Robins JM, van der Vaart A, Ventura V. The asymptotic distribution of p-values in composite null models. *Journal of the American Statistical Association* 2000;95:1143–1172.
- Ross, SM. *Introduction to Probability Models*. San Diego: Academic Press; 2003.
- Rubin DB. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics* 1984;12:1151–1172.
- Seber, GAF. *Multivariate Observations*. New York: John Wiley and Sons; 1984.
- Sibson R. Order invariance methods for data analysis. *Journal of the Royal Statistical Society Series B* 1972;34:311–349.
- Sinharay S, Johnson MS, Stern HS. Posterior predictive assessment of item response theory models. *Applied Psychological Measurement* 2006;30(4):298–321.
- Sinharay S, Stern HS. Posterior predictive model checking in hierarchical models. *Journal of Statistical Planning and Inference* 2003;11(1–2):209–221.
- Speed, T., editor. *Statistical Analysis of Gene Expression Microarray Data*. Boca Raton, FL: Chapman Hall/CRC CRC Press; 2003.
- Spiegelhalter D, Best N, Carlin B, van der Linde A. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B* 2002;64:583–639.
- Spiegelhalter, D.; Thomas, A.; Best, N.; Lunn, D., editors. *WinBUGS Version 1.4 Users Manual*. Cambridge, U.K.: MRC Biostatistics Unit; 2003.
- Wald A, Corey L, Cone R, Hobson A, Davis G, Zeh J. Frequent genital herpes virus 2 shedding in immunocompetent women: effect of acyclovir treatment. *Journal of Clinical Investigation* 1997;99:1092–1097. [PubMed: 9062368]
- Wald A, Zeh J, Barnum G, Davis LG, Corey L. Suppression of subclinical shedding of herpes simplex virus type 2 with acyclovir. *Annals of Internal Medicine* 1996;124:8–15. [PubMed: 7503497]

Figure 1.

Histograms showing the distribution of dissimilarity measures computed between observed and replicated data, $\{d^{obs,m}\}$, and among replicated data, $\{d^{rep,m}\}$. The figures show the results for all three models for Subject 21. In the Markov and semi-Markov models, the two histograms largely coincide.

Figure 2.

P-values based on Mann-Whitney-Wilcoxon tests of the null hypothesis that the distributions of $\{d^{obs,m}\}$ and $\{d^{rep,m}\}$ are equal for the 21 subjects. The frequency distribution of each subject's positive run lengths is provided below the plot. The subjects are in order of increasing

proportion of positive days. *P*-values were logit transformed as $\ln\left(\frac{p+.01}{1-p+.01}\right)$. The complete observed run length distributions were used in the PPMC but are truncated here at 6+ for ease of presentation.