Bayesian Analysis (2007) **2**, Number 4, pp. 735–760

Bayesian Diagnostic Techniques for Detecting Hierarchical Structure

Guofen Yan* and J. Sedransk†

Abstract. Motivated by an increasing number of Bayesian hierarchical model applications, the objective of this paper is to evaluate properties of several diagnostic techniques when the fitted model includes some hierarchical structure, but the data are from a model with additional, unknown hierarchical structure. Because there has been no apparent evaluation of Bayesian diagnostics used for this purpose, we start by studying the simple situation where the data come from a normal model with two-stage hierarchical structure while the fitted model does not have any hierarchical structure, and then extend this to the case where the fitted model has two-stage normal hierarchical structure while the data come from a model with three-stage normal structure. We use exact derivations, large sample approximations and numerical examples to evaluate the quality of the diagnostic techniques. Our investigation suggests two promising techniques: distribution of individual posterior predictive p values and the conventional posterior predictive p value with the F statistic as a checking function. We show that (at least) for large sample sizes these p values are uniformly distributed under the null model and are effective in detecting hierarchical structure not included in the null model. Finally, we apply these two techniques to examine the fit of a model for data from the Patterns of Care Study, a two-stage cluster sample of cancer patients undergoing radiation therapy.

Keywords: F statistic, model assessment, partial posterior predictive p value, posterior predictive distribution, posterior predictive p value.

1 Introduction

Often, a model that is fitted does not account for all of the hierarchical structure actually present. For example, the problem which motivated this study was making inference about age specific mortality rates for all cancer for white males (Nandram, Sedransk and Pickle 1999) and for chronic obstructive pulmonary disease (Nandram, Sedransk and Pickle 2000). In these investigations there were hierarchical models for the rates using the 798 Health Service Areas as the basic geographical units. However, there are many ways to choose the geographical units and it is not at all clear how many hierarchical levels are appropriate. In this paper we evaluate the ability of several Bayesian methods to identify unanticipated hierarchical structure. We consider model diagnostics rather than formal comparisons of alternative models.

*Division of Biostatistics and Epidemiology, Department of Public Health Sciences, University of Virginia, Charlottesville, VA, mailto:guofen.yan@virginia.edu

†Department of Statistics, Case Western Reserve University, Cleveland, OH, mailto:jxs123@case.edu

Preliminary work using the models in Nandram, Sedransk and Pickle (1999, 2000) clearly indicated that they were too complex to yield useful analytical results about the properties of the diagnostic techniques. Since there has been little research in this area, starting with a much simpler specification seemed sensible. That is, we assume that the data come from a model with a two-stage hierarchical normal structure while the fitted model does not have this hierarchical structure. We then extend our investigation to the case where the data come from a model with a three-stage normal structure, but the fitted model uses only two stages. It is encouraging that many of our findings are common to both cases, indicating that they are likely to apply to the situation where the fitted model includes some hierarchical structure but does not account for all of the stages actually present.

We use exact derivations, large sample approximations and numerical examples to evaluate two promising Bayesian diagnostic techniques. Our principal example is the Patterns of Care Study, a two-stage cluster sample of patients undergoing radiation therapy in 1978 for cervix cancer. The variable we examine is a (transformed) score measuring the quality of the workup received by a patient: See Calvin and Sedransk (1991) for additional details.

Section 2 describes our proposed evaluation method, and the models and diagnostic techniques under investigation. In Sections 3 and 4 we apply the evaluation method to study the quality of the two Bayesian diagnostic techniques in the important situation where the fitted model ignores hierarchical structure actually present. The results of our analysis of the Patterns of Care data are presented in Section 5. There is further discussion in Section 6, tying our results to the recent literature. Concluding remarks are in Section 7.

2 Models, diagnostic techniques and evaluation method

We first study the simple situation where the data come from a normal model with two-stage hierarchical structure while the fitted model does not have any hierarchical structure. We then extend this so that the fitted model has two-stage normal hierarchical structure while the data come from a model with three-stage normal structure.

Case I: Fit the non-hierarchical model

For the first case it is assumed that the fitted model has $E(y_{ij}) = \mu$ and $\text{var}(y_{ij}) = \phi$, i.e.,

$$\begin{aligned} y_{ij} \mid \mu, \phi &\overset{iid}{\sim} N(\mu, \phi)\,, \quad i = 1, \ldots, m\,,\ j = 1, \ldots, n\,; \\ \pi(\mu, \phi) &\propto \text{constant}. \end{aligned} \tag{2.1}$$

The data, $y = \{y_{ij} : i = 1, \ldots, m\,,\ j = 1, \ldots, n\}$, are assumed to come from the null model, (2.2), or from one of two alternative models, (2.3) and (2.4).

- Single level model

$$y_{ij} \mid \mu_0, \phi_0 \overset{iid}{\sim} N(\mu_0, \phi_0)\,,\ i = 1, \ldots, m\,,\ j = 1, \ldots, n\,. \tag{2.2}$$

- Two-level hierarchical model

$$\begin{aligned} y_{ij} \mid \theta_i, \phi_0 &\overset{iid}{\sim} N(\theta_i, \phi_0), \\ \theta_i \mid \mu_0, d_0 &\overset{iid}{\sim} N(\mu_0, d_0)\,,\ i = 1, \ldots, m\,,\ j = 1, \ldots, n\,. \end{aligned} \tag{2.3}$$

- Two-level hierarchical model with changepoint

$$\begin{aligned} y_{ij} \mid \theta_i, \phi_0 &\overset{iid}{\sim} N(\theta_i, \phi_0)\,,\ j = 1, \ldots, n, \\ \theta_i \mid \mu_{01}, d_0 &\overset{iid}{\sim} N(\mu_{01}, d_0)\,,\ i = 1, \ldots, g\,, \\ \theta_i \mid \mu_{02}, d_0 &\overset{iid}{\sim} N(\mu_{02}, d_0)\,,\ i = g+1, \ldots, m\,. \end{aligned} \tag{2.4}$$

The three models, (2.2)-(2.4), represent a range of deviations from the fitted model, (2.1). Since it is easy to detect deviations when the fitted model and the one generating the data are very different we emphasize situations where the fitted and actual models are somewhat similar. The fitted model, (2.1), is appropriate for data generated from (2.2), the null model, but not for data generated from (2.3) or (2.4): The data from (2.2) are iid observations with mean μ_0 and variance ϕ_0. In (2.3), $\text{var}(y_{ij}) = \phi_0 + d_0$, $\text{cov}(y_{ih}, y_{i'k}) = 0$ and $\text{cov}(y_{ih}, y_{ik}) = d_0$, thus $\text{cor}(y_{ih}, y_{ik}) = d_0/(\phi_0 + d_0)$. That is, the between-group observations are independent while the within-group observations are correlated. In (2.4) the data are also correlated, but some group means are clustered with expected value μ_{01} while the other means are clustered with expected value μ_{02}. If μ_{01} and μ_{02} are far apart, the overall variation of the data will be large.

Case II: Fit the two-stage hierarchical model

For the second case the two-stage hierarchical model that is fitted is

$$\begin{gathered} y_{ijk} \mid \mu_i, \phi \overset{iid}{\sim} N(\mu_i, \phi), \\ \mu_i \mid \nu, \gamma \overset{iid}{\sim} N(\nu, \gamma)\,, \\ \pi(\nu) \propto \text{constant}, \\ i = 1, \ldots, a\,,\ j = 1, \ldots, m\,,\ k = 1, \ldots, n\,. \end{gathered} \tag{2.5}$$

To permit analytical results the variances, ϕ and γ, are assumed known unless specified otherwise.

We consider the following two models to have generated the data $y = \{y_{ijk} : i = 1, \ldots, a, j = 1, \ldots, m, k = 1, \ldots, n\}$: The first is the null model while the second includes additional correlation structure.

- Two-level hierarchical model

$$\begin{aligned} y_{ijk} \mid \mu_i, \phi_0 &\overset{iid}{\sim} N(\mu_i, \phi_0), \\ \mu_i \mid \nu_0, \gamma_0 &\overset{iid}{\sim} N(\nu_0, \gamma_0)\,. \end{aligned} \tag{2.6}$$

- Three-level hierarchical model

$$\begin{aligned} y_{ijk} \mid \theta_{ij}, \phi_0 &\overset{iid}{\sim} N(\theta_{ij}, \phi_0), \\ \theta_{ij} \mid \mu_i, d_0 &\overset{iid}{\sim} N(\mu_i, d_0)\,, \\ \mu_i \mid \nu_0, \gamma_0 &\overset{iid}{\sim} N(\nu_0, \gamma_0)\,. \end{aligned} \tag{2.7}$$

All of the Bayesian model diagnostic techniques that we consider use the posterior distribution of predicted data under the fitted model (2.1) or (2.5). In general terms this density function is

$$p(\tilde{y}|y) = \int p(\tilde{y}|\theta)\, p(\theta|y)\, d\theta \tag{2.8}$$

where y are the observed data, $\tilde{y}$ the predicted data, θ the parameters of the fitted model and, given θ, y and $\tilde{y}$ are assumed to be independent. Early work in this area includes Box (1980), Rubin (1984) and Geisser (1993) who proposed diagnostics of this type.

Over the last ten years there has been important research evaluating posterior predictive p values and presenting alternatives (e.g., Bayarri and Berger 2000, Bayarri and Castellanos 2007, Hjort, Dahl and Steinbakk 2006, and Robins, van der Vaart and Ventura 2000). An objective of this research is evaluation of a procedure so that it has known (and desired) properties when data come from the "null model". For example, it is desirable that if a procedure is based on (2.1) and the data come from (2.2), the null model, the posterior predictive p value should be uniformly distributed. This line of research includes evaluation of existing procedures to investigate whether they are properly calibrated and to suggest and evaluate alternative methods.

Since the techniques we present use posterior predictive p values, we do this evaluation and show that the distribution of these p values is uniform (for moderate sample sizes) when the data are from a null model. Specifically, we first evaluate a diagnostic quantity, $d(y)$, under the assumption that the actual distribution of y is *consistent* with that postulated for the fitted model - to show the desired uniform distribution. Then we evaluate $d(y)$ assuming that the actual distribution of y is *different* from that of the fitted model - to investigate the conditions under which the diagnostic will reveal the existence (and nature) of a model different from that postulated for the fitted model. This means, for example, for Case I where the non-hierarchical model is fitted, the properties are evaluated under (2.2), (2.3) and (2.4).

We show in Section 6.2 why our findings are consistent with the results in Bayarri and Berger (2000) and Robins et al. (2000).

The two techniques that we investigate in detail are briefly described below. (This description is for the first case; the extension to the second case is straightforward.) We only consider diagnostic techniques that permit the use of noninformative prior distributions because this is, by far, the most common choice of prior, and introducing a general prior would make it impossible to draw useful general results. Bayarri and Castellanos (2007) comment that:" i) it might not be desired to carefully quantify a prior in these earlier stages of the analysis, since the model might well not be appropriate and hence the effort is wasted; ii) most importantly, model checking with informative priors can not separate inadequacy of the prior from inadequacy of the model."

Given the structure of the models in (2.2)-(2.4) and (2.6)-(2.7) we emphasize diagnostic techniques that are potentially useful for detecting incorrect specification of the variance (covariance) structure. First, consider the *ensemble* of $N = mn$ individual posterior predictive p values

$$p_{ij} = P(\tilde{y}_{ij} \leq y_{ij} \mid y)\,; \;\; i = 1, \ldots, m\,, \;\; j = 1, \ldots, n \tag{2.9}$$

evaluated under (2.1) (with a corresponding definition for p_{ijk} evaluated under (2.5)). Also, define

$$p^*_{ij} = P(\tilde{y}_{ij} \leq y_{ij} \mid y_{(ij)}) \tag{2.10}$$

where $y_{(ij)}$ denotes all of the data except for unit ij. Gelfand, Dey and Chang (1992) suggest that the set of N values of p^*_{ij} in (2.10) will be approximately uniformly distributed if the fitted model is the same as the model generating the data - with departures from uniformity indicating a lack of fit.

Our data models, (2.2)-(2.4), do not include outliers because we want to focus on detection of unanticipated hierarchical structure, not on detection of outliers: Including outliers in our models would make our conclusions much less clear. Since we do not include outliers in y, using p_{ij} in (2.9) rather than p^*_{ij} in (2.10) yields essentially the same results, as one may expect; see Yan (2003). Thus, we prefer to use (2.9) rather than (2.10) which permits us to obtain clearer analytical results.

In applications, using (2.10) is likely to be preferable to (2.9): Stern and Cressie (2000), considering posterior predictive model checks for disease mapping models, propose posterior predictive inference using cross-validation, as, e.g., in (2.10). This is especially important in disease mapping where there may be true extrema, corresponding to local "hot spots." They propose several methods to reduce the computational burden associated with calculation of quantities such as (2.10) while Marshall and Spiegelhalter (2003) suggest an approximation and compare it with the alternatives given by Stern and Cressie (2000).

The second class of techniques uses the checking function $T(\tilde{y})$ or the discrepancy measure $D(\tilde{y}, \theta)$ to obtain the posterior predictive p values

$$P(T(\tilde{y}) \geq T(y) \mid y) \tag{2.11}$$

and

$$P(D(\tilde{y}, \theta) \geq D(y, \theta) \mid y)\,. \tag{2.12}$$

An extreme p value leads to the conclusion that the fitted model is not consistent with the observed data. These techniques are described in detail in
Gelman, Carlin, Stern and Rubin (2004) and Gelman, Meng and Stern (1996).

Our most important finding is that using the set of individual posterior predictive p values is effective in detecting unanticipated hierarchical structure. Conversely, looking at each individual p value separately will not detect unanticipated hierarchical structure. In addition, the graphical displays are easy to interpret because the distribution of the set of individual p values is uniform (at least for moderate sample sizes) under the null model. Our results also support comments in the literature (e.g., Gelman et al. 2004, Sinharay and Stern 2003) that the choice of test quantity, $T(\tilde{y})$, or discrepancy measure, $D(\tilde{y}, \theta)$, is critical: Some natural choices are not effective in detecting hierarchical structure. By contrast, use of the F statistic as a test quantity *is* effective, and is appropriately calibrated.

Note that we are subjecting the diagnostic techniques to a stringent test because the fitted models ((2.1), (2.5)) and corresponding models assumed to have generated the observed data ((2.3), (2.7)) differ only mildly in their covariance structure.

In Sections 3 and 4 we evaluate the quality of the two Bayesian diagnostic techniques for the two cases described above.

3 Distribution of the set of individual posterior predictive p values

3.1 Numerical examples

We first present numerical examples for Case I to illustrate this method. We generated data $y = \{y_{ij} : i = 1, \ldots, m, j = 1, \ldots, n\}$ from (2.2), (2.3) and (2.4) with $m = 10$, $n = 50$ for $N = 500$, and $m = 5$ and $n = 10$ for $N = 50$. The means and variances are defined below. For data generated from (2.2), the single level model, the mean μ_0 is 20 and the variance ϕ_0 is 12. For data from (2.3), the two-level model, the parameters are different from the single level model only in the covariances - to permit fair comparisons of the two models. Thus, μ_0 is still 20 and $\phi_0 + d_0 = 12$. We considered two cases, one taking $\phi_0 = 10$, $d_0 = 2$ to represent weak within-cluster correlation (correlation = 1/6) and the other taking $\phi_0 = 1$, $d_0 = 11$ to represent strong correlation (correlation = 11/12). For data from (2.4), the two-level model with changepoint, $\phi_0 = d_0 = 6$, $\mu_{01} = 25$, $\mu_{02} = 5$. For $N = 500$, $g = 5$ while for $N = 50$, $g = 3$.

For each of the eight cases (two sample sizes, four models) we generated several data sets. For each data set we calculated $\{p_{ij} : i = 1, \ldots, m\,,\ j = 1, \ldots, n\}$ where p_{ij} is defined in (2.9). Each set of N values of p_{ij} is evaluated under the assumption that the *fitted model*, (2.1), holds.

Figures 1 and 2 are Q-Q plots (using the uniform as the reference distribution) of the p_{ij} for typical data sets, where Figure 1 is for $N = 500$ and Figure 2 for $N = 50$.

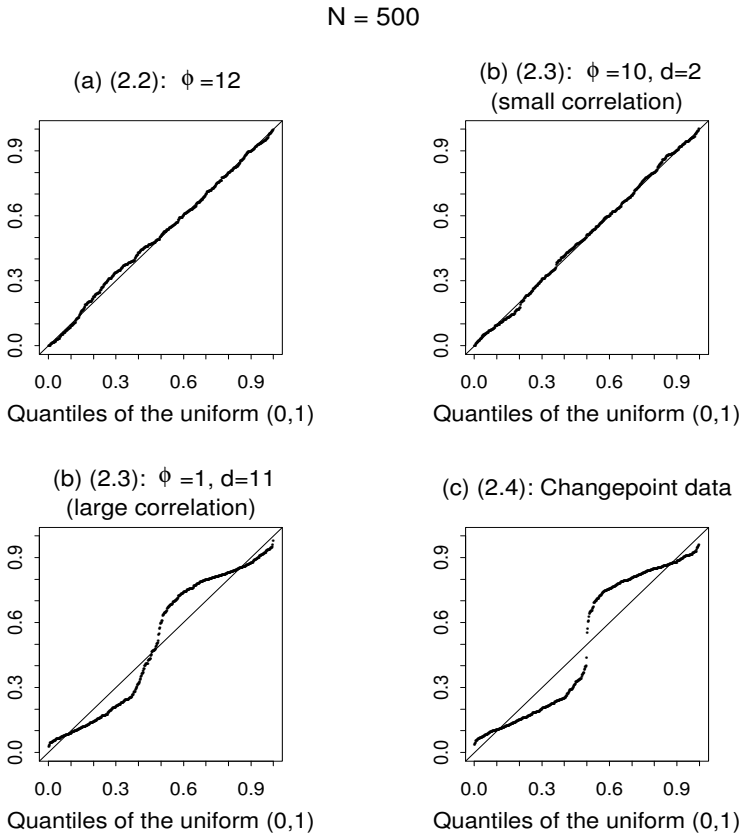

Figure 1: Q-Q plots of estimated distributions of individual posterior predictive p values vs. the uniform distribution

The four plots in each Figure correspond to the four models used to generate the data. For $N = 500$, the Q-Q plots show concordance between the fitted model, (2.1), and the single level model, (2.2), and the two-level model, (2.3), with small within-cluster correlation (1/6). That is, the distribution of the set of individual p values is uniform. There is clear evidence of lack of fit when the data are generated from (2.3) with large correlation (11/12) and (2.4). For the small sample, $N = 50$, the conclusions are similar.

3.2 Theoretical development

In (2.9), consider y as a random variable. If the distribution of y_{ij} is the same as the distribution of $\tilde{y}_{ij}$ conditional on y, $p_{ij} = P(\tilde{y}_{ij} \leq y_{ij}|y)$ is a uniform random variable, and a set of the p_{ij} should act like a uniformly distributed random sample. Conversely, if these two distributions are substantially different, then the set of p_{ij} should not look uniformly distributed.

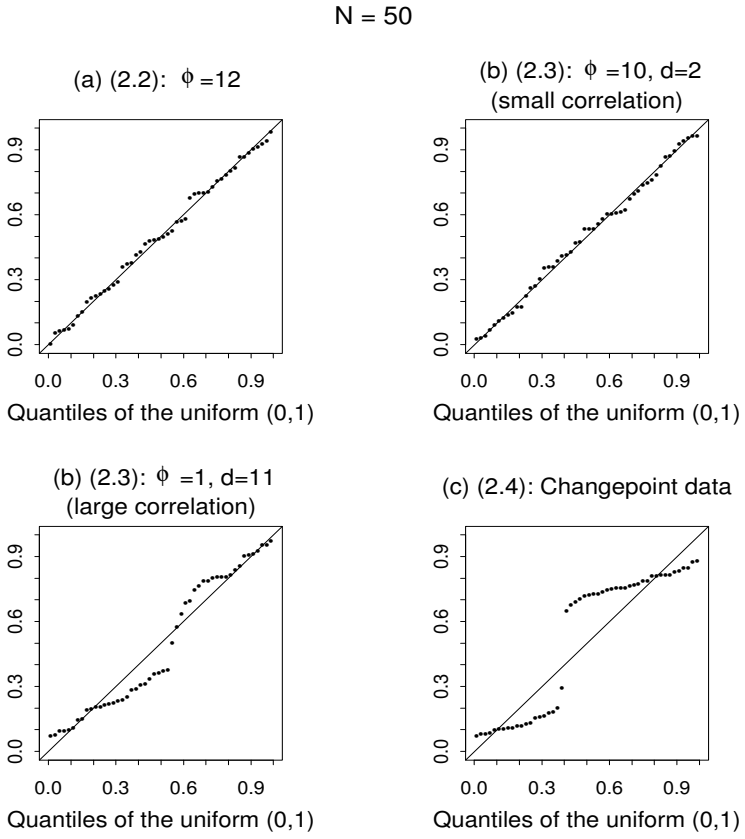

Figure 2: Q-Q plots of estimated distributions of individual posterior predictive p values vs. the uniform distribution

The basic result that we need is that, under the fitted model, (2.1), the distribution of $\tilde{y}$ given y is multivariate t, i.e.,

$$\tilde{y} \mid y \sim t_{N-3} \left[\bar{y} 1_N, \left(I_N + N^{-1} 1_N 1_N' \right) \frac{N-1}{N-3} S^2 \right] \tag{3.1}$$

with $E(\tilde{y}_{ij}|y) = \bar{y}$, $\text{var}(\tilde{y}_{ij}|y) = \{(N^2-1)/N(N-5)\}S^2$ and $\text{cov}(\tilde{y}_{ih}, \tilde{y}_{i'k} \mid y) = \text{cov}(\tilde{y}_{ih}, \tilde{y}_{ik} \mid y) = \{(N-1)/N(N-5)\}S^2$ where I_N is a $N \times N$ identity matrix, 1_N is a $N \times 1$ vector of 1's and S^2 is the sample variance.

3.2.1 Case I: Fit the non-hierarchical model

We consider two situations: the data are generated from (2.2), the single level model (consistent with the fitted model, (2.1)), and (2.3), the two-level hierarchical model. We do not consider (2.4), the two-level model with changepoint, because it is apparent from our numerical work (e.g., Figures 1 and 2) that deviations such as this can easily be detected. Applying our evaluation method we next present our investigation of the ensemble of p values (see (2.9)) by comparing the distributions of the y_{ij} with those of the $\tilde{y}_{ij}$ conditional on y, averaging over the actual distribution of y. That is, for the first situation, the moments of the y_{ij} use (2.2) and the moments of the $\tilde{y}_{ij}$ are obtained from (3.1) and then averaged over (2.2). For the second situation, replace (2.2) with (2.3).

The rationale for this evaluation approach is to ensure that conclusions from applying our procedure (i.e., "accept" the null model, "reject" the null model) are appropriate. We need to know how the procedure behaves under the null model so that we are unlikely to conclude that (2.2), the single level model, is not concordant with the observed data when, in fact, it is. The importance of such calibration has been documented in, e.g., Bayarri and Berger (2000) and Robins et al. (2000). Similarly, we need to know how the procedure behaves under the two-level model, (2.3). If (2.3) is the appropriate model is it likely that our procedure will detect this?

First, assume (2.2). Then, by definition

$$E(y_{ij}|\mu_0, \phi_0) = \mu_0, \quad \text{var}(y_{ij}|\mu_0, \phi_0) = \phi_0, \quad \text{and all of the } y_{ij} \text{ are independent}. \tag{3.2}$$

The moments in (3.2) are to be compared with the moments of $\tilde{y}_{ij}$ conditional on y, taken over the distribution of y in (2.2). For example, the first moment is $E_{(y|\mu_0,\phi_0)} E(\tilde{y}_{ij}|y)$ which we denote by $E(\tilde{y}_{ij}|\mu_0, \phi_0)$. Then

$$E(\tilde{y}_{ij}|\mu_0, \phi_0) = E_{(y|\mu_0,\phi_0)} E(\tilde{y}_{ij}|y) = E_{(y|\mu_0,\phi_0)} \bar{y} = \mu_0 . \tag{3.3}$$

Similarly, the variance of $\tilde{y}_{ij}$ is denoted by $\text{var}(\tilde{y}_{ij}|\mu_0, \phi_0)$ and

$$\begin{aligned} \text{var}(\tilde{y}_{ij}|\mu_0, \phi_0) &= E_{(y|\mu_0,\phi_0)} \text{var}(\tilde{y}_{ij}|y) + \text{var}_{(y|\mu_0,\phi_0)} E(\tilde{y}_{ij}|y) \\ &= [(N^2-1)/N(N-5)]\phi_0 + \phi_0/N . \end{aligned} \tag{3.4}$$

For any two units,

$$\begin{aligned}\operatorname{cov}(\tilde{y}_{ih}, \tilde{y}_{i'k}|\mu_0, \phi_0) &= E_{(y|\mu_0,\phi_0)}\operatorname{cov}(\tilde{y}_{ih}, \tilde{y}_{i'k}|y) + \operatorname{cov}_{(y|\mu_0,\phi_0)}[E(\tilde{y}_{ih}|y), E(\tilde{y}_{i'k}|y)] \\ &= [(N-1)/N(N-5)]\phi_0 + \phi_0/N\,. \end{aligned} \tag{3.5}$$

For large N it is clear from (3.2)-(3.5) that the distribution of y and that of $\tilde{y}$ are normal with the same moments. Thus, *the set of individual p values is uniformly distributed under the null model*, (2.2). For large N a Q-Q plot of the p_{ij} should be consistent with a uniform distribution, as we have seen in our examples (Section 3.1).

If the alternative model, (2.3), generates the data

$$\begin{aligned} E(y_{ij}|\mu_0, \phi_0, d_0) = \mu_0,\ \ \operatorname{var}(y_{ij}|\mu_0, \phi_0, d_0) = \phi_0 + d_0, \\ \operatorname{cov}(y_{ih}, y_{i'k}|\mu_0, \phi_0, d_0) = 0,\ \operatorname{cov}(y_{ih}, y_{ik}|\mu_0, \phi_0, d_0) = d_0\,. \end{aligned} \tag{3.6}$$

Proceeding as above it can be shown, after algebraic manipulation, that the moments of $\tilde{y}_{ij}$ under (2.3) are

$$\begin{aligned} E(\tilde{y}_{ij}|\mu_0, \phi_0, d_0) &= E_{(y|\mu_0,\phi_0,d_0)}\, E(\tilde{y}_{ij}|y) = E_{(y|\mu_0,\phi_0,d_0)}\, \bar{y} = \mu_0\,, \\ \operatorname{var}(\tilde{y}_{ij}|\mu_0, \phi_0, d_0) &= \frac{(N^2-1)}{N(N-5)}\left(\phi_0 + \frac{N-n}{N-1}d_0\right) + \frac{\phi_0 + nd_0}{N}\,, \\ \operatorname{cov}(\tilde{y}_{ih}, \tilde{y}_{ik}|\mu_0, \phi_0, d_0) &= \operatorname{cov}(\tilde{y}_{ih}, \tilde{y}_{i'k}|\mu_0, \phi_0, d_0) \\ &= \frac{N-1}{N(N-5)}\left(\phi_0 + \frac{N-n}{N-1}d_0\right) + \frac{\phi_0 + nd_0}{N}\,. \end{aligned} \tag{3.7}$$

Now consider the two cases

i. $N \to \infty$, $m \to \infty$ and n is fixed. Then

$$\begin{aligned} \lim_{m,N\to\infty} \operatorname{var}(\tilde{y}_{ij} \mid \mu_0, \phi_0, d_0) &= \phi_0 + d_0\,, \\ \lim_{m,N\to\infty} \operatorname{cov}(\tilde{y}_{ih}, \tilde{y}_{i'k} \mid \mu_0, \phi_0, d_0) &= \lim_{m,N\to\infty} \operatorname{cov}(\tilde{y}_{ih}, \tilde{y}_{ik} \mid \mu_0, \phi_0, d_0) = 0\,. \end{aligned} \tag{3.8}$$

ii. $N \to \infty$, $n \to \infty$ and m is fixed. Then

$$\begin{aligned} \lim_{n,N\to\infty} \operatorname{var}(\tilde{y}_{ij} \mid \mu_0, \phi_0, d_0) &= \phi_0 + d_0\,, \\ \lim_{n,N\to\infty} \operatorname{cov}(\tilde{y}_{ih}, \tilde{y}_{i'k}|\mu_0, \phi_0, d_0) &= \lim_{n,N\to\infty} \operatorname{cov}(\tilde{y}_{ih}, \tilde{y}_{ik}|\mu_0, \phi_0, d_0) = d_0/m\,. \end{aligned} \tag{3.9}$$

Comparing (3.6) with (3.7), (3.8) and (3.9), the expected values agree, the variances agree, but the covariances are different. Clearly, it is the ensemble of values of the p_{ij} that may permit us to distinguish the alternative model, (2.3), from the null model, (2.2).

Under case (i),

$$\text{cor}(y_{ih}, y_{ik} \mid \mu_0, \phi_0, d_0) = d_0/(d_0 + \phi_0) \text{ vs. } \text{cor}(\tilde{y}_{ih}, \tilde{y}_{ik} \mid \mu_0, \phi_0, d_0) \doteq 0\,,$$
$$\text{cor}(y_{ih}, y_{i'k} \mid \mu_0, \phi_0, d_0) = 0 \text{ vs. } \text{cor}(\tilde{y}_{ih}, \tilde{y}_{i'k} \mid \mu_0, \phi_0, d_0) \doteq 0\,.$$

Under case (ii),

$$\text{cor}(y_{ih}, y_{ik}|\mu_0, \phi_0, d_0) = \frac{d_0}{d_0 + \phi_0} \text{ vs. } \text{cor}(\tilde{y}_{ih}, \tilde{y}_{ik}|\mu_0, \phi_0, d_0) \doteq \frac{d_0}{m(d_0 + \phi_0)}, \quad (3.10)$$

$$\text{cor}(y_{ih}, y_{i'k}|\mu_0, \phi_0, d_0) = 0 \text{ vs. } \text{cor}(\tilde{y}_{ih}, \tilde{y}_{i'k}|\mu_0, \phi_0, d_0) \doteq \frac{d_0}{m(d_0 + \phi_0)}, \quad (3.11)$$

where the symbol $\doteq$ denotes equality in the limit with the conditions given in (i) and (ii), respectively, on the previous page.

For small d_0/ϕ_0 (i.e., small correlation), the two distributions (i.e., those of y and $\tilde{y}$) are not very different, in which case one would expect a set of the individual posterior p values, $\{p_{ij} = P(\tilde{y}_{ij} \leq y_{ij}|y)\}$, to be uniformly distributed. For large d_0/ϕ_0, the two distributions are very different, and the distribution of the p_{ij} will be less uniform. These results are consistent with our numerical results; i.e., the distribution looked uniform when d_0/ϕ_0 is small and less uniform when d_0/ϕ_0 is large. It is also apparent from (3.10) and (3.11) that it is easier to detect departures from the null model when m is small. This is so because there are many more observations with $i \neq i'$ than with $i = i'$. These results suggest that this technique is more effective to detect inadequate fit of the non-hierarchical model to data with two-stage structure when there is large within-cluster correlation and a small number of clusters.

3.2.2 Case II: Fit the two-stage hierarchical model

We proceed exactly as in Case I by first finding the posterior predictive distribution of $\tilde{y}$ given y assuming the fitted model, (2.5). Then we compare the moments of y and $\tilde{y}$ assuming the null model, (2.6). Finally, we compare the moments of y and $\tilde{y}$ assuming the three-stage model in (2.7). These derivations are given in Appendix A. The results are similar to those reported for Case I where a non-hierarchical model was fitted and the data were from a two-stage model, suggesting that these conclusions hold more generally. These results suggest that Q-Q plots, such as those in Figures 1 and 2, should be useful diagnostics with departures from the 45 degree line indicating nonconcordance between the actual and fitted distributions, thus a lack of fit of the fitted model.

4 Conventional posterior predictive p values

4.1 General checking functions

We next investigate for Case I the posterior predictive p values defined by (2.11) and (2.12). We first consider an ensemble of checking functions, $T(y)$, and discrepancy

measures, $D(y,\theta)$, that are general in nature; i.e., not motivated by the fitted distribution, (2.1), and alternative distributions, (2.3) and (2.4). For $T(y)$ we use the five quantities: minimum (i.e., min $\{y_{ij} : i = 1,\ldots,m,\ j = 1,\ldots,n\}$), median, maximum, mean and standard deviation. Defining the discrepancy measure for the (ij)th unit as $d_{ij} = [y_{ij} - E(y_{ij}|\mu,\phi)]/SD(y_{ij}|\mu,\phi)$, we use the same five quantities (applied to the $\{d_{ij} : i = 1,\ldots,m,\ j = 1,\ldots,n\}$) as the choices for $D(y,\theta)$. Here $E(y_{ij}|\mu,\phi) = \mu$ and $SD(y_{ij}|\mu,\phi) = \phi^{1/2}$, the moments of the sampling distribution in (2.1).

We have tested this technique using simulated data sets constructed as in Section 3.1. For small N (e.g., $N = 50$) none of these choices was effective, and for large N (e.g., $N = 500$) only a few were effective. These results support the view that the performance of the posterior predictive p value depends critically on the choice of the checking function (e.g., Gelman et al. 2004, Sinharay and Stern 2003).

We next show that with the choice of the F statistic as the checking function the associated posterior predictive p value is very effective for detecting hierarchical structure and also has the desired property of uniformity under the null model. This is to be expected when the investigator correctly guesses the nature of the hierarchical structure not included in the fitted model. However, we present results for Case II in Section 4.2 which suggest that use of the F statistic may also be beneficial when the hierarchical structure not included in the fitted model is *not* anticipated.

4.2 F statistic as a checking function

4.2.1 Case I: Fit the non-hierarchical model

We now consider as a checking function, the usual F statistic,

$$F(y) = \frac{\sum_{i=1}^{i=m} n(\bar{y}_i - \bar{y})^2/(m-1)}{\sum_{i=1}^{i=m}\sum_{j=1}^{j=n}(y_{ij} - \bar{y}_i)^2/(N-m)} \tag{4.1}$$

where $\bar{y}_i = \sum_{j=1}^{n} y_{ij}/n$. The posterior predictive p value is

$$P(F(\tilde{y}) \geq F(y) \mid y)\,. \tag{4.2}$$

As in Section 3.2 we apply our evaluation method to compare the distribution of $F(\tilde{y})$ given y with that of $F(y)$ when (2.2) and (2.3) generate the data. When the data are from the non-hierarchical model, (2.2), $F(y) \sim \mathcal{F}_{m-1,N-m}$. For large N, $\tilde{y}|y \sim N(\bar{y}1_N, S^2 I_N)$ from (3.1), and $F(\tilde{y})|y \sim \mathcal{F}_{m-1,N-m}$. Hence, the p value in (4.2) is *uniformly distributed under the null model*, as is appropriate for this case.

If the data are from the hierarchical alternative, (2.3), $F(y) \sim k\mathcal{F}_{m-1,N-m}$ where $k = (nd_0 + \phi_0)/\phi_0$, but still $F(\tilde{y})|y \sim \mathcal{F}_{m-1,N-m}$. Then the p values in (4.2) will be small if $k \gg 1$. This will occur if $d_0/(\phi_0/n)$ is large.

Our simulation study uses the same specifications as in Section 3.1 together with additional ones with different correlations. The results show that $F(y)$ is a powerful

checking function. As expected, the p values are moderately large (i.e., far from 0) when the data are from (2.2). The p values are almost always small (i.e., near 0) when: (a) N is large (e.g., $N = 500$) even if the correlation is as small as 1/6, (b) N is small (e.g., 50) and the data are from (2.4) or (2.3) with moderate correlation (e.g., > 0.50). Thus, if one anticipates clustering of the data such as (2.3) and (2.4), use of the F statistic as the checking function is a very effective diagnostic procedure.

4.2.2 Case II: Fit the two-stage hierarchical model

For the second case assume the fitted model (2.5) but with ν, ϕ and γ unknown and with any prior. We first consider the checking function

$$F(y) = \frac{n \sum_{i=1}^{a} \sum_{j=1}^{m} (\bar{y}_{ij.} - \bar{y}_{i..})^2 / a(m-1)}{\sum_{i=1}^{a} \sum_{j=1}^{m} \sum_{k=1}^{n} (y_{ijk} - \bar{y}_{ij.})^2 / am(n-1)} \,. \tag{4.3}$$

which is motivated by the three-stage alternative, (2.7).

It is easily shown that the posterior predictive p value in (4.2) using this F checking function is uniformly distributed (see Appendix B.1) when the data are from the two-stage model in (2.6). When the data are from the three-stage model, (2.7), it is easily shown (Appendix B.1) that the posterior predictive p value will be small if $d_0/(\phi_0/n)$ is large, a result analogous to that for Case I (Section 4.2) when (2.1) is fitted and the data are from (2.3). This result suggests that using F defined in (4.3) as the checking function is a useful technique to detect a lack of fit of the two-stage model if the data in fact have the three-stage structure.

Now, consider the alternative F statistic

$$F^*(y) = \frac{mn \sum_{i=1}^{a} (\bar{y}_{i..} - \bar{y}_{...})^2 / (a-1)}{\sum_{i=1}^{a} \sum_{j=1}^{m} \sum_{k=1}^{n} (y_{ijk} - \bar{y}_{i..})^2 / a(mn-1)} \,, \tag{4.4}$$

which does *not* take account of the three-stage model in (2.7).

Assuming the fitted model is (2.5) with ϕ and γ known, the posterior predictive distribution of $F^*(\tilde{y})$ given y cannot be obtained in an analytical form. Thus, we compare expected values; i.e., $E\{F^*(y)\}$ with $E_y\, E(F^*(\tilde{y})|y)$, and also must approximate $E(F^*(\tilde{y})|y)$ as the ratio of the expected values of the numerator and denominator of (4.4).

The results, given in Appendix B.2, suggest that: (a) When the data are from the two-stage hierarchical model in (2.6) (i.e., both the fitted and actual data models are two-stage), the p value, $P\{F^*(\tilde{y}) \geq F^*(y)|y\}$, will not be extreme, and (b) When the data are from the three-stage model in (2.7), F^*, which is not influenced by the three-stage cluster structure, will be an effective checking function when ϕ_0 or d_0 is large (as long as mn is not too large).

5 Analysis of Patterns of Care Study Data

The Patterns of Care Study is a sample survey to evaluate the quality of care received by cancer patients undergoing radiation therapy. The sample design is stratified, two-stage cluster sampling: Within each stratum a simple random sample of radiation therapy facilities is selected and then within each facility a simple random sample of patients undergoing radiation therapy within a specified time period is chosen. This is done for each of several disease sites. There are two scores measuring the quality of the patient's workup and therapy; each score, Y, is scaled so that $0 \leq Y \leq 1$. The analysis in Calvin and Sedransk (1991) shows that for the workup score for cervix cancer in the 1978 survey $W = Y^3$ transforms Y to approximate normality. It is also clear that the stratification effect is negligible, so the alternative models to be entertained are (2.2) and (2.3) with W_{ij} (jth patient in facility i) as the dependent variable. We now present our analysis of the two diagnostic techniques for the transformed workup score in the 1978 survey.

Fitting the non-hierarchical model, (2.1), the Q-Q plot in Figure 3a (using the uniform reference distribution) of the p_{ij} from (2.9) clearly shows a substantial departure from the uniform distribution and, thus, the fitted non-hierarchical model is inconsistent with the data. We then fit a two-stage model similar to (2.3), with the stages corresponding to patients and facilities; i.e.,

$$\begin{aligned} y_{ij} \mid \theta_i, \phi &\overset{iid}{\sim} N(\theta_i, \phi), \\ \theta_i \mid \mu, d &\overset{iid}{\sim} N(\mu, d) \end{aligned} \tag{5.1}$$

with independent, locally uniform priors on ϕ, μ, and d. The Q-Q plot in Figure 3b shows no apparent deviation from this fitted model. Thus, the model in (5.1) appears to be concordant with these data.

We have also evaluated the posterior predictive p value in (4.2). Fitting the non-hierarchical model, (2.1), and using the F statistic in (4.1) as the checking function, the posterior predictive p value is near 0 (see Figure 4a), clearly indicating that the data cluster within facilities and that the fitted, non-hierarchical model is inappropriate. Fitting the two-stage hierarchical model, (5.1), using the same F checking function yields a posterior predictive p value of 0.5 (see Figure 4b), suggesting that this two-stage hierarchical model is appropriate.

6 Discussion

6.1 Challenges in checking hierarchical structure

It is difficult to distinguish actual hierarchical structure (e.g., (2.3)) from the absence of such structure (e.g., (2.2)). One way to see this is to study properties of the posterior distribution of the scale parameter ϕ under the fitted model when the observed data, y, are generated from the hierarchical model, (2.3). Assuming (2.1) it is straightforward to show that $E(\phi|y) = (N-1)S^2/(N-5)$ where S^2 is the sample variance. Assuming

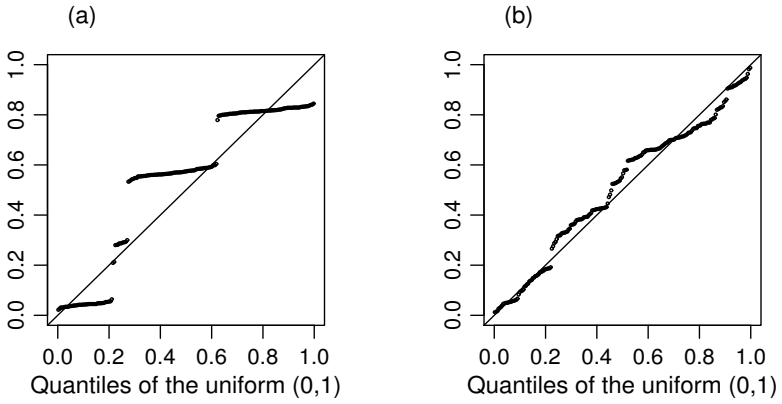

Figure 3: Patterns of Care Study: Q-Q plots of estimated distributions of individual posterior predictive p values vs. the uniform distribution for the transformed workup score. Plot (a) results from fitting the non-hierarchical model, (2.1) and plot (b) results from fitting the two-stage hierarchical model, (5.1).

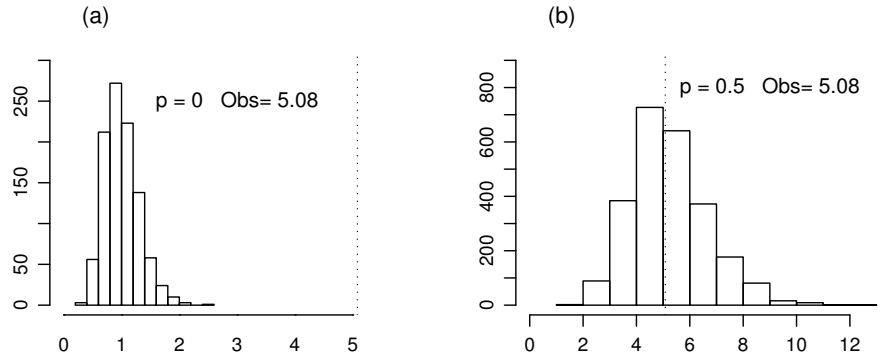

Figure 4: Patterns of Care Study: Estimated posterior predictive distributions of the F checking function for the transformed workup score. Plot (a) results from fitting the non-hierarchical model, (2.1) and plot (b) results from fitting the two-stage hierarchical model, (5.1). The observed value of $F(y)$, (4.1), is indicated by a dashed line and also by 'Obs' in the title. The posterior predictive p value, (4.2), is shown in the title.

that the data are from (2.3), it can be shown that

$$E(\phi|\mu_0,\phi_0,d_0) = E_{(y|\mu_0,\phi_0,d_0)}E(\phi|y) = \frac{(N-1)}{(N-5)}\left[\phi_0 + \frac{(N-n)}{(N-1)}d_0\right].$$

As $N, m \to \infty$, $E(\phi) \to \phi_0 + d_0$, and the distribution of $\tilde{y}_{ij}$ converges to $N(\mu_0, \phi_0 + d_0)$. That is, any posterior predictive assessment using the fitted model, (2.1), will reflect the true variation in data from the *hierarchical model*, (2.3). That is, the additional variability in (2.3) relative to (2.2) will not be revealed.

To be successful a general diagnostic technique (i.e., one not motivated by a specific alternative model) must reveal the different correlation structures of, for example, (2.2) and (2.3). We have shown in Sections 3 and 5 that the technique of individual posterior predictive p values does this.

Two papers that study diagnostics for hierarchical models are Sinharay and Stern (2003) and Bayarri and Castellanos (2007), but neither address the problem of detecting missing hierarchical structure in the fitted model. Sinharay and Stern (2003) fit a model similar to (2.3) and investigate the postulated normal assumption at the second stage. They investigate posterior predictive assessment as in (2.11) and (2.12), and use simulations (patterned on the SAT coaching example; see, e.g., Gelman et al. 2004, Section 5.5) to study the properties associated with several checking and discrepancy measures. Similarly, Bayarri and Castellanos (2007) assume a two-stage hierarchical normal model as the null model and investigate departures from it such as different distributions at the second stage. They investigate the performance of several p values, including the one in (2.11) and the corresponding partial posterior predictive p value (Bayarri and Berger 2000).

6.2 Partial posterior predictive p value and evaluation

Bayarri and Berger (2000) have criticized the "double use" of data in a posterior predictive p value, and have suggested as a practical alternative the partial posterior predictive p value. With our notation the partial posterior predictive p value of Bayarri and Berger (2000) is

$$P\Big\{T(\tilde{y}) \geq T(y) \mid y \setminus T(y)\Big\} = \int P\Big\{T(\tilde{y}) \geq T(y) \mid \theta\Big\}\, p\Big\{\theta|y \setminus T(y)\Big\}\, d\theta\,, \qquad (6.1)$$

where, by definition, $p\{\theta|y\setminus T(y)\} \propto h_1(y|\theta)h_2(\theta)/h_3\{T(y)|\theta\}$. Both Bayarri and Berger (2000) and Robins et al. (2000) have pointed to the need for calibration of a "p value," i.e., its properties under the "null model." Theoretical results in Bayarri and Berger (2000) and Robins et al. (2000) indicate that (at least) for large samples, the distribution of the partial posterior predictive p value is uniform under the null model while the distribution of the (conventional) posterior predictive p value may not be.

We have shown that the set of individual posterior predictive p values is uniformly distributed under the null models (2.2) and (2.6), and that the posterior predictive p value (with the F statistic as checking function) is also uniformly distributed under

(2.2) and (2.6). These results are consistent with those described above: First, let $T(y_{ij}) = y_{ij}$ in (2.11), so the individual posterior predictive p value is

$$P(\tilde{y}_{ij} \geq y_{ij} \mid y) = \int P(\tilde{y}_{ij} \geq y_{ij} \mid \theta)\, p_1(\theta|y)\, d\theta\,.$$

From (6.1), the corresponding individual partial posterior predictive p value is

$$P(\tilde{y}_{ij} \geq y_{ij} \mid y \setminus y_{ij}) = \int P(\tilde{y}_{ij} \geq y_{ij} \mid \theta)\, p_2(\theta|y \setminus y_{ij})\, d\theta\,.$$

However, under a fitted model such as (2.1), $p_2(\theta|y \setminus y_{ij})$ is essentially the same as $p_1(\theta|y)$ because the total sample size is assumed to be, at least, reasonably large and there are no outliers in our specification (see Section 2).

Taking $T(y) = F(y)$ with $F(y)$ defined in (4.1) it is easy to show that under the null model, (2.2), $p(\theta|y \setminus F(y)) = p(\theta|y)$ because $F(y) \sim \mathcal{F}_{m-1,N-m}$. Similarly, if $T(y) = F(y)$ with $F(y)$ defined by (4.3), it is easy to show that $p(\theta|y \setminus F(y)) = p(\theta|y)$ under the null model, (2.6), because $F(y) \sim \mathcal{F}_{a(m-1),\, am(n-1)}$. Thus, the posterior predictive p values with $T(y) = F(y)$ are the same as the partial posterior predictive p values.

Hjort et al. (2006) have recently proposed a method for calibrating posterior predictive p values. This research, though, assumes proper prior distributions which is both restrictive for practical applications and less desirable at the model checking stage (see Bayarri and Castellanos 2007, Section 1).

7 Conclusion

We have presented methodology to evaluate Bayesian diagnostic techniques that are based on the posterior distribution of predicted data, $\tilde{y}$, under a fitted model, i.e., the distribution with density $p(\tilde{y}|y)$ in (2.8). Assuming that the diagnostic quantity, $d(y)$, is a function only of the observed data, y, we first investigate $d(y)$ under the assumption that the actual distribution of y is consistent with that postulated for the fitted model - to show the desired uniform distribution. Then we investigate $d(y)$ assuming that the actual distribution of y is different from that of the fitted model - to investigate the conditions under which the diagnostic will reveal the existence (and nature) of a model different from that postulated for the fitted model. This approach can also be used to evaluate discrepancy measures, $D(y, \theta)$, which are functions of the observed data and parameters.

We have evaluated the ability of several Bayesian diagnostic techniques to detect hierarchical structure that is not accounted by the fitted model. Because there has been no apparent research of this type, we have started with the simple, yet important situation where the data come from a model with two-stage hierarchical structure while the fitted model does not have this structure. We then extended this analysis to the case where the fitted model is a two-stage hierarchical model while the data come from

a model with three stages. Since the results in the two cases are similar we expect that conclusions we have reached will apply to the situation where the fitted model includes some hierarchical structure but does not account for all of the stages actually present.

A general technique that we can recommend to detect hierarchical structure that is not included in the fitted model is to check the distribution of the ensemble of individual posterior predictive p values. Be careful, though, because looking at each individual p value separately will not detect unanticipated hierarchical structure.

If one has a structured situation, such as a set of hospitals (as illustrated by the Patterns of Care Study), evaluating the posterior predictive p value using the F statistic motivated by a specific *alternative* model should give a clear idea if the alternative model is appropriate. Our results for F^* for Case II in Section 4.2 suggest that the F statistic may also be useful in situations where the clustering of units is not clear a priori. All of these methods are effective in detecting hierarchical structure missing from the fitted model, and also are properly calibrated under the null model.

References

Bayarri, M.J. and Berger, J.O. (2000). "P-values for composite null models," *Journal of the American Statistical Association*, 95, 1127-1142.

Bayarri, M.J. and Castellanos, M.E. (2007). "Bayesian checking of hierarchical models," *Statistical Science*, in press.

Box, G.E.P. (1980). "Sampling and Bayes inference in scientific modeling and robustness," *Journal of the Royal Statistical Society A*, 143, 383-430.

Calvin, J.A. and Sedransk, J. (1991). "Bayesian and Frequentist Predictive Inference for the Patterns of Care Studies," *Journal of the American Statistical Association,* 86, 36-48.

Geisser, S. (1993). *Predictive Inference: An Introduction*, Chapman & Hall, New York, NY.

Gelfand, A.E., Dey, D.K., and Chang, H. (1992). "Model determination using predictive distributions with implementation via sampling-based methods," In *Bayesian Statistics 4*, eds. J.M. Bernardo et al., Oxford University Press, London, 147-167.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2004). *Bayesian Data Analysis*, 2nd Edition, Chapman & Hall/CRC, London.

Gelman, A., Meng, X.L., and Stern, H. (1996). "Posterior predictive assessment of model fitness via realized discrepancies (with discussion)," *Statistica Sinica*, 6, 733-807.

Hjort, N.L., Dahl, F.A., and Steinbakk, G. H. (2006). "Post-processing posterior predictive p values," *Journal of the American Statistical Association,* 101, 1157-1174.

Marshall, E.C. and Spiegelhalter, D.J. (2003). "Approximate cross-validatory predictive checks in disease mapping models," *Statistics in Medicine*, 22, 1649-1660.

Nandram, B., Sedransk, J., and Pickle, L. (1999). "Bayesian analysis of mortality rates for U.S. health service areas," *Sankhya B* 61, 145-165.

Nandram, B., Sedransk, J., and Pickle, L. (2000). "Bayesian analysis and mapping of mortality rates for chronic obstructive pulmonary disease," *Journal of the American Statistical Association,* 95, 1110-1118.

Robins, J.M., van der Vaart, A., and Ventura, V. (2000). "Asymptotic distribution of p-values in composite null models (with discussion)," *Journal of the American Statistical Association*, 95, 1143-1172.

Rubin, D.B. (1984). "Bayesianly justifiable and relevant frequency calculations for the applied statistician," *The Annals of Statistics*, 12, 1151-1172.

Sinharay, S. and Stern, H.S. (2003). "Posterior predictive model checking in hierarchical models," *Journal of Statistical Planning and Inference,* 111, 209-221.

Stern, H.S. and Cressie, N. (2000). "Posterior predictive model checks for disease mapping models," *Statistics in Medicine*, 19, 2377-2397.

Yan, G. (2003). *Evaluation of Bayesian Diagnostic Methods for Hierarchical Data*, Ph.D. dissertation, Department of Statistics, Case Western Reserve University, Cleveland, Ohio.

Appendix A Results for individual posterior predictive p values in Section 3.2: Case II, fitting the two-stage hierarchical model

Define $\bar{y}_i = \sum_j \sum_k y_{ijk}/mn$, $\bar{y} = (\bar{y}_1, \ldots, \bar{y}_a)'$, $\bar{y}_{\cdot\cdot\cdot} = \sum_i \bar{y}_i/a$, $\lambda = \gamma/(\gamma + \phi/mn)$, 1_b is a vector of b 1's and I_b is the $b \times b$ identity matrix. Assuming the fitted model, (2.5), the posterior predictive distribution of $\tilde{y}$, given y, is

$$\tilde{y}|y \quad \sim \quad N\Big(\nu(y), D(y)\Big) \tag{A.1}$$

where

$$\begin{aligned} \nu(y) &= \Big(I_a \otimes 1_{mn}\Big)\Big(\lambda\bar{y} + (1-\lambda)\bar{y}_{\cdot\cdot\cdot}1_a\Big)\,, \\ D(y) &= \lambda(\phi/mn)(I_a \otimes 1_{mn}1'_{mn}) + (1-\lambda)(\phi/amn)(1_{amn}1'_{amn}) + \phi I_{amn}\,. \end{aligned}$$

Assume, first, that the data come from the two stage hierarchical model in (2.6), the null model. Then $E(y_{ijk}) = \nu_0$, $\text{var}(y_{ijk}) = \phi_0 + \gamma_0$, $\text{cov}(y_{ijk}, y_{ijk'}) = \text{cov}(y_{ijk}, y_{ij'k}) = \gamma_0$ and $\text{cov}(y_{ijk}, y_{i'jk}) = 0$. Making the assumption that $\phi = \phi_0$ and $\gamma = \gamma_0$, it is straightforward to show that for any value of a, and for large m or n the moments of $\tilde{y}$ (i.e., the moments from (A.1) averaged over the two stage hierarchical model in (2.6)) agree with the moments of y given above. Thus, for large $N = anm$, *the set of individual p values is uniformly distributed under the null model*, (2.6), and a Q-Q plot of the p_{ijk} would be expected to be consistent with a uniform distribution.

Now, assuming the data coming from the three-stage model in (2.7)

$$\begin{aligned}
E(y_{ijk}|\phi_0, d_0, \gamma_0, \nu_0) &= \nu_0\,, \\
\text{var}(y_{ijk}|\phi_0, d_0, \gamma_0, \nu_0) &= \phi_0 + d_0 + \gamma_0, \\
\text{cor}(y_{ijk}, y_{ijk'}|\phi_0, d_0, \gamma_0, \nu_0) &= (d_0 + \gamma_0)/(\phi_0 + d_0 + \gamma_0), \\
\text{cor}(y_{ijk}, y_{ij'k}|\phi_0, d_0, \gamma_0, \nu_0) &= \gamma_0/(\phi_0 + d_0 + \gamma_0), \\
\text{cor}(y_{ijk}, y_{i'jk}|\phi_0, d_0, \gamma_0, \nu_0) &= 0\,.
\end{aligned} \tag{A.2}$$

After algebraic manipulation the moments of $\tilde{y}$ under (2.7) can be shown to be

$$E(\tilde{y}|\phi_0, d_0, \gamma_0, \nu_0) = 1_{amn}\nu_0\,, \tag{A.3}$$

$$\begin{aligned}
\text{var}(\tilde{y}|\phi_0, d_0, \gamma_0, \nu_0) &= \lambda\big(\phi/mn\big)\big(I_a \otimes 1_{mn}1'_{mn}\big) + \big(1-\lambda\big)\big(\phi/amn\big)\big(1_{amn}1'_{amn}\big) \\
&+ \phi I_{amn} + \big(\phi_0/mn + d_0/m + \gamma_0\big)\,\lambda^2\,\big(I_a \otimes 1_{mn}1'_{mn}\big) \\
&+ \big(\phi_0/amn + d_0/am + \gamma_0/a\big)\big(1-\lambda^2\big)\big(1_{amn}1'_{amn}\big)\,.
\end{aligned} \tag{A.4}$$

See a sketch of the proof in Appendix C.1.

Clearly, the expected values in (A.2) and (A.3) agree. Now, for any value of a,

$$\lim_{m\to\infty} \text{var}(\tilde{y}|\phi_0, d_0, \gamma_0, \nu_0) = \phi I_{amn} + \gamma_0\,(I_a \otimes 1_{mn}1'_{mn}) \tag{A.5}$$

and

$$\lim_{n\to\infty} \text{var}(\tilde{y}|\phi_0, d_0, \gamma_0, \nu_0) = \phi I_{amn} + \big(d_0/m + \gamma_0\big)(I_a \otimes 1_{mn}1'_{mn})\,. \tag{A.6}$$

To provide a clear comparison of (A.5) and (A.6) with (A.2) we make the *conservative* assumption that $\phi = \phi_0 + d_0$. (If $\phi \neq \phi_0 + d_0$ there will be greater differences between the predicted and observed values.) Then

$$\begin{aligned}
\lim_{m\to\infty} \text{var}(\tilde{y}_{ijk}|\phi_0, d_0, \gamma_0, \nu_0) &= \phi_0 + d_0 + \gamma_0\,, \\
\lim_{m\to\infty} \text{cor}(\tilde{y}_{ijk}, \tilde{y}_{ijk'}|\phi_0, d_0, \gamma_0, \nu_0) &= \gamma_0/(\phi_0 + d_0 + \gamma_0)\,, \\
\lim_{m\to\infty} \text{cor}(\tilde{y}_{ijk}, \tilde{y}_{ij'k}|\phi_0, d_0, \gamma_0, \nu_0) &= \gamma_0/(\phi_0 + d_0 + \gamma_0)\,, \\
\lim_{m\to\infty} \text{cor}(\tilde{y}_{ijk}, \tilde{y}_{i'jk}|\phi_0, d_0, \gamma_0, \nu_0) &= 0\,.
\end{aligned} \tag{A.7}$$

and

$$\begin{aligned}
\lim_{n\to\infty} \text{var}(\tilde{y}_{ijk}|\phi_0, d_0, \gamma_0, \nu_0) &= \phi_0 + d_0 + \gamma_0 + d_0/m\,, \\
\lim_{n\to\infty} \text{cor}(\tilde{y}_{ijk}, \tilde{y}_{ijk'}|\phi_0, d_0, \gamma_0, \nu_0) &= (\gamma_0 + d_0/m)/\{\phi_0 + d_0 + \gamma_0 + d_0/m\}\,, \\
\lim_{n\to\infty} \text{cor}(\tilde{y}_{ijk}, \tilde{y}_{ij'k}|\phi_0, d_0, \gamma_0, \nu_0) &= (\gamma_0 + d_0/m)/\{\phi_0 + d_0 + \gamma_0 + d_0/m\}\,, \\
\lim_{n\to\infty} \text{cor}(\tilde{y}_{ijk}, \tilde{y}_{i'jk}|\phi_0, d_0, \gamma_0, \nu_0) &= 0\,. \qquad \text{(A.8)}
\end{aligned}$$

With the *conservative* assumption that $\phi = \phi_0 + d_0$ and large values of m the two distributions in (A.2) and (A.7) are not very different if d_0/γ_0 is small, in which case the distribution of individual p values is expected to be approximately uniform. However, for large n and small m there are differences in three of the four terms (compare (A.2) and (A.8)). These differences depend on the magnitudes of d_0/γ_0 and d_0/m: The difference between the second expression in (A.8) and the comparable expression in (A.2) is large if d_0/γ_0 is large, especially if m is moderately large. The difference between the third expression in (A.8) and the comparable term in (A.2) is large if d_0/γ_0 is large but m is small.

Appendix B Results for the F statistic in Section 4.2: Case II, fitting the two-stage hierarchical model

B.1 Results for the F checking function in (4.3)

Assume the fitted model (2.5) but with ν, ϕ and γ unknown and with any prior. Consider the F statistic

$$F(y) = \frac{n\sum_{i=1}^{a}\sum_{j=1}^{m}(\bar{y}_{ij.} - \bar{y}_{i..})^2/a(m-1)}{\sum_{i=1}^{a}\sum_{j=1}^{m}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{ij.})^2/am(n-1)}\,. \qquad \text{(B.1)}$$

Since the distribution of $F(\tilde{y})$, given $\mu = (\mu_1, \ldots, \mu_a)$ and ϕ, is $\mathcal{F}_{a(m-1),\,am(n-1)}$, it follows that the posterior predictive distribution of $F(\tilde{y})$ given y is $\mathcal{F}_{a(m-1),\,am(n-1)}$. When the data are from the two-stage hierarchical model in (2.6), the null model, $F(y) \sim \mathcal{F}_{a(m-1),\,am(n-1)}$. Thus, the posterior predictive p value in (4.2) using this F checking function is uniformly distributed.

When the data are from the three-stage model, (2.7), $F(y) \sim \{(\phi_0 + nd_0)/\phi_0\}$ $\mathcal{F}_{a(m-1),\,am(n-1)}$. Since $F(\tilde{y})|y \sim \mathcal{F}_{a(m-1),\,am(n-1)}$, the posterior predictive p value in (4.2) will be small if $d_0/(\phi_0/n)$ is large. This result suggests that using F defined in (B.1) as the checking function is a useful technique to detect a lack of fit of the two-stage model if the data in fact have the three-stage structure. Note that this is the same result obtained in Case I (Section 4.2) when (2.1) is fitted and the data are from (2.3).

B.2 Results for the F^* checking function in (4.4)

Consider the alternative F statistic,

$$F^*(y) = \frac{mn\sum_{i=1}^{a}(\bar{y}_{i..} - \bar{y}_{...})^2/(a-1)}{\sum_{i=1}^{a}\sum_{j=1}^{m}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{i..})^2/a(mn-1)}\,, \tag{B.2}$$

which does not take account of the three-stage model in (2.7).

Assuming the fitted model is (2.5) with ϕ and γ known, the posterior predictive distribution of $F^*(\tilde{y})$ given y cannot be obtained in an analytical form. Thus, we compare expected values; i.e., $E\{F^*(y)\}$ with $E_y\,E(F^*(\tilde{y})|y)$.

First, we find the posterior predictive expectation $E(F^*(\tilde{y})|y)$. Approximating $E(F^*(\tilde{y})|y)$ as the ratio of the expected values of the numerator and denominator of (B.2), it can be shown, after extensive algebraic manipulation, that

$$E(F^*(\tilde{y})|y) \;\doteq\; (\lambda+1) + (\lambda^2/\phi)\,MS\mu(y) \tag{B.3}$$

where $\lambda = \gamma/(\gamma + \phi/mn)$ as in (A.1) and $MS\mu(y)$ is the numerator of (B.2). See a sketch of the proof in Appendix C.2.

Now assume that the data are from the two-stage hierarchical model in (2.6). Then $F^*(y) \sim \{(mn\gamma_0+\phi_0)/\phi_0\}\,\mathcal{F}_{(a-1),\,a(mn-1)}$ with (ignoring terms of $O(1/amn)$)

$$E\{F^*(y)|\phi_0,\gamma_0\} \doteq (mn\gamma_0+\phi_0)/\phi_0\,. \tag{B.4}$$

Assuming $\phi = \phi_0$ and $\gamma = \gamma_0$, defining $\lambda_0 = \gamma_0/(\gamma_0 + \phi_0/mn)$, it can be shown that (B.3) averaged over (2.6) is

$$E_{(y|\phi_0,\gamma_0)}\,E(F^*(\tilde{y})|y) \;\doteq\; (\lambda_0+1) + (\lambda_0^2/\phi_0)(mn\gamma_0+\phi_0) = (mn\gamma_0+\phi_0)/\phi_0. \tag{B.5}$$

The equality of (B.4) and (B.5) suggests that when the fitted and actual data models are both two-stage, the p value, $P\{F^*(\tilde{y}) \geq F^*(y)|y\}$, will not be extreme.

Next, assume that the data are from the three-stage hierarchical model, (2.7). Then it can be shown that

$$F^*(y) \sim \left\{\frac{(\phi_0+nd_0+mn\gamma_0)(mn-1)}{(m-1)(\phi_0+nd_0)+m(n-1)\phi_0}\right\}\mathcal{F}_{(a-1),\,a(mn-1)}\,. \tag{B.6}$$

See a sketch of the proof in Appendix C.3.

Assuming $a(mn-1)/\{a(mn-1)-2\} \doteq 1$ and $(m-1)/m \doteq 1$,

$$E\{F^*(y)|\phi_0,d_0,\gamma_0\} \;\doteq\; \left\{\frac{\phi_0+d_0+mn\gamma_0}{\phi_0+d_0}\right\} + \left\{\frac{(n-1)d_0}{\phi_0+d_0}\right\}. \tag{B.7}$$

Proceeding in a conservative manner by letting $\phi = \phi_0 + d_0$ and $\gamma = \gamma_0$, it can be shown that (B.3), with averaging over the three-stage model, (2.7), is

$$\begin{aligned} E_{(y|\phi_0,d_0,\gamma_0)}\,E(F^*(\tilde{y})|y) \;=\; & \left\{\frac{\phi_0+d_0+mn\gamma_0}{\phi_0+d_0}\right\} \\ & + \left\{\frac{(n-1)d_0}{\phi_0+d_0}\right\}\left\{\frac{m^2n^2\gamma_0^2}{(mn\gamma_0+d_0+\phi_0)^2}\right\}. \end{aligned} \tag{B.8}$$

Comparing (B.7) and (B.8), the second term in (B.8) is smaller than the comparable term in (B.7) and the difference will be large when ϕ_0 or d_0 is large (as long as mn is not too large). Even under these *conservative* assumptions (i.e., $\phi = \phi_0 + d_0$ and $\gamma = \gamma_0$) it appears that F^* is an effective checking function. Note that F^* is a "general" F-type checking function, i.e., not influenced by the three-stage cluster structure in the alternative model. Thus, the properties obtained here suggest that F-type checking functions will be effective in checking the fit of hierarchical models that do not account for all of the actual hierarchical stages.

Appendix C Sketch of proofs of (A.4), (B.3) and (B.6)

C.1 Proof of (A.4)

The variance of $\tilde{y}$ under the sampling distribution, (2.7), is

$$\begin{aligned} \mathrm{var}(\tilde{y}|\phi_0, d_0, \gamma_0, \nu_0) &= E_{(y|\phi_0,d_0,\gamma_0,\nu_0)}\, \mathrm{var}(\tilde{y}|y) + \mathrm{var}_{(y|\phi_0,d_0,\gamma_0,\nu_0)}\, E(\tilde{y}|y) \\ &= E_{(y|\phi_0,d_0,\gamma_0,\nu_0)}\, D(y) + \mathrm{var}_{(y|\phi_0,d_0,\gamma_0,\nu_0)}\nu(y)\,. \end{aligned} \tag{C.1}$$

Using (A.1), the first term is

$$E_{(y|\phi_0,d_0,\gamma_0,\nu_0)}\, D(y) = \lambda(\phi/mn)(I_a \otimes 1_{mn}1'_{mn}) + (1-\lambda)(\phi/amn)1_{amn}1'_{amn} + \phi I_{amn}\,.$$

Writing $\bar{y}_{\cdot\cdot\cdot} = 1'_a\,\bar{y}/a$, the second term is

$$\begin{aligned} \mathrm{var}_{(y|\phi_0,d_0,\gamma_0,\nu_0)}\nu(y) &= \mathrm{var}_{(y|\phi_0,d_0,\gamma_0,\nu_0)}\Big\{(I_a \otimes 1_{mn})\big(\lambda\bar{y} + (1-\lambda)\bar{y}_{\cdot\cdot\cdot}1_a\big)\Big\} \\ &= \mathrm{var}_{(y|\phi_0,d_0,\gamma_0,\nu_0)}\Big\{(I_a \otimes 1_{mn})\big(I_a\lambda + (1-\lambda)1_a1'_a/a\big)\,\bar{y}\Big\} \\ &= \mathrm{var}_{(y|\phi_0,d_0,\gamma_0,\nu_0)}A\,\bar{y} \end{aligned} \tag{C.2}$$

where $A = (I_a \otimes 1_{mn})\big(I_a\lambda + (1-\lambda)1_a1'_a/a\big)$. Since $\mathrm{var}(\bar{y}|\phi_0, d_0, \gamma_0, \nu_0) = \mathrm{var}(\bar{y}_1, \ldots, \bar{y}_a)'$ $= (\phi_0/mn + d_0/m + \gamma_0)\, I_a$,

$$\mathrm{var}_{(y|\phi_0,d_0,\gamma_0,\nu_0)}A\,\bar{y} = AA'\,(\phi_0/mn + d_0/m + \gamma_0)\,.$$

Using $(I_a \otimes 1_{mn})1_a = 1_{amn}$,

$$\begin{aligned} AA' &= \Big\{(I_a \otimes 1_{mn})\big(I_a\lambda + (1-\lambda)1_a1'_a/a\big)\Big\}\Big\{\big(I_a\lambda + (1-\lambda)1_a1'_a/a\big)(I_a \otimes 1'_{mn})\Big\} \\ &= \big\{\lambda(I_a \otimes 1_{mn}) + (1-\lambda)1_{amn}1'_a/a\big\}\big\{\lambda(I_a \otimes 1'_{mn}) + (1-\lambda)1_a1'_{amn}/a\big\} \\ &= \lambda^2(I_a \otimes 1_{mn}1'_{mn}) + 2\lambda(1-\lambda)\,1_{amn}1'_{amn}/a + (1-\lambda)^2\,1_{amn}1'_{amn}/a \\ &= \lambda^2(I_a \otimes 1_{mn}1'_{mn}) + (1-\lambda^2)\,1_{amn}1'_{amn}/a\,. \end{aligned}$$

Thus, (C.2) is

$$\begin{aligned} &\mathrm{var}_{(y|\phi_0,d_0,\gamma_0,\nu_0)}\,\nu(y) = (\phi_0/mn + d_0/m + \gamma_0)A\,A' \\ &= (\phi_0/mn + d_0/m + \gamma_0)\Big\{\lambda^2(I_a \otimes 1_{mn}1'_{mn}) + (1-\lambda^2)\,1_{amn}1'_{amn}/a\Big\}\,. \end{aligned}$$

Now, (C.6) is

$$\begin{aligned}\mathrm{var}(\tilde{y}|\phi_0,d_0,\gamma_0,\nu_0) &= \lambda(\phi/mn)(I_a\otimes 1_{mn}1'_{mn})+(1-\lambda)(\phi/amn)1_{amn}1'_{amn}\\ &+ \phi I_{amn}+(\phi_0/mn+d_0/m+\gamma_0)\lambda^2(I_a\otimes 1_{mn}1'_{mn})\\ &+ (\phi_0/amn+d_0/am+\gamma_0/a)(1-\lambda^2)1_{amn}1'_{amn}\,.\end{aligned}$$

C.2 Proof of (B.3)

Write $F^*(y)$ in (B.2) as

$$F^*(y)=\frac{MS\mu(y)}{\{SS\theta(y)+SSE(y)\}/a(mn-1)} \tag{C.3}$$

where $MS\mu(y)=SS\mu(y)/(a-1)$ and $SS\mu(y)=mn\sum_{i=1}^{a}(\bar{y}_{i..}-\bar{y}_{...})^2$, $SS\theta(y)=\sum_{i=1}^{a}\sum_{j=1}^{m}n(\bar{y}_{ij.}-\bar{y}_{i..})^2$, and $SSE(y)=\sum_{i=1}^{a}\sum_{j=1}^{m}\sum_{k=1}^{n}(y_{ijk}-\bar{y}_{ij.})^2$. Approximating $E(F^*(\tilde{y})|y)$ as the ratio of the expected values of the numerator and denominator, we obtain

$$E(F^*(\tilde{y})|y) \doteq \frac{E_{\tilde{y}|y}\,MS\mu(\tilde{y})}{E_{\tilde{y}|y}\left\{SS\theta(\tilde{y})+SSE(\tilde{y})\right\}/a(mn-1)}\,. \tag{C.4}$$

When the fitted model is (2.5), conditioning on ϕ and $\mu=(\mu_1,\ldots,\mu_a)$, $SS\theta(\tilde{y})/\phi \sim \chi^2_{a(m-1)}$, and independently, $SSE(\tilde{y})/\phi\sim\chi^2_{am(n-1)}$. Hence,

$$\{\big(SS\theta(\tilde{y})+SSE(\tilde{y})\big)/\phi\}\mid\mu,\phi\sim\chi^2_{a(mn-1)}\,.$$

It then follows that the denominator of (C.4) is

$$E_{\tilde{y}|y}\,\frac{SS\theta(\tilde{y})+SSE(\tilde{y})}{a(mn-1)} = E_{\mu,\nu|y}E_{\tilde{y}|\mu}\,\frac{SS\theta(\tilde{y})+SSE(\tilde{y})}{a(mn-1)} = E_{\mu,\nu|y}\,\phi=\phi\,. \tag{C.5}$$

To find the numerator of (C.4), we write

$$\begin{aligned}E_{\tilde{y}|y}\,(a-1)MS\mu(\tilde{y}) &= E_{\tilde{y}|y}\sum_{i=1}^{a}mn(\bar{y}_{i..}-\bar{y}_{...})^2 = E_{\tilde{y}|y}\,mn\Big(\sum_{i=1}^{a}\bar{y}^2_{i..}-a\bar{y}^2_{...}\Big)\\ &= E_{\tilde{y}|y}\left\{\tilde{y}'\Big((I_a\otimes 1_{mn}1'_{mn}/mn)-1_{amn}1'_{amn}/amn\Big)\tilde{y}\right\}\\ &= E_{\tilde{y}|y}\,(\tilde{y}'\,A\,\tilde{y})=tr\{A\,D(y)\}+\nu(y)'A\,\nu(y)\,,\end{aligned} \tag{C.6}$$

where $A=\{(I_a\otimes 1_{mn}1'_{mn}/mn)-1_{amn}1'_{amn}/amn\}$, $D(y)$ and $\nu(y)$ are given in (A.1). It can be shown that

$$\begin{aligned}&A\,D(y)=\left\{(I_a\otimes 1_{mn}1'_{mn}/mn)-1_{amn}1'_{amn}/amn\right\}\times\\ &\left\{\lambda(\phi/mn)(I_a\otimes 1_{mn}1'_{mn})+(1-\lambda)(\phi/amn)(1_{amn}1'_{amn})+\phi I_{amn}\right\}\\ &=\lambda(\phi/mn)(I_a\otimes 1_{mn}1'_{mn})+(1-\lambda)(\phi/amn)\,1_{amn}1'_{amn}+\phi(I_a\otimes 1_{mn}1'_{mn}/mn)\\ &-\lambda(\phi/amn)1_{amn}1'_{amn}-(1-\lambda)(\phi/amn)(1_{amn}1'_{amn})-(\phi/amn)\,1_{amn}1'_{amn}\,.\end{aligned}$$

After algebraic manipulation,

$$tr\{A\,D(y)\} = \phi(a-1)(\lambda+1)\,. \tag{C.7}$$

The second term in (C.6) is

$$\begin{aligned}\nu(y)'A\,\nu(y) &= \Big\{(\lambda\bar{y}+(1-\lambda)\bar{y}_{\cdots}1_a)'(I_a\otimes 1'_{mn})\Big\}\Big\{(I_a\otimes 1_{mn}1'_{mn}/mn)\\ &\quad - 1_{amn}1'_{amn}/amn\Big\}\Big\{(I_a\otimes 1_{mn})(\lambda\bar{y}+(1-\lambda)\bar{y}_{\cdots}1_a)\Big\}\\ &= mn(\lambda\bar{y}+(1-\lambda)\bar{y}_{\cdots}1_a)'(I_a-1_a1'_a/a)(\lambda\bar{y}+(1-\lambda)\bar{y}_{\cdots}1_a)\,.\end{aligned}$$

Letting $B = I_a - 1_a1'_a/a$,

$$\begin{aligned}\nu(y)'A\,\nu(y) &= mn\lambda^2\bar{y}'B\bar{y} + mn(1-\lambda)^2\bar{y}^2_{\cdots}1'_aB1_a + mn\lambda\,(1-\lambda)\bar{y}'B\bar{y}_{\cdots}1_a\\ &+ mn\lambda\,(1-\lambda)\bar{y}_{\cdots}1'_aB\bar{y}\\ &= mn\lambda^2\bar{y}'B\bar{y} + 0 + 0 + 0\,,\end{aligned}$$

where the second, third and fourth terms are zero because $1'_aB = 0$ and $\bar{y}'B\bar{y}_{\cdots}1_a = 0$. Hence,

$$\begin{aligned}\nu(y)'A\,\nu(y) &= mn\lambda^2\bar{y}'\big(I_a-1_a1'_a/a\big)\bar{y} = \lambda^2\sum_{i=1}^{a}mn(\bar{y}_{i\cdot\cdot}-\bar{y}_{\cdots})^2\\ &= \lambda^2(a-1)MS\mu(y)\,.\end{aligned} \tag{C.8}$$

Using (C.7) and (C.8), (C.6) is

$$E_{\tilde{y}|y}\,(a-1)MS\mu(\tilde{y}) = \phi(a-1)(\lambda+1) + \lambda^2(a-1)MS\mu(y)\,.$$

Thus, the numerator of (C.4) is

$$E_{\tilde{y}|y}\,MS\mu(\tilde{y}) = \phi(\lambda+1) + \lambda^2MS\mu(y)\,. \tag{C.9}$$

Finally, (C.5) and (C.9) yield the posterior predictive expectation in (C.4)

$$E_{\tilde{y}|y}\,F^*(\tilde{y}) \doteq (\lambda+1) + (\lambda^2/\phi)MS\mu(y)\,.$$

C.3 Proof of (B.6)

When the data are from the three-stage model, (2.7), note that

$$\begin{aligned}&E\left\{SS\theta(y) + SSE(y)\right\}/a(mn-1)\\ &= \Big\{a(m-1)E\,MS\theta(y) + am(n-1)E\,MSE(y)\Big\}/a(mn-1)\\ &= \left\{a(m-1)(nd_0+\phi_0) + am(n-1)\phi_0\right\}/a(mn-1)\,,\end{aligned}$$

where $SS\theta(y)$ and $SSE(y)$ are given in (C.3). Hence,

$$\frac{SS\theta(y)+SSE(y)}{\left\{a(m-1)(\phi_0+nd_0) + am(n-1)\phi_0\right\}/a(mn-1)} \sim \chi^2_{a(mn-1)}\,.$$

Also, it can be shown that

$$(a-1)MS\mu(y)/(\phi_0 + nd_0 + mn\gamma_0) \quad \sim \quad \chi^2_{a-1}\,.$$

Since the above two terms are independent, it follows immediately that

$$\begin{aligned} F^*(y) &= \frac{MS\mu(y)}{\{SS\theta(y) + SSE(y)\}/a(mn-1)} \\ &\sim \left\{\frac{(\phi_0 + nd_0 + mn\gamma_0)(mn-1)}{(m-1)(\phi_0 + nd_0) + m(n-1)\phi_0}\right\} \mathcal{F}_{(a-1),a(mn-1)}\,. \end{aligned}$$

Acknowledgments

The authors are grateful to Professors Jiming Jiang, Balgobin Nandram and Tom Greene, the Editor, Associate Editor and referee for their helpful comments.