Simulation Based Model Checking for Hierarchical Models

Dipak K. Dey, Alan E. Gelfand, Tim B. Swartz
and Pantelis K. Vlachos*

Abstract

Recent computational advances have made it feasible to fit hierarchical models in a wide range of serious applications. If one entertains a collection of such models for a given data set, the problems of model adequacy and model choice arise. We focus on the former.

While model checking usually addresses the entire model specification, model failures can occur at each hierarchical stage. Such failures include outliers, mean structure errors, dispersion misspecification, and inappropriate exchangeabilities.

We propose another approach which is entirely simulation based. It only requires the model specification and that, for a given data set, one be able to simulate draws from the posterior under the model. By replicating a posterior of interest using data obtained under the model we can "see" the extent of variability in such a posterior. Then, we can compare the posterior obtained under the observed data with this medley of posterior replicates to ascertain whether the former is in agreement with them and accordingly, whether it is plausible that the observed data came from the proposed model. This suggests the large scale use of Monte Carlo tests, each focusing on a potential model failure. It thus suggests the possibility of examining not only the overall adequacy of the hierarchical model but, using suitable posteriors, the adequacy of each stage.

This raises the question of when individual stages are separable and checkable which we explore in some detail. Finally, we develop this strategy in the context of generalized linear mixed models and offer a simulation study to demonstrate its capabilities.

Key words: Discrepancy measures; Generalized linear mixed models; Monte Carlo tests; Sampling-based model fitting; Stagewise model adequacy

*Dipak K. Dey and Alan E. Gelfand are professors and Pantelis K. Vlachos is a Ph.D. candidate in the Department of Statistics at the University of Connecticut, Storrs, CT, USA. Tim B. Swartz is an associate professor in the Department of Mathematics and Statistics at Simon Fraser University, Burnaby, B.C., Canada. The genesis of this work occurred during a sabbatical visit by the third author at the University of Connecticut in the fall of 1994. The research of the first author was supported in part by NSF SCREMS grant DMS-9506557. The research of the second and fourth authors was supported in part by NSF grant DMS-9301316. The research of the third author was supported in part by the Natural Sciences and Engineering Research Council of Canada.

1 Introduction

Recent computational advances have made it feasible to fit hierarchical models in a wide range of serious applications. As one entertains a collection of such models for a given data set, one needs to address the problem of model determination, i.e., model adequacy and model selection. In this regard, it seems as though model choice has received the greater attention. The scope of activity is revealed in recent work on Bayes factors by Kass and Raftery (1995) and Raftery (1993); on cross validation by Pettit and Young (1990), Gelfand, Dey and Chang (1992), Gelfand and Dey (1994) and Gelfand (1995); on intrinsic Bayes factors by Berger and Pericchi (1993, 1995); and on posterior Bayes factors by Aitkin (1991). Related work on choosing amongst a very large number of models has appeared from McCulloch and George (1993) and from Madigan and Raftery (1993). Sampling based fitting of formal Bayesian model choice which handles the change in dimensionality and interpretation of parameters across models has been discussed by Carlin and Chib (1995), Green (1994) and Phillips and Smith (1994).

For model adequacy much less has been said. Model checking typically addresses the entire hierarchical specification. Model failures can occur at each hierarchical stage. Such failures include outliers, mean structure errors, dispersion misspecifications and inappropriate exchangeabilities. These failures are not necessarily distinct, e.g., is an extreme response due to the fact that the provided stochastic mechanism permits inadequate dispersion or to an inadequate elaboration of the mean structure under this mechanism! It seems safer to attempt to identify the presence of a stagewise failure without attributing a unique explanation to it. Moreover, in providing the probabilistic components of a hierarchical model, we rarely believe that any of the distributions are correct. Indeed, those specifications further removed from the data are often intentionally made less precise, not because we *believe* them to be correct but in order to permit the data to drive the inference. However, what is *true* is apart from model checking. If we undertake model criticism we must examine the adequacy of what is specified. In this regard, it is apparent that we must assume proper priors (or else the observed data could not have arisen under the model) and that very vaguely specified hierarchical models will be difficult to criticize. Model choice criteria are required to eliminate such models. Also, in examining high dimensional models, often having more parameters than data points, it becomes silly to speak of adequacy. Such models must be adequate unless prior specifications are sufficiently precise to permit conflict with the data. In our work here we are interested in that *middle* range of models which are not so parsimonious as to prevent useful hierarchical modeling but are not so high dimensional as to render adequacy a nonissue. We attempt more formal clarification of this notion in section 3 but suspect that in applications many modelers and data analysts find that such models strike a sensible balance.

The formal Bayesian model adequacy criterion (as in Box, 1980) proposes that the marginal density of the data be evaluated at the observations. Large values support the model, small values do not. Assessment of the magnitude of this value could be facilitated by standardizing, using the maximum value or an average value of this density (Berger, 1985). In our view, such a checking procedure will be infeasible for the models we are interested in. The required high dimensional density most likely will be difficult to estimate well and hopeless to calibrate. In addition, with hierarchical models the procedure does not provide feedback regarding the adequacy of the stagewise specifications. Alternative approaches for model criticism are needed. We briefly review what has been proposed.

Chaloner and Brant (1988) and Chaloner (1994), focusing on outlier detection, suggest what is ultimately posterior-prior comparison. A somewhat general version of their idea is to identify random variables whose distribution, a priori, is a standard one. In particular, they choose these variables to be so-called realized residuals. Given the data, the posterior distribution of each such residual is obtained. Any posterior which differs considerably from its associated prior is claimed to indicate an outlying observation. In particular, tail area comparisons are proposed. This approach seems attractive since, if the entire model specification is correct we don't anticipate substantial prior-data disagreement whence the posterior should be roughly centered where the prior is but more concentrated. However, as we amplify in section 5, such comparisons will be successful on average but fail to recognize the *variability* in the posterior when the model is correct. In section 2 we show that such comparison can also be interpreted, using cross-validation, as a diagnostic for influence.

A second approach is often referred to as model expansion or elaboration. The idea is that model failures can be captured by specifying a more complex model using mixtures. This is well discussed in the case of outlier detection as in, e.g., Guttman, Dutter and Freeman (1978), Freeman (1980), Pettit and Smith (1985) and Sharples (1990). Recent work of Albert and Chib (1994) considers other model failures, in particular, exchangeability in the direction of partial exchangeability. Regardless, the model of interest becomes a reduced model nested within the expanded or full model so that formal or informal model choice procedures can be used to criticize the adequacy of the reduced one. Model checking is converted to model choice. Recent work of Müller and Parmigiani (1995) and Carota, Parmigiani and Polson (1993) combines elaboration with posterior-prior comparison using the Kullback-Leibler distance between these two distributions for the elaboration parameter. Implementation of this approach requires exact specification of the expanded model, i.e., exactly what sort of failure one is looking to detect. Even then, a variety of elaborations can capture a particular type of failure so which one should we choose? It may be preferable to discern whether the observed data, when the model of interest is fitted to it, can directly reveal the presence of a failure of the model and its nature.

A third approach is taken up in Gelman, Meng and Stern (1995) who propose a posterior predictive strategy. The authors avoid the posterior Bayes factor (Aitkin, 1991) which has been effectively criticized (Lindley, 1991; Smith, 1991). Rather, they define a discrepancy measure as a function of data and parameters treating both as unknown in one case, inserting the observed data in the other. They then compare the resulting posterior distributions given the observed data. Gelman et al argue that the posterior predictive distribution treats the prior as an outmoded first guess whereas the prior predictive distribution treats the prior as a true "population distribution". While few would argue that the prior distribution is ever *correct*, as noted above, model checking checks the acceptability of the model fitted to the data. Furthermore, this posterior predictive approach can be criticized, as was the posterior Bayes factor, for, in essence, using the data twice. The observed data, through the posterior, suggests which values of the parameter are likely under the model. Then, to assess adequacy, the observed data is checked against data generated using such parameter values apparently making it difficult to criticize the model. We shall observe this in our example of section 5. In other words, the approach of Gelman et al is based upon the posterior distribution of a discrepancy measure given both the model and the data. It may be preferable to study the nature of the posterior distribution of this discrepancy given only the model and then compare it with the posterior distribution of this discrepancy obtained under the observed data.

A fourth approach is developed in recent work of Hodges (1994). He shows, in the case where all hierarchical levels are Gaussian, how to reexpress linear hierarchical models as standard linear models with simple covariance structure. He then suggests that familiar linear models diagnostic tools be brought into play, e.g., residual plots, added variable plots, transformations, collinearity checks, case influence, etc. There is a bit of adhockery in tailoring some of these tools to the hierarchical structure. In addition, the restriction to normal linear models is somewhat limiting.

We propose yet another approach which is entirely simulation based. Though more computationally intensive than the foregoing approaches (which nowadays ought not be viewed as an impediment), it is conceptually quite simple. It only requires the model specification and that, for a given data set, one be able to simulate draws from the posterior under the model. If the model can be fitted in this way to the observed data, then it is reasonable to suppose that it can be thus fitted to data generated under the model. The strategy becomes clear. By replicating a posterior of interest using data obtained under the model we can "see" the extent of variability in such a posterior. Then, we can compare the posterior obtained under the observed data with this medley of posterior replicates to ascertain whether the former is in agreement with them and accordingly, whether it is plausible that the observed data came from the proposed model. This suggests the large scale use of Monte Carlo tests, each focusing on a potential model failure. It also suggests that not only may we examine the overall adequacy of the hierarchical model but, using suitable posteriors, the adequacy of each hierarchical stage. This raises the question of whether or not the hierarchical model itself permits stagewise checking.

To reiterate, our approach compares posteriors arising under the data with associated posteriors arising under the specified model. Since posteriors are the tool for inference such comparison seems natural for model criticism.

The plan of the paper is thus the following. In section 2 we consider the family of generalized linear mixed models (Breslow and Clayton, 1993) introducing suitable discrepancy functions for checking such models. In section 3 we examine the matter of stagewise checkability of a model. In section 4 we detail the proposed simulation-based approach for carrying out such checking. Finally, in section 5 we illustrate with a binomial regression model.

2 Hierarchical Models and Discrepancy Functions

Our simulation-based approach to model checking employs discrepancy measures. We elaborate such measures here confining ourselves to exemplification with a two stage hierarchical model, though all ideas extend directly to the case of more than two stages. Since the methods of Gelman, Meng and Stern (1995) and Chaloner and Brant (1988) are also discrepancy-based, we summarize them with pertinent comments.

Let $\mathbf{y}, n \times 1$, denote the data vector and let $\boldsymbol{\theta}, p \times 1$ denote the first stage parameters, i.e., the likelihood is $f(\mathbf{y}\ |\boldsymbol{\theta})$. Customarily, conditional independence is assumed in which case $f(\mathbf{y}\ |\boldsymbol{\theta}) = \prod\limits_{i=1}^{n} f_i(y_i\ |\boldsymbol{\theta})$. Subscripting by i suggests a regression context where there is a covariate vector $\mathbf{x}_i$ associated with observation y_i. Let $\boldsymbol{\eta}, r \times 1$, denote the second stage parameters, i.e., the second stage model sets the distribution of $\boldsymbol{\theta}$ to be $f(\boldsymbol{\theta}|\ \boldsymbol{\eta})$. Here again, some sort of conditional independence, often exchangeability is customarily assumed whence we factor a portion of $\boldsymbol{\theta}$ say $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_m)$ into conditionally independent pieces. Then $f(\boldsymbol{\beta} \mid \boldsymbol{\eta}) = \prod\limits_{j=1}^{m} f_j(\boldsymbol{\beta}_j \mid \boldsymbol{\eta})$. Now the subscript j suggests, for a vector $\boldsymbol{\beta}_j$, an associated covariate matrix $\mathbf{Z}_j$. Lastly $f(\boldsymbol{\eta})$ denotes the prior on the hyperparameters $\boldsymbol{\eta}$.

More specifically, we assume that the first stage model is captured through a GLM with the first stage mean for $\mathbf{y}$ denoted by $g(X\boldsymbol{\beta} + W\boldsymbol{\alpha})$ where X is the design matrix associated with the exchangeable effects $\boldsymbol{\beta}$ while $\boldsymbol{\alpha}$ is a vector of "population" effects with W its associated design matrix. With the possible addition of a first stage dispersion parameter, ϕ, we would have $\boldsymbol{\theta}= (\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi)$. The second stage model for the $\boldsymbol{\beta}_j$ is typically multivariate normal or a *mixture* of multivariate normals such as a multivariate t. So too, for $\boldsymbol{\alpha}$. The second stage specification for ϕ, if present, is typically inverse Gamma. Hence, $\boldsymbol{\eta}$ denotes the undetermined parameters of these distributions.

We let d denote a general discrepancy function. First stage discrepancies will be denoted by $d_1(\mathbf{y},\boldsymbol{\theta})$. Examples can be at the individual level such as a raw residual $y_i - g(\mathbf{x}_i^T\boldsymbol{\beta}+\mathbf{w}_i^T\boldsymbol{\alpha})$, perhaps standardized, or possibly a deviance residual or a more global level such as the negative of the full first stage log likelihood, $-\log \prod\limits_{i=1}^{n} f(y_i \mid \mathbf{x}_i^T\boldsymbol{\beta}+\mathbf{w}_i^T\boldsymbol{\alpha}, \phi)$ or perhaps a portion of this product, i.e., a product over a particular subset of the i's. Second stage discrepancies denoted by $d_2(\boldsymbol{\theta},\boldsymbol{\eta})$ offer similar possibilities as do marginal discrepancies $d_{2|1}(\mathbf{y}, \boldsymbol{\eta})$ associated with the marginal model $f(\mathbf{y} \mid \boldsymbol{\eta})f(\boldsymbol{\eta})$ obtained by integrating out $\boldsymbol{\theta}$. We clarify the usefulness of marginal discrepancies and justify their notation in section 3. Overall model discrepancies involve comparison of a new data vector, $\mathbf{y}_{new}$, predicted under the model with the observed data vector $\mathbf{y}_{obs}$ and will be denoted by $d_0(\mathbf{y}_{obs}, \mathbf{y}_{new})$. Such d_0 are referred to as checking functions. Examples include $y_{i,obs} - y_{i,new}$ and the indicator function $1_{[y_{iobs}-\epsilon, y_{i,obs}+\epsilon]}(y_{i,new})$ with full discussion appearing in Gelfand, Dey and Chang (1992). Other discrepancies may be of interest such as loss functions but they may comment more on the utility of the model than its adequacy.

A specific structure which we will use in the sequel presumes conditionally independent observations y_{ij}, $i = 1, ..., n$, $j = 1, ..., m_i$ from a GLM with mean $g(\mathbf{x}_{ij}^T \boldsymbol{\beta}_i)$. The $\boldsymbol{\beta}_i$ are conditionally independent, say multivariate normal, i.e., $\boldsymbol{\beta}_i \sim N(Z_i \boldsymbol{\delta}, \Omega)$. Hyperparameters $\boldsymbol{\delta}$ and Ω have a normal-inverse Wishart prior specification. At the first stage we use the residuals $y_{ij} - \mu_{ij}$ where $\mu_{ij} = g(x_{ij}^T \boldsymbol{\beta}_i)$ or the deviance residuals (McCullagh and Nelder, 1989), $dev(y_{ij}, \mu_{ij})$. To study the entire first stage fit we can look at $-\log \prod_{i,j} f(y_{ij} \mid \boldsymbol{\theta})$; to study the consonance of population i with the model, $-\log \prod_{j} f(y_{ij} \mid \boldsymbol{\theta})$. For the second stage we use the residuals $\beta_{i\ell} - \mathbf{z}_{i\ell}^T \boldsymbol{\delta}$. A discrepancy measure for the entire second stage is $-\log \prod_{i} f(\boldsymbol{\beta}_i \mid \boldsymbol{\eta})$. For a marginal residual we use $y_{ij} - g(\mathbf{x}_{ij}^T Z_i \boldsymbol{\delta})$. A more appealing form is $y_{ij} - E(\mu_{ij} \mid \boldsymbol{\delta}, \Omega)$ but, apart from the case of a Gaussian first stage, the expectation is not available explicitly. Overall model fit could look at $-\log\{\prod_{i} f(\boldsymbol{\beta}_i \mid \boldsymbol{\eta}) \cdot \prod_{i,j} f(y_{ij} \mid \boldsymbol{\theta})\}$ as well as the individual $y_{ij,obs} - y_{ij,new}$. Since $E(y_{ij,obs} - \mu_{ij} \mid \mathbf{y}_{obs}) = E(y_{ij,obs} - y_{ij,new} \mid \mathbf{y}_{obs})$ but, via the Rao-Blackwell theorem $var(\ y_{ij,obs} - \mu_{ij} \mid \mathbf{y}_{obs}) \leq var(y_{ij,obs} - y_{ij,new} \mid \mathbf{y}_{obs})$, more dispersed posteriors arise when examining overall residuals than first stage residuals though they share a common center.

Note that all of the foregoing d's are unobservable random variables. If d is to comment upon the adequacy of the model for the observed data, we should examine a conditional distribution, i.e., the posterior distribution of d given $\mathbf{y}$. We have claimed that this is natural since the posteriors are used for inference. Moreover, the unconditional or prior distribution of d provides a less sensitive measure; indeed, the unconditional distribution of $d_2(\boldsymbol{\theta}, \boldsymbol{\eta})$ is not affected by what we have observed.

Gelman, Meng and Stern (1995) also suggest the use of discrepancy measures within the context of what they term Bayes goodness-of-fit testing. For say $d_1(\mathbf{y}_{obs}, \boldsymbol{\theta})$, they develop the posterior $f(d_1(\mathbf{y}_{obs},\boldsymbol{\theta}) \mid \mathbf{y}_{obs})$ and compare it with the posterior $f(d_1(\mathbf{y}_{new},\boldsymbol{\theta}) \mid \mathbf{y}_{obs})$. They do so by looking at a sample $\boldsymbol{\theta}_\ell^*$, $\ell = 1, ..., B$ from $f(\boldsymbol{\theta}| \mathbf{y}_{obs})$ and then for each $\boldsymbol{\theta}_\ell^*$ drawing $\mathbf{y}_\ell^*$ from $f(\mathbf{y} \mid \boldsymbol{\theta}_\ell^*)$. Hence $d_1(\mathbf{y}_{obs},\boldsymbol{\theta}_\ell^*) \sim f(d_1(\mathbf{y}_{obs},\boldsymbol{\theta}) \mid \mathbf{y}_{obs})$ and $d_1(\mathbf{y}_\ell^*,\boldsymbol{\theta}_\ell^*) \sim f(d_1(\mathbf{y}_{new},\boldsymbol{\theta}) \mid \mathbf{y}_{obs})$. Clearly, given ℓ, $d_1(\mathbf{y}_{obs},\boldsymbol{\theta}_\ell^*)$ and $d_1(\mathbf{y}_\ell^*,\boldsymbol{\theta}_\ell^*)$ are dependent, but the pairs are independent across ℓ. Hence, scatter plots and other pairwise comparison methods can be used to compare the associated posteriors. We are uncomfortable with such comparison since the generated $(\mathbf{y}_\ell^*,\boldsymbol{\theta}_\ell^*)$ depend upon $\mathbf{y}_{obs}$. This would appear to encourage $\mathbf{y}_\ell^*$ too much like $\mathbf{y}_{obs}$ resulting in model checking which is too optimistic (see section 5 for quantitative support). It seems more appropriate to compare $f(d_1(\mathbf{y}_{obs}, \boldsymbol{\theta}) \mid \mathbf{y}_{obs})$ with $f(d_1(\mathbf{y}_{gen},\boldsymbol{\theta}) \mid \mathbf{y}_{gen})$ where $\mathbf{y}_{gen}$ is a vector of observations generated under the assumed model. Indeed, we would want to have replications of this latter posterior to acquire a better sense of the range of typical posterior distributions for $d_1(\mathbf{y},\boldsymbol{\theta})$ when the model is correct. Of course, such distributions depend only upon the assumed model, not upon the actual $\mathbf{y}_{obs}$. These distributions could then be compared with $f(d_1(\mathbf{y}_{obs},\boldsymbol{\theta}) \mid \mathbf{y}_{obs})$ to determine if the latter seems likely under this model.

Lastly, suppose we consider a cross validation approach letting $\mathbf{y}_{(i)}$ denote $\mathbf{y}$ with y_i deleted. Suppose that $d_1(\mathbf{y},\boldsymbol{\theta})$ depends upon $\mathbf{y}$ only through y_i, as it would for a residual. Consider then $f(d_1 \mid \mathbf{y}_{(i)}) = \int f(d_1 \mid \boldsymbol{\theta}, \mathbf{y}_{(i)}) \cdot f(\boldsymbol{\theta}| \mathbf{y}_{(i)})$. Generally, $f(d_1 \mid \boldsymbol{\theta},\mathbf{y}_{(i)})$ does not depend upon $\mathbf{y}_{(i)}$ but often does not depend upon $\boldsymbol{\theta}$ as well. Examples include scale and location regression models of the form $a(y_i, \boldsymbol{\alpha}) = b(\mathbf{x}_i; \boldsymbol{\beta}) + \sigma\epsilon_i$ where ϵ_i has a specified distribution symmetric about 0. Here $d_1(y_i,\boldsymbol{\theta}) = \sigma^{-1}(a(y_i; \boldsymbol{\alpha}) - b(\mathbf{x}_i; \beta))$. Also, included are proportional hazards models (e.g., Cox and Oakes, 1984) which, in the most general case, with time-varying covariates, specify a hazard function of the form $a(t; \boldsymbol{\alpha}) \cdot b(\mathbf{x}(t), \boldsymbol{\beta})$. Here the integrated hazard $d_1(y_i,\boldsymbol{\theta}) = \int_0^{y_i} a(t; \boldsymbol{\alpha}) b(\mathbf{x}_i(t), \boldsymbol{\beta}) dt$ is distributed as $Exp(1)$.

If $f(d_1 \mid \boldsymbol{\theta}, \mathbf{y}_{(i)})$ is free of $\mathbf{y}_{(i)}$ and $\boldsymbol{\theta}$ then $f(d_1 \mid \mathbf{y}_{(i)}) = f(d_1)$, the prior distribution of d_1. Recall that Chaloner and Brant (1988) suggest posterior to prior comparison between $f(d_1 \mid \mathbf{y})$ and $f(d_1)$ for outlier detection. In the instant case this is, in fact, a comparison of $f(d_1 \mid \mathbf{y})$ and $f(d_1 \mid \mathbf{y}_{(i)})$ which suggests an interpretation as an influence diagnostic, i.e., as a measure of the influence of y_i on the posterior. In fact the Kullback-Leibler divergence between these two densities has been proposed to quantify estimative/predictive influence in Geisser (1987) with more general ϕ-divergences studied in Peng and Dey (1995). In any event, with regard to model checking, examination solely of $f(d_1(y_i,\boldsymbol{\theta}) \mid \mathbf{y}_{(i)})$ in such cases can not be useful since this distribution is unaffected by the value of $\mathbf{y}_{(i)}$.

3 Stagewise Model Checking

In section 2 we suggested both first and second stage discrepancy measures with the intent of stagewise checking of model assumptions. This raises the obvious question of when the individual stages of a hierarchical model can be separated and checked. That is, in principle we can always examine the overall adequacy of a model but when can we examine it in stages? In this section we attempt to illuminate this query.

Clearly, such stagewise checkability is a property of the model specification. It has nothing to do with the values of the observed data to which the model might be fitted! Again, the remark in section 1 that we are interested in that middle range of model dimensionality is pertinent. If the model has "too many" parameters, then it must have "too many" parameters at some stage to permit checking of that stage. We take up two notions, the information about the r^{th} stage model specification in the r^{th} stage discrepancy measure and the role of inter-stage dependence of discrepancy functions. We define the former to be a relative efficiency, $I(d_r) = var(d_r)/E(var(d_r \mid \mathbf{y})) \geq 1$, $r = 1, 2$ whence the larger $I(d)$ is, the more the model is capable of informing about d. $I(d_{2|1})$ is also useful to compute since, as we clarify below, $d_{2|1}$ can be used to check the second stage specification given that the first is true. For the latter we compute $E[corr(d_1, d_2 \mid \mathbf{y})]$. Note that both are model properties, i.e., properties solely of the model specification. Two simple examples illustrating extreme cases provide clarification.

Example 1: Suppose y_i given θ_i are independent $\sim N(\theta_i, \sigma^2)$, $i = 1, 2, ..., n$ and θ_i given $\mu \sim N(\mu, \tau^2)$. Assume σ^2 and τ^2 known and, for convenience, place a flat hyperprior on μ. By routine calculation we may demonstrate that

$$\begin{aligned}
&f(y_i - \theta_i \mid \mathbf{y}) = N((1-A)(y_i - \overline{y}), (1-A)(\tfrac{\sigma^2}{n} + \tau^2)) \\
&f(\theta_i - \mu \mid \mathbf{y}) = N(A(y_i - \overline{y}), A(\sigma^2 + \tfrac{\tau^2}{n})) \\
&f(y_i - \mu \mid \mathbf{y}) = N(y_i - \overline{y}, \tfrac{\sigma^2+\tau^2}{n}) \\
&f(y_{i,obs} - y_{i,new} \mid \mathbf{y}_{obs}) = N((1-A)(y_{i,obs} - \overline{y}_{obs}), (1-A)(\tfrac{\sigma^2}{n} + \tau^2) + \sigma^2)
\end{aligned}$$

and

$$\begin{aligned}
&cov(y_i - \theta_i, \theta_i - \mu \mid \mathbf{y}) = -(1 - \tfrac{1}{n})A\sigma^2 \\
&cov(y_i - \theta_i, y_j - \theta_j \mid \mathbf{y}) = (1-A)\tfrac{\sigma^2}{n} \\
&cov(\theta_i - \mu, \theta_j - \mu \mid \mathbf{y}) = A\tfrac{\tau^2}{n} \\
&cov(y_i - \theta_i, \theta_j - \mu \mid \mathbf{y}) = A\tfrac{\sigma^2}{n}
\end{aligned}$$

where $A = \frac{\tau^2}{(\sigma^2+\tau^2)}$, the usual *shrinkage* factor.

The information about the first stage model specification contained in the residual $d_1 = y_i - \theta_i$ is $I(d_1) = (\sigma^2 + \tau^2)/(\frac{\sigma^2}{n} + \tau^2)$. If τ^2 is large relative to σ^2, this relative efficiency is approximately 1 whence the data provides little information to criticize the first stage specification. Also, as $n \to \infty$, $I(d_1)$ tends to $1 + \frac{\sigma^2}{\tau^2}$ so that the information or learning about the first stage is bounded. Of course, with one observation per parameter at the first stage it is not surprising that the first stage assumption is difficult to check.

The information about the second stage model specification in the residual $d_2 = \theta_i - \mu$ is $I(d_2) = (\sigma^2+\tau^2)/(\sigma^2+\frac{\tau^2}{n})$. Now, when τ^2 is large relative to σ^2 this efficiency is approximately n, i.e., the data can criticize the second stage model better than the first. Again, with "n observations" for one parameter at the second stage this seems reasonable. Of course, as $n \to \infty$, $I(d_2)$ tends to $1 + \frac{\tau^2}{\sigma^2}$ indicating that even with an infinite amount of data we can't check the second stage perfectly because we only observe y_i and not θ_i.

Next, note that the marginal density $f(y_i \mid \mu) = N(\mu, \sigma^2 + \tau^2)$ is identifiable in the prior $f(\theta_i \mid \mu) = N(\mu, \tau^2)$ given $f(y_i \mid \theta_i) = N(\theta_i, \sigma^2)$, as in Teicher (1963). Hence, checking the marginal model is equivalent to checking the second stage given the first is true, hence, clarifying the subscripting for marginal discrepancies. Also, $I(d_{2|1}) = n$ so we can check the second stage arbitrarily well given the first stage is true using $d_{2|1}$. Again, with only one observation per population perhaps we must assume that the first stage specification is correct.

Turning to the correlation structure, we find $E[corr(y_i - \theta_i, \theta_i - \mu \mid \mathbf{y})] = -(1 - \frac{1}{n})/\sqrt{(1 + \frac{\sigma^2}{n\tau^2})(1 + \frac{\tau^2}{n\sigma^2})}$. When n is large this correlation tends to -1. But then, the model does not permit stagewise checking. If, say y_i is extreme relative to the other y's it will be difficult to discern whether to view y_i as an outlying observation or θ_i as an outlying population. As might be anticipated, the other expected correlations are much weaker. In fact, $E[corr(y_i - \theta_i, y_j - \theta_j \mid \mathbf{y})] = \frac{\sigma^2}{\sigma^2 + n\tau^2} \rightarrow 0$ as $n \rightarrow \infty$, $E[corr(\theta_i - \mu, \theta_j - \mu \mid \mathbf{y})] = \frac{\tau^2}{n\sigma^2 + \tau^2} \rightarrow 0$ as $n \rightarrow \infty$ and $E[corr(y_i - \theta_i, \theta_j - \mu \mid \mathbf{y})] = \{(n + \frac{\sigma^2}{\tau^2})(n + \frac{\tau^2}{\sigma^2})\}^{-\frac{1}{2}} \rightarrow 0$ as $n \rightarrow \infty$.

Example 2: Suppose y_i given θ are independent $N(\theta, \sigma^2)$, $i = 1, 2, ..., n$, θ given $\mu \sim N(\mu, \tau^2)$. (This model is not really hierarchical!) Again, assume σ^2 and τ^2 known and a flat prior on μ. By routine calculation, we have

$$\begin{aligned}
&f(y_i - \theta \mid \mathbf{y}) = N(y_i - \overline{y}, \tfrac{\sigma^2}{n}) \\
&f(\theta - \mu \mid \mathbf{y}) = N(0, \tau^2) \\
&f(y_{i,obs} - y_{i,new} \mid \mathbf{y}_{obs}) = N(y_{i,obs} - \overline{y}_{obs}, \tfrac{\sigma^2}{n} + \sigma^2)
\end{aligned}$$

and

$$\begin{aligned}
&cov(y_i - \theta, \theta - \mu \mid \mathbf{y}) = 0 \\
&cov(y_i - \theta, y_j - \theta \mid \mathbf{y}) = \tfrac{\sigma^2}{n}.
\end{aligned}$$

The information about the first stage model specification contained in $d_1 = y_i - \theta$ is $I(d_1) = n$. Hence, we can check the first stage arbitrarily well, not surprising with n data points and one first stage parameter. For the second stage the posterior for $d_2 = \theta - \mu$ is the same as the prior, i.e., $I(d_2) = 1$. The data contains no information about $\theta - \mu$ so we can't learn about the second stage distribution using d_2. Again, this is expected given "one observation" for one parameter at the second stage. Interestingly, $I(d_{2|1}) = (\sigma^2 + \tau^2)/(\frac{\sigma^2}{n} + \tau^2)$

so we can learn something about the adequacy of the second stage given the first is true.

With regard to correlation structure, $E[corr(y_i - \theta, \theta - \mu \mid \mathbf{y})] = 0$ which might be expected because the second stage model does not involve i. Combined with insight gleaned from example 1 we suggest that low inter-stage correlation is only a necessary but not sufficient condition for stagewise checkability. If inter-stage correlation is high, the model will be difficult to check stagewise. Note that $corr(y_i - \theta, y_j - \theta \mid \mathbf{y}) = 1$ (and hence so is its expectation). That is, up to centering, there is only one first stage residual posterior distribution. But, of course, the centering can indicate whether a particular y_i is outlying.

These two examples, as well as intuition, suggest that the number of first stage observations relative to the number of first stage parameters affects the checkability of the first stage. Similarly, the number of second stage observations relative to the number of hyperparameters affects the checkability of the second stage. One might assert that, without first stage replications, we can not expect to check the first stage, without second stage exchangeability of some sort we can not expect to check the second stage. In particular, recalling the specific structure of section 2, at the first stage we would need m_i to be large relative to the dimension of $\boldsymbol{\beta}_i$ though we make this statement qualitatively rather than quantitatively. Similarly, at the second stage we would need n, the number of $\boldsymbol{\beta}_i$'s, large relative to r, the number of hyperparameters in the distributional specification for $\boldsymbol{\beta}_i$. In this regard, examples 1 and 2 present extreme first and second stage cases. In example 1, $m_i = 1$ and dim $\beta_i = 1$; in example 2, $m_i = n$ and dim$\beta_i = 1$. In example 1, $n = m$ and $r = 1$; in example 2, $n = 1$ and $r = 1$. In the simulated illustration of section 5 we move away from these extremes. We set $m_i = 20$ and dim $\beta_i = 2$ and set $n = 20$ and $r = 4$.

The following argument adds more formal support and clarification. Consider the joint distribution of all first stage and all second stage residuals. These can be usual deviations or deviance residuals. We illustrate with deviations. Employing, as in section 2, a GLM first stage with Gaussian second stage, we suppress $\boldsymbol{\alpha}$ and assume that X has full column rank, as it usually does in hierarchical modeling. Then we claim that

$$f(\mathbf{y} - g(X\boldsymbol{\beta}),\ \boldsymbol{\beta} - Z\boldsymbol{\delta} \mid \mathbf{y}) = f(\mathbf{y} - g(X\boldsymbol{\beta}) \mid \mathbf{y}) \cdot f(\boldsymbol{\beta} - Z\boldsymbol{\delta} \mid \boldsymbol{\beta}). \tag{1}$$

The proof is straightforward. We factor the left hand side of (1) as

$$f(\mathbf{y} - g(X\boldsymbol{\beta}) \mid \mathbf{y}) \cdot f(\boldsymbol{\beta} - Z\boldsymbol{\delta} \mid \mathbf{y} - g(X\boldsymbol{\beta}), \mathbf{y}).$$

But, conditioning on $\mathbf{y} - g(X\boldsymbol{\beta})$ and X is equivalent to conditioning on $\mathbf{y}, \boldsymbol{\beta}$ and X. (The fact that X is full column rank and that g is strictly increasing justifies this.) Finally the hierarchical structure implies that $f(\boldsymbol{\beta} - Z\boldsymbol{\delta} \mid \mathbf{y}, \boldsymbol{\beta}, X) = f(\boldsymbol{\beta} - Z\boldsymbol{\delta} \mid \boldsymbol{\beta})$.

Expression (1) provides an attractive factorization of the joint distribution on the left hand side. The first term on the right hand side is the usual posterior but the second term is also a usual posterior for the model where $\boldsymbol{\beta}$ is viewed as the "data". Extension to more than two hierarchical stages is obvious. The importance of (1) is to argue conceptually that, using residuals, the posterior for a hierarchical model can be factored into a product of posteriors for one stage models. Thus, the potential to check stagewise can be connected to the "amount of data" at a given stage relative to the number of parameters introduced at that stage.

We next turn to the asymptotic behavior of general $I(d)$, again using the specific structure of section 2. There are three distinct limiting cases of interest : (i) $n \to \infty$, all m_i fixed, (ii) all $m_i \to \infty$, n fixed, (iii) all $m_i \to \infty$ and $n \to \infty$. Assuming usual regularity conditions, e.g., those which insure an appropriate Bayesian central limit theorem, under (i) $f(\boldsymbol{\eta} \mid \mathbf{y}_{obs})$ becomes degenerate, under (ii) $f(\boldsymbol{\beta}_i \mid \mathbf{y}_{obs})$ becomes degenerate for each i and under (iii) $f(\boldsymbol{\eta} \mid \mathbf{y}_{obs})$ and the $f(\boldsymbol{\beta}_i \mid \mathbf{y}_{obs})$ become degenerate. In appendix 1 we argue that, as a result, under (i) $I(d_{2|1}) \to \infty$ while $I(d_1)$ and $I(d_2)$ remain bounded, under (ii) $I(d_1) \to \infty$ while $I(d_2)$ and $I(d_{2|1})$ remain bounded and under (iii) $I(d_1)$, $I(d_2)$ and $I(d_{2|1}) \to \infty$.

The emergent conclusions agree with the foregoing intuition: we can effectively check the first stage model using the d_1's provided that m_i are large. Turning to $d_{2|1}$, as in example

1, if $f(y_{ij} \mid \boldsymbol{\eta})$ is identifiable in $f(\mu_{ij} \mid \boldsymbol{\eta})$ given $f(y_{ij} \mid \mu_{ij})$ where μ_{ij} denotes the first stage mean of y_{ij} then the $d_{2|1}$ residual based upon y_{ij} checks the adequacy of $f(\mu_{ij} \mid \boldsymbol{\eta})$. Hence, we need the identifiability of an exponential family mixture in its mean. Discussion and conditions appear in Teicher (1961) and in Tallis and Chesson (1982). But then, if we accept the correctness of the first stage, the $d_{2|1}$ can effectively check the second stage specification given n is large. The d_2's in such cases will generally be less effective in checking the second stage since, in addition, they require large m_i.

Examples where $m_i = 1$ occur frequently. In particular, in many models by data reduction the first stage specification takes the form $f(t_i \mid \theta_i)$ with t_i and θ_i one dimensional. As in Efron (1995), inference proceeds from the likelihood as a function of θ_i given t_i with no other reference to the original data which gave these likelihoods. But then, such reduction suggests that the first stage is assumed and thus, under identifiability, the second stage can be checked marginally using the $d_{2|1}$'s.

To illustrate the above discussion consider a Poisson model, $y_i \sim P_0(\theta_i)$, $i = 1, 2, ..., n$, with $\gamma_i = \log \theta_i \sim N(\beta x_i, a)$ and $\Sigma x_i = 0$, $(n-1)^{-1}\Sigma x_i^2 = 1$. Finally, $\beta \sim N(1, b)$. We study $I(d)$ for $d_1 = y_i - \theta_i$, $d_2 = \gamma_i - \beta x_i$ and $d_{2|1} = y_i - e^{\beta x_i}$ for the choices $n = 10, 40, 100$ with $a = .1$, $b = .1$ and $a = 1$, $b = .1$. The results, displayed in Table 1, support our theoretical work showing, in addition, that with a tighter prior on the γ_i it will be easier to criticize the model. In practice, for any given model we can compute, using simulation, the $I(d)$'s to assess their diagnostic promise with regard to model adequacy.

Turning to the association between stagewise d's we ask when the necessary condition of low interstage correlation obtains. For the Gaussian hierarchical linear model with known covariances, we may study posterior association between residuals explicitly. Since all posterior associations are free of $\mathbf{y}$, we can drop the expectation over $\mathbf{y}$ when studying interstage dependence. In particular, let $\mathbf{y} \mid \boldsymbol{\beta} \sim N(X\boldsymbol{\beta}, C_1)$ and let $\boldsymbol{\beta} \mid \boldsymbol{\delta} \sim N(Z\boldsymbol{\delta}, C_2)$ where C_1 and C_2 are assumed known and we take a flat prior on $\boldsymbol{\delta}$. Typically $C_1 = \sigma^2 I$ and C_2 is block diagonal. Examples 1 and 2 are special cases. Let $\mathbf{r}_1 = \mathbf{y} - X\boldsymbol{\beta}$, $\mathbf{r}_2 = \boldsymbol{\beta} - Z\boldsymbol{\delta}$,

$F = X^T(C_1 + XC_2X^T)^{-1}X$ and $B = (X^TC_1^{-1}X + C_2^{-1})^{-1}$. From appendix 2 we have the following posterior covariances:

$$\Sigma(\mathbf{r}_1 \mid \mathbf{y}) = X[B + BC_2^{-1}Z(Z^TFZ)^{-1}Z^TC_2^{-1}B]X^T$$
$$\Sigma(\mathbf{r}_2 \mid \mathbf{y}) = B + BC_2^{-1}Z(Z^TFZ)^{-1}Z^TC_2^{-1}B + Z(Z^TFZ)^{-1}Z^T - 2BC_2^{-1}Z(Z^TFZ)^{-1}Z^T$$
$$\Sigma(\mathbf{r}_1, \mathbf{r}_2 \mid \mathbf{y}) = -X[B + BC_2^{-1}Z(Z^TFZ)^{-1}Z^TC_2^{-1}B] + XBC_2^{-1}Z(Z^TFZ)^{-1}Z^T.$$

For component r_{1s} of $\mathbf{r}_1$ and component r_{2t} of $\mathbf{r}_2$ the posterior correlation is

$$corr(r_{1s}, r_{2t} \mid \mathbf{y}) = (\Sigma(\mathbf{r}_1, \mathbf{r}_2) \mid \mathbf{y})_{st} / \sqrt{(\Sigma(\mathbf{r}_1 \mid \mathbf{y}))_{ss}(\Sigma(\mathbf{r}_2 \mid \mathbf{y}))_{tt}}\ . \qquad (2)$$

Calculations, again in appendix 2, enable us to study the behavior of (2) as we let the second stage precision grow small or as we let the first stage precision grow small. Setting $C_2 = \lambda_2 C_{2,0}$, and letting λ_2 grow large allows second stage precision to tend to 0 while retaining the same correlation structure. A similar comment holds for the first stage if we set $C_1 = \lambda_1 C_{1,0}$. We find (2) to be $O(\lambda_2^{-\frac{1}{2}})$ as $\lambda_2 \rightarrow \infty$ while (2) is $O(1)$ as $\lambda_1 \rightarrow \infty$. Hence, presuming that second stage variability is large compared to that of the first stage encourages necessary weak inter-stage association.

In practice C_1 and C_2 would not be known and would be modeled as random using a prior specification. But commonly, second stage exchangeabilities are modeled with less precision than first stage errors again encouraging weak inter-stage association, facilitating stagewise checking. Also, once we leave the Gaussian first stage, analytic investigation of correlation structure is intractable. However, using simulation we can compute any desired expected posterior correlations.

With very vague prior specification at each stage, it is clear that it will be hard to criticize other aspects of the model. Such specification implicitly assumes that model checking is not of much interest. A model adequacy investigation will be more illuminating when well-considered, informative prior specifications are employed. As a related point, under imprecise priors, using Markov chain Monte Carlo fitting of hierarchical models, problems with drifting of chains typically occurs and convergence is hard to achieve. More precise priors tend to eliminate these difficulties.

4 Simulation Based Model Checking

The foregoing discussion supports examination of the posterior distributions of various discrepancy measures to carry out model checking. More explicitly, for our observed data set and a given model which has been fitted to the data, we can obtain a collection of posterior distributions associated with first stage, second stage, marginal and overall discrepancies. How may we use these to assess whether the model is adequate for the data? We assert that, considered by themselves, it will be difficult to provide a satisfactory answer. What would be appealing is comparison of these posteriors respectively with corresponding posteriors arising from data obtained under the presumed model. Thus, a simulation based approach to model checking is suggested. Under the model, generate a data set analogous to the observed one and calculate the corresponding posteriors. Posterior vs posterior comparison would indicate whether that obtained from the observed data is similar to that obtained with data actually generated under the model.

But why generate only a single set of data under the model? More effective posterior-posterior comparison is achievable if we replicate, say R data sets, obtaining for each set the collection of posteriors of interest. Then, for any particular discrepancy function, we would have R posteriors against which to compare the one calculated from the observed data. If the latter appears consonant with the other R, then we would conclude that, for this discrepancy, the observed data does not criticize the model. Again, the benefit in replication is the opportunity to appreciate the variability in posteriors in order to better assess where the one for the observed data fits in.

In this regard, suppose $\mathbf{y}$ is generated under the model and consider the posterior probability that d is at most c, $\Pr(d \leq c \mid \mathbf{y})$. Then, averaging over $\mathbf{y}$, $E(\Pr(d \leq c \mid \mathbf{y})) = \Pr(d \leq c)$, the prior probability. In other words, the collection of R posteriors obtained for d vary about the prior, $f(d)$. But we do not advocate comparing $f(d \mid \mathbf{y}_{obs})$ with $f(d)$, which would return us to the approach of Chaloner and Brant (1988). Rather, we recognize that $f(d \mid \mathbf{y})$

will vary about $f(d)$ and use replication to learn how much variation we can expect when the model is correct.

In order to simulate data under the presumed model, in general we would draw $\boldsymbol{\eta}$, then $\boldsymbol{\theta}$ given $\boldsymbol{\eta}$, and lastly $\mathbf{y}$ given $\boldsymbol{\theta}$. Thus, all prior specifications be proper. This is not restrictive since otherwise how could the observed data have arisen under the presumed model? In practice we think improper priors are rarely needed or justified. Improper priors risk the possibility of improper posteriors, which may be hard to check for some hierarchical models. Moreover, as we noted in the previous section, improper priors often cause difficulties in using Markov chain Monte Carlo methods to fit hierarchical models.

Extending our earlier notation, let $\mathbf{y}^{(r)}$ denote the data set generated at the r^{th} replication and, for convenience, let $\mathbf{y}^{(0)}$ denote the observed data. Then for a given discrepancy d we seek to compare $f(d \mid \mathbf{y}^{(0)})$ with the set $f(d \mid \mathbf{y}^{(r)})$, $r = 1, 2, ..., R$. How shall we make this comparison? In this regard, note that, as a result of anticipated sampling based fitting required for the hierarchical model, rather than obtaining $f(d \mid \mathbf{y}^{(r)})$ we will have obtained a sample $d_\ell^{*(r)}$, $\ell = 1, 2, ..., B$ from this distribution. Hence, practically, what is required is comparison of R samples of size B with a single sample of size B obtained using the observed data.

Also note that, from section 2, we expect to study many d's, i.e., d_1's, d_2's, $d_{2|1}$'s. In fact, inclusion of additional discrepancy diagnostics adds little to the computational cost. The only concern is sufficient memory space. For instance, in the example of section 5 we have a model with y_{ij}, $i = 1, ..., 20$, $j = 1, ..., 20$ and $\boldsymbol{\beta}_i$ a 2×1 vector leading, as we shall see, to 420 d_1's, 40 d_2's and 400 $d_{2|1}$'s. Hence, we would have 860 sets of $R + 1$ replicates each of sample size B. With R and B typically in the range of 500 to (preferably) 1000, storage requires delicacy, bookkeeping requires care. As we indicated at the outset, such simulation based model checking is very computationally demanding!

Since specification of a distribution, in general, would be viewed as an infinite dimensional problem, formal comparison of a given density with a set of R densities is an infinite dimensional problem as well. (Were we to insist that all densities belong to a particular parametric family, we could simplify matters but such an assumption seems inappropriate.) We could attempt informal graphical comparison, plotting the R densities, $f(d \mid \mathbf{y}^{(r)})$ and overlaying $f(d \mid \mathbf{y}^{(0)})$. We suspect such a display would be hard to grasp visually and difficult to interpret. Moreover, while we might attempt such a plot for a particular discrepancy function d, it is clearly infeasible for say 860 different d's. What is needed is an automatable comparison procedure.

In this regard, Monte Carlo tests (Barnard 1963; Besag and Clifford, 1989) are standard procedures for comparing values generated under a given model with an observed value. But, Monte Carlo tests are univariate in nature since values must be ordered to determine whether an observed value is extreme relative to the generated values. They can not be directly applied to the $R+1$ B-dimensional data vectors which we have for each discrepancy function d. To implement such a test necessitates reducing each B-dimensional vector to a point, after which a customary Monte Carlo test could be conducted on the resulting $R+1$ points.

Obviously there is no unique reduction of a sample from a distribution to a single summary number for that distribution. Moreover, a sample of B points, apparently different from R other samples, need no longer look so when reduced to a single point. Hence, we may wish to run Monte Carlo tests on several different reductions. In fact, in the development of this work, we have experimented with several choices employing the following one in our illustrative example.

For any continuous discrepancy we use the B observations at replicate r to obtain five quantiles, say $q_{.05}^{(r)}$, $q_{.25}^{(r)}$, $q_{.5}^{(r)}$, $q_{.75}^{(r)}$ and $q_{.95}^{(r)}$ which we write as a vector $\mathbf{q}^{(r)}$. Next we obtain the average of each of these quantiles over r, $r = 1, 2, ..., R$, $\overline{q}_{.05}$, $\overline{q}_{.25}$, $\overline{q}_{.5}$, $\overline{q}_{.75}$, and $\overline{q}_{.95}$, which we also write as a vector $\overline{\mathbf{q}}$. (Following the earlier discussion these latter quantiles approximate those of the prior distribution for the discrepancy.) For each r, $r = 0, 1, ..., R$ we compute the Euclidean distance between $\mathbf{q}^{(r)}$ and $\overline{\mathbf{q}}$. For the resulting set of $r+1$ distances, we run a one-sided upper tail level .05 Monte Carlo test. We recognize that, because $\overline{\mathbf{q}}$ is computed from the $\mathbf{q}^{(r)}$, these distances are not independent. We choose not to be concerned about this since actual correlation between distances is very small, since our choice of reduction is arbitrary and since our use for these Monte Carlo tests is as an informal model criticism diagnostic.

We also ignore entirely the issue of simultaneous inference. That is, a Monte Carlo test will be run for each discrepancy of interest (860 in our example of section 5). Again, from a diagnostic perspective we do not care about the simultaneous correctness of all of these dependent tests. Our only calibration is the expected one-in-twenty rejection rate if the model specification is correct.

In the case of a discrete discrepancy, for instance a d_0 when y is discrete such as binomial or Poisson, we replace the five quantiles for replicate r with bin frequencies calculated after creating five bins for the possible values of d_0. The bin frequencies sum to B. Analogously, average bin frequencies can be calculated and thus, a Euclidean distance for each r from this vector of averages. Again, we run a one-sided Monte Carlo test on the resultant set of $R+1$ distances.

Lastly, we want to compare our adequacy approach to that of Gelman, Meng and Stern (1995) which, as noted in section 2, can be applied to certain discrepancy functions. As opposed to our foregoing Monte Carlo tests, the latter results in pairwise comparisons to compute a so-called Bayesian p-value, hence to determine rejection or acceptance. In our setting the method of Gelman et al is not applicable to the d_2's since d must involve $\mathbf{y}$. For the d_1's and the $d_{2|1}$'s the implementation is as follows. For d_1 we seek the one-sided p-value $P(d_1(\mathbf{y},\boldsymbol{\theta}) \geq d_1(\mathbf{y}_{obs},\boldsymbol{\theta}) \mid \mathbf{y}_{obs})$. From section 2, if we can calculate $P(d_1(\mathbf{y},\boldsymbol{\theta}_\ell^*) \geq d_1(\mathbf{y}_{obs},\boldsymbol{\theta}_\ell^*) \mid \boldsymbol{\theta}_\ell^*)$ for each ℓ we can average these values to obtain a Monte Carlo integration for the p-value. If not, then we must generate the $\mathbf{y}_\ell^*$'s and compute the proportion of times $d_1(\mathbf{y}_\ell^*,\boldsymbol{\theta}_\ell^*)$ exceeded $d_1(\mathbf{y}_{obs},\theta_\ell^*)$. For the $d_{2|1}$'s, apart from the Gaussian case, to calculate

$P(d_{2|1}(\mathbf{y},\boldsymbol{\eta}) \geq d_{2|1}(\mathbf{y}_{obs},\boldsymbol{\eta}) \mid \mathbf{y}_{obs})$ will always require the latter comparisons. In the case where d is a residual of the form $y_{ij} - (\cdot)$ all of the above events reduce to $P(y_{ij,new} \geq y_{ij,obs} \mid \mathbf{y}_{obs})$. Hence, using the approach of Gelman et al, d_1 does not check the first stage and $d_{2|1}$ does not check the second stage but rather we have a d_0 checking the entire model specification.

5 An Example

We consider an illustrative hierarchical binomial regression model of the following form. The first stage has $Y_{ij} \sim Bin(n_{ij}, p_{ij})$, $i = 1, ..., 20$ populations, $j = 1, ..., m_i$ with $m_i = 20$ and all $n_{ij} = 20$ where $p_{ij} \equiv p(\alpha_i, \beta_i) = \frac{e^{\alpha_i+\beta_i x_{ij}}}{1+e^{\alpha_i+\beta_i x_{ij}}}$ and $x_{ij} = \frac{j-\mu(j)}{\sigma(j)}$, $\mu(j) = \sum_{j=1}^{m_i} j/m_i$, and $\sigma(j) = \sqrt{\sum_{j=1}^{m_i} (j-\mu(j))^2/(m_i-1)}$. The second stage assumes $\binom{\alpha_i}{\beta_i} \sim N\left(\binom{\mu_\alpha}{\mu_\beta}, \begin{pmatrix} \tau_\alpha^2 & 0 \\ 0 & \tau_\beta^2 \end{pmatrix}\right)$, $i = 1, ..., 20$ with the illustrative informative priors, $\tau_\alpha^2 \sim IG(c, d)$ such that $E(\tau_\alpha^2) = 10$, $Var(\tau_\alpha^2) = 3$, $\tau_\beta^2 \sim IG(e, f)$ such that $E(\tau_\beta^2) = 1$, $Var(\tau_\beta^2) = 1$ and $\binom{\mu_\alpha}{\mu_\beta} \sim N\left(\binom{0}{2}, \begin{pmatrix} 10 & 0 \\ 0 & 1 \end{pmatrix}\right)$. Sampling based fitting of this model is accomplished using Metropolis steps within the Gibbs sampler. Table 2 summarizes the checkability of this model in terms of the $I(d)$ and the interstage correlations using 1000 replications each providing 1000 posterior samples. We see that associations are weak, that $d_{2|1}$ should be very effective, d_1 less so with the d_2's offering little promise.

We next obtained 3 observed data sets $\mathbf{y}^{(0)}$ arising under the following 3 models.

Case 1: $Y_{ij} = \frac{1}{2}Bin(n_{ij}, p_{ij}^{(1)}) + \frac{1}{2}Bin(n_{ij}, p_{ij}^{(2)})$ where $p_{ij}^{(r)} = \frac{k_r e^{\alpha_i+\beta_i x_{ij}}}{1+k_r e^{\alpha_i+\beta_i x_{ij}}}$ with $k_1 = 1$ and $k_2 = e$, a first stage failure of the distributional specification with the second stage correct.

Case 2: $p_{ij} = \frac{e^{\alpha_i+\beta_i x_{ij}}}{1+e^{\alpha_i+\beta_i x_{ij}}}$, $j = 1, ..., 10$ and $p_{ij} = \frac{e^{2+\alpha_i+\beta_i x_{ij}}}{1+e^{2+\alpha_i+\beta_i x_{ij}}}$, $j = 11, ..., 20$, a first stage failure of the mean specification with the second stage correct.

Case 3: First stage correct with $\binom{\alpha_i}{\beta_i} \sim N\left(\binom{\mu_\alpha}{\mu_\beta}, \begin{pmatrix} \tau_\alpha^2 & 0 \\ 0 & \tau_\beta^2 \end{pmatrix}\right)$, $i = 1, ..., 10$ and $\binom{\alpha_i}{\beta_i} \sim$

$N\left(\left(\begin{smallmatrix}\mu_\alpha \\ -\mu_\beta\end{smallmatrix}\right), \left(\begin{smallmatrix}\tau_\alpha^2 & 0 \\ 0 & \tau_\beta^2\end{smallmatrix}\right)\right)$, $i = 11, ..., 20$, a second stage failure in the form of partial exchangeability. Though artificial in practice, here, using known model failures is advantageous in enabling us to determine how well we can handle each type of violation.

Table 3 summarizes the findings. In Case 1, we (DGSV) accurately diagnose where the failure lies while Gelman et al (GMS) do not criticize the model. In Case 2, we again find the correct failure and now GMS do find failure with the entire model. In Case 3, we clearly find the second stage failure using $d_{2|1}$ as well as $d_2 = \beta_1 - \mu_\beta$ with the suggestion of an erroneous first stage failure. Again, GMS do not criticize the model.

References

[1] Aitkin, M. (1991). "Posterior Bayes factors" (with discussion), J.R. Statist. Soc., B, 53, 111-142.

[2] Albert, J.H. and Chib, S. (1994). "Bayesian tests and model diagnostics in conditionally independent hierarchical models", Technical Report, Department of Mathematics and Statistics, Bowling Green State University.

[3] Barnard, G.A. (1963). J.R. Statist. Soc., B, 25, p. 294-295 (in discussion).

[4] Berger, J.O. and Perrichi, L.R. (1993). "The intrinsic Bayes factor for model selection and prediction", Technical Report #93-43C, Department of Statistics, Purdue University.

[5] Berger, J.O. and Perrichi, L.R. (1995). "The intrinsic Bayes factor for linear models", In: *Bayesian Statistics 5*, Eds: J.M. Bernardo, et al, Oxford, U.K., Oxford University Press (to appear).

[6] Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*, New York: Springer-Verlag.

[7] Besag, J. and Clifford (1989). "Generalized Monte Carlo significance tests", Biometrika, 76, 633-642.

[8] Box, G.E.P. (1980). "Sampling and Bayes's inference in scientific modeling" (with discussion), J.F. Statist. Soc., A, 143, 383-430.

[9] Breslow, N.E. and Clayton, D.G. (1993). "Approximate inference in generalized linear mixed models", J. Amer. Statist. Assoc., 88, 9-25.

[10] Carlin, B.P. and Chib, S. (1995). "Bayesian model choice via Markov chain Monte Carlo", J.R. Statist. Soc., B, (to appear).

[11] Carota, C., Parmigiani, G. and Polson, N.G. (1993). "Diagnostic measures for model criticism", Technical Report #93-A20, ISDS, Duke University, Durham.

[12] Chaloner, K. (1994). "Residual analysis and outliers in Bayesian hierarchical models". In: *Aspects of Uncertainty*, eds. A.F.M. Smith and P.R. Freeman. Chichester, U.K., Wiley, 153-161.

[13] Chaloner, K. and Brant, R. (1988). "A Bayesian approach to outlier detection and residual analysis", Biometrika, 75, 651-659.

[14] Cox, D.R. and Oakes, D. (1984). *Analysis of Survival Data.* London, Chapman and Hall.

[15] Efron, B. (1995). "Empirical Bayes methods for combining likelihoods" (with discussion), J. Amer. Statist. Assoc., (to appear).

[16] Freeman, P.R. (1980). "On the number of outliers in data from a linear model:. In: *Bayesian Statistics*, eds. J.M. Bernardo et al., Valencia, University Press, 349-365.

[17] Geisser, S. (1980). J.R. Statist. Soc., A, 143, 416-417 (in discussion).

[18] Geisser, S. (1987). "Influential observations, diagnostics and discordancy tests", J. Appl. Statist., 14, 133-142.

[19] Gelfand, A.E. (1995). "Model determination arising sampling based methods. In: *Markov Chain Monte Carlo in Practice*, Eds: W. Gilks, S. Richardson, D. Spiegelhalter, London, Chapman and Hall (to appear).

[20] Gelfand, A.E. and Dey, D.K. (1994). "Bayesian model choice: asymptotics and exact calculations", J.R. Statist. Soc., B, 56, 501-514.

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

[22] Gelman, A., Meng, X.L. and Stern, H.S. (1995). "Bayesian model checking using tail area probability", J.R. Statist. Soc., B (to appear).

[23] Green, P.J. (1994). "Reversible jump MCMC computation and Bayesian model determination", Technical Report, Department of Mathematics, University of Bristol.

[24] Guttman, I., Dutter, R. and Freeman, P.R. (1978). "Care and handling of multivariate outliers in the general linear model to detect spuriosity - a Bayesian approach", Technometrics, 20, 187-193.

[25] Hodges, J. (1994). "Some algebra and geometry for hierarchical models applied to diagnostics", Technical Report, Division of Biostatistics, University of Minnesota.

[26] Kass, R.E. and Raftery, A.E. (1995). "Bayes factors", J. Amer. Statist. Assoc., 90, 773-795.

[27] Lindley, D.V. (1991). Discussion of "Posterior Bayes factors", J.R. Statist. Soc., B, 53, 130-131.

[28] Lindley, D.V. and Smith, A.F.M. (1972). "Bayes estimates for the linear model", J.R. Statist. Soc., B, 34, 1-41 (with discussion).

[29] Madigan, D. and Raftery, A.E. (1994). "Model selection and accounting for model uncertainty in graphical models using Occam's window", J. Amer. Statist. Assoc., 89, 1535-1546.

[30] McCulloch, R.E. and George, E.I. (1993). "Variable selection via Gibbs sampling", J. Amer. Statist. Assoc., 88, 881-889.

[31] McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models.* London, Chapman and Hall.

[32] Müller, P. and Parmigiani, G. (1995). Numerical dvaluation of information theoretic measures. In: *Bayesian Statistics and Econometrics: Essays in Honor of A. Zellner*, Eds: Berry, D.A, Chaloner, K.M., Geweke, J.F., New York: Wiley.

[33] Peng, F. and Dey, D.K. (1995). "Bayesian analysis of outlier problems using divergence measures", Canadian Journal of Statistics, 23, 199-213.

[34] Pettit, L.I. and Smith, A.F.M. (1985). Outliers and influential observations in linear models. In: *Bayesian Statistics 2*, J.M. Bernardo et al. eds., Amsterdam: North Holland, 473-494 (with discussion).

[35] Pettit, L.I. and Young, K.D.S. (1990). "Measuring the effect of observations on Bayes factors", Biometrika, 77, 455-466.

[36] Phillips, D.B. and Smith, A.F.M. (1994). "Bayesian model comparison via jump diffusion", Technical Report #94-20, Department of Mathematics, Imperial College.

[37] Raftery, A.E. (1993). "Approximate Bayes factors and accounting for model uncertainty in generalized linear models", Technical Report #255, Department of Statistics, University of Washington.

[38] Sharples, L.D. (1990). "Identification and accommodations of outliers in generalized hierarchical models", Biometrika, 77, 445-452.

[39] Smith, A.F.M. (1991). Discussion of "Posterior Bayes factors", J.R. Statist. Soc., B, 53, 132-133.

[40] Tallis, G.M. and Chesson, P. (1982). "Identifiability of mixtures", J. Australian Math. Soc., A, 339-348.

[41] Teicher, H. (1961). "Identifiability of mixtures", Annals Math. Statist., 32, 244-248.

[42] Teicher, H. (1963). "Identifiability of finite mixtures", Annals Math. Statist., 34, 1265-1269.

APPENDICES

Appendix 1: Asymptotic Behavior of $I(d)$

Following section 3 and assuming appropriate regularity conditions, in case (i) $n \to \infty$, m_i fixed implies $f(\boldsymbol{\eta} \mid \mathbf{y}_{obs})$ becomes degenerate but the $f(\boldsymbol{\beta}_i \mid \mathbf{y}_{obs})$ do not. In case (ii) all $m_i \to \infty$, n fixed implies the $f(\boldsymbol{\beta}_i \mid \mathbf{y}_{obs})$ become degenerate but $f(\boldsymbol{\eta} \mid \mathbf{y}_{obs})$ does not. In case (iii) all $m_i \to \infty$ and $n \to \infty$ implies the $f(\boldsymbol{\beta}_i \mid \mathbf{y}_{obs})$ become degenerate and so does $f(\boldsymbol{\eta} \mid \mathbf{y}_{obs})$.

For case (i) it is convenient to write $I(d) =$

$$\frac{var(d)}{E_{\mathbf{y}_{obs}} E_{\boldsymbol{\eta}|\mathbf{y}_{obs}} var_{\boldsymbol{\beta}|\boldsymbol{\eta}_{obs}}(d|\mathbf{y}_{obs},\eta)) + E_{\mathbf{y}_{obs}} var_{\boldsymbol{\eta}|\mathbf{y}_{obs}} E_{\boldsymbol{\beta}|\boldsymbol{\eta}_{obs}}(d|\mathbf{y}_{obs},\boldsymbol{\eta}))}. \tag{A1}$$

The variance in the second term in the denominator tends to 0, hence, the term itself does. For d_1 and d_2 the first term in the denominator does not approach 0. Hence, $I(d_1)$ and $I(d_2)$ remain bounded. For $d_{2|1}$ the variance in the first term is 0, hence the term is as well, so $I(d_{2|1}) \to \infty$.

For cases (ii) and (iii) it is more convenient to write $I(d) =$

$$\frac{var(d)}{E_{\mathbf{y}_{obs}} E_{\boldsymbol{\beta}|\mathbf{y}_{obs}} var_{\boldsymbol{\eta}|\boldsymbol{\beta},y_{obs}}(d|\boldsymbol{\beta},\mathbf{y}_{obs}) + E_{\mathbf{y}_{obs}} var_{\boldsymbol{\beta}|\mathbf{y}_{obs}} E_{\boldsymbol{\eta}|\boldsymbol{\beta},\mathbf{y}_{obs}}(d|\boldsymbol{\beta},\mathbf{y}_{obs})}. \tag{A2}$$

For case (ii), using similar argumentation, for d_1, both terms in the denominator tend to 0, so $I(d_1) \to \infty$. For d_2's and $d_{2|1}$'s, the first term in the denominator does not tend to 0, so $I(d_2)$ and $I(d_{2|1})$ remain bounded. For case (iii) for each of d_1, d_2 and $d_{2|1}$, both terms in the denominator tend to 0, so $I(d_1)$, $I(d_2)$ and $I(d_{2|1})$ tend to ∞.

Appendix 2: Derivation of Posterior Covariances

We assume $\mathbf{y} \mid \boldsymbol{\beta} \sim N(X\boldsymbol{\beta}, C_1)$, $\boldsymbol{\beta} \mid \boldsymbol{\delta} \sim N(Z\boldsymbol{\delta}, C_2)$ and $\boldsymbol{\delta}$ has a noninformative prior $\pi(\boldsymbol{\delta}) = 1$. It follows immediately that $\mathbf{y} \mid \boldsymbol{\delta} \sim N(XZ\boldsymbol{\delta}, D)$ where $D = C_1 + XC_2X^T$. Now, using Lindley and Smith (1972) it follows that $\boldsymbol{\delta} \mid \mathbf{y} \sim N((Z^TFZ)^{-1}Z^TX^TD\mathbf{y}, \, (Z^TFZ)^{-1})$ where $F = X^TD^{-1}X$. Furthermore, $\boldsymbol{\beta} \mid \boldsymbol{\delta}, \mathbf{y} \sim N(B(X^TC_1^{-1}\mathbf{y} + C_2^{-1}Z\boldsymbol{\delta}), \, B)$ where $B = (X^TC_1^{-1}X + C_2^{-1})^{-1}$. Now,

$$\Sigma(\boldsymbol{\beta} \mid \mathbf{y}) = E[\Sigma(\boldsymbol{\beta} \mid \boldsymbol{\delta}, \mathbf{y})] + \Sigma[E(\boldsymbol{\beta} \mid \boldsymbol{\delta}, \mathbf{y})] = B + BC_2^{-1}Z(Z^TFZ)^{-1}Z^TC_2^{-1}B. \quad \text{(A3)}$$

Similarly,

$$\Sigma(\boldsymbol{\beta}, \boldsymbol{\delta} \mid \mathbf{y}) = BC_2^{-1}Z(Z^TFZ)^{-1}. \qquad \text{(A4)}$$

Now, defining residual vectors at the two stages as $\mathbf{r}_1 = \mathbf{y} - X\boldsymbol{\beta}$ and $\mathbf{r}_2 = \boldsymbol{\beta} - Z\boldsymbol{\delta}$, it follows from (A3) that

$$\Sigma(\mathbf{r}_1 \mid \mathbf{y}) = X\Sigma_{(\boldsymbol{\beta}|\mathbf{y})}X^T = X[B + BC_2^{-1}Z(Z^TFZ)^{-1}Z^TC_2^{-1}B]X^T.$$

Further, we have, using (A4)

$$\Sigma(\mathbf{r}_2 \mid \mathbf{y}) = B + BC_2^{-1}Z(Z^TFZ)^{-1}Z^TC_2^{-1}B^T + Z(Z^TFZ)^{-1}Z^T - 2BC_2^{-1}Z(Z^TFZ)^{-1}Z^T.$$

Finally,

$$\Sigma(\mathbf{r}_1, \mathbf{r}_2 \mid \mathbf{y}) = -Z(B + BC_2^{-1}Z(Z^TFZ)^{-1}Z^TC_2^{-1}B) + XBC_2^{-1}Z(Z^TFZ)^{-1}Z^T.$$

(a,b)	$(0.1,0.1)$			$(1.0,0.1)$		
n	10	40	100	10	40	100
d_1	1.83×10^1	2.54×10^7	4.09×10^{14}	6.27	1.65×10^2	6.80×10^5
$d_{2\|1}$	2.61×10^1	3.36×10^7	5.95×10^{14}	1.55×10^2	1.59×10^4	3.05×10^7
d_2	5.18	5.62	5.92	5.60	6.20	9.43

Table 1: $I(d)$'s for Poisson Simulation (section 3)

$I(d)$	
$d_1 = Y_{ij} - mp(\alpha_i, \beta_i)$	3.378206
$d_{2\|1} = Y_{ij} - mp(\mu_\alpha, \mu_\beta)$	41.162387
$d_2 = \alpha_i - \mu_\alpha$	1.628587
$d_2 = \beta_i - \mu_\beta$	1.299362
Correlations	
$Corr(Y_{11} - mp(\alpha_1, \beta_1), \alpha_1 - \mu_\alpha)$	-0.043967
$Corr(Y_{11} - mp(\alpha_1, \beta_1), \beta_1 - \mu_\beta)$	0.130122
$Corr(Y_{11} - mp(\alpha_1, \beta_1), \alpha_2 - \mu_\alpha)$	0.005247
$Corr(Y_{11} - mp(\alpha_1, \beta_1), \beta_2 - \mu_\beta)$	-0.032818

Table 2: Checkability of the binomial regression model

	DGSV	DGSV	GMS		
popu-	$d_1 = Y_{ij} - mp(\alpha_i, \beta_i)$	$d_{2\vert 1} = Y_{ij} - mp(\mu_\alpha, \mu_\beta)$	$d_1 \Leftrightarrow d_{2\vert 1} \Leftrightarrow d_o$	$d_2 = \alpha_i - \mu_\alpha$	$d_2 = \beta_i - \mu_\beta$
lation	(# failures out of 20)	(# failures out of 20)	(# failures out of 20)	# of failures	# of failures
1	1	0	1	0	0
2	4	0	1	0	0
3	0	2	1	1	0
4	5	4	2	0	0
5	2	0	2	0	0
6	2	0	5	0	0
7	3	0	1	0	0
8	3	0	2	0	0
9	2	7	2	0	0
10	2	0	0	0	0
11	0	0	1	0	0
12	3	1	4	0	0
13	5	0	0	0	0
14	5	0	0	0	0
15	2	0	1	0	0
16	4	0	0	0	0
17	3	5	0	0	0
18	3	1	1	0	0
19	1	0	0	0	0
20	1	0	1	0	0
Total	51	20	25	1	0
Expected Total	20	20	20	1	1

Table 3.1: Case 1 - First stage violation, no second stage violation

	DGSV	DGSV	GMS		
popu-	$d_1 = Y_{ij} - mp(\alpha_i, \beta_i)$	$d_{2\|1} = Y_{ij} - mp(\mu_\alpha, \mu_\beta)$	$d_1 \Leftrightarrow d_{2\|1} \Leftrightarrow d_o$	$d_2 = \alpha_i - \mu_\alpha$	$d_2 = \beta_i - \mu_\beta$
lation	(# failures out of 20)	(# failures out of 20)	(# failures out of 20)	# of failures	# of failures
1	2	0	0	0	1
2	9	1	3	0	0
3	4	2	1	0	0
4	5	4	0	0	0
5	6	0	5	0	0
6	1	2	4	0	0
7	7	0	1	0	0
8	3	0	0	0	0
9	2	4	4	0	0
10	1	0	6	0	0
11	7	3	1	0	0
12	1	4	3	0	0
13	3	0	1	0	0
14	2	0	2	0	0
15	2	1	2	0	0
16	2	0	2	0	0
17	4	3	2	0	0
18	0	0	5	0	0
19	4	0	0	0	0
20	0	1	0	1	0
Total	65	25	42	1	1
Expected Total	20	20	20	1	1

Table 3.2: Case 2 - First stage violation, no second stage violation

	DGSV	DGSV	GMS		
popu-	$d_1 = Y_{ij} - mp(\alpha_i, \beta_i)$	$d_{2\|1} = Y_{ij} - mp(\mu_\alpha, \mu_\beta)$	$d_1 \Leftrightarrow d_{2\|1} \Leftrightarrow d_o$	$d_2 = \alpha_i - \mu_\alpha$	$d_2 = \beta_i - \mu_\beta$
lation	(# failures out of 20)	(# failures out of 20)	(# failures out of 20)	# of failures	# of failures
1	1	6	0	0	0
2	2	2	1	0	0
3	0	0	6	0	0
4	1	0	0	0	0
5	1	5	3	0	0
6	1	0	0	0	0
7	3	0	0	0	0
8	3	1	2	0	0
9	2	2	1	0	0
10	3	0	1	0	0
11	1	6	1	0	1
12	1	5	0	0	1
13	1	5	3	0	1
14	0	3	3	0	1
15	3	5	0	0	1
16	1	4	1	0	1
17	0	5	1	0	1
18	2	5	0	0	1
19	4	6	4	0	0
20	2	4	1	0	1
Total	32	64	28	0	9
Expected Total	20	20	20	1	1

Table 3.3: Case 3 - No first stage violation, second stage violation