December 28, 2003

MCMC Strategies for Computing Bayesian Predictive Densities for Censored Multivariate Data

J.R. Lockwood and Mark J. Schervish

Traditional criteria for comparing alternative Bayesian hierarchical models, such as cross validation sums of squares, are inappropriate for non-standard data structures. More flexible cross validation criteria such as predictive densities facilitate effective evaluations across a broader range of data structures, but do so at the expense of introducing computational challenges. This paper considers Markov Chain Monte Carlo strategies for calculating Bayesian predictive densities for vector measurements subject to differential component-wise censoring. It discusses computational obstacles in Bayesian computations resulting from both the multivariate and incomplete nature of the data, and suggests two Monte Carlo approaches for implementing predictive density calculations. It illustrates the value of the proposed methods in the context of comparing alternative models for joint distributions of contaminant concentration measurements.

KEY WORDS: cross validation, data augmentation, predictive density, marginal density, nested integration

AUTHOR FOOTNOTE: J.R. Lockwood is Associate Statistician, RAND Statistics Group, Pittsburgh, PA 15213 (email: `lockwood@rand.org`). Mark J. Schervish is Professor, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213 (email: `mark@stat.cmu.edu`).

ACKNOWLEDGMENTS: This research was supported by US EPA award No. R826890-01-0, the Cooperative Agreement CR825188-01-3 between US EPA Office of Policy, Planning and Evaluation (OPPE) and Carnegie Mellon University. Support for J.R. Lockwood was also provided by RAND. This paper has undergone neither US EPA nor RAND peer review, and no official endorsement should be inferred.

1. INTRODUCTION

Model comparison and selection are among the most common problems of statistical practice, with numerous procedures for choosing among a set of models proposed in the literature (see, for example, Kadane and Lazar (2001) and Rao and Wu (2001) for recent reviews). Advances in computational technologies provide practitioners the opportunity to use increasingly rich models, such as Bayesian hierarchical models, to explore structures in complicated data. However, comparing and selecting among complex models challenges standard procedures. Formal Bayesian methods for model selection or model averaging via posterior model probabilities (Kass and Raftery 1995; Hoeting, Madigan, Raftery, and Volinsky 1999; Han and Carlin 2001) present notorious computational difficulties that are exacerbated by richly parameterized models. As noted by Han and Carlin (2001), both direct approximations to marginal densities and model space searches require significant effort to implement, and in many cases may remain indeterminate.

Moreover, richly parameterized hierarchical models may also confound less computationally intensive model selection criteria such as AIC (Akaike 1973), BIC (Schwarz 1978; Kass and Raftery 1995) and CIC (Tibshirani and Knight 1999). These criteria penalize model fit by model complexity to favor parsimonious models in an effort to maintain predictive power. Such criteria may be misleading when applied to richly parameterized Bayesian models, because the terms that quantify model complexity are based on asymptotic approximations that replace the realized model dimension with the nominal dimension. When prior information restricts the ability of parameters to vary independently, the effective number of parameters fit by a model may be considerably less than the nominal number, leaving the appropriate measure of complexity ambiguous. Recent efforts have made progress in developing useful definitions of model complexity in these cases (Poskitt 1987; Ye 1998; Hansen and Yu 2001; Hodges and Sargent 2001; Spiegelhalter, Best, Carlin, and van der Linde 2002), but no consensus yet exists.

These factors make cross validation (Mosteller and Tukey 1977) and other informal predictive evaluations (Laud and Ibrahim 1995; Han and Carlin 2001) attractive alternatives. Unfortunately, traditional cross validation criteria – most notably sums of squared deviations from observed to predicted values – are not appropriate for complicated data structures. An example of such data is multivariate contaminant concentration data subject to censoring, for which the incompleteness of the data make sums of squares impossible to evaluate. An alternative, more flexible cross-validation criterion is the predictive density (Gelfand and Dey 1994; Alqallaf and Gustafson 2001). Cross-validation predictive densities compromise between formal Bayes factor calculations and less formal criteria not applicable to complicated data structures.

This paper discusses Markov Chain Monte Carlo (MCMC) methods for efficiently calculating predictive densities for multivariate data when data are subject to differential component-wise censoring. It reviews the role of censored data in sampling posterior distributions, and shows how extensions of standard methods for achieving such samples do not provide effective means of calculating predictive densities. It then presents two Monte Carlo methods for implementing the calculations. It focuses on multivariate Gaussian data in an arbitrary parametric model structure because of the importance of this widely applicable case, but the principles of the methods are extensible to more general structures. Moreover, because the predictive densities we consider have the marginal densities required for Bayes factors as a special case, the methods may have value for structuring Bayes factor calculations. The paper does not discuss the many important practical issues surrounding the use of cross-validation in model comparison, such as the appropriate number and size of data splits, but rather focuses on computational techniques that benefit all cross-validation schemes with censored multivariate data.

The remainder of the paper is organized as follows. Section 2 presents the Bayesian parametric structure for censored data in which the methods are derived. Section 3 reviews the use of predictive density as a cross validation criterion, and considers the role of censored data in the calculation of predictive densities. Section 4 presents two Monte Carlo approaches to calculating the predictive densities that address the computational obstacles presented by simpler approaches. Section 5 presents applications of the methods to a model comparison problem involving joint distributions of contaminant concentrations, and Section 6 summarizes and discusses possible extensions and directions for future work.

2. PARAMETRIC STRUCTURE FOR CENSORED DATA

Suppose that conditional on a value $\boldsymbol{\theta} \in \Omega$ of a vector parameter $\boldsymbol{\Theta}$, $\boldsymbol{Z}_i \sim N_k(\boldsymbol{\mu}_i(\boldsymbol{\theta}, \boldsymbol{x}_i), \boldsymbol{\Sigma}_i(\boldsymbol{\theta}, \boldsymbol{x}_i))$ independently for $i = 1, \ldots, n$, where $\boldsymbol{\mu}_i(\boldsymbol{\theta}, \boldsymbol{x}_i)$ is a k-dimensional mean vector and $\boldsymbol{\Sigma}_i(\boldsymbol{\theta}, \boldsymbol{x}_i)$ a $(k \times k)$ positive definite covariance matrix. The vectors $\boldsymbol{x}_i$ represent known covariate information and for the remainder of the paper all probability statements are implicitly conditional on the these covariates. The parameter $\boldsymbol{\Theta}$ is arbitrary, and we assume only that we are working in the Bayesian framework where $\boldsymbol{\Theta}$ has some (possibly hierarchical) probability model. This general structure encompasses a broad class of models, including Bayesian multivariate regression, Bayesian MANOVA, and more complex hierarchical, spatial or longitudinal models with multivariate responses.

The focus of the current study is where $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{ik})$ rather than $\boldsymbol{Z}_i$ is observed, where for $j = 1, \ldots, k$, either $Y_{ij} = Z_{ij}$ or $Y_{ij} = (c_{ij\ell}, c_{iju})$ (mnemonic devices for "lower" and "upper" endpoints). That is, for each component of $\boldsymbol{Z}_i$, we learn either the actual value Z_{ij}, or only the partial information that $Z_{ij} \in (c_{ij\ell}, c_{iju})$. The motivating example for this structure is multivariate concentration data for constituents in water, which are subject to censoring of sufficiently small concentrations because of limitations of measurement techniques. For generality we allow $c_{ij\ell} = -\infty$ and/or $c_{iju} = +\infty$ to cover cases where components are left censored ($c_{ij\ell} = -\infty$), right censored, ($c_{iju} = +\infty$), or completely unobserved ($c_{ij\ell} = -\infty$ and $c_{iju} = +\infty$). The double subscripting by i and j makes explicit the fact that the censoring patterns may vary across observation vectors i and components within observation vectors j. For notational convenience we partition the observation vectors $\boldsymbol{Y}_i$ as $(\boldsymbol{Y}_{io}, \boldsymbol{Y}_{ic})$ where $\boldsymbol{Y}_{io}$ are the observed coordinates and $\boldsymbol{Y}_{ic}$ are the censored coordinates. We let $d_i = \dim(\boldsymbol{Y}_{ic})$, $0 \leq d_i \leq k$, denote the number of censored coordinates. We analogously partition the "complete data" $\boldsymbol{Z}_i$ as $(\boldsymbol{Y}_{io}, \boldsymbol{Z}_{ic})$. We denote the region in which the unobserved $\boldsymbol{Z}_{ic}$ can take values by C_i, which by previous assumptions is a Cartesian product of the intervals $(c_{ij\ell}, c_{iju})$ over the censored coordinates. Finally, we follow common notational convention by denoting realizations of random vectors by bold lowercase, so that the observed data are denoted by $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$.

Because of the censoring, the distribution of $\boldsymbol{Y}_i$ is not absolutely continuous with respect to k-dimensional Lebesgue measure. Rather, the dominating measure is a mixture of Lebesgue measure on R^{k-d_i} and counting measure on the elements of the Cartesian product defining C_i. This complicates the likelihood function and thus poses the challenges in Bayesian computations that are the focus of this paper. Using p generically to denote a density function (precisely which density function is indicated by its arguments) and using ϕ to denote marginal, conditional or joint Gaussian densities conditional on $\boldsymbol{\Theta} = \boldsymbol{\theta}$, the density of observation i with respect to the dominating measure is of the form:

$$\begin{aligned} p(\boldsymbol{y}_i|\boldsymbol{\theta}) &= \int_{C_i} \phi(\boldsymbol{y}_{io}, \boldsymbol{z}_{ic}|\boldsymbol{\theta}) d\boldsymbol{z}_{ic} \\ &= \phi(\boldsymbol{y}_{io}|\boldsymbol{\theta}) \int_{C_i} \phi(\boldsymbol{z}_{ic}|\boldsymbol{y}_{io}, \boldsymbol{\theta}) d\boldsymbol{z}_{ic} \end{aligned}$$

This is a k-dimensional multivariate normal rectangle probability if all coordinates are censored, the usual k-dimensional normal density if all coordinates are observed, and the product of a $k - d_i$ dimensional multivariate normal density function (the marginal density of the observed coordinates) and a d_i-dimensional multivariate normal rectangle probability (the conditional probability of the censored coordinates being censored given the observed coordinates) if d_i coordinates are censored. By the assumed conditional independence of the observations, the joint conditional density of the observations is

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{\theta}) &= \prod_{i=1}^{n} \left(\phi(\boldsymbol{y}_{io}|\boldsymbol{\theta}) \int_{C_i} \phi(\boldsymbol{z}_{ic}|\boldsymbol{y}_{io}, \boldsymbol{\theta}) d\boldsymbol{z}_{ic} \right) \\ &= \prod_{i=1}^{n} \left(\phi(\boldsymbol{y}_{io}|\boldsymbol{\theta}) \Pr(\boldsymbol{Z}_{ic} \in C_i | \boldsymbol{y}_{io}, \boldsymbol{\theta}) \right) \qquad (1) \end{aligned}$$

The fundamental assumption of the current study is that it is either computationally infeasible or otherwise undesirable to perform exact calculations of the form given in Equation (1) within the MCMC setting. This is a reasonable assumption because the evaluation involves numerical integrations that are likely to be prohibitively expensive to carry out in realistic problems. Such "nested integration" also arises in constrained parameter applications, as discussed by Chen and Shao (1998). In the current application, if one complete cycle through the parameters (i.e. one MCMC step) requires h evaluations of the joint likelihood function, then obtaining M posterior samples requires $O(Mhn)$ numerical evaluations of multivariate normal rectangle probabilities. These probability calculations are notoriously challenging, particularly in high dimensions, and have received intense study. A battery of both analytical and stochastic methods for has been developed, including functional expansions of the integral (Dutt 1973), re-expression following by numerical integration (Schervish 1984), and numerous Monte Carlo integration procedures (Geweke 1989; Genz 1992; Keane 1994; Breslaw 1994; Hajivassiliou, McFadden, and Ruud 1996; Vijverberg 1997). The articles by Hajivassiliou *et al.* (1996) and Gassmann *et al.* (2002) provide excellent overviews of the available methods.

3. THE ROLE OF CENSORING IN BAYESIAN CROSS VALIDATION CALCULATIONS

This section first reviews cross validation and predictive density, and then discusses its implementation in the context of censored data. Cross validation begins by partitioning $\boldsymbol{y}$ into n_0 vectors of *training data* $\boldsymbol{y}^{(0)}$ and n_1 vectors of *testing data* $\boldsymbol{y}^{(1)}$, with $n_0 + n_1 = n$. In some cases, commonly called "leave one out" cross validation (Stone 1974; Stone 1977; Mosteller and Tukey 1977; Geisser and Eddy 1979; Gelfand, Dey, and Chang 1992), $n_1 = 1$. In other cases, the training data consist of approximately half of the data (Mosteller and Tukey 1977) or other unbalanced allocations (Alqallaf and Gustafson 2001). Then, each model of interest is fit to $\boldsymbol{y}^{(0)}$ and compared in some manner to $\boldsymbol{y}^{(1)}$ to assess the degree of agreement between the testing data and the inferences concerning those data that are made by the model. Most often the quantification of agreement is a scalar function of $h(\boldsymbol{y}) = E(\boldsymbol{Y}^{(1)}|\boldsymbol{y}^{(0)}) - \boldsymbol{y}^{(1)}$, such as the sum of squared deviations of predicted values from the observed testing values. Within a set of models being compared, models with smaller values of $h(\boldsymbol{y})$ are preferred, and $h(\boldsymbol{y})$ for the best model might be compared to an external standard depending on the substantive problem and the intended uses of the model inferences.

While such summaries can be a natural quantification of agreement with auspicious properties in some settings, they are not directly applicable to the censored data structures presented in Section 2 because the incomplete nature of the data make $h(\boldsymbol{y})$ uninterpretable. Although one could modify the criterion to treat the observed and censored observations separately, this is *ad hoc* and cumbersome when the patterns of censoring and the censoring points vary across vectors. The posterior predictive density of the testing data given the training data (see e.g. Schervish (1995)) is a viable alternative with a number of advantages over functions of $h(\boldsymbol{y})$. Under the assumption introduced in Section 2 that all components of $\boldsymbol{Y}$ are conditionally independent given their associated covariates and $\boldsymbol{\Theta}$, the density is given by

$$p(\boldsymbol{y}^{(1)}|\boldsymbol{y}^{(0)}) = \int_{\boldsymbol{\Omega}} \left(\prod_{i=1}^{n_1} p(\boldsymbol{y}_i^{(1)}|\boldsymbol{\theta}) \right) p(\boldsymbol{\theta}|\boldsymbol{y}^{(0)}) d\boldsymbol{\theta}. \tag{2}$$

The density function $p(\boldsymbol{y}^{(1)}|\boldsymbol{y}^{(0)})$ is defined over the space in which $\boldsymbol{Y}^{(1)}$ takes values. The term "density" is used in its more general sense with $p(\boldsymbol{y}^{(1)}|\boldsymbol{y}^{(0)})$ denoting the Radon-Nikodym derivative of the probability measure describing the distribution of $\boldsymbol{Y}^{(1)}$ given $\boldsymbol{y}^{(0)}$ with respect to some dominating measure ν. While ν is commonly (product) Lebesgue measure, as noted previously, the measures required for censored observations are more complicated.

Like $h(\boldsymbol{y})$, the predictive density can be used to quantify agreement between the predictions made by the model and the observed testing data, and allows comparisons of models with different parametric structures because the criterion integrates over the model parameters. More importantly, the predictive density is well-defined for even complex data structures, and is closely related to the Bayes factor. The ratio of predictive densities of the testing data for two different models is known as a "partial" Bayes factor (O'Hagan 1995) because it is equivalent to the Bayes factor assuming that the training data were used to form the prior. While partial Bayes factors are primarily motivated by applications with improper prior distributions, their generally more stable estimability makes them a pragmatic choice even with proper priors.

The most straightforward method of evaluating the predictive density in Equation (2) (outside of simple cases where the integral can be evaluated analytically) is to obtain a sample $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M$ from $p(\boldsymbol{\theta}|\boldsymbol{y}^{(0)})$ by MCMC methods and to approximate the predictive density by

$$p(\boldsymbol{y}^{(1)}|\boldsymbol{y}^{(0)}) \approx \frac{1}{M} \sum_{m=1}^{M} \left(\prod_{i=1}^{n_1} p(\boldsymbol{y}_i^{(1)}|\boldsymbol{\theta}_m) \right). \tag{3}$$

Additional computational stability is achieved by estimating $\log p(\boldsymbol{y}^{(1)}|\boldsymbol{y}^{(0)})$ via the values

$$v(\boldsymbol{\theta}_m) = \sum_{i=1}^{n_1} \log p(\boldsymbol{y}_i^{(1)}|\boldsymbol{\theta}_m) \tag{4}$$

and using as the final estimate

$$\max_m v(\boldsymbol{\theta}_m) + \log \left(\sum_{m=1}^{M} e^{v(\boldsymbol{\theta}_m) - \max_k v(\boldsymbol{\theta}_k)} \right) - \log M.$$

This form of the calculation is more numerically stable because it applies the exponential function to numbers that are location shifted relative to the maximum and thus less likely to result in numerical values of zero after exponentiation.

However, censoring presents computational challenges to this straightforward algorithm. Direct implementation of the estimate in Equation (3) requires evaluating conditional densities of the form given in Equation (1) both for obtaining the sample $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M$ from $p(\boldsymbol{\theta}|\boldsymbol{y}^{(0)})$ and for evaluating the conditional predictive density $p(\boldsymbol{y}^{(1)}|\boldsymbol{\theta}_m) = \prod_{i=1}^{n_1} p(\boldsymbol{y}_i^{(1)}|\boldsymbol{\theta}_m)$ for each $\boldsymbol{\theta}_m$. As noted in Section 2, we are assuming that it is not feasible to evaluate such densities. It is reasonably straightforward to sidestep the evaluation of Equation (1) while obtaining a sample from $p(\boldsymbol{\theta}|\boldsymbol{y}^{(0)})$ using data augmentation (Tanner and Wong 1987; Dyk and Meng 2001). Letting $\boldsymbol{z}^{(0)}$ denote the unobserved portions of the training data $(\boldsymbol{z}_{1c}^{(0)}, \ldots, \boldsymbol{z}_{n_1c}^{(0)})$, the logic of data augmentation is to obtain a sample from $p(\boldsymbol{\theta}|\boldsymbol{y}^{(0)})$ by sampling $p(\boldsymbol{\theta}, \boldsymbol{z}^{(0)}|\boldsymbol{y}^{(0)})$ and discarding the $\boldsymbol{z}^{(0)}$. The sample is obtained by iteratively simulating from $p(\boldsymbol{\theta}|\boldsymbol{z}^{(0)}, \boldsymbol{y}^{(0)})$ and $p(\boldsymbol{z}^{(0)}|\boldsymbol{\theta}, \boldsymbol{y}^{(0)})$ which by standard MCMC results converge in distribution to samples from $p(\boldsymbol{\theta}, \boldsymbol{z}^{(0)}|\boldsymbol{y}^{(0)})$.

Sampling $p(\boldsymbol{\theta}|\boldsymbol{z}^{(0)}, \boldsymbol{y}^{(0)})$ proceeds exactly as MCMC would normally proceed had the data been fully observed, with the usual multivariate normal likelihood function rather than the censored likelihood in Equation (1). Sampling $p(\boldsymbol{z}^{(0)}|\boldsymbol{\theta}, \boldsymbol{y}^{(0)})$ is slightly more complicated. Because the data vectors are assumed to be conditionally independent given $\boldsymbol{\Theta}$, we focus on a single observation vector without loss of generality. The conditional distribution of $\boldsymbol{z}_{ic}^{(0)}$ given $\boldsymbol{y}_{io}^{(0)}$ and $\boldsymbol{\theta}$ is a multivariate normal distribution restricted to lie in the set C_i; i.e. a truncated multivariate normal distribution. Obtaining exact samples from this distribution in an efficient manner is not easy. A naive rejection sampler will be exceedingly inefficient when $\Pr(C_i|\boldsymbol{y}_{io}^{(0)}, \boldsymbol{\theta})$ is small, while more sophisticated methods such as importance sampling will require evaluating the multivariate normal probabilities that the data augmentation algorithm is trying to avoid in the first place. Fortunately, a straightforward Gibbs sampling algorithm using the full conditional distribution of each censored coordinate can be used. The conditional distribution of a single censored coordinate j of $\boldsymbol{z}_{ic}^{(0)}$ given $\boldsymbol{\theta}$, $\boldsymbol{y}_{io}^{(0)}$ and imputed values of all of the other censored coordinates is a truncated univariate normal with range of the form $(c_{ij\ell}, c_{iju})$. This distribution is easily sampled using the inverse CDF method using a restricted uniform random variable. Thus, replacing the step of sampling from $p(\boldsymbol{z}^{(0)}|\boldsymbol{\theta}, \boldsymbol{y}^{(0)})$ with Gibbs sampling steps for each censored coordinate of each observation vector, when combined with iteratively sampling from $p(\boldsymbol{\theta}|\boldsymbol{z}^{(0)}, \boldsymbol{y}^{(0)})$, results in samples that converge in distribution to $p(\boldsymbol{\theta}, \boldsymbol{z}^{(0)}|\boldsymbol{y}^{(0)})$.

The other required part of the predictive density calculation is the evaluation of $p(\boldsymbol{y}^{(1)}|\boldsymbol{\theta}_m) = \prod_{i=1}^{n_1} p(\boldsymbol{y}_i^{(1)}|\boldsymbol{\theta}_m)$ for each $\boldsymbol{\theta}_m$. Although the imputation of censored coordinates via data augmentation avoids calculations of the form in Equation (1) and thus is valuable in obtaining a sample from $p(\boldsymbol{\theta}|\boldsymbol{y}^{(0)})$, unfortunately such methods provide little help in calculating predictive densities for censored data.

To demonstrate why imputation of censored coordinates in the testing data is not useful, first consider sampling the censored observations from their conditional distributions given the observed coordinates, $\boldsymbol{\theta}$, and the fact that the censored observations are known to lie in the censoring regions C_i. We call these the *constrained* conditional distributions of the censored coordinates; these are the distributions that are used in the data augmentation algorithm for sampling $p(\boldsymbol{\theta}|\boldsymbol{y}^{(0)})$. As noted previously, sampling directly from these multivariate truncated normal distributions is difficult, and this difficulty is further exacerbated by the fact that we cannot use the Gibbs sampling method discussed previously unless we nest that algorithm within the existing MCMC framework. That is, we would need to carry out an auxiliary Gibbs sampler for each sampled value of $\boldsymbol{\theta}_m$. More problematically, plugging the imputed values of the censored coordinates into the joint conditional density of the complete testing data given $\boldsymbol{\theta}$ (as if the data had been fully observed) implies that marginally we will be estimating

$$\int_{\boldsymbol{\Omega}} \left(\prod_{i=1}^{n_1} \int_{C_i} \phi(\boldsymbol{y}_{io}^{(1)}, \boldsymbol{z}_{ic}^{(1)}|\boldsymbol{\theta}) \left[\frac{\phi(\boldsymbol{z}_{ic}^{(1)}|\boldsymbol{y}_{io}^{(1)}, \boldsymbol{\theta})}{\Pr(\boldsymbol{Z}_{ic}^{(1)} \in C_i|\boldsymbol{y}_{io}^{(1)}, \boldsymbol{\theta})} \right] d\boldsymbol{z}_{ic}^{(1)} \right) p(\boldsymbol{\theta}|\boldsymbol{y}^{(0)}) d\boldsymbol{\theta}$$

or, equivalently,

$$\int_{\boldsymbol{\Omega}} \left(\prod_{i=1}^{n_1} \phi(\boldsymbol{y}_{io}^{(1)}|\boldsymbol{\theta}) \left[\frac{\int_{C_i} \phi^2(\boldsymbol{z}_{ic}^{(1)}|\boldsymbol{y}_{io}^{(1)}, \boldsymbol{\theta}) d\boldsymbol{z}_{ic}^{(1)}}{\Pr(\boldsymbol{Z}_{ic}^{(1)} \in C_i|\boldsymbol{y}_{io}^{(1)}, \boldsymbol{\theta})} \right] \right) p(\boldsymbol{\theta}|\boldsymbol{y}^{(0)}) d\boldsymbol{\theta}.$$

Because the term in square brackets does not equal $\Pr(\boldsymbol{Z}_{ic}^{(1)} \in C_i|\boldsymbol{y}_{io}^{(1)}, \boldsymbol{\theta})$ in general, this strategy does not perform the correct calculation.

A related strategy samples the censored coordinates from their *unconstrained* conditional distributions; that is, their conditional distributions given the observed coordinates and $\boldsymbol{\theta}$ but not the fact that the censored observations are known to lie in the censoring regions C_i. These can be used to calculate a simulation-consistent estimate of the predictive density using a "brute force" Monte Carlo algorithm, but this algorithm suffers from unacceptably large Monte Carlo variance. The approach is, for each observation vector, to sample from the unconstrained conditional distributions of the censored coordinates. If all coordinates happen to fall in C_i (i.e., if they happen to be valid plausible censored values), set $p(\boldsymbol{y}_i^{(1)}|\boldsymbol{\theta}) = \phi(\boldsymbol{y}_{io}^{(1)}|\boldsymbol{\theta})$; otherwise set $p(\boldsymbol{y}_i^{(1)}|\boldsymbol{\theta}) = 0$. Under this algorithm, the only way a non-zero value of $p(\boldsymbol{y}^{(1)}|\boldsymbol{\theta})$ is achieved for a particular iteration is when all simulated missing data from every observation vector, as sampled from their unconstrained distributions, happen to fall in their respective censoring regions. Such a strategy, while having the correct expected value (see Section 4.1), will suffer from extreme Monte Carlo variance. A similar strategy, in which we write

$$p(\boldsymbol{y}_i^{(1)}|\boldsymbol{\theta}) = \int 1\{\boldsymbol{z}_{ic}^{(1)} \in C_i\} \phi(\boldsymbol{y}_{io}^{(1)}|\boldsymbol{z}_{ic}^{(1)}, \boldsymbol{\theta}) \phi(\boldsymbol{z}_{ic}^{(1)}|\boldsymbol{\theta}) d\boldsymbol{z}_{ic}^{(1)}$$

and sample from the unconstrained marginal distribution of the censored coordinates, has the same shortcoming.

4. TWO MONTE CARLO APPROACHES

We provide two strategies to reduce the Monte Carlo variance in the calculation of predictive densities for multivariate censored data. The first is a refined version of the brute force method described in the last section, using a more efficient Monte Carlo estimator for the conditional density in Equation (1). It is most useful when the effective parameter space is small enough to permit stable Monte Carlo integration over the entire space. When this fails, the second method that conditions sequentially on components of the data (thus capitalizing on the strengths of data augmentation) can be used for more stable estimation.

4.1: (Method 1) Direct Predictive Density Calculation

The brute force method of Section 3 is the most naive implementation of a more general class of methods that, rather than evaluating $p(\boldsymbol{y}_i^{(1)}|\boldsymbol{\theta}) = \phi(\boldsymbol{y}_{io}^{(1)}|\boldsymbol{\theta})\Pr(\boldsymbol{Z}_{ic}^{(1)} \in C_i|\boldsymbol{y}_{io}^{(1)}, \boldsymbol{\theta})$ exactly, use Monte Carlo estimates of the multivariate normal rectangle probabilities. Such methods introduce a function $h_i(\boldsymbol{W}_i; \boldsymbol{y}_i^{(1)}, \boldsymbol{\theta})$ of an auxiliary random variable $\boldsymbol{W}_i$ such that $p(\boldsymbol{W}_i|\boldsymbol{y}_i^{(1)}, \boldsymbol{\theta})$ is easy to sample, $h_i(\boldsymbol{W}_i; \boldsymbol{y}_i^{(1)}, \boldsymbol{\theta})$ is easy to calculate, and $E\left(h_i(\boldsymbol{W}_i; \boldsymbol{y}_i^{(1)}, \boldsymbol{\theta})|\boldsymbol{y}_i^{(1)}, \boldsymbol{\theta}\right) = \Pr(\boldsymbol{Z}_{ic}^{(1)} \in C_i|\boldsymbol{y}_{io}^{(1)}, \boldsymbol{\theta})$. Thus $\phi(\boldsymbol{y}_{io}^{(1)}|\boldsymbol{\theta})h_i(\boldsymbol{W}_i; \boldsymbol{y}_i^{(1)}, \boldsymbol{\theta})$ is unbiased for $p(\boldsymbol{y}_i^{(1)}|\boldsymbol{\theta})$, and because the random variables $h_i(\boldsymbol{W}_i; \boldsymbol{y}_i^{(1)}, \boldsymbol{\theta})$, $i = 1, \ldots, n_1$, are conditionally independent given $\boldsymbol{y}^{(1)}$ and $\boldsymbol{\theta}$,

$$E\left(\prod_{i=1}^{n_1}[\phi(\boldsymbol{y}_{io}^{(1)}|\boldsymbol{\theta})h_i(\boldsymbol{W}_i; \boldsymbol{y}_i^{(1)}, \boldsymbol{\theta})]|\boldsymbol{y}_i^{(1)}, \boldsymbol{\theta}\right) = \prod_{i=1}^{n_1} p(\boldsymbol{y}_i^{(1)}|\boldsymbol{\theta}). \tag{5}$$

The law of iterated expectations implies that the marginal mean of these random variables is the desired predictive density in Equation (2)

The brute force method takes $p(\boldsymbol{W}_i|\boldsymbol{y}_i^{(1)}, \boldsymbol{\theta})$ to be the unconstrained conditional distribution of the censored coordinates given the observed coordinates and $\boldsymbol{\theta}$, and sets $h_i(\boldsymbol{W}_i; \boldsymbol{y}_i^{(1)}, \boldsymbol{\theta}) = 1\{\boldsymbol{W}_i \in C_i\}$. An obvious improvement is to use more efficient, unbiased estimates of the multivariate rectangle probabilities that also can be obtained without extensive computing effort. While many Monte Carlo methods are available (Gassmann, Deak, and Szantai 2002; Hajivassiliou, McFadden, and Ruud 1996), the method that we propose is the "Geweke-Hajivassiliou-Keane (GHK)" Monte Carlo simulator for multivariate normal rectangle probabilities (Hajivassiliou, McFadden, and Ruud 1996). This method was derived independently by Geweke (1989) and Keane (1994), and was improved by Hajivassiliou (1994). Both the study by Hajivassiliou *et al.* (1996) and another study by Geweke *et al.* (1997) indicate that this simulator offers an effective balance of accuracy and computational speed. Details on the method (including the f and h functions) are provided in the Appendix.

The primary benefit of the proposed method is that it maintains the relative simplicity and computational efficiency of the brute force method, while providing a more statistically efficient estimator because $0 < h_i(\boldsymbol{W}_i; \boldsymbol{y}_i^{(1)}, \boldsymbol{\theta}) < 1$. Additional statistical efficiency can be achieved by setting $h_i(\boldsymbol{W}_i; \boldsymbol{y}_i^{(1)}, \boldsymbol{\theta})$ to the average of the function over K realizations from the GHK simulator. Because the GHK simulator is consistent, $K \rightarrow \infty$ provides an exact evaluation of the multivariate normal probabilities and thus of the integrand in Equation (2). As noted previously, such exact evaluations will be infeasible in many realistic problems. Moreover, we performed some empirical investigations with various values of $K > 1$ in the context of the example presented in Section 5 and found the reduction in the Monte Carlo standard error of the predictive density estimate was modest relative to the additional computational burden. This is because many, and in some cases most, sampled values $\boldsymbol{\theta}_m$ contribute little to the overall estimate, and thus obtaining improved precision of integrals conditional on these parameters does not provide much additional stability to the overall calculation.

4.2: (Method 2) Sequential Predictive Density Calculation

The direct predictive density calculation is carried out via a single Monte Carlo integral over the entire parameter space. However, in some cases the predictive density estimates are subject to considerable Monte Carlo variability, even in long chains. The regions of the parameter space that result in high conditional densities for the testing data can have low probabilities under the posterior distribution based on the training data. Thus, these regions may be visited only rarely in a typical MCMC analysis, similar to the situation faced when trying to calculate the marginal density of the data (e.g., for Bayes factors) using a Monte Carlo sample from the prior distribution. (This is why expending computing resources to obtain precise estimates of the conditional predictive density $p(\boldsymbol{y}^{(1)}|\boldsymbol{\theta}_m)$ for every $\boldsymbol{\theta}_m$ in the direct calculation is not very effective.) Part of the reason for using cross validation rather than Bayes factors is that this should be less likely to happen; nevertheless, it may still manifest, especially in richly parameterized models.

Moreover, the problem is exacerbated as the dimension k of the data vectors increases. Heuristically, in order to assign a high predictive density to a given data vector, a sampled parameter $\boldsymbol{\theta}$ must be consistent with each of the p coordinates. If M iterations are required to calculate the predictive density within a given level of Monte Carlo variability when $k = 1$, then $O(M^k)$ iterations may be necessary to achieve the same order of control of Monte Carlo error when the dimension is $k > 1$. This is especially true for cases in which the dimension of the parameter space increases with k, which is extremely common in parametric models that include additional parameters for means, variances and correlations with each new coordinate. The additional size of the parameter space may make the direct predictive density calculations subject to unacceptably large Monte Carlo variability; examples of this behavior are presented in Section 5. This section presents an alternative method for estimating the predictive density that can help substantially to reduce the Monte Carlo variability. Although motivated by predictive evaluations for censored observations, the concepts and methods are applicable to even fully observed multidimensional data structures.

Let $\boldsymbol{y}_{\cdot j}^{(1)} = (y_{1j}^{(1)}, \ldots, y_{n_1 j}^{(1)})$ denote the vector of n_1 observations of coordinate j from the testing data, and let $\boldsymbol{y}_{\cdot <j}^{(1)}$ denote the collection of coordinates $1, \ldots, j-1$ from these observations, defined to

be empty for $j = 1$. Then, the predictive density for the testing data given the training data can be expressed as a product of conditional predictive densities:

$$p(\boldsymbol{y}^{(1)}|\boldsymbol{y}^{(0)}) = \prod_{j=1}^{k} p(\boldsymbol{y}_{\cdot j}^{(1)}|\boldsymbol{y}_{\cdot <j}^{(1)}, \boldsymbol{y}^{(0)})$$

where, through conditional independence, each term on the RHS can be expressed as the following integral over the parameter space:

$$p(\boldsymbol{y}_{\cdot j}^{(1)}|\boldsymbol{y}_{\cdot <j}^{(1)}, \boldsymbol{y}^{(0)}) = \int \left(\prod_{i=1}^{n_1} p(y_{ij}^{(1)}|\boldsymbol{y}_{i<j}^{(1)}, \boldsymbol{\theta}) \right) p(\boldsymbol{\theta}|\boldsymbol{y}_{\cdot <j}^{(1)}, \boldsymbol{y}^{(0)}) d\boldsymbol{\theta}.$$

That is, the desired predictive density can be calculated as a product of k separate integrals rather than one omnibus integral (or in practice, as a sum of the logarithms of k separate integrals as discussed in Section 3). Expressing it in this manner provides several advantages that should help to reduce Monte Carlo variability. Because each successive integral conditions on more of the testing data, the posterior distribution concentrates sequentially on regions of the parameter space which are likely to result in high conditional density for the next coordinates. Equally important is that the sequential calculation obviates the need for calculating multivariate normal rectangle probabilities, because each integral involves only products of univariate conditional densities. The sequential calculation is valid for any ordering of the coordinates, and more generally, is not restricted to the prediction of a single coordinate given all previous coordinates. Instead, the vectors could be partitioned into blocks of arbitrary sizes.

The primary disadvantage of the sequential calculation is that it requires the model to be fit k times. The first fits the model to only the training data and calculates the predictive density of the collection of first coordinates of the testing data. The remaining $k-1$ fits use the training data and j coordinates of the testing data, treating the remaining $k-j$ coordinates of the testing data as missing, for $j = 1, \ldots, k-1$. However, the inconvenience of the multiple model fits is minor in cases where direct calculations result in estimates too variable to be informative.

In addition, note that in general it will be necessary to condition on censored or missing coordinates in the sequential calculation. To account for the censoring, we again use data augmentation. Let $\boldsymbol{z}_{\cdot <j}^{(1)}$ denote the unobserved values of the censored coordinates in $\boldsymbol{y}_{\cdot <j}^{(1)}$, and note that

$$p(\boldsymbol{y}_{\cdot j}^{(1)}|\boldsymbol{y}_{\cdot <j}^{(1)}, \boldsymbol{y}^{(0)}) = \int\int \left(\prod_{i=1}^{n_1} p(y_{ij}^{(1)}|\boldsymbol{y}_{i<j}^{(1)}, \boldsymbol{\theta}, \boldsymbol{z}_{i<j}^{(1)}) \right) p(\boldsymbol{\theta}, \boldsymbol{z}_{\cdot <j}^{(1)}|\boldsymbol{y}_{\cdot <j}^{(1)}, \boldsymbol{y}^{(0)}) d\boldsymbol{\theta} d\boldsymbol{z}_{\cdot <j}^{(1)} \quad (6)$$

The terms $p(y_{ij}^{(1)}|\boldsymbol{y}_{i<j}^{(1)},\boldsymbol{\theta},\boldsymbol{z}_{i<j}^{(1)})$ are substantially easier to handle than $p(y_{ij}^{(1)}|\boldsymbol{y}_{i<j}^{(1)},\boldsymbol{\theta})$ because the former conditions on the complete data. In particular, $p(y_{ij}^{(1)}|\boldsymbol{y}_{i<j}^{(1)},\boldsymbol{\theta},\boldsymbol{z}_{i<j}^{(1)})$ is either a univariate normal density function (if $y_{ij}^{(1)}$ is observed) or a univariate normal integral (if $y_{ij}^{(1)}$ is censored), both of which can be calculated efficiently with standard routines. We use data augmentation to sample from $p(\boldsymbol{\theta},\boldsymbol{z}_{\cdot<j}^{(1)}|\boldsymbol{y}_{\cdot<j}^{(1)},\boldsymbol{y}^{(0)})$ in much the same way as we did to sample $p(\boldsymbol{\theta}|\boldsymbol{y}^{(0)})$. We use Gibbs sampling steps to impute each censored coordinate from its conditional distribution given the observed coordinates, the imputed values of all other censored coordinates, and $\boldsymbol{\theta}$. Then, conditional on the complete data, we sample $\boldsymbol{\theta}$ from $p(\boldsymbol{\theta}|\boldsymbol{z}_{\cdot<j}^{(1)},\boldsymbol{y}_{\cdot<j}^{(1)},\boldsymbol{y}^{(0)},\boldsymbol{z}^{(0)})$ exactly as if the data had been fully observed. The integral in Equation (6) can be approximated by

$$p(\boldsymbol{y}_{\cdot j}^{(1)}|\boldsymbol{y}_{\cdot<j}^{(1)},\boldsymbol{y}^{(0)}) \approx \frac{1}{M}\sum_{m=1}^{M} p(\boldsymbol{y}_{\cdot j}^{(1)}|\boldsymbol{y}_{\cdot<j}^{(1)},\boldsymbol{\theta}_m,\boldsymbol{z}_{\cdot<j,m}^{(1)})$$

where $(\theta_m,\boldsymbol{z}_{\cdot<j,m}^{(1)})$, $m=1,\ldots,M$ is a sample from the distribution $p(\theta,\boldsymbol{z}_{\cdot<j}^{(1)}|\boldsymbol{y}_{\cdot<j}^{(1)},\boldsymbol{y}^{(0)})$ (as before, practical applications should work on the logarithm scale as in Equation (4)).

5. EXAMPLE

This section presents an application of the methods to the comparison of models for joint distributions of contaminants in community water systems. Such models are used to predict raw (untreated) water contaminant profiles in all community water systems in the country, helping to quantify uncertainty during the regulatory decision process that sets maximum contaminant levels for drinking water supplies. The cross-validation approach to comparing models is particularly appropriate because predictive validity is central to the application. The details of the substantive problem, the data sources, and the class of models under consideration can be found in Lockwood *et al* (2003) and Gurian *et al* (2001), and are only briefly summarized here.

The data, from the National Arsenic Occurrence Survey (Frey and Edwards 1997), consist of raw water arsenic (As), manganese (Mn) and uranium (U) and concentrations from 482 of the approximately 55,000 U.S. community water systems. For each system we know the U.S. state and the source water type (ground water versus surface water), and the goal is to use these limited data to estimate joint raw water distributions of the three substances as a function of location and source water type. That is, we would like to estimate a distribution in each of the 100 cells defined by the 50 states and 2 source water types. This challenge is more formidable given that the observation vectors are subject to considerable censoring. The observations in the data set were measured with a standardized protocol that resulted in left censoring points of 0.5 $\mu g/L$ for each of the three substances. Less than 30% of the observation vectors have fully observed coordinates; Figure 1 presents scatterplots of the log contaminant concentrations

Estimating 100 multivariate distributions from 482 data vectors while maintaining predictive validity requires choosing model complexity that respects the inferential limitations of the available data. The two primary aspects of model complexity that we explored are the spatial resolution of the model and whether the models do or do not include parameters that model residual covariance among measurements. In particular we present comparisons of ten different models organized in a two by five design. The first factor is whether the contaminants are modeled with independent lognormal distributions within each cell, or with general multivariate lognormal distributions within each cell. The other factor compares a sequence of five increasingly rich spatial resolutions for the model as follows:

- 1 Region (no spatial differentiation)
- 2 Regions (Eastern and Western regions of the U.S)
- 7 Regions (defined by the NAOS surveyors (Frey and Edwards 1997))
- 50 States with a distance-based spatial correlation structure for the contaminant means and variances
- 50 States with no spatial correlation

In all cases we allow for different distributions in ground water and surface water. In addition, for all of the multivariate lognormal models that allow non-diagonal covariance matrices within cells, separate correlation matrices for ground water and surface water were estimated, but these correlation matrices were assumed common across locations. Full details of the model structure, prior distributions, estimation algorithms and diagnostics can be found in Lockwood *et al* (2003).

We use cross-validation predictive densities to help guide our choice about the most effective model structure given the resolution of the data. We randomly split the 482 observations in half into training and testing data sets, and used MCMC to fit each of the ten models under consideration to the training data. We then calculated the predictive density for the testing data using the methods discussed previously. For the five models that treated the contaminants independently within cells, we fit separate models of the appropriate spatial structure to each contaminant, with the predictive density for each contaminant estimated using the straightforward methods of Section 2. The overall predictive density for the testing data was then estimated as the product of the individual predictive densities for each contaminant. For the five models that explicitly respected the multivariate structure of the data, the predictive densities were estimated using the two different methods in Section 3.

We also compare our cross-validation predictive density results to those obtained by the Deviance Information Criterion (DIC) (Spiegelhalter, Best, Carlin, and van der Linde 2002), which generalizes AIC to richly parameterized Bayesian hierarchical models. DIC requires the posterior mean of the log likelihood function

$$\bar{L} = \int_{\boldsymbol{\Omega}} \left(\sum_{i=1}^{n} \log p(\boldsymbol{y}_i|\boldsymbol{\theta}) \right) p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta} \tag{7}$$

as well as the log likelihood function evaluated at the posterior mean of $\boldsymbol{\theta}$, denoted $L(\bar{\boldsymbol{\theta}})$. Apart from a normalizing constant that depends on only the observed data and thus does not affect model comparison, DIC is given by $-4\bar{L} + 2L(\bar{\boldsymbol{\theta}})$ and models with smaller values are favored. One of the strengths of DIC is that it is relatively easy to calculate. However, censored observations provide additional complication because the Monte Carlo estimates of the multivariate normal rectangle probabilities that were used to sidestep the exact calculation of the conditional densities are not unbiased estimates of the logs of the probabilities. By Jensen's Inequality, the log of an estimator such as that on the LHS of Equation (5) will be negatively biased for $\log p(\boldsymbol{y}_i|\boldsymbol{\theta})$. Thus it is necessary to perform exact (or nearly exact) calculations of the multivariate normal probabilities for each sampled parameter. As discussed previously, this is assumed to be computationally infeasible. Fortunately, because the integrand in Equation (7) is generally largest near the mode of the posterior distribution, the number of sampled parameters required to get stable estimates of DIC is lower than that for predictive densities. In the current application we estimated $\bar{L}$ using every 1000^{th} parameter vector from our MCMC samples. Repeated applications of this procedure indicated that this was sufficient to get a reasonably stable estimate of DIC for each model.

The results are summarized in Figure 2, which provides the estimated log predictive density (LPD) of the testing data under the 10 models. In the figure, "I" indicates the "Independence" models; "J Direct" and "J Sequential" refer to the estimates under the "Joint" models as calculated by the direct and sequential methods, respectively. The six different sequential estimates for each spatial complexity derive from the six possible orderings of the three contaminants. The LPD estimates are scaled relative to the lowest LPD among all models (the Independence model with no spatial differentiation). The DIC values for each of the ten models are provided in the right frame of the figure, again scaled relative to the least effective model (with the highest DIC). In all cases the estimates are based on 3 million parameter vectors obtained via MCMC from the appropriate posterior distribution given the training data. The boxplots represent variability of the estimates in consecutive blocks of 100,000 parameter vectors (Han and Carlin 2001). All models were run on a 1.5 GHz PC running Linux; computing times for each model ranged from approximately 600 minutes to 1500 minutes depending on the complexity of the model.

The two primary questions that underly the model comparison are 1) are the contaminants sufficiently highly correlated that explicitly modeling the multivariate structure is advantageous; and 2) what is a pragmatic degree of spatial differentiation for the contaminant distributions given the coarse resolution of the data? The general trends evident in the figure are that for the models with a simple spatial structure (one or two regions), there is a clear predictive advantage to modeling additional correlation among the contaminants via the multivariate lognormal, and that the models with the richer spatial structure offer a clear predictive advantage relative to the simpler spatial models.

However, there is a substantively critical interplay between these inferences and the method used to estimate the LPD for the joint models. The direct calculation method seems to suggest that conditional on a richer spatial resolution of the model, there is no additional benefit to modeling the residual correlations between the contaminants. In fact, it would appear that the fully multivariate models predict the testing data more poorly than the independence models when the state-based spatial structure is used. However, notice that there is a much higher degree of Monte Carlo variability in the LPD calculations for the spatially rich models. Extensive empirical investigation of the root of this variability revealed the parameter spaces for the richer spatial models are so large that Monte Carlo integrals that try to average over the whole space simultaneously are subject to an overwhelming degree of Monte Carlo variability. The regions of the parameter space which concurrently make effective predictions for all contaminants are so small that even three million iterations are not sufficient to visit them frequently (or at all), and thus the integral is estimated inaccurately. This is why the direct calculation would imply that it is more effective to model the contaminants independently when there is richer spatial structure.

This motivates the sequential calculation in which each successive model fit spends more time in the parts of the parameter space most consistent with the testing data because it learns portions of the testing data in turn. This can dramatically improve estimation precision, as shown in the figure. When the effective parameter space is reasonably small, the direct and sequential methods give the same results, as expected. However, when the model complexity grows, the estimates from the sequential method are shifted toward higher values than those from the direct methods because they concentrate more heavily on the relevant portions of the parameter space. These regions are never even visited by the direct calculation, explaining the almost complete lack of overlap of the boxplots for the two methods. The improved estimates imply different substantive conclusions as well, as the LPD for the joint model is higher than that for the independence model for all spatial resolutions. Thus even with the richer spatial structure, there is additional predictive power in modeling residual correlation between the contaminants. Most interestingly, the calculations provide insight into the appropriate degree of spatial complexity for the model. Very little spatial differentiation is suboptimal, but so is ignoring large-scale spatial trends in the contaminant distributions. Exploiting these spatial trends by using data from surrounding states to inform the distributions for a given state provides more efficient estimates, and ultimately, more effective predictions of external data.

The sequential estimates illustrate an interesting aspect of the role of missing data in the sequential calculation. The primary advantage that the direct calculation has over the sequential calculation is that all composite terms (e.g. all summands in Equation (4)) involve a conditional density of censored coordinates given observed coordinates. As discussed in Section 3, the sequential calculation must calculate the conditional density of an observed coordinate given imputed values of missing coordinates obtained via data augmentation. Although the expected value of the sequential estimate is invariant to this as well as the the order of conditioning, certain orders of conditioning may be more advantageous than others depending on the particular data and models under consideration. As is evident in the figure, three orders of conditioning result in higher LPD estimates than the other three orders for all but the model with no spatial differentiation. In-depth examination of this case revealed that there was an influential observation with a relatively large arsenic concentration, but for which the uranium coordinate was censored. The estimated contribution of that observation to the overall LPD was sensitive to whether the calculation used the conditional density of the censored coordinate given the observed one, or used the conditional density of the observed one given imputed values of the censored one. The sensitivity was exacerbated by the relatively strong within-cell correlation of these two contaminants. In this case it seems likely that the three orders of conditioning that result in higher LPD values are providing estimates that are closer to the truth because the distributions for the lower three orders have higher variances and generally have upper tails that support values that are consistent with the estimates provided by the higher orders. As a practical matter, it is thus advisable to try different orderings of conditioning for the coordinates, as these may reveal subtly influential observation vectors.

Finally, it is interesting to note that while the DIC and sequential predictive density criteria agree that both the independence models and the models that do not provide the full 50 state spatial differentiation are inferior, they disagree somewhat on which of the two 50 state models is preferred. The sequential predictive density calculation favors the 50 state model with spatial correlation, while DIC slightly favors the 50 state model without spatial correlation. Of course, this is based on only a single split of the data, but does raise a general question about the circumstances under which the two criteria might provide different results.

6. SUMMARY AND DISCUSSION

Cross validation is an effective means to compare complex Bayesian models in which formal evaluations may be more difficult. Multivariate censored data present additional challenges by confounding traditional cross validation criteria. This study focused on the posterior predictive density of the testing data given the training data as an alternative criterion, presenting two approaches for calculating predictive densities for censored data in MCMC applications. Which method is most appropriate may depend on the complexity of the model; the sequential approach helps to reduce Monte Carlo variability and can be used to make decisions where the direct methods are unclear or potentially misleading.

Both computational approaches apply to all predictive density calculations of the form $p(y_r|y_{-r})$ (e.g. the conditional predictive ordinates of Geisser and Eddy (1979) and others reviewed by Gelfand and Dey (1994)), not only those involving half splits of the data, and thus have direct applicability to formal Bayes factor calculations. In models with sufficiently small parameter spaces and/or sufficiently informative prior distributions, the methods may be applied exactly as described here with a null training data set (e.g. by direct integration with respect to the prior distribution). The methods also can be coupled straightforwardly with more advanced importance sampling or derivative methods for calculating marginal densities (Han and Carlin 2001) in more complex settings.

As noted, the sequential approach is useful even with fully observed data; it is similar in spirit to calculating the marginal densities one observation at a time through sequentially conditioning on observation vectors. Partitioning the data by components rather than by observation vectors (akin to Chib's (1995) method for calculating marginal densities from the Gibbs sampler) limits the number of integrals to k rather than n, and is particularly well-suited to models where the parameter space is approximately partitioned by coordinate.

A promising avenue for future research is the exploration of the trade-off between Monte Carlo errors for the different levels of nesting required for the density calculation. Our experience in this problem indicated that it is likely more valuable to spend time sampling additional parameter vectors rather than obtaining more precise Monte Carlo integral estimates for a given parameter vector. However, the value (in terms of reducing overall Monte Carlo variance of the final estimate) of precise estimates of the "inner" or conditional integral depends on the particular value of the parameter vector. It would be useful to this problem and the numerous problems like it to develop adaptive algorithms that would spend time obtaining precise estimates for only those values of the parameter vector that have high leverage in the overall estimate.

Finally, although the results are presented for censored multivariate Gaussian data, the principles carry over to non-Gaussian data, but may present difficulties if conditional densities cannot be evaluated in closed form. Further exploration of this issue, and empirical investigations of performance in higher dimensional Gaussian problems, are ongoing.

APPENDIX: GHK Simulator

Let $\boldsymbol{Z}$ be a d-dimensional random vector with a non-singular $N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution. (Note that in the context of the current study, this should be considered the conditional distribution of the d censored coordinates given the $k - d$ observed coordinates and the current value of $\boldsymbol{\theta}$ in the MCMC). The GHK simulator is a Monte Carlo method for estimating $\Pr(\boldsymbol{Z} \in C)$, where $C = \{(z_1, \ldots, z_d) : c_{j\ell} \leq z_j \leq c_{ju}, j = 1, \ldots, d\}$. We summarize the method as it is defined by Hajivassiliou *et al.* (1996). Allow the endpoints of the marginal intervals $c_{j\ell}$ and c_{ju} to take the values $\pm\infty$, with the convention that adding finite numbers to or multiplying finite non-zero numbers by $\pm\infty$ still results in $\pm\infty$. For a vector $(x_1, \ldots, x_d)$, let $\boldsymbol{x}_{<j}$ denote the row vector $(x_1, \ldots, x_{j-1})$, and for a matrix $\boldsymbol{X} = ((x_{ij}))$, let $\boldsymbol{X}_{j,<j}$ denote the row vector consisting of the first $j - 1$ elements of row j. In both cases, adopt the convention that the vector is null for $j = 1$. Let $\boldsymbol{L} = ((\ell_{ij}))$ be the lower triangular Cholesky factor of

the covariance matrix $\boldsymbol{\Sigma}$; i.e., $\boldsymbol{L}\boldsymbol{L'} = \boldsymbol{\Sigma}$. Also, let $\boldsymbol{U}$ be a d-dimensional random vector with a $N(0, \boldsymbol{I}_d)$ distribution, and let $\boldsymbol{u} = (u_1, \ldots, u_d)'$ denote a general realization of $\boldsymbol{U}$. Finally, let ϕ and Φ denote the standard univariate normal density and CDF, respectively. Given the set C defined previously and a vector $\boldsymbol{u}$, consider a collection of sets $C_j(u_{<j}), j = 1, \ldots, d$ defined by

$$C_j(u_{<j}) = \{u_j : (c_{j\ell} - \mu_j - \boldsymbol{L}_{j,<j}\boldsymbol{u}_{<j})/\ell_{jj} \leq u_j \leq (c_{ju} - \mu_j - \boldsymbol{L}_{j,<j}\boldsymbol{u}_{<j})/\ell_{jj}\} \tag{8}$$

Then

$$\begin{aligned} \Pr(\boldsymbol{Z} \in C) &= \Pr(\boldsymbol{L}\boldsymbol{U} + \boldsymbol{\mu} \in C) && (9) \\ &= \Pr\left(U_1 \in C_1(U_{<1}), U_2 \in C_2(U_{<2}), \ldots, U_d \in C_d(U_{<d})\right) && (10) \\ &= \int_{\mathcal{R}^d} \left[\prod_{j=1}^{d} I_{C_j(u_{<j})}(u_j)\phi(u_j)\right] d\boldsymbol{u} && (11) \\ &= \int_{\mathcal{R}^d} \left[\prod_{j=1}^{d} \Phi(C_j(u_{<j}))\right] \left[\prod_{j=1}^{d} \frac{I_{C_j(u_{<j})}(u_j)\phi(u_j)}{\Phi(C_j(u_{<j}))}\right] d\boldsymbol{u} && (12) \end{aligned}$$

The purpose of this sequence of equations is to transform the probability calculation under the distribution of $\boldsymbol{Z}$ into an expected value under the distribution of a different collection of random variables $\boldsymbol{W} = (W_1, \ldots, W_d)$. These random variables are defined recursively: W_1 has a static truncated univariate normal distribution, while the distribution of W_j is a truncated univariate normal distribution that is a function of $W_{<j}$. The joint density of the random variables defined in this manner is equal to the second bracketed term in the integrand of Equation (12) (which is the f function noted in Section 4.1). A realization $\boldsymbol{w}$ of the random vector $\boldsymbol{W}$ is obtained by recursively sampling from the proper univariate truncated normal distributions specified in Equation (12), and this realization is used to calculate the quantity $\prod_{j=1}^{d} \Phi(C_j(w_{<j}))$ (which is the h function noted in Section 4.1). The GHK method is to estimate the desired probability as the average of this quantity over $K >= 1$ realizations of $\boldsymbol{w}$.

References

Akaike, H. (1973). Theory and extension of the maximum likelihood principle. In B. Petrov and F. Csaki (Eds.), *Second International Symposium on Information Theory.* Budapest: Akademiai Kiado.

Alqallaf, F. and P. Gustafson (2001). On cross-validation of Bayesian models. *The Canadian Journal of Statistics 29*(2), 333–340.

Breslaw, J. (1994). Evaluation of multivariate normal probability integrals using a low variance simulator. *The Review of Economics and Statistics 76*, 673–682.

Chen, M.-H. and Q.-M. Shao (1998). Monte Carlo methods for Bayesian analysis of constrained parameter problems. *Biometrika 85*, 73–87.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association 90*, 1313–1321.

Dutt, J. (1973). A representation of multivariate normal probability integrals by integral transforms. *Biometrika 60*, 637–645.

Dyk, D. v. and X. Meng (2001). The art of data augmentation (with discussion). *Journal of Computational and Graphical Statistics 10*(1), 1–111.

Frey, M. and M. Edwards (1997). Surveying arsenic occurrence. *Journal of the American Water Works Association 89*(3), 105–117.

Gassmann, H., I. Deak, and T. Szantai (2002). Computing multivariate normal probabilities: A new look. *Journal of Computational and Graphical Statistics 11*(4), 920–949.

Geisser, S. and W. Eddy (1979). A predictive approach to model selection. *Journal of the American Statistical Association 74*(365), 153–160.

Gelfand, A., D. Dey, and H. Chang (1992). Model determination using predictive distributions with implementation via sampling-based methods (with discussion). In J. Bernardo, J. Berger, A. Dawid, and A. Smith (Eds.), *Bayesian Statistics 4*, pp. 147–167. Oxford: Oxford University Press.

Gelfand, A. E. and D. K. Dey (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society, Series B: Statistical Methodology 56*, 501–514.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics 1*, 141–149.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica 57*, 1317–1339.

Geweke, J. F., M. P. Keane, and D. E. Runkle (1997). Statistical inference in the multinomial multiperiod probit model. *Journal of Econometrics 80*, 125–165.

Gurian, P., M. Small, J. Lockwood, and M. Schervish (2001). Benefit-cost estimation for alternative drinking water MCLs. *Water Resources Research 37*(8), 2213–2226.

Hajivassiliou, V., D. McFadden, and P. Ruud (1996). Simulation of multivariate normal rectangle probabilities and their derivatives: Theoretical and computational results. *Journal of Econometrics 72*, 85–134.

Han, C. and B. Carlin (2001). Markov Chain Monte Carlo methods for computing Bayes factors: A comparative review. *Journal of the American Statistical Association 96*(455), 1122–1132.

Hansen, M. and B. Yu (2001). Model selection and the principle of minimum description length. *Journal of the American Statistical Association 96*(454), 746–774.

Hodges, J. and D. Sargent (2001). Counting degrees of freedom in hierarchical and other richly-parameterized models. *Biometrika 88*(2), 367–379.

Hoeting, J., D. Madigan, A. Raftery, and C. Volinsky (1999). Bayesian modeling averaging: A tutorial. *Statistical Science 14*(4), 382–417.

Kadane, J. and N. Lazar (2001). Methods and criteria for model selection. Technical Report 759, Carnegie Mellon University.

Kass, R. and A. Raftery (1995). Bayes factors. *Journal of the American Statistical Association 90*(430), 773–795.

Keane, M. (1994). A computationally practical simulation estimator for panel data. *Econometrica 62*, 95–116.

Laud, P. and J. Ibrahim (1995). Predictive model selection. *Journal of the Royal Statistical Society, Series B: Statistical Methodology 57*, 247–262.

Lockwood, J., M. Schervish, P. Gurian, and M. Small (2003). Analysis of contaminant co-occurrence in community water systems. Conditionally accepted for publication by *The Journal of the Americal Statistical Association*.

Mosteller, F. and J. Tukey (1977). *Data Analysis and Regression: A Second Course in Statistics*. Reading, MA: Addison-Wesley Publishing Company.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison (with discussion). *Journal of the Royal Statistical Society, Series B: Statistical Methodology 57*, 99–138.

Poskitt, D. (1987). Precision, complexity and Bayesian model determination. *Journal of the Royal Statistical Society, Series B: Statistical Methodology 49*(2), 199–208.

Rao, C. and Y. Wu (2001). On model selection (with discussion). In P. Lahiri (Ed.), *Model Selection*, Volume 38 of *IMS Lecture Notes - Monograph Series*. Institute of Mathematical Statistics.

Schervish, M. (1984). Multivariate normal probabilities with error bound. *Journal of the Royal Statistical Society, Series C: Applied Statistics 33*, 81–87.

Schervish, M. (1995). *Theory of Statistics* (Second ed.). New York: Springer-Verlag.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*(2), 461–464.

Spiegelhalter, D., N. Best, B. Carlin, and A. van der Linde (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B: Statistical Methodology 64*, 583–639.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society, Series B: Statistical Methodology 36*(2), 111–147.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society, Series B: Statistical Methodology 39*(1), 44–47.

Tanner, M. and W. Wong (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association 82*(398), 528–550.

Tibshirani, R. and K. Knight (1999). The covariance inflation criterion for adaptive model selection. *Journal of the Royal Statistical Society, Series B: Statistical Methodology 61*(3), 529–546.

Vijverberg, W. (1997). Monte Carlo evaluation of multivariate normal probabilities. *Journal of Econometrics 76*, 281–307.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association 93*(441), 120–131.

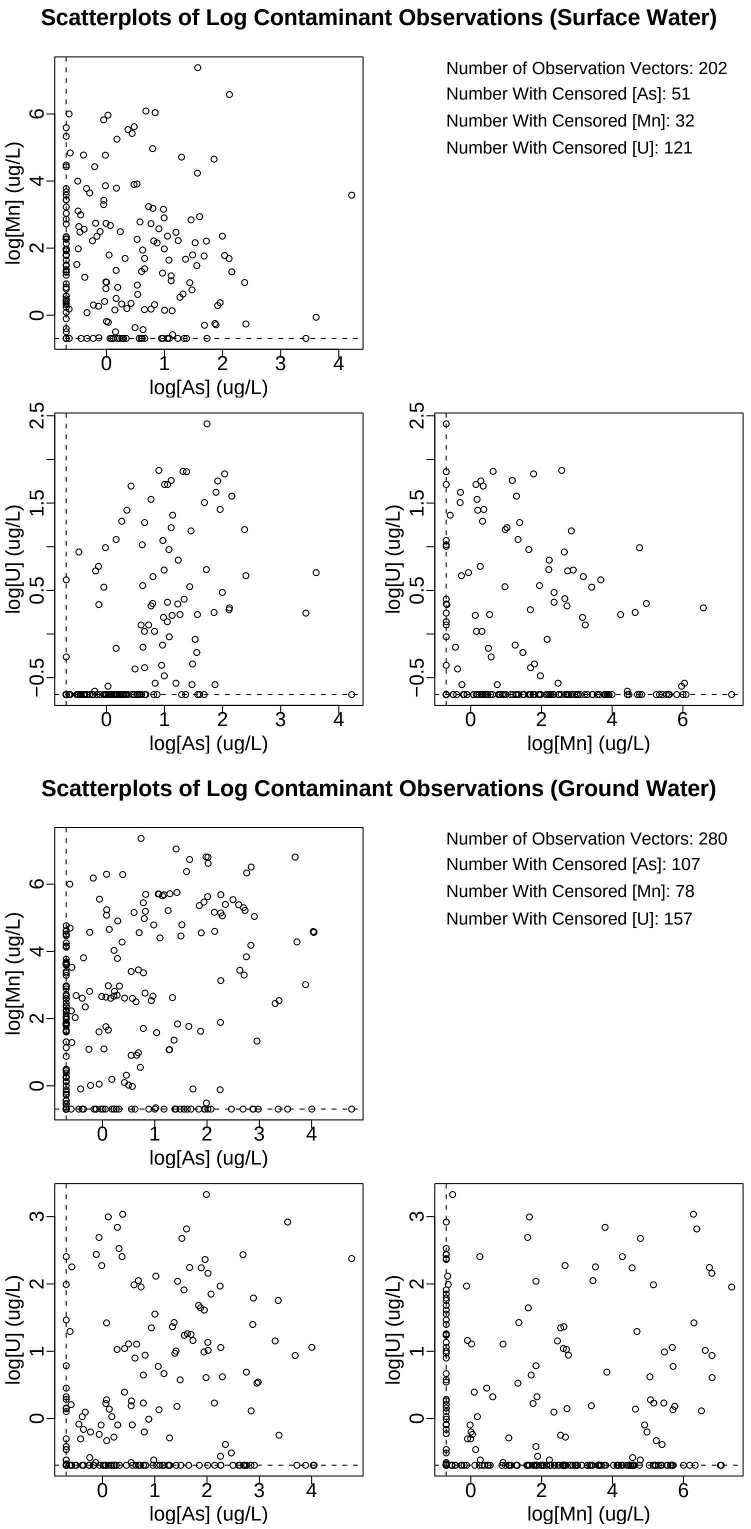

Figure 1: Scatterplots of log concentrations for As, Mn and U separately for surface and ground water. Censored observations fall along dotted lines.

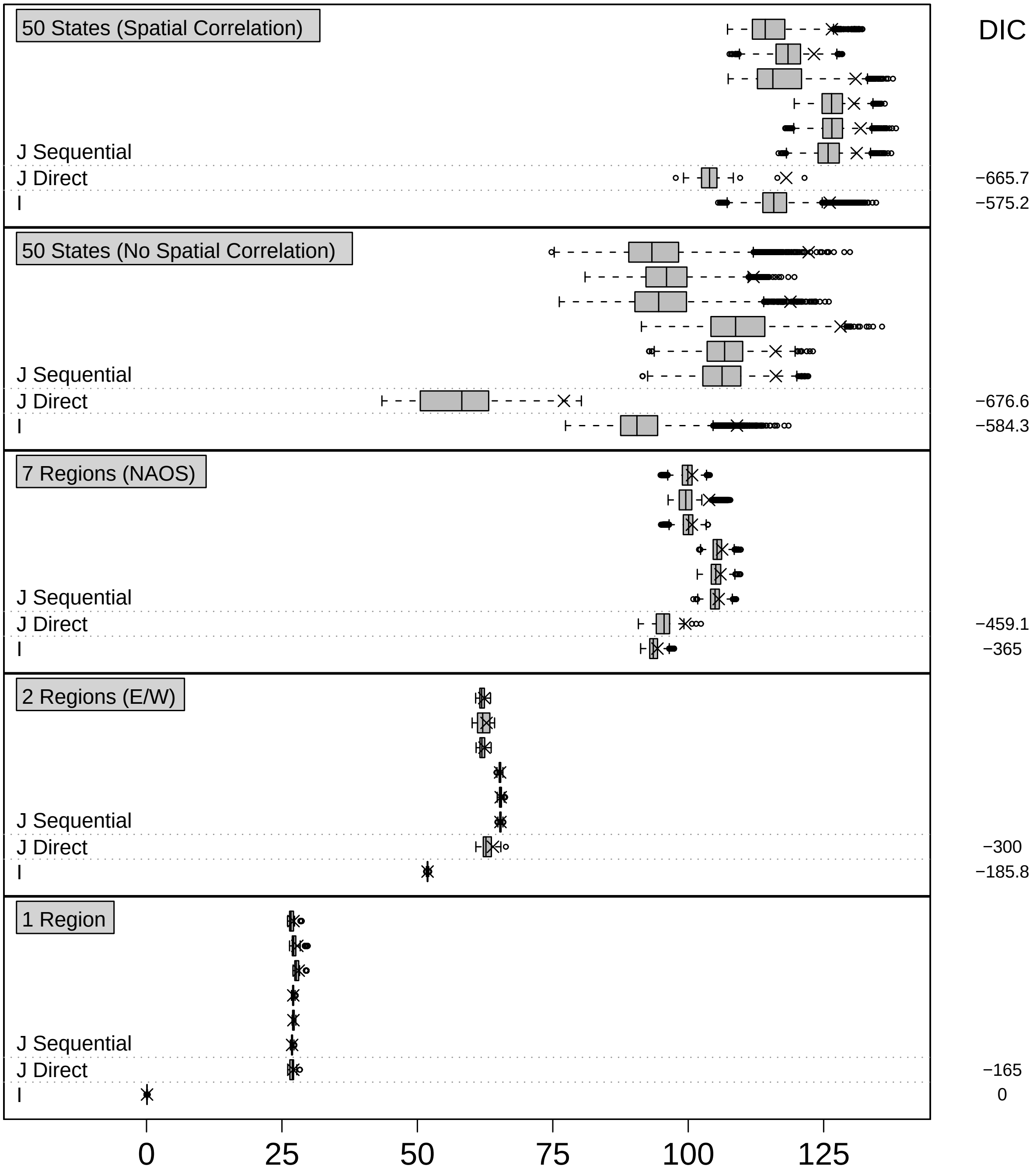

Figure 2: Estimated log predictive density (LPD) of the testing data under 10 models. “I” indicates the “Independence” models; “J Direct” and “J Sequential” refer to the estimates under the “Joint” models as calculated by the direct and sequential methods, respectively. The six different sequential estimates for each spatial complexity derive from the six possible orderings of the three contaminants. The LPD point estimates, based on 3 million MCMC iterations per model, are indicated by “x”, with the boxplots representing variability across contiguous blocks of 100,000 parameter vectors.