STATISTICS IN TRANSITION new series and *SURVEY METHODOLOGY*
Joint Issue: Small Area Estimation 2014
Vol. 17, No. 1, pp. 91–104

VARIATIONAL APPROXIMATIONS FOR SELECTING HIERARCHICAL MODELS OF CIRCULAR DATA IN A SMALL AREA ESTIMATION APPLICATION

Daniel Hernandez-Stumpfhauser[1], **F. Jay Breidt**[2], **Jean D. Opsomer**[3]

ABSTRACT

We consider hierarchical regression models for circular data using the projected normal distribution, applied in the development of weights for the Access Point Angler Intercept Survey, a recreational angling survey conducted by the US National Marine Fisheries Service. Weighted estimates of recreational fish catch are used in stock assessments and fisheries regulation. The construction of the survey weights requires the distribution of daily departure times of anglers from fishing sites, within spatio-temporal domains subdivided by the mode of fishing. Because many of these domains have small sample sizes, small area estimation methods are developed. Bayesian inference for the circular distributions on the 24-hour clock is conducted, based on a large set of observed daily departure times from another National Marine Fisheries Service study, the Coastal Household Telephone Survey. A novel variational/Laplace approximation to the posterior distribution allows fast comparison of a large number of models in this context, by dramatically speeding up computations relative to the fast Markov Chain Monte Carlo method while giving virtually identical results.

Key words: deviance information criterion, Laplace approximation, model selection, projected normal distribution.

1. Introduction

In the United States, the Marine Recreational Fisheries Statistics Survey (MRFSS) has been the traditional source of information on recreational fishing in saltwater. The key question for stock assessment and fisheries regulation is the amount of recreational fishing catch, determined from the simple relationship

$$\text{(recreational catch)} = \text{(catch per angler-trip)} \times \text{(number of angler-trips)}.$$

[1]University of North Carolina–Chapel Hill. E-mail: danielhs@live.unc.edu
[2]Colorado State University. E-mail: jbreidt@stat.colostate.edu
[3]Colorado State University. E-mail: jopsomer@stat.colostate.edu

Due to a number of coverage and measurement issues, the two factors in the above expression are measured using different surveys: (catch per angler-trip) is measured by an on-site survey called the Access Point Angler Intercept Survey (APAIS), while the number of angler-trips is measured by an off-site survey called the Coastal Household Telephone Survey (CHTS). Data from these two surveys are combined to estimate the recreational catch in 17 US states along the coast of the Atlantic Ocean and the Gulf of Mexico, during six two-month waves (January–February, March–April,..., November–December), in four different fishing modes (from the shoreline, from a private boat, from a small guided vessel called a charter boat, or from a large guided vessel called a party boat). Because the state of Florida is divided into its Atlantic coast and its Gulf of Mexico coast, we will refer to 18 "states" instead of 17.

As part of the weighting procedure for the APAIS, estimates are needed for the fraction of anglers who leave the fishing site during a prespecified time interval on a selected day. In principle, these estimates could be readily obtained from extensive historical data from the CHTS, consisting of reports on 980,000 trips between 1990 and 2008. These data include the angler's departure time (on a 24-hour clock) from the fishing site, the mode of fishing, the fishing date (from which we determine the two-month wave), and the fishing site (from which we determine the state). Figure 1 shows these data in histogram form for the state of Alabama. There are 24 histograms, corresponding to six waves by four fishing modes. The bars in the histograms, when normalized by sample sizes, can be regarded as direct estimates $\widehat{F}^{\text{direct}}_{hijk}$ of the hourly fractions of daily departures by state, wave, and mode:

$$F_{hijk} = \text{fraction of a day's anglers leaving a site during hour } h \text{ in state } i, \text{ wave } j, \text{ mode } k.$$

The fraction for any prespecified block of hours is then modeled as $\sum_h F_{hijk}$, where the sum is over all hours h in that block. Other time intervals are rounded to the nearest whole hours, for simplicity.

The direct estimates $\widehat{F}^{\text{direct}}_{hijk}$ from the off-site CHTS data are unbiased, but have a small (or even zero) sample size in many of the (h,i,j,k) cells, of which there are

$$(24 \text{ hours}) \times (18 \text{ states}) \times (6 \text{ waves}) \times (4 \text{ modes}) = (10368 \text{ cells}).$$

We therefore consider the small area estimation approach, combining the direct estimates with modeled estimates using the Fay and Herriot (1979) estimation method-

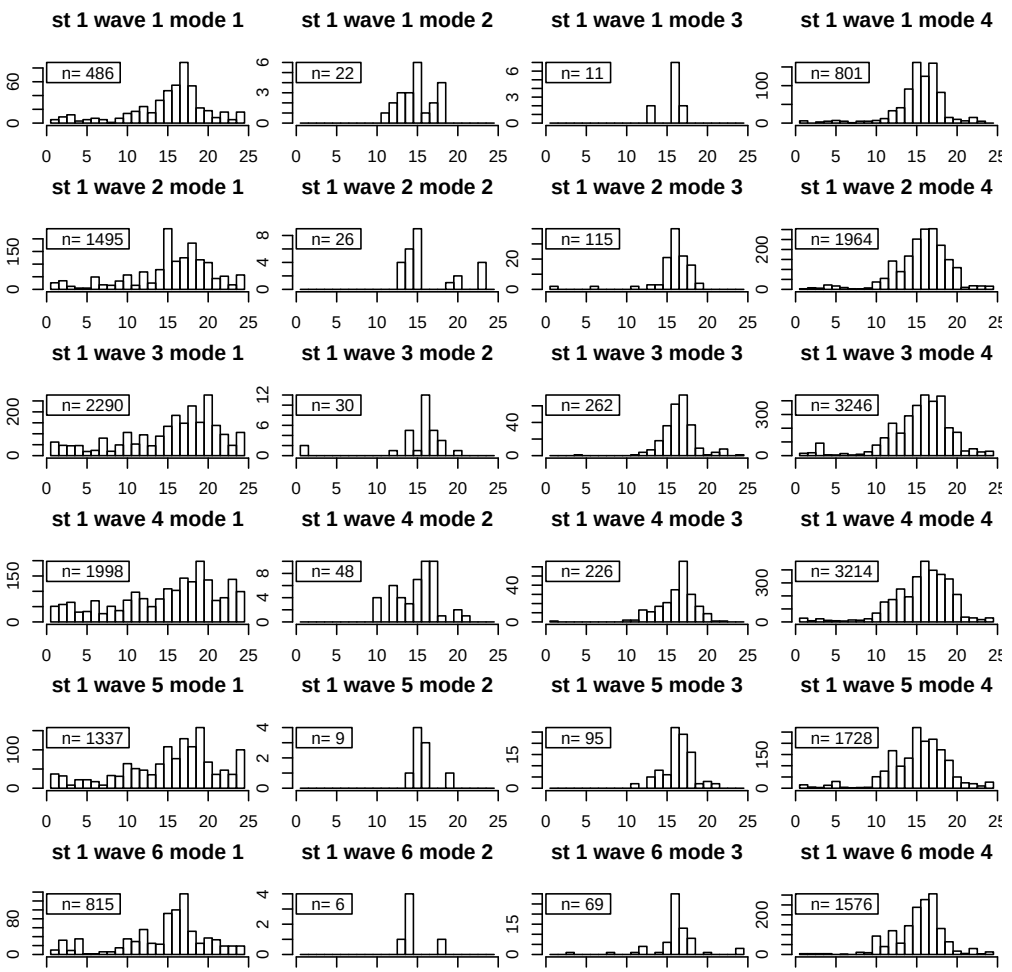

Figure 1: Histograms of trip departure times from the Coastal Household Telephone Survey for the state of Alabama (st 1) in six waves (top row = wave 1 = January–February, ..., bottom row = wave 6 = November–December) and four modes (column 1 = shoreline, 2 = private boat, 3 = charter boat, 4 = party boat).

ology. Briefly, we consider an area-level linear mixed model

$$\widehat{F}_{hijk}^{\text{direct}} = F_{hijk} + e_{hijk} = F_{hijk}^{\text{model}} + u_{hijk} + e_{hijk}$$

for $h = 1, \ldots, 23$ hours, where the sampling errors are assumed to be

$$\mathbf{e}_{ijk} = (e_{1ijk}, e_{2ijk}, \ldots, e_{23,ijk})^T \sim \text{independent } \mathcal{N}(0, \Psi_{ijk}),$$

with Ψ_{ijk} known, and where the model errors are assumed to be

$$\mathbf{u}_{ijk} = (u_{1ijk}, u_{2ijk}, \ldots, u_{23,ijk})^T \sim \text{independent } \mathcal{N}(0, \sigma^2\Delta_{ijk}),$$

with Δ_{ijk} of known form. Sampling errors and model errors are assumed to be independent. To implement the estimation strategy, we replace Ψ_{ijk} by design-based variance estimates and we choose Δ_{ijk} to be the variance of a scaled multinomial random vector, specified as follows. Consider a vector of 24 independent normal random variables with covariance matrix

$$\begin{aligned}\sigma^2 \text{diag}\left\{G_{1ijk}^{\text{model}}, \ldots, G_{23ijk}^{\text{model}}, G_{24ijk}^{\text{model}}\right\} \\ = \quad \sigma^2 \text{diag}\left\{F_{1ijk}^{\text{model}}\left(1 - F_{1ijk}^{\text{model}}\right), \ldots, F_{24ijk}^{\text{model}}\left(1 - F_{24ijk}^{\text{model}}\right)\right\}.\end{aligned}$$

Then $\sigma^2\Delta_{ijk}$ is the covariance matrix of the first 23 elements of the vector, conditioned on the sum of the 24 elements being equal to one; namely,

$$\begin{aligned}\sigma^2\Delta_{ijk} \quad = \quad & \sigma^2 \text{diag}\left\{G_{1ijk}^{\text{model}}, \ldots, G_{23ijk}^{\text{model}}\right\} \\ & - \frac{\sigma^2}{\sum_{\tau=1}^{24} G_{\tau ijk}^{\text{model}}} \begin{bmatrix} G_{1ijk}^{\text{model}} \\ \vdots \\ G_{23ijk}^{\text{model}} \end{bmatrix} \left[G_{1ijk}^{\text{model}}, \cdots, G_{23ijk}^{\text{model}}\right]. \qquad (1)\end{aligned}$$

We use a projected normal model for F_{hijk}^{model} to account for the circular nature of the time-of-day departure data, replacing F_{hijk}^{model} by posterior means $\text{E}\left[F_{hijk}^{\text{model}} \mid D\right]$ and also G_{hijk}^{model} by $\left(\text{E}\left[F_{hijk}^{\text{model}} \mid D\right]\right)\left(1 - \text{E}\left[F_{hijk}^{\text{model}} \mid D\right]\right)$ for implementation. The mean vector in the projected normal includes state, wave, and mode effects to account for the spatial and temporal distribution of fishing behavior. Since we consider various interactions among the effects as well as placement within the hierarchy (essentially, specifying whether a given effect is treated as fixed or random), we are interested in conducting model selection.

The main contribution of the present paper is to show that in this small area

estimation context, with a model somewhat more complex than a hierarchical linear model (due to the embedding in a projected normal model), fast and accurate model selection can be accomplished with a Laplace/variational approximation. Specifically, we show that a simple and fast deterministic approximation can replace a sophisticated Markov Chain Monte Carlo (MCMC) sampler, giving results that are essentially identical at a far lower computational cost. In this paper, we emphasize model selection as both the motivation for the deterministic approximation and the evaluation of its accuracy. However, the Laplace/variational approximation can also be used effectively in model estimation and inference even when no model selection is needed.

In §2.1, we briefly review the projected normal distribution. The MCMC procedure that serves as the benchmark for comparison is presented in §2.2. The variational approximation is given in §3.1 with its Laplace refinement in §3.2. Model selection criteria based on MCMC and on the Laplace/variational approximation are compared in §4; discussion follows in §5.

2. Inference for the projected normal distribution

2.1. The projected normal distribution

Suppose $X = (X_1, X_2)^T \sim \mathcal{N}(\mu, I_2)$, the bivariate normal distribution with mean vector μ and identity covariance matrix I_2. Writing X in polar coordinates, we have

$$X_1 = \|X\| \cos D = R \cos D, \quad X_2 = \|X\| \sin D = R \sin D.$$

Discarding the random length $R \in (0, \infty)$, the random angle $D \in [0, 2\pi)$ has a projected normal distribution, $\mathscr{PN}(\mu, I_2)$. As illustrated in Figure 2, the parameter vector μ plays the role of both "location" and "spread" for the projected normal: the further μ lies from the origin, the more concentrated the $\mathscr{PN}$ distribution around the direction determined by μ. As $\mu \to 0$, the $\mathscr{PN}$ distribution converges to the uniform distribution on the unit circle. In our application, the departure time d_{ijkt} for trip t in state i, wave j, mode k is on the 24-hour clock. Converting clock time to $[0, 2\pi)$, we model $D_{ijkt} = 2\pi d_{ijkt}/24$ as independent and identically distributed projected normals within state×wave×mode cells. For observations following a projected normal distribution, the fraction F_{hijk} for a given hour h is the integral of the projected normal probability density function over the interval $(2\pi(h-1)/24, 2\pi h/24]$.

Presnell, Morrison and Littell (1998) used the projected normal distribution as the basis for the Spherically Projected Multivariate Linear Model (SPMLM) for

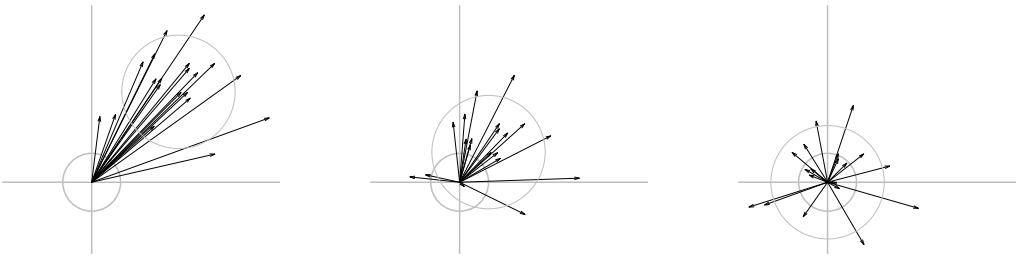

Figure 2: Realizations ($n = 20$) from three projected normal distributions. The large circle is centered at mean vector μ of bivariate normal $\mathcal{N}(\mu, I_2)$ and contains 95% of its probability. Arrows are the realized bivariate normal random vectors $(R\cos D, R\sin D)$. Projected normal random variables are the angles D, or the intersections of the normal random vectors with the unit circle (small circle), scaled to $[0, 2\pi)$. Left: Projected normal distribution with mode equal to $\pi/4$ and with low variance. Middle: Projected normal distribution with mode equal to $\pi/4$ and with high variance. Right: Projected normal distribution that is uniform on the unit circle.

directional data, specifying μ as a linear model. Parameters of the model were estimated with the maximum likelihood and the EM algorithm in Presnell et al. (1998). In the current paper, we specify hierarchical linear models for μ_{ijk} in terms of categorical covariates for the state, wave and mode. We conduct Bayesian inference for the model, comparing approximate posterior inference based on Markov Chain Monte Carlo to approximate inference based on deterministic approximations.

2.2. Markov Chain Monte Carlo for the projected normal distribution

The key step in conducting Bayesian inference under the SPMLM is to augment the observed angles $\{D_{ijkt}\}$ with the latent lengths $\{R_{ijkt}\}$, so that the structure of the complete data is simply that of a normal linear model. See Nuñez-Antonio and Gutiérrez-Peña (2005), Nuñez-Antonio, Gutiérrez-Peña, and Escalera (2011), and Hernandez-Stumpfhauser (2012) for details.

The likelihood for the complete-data model is the product of the joint densities of $\left(R_{ijkt}, D_{ijkt}\right)$ which can be obtained by a change of variables $X_{ijkt} = R_{ijkt}A_{ijkt}$, where X_{ijkt} is distributed as $\mathcal{N}\left(\mu_{ijk}, I_2\right)$ and $A_{ijkt} = \left(\cos\left(D_{ijkt}\right), \sin\left(D_{ijkt}\right)\right)^T$:

$$p\left(R_{ijkt}, D_{ijkt} \mid \mu_{ijk}\right) = \frac{1}{2\pi} r_{ijkt} \exp\left\{-\frac{1}{2}\left(R_{ijkt}A_{ijkt} - \mu_{ijk}\right)^T \left(R_{ijkt}A_{ijkt} - \mu_{ijk}\right)\right\}.$$

We specify conjugate normal priors for μ_{ijk}. For example, for a model specified as $\mu_{ijk} = \mu + m_k + s_i + w_j$, we set vague normal priors for the overall mean μ and mode effects m_k, and mean-zero normal priors with inverse gamma variances for the random state effects s_i and wave effects w_j.

In this work, we draw the latent lengths using a slice sampler (Neal 2003). Given the latent lengths and the conjugate priors, the full conditionals of the model parameters all have closed forms, and so the Gibbs sampler is fast and easy to conduct. Nonetheless, the large number of models to be evaluated led us to consider fast, deterministic approximations to the posterior distribution. This is the subject of the next section.

3. Deterministic approximations to the posterior

3.1. Variational approximation

In this context, a carefully-developed MCMC works well and serves as a benchmark for comparison. But it is extremely slow, given the very large size of the off-site CHTS data set. Because we wanted to compare a number of different model specifications, we investigated replacing the MCMC approximation of the full posterior distribution by a deterministic "variational approximation" that is easier to compute.

The variational idea is to find the best approximation of the posterior within a class of densities $\mathscr{Q}$, which is chosen so that the densities in the class are more analytically tractable than the posterior density itself. A natural choice for the "best" approximating density in $\mathscr{Q}$, and the one most commonly used, is the density that minimizes the Kullback-Leibler (KL) distance between it and the posterior density. Let D denote the observed data and ω denote the unknown parameters, so that $p(D,\omega)$ is their joint density and $p(\omega \mid D)$ is the unknown posterior density. Let $q(\omega)$ denote a density in $\mathscr{Q}$. Finding q that minimizes the KL distance to $p(\omega \mid D)$ is equivalent to maximizing the *variational lower bound*, denoted by

$$\underline{p}(D;q) = \exp\left[\int q(\omega)\log\left\{\frac{p(D,\omega)}{q(\omega)}\right\} d\omega\right]. \tag{2}$$

Let

$$q_* = \max_{q\in\mathscr{Q}} \underline{p}(D;q).$$

If $\mathscr{Q} = \{\text{all densities } q\}$, then $q_*(\omega) = p(\omega \mid D)$, the true posterior of ω given D. If $\mathscr{Q}$ is a sufficiently rich class of densities, then q_* should be a good approximation to the true posterior. In practice, the approximation method is necessarily of limited

accuracy, so q_* will not converge to the true posterior if the true posterior $\notin \mathcal{Q}$.

If $\mathcal{Q} = \left\{q : q(\omega) = \prod_{m=1}^{M} q_{\omega_m}(\omega_m)\right\}$, then $q_*(\omega) = \prod_{m=1}^{M} q_{*\omega_m}(\omega_m)$, which is called a "mean field variational approximation" (Bishop 2006; Ormerod and Wand 2010). This approximation can be computed efficiently, even for very large samples. See Ormerod and Wand (2010) for an excellent review. The solution satisfies

$$\begin{aligned} q_{*\omega_1}(\omega_1) &\propto \exp\{\mathrm{E}_{-\omega_1} \log p(\omega_1 \mid \omega_2, \ldots, \omega_M, D)\} \\ q_{*\omega_2}(\omega_2) &\propto \exp\{\mathrm{E}_{-\omega_2} \log p(\omega_2 \mid \omega_1, \omega_3, \ldots, \omega_M, D)\} \\ &\vdots \\ q_{*\omega_M}(\omega_M) &\propto \exp\{\mathrm{E}_{-\omega_M} \log p(\omega_M \mid \omega_1, \ldots, \omega_{M-1}, D)\}, \end{aligned}$$

where $\mathrm{E}_{-\omega_m}[\cdot]$ denotes expectation with respect to all of the variational component distributions except $q_{*\omega_m}$.

In our setting, $q_{*\omega_m}(\cdot)$ and $\mathrm{E}_{-\omega_m}[\cdot]$ have simple parametric forms; iteratively updating the parameters leads to the solution. Convergence is assessed by monitoring the change in the lower bound $\underline{p}(D;q)$ from (2).

For simplicity, we begin by considering $\{D_t\}$ independent and identically distributed $\mathscr{PN}(\mu, I_2)$, with prior $p(\mu) = \mathcal{N}_2\left(\mu_0, \sigma_0^2 I_2\right)$. Denoting the observed data as $A_t^T = (\cos D_t, \sin D_t)$, the mean field variational approximation satisfies

$$q_{*\mu}(\mu) = \mathcal{N}\left(\frac{\mu_0/\sigma_0^2 + \sum_{t=1}^{n} \mathrm{E}_{-r_t}(r_t) A_t}{n + \left(1/\sigma_0^2\right)}, \frac{1}{n + \left(1/\sigma_0^2\right)} I_2\right) \tag{3}$$

$$q_{*r_t}(r_t) \propto r_t \exp\left(-\frac{1}{2} r_t^2 + r_t A_t^T \mathrm{E}_{-\mu}(\mu)\right), \tag{4}$$

where the expectations in these expressions are computed iteratively:

$$\begin{aligned} \mathrm{E}_{-\mu}(\mu) &\leftarrow \frac{\mu_0/\sigma_0^2 + \sum_{t=1}^{n} \mathrm{E}_{-r_t}(r_t) A_t}{n + \left(1/\sigma_0^2\right)} \\ b_t &\leftarrow A_t^T \mathrm{E}_{-\mu}(\mu) \\ \mathrm{E}_{-r_t}(r_t) &\leftarrow b_t + \frac{\sqrt{2\pi}\, \exp(b_t^2/2)\, \Phi(b_t)}{1 + \sqrt{2\pi}\, b_t\, \exp(b_t^2/2)\, \Phi(b_t)}. \end{aligned}$$

The mean field variational approximation yields a highly tractable approximate posterior, and the iterative solution is simple to compute and fast to converge. In fact, it is proved in Hernandez-Stumpfhauser (2012) that the parameter iterations for μ converge to the posterior mode of the parameters of the projected normal distribution, denoted here by $\mu^\dagger$. Extension of the mean field variational approx-

imation to the case of a linear model for μ is straightforward, and involves updates of means and variances of each one of the fixed and random effects as well as updates of the means of inverse variances of the random effects. See Hernandez-Stumpfhauser (2012) for details.

3.2. Refinement via Laplace approximation

While the mean field variational approximation of the previous section is simple and fast, it is not very accurate. Indeed, the approximate posterior variance for μ in (3) depends on the sample size but not on the data, and so cannot be accurate except in simple cases. Our approach to improving the accuracy of the variational approximation is to replace $q_{*\mu}(\mu)$ by a Laplace approximation $\mathcal{N}(\mu^\dagger, V^\dagger)$, where $\mu^\dagger$ is the posterior mode and the covariance matrix $V^\dagger$ is the inverse of minus the Hessian of the log posterior distribution evaluated at the mode,

$$V^\dagger = \left(- \left[\begin{array}{cc} \frac{\partial^2}{\partial \mu_1^2} \log p(\mu \mid D) & \frac{\partial^2}{\partial \mu_1 \mu_2} \log p(\mu \mid D) \\ \frac{\partial^2}{\partial \mu_1 \mu_2} \log p(\mu \mid D) & \frac{\partial^2}{\partial \mu_2^2} \log p(\mu \mid D) \end{array} \right] \Bigg|_{\mu=\mu^\dagger} \right)^{-1}.$$

The log posterior distribution is

$$\log p(\mu \mid D) = \log \mathcal{N}\left(\mu_0, \sigma_0^2 I_2\right) + \sum_{t=1}^{n} \log \mathscr{P}\mathcal{N}(D_t; \mu, I_2) + C,$$

where C is a term that does not depend on μ, and the calculations to compute the Hessian are given in Hernandez-Stumpfhauser (2012). This Laplace refinement to the variational approximation greatly improves the quality of the original approximation, as is shown in Hernandez-Stumpfhauser (2012) by comparing the variational approximation and the variational/Laplace approximation to the output of the Gibbs sampler. Similar results hold in the regression case: the Laplace refinement substantially improves the quality of the variational approximation.

4. Comparing model selection via Gibbs, variational, and variational/Laplace

For a general Bayesian estimation problem, the deviance is defined as $\Delta(D, \omega) = -2 \ln p(D \mid \omega)$ where D are the data, ω are the unknown parameters and $p(D \mid \omega)$ is the likelihood function (Gelman et al. 2004, p. 179–184). The expected deviance $\mathrm{E}[\Delta(D, \omega) \mid D]$ is a measure of how well the model fits and it can be estimated by the posterior mean deviance $\overline{\Delta(D)} = B^{-1} \sum_{b=1}^{B} \Delta(D, \omega^{(b)})$, where $\{\omega^{(b)}\}_{b=1}^{B}$ are

random draws from the posterior distribution. The difference between the posterior mean deviance and the deviance at the posterior mean, estimated as

$$p_\Delta = \overline{\Delta(D)} - \Delta(D, \bar{\omega})$$

where $\bar{\omega} = B^{-1} \sum_{b=1}^{B} \omega^{(b)}$, is often interpreted as a measure of the effective number of parameters of a Bayesian model. More generally, p_Δ can be thought of as the number of "unconstrained" parameters in the model, where a parameter counts as 1 if it is estimated without constraints or prior information, 0 if it is fully constrained or if all the information about the parameter comes from the prior distribution, or an intermediate value if both the data and prior distributions are contributing.

We used the Deviance Information Criterion,

$$\text{DIC} = 2\overline{\Delta(D)} - \Delta(D, \bar{\omega}) = \overline{\Delta(D)} + p_\Delta,$$

to compare different model specifications for the departure time data. The DIC can be interpreted as a measure of goodness-of-fit (the estimated expected deviance) plus a penalty for model complexity in the form of the total number of effective parameters. Lower values of DIC correspond to more preferable tradeoffs between fit and model complexity.

We evaluated a large number of different model specifications for the mean of the projected normal distribution, including fixed and random effects for the states, waves and modes as well as for interactions between these factors. As an example of the type of models compared, the following is the full hierarchical specification for a model with mode as fixed effect and state and wave as random effects:

$$\begin{aligned}
D_{ijkt} &\sim \mathscr{PN}(\mu_{ijk}, I_2) \\
\mu_{ijk} &= \mu + m_k + s_i + w_j \\
\mu &\sim \mathscr{N}(0, 10^6 I_2) \\
m_k &\sim \mathscr{N}(0, 10^6 I_2) \\
s_i &\sim \mathscr{N}(0, \sigma_s^2 I_2) \\
w_j &\sim \mathscr{N}(0, \sigma_w^2 I_2) \\
\sigma_s^2 &\sim \mathscr{IG}(0.001, 0.001) \\
\sigma_w^2 &\sim \mathscr{IG}(0.001, 0.001).
\end{aligned}$$

In this specification, μ, m_k have vague priors while those of s_i, w_j are determined by their variance parameters, which follow pre-specified inverse gamma hyper-priors. This hierarchical set-up is similar to the usual Bayesian normal regression model.

We computed DIC and p_Δ values using Gibbs sampling, variational and variational/Laplace. Table 1 shows the Gibbs DIC values for different models applied to the departure time data. In Table 1, the models containing all three factors (mode, state, wave) consistently achieve lower DIC values than the models that excluded any of those factors. While not shown here, models with mode as random effect performed worse than models with mode as fixed effect. In contrast, very similar DIC values were obtained with the state and wave treated as either fixed or random. When we investigated models with interactions between the three factors, those with state-wave interactions scored better than any other arrangement of two-way interactions. Among the various models considered, DIC leads to selection of

$$\mu_{ijk} = \mu + m_k + sw_{ij},$$

with m_k a fixed mode effect and sw_{ij} a random interaction effect between the state and wave, with 99 total levels. Note that there are $6 \times 18 = 108$ possible state-wave combinations, so that there are nine state-wave combinations without observations where a mode-only model was applied. This was the final model used for purposes of small area estimation.

The effective number of parameters p_Δ for each model, computed via Gibbs sampling, are shown in Table 2. In interpreting these values, it should be noted that one level of a factor is represented by two parameters. Hence, in a model with only a mode effect there are eight parameters: two for the overall mean and six more for the three remaining free mode levels. The model with only a mode effect has p_Δ values (in the first row of Table 1) very close to eight. The final selected model has $p_\Delta = 191.5$.

We now turn to a comparison of the computation of DIC and p_Δ using Gibbs, variational and variational/Laplace. All three methods yield essentially identical posterior means $\bar{\omega}$, so $\Delta(D, \bar{\omega})$ is also essentially identical across methods. The differences in DIC across methods, displayed in Table 1, and differences in p_Δ across methods, displayed in Table 2, therefore come from differences in the posterior mean deviance $\overline{\Delta(D)}$ across methods. As can be seen from the two tables, the variational approximation without Laplace refinement significantly underestimates the posterior mean deviance, resulting in large negative differences in both DIC and p_Δ values. By contrast, Gibbs and variational/Laplace yield nearly identical estimates of the posterior mean deviance, hence virtually identical DIC and p_Δ values. For the tabled results, iterating the variational/Laplace approximation to convergence is about 15 times faster than 5000 iterates of Gibbs sampling. For purposes of model selection, therefore, the variational/Laplace approximation performs extremely well in this example.

Table 1: DIC values from Gibbs sampler for ten different projected normal model specifications, along with comparisons to DIC computed via other methods: Variational DIC minus Gibbs DIC and Variational/Laplace DIC minus Gibbs DIC.

Fixed Effects	Random Effects	Gibbs DIC	Variational − Gibbs DIC	Variational/Laplace − Gibbs DIC
mode		2642714.6	−2.7	−0.5
mode; wave		2631925.2	−4.3	0.1
mode	wave	2631925.9	−5.1	0.0
mode; state		2626382.7	−18.1	0.7
mode	state	2626383.6	−20.0	0.1
mode; wave; state		2616177.1	−23.4	0.2
mode; state	wave	2616177.2	−23.5	−1.7
mode; wave	state	2616175.4	−21.5	1.3
mode	state; wave	2616176.3	−22.9	−0.4
mode	state×wave	2613338.4	−105.9	−0.4

Table 2: Effective number of parameters p_Δ values from Gibbs sampler for ten different projected normal model specifications, along with comparisons to effective number of parameters computed via other methods: Variational p_Δ minus Gibbs p_Δ and Variational/Laplace p_Δ minus Gibbs p_Δ.

Fixed Effects	Random Effects	Gibbs p_Δ	Variational − Gibbs p_Δ	Variational/Laplace − Gibbs p_Δ
mode		8.3	−1.3	−0.2
mode; wave		17.7	−2.1	0.1
mode	wave	18.0	−2.5	0.0
mode; state		41.4	−9.0	0.4
mode	state	41.8	−9.9	0.1
mode; wave; state		52.5	−11.7	0.1
mode; state	wave	52.5	−11.7	−0.8
mode; wave	state	51.5	−10.7	0.7
mode	state;wave	52.0	−11.4	−0.2
mode	state×wave	191.5	−53.4	−0.9

5. Discussion

In this paper, we have briefly described an important small area estimation problem in which a hierarchical linear model is embedded in a nonlinear, projected normal model. A massive data set is considered, for which MCMC is feasible but slow. A large number of models are compared. Though a mean field variational approximation is not very accurate in this problem, it can be refined substantially by using a Laplace approximation, and the resulting variational/Laplace approximation is both accurate and extremely fast to compute. In particular, model selection results are virtually indistinguishable between the MCMC and the variational/Laplace approaches. While these results are limited to the particular problem under consideration, they do suggest that there is considerable promise for variational/Laplace approximations in model selection and inference in small area estimation problems.

REFERENCES

BISHOP, C. M., (2006). *Pattern Recognition and Machine Learning*. Springer.

FAY, R. E., HERRIOT, R. A., (1979). Estimation of Income from Small Places: An Application of James-Stein Procedures to Census Data. *Journal of the American Statistical Association 74*, 269–277.

GELMAN, A., CARLIN, J. B., STERN, H. S., RUBIN, D. B., (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC.

HERNANDEZ-STUMPFHAUSER, D., (2012). *Topics in Design-Based and Bayesian Inference for Surveys*. Ph. D. thesis, Colorado State University.

NEAL, R. M., (2003). Slice Sampling. *The Annals of Statistics 31*, 705–741.

NUÑEZ-ANTONIO, G., GUTIÉRREZ-PEÑA, E., (2005). A Bayesian Analysis of Directional Data Using the Projected Normal Distribution. *Journal of Applied Statistics 32*(10), 995–1001.

NUÑEZ-ANTONIO, G., GUTIÉRREZ-PEÑA, ESCALERA, E. G., (2011). A Bayesian Regression Model for Circular Data Based on the Projected Normal Distribution. *Statistical Modeling 11*, 185–201.

ORMEROD, J. T., WAND, M. P., (2010). Explaining Variational Approximations. *The American Statistician 64*, 140–153.

PRESNELL, B., MORRISON, S. P., LITTELL, R. C., (1998). Projected Multivariate Linear Models for Directional Data. *Journal of the American Statistical Association 93*(443), 1068–1077.