Independent Metropolis-Hastings Steps for Generalized Linear Models with Latent Gaussian Processes via Global Conditional Laplace Approximations

Abstract

KEY WORDS:

1 Introduction

2 Fitting GLMMs with LGPs

We consider a class of generalized linear mixed models (GLMMs; Stroup, 2012) with latent Gaussian processes (LGPs). We will conceptualize our class of models using the strategy of Berliner (1996) and Wikle et al. (2003), that is hierarchically with a data model conditional on parameters and a latent process, a process model conditional on parameters, and finally a parameter model. Suppose we observe an m-dimensional vector $\boldsymbol{z}_i$ for each observational unit i, and let $\boldsymbol{z}_{1:n} = (\boldsymbol{z}_1', \boldsymbol{z}_2', \dots, \boldsymbol{z}_n')'$ where n is the number of observational units. We assume these observations are conditionally independent given a set of observation-specific location parameters and some common data level parameter $\boldsymbol{\phi}$. Let $\boldsymbol{\mu}_i$ denote the location parameter for observation i and $\boldsymbol{\mu}_{1:n} = (\boldsymbol{\mu}_1', \boldsymbol{\mu}_2, \dots, \boldsymbol{\mu}_n')'$. Then the data model is $\boldsymbol{z}_i|\boldsymbol{\mu}, \boldsymbol{\phi} \overset{ind}{\sim} [\boldsymbol{z}|\boldsymbol{\mu}_i, \boldsymbol{\phi}]$ where $[\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\phi}]$ is some known parameterized family of densities for the $\boldsymbol{z}_i$s indexed by $\boldsymbol{\mu}$ and $\boldsymbol{\phi}$; often an exponential family. The location parameters are a known function of a LGP, i.e. $\boldsymbol{\mu}_i = \boldsymbol{g}(\boldsymbol{y}_i)$ where $\boldsymbol{g}$ is known and $\boldsymbol{y}_{1:n}$ is a LGP where $\boldsymbol{y}_{1:n} = (\boldsymbol{y}_1', \boldsymbol{y}_2', \dots, \boldsymbol{y}_n')'$. Then at the process level we model $\boldsymbol{y}_{1:n}$ as sum of function of fixed effects and random effects, namel $\boldsymbol{y}_{1:n} = \boldsymbol{X}_{1:n}\boldsymbol{\beta} + \boldsymbol{S}_{1:n}\boldsymbol{\delta}$ where $\boldsymbol{X}_{1:n}$ and $\boldsymbol{S}_{1:n}$ are known $mn \times p$ and $mn \times r$ matrices respectively, i.e. $\boldsymbol{X}_{1:n} = (\boldsymbol{X}_1', \boldsymbol{X}_2', \dots, \boldsymbol{X}_n')'$ where $\boldsymbol{X}_i$ a $m \times p$ matrix and similarly for $\boldsymbol{S}_{1:n}$. In addition $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ are unknown p-dimensional and r-dimensional column vectors respectively. Further, we assume $\boldsymbol{\delta} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$ The process model can be more succinctly written as $\boldsymbol{y}_{1:n} \sim N(\boldsymbol{X}_{1:n}\boldsymbol{\beta}, \boldsymbol{S}_{1:n}\boldsymbol{\Sigma}\boldsymbol{S}_{1:n}')$ so that $\boldsymbol{y}_{1:n}$s mean is structured by $\boldsymbol{X}_{1:n}$ and its covariance matrix is structured by $\boldsymbol{S}_{1:n}$. Further, we assume that $\boldsymbol{\Sigma}$ depends on the unknown parameter $\boldsymbol{\phi}$ to allow for, e.g., parsimonious representations of $\boldsymbol{\Sigma}$. We allow $\boldsymbol{\phi}$ to enter both the data model and $\boldsymbol{\Sigma}$, though in many applications they will depend on distinct components of $\boldsymbol{\phi}$. The leaves the unknown parameters $(\boldsymbol{\phi}, \boldsymbol{\beta})$, which in the parameter model obtain a joint prior distribution denoted by $[\boldsymbol{\phi}, \boldsymbol{\beta}]$. For notational simplicity, let $\boldsymbol{D}_n = (\boldsymbol{z}_{1:n}, \boldsymbol{X}_{1:n}, \boldsymbol{S}_{1:n})$ denote all data in the model and $\boldsymbol{\theta} = (\boldsymbol{\phi}, \boldsymbol{\beta})$ denote all model parameters.

A common MCMC approach in these models is data augmentation (DA; Tanner and Wong, 1987). That is, a two step Gibbs sampler consiting of a draw from $[\boldsymbol{\theta}|\boldsymbol{y}_{1:n}, \boldsymbol{D}_n]$ and a draw from $[\boldsymbol{y}_{1:n}|\boldsymbol{\theta}, \boldsymbol{D}_n]$ iterated repeatedly. In GLMMs with non-Gaussian data models, $[\boldsymbol{y}_{1:n}|\boldsymbol{\theta}, \boldsymbol{D}_n]$ is typically intractable and thus requires a MH step. Often a random walk proposal is used here, but if a good approximation for $[\boldsymbol{y}_{1:n}|\boldsymbol{\theta}, \boldsymbol{D}_n]$ is available, then an independent MH (IMH) Gibbs step is attractive. Laplace approximations (LAs) provide a useful approach for constructing a good IMH proposal density for $\boldsymbol{y}_{1:n}$'s conditional posterior.

3 Laplace Approximations

Let $\boldsymbol{\omega}$ be the model parameter and $\boldsymbol{D}_n$ the observed data so that $[\boldsymbol{\omega}|\boldsymbol{D}_n]$ is the posterior distribution, available up to a normalizing constant. Further let $\boldsymbol{\omega}_n^*$ denote the posterior mode of $[\boldsymbol{\omega}|\boldsymbol{D}_n]$ and let $\boldsymbol{H}_n(\boldsymbol{\omega}_n^*)$ denote the Hessian matrix of $\log[\boldsymbol{\omega}|\boldsymbol{D}_n]$ evaluated at $\boldsymbol{\omega}_n^*$. Then under suitable regularity conditions $\boldsymbol{\omega}$'s posterior distribution is asymptotically normal (Schervish, 1997, Sections 7.4.2 and 7.4.3), so for a fixed but large value of n, $\boldsymbol{\omega}|\boldsymbol{D}_n \overset{a}{\sim} N(\boldsymbol{\omega}_n^*, -\boldsymbol{H}_n^{-1}(\boldsymbol{\omega}_n^*))$ where the notation $\overset{a}{\sim}$ means "approximately distributed as." We do not discuss the precise technical conditions necessary for this result here, but intuitively for the approximation to be good each element of $\boldsymbol{\omega}$ must have enough "observations" for asymptotics to kick in. We put "observations" in scare quotes because often rather than data, it is other elements of $\boldsymbol{\omega}$ are directly informing on another element, e.g. in the GLMM case above $\boldsymbol{\omega} = (\boldsymbol{\theta}, \boldsymbol{y}_{1:n})$ and $\boldsymbol{y}_{1:n}$ is directly informing on $\boldsymbol{\theta}$. In that case the asymptotic argument does not apply to $\boldsymbol{\omega}$, but it typically does apply for $\boldsymbol{\theta}$ alone so that if $\boldsymbol{y}_{1:n}$ assumed to be Gaussian conditional on $\boldsymbol{\theta}$, then a normal approximation may still be reasonable for the posterior of $\boldsymbol{\omega}$. It "may" be reasonable because the non-Gaussian data model implies that $\boldsymbol{y}_{1:n}$ is non-Gaussian in its full conditional posterior, but it is still often approximately Gaussian. We call this LA a global LA (GLA) since it approximates the full joint posterior of $\boldsymbol{\omega} = (\boldsymbol{\theta}, \boldsymbol{y}_{1:n})$.

If the GLA is good, it can be used as a proposal distribution in a single step IMH algorithm to draw from $\boldsymbol{\omega}$'s posterior distribution. Typically a t_{df} proposal is used instead of a normal with *df* set small enough so that its tails dominate the posterior's, which ensures that the Markov chain is uniformly ergodic (Robert and Casella, 2013, Theorem 7.8). Often $\boldsymbol{\theta}$ constains too many parameters relative to the amount of data available so that $\boldsymbol{\theta}$ is not approximately normal in its marginal posterior, but because of the latent Gaussian assumption $\boldsymbol{y}_{1:n}$ is still approximately normal in it full conditional posterior. In this context, a conditional LA can be constructed to the density $[\boldsymbol{y}_{1:n}|\boldsymbol{\theta}, \boldsymbol{D}_n]$ directly. We call this a *local* conditional LA (LCLA) since the LA approximation is recomputed for each value of $\boldsymbol{\theta}$. Using the LCLA in the context of a Gibbs sampler as the IMH proposal for $\boldsymbol{y}_{1:n}$ while sampling $\boldsymbol{theta}$ in one or more hopefully conjugate form Gibbs steps is attractive when the LCLA is good since the acceptance rate of the Metropolis step will be high. However, constructing any of these LAs typically requires numerical optimization since the data model is non-Gaussian. For the GLA this optimization only needs to be performed once, so as long as it is doable it only represents a fixed cost of running the resulting IMH algorithm. However, the IMHwG algorithm that results from the LCLA requires a numerical optimization to be run each MCMC iteration in order to update the conditional posterior mode and approximate covariance matrix. So while IMWwG using the LCLA may have nice MCMC properties, the computational cost can make it significantly less attractive and in some cases even prohibitive.

We propose using a simple method in order to keep the fixed cost associated with the GLA from becoming a variable cost when apply to a conditional distribution: use the conditional distribution implied by the GLA as the IMH proposal. We call this a global *conditional* LA

(GCLA) since it is an approximation to $\boldsymbol{y}_{1:n}$'s conditional posterior that relies on the GLA. More formally, suppose the GLA is given by the posterior mode $(\boldsymbol{\theta}^*, \boldsymbol{y}^*_{1:n})$ with approximate covariance matrix $\boldsymbol{\Omega}^*$ given by

$$\boldsymbol{\Omega}^* = (-\boldsymbol{H}^*)^{-1} = \begin{bmatrix} \boldsymbol{\Omega}^*_\theta & \boldsymbol{\Omega}^*_{\theta y} \\ \boldsymbol{\Omega}^*_{y\theta} & \boldsymbol{\Omega}^*_y \end{bmatrix}.$$

Then the GCLA to $\boldsymbol{y}_{1:n}$'s conditional posterior is $\boldsymbol{y}_{1:n}|\boldsymbol{\theta}, \boldsymbol{D}_n \overset{a}{\sim} N(\widetilde{\boldsymbol{y}}_{1:n}, \widetilde{\boldsymbol{\Omega}}_y)$, where $\widetilde{\boldsymbol{\theta}}_y = \boldsymbol{y}^*_{1:n} + \boldsymbol{\Omega}^*_{y\theta}(\boldsymbol{\Omega}^*_\theta)^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$ and $\widetilde{\boldsymbol{\Omega}}_y = \boldsymbol{\Omega}^*_y - \boldsymbol{\Omega}^*_{y\theta}(\boldsymbol{\Omega}^*_\theta)^{-1}\boldsymbol{\Omega}^*_{\theta y}$. The GCLA may be worse than the LCLA since the LCLA directly approximates the target conditional posterior, but the computation savings realized from only having to run the optimization algorithm once are typically significant and make it worthwhile. A problem with using either approximation is that by moving from an IMH algorithm to an IMHwG algorithm the mixing and convergence properties of the chain do deteriorate to the extent that $\boldsymbol{\theta}$ and $\boldsymbol{y}_{1:n}$ are dependent in the posterior. In practice this can often be solved by reparameterizing the model, e.g. using a non-centered parameterization (Gelfand et al., 1995; Roberts and Sahu, 1997; Van Dyk and Meng, 2001; Bernardo et al., 2003).

4 Spatially Modeling County Population Estimates

The American Community Survey (ACS) provides 5-year period estimates of county populations as recently as 2014. In 2014 there were 3,142 counties in the United States, including the District of Columbia and counties in Alaska and Hawaii. We use two separate data models in order to illustrate when both the joint and global conditional Laplace approximations work well. The first is a Poisson data model, i.e., $Z_i \sim \text{Poisson}(\lambda_i)$ where $\lambda_i = \exp(Y_i)$. Through visual inspection of a histogram, it was determined that log county populations look approximately normally distributed, so our second data model is $\log Z_i \sim N(\mu_i, \phi^2)$, where $\mu_i = Y_i$.

The process model in both cases is a reduced rank spatial model $Y = \boldsymbol{X\beta} + \boldsymbol{S\delta}$, where $\boldsymbol{X\beta}$ represents the process mean at each county and $\boldsymbol{S\delta}$ implies the spatial correlation across counties. The spatial correlation term consists of a set of r basis functions evaluated at each of the $n = 3,142$ counties, denoted by the $n \times r$ matrix $\boldsymbol{S}$, and a common random effect $\boldsymbol{\delta}$. We assume that $\boldsymbol{\delta}$ is r-dimensional with $r \ll n$ so that the model is reduced rank. Any set of spatial basis functions could be used for $\boldsymbol{S}$ but we use the Moran's I (MI) basis set, described below, but see Hughes and Haran (2013), Porter et al. (2015), Bradley et al. (2015) and references therein for additional discussion. Another possibility is to define a reduced rank model for a point-level spatial process using a basis function expansion and compute the implied set of basis functions for each of the census tracts by integrating the point level basis functions appropriately. See Sections 2.1, 3.1, and 4 of Bradley et al. (2016) for details.

The MI basis functions are defined through the orthogonal projection matrix $\boldsymbol{P_X} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. Let $\boldsymbol{A}$ denote the binary adjacency matrix with $a_{ij} = 1$ if counties i and j are neighbors, $a_{ij} = 0$ otherwise, and $a_{ii} = 0$ along the diagonal, and define the MI operator $\boldsymbol{G}$ as

$$\boldsymbol{G} = (\boldsymbol{I}_n - \boldsymbol{P_X})\boldsymbol{A}(\boldsymbol{I}_n - \boldsymbol{P_X})$$

where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix. The spectral decomposition of $\boldsymbol{G}$ is $\boldsymbol{G} = \boldsymbol{\Phi\Lambda\Phi}'$. To use a reduced rank version of the MI basis functions we truncate the basis function expansion and take $\boldsymbol{S}$ to be the $n \times r$ matrix formed by the r columns of $\boldsymbol{\Phi}$ corresponding to the r largest in magnitude eigenvalues of $\boldsymbol{G}$. The random effect $\boldsymbol{\delta}$ is further modeled as $\boldsymbol{\delta} \sim N(\boldsymbol{0}_r, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$ where $\boldsymbol{0}_r$ denotes an r-dimensional vector of zeroes and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is an unknown covariance matrix that depends on the parameter $\boldsymbol{\theta}$. The covariance matrix of $\boldsymbol{S\delta}$ is then $\boldsymbol{S\Sigma}(\boldsymbol{\theta})\boldsymbol{S}'$.

The process model depends on choices for $\boldsymbol{X}$ and r. In practice r can be chosen using a sensitivity analysis. Since our goal is to illustrate computational methods, we will elide choosing r in a principled way and instead use several values for r in order to illustrate when the parameter space becomes too high dimensional for our method to be advantageous. For simplicity we choose an intercept only model, but all derivations for the model here and in Appendix A assume that $\boldsymbol{X}$ is $n \times p$.

Finally, we consider two extreme parameterizations of $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ — the iid parameterization where $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_r$ and the full parameterization where $\boldsymbol{\Sigma}$ is fully parameterized. In the iid parameterization we assume that $\sigma^2 \sim IG(a_\sigma, b_\sigma)$ in the prior, while in the full parameterization we assume that $\boldsymbol{\Sigma} \sim IW(d, \boldsymbol{E})$. The prior for $\boldsymbol{\beta}$ in all models is $\boldsymbol{\beta} \sim N(\boldsymbol{b}, v^2 \boldsymbol{I}_p)$, and in the lognormal models the prior for ϕ^2 is $\phi^2 \sim IG(a_\phi, b_\phi)$. We assume that the parameters $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$, and when applicable ϕ^2 are mutually independent in the prior. For our examples we assume that $a_\sigma = a_\phi = b_\sigma = b_\phi = 1$, $\boldsymbol{b} = \boldsymbol{0}_p$, $v = 10$, $d = r + 1$ and $\boldsymbol{E} = \boldsymbol{I}_p$. Often a more complicated prior is appropriate on variance or covariance matrix parameters so that the marginal posterior is not sensitive to arbitrary choices in the prior (Gelman, 2006). We use these conditionally conjugate priors because they allow for a fair comparison between MCMC algorithms — most alternatives will complicate alternative Gibbs samplers with extra steps or necessitate Metropolis steps making IMH or IMHwG relatively more attractive.

Between the two possible parameterizations of $\boldsymbol{\Sigma}$ and the two choices for the data model we consider four possible classes of models. Both for implementing PSO algorithms to find the Laplace approximations and for running the IMH and IMHwG algorithms, we reparameterize the variance and covariance matrix parameters so that they have support on an unconstrained space. For σ^2 and ϕ^2 we use the log transformation, and for $\boldsymbol{\Sigma}$ we use the Cholesky factorization of $\boldsymbol{\Sigma}^{-1}$ allowing the diagonal elements to be negative. See Appendix A for details about the posterior distributions, including this transformation and relevant full conditionals.

5 Predicting the 1988 presidential election

Gelman and Hill (2006, Chapter 14) describes a model used to predict state-level opinions about the 1988 presidential candidates from national polls in order to predict the outcome of the election. They model the responses to a series of seven polls conducted by CBS News during the week before the 1988 presidential election. The variable of interest is binary: $Z_i = 1$ if the ith respondent said they supported the Republican candidate and $Z_i = 0$ if they said they supported the Democratic candidate, with undecideds being excluded. Focusing on the last poll, they ultimately estimate a logistic regression model with fixed effects for race (whether the respondent was African American or not), sex, and race×sex, and random effects for four age categories, four education categories, and 16 age×education categories, as well as for the respondent's state of residence (including District of Columbia). The mean of the state random effect distribution is one of five region random effects plus the proportion of the state that voted republican in the last election times a slope coefficient.

The model is somewhat overparameterized since it has age, education, and age×education random effects, so we reduce its size by omitting the age and education random effects. In our model the age×education random effects now represent the random effects for each age×education category rather than an interaction term. The single poll data model is

$$\begin{aligned} &P(Z_i = 1) = \theta_i, \quad \theta_i = \exp(Y_i)/\{1 + \exp(Y_i)\}, \\ &\qquad Y_i = \beta_0 + f_i\beta_f + b_i\beta_b + f_ib_i\beta_{fb} + \alpha_{ae}[ae_i] + \alpha_s[s_i] \quad \text{(single poll data model)}, \end{aligned}$$

where f_i indicates whether respondent i identified as female, b_i indicates whether respondent i identified as African American, ae_i indicates respondent i's age×education category, and s_i indicates respondent i's state of residence. Here we use $\alpha_s[k]$ to denote the kth element of the vector $\boldsymbol{\alpha}_s$, so $\boldsymbol{\alpha}_{ae}$ contains 16 elements, and $\boldsymbol{\alpha}_s$ contains 51 elements (50 states plus the District of Columbia). The single poll process model is

$$\begin{aligned} \alpha_s[k] &\overset{ind}{\sim} N(\alpha_r[r_k] + prev_k\beta_{prev}, \sigma_s^2) \text{ for } k = 1, 2, \ldots, 51, \\ \alpha_{ae}[k] &\overset{iid}{\sim} N(0, \sigma_{ae}^2) \text{ for } k = 1, 2, \ldots, 16, \\ \alpha_r[k] &\overset{iid}{\sim} N(0, \sigma_r^2) \text{ for } k = 1, 2, \ldots, 5 \quad \text{(single poll process model)}, \end{aligned}$$

where $\alpha_r[r_k]$ denotes the region containing state k and $prev_k$ denotes the average vote share for the Republicans in the previous three presidential elections. This model expands the class of models discussed in Section 2 by allowing the mean of $\boldsymbol{\delta}$ to depend on random effects that are further modeled, but adding a level to the hierarchy does not fundamentally change the applicability of the Laplace approximations.

The last poll had 2,015 respondents, but together all seven polls have 11,566 respondents. Using each poll with a minimal number of additional parameters to account for poll to poll variability should increase the quality of the model and result in a posterior with better Laplace approximations since there is so much more data. We analyze a model for all of the polls using the following data model

$$\begin{aligned} P(Z_i = 1) = \theta_i, &\quad \theta_i = \exp(Y_i)/\{1 + \exp(Y_i)\}, \\ Y_i &= \beta_0 + f_i\beta_f + b_i\beta_b + f_ib_i\beta_{fb} + \alpha_{ae}[ae_i] + \alpha_s[s_i] + \alpha_p[p_i] \quad \text{(all polls data model)}, \end{aligned}$$

where p_i denotes which poll respondent i was surveyed in. The process model is given by

$$\begin{aligned} \alpha_s[k] &\overset{ind}{\sim} N(\alpha_r[r_k] + prev_k\beta_{prev}, \sigma_s^2) \text{ for } k = 1, 2, \ldots, 51, \\ \alpha_{ae}[k] &\overset{iid}{\sim} N(0, \sigma_{ae}^2) \text{ for } k = 1, 2, \ldots, 16, \\ \alpha_r[k] &\overset{iid}{\sim} N(0, \sigma_r^2) \text{ for } k = 1, 2, \ldots, 5, \\ \alpha_p[k] &\overset{iid}{\sim} N(0, \sigma_p^2) \text{ for } k = 1, 2, \ldots, 7 \quad \text{(all polls process model)}. \end{aligned}$$

In both models we assume each of the βs have independent $N(0, 1000)$ priors, and each of the σ^2s have $IG(1, 1)$ priors. Including the random effects the single poll model contains

80 parameters while the all polls model contains 89 parameters. Writing down the log posteriors and deriving the Hessians is straightforward but tedious for these models, so we omit these steps, though note that for both the PSO and IMH algorithms the variances should be transformed to the log scale. For MCMC algorithms with multiple Gibbs steps, in single poll model of Section 5, we draw $(\beta_0, \beta_f, \beta_b, \boldsymbol{\alpha}_s, \boldsymbol{\alpha}_{ae})$ in the Metropolis step, while in the all polls model we draw all of those parameters and additionally $\boldsymbol{\alpha}_p$ in the Metropolis step. Then for both models there are two additional Gibbs steps — one where each random effect variance is drawn, and one where $(\beta_{prev}, \boldsymbol{\alpha}_r)$ is drawn. These full conditionals are straightforward to derive, so we do not reproduce them here. [SHOULD WE ADD THESE DETAILS INTO AN APPENDIX?]

6 Discussion

References

Berliner, L. M. (1996). "Hierarchical Bayesian time series models." In *Maximum entropy and Bayesian methods*, 15–22. Springer.

Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M. (2003). "Non-centered Parameterisations for Hierarchical Models and Data Augmentation." In *Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting*, 307–326. London: Oxford University Press.

Bradley, J. R., Holan, S. H., and Wikle, C. K. (2015). "Multivariate spatio-temporal models for high-dimensional areal data with application to longitudinal employer-household dynamics." *Annals of Applied Statistics*, 9, 4, 1761–1791.

Bradley, J. R., Wikle, C. K., and Holan, S. H. (2016). "Regionalization of multiscale spatial processes by using a criterion for spatial aggregation error." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1995). "Efficient Parametrisations for Normal Linear Mixed Models." *Biometrika*, 82, 3, 479–488.

Gelman, A. (2006). "Prior Distributions for Variance Parameters in Hierarchical Models (Comment on Article by Browne and Draper)." *Bayesian analysis*, 1, 3, 515–534.

Gelman, A. and Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Hughes, J. and Haran, M. (2013). "Dimension reduction and alleviation of confounding for spatial generalized linear mixed models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75, 1, 139–159.

Porter, A. T., Holan, S. H., and Wikle, C. K. (2015). "Bayesian semiparametric hierarchical empirical likelihood spatial models." *Journal of Statistical Planning and Inference*, 165, 78–90.

Robert, C. and Casella, G. (2013). *Monte Carlo statistical methods*. 2nd ed. Springer Science & Business Media.

Roberts, G. O. and Sahu, S. K. (1997). "Updating Schemes, Correlation Structure, Blocking and Parameterization for the Gibbs Sampler." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59, 2, 291–317.

Schervish, M. J. (1997). *Theory of statistics*. Springer Science & Business Media.

Stroup, W. W. (2012). *Generalized linear mixed models: modern concepts, methods and applications*. CRC press.

Tanner, M. A. and Wong, W. H. (1987). "The calculation of posterior distributions by data augmentation." *Journal of the American statistical Association*, 82, 398, 528–540.

Van Dyk, D. and Meng, X.-L. (2001). "The Art of Data Augmentation." *Journal of Computational and Graphical Statistics*, 10, 1, 1–50.

Wikle, C. K. et al. (2003). "Hierarchical Models in Environmental Science." *International Statistical Review*, 71, 2, 181–199.