Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Bayesian Hierarchical Models With Conjugate Full-Conditional Distributions for Dependent Data From the Natural Exponential Family

Jonathan R. Bradley , Scott H. Holan & Christopher K. Wikle

To cite this article: Jonathan R. Bradley , Scott H. Holan & Christopher K. Wikle (2020) Bayesian Hierarchical Models With Conjugate Full-Conditional Distributions for Dependent Data From the Natural Exponential Family, Journal of the American Statistical Association, 115:532, 2037-2052, DOI: 10.1080/01621459.2019.1677471

To link to this article: https://doi.org/10.1080/01621459.2019.1677471

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2020, VOL. 115, NO. 532, 2037–2052: Theory and Methods
https://doi.org/10.1080/01621459.2019.1677471

Bayesian Hierarchical Models With Conjugate Full-Conditional Distributions for Dependent Data From the Natural Exponential Family

Jonathan R. Bradley[a], Scott H. Holan[b,c], and Christopher K. Wikle[b]

[a]Department of Statistics, Florida State University, Tallahassee, FL; [b]Department of Statistics, University of Missouri, Columbia, MO; [c]U.S. Census Bureau, Washington, DC

ABSTRACT

We introduce a Bayesian approach for analyzing (possibly) high-dimensional dependent data that are distributed according to a member from the natural exponential family of distributions. This problem requires extensive methodological advancements, as jointly modeling high-dimensional dependent data leads to the so-called "big *n* problem." The computational complexity of the "big *n* problem" is further exacerbated when allowing for non-Gaussian data models, as is the case here. Thus, we develop new computationally efficient distribution theory for this setting. In particular, we introduce the "conjugate multivariate distribution," which is motivated by the Diaconis and Ylvisaker distribution. Furthermore, we provide substantial theoretical and methodological development including: results regarding conditional distributions, an asymptotic relationship with the multivariate normal distribution, conjugate prior distributions, and full-conditional distributions for a Gibbs sampler. To demonstrate the wide-applicability of the proposed methodology, we provide two simulation studies and three applications based on an epidemiology dataset, a federal statistics dataset, and an environmental dataset, respectively. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received July 2016
Accepted September 2019

KEYWORDS
Bayesian hierarchical model; Big data; Exponential family; Gibbs sampler; Markov chain Monte Carlo; Non-Gaussian

1. Introduction

The multivariate normal distribution has become a fundamental tool for statisticians, as it provides a way to incorporate dependence for Gaussian and non-Gaussian data alike. Notice that many statistical models are defined hierarchically, where the joint distribution of the data, latent processes, and unknown parameters are written as the product of a data model, a latent Gaussian process model, and a parameter model (see, e.g., Cressie and Wikle 2011; Banerjee, Carlin, and Gelfand 2015, among others). Jointly modeling a member from the exponential family may be seen as straightforward to some. That is, one can simply define the data model to be the appropriate member of the exponential family and define latent Gaussian processes using the hierarchical modeling framework. Models of this form are often referred to as *latent Gaussian process* (LGP) models (see Diggle, Tawn, and Moyeed 1998; Rue, Martino, and Chopin 2009; Cressie and Wikle 2011, secs. 4.1.2 and 7.1.5; Holan and Wikle 2016, among others).

In the Bayesian context, LGPs can be nontrivial to implement using standard Markov chain Monte Carlo (MCMC) procedures when the dataset is high-dimensional. This is primarily because big data can lead to big parameter spaces, which allows parameters to be highly correlated. This in turn, creates a challenge for defining useful proposal distributions, tuning these proposal distributions, and assessing convergence of the Markov chain (see, e.g., Rue, Martino, and Chopin 2009; Bradley, Holan, and Wikle 2018 for a discussion on convergence issues of MCMC algorithms for LGPs). In this article, our primary goal is to introduce new distribution theory that facilitates Bayesian inference of dependent non-Gaussian data. In particular, we introduce a multivariate distribution that leads to conjugate forms of the full conditional distributions within a Gibbs sampler.

We provide a multivariate extension of the class of distributions introduced the seminal paper by Diaconis and Ylvisaker (1979), who developed the conjugate prior for distributions from the natural exponential family (EF), which leads to the well-known Poisson/gamma, binomial/beta, negative binomial/beta, and gamma/inverse-gamma hierarchical models. In this article, we develop a multivariate version of this distribution, which we call the *conjugate multivariate* (CM) distribution. Similar to the special cases that emerged from Diaconis and Ylvisaker (1979) and Chen and Ibrahim (2003) we obtain Poisson/multivariate log gamma (MLG), binomial/multivariate logit beta, negative binomial/multivariate logit beta, and gamma/multivariate negative-inverse-gamma hierarchical models. The hierarchical model that specifies the data model to be from the natural exponential family, and the latent process to be a CM distribution is referred to as a latent CM process (LCM) model. The LCM model constitutes a more general paradigm for modeling dependent data than LGPs, since the LGP is a special case of the LCM model. An important motivating feature of this more general framework is that the LCM model incorporates dependency and results in full-conditional distributions (within a Gibbs sampler) that are

CONTACT Jonathan R. Bradley bradley@stat.fsu.edu Department of Statistics, Florida State University, 117 N. Woodward Ave, Tallahassee, FL 32306.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

easy to simulate from. This allows one to avoid computationally inefficient and subjective tuning methods.

An immediate issue that arises with the introduction of the LCM model is the need to define flexible prior distributions. One goal of this article is to describe the *fully conjugate* Bayesian hierarchical model that has a data model that belongs to the natural exponential family. By "fully conjugate" we mean that each full conditional distribution, within a Gibbs sampler, falls in the same class of distributions of the associated process or parameter models. To derive a fully conjugate statistical model, we introduce the LCM analogue to the prior distributions used in Daniels and Pourahmadi (2002), Chen and Dunson (2003), and Pourahmadi, Daniels, and Park (2007) for covariance parameters. Additionally, extensions of the standard inverse-gamma priors for variances of a normal random variable (Gelman 2006) are discussed in context of the LCM.

There is an added benefit of the CM distribution besides providing conjugacy in the non-Gaussian dependent data setting. Namely, LGPs are not necessarily realistic for *every* dataset. For example, De Oliveira (2013) shows that there are parametric limitations to the LGP paradigm for count-valued data (e.g., when spatial overdispersion is small). We support this claim by showing that if certain hyperparameters (defined in Section 2) of the CM distribution are "large" then the corresponding CM distribution gives a very good approximation to a Gaussian distribution. This indicates that if the data suggests small values of these hyperparameters, then the CM distribution should be used in place of the multivariate normal distribution.

Reduced rank methods are extremely prevalent in the more general "dependent data" setting. For example, reduced rank assumptions are crucial for principle component analysis, which has become an established technique in multivariate data analysis (see, e.g., Jolliffe 2002; Cox 2005; Everitt and Hothorn 2011, among others). Additionally, reduced rank models have been used to great effect within spatial and spatio-temporal settings to obtain precise predictions in a computationally efficient manner (see, e.g., Wikle and Cressie 1999; Cressie and Johannesson 2006; Shi and Cressie 2007; Banerjee et al. 2008; Cressie and Johannesson 2008; Finley et al. 2009; Cressie, Shi, and Kang 2010a, 2010b; Kang and Cressie 2011; Katzfuss and Cressie 2011, 2012; Bradley, Holan, and Wikle 2015). Thus, an additional motivating feature of the LCM model is that it can easily be cast within the reduced rank modeling framework to obtain further computational gains. The ability to specify a reduced rank LCM does not imply that the LCM can handle *all* types of "big data" problems. One type of big data problem that we do not consider is the "big p" problem (Hastie, Tibshirani, and Friedman 2009; Matloff 2016). Here, our focus is on difficulties with incorporating dependence when n is large. Specifically, inverses of $n \times n$ matrices often manifest in dependent data settings (see, e.g., Sun and Li 2012, among others). Our incorporation of reduced rank modeling allows one to avoid order n^3 computations needed for matrix inversion, and allows one to avoid storage of large $n \times n$ matrices.

This computationally efficient fully conjugate distribution theory could have an important impact on a number of different communities within and outside statistics. High-dimensional non-Gaussian data are pervasive in official statistics (see, e.g., Bradley, Holan, and Wikle 2018), ecology (see, e.g., Hooten, Larsen, and Wikle 2003; Wu, Holan, and Wikle 2013, among others), climatology (see, e.g., Wikle and Anderson 2003), atmospheric sciences (see, e.g., Sengupta et al. 2012), statistical genetics (see, e.g., Lange et al. 2014, and the references therein), neuroscience (see, e.g., Zhang, Guindani, and Vannucci 2015; Castruccio, Ombao, and Genton 2016), and many other domains. The size of modern datasets is becoming more and more high-dimensional, and the aforementioned computational difficulties with LGPs suggest that there is a growing need to develop methods that are straightforward to implement (see, e.g., Bradley, Cressie, and Shi 2016, for a discussion). Hence, the methodology presented here offers an exciting avenue that makes new applied research for modeling dependent non-Gaussian data practical for modern big datasets.

The LCM model is a type of hierarchical generalized linear model (HGLM) from Lee and Nelder (1996). However, the current HGLM literature specifies an LGP for the dependent data setting (Lee and Nelder 2000, 2001). Additionally, there are other alternatives to a Gibbs sampler with Metropolis–Hastings updates; in particular, integrated nested Laplace approximations (INLA) (Rue, Martino, and Chopin 2009) and Hamiltonian MCMC have proven to be useful tools in the literature. These approaches can easily be applied to our new proposed distribution theory; however, the need to adapt INLA and Hamiltonian MCMC (Neal 2011) to the LCM is not immediately necessary since the full conditional distributions are straightforward to simulate from in this setting.

For Poisson counts there are a number of choices besides the LGP strategy available to incorporate dependence (see, e.g., Lee and Nelder 1974; Kotz, Balakrishnan, and Johnson 2000; Demirhan and Hamurkaroglu 2011, among others). For example, Wolpert and Ickstadt (1998), introduced a spatial convolution of gamma random variables, and provide a data augmentation scheme for Gibbs sampling that produces spatial predictions. Similarly, Frühwirth-Schnatter and Wagner (2006) have an approximate Bayesian method for Poisson counts with latent Gaussian random variables. The recently proposed multivariate log-gamma distribution of Bradley, Holan, and Wikle (2018) results in a special case of our modeling approach when the data model is Poisson, and the latent processes are distributed according to a type of CM distribution. Additionally, in more specific settings (e.g., Pareto data spatio-temporal data), conjugate distribution theory has been developed (Nieto-Barajas and Huerta 2017; Hu and Bradley 2018).

The remainder of this article is organized as follows. In Section 2, we introduce the conjugate multivariate distribution and provide the necessary technical development for fully Bayesian inference of dependent data from the natural exponential family. Specifically, we define the CM distribution, give the specification of the LCM model, discuss important methodological properties, introduce additional hyperpriors, and derive the full conditional distributions for a Gibbs sampler. Then, in Section 3, we provide a simulated example and an in-depth simulation study to show the performance of the LCM model compared to LGPs. Several illustrations from different subject matter areas are also presented in Section 3, which is done in an effort to demonstrate the wide-applicability of the LCM. Specifically, we provide an example analyzing an epidemiology dataset, a federal statistics dataset, and an environmental dataset. Finally, Section 4

contains discussion. For convenience of exposition, proofs of the technical results, Matlab and R code, and instructions on implementation are given in the Supplemental Appendix.

2. Distribution Theory for Dependent Data From the Natural Exponential Family

In this section, we propose methodology for Bayesian analysis of non-Gaussian dependent data from the natural exponential family. In Section 2.1, we review and develop the univariate distribution introduced in Diaconis and Ylvisaker (1979). Then, in Section 2.2, this univariate distribution is used as the rudimentary quantity to develop the CM distribution. This new multivariate distribution theory is incorporated within a Bayesian hierarchical model (i.e., the aforementioned LCM model) in Section 2.3, and the corresponding methodological properties are discussed in Section 2.4. A collapsed Gibbs sampler is derived in Section 2.5, and additional properties associated with the Gibbs sampler are discussed in Section 2.6. Finally, prior distributions on remaining parameters are discussed in Sections 2.7 and 2.8.

2.1. The Diaconis and Ylvisaker Conjugate Distribution

Suppose Z is distributed according to the natural exponential family (Diaconis and Ylvisaker 1979; Lehmann and Casella 1998), then

$$f(Z|Y) = \exp\{ZY - b\psi(Y) + c(Z)\};\ Z \in \mathcal{Z}, Y \in \mathcal{Y}, \quad (1)$$

where f will be used to denote a generic probability density function/probability mass function (pdf/pmf), $Z \in \mathcal{Z}$, $\mathcal{Z}$ is the support of Z, $\mathcal{Y}$ is the support of Y, b is possibly unknown, and both $\psi(\cdot)$ and $c(\cdot)$ are known real-valued functions. The function $b\psi(Y)$ is often called the log partition function (Lehmann and Casella 1998). It will be useful for us to discuss $\psi(Y)$ and not $b\psi(Y)$; hence, we refer to $\psi(Y)$ as the "unit log partition function" because its coefficient is one and not b. Let $\text{EF}(Y;\ \psi)$ denote a shorthand for the pdf/pmf in (1). It follows from Diaconis and Ylvisaker (1979) that the conjugate prior distribution for Y is given by,

$$f(Y|\alpha,\kappa) = K(\alpha,\kappa)\exp\{\alpha Y - \kappa\psi(Y)\};\ Y \in \mathcal{Y}, \frac{\alpha}{\kappa} \in \mathcal{Z}, \kappa > 0, \quad (2)$$

where $K(\alpha,\kappa)$ is a normalizing constant. Let $\text{DY}(\alpha,\kappa;\ \psi)$ denote a shorthand for the pdf in (2). Here "DY" stands for "Diaconis–Ylvisaker," and we will refer to Y as either a Diaconis–Ylvisaker random variable or a DY random variable. Diaconis and Ylvisaker (1979) proved that the pdf in (2) is proper (i.e., yields a probability measure). We also call α and κ "DY parameters."

Multiplying both sides of (2) by $\exp(tY)$ and integrating, gives the moment generating function

$$E[\exp(tY)|\alpha,\kappa] = \frac{K(\alpha,\kappa)}{K(\alpha+t,\kappa)}, \quad (3)$$

which exists provided that $(\alpha + t)/\kappa \in \mathcal{Z}$, $\kappa > 0$, and the corresponding values of $K(\alpha + t,\kappa)$ and $K(\alpha,\kappa)$ are strictly positive and finite. This gives us that the mean and variance of Y is

$$E(Y|\alpha,\kappa) = K(\alpha,\kappa)K^{(1)}(\alpha,\kappa) \quad (4)$$

$$\text{var}(Y|\alpha,\kappa) = K(\alpha,\kappa)K^{(2)}(\alpha,\kappa) - K(\alpha,\kappa)^2K^{(1)}(\alpha,\kappa)^2, \quad (5)$$

assuming that the moment generating function exists at $t = 0$, where $K^{(1)}(\alpha,\kappa) \equiv \left[\frac{d}{dt}\frac{1}{K(\alpha+t,\kappa)}\right]_{t=0}$ and $K^{(2)}(\alpha,\kappa) \equiv \left[\frac{d^2}{dt^2}\frac{1}{K(\alpha+t,\kappa)}\right]_{t=0}$.

Finally, it is immediate from (1) and (2) that

$$Y|Z,\alpha,\kappa \sim \text{DY}(\alpha + Z, \kappa + b;\ \psi). \quad (6)$$

This conjugacy motivates the development of a multivariate version of the DY random variable to model dependent non-Gaussian data from the natural exponential family. Thus, in this section, we define a conjugate multivariate distribution and develop a distribution theory that we find useful for fully Bayesian analysis in the dependent non-Gaussian (natural exponential family) data setting.

2.2. The Conjugate Multivariate (CM) Distribution

Bradley, Holan, and Wikle (2018) use a linear combination of independent log-gamma random variables to build their multivariate log-gamma distribution. In a similar manner, we take linear combinations of DY random variables to generate a conjugate version of the DY distribution. Specifically, let the n-dimensional random vector $\mathbf{w} = (w_1,\ldots,w_n)'$ consist of n mutually independent DY random variables such that $w_i \sim \text{DY}(\alpha_i,\kappa_i;\ \psi)$ for $i = 1,\ldots,n$. Then, define $\mathbf{Y} \equiv (Y_1,\ldots,Y_n)'$ such that

$$\mathbf{Y} = \boldsymbol{\mu} + \mathbf{V}\mathbf{w}, \quad (7)$$

where $\mathbf{Y} \in \mathcal{M}^n$, the matrix $\mathbf{V} \in \mathbb{R}^n \times \mathbb{R}^n$, and $\boldsymbol{\mu} \in \mathbb{R}^n$. The space $\mathcal{M}^n$ is not necessarily equal to $\mathcal{Y}^n \equiv \{\mathbf{Y} = (Y_1,\ldots,Y_n)' : Y_i \in \mathcal{Y}, i = 1,\ldots,n\}$; for example, if $\mathcal{Y}$ is strictly positive, we obtain a $\mathbf{Y}$ that can have negative components since $\mathbf{V} \in \mathbb{R}^n \times \mathbb{R}^n$. Call $\mathbf{Y}$ in (7) a *conjugate multivariate* (CM) random vector. A special case of the CM random vector is the multivariate normal random vector. To see this, let $\alpha_i \equiv 0$, $\kappa_i \equiv 1/2$, and $\psi(Y) = Y^2$ for $Y \in \mathbb{R}$. Then, it follows that (7) is a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{V}\mathbf{V}'$, since the elements of $\mathbf{w}$ consist of iid standard normal random variables. Additionally, the aforementioned MLG distribution can be written as a CM distribution when $\alpha > 0$, $\kappa > 0$, and $\psi(Y) = \exp(Y)$.

To use the CM distribution in a Bayesian context, we require its pdf, which is formally stated below.

Theorem 1. Let $\mathbf{Y} = \boldsymbol{\mu} + \mathbf{V}\mathbf{w}$, where $\mathbf{Y} = (Y_1,\ldots,Y_n)'$, $\boldsymbol{\mu} \in \mathbb{R}^n$, the $n \times n$ real valued matrix $\mathbf{V}$ is invertible, and the n-dimensional random vector $\mathbf{w} = (w_1,\ldots,w_n)'$ consists of n mutually independent DY random variables such that $w_i \sim \text{DY}(\alpha_i,\kappa_i;\ \psi)$ for $i = 1,\ldots,n$.

(i) Then $\mathbf{Y}$ has the following pdf:

$$f(\mathbf{Y}|\boldsymbol{\mu},\mathbf{V},\boldsymbol{\alpha},\boldsymbol{\kappa}) = \det(\mathbf{V}^{-1})\left\{\prod_{i=1}^{n} K(\kappa_i,\alpha_i)\right\}\exp\left[\boldsymbol{\alpha}'\mathbf{V}^{-1}(\mathbf{Y}-\boldsymbol{\mu}) - \boldsymbol{\kappa}'\psi\left\{\mathbf{V}^{-1}(\mathbf{Y}-\boldsymbol{\mu})\right\}\right] I(\mathbf{Y}\in\mathcal{M}^n), \quad (8)$$

where $I(\cdot)$ is the indicator function, the jth element of $\psi\left\{\mathbf{V}^{-1}(\mathbf{Y}-\boldsymbol{\mu})\right\}$ contains ψ evaluated at the jth element of the n-dimensional vector $\mathbf{V}^{-1}(\mathbf{Y}-\boldsymbol{\mu})$, "det" denotes the determinant function, $\boldsymbol{\alpha} \equiv (\alpha_1,\ldots,\alpha_n)'$, and $\boldsymbol{\kappa} \equiv (\kappa_1,\ldots,\kappa_n)'$.

(ii) The mean and variance of $\mathbf{Y}$ is given by,

$$E(\mathbf{Y}|\boldsymbol{\alpha},\boldsymbol{\kappa}) = \boldsymbol{\mu} + \mathbf{V}\mathbf{k}(\boldsymbol{\alpha},\boldsymbol{\kappa})$$
$$\text{cov}(\mathbf{Y}|\boldsymbol{\alpha},\boldsymbol{\kappa}) = \mathbf{V}\mathbf{K}(\boldsymbol{\alpha},\boldsymbol{\kappa})\mathbf{V}', \quad (9)$$

where, the n-dimensional real-valued vector

$$\mathbf{k}(\boldsymbol{\alpha},\boldsymbol{\kappa}) = \Big(K(\alpha_1,\kappa_1)K^{(1)}(\alpha_1,\kappa_1),\ldots, \times K(\alpha_n,\kappa_n)K^{(1)}(\alpha_n,\kappa_n)\Big)',$$

and the $n \times n$ diagonal matrix $\mathbf{K}(\boldsymbol{\alpha},\boldsymbol{\kappa}) \equiv \text{diag}\left\{K(\alpha_i,\kappa_i)K^{(2)}(\alpha_i,\kappa_i) - K(\alpha_i,\kappa_i)^2K^{(1)}(\alpha_i,\kappa_i)^2\right\}$.

The proof of Theorem 1(i) can be found in the Supplemental Appendix. In general, we let $\text{CM}(\boldsymbol{\mu},\mathbf{V},\boldsymbol{\alpha},\boldsymbol{\kappa};\psi)$ denote the pdf in (8). Theorem 1(ii) follows immediately from Equations (4) and (5), and thus, Theorem 1(ii) is stated without proof.

When comparing (1), (2), and (8) we see that the univariate natural exponential family, the DY pdf, and the CM pdf share a basic structure. Specifically, all three distributions have an exponential term and an "exponential of $-\psi$ term." This pattern is the main reason why conjugacy exists between the distributions from the natural exponential family and the DY distribution, which we take advantage of in subsequent sections. Also, Theorem 1(ii) shows that if we restrict $\mathbf{V}$ (or equivalently $\mathbf{V}^{-1}$) to be a lower unit triangle matrix, then the expression of the covariance matrix of $\mathbf{Y}$ in (9) is a type of LDL decomposition (Ravishanker and Dey 2002). Hence, in subsequent sections we assume that $\mathbf{V}$ is lower unit triangular.

Bayesian inference not only requires the pdf of $\mathbf{Y}$, but also requires simulating from conditional distributions of $\mathbf{Y}$.

Theorem 2. Let $\mathbf{Y} \sim \text{CM}(\boldsymbol{\mu},\mathbf{V},\boldsymbol{\alpha},\boldsymbol{\kappa};\psi)$, and let $\mathbf{Y} = (Y_1,\ldots,Y_n)' = (\mathbf{Y}_1',\mathbf{Y}_2')'$, so that $\mathbf{Y}_1$ is r-dimensional and $\mathbf{Y}_2$ is $(n-r)$-dimensional. In a similar manner, partition $\mathbf{V}^{-1} = [\mathbf{H}\ \mathbf{B}]$ into an $n \times r$ matrix $\mathbf{H}$ and an $n \times (n-r)$ matrix $\mathbf{B}$. Also let $\boldsymbol{\mu}^* = \mathbf{V}^{-1}\boldsymbol{\mu} - \mathbf{B}\mathbf{d}$ for $\mathbf{d} \in \mathbb{R}^{n-r}$. Then, the conditional distribution $\mathbf{Y}_1|\mathbf{Y}_2 = \mathbf{d},\boldsymbol{\mu}^*,\mathbf{H},\boldsymbol{\alpha},\boldsymbol{\kappa}$ is given by

$$f(\mathbf{Y}_1|\mathbf{Y}_2 = \mathbf{d},\boldsymbol{\mu}^*,\mathbf{H},\boldsymbol{\alpha},\boldsymbol{\kappa}) = M\ \exp\left\{\boldsymbol{\alpha}'\mathbf{H}\mathbf{Y}_1 - \boldsymbol{\alpha}'\boldsymbol{\mu}^* - \boldsymbol{\kappa}'\psi(\mathbf{H}\mathbf{Y}_1 - \boldsymbol{\mu}^*)\right\} \times I\{(\mathbf{Y}_1',\mathbf{d}')' \in \mathcal{M}^n\}, \quad (10)$$

where M is a strictly positive and finite normalizing constant. Let $\text{CM}_c(\boldsymbol{\mu}^*,\mathbf{H},\boldsymbol{\alpha},\boldsymbol{\kappa};\psi)$ be a shorthand for the pdf in (10), where the subscript "c" represents the word "conditional."

In Supplemental Appendix A, we describe technical results related to simulating from the conditional CM distribution.

In this article, we consider CM distributions that are implied by the unit log partition function of the data model including: the gamma data model, binomial data model, negative binomial data model, the Poisson data model, and the normal data model (see Tables 1 and 2). In the univariate case, each of these special cases lead to well-known hierarchical models (i.e., gamma/inverse-gamma, (negative) binomial/beta, Poisson/log-gamma, and normal/normal models) (Diaconis and Ylvisaker 1979). To delineate from the univariate setting, we shall refer to $\text{CM}(\boldsymbol{\mu},\mathbf{V},\boldsymbol{\alpha},\boldsymbol{\kappa};\psi_j)$ for $j = 1,\ldots,4$ (see Table 1 for the definitions of ψ_1, ψ_2, ψ_3, and ψ_4) as the multivariate negative-inverse-gamma distribution, multivariate logit-beta distribution, the multivariate log-gamma, and the multivariate normal distribution, respectively.

These choices of the CM distribution are themselves general. For example, when $\boldsymbol{\alpha} = \mathbf{J}_{n,1}$, we obtain an exponential/multivariate negative-inverse-gamma model. Similarly, the binomial/multivariate logit-beta model has a Bernoulli/multivariate logit-beta model as a special case, which occurs when the number of Bernoulli trials that define the binomial distribution is equal to one. Likewise, when the number of successful Bernoulli trials is equal to one, the negative binomial/multivariate logit-beta model reduces to a geometric/multivariate logit-beta specification. This creates opportunity for analyzing many different types of dependent data.

Table 1. Special cases: we list the form of the CM distribution by ψ_j for $j = 1,\ldots,4$.

Unit log partition function (i.e., ψ)	CM distribution (i.e., $f(\mathbf{Y}\|\boldsymbol{\mu},\mathbf{V},\boldsymbol{\alpha},\boldsymbol{\kappa})$)
$\psi_1(Y) = \log\left(-\frac{1}{Y}\right)$	$\det(\mathbf{V}^{-1})\left\{\prod_{i=1}^{n}\frac{\alpha_i^{\kappa_i+1}}{\Gamma(\kappa_i+1)}\right\}\exp\left[\boldsymbol{\alpha}'\mathbf{s} - \boldsymbol{\kappa}'\log\left(-\mathbf{s}^{(-1)}\right)\right] I(-\mathbf{s}\in\mathbb{R}_n^+)$
$\psi_2(Y) = \log(1+\exp(Y))$	$\det(\mathbf{V}^{-1})\left\{\prod_{i=1}^{n}\frac{\Gamma(\kappa_i)}{\Gamma(\alpha_i)\Gamma(\kappa_i-\alpha_i)}\right\}\exp\left[\boldsymbol{\alpha}'\mathbf{V}^{-1}(\mathbf{Y}-\boldsymbol{\mu}) - \boldsymbol{\kappa}'\log\left[\mathbf{J}_{n,1}+\exp\left\{\mathbf{V}^{-1}(\mathbf{Y}-\boldsymbol{\mu})\right\}\right]\right] I(\mathbf{Y}\in\mathbb{R}^n)$
$\psi_3(Y) = \exp(Y)$	$\det(\mathbf{V}^{-1})\left\{\prod_{i=1}^{n}\frac{\kappa_i^{\alpha_i}}{\Gamma(\alpha_i)}\right\}\exp\left[\boldsymbol{\alpha}'\mathbf{V}^{-1}(\mathbf{Y}-\boldsymbol{\mu}) - \boldsymbol{\kappa}'\exp\left\{\mathbf{V}^{-1}(\mathbf{Y}-\boldsymbol{\mu})\right\}\right] I(\mathbf{Y}\in\mathbb{R}^n)$
$\psi_4(Y) = Y^2$	$\det(\mathbf{V}^{-1})\left\{\prod_{i=1}^{n}\left(\frac{\kappa_i}{\pi}\right)^{1/2}\right\}\exp\left\{-(\mathbf{Y}-\boldsymbol{\mu}-\boldsymbol{\gamma})'\mathbf{V}^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{V}^{-1\prime}(\mathbf{Y}-\boldsymbol{\mu}-\boldsymbol{\gamma})/2\right\} I(\mathbf{Y}\in\mathbb{R}^n)$

NOTE: The first column has the unit log partition function ψ_j, and the second column has the form of the CM distribution with generic $\mathbf{V}^{-1}\in\mathbb{R}^n\times\mathbb{R}^n$. Let $\mathbf{J}_{m,g}$ denote a $m \times g$ matrix of ones, $\mathbf{s} = (s_1,\ldots,s_n)' \equiv \mathbf{V}^{-1}(\mathbf{Y}-\boldsymbol{\mu})$, $\boldsymbol{\gamma} = \left(\frac{\alpha_1}{2\kappa_1},\ldots,\frac{\alpha_n}{2\kappa_n}\right)'$, $\mathbf{s}^{(-1)} = (1/s_1,\ldots,1/s_n)'$, and $\boldsymbol{\Sigma} \equiv \text{diag}\left(\frac{1}{2\kappa_i} : i = 1,\ldots,n\right)$.

Table 2. A cross-tabulation of a hold-out dataset with 216 observations and the corresponding rounded predicted values (i.e., the posterior mean estimated from the Gibbs sampler).

		Hold-out data value							
		0	1	2	3	4	5	6	7
Rounded Poisson LCM predictions	0	**18**	0	0	0	0	0	0	0
	1	0	**75**	14	0	0	0	0	0
	2	0	14	**43**	15	1	0	0	0
	3	0	0	3	**13**	8	1	0	0
	4	0	0	0	0	**9**	0	0	0
	5	0	0	0	0	4	**3**	0	0
	6	0	0	0	0	0	1	**2**	0
	7	0	0	0	0	0	0	1	**0**
	8	0	0	0	0	0	0	1	0
	9	0	0	0	0	0	0	0	1

NOTE: These predictions are rounded to the nearest integer, since the hold-out dataset is known to be integer-valued. The bold values indicate that the rounded predictions and the hold-out data exactly agree.

2.3. The LCM Model

The LCM model is proportional to the product of the following conditional and marginal distributions:

$$\begin{aligned}
&\text{Data Model}: Z_i|\boldsymbol{\beta},\boldsymbol{\eta},\xi_i \overset{\text{ind}}{\sim} \text{EF}\left(\mathbf{x}_i'\boldsymbol{\beta}+\boldsymbol{\phi}_i'\boldsymbol{\eta}+\xi_i;\ \psi_j\right);\\
&i=1,\ldots,n, j=1,\ldots,4\\
&\text{Process Model 1}: \boldsymbol{\eta}|\mathbf{V},\alpha_\eta,\kappa_\eta \sim \text{CM}\left(\mathbf{0}_{r,1},\mathbf{V},\boldsymbol{\alpha}_\eta,\boldsymbol{\kappa}_\eta;\ \psi_k\right);\\
&\text{Process Model 2}: \boldsymbol{\xi}|\alpha_\xi,\kappa_\xi \sim \text{CM}\left(\mathbf{0}_{n,1},\mathbf{V}_\xi,\boldsymbol{\alpha}_\xi,\boldsymbol{\kappa}_\xi;\ \psi_k\right);\\
&\text{Parameter Model 1}: b|\alpha_b,\kappa_b \sim \text{CM}\left(0,1,\alpha_b,\kappa_b;\ \psi_k\right) I(b>0)\\
&\text{Parameter Model 2}: \boldsymbol{\beta}|\alpha_\beta,\kappa_\beta \sim \text{CM}\left(\mathbf{0}_{p,1},\mathbf{V}_\beta,\boldsymbol{\alpha}_\beta,\boldsymbol{\kappa}_\beta;\ \psi_k\right);\\
&k=1,\ldots,4, \qquad (11)
\end{aligned}$$

where ψ_j and ψ_k (for $j,k=1,\ldots,4$) are defined in Table 1 and the elements of n-dimensional vector $\mathbf{Z}\equiv(Z_1,\ldots,Z_n)'$ represent data that can be reasonably modeled using a member from the natural exponential family. Additionally for each i, $\mathbf{x}_i$ is a known p-dimensional vector of covariates, $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_p)'\in\mathbb{R}^p$ is an unknown vector interpreted as fixed effects, $\boldsymbol{\phi}_i$ is a known r-dimensional real-valued vector (see Section 3.5 for an example), and the r-dimensional vector $\boldsymbol{\eta}=(\eta_1,\ldots,\eta_r)'$ and n-dimensional vector $\boldsymbol{\xi}\equiv(\xi_1,\ldots,\xi_n)'$ are interpreted as real-valued random effects. We have not yet provided specifications of the hyperparameters and variance parameters: $\boldsymbol{\alpha}_\beta=(\alpha_{\beta,1},\ldots,\alpha_{\beta,p})'$, $\boldsymbol{\alpha}_\eta=(\alpha_{\eta,1},\ldots,\alpha_{\eta,r})'$, $\boldsymbol{\alpha}_\xi=(\alpha_{\xi,1},\ldots,\alpha_{\xi,n})'$, $\boldsymbol{\kappa}_\beta=(\kappa_{\beta,1},\ldots,\kappa_{\beta,p})'$, $\boldsymbol{\kappa}_\eta=(\kappa_{\eta,1},\ldots,\kappa_{\eta,r})'$, $\boldsymbol{\kappa}_\xi=(\kappa_{\xi,1},\ldots,\kappa_{\xi,n})'$, $\mathbf{V}_\beta\in\mathbb{R}^p\times\mathbb{R}^p$, $\mathbf{V}\in\mathbb{R}^r\times\mathbb{R}^r$, and $\mathbf{V}_\xi\in\mathbb{R}^n\times\mathbb{R}^n$, where $\alpha_{\beta,i}/\kappa_{\beta,i}\in\mathcal{Y}$, $\alpha_{\eta,j}/\kappa_{\eta,j}\in\mathcal{Y}$, $\alpha_{\xi,k}/\kappa_{\xi,k}\in\mathcal{Y}$, $\kappa_{\beta,i}>0$, $\kappa_{\eta,j}>0$, and $\kappa_{\xi,k}>0$; $i=1,\ldots,p$, $j=1,\ldots,r$, $k=1,\ldots,n$. These details are presented in Sections 2.7 and 2.8.

Parameter Model 1 in (11) is only included when b is unknown (i.e., when the data model is specified to be either the negative binomial or gamma distributions). The truncated CM distribution is chosen because it is conjugate; see details in the Supplemental Appendix. In our experience (see Section 3.4), b is difficult to learn, and the results are extremely sensitive to the choice of α_b and κ_b. Several priors have been suggested for the overdispersion parameter when the data are distributed according to a negative binomial distribution (see, e.g., Gelman 2006, among others), some of which have been developed based on the gamma-Poisson interpretation of the negative binomial distribution (see Zhou and Carin 2015, and the references therein). In this article, we focus on using CM priors, and hence, other choices of priors on b (for LCM models) may lead to better results. We have found that the results are more favorable when specifying a different data model for the settings where b is unknown. Specifically, for the negative binomial setting we suggest using a Poisson distribution, and when the data is distributed as gamma we suggest taking the log transform and using a normal distribution.

Another important quantity that needs to be specified are the basis functions $\{\boldsymbol{\phi}_i\}$. This choice is very important and requires careful consideration. To illustrate the generality of our proposed model we consider three classes of basis function, each of which are demonstrated in Sections 3.3–3.5, respectively. Many analyses let $\{\boldsymbol{\phi}_i\}$ consist of known covariates (see, e.g., Wilson and Reich 2014, for a recent example). Another choice is to specify latent classes to model within-subject variability; in this setting, $\{\boldsymbol{\phi}_i\}$ is sometimes referred to as a "random effect design matrix" (see, e.g., Hodges 2013, chap. 1 for a discussion). Consider the example where $g_k\subset\{1,\ldots,n\}$ represents the kth group. In Section 3.3, g_k represents the kth herd of cows, and each element in g_k represents a specific cow in the sample. Here, we shall specify $\boldsymbol{\phi}_i=\left(I(i\in g_1),\ldots,I(i\in g_r)\right)'$. For spatial and time-series datasets, it is often assumed that $\{\boldsymbol{\phi}_i\}$ consists of spatial/temporally varying functions, referred to as "basis functions." For example, Fourier basis functions/wavelets are often used in the image analysis literature (see, e.g., Donoho and Johnstone 1994, for a classic reference). Similarly, radial basis functions, empirical orthogonal functions, and splines have been used to great effect in the spatial statistics, time-series, and spatio-temporal statistics literature (see, e.g., Wahba 1990; Bradley, Cressie, and Shi 2016; Bradley, Wikle, and Holan 2017; Wikle 2010, for a different choices of basis functions).

The value of r is a feature of the observed dataset when specifying $\{\boldsymbol{\phi}_i\}$ to be either covariates or a random effects design matrix (see Sections 3.3 and 3.4 for examples). However, when using a known class of basis functions, r must be specified. In this setting, selection criteria are often used to investigate both the sensitivity to the choice of r and how many are necessary to give reasonable predictions (see, e.g., Wahba 1990; Henao 2009; Bradley, Cressie, and Shi 2011, among others). Spike and slab, horseshoe priors, and SSVS (among other similar techniques) are extensions of the LGP, which one might adapt to the LCM to select covariates and basis functions (O'Hara and Sillanpaa 2009); however, we do not consider these extensions of the LCM in this article. When spatial basis functions depend on knot locations (thin-plate splines), a common rule-of-thumb is to specify equally spaced knots over the spatial domain (see, e.g., Nychka 2001, among others). In Section 3.5, we demonstrate the use of a known kernel using a big Bernoulli dataset consisting of cloud fractions. Here, we use the same basis functions specified in Sengupta et al. (2012), where the knots were chosen to be equally spaced.

2.4. Methodological Properties of the LCM

An important point argued in Section 1 is that the LGP model is a *special case* of an LCM. This can now easily be seen by letting $j=1,\ldots,4$, $k=4$, $\boldsymbol{\alpha}_\beta=\mathbf{0}_{p,1}$, $\boldsymbol{\alpha}_\eta=\mathbf{0}_{r,1}$, and $\boldsymbol{\alpha}_n=\mathbf{0}_{n,1}$. This specification yields an LGP model. A difficulty with this

specification is that we lose conjugacy by specifying $j \neq k$. Bradley, Holan, and Wikle (2018) showed that the multivariate log-gamma distribution they proposed can be made arbitrarily close to a multivariate normal distribution by specifying the shape and scale parameters to be large. This essentially allows one to use a LGP specification with a Poisson data model, and *also use* the conjugacy that arises from the MLG distribution when $j = k = 3$ in (11). This important property of the MLG distribution can be extended to the more general CM distribution.

Theorem 3. Suppose that $\psi \neq \psi_4$, and denote the first and second derivatives with ψ' and ψ'', $0 < \psi' < \infty$, and $0 < \psi'' < \infty$. Let the n-dimensional random vector $\mathbf{Y}$ distributed according $\text{CM}(\boldsymbol{\mu}, \left(\psi''(0)/\psi'(0)\right)^{1/2}\alpha^{1/2}\mathbf{V}, \alpha\mathbf{J}_{n,1}, \frac{\alpha}{\psi'(0)}\mathbf{J}_{n,1}; \psi)$ ignoring proportionality constants. Then $\mathbf{Y}$ converges in distribution to a multivariate normal random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{V}\mathbf{V}'$ as α approaches infinity.

The restriction of $\psi \neq \psi_4$ is sensible, since $\psi = \psi_4$ yields a CM exactly equal to a multivariate normal distribution. Also, Theorem 3 does not hold for the multivariate negative-inverse-gamma distribution, since $\psi_1'(0) = -\infty$.

The "best" DY parameters, for the multivariate logit-beta distribution and the MLG distribution, might not lead to something that looks Gaussian. That is, we should be able to learn whether or not the multivariate normal distribution is appropriate for latent processes of binomial and Poisson data by observing whether or not posterior replicates of the DY parameters (i.e., α and κ) are large (which would invoke Theorem 3). Hence, from this point-of-view, it is very important that we place prior distributions on the DY parameters, as we describe in Section 2.8.

These connections to the Gaussian distribution are important because it shows potential for the LCM to outperform a latent Gaussian process model. However, for the LCM to be as widely applicable as an LGP, we also require an important theoretical property referred to as Kolmogorov consistency (Daniell 1919; Kolmogorov 1933). That is, if the index on Z_i is defined over space or time, for example, then we need the CM distribution to be well defined for every possible subset of locations (Gelfand and Schliep 2016).

Theorem 4. The CM distribution, as defined in Theorem 1, is Kolmogorov consistent.

Theorems 3 and 4 are important methodological properties; however, if it is more difficult to implement LCM over the LGP, then these results may have less of an impact in practice. In Section 2.5, we show that it is rather straightforward to implement the LCM using a collapsed Gibbs sampler.

2.5. An Example Gibbs Sampler for the LCM

To simulate from a posterior distribution that is proportional to (11) we consider the following likelihood:

$$\begin{aligned}
&\text{Data Model}: Z_i|\boldsymbol{\beta},\boldsymbol{\eta},\xi_i \overset{\text{ind}}{\sim} \text{EF}\left(\mathbf{x}_i'\boldsymbol{\beta} + \boldsymbol{\phi}_i'\boldsymbol{\eta} + \xi_i + \mathbf{b}_{\beta,i}'\mathbf{q}_\beta \right.\\
&\qquad \left. + \mathbf{b}_{\eta,i}'\mathbf{q}_\eta + \mathbf{b}_{\xi,i}'\mathbf{q}_\xi;\ \psi_j\right).\\
&\text{Process Model 1}: \boldsymbol{\eta}|\mathbf{V},\alpha_\eta,\kappa_\eta,\mathbf{q}_\eta\\
&\qquad \sim \text{CM}\left(-\mathbf{V}\mathbf{B}_\eta\mathbf{q}_\eta, \mathbf{V}, \alpha_\eta, \kappa_\eta;\ \psi_k\right);\\
&\text{Process Model 2}: \boldsymbol{\xi}|\alpha_\xi,\kappa_\xi,\mathbf{q}_\xi\\
&\qquad \sim \text{CM}\left(-\mathbf{V}_\xi\mathbf{B}_\xi\mathbf{q}_\xi, \mathbf{V}_\xi, \alpha_\xi, \kappa_\xi;\ \psi_k\right);\\
&\text{Parameter Model 1}: b|\alpha_b,\kappa_b \sim \text{CM}\left(0,1,\alpha_b,\kappa_b;\ \psi_k\right) I(b>0)\\
&\text{Parameter Model 2}: \boldsymbol{\beta}|\alpha_\beta,\kappa_\beta,\mathbf{q}_\beta\\
&\qquad \sim \text{CM}\left(-\mathbf{V}_\beta\mathbf{B}_\beta\mathbf{q}_\beta, \mathbf{V}_\beta, \alpha_\beta, \kappa_\beta;\ \psi_k\right)\\
&\text{Parameter Model 3}: f(\mathbf{q}_\beta) = 1\\
&\text{Parameter Model 4}: f(\mathbf{q}_\eta) = 1\\
&\text{Parameter Model 5}: f(\mathbf{q}_\xi) = 1;\\
&i = 1,\ldots,n, j = 1,\ldots,4, k = 1,\ldots,4, \qquad (12)
\end{aligned}$$

where $\mathbf{b}_{\beta,i}$, $\mathbf{b}_{\eta,i}$, and $\mathbf{b}_{\xi,i}$ are prespecified n-dimensional vectors and the $p \times n$ matrix $\mathbf{B}_\beta$, the $r \times n$ matrix $\mathbf{B}_\eta$, and the $n \times n$ matrix $\mathbf{B}_\xi$ are also prespecified. There is an immediate connection between (11) and (12), which introduces the improper n-dimensional random vector $\mathbf{q}_\beta$, n-dimensional random vector $\mathbf{q}_\eta$, and n-dimensional random vector $\mathbf{q}_\xi$. Specifically, when conditioning (12) on the events $\mathbf{q}_\beta = \mathbf{0}_{n,1}$, $\mathbf{q}_\eta = \mathbf{0}_{n,1}$, and $\mathbf{q}_\xi = \mathbf{0}_{n,1}$, we obtain a likelihood that is proportional to (11). Consequently, we suggest implementing the collapsed Gibbs sampler (Liu 1994) outlined in the pseudo-code. In general, one can interpret $\mathbf{q}_\beta$, $\mathbf{q}_\eta$, and $\mathbf{q}_\xi$ as location parameters for $\boldsymbol{\beta}$, $\boldsymbol{\eta}$, and $\boldsymbol{\xi}$, and are given non-informative priors.

As an example, consider deriving the full-conditional distribution in Step 2. Write the data model in (12) as

$$\begin{aligned}
f(\mathbf{Z},\mathbf{q}_\beta|\cdot) \underset{\mathbf{Z}}{\propto} \exp\Big\{&\mathbf{Z}'\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}'\boldsymbol{\Phi}\boldsymbol{\eta} + \mathbf{Z}'\boldsymbol{\xi} + \mathbf{Z}'\mathbf{B}_{\beta,1}\mathbf{q}_\beta\\
&- b\mathbf{J}_{n,1}'\psi(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\Phi}\boldsymbol{\eta} + \boldsymbol{\xi} + \mathbf{B}_{\beta,1}\mathbf{q}_\beta)\Big\}\, h, \qquad (13)
\end{aligned}$$

where the $n \times n$ matrix $\mathbf{B}_{\beta,1} = \left(\mathbf{b}_{\beta,1},\ldots,\mathbf{b}_{\beta,n}\right)'$, $n \times p$ matrix $\mathbf{X} \equiv (\mathbf{x}_1,\ldots,\mathbf{x}_n)'$, the $n \times r$ matrix $\boldsymbol{\Phi} \equiv \left(\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_n\right)'$, $h = \left\{\prod_{i=1}^n I(\mathbf{x}_i'\boldsymbol{\beta} + \boldsymbol{\phi}_i'\boldsymbol{\eta} + \xi_i \in \mathcal{Y})\right\}$, and $\underset{\mathbf{Z}}{\propto}$ denotes the "proportional to as a function of $\mathbf{Z}$" symbol. Using (12) and Parameter Model 2 in (13) we have that

$$f(\boldsymbol{\beta},\mathbf{q}_\beta|\cdot) \underset{\boldsymbol{\beta}}{\propto} \text{CM}\left(\boldsymbol{\mu}_\beta^*, \mathbf{V}_\beta^*, \boldsymbol{\alpha}_\beta^*, \boldsymbol{\kappa}_\beta^*;\ \psi\right) h, \qquad (14)$$

where $\boldsymbol{\mu}_\beta^* = \mathbf{V}_\beta^*\left(-\boldsymbol{\eta}'\boldsymbol{\Phi}' - \boldsymbol{\xi}', \mathbf{0}_p'\right)'$, $\boldsymbol{\alpha}_\beta^* = (\mathbf{Z}', \boldsymbol{\alpha}_\beta')'$, $\boldsymbol{\kappa}_\beta^* = (b\mathbf{J}_{n,1}, \boldsymbol{\kappa}_\beta')'$, $\mathbf{V}_\beta^{*-1} = (\mathbf{H}_\beta, \mathbf{Q}_\beta)$, the $(n+p) \times p$ matrix $\mathbf{H}_\beta = (\mathbf{X}', \mathbf{V}_\beta')'$, and the $(n+p) \times n$ matrix $\mathbf{Q}_\beta = (\mathbf{B}_{\beta,1}', \mathbf{B}_\beta')'$. See the Supplemental Appendix for the algebra leading to (14). The full-conditional distribution in (14) is not well defined when $Z_i = 0$ for some i, because this produces a zero shape parameter. In this setting one can add an "ϵ" to the elements of $\mathbf{Z}$ to force nonzero shape parameters. However, this choice changes the prior from a CM distribution to a CM_c distribution, and a considerable amount of book-keeping is required to derive the full-conditional distributions. For ease of exposition, we put these details in the Supplemental Appendix C.

Algorithm 1 Pseudo-code: Collapsed Gibbs sampler for the model in (12)

1: Set $b = 1$ and initialize $\boldsymbol{\beta}^{[0]}$, $\boldsymbol{\eta}^{[0]}$, and $\boldsymbol{\xi}^{[0]}$.
2: Sample $\boldsymbol{\beta}^{[g]}$ from $f(\boldsymbol{\beta}|\mathbf{Z}, \boldsymbol{\eta}^{[g-1]}, \boldsymbol{\xi}^{[g-1]}, b^{[g-1]}, \mathbf{q}_\eta = \mathbf{0}_n, \mathbf{q}_\xi = \mathbf{0}_n)$.
3: Sample $\boldsymbol{\eta}^{[g]}$ from $f(\boldsymbol{\eta}|\mathbf{Z}, \boldsymbol{\beta}^{[g]}, \boldsymbol{\xi}^{[g-1]}, b^{[g-1]}, \mathbf{q}_\beta = \mathbf{0}_n, \mathbf{q}_\xi = \mathbf{0}_n)$.
4: Sample $\boldsymbol{\xi}^{[g]}$ from $f(\boldsymbol{\xi}|\mathbf{Z}, \boldsymbol{\beta}^{[g]}, \boldsymbol{\eta}^{[g]}, b^{[g-1]}, \mathbf{q}_\beta = \mathbf{0}_n, \mathbf{q}_\eta = \mathbf{0}_n)$.
5: Sample $b^{[g]}$ from $f(b|\mathbf{Z}, \boldsymbol{\beta}^{[g]}, \boldsymbol{\eta}^{[g]}, \boldsymbol{\xi}^{[g]}, \mathbf{q}_\beta = \mathbf{0}_n, \mathbf{q}_\eta = \mathbf{0}_n, \mathbf{q}_\xi = \mathbf{0}_n)$.
6: Repeat Steps 2, 3, and 4 until $g = G$ for a prespecified value of G.

If we prespecify $\mathbf{Q}_\beta$ so that it is equal to the basis for the null space of $\mathbf{H}_\beta$ (i.e., $\mathbf{Q}_\beta'\mathbf{Q}_\beta = \mathbf{I}_n$, $\mathbf{Q}_\beta'\mathbf{H}_\beta = \mathbf{0}_{n,p}$, and $\mathbf{H}_\beta(\mathbf{H}_\beta'\mathbf{H}_\beta)^{-1}\mathbf{H}_\beta' + \mathbf{Q}_\beta\mathbf{Q}_\beta' = \mathbf{I}_{n+p}$). Then,

$$\mathbf{V}_\beta^* = (\mathbf{H}_\beta, \mathbf{Q}_\beta)^{-1} = \begin{pmatrix} (\mathbf{H}_\beta'\mathbf{H}_\beta)^{-1}\mathbf{H}_\beta' \\ \mathbf{Q}_\beta' \end{pmatrix}. \tag{15}$$

From (7) we see that to sample a value from $f(\boldsymbol{\beta}, \mathbf{q}_\beta|\mathbf{Z}, \boldsymbol{\eta}, \boldsymbol{\xi}, b, \mathbf{q}_\eta = \mathbf{0}_n, \mathbf{q}_\xi = \mathbf{0}_n)$ we can compute

$$\begin{pmatrix} \boldsymbol{\beta} \\ \mathbf{q}_\beta \end{pmatrix} = -\begin{pmatrix} (\mathbf{H}_\beta'\mathbf{H}_\beta)^{-1}\mathbf{H}_\beta'(\boldsymbol{\Phi}\boldsymbol{\eta} + \boldsymbol{\xi}) \\ \mathbf{0}_n \end{pmatrix} + \begin{pmatrix} (\mathbf{H}_\beta'\mathbf{H}_\beta)^{-1}\mathbf{H}_\beta'\mathbf{w} \\ \mathbf{Q}_\beta'\mathbf{w} \end{pmatrix}, \tag{16}$$

where $\mathbf{w} \sim \text{CM}\left(\mathbf{0}_{n+p}, \mathbf{I}_{n+p}, \boldsymbol{\alpha}_\beta^*, \boldsymbol{\kappa}_\beta^*; \psi\right)$, which can easily be generated using (7). Thus, to simulate according to Step 2 of the collapsed Gibbs sampler we can compute,

$$\boldsymbol{\beta} = -(\mathbf{H}_\beta'\mathbf{H}_\beta)^{-1}\mathbf{H}_\beta'(\boldsymbol{\Phi}\boldsymbol{\eta} + \boldsymbol{\xi}) + (\mathbf{H}_\beta'\mathbf{H}_\beta)^{-1}\mathbf{H}_\beta'\mathbf{w}. \tag{17}$$

It is (computationally) easy to simulate in this manner provided that $p \ll n$. Recall that $\mathbf{H}_\beta$ is $n \times p$, which implies that computing the $p \times p$ matrix $(\mathbf{H}_\beta'\mathbf{H}_\beta)^{-1}$ is computationally feasible when p is "small." By small we mean a value such that the Gauss–Jordan elimination method for the inverse of a $p \times p$ matrix can be computed in real-time. Furthermore, the p-dimensional random vector $\boldsymbol{\beta}$ is an orthogonal projection of the n-dimensional random vector $\mathbf{w}$ onto the column space spanned by the columns of $\mathbf{H}_\beta$. This provides a geometric interpretation of random vectors generated according to (17).

2.6. Properties of the Augmented LCM Model

As discussed in Section 2.2 and Supplementary Appendix A, it is difficult to simulate directly from a CM_c distribution since $\mathbf{H}$ in (10) is not square, and hence, one cannot use Equation (7). In Section 2.5, we instead consider simulating from CM_c after marginalizing across a location parameter with improper prior. This leads to the following result.

Theorem 5. Let $\mathbf{q}_1|\mathbf{c}, \mathbf{H}, \boldsymbol{\alpha}, \boldsymbol{\kappa} \sim \text{CM}_c(\mathbf{c}, \mathbf{H}, \boldsymbol{\alpha}, \boldsymbol{\kappa})$, where $\mathbf{H} \in \mathbb{R}^M \times \mathbb{R}^r$ is full column rank, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_M)'$, $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_M)'$, $\alpha_i/\kappa_i \in \mathcal{Y}$, and $\kappa_i > 0$ for $i = 1, \ldots, M$. Assume a reparameterized value of $\mathbf{c} = -\mathbf{B}\mathbf{q}_2 + \boldsymbol{\mu}$, and the improper prior $f(\mathbf{q}_2|\mathbf{c}, \mathbf{H}, \mathbf{B}, \boldsymbol{\alpha}, \boldsymbol{\kappa}) \propto 1$, where $\mathbf{q}_2$ is $(M - r)$-dimensional. Also let $\mathbf{B}, \in \mathbb{R}^M \times \mathbb{R}^{M-r}$ be the orthonormal basis for the null space of $\mathbf{H}$, $\mathbf{q} = (\mathbf{q}_1', \mathbf{q}_2')'$, $\boldsymbol{\mu} \in \mathbb{R}^M$, $\mathbf{I}_n$ be an $n \times n$ identity matrix, and let $\mathbf{w} \sim \text{CM}_c(\boldsymbol{\mu}, \mathbf{H}, \boldsymbol{\alpha}, \boldsymbol{\kappa})$. Define $\mathbf{V}^{-1} = (\mathbf{H}, \mathbf{B})$.

(i) Then,

$$\int f(\mathbf{q}_1|\mathbf{c} = -\mathbf{B}\mathbf{q}_2 + \boldsymbol{\mu}, \mathbf{H}, \mathbf{B}, \boldsymbol{\alpha}, \boldsymbol{\kappa}) d\mathbf{q}_2 \propto \int \exp\left\{\boldsymbol{\alpha}'\mathbf{V}^{-1}\mathbf{q} - \boldsymbol{\kappa}'\psi\left(\mathbf{V}^{-1}\mathbf{q} - \boldsymbol{\mu}\right)\right\} d\mathbf{q}_2, \tag{18}$$

where ψ is a unit log-partition function and the integrand on the right hand side of (18) is proportional to $\text{CM}(\mathbf{V}\boldsymbol{\mu}, \mathbf{V} = (\mathbf{H}, \mathbf{B})^{-1}, \boldsymbol{\alpha}, \boldsymbol{\kappa})$. Furthermore, the affine transformation $(\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'\mathbf{w}$ is a draw from the density in (18).

(ii) The conditional mean and covariance can be computed as

$$E(\mathbf{q}_1|\mathbf{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa}) = (\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'\mathbf{k}(\boldsymbol{\alpha}, \boldsymbol{\kappa})$$
$$\text{cov}(\mathbf{q}_1|\mathbf{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa}) = (\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'\mathbf{K}(\boldsymbol{\alpha}, \boldsymbol{\kappa})\mathbf{H}(\mathbf{H}'\mathbf{H})^{-1},$$

where we have integrated across $\mathbf{q}_2$.

The proof of Theorem 5(i) is given in the Supplemental Appendix. The proof of Theorem 5(ii) follows immediately from Theorem 1(ii) and Theorem 5(i). Thus, we state Theorem 5(ii) without proof.

Theorem 5(i) offers a more formal statement of a heuristic described in the rejoinder of Bradley, Holan, and Wikle (2018) for the MLG distribution. Thus, this result is an important contribution as it provides the necessary conditions required to argue the use of the sampler described in Section 2.5. As discussed at the end of Section 2.4, computational considerations are extremely important when proposing a new complex model. A collapsed Gibbs sampler will allow one to avoid Metropolis–Hastings updates, which in turn, increases the effective sample size and, consequently, the computational performance of the method.

The integrand on the left-hand-side of (18) is proportional to a CM_c, and is of the same form as the full-conditional distributions that arise in the LCM in Section 2.5. Theorem 5 shows that it is (computationally) easy to simulate from the left-hand-side of (18) provided that $r \ll n$ and that $\mathbf{q}_2$ is marginalized. Recall that $\mathbf{H}$ is $n \times r$, which implies that computing the $r \times r$ matrix $(\mathbf{H}'\mathbf{H})^{-1}$ is computationally feasible when r is "small." By small we mean a value such that the Gauss–Jordan elimination method for the inverse of a $r \times r$ matrix can be computed in real-time. Furthermore, Theorem 5 shows that the r-dimensional random vector $\mathbf{q}$ is an orthogonal projection of the n-dimensional random vector $\mathbf{w}$ onto the column space spanned by the columns of $\mathbf{H}$. This provides a geometric interpretation of random vectors generated from CM_c after marginalizing $\mathbf{q}_2$.

2.7. Prior Distributions on Covariance Parameters

A critical feature of our proposed distribution theory is the incorporation of dependence in non-Gaussian data from the exponential family. From this point-of-view it is especially

important to learn about these dependencies, which are quantified by the unknown $n \times n$ real-valued matrix $\mathbf{V}$ (or equivalently $\mathbf{V}^{-1}$). Thus, we place a prior distribution on $\mathbf{V}^{-1}$. Specifically, let $\mathbf{V}^{-1}$ be an unknown lower unit triangle matrix. That is, let $\mathbf{V}^{-1} \equiv \{v_{i,j}\}$, where $v_{i,j} = 1$ for $j = i$, $v_{i,j} = 0$ for $j > i$, and $v_{i,j} \in \mathbb{R}$ for $j < i$. It will be useful to organize the elements below the lower main diagonal into the $(i-1)$-dimensional vectors $\mathbf{v}_i \equiv (v_{i,j} : j = 1, \ldots, i-1)'$ for $i = 2, \ldots, n$.

We place a CM prior distribution on $\mathbf{v}_i$ for each i. Specifically, let

$$\mathbf{v}_i \overset{\text{ind}}{\sim} \text{CM}(\mathbf{0}_{i-1}, \mathbf{C}_i, \boldsymbol{\alpha}_i, \boldsymbol{\kappa}_i; \psi); \quad i = 2, \ldots, n, \tag{19}$$

where, in practice, the $(i-1) \times (i-1)$ matrix $\mathbf{C}_i$ is set equal to $\sigma_v \mathbf{I}_{i-1}$, and $\boldsymbol{\alpha}_i$, $\boldsymbol{\kappa}_i$, and σ_v are specified such that (19) is relatively "flat." This specification leads to a conjugate full-conditional distribution within a Gibbs sampler (see Supplemental Appendix C for the derivation).

The CM prior distribution on the modified Cholesky decomposition of the precision matrix is similar to priors considered by Daniels and Pourahmadi (2002), Chen and Dunson (2003), and Pourahmadi, Daniels, and Park (2007) in the Gaussian setting. In fact, when $\psi = \psi_4$ the prior distribution in (19) reduces to the prior distributions used in Daniels and Pourahmadi (2002), Chen and Dunson (2003), and Pourahmadi, Daniels, and Park (2007). Thus, (19) constitutes a general non-Gaussian (natural exponential family) extension of such priors on modified Cholesky decompositions of precision and covariance matrices.

There are certainly other prior distributions for $\mathbf{V}^{-1}$ that may be more appropriate. For example, see Yang and Berger (1994) and Bradley, Wikle, and Holan (2016, for the spatial setting) for a Givens angle prior on covariance parameters. The Wishart and inverse Wishart are also common alternatives (see, e.g., Gelman et al. 2013, for a standard reference). However, conjugacy may not always be present depending on the choice of CM distribution. Thus, in this article, we investigate the fully conjugate form of the LCM and specify the prior for $\mathbf{V}^{-1}$ as stated in (19).

2.8. Prior Distributions on DY Parameters

Following the theme of the previous sections, we define conjugate priors for the DY parameters by defining a distribution with an exponential term and an exponential to the negative unit log partition function. That is, consider

$$f(\alpha, \kappa | \gamma_1, \gamma_2, \rho) \propto \exp\left[\gamma_1 \alpha + \gamma_2 \kappa - \rho \log\left\{\frac{1}{K(\alpha, \kappa)}\right\}\right], \tag{20}$$

where γ_1, γ_2, and ρ are hyperparameters. The parameter space for γ_1, γ_2, and ρ that ensures that (20) is proper (i.e., can be normalized to define a probability measure) is an immediate consequence of a result from Diaconis and Ylvisaker (1979). In particular, from Theorem 1 of Diaconis and Ylvisaker (1979), the distribution in (20) is proper provided that $\mathcal{Y}$ is a nonempty real-valued open set, the range of ψ is a nonempty real-valued open set, $\gamma_1/\rho \in \mathcal{Y}$, $\gamma_2/\rho \in \mathcal{Y}_\psi$, and $\rho > 0$, where $\mathcal{Y}_\psi \equiv \{M : M = -\psi(Y), Y \in \mathcal{Y}\}$. For the CM distribution associated with ψ_1 we see that $\mathcal{Y} = \{Y : Y < 0\}$ and the range of ψ is $\mathbb{R}$; thus, for this setting $\gamma_1 < 0$, $\gamma_2 \in \mathbb{R}$, and $\rho > 0$ results in a proper prior in (20). For the CM distributions associated with ψ_2, ψ_3, and ψ_4 we have that $\mathcal{Y} = \mathbb{R}$ and ψ is a strictly positive; thus, for this setting $\gamma_1 \in \mathbb{R}$, $\gamma_2 < 0$, and $\rho > 0$ ensures propriety of (20).

There are many interesting special cases of the prior distribution in (20). For example, when α is integer-valued and $\psi = \psi_3$ then the prior in (20) has a relationship with the Conway–Maxwell–Poisson distribution (Conway and Maxwell 1962) and the gamma distribution. These special cases (listed in Table 3 of the Supplemental Appendix C) are particularly useful because they give rise to interpretations of the hyperparameters. In particular, for $\psi = \psi_1$, we have that ρ can be interpreted as a dispersion parameter (in relation to the dispersion parameter of a Conway–Maxwell–Poisson distribution), γ_1 can be interpreted as a location parameter, and γ_2 can be interpreted as a scale parameter. When $\psi = \psi_2$ we have that γ_1 and γ_2 can be interpreted as functions of a proportion (i.e., the inverse logit or log of a proportion). For $\psi = \psi_3$, we have that ρ can be interpreted as a dispersion parameter, γ_2 can be interpreted as a location parameter, and γ_1 can be interpreted as a scale parameter. Finally, when $\psi = \psi_4$ we have that γ_1 is interpreted as a location parameter, ρ represents a shape parameter, and γ_2 represents a scale parameter.

The most familiar special case occurs when $\psi = \psi_4$ (i.e., a normal data model) and $\alpha = 0$. Namely, (20) reduces to independent gamma prior distributions on κ with shape parameter $\rho/2 + 1$, and scale parameter $-\gamma_2$. When recognizing that κ is equal to one-half the unknown variance of a normal random variable (see Table 1), we see that the conjugate prior distribution implies an inverse gamma distribution for the variance parameter, which is a common choice of a prior distribution on the variance parameter for normally distributed data (Gelman 2006).

3. Empirical Results

In Sections 3.1 and 3.2, we use simulations to demonstrate the performance of the LCM when analyzing binomial and Poisson data. To demonstrate the wide-applicability of the CM distribution, we also give several illustrations from a variety of disciplines; namely, we analyze an epidemiology dataset (Section 3.3), a federal statistics dataset (Section 3.4), and an environmental dataset (Section 3.5). Our computations were performed on a dual 10 core 2.8 GHz Intel Xeon E5-2680 v2 processor with 256 GB of RAM. All R code and Matlab code used in these examples are provided in the Supplemental Appendix. User-friendly R code is provided at: *https://github.com/JonathanBradley28/CM*.

3.1. Simulation Example

We compare predictions using the LGP versus predictions based on a LCM. As discussed in Section 1, the LGP is the standard approach for Bayesian analysis of dependent data, and thus, the results in this section are meant to provide one comparison of the LCM to the current state-of-the-art. It is important to emphasize that if the LGP is more appropriate than the LCM, our model will be able to identify this for some settings because of Theorem 3; that is, if the posterior replicates of the DY

parameters are large then Theorem 3 suggests that the latent processes are approximately Gaussian.

The $n \times p$ matrix $\mathbf{X} \equiv (\mathbf{x}_1, \ldots, \mathbf{x}_p)'$, the $r \times r$ matrix $\boldsymbol{\Phi} \equiv (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_r)'$, and the $r \times r$ lower unit triangle matrix $\mathbf{V}^{-1}$ are randomly generated with $p = 500$ and $r = 10$. The choices for p and r were made to represent realistic values that one might see in practice. For example, see Matloff (2016, chap. 2), where they consider a dataset taken from a "data exposition" provided by the American Statistical Association's Sections on Statistical Computing and Statistical Graphics. This example had $n = 500{,}000$ and $p = 29$ and was considered to be a moderate p and large n setting. Also, see Huang and Sun (2018) for a recent example where $n = 2{,}153{,}888$ and $r = 60$ is considered to be a moderate r and large n setting. Each element of the $n \times p$ matrix $\mathbf{X}$, the $n \times r$ matrix $\boldsymbol{\Phi}$, and the $r \times r$ matrix $\mathbf{V}^{-1}$ are selected from a standard normal distribution. The elements of the fixed and random effects $\boldsymbol{\beta}$, $\boldsymbol{\eta}$, and $\boldsymbol{\xi}$ are randomly selected from a standard normal distribution as well. Then, we define $p_i = \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta}+\boldsymbol{\phi}_i'\boldsymbol{\eta}+\xi_i)}{1+\exp(\mathbf{x}_i'\boldsymbol{\beta}+\boldsymbol{\phi}_i'\boldsymbol{\eta}+\xi_i)}$ for $i = 1, \ldots, n$. We consider two different data models in this section. In particular, we consider observations Z_i generated from a binomial distribution with sample size t_i (generated from a Poisson with mean 40) and probability of success p_i for $i = 1, \ldots, n$. For example, consider n households, where for each household i there are t_i individuals, and let p_i represent the probability that an individual is female. Then Z_i would represent the number of women living in household i. Here, one might choose $\boldsymbol{\phi}_i = (1, 0)'$ if the total income of the household is below the poverty line, and $\boldsymbol{\phi}_i = (0, 1)'$ otherwise. Similarly, we consider observations Z_i generated from a Poisson distribution with mean $\exp(\mathbf{x}_i'\boldsymbol{\beta} + \boldsymbol{\phi}_i'\boldsymbol{\eta} + \xi_i)$ for $i = 1, \ldots, n$.

Using the Gibbs sampler outlined in Supplemental Appendix C we implement the LCM in Model 1 with $j = k$ and use the appropriate data model (i.e., binomial or Poisson). We use the R-package `lme4` to implement a LGP. Default choices were used when possible in when using `lme4`. For each i, denote the posterior mean of p_i with $\widehat{p}_i$, the posterior mean of μ with $\widehat{\mu}_i$, and define the total squared prediction errors to be

$$\sum_i (tp_i - t\widehat{p}_i)^2$$
$$\sum_i (\mu_i - \widehat{\mu}_i)^2, \tag{21}$$

used for the binomial and Poisson settings, respectively. We used a burn-in of 10,000 and generate $B = 20{,}000$ posterior replications for both data models that are considered.

We consider a large sample size of $n = 100{,}000$ and simulate $\{Z_1, \ldots, Z_n\}$ 50 times from the binomial distribution, and simulate another 50 independent replications of $\{Z_1, \ldots, Z_n\}$ from the Poisson distribution. In Figures 1(a) and (b), we plot the difference in log mean squared prediction error (MSPE) error of the LCM model and the total squared prediction error of the LGP model over the 50 independent replicates. A difference greater than zero indicates that the LCM has smaller total square prediction error. Here, we see that the differences in log MSPE are consistently larger than zero, and hence, the LCM clearly outperforms the LGP for this simulation design for both the binomial and Poisson settings. Thus, not only does the LCM lead to practical advantages (no tuning is involved) over the LGP for this example, there are also clear gains in predictive performance. Thus, this simulation suggests that the LCM model yields precise predictions, and is computationally feasible for a large dataset with moderate values for p and r. Note that the high predictive performance of the LCM model occurs in a setting where we do not generate the truth from a multivariate logit-beta distribution.

3.2. A Simulation Study

Real datasets often do not perfectly reflect the statistical model used for implementation. As such, it is necessary to provide evidence of the robustness (of prediction) to model misspecification through simulation studies. We do this by considering several specifications of the simulation model in Section 3.1 and of the fitted model used to analyze the simulated data with $n = 100$. Specifically, we consider the following factors in an analysis of variance (ANOVA) experiment:

- *Factor 1 (Random effects in the simulation model):* The simulated data are generated from the Poisson distribution in the same way as in Section 3.1 with: (*Level 1*) Gaussian random effects and (*Level 2*) multivariate log-gamma random effects.
- *Factor 2 (Number of covariates in the simulation model):* The simulated data are generated from the Poisson distribution in the same way as in Section 3.1 with: (*Level 1*) $p = 10$ and (*Level 2*) $p = 50$.

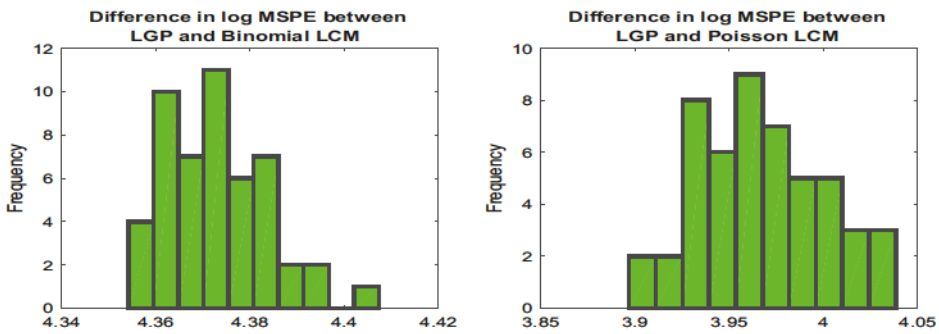

Figure 1. Histogram over the difference in log MSPE: For each of the fifty realizations of $\{Z_1, \ldots, Z_n\}$ from a (Poisson distribution) binomial distribution, we produce $(\{\widehat{\mu}_i\})$ $\{\widehat{p}_i\}$ using the appropriate LCM model, produce $(\{\widehat{\mu}_i\})$ $\{\widehat{p}_i\}$ using the LGP model, and compute the difference in log MSPE. The difference in log MSPE for the (Poisson distribution) binomial setting is the log of the total squared prediction error of $(\{\widehat{\mu}_i\})$ $\{t\widehat{p}_i\}$ from the LGP model minus the log total squared prediction error of $(\{\widehat{\mu}_i\})$ $\{t\widehat{p}_i\}$ computed using the LCM model. The histogram in the (right) left panel is over the 50 independent replicates from the (Poisson distribution) binomial distribution.

- *Factor 3 (Number of basis functions in the simulation model):* The simulated data are generated from the Poisson distribution in the same way as in Section 3.1 with: (*Level 1*) $r = 10$ and (*Level 2*) $r = 50$.
- *Factor 4 (Distributional assumptions of the fitted model):* We make the following distributional assumptions when fitting a model to the simulated data: (*Level 1*) a Poisson LGP model, (*Level 2*) a Poisson LCM model with $j = k = 3$, and (*Level 3*) a negative binomial LCM with $j = k = 2$.
- *Factor 5 (Number of covariates in the fitted model):* We make the following assumptions when fitting a model to the simulated data: (*Level 1*) $p = 10$ and (*Level 2*) $p = 50$.
- *Factor 6 (Number of basis functions in the fitted model):* We make the following distributional assumptions when fitting a model to the simulated data: (*Level 1*) $r = 10$ and (*Level 2*) $r = 50$.

There are a total of $2^5 \times 3 = 96$ factor level combinations (Factor 4 has three levels). The response in this experiment is the log total prediction error in (21). The log transformation is done to aid in producing normality in an analysis of variance (ANOVA) experiment. Within each factor-level-combination we simulate 10 independent replicates of $\{Z_1, \ldots, Z_{100}\}$, and compute the log total prediction error in (21). This leads to a total of $10 \times 96 = 960$ observations used in our ANOVA. Notice that we consider cases were we both correctly and incorrectly specify the covariates, basis functions, and distributional assumptions. This is done in an effort to assess robustness (of prediction) to model misspecification.

We implement an ANOVA with up to two-way interactions between the factors defined in the bulleted list above. The ANOVA table is provided in Supplemental Appendix D. The first and fourth main effect, and their interaction, have large F statistics. The remaining F statistics are not "significant." This suggests that, for our simulation setup, the proposed model is fairly robust to misspecification of covariates and basis functions. However, the specification of the fitted model (i.e., an LGP or LCM) appears to explain most of the variability in the log total prediction error. In Figure 2, we plot the interaction plot associated with Factors 1 and 4. Here we see that even when the data are simulated with Gaussian random effects, we appear to outperform the LGP with either LCM. Both LCMs perform similarly in this setting. When the data are simulated with multivariate log-gamma random effects the ANOVA suggests that the Poisson LCM performs considerably better than the LGP, and slightly outperforms the negative binomial LCM.

Further details on how the LCM outperforms the LGP (in terms of prediction) in several settings are provided in Figure 3. Here, we isolate 9 factor level combinations (or settings) of interest that compare the predictive performance of the LGP and the LCM. Specifically, Figure 3 displays boxplots of the log MSPE (of both the LGP and LCM) when the covariates and basis functions are either correctly specified or misspecified. In general, we see that both the Poisson and negative binomial LCMs consistently outperform the LGP in each setting (i.e., when the covariates and basis functions are either correctly or incorrectly specified). For the LGP it appears that the log MSPE is larger when too few covariates and basis functions are specified, but is fairly robust to the case when too many are specified. This is consistent with

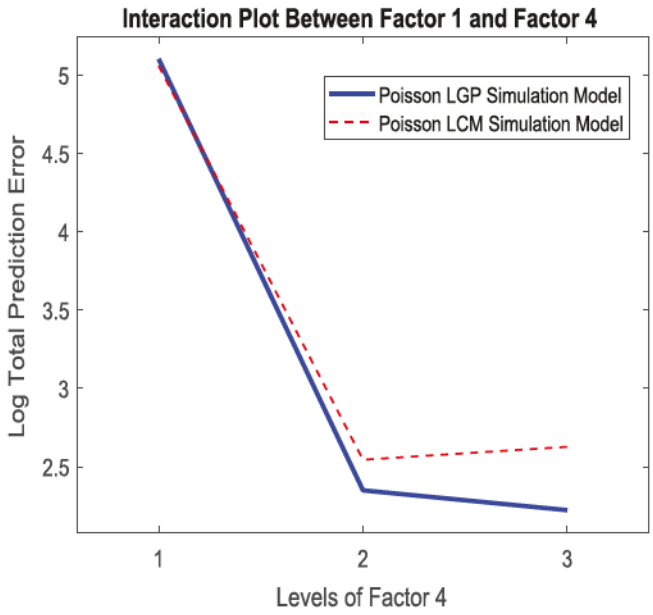

Figure 2. A two-way interaction plot for Factors 1 and 4, using the log total prediction error in (21) as the response. The levels of Factor 1 are given in the legend, and the levels of factor one are listed on the x-axis. Notice that the models that are implemented (labeled on the x-axis), may be different from the models that the data are simulated from (indicated by the solid blue, and dashed red lines). Regardless of how the data are generated, fitting the LCM (levels 2 and 3 of Factor 4) appears leads to smaller log total prediction error than when fitting the Poisson LGP on average.

previous results in the literature (see, e.g., Bradley, Cressie, and Shi 2011, among others). However, the predictive performance of the LCM models appears less sensitive to when fewer covariates and basis functions are present.

3.3. An Application to Contagious Bovine Pleuropneumonia in Ethiopian Highlands

Contagious bovine pleuropneumonia (CBPP) has been classified as a list-A disease by the World Organization for Animal Health. For this reason, Lesnoff et al. (2004) conducted an extensive study on herds of cows located within the Boji district of West Wellega, Ethiopia. They collected the incidence of CBPP among 15 herds over four time periods that span 16 months. They were interested in tracking the probability of contracting the disease as a function of time, and considered a generalized linear mixed model to assess this. Time was found to be an important fixed effect, and the herds were found to be an important random effect (Lesnoff et al. 2004). This is a small dataset consisting of 54 observations.

We fit a binomial LCM to these data, where the response is the number of cows infected and the total number cows in the heard is known (i.e., t_i). We define $\mathbf{x}_i$ to consist of indicators of the different time-periods. Let g_k represent the kth herd of cows, and let each element in g_k represents a specific cow in the sample. Specify $\boldsymbol{\phi}_i = \big(I(i \in g_1), \ldots, I(i \in g_r)\big)'$ for cows $i = 1, \ldots, 54$. We compare our results to a LGP fitted using a standard R-package for generalized linear models; namely, the R-package `lme4`, and using the function "glmer" (Bates et al. 2017). In Figure 4, we plot the ratio of the mean squared prediction errors (i.e., MSPE associated with LGP and the MSPE associated with the LCM). Here, the paired t-test resulted in a p-value of

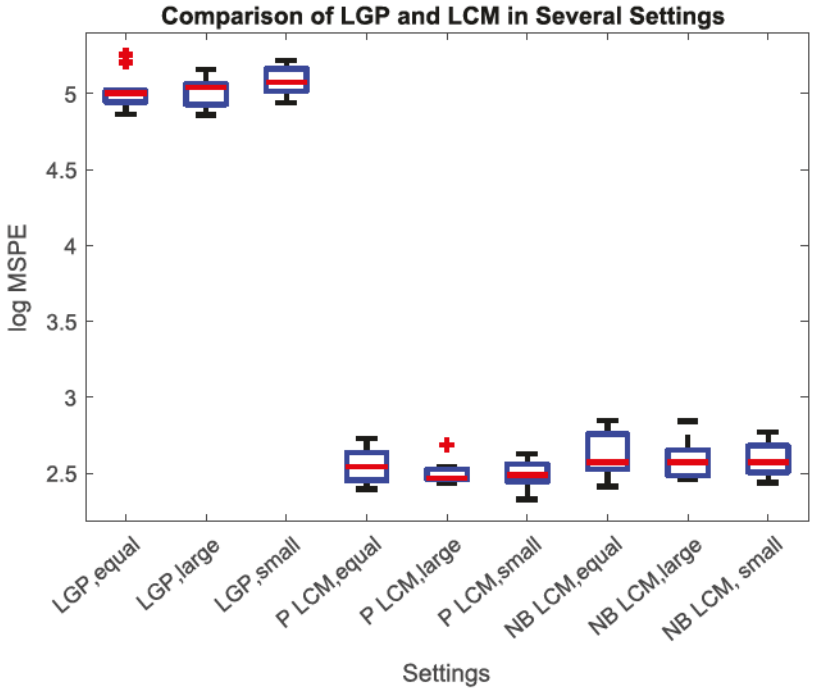

Figure 3. We provide boxplot of the log MSPE in several settings. These settings represent 9 factor level combinations of the 96 available. In particular Factor 1 is set equal to Level 1, where the Gaussian random effects are specified. The labels along the *x*-axis indicates the values for Factors 2 through 6. Here, "LGP" indicates Factor 4 is set to Level 1, "P LCM" indicates that Factor 4 is set to Level 2, and "NB LCM" indicates that Factor 4 is set to Level 3. When the label contains the word "equal" Factor 2, 3, 5, and 6 are all set to Level 1 so that the choice of p and r in the fitted model is equal to the value of p and r generating the data. When the label contains the word "large" ("small") Factors 2 and 3 are set to Level 1 (Level 2) and Factors 5 and 6 are all set to Level 2 (Level 1) so that the choice of p and r in the fitted model is larger (smaller) than the value of p and r generating the data.

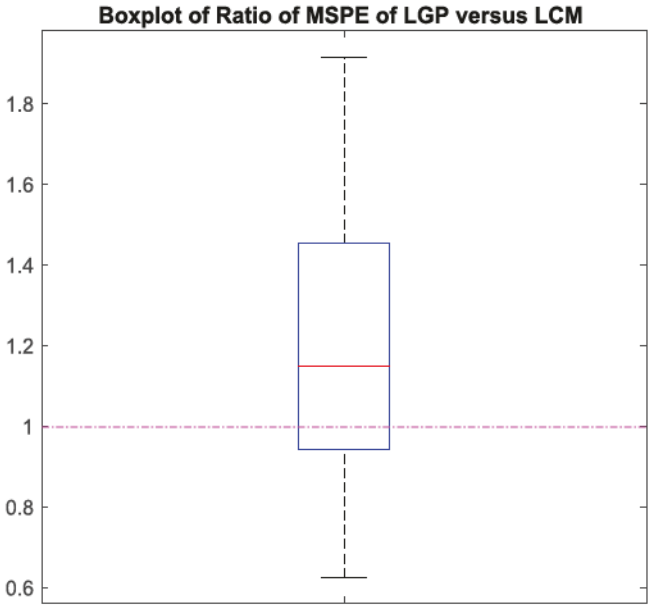

Figure 4. The response is the ratio between mean squared prediction error using the LGP model and the mean squared prediction error of the LCM. We hold out roughly 5% of the observations. A boxplot is displayed over 50 different hold-out observations. The mean squared prediction error (MSPE) is between the predicted mean (e.g., posterior mean of $t_i \exp\{\mathbf{x}_i'\boldsymbol{\beta} + \boldsymbol{\phi}_i'\boldsymbol{\eta} + \xi_i\} / [1 + \exp\{\mathbf{x}_i'\boldsymbol{\beta} + \boldsymbol{\phi}_i'\boldsymbol{\eta} + \xi_i\}]$) and the hold out dataset. Values greater than one (indicated by the dashed-dotted magenta line) suggest that the binomial LCM outperforms the binomial LGP.

1.27×10^{-5}, which suggests that the LCM is outperforming the GLM in this setting. However, visually Figure 4 suggests that the LCM and LGP give similar results for this example. The LGP and LCM both were computed in under a minute on average (across holdout data). The LGP on average took 0.07 min, while the LCM was took slightly longer at 0.86 min. The fast speed of both methods is not surprising since this is a very small dataset.

3.4. An Application to Count-Valued ACS Public-Use Micro-Data

The US Census Bureau has replaced the decennial census long-form with the American Community Survey (ACS), which is an ongoing survey that collects an enormous amount of information on US demographics. (To date there are over 64,000 variables published through the ACS.) The estimates published from the ACS have a unique multiyear structure. Specifically, the ACS produces 1-year and 5-year period estimates of US demographics, where 1-year period estimates are summaries (e.g., median income of a particular county) made available over populations over 65,000 and 5-year period estimates are made available for all published geographies (see, e.g., Torrieri 2007, for more information).

A difficulty with using ACS period estimates published over predefined geographies is that it is difficult to infer fine-level (i.e., household) information. As a result, the ACS provides a public-use micro-sample (PUMS) over public-use micro-areas (PUMAs). PUMS consists of individual and household information within each PUMA, where the location of the household within the PUMA is not released to the public. In this section, we focus on household level PUMS found within one particular PUMA; namely the PUMA that covers the metropolitan area of Tallahassee Florida (labeled as PUMA number 00701).

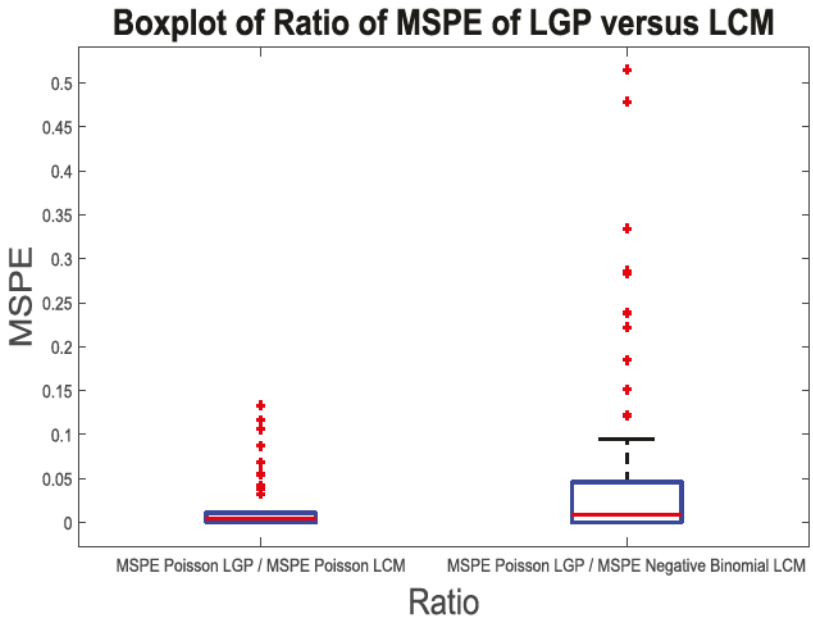

Figure 5. The response is the ratio between mean squared prediction error using the LGP model and the mean squared prediction error of the LCM. We hold out roughly 5% of the observations. A boxplot is displayed over 50 different hold-out observations. Values less than one suggest that the LCM outperforms the LGP. The left boxplot represents the 50 ratios of the MSPE using the Poisson LCM and the MSPE using the Poisson LGP. The right boxplot represents the 50 ratios of the MSPE using the negative binomial LCM and the MSPE using the Poisson LGP.

Consider 2005–2009 PUMS estimates of the number of individuals living in a household contained within PUMA 00701. This is a fairly large dataset (for multivariate statistics) consisting of 4537 observations. An important inferential goal, besides giving an illustration of the LCM, is to accurately predict vacant households (i.e., predict zero people living in a household). Vacant households exhaust resources for those conducting surveys, and is of practical interest to the US Census Bureau (see *http://www.census.gov/en.html*).

We would expect the number of individuals living in a household to be spatially correlated, since certain neighborhoods within Tallahassee are known to be more attractive for those with a family, and hence, have more people living within a household in these neighborhoods. However, the spatial correlation cannot be leveraged, since the location *within* PUMAs are not publicly available. Consequently, we model the dependencies within the PUMS using a generic multivariate distribution; namely, the CM distribution. In particular, we assume that the data follows a LCM. We consider three types of LCMs: the first is a Poisson LCM (i.e., $j = k = 3$), the second is a negative binomial LCM (i.e., $j = k = 2$), and the third is a Poisson LGP (i.e., $j = 3$ and $k = 4$). There are a large number of potential covariates (there are 358 in total) including fuel cost of the household, number of bedrooms in the household, and lot size, among others. For illustration, we picked a small subset of covariates using least angle regression (Efron et al. 2004), which lead to 41 covariates. We consider defining each covariate as the coefficient of the random effects (i.e., a column of $\boldsymbol{\Psi}$) so that $r = 41$. Additionally, we include an intercept as a fixed effect (i.e., $\mathbf{X} = \mathbf{J}_{p,1}$). Convergence of the MCMC algorithm was assessed visually using trace plots, and no lack of convergence was detected.

To assess the quality of the predictions we randomly selected 216 observations (roughly 5% of the data). Using the remaining data, we produce estimates of the mean number of individuals living in a household for the 216 observations. As an example, see Table 2 where we display the hold-out dataset and the corresponding Poisson LCM predictions that were based on the remaining 4321 observations. Here, we see that a majority of the rounded (to the nearest integer) predictions are *exactly equal* to the hold-out data. In fact, $163/226 \approx 72\%$ of the 226 hold-out dataset are *exactly* equal to the corresponding rounded predicted value, and the remaining 30% are within two counts of the corresponding hold-out data value. Furthermore, we are able to very accurately predict an empty household, which may have implications for sampling done by the US Census Bureau.

This hold-out study was repeated 50 times, and the results of the MSPE between the hold-out data and the rounded predictions are presented in Figure 5. Here, we fit the LGP (or a Bayesian GLM) using the R-package `MCMCglmm` and the function "MCMCglmm" (Hadfield 2016), and the remaining models were fitted using the Matlab (Version 9) code in the supplementary materials. Here, we see that both the Poisson LCM and the negative binomial LCM outperforms the Poisson LGP. However, the negative binomial LCM performs worse than the Poisson LCM. The pairwise p-values for a paired t-tests (using 50 MSPE values as the response) are as follows: a one-sided test between the Poisson LCM and the Poisson LGP resulted in a p-value of 0.033; a one-sided test between the negative binomial LCM and the Poisson LGP resulted in a p-value of 0.03; and a one-sided test between the negative binomial LCM and the Poisson LCM resulted in a p-value of 4.11×10^{-24}. We found that the results for the negative binomial LCM to be sensitive to the prior on b (i.e., the coefficient of the unit log-partition function); hence, we suggest using the Poisson LCM instead of the negative binomial LCM. The LGP and LCM both were computed in under or slightly over a minute on average (across holdout data). The LGP on average took 0.9 min, while

the Poisson LCM and the negative binomial LCM took slightly longer at 0.93 min and 3.2 min, respectively. The speed of all three methods appears comparable.

3.5. An Application to Moderate Resolution Imaging Spectroradiometer Cloud Data

On December 18, 1999 the National Aeronautics and Space Administration (NASA) launched the Terra satellite, which is part of the Earth Observing System (EOS). The Moderate Resolution Imaging Spectroradiometer (MODIS) is a remote sensing instrument attached to the Terra satellite and collects information on many environmental processes. In particular, the MODIS instrument converts spectral radiances into a level-2 (i.e., 1 km $\times$ 1 km spatial resolution) cloud mask using cloud detection algorithms. These cloud detection algorithms cannot perfectly identify the presence of a cloud at each 1 km $\times$ 1 km region. Sengupta et al. (2012) cast this as a big spatial data problem as, visually speaking, spatial correlations appear to be present (i.e., nearby observations tend to be more similar) and $n = 2{,}748{,}620$ is large.

In this article, we consider fitting a Bernoulli LCM (i.e., $j = k = 2$ and $b = 1$) to the MODIS level-2 cloud mask data from Sengupta et al. (2012). This LCM is implemented using the code in the supplementary materials. We use the same covariates and radial basis functions in Sengupta et al. (2012). Specifically, let $\mathbf{s}_1, \ldots, \mathbf{s}_n \in \mathrm{R}^2$ represent the observed data locations (latitude/longitude) seen in the left-panel in Figure 6. Set $\boldsymbol{\phi}_i = \left(\phi_1(\mathbf{s}_i), \ldots, \phi_j(\mathbf{s}_i)\right)'$, where

$$\phi_j(\mathbf{s}) = \left\{1 - \frac{||s - \mathbf{g}_j||}{w_j}\right\}^2 I(||\mathbf{s} - \mathbf{g}_j|| < w_j); \quad j = 1, \ldots, r,$$

where $\mathbf{g}_j$, $j = 1, \ldots, r$, is the aforementioned knot points. This radial basis function is referred to as a bisquare function (Cressie and Johannesson 2008). The knot locations are divided into three groups called "resolutions." Then w_j is set equal to 1.5 times the shortest great arc distance between the points that are in the same resolution as $\mathbf{g}_j$. Sengupta et al. (2012) chose $r = 137$ knots to have a "quad-tree" structure (or equally spaced structure), where the knots of the different resolution all differ from one another (see, e.g., Cressie, Shi, and Kang 2010b, among others).

The posterior predicted value of the probability of a clear sky is given in Figure 6 along with the posterior variance. The posterior predicted probabilities reflect the general pattern of the data. Also, the posterior variances are larger at larger posterior predicted probabilities, which is to be expected as more extreme probabilities tend to be more difficult to estimate. Thus, Figure 6 shows that it is possible to fit an LCM to a high-dimensional spatial dataset (with small p) and obtain reasonable in-sample results.

We consider only a single hold-out sample in this section, since the computation times for this dataset are so demanding. Here, we hold out 5% of the observations from the left-most panel of Figure 6 and produced the posterior expected value of the probability of clear skies. We threshold the values of these posterior probabilities around the midpoint of the range to classify either clear sky or cloudy. The LCM took approximately 12 hr to run, the false positive rate is 0.22, and the false negative rate is moderately large at 0.28. We chose to compare these values to the misclassification rates using a standard binary classifier, support vector machines (SVM, Hastie, Tibshirani, and Friedman 2009) fitted using Matlab's "fitcsvm" function. SVM took approximately three days to run, the false positive rate is smaller at 0.11, and the false negative rate is much larger at 0.53. Thus, this single hold-out study suggests that the Bernoulli LCM leads to a classifier that is comparable to the current industry standard, SVM. Moreover, we are able to provide prediction uncertainty. The computation time of SVM (3 days) is also considerably longer than the computation time of the LCM (12 hr).

4. Discussion

We have introduced methodology for jointly modeling dependent non-Gaussian data within the Bayesian framework. This methodology is rooted in the development of new distribution theory for dependent data that makes Bayesian inference possible to implement using a Gibbs sampler; hence, computationally intensive and ad hoc approaches needed for tuning and specifying proposal distributions are not needed. Specifically, we propose a multivariate version of the prior distributions introduced by Diaconis and Ylvisaker (1979). Furthermore, the prior distributions similar to those used by Daniels and Pourahmadi (2002), Chen and Dunson (2003), and Pourahmadi, Daniels, and Park (2007) are adapted to the non-Gaussian setting.

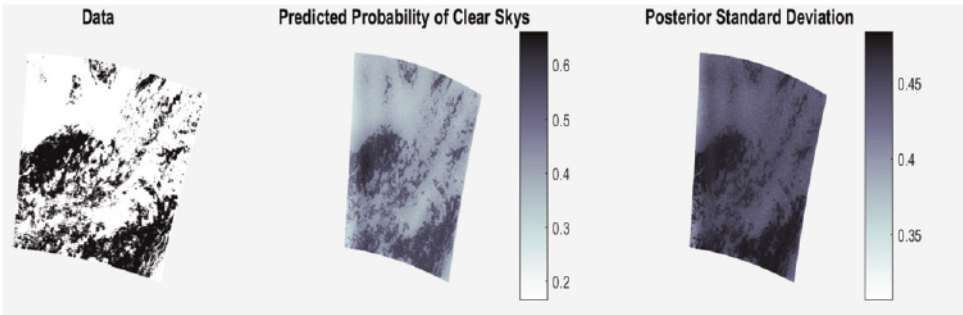

Figure 6. In the left-most panel we have a plot of the data. White locations are observed clouds and black locations are observed clear skies. The middle panel are the posterior expected value of the probability of clear skies using the Bernoulli LCM, and the right-most panel is the corresponding posterior variance of the probability of clear skies. Posterior expected values and variances were computed using a training dataset consisting of 95% of the points in the left-most plot (these points were randomly selected).

Several theoretical results were required to derive this *conjugate multivariate distribution* (CM), and to develop its use for Bayesian inference of dependent data from the natural exponential family. The latter is facilitated through the introduction of the LCM model. In particular, we show that full conditional distributions are of the same form of a conditional distribution of a CM random vector, and provide a way to simulate from a collapsed full conditional distribution. Relationships between the LCM and the LGP also provide motivation for the use of the LCM. In particular, the latent Gaussian process (LGP) is a special case of the LCM. Furthermore, many types of LCMs can be well approximated by a LGP, by specifying certain parameters of a LCM to be "large." This result shows that the LCM is not only computationally easier to implement, but is also more flexible than a LGPs.

Empirical exploration of the Poisson, binomial, Bernoulli, and negative binomial special cases were performed through simulations studies and through analyses of several datasets from a variety of disciplines. These examples indicate very small out-of-sample error when using LCM for prediction, and show gains in predictive performance over the LGP. Additionally, the LCM model is applicable for large datasets (in the application we implemented the LCM on a MODIS level-2 cloud-mask data of size 2,748,620). In the first example, we considered a small dataset of binomial counts of CBPP among herds of cows. We obtained precise predictions and outperformed the LGP computed using a standard R-package. In the second real data analysis section, we predict the number of individuals within a household over the U.S. city of Tallahassee, FL, and obtain very precise estimates (in terms of hold-out error). The predictions were very accurate even though $18/226 \approx 8\%$ of the hold-out dataset consisted of zero counts, which is known to cause difficulties in an LGP (Lambert 2006). In the third application, we obtain posterior predicted probabilities that reflected the pattern of data at observed locations, and a binary classifier that has misclassification rates that are comparable to support vector machines.

Although there are many settings where the LCM improves both precision and computation, there are settings where it would not be feasible to implement the LCM. In particular, we consider one choice of ψ that results in a case where $\mathcal{M}_p^{n+p}$, $\mathcal{M}_r^{n+r}$, and $\mathcal{M}_n^{2n}$ does not guarantee that $\mathbf{x}_i'\boldsymbol{\beta} + \boldsymbol{\phi}_i'\boldsymbol{\eta} + \xi_i \in \mathcal{Y}$ for each i; namely ψ_1, which is the unit log partition function of a gamma data model. In this case, the full conditional distributions are *truncated* CM_c distributions. Thus, in this setting the LCM is most easily implemented by doing a Gibbs sampler with component-wise updates due to the truncated support of the natural parameter. This is computationally less efficient than simply transforming the gamma data to the log scale and fitting an LGP, which can give precise predictions. Additionally, we found that the negative binomial LCM to give poorer predictive results than the Poisson LCM. Thus, we suggest using the Poisson LCM when analyzing unbounded count values instead of the negative binomial LCM.

As discussed in Section 1, a general modeling framework for dependent data that can model non-Gaussian (natural exponential family) data as easily as Gaussian data, has important implications for applied statistics. Nevertheless, there are also many opportunities for new methodological results that are exciting, since a special case of our framework (i.e., the LGP) has been the central methodological tool used in the dependent data literature. In particular, we are interested in developing the LCM model within "more specific" dependent data settings such as time-series, spatial, spatio-temporal, and multivariate spatio-temporal arenas.

Acknowledgments

We would like to express our sincere gratitude to the editors, the associate editor, and the referees for their very helpful comments that improved this manuscript. We would also like to thank Drs. Matthew Simpson of SAS Inc. and Erin Schliep at the University of Missouri for helpful discussions.

Funding

This research was partially supported by the U.S. National Science Foundation (NSF) and the U.S. Census Bureau under NSF grant SES-1132031, funded through the NSF-Census Research Network (NCRN) program. This article is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the authors and not those of the NSF or the U.S. Census Bureau.

References

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2015), *Hierarchical Modeling and Analysis for Spatial Data*, London, UK: Chapman and Hall. [2037]

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008), "Gaussian Predictive Process Models for Large Spatial Data Sets," *Journal of the Royal Statistical Society*, Series B, 70, 825–848. [2038]

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, G., Dai, B., Grothendieck, G., and Green, P. (2017), "Package 'lme4'," available at *https://cran.r-project.org/web/packages/lme4/lme4.pdf*. [2046]

Bradley, J. R., Cressie, N., and Shi, T. (2011), "Selection of Rank and Basis Functions in the Spatial Random Effects Model," in *Proceedings of the 2011 Joint Statistical Meetings*, Alexandria, VA: American Statistical Association, pp. 3393–3406. [2041,2046]

——— (2016), "A Comparison of Spatial Predictors When Datasets Could Be Very Large," *Statistics Surveys*, 10, 100–131. [2038,2041]

Bradley, J., Holan, S., and Wikle, C. (2015), "Multivariate Spatio-Temporal Models for High-Dimensional Areal Data With Application to Longitudinal Employer-Household Dynamics," *The Annals of Applied Statistics*, 9, 1761–1791. [2038]

——— (2018), "Computationally Efficient Distribution Theory for Bayesian Inference of High-Dimensional Dependent Count-Valued Data" (with discussion), *Bayesian Analysis*, 13, 253–310. [2037,2038, 2039,2042,2043]

Bradley, J., Wikle, C., and Holan, S. (2016), "Bayesian Spatial Change of Support for Count-Valued Survey Data," *Journal of the American Statistical Association*, 111, 472–487. [2044]

——— (2017), "Regionalization of Multiscale Spatial Processes Using a Criterion for Spatial Aggregation Error," *Journal of the Royal Statistical Society*, Series B, 79, 815–832. [2041]

Castruccio, S., Ombao, H., and Genton, M. G. (2016), "A Scalable Multi-Resolution Spatio-Temporal Model for Brain Activation and Connectivity in fMRI Data," arXiv no. 1602.02435. [2038]

Chen, M. H., and Ibrahim, J. G. (2003), "Conjugate Priors for Generalized Linear Models," *Statistica Sinica*, 13, 461–476. [2037]

Chen, Z., and Dunson, D. B. (2003), "Random Effects Selection in Linear Mixed Models," *Biometrics*, 59, 762–769. [2038,2044,2049]

Conway, R. W., and Maxwell, W. L. (1962), "A Queuing Model With State Dependent Service Rates," *Journal of the American Statistical Association*, 12, 132–136. [2044]

Cox, T. F. (2005), *An Introduction to Multivariate Data Analysis*, London: Hodder Arnold. [2038]

Cressie, N., and Johannesson, G. (2006), "Spatial Prediction for Massive Data Sets," in *Australian Academy of Science Elizabeth and Frederick White Conference*, Australian Academy of Science, Canberra, pp. 1–11. [2038]

——— (2008), "Fixed Rank Kriging for Very Large Spatial Data Sets," *Journal of the Royal Statistical Society*, Series B, 70, 209–226. [2038,2049]

Cressie, N., Shi, T., and Kang, E. L. (2010a), "Fixed Rank Filtering for Spatio-Temporal Data," *Journal of Computational and Graphical Statistics*, 19, 724–745. [2038]

——— (2010b), "Using Temporal Variability to Improve Spatial Mapping With Application to Satellite Data," *Canadian Journal of Statistics*, 38, 271–289. [2038,2049]

Cressie, N., and Wikle, C. K. (2011), *Statistics for Spatio-Temporal Data*, Hoboken, NJ: Wiley. [2037]

Daniell, P. J. (1919), "Integrals in an Infinite Number of Dimensions," *Annals of Mathematics*, 20, 281–288. [2042]

Daniels, M. J., and Pourahmadi, M. (2002), "Dynamic Models and Bayesian Analysis of Covariance Matrices in Longitudinal Data," *Biometrika*, 89, 553–566. [2038,2044,2049]

De Oliveira, V. (2013), "Hierarchical Poisson Models for Spatial Count Data," *Journal of Multivariate Analysis*, 122, 393–408. [2038]

Demirhan, H., and Hamurkaroglu, C. (2011), "On a Multivariate Log-Gamma Distribution and the Use of the Distribution in the Bayesian Analysis," *Journal of Statistical Planning and Inference*, 141, 1141–1152. [2038]

Diaconis, P., and Ylvisaker, D. (1979), "Conjugate Priors for Exponential Families," *The Annals of Statistics*, 17, 269–281. [2037,2039,2040,2044, 2049]

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998), "Model-Based Geostatistics," *Journal of the Royal Statistical Society*, Series C, 47, 299–350. [2037]

Donoho, D., and Johnstone, I. (1994), "Ideal Spatial Adaptation by Wavelet Shrinkage," *Biometrika*, 81, 425–455. [2041]

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," *Annals of Statistics*, 32, 407–499. [2048]

Everitt, B., and Hothorn, T. (2011), *An Introduction to Applied Multivariate Analysis With R*, New York: Springer. [2038]

Finley, A. O., Sang, H., Banerjee, S., and Gelfand, A. E. (2009), "Improving the Performance of Predictive Process Modeling for Large Datasets," *Computational Statistics and Data Analysis*, 53, 2873–2884. [2038]

Frühwirth-Schnatter, S., and Wagner, H. (2006), "Auxiliary Mixture Sampling for Parameter-Driven Models of Time Series of Counts With Applications to State Space Modelling," *Biometrika*, 93, 827–841. [2038]

Gelfand, A. E., and Schliep, E. M. (2016), "Spatial Statistics and Gaussian Processes: A Beautiful Marriage," *Spatial Statistics*, 18, 86–104. [2042]

Gelman, A. (2006), "Prior Distributions for Variance Parameters in Hierarchical Models," *Bayesian Analysis*, 1, 515–533. [2038,2041,2044]

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013), *Bayesian Data Analysis* (3rd ed.), Boca Raton, FL: Chapman and Hall/CRC. [2044]

Hadfield, J. (2016), "Package 'MCMCglmm,'" available at *https://cran.r-project.org/web/packages/MCMCglmm/MCMCglmm.pdf*. [2048]

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York: Springer. [2038,2049]

Henao, R. G. (2009), "Geostatistical Analysis of Functional Data," Ph.D. thesis, Universitat Politècnica de Catalunya. [2041]

Hodges, J. (2013), *Richly Parameterized Linear Models: Additive, Time Series, and Spatial Models Using Random Effects*, Boca Raton, FL: Chapman and Hall/CRC. [2041]

Holan, S. H., and Wikle, C. K. (2016), "Hierarchical Dynamic Generalized Linear Mixed Models for Discrete-Valued Spatio-Temporal Data," in *Handbook of Discrete-Valued Time Series*, eds. R. A. Davis, S. H. Holan, R. Lund, and N. Ravishanker, Boca Raton, FL: CRC Press. [2037]

Hooten, M. B., Larsen, D. R., and Wikle, C. K. (2003), "Predicting the Spatial Distribution of Ground Flora on Large Domains Using a Hierarchical Bayesian Model," *Landscape Ecology*, 18, 487–502. [2038]

Hu, G., and Bradley, J. R. (2018), "A Bayesian Spatial–Temporal Model With Latent Multivariate Log-Gamma Random Effects With Application to Earthquake Magnitudes," *Stat*, 7, e179. [2038]

Huang, H., and Sun, Y. (2003), "Hierarchical Low Rank Approximation of Likelihoods for Large Spatial Datasets," *Journal of Computational and Graphical Statistics*, 27, 110–118. [2045]

Jolliffe, I. T. (2002), *Principal Components Analysis* (2nd ed.), New York: Springer Verlag. [2038]

Kang, E. L., and Cressie, N. (2011), "Bayesian Inference for the Spatial Random Effects Model," *Journal of the American Statistical Association*, 106, 972–983. [2038]

Katzfuss, M., and Cressie, N. (2011), "Spatio-Temporal Smoothing and EM Estimation for Massive Remote-Sensing Data Sets," *Journal of Time Series Analysis*, 32, 430–446. [2038]

——— (2012), "Bayesian Hierarchical Spatio-Temporal Smoothing for Very Large Datasets," *Environmetrics*, 23, 94–107. [2038]

Kolmogorov, A. N. (1933), *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Berlin: Springer. [2042]

Kotz, S., Balakrishnan, N., and Johnson, N. (2000), *Continuous Multivariate Distributions, Volume 1: Models and Applications*, New York: Wiley. [2038]

Lambert, D. (2006), "Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing," *Technometrics*, 34, 1–14. [2050]

Lange, K., Papp, J. C., Sinsheimer, J. S., and Sobel, E. M. (2014), "Next-Generation Statistical Genetics: Modeling, Penalization, and Optimization in High-Dimensional Data," *Annual Review of Statistics and Its Application*, 1, 279–300. [2038]

Lee, Y., and Nelder, J. A. (1974), "Double Hierarchical Generalized Linear Models With Discussion," *Applied Statistics*, 55, 129–185. [2038]

——— (1996), "Hierarchical Generalized Linear Models" (with discussion), *Journal of the Royal Statistical Society*, Series B, 58, 619–678. [2038]

——— (2000), "HGLMs for Analysis of Correlated Non-Normal Data," in *COMPSTAT: Proceedings in Computational Statistics 14th Symposium, 2000*, Utrecht, the Netherlands, eds. J. G. Bethlehem and P. G. M. van der Heijden, pp. 97–107. [2038]

——— (2001), "Modelling and Analysing Correlated Non-Normal Data," *Statistical Modelling*, 1, 3–16. [2038]

Lehmann, E., and Casella, G. (1998), *Theory of Point Estimation* (2nd ed.), New York: Springer. [2039]

Lesnoff, M., Laval, G., Bonnet, P., Abdicho, S., Workalemahu, A., Kifle, D., Peyraud, A., Lancelot, R., and Thiaucourt, F. (2004), "Within-Herd Spread of Contagious Bovine Pleuropneumonia in Ethiopian Highlands," *Preventive Veterinary Medicine*, 64, 27–40. [2046]

Liu, J. S. (1994), "The Collapsed Gibbs Sampler in Bayesian Computations With Applications to a Gene Regulation Problem," *Journal of the American Statistical Association*, 89, 958–966. [2042]

Matloff, N. (2016), "Big-n Versus Big-p in Big Data," in *Handbook of Big Data*, eds. P. Buhlmann, P. Drineas, M. Kane, and M. van der Laan, Boca Raton, FL: Chapman and Hall, pp. 21–31. [2038,2045]

Neal, R. M. (2011), "MCMC Using Hamiltonian Dynamics," in *Handbook of Markov Chain Monte Carlo*, eds. S. Brooks, A. Gelman, G. L. Jones, and X. Meng, Boca Raton, FL: Chapman and Hall, pp. 113–160. [2038]

Nieto-Barajas, L. E., and Huerta, G. (2017), "Spatio-Temporal Pareto Modelling of Heavy-Tail Data," *Spatial Statistics*, 20, 92–109. [2038]

Nychka, D. W. (2001), "Spatial Process Estimates as Smoothers," in *Smoothing and Regression: Approaches, Computation and Applications* (rev. ed.), ed. M. G. Schmiek, New York: Wiley, pp. 393–424. [2041]

O'Hara, R. B., and Sillanpaa, M. J. (2009), "A Review of Bayesian Variable Selection Methods: What, How and Which," *Bayesian Analysis*, 4, 85–118. [2041]

Pourahmadi, M., Daniels, M. J., and Park, T. (2007), "Simultaneous Modelling of the Cholesky Decomposition of Several Covariance Matrices," *Journal of Multivariate Analysis*, 98, 568–587. [2038,2044,2049]

Ravishanker, N., and Dey, D. K. (2002), *A First Course in Linear Model Theory*, Boca Raton, FL: Chapman and Hall/CRC. [2040]

Rue, H., Martino, S., and Chopin, N. (2009), "Approximate Bayesian Inference for Latent Gaussian Models Using Integrated Nested Laplace Approximations," *Journal of the Royal Statistical Society*, Series B, 71, 319–392. [2037,2038]

Sengupta, A., Cressie, N., Frey, R., and Kahn, B. (2012), "Statistical Modeling of MODIS Cloud Data Using the Spatial Random Effects Model," in *Proceedings of the Joint Statistical Meetings*, Alexandria, VA: American Statistical Association, pp. 3111–3123. [2038,2041,2049]

Shi, T., and Cressie, N. (2007), "Global Statistical Analysis of MISR Aerosol Data: A Massive Data Product From NASA's Terra Satellite," *Environmetrics*, 18, 665–680. [2038]

Sun, Y., and Li, B. (2012), "Geostatistics for Large Datasets," in *Space-Time Processes and Challenges Related to Environmental Problems*, eds. E. Porcu, J. M. Montero, and M. Schlather, Berlin, Heidelberg: Springer, pp. 55–77. [2038]

Torrieri, N. (2007), "America Is Changing, and So Is the Census: The American Community Survey," *The American Statistician*, 61, 16–21. [2047]

Wahba, G. (1990), *Spline Models for Observational Data*, Philadelphia, PA: Society for Industrial and Applied Mathematics. [2041]

Wikle, C. K. (2010), "Low-Rank Representations for Spatial Processes," in *Handbook of Spatial Statistics*, eds. A. E. Gelfand, P. J. Diggle, M. Fuentes, and P. Guttorp, Boca Raton, FL: Chapman & Hall/CRC Press, pp. 107–118. [2041]

Wikle, C. K., and Anderson, C. J. (2003), "Cllimatological Analysis of Tornado Report Counts Using a Hierarchical Bayesian Spatio-Temporal Model," *Journal of Geophysical Research-Atmospheres*, 108, 9005. [2038]

Wikle, C. K., and Cressie, N. (1999), "A Dimension-Reduced Approach to Space-Time Kalman Filtering," *Biometrika*, 86, 815–829. [2038]

Wilson, A., and Reich, B. J. (2014), "Confounder Selection via Penalized Credible Regions," *Biometrics*, 70, 852–861. [2041]

Wolpert, R., and Ickstadt, K. (1998), "Poisson/Gamma Random Field Models for Spatial Statistics," *Biometrika*, 85, 251–267. [2038]

Wu, G., Holan, S. H., and Wikle, C. K. (2013), "Hierarchical Bayesian Spatio-Temporal Conway-Maxwell Poisson Models With Dynamic Dispersion," *Journal of Agricultural, Biological, and Environmental Statistics*, 18, 335–356. [2038]

Yang, R., and Berger, J. (1994), "Estimation of a Covariance Matrix Using the Reference Prior," *Annals of Statistics*, 22, 1195–1211. [2044]

Zhang, L., Guindani, M., and Vannucci, M. (2015), "Bayesian Models for fMRI Data Analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, 7, 21–41. [2038]

Zhou, M., and Carin, L. (2015), "Negative Binomial Process Count and Mixture Modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37, 307–320. [2041]