Bayesian Analysis (2018) **13**, Number 1, pp. 253–310

Computationally Efficient Multivariate Spatio-Temporal Models for High-Dimensional Count-Valued Data (with Discussion)

Jonathan R. Bradley[*], Scott H. Holan[†,‡], and Christopher K. Wikle[§]

Abstract. We introduce a computationally efficient Bayesian model for predicting high-dimensional dependent count-valued data. In this setting, the Poisson data model with a latent Gaussian process model has become the de facto model. However, this model can be difficult to use in high dimensional settings, where the data may be tabulated over different variables, geographic regions, and times. These computational difficulties are further exacerbated by acknowledging that count-valued data are naturally non-Gaussian. Thus, many of the current approaches, in Bayesian inference, require one to carefully calibrate a Markov chain Monte Carlo (MCMC) technique. We avoid MCMC methods that require tuning by developing a new conjugate multivariate distribution. Specifically, we introduce a multivariate log-gamma distribution and provide substantial methodological development of independent interest including: results regarding conditional distributions, marginal distributions, an asymptotic relationship with the multivariate normal distribution, and full-conditional distributions for a Gibbs sampler. To incorporate dependence between variables, regions, and time points, a multivariate spatio-temporal mixed effects model (MSTM) is used. To demonstrate our methodology we use data obtained from the US Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) program. In particular, our approach is motivated by the LEHD's Quarterly Workforce Indicators (QWIs), which constitute current estimates of important US economic variables.

MSC 2010 subject classifications: Primary 62H11; secondary 62P12.

Keywords: aggregation, American Community Survey, Bayesian hierarchical model, big data, Longitudinal Employer-Household Dynamics (LEHD) program, Markov chain Monte Carlo, non-Gaussian, Quarterly Workforce Indicators.

1 Introduction

Latent Gaussian process (LGP) models have become a standard tool for modeling dependencies in count-valued and other non-Gaussian datasets; see Diggle et al. (1998), Gelfand and Smith (2007), Rue et al. (2009), Sections 4.1.2 and 7.1.5 of Cressie and Wikle (2011), Holan and Wikle (2016), Gelfand and Schliep (2016), and the references therein. One standard LGP for modeling dependent count-data defines a Poisson data

[*] Corresponding author. Department of Statistics, Florida State University, 117 N. Woodward Ave., Tallahassee, FL 32306-4330, bradley@stat.fsu.edu

[†] Department of Statistics, University of Missouri, 146 Middlebush Hall, Columbia, MO 65211-6100

[‡] U.S. Census Bureau, 4600 Silver Hill Road, Washington, D.C., 20233-9100

[§] Department of Statistics, University of Missouri, 146 Middlebush Hall, Columbia, MO 65211-6100

https://doi.org/10.1214/17-BA1069

model (i.e., the data given latent processes), and a Gaussian process model. This strategy allows one to combine a familiar model for counts (i.e., a Poisson distribution), and a familiar model for dependent data (i.e., a Gaussian process). The LGP strategy is extremely prevalent in the dependent data literature, and branches out to many other subdisciplines in statistics (e.g., see Lee and Nelder (2000) and Lee and Nelder (2001) in the generalized linear models literature, and Lawson (2006) in the disease mapping literature).

Unfortunately, this pervasive strategy is becoming increasingly difficult to use, since the size of modern datasets are typically extremely large (e.g., see Bradley et al., 2014, for a discussion). In the Bayesian setting, this leads to all-too-common computational difficulties. Namely, defining proposal distributions for Markov chain Monte Carlo (MCMC) algorithms can be extremely difficult in the high-dimensional setting (e.g., see Rue et al., 2009, for a discussion on convergence issues of MCMC algorithms for LGPs). Thus, the primary goal of this article is to introduce new multivariate distribution theory for count-data that leads to a Gibbs sampler with full-conditional distributions that are straightforward to simulate from.

Ideally, we would like to obtain convergence of the MCMC algorithm the first time it is executed, low prediction error, and a wait-time that rivals the Gaussian data setting. A multivariate spatio-temporal model for high-dimensional count-valued data that achieves this wish-list requires significant methodological development to avoid the computational issues that naturally arise when using various *ad-hoc* Metropolis-Hastings algorithms. Specifically, we introduce a *multivariate log-gamma distribution*[1] to use in place of the multivariate normal distribution within the LGP paradigm. The main motivating feature of this modeling framework is that it incorporates dependency and results in full-conditional distributions (within a Gibbs sampler) that are easy to simulate from.

Tuning MCMC algorithms for implementing an LGP is a well-known and recurring problem for applied Bayesian statisticians (e.g., see Rue et al., 2009). Consequently, the computationally efficient distribution theory proposed here could have an important impact on a number of different communities, and is therefore of independent interest. For example, count-valued datasets are ubiquitous within the official statistics setting. A clear majority of the US Census Bureau's American Community Survey (ACS) period estimates are count-valued (e.g., see `http://factfinder.census.gov/`). High-dimensional count-valued data are also widespread in ecology (e.g., see Royle and Wikle, 2005; Wu et al., 2013, among others) and climatology (e.g., see Wikle and Anderson, 2003). Hence, the methodology presented here offers an exciting avenue that makes new research for modeling correlated count-valued data practical for modern big datasets.

There are other choices besides the LGP strategy available in the literature (e.g., see Nieto-Barajas and Huerta (2017) for an example model for pareto data). For count-valued data, an important alternative to the LGP paradigm was proposed by Wolpert and Ickstadt (1998), who introduced a (spatial) convolution of gamma random variables,

[1] To avoid confusion, we note that a "log-gamma" random variable is the natural log of a gamma random variable, while a "log-normal" random variable represents the exponential of a normal random variable.

and provide a data augmentation scheme for Gibbs sampling in the spatial-only setting. In fact there are strong connections between our approach and the method in Wolpert and Ickstadt (1998). In particular, Wolpert and Ickstadt (1998) convolve gamma random variables using a spatial kernel, while we work on a transformed space and take a linear combination of log-gamma random variables. However, their framework can only be applied for smaller-dimensional spatial-only settings.

In the non-spatially referenced settings some have used different types of multivariate log-gamma (and gamma) distributions as an alternative to the multivariate normal distribution (e.g., see Lee and Nelder, 1974; Kotz et al., 2000; Demirhan and Hamurkaroglu, 2011). The common formulation of a multivariate log-gamma distribution, starts with defining a multivariate gamma distribution, which is then transformed to the log-scale (Demirhan and Hamurkaroglu, 2011). Multivariate gamma distributions have a rich history (e.g., see Johnson and Wichern, 1999), and are often formulated by defining a multivariate moment generating function and using the inverse Laplace transform to define a probability density function (for a commonly used multivariate gamma moment generating function see Vere-Jones (1967), Moran and Vere-Jones (1969), and Griffiths (1984), and see Bernardoff (2006) for a more general formulation). However, we have found that transforming a multivariate gamma distribution leads to complications for Gibbs sampling. For example, in Demirhan and Hamurkaroglu (2011), their full-conditional distributions only have a known form when performing component-wise updating, and these component-wise full-conditional distributions are approximated. Instead, we develop a multivariate log-gamma distribution by defining a discrete convolution of independent log-gamma random variables. This approach leads to block-wise full-conditional distributions that are easy to simulate from.

It is not our intent to hold Gibbs sampling as an ideal. There are several Bayesian computation techniques that have shown to perform very well. For example, Hamiltonian MCMC (HMC; Neal, 2011) and Integrated nested Laplace approximations (INLA; Rue et al., 2009) are two such techniques. The use of HMC can lead to a noticeable increase in the efficiency. This is partially because HMC allows one to jointly update all parameters, while the Gibbs sampler used in this paper imposes block updates. INLA is not an MCMC approach, and involves fast numerical integration of a Laplace approximate. However, in this article we find the Gibbs sampler an attractive approach because it is simple to implement. Furthermore, our proposed distribution theory allows us to capitalize on this simplicity.

We use a motivating dataset to demonstrate the wide range of complex and modern problems that this new distribution theory can handle. Specifically, the Longitudinal Employer Household Dynamics (LEHD) program's Quarterly Workforce Indicators (QWIs), which has become a key data source for understanding US economy (e.g., see Abowd et al. (2009), Abowd et al. (2013), and Bradley et al. (2015), and the references therein). Many of the QWIs are suppressed due to disclosure limitations and because some states fail to sign the required Memorandum of Understanding (MOU) for a given year (Abowd et al., 2009, Sections 5.5.1 and 5.6). Additionally, the QWIs currently do not have measures of uncertainty associated with them, which limits their use. (By "uncertainty," we mean a margin of error or mean squared prediction error associated with the predictions.) Thus, there is a need to estimate missing values and measures of error.

Recently, Bradley et al. (2015) efficiently modeled 7,530,037 *Gaussian* QWIs jointly, over 40 variables (20 industries and two genders), 92 time-points, and 3,145 different counties. To do this, they developed a type of dynamic spatio-temporal model that is referred to as the multivariate spatio-temporal mixed effects model (MSTM). The MSTM has been used to produce estimates of continuous QWIs (e.g., average quarterly income) that have complete spatio-temporal coverage and corresponding measures of uncertainty. These advancements, although important for the LEHD community, have limited utility on QWIs. This is because approximately 70% of the QWIs are count-valued (e.g., county-level beginning of quarter employment), which implies that the MSTM is applicable to a small portion of the entire scope of the QWIs.

There are many ways that one may use the multivariate log gamma distribution to model the multivariate spatio-temporal dependencies within count-valued QWIs. For example, a naive approach (for this particular dataset) would be to assume separability between each variable, region, and time (Daniels et al., 2006). However, QWIs exhibit complex dependencies that are non-separable, asymmetric, and non-stationary (Bradley et al., 2015). Thus, we incorporate the Moran's I (MI) basis functions, MI propagator matrix, and MI prior distribution from Bradley et al. (2015) to better describe the dependency of latent processes. The resulting hierarchical statistical model is called the *Poisson multivariate spatio-temoral mixed effects model* (P-MSTM).

The remainder of this article is organized as follows. In Section 2 we introduce a multivariate log-gamma distribution and provide the necessary technical development for this distribution. In Section 3, we use this new distribution theory to define the P-MSTM. Section 4 provides a simulation study, and an illustration where we efficiently *jointly analyze/model* 4,089,755 count-valued QWIs obtained from the US Census Bureau's LEHD program. Finally, Section 5 contains discussion. For convenience of exposition, proofs of the technical results are provided in Supplemental Materials (Bradley et al., 2017).

2 Conjugate Distributions for Correlated Poisson Data

The rudimentary quantity in our development of the multivariate log-gamma distribution is the (univariate) log-gamma random variable q (Prentice, 1974; Kotz et al., 2000; Crooks, 2015), where

$$q \equiv \log(\gamma), \tag{1}$$

and γ is a gamma random variable with shape parameter $\alpha > 0$ and rate parameter $\kappa > 0$. There are many relationships between the log-gamma distribution and other distributions including the Gumbel distribution, the Amoroso distribution, and the normal distribution (e.g., see Crooks, 2015). These relationships are derived by considering special cases of the probability density function (pdf) associated with q in (1). The mean and variance of the log gamma random variables are well known (e.g., see Prentice, 1974, among others) and given by

$$\begin{aligned} E[q] &= \omega_0(\alpha) + \log(\kappa) \\ Var[q] &= \omega_1(\alpha). \end{aligned}$$

The function ω_k, for non-negative integer k, is the polygamma function, and for a real value z we have that $\omega_k(z) \equiv \frac{d^{k+1}}{dz^{k+1}}\log(\Gamma(z))$.

Straightforward change-of-variable techniques lead to the following expression for the pdf of q,

$$f(q|\alpha, \kappa) = \frac{\kappa^{\alpha}}{\Gamma(\alpha)}\exp\left\{\alpha q - \kappa\exp(q)\right\}; \quad q \in \mathbb{R}, \tag{2}$$

where f will be used to denote a generic pdf and $\mathrm{LG}(\alpha, \kappa)$ denotes a shorthand for the pdf in (2). The importance of the log-gamma random variable for our purpose of modeling count-valued data is transparent in the univariate setting. Let $Z|q \sim \mathrm{Pois}\{\exp(q)\}$, and notice that

$$f(Z|q) \propto \exp\left\{Zq - \exp(q)\right\}. \tag{3}$$

It is immediate from (2) and (3) that

$$q|Z, \alpha, \kappa \sim \mathrm{LG}\left\{Z + \alpha, 1 + \kappa\right\}. \tag{4}$$

This conjugacy between the Poisson distribution and the log-gamma distribution motivates us to develop a multivariate version of the log-gamma distribution to model multivariate spatio-temporal count-valued data.

2.1 The Multivariate Log-Gamma Distribution

Let the m-dimensional random vector $\mathbf{w} = (w_1, \ldots, w_m)'$ consist of m mutually independent log-gamma random variables such that $w_i \sim \mathrm{LG}(\alpha_i, \kappa_i)$ for $i = 1, \ldots, m$. Then, define

$$\mathbf{q} = \mathbf{c} + \mathbf{V}\mathbf{w}, \tag{5}$$

where the matrix $\mathbf{V} \in \mathbb{R}^m \times \mathbb{R}^m$ and $\mathbf{c} = (c_1, \ldots, c_m)' \in \mathbb{R}^m$. Call $\mathbf{q}$ in (5) a multivariate log-gamma (MLG) random vector. The linear combination in (5) is similar to the derivation of the multivariate normal distribution; that is, if one replaces $\mathbf{w}$ with an m-dimensional random vector consisting of independent and identically standard normal random variables, one obtains the multivariate normal distribution (e.g., see Anderson, 1958; Johnson and Wichern, 1999, among others) with mean $\mathbf{c}$ and covariance $\mathbf{V}\mathbf{V}'$. However, an important difference in our approach is that we have additional shape and rate parameters associated with each component, which could possibly be different. This suggests that the MLG distribution is more flexible than the Gaussian distribution.

Equation (5) is inspired by the method presented in Wolpert and Ickstadt (1998). Specifically, if the integral in Wolpert and Ickstadt (1998)'s (3.1) is discretized then you obtain $\mathbf{v}'\mathbf{w}^*$, where $\mathbf{v}$ is an n-dimensional vector that consists of spatial kernel operators evaluated at a particular spatial location and $\mathbf{w}^*$ consists of independent gamma (not log-gamma) random variables. Our choice to define the linear combination in log space is important because it will allow us to avoid computationally expensive data augmentation steps within a Gibbs Sampler.

Now, to use the MLG distribution in a Bayesian context, we require its pdf, which is formally stated in Theorem 1.

Theorem 1. *Let $\boldsymbol{q} = \boldsymbol{c} + \boldsymbol{V}\boldsymbol{w}$, where $\boldsymbol{c} \in \mathbb{R}^m$, the $m \times m$ real valued matrix $\boldsymbol{V}$ is invertible, and the m-dimensional random vector $\boldsymbol{w} = (w_1, \ldots, w_m)'$ consists of m mutually independent log-gamma random variables such that $w_i \sim \mathrm{LG}(\alpha_i, \kappa_i)$ for $i = 1, \ldots, m$.*

i. Then $\boldsymbol{q}$ has the following pdf:

$$f(\boldsymbol{q}|\boldsymbol{c}, \boldsymbol{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa}) = \frac{1}{\det(\boldsymbol{V}\boldsymbol{V}')^{1/2}} \left(\prod_{i=1}^{m} \frac{\kappa_i^{\alpha_i}}{\Gamma(\alpha_i)} \right) \exp\left[\boldsymbol{\alpha}' \boldsymbol{V}^{-1}(\boldsymbol{q} - \boldsymbol{c}) - \boldsymbol{\kappa}' \exp\left\{ \boldsymbol{V}^{-1}(\boldsymbol{q} - \boldsymbol{c}) \right\} \right]; \quad \boldsymbol{q} \in \mathbb{R}^m, \tag{6}$$

where "det" represents the determinant function, $\boldsymbol{\alpha} \equiv (\alpha_1, \ldots, \alpha_m)'$ and $\boldsymbol{\kappa} \equiv (\kappa_1, \ldots, \kappa_m)'$.

ii. The mean and variance of $\boldsymbol{q}$ is given by,

$$\begin{aligned} E[\boldsymbol{q}] &= \boldsymbol{c} + \boldsymbol{V}(\omega_0(\boldsymbol{\alpha}) - \log(\boldsymbol{\kappa})) \\ \mathrm{cov}[\boldsymbol{q}] &= \boldsymbol{V} \,\mathrm{diag}(\omega_1(\boldsymbol{\alpha}))\, \boldsymbol{V}', \end{aligned} \tag{7}$$

where for a generic m-dimensional real-valued vector $\boldsymbol{k} = (k_1, \ldots, k_m)'$ let $\mathrm{diag}(\boldsymbol{k})$ be an $m \times m$ dimensional diagonal matrix with main diagonal equal to $\boldsymbol{k}$.

Proof. See Supplemental Appendix A in Bradley et al. (2017). □

Let $\mathrm{MLG}(\mathbf{c}, \mathbf{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa})$ be shorthand for the pdf in (6). When comparing (2), (3), and (6) we see that the univariate log-gamma pdf, the Poisson pdf, and the multivariate log-gamma pdf share a basic structure. Specifically, all three pdfs have an exponential term and a double exponential term. This pattern is the main reason why conjugacy exists between the Poisson distribution and the log gamma distribution, which we take advantage of in subsequent sections.

2.2 Conditional Distributions for Multivariate Log-Gamma Random Vectors

Gibbs sampling from full-conditional distributions will require simulating from conditional distributions of multivariate log-gamma random vectors. Thus, we provide the technical results needed to simulate from these conditional distributions.

Proposition 1. *Let $\boldsymbol{q} \sim \mathrm{MLG}(\boldsymbol{c}, \boldsymbol{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa})$, and let $\boldsymbol{q} = (\boldsymbol{q}_1', \boldsymbol{q}_2')'$, so that $\boldsymbol{q}_1$ is g-dimensional and $\boldsymbol{q}_2$ is $(m-g)$-dimensional. In a similar manner, partition $\boldsymbol{V}^{-1} = [\boldsymbol{H}\ \boldsymbol{B}]$ into an $m \times g$ matrix $\boldsymbol{H}$ and an $m \times (m-g)$ matrix $\boldsymbol{B}$. Then, the conditional pdf of $\boldsymbol{q}_1 | \boldsymbol{q}_2 = \boldsymbol{d}, \boldsymbol{c}, \boldsymbol{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa}$ is given by*

$$f(\boldsymbol{q}_1 | \boldsymbol{q}_2 = \boldsymbol{d}, \boldsymbol{c}, \boldsymbol{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa}) = f(\boldsymbol{q}_1 | \boldsymbol{H}, \boldsymbol{\alpha}, \boldsymbol{\kappa}_{1.2}) = M \ \exp\left\{ \boldsymbol{\alpha}' \boldsymbol{H} \boldsymbol{q}_1 - \boldsymbol{\kappa}_{1.2}' \exp(\boldsymbol{H}\boldsymbol{q}_1) \right\}, \tag{8}$$

where $\boldsymbol{\kappa}_{1.2} \equiv \exp\{\boldsymbol{Bd} - \boldsymbol{V}^{-1}\boldsymbol{c} - \log(\boldsymbol{\kappa})\}$ *and the normalizing constant* M *is*

$$M = \frac{1}{\det(\boldsymbol{V}\boldsymbol{V}')^{1/2}} \left(\prod_{i=1}^{m} \frac{\kappa_i^{\alpha_i}}{\Gamma(\alpha_i)} \right) \frac{\exp\left(\boldsymbol{\alpha}'\boldsymbol{Bd} - \boldsymbol{\alpha}'\boldsymbol{V}^{-1}\boldsymbol{c}\right)}{\left[\int f(\boldsymbol{q}|\boldsymbol{c}, \boldsymbol{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa}) d\boldsymbol{q}_1\right]_{\boldsymbol{q}_2 = \boldsymbol{d}}}.$$

Proof. See Supplemental Appendix A in Bradley et al. (2017). □

Let cMLG$(\mathbf{H}, \boldsymbol{\alpha}, \boldsymbol{\kappa}_{1.2})$ be a shorthand for the pdf in (8), where "cMLG" stands for "conditional multivariate log-gamma." Proposition 1 shows that cMLG does not fall within the same class of pdfs as the joint distribution given in (6). This is primarily due to the fact that the $m \times g$ real-valued matrix $\mathbf{H}$, within the expression of cMLG in (8), is not square. This property is different from the multivariate normal distribution, where both marginal and conditional distributions obtained from a multivariate normal random vector, are multivariate normal (e.g., see Anderson, 1958; Johnson and Wichern, 1999, among others). The fact that cMLG in (8) is not MLG is especially important for Gibbs sampling because we will need to simulate from cMLG, and we cannot use (5) to do this. Thus, we require an additional result that allows us to simulate from cMLG.

Theorem 2. *Let* $\boldsymbol{q} \sim \mathrm{MLG}(\mathbf{0}_m, \boldsymbol{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa})$*, and partition this* m*-dimensional random vector so that* $\boldsymbol{q} = (\boldsymbol{q}_1', \boldsymbol{q}_2')'$*, where* $\boldsymbol{q}_1$ *is* g*-dimensional and* $\boldsymbol{q}_2$ *is* $(m-g)$*-dimensional. Additionally, consider the class of MLG random vectors that satisfy the following:*

$$\boldsymbol{V}^{-1} = \left[\begin{array}{cc} \boldsymbol{Q}_1 & \boldsymbol{Q}_2 \end{array} \right] \left[\begin{array}{cc} \boldsymbol{R}_1 & \mathbf{0}_{g,m-g} \\ \mathbf{0}_{m-g,g} & \frac{1}{\sigma_2}\boldsymbol{I}_{m-g}, \end{array} \right], \tag{9}$$

where in general $\mathbf{0}_{k,b}$ *is a* $k \times b$ *matrix of zeros;* $\boldsymbol{I}_{m-g}$ *is a* $(m-g) \times (m-g)$ *identity matrix;*

$$\boldsymbol{H} = \left[\begin{array}{cc} \boldsymbol{Q}_1 & \boldsymbol{Q}_2 \end{array} \right] \left[\begin{array}{c} \boldsymbol{R}_1 \\ \mathbf{0}_{m-g,g}, \end{array} \right]$$

is the QR decomposition of the $m \times g$ *matrix* $\boldsymbol{H}$*; the* $m \times g$ *matrix* $\boldsymbol{Q}_1$ *satisfies* $\boldsymbol{Q}_1'\boldsymbol{Q}_1 = \boldsymbol{I}_g$*, the* $m \times (m-g)$ *matrix* $\boldsymbol{Q}_2$ *satisfies* $\boldsymbol{Q}_2'\boldsymbol{Q}_2 = \boldsymbol{I}_{m-g}$*, and* $\boldsymbol{Q}_2'\boldsymbol{Q}_1 = \mathbf{0}_{m-g,g}$*;* $\boldsymbol{R}_1$ *is a* $g \times g$ *upper triangular matrix; and* $\sigma_2 > 0$*. Then, the following statements hold.*

(i) The marginal distribution of the g*-dimensional random vector* $\boldsymbol{q}_1$ *is given by*

$$f(\boldsymbol{q}_1|\boldsymbol{H}, \boldsymbol{\alpha}, \boldsymbol{\kappa}) = M_1 \ \exp\left\{\boldsymbol{\alpha}'\boldsymbol{H}\boldsymbol{q}_1 - \boldsymbol{\kappa}'\exp(\boldsymbol{H}\boldsymbol{q}_1)\right\}, \tag{10}$$

where the normalizing constant M_1 *is*

$$M_1 = \det\left([\boldsymbol{H} \ \ \boldsymbol{Q}_2]\right) \left(\prod_{i=1}^{m} \frac{\kappa_i^{\alpha_i}}{\Gamma(\alpha_i)} \right) \frac{1}{\left[\int f(\boldsymbol{q}|\mathbf{0}_m, \boldsymbol{V} = [\boldsymbol{H} \ \ \boldsymbol{Q}_2]^{-1}, \boldsymbol{\alpha}, \boldsymbol{\kappa}) d\boldsymbol{q}_1\right]_{\boldsymbol{q}_2 = \mathbf{0}_{m-g}}}.$$

(ii) The g*-dimensional random vector* $\boldsymbol{q}_1$ *is equal in distribution to* $(\boldsymbol{H}'\boldsymbol{H})^{-1}\boldsymbol{H}'\boldsymbol{w}$*, where the* m*-dimensional random vector* $\boldsymbol{w} \sim \mathrm{MLG}(\mathbf{0}_m, \boldsymbol{I}_m, \boldsymbol{\alpha}, \boldsymbol{\kappa})$*.*

Proof. See Supplemental Appendix A in Bradley et al. (2017). □

From Proposition 1 and Theorem 2(i) it is evident that *this particular class* of marginal distributions (defined in Theorem 2) falls into the *same* class of distributions as the conditional distribution of $\mathbf{q}_1$ given $\mathbf{q}_2$. That is, from Proposition 1 and Theorem 2(i), cMLG$(\mathbf{H}, \boldsymbol{\alpha}, \boldsymbol{\kappa})$ is equal to the pdf in (10). This equality is important because Theorem 2(ii) provides a way to simulate from cMLG. Furthermore, Theorem 2(ii) shows that it is (computationally) easy to simulate from cMLG provided that $g \ll m$. Recall that $\mathbf{H}$ is $m \times g$, which implies that computing the $g \times g$ matrix $(\mathbf{H}'\mathbf{H})^{-1}$ is computationally feasible when g is "small."

It is important to emphasize that Theorem 2 gives an equivalence between the cMLG and a *very specific class* of marginal distributions. That is, there are a number of restrictions (i.e., (9)) that defines the class of marginal distributions from the multivariate log-gamma that are equivalent to a cMLG. The cMLG is equivalent to the marginal distribution that can be interpreted as an orthogonal projection of the m-dimensional vector of independent log-gamma random variables $\mathbf{w}$ onto the column space spanned by the columns of $\mathbf{H}$.

2.3 Multivariate Log Gamma Approximation of the Multivariate Normal Distribution

An extremely common model to fit dependent count-valued data is the aforementioned LGP. Recently, Gelfand and Schliep (2016) reviewed and discussed why LGPs have become an industry standard in spatial statistics. This motivates us to investigate a connection between the multivariate log-gamma distribution and the multivariate normal distribution.

Proposition 2. *Let* $\boldsymbol{q} \sim \mathrm{MLG}(\boldsymbol{c}, \alpha^{1/2}\boldsymbol{V}, \alpha\mathbf{1}, \alpha\mathbf{1})$. *Then* $\boldsymbol{q}$ *converges in distribution to a multivariate normal random vector with mean* $\boldsymbol{c}$ *and covariance matrix* $\boldsymbol{V}\boldsymbol{V}'$ *as* α *goes to infinity.*

Proof. See Supplemental Appendix A in Bradley et al. (2017). □

The asymptotic result in Proposition 2 is on the shape parameter, which can be specified as any value that one would like. Thus, Proposition 1 provides motivation for those who prefer to use the LGP, since the multivariate log-gamma distribution can be specified to be "arbitrarily close" to the commonly used multivariate normal distribution. In practice, we have found that $\alpha = 1000$ to be sufficiently large; however, one must verify an appropriate value of α for their specific setting. Specifically, a tuning stage could be incorporated into the Gibbs sampler, and α can be increased until the ratio of the Gaussian pdf to the MLG pdf is "close to 1."

3 Modeling Dependent Count-Valued Data

We now introduce the use of the MLG distribution to model dependent count-valued data. To demonstrate the wide use of the MLG we consider many sources of dependency including space, time, and generic multivariate dependence.

3.1 The Poisson Multivariate Spatio-Temporal Mixed Effects Model

Consider Poisson count-data that are recorded over $\ell = 1, \ldots, L$ different variables, $t = T_{\mathrm{L}}^{(\ell)}, \ldots, T_{\mathrm{U}}^{(\ell)}$ different time points, and $N_t^{(\ell)}$ areas from the set $D_{t,\mathrm{P}}^{(\ell)} \equiv \{A_i : i = 1, \ldots, N_t^{(\ell)}\}$, where $A_i \subset \mathbb{R}^d$ is a region (e.g., a US county) and the subscript "P" stands for "prediction regions." Let $D_{t,\mathrm{P}}^{(\ell)}$ consist of disjoint areal units; that is, $A_i \cap A_j = \emptyset$ $(i \neq j)$. In practice, all possible prediction regions are not observed, and hence, we denote the set of $n_t^{(\ell)}$ areal units that are associated with observed data with $D_{t,\mathrm{O}}^{(\ell)} \subset D_{t,\mathrm{P}}^{(\ell)}$, where the subscript "O" stands for "observed regions."

Denote a count-value located at areal unit A, time point t, and variable ℓ with $Z_t^{(\ell)}(A)$. Here, $Z_t^{(\ell)}(A)$ is assumed to have the following conditional distribution:

$$Z_t^{(\ell)}(A) | Y_t^{(\ell)}(A) \overset{\text{ind}}{\sim} \text{Pois}\left(\exp\left\{Y_t^{(\ell)}(A)\right\}\right); \;\; \ell = 1, \ldots, L, \;\; t = T_{\mathrm{L}}^{(\ell)}, \ldots, T_{\mathrm{U}}^{(\ell)}, \;\; A \in D_{t,\mathrm{P}}^{(\ell)}, \tag{11}$$

where the canonical log-link is used. The latent process $\{Y_t^{(\ell)}(A)\}$ is assumed to have the following mixed effects model representation:

$$Y_t^{(\ell)}(A) = \mathbf{x}_t^{(\ell)}(A)'\boldsymbol{\beta} + \boldsymbol{\psi}_t^{(\ell)}(A)'\boldsymbol{\eta}_t + \xi_t^{(\ell)}(A); \quad \ell = 1, \ldots, L, \quad t = 1, \ldots, T, \quad A \in D_{t,\mathrm{P}}^{(\ell)}, \tag{12}$$

where $\mathbf{x}_t^{(\ell)}$ is a p-dimensional vector of known multivariate spatio-temporal covariates, and $\boldsymbol{\beta} \in \mathbb{R}^p$ is an unknown vector-valued parameter specified to have a MLG prior with mean zero, variance parameter σ_β^2, shape parameter $\alpha_\beta > 0$, and rate parameters $\kappa_\beta > 0$. The hyperparameters are chosen so that the prior on $\boldsymbol{\beta}$ is "flat," and we have found that results are robust to this specification. Specifically, in Section 4 we set $\sigma_\beta = \alpha_\beta = \kappa_\beta = 1000$. The r-dimensional MLG random vectors in the set $\{\boldsymbol{\eta}_t\}$ are assumed to be mean-zero, have an unknown covariance parameters $\{\sigma_{\mathrm{K},t}^2 \mathbf{K}_t\}$, and unknown shape and rate parameters $\{\alpha_t\}$ and $\{\kappa_t\}$. The set $\{\xi_t^{(\ell)}(A)\}$ consists of independent log-gamma random variables with mean zero and unknown variance parameter $\sigma_{\xi,t}^2$, and unknown shape and rate parameters $\{\tau_t\}$ and $\{\theta_t\}$.

The real-valued r-dimensional multivariate spatio-temporal basis functions (denoted with $\boldsymbol{\psi}$) has rank r (with $r \ll n$). In Section 4, we use the Moran's I (MI) basis functions (Griffith, 2000, 2002, 2004), which has useful properties related to spatial confounding (Clayton et al., 1993; Reich et al., 2006; Hodges and Reich, 2011), dimension reduction, and the Moran's I statistic (Moran, 1950) (see the Supplemental Appendix B in Bradley et al., 2017 for more details). Our use of the MI basis function is motivated primarily in it's use for dimension reduction (e.g., see Hughes and Haran, 2013; Bradley et al., 2015, among others), since we are interested in big data problems.

The random effects $\{\boldsymbol{\eta}_t\}$ are specified to have a first-order vector autoregressive structure. Bradley et al. (2015) has provided a class of real-valued propagator matrices $\{\mathbf{M}_t\}$. These propagator matrices lead to an extremely parsimonious model since they can be assumed *known* and are specified based on the angle between covariates (see the Supplemental Appendix B in Bradley et al., 2017 for more details). Additionally, the covariance parameters $\{\sigma_{K,t}^2 \mathbf{K}_t\}$ are highly structured, with $\{\sigma_{K,t}\}$ as the only unknown parameters (e.g., see Bradley et al., 2015). Specifically, let $\sigma_{\mathrm{K},t} > 0$, and for each t let $\mathbf{K}_t^{-1} \equiv \arg\min_{\mathbf{C}}\{||\mathbf{Q}_t - \mathbf{\Psi}_t^{\mathrm{P}}\mathbf{C}^{-1}\mathbf{\Psi}_t^{\mathrm{P}\prime}||_{\mathrm{F}}^2\}$, where $\mathbf{\Psi}_t^P \equiv (\psi_t^{(\ell)}(A) : \ell = 1, \ldots, L, A \in D_{\mathrm{P},t}^{(\ell)})'$. Here, $\mathbf{Q}_t = \mathbf{I}_{N_t} - \mathbf{A}_t$ is the precision associated with an intrinsic conditionally autoregressive model. Additionally, the minimization to obtain $\mathbf{K}_t$ is among $r \times r$ positive semi-definite matrices $\mathbf{C}$, and for a generic real-valued square matrix $\mathbf{H}$ the Frobenius norm is defined as $||\mathbf{H}||_F^2 = \mathrm{trace}(\mathbf{H}'\mathbf{H})$. We let $\mathbf{A}_t$ be the adjacency matrix corresponding to the edges formed by $\{D_{t,\mathrm{P}}^{(\ell)} : \ell = 1, \ldots, L\}$.

Parameter models are required for $\sigma_{\mathrm{K},t}$ and $\sigma_{\xi,t}$. We show (in Section 3.2) that full-conditionals are known for $\sigma_{\mathrm{K},t}$ and $\sigma_{\xi,t}$ if the priors on $1/\sigma_{\mathrm{K},t}$ and $1/\sigma_{\xi,t}$ are truncated log-gamma distributions. (These priors are truncated below by zero to ensure that $\sigma_{\mathrm{K},t}$ and $\sigma_{\xi,t}$ are positive.) In general, let "TruncLG(ω, ρ, h)" be a shorthand for a log-gamma distribution with shape $\omega > 0$ and rate $\rho > 0$, truncated below by h. Similarly, the gamma priors on κ_t and $\kappa_{\xi,t}$ results in gamma full-conditional distributions. Let "Gamma(ζ, δ)" be a shorthand for a gamma distribution with shape $\zeta > 0$ and rate $\delta > 0$. The hyperparameters are chosen so that the priors are "flat," and we have found that results are robust to this specification. Specifically, in Section 4 we set $\omega = \rho = \zeta = \delta = 1000$.

Finally, we consider discrete uniform priors for α_t and τ_t. That is, it is assumed that

$$\begin{aligned} f(\alpha_t) &= \frac{1}{U}; \quad \alpha_t = a_1, \ldots, a_U \\ f(\tau_t) &= \frac{1}{U}; \quad \tau_t = a_1, \ldots, a_U; \;\; 1 \leq t \leq T, \end{aligned} \tag{13}$$

where for our results in Section 4 many different choices for $a_1, \ldots, a_U$ were considered, and we found that $a_1 = 200, a_2 = 210, \ldots, a_{200} = 10,000$ is appropriate for that application. Any number of different parameter models may be considered, and we suggest that one seriously considers alternatives to what we use in (13). However, for our purpose of prediction, the simple discrete uniform prior is appropriate. Let "DU$(a_1, \ldots, a_B)$" be the discrete uniform distribution over the values $a_1, \ldots, a_B$.

The culmination of (11) through (13) leads to what we call the *Poisson multivariate spatio-temporal mixed effects model (P-MSTM)*. To aid the reader an outline of the P-MSTM is presented in Model 1.

3.2 Model Implementation

The P-MSTM is extremely general, and can be adapted in variety of ways. In particular, the P-MSTM is flexible enough to handle different basis functions, propagator matrices, and parameter models that may be more suitable in the context of different problems. This includes point referenced basis functions, which are used to model geostatistical data (e.g., see Wikle et al., 2001; Cressie and Johannesson, 2008, among others). Additionally, Model 1 is well defined for the case when $L = 1$, $T = 1$, and/or $|D_{t,\mathrm{P}}^{(\ell)}| = 1$ for each t and ℓ. This implies that our modeling framework can be readily applied to

Model 1: Latent MLG Poisson Multivariate Spatio-Temporal Mixed Effects Model

$$
\begin{aligned}
&\text{Data Model:}\\
&Z_t^{(\ell)}(A)|\boldsymbol{\beta},\boldsymbol{\eta}_t,\xi_t^{(\ell)}(A) \overset{\text{ind}}{\sim} \text{Pois}\left[\exp\left\{\mathbf{x}_t^{(\ell)}(A)'\beta+\boldsymbol{\psi}_t^{(\ell)}(A)'\boldsymbol{\eta}_t+\xi_t^{(\ell)}(A)\right\}\right];\\
&\qquad\qquad \ell=1,\ldots,L, t=T_L^{(\ell)},\ldots,T_U^{(\ell)}, A\in D_{t,O}^{(\ell)}\\
&\text{Process Model 1: } \boldsymbol{\eta}_t|\boldsymbol{\eta}_{t-1},\sigma_{\text{K},t},\alpha_t,\kappa_t \sim \text{MLG}\left(\mathbf{M}_t\boldsymbol{\eta}_{t-1},\sigma_{\text{K},t}\mathbf{K}_t^{1/2},\alpha_t\mathbf{1}_r,\kappa_t\mathbf{1}_r\right);\\
&\qquad\qquad 2\le t\le T, T>1\\
&\text{Process Model 2: } \boldsymbol{\eta}_1|\sigma_{\text{K},1},\alpha_1,\kappa_1 \sim \text{MLG}\left(\mathbf{0}_{r,1},\sigma_{\text{K},1}\mathbf{K}_1^{1/2},\alpha_1\mathbf{1}_r,\kappa_1\mathbf{1}_r\right);\\
&\text{Process Model 3: } \boldsymbol{\xi}_t|\sigma_{\xi,t},\tau_t,\theta_t \sim \text{MLG}\left(\mathbf{0}_{n_t,1},\sigma_{\xi,t}\mathbf{I}_{n_t},\tau_t\mathbf{1}_{n_t},\theta_t\mathbf{1}_{n_t}\right);\quad 1\le t\le T\\
&\text{Parameter Model 1: } \boldsymbol{\beta}\sim\text{MLG}\left(\mathbf{0}_{p,1},\sigma_\beta\,\mathbf{I}_p,\alpha_\beta\,\mathbf{1}_p,\kappa_\beta\,\mathbf{1}_p\right);\\
&\text{Parameter Model 2: } \frac{1}{\sigma_{\text{K},t}}\sim\text{TruncLG}\left(\omega,\rho,0\right);\quad 1\le t\le T\\
&\text{Parameter Model 3: } \frac{1}{\sigma_{\xi,t}}\sim\text{TruncLG}\left(\omega,\rho,0\right);\quad 1\le t\le T, \omega>0,\rho>0\\
&\text{Parameter Model 4: } \alpha_t\sim\text{DU}(a_1,\ldots,a_U);\;\; 1\le t\le T\\
&\text{Parameter Model 5: } \tau_t\sim\text{DU}(a_1,\ldots,a_U);\;\; 1\le t\le T\\
&\text{Parameter Model 6: } \kappa_t\sim\text{Gamma}(\zeta,\delta);\;\; 1\le t\le T\\
&\text{Parameter Model 7: } \theta_t\sim\text{Gamma}(\zeta,\delta);\;\; 1\le t\le T,\zeta>0,\delta>0
\end{aligned}
\tag{14}
$$

multivariate-only, spatial-only, times series, multivariate spatial, multivariate time series, and spatio-temporal datasets (in addition to multivariate spatio-temporal data). This generality is especially notable because it is rather straightforward to simulate from the full-conditional distributions implied by Model 1.

Proposition 3. *Suppose the n-dimensional data vector $\boldsymbol{Z}$ follows the P-MSTM distribution given in Model 1. Then, we have the following full conditional distribution for the unknown latent random vectors, and unknown parameters.*

$$
\begin{aligned}
&f(\boldsymbol{\beta}|\cdot)=\text{cMLG}(\boldsymbol{H}_\beta,\boldsymbol{\alpha}_\beta,\boldsymbol{\kappa}_\beta)\\
&f(\boldsymbol{\eta}_t|\cdot)=\text{cMLG}(\boldsymbol{H}_{\eta,t},\boldsymbol{\alpha}_{\eta,t},\boldsymbol{\kappa}_{\eta,t});\;\; 2\le t\le T-1 \qquad (\textit{provided}\, T>1)\\
&f(\boldsymbol{\eta}_1|\cdot)=\text{cMLG}(\boldsymbol{H}_{\eta,1},\boldsymbol{\alpha}_{\eta,1},\boldsymbol{\kappa}_{\eta,1}) \qquad\qquad\qquad (\textit{provided}\, T>1)\\
&f(\boldsymbol{\eta}_T|\cdot)=\text{cMLG}(\boldsymbol{H}_{\eta,T},\boldsymbol{\alpha}_{\eta,T},\boldsymbol{\kappa}_{\eta,T})\\
&f(\boldsymbol{\xi}_t|\cdot)=\text{cMLG}(\boldsymbol{H}_{\xi,t},\boldsymbol{\alpha}_{\xi,t},\boldsymbol{\kappa}_{\xi,t});\;\; 1\le t\le T\\
&f\left(1/\sigma_{\text{K},t}|\cdot\right)\propto\text{cMLG}(\boldsymbol{H}_{\text{K},t},\boldsymbol{\omega}_t,\boldsymbol{\rho}_t)I(\sigma_{\text{K},t}>0);\;\; 1\le t\le T\\
&f\left(1/\sigma_{\xi,t}|\cdot\right)\propto\text{cMLG}(\boldsymbol{H}_{\sigma,t},\boldsymbol{\omega}_{\xi,t},\boldsymbol{\rho}_{\xi,t})I(\sigma_{\xi,t}>0);\;\; 1\le t\le T\\
&f\left(\kappa_t|\cdot\right)=\text{Gamma}(\zeta_t,\delta_t);\;\; 1\le t\le T
\end{aligned}
$$

$$
\begin{aligned}
&f\,(\theta_t|\cdot) = \mathrm{Gamma}(\zeta_{\xi,t}, \delta_{\xi,t});\;\; 1 \leq t \leq T\\
&f(\alpha_t = a_i|\cdot) = \frac{f\left(\boldsymbol{\eta}_t|\boldsymbol{c} = \boldsymbol{M}_t\boldsymbol{\eta}_{t-1},\, \boldsymbol{V} = \sigma_{\mathrm{K},t}\boldsymbol{K}_t^{1/2}, \boldsymbol{\alpha} = a_i\mathbf{1}_r, \boldsymbol{\kappa} = \kappa_t\mathbf{1}_r\right)}{\sum_{b=1}^{U} f\left(\boldsymbol{\eta}_t|\boldsymbol{c} = \boldsymbol{M}_t\boldsymbol{\eta}_{t-1},\, \boldsymbol{V} = \sigma_{\mathrm{K},t}\boldsymbol{K}_t^{1/2}, \boldsymbol{\alpha} = a_b\mathbf{1}_r, \boldsymbol{\kappa} = = \kappa_t\mathbf{1}_r\right)};\\
&\qquad\qquad\qquad\qquad\qquad\qquad 1 \leq t \leq T,\; i = 1, \ldots U,\\
&f(\tau_t = a_i|\cdot) = \frac{f\,(\boldsymbol{\xi}_t|\boldsymbol{c} = \mathbf{0}_{n_t,1},\, \boldsymbol{V} = \sigma_{\xi,t}\boldsymbol{I}_{n_t}, \boldsymbol{\alpha} = a_i\mathbf{1}_{n_t}, \boldsymbol{\kappa} = \theta_t\mathbf{1}_{n_t})}{\sum_{b=1}^{U} f\,(\boldsymbol{\xi}_t|\boldsymbol{c} = \mathbf{0}_{n_t,1},\, \boldsymbol{V} = \sigma_{\xi,t}\boldsymbol{I}_{n_t}, \boldsymbol{\alpha} = a_b\mathbf{1}_{n_t}, \boldsymbol{\kappa} = \theta_t\mathbf{1}_{n_t})};\\
&\qquad\qquad\qquad\qquad\qquad\qquad 1 \leq t \leq T,\; i = 1, \ldots U, \qquad (15)
\end{aligned}
$$

where $f(\boldsymbol{\beta}|\cdot)$ represents the pdf of $\boldsymbol{\beta}$ given all other process variables, parameters, and the data. For each t, we define $f(\boldsymbol{\eta}_t|\cdot)$, $f(\boldsymbol{\xi}_t|\cdot)$, $f(1/\sigma_{\mathrm{K},t}|\cdot)$, $f(1/\sigma_{\xi,t}|\cdot)$, $f(\kappa_t|\cdot)$, $f(\kappa_{\xi,t}|\cdot)$, $f(\alpha_t|\cdot)$, and $f(\tau_t|\cdot)$ in a similar manner. For ease of exposition, in Table 1 we provide the definitions of each quantity in (15).

Proof. See Supplemental Appendix A in Bradley et al. (2017). □

The proof of Proposition 3 is given in Bradley et al. (2017). Additionally, the step-by-step instructions outlining the implementation of the Gibbs sampler based on (15) is given in Algorithm 1. Notice that it is relatively easy to simulate from the cMLG distributions in (15) using Theorem 2(*ii*); provided that $r \ll n$ and $p \ll n$. That is, from Theorem 2(*ii*) simulating from the cMLG full-conditionals in (15) involves computing the inverse of $p \times p$ and $r \times r$ matrices, which involves computations on the order of p^3 and r^3, respectively. Thus, joint samples are taken from the cMLG distributions stated in Proposition 3 using Theorem 2(*ii*). This is a particularly important point because there exists other multivariate log-gamma approaches that result in component-wise updating (Demirhan and Hamurkaroglu, 2011).

Proposition 3 can be applied to the aforementioned special cases of multivariate spatio-temporal data (i.e., spatial-only, times series, multivariate-only, spatio-temporal, multivariate spatial, and multivariate times series datasets). Thus, the implications of Proposition 3 are enormous, as it provides a way to *efficiently model* a wide range of less general but interesting special cases not explicitly considered in this manuscript. As an example, the full conditional distributions for the multivariate-only setting (i.e., when $L > 1$, $T = 1$, and $D_1^{(\ell)} \equiv 1$) are presented in Supplemental Appendix D (Bradley et al., 2017). Additionally, the full conditional distributions for the spatial-only setting (i.e., when $L = 1$, $T = 1$, and when $|D_1^{(1)}| > 1$) and a demonstration using spatially-referenced US Census estimates obtained from the American Community Survey are presented in Supplemental Appendix E (Bradley et al., 2017).

The P-MSTM captures a balance between modeling complex dependencies, and the computational needs required to jointly model datasets. In particular, many datasets express non-stationarity in both space *and* time, and space-time interactions. Allowing for non-stationarity and space-time interactions eliminates the possibility of using computationally advantageous methodologies based on the linear models for coregionalization and separability (e.g., see Gelfand and Vounatsou, 2003; Daniels et al., 2006;

Definition	Additional Notes
$\mathbf{H}_{\beta} = (\mathbf{X}_1', \ldots, \mathbf{X}_T', \sigma_{\beta}^{-1}\mathbf{I}_p)'$	
$\mathbf{H}_{\eta,t} = (\boldsymbol{\Psi}_t', \sigma_{\mathrm{K},t}^{-1}\mathbf{K}_t^{-1/2}, -\sigma_{\mathrm{K},t+1}^{-1}\mathbf{K}_{t+1}^{-1/2}\mathbf{M}_{t+1})'$	$1 \leq t < T$
$\mathbf{H}_{\eta,T} = (\boldsymbol{\Psi}_T', \sigma_{\mathrm{K},T}^{-1}\mathbf{K}_T^{-1/2})'$	
$\mathbf{H}_{\xi,t} = (\mathbf{I}_{n_t}, \sigma_{\xi,t}^{-1}\mathbf{I}_{n_t})'$	$1 \leq t \leq T$
$\mathbf{H}_{\mathrm{K},t} = (\boldsymbol{\eta}_t'\mathbf{K}_t^{1/2\prime} - \boldsymbol{\eta}_{t-1}'\mathbf{M}_t\mathbf{K}_t^{1/2\prime}, 1)'$	$1 < t \leq T$
$\mathbf{H}_{\mathrm{K},1} = (\boldsymbol{\eta}_1'\mathbf{K}_1^{1/2\prime}, 1)'$	$1 < t \leq T$
$\mathbf{H}_{\sigma,t} = (\boldsymbol{\xi}_t, 1)'$	$1 \leq t \leq T$
$\mathbf{H}_{\xi,t} = (\mathbf{I}_{n_t}, \sigma_{\xi,t}^{-1}\mathbf{I}_{n_t})'$	$1 \leq t \leq T$
$\boldsymbol{\kappa}_{\beta} = \left\{\exp(\boldsymbol{\Psi}_1\boldsymbol{\eta}_1 + \boldsymbol{\xi}_1)', \ldots, \exp(\boldsymbol{\Psi}_T\boldsymbol{\eta}_T + \boldsymbol{\xi}_T)', \kappa_{\beta}\mathbf{1}_p'\right\}'$	
$\boldsymbol{\kappa}_{\eta,t} = \left\{\exp(\mathbf{X}_t\boldsymbol{\beta} + \boldsymbol{\xi}_t)', \kappa_t \exp(-\sigma_{\mathrm{K},t}^{-1}\mathbf{K}_t^{-1/2}\mathbf{M}_t\boldsymbol{\eta}_{t-1})', \kappa_{t+1} \exp(\sigma_{\mathrm{K},t+1}^{-1}\mathbf{K}_{t+1}^{-1/2}\boldsymbol{\eta}_{t+1})'\right\}'$	$1 < t < T$
$\boldsymbol{\kappa}_{\eta,1} = \left\{\exp(\mathbf{X}_1\boldsymbol{\beta} + \boldsymbol{\xi}_1)', \kappa_1\mathbf{1}_r', \kappa_2 \exp(\sigma_{\mathrm{K},t}^{-1}\mathbf{K}_2^{-1/2}\boldsymbol{\eta}_2)'\right\}'$	Provided $T > 1$
$\boldsymbol{\kappa}_{\eta,T} = \left\{\exp(\mathbf{X}_T\boldsymbol{\beta} + \boldsymbol{\xi}_T)', \kappa_T \exp(-\sigma_{\mathrm{K},T}^{-1}\mathbf{K}_T^{-1/2}\mathbf{M}_T\boldsymbol{\eta}_{T-1})'\right\}'$	If $T = 1$ then replace $\mathbf{M}_T$ and $\boldsymbol{\eta}_{T-1}$ with $\mathbf{0}_{r,r}$ and $\mathbf{0}_r$, respectively.
$\boldsymbol{\kappa}_{\xi,t} = \left\{\exp(\mathbf{X}_t\boldsymbol{\beta} + \boldsymbol{\Psi}_t\boldsymbol{\eta}_t)', \theta_t\mathbf{1}_{n_t}'\right\}'$	$1 \leq t \leq T$
$\boldsymbol{\rho}_t = (\kappa_t\mathbf{1}_r', \rho)'$	$1 \leq t \leq T$
$\boldsymbol{\rho}_{\xi,t} = \left(\theta_t\mathbf{1}_{n_t}', \rho\right)'$	$1 \leq t \leq T$
$\delta_t = \delta + \mathbf{1}_r'\exp\left\{\sigma_{\mathrm{K},t}^{-1}\mathbf{K}_t^{-1/2}\left(\boldsymbol{\eta}_t - \mathbf{M}_t\boldsymbol{\eta}_{t-1}\right)\right\}$	$1 < t \leq T$
$\delta_1 = \delta + \mathbf{1}_r'\exp\left(\sigma_{\mathrm{K},1}^{-1}\mathbf{K}_1^{-1/2}\boldsymbol{\eta}_1\right)$	

Table 1: A comprehensive list of matrices, vectors, and constants used within the Proposition 3. If there are no zero counts within the dataset, then set $d_{\beta} = d_{\eta,1} = \cdots = d_{\eta,T} = d_{\xi,1} = \cdots = d_{\xi,T} = 0$.

$\delta_{\xi,t} = \delta + \mathbf{1}_{n_t}' \exp\left(\sigma_{\xi,t}^{-1}\boldsymbol{\xi}_t\right)$	$1 \leq t \leq T$
$\boldsymbol{\alpha}_\beta = \left\{\mathbf{Z}_1' + d_\beta \mathbf{1}_{n_1}', \ldots, Z_T' + d_\beta \mathbf{1}_{n_T}', \alpha_\beta \mathbf{1}_p' - d_\beta \sigma_\beta \sum_{t=1}^T \mathbf{1}_{n_t}' \mathbf{X}_t\right\}'$	
$\boldsymbol{\alpha}_{\eta,t} = \left\{\mathbf{Z}_t' + d_{\eta,t}\mathbf{1}_{n_t}', \alpha_t \mathbf{1}_r' - \frac{d_{\eta,t}\sigma_{\mathrm{K},t}}{2}\mathbf{1}_{n_t}'\boldsymbol{\Psi}_t\mathbf{K}_t^{1/2}, \alpha_{t+1}\mathbf{1}_r' + \frac{d_{\eta,t}\sigma_{\mathrm{K},t+1}}{2}\mathbf{1}_{n_t}'\boldsymbol{\Psi}_t\mathbf{M}_{t+1}'\mathbf{K}_{t+1}^{1/2}\right\}'$	$1 < t < T$
$\boldsymbol{\alpha}_{\eta,1} = \left\{\mathbf{Z}_1' + d_{\eta,1}\mathbf{1}_{n_1}', \alpha_1 \mathbf{1}_r' - \frac{d_{\eta,1}\sigma_{\mathrm{K},1}}{2}\mathbf{1}_{n_1}'\boldsymbol{\Psi}_1\mathbf{K}_1^{1/2}, \alpha_2\mathbf{1}_r' + \frac{d_{\eta,1}\sigma_{\mathrm{K},2}}{2}\mathbf{1}_{n_1}'\boldsymbol{\Psi}_1\mathbf{M}_2'\mathbf{K}_2^{1/2}\right\}'$	Provided $T > 1$
$\boldsymbol{\alpha}_{\eta,T} = \left\{\mathbf{Z}_T' + d_{\eta,T}\mathbf{1}_{n_T}', \alpha_T \mathbf{1}_r' - d_{\eta,1}\sigma_{\mathrm{K},T}\mathbf{1}_{n_T}'\boldsymbol{\Psi}_T\mathbf{K}_T^{1/2}\right\}'$	
$\boldsymbol{\alpha}_{\xi,t} = \left\{\mathbf{Z}_t' + d_{\xi,t}\mathbf{1}_{n_T}', \tau_t\mathbf{1}_{n_t}' - d_{\xi,t}\sigma_{\xi,t}\mathbf{1}_{n_t}'\right\}'$	$1 \leq t \leq T$
$\boldsymbol{\omega}_t = \left(\alpha_t \mathbf{1}_r', \omega\right)'$	$1 \leq t \leq T$
$\boldsymbol{\omega}_{\xi,t} = \left(\tau_t \mathbf{1}_{n_t}', \omega\right)'$	$1 \leq t \leq T$
$\zeta_{\xi,t} = \zeta + n_t\tau_t$	$1 \leq t \leq T$
$\zeta_t = \zeta + r\alpha_t$	$1 \leq t \leq T$
$d_\beta = \alpha / \left[1 + \max\left\{\mathrm{abs}\left(\sigma_\beta \sum_{t=1}^T \mathbf{1}_{n_t}'\mathbf{X}_t\right)\right\}\right]$	
$d_{\eta,t} = \alpha / \left(1 + \max\left[\mathrm{abs}\left\{(\sigma_{\mathrm{K},t}\mathbf{1}_{n_t}'\boldsymbol{\Psi}_t\mathbf{K}_t^{1/2}, -\sigma_{\mathrm{K},t+1}\mathbf{1}_{n_t}'\boldsymbol{\Psi}_t\mathbf{M}_{t+1}'\mathbf{K}_{t+1}^{1/2})\right\}\right]\right)$	$1 \leq t < T$
$d_{\eta,T} = \alpha / \left(1 + \max\left[\mathrm{abs}\left\{\sigma_{\mathrm{K},T}\mathbf{1}_{n_t}'\boldsymbol{\Psi}_t\mathbf{K}_t^{1/2}\right\}\right]\right)$	
$d_{\xi,t} = \alpha / \left(1 + \sigma_{\xi,t}\right)$	$1 \leq t \leq T$

Table 1: (Continued).

Algorithm 1: The Gibbs Sampler for the P-MSTM

```
1. Initialize $\boldsymbol{\beta}$, $\sigma^2_{\mathrm{K},t}$, $\sigma^2_{\xi,t}$, and $\boldsymbol{\xi}_t$ and $\boldsymbol{\eta}_t$ for each $t$. Denote these initializations with $\boldsymbol{\beta}^{[0]}$, $\sigma^{2[0]}_{\mathrm{K},t}$, $\sigma^{2[0]}_{\xi,t}$, $\sigma^{2[0]}_{\beta}$, and $\boldsymbol{\xi}^{[0]}_t$ and $\boldsymbol{\eta}^{[0]}_t$ for each $t$. Set $b = 1$.
2. Set $\boldsymbol{\beta}^{[b]}$ equal to a draw from cMLG$(\mathbf{H}_\beta, \boldsymbol{\alpha}_\beta, \boldsymbol{\kappa}_\beta)$ using Theorem 2$(ii)$.
3. If $t < T$, then set $\boldsymbol{\eta}^{[b]}_t$ equal to a draw from cMLG$(\mathbf{H}_{\eta,t}, \boldsymbol{\alpha}_{\eta,t}, \boldsymbol{\kappa}_{\eta,t})$ using Theorem 2$(ii)$.
4. Set $\boldsymbol{\eta}^{[b]}_T$ equal to a draw from cMLG$(\mathbf{H}_{\eta,T}, \boldsymbol{\alpha}_{\eta,T}, \boldsymbol{\kappa}_{\eta,T})$ using Theorem 2$(ii)$.
5. For each $t$ let $\boldsymbol{\xi}^{[b]}_t$ be a draw from cMLG$(\mathbf{H}_{\xi,t}, \boldsymbol{\alpha}_{\xi,t}, \boldsymbol{\kappa}_{\xi,t})$ using Theorem 2$(ii)$.
6. For each $t$, repeatedly simulate $\frac{1}{\sigma^{[b]}_{\mathrm{K},t}}$ from cMLG$(\mathbf{H}_{\mathrm{K},t}, \boldsymbol{\omega}_t, \boldsymbol{\rho}_t)$ until $\frac{1}{\sigma^{[b]}_{\mathrm{K},t}}$ is positive.
7. For each $t$, repeatedly simulate $\frac{1}{\sigma^{[b]}_{\xi,t}}$ from cMLG$(\mathbf{H}_{\sigma,t}, \boldsymbol{\omega}_{\xi,t}, \boldsymbol{\rho}_{\xi,t})$ until $\frac{1}{\sigma^{[b]}_{\xi,,t}}$ is positive.
8. For each $t$, draw $\alpha^{[b]}_t = a_i$ with probability,
   $\frac{f\left(\boldsymbol{\eta}^{[b]}_t | \mathbf{c} = \mathbf{M}_t \boldsymbol{\eta}^{[b]}_{t-1}, \mathbf{V} = \sigma^{[b]}_{\mathrm{K},t} \mathbf{K}^{1/2}_t, \boldsymbol{\alpha} = a_i \mathbf{1}_r, \boldsymbol{\kappa} = \kappa^{[b-1]}_t \mathbf{1}_r\right)}{\sum_{b=1}^{U} f\left(\boldsymbol{\eta}^{[b]}_t | \mathbf{c} = \mathbf{M}_t \boldsymbol{\eta}^{[b]}_{t-1}, \mathbf{V} = \sigma^{[b]}_{\mathrm{K},t} \mathbf{K}^{1/2}_t, \boldsymbol{\alpha} = a_b \mathbf{1}_r, \boldsymbol{\kappa} = \kappa^{[b-1]}_t \mathbf{1}_r\right)}; \quad i = 1, \ldots U,$
   where recall that $f$ is defined in Theorem 1$(i)$, and let $\mathbf{1}_r$ be an $r$-dimensional vector of ones.
9. For each $t$, draw $\tau^{[b]}_t = a_i$ with probability,
   $\frac{f\left(\boldsymbol{\xi}^{[b]}_t | \mathbf{c} = \mathbf{0}_{n_t,1}, \mathbf{V} = \sigma^{[b]}_{\xi,t} \mathbf{I}_{n_t}, \boldsymbol{\alpha} = a_i \mathbf{1}_{n_t}, \boldsymbol{\kappa} = \theta^{[b-1]}_t \mathbf{1}_{n_t}\right)}{\sum_{b=1}^{U} f\left(\boldsymbol{\xi}^{[b]}_t | \mathbf{c} = \mathbf{0}_{n_t,1}, \mathbf{V} = \sigma^{[b]}_{\xi,t} \mathbf{I}_{n_t}, \boldsymbol{\alpha} = a_b \mathbf{1}_{n_t}, \boldsymbol{\kappa} = \theta^{[b-1]}_t \mathbf{1}_{n_t}\right)}; \quad i = 1, \ldots U,$
   where recall that $f$ is defined in Theorem 1$(i)$.
10. For each $t$, simulate $\kappa^{[b]}_t$ from Gamma$(\zeta_t, \delta_t)$.
11. For each $t$, simulate $\theta^{[b]}_t$ from Gamma$(\zeta_{\xi,t}, \delta_{\xi,t})$.
12. Set $b = b + 1$.
13. Repeat steps 2 through 12 until $b$ is equal to the desired value (i.e., convergence is achieved).
```

Jin et al., 2007). Thus, other features of the P-MSTM are specified to allow for high-dimensional data. In particular, a reduced rank approach (Cressie and Johannesson, 2008) is assumed (i.e., $r \ll n$) and confounded random effects are removed (Hughes and Haran, 2013). The literature on non-stationarity in both space and time is relatively new (e.g., see Ma, 2002; Huang and Hsu, 2004; Sigrist et al., 2011; Garg et al., 2012; Bradley et al., 2015), and the P-MSTM provides a viable approach for modeling non-stationarity in count-valued data.

3.3 Overdispersion Properties of the P-MSTM

A reoccurring modeling question involved with Poisson spatial models is the characterization of overdispersion (e.g., see De Oliveira, 2003, 2013). Following the notation of (De Oliveira, 2013) we define the relative overdispersion at time t, variable ℓ, and location $A \in D_{P,t}^{(\ell)}$ with

$$\mathrm{OD}_t^{(\ell)}(A) = \frac{\mathrm{var}\left\{Z_t^{(\ell)}(A)\right\} - E\left\{Z_t^{(\ell)}(A)\right\}}{E\left\{Z_t^{(\ell)}(A)\right\}}.$$

In Proposition 4, we state the expression of $\mathrm{OD}_t^{(\ell)}(A)$ for the P-MSTM.

Proposition 4. *Suppose the n-dimensional data vector $\boldsymbol{Z}$ follows the P-MSTM distribution given in Model 1. For a given time t, variable ℓ, and location $A \in D_{P,t}^{(\ell)}$, let $(k_1, \ldots, k_r) = \sigma_{\mathrm{K}} \boldsymbol{\psi}_t^{(\ell)}(A)' \boldsymbol{K}_t^{-1/2}$. Suppose α_β is strictly larger than the absolute value of the smallest negative element in the p-dimensional vector $\boldsymbol{X}_t^{(\ell)}(A) = (X_{t,1}^{(\ell)}(A), \ldots, X_{t,p}^{(\ell)}(A))'$. Likewise, let α_t be strictly larger than the absolute value of the smallest negative element in the r-dimensional vector $(k_1, \ldots, k_r)'$. Then, we have the following expression for the relative overdispersion at time t, variable ℓ, and location $A \in D_{P,t}^{(\ell)}$*

$$\begin{aligned}\mathrm{OD}_t^{(\ell)}(A) = &\left(\frac{1}{\kappa_\beta^{\sum_{i=1}^p X_{t,i}^{(\ell)}(A)\sigma_\beta}\, \kappa_t^{\sum_{i=1}^r k_i}\, \kappa_{\xi,t}}\right) \\ &\times \left[\left\{\prod_{i=1}^p \frac{\Gamma(\alpha_\beta + 2\, X_{t,i}^{(\ell)}(A)\sigma_\beta)}{\Gamma(\alpha_\beta + X_{t,i}^{(\ell)}(A)\sigma_\beta)}\right\}\left\{\prod_{i=1}^r \frac{\Gamma(\alpha_t + 2\, k_i)}{\Gamma(\alpha_t + k_i)}\right\}\left\{\frac{\Gamma(\alpha_{\xi,t}+2)}{\Gamma(\alpha_{\xi,t}+1)}\right\}\right. \\ &\left. - \left\{\prod_{i=1}^p \frac{\Gamma(\alpha_\beta + X_{t,i}^{(\ell)}(A)\sigma_\beta)}{\Gamma(\alpha_\beta)}\right\}\left\{\prod_{i=1}^r \frac{\Gamma(\alpha_t + k_i)}{\Gamma(\alpha_t)}\right\}\left\{\frac{\Gamma(\alpha_{\xi,t}+1)}{\Gamma(\alpha_{\xi,t})}\right\}\right]. \qquad (16)\end{aligned}$$

Proof. See Supplemental Appendix A in Bradley et al. (2017). □

It is immediately apparent from (16) that the shape and rate parameters of the P-MSTM (along with the covariates, basis functions, and covariances) are important for understanding the relative overdispersion of a count-valued observation from Model 1. As we see below, this connection between the relative overdispersion and the shape and rate parameters, also provides a connection between the shape and rate parameters and the correlations induced by the P-MSTM.

Proposition 5. *Suppose that for a given time t, variable ℓ, and location $A \in D_{P,t}^{(\ell)}$, $Z_t^{(\ell)}(A) \sim$ Pois$[\exp\{Y_t^{(\ell)}(A)\}]$. Assume that $Z_t^{(\ell)}(A)$ is conditionally independent of $Z_h^{(m)}(B)$ for $t, h = 1, \ldots, T$, $\ell, m = 1, \ldots, L$, $A \in D_{\mathrm{P},t}^{(\ell)}$, $B \in D_{\mathrm{P},h}^{(m)}$, and $t \neq h$, $\ell \neq m$, or $A \neq B$. Let $Y_t^{(\ell)}(A)$ be a measurable random variable for every $t = 1, \ldots, T$, $\ell =$*

$1,\ldots,L$, $A \in D_{\mathrm{P},t}^{(\ell)}$ *where the mean and variance of* $\exp\{Y_t^{(\ell)}(A)\}$ *are finite, and the covariogram of* $\exp\{Y_t^{(\ell)}(A)\}$ *is positive semi-definite. Then,*

$$\mathrm{corr}\left\{Z_t^{(\ell)}(A), Z_h^{(m)}(B)\right\} = \mathrm{corr}\left[\exp\left\{Y_t^{(\ell)}(A)\right\}, \exp\left\{Y_h^{(m)}(B)\right\}\right] H_{t,h}^{(\ell,m)}(A,B),$$

where

$$H_{t,h}^{(\ell,m)}(A,B) = \left[\left\{1+\frac{1}{\mathrm{OD}_t^{(\ell)}(A)}\right\}\left\{1+\frac{1}{\mathrm{OD}_h^{(m)}(B)}\right\}\right]^{-1/2},$$

"corr" is the correlation function, $t,h=1,\ldots,T$, $\ell,m=1,\ldots,L$, $A \in D_{\mathrm{P},t}^{(\ell)}$, $B \in D_{\mathrm{P},h}^{(m)}$, *and* $t \neq h$, $\ell \neq m$, *or* $A \neq B$.

Proof. See Supplemental Appendix A in Bradley et al. (2017). □

Proposition 5 shows that $\mathrm{corr}\{Z_t^{(\ell)}(A), Z_h^{(m)}(B)\}$ is bounded between $-H_{t,h}^{(\ell,m)}(A,B)$ and $H_{t,h}^{(\ell,m)}(A,B)$. Furthermore, Proposition 5 shows that as the relative overdispersion decreases the closer to zero $H_{t,h}^{(\ell,m)}(A,B)$ becomes. Thus, low overdispersion implies a very strong restriction on the range of possible values for $\mathrm{corr}\{Z_t^{(\ell)}(A), Z_h^{(m)}(B)\}$. This is true for *every choice of correlation function* for $\exp\{Y_t^{(\ell)}(A)\}$. (Note that we are not assuming the P-MSTM is true in Proposition 5.) This suggests that Parameter Models $4-7$ are valuable, since the resulting posterior distributions will allow the data to inform the appropriate range of values for $\mathrm{corr}\{Z_t^{(\ell)}(A), Z_h^{(m)}(B)\}$.

The implications of Propositions 4 and 5, suggest that we should investigate which values of the shape and rate produce low and high relative overdispersion

Proposition 6. *Suppose the* n*-dimensional data vector* $\boldsymbol{Z}$ *follows the P-MSTM distribution given in Model 1. For a given time* t*, variable* ℓ*, and location* $A \in D_{\mathrm{P},t}^{(\ell)}$*, let* $(k_1,\ldots,k_r) = \sigma_{\mathrm{K}}\boldsymbol{\psi}_t^{(\ell)}(A)'\boldsymbol{K}_t^{-1/2}$. *Suppose* α_β *is strictly larger than the absolute value of the smallest negative element in the* p*-dimensional vector* $\boldsymbol{X}_t^{(\ell)}(A) = (X_{t,1}^{(\ell)}(A),\ldots,X_{t,p}^{(\ell)}(A))'$. *Likewise, let* α_β *be strictly larger than the absolute value of the smallest negative element in the* r*-dimensional vector* $(k_1,\ldots,k_r)'$. *Then, for a given time* t*, variable* ℓ*, and location* $A \in D_{\mathrm{P},t}^{(\ell)}$

(i) If $\alpha_\beta \to \infty$, $\alpha_t \to \infty$, *and* $\alpha_{\xi,t} \to \infty$ *then* $\mathrm{OD}_t^{(\ell)}(A) \to 0$.

(ii) If $\kappa_\beta^{\sum_{i=1}^p X_{t,i}^{(\ell)}(A)\sigma_\beta}\kappa_t^{\sum_{i=1}^r k_i}\kappa_{\xi,t} \to \infty$ *then* $\mathrm{OD}_t^{(\ell)}(A) \to 0$.

(iii) If $\kappa_\beta^{\sum_{i=1}^p X_{t,i}^{(\ell)}(A)\sigma_\beta}\kappa_t^{\sum_{i=1}^r k_i}\kappa_{\xi,t} \to 0$ *then* $\mathrm{OD}_t^{(\ell)}(A) \to \infty$.

Proof. Proposition 6(i) follows immediately from Stirling's formula (shown in Bradley et al., 2017). Proposition 6(ii) and Proposition 6(iii) follow immediately from Proposition 4. □

Propositions 5 and 6 show that the value of the correlation between two different counts can be controlled by either inflating or deflating the shape and rate parameters. This suggests that the P-MSTM can model the correlation between two different Poisson counts in a very flexible manner.

4 Results: LEHD Simulations and Analysis

The QWIs, which can be accessed at http://www.census.gov/, are extremely comprehensive including important economic indicators over 96 quarters, every US county, and every industry as defined by the North Atlantic Classification System (NAICS). To date, there are no alternative data sources that measure US economic variables on as fine a spatio-temporal resolution for each NAICS industry. Thus, the extent of the QWIs coupled with the lack of alternative data sources makes the QWIs especially valuable for US economists. Furthermore, these factors motivate the need to obtain high-quality predictions of QWIs at every county along with measures of error.

The P-MSTM provides a way to estimate missing QWIs and provide measures of uncertainty. Nevertheless, there are several features of the QWIs that have not been incorporated into the P-MSTM, which could potentially be used to improve upon our analyses. In particular, similar to survey statistics there may be issues surrounding censoring and modifications due to non-response and disclosure (e.g., see Lohr (1999) for a standard reference, and Quick et al. (2015) for a recent paper with spatial data). In principle, incorporating these types of features into the P-MSTM would more realistically represent the QWIs; however, in general, this information is not available to data users. Although there is potential to develop the P-MSTM in this direction, these extensions are outside the scope of this paper. Consequently, the results in Section 4.2 should be interpreted as an illustration of the use of the P-MSTM to model a large dependent dataset. In what follows, we evaluate the performance of the P-MSTM through an empirical simulation study (Section 4.1) using a subset of the QWI dataset, and an analysis of the beginning of the quarter employment QWI (Section 4.2).

All computations were computed using Matlab (Version 8.0) on a dual 10 core 2.8 GHz Intel Xeon E5-2680 v2 processor, with 256 GB of RAM.

4.1 A Simulation Study

We choose to calibrate our simulation model towards QWIs. That is, we set the mean of a Poisson random variable equal to a count-valued QWI and use this distribution to generate a "pseudo data-value." Then, the pseudo data and the P-MSTM are used to predict the QWIs. This empirical simulation study design is similar to what is done in Bradley et al. (2015), and is motivated as a way to produce simulated data that behave similar to what one might observe in practice.

Let $Z_t^{(\ell)}(A)$ represent the number of individuals employed at the beginning of the quarter, for industry ℓ, Minnesota county A, and quarter t. Then, simulate pseudo-data

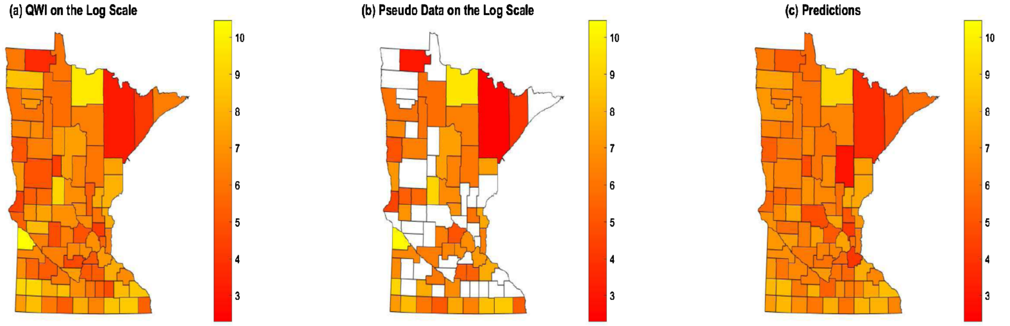


Figure 1: (a), The LEHD estimated number of individuals employed in the beginning of the 4-th quarter of 2013 within the information industry (i.e., $\{Z_{96}^{(1)}(A)\}$) in Minnesota. For comparison, a map of the pseudo-data is $\{R_{96}^{(1)}(A)\}$ computed using (17) is given in (b). The white areas indicate "suppressed" QWIs. In (c), we provide the predictions of $\{\widehat{Z}_{96}^{(1)}(A)\}$ that are computed using P-MSTM and the pseudo-data $\{R_t^{(\ell)}(A)\}$ from (17).

as follows,

$$R_t^{(\ell)} \sim \text{Pois}(Z_t^{(\ell)}(A)+1);\ \ \ell = 1,2,\ \ t = 76,\ldots,96,\ \ A \in D_{\text{MN},t}^{(\ell)}, \qquad (17)$$

where $D_{\text{MN},t}^{(\ell)}$ represents the set of counties in Minnesota (MN) that have available QWIs, $\ell = 1$ denotes the information industry, and $\ell = 2$ represents the professional, scientific and technical services industry. These two industries were chosen for this simulation study since they are highly correlated. Notice that we add 1 in (17) so that the mean of the Poisson random variables are also greater than 0.

Randomly select 65% of the areal units in $D_{\text{MN},t}^{(\ell)}$ to be "observed," and denote this new set with $D_{\text{MN,O},t}^{(\ell)}$. For illustration, we use the following covariates $\mathbf{x}_t^{(\ell)}(A) = (1, I(\ell = 1), \ldots, I(\ell = 19), |A|, I(t = 1), \ldots, I(t = 1, \ldots, 95), \text{population}(A))'$, where population$(A)$ is the 2010 decennial Census value of the population of county A and $I(\cdot)$ is the indicator function. Following Hughes and Haran (2013)'s rule of thumb for specifying r (i.e., set r equal to approximately the top 10% of the available basis functions), we set $r = 42$ (see the Supplemental Appendix B in Bradley et al., 2017). Thus, $L = 2$, $T = 20$, and $\mathbf{K}_t$ is 42×42.

In Figure 1, we present the QWIs (panel a), the pseudo data (panel b), and the predictor (panel c) given by

$$E\left[Z_t^{(\ell)}(A) | \{R_t^{(\ell)}(A) : \ell = 1,2, t = 76,\ldots,96, A \in D_{\text{MN,O},t}^{(\ell)}\}\right];$$
$$\ell = 1,2,\ \ t = 96,\ \ A \in D_{\text{MN},t}^{(\ell)},$$

where the expectation is obtained using the P-MSTM and Algorithm 1. In general, the predictions reflect the overall pattern of the data. This is further supported in

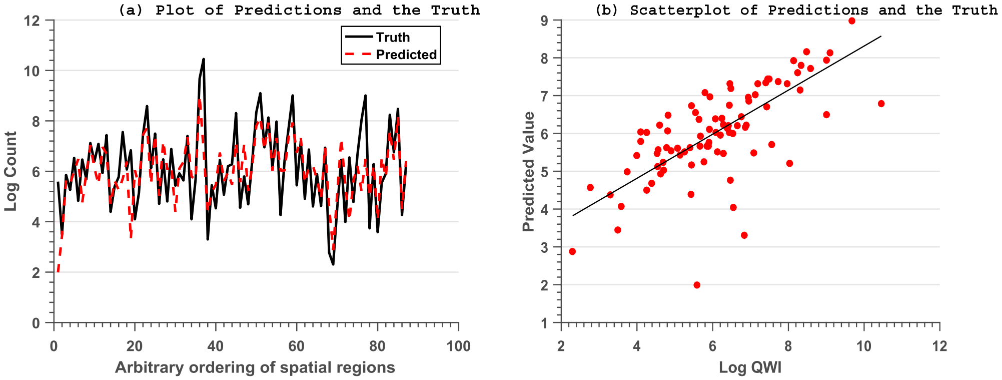


Figure 2: In (a), we plot the LEHD estimated number of individuals employed in the beginning of the 4-th quarter of 2013 within the information industry in Minnesota, and the predicted values. In (b), we produce scatterplots of the LEHD estimated number of individuals employed in the beginning of the 4-th quarter of 2013 within the information industry in Minnesota, versus the predicted values.

Figure 2(a), where we plot the log QWIs and the log predictions over an arbitrary ordering of the regions. Again we see that the predictions tend to track the truth fairly closely. In Figure 2(b) we provide a scatter-plot of the log QWIs versus the log predictions. Here, the predictions are similar to the truth (the correlation is 0.69).

Now, consider 100 independent replications of the set $\{R_{t,j}^{(\ell)}(A) : \ell = 1,2,\; t = 76,\ldots,96,\; A \in D_{\mathrm{MN,O},t}^{(\ell)}\}$, where $j = 1,\ldots,100$ and for each j we have that $R_{t,j}^{(\ell)}(A)$ is simulated according to (17). The results (not shown) indicate that the P-MSTM has a high predictive performance similar to the results based on a single replicate presented in Figures 1 and 2.

The computational performance of the P-MSTM is of particular interest. To evaluate the Markov chain we use the effective sample size (ESS). Specifically, for each A, ℓ, and t we compute the effective sample size, denoted by $\mathrm{ESS}_t^{(\ell)}(A)$. The ESS is computed as the number of MCMC replicates times the ratio of the within chain variance and the between chain variance (e.g., see Kass et al. (2016), Liu (2008), Robert and Casella (2013), and Gong and Flegal (2016) for component-wise ESS, and Vats et al. (2016) for a multivariate ESS). If the ESS is smaller (larger) than the number of replicates, this suggests that the MCMC chain has positive (negative) correlations between values in the chain. Thus, ESS close to the total number of MCMC replicates computed implies an efficient MCMC. In Figure 3, we show a boxplot, over the 100 replicates of the median $\mathrm{ESS}_t^{(\ell)}(A)$ across all $t = 76,\ldots,96$, $\ell = 1,2$ and $A \in D_{\mathrm{MN},t}^{(\ell)}$. Here, the Markov chain involved 10,000 iterations and we see that the component-wise ESS tends to be 7,800 indicating a computationally efficient Markov chains. To further evaluate the computational performance, we compare to the ESS from a LGP version of Model 1. Specifically, consider the model that replaces the truncated log-gamma priors

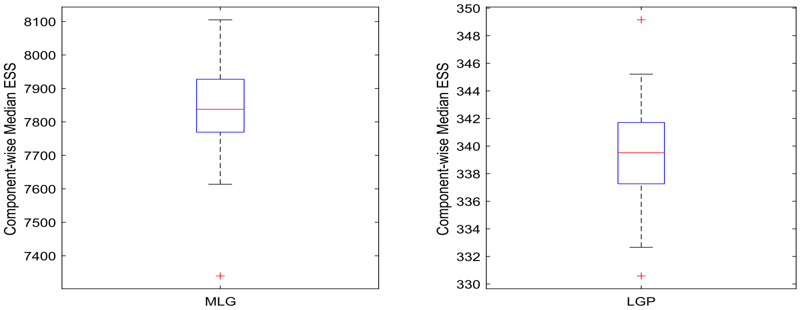


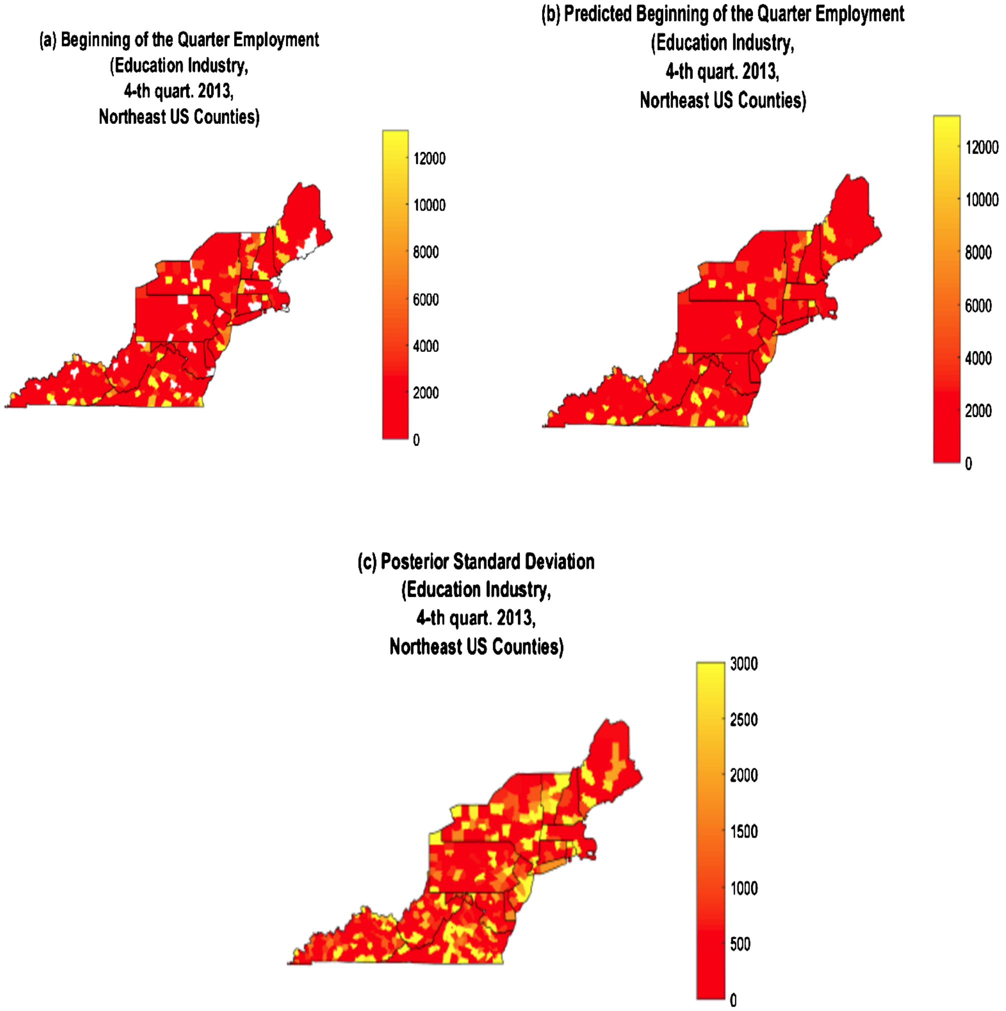
Figure 3: Boxplots of the median ESS, where the median is taken across all $t = 76, \ldots, 96$, $\ell = 1, 2$ and $A \in D_{\mathrm{MN},t}^{(\ell)}$. The boxplot is plotted over the 100 replicate simulations. The left panel gives the boxplot associated with the P-MSTM, and the right panel gives the boxplot associated with the LGP (details are left to Supplemental Appendix C).

with inverse-gamma priors, removes parameter models 4−7 and replaces MLG distributions with Gaussian distributions. The proposal distribution for the Metropolis step was derived using a Taylor series expansion on the Poisson likelihood. This Taylor series expansion leads to a Gaussian likelihood that is used as the proposal distribution (see Supplemental Appendix C in Bradley et al., 2017, for more details). The ESS is also presented in Figure 3. Here, the ESS is consistently lower than 10,000 (the median is approximately 340) suggesting that the Metropolis-within-Gibbs sampler used for this LGP is inefficient.

4.2 Predicting the Mean Beginning of the Quarter Employment

We now show that one can obtain reasonable predictions of the mean number of individuals employed at the beginning of a quarter, over all 3,145 US counties, $L = 20$ NAICS sectors, and $T = 96$ quarters, using the high-dimensional QWI dataset of size 4,089,755 (partially presented in Figure 4(a)). The P-MSTM should be used in settings where there appears to be dependence. As an informal analysis, Moran's I statistics (Moran, 1950) were computed for each spatial field in this dataset. A clear majority of these statistics suggested that spatial dependence is present.

For illustration, we again use the following covariates $\mathbf{x}_t^{(\ell)}(A) = (1, I(\ell = 1), |A|, I(t = 1), \ldots, I(t = 95), \mathrm{population}(A))'$ and set $r = 42$. This choice was also supported using Spiegelhalter et al. (2002)'s deviance information criterion (DIC), where we considered $r = 38, \ldots, 46$, computed the DIC associated with each choice of r, and found that $r = 42$ performs the best in this range of values of r. Thus, there are a total of $p = 99$ large-scale parameters (i.e., $\boldsymbol{\beta}$), $96 \times 42 = 4,032$ small-scale random effects (i.e., $\{\boldsymbol{\eta}_t\}$), a total of 4,089,755 fine-scale random effects (i.e., $\{\boldsymbol{\xi}_t\}$), and $96 \times 6 = 576$ additional parameters (i.e., $\{\sigma_{\mathrm{K},t}\}$, $\{\sigma_{\xi,t}\}$, $\{\alpha_t\}$, $\{\tau_t\}$, $\{\kappa_t\}$, and $\{\theta_t\}$). The analogous LGP posterior distribution is of a similar dimension as the posterior distribution associated with



(a) Beginning of the Quarter Employment (Education Industry, 4-th quart. 2013, Northeast US Counties)

(b) Predicted Beginning of the Quarter Employment (Education Industry, 4-th quart. 2013, Northeast US Counties)

(c) Posterior Standard Deviation (Education Industry, 4-th quart. 2013, Northeast US Counties)

Figure 4: (a), Map of the LEHD estimated number of individuals employed in the beginning of the 4-th quarter of 2013, within the information industry (i.e., $\{Z_{96}^{(1)}(A)\}$), and counties within the Northeast US. The state borders are highlighted as a reference. In (b) and (c), we present the predictions and standard deviations, respectively. Note that (a,b,c) are only a subset of the available QWIs, predictions, and posterior standard deviations. Specifically, there are QWIs available over the 20 NAICS sectors, all US counties, and 96 quarters. Additionally, the predictions and posterior standard deviations have complete coverage over all 20 sectors, 3,145 US counties, and 96 quarters.

the P-MSTM (i.e., see Model 2 in Supplemental Appendix C and compare to Model 1 in the Main-Text). The primary difference is that the LGP does not have shape and rate parameters (i.e., $\{\alpha_t\}$, $\{\tau_t\}$, $\{\kappa_t\}$, and $\{\theta_t\}$), and hence has $96 \times 4 = 384$ fewer parameters. As discussed in Section 3.3, the shape and rate parameters allow for greater flexibility in modeling multivariate spatio-temporal overdispersion.

It consistently takes approximately 5 seconds to compute 1 MCMC iteration from the P-MSTM. Furthermore, the entire chain (of 10,000 iterations) took approximately 14.5 hours to compute, and the preprocessing time (i.e., computing the basis functions, propagator matrices, etc.) took approximately 7 hours to compute. The median ESS for this illustration is approximately 8,121, which indicates that we are obtaining an efficient Markov chain. Moreover, we check batch mean estimates of Monte Carlo error (with batch size 50) (e.g., see Roberts, 1996; Jones et al., 2006), and compute Gelman–Rubin diagnostics based on three independent chains initialized at draws from the prior distribution (e.g., see Gelman and Rubin, 1992). These Gelman–Rubin diagnostics were consistently less than 1.03. All of these diagnostics provide evidence to suggest that there is no lack of convergence of the MCMC algorithm.

In Figures 4(a,b,c), we plot the beginning of quarter employment, the corresponding predictions, and the associated posterior standard deviation. The maps in Figure 4 are for the 4-th quarter of 2013, the education industry, and for counties in Northeast US. It should be emphasized that predictions have been made over all 3,145 US counties, 20 NAICS sectors, and 96 quarters. Upon comparison of Figure 4(a) to Figure 4(b) we see that the predictions reflect the general patterns of the data. Furthermore, the posterior standard deviations in Figure 4(c) are very small (the median is approximately 515) considering that Poisson random variables have their mean equal to their variance. These patterns are consistent across different industries and times. Thus, we see that the in-sample error of the predictors based on the P-MSTM tends to be small and have relatively little bias.

5 Discussion

In this article, we propose a fully Bayesian approach to efficiently model count-valued data jointly over different variables, regions, and/or time-points. To do this, we have introduced a comprehensive framework for jointly modeling Poisson data that could possibly be referenced over different variables, regions, and times. This methodology is rooted in the development of new distribution theory that makes Gibbs sampling for correlated count-valued data computationally feasible. Specifically, we propose a multivariate log-gamma distribution. The MLG distribution leads to computationally efficient sampling of full conditional distributions within a Gibbs sampler. Also, this MLG specification is used within a multivariate spatio-temporal mixed effects model specification, which incorporates non-separable asymmetric non-stationary dependencies.

There are many implications of a general (easy to fit) model for multivariate spatio-temporal count-valued data. First, it is well-known that it is generally more difficult to fit correlated Poisson data than correlated Gaussian data, since Poisson generalized linear mixed models often involve computational expensive Metropolis-Hasting updates within a Gibbs sampler. However, this is no longer the case as Proposition 3 shows that the multivariate log-gamma distribution leads to full-conditional distributions that are easy to simulate from. Moreover, we show that the MLG distribution offers flexibility in modeling overdispersion through shape and rate parameters and the MLG can approximate a Gaussian distribution. Another important implication of the P-MSTM is that it can be used in a wide range of special cases including: spatial-only, times series, multivariate-only, spatio-temporal, multivariate spatial, and multivariate times series datasets.

The generality of the P-MSTM is especially notable considering that the P-MSTM can be applied to "big datasets." It is absolutely crucial that modern statistical methodology be computationally feasible, since "big data" has become the norm with sizes that are ever-increasing. Thus, in this article we demonstrated that the P-MSTM is computationally feasible for a big dataset (of 4,089,755 observations) consisting of count-valued QWIs obtained from US Census Bureau's LEHD program. Furthermore, the P-MSTM was shown to give small in-sample errors. Using an empirically motivated simulation study, we also show that the P-MSTM leads to small out-of-sample errors, and is computationally efficient (in terms of median component-wise ESS).

The P-MSTM is flexible enough to allow for many different specifications. For example, one could use a different class of areal basis functions, point referenced basis functions for lattice data defined on a continuous spatial domain, a different class of propagator matrices, and different parameter models (or even estimates) for covariances. Thus, there are many exciting open research directions, that build on this new distributional framework for count-valued data.

It is important to note that the latent random variables are defined on a lattice (i.e., not in continuous space). This places restrictions on our model. Specifically, questions of spatial coherence (on continuous spatial domains) and relationships to a Poisson process can not be developed for our model (e.g., see Wolpert and Ickstadt, 1998, where they develop these properties for a Poisson-gamma random field). Thus, an important area of future research will be to consider a continuous (i.e., not on a lattice) version of our model, so that these properties of the log-gamma distribution can be investigated.

Supplementary Material

Supplemental Materials: Computationally Efficient Multivariate Spatio-Temporal Models for High-Dimensional Count-Valued Data (DOI: 10.1214/17-BA1069SUPP; .pdf).

References

Abowd, J., Schneider, M., and Vilhuber, L. (2013). "Differential privacy applications to Bayesian and linear mixed model estimation." *Journal of Privacy and Confidentiality*, 5: 73–105. 255

Abowd, J., Stephens, B., Vilhuber, L., Andersson, F., McKinney, K., Roemer, M., and Woodcock, S. (2009). "The LEHD infrastructure files and the creation of the Quarterly Workforce Indicators." In Dunne, T., Jensen, J., and Roberts, M. (eds.), *Producer Dynamics: New Evidence from Micro Data*, 149–230. Chicago: University of Chicago Press for the National Bureau of Economic Research. 255

Anderson, T. (1958). *Introduction to Multivariate Statistical Analysis*. Canada: Wiley and Sons. MR0091588. 257, 259

Bernardoff, P. (2006). "Which multivariate gamma distributions are infinitely divisible?" *Bernoulli*, 12: 169–189. MR2202328. 255

Bradley, J. R., Cressie, N., and Shi, T. (2014). "A comparison of spatial predictors when datasets could be very large." *Statistics Surveys*, 10: 100–131. MR3527662. doi: https://doi.org/10.1214/16-SS115. 254

Bradley, J. R., Holan, S. H., and Wikle, C. K. (2015). "Multivariate spatio-temporal models for high-dimensional areal data with application to Longitudinal Employer-Household Dynamics." *The Annals of Applied Statistics*, 9: 1761–1791. MR3456353. doi: https://doi.org/10.1214/15-AOAS862. 255, 256, 261, 262, 267, 270

Bradley, J. R., Holan, S. H., and Wikle, C. K. (2017). "Supplemental Materials: Computationally Efficient Multivariate Spatio-Temporal Models for High-Dimensional Count-Valued Data." *Bayesian Analysis*. doi: https://doi.org/10.1214/17-BA1069SUPP. 256, 258, 259, 260, 261, 264, 268, 269, 271, 272

Clayton, D., Bernardinelli, L., and Montomoli, C. (1993). "Spatial correlation in ecological analysis." *International Journal of Epidemiology*, 6: 1193–1202. 261

Cressie, N. and Johannesson, G. (2008). "Fixed rank kriging for very large spatial data sets." *Journal of the Royal Statistical Society, Series B*, 70: 209–226. MR2412639. doi: https://doi.org/10.1111/j.1467-9868.2007.00633.x. 262, 264

Cressie, N. and Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data*. Hoboken, NJ: Wiley. MR2848400. 253

Crooks, G. (2015). "The Amoroso distribution." *arXiv preprint: 1005.3274*. 256

Daniels, M., Zhou, Z., and Zou, H. (2006). "Conditionally specified space–time models for multivariate processes." *Journal of Computational and Graphical Statistics*, 15: 157–177. MR2269367. doi: https://doi.org/10.1198/106186006X100434. 256, 264

De Oliveira, V. (2003). "A note on the correlation structure of transformed Gaussian random fields." *Australian and New Zealand Journal of Statistics*, 45: 353–366. MR1999517. doi: https://doi.org/10.1111/1467-842X.00289. 267

De Oliveira, V. (2013). "Hierarchical Poisson models for spatial count data." *Journal of Multivariate Analysis*, 122: 393–408. MR3189330. doi: https://doi.org/10.1016/j.jmva.2013.08.015. 267

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. MR2739155. doi: https://doi.org/10.1016/j.jspi.2010.09.015. 255, 264

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. MR1626544. doi: https://doi.org/10.1111/1467-9876.00113. 253

Garg, S., Singh, A., and Ramos, F. (2012). "Learning non-stationary space-time models for environmental monitoring." In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*. 267

Gelfand, A. and Vounatsou, P. (2003). "Proper multivariate conditional autoregressive models for spatial data analysis." *Biostatistics*, 4: 11–15. 264

Gelfand, A. E. and Schliep, E. M. (2016). "Spatial statistics and Gaussian processes: a beautiful marriage." *Spatial Statistics*, 18: 86–104. MR3573271. doi: https://doi.org/10.1016/j.spasta.2016.03.006. 253, 260

Gelfand, A. E. and Smith, A. (2007). "Disease mapping and spatial regression with count data." *Biostatistics*, 8: 158–183. 253

Gelman, A. and Rubin, D. (1992). "Inference from iterative simulation using multiple sequences." *Statistical Science*, 7: 473–511. 275

Gong, L. and Flegal, J. M. (2016). "A practical sequential stopping rule for highdimensional Markov chain Monte Carlo." *Journal of Computational and Graphical Statistics*, 25: 684–700. MR3533633. doi: https://doi.org/10.1080/10618600.2015.1044092. 272

Griffith, D. (2000). "A linear regression solution to the spatial autocorrelation problem." *Journal of Geographical Systems*, 2: 141–156. 261

Griffith, D. (2002). "A spatial filtering specification for the auto-Poisson model." *Statistics and Probability Letters*, 58: 245–251. MR1920751. doi: https://doi.org/10.1016/S0167-7152(02)00099-8. 261

Griffith, D. (2004). "A spatial filtering specification for the auto-logistic model." *Environment and Planning A*, 36: 1791–1811. 261

Griffiths, R. C. (1984). "Characterization of infinitely divisible multivariate gamma distribution." *Journal of Multivariate Analysis*, 15: 13–20. MR0755813. doi: https://doi.org/10.1016/0047-259X(84)90064-2. 255

Hodges, J. S. and Reich, B. J. (2011). "Adding Spatially-Correlated Errors Can Mess Up the Fixed Effect You Love." *The American Statistician*, 64: 325–334. MR2758564. doi: https://doi.org/10.1198/tast.2010.10052. 261

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*. R. A. Davis, S. H. Holan, R. Lund, and N. Ravishanker (eds). CRC Press. 253

Huang, H. C. and Hsu, N. J. (2004). "Modeling transport effects on ground-level ozone using a non-stationary space-time model." *Environmetrics*, 15: 251–268. 267

Hughes, J. and Haran, M. (2013). "Dimension reduction and alleviation of confounding for spatial generalized linear mixed model." *Journal of the Royal Statistical Society, Series B*, 75: 139–159. MR3008275. doi: https://doi.org/10.1111/j.1467-9868.2012.01041.x. 261, 267, 271

Jin, X., Banerjee, S., and Carlin, B. (2007). "Order-free coregionalized lattice models with application to multiple disease mapping." *Journal of the Royal Statistical Society series B*, 69: 817–838. MR2368572. doi: https://doi.org/10.1111/j.1467-9868.2007.00612.x. 264

Johnson, R. and Wichern, D. (1999). *Applied Multivariate Statistical Analysis, 3rd ed.*. Englewood Cliffs, New Jersey: Prentice Hall, Inc. MR0653327. 255, 257, 259

Jones, G., Haran, M., Caffo, B., and Neath, R. (2006). "Fixed-width output analysis for Markov chain Monte Carlo." *Journal of the American Statistical Association*, 101: 1537–1547. MR2279478. doi: https://doi.org/10.1198/016214506000000492. 274

Kass, R. E., Carlin, B. P., and Neal, R. M. (2016). "Markov chain Monte Carlo in practice: a roundtable discussion." *The American Statistician*, 52: 93–100. MR1628427. doi: https://doi.org/10.2307/2685466. 272

Kotz, S., Balakrishnan, N., and Johnson, N. (2000). *Continuous Multivariate Distributions, Volume 1: Models and Applications*. New York, NY: Wiley. MR1788152. doi: https://doi.org/10.1002/0471722065. 255, 256

Lawson, A. B. (2006). *Statistical Methods in Spatial Epidemiology, 2nd edn.*. New York, NY: Wiley. MR2243369. doi: https://doi.org/10.1002/9780470035771. 254

Lee, Y. and Nelder, J. (1974). "Double hierarchical generalized linear models with discussion." *Applied Statistics*, 55: 129–185. MR2226543. doi: https://doi.org/10.1111/j.1467-9876.2006.00538.x. 255

Lee, Y. and Nelder, J. A. (2000). "HGLMs for analysis of correlated non-normal data." In Bethlehem, J. G. and van der Heijden, P. G. M. (eds.), *COMPSTAT: Proceedings in Computational Statistics 14th Symposium held in Utrecht, The Netherlands, 2000*, 97–107. Utrecht, the Netherlands. 254

Lee, Y. and Nelder, J. A. (2001). "Modelling and analysing correlated non-normal data." *Statistical Modelling*, 1: 3–16. 254

Liu, J. S. (2008). *Monte Carlo Strategies in Scientific Computing*. New York, NY: Springer. MR2401592. 272

Lohr, S. (1999). *Sampling Design and Analysis*. Pacific Grove, CA, USA: Brooks/Cole Publishing Company. MR3057878. 270

Ma, C. (2002). "Spatio-temporal covariance functions generated by mixtures." *Mathematical Geology*, 34: 965–975. MR1951438. doi: https://doi.org/10.1023/A:1021368723926. 267

Moran, P. A. P. (1950). "Notes on Continuous Stochastic Phenomena." *Biometrika*, 37: 17–23. MR0035933. doi: https://doi.org/10.1093/biomet/37.1-2.17. 261, 273

Moran, P. A. P. and Vere-Jones, D. (1969). "The infinite divisibility of multivariate gamma distributions." *Sankhya. Series A*, 40: 393–398. 255

Neal, R. M. (2011). "MCMC Using Hamiltonian Dynamics." In Brooks, S., Gelman, A., Jones, G. L., and Meng, X. (eds.), *Handbook of Markov Chain Monte Carlo*, 113–160. Chapman and Hall. MR2858447. 255

Nieto-Barajas, L. E. and Huerta, G. (2017). "Spatio-temporal pareto modelling of heavy-tail data." *Spatial Statistics*, 20: 92–109. MR3654005. doi: https://doi.org/10.1016/j.spasta.2017.02.003. 254

Prentice, R. (1974). "A log gamma model and its maximum likelihood estimation." *Biometrika*, 61: 539–544. MR0378212. doi: https://doi.org/10.1093/biomet/61.3.539. 256

Quick, H., Holan, S. H., and Wikle, C. K. (2015). "Zeros and ones: a case for suppressing zeros in sensitive count data with an application to stroke mortality." *Stat*, 4: 227–234. MR3405403. doi: https://doi.org/10.1002/sta4.92. 270

Reich, B. J., Hodges, J. S., and Zadnik, V. (2006). "Effects of Residual Smoothing on the Posterior of the Fixed Effects in Disease-Mapping Models." *Biometrics*, 62: 1197–1206. MR2307445. doi: https://doi.org/10.1111/j.1541-0420.2006.00617.x. 261

Robert, C. P. and Casella, G. (2013). *Monte Carlo Statistical Methods*. New York, NY: Springer. MR1707311. doi: https://doi.org/10.1007/978-1-4757-3071-5. 272

Roberts, G. (1996). "Markov chain concepts related to sampling algorithms." In Gilks, W., Richardson, S., and Spiegelhalter, D. (eds.), *Markov Chain Monte Carlo in Practice*, 45–57. Chapman and Hall, Boca Raton. MR1397967. 274

Royle, J. A. and Wikle, C. K. (2005). "Efficient statistical mapping of avian count data." *Environmental and Ecological Statistics*, 12: 225–243. MR2144403. doi: https://doi.org/10.1007/s10651-005-1043-4. 254

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. MR2649602. doi: https://doi.org/10.1111/j.1467-9868.2008.00700.x. 253, 254, 255

Sigrist, F., Kunsch, H. R., and Stehel, W. A. (2011). "A dynamic nonstationary spatiotemporal model for short term prediction of precipitation." *The Annals of Applied Statistics*, 6: 1452–1477. MR3058671. doi: https://doi.org/10.1214/12-AOAS564. 267

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). "Bayesian measures of model complexity and fit." *Journal of the Royal Statistical Society, Series B*, 64: 583–616. MR1979380. doi: https://doi.org/10.1111/1467-9868.00353. 273

Vats, D., Flegal, J. M., and Jones, G. L. (2016). "Multivariate Output Analysis for Markov Chain Monte Carlo." *arXiv preprint: 1512.07713*. MR3653667. 272

Vere-Jones, D. (1967). "The infinite divisibility of a bivariate gamma distribution." *Sankhya. Series A*, 29: 421–422. MR0226704. 255

Wikle, C. K. and Anderson, C. J. (2003). "Limatological analysis of tornado report counts using a hierarchical Bayesian spatio-temporal model." *Journal of Geophysical Research-Atmospheres*, 108: 9005. 254

Wikle, C. K., Milliff, R. F., Nychka, D., and Berliner, L. M. (2001). "Spatiotemporal hierarchical Bayesian modeling tropical ocean surface winds." *Journal of the American Statistical Association (Theory and Methods)*, 96: 382–397. MR1939342. doi: https://doi.org/10.1198/016214501753168109. 262

Wolpert, R. and Ickstadt, K. (1998). "Poisson/gamma random field models for spatial statistics." *Biometrika*, 85: 251–267. MR1649114. doi: https://doi.org/10.1093/biomet/85.2.251. 254, 257, 276

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. MR3110897. doi: https://doi.org/10.1007/s13253-013-0141-2. 254

Acknowledgments

We would like to express our gratitude to the editor-in-chief Bruno Sansó, the editor, the associate editor, and the referees for their very helpful comments that improved this manuscript. 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 paper 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 necessarily those of the NSF or the U.S. Census Bureau.

Invited comment on Article by Bradley, Holan, and Wikle

Stefano Castruccio[*]

Discussion

It is both a pleasure and an honor to be called to discuss a work with such far-reaching implications in spatio-temporal modeling. This paper can indeed set a new (theoretical and applied) standard for modeling dependent counts, so I share the authors' enthusiasm and hopes to propose a considerable more efficient alternative to Gaussian latent processes.

My discussion is divided into two parts, in the first I will discuss modeling aspects that could be implemented as future directions of research to expand this work, while in the second I will focus on general practical issues to allow a wide dissemination of this approach across the statistical community.

1 Modeling considerations

My first comment is about nonstationarity and nonseparability. While the Poisson Multivariate Spatio-Temporal Mixed Effects Model (P-MSTM) is not constrained by stationarity and/or separability, its dependence structure is implied, and not explicitly defined, from the distributional assumptions of the model and the prior. So would it be possible to find particular subclasses of Multivariate Log-Gamma that would allow to capture (more or less abrupt) changes in the dependence structure as dictated by external factors such as (static or dynamic) geographical descriptors? The structure of some count processes from physical science are indeed influenced by sharp natural boundaries such as mountain regions or land/ocean domains.

Secondly, there are applications where inference on a temporal scale smaller than the sampling frequency, or on the gradient might be of interest. While this is an area of spatio-temporal Statistics that is relatively unexplored compared to the continuous space/discrete time, there are some recent examples in the Gaussian setting, e.g. Quick et al. (2013), that prompts me to ask the authors if and how the P-MSTM could be generalized in this direction, and to what extent scalability for massive data sets can be preserved. Could the first order autoregressive structure predicated here for $\boldsymbol{\eta}_t$, encapsulated in the propagator matrix $\mathbf{M}_t$, be generalized to a continuous state process without resorting to linear models of coregionalizations, which would disrupt the scalability?

[*]Department of Applied and Computational Mathematics and Statistics, 153 Hurley Hall, Notre Dame, IN 46556, United States, scastruc@nd.edu

A final and related point is about embedding the lattice into a continuous process. I do agree with the authors on the importance of this, and I believe this is going to be one key point to convince the broad community to use this methodology extensively. The existence of a continuous Gaussian measure in space and time has allowed the development of many theoretical results for both infill and increasing domain asymptotics that would be very valuable to investigate in this setting as well.

2 Inference and dissemination

Inference is efficient, 5 seconds per Markov Chain Monte Carlo (MCMC) iteration for a data set of 4 million observation is noticeable, but 14.5 hours for an analysis is still a long time. So I wonder how and to what extent the algorithm can be parallelized, and if yes, how can this be implemented in a software package (see my next comment), and for which data sets and computer architectures (many fast cores vs fewer slow cores) distributed computing could be beneficial. In the long term, Graphics Processing Unit (GPU) computing could also be explored (and will likely be more beneficial), but I suspect the software is not mature and flexible enough to be able to implement the algorithm in this work.

As a final note, among the fundamental novelties in spatio-temporal Statistics over the last decades, Integrated Nested Laplace Approximation (INLA, Rue et al. (2009)), has played a key role. While the methodological innovation was sizable, a key factor driving its success is the proposal of a comprehensive R package with an ever-increasing set of case studies and automatic tools for inference (Lindgren and Rue, 2015), which fundamentally changed the practice of modeling latent Gaussian processes in space and time, and allowed dissemination far beyond the topical boundaries. In the same spirit, I believe the availability of an R (and possibly MATLAB) package would be a necessary condition for the success of such approach, with a set of appropriate case studies for all settings (multivariate, space, space/time, etc.) and, perhaps most importantly, automatic methods for prior calibration for practitioners with limited Statistics training.

References

Lindgren, F. and Rue, H. (2015). "Bayesian Spatial Modelling with R-INLA." *Journal of Statistical Software, Articles*, 63(19): 1–25. MR2490553. doi: https://doi.org/10.1016/S0169-7161(05)25033-2. 283

Quick, H., Banerjee, S., and Carlin, B. P. (2013). "Modeling Temporal Gradients in Regionally Aggregated California Asthma Hospitalization Data." *Annals of Applied Statistics*, 7: 154–176. MR3086414. doi: https://doi.org/10.1214/12-AOAS600. 282

Rue, H., Martino, S., and Chopin, N. (2009). "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations." *Journal of the Royal Statistical Society: Series B*, 71(2): 319–392. MR2649602. doi: https://doi.org/10.1111/j.1467-9868.2008.00700.x. 283

Contributed comment on Article by Bradley, Holan, and Wikle

William Weimin Yoo[*]

Abstract. I begin my discussion by summarizing the methodology proposed and new distributional results on multivariate log-Gamma derived in the paper. Then, I draw an interesting connection between their work with mean field variational Bayes. Lastly, I make some comments on the simulation results and the performance of the proposed Poisson multivariate spatio-temporal mixed effects model (P-MSTM).

Keywords: multivariate log-Gamma, spatio-temporal, variational Bayes, mean field, Latent Gaussian Process.

I would like to congratulate the authors for such an interesting and important work in spatio-temporal statistics. Indeed, high-dimensional count-valued data is a norm in large-scale census studies across the world, and the authors proposed an efficient and innovative procedure to model this complex data that scales well with its size. Let me briefly summarize their methodology before I begin my discussion. At the highest hierarchy, counts are modelled using a Poisson distribution and the log-link is used to link its mean with the underlying latent process. This latent process in turn has a mixed effects model representation, where the fixed effect is a linear combination of spatio-temporal covariates and the random effect part consisting of spatio-temporal basis functions. The authors took a departure from the Latent Gaussian Process approach (widely regarded as the industry standard), by modeling the fixed and random effects coefficients with multivariate log-Gamma (MLG) priors.

As its name suggests, the log-Gamma is simply the logarithm of a Gamma distributed random variable. The authors then took the opportunity to develop new distributional theory for MLG's. In particular, they derived probability density function of MLG under affine transformation and also their conditional distributions. The most striking result here is Theorem 2, where they established equivalence between the conditional MLG to certain classes of marginal MLG. This then enables them to sample efficiently from conditional MLG's and they designed a fast Gibbs sampler based on this new sampling scheme.

The strategy of reducing the simulation of a complicated conditional MLG to simulation using its equivalent marginal distribution, is reminiscence to another class of methods called Variational Bayes (VB) used especially in the machine learning community for massive data problems. As opposed to Markov Chain Monte Carlo (MCMC) algorithms, VB seeks an analytic approximation to the posterior such that this approximation is close to the posterior in Kullback–Leibler divergence. A widely used strategy in VB is to assume that the approximating multivariate distribution has a factorised

[*] Mathematical Institute, Leiden University, The Netherlands, yooweimin0203@gmail.com

form (mean field VB), e.g., product of marginals across parameters and latent variables. It is conceivable that for a Poisson likelihood and MLG priors as considered in this paper, the resulting best approximating marginals for the parameters will also be a MLG due to conjugacy and I think they will have the same form as in (10) of Theorem 2 in Bradley et al. (2018). As a result, the mean field VB will also produce an iterative procedure much like the Gibbs sampler algorithm proposed by the authors, but it will be a set of circular equations updating the hyperparameters (scale and rate) of the MLG marginal approximations. Although there are not much theory on VB, but empirical studies in real-world massive data applications seem to show that VB has comparable estimation performance as MCMC and is several magnitudes faster (Giordano et al. (2017)).

My other comment centers around the out-of-sample simulation experiment. In Figure 1, the proposed model captures the global spatial pattern well but seems to underestimate regions with high employment, and I was wondering how could one fine-tune the model to better capture these local county-level characteristics. A very natural idea is to include economic indicators for a county (if available) or some seasonality correction term in the fixed effect covariates, since employment numbers depend on economic/commercial activities of a county and they tend to follow job market seasons. My last point is about the performance between the proposed P-MSTM model and the "industry standard" LGP (Latent Gaussian Process). It was discussed in the paper that LGP is inefficient compared to P-MSTM, but I am also curious about the predictive performance of P-MSTM in comparison to LGP for the simulation and the actual data analysis considered in this paper. In particular, whether P-MSTM achieves the same accuracy as LGP using much less computer running time.

Massive and high-dimensional data is now a norm in spatio-temporal statistics, and this paper, through the development of new distributional theory, opens up a way to model and compute these complex datasets. Interesting future work might include generalizing the proposed modeling strategy to encompass both count (discrete) and continuous data. I envision that this paper will inspire new research activities by encouraging statisticians to explore models beyond Gaussian Process and stationarity.

References

Bradley, J. R., Holan, S. H., and Wikle, C. K. (2018). "Computationally Efficient Multivariate Spatio-Temporal Models for High-Dimensional Count-Valued Data." *Bayesian Analysis*, 1–29. Advance publication. 285

Giordano, R., Broderick, T., and Jordan, M. I. (2017). "Covariances, Robustness, and Variational Bayes." ArXiv:1709.02536 [stat.ME]. 285

Contributed comment on Article by Bradley, Holan, and Wikle

Andrew Hoegh[*], Kenneth Flagg[†], and Christian Stratton[‡]

Abstract. Bradley, Holan, and Wikle detail a novel approach for jointly modeling correlated, high-dimensional multivariate count data. The development of the multivariate log-gamma distribution enables conjugate prior specification and efficient Gibbs sampling for a variety of high-dimensional multivariate count data settings. We discuss one small addition that would enable this method to be used with sparse counts in a multivariate zero-inflated Poisson setting.

Keywords: multivariate count data, zero inflated Poisson.

1 Introduction

Bradley, Holan, and Wikle implement and detail a novel methodological approach for jointly modeling correlated, high-dimensional multivariate count data using a Poisson multivariate spatio-temporal mixed effects model (P-MSTM). The development of the multivariate log-gamma distribution in combination with Moran‘s – I basis functions and propagator matrices enable conjugate prior specification and efficient Gibbs sampling. The P-MSTM model is useful for a variety of multivariate count data settings including spatio-temporal structures and more general high-dimensional count data frameworks. This work is a large step forward for dealing with high-dimensional data and we look forward to incorporating it into our future work.

A Poisson sampling model is a convenient option for count data; however, two issues typically require more complicated modeling: overdispersion and excess zeros. The P-MSTM is well equipped to handle overdispersion as detailed in Section 3.3, but the framework cannot directly handle excess zeros. We highlight an extension to the P-MSTM framework that can handle an excess of zeros.

2 Modeling Sparse Counts

Consider a dataset, presented in Hoegh et al. (2015, 2016), containing daily counts of civil unrest protest events in Central and South America. Protests are grouped by the following categories: type of protest (6 levels), violent protest (2 levels) and group protesting (11 levels), which result in a 132-dimensional count vector for protests at each areal location and date. There are 219 distinct areal units composed of states or provinces across the region of interest. Using counts from June 2013 results in a total

[*] Department of Mathematical Sciences, Montana State University, andrew.hoegh@montana.edu
[†] Department of Mathematical Sciences, Montana State University, kenneth.flagg@msu.montana.edu
[‡] Department of Mathematical Sciences, Montana State University, christianstratton@montana.edu

of 867,240 values of which less than 1,000 are non-zero. An abundance of zeros are still present when aggregating data, as the monthly count of protests at each state/province contains roughly 1/3 zeros. The zeros are not directly a problem; however, the proportion of zeros can be incompatible with a Poisson distribution with a specified mean.

A common solution for modeling excess zeros is a mixture distribution of a point mass at zero and a Poisson distribution known as a zero-inflated Poisson (ZIP) (Lambert, 1992). The underlying log-gamma framework can easily be adapted for this model. Consider a univariate setting where $y_j = \tilde{x}^T\tilde{\beta} + \epsilon_j$ is a latent log-gamma random variable from a simplified version of the P-MSTM. Then Z comes from a ZIP if

$$P(Z = k) = \begin{cases} p_0 + (1 - p_0)\exp(-\exp(y)) & \text{if } k = 0 \\ (1 - p_0)\frac{\exp(ky - \exp(y))}{k!} & \text{if } k > 1 \end{cases}$$

where p_0 is the probability of the excess zeros. The probability p_0 can be modeled as a function of covariates, using a generalized linear model framework.

$$g(p_0) \quad = \quad \tilde{x^*}^T\tilde{\gamma}$$

Using log-gamma priors on $\tilde{\beta}$ and ϵ_j and normal priors on $\tilde{\gamma}$ coupled with the inverse Cumulative Distribution Function (CDF) of a normal model as a link function, as in Albert and Chib (1993), permits Gibbs sampling for all parameters in the model. Note this model assumes a different parameter vector is used in the point mass probability and the Poisson mean term; otherwise, Gibbs samples would no longer be possible.

The formulation for a multivariate ZIP is considerably more complicated, but the P-MSTM can be used to improve the computational efficiency to make this model more feasible in higher dimensions. Consider a setting with $p = 3$, then the response for each individual component can come from a point mass at zero or from a Poisson distribution using the P-MSTM framework. With $p = 3$, eight mixture components are necessary to enumerate all of the combinations of point mass terms and Poisson distributions. Each mixture component contains a mixture probability and potentially a P-MSTM term. For instance in the case where the $z = \{0, 0, 1\}$, each zero could come from a point mass term or a Poisson distribution so this response could come from four possible components shown below.

$$\begin{cases} p_0\{Pois(\exp(y_1)), Pois(\exp(y_2)), Pois(\exp(y_3))\} \\ p_{11}\{\delta(z_1 = 0), Pois(\exp(y_2)), Pois(\exp(y_3))\} \\ p_{12}\{Pois(\exp(y_1)), \delta(z_2 = 0), Pois(\exp(y_3))\} \\ p_{2_{12}}\{\delta(z_1 = 0), \delta(z_2 = 0), Pois(\exp(y_3))\} \end{cases}$$

Similar to the univariate case, the set of mixture probabilities could be computed as a function of covariates. A computationally efficient approach using a multivariate probit model is described in more detail in Hoegh and Leman (2017).

In dimensions higher than 3, this approach would be much more computationally expensive, but the efficiency provided by the multivariate log-gamma distribution and the P-MSTM is a big step forward for modeling data of this type. Another avenue to explore would be using the random effect vector to model the mixture probabilities and the latent process in the multivariate log-gamma framework. The priors specified above would not permit Gibbs sampling, but perhaps the CDF of the multivariate log-gamma could be used in a creative way as the link function to enable efficient computing similar to multivariate probit model in Chib and Greenberg (1998).

References

Albert, J. H. and Chib, S. (1993). "Bayesian analysis of binary and polychotomous response data." *Journal of the American statistical Association*, 88(422): 669–679. MR1224394. 287

Chib, S. and Greenberg, E. (1998). "Analysis of multivariate probit models." *Biometrika*, 85(2): 347–361. 288

Hoegh, A., Ferreira, M. A., and Leman, S. (2016). "Spatiotemporal model fusion: multiscale modelling of civil unrest." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 65(4): 529–545. 286

Hoegh, A. and Leman, S. (2017). "Correlated model fusion." *Applied Stochastic Models in Business and Industry*, n/a–n/a. Asmb.2261. URL http://dx.doi.org/10.1002/asmb.2261 287

Hoegh, A., Leman, S., Saraf, P., and Ramakrishnan, N. (2015). "Bayesian model fusion for forecasting civil unrest." *Technometrics*, 57(3): 332–340. MR3384948. doi: https://doi.org/10.1080/00401706.2014.1001522. 286

Lambert, D. (1992). "Zero-inflated Poisson regression, with an application to defects in manufacturing." *Technometrics*, 34(1): 1–14. 287

Contributed comment on Article by Bradley, Holan, and Wikle

Kevin He[*] and Jian Kang[†]

We congratulate the authors on their excellent work. Our comments will focus on the following three aspects: model flexibility, computation efficiency and a potential application.

Model flexibility The proposed multivariate log-gamma distribution is a useful method that generates the dependence among multiple gamma random variables which can be used to specify the prior for the Poisson models. It is also a general multivariate continuous distribution, which may provide extra flexibility compare to the multivariate Gaussian distribution. A natural question is that how flexible the mean and covariance structure of this model can be. Let $\mathbb{R}^m$ be an Euclidean vector space of dimension m and SPD^m represents an $m \times m$ symmetric positive definite matrix. It is well known that for any $\boldsymbol{\mu} \in \mathbb{R}^m$ and $\boldsymbol{\Sigma} \in \mathrm{SPD}^m$, we can uniquely determine a m-variate Gaussian distribution with the mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. According to Theorem 1, for $\mathbf{q} \sim \mathrm{MLG}(\mathbf{c}, \mathbf{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa})$, we may re-parameterize the distribution of $\mathbf{q}$ based on its mean and covariance structure. We need to find the log-gamma distribution parameters $\mathbf{c}, \mathbf{V}, \boldsymbol{\alpha}$ and $\boldsymbol{\kappa}$ such that $\mathrm{E}(\mathbf{q}) = \boldsymbol{\mu}$ and $\mathrm{Cov}(\mathbf{q}) = \boldsymbol{\Sigma}$, We describe one approach here. We first perform the eigen decomposition of covariance matrix $\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}}$, where $\boldsymbol{\Lambda} = \mathrm{diag}\{\boldsymbol{\lambda}\}$ and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)^{\mathrm{T}}$ with λ_i being the ith largest eigenvalue for $i = 1, \ldots m$. Then for any $\boldsymbol{\kappa} \in \mathbb{R}^{+m}$, we have

$$\mathbf{V} = \mathbf{U}, \qquad \boldsymbol{\alpha} = \omega_1^{-1}(\boldsymbol{\lambda}), \qquad \mathbf{c} = \boldsymbol{\mu} - \mathbf{U}[\omega_0\{\omega_1^{-1}(\boldsymbol{\lambda})\} - \log(\boldsymbol{\kappa})]. \tag{1}$$

Obviously, the mean and covariance structure do not uniquely determine all the parameters in the multivariate log-gamma distribution. What is the gain of the extra flexibility of the log-gamma distribution compared to the multivariate Gaussian distribution? In some applications, we would like to specify a particular correlation structure for the spatial and/or temporal random effects for the model simplicity and interpretability. For example, if we would like to assume the covariance of $\mathbf{q}$ is compound symmetry, that is, $\mathrm{Cov}(\mathbf{q}) = \sigma^2\{(1-\rho)\mathbf{I}_m + \rho\mathbf{1}_m\mathbf{1}_m^{\mathrm{T}}\}$. In this case, we may use (1) to specify $\mathrm{MLG}(\mathbf{c}, \mathbf{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa})$ and write $\mathbf{c}, \mathbf{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa}$ as a function of ρ and σ^2, while it is worth learning if there exists an easier or more intuitive way to perform structural covariance specifications.

Computational efficiency As mentioned by the authors, the latent Gaussian processes (LGP) and Poisson gamma random fields (PGRF) (Wolpert and Ickstadt, 1998) have been widely used for modeling the dependence of count-value data using a Poisson

[*] Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, kevinhe@umich.edu

[†] Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, jiankang@umich.edu

model from the Bayesian perspective. However, the posterior computations of LGP and PGRF can be quite challenging. The main reason is that the full conditionals of the posterior distributions are intractable for both models; and thus the Metropolis-Hasting type algorithms such as the Hamiltonian Monte Carlo (HMC) and its variations (Girolami and Calderhead, 2011) can be adopted. The convergence of Markov chain Monte Carlo (MCMC) algorithms for the count-value Bayesian models can still be slow in some cases, especially for high-dimensional problems (Ge et al., 2014; Kang et al., 2014; Kang and Johnson, 2014). For the proposed Poisson multivariate spatio-temporal mixed effects model (PMSTM) with the multivariate log-gamma priors, the full conditionals are available, which makes the implementations of the posterior computation straightforward. What about the convergence of this Gibbs sampler? It is well known that the Gibbs sampler does not converge faster than a simple random walk when the target distribution is a highly correlated bivariate Gaussian distribution (Liu, 2008). For the proposed PMSTM with log-gamma priors, is the convergence of the Gibbs samplers faster than an HMC algorithm? To answer this question, it may be helpful to study the effective sample size (Gelman et al., 2014) that the two algorithms can generate during a fixed computing time period.

Applications The proposed PMSTM framework is general and can have many different applications. We focus here on a study of Chronic Kidney Disease (CKD), which has emerged as a major non-communicable disease (NCD) with public health importance, affecting more than 5% of population around the world (Couser et al., 2011). In rapidly developing nations such as China, risk factor profiles of the population are constantly evolving, resulting in increasing likelihood of rising burden of multiple comorbid conditions such as obesity, diabetes, hypertension, cardiovascular diseases, cancer and kidney disease (Zhang et al., 2012). In an effort to control and manage the kidney disease, many nations including both the United States (US) and China have initiated comprehensive CKD Surveillance. The number of CKD cases with different stages (ranging from 1 to 4) are commonly collected for many small regions in those nations. Compared with data from developed countries, the spectrum of CKD in China shows an interesting pattern. Although the overall prevalence is similar, the prevalence of stage 3 and stage 4 CKD in China are lower than those in developed countries. For example, the prevalence of stage 3 CKD was 1.6%, compared with 7.7% in the US. Furthermore, despite that half of dialysis patients were diagnosed as glomerulonephritis, population-based studies revealed that risk factors for CKD were hypertension and diabetes, which are similar to studies from developed countries. One hypothesis is that rapidly increased prevalence of hypertension and diabetes during the last 20 years has led to larger numbers of patients with early stage CKD in China, and it will take a longer time to observe their effect on later stages of CKD. One question of interest is to apply the PMSTM to assess the change in risk factor pattern and distribution of prevalence of CKD (by stage) in the US, and predict the future of CKD burden in China, given potential change in risk factor burden in that country. The model can be fairly complex and may include hierarchical random effects at the patient level, at the region level as well as at the nation level. The efficiency of the posterior computation becomes really important for the practical use of those models. The prior specifications and posterior computation algorithms proposed in this article will provide a promising solution.

Conclusion The computational and theoretical results of the PMSTM shed new light on modeling large-scale multivariate spatial-temporal count-valued data. We believe that this method is useful and applicable in many settings. We hope our discussion convey this message successfully.

References

Couser, W. G., Remuzzi, G., Mendis, S., and Tonelli, M. (2011). "The contribution of chronic kidney disease to the global burden of major noncommunicable diseases." *Kidney International*, 80(12): 1258–1270. 290

Ge, T., Müller-Lenke, N., Bendfeldt, K., Nichols, T. E., and Johnson, T. D. (2014). "Analysis of multiple sclerosis lesions via spatially varying coefficients." *The Annals of Applied Statistics*, 8(2): 1095. 290

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian data analysis*, volume 2. CRC press Boca Raton, FL. MR3235677. 290

Girolami, M. and Calderhead, B. (2011). "Riemann manifold langevin and hamiltonian monte carlo methods." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2): 123–214. MR2814492. doi: `https://doi.org/10.1111/j.1467-9868.2010.00765.x`. 290

Kang, J. and Johnson, T. D. (2014). "A slice sampler for the hierarchical Poisson/Gamma random field model." *Perspectives on Big Data Analysis: Methodologies and Applications*, 622: 21. 290

Kang, J., Nichols, T. E., Wager, T. D., and Johnson, T. D. (2014). "A Bayesian hierarchical spatial point process model for multi-type neuroimaging meta-analysis." *The Annals of Applied Statistics*, 8(3): 1800. 290

Liu, J. S. (2008). *Monte Carlo strategies in scientific computing*. Springer Science & Business Media. MR2401592. 290

Wolpert, R. L. and Ickstadt, K. (1998). "Poisson/gamma random field models for spatial statistics." *Biometrika*, 85(2): 251–267. 289

Zhang, L., Wang, F., and Wang, L. e. a. (2012). "Prevalence of chronic kidney disease in China: a cross-sectional survey." *Lancet*, 379: 815–822. 290

Invited comment on Article by Bradley, Holan, and Wikle

Catherine A. Calder[*][‡] and Candace Berrett[†]

Abstract. We provide a discussion of the article "Computationally Efficient Multivariate Spatio-Temporal Models for High-Dimensional Count-Valued Data" by Bradley, Holan, and Wikle. In our opinion, this work constitutes a major contribution to the field of spatio-temporal statistics and contains distribution theory that should be broadly applicable. In this note, we reflect on modeling decisions made by the authors. We include a small set of simulation results to illustrate the effect of one aspect of the proposed model.

MSC 2010 subject classifications: Primary 62H11.

Keywords: convolution prior, parameter identifiability, model approximation, spatial statistics.

In their article "Computationally Efficient Multivariate Spatio-Temporal Models for High-Dimensional Count-Valued Data," Bradley, Holan, and Wikle (BHW) tackle a collection of statistical challenges that are at the forefront of research on spatio-temporal modeling: efficient computation in high-dimensional data settings, non-Gaussian data, multivariate responses, and spatial confounding. While contributions in any one of these areas are valuable, this paper simultaneously advances the field in each of these key directions making it a real tour de force. Beyond the importance to the field of spatio-temporal statistics, the distribution theory presented in the paper is important in its own right. We hope this paper reaches a broad audience as there is much food-for-thought for statisticians.

In the sections below, we comment on several aspects of this paper. None of these comments are intended to be critical. Rather they summarize our reflections on a few of the paper's many contributions and reflect on modeling decisions of interest to us and, perhaps, suggest directions for future research.

1 Uncorrelated heterogeneity?

A longstanding debate in the disease mapping literature – arguably, the setting that has served as the testbed for the development of spatio-temporal generalized linear mixed models for count data for the last 25 years – is whether to include both spatially-

[*] Department of Statistics, The Ohio State University, 404 Cockins Hall, Columbus, OH 43210, calder@stat.osu.edu

[†] Department of Statistics, Brigham Young University, 223 TMCB, Provo, UT 84602, cberrett@stat.byu.edu

[‡] Support for this research was provided in part by NIH-NICHD R01 HD088545 and The Ohio State University Institute for Population Research (NIH-NICHD P2CHD058484).

structured and non-spatial (i.e., exchangable) random effects in the model for the log relative risk (LRR). In their seminal paper, Besag et al. (1991) advocate for a "convolution prior" for the component of the LRR capturing unobserved heterogeneity: this component for region i is modeled as the sum of u_i and v_i, where the u_is (spatial components) follow an intrinsic autoregression model and the v_is (non-spatial components) are assumed to be independent and identically distributed, conditional on an unknown variance parameter. The u_is capture unobserved spatially-structured factors influencing the LRR, while the v_is capture remaining unobserved sources of uncorrelated heterogeneity.

Besag et al. (1991) acknowledge, "In practice, it will often be the case that either u or v dominates the other but which one will not usually be known in advance." Hence, the convolution prior, in theory, allows the data to decide the relative importance of the two sources of heterogeneity. It has long been recognized, however, that the hyperparameters in convolution priors are only weakly identifiable. Numerous solutions to this identifiability issue have been proposed. For example, Bernardinelli et al. (1995) suggest taking the marginal standard deviation of v_i to be equal to the conditional standard deviation of $u_i | u_{j \neq i}$ divided by 0.7. Alternatively, Rue and Held (2005) propose a strategy based on calculating the marginal variances of the u_is implied by a conditional autoregressive prior for the u_is.

In BHW's paper, they too propose a convolution-style prior. For example, consider the purely spatial version of the Poisson multivariate spatio-temoral mixed effects model (P-MSTM) presented in Appendix E. In this case, the random component of the log expected count for area A is

$$\underbrace{\boldsymbol{\psi}_1^{(1)\prime}(A)\,\boldsymbol{\eta}_1}_{\text{(spatial)}} + \underbrace{\boldsymbol{\xi}_1^{(1)}(A)}_{\text{(non-spatial)}}. \tag{1}$$

Note that this special case differs from the count-data version of Hughes and Haran (2013)'s dimension-reduced Spatial Generalized Linear Mixed Model (SGLMM; described in their Section 6.2). In Hughes and Haran (2013)'s model, where dimension reduction via truncated Moran basis functions was first introduced, a term allowing for uncorrelated (i.e., non-spatial) heterogeneity is not included. Instead of a random component in the form of a sum as in (1), Hughes and Haran (2013)'s model includes only a spatial random effect, $\boldsymbol{M}\boldsymbol{\delta}_S$, defined in more detail below.

To explore the implications of including the non-spatial component in the model described in BHW's Appendix E, we simulated data from Hughes and Haran (2013)'s Gaussian SGLMM for count data and an analogous log gamma model. To establish a unified notation for these models, we let $Z_i | \lambda_i \sim \text{Pois}(\lambda_i)$, independently for $i = 1, \ldots, n$, where the Z_is are count-valued random variables associated with the nodes of a square grid graph, $\mathcal{G}$, wrapped onto a torus so that every node in $\mathcal{G}$ has four neighbors. Let $\boldsymbol{A}$ denote the adjacency matrix corresponding to $\mathcal{G}$. The elements of $\boldsymbol{A}$, $\boldsymbol{A}_{ij}$, are equal to zero unless nodes i and j are first-order neighbors in which case they equal one. We consider the following four models for $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_n)'$:

	Spatial + Non-Spatial	*Spatial Only*
Gaussian	$\boldsymbol{\lambda} = \exp\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{M}^{(r)}\boldsymbol{\delta} + \boldsymbol{\epsilon}\right)$	$\boldsymbol{\lambda} = \exp\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{M}^{(r)}\boldsymbol{\delta}\right)$
Log Gamma	$\boldsymbol{\lambda} = \exp\left(\boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{M}^{(r)}\boldsymbol{\delta}^* + \boldsymbol{\epsilon}^*\right)$	$\boldsymbol{\lambda} = \exp\left(\boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{M}^{(r)}\boldsymbol{\delta}^*\right)$

where exp() here denotes the element-wise exponent of its vector argument, $\boldsymbol{X}$ is an $n \times p$ fixed design matrix with corresponding coefficient $\boldsymbol{\beta}$ or $\boldsymbol{\beta}^*$, and $\boldsymbol{M}^{(r)}$ is a matrix consisting of the first r eigenvectors of the Moran operator for $\boldsymbol{X}$ with respect to $\mathcal{G}$ (as defined in Hughes and Haran, 2013). The random term in the spatial component of the Gaussian models, $\boldsymbol{\delta}$, is an r-dimensional normally distributed random vector with mean zero and covariance matrix $(3\boldsymbol{Q}_S)^{-1}$, where $\boldsymbol{Q}_S = \boldsymbol{M}^{(r)\prime}\boldsymbol{Q}\boldsymbol{M}^{(r)}$, $\boldsymbol{Q}$ is the Laplacian of $\mathcal{G}$ and equals diag$(\boldsymbol{A}\mathbf{1}_n) - \boldsymbol{A}$, $\mathbf{1}_n$ is the n-dimensional vector consisting of all ones, and diag() denotes the diagonal matrix with main diagonal equal to its argument. For the log gamma models, $\boldsymbol{\delta}^*$ is an r-dimensional random vector assumed to follow the multivariate log gamma distribution MLG$(\mathbf{0}_r, \boldsymbol{V}, \alpha_{\delta^*}\mathbf{1}_r, \omega_{\delta^*}\mathbf{1}_r)$, as defined in BHW's Section 2.1, $\boldsymbol{V}$ is the lower Cholesky factor of the matrix $(3\boldsymbol{Q})^{-1}$, and α_{δ^*} and ω_{δ^*} are scalars. In the Gaussian spatial+non-spatial model, $\boldsymbol{\epsilon}$ is an n-dimensional random vector with mean zero and covariance matrix $\sigma^2_\epsilon \boldsymbol{I}_r$, where σ^2_ϵ is a scalar and $\boldsymbol{I}_r$ is the $r \times r$ identity matrix. In the log gamma spatial+non-spatial model, $\boldsymbol{\epsilon}^*$ is an n-dimensional random vector assumed to follow the multivariate log gamma distribution MLG$(\mathbf{0}_n, \mathbf{1}_n, \boldsymbol{I}_n, \alpha_{\epsilon^*}\mathbf{1}_m, \omega_{\epsilon^*}\mathbf{1}_r)$, where α_{ϵ^*}, and ω_{ϵ^*}, are scalars.

Setting $n = 400$ (equivalently, a 20×20 grid), we simulated 10 realizations of $\boldsymbol{\lambda}$ from both the Gaussian and log gamma models, with and without the non-spatial components. $\boldsymbol{X}$ was taken to be a column of ones, $\boldsymbol{\beta} = 1$ and $\boldsymbol{\beta}^* = 1$, $\sigma^2_\delta = 1$, $\alpha_{\delta^*} = 1.5$ and $\omega_{\delta^*} = 1$ (making E$[\delta^*_u] \approx 0$ and var$[\delta^*_u] \approx 1$, for $u = 1, \ldots, r$)[1], $\sigma^2_\epsilon = 0.01$, and $\alpha_{\epsilon^*} = 100$ and $\omega_{\epsilon^*} = 100$ (making E$[\epsilon^*_i] \approx 0$ and var$[\epsilon^*_i] \approx 0.01$, for $i = 1, \ldots, n$). These choices imply a roughly one order of magnitude difference in the standard deviation of the spatial signal (δ_u and δ^*_u, for $u = 1, \ldots r$) and non-spatial signal (ϵ_i and ϵ^*_i, for $i = 1, \ldots, n$). We then calculated cor$[\lambda_i, \lambda_j]$ for all pairs i, j that are mth order spatial neighbors, where $m = 1, \ldots, 10$ is the "spatial lag."

Figure 1 summarizes the empirical spatially-lagged correlations between the elements of $\boldsymbol{\lambda}$ as a function of the spatial lag, m. The columns in Figure 1 indicate whether $\boldsymbol{\epsilon}$ or $\boldsymbol{\epsilon}^*$ are included in the models and the rows correspond to the amount of dimension reduction. Here, $r = 200$ implies that all Moran basis functions with a corresponding non-negative eigenvalue are retained, while $r = 100$ and $r = 20$ correspond to more extensive amounts of dimension reduction. For the spatial-only simulations (right column plots), the correlation as a function of spatial lag appear similar for the Gaussian and log gamma models. That is, the variability across simulated datasets is similar to the variability across models. The only discernible difference across the rows in the right column is in the slope of the empirical spatially-lagged correlation function, which appears to die off more slowly for smaller values of r. On the other hand, for each value of r, the Gaussian and log gamma models that include a non-spatial component are apparently

[1] We note that the expression for the expected value of q, a log gamma random variable, should be E$[q] = \omega_0(\alpha) - \log(\kappa)$, instead of E$[q] = \omega_0(\alpha) + \log(\kappa)$ as it appears in Section 2 of BHW's paper.

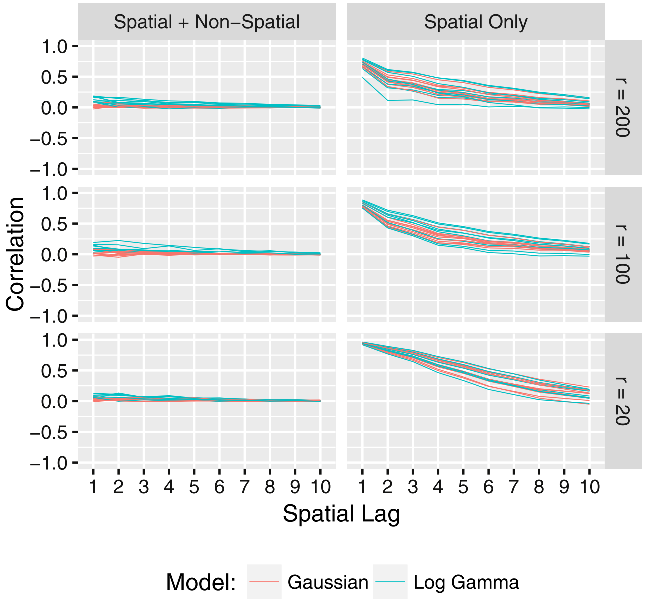


Figure 1: Empirical correlation between pairs of λ_i and λ_j as a function of the spatial lag for spatial+non-spatial and spatial-only versions of Gaussian and log gamma models for various amounts of dimension reduction indicated by the value of r.

different in that the correlations at smaller lags are larger for the log gamma model than they are for the Gaussian model. Unlike the spatial-only models, the variability across simulated data is smaller than the variability across the models.

What do our model comparisons imply? For the spatial-only models, the log gamma model appears to be able to capture the same spatial-dependence structure implied by the Gaussian model. This is not the case in the spatial+non-spatial models. Of course, it is conceivable that different choices of model hyperparameters could render the spatially-lagged correlation functions to be nearly identical. The point, however, is that the convolution-style specification of the log gamma model requires some thought. Unlike the corresponding Gaussian model where the sum of the spatial and non-spatial components follows a multivariate normal distribution, the distribution of the sum of these terms in the log gamma model are not multivariate log gamma. This complicates the study of identifiability of the log gamma hyperparameters. Furthermore, it is not immediately clear that $\mathrm{E}[\boldsymbol{\epsilon}_i^*]$ can equal zero under BHW's specification for $\boldsymbol{\epsilon}^* \equiv \boldsymbol{\xi}_1$:

$$\boldsymbol{\xi}_1|\sigma_{\xi,1} \sim \mathrm{MLG}\left(\mathbf{0}, \alpha^{1/2}\sigma_{\xi,1}\boldsymbol{I}, \alpha\mathbf{1}, \frac{1}{\alpha}\mathbf{1}\right),$$

since for $\theta \sim \mathrm{Ga}(\alpha, 1/\alpha)$, $\log(\mathrm{E}[\theta]) = 0$ but, of course, $\log(\mathrm{E}[\theta]) \neq \mathrm{E}[\log(\theta)]$ in general. In any case, before the "spatial-only" version of the P-MSTM is rolled-out as a general framework for spatial modeling of count data, additional thought is needed about the specification of process models and prior distributions for the hyperparameters in them.

2 Further thoughts

In the spatial statistics literature, there has been an ongoing discussion about whether it is preferable to approximate a model to facilitate inference or, alternatively, perform approximate inference for the exact (i.e., not approximated) model. If the Poisson STGLMM (temporal extension of the Poisson SGLMM) is thought to be the exact model and BHW's log gamma model is thought to approximate it, then avoiding "*ad hoc* Metropolis-Hastings algorithms" – a stated goal of BHW – can be viewed as favoring exact inference for an approximate model over approximate inference for the exact model. Of course, Metropolis-Hastings algorithms are not "approximate" in the same sense as other methods for fitting Bayesian models that are generally referred to as approximate (e.g., variational Bayes, integrated nested Laplace approximations (INLA)). But, perhaps, a poorly mixing Metropolis-Hastings algorithm could be viewed as providing approximate inference, albeit not in a rigorous sense. Such approximate inference for the exact model can certainly be avoided by specifying a P-MSTM instead of a traditional Gaussian process model. Alternatively, like Kaufman et al. (2011)'s argument for the model they propose for cosmology computer experiments, the P-MSTM can be viewed simply as a different model, as opposed to an approximate model, which happens to allow for inference via a Gibbs sampler. BHW's paper does not spell out this argument explicitly, leading us to question whether the authors view P-MSTM as an approximate model or whether in certain situations it is the preferred exact model.

Lastly, we note the many robustness checks and modeling decisions BHW make in their analysis of Quarterly Workforce Indicators (QWIs). These include the use of a discrete uniform prior distribution with a known upper bound for certain parameters and the need to fit multiple versions of their model to assess the sensitivity of the results to fixed hyperparameters. It appears that more of these types of decisions are needed for the P-MSTM than for the Poisson STGLMM. Based on BHW's experience performing these sensitivity checks in their analysis of QWI data, we wonder if they view the extra flexibility of their model as an advantage or as a nuisance in that using it requires more extensive robustness checks.

References

Bernardinelli, L., Clayton, D., and Montomoli, C. (1995). "Bayesian estimates of disease maps: How important are priors? Bayesian estimates of disease maps: How important are the priors?" *Statistics in Medicine*, 14(2411–2431). 293

Besag, J., York, J., and Mollié, A. (1991). "Bayesian image restoration, with two applications in spatial statistics." *Annals of the Institute of Mathematical Statistics*, 43(1): 1–59. MR1105822. doi: https://doi.org/10.1007/BF00116466. 293

Hughes, J. and Haran, M. (2013). "Dimension reduction and alleviation of confounding for spatial generalized linear mixed models." *Journal of the Royal Statistical Society, Series B*, 75(139–159). MR3008275. doi: https://doi.org/10.1111/j.1467-9868.2012.01041.x. 293, 294

Kaufman, C., Bingham, D., Habib, S., Heitmann, K., and Frieman, J. (2011). “Efficient emulators of computer experiments using compactly supported correlation functions, with an application to cosmology.” *The Annals of Applied Statistics*, 5(4): 2470–2492. MR2907123. doi: `https://doi.org/10.1214/11-AOAS489`. 296

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Boca Raton, Florida: Chapman & Hall/CRC. 293

Invited comment on Article by Bradley, Holan, and Wikle

D. Gamerman

1 Introduction

The paper presents an approach for inference on spatio-temporal data that is alternative to the usual approach based on latent Gaussian processes (LGP). Their approach is based on the multivariate Gamma distribution (MGD) that conjugates with Poisson distribution with logarithmic link. The convenience of conjugacy is explored and highlighted in terms of easily specified full conditional distributions and hence easy sampling for use with Markov chain Monte Carlo (MCMC). The authors make it clear that their approach is restricted to Poisson data and is geared towards highly dimensional data sets.

My contribution to the discussion is organized in sections for modelling, computation and application. The sections are obviously not mutually exclusive but I will try to separate them as much as possible for the sake of clarity.

2 Modelling

The introduction of the MGD as prior distribution for analysis is a welcomed addition for the tool set of practitioners. The fact that they *might* be more useful than LGP because they introduce additional shape and scale parameters could be more clearly identified. I suspect that some Poisson data will confirm that and some won't. It would be nice to identify situations were each is more advantageous.

In terms of the additional parameters (with respect to LGP) it seems clear that shape parameters play a different role but the picture does not seem so clear for the scale parameters. The expressions of the density and the mean of MGD suggest that there may be identification issues between $\boldsymbol{\kappa}$ and $\mathbf{c}$ (or between $\log \boldsymbol{\kappa}$ and $\mathbf{V}^{-1}\mathbf{c}$).

The authors apply dimension reduction to their spatial component possibly due to very large dimension of the applications they have in mind. An alternative, more basic formulation for this component is with random terms whose correlation matrix is defined via a correlation function depending on the distance or neighborhood structure between observation sites. Would it be possible to accommodate the above specification in their framework?

The authors proposed the use of discrete uniform distributions for the shape parameters of the Gamma distributions, albeit with some cautionary remarks. It would

Departmento de Métodos Estatisticos, Instituto de Matemática, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil, dani@im.ufrj.br

be nice to have an explanation for this choice. I suspect it is due to the difficulties associated with estimation of these quantities but I prefer the authors to register their more knowledged experience on the subject, specially with their evaluation on other (continuous) distribution they may have tried. The smaller value they consider for the shape parameter in their application is 200. Any guidance on why can one safely start at such a (high?) value and why do we have to proceed up until the value 10,000? Even more generally, is there information in the data for these parameters? Plots and other summaries of (marginal) posterior densities for some of them could also be useful.

The similarity between many features of Poisson data and point pattern data begs for an extension to handle Poisson process data. In that context, LGP are particularly useful for their characterization in such infinite-dimension settings. Gonçalves and Gamerman (2018) explore this possibility in the realm of Bayesian inference for Poisson process data via MCMC. I wonder if the authors can envisage the extension of their approach towards point pattern analysis (via Poisson processes) and/or an infinite dimension extension of their multivariate Gamma distributions into stochastic processes.

3 Computation

The authors indicate the use of Gibbs sampling as an advantage in terms of simplicity of their sampling scheme as opposed to the need for tuning of other MCMC schemes and the elaboration required for Hamiltonian Monte Carlo. It is useful to point out that there are other simple, model-based MCMC schemes based on proposals that do not require tuning. Gamerman (1997) is an example in a similar setting of this paper.

The discretised prior adopted for the shape parameters introduces tuning problems associated with the choice of number and values of points considered. Although different in nature, this discretisation brings back the tuning problems they seem to have avoided with the closed-form expressions for all full conditional distributions of their unknowns.

The models of the paper handle temporal dependence via the state-space formulation for the time dependent parameters $\boldsymbol{\eta}_t$. These components are sampled separately for each time point considered. There is a well documented literature about the difficulty in achieving efficient MCMC schemes when sampling each time parameter separately for Gaussian (see Carter and Kohn, 1996; Frühwirth-Schnatter, 1994) and for non-Gaussian (Gamerman, 1998) data. In fact, the difficulties are not associated with the nature of the likelihood but with the strength of the association in the latent state-space component. So, I would expect that the same troubles would have appeared here.

Nevertheless, the results of Figure 1 seem to indicate an efficient MCMC scheme. Reasons for it may be a weak temporal correlation or a relatively small number of time points for their specific application. Even so, these features may not stand for all other Poisson applications. So I wonder whether it is possible to devise alternative sampling schemes based on block sampling and/or reparametrisation (see De Jong and Shephard, 1995; Gamerman, 1998). Plots of the autocorrelation function of $\boldsymbol{\eta}_t$'s or related quantities would be useful additions for the assessment of the efficiency of the MCMC scheme they proposed.

The approach proposed in the paper is contrasted against the currently prevailing approach of LGP only in terms of the computations. They concentrate on computing performance evaluated via effective sample sizes (ESS). It was not surprising to see the preference for the MGD-based approach but for the order of magnitude of the difference. Any explanation for such a strong advantage for the criteria used? A more directly interpretable definition of ESS is based on the (estimated) variance of mean trajectories of Markov chains (see Gamerman and Lopes, 2006), taking into account the (estimated) chain autocorrelations. It would be interesting to see the results of this comparison via this statistic and other statistics related to efficiency evaluation of the sampling scheme.

4 Application

Predictive assessment is crucial for evaluation of any modeling strategy. The authors use visual assessment to check the adequacy of their approach. There are many tools developed for this task, starting from aggregated values of the point predictions but in any case going beyond visual inspections (e.g. Gneiting et al., 2007). These would allow for a more comprehensive evaluation of the relative merits of each assessed model. Also, it would be important to compare their predictions against those for competing models, eg LGP-based.

In that sense, Figure 1 and 2 could incorporate results of the corresponding LGP model. This would be a useful addition, providing hints of the relative merits of both approaches. This comparison could shed some light on why/where one approach is performing better than the other one. By doing that, the authors could not only inform the readers on whether they consider their prediction adequate but also whether their prediction are better than existing approaches.

It would be nice to see more results (such as fig 2 and other predictive summaries) for the real-data application, specially estimation of some model parameters: regression coefficients, variance components and shape parameters. Once again and even more so for the large scale application, comparison against results obtained for LGP-based would be useful.

Finally, time-varying regression coefficients could have been introduced following the same temporal evolution adopted for the random effects. This extension seem a natural approach for such econometric applications (see Min and Zellner, 1993).

5 Conclusion

The paper is a useful addition to the literature and I congratulate the authors on their efforts in terms of modelling, computation and application of their proposal. I thank the Editor for allowing me the opportunity to contribute to the discussion. I enjoyed reading the paper, specially the distributional properties of their proposed MGD. My comments were basically associated with my lack of knowledge of some aspects of their work. My request for additional theoretical and empirical evidence is addressed at making the material more accessible to a wider audience.

In that respect, making their software user-friendly and available for general use with some guidance on the choices they make may boost their reach towards the end users.

References

Carter, C. K. and Kohn, R. (1996). "Markov chain Monte Carlo in conditionally Gaussian state space models." *Biometrika*, 83(3): 589–601. 299

De Jong, P. and Shephard, N. (1995). "The simulation smoother for time series models." *Biometrika*, 82(2): 339–350. 299

Frühwirth-Schnatter, S. (1994). "Data augmentation and dynamic linear models." *Journal of Time Series Analysis*, 15(2): 183–202. 299

Gamerman, D. (1997). "Sampling from the posterior distribution in generalized linear mixed models." *Statistics and Computing*, 7(1): 57–68. 299

Gamerman, D. (1998). "Markov chain Monte Carlo for dynamic generalised linear models." *Biometrika*, 85(1): 215–227. 299

Gamerman, D. and Lopes, H. F. (2006). *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. CRC Press. 300

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). "Probabilistic forecasts, calibration and sharpness." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2): 243–268. 300

Gonçalves, F. B. and Gamerman, D. (2018). "Exact Bayesian inference in spatiotemporal Cox processes driven by multivariate Gaussian processes." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1): 157–175. 299

Min, C.-k. and Zellner, A. (1993). "Bayesian and non-Bayesian methods for combining models and forecasts with applications to forecasting international growth rates." *Journal of Econometrics*, 56(1–2): 89–118. 300

Rejoinder

Jonathan R. Bradley[*], Scott H. Holan[†,‡], and Christopher K. Wikle[†]

1 Introduction

We would like to thank the discussants: Stefano Castruccio (C); Dani Gamerman (G); Catherine A. Calder and Candace Berrett (CB); William Weimin Yoo (Y); Andrew Hoegh, Kenneth Flagg, and Christian Stratton (HFS); and Kevin He and Jian Kang (HK). Their time and insight has given us opportunities to both build-on and clarify the Poisson multivariate spatio-temporal mixed effects (P-MSTM) model, and has highlighted key issues to consider in future work.

Bradley et al. (2017b, BHW) provided an extension of the multivariate spatio-temporal mixed effects model (MSTM; Bradley et al., 2015) to Poisson data. Our starting point to solve this problem was to replace the Gaussian data model in Bradley et al. (2015) with a Poisson data model using the log-link. This model is explicitly written as "Model 2" and is stated in the Supplemental Appendix of BHW. The MSTM matched the correlation structure we were seeing in our exploratory analysis of the Quarterly Workforce Indicators (QWI) and was flexible enough for other practitioners to adapt to their setting. For example, in Table 1 of the Supplemental Appendix one can define the target covariance matrix with a covariance that includes marked changes; i.e., the scenario posed by C.

We were interested in predicting the mean number of people employed at the beginning of a quarter, over all 3,145 US counties, 20 industries, and 96 quarters using Model 2 and a dataset consisting of 4,089,755 QWIs. We used a Gibbs sampler with Metropolis-Hastings updates when necessary. There are many choices that one can use to tune this MCMC algorithm, and we used the Metropolis Adjusted Langevin (MALA) (Roberts and Tweedie, 1996) algorithm, adaptive proposals based on the Robbins-Monroe process (Garthwaite et al., 2010), and Log-Adaptive Proposals (LAP) (Shaby and Wells, 2011). The acceptance rates were extremely small and convergence was not obtained using any of these tuning strategies. G and HK asked why the effective sample sizes in Section 4.1 of BHW were so small for the LGP model. In our experience, the acceptance rates tend to be extremely small when fitting Model 2, which induces strong positive autocorrelation in the Markov chain, and hence, small effective sample sizes.

Motivated by convergence issues we pursued the development of a multivariate log-gamma (MLG) distribution. Specifically, in BHW we derived the MLG distribution which results in full-conditional distributions that are of the same form as a conditional MLG distribution. Importantly, there is a particular class of marginal distributions that has a density proportional to the conditional MLG, which can be easily simulated from.

[*] Corresponding author. Department of Statistics, Florida State University, 117 N. Woodward Ave., Tallahassee, FL 32306-4330, bradley@stat.fsu.edu

[†] Department of Statistics, University of Missouri, 146 Middlebush Hall, Columbia, MO 65211-6100

[‡] U.S. Census Bureau, 4600 Silver Hill Road, Washington, D.C., 20233-9100

Thus, the MLG allows for exact sampling from the full-conditional distributions (i.e., no Metropolis Hastings steps), and we were finally *able* to produce reasonable predictions that use the complex structure of the MSTM.

There were consequences of restricting ourselves to the high-dimensional setting. Namely, the LGP was extremely difficult to fit, and consequently, our comparisons to the LGP were primarily methodological (see Section 3 of BHW) and computational (see Section 4.1 of BHW) in nature. Consequently, we refer G, HK, Y, and the reader to Bradley et al. (2018a) for additional empirical comparisons between a latent MLG model and the LGP in lower-dimensional settings. Similarly, the computational motivations of BHW were driven by the difficulties in fitting an LGP using a Gibbs sampler with Metropolis updates. Thus, in Section 2, we provide an additional discussion on the computational performance of the MLG/P-MSTM beyond its comparison to the LGP.

We are encouraged and excited by the discussants exploration into the properties of the MLG distribution and on promising extensions of the MLG/P-MSTM. For example, HFS provided the beginnings of a zero-inflated Poisson (ZIP) extension of the P-MSTM, which is especially prudent in our setting since we know that small counts with high spatial dependence can create difficulty for the P-MSTM (De Oliveira, 2013; Hoegh et al., 2016). A majority of the ideas posed by the discussants involved defining a spatial random process, and thus, we focus more on issues related to these extensions. Thus, in Section 3 we add some discussion surrounding a process definition of the MLG. Then, in Section 4 we discuss the role of the shape and rate parameters in determining the properties of the MLG distribution.

2 Computational Considerations

We start this section with a question posed by CB: Is the P-MSTM an approximation of Model 2, or is it an “exact model?” We can see a case for either interpretation because of the Taylor series argument (i.e., Proposition 2) that provides a relationship between the multivariate normal distribution and the MLG. However, from our point-of-view, the MLG distribution is an “exact model,” and Proposition 2 provides an argument that the MLG distribution is a more general model than the multivariate normal distribution.

This point is important when considering many of the computational considerations brought up by Y, G, and C. That is, the MLG distribution is more than a choice that aids in Gibbs sampling, but it is also a flexible multivariate distribution. As discussed in Section 1 of BHW (and discussed by G) Gibbs sampling should not be held as an ideal because there are many other computational tools available in the literature. However, the flexibility of the MLG makes it a reasonable consideration even when one chooses other computational tools (e.g., see Gamerman, 1997; Lindgren et al., 2011; Neal, 2011; Giordano et al., 2013, among others). This is worth emphasizing because many of these computational tools are preferable to Gibbs sampling. For example, HK was curious on whether HMC provides faster convergence than Gibbs sampling. We suspect that it does, however, in high-dimensional settings it is difficult to implement HMC. Also, as G discusses, updating each random effect $\boldsymbol{\eta}_t$ separately over time t, may lead to problems with mixing of the MCMC. Thus, a joint update of all $\{\boldsymbol{\eta}_t : t = 1, \ldots, T\}$ would be preferable when T is large.

We are excited about the potential variational Bayes extension of the P-MSTM. A variational Bayes implementation of the P-MSTM with C's suggestion of parallelization has the potential to allow one to analyze much higher-dimensional multivariate spatio-temporal count data than what is presented in BHW. Y's discussion on the relationship between our Gibbs sampler and variational Bayes is intriguing. To better understand this relationship, one would need to develop the variational Bayes algorithm for the multivariate version of the P-MSTM. This is subject of current research.

3 A Spatial Random Process Definition of the MLG

A clear majority of the extensions proposed by the discussants involved a random process definition of the MLG. For example, G and C inquired about using a MLG distribution with a stationary covariance. G asked a question about using the MLG as a prior for Poisson point patterns. C also discussed a continuous propagator version of the MSTM, our need for an embedded lattice, and modeling gradients similar to Quick et al. (2013). However, whenever a new multivariate distribution is proposed there is a certain property that should be checked before this distribution is used for processes (say, $q(\mathbf{s})$ for $\mathbf{s} \in D \subset \mathbb{R}^d$). The property we are referring to is Kolmogorov Consistency, which ensures that $\{q(\mathbf{s}_1), \ldots, q(\mathbf{s}_K)\}$ has a "well defined" probability measure for *any* collection of locations $\{\mathbf{s}_1, \ldots, \mathbf{s}_K\} \subset D$ (Kolmogorov, 1933). The spatial domain in BHW is defined to be a finite lattice, and hence our response is a random vector and not a random process defined on a possibly uncountably infinite spatial domain.

To assess Kolmogorov Consistency in this setting a careful understanding of Theorem 2 in BHW is needed. To aid in this effort, we find it necessary to give some clarification surrounding Theorem 2, and in Section 5 of this Rejoinder we give a re-statement of Theorem 2 to avoid any potential confusion. Specifically, the marginal distribution of $\mathbf{q}_1$ not only has a specific form of $\mathbf{V}$, but also is a *limiting* case of an unnormalized MLG distribution. Specifically, let ρ be an unnormalized MLG distribution with mean zero and covariance parameter $\mathbf{V}^{-1} = [\mathbf{H}, \frac{1}{\sigma_2}\mathbf{Q}_2]$, where $\mathbf{Q}_2$ is the basis for the null space of $\mathbf{H}$. Then,

$$\begin{aligned}\lim_{\sigma_2 \to \infty} \rho(\mathbf{q}_1, \mathbf{q}_2 | \mathbf{c} = \mathbf{0}_m, \mathbf{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa}) &= \exp\{\boldsymbol{\alpha}'\mathbf{H}\mathbf{q}_1 - \boldsymbol{\kappa}'\exp(\mathbf{H}\mathbf{q}_1)\} \\ &= f(\mathbf{q}_1 | \mathbf{c} = \mathbf{0}_m, \mathbf{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa}) f(\mathbf{q}_2 | \mathbf{c} = \mathbf{0}_m, \mathbf{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa}),\end{aligned}$$

where

$$f(\mathbf{q}_1 | \mathbf{c} = \mathbf{0}_m, \mathbf{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa}) = \exp\{\boldsymbol{\alpha}'\mathbf{H}\mathbf{q}_1 - \boldsymbol{\kappa}'\exp(\mathbf{H}\mathbf{q}_1)\} \tag{1}$$

$$f(\mathbf{q}_2 | \mathbf{c} = \mathbf{0}_m, \mathbf{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa}) = 1. \tag{2}$$

This implies that $\mathbf{q}_1$ is independent of $\mathbf{q}_2$ as σ_2 approaches infinity, and hence, the marginal distribution of $\mathbf{q}_1 = (\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'\mathbf{w}$ (in the limit) is given by (1). Although the pdf in (1) is proper (see Proposition 2 in BHW), it is crucial that we recognize that $\mathbf{q}_1$ is extended by an improper $\mathbf{q}_2$. This *improper extension* leads to issues with Kolmogorov Consistency.

Kolmogorov Consistency requires two properties. The first is permutation invariance; that is if we change the order of the elements within $\mathbf{q}_1$, we obtain the same density. The second criteria for Kolmogorov Consistency is extension. That is, if we extend $\mathbf{q}_1$ by any vector $\mathbf{q}_2$ the marginal distribution stays the same regardless of the choice of $\mathbf{q}_2$. This may appear to hold trivially, but in many settings this is not the case. For example, Minozzo and Ferracuti (2012) show that the marginal distribution of the multivariate skew normal distribution from Kim and Mallick (2004) is different from the marginal distribution based on the direct transformation.

To check for Kolmogorov Consistency, we consider a similar argument to Minozzo and Ferracuti (2012). Consider the special case where $\mathbf{H} = (1,1)'$. The transformation $(\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'\mathbf{w} = \frac{1}{2}\mathrm{w}_1 + \frac{1}{2}\mathrm{w}_2$. Denote the independent gamma random variables, $\exp(w_i) = \gamma_i$ for $i = 1, 2$. Then $(\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'\mathbf{w} = \sqrt{\gamma_1\gamma_2}$, which has a K-distribution (e.g., see Jakeman and Pusey, 1978, among others). Now, extend q_1 by our q_2 according to (1) and (2). Letting $\boldsymbol{\alpha} = \alpha(1,1)'$, and $\boldsymbol{\kappa} = \kappa(1,1)'$. We have from Theorem 2(*ii*) that the marginal distribution is a log-gamma distribution with shape and rate equal to 2. Transforming to the exponential scale, we have that $\exp(q_1)$ is gamma with shape and rate parameters equal to 2. Now, since the gamma distribution differs from the K-distribution, we do not have extension, and hence, we do not have Kolmogorov Consistency in the setting of a fixed shape and scale parameter. [1]

Note that Kolmogorov Consistency holds under *proper extensions* of $\mathbf{q}_1$ (see Bradley et al., 2018a, for a proof). This point is especially important when one considers placing a prior distribution on the rate parameter. To see this, let g be used to denote proper densities, $\mathbf{V}^{-1} = [\mathbf{H}, \mathbf{B}]$ be invertible, $g(\boldsymbol{\kappa})$ be the density for $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_M)'$, and let $\boldsymbol{\kappa}_1 = (\kappa_{11}, \ldots, \kappa_{M1})'$ be independent and identically distributed as $\boldsymbol{\kappa}$. For a given $\mathbf{q}_3 \in \mathbb{R}^m$,

$$\begin{aligned}
&g(\mathbf{q}_1|\mathbf{q}_2 = \mathbf{0}_m, \boldsymbol{\mu} = \mathbf{0}_m, \mathbf{V}, \boldsymbol{\alpha})\\
&\quad\propto \int \frac{g(\boldsymbol{\kappa}_1)\left(\prod_{i=1}^m \kappa_i^{\alpha_i}\right)\exp\left\{-\mathbf{1}_m'\mathbf{B}\mathbf{q}_3 + \boldsymbol{\alpha}'\mathbf{H}\mathbf{q}_1 - \boldsymbol{\kappa}'\exp\left(\mathbf{H}\mathbf{q}_1\right)\right\}}{\left(\prod_{i=1}^m \kappa_{i1}^{\alpha_i}\right)M(\mathbf{H}, \boldsymbol{\alpha}, \boldsymbol{\kappa}_1)g_\kappa(\boldsymbol{\kappa} = \exp(\mathbf{B}\mathbf{q}_3 + \log(\boldsymbol{\kappa}_1)))} g(\boldsymbol{\kappa})d\boldsymbol{\kappa},
\end{aligned}$$

where $M(\mathbf{H}, \boldsymbol{\alpha}, \boldsymbol{\kappa}_1)$ is the marginalizing constant of a conditional MLG with parameters $\mathbf{H}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\kappa}_1$ and the integrand is proportional (as a function of $\mathbf{q}_1$) to a conditional MLG with parameters $\mathbf{H}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\kappa}$. Now, consider the change of variables $\boldsymbol{\kappa} = \exp(\mathbf{B}\mathbf{q}_3 + \log(\boldsymbol{\kappa}_1))$. The Jacobian for a given $\mathbf{q}_3$ is given by $\exp(-\mathbf{1}_m'\mathbf{B}\mathbf{q}_3)$. Thus,

$$\begin{aligned}
&= \int \frac{1}{M(\mathbf{H}, \boldsymbol{\alpha}, \boldsymbol{\kappa}_1)}\exp\left\{\boldsymbol{\alpha}'\mathbf{V}^{-1}\mathbf{q} - \boldsymbol{\kappa}_1'\exp\left(\mathbf{V}^{-1}\mathbf{q}\right)\right\} g(\boldsymbol{\kappa}_1)d\boldsymbol{\kappa}_1\\
&\propto g(\mathbf{q}_1|\mathbf{q}_2 = \mathbf{q}_3, \boldsymbol{\mu} = \mathbf{0}_m, \mathbf{V}, \boldsymbol{\alpha}).
\end{aligned}$$

Hence, for every $\mathbf{q}_3$, $g(\mathbf{q}_1|\mathbf{q}_2 = \mathbf{0}_m, \boldsymbol{\mu} = \mathbf{0}_m, \mathbf{V}, \boldsymbol{\alpha}) \propto g(\mathbf{q}_1|\mathbf{q}_2 = \mathbf{q}_3, \boldsymbol{\mu} = \mathbf{0}_m, \mathbf{V}, \boldsymbol{\alpha})$, where notice that we marginalize across $\boldsymbol{\kappa}$. Since $g(\mathbf{q}_1|\mathbf{q}_2 = \mathbf{0}_m, \boldsymbol{\mu} = \mathbf{0}_m, \mathbf{V}, \boldsymbol{\alpha})$ is not a function of $\mathbf{q}_3$,

$$g(\mathbf{q}_1|\mathbf{q}_2 = \mathbf{0}_m, \boldsymbol{\mu} = \mathbf{0}_m, \mathbf{V}, \boldsymbol{\alpha})$$

[1] We would like to thank Antonio Linero at Florida State University for alerting us to this counter-example to Kolmogorov Consistency.

$$\begin{aligned}
&= E[g(\mathbf{q}_1|\mathbf{q}_2 = \mathbf{0}_m, \boldsymbol{\mu} = \mathbf{0}_m, \mathbf{V}, \boldsymbol{\alpha})] \\
&= E[g(\mathbf{q}_1|\mathbf{q}_2, \boldsymbol{\mu} = \mathbf{0}_m, \mathbf{V}, \boldsymbol{\alpha})] \\
&= g(\mathbf{q}_1|\boldsymbol{\mu} = \mathbf{0}_m, \mathbf{V}, \boldsymbol{\alpha}), \qquad (3)
\end{aligned}$$

where the expectation is with respect to the joint $\mathbf{q}_3$, and $\mathbf{q}_3$ is assumed to follow $g(\mathbf{q}_3|\boldsymbol{\mu} = \mathbf{0}_m, \mathbf{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa})$.

Thus, after marginalizing across the rate parameter, the cMLG is equal in distribution to $\mathbf{q}_1|\boldsymbol{\mu} = \mathbf{0}_m, \mathbf{V}, \boldsymbol{\alpha}$. A composite sampling approach can be used to simulate from a cMLG distribution after marginalizing out $\boldsymbol{\kappa}$. That is, first simulate $\boldsymbol{\kappa}$ and then simulate from $\mathbf{q}_1|\boldsymbol{\mu} = \mathbf{0}_m, \mathbf{V}^{-1} = [\mathbf{H}, \mathbf{Q}_2], \boldsymbol{\alpha}, \boldsymbol{\kappa}$ using the transformation $(\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'\mathbf{w}$, where $\mathbf{w} \sim \text{MLG}(\mathbf{0}_m, \mathbf{I}_m, \boldsymbol{\alpha}, \boldsymbol{\kappa})$. This result is possible since $\mathbf{B}\mathbf{q}_2$ is confounded with $\log(\boldsymbol{\kappa})$ (as similarly noted by G). This small technical result, leads one to simulate $\mathbf{q}_1$ in the same way as in BHW (since we place a prior on $\boldsymbol{\kappa}$ in BHW), but does not require an improper extension of $\mathbf{q}_1$. Thus, the result on Kolmogorov Consistency in Bradley et al. (2018a) and (3), suggests that our implementation satisfies Kolmogorov Consistency provided that $\boldsymbol{\kappa}$ has a prior distribution and is marginalized. That is, one can use our implementation (i.e., the transformation $(\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'\mathbf{w}$), and develop process versions of the MLG, provided that $\boldsymbol{\kappa}$ is marginalized.

4 The Shape and Rate Parameters of the MLG

CB provided an interesting discussion on whether or not to include uncorrelated random effects in a Poisson spatial GLM. In their simulation study they found an example where the spatial model with uncorrelated LG random effects had different autocorrelations from the spatial model with uncorrelated normal random effects. We suspect that if you use HK's parameterization that you would obtain similar autocorrelations between the two models (a realization also suggested by CB). However, CB's simulation results showed that the role of the shape and rate parameters needs to be developed before implementing a spatial-only special case of the P-MSTM. Section 2 of this rejoinder is especially pertinent to CB's exploration into the use of uncorrelated random effects within the P-MSTM, since the uncorrelated random effects ($\boldsymbol{\xi}_t$) are confounded with $\log(\boldsymbol{\kappa}_t)$. Since we can not separate $\boldsymbol{\xi}_t$ and $\log(\boldsymbol{\kappa}_t)$, we implicitly have two sources of uncorrelated random effects. This might explain why CB's results show extra variability at spatial lags near zero. The consequence of this confounding issue provides additional motivation for treating the rate parameter as a nuisance by marginalizing it.

In Section 3 of BHW, we offered some guidance on *when* it matters that the presence of the shape parameter makes the MLG more flexible than the multivariate normal (also see questions posed by G and HK). Specifically, when the relative overdispersion is small then any multivariate spatio-temporal correlation function is severely restricted. Both CB and G commented on the choice of the discrete uniform prior for the shape parameters. In general, the shape parameters appear weakly identified and can be difficult to update. In Bradley et al. (2018a), we have found that the Diaconis and Ylvishaker prior (Diaconis and Ylvisaker, 1979) is less restrictive than the discrete uniform distribution, and can perform well in practice.

5 A Re-Statement of Theorem 2

Re-Statement of Theorem 2: Partition the m-dimensional random vector so that $\boldsymbol{q} = (\boldsymbol{q}_1', \boldsymbol{q}_2')'$*, where* $\boldsymbol{q}_1$ *is g-dimensional and* $\boldsymbol{q}_2$ *is* $(m-g)$*-dimensional. Define the following matrix:*

$$\boldsymbol{V}^{-1} = \left[\begin{array}{cc} \boldsymbol{Q}_1 & \boldsymbol{Q}_2 \end{array}\right] \left[\begin{array}{cc} \boldsymbol{R}_1 & \mathbf{0}_{g,m-g} \\ \mathbf{0}_{m-g,g} & \frac{1}{\sigma_2}\boldsymbol{I}_{m-g}, \end{array}\right], \tag{4}$$

where in general $\mathbf{0}_{k,b}$ *is a* $k \times b$ *matrix of zeros;* $\boldsymbol{I}_{m-g}$ *is a* $(m-g) \times (m-g)$ *identity matrix;*

$$\boldsymbol{H} = \left[\begin{array}{cc} \boldsymbol{Q}_1 & \boldsymbol{Q}_2 \end{array}\right] \left[\begin{array}{c} \boldsymbol{R}_1 \\ \mathbf{0}_{m-g,g}, \end{array}\right]$$

is the QR decomposition of the $m \times g$ *matrix* $\boldsymbol{H}$*; the* $m \times g$ *matrix* $\boldsymbol{Q}_1$ *satisfies* $\boldsymbol{Q}_1'\boldsymbol{Q}_1 = \boldsymbol{I}_g$*, the* $m \times (m-g)$ *matrix* $\boldsymbol{Q}_2$ *satisfies* $\boldsymbol{Q}_2'\boldsymbol{Q}_2 = \boldsymbol{I}_{m-g}$*, and* $\boldsymbol{Q}_2'\boldsymbol{Q}_1 = \mathbf{0}_{m-g,g}$*;* $\boldsymbol{R}_1$ *is a* $g \times g$ *upper triangular matrix; and* $\sigma_2 > 0$*. Let,* $\boldsymbol{q}$ *be distributed as*

$$\rho(\boldsymbol{q}_1, \boldsymbol{q}_2 | \boldsymbol{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa}) = \exp\left\{\boldsymbol{\alpha}'\boldsymbol{H}\boldsymbol{q}_1 + \frac{1}{\sigma_2}\boldsymbol{\alpha}'\boldsymbol{Q}_2\boldsymbol{q}_2 - \boldsymbol{\kappa}'\exp\left(\boldsymbol{H}\boldsymbol{q}_1 + \frac{1}{\sigma_2}\boldsymbol{Q}_2\boldsymbol{q}_2\right)\right\} \tag{5}$$

which is the unnormalized $\mathrm{MLG}(\mathbf{0}_m, \boldsymbol{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa})$*. Then, the following statements hold.*

(i) The marginal distribution of $\boldsymbol{q}_1$ *from* $\lim_{\sigma_2 \to \infty} \rho(\boldsymbol{q}_1, \boldsymbol{q}_2, \boldsymbol{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa})$ *is given by*

$$f(\boldsymbol{q}_1 | \boldsymbol{H}, \boldsymbol{\alpha}, \boldsymbol{\kappa}) = \exp\left\{\boldsymbol{\alpha}'\boldsymbol{H}\boldsymbol{q}_1 - \boldsymbol{\kappa}'\exp(\boldsymbol{H}\boldsymbol{q}_1)\right\}, \tag{6}$$

which has a normalizing constant,

$$\frac{1}{\int \exp\left\{\boldsymbol{\alpha}'\boldsymbol{H}\boldsymbol{q}_1 - \boldsymbol{\kappa}'\exp(\boldsymbol{H}\boldsymbol{q}_1)\right\} d\boldsymbol{q}_1}.$$

(ii) The g-dimensional random vector $\boldsymbol{q}_1$ *obtained from Theorem 2(i) is equal in distribution to* $(\boldsymbol{H}'\boldsymbol{H})^{-1}\boldsymbol{H}'\boldsymbol{w}$*, where the m-dimensional random vector* $\boldsymbol{w} \sim \mathrm{MLG}(\mathbf{0}_m, \boldsymbol{I}_m, \boldsymbol{\alpha}, \boldsymbol{\kappa})$*.*

Proof of Theorem 2 Much of this proof is the same as the proof stated in Bradley et al. (2017c). We simply add statements to clarify steps of this proof.

Theorem 2(ii): Notice that

$$\mathbf{V} = \left[\begin{array}{c} (\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}' \\ \sigma_2\mathbf{Q}_2' \end{array}\right].$$

From (2.5) of the main text we see that

$$\left[\begin{array}{c} \mathbf{q}_1 \\ \mathbf{q}_2 \end{array}\right] = \left[\begin{array}{c} (\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'\mathbf{w} \\ \sigma_2\mathbf{Q}_2'\mathbf{w} \end{array}\right], \tag{7}$$

where the m-dimensional random vector $\mathbf{w} \sim \text{MLG}(\mathbf{0}_m, \mathbf{I}_m, \boldsymbol{\alpha}, \boldsymbol{\kappa})$. Multiplying both sides of (7) by $[\mathbf{I}_g, \mathbf{0}_{g,m-g}]$ we have

$$\mathbf{q}_1 = (\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'\mathbf{w}.$$

This is true regardless of the value of σ_2.

Theorem 2(i): We have

$$\lim_{\sigma_2 \to \infty} \rho(\mathbf{q}_1, \mathbf{q}_2 | \mathbf{c} = \mathbf{0}_m, \mathbf{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa}_{1.2}) = \exp\left\{\boldsymbol{\alpha}'\mathbf{H}\mathbf{q}_1 - \boldsymbol{\kappa}'\exp(\mathbf{H}\mathbf{q}_1)\right\}, \tag{8}$$

where from Proposition 1, is proper. Let M_1 be the normalizing constant defined in the restatement of Theorem 2(ii). Notice that the limit in (8) does not depend on $\mathbf{q}_2$. In general, for two random variables q and X, if $f(q, X) = f(q)f(X)$ then q is independent of X and the marginal distribution of q is $f(q)$. Furthermore, it follows from (8) that, in the limit, the marginal distribution for $\mathbf{q}_2$ is $f(\mathbf{q}_2 | \mathbf{c} = \mathbf{0}_m, \mathbf{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa}) = 1$. This completes the result. However, we show that $\mathbf{q}_2$ can be marginalized.

The proof of Theorem 2 in Bradley et al. (2017c) suppressed limits and marginalization constants for simplicity. To avoid any potential confusion we include these terms. This implies that

$$\begin{aligned}
f(\mathbf{q}_1 | \mathbf{c} = \mathbf{0}_m, \mathbf{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa}) &= \lim_{\sigma_2 \to \infty} \frac{1}{M_1 \int 1 d\mathbf{q}_2} \int \lim_{\sigma_2 \to \infty} \rho(\mathbf{q}_1, \mathbf{q}_2 | \mathbf{c} = \mathbf{0}_m, \mathbf{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa}) d\mathbf{q}_2 \\
&= \frac{1}{M_1 \int 1 d\mathbf{q}_2} \int f(\mathbf{q}_1 | \mathbf{H}, \boldsymbol{\alpha}, \boldsymbol{\kappa}) d\mathbf{q}_2 \\
&= \frac{1}{M_1 \int 1 d\mathbf{q}_2} f(\mathbf{q}_1 | \mathbf{H}, \boldsymbol{\alpha}, \boldsymbol{\kappa}) \int 1 d\mathbf{q}_2 \\
&= \frac{1}{M_1} f(\mathbf{q}_1 | \mathbf{H}, \boldsymbol{\alpha}, \boldsymbol{\kappa}),
\end{aligned}$$

which is the desired result.

6 Discussion

We would like to express our appreciation to the discussants, two anonymous reviewers, the anonymous associate editor, and Bruno Sansó for their time and input on BHW. This has been an extremely productive discussion. In particular, the comments from the discussants has led us to several motivations for marginalizing $\boldsymbol{\kappa}$. We are excited and humbled by the discussants ideas for future work in this area, and we currently are in the process of producing public use code to help promote extensions of BHW.

References

Bradley, J., Holan, S., and Wikle, C. (2018a). "Bayesian Hierarchical Models with Conjugate Full-Conditional Distributions for Dependent Data from the Natural Exponential Family." *arXiv preprint: 1701.07506*. 303, 305, 306

Bradley, J. R., Holan, S. H., and Wikle, C. K. (2015). "Multivariate Spatio- Temporal Models for High-Dimensional Areal Data with Application to Longitudinal Employer-Household Dynamics." *The Annals of Applied Statistics*, 9: 1761–1791. 302

Bradley, J. R., Holan, S. H., and Wikle, C. K. (2017b). "Computationally Efficient Multivariate Spatio-Temporal Models for High-Dimensional Count-Valued Data." *Bayesian Analysis*, DOI: 10.1214/17-BA1069: 1–33. 302

Bradley, J. R., Holan, S. H., and Wikle, C. K. (2017c). "Supplemental Materials: Computationally Efficient Multivariate Spatio-Temporal Models for High-Dimensional Count-Valued Data." *Bayesian Analysis*, doi: http://dx.doi.org/10.1214/17-BA1069SUP. 307, 308

De Oliveira, V. (2013). "Hierarchical Poisson models for spatial count data." *Journal of Multivariate Analysis*, 122: 393–408. MR3189330. doi: `https://doi.org/10.1016/j.jmva.2013.08.015`. 303

Diaconis, P. and Ylvisaker, D. (1979). "Conjugate priors for exponential families." *The Annals of Statistics*, 17: 269–281. 306

Gamerman, D. (1997). "Sampling from the posterior distribution in generalized linear mixed models. Statistics and Computing." *Statistics and Computing*, 7: 57–68. 303

Garthwaite, P., Fan, Y., and Sisson, S. (2010). "Adaptive optimal scaling of Metropolis-Hastings algorithms using the Robbins-Monro process." *arXiv preprint: 1006.3690*. 302

Giordano, R., Broderick, T., and Jordan, M. I. (2013). "Covariances, Robustness, and Variational Bayes." *arXiv preprint: 1709.02536*. 303

Hoegh, A., Ferreira, M. A., and Leman, S. (2016). "Spatiotemporal model fusion: multiscale modelling of civil unrest." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 65: 529–545. 303

Jakeman, E. and Pusey, P. N. (1978). "Significance of K distributions in scattering experiments." *Physical Review Letters*, 40(9): 546. 305

Kim, H.-M. and Mallick, B. K. (2004). "A Bayesian prediction using the skew Gaussian distribution." *Journal of Statistical Planning and Inference*, 120(1): 85–101. 305

Kolmogorov, A. N. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Berlin: Springer. 304

Lindgren, F., Rue, H., and Lindström, J. (2011). "An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach." *Journal of the Royal Statistical Society, Series B*, 73: 423–498. 303

Minozzo, M. and Ferracuti, L. (2012). "On the existence of some skew-normal stationary processes." *Chilean Journal of Statistics (ChJS)*, 3(2). 305

Neal, R. M. (2011). "MCMC Using Hamiltonian Dynamics." In Brooks, S., Gelman, A., Jones, G. L., and Meng, X. (eds.), *Handbook of Markov Chain Monte Carlo*, 113–160. Chapman and Hall. 303

Quick, H., Banerjee, S., and Carlin, B. P. (2013). "Modeling Temporal Gradients in Regionally Aggregated California Asthma Hospitalization Data." *Annals of Applied Statistics*, 7: 154–176. MR3086414. doi: https://doi.org/10.1214/12-AOAS600. 304

Roberts, G. and Tweedie, R. (1996). "Exponential convergence of Langevin distributions and their discrete approximations." *Bernoulli*, 2: 341–363. 302

Shaby, B. and Wells, M. (2011). "Exploring an adaptive Metropolis algorithm." In *Technical Report*. Department of Statistics: Duke University. 302

Acknowledgments

We would like to express our gratitude to the editor-in-chief Bruno Sansó, the editor, the associate editor, the referees, and the discussants for their very helpful comments. 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 paper 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 necessarily those of the NSF or the U.S. Census Bureau.