Bayesian Analysis (2015) **TBA**, Number TBA, pp. 1–26

Restricted Covariance Priors with Applications in Spatial Statistics

Theresa R. Smith*, Jon Wakefield†, and Adrian Dobra‡

Abstract. We present a Bayesian model for area-level count data that uses Gaussian random effects with a novel type of G-Wishart prior on the inverse variance–covariance matrix. Specifically, we introduce a new distribution called the truncated G-Wishart distribution that has support over precision matrices that lead to positive associations between the random effects of neighboring regions while preserving conditional independence of non-neighboring regions. We describe Markov chain Monte Carlo sampling algorithms for the truncated G-Wishart prior in a disease mapping context and compare our results to Bayesian hierarchical models based on intrinsic autoregression priors. A simulation study illustrates that using the truncated G-Wishart prior improves over the intrinsic autoregressive priors when there are discontinuities in the disease risk surface. The new model is applied to an analysis of cancer incidence data in Washington State.

Keywords: G-Wishart distribution, Markov chain Monte Carlo (MCMC), spatial statistics, disease mapping.

1 Introduction

Spatial data arise when outcomes and predictors of interest are observed at particular points or regions inside a defined study area. Spatial data sets are common in many fields including environmental science, economics, and epidemiology. In epidemiology, understanding the underlying spatial patterns of a disease is an important starting point for further investigations. The risk of disease inherently varies in space because the risk factors are non-uniformly distributed in space. Such risk factors may include lifestyle variables such as alcohol and tobacco use or exposure levels of environmental causes of disease such as air pollution or UV radiation. We expect that these risk factors are positively correlated in space meaning that nearby areas will have similar exposure levels or underlying characteristics. That is, we assume risk factors obey Tobler's first law of geography: "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970).

In many studies, underlying disease risk factors are unknown or unmeasured. Bayesian models account for unknown or unmeasured risk factors using priors chosen to mimic their correlation structure. The most common Bayesian framework for area-level spatial data uses Gaussian random effects with a covariance structure that imposes positive

*Department of Statistics, University of Washington, Seattle, WA 98195, tsmith7@uw.edu
†Departments of Statistics and Biostatistics, University of Washington, Seattle, WA 98195, jonno@uw.edu
‡Departments of Statistics, Biobehavioral Nursing, and Health Systems and the Center for Statistics and the Social Sciences, University of Washington, Seattle, WA 98195, adobra@uw.edu

 DOI: 10.1214/14-BA927

spatial dependence between random effects of neighboring or near-by areas (Besag et al., 1991; Diggle et al., 1998; Banerjee et al., 2004). The non-Gaussian spatial clustering and Potts model based priors also impose positive dependence in the relative risks of neighboring areas (Knorr-Held and Best, 2001; Green and Richardson, 2002). More recently, several authors have developed modifications to existing models, specifically to preserve positive dependence for spatial statistics applications (Wang and Pillai, 2013; Hughes and Haran, 2013). Further, positive spatial dependence is usually imposed in geostatistical models for data observed point-wise rather than area-wise. For example, the Matérn family of marginal covariance functions for Gaussian random fields yields positive correlations between observations at two locations locations, with the magnitude of the correlation decreasing with distance (Stein, 1999; Diggle and Ribeiro, 2007).

We present a Bayesian model for area-level count data that uses Gaussian random effects with a novel type of G-Wishart prior on the inverse variance–covariance matrix. The usual G-Wishart or hyper inverse Wishart prior restricts off-diagonal elements of the precision matrix to 0 according to the edges in an undirected graph (Dawid and Lauritzen, 1993; Roverato, 2002). Dobra et al. (2011) use the G-Wishart prior to analyze mortality counts for ten cancers in the United States using a Bayesian hierarchical model incorporating Gaussian random effects with a separable covariance structure. Their comparisons show that allowing different strengths of association between paris of neighboring states can have advantages over traditional conditional autoregressive priors that assume the same strength of conditional association across the study region. However, the G-Wishart prior allows for both positive and negative conditional associations between neighboring areas.

The truncated G-Wishart distribution that we introduce only has support over precision matrices that lead to positive conditional associations. We describe Markov chain Monte Carlo (MCMC) algorithms for this new prior and construct a Bayesian hierarchical model for areal count data that uses the truncated G-Wishart prior for the precision matrix of Gaussian random effects. We show via simulation studies that risk estimates based on a model using the truncated G-Wishart prior are better than those based on conditional autoregression when the outcome is rare and the risk surface is not smooth. For univariate data, there is little information to identify the parameters of the spatial precision matrix; however, we can share information across outcomes in a multivariate model by assuming a separable covariance structure. We illustrate the improvement of using the truncated G-Wishart prior in a separable model (measured via cross-validation) using cancer incidence data from the Washington State Cancer Registry.

The structure of this paper is as follows. In Section 2, we present our modeling framework and give a brief overview of conditional autoregressive models. In Section 3, we define the truncated G-Wishart distribution and give the details of an MCMC sampler for estimating relative risks in a spatial statistics context. In Section 4, we present a simulation study based on univariate disease mapping using the geography of the counties of Washington State. Finally, in Section 5, we extend the univariate truncated G-Wishart model to multivariate disease mapping using the separable Gaussian graphical model framework of Dobra et al. (2011).

2 Background

2.1 Notation

Let $\mathcal{A} = \{A_1, \dots A_n\}$ be a set of non-overlapping geographical areas, and let $\mathbf{y} = \{y_1, \dots, y_n\}$ represent the set of counts of the observed number of health events in these areas. Possible health events include deaths from a disease, incident cases of a disease, or hospital admissions with specific symptoms of a disease. Next, let $\mathbf{E} = \{E_1, \dots, E_n\}$ be the set of expected counts and $\mathbf{X} = \{\mathbf{x}_1 \dots \mathbf{x}_n\}$ be a matrix where $\mathbf{x}_i$ is a vector of suspected risk factors measured in area i. The expected counts account for differences in known demographic risk factors. If the population in each area is stratified into J groups (e.g., gender and 5 year age-band combinations), then the expected count for each area is

$$E_i = \sum_{j=1}^{J} q_j P_{ij},$$

where P_{ij} is the population in area i in demographic group j and q_j is the rate of disease in group j. The rates q_j may be estimated from the data if the disease counts are available by strata (internal standardization) or they may be previously published estimates for the rates of disease (external standardization).

A generic Bayesian hierarchical model for data of this type is:

$$\begin{aligned} y_i \mid \mathbf{y}_{-i}, E_i, \theta_i &\sim \mathsf{Poi}(E_i\theta_i), \\ \log(\theta_i) &= \mathbf{x}_i^T \beta + u_i, \\ \pi(\mathbf{u}) &= H, \end{aligned}$$

where $\mathbf{y}_{-i}$ is the vector of counts with area i excluded and H is a probability distribution with spatial structure. Most choices of H encode the belief that the residual spatial random effects, $\mathbf{u}$, of nearby areas have similar values. This restriction follows from the interpretation of the random effects as surrogates for unmeasured risk factors, which are generally assumed to be positively correlated in space. The inclusion of H produces smoother (though biased) estimates of the vector of relative risks, $\boldsymbol{\theta}$, with reduced variability compared to the maximum likelihood estimates $\widehat{\boldsymbol{\theta}} = \mathbf{y}/\mathbf{E}$. These maximum likelihood estimates, called standardized incidence ratios (SIRs) or standardized mortality/morbidity ratios (SMRs), have large sampling variances when the expected counts are small. A key task in modeling areal count data is to choose a prior H that is flexible enough to adapt to the smoothness of the risk surface.

2.2 Existing Models for Areal Count Data

The most common choice for H is the Gaussian conditional autoregression or CAR prior (Besag, 1974; Rue and Held, 2005), which is a type of Gaussian Markov random field. The CAR model for a vector of Gaussian random variables is defined by a set

of conditional distributions. The conditional distribution for the random variable, u_i, given the other variables, $\mathbf{u}_{-i}$, is

$$u_i \mid \mathbf{u}_{-i} \sim \mathsf{N}\left(\sum_{j:j\neq i} b_{ij}u_j, \tau_i^2\right).$$

The joint distribution of the vector $\mathbf{u}$ is a mean-zero multivariate normal distribution with precision $\mathbf{D}^{-1}(\mathbf{I}-\mathbf{B})$, where $B_{ij} = b_{ij}$, $B_{ii} = 0$, and $D_{ii} = \tau_i^2$. This is a proper joint distribution if $\mathbf{D}^{-1}(\mathbf{I}-\mathbf{B})$ is a symmetric, positive definite matrix (Banerjee et al., 2004).

The *intrinsic conditional autoregression* or ICAR prior is the most commonly used prior for spatial random effects within the class of CAR priors. Under the ICAR prior, the conditional mean for a given random effect is the weighted average of the neighboring random effects, and the conditional variance is inversely proportion to the sum of these weights:

$$u_i \mid \mathbf{u}_{-i} \sim \mathsf{N}\left(\frac{1}{\omega_{i+}}\sum_{j:j\neq i}\omega_{ij}u_j, \frac{\tau_u^2}{\omega_{i+}}\right). \tag{1}$$

Here ω_{ij} is nonzero if regions i and j are neighbors (i.e., share a border) and 0 otherwise; ω_{i+} is the sum of all of the weights for a specific area. A binary specification for $\mathbf{W} = \{\omega_{ij}; i, j = 1, \ldots, n\}$ is frequently used, though other weights that incorporate the distance between areas can also be used (White and Ghosh, 2009). In the binary case, $\omega_{ij} = 1$ for neighboring regions and $\omega_{i+} = n_i$, the number of regions that border area i. Under this specification, the conditional mean for a particular random effect is the average value of the random effects for the neighboring regions, and the conditional variance is inversely proportional to the number of neighbors of the area.

Besag et al. (1991) use a CAR prior for spatial random effects in a disease mapping context in what has become known as the *convolution model*:

$$\log(\theta_i) = \mathbf{x}_i^T\beta + v_i + u_i.$$

Here v_i is a non-spatial random effect and u_i is a spatial random effect. The prior for $\mathbf{v}$ is $\mathsf{N}(0, \sigma_v^2\mathbf{I})$, and the prior for $\mathbf{u}$ is the ICAR prior.

Though popular, the convolution model has several drawbacks. First, there are only two parameters (σ_v^2 and τ_u^2) to control the level of smoothing with only one of these (τ_u^2) contributing to the spatial portion of the model. This parsimony is ideal for estimating a smooth risk surface in the presence of large sampling variably, which is a common issue for rare diseases or for small area estimation. However, using ICAR random effects can lead to over-smoothing, which masks interesting features of the risk surface, including sharp changes. Several authors have addressed this issue by incorporating flexibility in the conditional independence structure of the relative risks (Knorr-Held and Raßer, 2000; Green and Richardson, 2002; Lee and Mitchell, 2013;

Lee et al., 2014). These approaches are fairly parsimonious, but estimating the parameters requires careful reversible jump MCMC or access to data from previous years. In contrast, we develop a locally-adaptive approach with a separate parameter for the strength of spatial association between each pair of neighboring areas while preserving the conditional independence structure.

A second drawback is that the ICAR prior is improper. The joint distribution implied by the conditional specification in (1) is a singular multivariate normal distribution with precision matrix $\tau_u^2(\mathbf{D}_\omega - \mathbf{W})$, where $\mathbf{D}_\omega$ is a diagonal matrix with elements $D_{ii} = \omega_{i+}$. Since each row of $\mathbf{D}_\omega - \mathbf{W}$ sums to 0, this precision matrix does not have full rank, and the joint prior for $\mathbf{u}$ is improper. One way to alleviate both the over smoothing and the singularity issues is through the addition of a spatial autocorrelation parameter ρ:

$$u_i \mid \mathbf{u}_{-i} \sim \mathsf{N}\left(\frac{\rho}{\omega_{i+}} \sum_{j:j\neq i} \omega_{ij} u_j, \frac{\sigma_u^2}{w_{i+}}\right).$$

This specification is called the proper CAR because it gives rise to a proper joint distribution as long as ρ is between the reciprocals of the largest and smallest eigenvalues of $\mathbf{D}_\omega^{-1/2}\mathbf{W}\mathbf{D}_\omega^{-1/2}$ (Banerjee et al., 2004). For the binary specification of $\mathbf{W}$, this always includes $\rho \in [0, 1)$. The relationship between ρ and the overall level of spatial smoothing in the proper CAR prior is complex. The prior marginal correlations between the random effects of neighboring areas increase very slowly as ρ increases, with substantial correlation obtained only when ρ is very close to 1 (Besag and Kooperberg, 1995). Further, as ρ increases, the ordering of these marginal correlations is not fixed (Wall, 2004).

Nonetheless, the ICAR prior remains a popular choice for spatially correlated errors in many applied settings. The conditional specification in (1) is parsimonious, and one only needs to specify a single prior for the precision of the spatial random effects. Prior specification has received some attention in the literature (Fong et al., 2009; Sørbye and Rue, 2014). Further, off-the-shelf MCMC routines for the ICAR and convolution models are available in WinBUGS (Lunn et al., 2000) and various `R` packages. Fast computation of approximate marginal posterior summaries is available using integrated nested Laplace approximation (INLA) (Rue et al., 2009).

3 Methodology

An alternative to specifying the prior for spatial random effects based on a set of conditional distributions is to work directly with the joint distribution. A Gaussian graphical model or covariance selection model is a set of joint multivariate normal distributions that obey the pairwise conditional independence properties encoded by an undirected graph, G (Dempster, 1972; Lauritzen, 1996). This graph has two elements: the vertex set V and the edge list E. The absence of an edge between two vertices corresponds to conditional independence and implies a specific structure for the precision matrix of the joint distribution. If $\mathbf{u}$ follows a multivariate normal distribution with precision matrix $\mathbf{K}$, then $\mathbf{u}$ follows a Gaussian graphical model if $u_i \perp\!\!\!\perp u_j \mid \mathbf{u}_{V\setminus(i,j)} \iff (i,j) \notin$

$E \implies K_{ij} = 0$ for any pairs i and j. Here $\mathbf{u}_{V\setminus(i,j)}$ is the vector $\mathbf{u}$ excluding the ith and jth elements.

The conjugate prior for the precision matrix in the Gaussian setting is the Wishart distribution, which is a distribution over all symmetric, positive definite matrices of a fixed dimension. The Wishart distribution has two parameters. The first is a scaler $\delta > 2$, which controls the spread of the distribution. The second is an $n \times n$ matrix $\mathbf{D}$, which is related to the location of the distribution. For $\mathbf{K} \sim \mathsf{Wis}(\delta, D)$, $\mathsf{E}(\mathbf{K}) = (\delta + n - 1)\mathbf{D}^{-1}$ and $\text{mode}(\mathbf{K}) = (\delta - 2)\mathbf{D}^{-1}$. The G-Wishart distribution is the conjugate prior for the precision matrix in a Gaussian graphical model (Dawid and Lauritzen, 1993; Roverato, 2002). The G-Wishart distribution is a distribution over $\mathsf{P}^+(G)$, the set of all symmetric, positive definite matrices with zeros in the off-diagonal elements that correspond to missing edges in G. The density of the G-Wishart distribution for a matrix $\mathbf{K}$ is

$$\mathsf{Pr}(\mathbf{K} \mid \delta, \mathbf{D}, G) = \frac{1}{I_1(G, \delta, \mathbf{D})} |\mathbf{K}|^{(\delta-2)/2} \exp\left(-\frac{1}{2}\langle \mathbf{K}, \mathbf{D}\rangle\right) \mathbf{1}_{\mathbf{K}\in\mathsf{P}^+(G)}, \tag{2}$$

where $\langle A, B\rangle$ is the trace of $A^T B$. The normalizing constant $I_1(G, \delta, \mathbf{D})$ has a closed form when G is a decomposable graph and can be estimated for general graphs using the Monte Carlo method proposed by Atay-Kayis and Massam (2005).

3.1 Truncated G-Wishart Distribution

We propose a new G-Wishart distribution called the truncated G-Wishart distribution that imposes additional constraints on $\mathbf{K}$. This is a distribution over positive definite matrices where the off-diagonal elements that correspond to (non-missing) edges in E are less than 0. This restriction means that all pairwise conditional (or partial) correlations are positive because

$$\mathsf{cor}\left(u_i, u_j \mid \mathbf{u}_{V\setminus\{i,j\}}\right) = \frac{-K_{ij}}{\sqrt{K_{ii}K_{jj}}}.$$

This restriction is attractive in a spatial statistics context where we believe neighboring areal units are likely to be similar to each other, given the other areas.

If $\mathbf{K}$ follows a truncated G-Wishart distribution, then

$$\mathsf{Pr}(\mathbf{K} \mid G, \delta, \mathbf{D}) = \frac{1}{I_2(G, \delta, \mathbf{D})} |\mathbf{K}|^{(\delta-2)/2} \exp\left(-\frac{1}{2}\langle \mathbf{K}, \mathbf{D}\rangle\right) \mathbf{1}_{\mathbf{K}\in\mathsf{P}^+(G)\cap\mathcal{S}^0}. \tag{3}$$

Here $I_2(G, \delta, \mathbf{D})$ is the unknown normalizing constant, and $\mathcal{S}^0$ is the set of matrices with negative off-diagonal elements. The normalizing constant in (2) is finite as long as $\delta > 2$ and $\mathbf{D}^{-1} \in \mathsf{P}^+(G)$ (Atay-Kayis and Massam, 2005). The normalizing constant in (3) is finite under the same conditions because the support of the truncated G-Wishart is a subset of the support of the G-Wishart distribution. The mode of the truncated G-Wishart is again $(\delta-2)\mathbf{D}^{-1}$, and for this reason we only consider $\mathbf{D}^{-1} \in \mathsf{P}^+(G)\cap\mathcal{S}^0$. In this paper, we write TWis_G for the truncated G-Wishart distribution and Wis_G for the G-Wishart distribution.

Atay-Kayis and Massam (2005) and Dobra et al. (2011) transform $\mathbf{K}$ to the Cholesky square root, which we call $\mathbf{\Phi}$, because it is easier to handle the positive definite constraint in the transformed space. In the G-Wishart case, the elements of $\mathbf{\Phi}$ are either variation independent or are deterministic functions of other elements. We call the off-diagonal elements of $\mathbf{\Phi}$ that correspond to missing edges in the graph G "non-free." These are deterministic functions of the "free" elements: the diagonal elements and the off-diagonal elements corresponding to edges in G. If we restrict $\mathbf{K}$ to the space $\mathsf{P}^+(G) \cap \mathcal{S}^0$, we have the following constraints on the off-diagonal elements of the Cholesky square root $\mathbf{\Phi}$:

$$\Phi_{ii} > 0 \text{ for } i = 1, \ldots, n, \tag{4}$$

$$\Phi_{ij} = -\frac{1}{\Phi_{ii}} \sum_{d=1}^{i-1} \Phi_{di}\Phi_{dj} \text{ for } (i,j) \notin E, \tag{5}$$

$$\Phi_{ij} < -\frac{1}{\Phi_{ii}} \sum_{d=1}^{i-1} \Phi_{di}\Phi_{dj} \text{ for } (i,j) \in E. \tag{6}$$

The first two conditions guarantee that $\mathbf{\Phi}^T\mathbf{\Phi} \in \mathsf{P}^+(G)$. The addition of the third inequality guarantees that $\mathbf{\Phi}^T\mathbf{\Phi} \in \mathcal{S}^0$; however, this restriction comes at the cost of losing variation independence (i.e., the parameters space of $\mathbf{\Phi}$ is no longer rectangular).

3.2 Sampling from the Truncated G-Wishart Distribution

We sample from the truncated G-Wishart distribution using a random walk Metropolis–Hastings algorithm similar to the sampler proposed by Dobra et al. (2011). We sequentially perturb one free element $\Phi_{i_0j_0}$ at a time, holding the other free elements constant. In doing so, we must find the support of the conditional distribution of $\Phi_{i_0j_0}$ given the other elements. The support of this conditional distribution is the set of $\Phi_{i_0j_0}$ that satisfy inequalities (4)–(6) when the free elements, the left-hand sides of (4) and (6), are fixed.

For each specific graph and fixed pair (i_0, j_0), we can write the inequalities in (6) as

$$\Phi_{ij} < g_{ij}\left(\Phi_{i_0j_0}, \mathcal{F}_{-(i,j)}\right) \text{ for } (i,j) \in E,$$

where $\mathcal{F}_{-(i,j)}$ is the set of fixed, free elements of $\mathbf{\Phi}$ excluding Φ_{ij} and $\Phi_{i_0j_0}$. We construct g_{ij} by substituting the equalities from (5) for all of the non-free elements that depend on $\Phi_{i_0j_0}$. Each g is (at worst) a quadratic function of $\Phi_{i_0j_0}$. When g is a linear function, solving g for $\Phi_{i_0j_0}$ gives a solution set of the form $g_{ij}^{-1}(\Phi_{ij}, \mathcal{F}_{-(i,j)}) = \{\Phi_{i_0,j_0} \in (L_{ij}, \infty)\}$, where $L_{ij} < 0$. When g is quadratic, the solution set is $g_{ij}^{-1}(\Phi_{ij}, \mathcal{F}_{-(i,j)}) = \{\Phi_{i_0j_0} \in (L_{ij}, U_{ij})\}$, where L_{ij} is again negative.

If $(i,j) \prec (i_0,j_0)$ in lexicographical order, then the upper bound for Φ_{ij} cannot depend on $\Phi_{i_0j_0}$. Depending on the graphical structure, there are pairs $(i,j) \succ (i_0,j_0)$ such that the bound for Φ_{ij} does not depend on $\Phi_{i_0j_0}$. In these cases $g_{ij}^{-1}(\Phi_{ij}, \mathcal{F}_{-(i,j)}) = (-\infty, \infty)$.

Theorem 1. *The conditional distribution of a free element $\Phi_{i_0 j_0}$, $i_0 \neq j_0$ given all other free elements is a continuous distribution over an open subinterval of $\mathbb{R}^-$ given by*

$$\bigcap_{(i,j)\in E} g_{ij}^{-1}\left(\Phi_{ij}, \mathcal{F}_{-(i,j)}\right) \cap \left(-\infty, \frac{-1}{\Phi_{i_0 i_0}} \sum_{d=1}^{i_0-1} \Phi_{d i_0}\Phi_{d j_0}\right).$$

We now give the analogous theorem for free, diagonal elements:

Theorem 2. *The conditional distribution of a free element $\Phi_{i_0 i_0}$ given other free elements is a continuous distribution over a subinterval of $\mathbb{R}^+$ given by*

$$\Phi_{i_0 i_0} \in \left(\max_{i_0 < k \leq p, (i_0,k)\in E} \left\{-\frac{\sum_{d=1}^{i_0-1}\Phi_{d i_0}\Phi_{dk}}{\Phi_{i_0 k}}\right\}, \infty\right) \text{ for } 1 < i_0 < n,$$
$$\Phi_{i_0 i_0} \in (0, \infty) \text{ for } i_0 = 1, n.$$

For proofs, see the supplementary material (Smith et al., 2015).

We use these bounds to construct a Markov chain with stationary distribution equal to the truncated G-Wishart distribution. Suppose $\mathbf{\Phi}^t$ is an upper-triangular matrix at iteration t such that $(\mathbf{\Phi}^t)^T\mathbf{\Phi}^t \in \mathsf{P}^+(G) \cap \mathcal{S}^0$. For each free element in $\Phi^t_{i_0 j_0}$ do the following:

1. Calculate the upper and lower limits for $\Phi^t_{i_0 j_0}$ as described above.
2. Sample from a truncated normal with these limits, mean $\Phi^t_{i_0 j_0}$, and standard deviation σ_m.
3. Update the non-free elements in lexicographical order. These steps give a proposal $\mathbf{K}' = (\mathbf{\Phi}')^T\mathbf{\Phi}'$ where the free elements in $\mathbf{\Phi}'$ equal to the free elements of $\mathbf{\Phi}^t$ except in the (i_0, j_0) entry.
4. Accept according to the acceptance probability $\alpha = \min(1, R_m)$, where

$$\begin{aligned} R_m &= \frac{\pi(\mathbf{K}' \mid \mathbf{D}, \delta, G) q(\mathbf{K}^t \mid \mathbf{K}')}{\pi(\mathbf{K}^t \mid \mathbf{D}, \delta, G) q(\mathbf{K}' \mid \mathbf{K}^t)} \\ &= \left(\frac{\Phi'_{i_0 i_0}}{\Phi^t_{i_0 i_0}}\right)^{\delta+\nu_i(G)-1} \exp\left(-\frac{1}{2}\left\langle \mathbf{K}' - \mathbf{K}^t, \mathbf{D}\right\rangle\right) \\ &\quad \times \frac{\mathrm{TNorm}(\Phi^t_{i_0 j_0}; \Phi'_{i_0 j_0}, \sigma_m, l_{i_0 j_0}, u_{i_0 j_0})}{\mathrm{TNorm}(\Phi'_{i_0 j_0}; \Phi^t_{i_0 j_0}, \sigma_m, l_{i_0 j_0}, u_{i_0 j_0})}. \end{aligned}$$

$\mathrm{TNorm}(\cdot; \mu, \sigma, l, u)$ is the density of a normal distribution with mean μ and standard deviation σ truncated to the interval (l, u), and $\nu_i(G)$ is the number of areas that are neighbors of area i but have larger index numbers, that is, $\nu_i(G) = \#\{j : \omega_{ij} = 1 \text{ and } i < j\}$.

The speed of this sampler depends on both the number of areas and on the number of edges in the adjacency graph. These determine the number of non-zero elements in $\mathbf{K}$

and the number of non-zero elements in $\mathbf{\Phi}$. The elements of $\mathbf{K}$ can be reordered to form a banded matrix. The size of the bandwidth depends on the proportion of non-missing edges (i.e., the edge density), and the bandwidth of $\mathbf{\Phi}$ is the same as $\mathbf{K}$ (Rue and Held, 2005). Thus reordering the elements of $\mathbf{K}$ can create sparsity in $\mathbf{\Phi}$, which reduces the number of nonzero terms in (6). Figure 1 shows the time to one thousand iterations for graphs with different numbers of nodes and edges, averaging over 50 simulated networks for each size-density combination. For each simulation, we randomly sample networks with a given size and density and reorder the elements using a bandwidth-decreasing algorithm (the reverse Cuthill–McKee algorithm, available in the `spam` package). The sampler scales well for very sparse networks, but the time to 1000 iterations grows quickly when the edge density is over 20%. The edge densities of the counties in Washington State and the states in the continental US are 0.123 and 0.093, respectively.

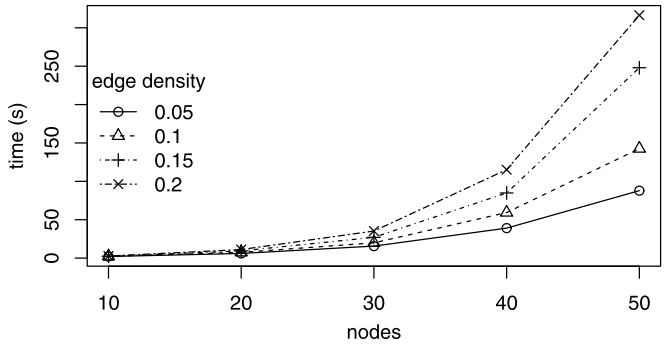

Figure 1: Time to 1000 iterations by edge density and number of nodes. For reference, the edge density of the counties in Washington State is 0.123, and the edge density of the continental US and the District of Columbia is 0.093.

3.3 Using the Truncated G-Wishart in a Hierarchical Model

We use truncated G-Wishart prior within the generic Bayesian hierarchical model for areal counts given in Section 2:

$$\begin{aligned}
\log(\theta_i) &= \mathbf{x}_i^T \beta + u_i, \\
\pi(\mathbf{u} \mid \alpha, \tau_u, \mathbf{K}) &= \mathsf{N}\left(\alpha \mathbf{1}, (\tau_u^2 \mathbf{K})^{-1}\right), \\
\pi(\alpha) &= \mathsf{N}(0, \sigma_\alpha^2), \\
\pi(\beta) &= \mathsf{N}(0, \sigma_\beta^2 \mathbf{I}), \\
\pi(\tau_u^2 \mid a, b) &= \mathsf{Gam}(a, b), \\
\pi\left(\mathbf{K} \mid G, \delta, \mathbf{D}\right) &= \mathsf{TWis}_G\left(\delta, (\delta - 2)\mathbf{D}(\rho)\right) \text{ with } K_{11} = W_{1+}, \\
\mathbf{D}^{-1}(\rho) &= \mathbf{D}_W - \rho W, \\
\pi(\rho) &= \text{Unif}(0, 0.05, 0.1, \ldots, \\
&\quad 0.8, 0.82, \ldots, 0.90, 0.91, \ldots, 0.99).
\end{aligned}$$

We suggest choosing the hyper parameters for the priors on α and τ^2 by first specifying a reasonable range for the average relative risk and then finding values of σ_α^2 and (a, b) that match this range for a fixed value of $\mathbf{K}$. For fixed $\mathbf{K}$, the distribution of $\overline{\mathbf{u}} = 1/n \sum_{i=1}^{n} u_i$ is a univariate normal distribution depending on α and τ^2. Using the adjacency matrix of Washington State as an example and letting $\mathbf{K} = \mathbf{D}^{-1}(0.99)$, 95% of the prior on $\exp(\overline{\mathbf{u}})$ is between $(1/8, 8)$ when $\sigma_\alpha^2 = 1$ and $(a, b) = (0.5, 0.0015)$. For a more informative prior, setting $\sigma_\alpha^2 = 1/4$ gives a range of $(1/2, 2)$. More details of this prior specification framework are in the supplementary material.

The prior on the spatial autocorrelation parameter ρ was introduced by Gelfand and Vounatsou (2003) for computational convenience and to reflect the fact that large values of ρ are needed to achieve non-negligible spatial dependence in the proper CAR prior. Jin et al. (2007) use a continuous uniform prior on $(0, 1)$ and a $\mathsf{Beta}(18, 2)$ prior in a similar multivariate context. For our purposes, using a discrete prior for ρ is essential for carrying out MCMC because ρ appears in the normalizing constant of the prior on $\mathbf{K}$. That is, the normalizing constant in (3) becomes $I_2(G, \delta, \mathbf{D}(\rho))$. As will be shown below, we pre calculate ratios of these normalizing constants in advance. It is not practical to repeat this process at each step of the MCMC.

We estimate the posterior distribution of the relative risks, $\boldsymbol{\theta}$, using MCMC. Most of the transitions are standard Metropolis or Gibbs updates (see supplementary material) except for the updates on the precision matrix $\mathbf{K}$ and the autocorrelation parameter ρ. We update $\mathbf{K}$ as described in Section 3.3, skipping over Φ_{11} to preserve the restriction on K_{11}. We update ρ by choosing the next smallest or largest value in $\{0, 0.05, 0.1, \ldots, 0.8, 0.82, \ldots, 0.90, 0.91, \ldots, 0.99\}$, each with probability 1/2. If ρ_t and ρ' are not on the boundary of this list, then the acceptance probability is $\alpha_\rho = \min(1, R_m)$ where

$$\begin{aligned} \log(R_m) = -1/2\mathrm{tr}\left[(\delta-2)\mathbf{K}\left\{(\mathbf{D}_w - \rho'\mathbf{W})^{-1} - (\mathbf{D}_w - \rho_t\mathbf{W})^{-1}\right\}\right] \\ + \log\left[I_2\left(G, \delta, (\delta-2)\mathbf{D}(\rho_t)\right)\right] - \log\left[I_2\left(\delta, (\delta-2)\mathbf{D}(\rho')\right)\right]. \end{aligned} \tag{7}$$

If either ρ_t or ρ' is on the boundary, there is an extra factor of 2 because the proposal is not symmetric: if $\rho_t = 0$, we propose $\rho' = 0.05$ with probability 1. Because the graph G is constant, the normalizing constants in (7) only depend on ρ. We estimate the necessary ratios of normalizing constants and store them in a table prior to running the full MCMC.

For two densities of the form $\pi_1(\eta) = c_1 q_1(\eta)$ and $\pi_2(\eta) = c_2 q_2(\eta)$ with normalizing constants c_1 and c_2, the ratio of normalizing constants is given by $r = c_1/c_2 = \mathsf{E}_2[q_1(\eta)/q_2(\eta)]$ when the support of the two distributions are the same (Chen et al., 2000). Here E_2 is the expectation under the second density. We estimate this expectation for each consecutive pair $\rho_1 > \rho_2$ using MCMC. Here we give the details for estimating the normalizing constants of a set of G-Wishart distributions without restrictions on the K_{11} element and with $\delta = 3$. However, the same process will work for the truncated G-Wishart and with the restriction that $K_{11} = W_{1+}$.

- Generate a Markov chain $\mathbf{K}_1, \mathbf{K}_2, \ldots, \mathbf{K}_S$ with stationary distribution $\mathsf{Wis}_G(3, (\mathbf{D}_w - \rho_2\mathbf{W})^{-1})$.

- For each state, let $Z_i = -1/2\text{tr}[\mathbf{K}_i((\mathbf{D}_w - \rho_1\mathbf{W})^{-1} - (\mathbf{D}_w - \rho_2\mathbf{W})^{-1})]$.
- Estimate $\log[I_1(G, 3, \mathbf{D}(\rho_1))] - \log[I_1(G, 3, \mathbf{D}(\rho_2))]$ by $\log[\frac{1}{S}\sum_{i=1}^{S}\exp(Z_i)]$.

For each pair (ρ_1, ρ_2), we average over the estimates from 10 parallel chains of 100,000 iterations. Figure 6 in the supplementary material shows the evolution of the estimates of $\log[I_1(G, 3, (\mathbf{D}_w - 0.99\mathbf{W})^{-1})] - \log[I_1(G, 3, (\mathbf{D}_w - 0.98\mathbf{W})^{-1})]$ using the adjacency graph of the counties in Washington State.

3.4 Multivariate Disease Mapping

In Section 5, we use the truncated G-Wishart prior to analyze incidence data from the Washington State Cancer Registry. In doing so, we adopt the same framework as Dobra et al. (2011) and assign a matrix normal prior with a separable covariance structure to the log relative risks. This means we assume that the covariance in the log relative risks factors into a purely spatial portion and a purely between-outcomes portion. This assumption is common for modeling two-way data including multivariate spatial data (Gelfand and Vounatsou, 2003; Carlin and Banerjee, 2003; Jin et al., 2007) and spatio-temporal data (Knorr-Held, 2000; Stein, 2005; Quick et al., 2013) as well as multi-way data (Mardia and Goodall, 1993; Fosdick and Hoff, 2014).

Here we assume that there are n areas with counts for C cancer sites (site of primary origin of the cancer) observed in each area. If $\mathbf{Y} = \{y_{ic} : i = 1, \ldots, n,\ c = 1, \ldots, C\}$ is a matrix of observed counts and $\mathbf{E} = \{E_{ic} : i = 1, \ldots, n,\ c = 1, \ldots, C\}$ is a matrix of expected counts, then we have

$$
\begin{aligned}
y_{ic}|E_{ic},\, \theta_{ic} &\sim \mathsf{Poi}\left(E_{ic}\theta_{ic}\right), \\
\log(\boldsymbol{\Theta}) &= \mathbf{U}, \\
\mathbf{U} &\sim \mathsf{MN}\left(\mathbf{M}, \mathbf{K}_C^{-1}, \mathbf{K}_R^{-1}\right), \\
M_c &\sim \mathsf{N}\left(0, \sigma_M^2\right) \text{ for } c = 1, \ldots, C, \\
\mathbf{K}_C &\sim \mathsf{Wis}\left(\delta_C, (\delta_C - 2)\mathbf{I}\right) \text{ or } \mathsf{Wis}_{G_C}\left(\delta_C, (\delta_C - 2)\mathbf{I}\right), \\
\mathbf{K}_R &\sim \mathsf{TWis}_{G_R}\left(\delta_R, (\delta_R - 2)\mathbf{D}_R^{-1}\right).
\end{aligned}
$$

We use $\mathsf{MN}(\mathbf{M}, \boldsymbol{\Sigma}_C, \boldsymbol{\Sigma}_R)$ to denote the matrix normal distribution with separable covariance structure (Dawid, 1981). That is $\text{vec}(\mathbf{U}) \mid \mathbf{M}, \boldsymbol{\Sigma}_R, \boldsymbol{\Sigma}_C \sim \mathsf{N}(\text{vec}\{\mathbf{M}\}, \boldsymbol{\Sigma}_C \otimes \boldsymbol{\Sigma}_R)$, where "$\otimes$" is the Kronecker product. In the absence of any information on cancer risk factors such as smoking rate or a socioeconomic summary measure, we only include an overall rate for each cancer in the mean model, that is, $M_{ic} = M_c$. The row covariance $\boldsymbol{\Sigma}_R$ describes the spatial covariance structure of the log relative risks. The column covariance matrix $\boldsymbol{\Sigma}_C$ describes the covariance between the cancers.

We incorporate the truncated G-Wishart distribution as the prior for the spatial precision matrix $\boldsymbol{\Sigma}_R^{-1} = \mathbf{K}_R$, and we use a G-Wishart or Wishart prior with mode equal to the identity matrix for $\boldsymbol{\Sigma}_C^{-1} = \mathbf{K}_C$. When the prior on $\mathbf{K}_C$ is a G-Wishart prior, we incorporate uncertainty in the between-cancer conditional independence graph G_C

using a uniform prior over all graphs. For both priors, we restrict $(\mathbf{K}_C)_{11} = 1$ for identifiability. Finally, we use an independent normal prior on each M_c. We estimate the relative risks under this model using an MCMC sampler identical to that in Dobra et al. (2011), substituting in the sampler from Section 3.2 for the update on $\mathbf{K}_R$.

The assumption of separability yields a more parsimonious covariance structure and can yield more stable estimation than with a full, unstructured covariance matrix. Conditioning on one precision matrix forms 'replicates' for estimating the other:

$$\text{vec}(\mathbf{U}) \sim \mathsf{N}(0, \mathbf{K}_C^{-1} \otimes \mathbf{K}_R^{-1}) \implies (\mathbf{I}_C \otimes \mathbf{\Phi}_R) \cdot \text{vec}(\mathbf{U}) \sim \mathsf{N}(0, \mathbf{K}_C^{-1} \otimes \mathbf{I}_R).$$

Thus, if $\mathbf{K}_C$ is known, then the sample size for estimating $\mathbf{K}_R$ is equal to the number of rows, and similarly, if $\mathbf{K}_R$ is known, the sample size for estimating $\mathbf{K}_C$ (and G_C) is equal to the number of columns. This factorization appears in the iterative algorithm for finding the maximum likelihood estimates of the matrix normal distribution (Dutilleul, 1999) as well as in the Gibbs sampler when using the conjugate Wishart prior with matrix or array normal data (Hoff, 2011).

4 Simulation Study

We compare the univariate disease mapping model using the truncated G-Wishart prior to three other models in a simulation study based on a similar study in Lee et al. (2014). The purpose of this simulation study is to investigate the potential of the truncated G-wishart prior in a Bayesian hierarchical model for a single realization of a disease outcome in each area. We also directly compare the G-Wishart to the standard Gaussian Markov random field formulation in a univariate context, which has not previously been done in the literature. We find that the more flexible G-Wishart priors can be advantageous when the underlying disease risk surface has sharp changes, but there are serious concerns related to estimating a large number of covariance parameters ($39 + 93 = 132$ for our example). We illustrate a more realistic example relying on the assumption of separability in Section 5.

4.1 Data Generation

We use the 39 counties in Washington State as our study region and generate expected counts based on the age-gender structure of these counties in the 2010 Census and published rates for larynx, ovarian, and lung cancer in the United Kingdom in 2008 (Cancer Research UK, 2013). These three cancers are chosen to represent a range of disease incidence from rare to common. A map of the counties with the underlying undirected graph is shown in Figure 2, and the distributions of expected counts for each cancer are shown on the log scale in Figure 3.

We generate the risk surface as the combination of a globally-smooth surface and a locally-constant surface. We label each area -1, 0 or 1 using a Potts model (Green and Richardson, 2002) so that neighboring areas are more likely to have the same label. The label allocation for this simulation study is shown in Figure 4. For each simulation, we

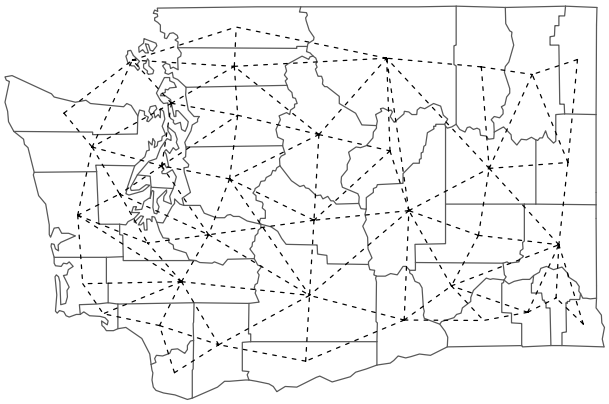

Figure 2: Washington counties and adjacency graph: 39 areas, 93 edges, 648 missing edges.

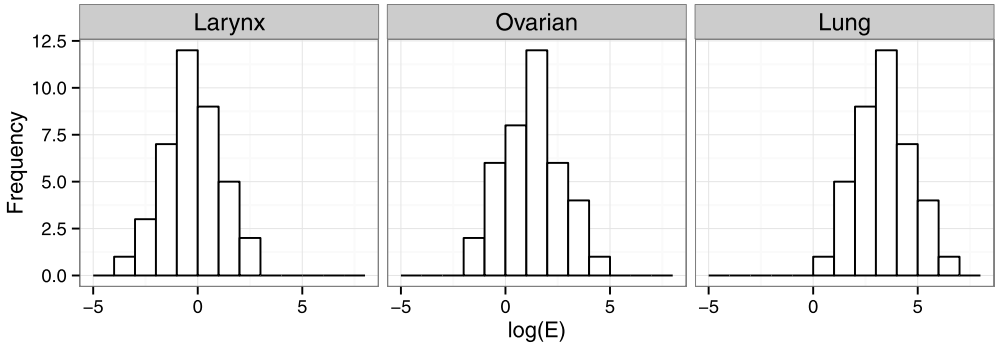

Figure 3: Distribution of the log expected counts. Expected counts are based on the 2010 population in each county and published rates for larynx, ovarian, and lung cancer in the UK. These three cancers represent a range of disease incidence from rare to common.

generate

$$y_i = \mathsf{Poi}(E_i\theta_i),$$
$$\log(\theta_i) = 0.1x_i + (M \times L_i + u_i),$$

where L_i is the label assigned to county i. We simulate x_i and u_i independently from multivariate normal distributions with Matérn covariance function with smoothness parameter 2.5 and range chosen so that the median marginal correlation is 0.5. Thus, each of the vectors $\mathbf{x}$ and $\mathbf{u}$ are realizations of a smooth spatial process observed at a finite set of points. In different simulations, we set M to 0.5, 1, or 1.5. Larger values of M lead to a risk surface with more discontinuities. We generate 50 realizations from each combination of M and the three sets of expected counts.

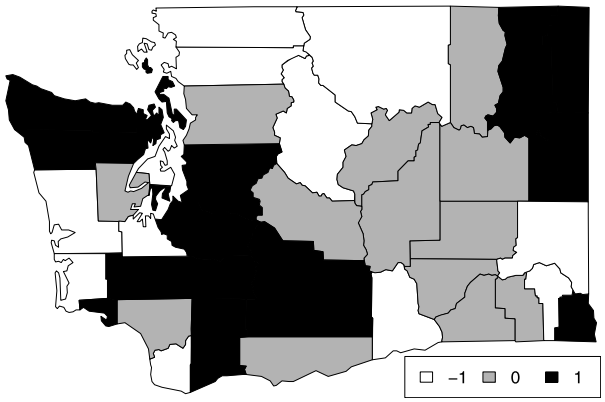

Figure 4: Labels (L_i) for simulation study.

For the simulation results described below, we run each chain for 100,000 iterations, discarding the first half as burn in. We set the prior parameters for the model in Section 3.3 to $\sigma_\alpha = 1$, $\sigma_\beta = 10$, $(a, b) = (0.5, 0.0015)$, and $\delta = 3$. Figure 3 in the supplementary material shows the evolution of the posterior mean for 10 different chains for two elements of the Cholesky square root and two random effects. In all cases, we reach convergence in about 10,000 iterations.

4.2 Results

We compare the model using the truncated G-Wishart prior to three other models. The model using the G-Wishart prior is identical to the model from Section 3.3 except that the prior on the precision matrix $\mathbf{K}$ is the G-Wishart prior instead of the truncated G-Wishart prior. We also compare against the convolution model from Section 2.2 and a similar model that includes only spatial random effects with an ICAR prior. In the convolution and ICAR models, we estimate the posterior mean and variance of the relative risks using INLA. For the models using truncated G-Wishart and G-Wishart priors, we explore the posterior distributions using MCMC.

In Figure 5, we compare the true spatial random effects $\mathbf{u}$ against the posterior estimates of the random effects for the truncated G-Wishart model and the G-Wishart models from one simulation for each set of expected counts. The estimates of the random effects are similar to the true values when the expected counts are high, but there is substantial shrinkage toward the prior mean of zero when the expected counts are small. This reflects the fact that there is much more information about the relative risks when the counts $\mathbf{y}$ are larger, and we see the same relationship in other disease mapping models.

We compare the four methods using the root-averaged mean squared error (RAMSE) of the posterior mean of each relative risk θ_i. This is the square root of the mean squared error averaged over all simulations and all areas. For S simulations and B iterations of

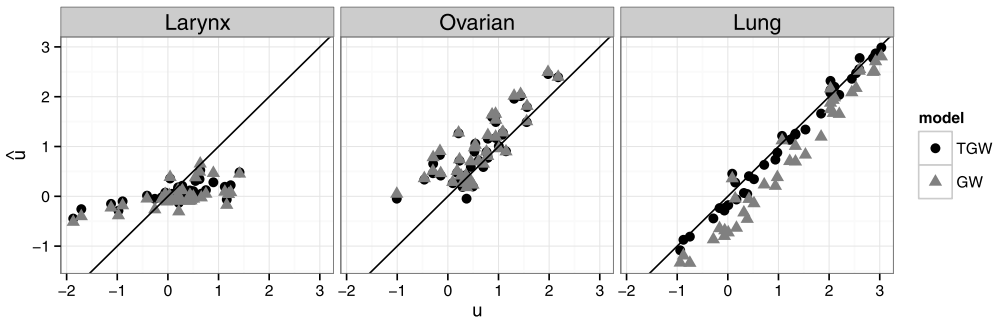

Figure 5: Simulated versus estimated spatial random effects for one simulation from each of set of expected counts with M = 0.5. The estimates are the posterior means of **u** under the truncated G-Wishart (TGW) and G-Wishart (GW) models. The posterior estimates shrink toward the prior mean of zero as the expected counts decrease.

the MCMC sampler, the RAMSE is

$$\text{RAMSE} = \sqrt{\frac{1}{39 \times S \times B} \sum_{i=1}^{39} \sum_{s=1}^{S} \sum_{b=1}^{B} (\theta_{is}^{(b)} - \theta_{is})^2},$$

where θ_{is} it the true relative risk for area i in simulation s and $\theta_{is}^{(b)}$ is the corresponding value at iteration b of the MCMC. The results of this simulation are shown in Figure 6, and the triangle indicates the lowest RAMSE within each scenario.

In general, the RAMSE decreases for all four models when the expected counts increase, and the RAMSE increases when the level of smoothing decreases (i.e., M increases). The model using the truncated G-Wishart prior performs the best in six out of nine scenarios, and we see the greatest benefit in the larynx, $M = 1.5$ simulation when the expected counts are low and the local discontinuities in the risk surface are most prominent.

While the truncated G-Wishart and G-Wishart priors for the spatial covariance appear advantageous in this simulation study, there is little information in a single sample for estimating the full covariance matrix. Figure 7 shows that the posterior distributions of the elements of the Chokesly square root are nearly identical to the prior distributions. This suggests that prior parameter choice plays a substantial role in the results from the TGW and GW models. Furthermore, the TGW and GW models should struggle when the risk surface is smoothly varying and the degree of smoothness is common across the study region. Table 2 in the supplementary material shows that the convolution model outperforms the TGW and GW models when there is no spatial association (the log relative risks are generated independently) and when the underlying risk surface is smooth (the log relative risks are generated directly from the ICAR prior). In general, the TGW and GW results are comparable with the convolution model when

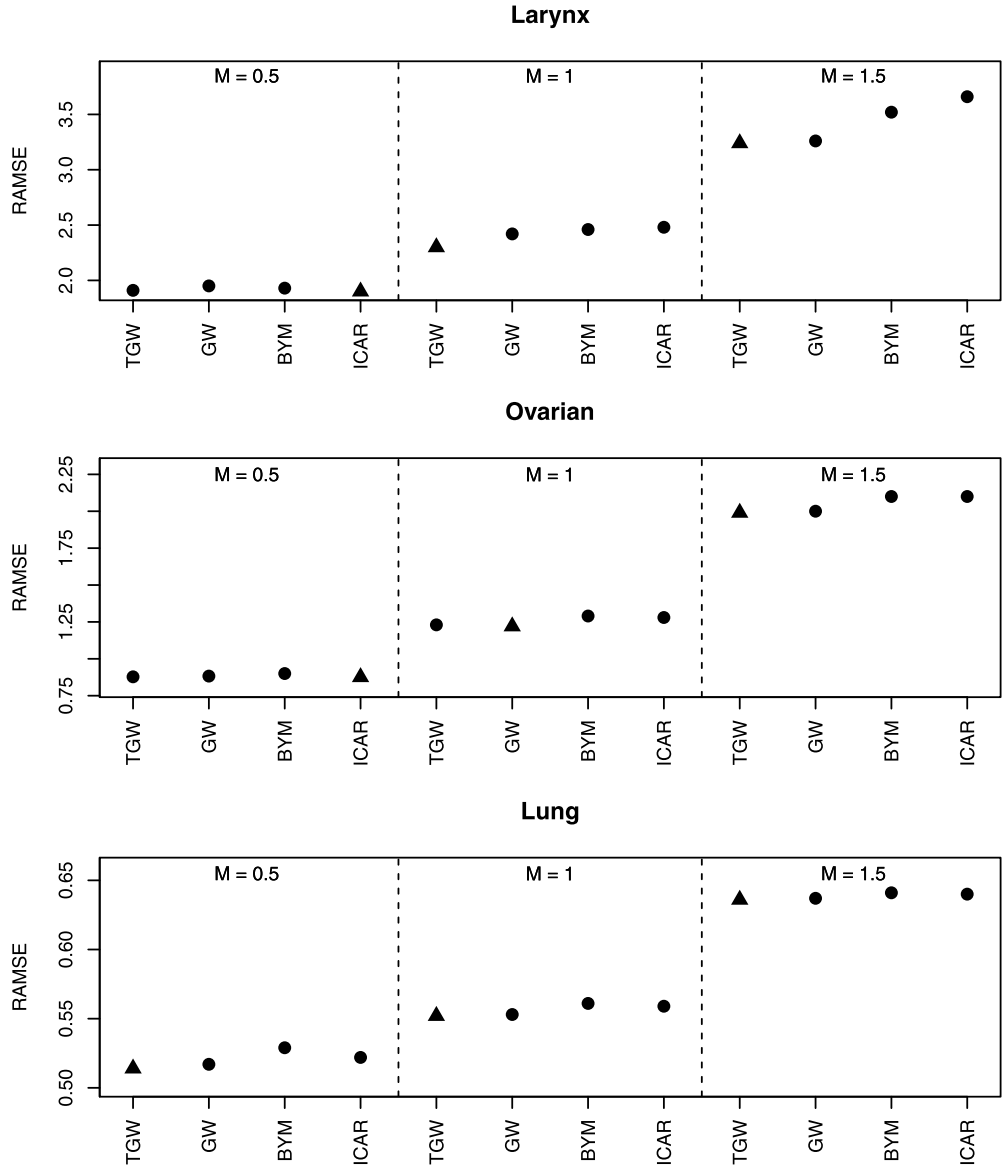

Figure 6: Root average mean squared error (RAMSE) for relative risks $\boldsymbol{\theta}$. The triangle signifies the smallest value for each experiment. The four models are: TGW, truncated G-Wishart prior on the precision matrix for the spatial random effects; GW, G-Wishart prior on the precision matrix for the spatial random effects; BYM, convolution model with independent and ICAR random effects; ICAR, only ICAR random effects. All models show increased RAMSE with increased spatial discontinuities (large M) and increased RAMSE with smaller expected counts. The TGW prior performs the best in six out of nine scenarios with the greatest benefit in the larynx, $M = 1.5$ experiment.

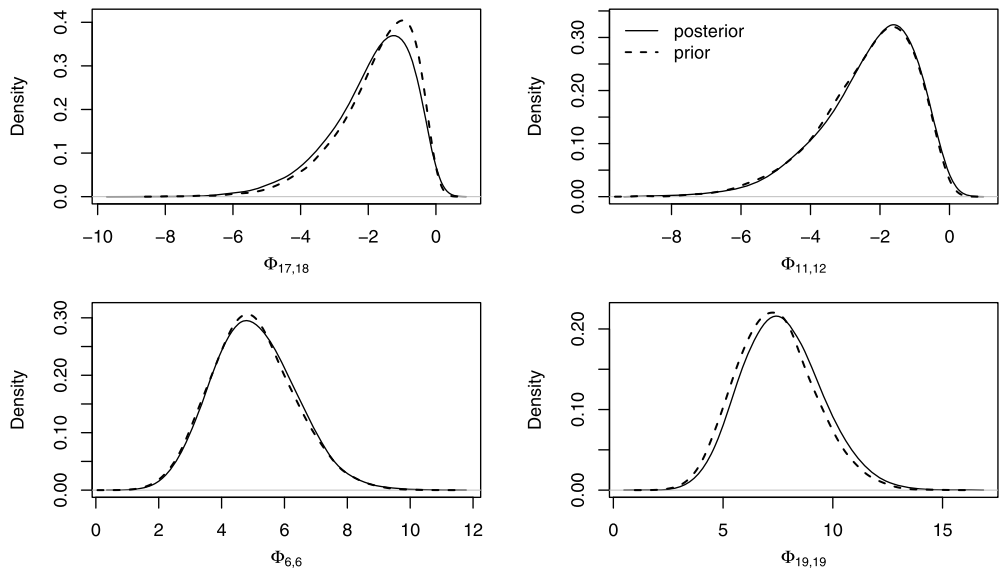

Figure 7: Comparison of the prior versus the posterior distribution of the elements of the Cholesky square root for one realization of the univariate simulation study with the TGW model.

the expected counts were larger and there was some spatial structure in the risk surface. However, with large expected counts (e.g., the lung cancer scenario), most reasonable methods will perform adequately.

5 Multiway Disease Mapping

In this section, we use the truncated G-Wishart prior in a multivariate disease mapping context using cancer incidence data from the Washington State Cancer Registry. Let $\mathbf{Y} = \{y_{ic} : i = 1, \ldots, 39,\ c = 1, \ldots, 10\}$ be a 39×10 matrix of incidence for 10 cancers in each county in Washington State in 2010. These 10 cancers have the largest incidence across the state in 2010. The expected counts E_{ic} are calculated separately for each cancer using internal standardization based on sex and 5-year age bands. The standardized incidence ratios (SIRs $= \mathbf{Y}/\mathbf{E}$) for these data are between 0 and 3.91, and the range of the empirical correlations between the SIRs of the different cancers (not taking into account spatial dependence) is $(-0.203, 0.477)$. Just over 20% of the counts are under 5, but we do not treat small counts as missing in this analysis.

We use cross-validation to compare the model in Section 3.4 to models using the G-Wishart prior (Dobra et al., 2011) and using the proper CAR form for $\mathbf{K}_R$ (Gelfand and Vounatsou, 2003). We compare 3 different choices for the prior on $\mathbf{K}_R$ and two choices for the prior on $\mathbf{K}_C$. For the truncated G-Wishart and G-Wishart priors on $\mathbf{K}_R$, we set $\delta_R = 3$ and $\mathbf{D}_R = \mathbf{D}(\rho) = (\mathbf{D}_\omega - \rho\mathbf{W})^{-1}$, where the prior on ρ is the

$\times 10^5$	BIAS^2	VAR	MSE	$\pi(\mathbf{K}_C)$	$\pi(\mathbf{K}_R)$
GGM	2.18	1.06	3.23	G-Wis	G-Wis
TGGM	**1.25**	0.73	**1.98**	G-Wis	NG-Wis
FULL	2.40	0.99	3.39	Wis	G-Wis
TFULL	1.61	**0.69**	2.29	Wis	NG-Wis
MCAR	1.31	0.82	2.13	Wis	CAR

Table 1: Ten-fold cross-validation results for the Washington State cancer incidence data. The five models use the matrix normal random effects model from Section 3.4. The priors on the precision matrices are: GGM, G-Wishart priors on $\mathbf{K}_R$ and $\mathbf{K}_C$; TGGM, truncated G-Wishart prior on $\mathbf{K}_R$ and G-Wishart prior on $\mathbf{K}_C$; FULL, G-Wishart prior on $\mathbf{K}_R$ and Wishart prior on $\mathbf{K}_C$; TFULL, truncated G-Wishart prior on $\mathbf{K}_R$ and Wishart prior on $\mathbf{K}_C$; MCAR, proper CAR prior on $\mathbf{K}_R$ and Wishart prior on $\mathbf{K}_C$. In the GGM and TGGM models, the cancer conditional independence graph G_C is random. In the other three models, G_C is a complete graph.

same as in Section 3.3. The MCAR prior on $\mathbf{K}_R$ is simply $\mathbf{K}_R = \mathbf{D}(\rho)^{-1}$. For both the Wishart and the G-Wishart priors on $\mathbf{K}_C$, we set $\delta_C = 3$ and $\mathbf{D}_C = \mathbf{I}$.

We randomly split all observations into 10 bins and create 10 data sets, each with one bin of counts held out. We impute the missing counts as part of the MCMC and compare the models based on average predictive squared bias (BIAS^2) and average predictive variance (VAR). Let $E_{\mathcal{M}}(Y_{ic})$ be the predicted value under model $\mathcal{M}$, $\text{var}_{\mathcal{M}}(Y_{ic})$ be the variance of the posterior predictive distribution, and Y_{ic} be the observed count. The comparison criteria are

$$\text{BIAS}^2_{\mathcal{M}} = \frac{1}{39 \times 10} \sum_{Y_{ic}} \left(E_{\mathcal{M}}(Y_{ic}) - Y_{ic} \right)^2,$$

$$\text{VAR}_{\mathcal{M}} = \frac{1}{39 \times 10} \sum_{Y_{ic}} \text{var}_{\mathcal{M}}(Y_{ic}).$$

The results (based on running each MCMC for 200,000 iterations) are given in Table 1. The truncated G-Wishart model with a G-Wishart prior on $\mathbf{K}_C$ performs best in terms of bias, and the truncated G-Wishart model with a Wishart prior on $\mathbf{K}_C$ performs best in terms of predictive variance. Using the truncated G-Wishart prior for the spatial precision matrix improves over the G-Wishart prior for both choices of prior for $\mathbf{K}_C$. The MCAR model is the second best model in terms of MSE (the sum of BIAS^2 and VAR).

Figure 8 shows the estimated posterior distribution of the spatial autocorrelation parameter ρ for the five models. Under the G-Wishart prior on $\mathbf{K}_R$, the posterior for ρ is much more concentrated near zero than with a truncated G-Wishart prior (regardless of the prior on $\mathbf{K}_C$). The posterior median for ρ when using CAR prior on $\mathbf{K}_R$ is between the estimates from the G-Wishart and truncated G-Wishart priors. Figure 9 shows the estimated posterior probabilities of including edges in G_C for two different priors on $\mathbf{K}_R$. The upper and lower triangles are quite similar, indicating that inference

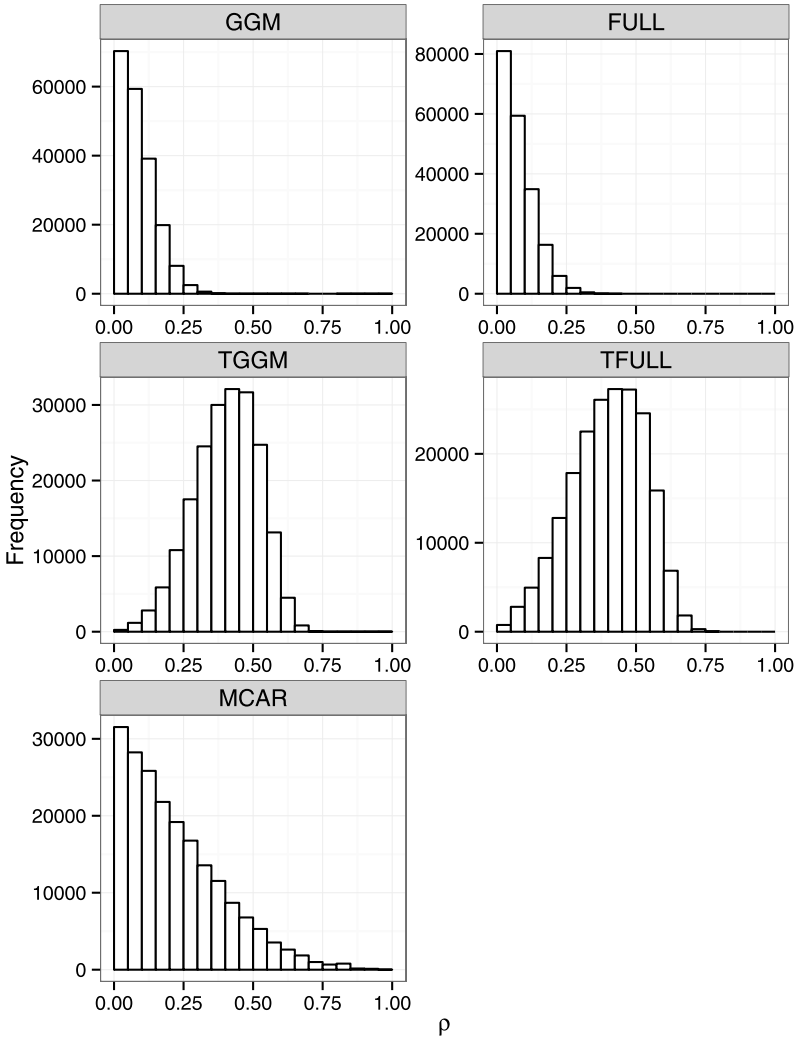

Figure 8: Posterior distribution of the spatial autocorrelation parameter ρ under the five models considered.

on the between-cancer conditional independence graph is not sensitive to the choice of prior on $\mathbf{K}_R$. The Lung–Leukemia, Bladder–Non-Hodgkin lymphoma, and Colon–Breast cancer edges have the biggest posterior edge inclusion probabilities.

Finally, we compare the GGM, TGGM, and MCAR models using within-sample fit for the complete data. Table 2 shows the average coverage and length of 95% posterior predictive intervals as well as two measures of the effective number of parameters: p_{DIC} (Spiegelhalter et al., 2002) and p_{WAIC} (Gelman et al., 2013):

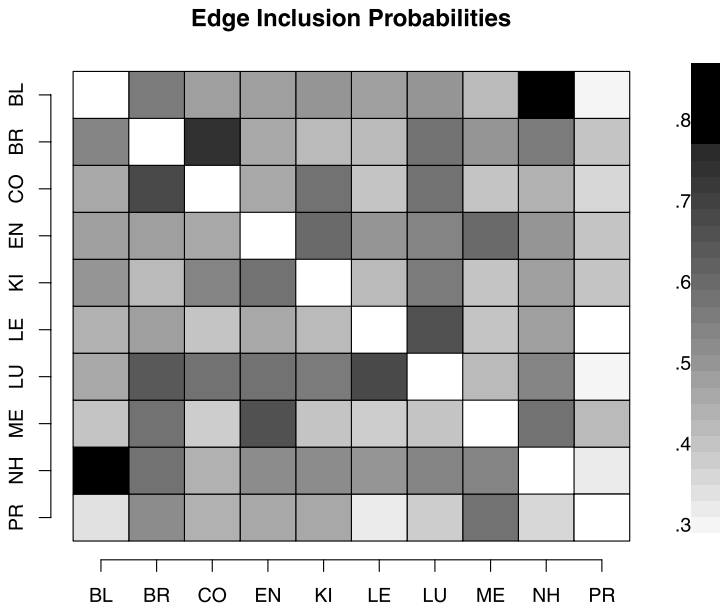

Figure 9: Pairwise edge inclusion probabilities for G_C when the prior on $\mathbf{K}_R$ is G-Wishart (upper triangle) or truncated G-Wishart (lower triangle). The abbreviations are: BL, Bladder; BR, Breast; CO, Colorectal; EN, Endometrial; KI, Kidney; LE, Leukemia; LU, Lung; ME, Melanoma of the skin; NH, Non-Hodgkin lymphoma; PR, Prostate. The Lung–Leukemia, Bladder–Non-Hodgkin lymphoma, and Colon–Breast cancer edges have the biggest posterior edge inclusion probabilities in both models.

$$p_{\text{DIC}} = 2\left(\log p(\mathbf{Y} \mid \hat{\theta}_{\text{post}}) - \mathsf{E}_{\text{post}} \log p(\mathbf{Y} \mid \boldsymbol{\Theta})\right),$$
$$p_{\text{WAIC}} = \sum_{i=1}^{n}\sum_{c=1}^{C} \text{VAR} \log p(Y_{ic} \mid \theta).$$

While all models have approximately the correct coverage, the posterior predictive intervals from the truncated G-Wishart model are slightly smaller. This remains true when averaging over the predictive intervals for small counts (≤ 5) or larger counts (≥ 20). The MCAR model has the fewest number of effective parameters by both measures, which is consistent with the parsimonious form of the spatial covariance in the MCAR model. The G-Wishart model has the largest number of effective parameters under p_{DIC} but the truncated G-Wishart has the largest number under p_{WAIC}. This inconsistency in the ordering is likely a result of differences in the shapes of the posterior predictive distributions under the GGM and TGGM models. Both p_{DIC} and p_{WAIC} measure the spread in the log posterior predictive density, but the estimators are affected differently by features such as longer tails.

The cross-validation results are somewhat sensitive to the choice of prior on ρ. We investigated fixing ρ to 0.99 or 0.9 (the mean of the $\mathsf{Beta}(18, 2)$ prior used in Jin et al.

	COV	LEN	$\text{LEN}_{\leq 5}$	$\text{LEN}_{\geq 20}$	p_{DIC}	p_{WAIC}
GGM	0.959	31.33	7.47	51.82	184.1	133.1
TGGM	0.954	**31.27**	**7.36**	**51.79**	182.2	135.3
MCAR	0.956	31.31	7.41	51.85	181.3	131.3

Table 2: Coverage rates (COV) and mean length (LEN) of the in-sample 95% credible intervals. Mean lengths are also give by ranges of observed counts. p_{DIC} and p_{WAIC} are two measures of the effective number of parameters.

(2007)) as well as using a discrete uniform prior on $\{0.05, 0.1, \ldots, 0.9, 0.95, 0.99\}$. In some cases, the predictive variance is substantially smaller than the variance in Table 1, but this comes at the cost of greater bias. The best method in terms of overall MSE is still the TGGM model where the prior on ρ is discrete uniform with additional values closer to 1. Full cross-validation results for the three additional priors on ρ are in the supplementary material.

6 Discussion

This article presents a novel extension of the G-Wishart prior for the precision matrix of spatial random effects. In a simulation study, the truncated G-Wishart prior is able to better estimate the relative risks when the outcomes are rare (i.e., the expected counts are small) and when the risk surface is not smooth. However, we found that there is not enough information in a single outcome to estimate the spatial correlation structure. The restriction of the G-Wishart prior was shown to be advantageous when used in a multivariate disease mapping context with incidence data from the Washington State Cancer Registry.

The multivariate model relies on the assumption of separability to estimate the rich correlation structure by pooling information across outcomes. The validity of the separability assumption has been carefully considered for spatiotemporal applications (Stein, 2005; Fuentes, 2006), and alternative, non-separable space–time covariance models have been proposed for Gaussian processes (Gneiting, 2002; Gneiting and Guttorp, 2010) and Gaussian Markov random fields (Knorr-Held, 2000). Gelfand and Vounatsou (2003) extend the MCAR to allow for different spatial autocorrelation parameters for each outcome, yielding non-separable model that is still relatively parsimonious, and Jin et al. (2005, 2007) further extend the MCAR paradigm by including parameters that a directly represent the correlation between different outcomes in neighboring areas. Ultimately, these MCAR extensions still make an assumption similar to separability in that the correlation between outcomes within a single areas is the same for all areas.

As mentioned in Section 2.1, others have approached this problem by directly altering the conditional independence structure (Knorr-Held and Raßer, 2000; Green and Richardson, 2002; Lee and Mitchell, 2013; Lee et al., 2014). Given that these models have been shown to outperform the traditional convolution model in some scenarios and are fairly parsimonious, these methods may be better for univariate outcomes than our

TGW model. One direction for future research is to incorporate the locally adaptive CAR (Lee and Mitchell, 2013) in the matrix variate random effect framework of Sections 3.4 and 5.

There are a number of computation issues when using the truncated G-Wishart and G-Wishart priors. Each MCMC run for the univariate truncated G-Wishart model in Section 4 takes approximately 1.5 hours to complete on a 2.5 GHz Intel Xeon E5-2640 processor, and, with the exception of the MCAR model, the MCMC for each model in Section 5 takes about 6.5 hours to complete. In contrast, estimating the convolution and ICAR models from Section 4 takes a matter of seconds in INLA. We have found that the proposal variance for updates of the Cholesky square (Section 3.3) and the random effects (see supplementary material) must be chosen carefully to avoid poor convergence. In both Sections 4 and 5, we used $s = 2$ for updating $\mathbf{\Phi}$ and $s = 0.1$ for updating $\mathbf{u}$. While the computation time for the models detailed here are not prohibitive, they may pose a challenge as we extend to more complicated datasets, such as those including multiple diseases in time and space.

R code for the simulation in Section 4 and C++ code for the analysis in Section 5 are available at http://www.lancaster.ac.uk/staff/smithtr/NGWSource.zip. Included here are the expected counts and labeling scheme for Section 4 and prototypical data for Section 5. A censored version of the data used in Section 5 is available from https://fortress.wa.gov/doh/wscr/WSCR/Query.mvc/Query.

Supplementary Material

Supplementary Material for "Restricted Covariance Priors with Applications in Spatial Statistics" (DOI: 10.1214/14-BA927SUPP; .pdf).

References

Atay-Kayis, A. and Massam, H. (2005). "A Monte Carlo method for computing the marginal likelihood in nondecomposable Gaussian graphical models." *Biometrika*, 92: 317–335. MR2201362. doi: http://dx.doi.org/10.1093/biomet/92.2.317. 6, 7

Banerjee, S., Carlin, B., and Gelfand, A. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC. 2, 4, 5

Besag, J. (1974). "Spatial interaction and the statistical analysis of lattice systems." *Journal of the Royal Statistical Society: Series B (Methodological)*, 36: 192–236. MR0373208. 3

Besag, J. and Kooperberg, C. (1995). "On conditional and intrinsic autogregressions." *Biometrika*, 82: 733–746. MR1380811. 5

Besag, J., York, J., and Mollié, A. (1991). "Bayesian image restoration, with two applications in spatial statistics." *Annals of the Institute of Statistical Mathematics*, 43: 1–59. MR1105822. doi: http://dx.doi.org/10.1007/BF00116466. 2, 4

Cancer Research UK (2013). "Cancer statistics by type." http://www.cancerresearchuk.org/cancer-info/cancerstats/types/. Last visited on 01/05/2013. 12

Carlin, B. and Banerjee, S. (2003). "Hierarchical multivariate CAR models for spatially correlated survival data." In: *Bayesian Statistics 7*, 45–65. Oxford University Press. MR2003166. 11

Chen, M.-H., Shao, Q.-M., and Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*. Springer: New York. MR1742311. doi: http://dx.doi.org/10.1007/978-1-4612-1276-8. 10

Dawid, A. (1981). "Some matrix-variate distribution theory: notational considerations and a Bayesian application." *Biometrika*, 68: 265–274. MR0614963. doi: http://dx.doi.org/10.1093/biomet/68.1.265. 11

Dawid, A. P. and Lauritzen, S. L. (1993). "Hyper Markov laws in the statistical analysis of decomposable graphical models." *The Annals of Statistics*, 21: 1272–1317. MR1241267. doi: http://dx.doi.org/10.1214/aos/1176349260. 2, 6

Dempster, A. P. (1972). "Covariance selection." *Biometrics*, 28: 157–175. doi: http://dx.doi.org/10.2307/2528966. 5

Diggle, P. and Ribeiro, P. J. (2007). *Model-based Geostatistics*. Springer. MR2293378. 2

Diggle, P., Tawn, J., and Moyeed, R. (1998). "Model-based geostatistics." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47: 299–350. MR1626544. doi: http://dx.doi.org/10.1111/1467-9876.00113. 2

Dobra, A., Lenkoski, A., and Rodriguez, A. (2011). "Bayesian inference for general Gaussian graphical models with applications to multivariate lattice data." *Journal of the American Statistical Association*, 106: 1418–1433. MR2896846. doi: http://dx.doi.org/10.1198/jasa.2011.tm10465. 2, 7, 11, 12, 17

Dutilleul, P. (1999). "The MLE algorithm for the matrix normal distribution." *Journal of Statistical Computation and Simulation*, 64: 105–123. doi: http://dx.doi.org/10.1080/00949659908811970. 12

Fong, Y., Rue, H., and Wakefield, J. (2009). "Bayesian inference for generalized linear mixed models." *Biostatistics*, 11: 397–412. doi: http://dx.doi.org/10.1093/biostatistics/kxp053. 5

Fosdick, B. K. and Hoff, P. (2014). "Separable factor analysis with applications to mortality data." *The Annals of Applied Statistics*, 8: 120–147. MR3191985. doi: http://dx.doi.org/10.1214/13-AOAS694. 11

Fuentes, M. (2006). "Testing for separability of spatial–temporal covariance functions." *Journal of Statistical Planning and Inference*, 136: 447–466. MR2211349. doi: http://dx.doi.org/10.1016/j.jspi.2004.07.004. 21

Gelfand, A. and Vounatsou, P. (2003). "Proper multivariate conditional autoregressive models for spatial data analysis." *Biostatistics*, 4: 11–25. doi: http://dx.doi.org/10.1093/biostatistics/4.1.11. 10, 11, 17, 21

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013). *Bayesian data analysis*. CRC press. MR3235677. 19

Gneiting, T. (2002). "Nonseparable, stationary covariance functions for space–time data." *Journal of the American Statistical Association*, 97: 590–600. MR1941475. doi: http://dx.doi.org/10.1198/016214502760047113. 21

Gneiting, T. and Guttorp, P. (2010). "Continuous parameter spatio-temporal processes." In: Gelfand, A., Diggle, P., Guttorp, P., and Fuentes, M. (eds.), *Handbook of Spatial Statistics*, 427–436. CRC Press. MR2730958. doi: http://dx.doi.org/10.1201/9781420072884-c23. 21

Green, P. and Richardson, S. (2002). "Hidden Markov models and disease mapping." *Journal of the American Statistical Association*, 97: 1055–1070. MR1951259. doi: http://dx.doi.org/10.1198/016214502388618870. 2, 4, 12, 21

Hoff, P. D. (2011). "Separable covariance arrays via the Tucker product, with applications to multivariate relational data." *Bayesian Analysis*, 6: 179–196. MR2806238. doi: http://dx.doi.org/10.1214/11-BA606. 12

Hughes, J. and Haran, M. (2013). "Dimension reduction and alleviation of confounding for spatial generalized linear mixed models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75: 139–159. MR3008275. doi: http://dx.doi.org/10.1111/j.1467-9868.2012.01041.x. 2

Jin, X., Banerjee, S., and Carlin, B. P. (2007). "Order-free co-regionalized areal data models with application to multiple-disease mapping." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69: 817–838. MR2368572. doi: http://dx.doi.org/10.1111/j.1467-9868.2007.00612.x. 10, 11, 20, 21

Jin, X., Carlin, B., and Banerjee, S. (2005). "Generalized hierarchical multivariate CAR models for areal data." *Biometrics*, 61: 950–961. MR2216188. doi: http://dx.doi.org/10.1111/j.1541-0420.2005.00359.x. 21

Knorr-Held, L. (2000). "Bayesian modelling of inseparable space–time variation in disease risk." *Statistics in Medicine*, 19: 2555–2568. doi: http://dx.doi.org/10.1002/1097-0258(20000915/30)19:17/18<2555::AID-SIM587>3.0.CO;2-#. 11, 21

Knorr-Held, L. and Best, N. (2001). "A shared component model for detecting joint and selective clustering of two diseases." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 164: 73–85. MR1819023. doi: http://dx.doi.org/10.1111/1467-985X.00187. 2

Knorr-Held, L. and Raßer, G. (2000). "Bayesian detection of clusters and discontinuities in disease maps." *Biometrics*, 56: 13–21. doi: http://dx.doi.org/10.1111/j.0006-341X.2000.00013.x. 4, 21

Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press. MR1419991. 5

Lee, D. and Mitchell, R. (2013). "Locally adaptive spatial smoothing using conditional auto-regressive models." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62: 593–608. MR3083913. doi: http://dx.doi.org/10.1111/rssc.12009. 4, 21, 22

Lee, D., Rushworth, A., and Sahu, S. (2014). "A Bayesian localized conditional autoregressive model for estimating the health effects of air pollution." *Biometrics*, 70: 419–429. doi: http://dx.doi.org/10.1111/biom.12156. 4, 12, 21

Lunn, D., Thomas, A., Best, N., and Spiegelhalter, D. (2000). "WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility." *Statistics and Computing*, 10: 325–337. 5

Mardia, K. and Goodall, C. (1993). "Spatial-temporal analysis of multivariate environmental monitoring data." In: Patil, G. and Rao, C. (eds.), *Multivariate Environmental Statistics*, 347–385. Elsevier. MR1268443. 11

Quick, H., Banerjee, S., and Carlin, B. (2013). "Modeling temporal gradients in regionally aggregated California asthma hospitalization data." *The Annals of Applied Statistics*, 7: 154–176. MR3086414. doi: http://dx.doi.org/10.1214/12-AOAS600. 11

Roverato, A. (2002). "Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models." *Scandinavian Journal of Statistics*, 29: 391–411. MR1925566. doi: http://dx.doi.org/10.1111/1467-9469.00297. 2, 6

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields*. Chapman & Hall/CRC. MR2130347. doi: http://dx.doi.org/10.1201/9780203492024. 3, 9

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 (Statistical Methodology)*, 71: 319–392. MR2649602. doi: http://dx.doi.org/10.1111/j.1467-9868.2008.00700.x. 5

Smith, T. R., Wakefield, J., and Dobra, A. (2015). "Supplement to "Restricted Covariance Priors with Applications in Spatial Statistics"." doi: http://dx.doi.org/10.1214/14-BA927SUPP. 8

Sørbye, S. H. and Rue, H. (2014). "Scaling intrinsic Gaussian Markov random field priors in spatial modelling." *Spatial Statistics*, 8: 39–51. doi: http://dx.doi.org/10.1016/j.spasta.2013.06.004. 5

Spiegelhalter, D., Best, N., Carlin, B., and Van Der Linde, A. (2002). "Bayesian measures of model complexity and fit." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64: 583–639. MR1979380. doi: http://dx.doi.org/10.1111/1467-9868.00353. 19

Stein, M. L. (1999). *Interpolation of spatial data: some theory for Kriging*. Springer. MR1697409. doi: http://dx.doi.org/10.1007/978-1-4612-1494-6. 2

— (2005). "Space–time covariance functions." *Journal of the American Statistical Association*, 100: 310–321. MR2156840. doi: http://dx.doi.org/10.1198/016214504000000854. 11, 21

Tobler, W. R. (1970). "A computer movie simulating urban growth in the Detroit region." *Economic Geography*, 46: 234–240. doi: http://dx.doi.org/10.2307/143141. 1

Wall, M. (2004). "A close look at the spatial structure implied by the CAR and SAR models." *Journal of Statistical Planning and Inference*, 121: 311–324. MR2038824. doi: http://dx.doi.org/10.1016/S0378-3758(03)00111-3. 5

Wang, H. and Pillai, N. S. (2013). "On a class of shrinkage priors for covariance matrix estimation." *Journal of Computational and Graphical Statistics*, 22: 689–707. MR3173737. doi: http://dx.doi.org/10.1080/10618600.2013.785732. 2

White, G. and Ghosh, S. K. (2009). "A stochastic neighborhood conditional autoregressive model for spatial data." *Computational Statistics & Data Analysis*, 53: 3033–3046. MR2667608. doi: http://dx.doi.org/10.1016/j.csda.2008.08.010. 4

Acknowledgments

TS and AD were supported in part by the National Science Foundation (DMS 1120255). JW was supported by 2R01CA095994-05A1 from the National Institutes of Health. The authors thank the Washington State Cancer Registry for providing the cancer incidence data and the referees for their helpful comments.