Journal of Statistical Planning and Inference 165 (2015) 78–90

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Bayesian semiparametric hierarchical empirical likelihood spatial models

Aaron T. Porter [a,*], Scott H. Holan [b], Christopher K. Wikle [b]

[a] Department of Applied Mathematics and Statistics, Colorado School of Mines, 1500 Illinois St., Golden, CO 80401, United States

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

ARTICLE INFO

Article history:
Received 6 May 2014
Received in revised form 5 December 2014
Accepted 10 April 2015
Available online 24 April 2015

Keywords:
Conditional autoregressive model
Fay–Herriot model
Kriging
Random field
Small area estimation

ABSTRACT

We introduce a general hierarchical Bayesian framework that incorporates a flexible nonparametric data model specification through the use of empirical likelihood methodology, which we term semiparametric hierarchical empirical likelihood (SHEL) models. Although general dependence structures can be readily accommodated, we focus on spatial modeling, a relatively underdeveloped area in the empirical likelihood literature. Importantly, the models we develop naturally accommodate spatial association on irregular lattices and irregularly spaced point-referenced data. We illustrate our proposed framework by means of a simulation study and through three real data examples. First, we develop a spatial Fay–Herriot model in the SHEL framework and apply it to the problem of small area estimation in the American Community Survey. Next, we illustrate the SHEL model in the context of areal data (on an irregular lattice) through the North Carolina sudden infant death syndrome (SIDS) dataset. Finally, we analyze a point-referenced dataset from the North American Breeding Bird Survey that considers dove counts for the state of Missouri. In all cases, we demonstrate superior performance of our model, in terms of mean squared prediction error, over standard parametric analyses.

1. Introduction

The empirical likelihood (EL) dates back to the seminal work of Owen (1988) and has become increasingly popular in recent years, as a result of Owen (2001), which placed many of the fundamental concepts in a single text. Early work by Qin and Lawless (1994) greatly expanded the use of EL by placing it in the context of estimating equations. Kolaczyk (1994) derived general conditions for the use of estimating equations for the EL that are applicable to many types of linear, nonlinear, and semiparametric models. Many EL-type estimators have since been derived, known as Generalized Empirical Likelihood (GEL) estimators. Newey and Smith (2004) provides an excellent overview of these estimators and their higher order properties.

Lazar (2003) provides evidence, by means of a simulation study, that the EL framework is appropriate for Bayesian inference. Making use of a result from Monahan and Boos (1992) that yields conditions by which a likelihood can be determined suitable for Bayesian inference, this paper initiated Bayesian research on EL and GEL estimators. Schennach (2005) derived a Bayesian GEL estimator by means of nonparametric priors and further extended their approach in Schennach (2007). Fang and Mukerjee (2006) derived the asymptotic frequentist coverage properties of the Bayesian

* Corresponding author.
E-mail address: aporter@mines.edu (A.T. Porter).

http://dx.doi.org/10.1016/j.jspi.2015.04.002
0378-3758/

credible intervals for the mean parameters of a wide class of EL-type likelihoods, and demonstrated undercoverage for credible intervals for parametric means generated by GEL estimators. Additional work comparing the properties of credible intervals for specific types of EL-type likelihoods can be found in Chang and Mukerjee (2008). In particular, this work demonstrates favorable coverage rates for the traditional EL of Owen (1988).

Bayesian hierarchical modeling (BHM) has become an expansive field. When modeling complex stochastic phenomena within the BHM framework, typically at least three levels of model hierarchy are considered, which are the data model, process model, and parameter model (Berliner, 1996; Wikle, 2003). Subsequently, modeling typically proceeds by selecting parametric distributions for each stage of the hierarchy. As demonstrated in Cressie and Wikle (2011), this framework advantageously also allows for scientifically motivated process models to be utilized at the latent stage. One aspect of this approach is that model implementation typically requires selection of an appropriate data distribution (likelihood) for the observations.

Our approach extends the general applicability of BHMs by broadly placing them in the context of the empirical likelihood. The model we propose can be viewed as a semiparametric hierarchical empirical likelihood (SHEL) model and utilizes either EL estimators or GEL estimators at the data stage of the model hierarchy. Parametric process models can then be utilized to handle the potentially complex underlying dependence structures. By placing the EL in the context of Bayesian hierarchical modeling, we alleviate the issues of modeling the dependency in the observations, which is often difficult to handle in the usual observation-driven EL framework and generally utilizes restrictive blocking arguments. Specifically, we expand the BHM framework to allow empirical data models, rather than requiring the user to select a parametric structure for the data.

Hierarchical approaches to empirical likelihood have been recently considered, but still remain largely underdeveloped, with no general framework to date. Chaudhuri and Ghosh (2011) proposed using the EL in a semiparametric hierarchical nested error regression model for small area estimations (SAE). The model they developed extends the traditional Fay–Herriot (FH) model (Fay and Herriot, 1979) to the EL framework. Although Chaudhuri and Ghosh (2011) demonstrate good model performance, their implementation utilized informative priors for some of the model parameters, and they noted sensitivity to these specifications. The general approach they propose allows for both semiparametric and nonparametric specifications of the model for the superpopulation mean, with the nonparametric specification relying on a Bayesian nonparametric formulation (i.e., a Dirichlet process mixture with Gaussian base measure). We pursue a more complete development of EL in the context of BHMs. The model we propose here is of independent interest and readily allows for various other hierarchical and/or dependence structures, such as temporal and/or spatio-temporal dependencies. However, for the sake of brevity, subsequent exposition focuses on spatially correlated data.

Based on blocking arguments originally developed for time series by Kitamura (1997), Nordman and Caragea (2008) developed a point referenced spatial model in the frequentist EL framework that considers variogram fitting for data collected on a regular grid, and assumes stationarity. Utilizing a similar blocking argument, Nordman (2008) considered an observation-driven model for spatial data on a regular lattice using the EL framework that does not require stationarity. To the best of our knowledge, hierarchical models for spatial data on an irregular lattice that explicitly account for the underlying spatial structure in the data do not exist in the current literature. A recent advancement in the spatial EL literature is Bandyopadhyay et al. (2012), in which irregularly spaced spatial data is modeled using frequency domain techniques. Their framework greatly expands EL methodology for point referenced spatial data but is based on different assumptions than those presented herein and does not immediately extend to the lattice case, where distances are not uniquely defined.

The structure of this paper is as follows. Section 2 develops methodology that will be needed for the general specification of the SHEL model. Section 3 discusses technical details related to the Bayesian estimation of the SHEL model, and provides the Markov chain Monte Carlo (MCMC) algorithm that we propose. Section 4 presents three case studies: the FH model for SAE in the context of the American Community Survey (ACS), the North Carolina SIDS data (areal data), and a point referenced dataset from the North American Breeding Bird Survey that considers dove counts for the state of Missouri. Section 5 provides concluding discussion. For ease of exposition, the results of two simulation studies, illustrating the effectiveness of our approach, are left to an Appendix.

2. Spatial SHEL models

2.1. The SHEL framework

Let $\mathbf{Z}$ be an n_Z-dimensional vector of observations, $\mathbf{Y}$ be an n_Y-dimensional vector corresponding to an unobserved process, and $\boldsymbol{\xi}$ be a set of parameters related to both the data model and process model. Here, $\mathbf{Z}$ and $\mathbf{Y}$ do not need to be of the same dimension. For example, the observations could be mapped to the unobserved process through a matrix that accounts for change-of-support or aggregation (Wikle and Berliner, 2005). However, for ease of notation, we assume $n_Z = n_Y \equiv n$, unless specified otherwise. Further, let $[\mathbf{Z}|\mathbf{Y}]$ denote the conditional distribution of $\mathbf{Z}$ given $\mathbf{Y}$ and $[\mathbf{Y}]$ denote the marginal distribution of $\mathbf{Y}$. We propose a general setup for the SHEL framework that considers a data model $[\mathbf{Z}|\mathbf{Y}, \boldsymbol{\xi}_D]$, process model $[\mathbf{Y}|\boldsymbol{\xi}_P]$, and parameter model $[\boldsymbol{\xi}] = [\boldsymbol{\xi}_D, \boldsymbol{\xi}_P]$, with $[\boldsymbol{\xi}_D]$ being the joint prior distribution of the data model parameters and $[\boldsymbol{\xi}_P]$ being the joint prior distribution of the process model parameters. The framework we propose here is not unique to spatial data, and any process model in which $[\mathbf{Y}, \boldsymbol{\xi}]$ is proper can be utilized.

The hierarchical framework that we propose is motivated by the parametric counterpart (e.g., Berliner, 1996; Wikle, 2003), but with increased flexibility from relaxing the parametric data model assumption. The SHEL structure hierarchy can be written as

Empirical Data Model: $[\mathbf{Z}|\mathbf{Y}, \boldsymbol{\xi}_D]$
Process Model: $[\mathbf{Y}|\boldsymbol{\xi}_P]$
Parameter Model: $[\boldsymbol{\xi}_D, \boldsymbol{\xi}_P]$,

where the underlying distribution $[\mathbf{Z}|\mathbf{Y}, \boldsymbol{\xi}_D]$ is assumed to have two finite moments. Critically, we further assume $E(\mathbf{Z}|\mathbf{Y}, \boldsymbol{\xi}_D) = g(\mathbf{X}\boldsymbol{\beta} + \mathbf{Y})$ and $E(\mathbf{Z}^2|\mathbf{Y}, \boldsymbol{\xi}_D) = h(\mathbf{X}\boldsymbol{\beta} + \mathbf{Y})$ for g and h known, with $\mathbf{X}$ being an $n \times m$ design matrix of fixed and known covariate information. These relationships will serve to inform a set of estimating equations utilized in estimating the parameters of the empirical data model.

When utilizing the SHEL framework, $[\mathbf{Z}|\mathbf{Y}, \boldsymbol{\xi}_D]$ will be modeled empirically, using the EL. As a result, our approach typically allows for the data to be modeled directly. This avoids the need to identify an appropriate transformation in order to model data that do not follow a known distribution and allows for model development to proceed in cases where no appropriate transformation exists. The spatial SHEL model we propose creates a unifying model for empirical likelihood-based Bayesian hierarchical spatial modeling.

One of the main advantages of working in the hierarchical paradigm with an EL data model is the ability to introduce conditional independence in a natural way, specifying the dependence structure at a higher level in the model hierarchy. That is, dependence among outcomes in a spatial (and/or temporal) setting is handled by conditioning on a latent spatial (and/or temporal) process. By taking a conditional approach, the original formulation of the EL, which assumes independent and identically distributed (i.i.d.) observations, becomes immediately applicable — although the assumption of independent observations could be relaxed (e.g., see Owen, 2001, Chapter 4). In other words, the SHEL framework effectively utilizes the conditional model specification inherent to BHMs to extend the applicability of the EL to a broad range of analyses. In doing so, we alleviate some of the strict assumptions often required of the blocking arguments used in EL modeling of dependent data, such as those in Kitamura (1997), Nordman and Caragea (2008), Nordman (2008), and Kaiser and Nordman (2012).

2.2. Empirical likelihood

The use of estimating equations in the EL framework (Qin and Lawless, 1994) has recently been used in the FH model by Chaudhuri and Ghosh (2011) and represents an attractive way to employ EL in the BHM framework. Generally, the EL of a vector of functionals $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_R\}$ given independent and identically distributed observations $Z_1, \ldots, Z_n$, can be computed as

$$L(\boldsymbol{\theta}) \propto \prod_{i=1}^{n} w_i(\boldsymbol{\theta}), \tag{1}$$

where $L(\boldsymbol{\theta})$ is maximized over the simplex

$$W_\theta = \left\{ \sum_{i=1}^{n} w_i = 1;\ w_i > 0 \text{ for all } i\ ;\ \sum_{i=1}^{n} w_i m_j(z_i, \theta_i) = 0 \text{ for all } j \right\} \tag{2}$$

and R is the number of functionals to be estimated. Here, for i in $1, \ldots, n$, $\{m_j(z_i, \theta_i)\}_{j=1,\ldots,J}$ are a set of J estimating equations and $\boldsymbol{\theta} \in \mathbb{R}^J$ are of the form $k_j(\sum_{i=1}^{n} w_i z_i) = \theta_j$, for known functions $k_j(\cdot)$, where we have assumed $J = R$, i.e., that unstructured $\boldsymbol{\theta}$ is not under- or overspecified. Without covariate information, one cannot estimate more parameters than the number of estimating equations. However, Chaudhuri and Ghosh (2011) suggest utilizing structured $\boldsymbol{\theta}$, by which each location $i = 1, \ldots, n$ has a unique mean and variance. Covariate information is then used to provide structure to a set of mean parameters $\{\theta_1, \ldots, \theta_n\}$, where θ_i is modeled based on auxiliary information $\mathbf{x}_i$. This covariate information allows the dimension of $\boldsymbol{\theta}$ to be greater than J. The estimating equations Chaudhuri and Ghosh (2011) suggest have the form

$$\begin{aligned} &\sum_{i=1}^{n} w_i\{z_i - \theta_i\} = 0 \\ &\sum_{i=1}^{n} \{w_i(z_i - \theta_i)^2 / V(\theta_i)\} - 1 = 0, \end{aligned} \tag{3}$$

which are derived based on the exponential family. In the exponential family we define θ_i to be mean of Z_i and $V(\theta_i)$ to be the variance of $Z_i|\theta_i$. These easily extend to the GEL framework, but $V(\theta_i)$ is no longer properly considered a variance, instead serving as a scale parameter.

In the SHEL framework, θ_i will denote the conditional mean of $Z_i|Y_i$. The estimating equations approach is natural for the SHEL framework because one can compute the EL based on known formulas given proposed values for $\{\theta_i\}$. When utilizing the estimating equation approach to the EL, the model weights can be computed as

$$w_i = \frac{1}{n}\left(\frac{1}{1+\sum_{j=1}^{J}\lambda_j m_j(z_i,\theta_i)}\right), \tag{4}$$

where $\lambda_j, j = 1,\ldots,J$ satisfies

$$\sum_{i=1}^{n}\frac{m_j(z_i,\theta_i)}{1+\sum_{j=1}^{J}\{\lambda_j m_j(z_i,\theta_i)\}} = 0$$

for all j, and $\{z_i\}$ denote the observations. Clearly, these weights are monotone in each element of $\boldsymbol{\lambda} = \{\lambda_1,\ldots,\lambda_J\}$.

The likelihood can be extended to a set of GEL estimators (Smith, 1997) by the function

$$L(\theta) \propto \prod_{i=1}^{n}\widehat{w}_i, \tag{5}$$

where $\widehat{w}(\boldsymbol{\theta}) = \text{argmax}_{W_\theta}\sum_{i=1}^{n} f\{w_i(\boldsymbol{\theta})\}$ for a known function $f_\theta(w_i)$. Two notable choices include $f_\theta(w_i) = \log(w_i)$, which is the traditional EL function first introduced by Owen (1988), and $f_\theta(w_i) = -w_i\log(w_i)$, which was introduced by Schennach (2005) and represents the exponentially tilted empirical likelihood (ETEL) estimator. Henceforth, we utilize only the traditional EL of Owen (1988) throughout the methodological development, but note that other choices of $f_\theta(\cdot)$ in the GEL family could also be used.

An important observation of Chaudhuri and Ghosh (2011) is that the non-analytic form of the posterior distributions introduced by the EL makes verification of propriety of these models difficult. Therefore, improper priors should generally not be used in the SHEL framework, as only proper priors can guarantee propriety of the posterior parameter distributions.

2.3. Lattice priors for the SHEL framework

Intrinsic Gaussian Markov Random Fields (IGMRFs) (Rue and Held, 2005), such as the intrinsic conditional autoregressive model (ICAR) (Besag et al., 1991), may seem to be a poor choice for a SHEL prior due to the impropriety implicit to these models. However, recent developments in lattice priors allow for modification of the ICAR to yield a proper prior, while avoiding some of the common difficulties of proper CAR models.

A common ICAR model specification is given by

$$Y_i \sim N\left(\sum_{j\in ne(i)}\left\{\frac{b_{ij}}{\sum_{j\in ne(i)} b_{ij}}y_j\right\}, \frac{\sigma^2}{\sum_{j\in ne(i)} b_{ij}}\right),$$

where $b_{ij} = 1$ if locations i and j are neighbors and 0 otherwise, and $j \in ne(i)$ indicates that locations i and j are neighbors. This yields a probability density function for $\mathbf{Y} = (Y_1,\ldots,Y_n)'$ given by

$$\pi(\mathbf{Y}=\mathbf{y}) \propto \exp\left\{-\frac{1}{2}\mathbf{y}'\tau(\mathbf{B}_+ - \mathbf{B})\mathbf{y}\right\},$$

where $\mathbf{B}$ is a matrix with $\{B\}_{(i,j)} = b_{ij}$ and $\mathbf{B}_+$ is a diagonal matrix with $\{B_+\}_{i,i} = \sum_{j\in ne(i)} b_{ij}$. The IGMRF specification of this model would therefore imply the log-density of $\mathbf{y}$ as

$$\pi(\mathbf{Y}=\mathbf{y}) = -\frac{n-1}{2}\log(2\pi) + \frac{1}{2}\sum_{i=1}^{n-1}\log(\lambda_i) - \frac{1}{2}\mathbf{y}'\tau(\mathbf{B}_+ - \mathbf{B})\mathbf{y}, \tag{6}$$

where $\lambda_1 \geq \cdots \geq \lambda_n$ are the ordered eigenvalues of $(\mathbf{B}_+ - \mathbf{B})$.

Because $(\mathbf{B}_+ - \mathbf{B})\mathbf{1} = \mathbf{0}$, where $\mathbf{1}$ is a vector of ones, we see that the precision matrix $\tau(\mathbf{B}_+ - \mathbf{B})$ is singular, and the ICAR can only be utilized as an improper prior. One possible solution is to modify the matrix $\tau(\mathbf{B}_+ - \mathbf{B})$, by adding a spatial dependency parameter. For this ICAR parameterization, the matrix $\tau(\mathbf{B}_+ - \rho\mathbf{B})$ is guaranteed to be positive definite for $\rho \in (-1, 1)$. However, there are two major drawbacks to introducing a spatial dependency parameter ρ. First, ρ must be quite large to generate significant spatial dependency, and a uniform prior distribution often leads to diffuse posterior distributions for ρ. Second, Wall (2004) notes undesirable properties of the pairwise correlations of the locations on an irregular lattice as ρ is varied throughout the space $(-1, 1)$.

Hughes and Haran (2013) utilize an orthogonalization argument derived in Reich et al. (2006) by considering orthogonal spatial smoothing using a generalized Moran basis. Hughes and Haran (2013) smooth orthogonal to $\mathbf{X}$ by considering an eigenvector basis of $\mathbf{P}_c\mathbf{B}\mathbf{P}_c$ for the latent process space, where $\mathbf{P}_c = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. This allows orthogonal smoothing to $\mathbf{X}$ while accounting for the underlying lattice structure of the data. In our formulation of a SHEL model on a lattice, we utilize this structure. We define $\mathbf{M}$ as an $n \times q$ matrix with the columns being the eigenvectors corresponding to the q largest nonzero eigenvalues of the matrix $\mathbf{P}_c\mathbf{B}\mathbf{P}_c$. The process $\mathbf{Y}_n$ can then be modeled in a rank-reduced form, $\mathbf{Y}_n = \mathbf{M}_{n\times q}\mathbf{Y}_q^*$, where $\mathbf{Y}_q^*$ is the rank-reduced process. This model is useful because, under weak conditions, the prior of this Moran basis prior is proper. Hughes and Haran (2013) illustrate that the eigenvectors associated with the positive eigenvalues of this matrix are associated with positive spatial autocorrelation, and that the larger positive eigenvalues are associated with larger dependence structures. The eigenvectors associated with negative eigenvalues behave similarly, but handle negative spatial autocorrelation. Negative spatial autocorrelation is rarely desirable in lattice modeling, so selecting the first q eigenvectors to form the Moran basis serves to respect the underlying lattice without the need to introduce new parameters. We now provide a sufficient condition for $\mathbf{M}'(\mathbf{B}_+ - \mathbf{B})\mathbf{M}$ to be positive definite:

Theorem 1. *Consider a Bayesian hierarchical model in which the data model has two finite moments* $E(\mathbf{Z}|\mathbf{Y}, \boldsymbol{\xi}_D) = g(\mathbf{X}\boldsymbol{\beta}+\mathbf{M}\mathbf{Y}^*)$ *and* $E(\mathbf{Z}^2|\mathbf{Y}, \boldsymbol{\xi}_D) = h(\mathbf{X}\boldsymbol{\beta} + \mathbf{M}\mathbf{Y}^*)$ *with g and h being known functions. Let the process* $\mathbf{Y}^*$ *be given a Moran basis prior of the form* $\pi(\mathbf{Y}^* = \mathbf{y}^*) \propto \tau^{q/2}\exp\{-\frac{1}{2}\tau\mathbf{y}^{*'}\mathbf{M}'(\mathbf{B}_+ - \mathbf{B})\mathbf{M}\mathbf{y}^*\}$*, where rank* $(\mathbf{M}) \leq n-1$*. Assume that* $\mathbf{B}$ *is the adjacency matrix for a first order IGMRF (i.e., rank* $(\mathbf{B}) = n-1$*). Then, a sufficient condition for* $\mathbf{M}'(\mathbf{B}_+ - \mathbf{B})\mathbf{M}$ *to be positive definite is that the design matrix* $\mathbf{X}$ *contains a column corresponding to an intercept term.*

A proof of Theorem 1 can be found in Appendix A. Theorem 1 implies that the Moran basis lattice prior will yield a proper prior suitable for use with EL methods on a lattice whenever one includes an intercept term in the design matrix. The main advantage of this model over other lattice priors is that this basis simultaneously allows for dimension reduction. Let $\lceil x \rceil$ denote the ceiling of x — the smallest integer greater than or equal to x. The recommendation of Hughes and Haran (2013) is that the eigenvectors associated with the largest $q = \lceil 0.1n \rceil$ eigenvalues of the matrix $\mathbf{P}_c\mathbf{B}\mathbf{P}_c$ are typically sufficient to allow accurate estimation of the fixed effects, though there is some sensitivity to the actual proportion used. In our analyses, which are of much lower dimensionality than those considered in Hughes and Haran (2013), we have found that the prediction is markedly better in terms of mean squared prediction error (MSPE) when we utilize every eigenvector of $\mathbf{P}_c\mathbf{B}\mathbf{P}_c$ associated with a positive eigenvalue. This strategy leads to substantially decreased computation time in the SHEL framework, along with simpler tuning of the Markov chain Monte Carlo (MCMC) algorithms employed in this model relative to the full-rank implementation.

3. Bayesian model estimation

3.1. Computational details

EL computation is well established. As early as 2001, several methods had been developed (Owen, 2001), with additional methods building off of this early research. Chen et al. (2002) is notable in that it provides a method for computing the EL with guaranteed convergence. We propose a straightforward approach that allows the built-in optimization functionality of the R programming language (R Core Team, 2013) to be utilized for fast computation.

An issue to overcome is selecting starting parameters and latent values that allow the EL to be computed. We propose setting the process model values to zero, and utilizing the maximum empirical likelihood estimates (MELEs) of the fixed effects as the starting values of the chain. The *gmm* package in R (Chaussé, 2010) can be used to rapidly obtain these starting values. MCMC computations can then proceed via standard Metropolis–Hastings methodology for any parameter appearing in the estimating equations for the EL portion of the model.

In the case where the model defined by the estimating equations approach to EL as outlined in (3) is not over- or under-determined, the solution for $\boldsymbol{\lambda} = \{\lambda_1, \ldots, \lambda_J\}$, if it exists, is unique for a given value of $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_R\}$. The EL constraints for $\boldsymbol{\lambda}$ are $\left\{\sum_{i=1}^n w_i m_j(z_i, \theta_j) = 0\right\}_{j=1,\ldots,J}$. This structure can be exploited by using the *optim* function in R in order to find the minimum of $\sum_{j=1}^J\{\sum_{i=1}^n w_i m_j(z_i, \theta_j)\}^2$. If the value of this objective function is zero, we can verify that the solution for $\{\lambda_1, \ldots, \lambda_J\}$ yields a set of weights $\{w_i, i = 1, \ldots, n\}$ in the simplex of (2), by checking that $\sum_{i=1}^n w_i = 1$ and that $w_i > 0$ for all i. When these conditions are met, we have the value of the EL as $\prod_{i=1}^n w_i$. When using the *optim* function, which is the default fitting method for the *gmm* package, one must decide on a numerical threshold for deciding when $\left\{\sum_{i=1}^n w_i m_j(z_i, \theta_j) = 0\right\}_{j=1,\ldots,J}$ and $\sum_{i=1}^n w_i = 1$ are satisfied. We have had success in evaluating this term by considering $\left\{\sum_{i=1}^n w_i m_j(z_i, \theta_j) < \epsilon\right\}_{j=1,\ldots,J}$ and $(\sum_{i=1}^n w_i) - 1 < \epsilon$ where $\epsilon = 5 \times 10^{-3}$.

Because the data model is non-analytic, Gibbs sampling is not possible for any of the parameters, as none of the full conditional distributions are of standard form. Therefore, we utilize Metropolis–Hastings within Gibbs (MH) sampling for all of the model parameters. Specifically, for our analyses, we use a random walk MH sampling algorithm having Gaussian proposals with variances tuned based on the empirical covariances from a pilot chain (Gelman et al., 2013). An example of the algorithm can be found in Section 3.2.

3.2. MCMC sampling algorithm

Herein, we provide the sampling algorithm used to sample the SHEL Fay–Herriot model of Section 4.1. Sampling algorithms for the other models discussed are similar and proceed in a straightforward manner. The sampling algorithm proceeds as follows.

1. Utilizing the estimating equations

$$\sum_{i=1}^{n} w_i\{z_i - \mathbf{x}_i'\boldsymbol{\beta}\} = 0,$$
$$\sum_{i=1}^{n} \{w_i(z_i - \mathbf{x}_i'\boldsymbol{\beta})^2/\sigma_i^2\} - 1 = 0,$$

and the gmm package in the R programming language, generate the MELEs for $\boldsymbol{\beta}$ given that the latent process, $\mathbf{Y}_q^*$, is set identically equal to zero. Next set the initial values for $\boldsymbol{\beta}$ to the MELE values and set $\mathbf{Y}_q^* = \mathbf{0}$. This provides starting values for $\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{M}\mathbf{Y}_q^*$ that generate a set of weights $\{w_i\}$ guaranteed to be in the simplex

$$W_\theta = \left\{ \sum_{i=1}^{n} w_i = 1;\ w_i > 0 \text{ for all } i\ ;\ \sum_{i=1}^{n} w_i m_j(z_i, \boldsymbol{\theta}) = 0 \text{ for all } j \right\}. \tag{7}$$

2. **Sampling $\mathbf{Y}_q^*$**
In blocks of size B (we use $B = 15$) we sample $\mathbf{Y}_q^*$ using a random walk Metropolis–Hastings step with a multivariate normal proposal for block k, $\widetilde{\mathbf{y}}_{q,k}^* \sim N(\mathbf{y}_{q,k}^*, \boldsymbol{\Sigma}_{Y_{q,k}})$, where the proposal covariance $\boldsymbol{\Sigma}_{Y_{q,k}}$ is tuned based on pilot chains (Gelman et al., 2013). We utilize the proposed values with the estimating equations

$$\sum_{i=1}^{n} w_i\{z_i - \mathbf{x}_i'\boldsymbol{\beta} - \mathbf{M}_i\widetilde{\mathbf{y}}_q^*\} = 0$$
$$\sum_{i=1}^{n} \{w_i(z_i - \mathbf{x}_i'\boldsymbol{\beta} - \mathbf{M}_i\widetilde{\mathbf{y}}_q^*)^2/\sigma_i^2\} - 1 = 0$$

to generate a set of weights $\{\widetilde{w}_i\}$, where $\mathbf{M}_i$ is the ith row of $\mathbf{M}$, and the elements of $\widetilde{\mathbf{y}}_q^*$ in block k are set to $\{\widetilde{\mathbf{y}}_{k,q}^*\}$, and the elements of $\widetilde{\mathbf{y}}_q^*$ that are not in block k are left as $\{\mathbf{y}_q^*\}$. Once generated, we verify that $\{\widetilde{w}_i\}$ satisfies (7). If it does not, the block of B elements of $\{\mathbf{y}_q^*\}$ remains at their previous values, and we move to the next block of B elements of $\{\mathbf{y}_q^*\}$. Otherwise, perform a Metropolis–Hastings step with the posterior density ratio

$$\Upsilon_{\mathbf{y}_q} = \frac{p(\mathbf{Z}|\widetilde{\mathbf{y}}_q^*, \boldsymbol{\beta})\pi(\widetilde{\mathbf{y}}_q^*|\tau)}{p(\mathbf{Z}|\mathbf{y}_q^*, \boldsymbol{\beta})\pi(\mathbf{y}_q^*|\tau)}$$
$$\Upsilon_{\mathbf{y}_q} = \frac{\prod_{i=1}^{n}(\widetilde{w}_i)\exp(-\frac{1}{2}\widetilde{\mathbf{y}}_q^{*'}\mathbf{M}'\{\mathbf{B}_+ - \mathbf{B}\}\mathbf{M}\widetilde{\mathbf{y}}_q^*\tau)}{\prod_{i=1}^{n}(w_i)\exp(-\frac{1}{2}\mathbf{y}_q^{*'}\mathbf{M}'\{\mathbf{B}_+ - \mathbf{B}\}\mathbf{M}\mathbf{y}^*\tau)}.$$

We accept $\widetilde{\mathbf{y}}_q^*$ if $\Upsilon_{\mathbf{y}_q} > u_{\mathbf{y}_q}$, where $u_{\mathbf{y}_q} \sim \text{Unif}(0, 1)$. Repeat this process for every block of B elements of $\{y_q^*\}$ until the entire set has been considered.

3. **Sampling $\boldsymbol{\beta}$**
We sample $\boldsymbol{\beta}$ using a random walk Metropolis–Hastings step with a multivariate normal proposal $\widetilde{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \Sigma_\beta)$, where the proposal covariance Σ_β is tuned based on pilot chains. We use the estimating equations

$$\sum_{i=1}^{n} w_i\{z_i - \mathbf{x}_i'\widetilde{\boldsymbol{\beta}} - \mathbf{M}_i\mathbf{y}_q^*\} = 0$$
$$\sum_{i=1}^{n} \{w_i(z_i - \mathbf{x}_i'\widetilde{\boldsymbol{\beta}} - \mathbf{M}_i\mathbf{y}_q^*)^2/\sigma_i^2\} - 1 = 0$$

to generate a set of weights $\{\widetilde{w}_i\}$. Once generated, we verify that $\{\widetilde{w}_i\}$ satisfies (7). If it does not, we set $\boldsymbol{\beta}$ to the previous values. Otherwise, perform a Metropolis–Hastings step with the posterior density ratio

$$\Upsilon_\beta = \frac{p(\mathbf{Z}|\mathbf{y}_q^*, \widetilde{\boldsymbol{\beta}})\pi(\widetilde{\boldsymbol{\beta}}|\tau)}{p(\mathbf{Z}|\mathbf{y}_q^*, \boldsymbol{\beta})\pi(\boldsymbol{\beta}|\tau)}$$

$$\Upsilon_\beta = \frac{\prod_{i=1}^{n}(\widetilde{w}_i)\exp(-\frac{1}{2}\{\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\}'\{\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\}g\tau)}{\prod_{i=1}^{n}(w_i)\exp(-\frac{1}{2}\{\boldsymbol{\beta}-\boldsymbol{\beta}^*\}'\{\boldsymbol{\beta}-\boldsymbol{\beta}^*\}g\tau)},$$

where g is Zellner's g prior and $\boldsymbol{\beta}^*$ are the weighted least squares estimates of $\boldsymbol{\beta}$. We accept $\widetilde{\boldsymbol{\beta}}$ if $\Upsilon_\beta > u_\beta$, where $u_\beta \sim \text{Unif}(0, 1)$.

4. **Sampling** τ

We sample τ using a random walk Metropolis–Hastings step with a normal proposal $\widetilde{\tau}^* \sim N(\tau, \Sigma_\tau)$, where the proposal variance Σ_τ is tuned based on pilot chains. We then perform a Metropolis–Hastings with the posterior density ratio

$$\Upsilon_\tau = \frac{\pi(\boldsymbol{\beta}|\widetilde{\tau})\pi(\mathbf{Y}_q^*|\widetilde{\tau})\pi(\widetilde{\tau})}{\pi(\boldsymbol{\beta}|\tau)\pi(\mathbf{Y}_q^*|\tau)\pi(\tau)}$$

$$\Upsilon_\tau = \frac{\widetilde{\tau}^{\frac{q+p}{2}}\exp(-\frac{1}{2}\mathbf{y}_q^{*'}\mathbf{M}'\{\mathbf{B}_+-\mathbf{B}\}M\mathbf{y}_q^*\widetilde{\tau})\exp(-\frac{1}{2}\{\boldsymbol{\beta}-\boldsymbol{\beta}^*\}'\{\boldsymbol{\beta}-\boldsymbol{\beta}^*\}\widetilde{\tau}g)\widetilde{\tau}^{-(1+\alpha_1)}\exp(-\frac{\alpha_2}{\widetilde{\tau}})}{\tau^{\frac{q+p}{2}}\exp(-\frac{1}{2}\mathbf{y}_q^{*'}\mathbf{M}'\{\mathbf{B}_+-\mathbf{B}\}M\mathbf{y}_q^*\tau)\exp(-\frac{1}{2}\{\boldsymbol{\beta}-\boldsymbol{\beta}^*\}'\{\boldsymbol{\beta}-\boldsymbol{\beta}^*\}\tau g)\tau^{-(1+\alpha_1)}\exp(-\frac{\alpha_2}{\tau})},$$

where we have used an IG(α_1, α_2) prior distribution for τ. We accept $\widetilde{\tau}$ if $\Upsilon_\tau > u_\tau$, where $u_\tau \sim \text{Unif}(0, 1)$.

5. Utilizing (7), Steps 2–4 are repeated until convergence.

4. Case studies

4.1. A SHEL Fay–Herriot model

The FH model (Fay and Herriot, 1979) is a SAE model and can be written as

$$\begin{aligned} Z_i &= \theta_i + \epsilon_i \\ \theta_i &= \mathbf{x}_i'\boldsymbol{\beta} + y_i, \end{aligned} \tag{8}$$

where Z_i is a design unbiased survey estimate of θ_i, the superpopulation parameter of interest at location i, and ϵ_i is a spatially referenced sampling error with mean zero and known variance σ_i^2. Auxiliary information at location i is denoted by $\mathbf{x}_i$, and $\mathbf{y} = (y_1, \ldots, y_n)'$ denotes a vector of spatially referenced random effects.

Additionally, one typically assumes that, for $i = 1, \ldots, n$, ϵ_i are independent and that $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)'$ follows a multivariate normal distribution. Chaudhuri and Ghosh (2011) employed a Bayesian nested error regression in the FH framework that relaxed this assumption. Their analysis is demonstrated using two possible priors on $\mathbf{y}$. The first prior is an independent and identically distributed (i.i.d.) Gaussian distribution, whereas the second prior is a Dirichlet process (DP) prior with a Gaussian base measure. The actual estimating equations we utilize in the EL for estimating $\{\boldsymbol{\beta}, Y\}$ are:

$$\sum_{i=1}^{n} w_i\{z_i - \theta_i\} = 0$$

$$\sum_{i=1}^{n}\{w_i(z_i-\theta_i)^2/\sigma_i^2\} - 1 = 0.$$

In our FH analysis, we consider the parameter of interest to be the 2010 five year period estimate of mean per capita income in Missouri counties, obtained from the American Community Survey (ACS) (www.census.gov/ACS), which was scaled by 10,000 for numerical stability. We utilize the percentage of unemployed individuals in each county as auxiliary information, also obtained from the ACS. The data are not normally distributed, and neither a log nor a Box–Cox transformation yielded normality.

For the SHEL analysis, the prior on $\mathbf{y}^*$ (the reduced-rank process) is taken as $N(0, \{\tau\mathbf{M}'(B_+ - B)\mathbf{M}\}^{-1})$, where $\mathbf{M}$ is a matrix that contains the eigenvectors of the Moran basis associated with the positive eigenvalues of the matrix $\mathbf{P}_c\mathbf{B}\mathbf{P}_c$ as columns. We compare our model to the independence model of Chaudhuri and Ghosh (2011), and a model using the DP prior. For these data, there was no available transformation that satisfied the normality assumption of the data, but we perform a naïve parametric FH analysis modeling $\boldsymbol{\epsilon}$ as independent and normally distributed random errors for comparison.

For the independence prior, we used the specification $y_i \sim \text{N}(0, A)$ with $A \sim \text{IG}(1, 1)$, $\boldsymbol{\beta} \sim \text{N}(\boldsymbol{\beta}^*, g^{-1}A\mathbf{I}_2)$. The constant g represents Zellner's g prior (Zellner, 1986), here set to 10. The prior means, $\boldsymbol{\beta}^*$, are the weighted least squares (WLS) estimates from a regression of the auxiliary information on the data assuming no latent effects are present. These prior specifications represent an identical formulation as in Chaudhuri and Ghosh (2011). For comparison, the SHEL model utilizes the Moran basis, with the vague priors $\tau \sim \text{Gamma}(1, 1)$, $\boldsymbol{\beta} \sim \text{N}(\boldsymbol{\beta}^*, g^{-1}\tau^{-1}\mathbf{I}_2)$, and $\mathbf{y} = \mathbf{M}\mathbf{y}^*$ where $\mathbf{y}^* \sim \text{N}(\mathbf{0}, \tau\mathbf{M}'\{\mathbf{B}_+ - \mathbf{B}\}\mathbf{M})$.

Table 1
Posterior medians and 95% (central) credible intervals for the FH example (Section 4.1). A represents the variance of $\mathbf{y}$ in Chaudhuri and Ghosh (2011) parameterizations, and τ^{-1} for the SHEL parameterization. MPV is the mean posterior variance of $\boldsymbol{\theta}$ for each model.

Model	β_0	β_1	A	MSPE
SHEL	2.164	−0.042	0.287	0.066
	(2.051, 2.256)	(−0.063, −0.015)	(0.157, 0.628)	
Independence EL	2.230	−0.077	0.008	0.128
	(2.210, 2.364)	(−0.095, −0.058)	(0.004, 0.015)	
DP EL	2.331	−0.0375	0.049	0.128
	(2.170, 2.474)	(−0.069, −0.002)	(0.006, 0.745)	
Independence Parametric	2.094	−0.006	0.142	0.130
	(1.971, 2.217)	(−0.027, 0.015)	(0.109 0.187)	
Spatial Parametric	2.327	−0.058	0.503	0.076
	(2.284, 2.370)	(−0.067, −0.050)	(0.345, 0.765)	

The prior for the DP process prior was $y_i|G \sim G$, $G|A \sim \mathrm{DP}(\alpha, \mathcal{G})$, where $\alpha \equiv 1$ as in Chaudhuri and Ghosh (2011) and $\mathcal{G}$ represents a Gaussian base measure. For computational reasons, we approximate the DP prior using a finite mixture of normals. For these data, we considered possible cluster counts of 20, 50, and 115 (the full data size), and found our results to be robust in terms of MSPE to the number of clusters we select *a priori*. However, Chaudhuri and Ghosh (2011) utilize an informative prior on A, and we note sensitivity to this prior specification for our data. We assumed several prior specifications for A in their framework, and we present the results for $A \sim IG(2, 1000)$, which yielded the lowest MSPE of any prior we tried (MSPE of 0.128). We define MSPE as $\sum_{i=1}^{115}(Z_i - \widehat{Z}_{(-i)})^2/115$ with $\widehat{Z}_{(-i)}$ being the prediction at location i when the data at location i is treated as missing. We note that the performance of both the DP prior and the Moran's I prior may be influenced by the scale of the underlying lattice. In order to assess the relative importance of the spatial structure, we additionally consider a spatial parametric FH model with the exact same prior specifications as our SHEL model but with a Gaussian data distribution.

Table 1 reports summary statistics for the posterior distributions of these models, as well as the mean posterior variances for the mean posterior predicted variances for $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_n\}$. All model results are based on 11,000 MCMC iterations with the first 1000 iterations discarded for burn-in (i.e., 10,000 iterations total). Convergence was assessed through visual inspection of the sample chains, with no deviations from convergence detected.

We additionally performed a leave-one-out MSPE analysis for each model. The parametric model performs nearly as well in terms of MSPE as the models of Chaudhuri and Ghosh (2011). The DP prior model of Chaudhuri and Ghosh (2011), performs nearly equivalently to their independence model in terms of MSPE. In summary, the SHEL model, which explicitly accounts for the spatial correlation in these data, performs markedly better than all three other models, and yields a MSPE of 0.066, while the best fitting model of Chaudhuri and Ghosh (2011) yields an MSPE of 0.128, which is a reduction in MSPE of 48.4%. These results strongly indicate that the SHEL model with the Hughes and Haran (2013) lattice prior is the preferred model for these data. Additionally, the spatial parametric FH model yielded a MSPE of 0.076, underscoring the importance of accounting for the spatial correlation in these data. The differences in MSPE for each location are plotted spatially in Fig. 1 and clearly illustrate that the SHEL model provides estimates that deviate less from the observed data in the high population areas near St. Louis, MO and Kansas City, MO. These cities greatly influence the surrounding areas, and the explicit spatial autocorrelation embedded in the SHEL FH model greatly aids in the estimation of these areas. Due to the similarity in spatial performance, the spatial parametric FH is not shown in this figure.

4.2. The north Carolina SIDS dataset

The North Carolina Sudden Infant Death Syndrome (SIDS) dataset is a frequently analyzed areal dataset in spatial modeling. We utilize the data collected over the period from 1974 to 1978. After accounting for the counts of live births in North Carolina, there is still a significant clustering of events (e.g., Getis and Ord, 1992; Kulldorf, 1997). For this particular dataset, several parametric models have been considered. For example, Symons et al. (1983) first attempted to model the spatial structure in these data based on high risk and low risk populations. More recent work has considered models with more explicit formulations. The parametric model we utilize is

$$Z_i|\lambda_i \sim \mathrm{Poisson}(\lambda_i), \quad i = 1, \ldots, n;$$
$$\log(\lambda_i) = log(E_i) + \mathbf{x}_i'\boldsymbol{\beta} + y_i,$$

which is suggested by Cressie and Chan (1989). Here E_i is the expected SIDS count in each county, which is computed as $N_i\{\sum_{i=1}^{n}(Z_i)/\sum_{i=1}^{n}(N_i)\}$, where N_i is the total number of births in county i. We utilize an intercept, and the proportion of births in each county that resulted in non-white children as covariates. A Tukey–Freeman transformation was applied to the proportion of non-white births, as suggested by Cressie and Chan (1989). This data is well modeled by an overdispersed Poisson distribution, with the exception of a single extreme outlier, which is Anson county. In the analysis performed by Cressie and Chan (1989), this location was left out of the analysis. More recently, Sengupta and Cressie (2013) dealt with this outlier in the empirical Bayesian hierarchical model setting by modeling the data through a non-stationary spatial

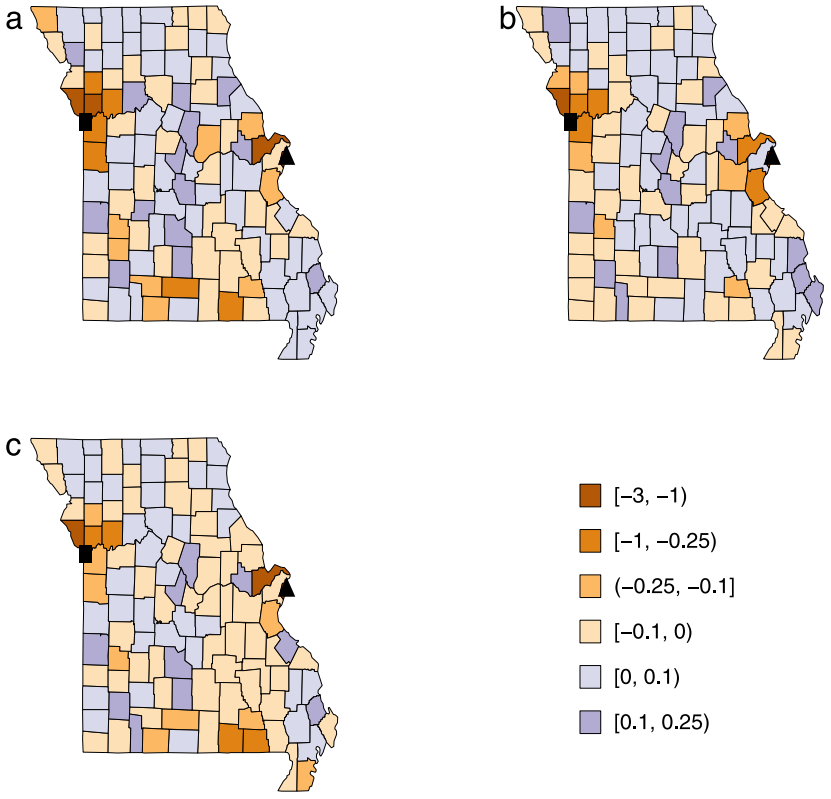

Fig. 1. The difference of the squared deviations $(Y_i - \widehat{Y}_{(-i)})^2$ for each location of estimated per capita income for (a) the SHEL model versus Chaudhuri and Ghosh (2011) independence model, (b) the SHEL model versus Chaudhuri and Ghosh (2011) DP model, (c) the SHEL model versus the parametric model. The square represents Kansas City, MO and the triangle represents St. Louis, MO.

Table 2
Posterior medians and 95% (central) credible intervals for the SIDS example (Section 4.2).

Model	β_0	β_1	τ	MSPE
Parametric	−1.071	1.899	1.102	54.4
	(−1.441, −0.724)	(1.322,2.494)	(0.602, 2.050)	
SHEL	−0.971	1.723	0.289	12.0
	(−1.540, −0.404)	(0.794, 2.659)	(0.142, 0.635)	

process over 13 regions, where the correlation between these spatial regions was built on Euclidean distances. In this analysis, Anson county was considered its own region and included in the model. Since our goal is to demonstrate the robustness of the SHEL model (in terms of MSPE) to extreme outliers, we compare our approach to Cressie and Chan (1989), as they removed the outlier from their analysis. That is, we propose a relatively simple way to handle this outlier: depart from the Poisson distribution, and use EL methods to obtain estimates for $\{\boldsymbol{\beta}, \mathbf{y}\}|\mathbf{Z}$. We propose estimating equations based on the identities $\theta_i = V(\theta_i) = \exp(\mathbf{x}_i'\boldsymbol{\beta} + y_i)$, linking the SHEL model to the Poisson distribution.

We compare the SHEL methodology to the overdispersed Poisson suggested above, but with all of the locations considered in the analysis. For both the SHEL and parametric models, $\mathbf{y} = (y_1, \ldots, y_n)'$ is modeled according to the basis of Hughes and Haran (2013), using the eigenvectors associated with all of the positive eigenvalues of $\mathbf{P}_c\mathbf{B}\mathbf{P}_c$. The fixed effects parameters $\boldsymbol{\beta}$ are given a MVN$(\mathbf{0}, 100^2\mathbf{I}_2)$ prior, and τ is given an Unif(0.01, 100) prior, both of which are intentionally vague. The parametric model uses identical prior specifications.

Parameter posterior summaries and the results of a leave-one-out MSPE experiment can be found in Table 2. The results show similar medians for the posteriors of the parameters, but the credible interval for τ (the spatial precision parameter) in the overdispersed Poisson is much larger than the SHEL model. Additionally, the leave-one-out MSPE for the SHEL model is 12.0, approximately 78% lower than the MSPE of 54.4 for the parametric model. Additionally, not only does the parametric model poorly estimate Anson county, but it also poorly estimates the counties adjacent to it. The SHEL model is clearly more accurate in out-of-sample prediction. This is a case of SHEL methodology fitting the data much better when a parametric model appears to be suggested by the data. The results for both models are based on 10,000 MCMC iterations after 1000 iterations of burn-in. Convergence was assessed through visual inspection of the sample chains, with no deviations from

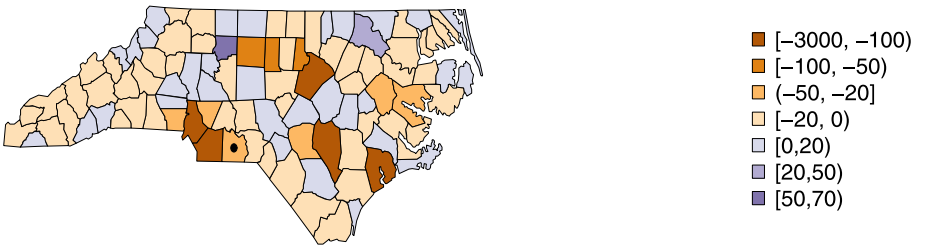

Fig. 2. The difference of the squared deviations $(Y_i - \widehat{Y}_{(-i)})^2$ for each location of the SHEL model versus the parametric model for the SIDS dataset. The circle indicates Anson county.

Table 3
Posterior medians and 95% (central) credible intervals for the North American Breeding Bird Survey example (Section 4.3).

Model	β_0	σ_u^2	ϕ	MSPE
Parametric	3.322	0.377	0.587	228.5
	(3.261 3.384)	(0.157, 1.272)	(0.028, 2.937)	
SHEL	3.390	0.230	1.580	195.4
	(3.277, 3.503)	(0.042, 0.751)	(0.082, 3.868)	

convergence detected. The results are displayed in Fig. 2, and demonstrate that the SHEL model provides superior estimates in terms of MSPE in the majority of locations, but especially in Anson county and the surrounding region.

4.3. North American breeding birds survey

To illustrate an example with continuous spatial reference and non-Gaussian data we utilize the example described in Wikle (2010). Counts of mourning doves in and near Missouri in 2007 from the North American Breeding Bird Survey represent a highly overdispersed spatially point referenced count dataset (mean = 30.8, variance = 221.7). Counts are collected on 44 sampling routes containing 50 stops each. All routes are 39.2 km in length and each count is assigned to the centroid of the route (see Robbins et al., 1986, for a comprehensive description). These data have been previously analyzed using a generalized linear mixed model (GLMM) framework by considering an overdispersed Poisson outcome, with the amount of overdispersion dictated by a latent Gaussian spatial process (Wikle, 2010). To assess the performance of the SHEL model relative to a parametric specification, we use the following model for comparison

$$\begin{aligned} &Z(s_i)|\lambda(s_i) \sim ind\,\text{Poisson}(\lambda(s_i)), \quad i = 1, \ldots, n; \\ &\log\{\lambda(s_i)\} = \beta + y(s_i). \end{aligned} \tag{9}$$

We modeled $\mathbf{y} = (y_1(s_1), \ldots, y_n(s_n))'$ as multivariate Gaussian with mean zero and covariance function $\sigma_y^2 r(s_i, s_j; \phi)$, where $r(s_i, s_j; \phi) = \exp(-\|s_i - s_j\|/\phi)$. We placed a $N(0, 100^2)$ prior on β, a Unif(0.01, 100) prior on σ_y^2 and, similar to Wikle (2010), a Unif(0, 4) prior on ϕ.

The estimating equations for the SHEL model are based on the identities $\theta_i = V(\theta_i) = \exp\{\beta + y(s_i)\}$ in Eq. (3). This is a SHEL specification based on an overdispersed Poisson model, where we have the conditional mean, θ_i, and variance of $Z_i|\theta_i$ equal. The results for both models are based on 10,000 MCMC iterations after 1000 iterations of burn-in, again convergence was assessed through visual inspection of the sample chains with no deviation from convergence detected.

Results from both the parametric and SHEL models can be found in Table 3. There are two main differences in the model outcomes. The point estimates of the SHEL model indicate lower spatial variance as well as increasing spatial decay as compared to the parametric model. This would argue that the SHEL analysis detects less spatial structure than the parametric method. It is worth noting that this is likely due to the flexibility of the empirical data model, which serves to account for some of the spatial structure of the data. Secondly, we again see an improved predictive ability of the SHEL framework for these data, as indicated by the leave-one-out MSPEs. This decrease is noteworthy, with the leave-one-out MSPE for the SHEL model being 195.4, nearly a 15% reduction over the 228.5 for parametric model. The results are displayed in Fig. 3 and again demonstrate superior prediction in terms of MSPE in the majority of locations analyzed.

5. Discussion

In this paper, we have proposed a general framework for including empirical data models in the BHM framework. We have shown that the SHEL model can explicitly accommodate spatial correlation on irregular lattices as well as handle spatial point-referenced data not collected on a regular grid, both of which are novel models. In order for the models we propose to be useful in practice, we have provided detailed discussion regarding sampling and computational considerations. Additionally, the Appendix contains two brief simulation studies exploring prediction in the geostatistical and lattice case.

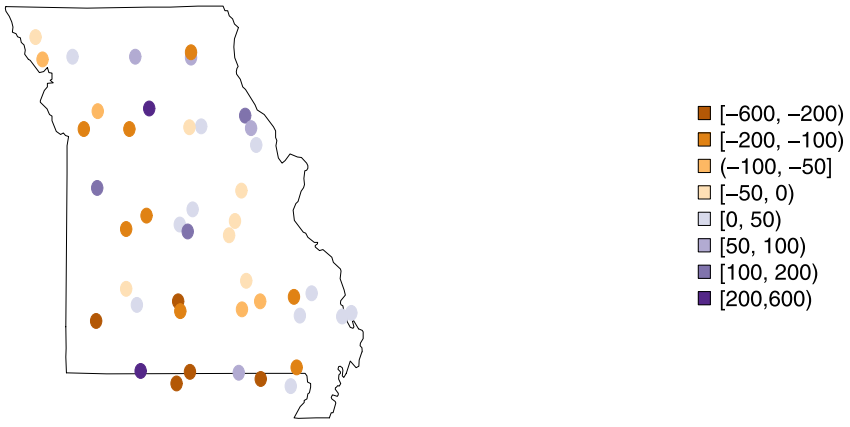

Fig. 3. The difference of the squared deviations $(Y_i - \widehat{Y}_{(-i)})^2$ for each location of the SHEL model and the parametric model for the North American Breeding Birds Survey example.

Both simulation studies demonstrate improved predictive performance over similar parametric models, and corroborate the results of Sections 4.1 and 4.3.

Importantly, we have shown that the SHEL framework outperforms standard parametric analyses in three distinct and unrelated case studies. In every case, the SHEL model has outperformed parametric models in terms of out of sample prediction as measured by reduction in MSPE of at least 15%. In the case of the SIDS data and the ACS data, we have outperformed a standard analyses by a reduction of 30% in terms of MSPE. While the SHEL paradigm can certainly be used for inference, EL methods are known to produce asymptotic credible intervals that slightly undercover the true parameter values in the mean structure of multiple regression models (Fang and Mukerjee, 2006). Therefore, one should take care when interpreting the credible intervals produced by such methods.

The SHEL model overcomes one of the main difficulties in standard EL analysis, which is handling dependence in the outcomes. That is, the SHEL model places the dependence structure at the process and parameter stages of the hierarchy. This makes the framework extremely advantageous for a wide range of problems where parametric modeling assumptions may be difficult to verify. Accordingly, the SHEL model provides a unified BHM framework that is capable of handling a broad range of dependence structures, including spatial dependence, as illustrated here. In short, by casting the SHEL model within the BHM paradigm we provide an extremely flexible approach that takes advantage of conditional thinking and is, therefore, capable of effectively modeling parameters. In addition, as a byproduct of the BHM specification, we are easily able to incorporate relevant scientific information, while providing a quantification of uncertainty of our predictions.

Acknowledgments

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. We thank an anonymous referee for providing valuable comments that have helped strengthen this manuscript.

Appendix A. Proof of Theorem 1

Let $\mathcal{C}(\mathbf{A})$ denote the column space of a matrix A and $\mathcal{N}(\mathbf{A})$ represent the null space. Assume that $\mathbf{X}$ contains a column equal to the one vector, which implies that the model contains an intercept. We proceed by contradiction. First, suppose there exists $\mathbf{v} \neq \mathbf{0}$ such that

$$\mathbf{v}'\mathbf{M}'(\mathbf{B}_+ - \mathbf{B})\mathbf{M}\mathbf{v} = 0.$$

Let $\mathbf{P}\boldsymbol{\Lambda}\mathbf{P}'$ represent the eigenspace decomposition of $(\mathbf{B}_+ - \mathbf{B})$. Then we have

$$\mathbf{v}'\mathbf{M}'\mathbf{P}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{P}'\mathbf{M}\mathbf{v} = 0,$$

for some $\mathbf{v} \neq \mathbf{0}$. This implies

$$\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{P}'\mathbf{M}\mathbf{v} = 0 \tag{A.1}$$

for this choice of $\mathbf{v}$. Now, we know that, for the ICAR specification we have chosen,

$$\begin{aligned} 1 &= \text{nullity}(\mathbf{B}_+ - \mathbf{B}) \\ &= \text{nullity}(\mathbf{P}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{P}') \\ &\geq \text{nullity}(\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{P}'). \end{aligned} \tag{A.2}$$

Note that

$$\begin{aligned}&\mathbf{1}'(\mathbf{B}_+ - \mathbf{B})\mathbf{1} = 0\\ &\Rightarrow \mathbf{1}'(\mathbf{P}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{P}')\mathbf{1} = 0\\ &\Rightarrow \boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{P}'\mathbf{1} = 0. \end{aligned} \tag{A.3}$$

Together with (A.2), (A.3) implies $\mathcal{N}(\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{P}') = \{\mathbf{0}, \mathbf{1}\}$, as nullity$(\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{P}') \leq 1$ and we have demonstrated that $\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{P}'\mathbf{1} = \mathbf{0}$. $\mathbf{M}$ is full rank by construction; so, $\mathbf{Mv} \neq \mathbf{0}$ for $\mathbf{v} \neq \mathbf{0}$. So, if $\mathbf{v}'\mathbf{M}'(\mathbf{B}_+ - \mathbf{B})\mathbf{Mv} = 0$, which in turn implies $\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{P}'\mathbf{Mv} = 0$, we must have that $\mathbf{1} \in \mathcal{C}(\mathbf{M})$ if $\mathbf{v} \neq 0$. However, $\mathcal{C}(\mathbf{M}) \perp \mathcal{C}(\mathbf{X})$ and $\mathbf{1} \in \mathcal{C}(\mathbf{X})$, which is a contradiction. Therefore, $\mathbf{v}'\mathbf{M}'(\mathbf{B}_+ - \mathbf{B})\mathbf{Mv} = 0$ implies $\mathbf{v} = \mathbf{0}$, and we have that $\mathbf{M}'(\mathbf{B}_+ - \mathbf{B})\mathbf{M}$ is positive definite.

Appendix B. Simulations

Of particular interest is the performance of the SHEL paradigm in spatial prediction, and so we conduct a simulation study to assess the predictive performance of the SHEL framework as compared to parametric models.

B.1. Study 1: The SHEL Fay–Herriot model

In the simulation study presented here, we compare the prediction of the SHEL FH model and the independence model of Chaudhuri and Ghosh (2011) for data that behave similar to those of our data analysis in the FH analysis of Section 4.1. We do not utilize their DP prior model due to concerns of computational considerations associated with repeated estimation within a full simulation study and the fact that the DP process model performs similar to the independence model in the analysis of Chaudhuri and Ghosh (2011). To simulate data, random effects y_i are generated based on a Hughes and Haran (2013) lattice prior with a precision parameter equal to the posterior mean of τ in the analysis of Section 4.1. Then data model weights $\{w_i\}$ are generated based on the posterior means of the fixed effects parameters of that analysis. This gives an EL to generate data that will have similar properties to the data in Section 4.1. Prior specifications are identical to that analysis for both models.

We generate 125 datasets in this way and perform a leave-one-out MSPE analysis on each dataset. For each location within a given dataset, the model is run for 11,000 iterations, with 1000 iteration discarded as burn-in (i.e., 10,000 used for our analysis). To assess convergence, we visually inspect a random subset of sample chains from the 125×115 analyses and note that no lack of convergence was detected.

Over all 125 simulated datasets, the SHEL FH model provides an average MSPE of 0.163, while the independence model of Chaudhuri and Ghosh (2011) provides an average MSPE of 0.239. This represents a 31.6% average reduction in MSPE. Notably, we see similar results in terms of MSPE reduction in Section 4.1, and the results corroborate one another.

B.2. Study 2: breeding birds

In the simulation study presented here, we compare the prediction of the SHEL model to the parametric model for data that behave similar to those of our data analysis in Section 4.3. We assess the predictive performance of the model by means of a leave-one-out mean squared prediction error MSPE experiment. In order to generate data that have similar properties to the Dove data, we first analyzed the data according to the SHEL model we propose. Next, we computed the posterior means, $\widehat{\beta}$ and $\widehat{\mathbf{y}}$, from this analysis. These values were then used to compute average weights $\{w_i\}$ which correspond to an EL based on the posterior parameter means. These weights were then used, in turn, to generate new data. New random effects were generated from the spatial prior used in the analysis with σ_y^2 and ϕ set at their respective mean posterior values in the analysis in Section 4.3. Prior specifications are identical to that analysis for both models.

We generated 250 datasets in this way and performed a leave-one-out MSPE experiment in which we analyze each dataset 44 times, each time with a different location left out of the analysis. For each dataset, each analysis was run for 11,000 iterations, with 1000 iterations for burn-in, resulting in 10,000 MCMC iterations which were used for analysis. We visually inspected all 47 sample chains (44 random effects and 3 parameters) for 10 random analyses and found no evidence of non-convergence.

The SHEL model yields a MSPE of 331.4 when averaged across all 250 simulations, while the previously proposed Poisson model of Wikle (2010) yields a MSPE of 400.0. This constitutes a 24.7% average MSPE reduction and strongly indicates that the SHEL model provides superior performance in this context.

References

Bandyopadhyay, S., Lahiri, S.N., Nordman, D., 2014. Frequency domain empirical likelihood method for irregularly spaced spatial data. Ann. Statist. 43, 519–545.
Berliner, L.M., 1996. Hierarchical Bayesian time series models. In: Maximum Entropy and Bayesian Methods. Springer, pp. 15–22.
Besag, J., York, J., Mollié, A., 1991. Bayesian image restoration with two applications in spatial statistics (with discussion). Ann. Inst. Statist. Math. 43, 1–59.

Chang, I., Mukerjee, R., 2008. Bayesian and frequentist confidence intervals arising from empirical-type likelihoods. Biometrika 95 (1), 139–147.
Chaudhuri, S., Ghosh, M., 2011. Empirical likelihood for small area estimation. Biometrika 98 (2), 473–480.
Chaussé, P., 2010. Computing generalized method of moments and generalized empirical likelihood with R. J. Stat. Softw. 34 (11), 1–35.
Chen, J., Sitter, R., Wu, C., 2002. Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. Biometrika 89 (1), 230–237.
Cressie, N., Chan, N.H., 1989. Spatial modeling of regional variables. J. Amer. Statist. Assoc. 84 (406), 393–401.
Cressie, N., Wikle, C.K., 2011. Statistics for Spatio-Temporal Data. John Wiley and Sons, Hoboken, NJ.
Fang, K., Mukerjee, R., 2006. Empirical-type likelihoods allowing posterior credible sets with frequentist validity: Higher-order asymptotics. Biometrika 93 (3), 723–733.
Fay, R., Herriot, R., 1979. Estimates of income for small places: an application of James-Stein procedures to census data. J. Amer. Statist. Assoc. 74, 269–277.
Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2013. Bayesian Data Analysis, third ed. CRC Press, Boca Raton, FL.
Getis, A., Ord, J., 1992. The analysis of spatial association by use of distance statistics. Geograph. Anal. 23 (3), 190–205.
Hughes, J., Haran, M., 2013. Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. J. R. Stat. Soc. Ser. B Stat. Methodol. 75 (1), 139–159.
Kaiser, M.S., Nordman, D.J., 2012, Blockwise empirical likelihood for spatial Markov model assessment. Unpublished manuscript http://streaming.stat.iastate.edu/~stat506/papers/SBEL.pdf.
Kitamura, Y., 1997. Empirical likelihood methods with weakly dependent processes. Ann. Statist. 25 (5), 2084–2102.
Kolaczyk, E.D., 1994. Empirical likelihood for generalized linear models. Statist. Sinica 4, 199–218.
Kulldorf, M., 1997. A spatial scan statistic. Comm. Statist. Theory Methods 26 (6), 1481–1496.
Lazar, N., 2003. Bayesian empirical likelihood. Biometrika 90 (2), 319–326.
Monahan, J., Boos, D., 1992. Proper likelihoods for Bayesian analysis. Biometrika 79 (2), 271–278.
Newey, W., Smith, R., 2004. Higher order properties of GMM and generalized empirical likelihood estimators. Econometrica 72 (1), 219–255.
Nordman, D., 2008. An empirical likelihood method for spatial regression. Metrika 68 (3), 351–363.
Nordman, D.J., Caragea, P.C., 2008. Point and interval estimation of variogram models using spatial empirical likelihood. J. Amer. Statist. Assoc. 103 (481), 350–361.
Owen, A., 1988. Empirical likelihood ratio confidence intervals for a single functional. Biometrika 75 (2), 237–249.
Owen, A.B., 2001. Empirical Likelihood. Chapman and Hall/CRC, Boca Raton, FL.
Qin, J., Lawless, J., 1994. Empirical likelihood and general estimating equations. Ann. Statist. 22 (1), 300–325.
R Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Reich, B., Hodges, J., Zadnik, V., 2006. Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. Biometrics 62 (4), 1197–1206.
Robbins, C., Bystrak, D., Geissler, P., 1986. The Breeding Birds Survey: Its First Fifteen Years, 1965–1979, vol. 157. USDOI, Fish and Wildlife Resource Publication, Washington, D.C.
Rue, H., Held, L., 2005. Gaussian Markov Random Fields: Theory and Applications. Chapman & Hall/CRC, Boca Raton, FL.
Schennach, S., 2005. Bayesian exponentially tilted empirical likelihood. Biometrika 92 (1), 31–46.
Schennach, S., 2007. Point estimation with exponentially tilted empirical likelihood. Ann. Statist. 35 (2), 634–672.
Sengupta, A., Cressie, N., 2013. Empirical hierarchical modelling for count data using the spatial random effects model. Spatial Econom. Anal. 8 (3), 389–418.
Smith, R., 1997. Alternative semi-parametric likelihood approaches to generalised method of moments estimation. Econom. J. 107 (441), 503–519.
Symons, M.J., Grimson, R.C., Yuan, Y.C., 1983. Clustering of rare events. Biometrics 39 (1), 193–205.
Wall, M., 2004. A close look at the spatial structure implied by the CAR and SAR models. J. Statist. Plann. Inference 121 (2), 311–324.
Wikle, C.K., 2003. Hierarchical Bayesian models for predicting the spread of ecological processes. Ecology 84 (6), 1382–1394.
Wikle, C.K., 2010. Hierarchical modeling with spatial data. In: Gelfand, A., Diggle, P.J., Guttorp, P., Fuentes., M. (Eds.), Handbook of Spatial Statistics. CRC Press, Boca Raton, FL.
Wikle, C.K., Berliner, L.M., 2005. Combining information across spatial scales. Techonmetrics 47, 80–91.
Zellner, A., 1986. Bayesian estimation and prediction using asymmetric loss functions. J. Amer. Statist. Assoc. 81 (394), 446–451.