Parametric and Nonparametric Bayesian Models for Ecological Inference in 2×2 Tables[*]

Kosuke Imai[†]
Department of Politics, Princeton University

Ying Lu[‡]
Office of Population Research, Princeton University

November 18, 2004

Abstract

The ecological inference problem arises when making inferences about individual behavior from aggregate data. Such a situation is frequently encountered in the social sciences and epidemiology. In this article, we propose a Bayesian approach based on data augmentation. We formulate ecological inference in 2×2 tables as a missing data problem where only the weighted average of two unknown variables is observed. This framework directly incorporates the deterministic bounds, which contain all information available from the data, and allow researchers to incorporate the individual-level data whenever available. Within this general framework, we first develop a parametric model. We show that through the use of an *EM* algorithm, the model can formally quantify the effect of missing information on parameter estimation. This is an important diagnostic for evaluating the degree of aggregation effects. Next, we introduce a nonparametric model using a Dirichlet process prior to relax the distributional assumption of the parametric model. Through simulations and an empirical application, we evaluate the relative performance of our models in various situations. We demonstrate that in typical ecological inference problems, the fraction of missing information often exceeds 50 percent. We also find that the nonparametric model generally outperforms the parametric model, although the latter gives reasonable in-sample predictions when the bounds are informative. C-code, along with an R interface, is publicly available for implementing our Markov chain Monte Carlo algorithms to fit the proposed models.

Key Words: Aggregate data, *EM* algorithm, Missing information principle, Data augmentation, Political science, Racial voting.

[*] We thank Wendy Tam Cho, Gary King, Xiao-li Meng, Kevin Quinn, Phil Shively, David van Dyk, Jon Wakefield, and participants of the Office of Population Research seminar at Princeton University for helpful comments. We also thank Jianqing Fan for his advice and support. An earlier version of this article was presented at the 2004 Joint Statistical Meetings, the Second Cape Cod Monte Carlo Workshop, and the 2004 Annual Political Methodology Summer Meeting.

[†] Assistant Professor, Department of Politics, Princeton University, Princeton NJ 08544. Phone: 609–258–6610, Fax: 973–556–1929, Email: kimai@Princeton.Edu, URL: www.princeton.edu/˜ kimai

[‡] Ph.D. candidate, Office of Population Research, Woodrow Wilson School of Public and International Affairs, Princeton University, Princeton NJ 08544. Phone: 609–258–5508, Email: yinglu@Princeton.Edu

1 Introduction

The ecological inference problem arises when making inferences about individual behavior from aggregate data. Such a situation is frequently encountered in the social sciences and epidemiology (e.g., Greenland and Robins, 1994; Achen and Shively, 1995). Although the ecological regression suggested by Goodman (1953) and an alternative model developed by King (1997) have been widely used in practice, a number of other approaches have been recently proposed (e.g., King, Rosen, and Tanner, 1999; Wakefield, 2004a). While some highlight the limitations of ecological inference (e.g., Gelman *et al.*, 2001), progress has also been made, and there now exist a growing number of new statistical techniques based on a variety of assumptions (e.g., King, Rosen, and Tanner, 2004).

In this article, we propose an approach based on data augmentation (Tanner and Wong, 1987). We formulate ecological inference in 2×2 tables as a missing data problem where only the weighted average of two unknown variables is observed. This framework directly incorporates the deterministic bounds, which contain all information available from the data, and allow researchers to use the individual-level data whenever available. Many have shown that incorporating such auxiliary information is essential for reliable ecological inference in some situations (e.g., Wakefield, 2004a).

Within this general framework, we first develop a parametric model. We show that through the use of an EM algorithm and its extension (Dempster, Laird, and Rubin, 1977; Meng and Rubin, 1991), the model can formally quantify the effect of missing information due to aggregation. This formal evaluation of aggregation effects on parameter estimation is an important contribution to the literature because the previous works rely solely on informal, graphical diagnostics (e.g., King, 1997; Gelman *et al.*, 2001; Wakefield, 2004a). Our method can also suggest the degree to which one's ecological inference relies on the parametric assumptions, an essential consideration for empirical researchers in light of the debates about the appropriateness of particular parametric assumptions (e.g., Freedman *et al.*, 1998; King, 1999; Cho and Gaines, 2004).

Second, using this parametric model as a base model, we develop a nonparametric Bayesian model using a Dirichlet process prior (Ferguson, 1973) in order to relax the distributional assumption of the parametric model. One common feature of many existing models is that they make parametric assumptions. For example, in his exchange with Freedman *et al.* (1998), King (1999) concludes that "open issues . . . include . . . flexible distributional and functional form specifications" (p.354). We take up this challenge by relaxing the distributional assumption of our parametric model, and examine the relative advantages of the nonparametric model through simulation studies and an empirical example.

Our method deals with ecological inference in 2×2 tables and is motivated by the racial voting example, which is of particular importance in the social sciences. Every ten years, congressional districts are redrawn based on the most recent census counts, providing politicians with the opportunity of gerrymandering and yielding many litigations by Democratic and Republican parties. The rulings of these court cases typically rely on the empirical estimates of racial voting behavior under various redistricting plans that are provided by expert witnesses. Given their importance in court rooms, the statistical procedures that are used to produce these estimates are often disputed (e.g., Freedman *et al.*, 1991; Grofman, 1991; Lichtman, 1991; Freedman *et al.*, 1998; King, 1999).

To understand a typical racial voting problem, suppose that we observe the number of registered white and black voters for each geographical unit (e.g., a county). The election results reveal the total number of votes for all geographical units. Given this information, we wish to infer the number of black and white voters who turned out. Table 1 presents a 2×2 ecological table of the racial voting example where counts are transformed into proportions. In typical racial voting examples, the number of voters within each geographical unit is very large. Hence, many previous methods directly modeled proportions rather than treating them as parameters (e.g., Goodman, 1953; Freedman *et al.*, 1991; King, 1997). We follow this practice in our article, hence our proposed method may not be applicable to the situations where the number of voters is small. In contrast, King *et al.* (1999) and Wakefield (2004a) model counts although Wakefield (2004a) suggests the use of normal approximation based on proportions when the counts are large (see also Brown and Payne, 1986).

For every geographical unit $i = 1, \ldots, n$, such a 2×2 ecological table is available. Given the total turnout rate Y_i and the proportion of black voters X_i, one seeks to infer the proportions of black and white voters who turned out, W_{1i} and W_{2i} respectively. We may also be interested in

	black voters	white voters	
Voted	W_{1i}	W_{2i}	Y_i
Not Voted	$1 - W_{1i}$	$1 - W_{2i}$	$1 - Y_i$
	X_i	$1 - X_i$	

Table 1: 2×2 Ecological Table for the Racial Voting Example. X_i, Y_i, W_{1i}, and W_{2i} are proportions, and hence lie between 0 and 1. The unit of observation is typically a geographical unit and is denoted by i.

related quantities such as the average of these probabilities, weighted by the total number of black or white voters in each geographical unit.

Typically, the primary goal of ecological inference is to obtain *in-sample* and *out-of-sample* predictions of these quantities. In-sample predictions are of interest when researchers simply wish to predict the missing inner cell proportions of ecological tables in a given sample, thereby limiting inferences to voting behavior of blacks and whites in a particular election. In contrast, out-of-sample predictions enable inferences about the underlying population from which the sample is assumed to be randomly drawn. This distinction between the in-sample and population inferences has been often neglected.

The rest of the article is organized as follows. In Section 2, we briefly review the approaches that have been previously proposed in the literature. In Section 3, we introduce our parametric and nonparametric methods. In Section 4, we evaluate the performance of our models through a variety of simulations. Section 5 presents an empirical application of the voter registration data in four U.S. southern states. Finally, Section 6 gives concluding remarks.

2 Previous Approaches

In this section, we briefly review some of the previously proposed approaches to motivate our method. For a more comprehensive survey of the literature, readers may wish to consult Cleave, Brown, and Payne (1995), King (1997), and Wakefield (2004a), among others.

2.1 Method of Bounds

Suppose that in a simple random sample of size n from a population, we observe the margins of Table 1 for each county i. The method of bounds is based on the following deterministic relationship,

$$Y_i = W_{1i} X_i + W_{2i}(1 - X_i), \quad \text{for} \quad i = 1, 2, \ldots, n \tag{1}$$

where $X_i, Y_i, W_{1i}, W_{2i} \in [0, 1]$. When Y_i is equal to either 0 or 1, W_{1i} and W_{2i} are completely known. If $X_i = 1$, then $W_{1i} = Y_i$ but W_{2i} does not exist. Similarly, if $X_i = 0$, then $W_{2i} = Y_i$ but W_{1i} does not exist. King (1997) called equation 1 a tomography line. For every i, this tomography line defines a *deterministic* relationship between the missing data, $W_i = (W_{1i}, W_{2i})$ and the observed data, (Y_i, X_i). Duncan and Davis (1953) first recognized that with equation 1, one can narrow the original bound of $[0, 1]$ for W_i to the following intervals,

$$W_{1i} \in \left[\max\left(0, \frac{X_i + Y_i - 1}{X_i}\right), \ \min\left(1, \frac{Y_i}{X_i}\right)\right], \tag{2}$$

$$W_{2i} \in \left[\max\left(0, \frac{Y_i - X_i}{1 - X_i}\right), \ \min\left(1, \frac{Y_i}{1 - X_i}\right)\right]. \tag{3}$$

Given these bounds for each i (e.g., a county), the analysis of larger units (e.g, a state) can be carried out by simply aggregating the upper and lower bounds with appropriate weights; $N_i X_i$ and $N_i(1 - X_i)$ for W_{1i} and W_{2i}, respectively, where N_i is the total number of voters in county i. When the resulting bounds are sufficiently narrow, researchers can make reasonable in-sample inferences.

Although applied researchers often find the bounds too wide for their purposes, the method of bounds shows the identifying power of the data without any statistical assumption. That is, equation 1 implies the exact degree to which the data are informative about W_i. For this reason, statistical analysis that does not incorporate this deterministic relationship is likely to be sensitive to modeling assumptions (King, 1997).

2.2 The Ecological Regression and Related Methods

Goodman (1953, 1959) proposes the use of a linear regression to model *population* means of (W_{1i}, W_{2i}) (see also Achen and Shively, 1995; Gelman *et al.*, 2001). This ecological regression assumes that the population average turnout for each racial group is fixed and does not vary from one county to another. Namely, $E(W_{1i} \mid X_i) = E(W_{1i})$ and $E(W_{2i} \mid X_i) = E(W_{2i})$ for all i. With this constancy assumption, taking the conditional expectation of both sides of equation 1 and rearranging the right hand side yield $E(Y_i \mid X_i) = E(W_{2i}) + E(W_{1i} - W_{2i})\, X_i$. If we further assume that the error term $\epsilon_i = Y_i - E(Y_i \mid X_i)$ is uncorrelated with X_i, the unbiased estimates of $E(W_{ji})$ can be obtained from the following least squares regression,

$$Y_i \;=\; \alpha + \beta X_i + \epsilon_i. \tag{4}$$

Although the ecological regression estimates the marginal means, it does not estimate other features of the population distribution. Furthermore, the ecological regression often fails to provide reasonable in-sample inferences. Typically, the estimated population means are used as the in-sample predictions of W_{1i} and W_{2i}. The problem is, however, that these "in-sample" predictions are not consistent with the deterministic bounds of equations 2 and 3.

Freedman *et al.* (1991) introduce a related method, called the neighborhood model, that does not rely upon the assumption of homogeneity within a racial group. The neighborhood model makes an alternative assumption that within each county both black and white voters behave in the same way, i.e., $W_{1i} = W_{2i} = \gamma_i$ and hence $Y_i = \gamma_i$ for all i. In the context of racial voting, however, some argue that this assumption is unrealistic (e.g., Grofman, 1991). The linear neighborhood model estimates γ_i by first fitting the ecological regression of equation 4 and then using $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$ as the estimate of γ_i. The nonlinear neighborhood model uses the observed values of Y_i directly as the estimate of γ_i. Although these in-sample predictions are consistent with the bounds, the neighborhood models do not allow for population inferences.

2.3 Recent Advances

Recently, a number of parametric models for ecological inference have been proposed. In contrast to the ecological regression, these models specify a parametric distribution for the missing data W_i and therefore yield the in-sample predictions of W_i as well as the population-level estimates. King (1997) proposes a model where W_i is assumed to follow a truncated bivariate normal distribution.

The main contribution of his model is to incorporate the bounds into statistical estimation. Similarly, Wakefield (2004a) suggests a binomial convolution model that respects the bounds of W_i. King *et al.* (1999) propose a hierarchical Bayesian model that adds greater flexibility to distributional assumptions, but only incorporates the bounds in expectation. Its in-sample predictions are therefore inconsistent with the bounds.

3 Method

The review of previous approaches in Section 2 reveals our modeling tasks. First, a model must directly incorporate bounds. Second, strong distributional assumptions should be avoided. In this section, we propose the parametric and nonparametric methods to address these issues. Our method is based on data augmentation (Tanner and Wong, 1987). We formulate ecological inference as a missing data problem where only the weighted average of two unknown variables is observed (see Wakefield, 2004a, for a related approach).

3.1 A Parametric Model

We first present our parametric model and show that it can incorporate additional information whenever available. A similar parametric model has appeared in the literature (King, 1997; Wakefield, 2004a).

We first take the logit transformation of W_i; $W^*_{ji} = \text{logit}(W_{ji})$, $j = 1, 2$, where $\text{logit}(t) = \log\{t/(1-t)\}$. The deterministic relationship of equation 1 implies,

$$Y_i \quad = \quad \text{logit}^{-1}(W^*_{1i})\, X_i + \text{logit}^{-1}(W^*_{2i})\,(1 - X_i).$$

Next, we assume that $W^*_i = (W^*_{1i}, W^*_{2i})$ follows a bivariate normal distribution,

$$W^*_i \mid \mu, \Sigma \quad \sim \quad \mathcal{N}(\mu, \Sigma), \tag{5}$$

where μ is a (2×1) vector of the population means and Σ is the (2×2) positive definite population covariance matrix. Similar to the model of King (1997), this model allows W_1 and W_2 to be correlated with each other (through their logit transformations). In the racial voting example, the turnout rates of black and white voters in each county may be correlated with one another.

It is possible to obtain the maximum likelihood (ML) estimate of μ and Σ via *EM* algorithm Dempster *et al.* (1977). In this case, the E-step requires the numerical calculation of one-dimensional integrals (e.g., Lange, 1999, chap. 16), which is done by the trapezoidal approximation. Alternatively, one can formulate a Bayesian model by placing the following conjugate prior distribution on (μ, Σ),

$$\mu \mid \Sigma \ \sim \ \mathcal{N}(\mu_0, \Sigma/\tau_0^2), \quad \text{and} \quad \Sigma \sim \text{InvWish}\,(\nu_0,\ S_0^{-1}), \tag{6}$$

where μ_0 is a (2×1) vector of the prior mean, τ_0 is a scalar, ν_0 is the prior degrees of freedom parameter, and S_0 is a (2×2) positive definite prior scale matrix. The Gibbs sampling algorithm described in Appendix A, is used to fit this Bayesian model.

An advantage of the parametric Bayesian model is that the posterior draws of (W_{1i}, W_{2i}) are readily available as a part of the Gibbs sampler and respect the deterministic relationship of equation 1. Indeed, the observed data (Y_i, X_i) enter the likelihood only through equation 1. We also emphasize that, like the ecological regression and many other existing models in the literature, our parametric model assumes no contextual effect. That is, W_1 and W_2 are independent of X.

3.2 Quantifying the Amount of Missing Information due to Aggregation

An important advantage of our parametric model is that it is possible to formally quantify the amount of information lost due to the aggregation process in ecological inference. This is useful because the large amount of missing information, relative to the observed information, implies that the resulting estimates may be driven by the parametric assumption of the model. In contrast, previous studies have used informal graphical methods to examine the informativeness of bounds (e.g., King, 1997; Gelman *et al.*, 2001; Cho and Gaines, 2004; Wakefield, 2004a).

We use the missing information principle of Orchard and Woodbury (1972): *observed information = complete information − missing information.* Formally, $I_o = I_{oc} - I_{om}$ where I_o represents the negative second derivative of the observed log-likelihood (or the observed Fisher information matrix), I_{oc} denotes the expected information matrix from the complete log-likelihood, and I_{om} can be viewed as the missing information. Dempster *et al.* (1977) show that the asymptotic variance-covariance matrix can be written as $I_o^{-1} = I_{oc}^{-1} + I_{oc}^{-1} DM(I - DM)^{-1}$ where DM is a Jacobian matrix associated with the rate of convergence of the EM algorithm. The diagonal elements of the difference between I_o^{-1} and I_{oc}^{-1} quantifies the increased asymptotic variance for each parameter due to the information lost through the aggregation process. Meng and Rubin (1991) introduce the Supplemented EM algorithm that can be used to compute the DM matrix. We use this algorithm to estimate the fraction of missing information for each model parameter, i.e., the observed data information divided by the complete data information. This gives a formal evaluation of aggregation effects on parameter estimation under our model. We illustrate our method using simulated datasets in Section 4.1 and a real dataset in Section 5.

3.3 A Nonparametric Model

Like other parametric models in the literature, the model introduced in Section 3.1 makes a specific distributional assumption. To relax this assumption, we apply a Dirichlet process prior and model the unknown population distribution as a mixture of bivariate normal distributions (Ferguson, 1973, 1974, 1983) (see Imai and King (2004) for an alternative approach based on the Bayesian model averaging). The resulting model is nonparametric in the sense that no distributional assumption is made, and its in-sample predictions respect the deterministic bounds. Although its applications were limited in the past, rapid development of MCMC algorithms made it feasible to employ a Dirichlet process prior for Bayesian density estimation (e.g., Escobar, 1994; Escobar and West, 1995), and other nonparametric and semiparametric problems (e.g., Mukhopadhyay and Gelfand, 1997; Dey, Müller, and Sinha, 1998).

The basic idea is to model the parameters, in our case (μ_i, Σ_i), with an unknown (random) distribution function G rather than a known (fixed) one such as the normal/inverse-Wishart distribution. We then place a prior distribution on G over all possible probability measures. Such a prior distribution is called a Dirichlet process prior and is denoted by $G \sim \mathcal{D}(G_0, \alpha)$, where $G_0(\cdot)$ is the known base prior distribution and is also the prior expectation of $G(\cdot)$; $E(G(\mu, \Sigma)) = G_0(\mu, \Sigma)$

for all (μ, Σ) in its parameter space. A positive scalar α is a concentration parameter. Ferguson (1973) established that given any measurable partition $(A_1, A_2, \ldots, A_k)$ on the support of G_0, the random vector of probabilities $(G(A_1), G(A_2), \ldots, G(A_k))$ follows a Dirichlet distribution with parameter $(\alpha G_0(A_1), \alpha G_0(A_2), \ldots, \alpha G_0(A_k))$. A large value of α suggests that G is likely to be close to G_0, and hence, to yield the results that are similar to those obtained from the parametric model with the prior distribution G_0. On the other hand, a small value of α implies that G is likely to place most of the probability mass on a few partitions. This setup allows the unknown distribution function G to be nonparametrically estimated from the data.

We model W^* with the unknown distribution that can be in general characterized by a normal mixture. To do this, we specify a Dirichlet process prior on the unknown distribution function of the population parameters (μ, Σ), using the same conjugate normal/inverse-Wishart prior distribution as the base prior distribution. Finally, we place a gamma prior on the concentration parameter α. Then, our nonparametric model is given by,

$$\begin{aligned} Y_i &= \text{logit}^{-1}(W^*_{1i})\, X_i + \text{logit}^{-1}(W^*_{2i})\,(1 - X_i), \\ W^*_i \mid \mu_i, \Sigma_i &\sim \mathcal{N}(\mu_i,\ \Sigma_i), \\ \mu_i, \Sigma_i \mid G &\sim G, \\ G \mid \alpha &\sim \mathcal{D}(G_0,\ \alpha), \\ \alpha &\sim \mathcal{G}(a_0,\ b_0), \end{aligned}$$

where under G_0, (μ_i, Σ_i) is distributed as

$$\mu_i \mid \Sigma_i \ \sim\ \mathcal{N}\left(\mu_0,\ \frac{\Sigma_i}{\tau_0^2}\right), \quad \text{and} \quad \Sigma_i \ \sim\ \text{InvWish}\,(\nu_0,\ S_0^{-1}).$$

Appendix B describes our Gibbs sampling algorithm that is used to fit the model.

To illustrate how our model relates to a normal mixture, we follow Ferguson (1973) and Escobar and West (1995) to compute the conditional prior, $p(\mu_i, \Sigma_i \mid \mu^{(i)}, \Sigma^{(i)}, \alpha)$ where $\mu^{(i)} = \{\mu_1, \ldots, \mu_{i-1}, \mu_{i+1}, \ldots, \mu_n\}$ and $\Sigma^{(i)} = \{\Sigma_1, \ldots, \Sigma_{i-1}, \Sigma_{i+1}, \ldots, \Sigma_n\}$.

$$\mu_i, \Sigma_i \mid \mu^{(i)}, \Sigma^{(i)}, \alpha \ \sim\ \alpha\, a_{n-1}\, G_0(\mu_i, \Sigma_i) \ +\ a_{n-1} \sum_{j=1, j\neq i}^{n} \delta_{(\mu_j, \Sigma_j)}\,(\mu_i, \Sigma_i) \quad \text{for} \quad i = 1, \ldots, n, \quad (7)$$

where $\delta_{(\mu_j,\Sigma_j)}(\mu_i,\Sigma_i)$ is a degenerate distribution whose entire probability mass is concentrated at $(\mu_i,\Sigma_i) = (\mu_j,\Sigma_j)$ and $a_{n-1} = 1/(\alpha+n-1)$. Equation 7 shows that given any $(n-1)$ values of (μ_i,Σ_i), there is a positive probability of coincident values, and that as α tends to ∞, the distribution approaches to G_0.

Similarly, a future replication draw of (μ_{n+1},Σ_{n+1}) given $\mu = \{\mu_1,\ldots,\mu_n\}$ and $\Sigma = \{\Sigma_1,\ldots,\Sigma_n\}$ has the following distribution,

$$\mu_{n+1},\Sigma_{n+1} \mid \mu,\Sigma,\alpha \quad \sim \quad \alpha\, a_n\, G_0(\mu_{n+1},\Sigma_{n+1}) \; + \; a_n \sum_{i=1}^{n} \delta_{(\mu_i,\Sigma_i)}\,(\mu_{n+1},\Sigma_{n+1}),$$

where $a_n = 1/(\alpha+n)$. We then compute the predictive distribution of a future observation W^*_{n+1} given (μ,Σ,α) which forms the basis of Bayesian density estimation. In particular, we evaluate $\int p(W^*_{n+1} \mid \mu_{n+1},\Sigma_{n+1},\alpha)d\,P(\mu_{n+1},\Sigma_{n+1} \mid \mu,\Sigma,\alpha)$, which yields,

$$W^*_{n+1} \mid \mu,\Sigma,\alpha \quad \sim \quad \alpha\, a_n\, \mathcal{T}_{\nu_0}\,(\mu_0,S) \; + \; a_n \sum_{i=1}^{n} \mathcal{N}\,(\mu_i,\Sigma_i), \tag{8}$$

where $\mathcal{T}_{\nu_0}(\mu_0,S)$ is a bivariate t distribution with ν_0 degrees of freedom, the location parameter μ_0 and the scale matrix $S = \tau_0^2\nu_0 S_0/(1+\tau_0^2)$. Equation 8 shows that when the value of α is small the predictive distribution is equivalent to a normal mixture. This setup resembles the standard kernel density estimator with a bivariate normal kernel. α plays a role similar to the bandwidth parameter, which controls the degree of smoothness.

Our nonparametric model, therefore, in principle can provide flexible estimation of bivariate density functions for ecological inference problems. However, because we do not directly observe W_{1i} and W_{2i}, the density estimation problem for ecological inference is much more difficult. Therefore, bounds must be sufficiently informative in order for the nonparametric model to be able to recover the underlying population distribution. We empirically investigate this issue through various simulations in Section 4.2 and a real data example in Section 5. Finally, although the nonparametric model relaxes the distributional assumptions of parametric models, it maintains the assumption of no contextual effects.

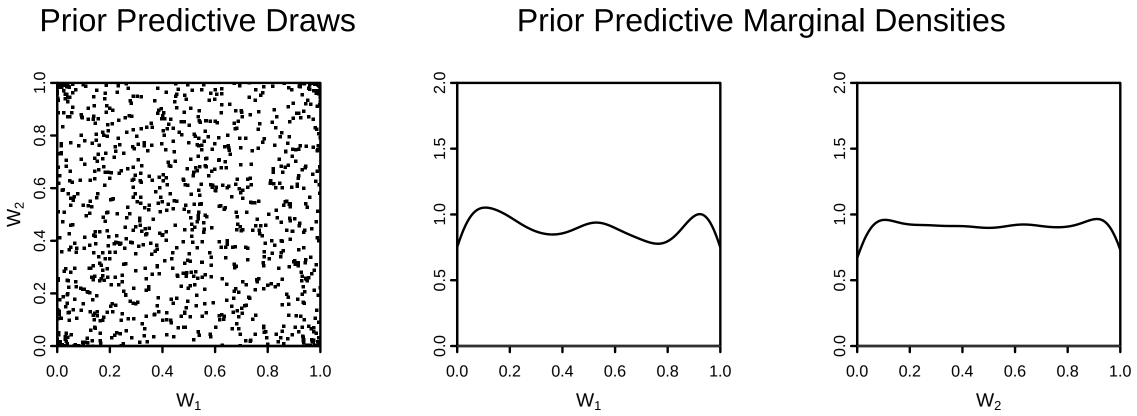
3.4 Incorporating Individual-level Data

When bounds are not informative, ecological inference is extremely difficult. The parametric inference will be sensitive to modeling assumptions, and the nonparametric model will not be able to recover the underlying distribution. Therefore, as emphasized by Wakefield (2004a), incorporating individual-level data may be helpful whenever such additional information is available. In our data augmentation approach, this is straightforward to accomplish. Suppose that for some counties we observe the true values or good estimates of W_i. For example, one might conduct a survey in randomly selected counties to obtain such information (One can also survey only one ethnic group). Sometimes, a small scale survey can be conducted to get rough estimates of W_i for some counties, and incorporating such auxiliary information can also be helpful (Wakefield, 2004a). For any of these cases, the Gibbs samplers described in Appendices can be modified slightly to incorporate the available information. In Sections 4 and 5, we investigate how additional individual-level data affect in-sample and population inferences.

3.5 Prior and Posterior Inferences

Our models require the specification of prior distributions. For the parametric model, $p(\mu, \Sigma)$ must be specified, whereas for the nonparametric model, the base prior G_0 as well as $p(\alpha)$ are required. Since we employ the conjugate normal/inverse-Wishart prior for both $p(\mu, \Sigma)$ and G_0, μ_0, S_0, τ_0, ν_0 need to be specified. In addition, the nonparametric model requires the specification of a_0 and b_0 if we choose a gamma prior distribution $\mathcal{G}(a_0, b_0)$ for α.

When strong prior information is available from previous studies or elsewhere, we specify these prior parameters so that the prior knowledge is properly approximated. When such information is not available, however, we consider a non-informative prior where the prior predictive distribution of (W_1, W_2) is approximately uniform. This leads to our choice of the prior parameters for the parametric model and the base prior of the nonparametric model; $\mu_0 = \mathbf{0}$, $S_0 = 8I_2$, $\tau_0 = 1$, and $\nu_0 = 4$. The left panel of Figure 1 shows the prior predictive draws of (W_1, W_2), and the middle and right panels show the marginal distributions of W_1 and W_2. The figure illustrates that the prior


Prior Predictive Draws

Prior Predictive Marginal Densities

Figure 1: Prior Predictive Distribution of W_1 and W_2. The prior distribution is an inverse logit transformation of normal/inverse-Wishart distribution with $\mu_0 = \mathbf{0}$, $S_0 = 8I_2$, $\tau_0 = 1$, and $\nu_0 = 4$. The three graphs are, from left to right, the scatter-plot of (W_1, W_2) prior predictive draws, the marginal density of W_1 and the marginal density of W_2.

predictive distribution of (W_1, W_2) is approximately uniform. For the nonparametric model, we use a diffuse prior, $\mathcal{G}(1, 0.1)$, with a mean of 10 and variance 100 for the concentration parameter, α. According to Antoniak (1974), the expected number of clusters given α and the sample size n is approximately $\alpha \log(1 + n/\alpha)$. With this choice of prior distribution for α, the prior expected number of clusters is approximately 27. Since the concentration parameter plays an important role in the density estimation with Dirichlet processes, a sensitivity analysis should be conducted in order to assess the influence of prior specification on posterior inferences.

The posterior inferences are based on the MCMC draws from the joint posterior distribution (see Appendices). For both the parametric and nonparametric models, the posterior draws of W_i are used to make in-sample predictions (or finite population inferences). These draws are readily available as a part of the Markov chain and respect the deterministic relationship of equation 1. To make out-of-sample predictions (or population inferences), we draw from the posterior *predictive* distribution of (W_1, W_2) without being subject to bounds. Specifically, we first sample the posterior predictive draws of (W_1^*, W_2^*), and then use the inverse logit transformation to obtain (W_1, W_2).

4 Simulation Studies

In this section, we assess the performance of our proposed method using simulated data. We emphasize that there are three factors that influence ecological inference: *aggregation effects* produced by different distributions of the observed weight variable X_i, *distributional effects* produced by different distributions of the unobserved variables W_i, and *contextual effects* produced by the possible dependence between X_i and W_i. The literature often does not make the distinction between aggregation and contextual effects, which are commonly lumped together and referred to as "aggregation bias."

Our simulation studies are designed to assess aggregation and distributional effects on ecological inference. We do not report our simulation results regarding contextual effects. Instead, we only note that most models, including ours, assume no contextual effect. And, without additional information, it is difficult to control contextual effects. In this section, we first show that even without contextual effects, aggregation effects can greatly influence ecological inference. Second, we investigate the sensitivity of the parametric model to its distributional assumption and the performance of the nonparametric model. For all simulations, we use the prior distribution described in Section 3.5. To fit our models, we use the Gibbs samplers described in Appendices. After running 20,000 MCMC iterations, we discard the initial 5,000 draws and take every fifth draw.

4.1 Aggregation Effects

For the ecological inference problem in 2 × 2 tables, we only observe Y_i, which is the weighted average of the two unknown variables, W_{1i} and W_{2i}, with known weight X_i. This means that even with the same values of W_i, a different distribution of X_i can lead to a different distribution of Y_i, consequently leading to different bounds conditions about W_i. Here, we evaluate our method against a variety of such aggregation effects by generating X_i from different distributions while keeping the same sample values of W_i. In particular, we draw (W^*_{1i}, W^*_{2i}) from a bivariate normal distribution with a mean $(0, 1.4)$, variance $(0.5, 0.5)$ and covariance 0.2. We then draw X_i, the ratio of black voters in each county, independently from the following distributions. We consider the sample size of 200.

- **Simulation I:** X_i is drawn independently from $\mathcal{B}(0.5, 2)$. This distribution is skewed to the right, implying that many counties have a high percentage of white voters.

- **Simulation II:** X_i is drawn independently from $\mathcal{B}(0.5, 0.5)$. This distribution corresponds to a polarized situation, where some counties are predominantly black, while others are white. Relatively few counties have similar ratios of black and white voters.

- **Simulation III:** X_i is drawn independently from $\mathcal{B}(2, 0.5)$. This distribution is skewed left, implying a situation where many counties have a high percentage of black voters.

Figure 2 presents the tomography plots and the bounds of W_1 and W_2 across the four simulation examples. Given W_{1i} and W_{2i}, the value of X_i determines the slope $-\frac{X_i}{1-X_i}$ and the intercept $\frac{Y_i}{X_i}$ of a tomography line. Therefore, applying different distributions of X_i yields four considerably different tomography plots even when the true values of W_i are identical. The second and third columns of Figure 2 illustrate how the bounds of W_{1i} and W_{2i} vary when different distributions of X_i are used. In general, when the value of X_i is close to 1, the bound for W_{1i} is likely to be narrow. Conversely, when the value of X_i is close to 0, the bound tends to be informative for W_{2i}. For example, in Simulation I where more counties have a high percentage of white voters, the bounds of W_2 are narrow. In contrast, Simulation III consists of more counties with a high percentage of black voters, and hence the bounds of W_1 are more informative. When X is more symmetrically distributed as in Simulation II, the amount of information contained in the bounds of W_1 and W_2 is approximately equal. For Simulation II, the distribution of X is chosen to be bimodal. In this case, both informative and non-informative bounds exist for W_1 and W_2.

An examination of the length of the bounds via summary statistics or graphs, such as those in Figure 2, gives us only a rough idea of the severity of aggregation effects. Instead, it is desirable to formally evaluate aggregation effects on parameter estimation. As discussed in Section 3.2, using the Supplemented *EM* algorithm, we can quantify the fraction of missing information for different aggregation patterns. Table 2 presents the ML estimates of the model parameters, their asymptotic standard errors, as well as the fraction of missing information for each simulation. In all cases,

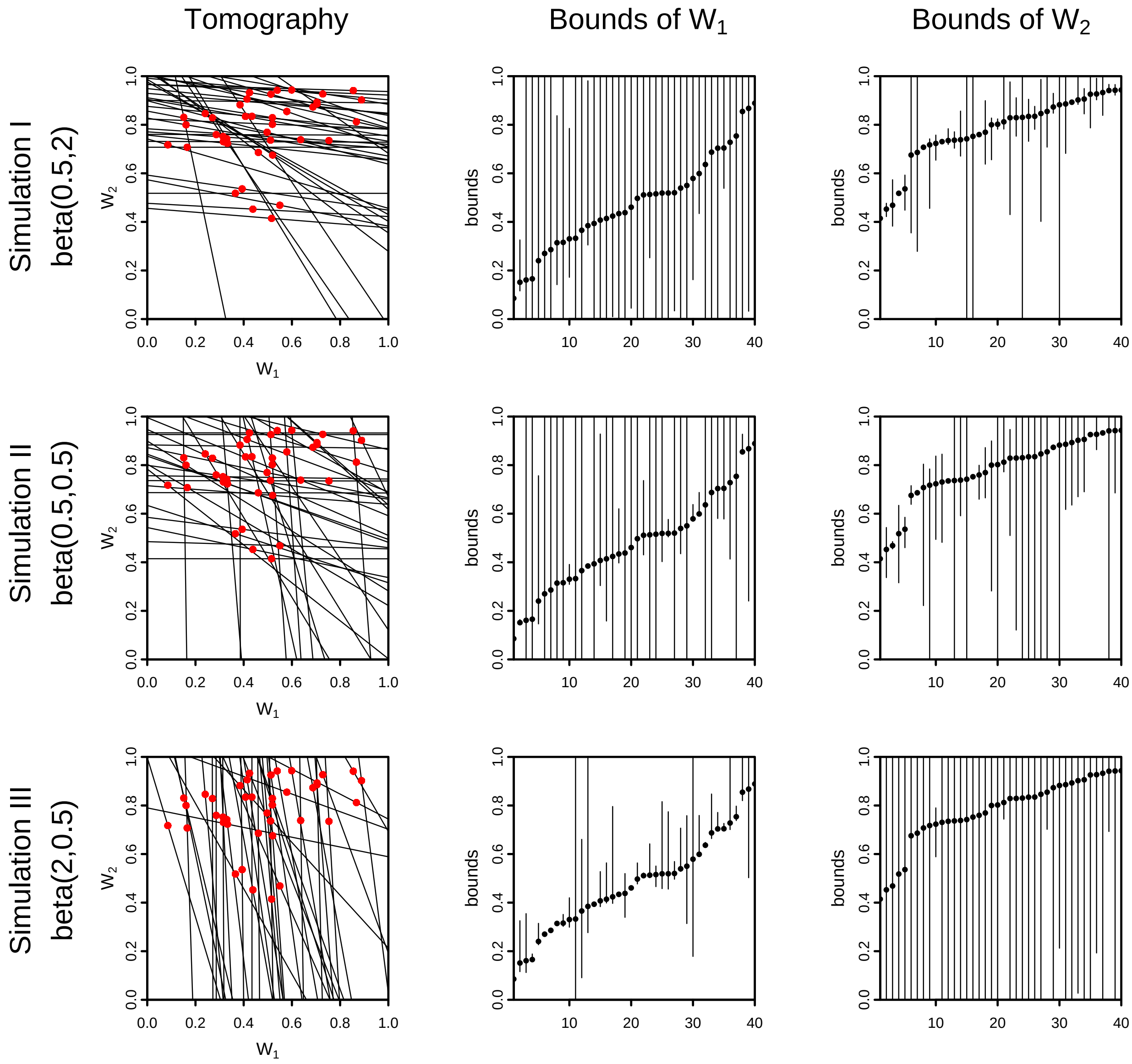


Figure 2: Aggregation Effects Due to Different Distributions of X. The four simulation setups with different distributions of X are illustrated in each row using randomly selected counties. The first column presents tomography plots. The second and third columns illustrate that the amount of information contained in the deterministic bounds depends on the distribution of X. The same set of the true values is used for W_1 and W_2 in all simulations.

		Simulation I		Simulation II		Simulation III	
	Complete-data	without	10%	without	10%	without	10%
	MLE	survey	survey	survey	survey	survey	survey
μ_1							
estimate	0.043	−0.079	−0.098	0.159	0.080	0.134	0.086
standard error	(0.068)	(0.166)	(0.118)	(0.100)	(0.087)	(0.080)	(0.069)
% miss. info.		*87.01*	*75.19*	*60.54*	*50.76*	*33.31*	*17.33*
μ_2							
estimate	1.386	1.414	1.393	1.242	1.267	1.028	1.095
standard error	(0.048)	(0.068)	(0.058)	(0.087)	(0.072)	(0.247)	(0.135)
% miss. info.		*47.74*	*34.46*	*62.11*	*53.04*	*91.07*	*80.01*
σ_{11}							
estimate	0.918	0.716	0.760	0.789	0.824	0.862	0.868
standard error	(0.091)	(0.219)	(0.164)	(0.145)	(0.123)	(0.111)	(0.097)
% miss. info.		*89.26*	*80.54*	*70.50*	*59.01*	*40.05*	*27.56*
σ_{22}							
estimate	0.467	0.490	0.480	0.570	0.542	1.086	0.801
standard error	(0.046)	(0.072)	(0.061)	(0.098)	(0.080)	(0.376)	(0.180)
% miss. info.		*53.57*	*42.93*	*66.22*	*58.02*	*91.67*	*82.01*
σ_{12}							
estimate	0.254	0.254	0.283	0.291	0.292	0.126	0.161
standard error	(0.038)	(0.240)	(0.150)	(0.180)	(0.126)	(0.209)	(0.134)
% miss. info.		*92.56*	*82.84*	*87.01*	*75.96*	*88.93*	*75.92*

Table 2: Aggregation Effects on Parameter Estimation in Four Simulations. The complete-data ML estimates of the model parameters are listed in the first column, while the remaining columns show the observed-data MLE using the EM algorithm. The standard errors of the estimates are in parentheses, while the numbers in italics represent the fraction of missing information. The second column for each simulation gives the results based on the addition of 10% individual-level data.

aggregation effects greatly impact parameter estimation; the fraction of missing information often exceeds 50%. In particular, there is hardly any observed information for estimating the covariance; the fraction of missing information is higher than 90% in all cases. On the other hand, there also exists a significant variation across different simulation setups. For example, the fraction of missing information for μ_2, the mean of W_2^*, is 48% in Simulation I while it is 91% in Simulation III. Indeed, the estimates of mu_2 from Simulation I are much closer to the complete-data ML estimates than those from Simulation III. Similar patterns are observed for the other parameter estimates.

We also examine the extent to which additional individual-level data can improve estimation. To do this, we add 20 counties using the same underlying distribution and assume that the values of W_{1i} and W_{2i} are known for these counties. The sample size of the resulting simulated datasets

	Simulation I		Simulation II		Simulation III	
	W_1	W_2	W_1	W_2	W_1	W_2
Bias						
Parametric Model						
Without survey	-0.012	0.001	0.023	-0.024	0.010	-0.031
With 10% survey	-0.023	0.002	0.015	-0.020	0.010	-0.028
Ecological regression	-0.037	0.004	0.010	-0.012	-0.010	0.053
RMSE						
Parametric Model						
Without survey	0.175	0.044	0.126	0.080	0.053	0.116
With 10% survey	0.171	0.044	0.123	0.078	0.050	0.112
Ecological regression	0.204	0.113	0.201	0.114	0.201	0.125

Table 3: In-sample Predictive Performance of the Parametric and Nonparametric Models When Subject to Different Aggregation Effects. The bias for W_j is calculated as $\sum_{i=1}^{n}(\widehat{W}_{ji} - W_{ji})/n$ for $j = 1, 2$, where $\widehat{W_{ji}}$ denotes the in-sample predictions of W_{ji}, and W_{ji} is the true value. Similarly, the root mean squared error (RMSE) is defined as $\sqrt{\sum_{i=1}^{n}(\widehat{W}_{ji} - W_{ji})^2/n}$.

is 220. In Table 2 these results are listed in the second column for each simulation. Overall, adding the survey data greatly reduces the fraction of missing information for all the estimates, resulting in smaller standard errors; the discrepancy between the observed-data and complete-data ML estimates may not necessarily be reduced due to sample variability. Even with the survey data, however, estimating the covariance between the two unknown variables is still a difficult task as indicated by its high fraction of missing information.

Finally, Table 3 presents the bias and the root mean squared error (RMSE) of the *in-sample* predictions for the parametric Bayesian model and the ecological regression. Bias is computed as the average difference between the in-sample predictions and their corresponding true values, i.e., $\sum_{i=1}^{n}(\widehat{W}_{ji} - W_{ji})/n$ for $j = 1, 2$ where $\widehat{W}_{ji}$ denotes the in-sample prediction of W_{ji}. RMSE is defined as the square root of the mean square error, $\sqrt{\sum_{i=1}^{n}(\widehat{W}_{ji} - W_{ji})^2/n}$. Although the ecological regression only estimates $E(W_j)$, this estimate is used as the "in-sample" prediction for all i as it is often done in practice. Under the assumption of no contextual effects, the ecological regression is known to estimate $E(W_i)$ without bias, and hence is shown here as a baseline.

In all simulations, our parametric model yields the magnitude of bias similar to that of the ecological regression. More importantly, when the amount of missing information is small, the parametric model provides in-sample predictions with smaller bias and RMSE. For example, in Simulation I, the bounds of W_2 are more informative than those of W_1, and the fraction of missing information for μ_1 is 87% as opposed to 48% for μ_2. Hence, the in-sample values of W_{2i}s are better predicted by the parametric model as well as the ecological regression; the same pattern is observed for Simulation III. Adding the individual-level data improves the overall prediction by reducing the RMSE and, in most cases, the bias, though the magnitude of improvement is rather small.

4.2 Distributional Effects

To investigate distributional effects, we use X_i from the dataset analyzed by Burden and Kimball (1998) which has a sample size of 361. Although this dataset is not about racial voting, for simplicity, we use the notation of Table 1 and refer to X_i as the proportion of black voters and Y_i as the overall turnout rate for each county i. The unknown inner cells (W_{1i}, W_{2i}) are the fractions of those who voted among black and white voters, respectively. To construct different simulation settings, we draw (W_{1i}, W_{2i}) independently from the following three distributions, while maintaining the same racial composition X_i,

- **Simulation IV:** (W^*_{1i}, W^*_{2i}) is independently drawn from the same bivariate normal distribution used in Section 4.1. This simulation setup follows the parametric model, and yields the average turnout of about 50 and 80 percent for black and white voters, respectively.

- **Simulation V:** (W^*_{1i}, W^*_{2i}) is independently drawn from a mixture of two bivariate normal distributions with the mixing probability vector $(0.6, 0.4)$. The first distribution has mean $(-0.4, 1.4)$, variance $(0.2, 0.1)$, and covariance 0. The second distribution has a different mean $(-0.4, -1.4)$, but the same covariance matrix. This simulation yields the average turnout of roughly 40 percent for black voters. For white voters, the average turnout is approximately 80 percent for three-fifths of the counties, and about 20 percent for the other counties.

- **Simulation VI:** (W^*_{1i}, W^*_{2i}) is independently drawn from a mixture of two bivariate normal distributions. The mixing probability vector is $(0.6, 0.4)$. The first distribution has mean $(-1.4, 1.4)$, variance $(0.1, 0.1)$, and covariance 0. The second distribution has a different

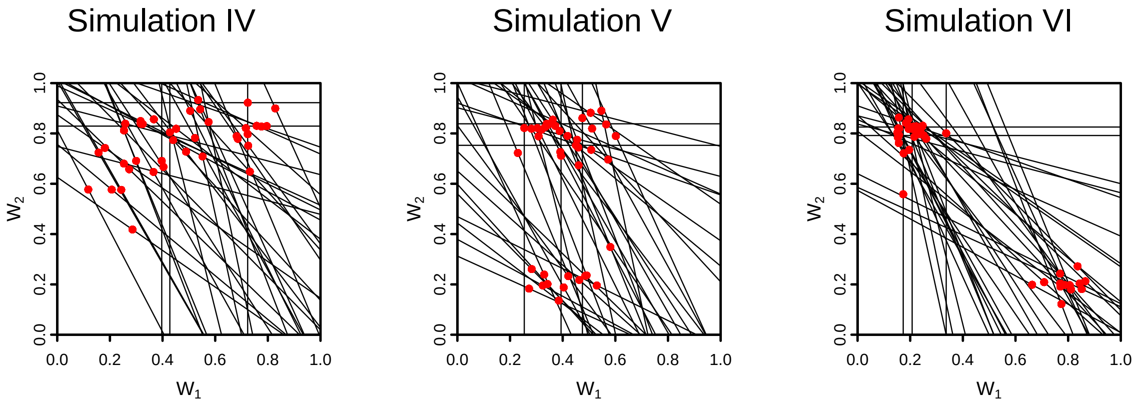


Figure 3: Tomography Plots of Simulations IV, V, and VI. The solid lines illustrate the deterministic relationship of equation 1, and the dots represent the true values of (W_{1i}, W_{2i}), for randomly selected 40 counties.

mean $(1.4, -1.4)$, but the same covariance matrix. In 60 percent of the counties, the average turnout is 20 percent for blacks and 80 percent for whites, while in the rest of the counties this pattern is reversed.

For each simulation, the inverse logit transformation gives the values of (W_{1i}, W_{2i}), and Y_i is computed given X_i and W_i using equation 1. Note that in Simulations V only the marginal distribution of W_{2i} is bimodal, while in Simulation VI the marginal distributions of both W_{1i} and W_{2i} are bimodal. It is of particular interest to see whether the nonparametric method can recover such distributions. In all three simulations, we assume no contextual effect.

Figure 3 presents the tomography plots of the simulated datasets with the true values of W_i. The graphs illustrate the bounds for W_{1i} and W_{2i}, which can be obtained by projecting tomography lines onto the horizontal and vertical axes. Using equations 2 and 3, we compute the length of bounds. The average length of bounds for W_{1i} in Simulations IV, V, and VI is 0.55, 0.58, and 0.64, while that for W_{2i} is 0.71, 0.73, and 0.78, respectively. This indicates that in all three simulations, the bounds are not particularly informative.

Treating X_i and Y_i as observed and W_i as unknown, we fit our parametric and nonparametric models and assess their relative performance in terms of both in-sample and out-of-sample predictions. Table 4 numerically summarizes the in-sample predictive performance. In Simulations V and VI, the RMSE of our nonparametric model is smaller than that of the parametric

	Simulation IV		Simulation V		Simulation VI	
	W_1	W_2	W_1	W_2	W_1	W_2
Bias						
Parametric model	-0.007	0.004	0.001	-0.009	-0.009	0.009
Nonparametric model	-0.006	0.003	0.004	-0.006	-0.007	0.007
Ecological regression	-0.011	0.011	-0.003	0.006	-0.027	0.029
RMSE						
Parametric model	0.083	0.080	0.095	0.162	0.133	0.135
Nonparametric model	0.083	0.080	0.078	0.147	0.112	0.105
Ecological regression	0.164	0.113	0.102	0.288	0.293	0.291

Table 4: In-sample Predictive Performance with Different Distributions of (W_1, W_2). The bias for W_j is calculated as $\sum_{i=1}^{n}(\widehat{W}_{ji} - W_{ji})/n$ for $j = 1, 2$, where $\widehat{W}_{ji}$ denotes the in-sample predictions of W_{ji}, and W_{ji} is the true value. Similarly, the root mean squared error (RMSE) is defined as $\sqrt{\sum_{i=1}^{n}(\widehat{W}_{ji} - W_{ji})^2/n}$.

model. Nevertheless, even when the true distribution is bimodal, the in-sample predictions from our parametric model are reasonable. This is because the parametric model yields the in-sample predictions that respect the bound conditions. The ecological regression yields relatively small bias in these simulations, but its RMSE is much larger than the other two methods.

Finally, we examine the out-of-sample predictive performance which is of importance for population inferences. Figure 4 compares the true distribution with the estimated marginal density based on out-of-sample predictions from our models. In Simulation IV, our nonparametric and parametric models give essentially identical estimates and approximate the marginal distributions well. Indeed, the number of clusters for the nonparametric model reduces to one. In our setup, the nonparametric model with one cluster is identical to the parametric model. This is not surprising given that this dataset is generated using the parametric model. The other two simulations, however, demonstrate the clear advantage of the nonparametric model. The nonparametric model captures the bimodality feature of the marginal distributions, while the parametric model fails to approximate the true distribution as expected. We have also investigated the effect of survey data on model performance. In this particular case, the additional individual data do not seem to matter much for the nonparametric model though they slightly improve the in-sample prediction of the parametric model (See Section 5 for more discussion).

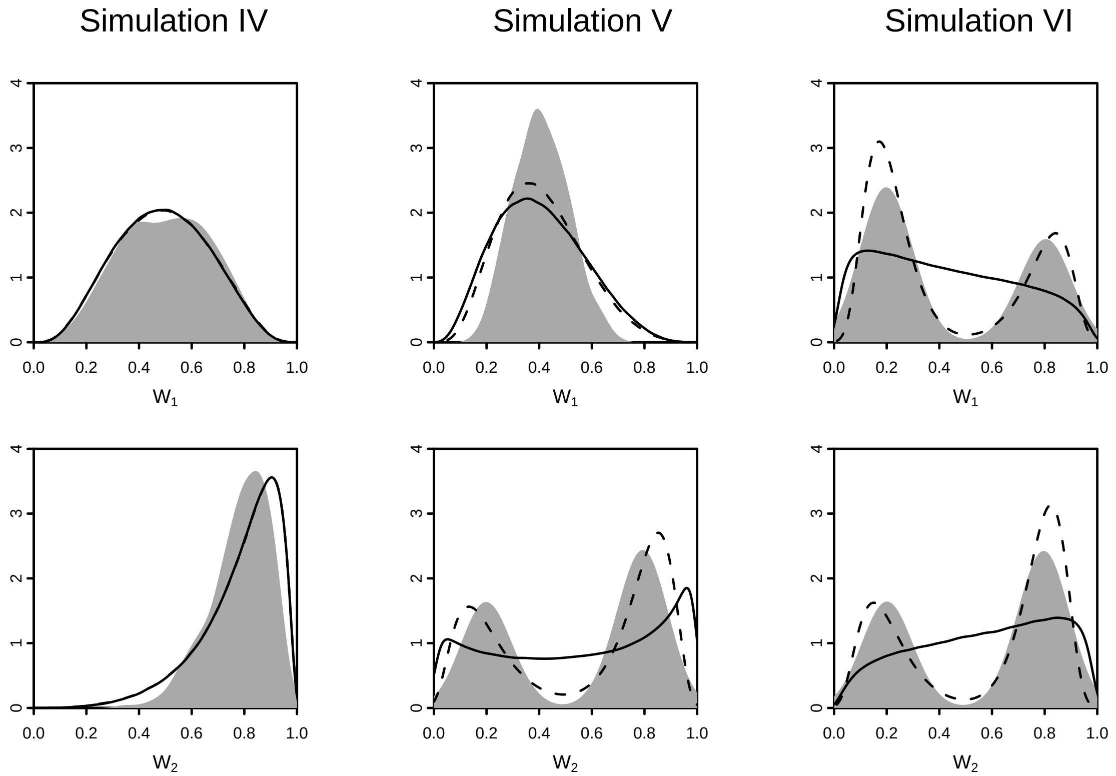


Figure 4: Out-of-sample Predictive Performance with Different Distributions of (W_1, W_2). The true marginal distributions are shown as shaded areas. The solid line represents the estimated density from the parametric model, whereas the dashed line represents that from the nonparametric model.

5 Empirical Application: Voter Registration in US Southern States

5.1 Data

In this section, we analyze voter registration data from 275 counties of four Southern states in the United States: Florida, Louisiana, North Carolina, and South Carolina. This dataset is first studied by King (1997) and subsequently analyzed by others (King *et al.*, 1999; Wakefield, 2004b). For each county, X_i represents the proportion of black voters, Y_i denotes the registration rate, W_{1i} and W_{2i} represent the registration rates of black and white voters. In this example, the true values of W_{1i} and W_{2i} are known, which allows us to compare the performance of our method with that of existing models in the literature.

Figure 5 presents a graphical summary of the data. The upper-left panel plots the true values of W_{1i} and W_{2i}. The registration rates among white voters are high in many counties with an

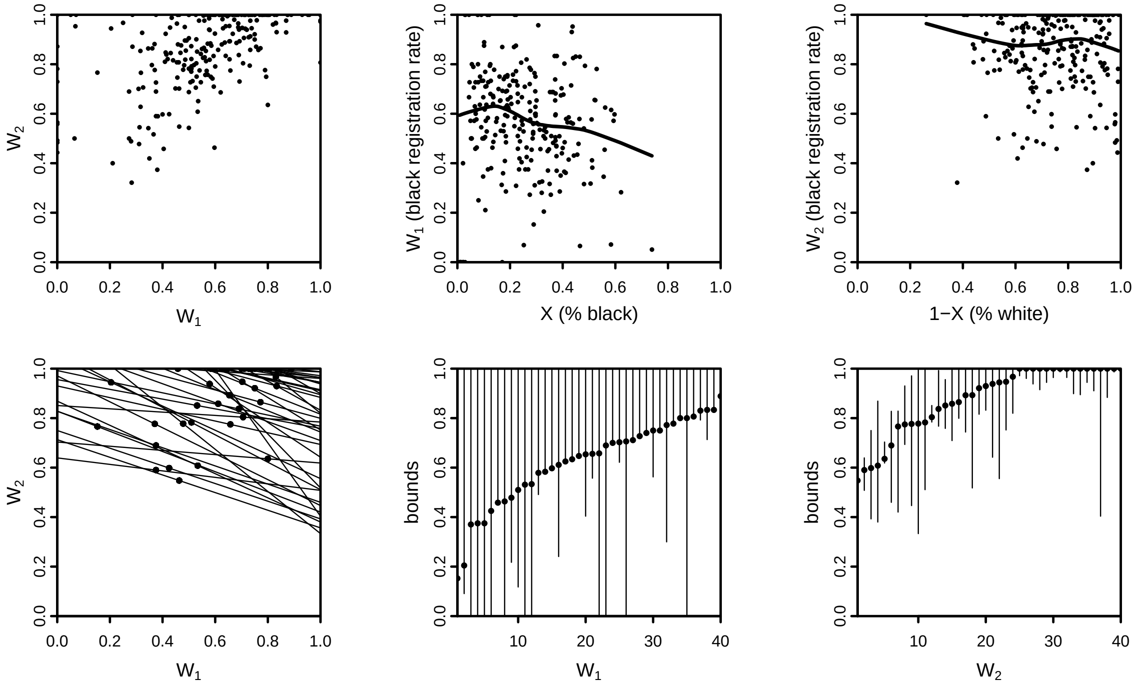


Figure 5: Summary of the Voter Registration Data from Four US Southern States. The upper-left graph is the scatter-plot of the true values of W_{1i} and W_{2i}. The upper-middle graph is the scatter-plot of black registration rate, W_{1i}, and the ratio of black voters, X_i. The solid line represents a LOWESS curve. The upper-right graph presents the same figure for white voters. The lower-left graph is the tomography plot with the true values indicated as dots. The lower middle and right graphs plot the bounds of W_1 and W_2, respectively.

average of 86 percent. In contrast, black registration rates are much lower, with an average of 56 percent. The sample variances of registration rates are 0.044 and 0.024 for black and white voters, respectively. The other two graphs in the upper panel are the scatter-plots of the registration rates and the proportions for black and white voters. In this dataset, the correlation between X and W_1 is -0.08, while the correlation between X and W_2 is only 0.01, implying minor contextual effects.

The lower panel of Figure 5 presents the tomography plots for a random subset of the counties. The bounds reveal asymmetric information about W_1 and W_2 and they are more informative for W_2 than for W_1. Moreover, for 30 percent of W_2, the true values are equal to 1. As a result, the true values of the corresponding W_1 lie at the lower end of the bounds. This may pose some difficulty for in-sample predictions, especially for the counties whose bounds are wide.

	Bias		RMSE		MAE	
	W_1	W_2	W_1	W_2	W_1	W_2
Without survey data						
Parametric model	−0.173	0.058	0.286	0.096	0.209	0.065
Nonparametric model	0.002	−0.001	0.163	0.049	0.111	0.032
With survey data						
Parametric model	−0.055	0.021	0.199	0.066	0.157	0.047
Nonparametric model	0.026	−0.011	0.146	0.054	0.095	0.029
Other methods						
Ecological regression	−0.059	0.016	0.226	0.156	0.177	0.121
King's EI model	0.093	−0.031	0.175	0.065	0.127	0.041
Wakefield's hierarchical model	0.045	−0.013	0.193	0.064	0.145	0.045
Neighborhood method	0.220	−0.077	0.311	0.182	0.247	0.158
Nonlinear neighborhood method	0.220	−0.077	0.269	0.111	0.224	0.078
Midpoints of bounds	0.099	−0.049	0.185	0.092	0.148	0.057

Table 5: In-sample Predictive Performance of Various Models. The bias for W_j is calculated as $\sum_{i=1}^{n}(\widehat{W}_{ji} - W_{ji})/n$ for $j = 1, 2$, where $\widehat{W}_{ji}$ denotes the in-sample predictions of W_{ji}, and W_{ji} is the true value. Similarly, the root mean squared error (RMSE) is defined as $\sqrt{\sum_{i=1}^{n}(\widehat{W}_{ji} - W_{ji})^2/n}$ and the mean absolute error (MAE) is given by $\sum_{i=1}^{n}|\widehat{W}_{ji} - W_{ji}|/n$.

5.2 Analysis

By treating W_1 and W_2 as unknown, we fit both our parametric and nonparametric models to 250 counties that are randomly selected without replacement. We also examine the model performance by adding the individual-level data of the remaining 25 counties. Finally, we compare the results with other methods in the literature, including the ecological regression, the linear and nonlinear neighborhood models, the midpoints of bounds, King's EI model, and Wakefield's hierarchical model. To fit King's EI model, we use the publicly available software, EzI (version 2.7) by Benoit and King, with its default specifications. To fit Wakefield's binomial convolution model, we use his WinBUGS code (Wakefield, 2004b) which fits the model based on normal approximation. We specify prior distributions such that the implied prior predictive distribution of W_i is approximately uniform. Specifically, we use $\mu_0 \sim \text{logistic}(0, 1)$, $\mu_1 \sim \text{logistic}(0, 1)$, $\sigma_0^{-2} \sim \mathcal{G}(1, 100)$, and $\sigma_1^{-2} \sim \mathcal{G}(1, 100)$. After 50,000 iterations, we discard the initial 20,000 draws and take every tenth draw.

Table 5 summarizes the in-sample predictive performance. For this dataset, our nonparametric model significantly outperforms our parametric model in all discrepancy measures by a magnitude

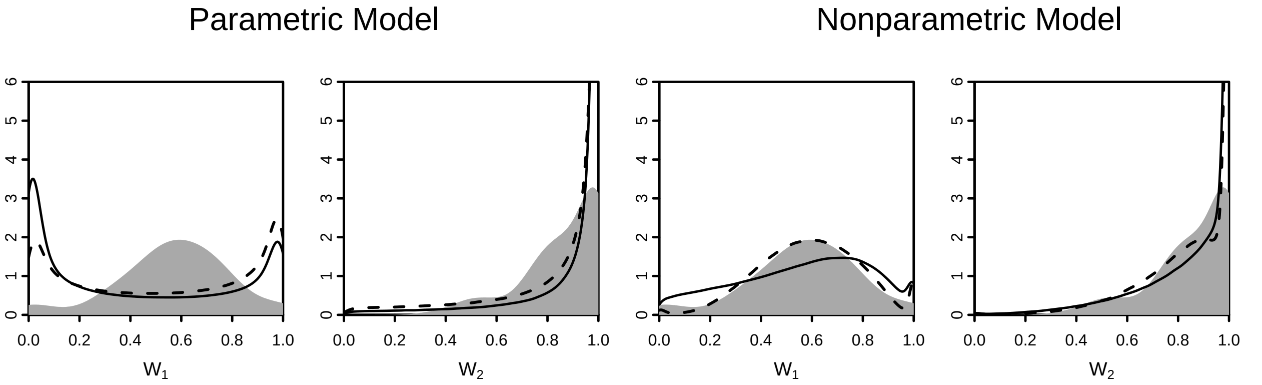


Figure 6: Out-of-sample Predictive Performance of Selected Models. The true density is represented by the shaded area. The solid and dashed lines represent the estimated density without and with the additional survey data information, respectively.

that is much greater than what we have seen in our simulation examples. With the addition of the individual-level data, however, the in-sample predictions of the parametric model improve substantially. Furthermore, the predictions of the nonparametric model are more accurate than those of existing methods. For example, the biases of the ecological regression are more than ten times as large as those of the nonparametric model. The performance of King's EI model and Wakefield's model is comparable with that of the nonparametric model in terms of RMSE and MAE (mean absolute error), but the biases of both models are larger. Finally, the neighborhood models do not work well in this application, and simply using the midpoint of a bound as an estimate gives better results than some methods.

For our two models, the posterior predictive distribution serves as a basis for population inferences. Figure 6 compares the out-of-sample predictive performance of our models, with and without the help of individual level data. In this application, the true distribution of W_1 and W_2 is unknown, so we approximate it by a kernel smoothing technique using the sample values (Wand and Jones, 1995). The nonparametric model estimates the marginal density of W_2 very well, while its density estimate for W_1 is slightly off. (Note that the bounds of W_2 are more informative than those of W_1.) In contrast, the estimated marginal densities based on our parametric model are not accurate. With the addition of the individual-level data, the nonparametric model now recovers the density of W_1, and the density estimation of W_2 is further improved. The parametric model still gives a poor estimate even after adding the individual-level data.

6 Concluding Remarks

Since Robinson's (1950) article, the ecological inference problem has attracted the attention of many social scientists, epidemiologists, and statisticians. In this article, we propose parametric and nonparametric models for ecological inference in 2 × 2 tables. Both models provide in-sample predictions that are consistent with deterministic bounds. At the same time, they also give out-of-sample predictions that can be used to make population inferences. Although the distinction between in-sample and out-of-sample inferences is rarely made in the literature, it is essential for evaluating ecological inference models.

In addition, the proposed parametric model allows one to formally identify the amount of missing information. The simulation study shows that the amount of missing information depends highly on the distribution of the racial composition variable. In many scenarios, aggregation effects are so severe that more than half of the information is lost, yielding estimates with little precision. Moreover, if the distributional assumption is not satisfied, the resulting inferences may be even more unreliable. Our nonparametric Bayesian model relaxes the distributional assumption of the parametric model. Through simulation studies and an empirical example, we find that in general the nonparametric model outperforms parametric models. Therefore, our nonparametric model offers the advantage of flexible estimation, which is important given the inherent uncertainty about underlying distribution in ecological inference.

Appendices: Computational Details

A Gibbs Sampler for the Parametric Model

To sample from the joint posterior distribution $p(W_i^*, \mu, \Sigma \mid Y, X)$, we construct a Gibbs sampler. First, we draw W_i from its conditional posterior density, which is proportional to,

$$\frac{\mathbf{1}\{W_i : Y_i = W_{1i}X_i + W_{2i}(1-X_i)\}}{\sqrt{2\pi|\Sigma|}\, W_{1i}W_{2i}(1-W_{1i})(1-W_{2i})} \exp\left[-\frac{1}{2}\left\{\text{logit}(W_i) - \mu\right\}^\top \Sigma^{-1}\left\{\text{logit}(W_i) - \mu\right\}\right], \tag{9}$$

if $(W_{1i}, W_{2i}) \in (0,1)$, otherwise the density is equal to zero. Although equation 9 is not the density of a standard distribution, it has a bounded support because (W_{1i}, W_{2i}) lies on a bounded line segment. Therefore, we can use the inverse CDF method by evaluating equation 9 on a grid of equidistant points on a tomography line. Given a sample of W_i, we then obtain W_i^* via the logit transformation. Alternatively, Metropolis-Hastings or importance sampling algorithms can be used, although they require separate tuning parameters or target densities for each observation.

When $X_i = 1$, we know W_{1i} exactly, and so we take the logit transformation of the observed value. In this situation, however, W_{2i} does not exist. We therefore impute W_{2i}^*, conditioning on the observed W_{1i}^*, from $\mathcal{N}[\mu_2 + \sigma_{12}(W_{1i}^* - \mu_1)/\sigma_{11}),\ \sigma_{22}(1-\rho^2)]$ where σ_{jk} is the (j,k) element of Σ and $\rho = \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$. The inverse logit transformation gives a draw of W_{2i}. When $X = 0$, W_{2i} is observed, and a similar method is used to draw W_{1i}. Moreover, if $Y_i = 1$, then $W_{1i} = W_{2i} = 1$. In this case, we set $W_{1i}^* = W_{2i}^* = \text{logit}(1-\epsilon)$, where ϵ is a small positive number. If $Y_i = 0$, on the other hand, we set $W_{1i}^* = W_{2i}^* = \text{logit}(\epsilon)$. Alternatively, one can exclude these observations from the sample since no internal cell needs to be estimated.

Next, we draw (μ, Σ) from their conditional posterior distributions. Note that the observed data, (Y_i, X_i), are redundant given W_i^*. The augmented-data conditional posterior distribution has the form of a standard bivariate normal/inverse-Wishart model, $p(\mu, \Sigma \mid W_i^*) \propto p(\mu \mid \Sigma)\, p(\Sigma) \prod_{i=1}^n p(W_i^* \mid \mu, \Sigma)$. This implies that conditioning on W_i^*, sampling (μ, Σ) can be done using the following standard distributions,

$$\mu \mid W^*, \Sigma \ \sim\ \mathcal{N}\left(\frac{\tau_0^2\mu_0 + n\overline{W}^*}{\tau_0^2 + n},\ \frac{\Sigma}{\tau_0^2 + n}\right), \quad \text{and} \quad \Sigma \mid W^* \ \sim\ \text{InvWish}\left(\nu_0 + n,\ S_n^{-1}\right),$$

where $W^* = \{W_i^*, \ldots, W_n^*\}$, $\overline{W}^* = \sum_{i=1}^n W_i^*/n$, and $S_n = S_0 + \sum_{i=1}^n (W_i^* - \overline{W}^*)(W_i^* - \overline{W}^*)^\top + \frac{\tau_0^2 n}{\tau_0^2 + n}(\overline{W}^* - \mu_0)(\overline{W}^* - \mu_0)^\top$.

B Gibbs Sampler for the Nonparametric Model

We construct a Gibbs sampler in order to sample from the joint posterior distribution $p(W^*, \mu, \Sigma, \alpha \mid Y)$. First, we independently sample W_i for each i and transform it to obtain W_i^* in the same way as above, but we replace (μ, Σ) with (μ_i, Σ_i) in equation 9. Then, given the draw of W_i^*, the augmented-data model can be estimated through a multivariate generalization of the density estimation method of Escobar and West (1995). In our Gibbs sampler, we sample (μ_i, Σ_i) given $(\mu^{(i)}, \Sigma^{(i)}, W^*, \alpha)$ for each i, and then update α based on the new values of (μ_i, Σ_i).

An application of the usual calculation due to Antoniak (1974) shows that the conditional posterior distribution of (μ_i, Σ_i) given W_i^* is given by the following mixture of Dirichlet processes,

$$(\mu_i, \Sigma_i) \mid \mu^{(i)}, \Sigma^{(i)}, W_i^* \quad \sim \quad q_0 \; G_i(\mu_i, \Sigma_i) \; + \sum_{j=1, j\neq i}^{n} q_j \; \delta_{(\mu_j, \Sigma_j)}(\mu_i, \Sigma_i),$$

where $G_i(\mu_i, \Sigma_i)$ is the posterior distribution under G_0 which is a normal/inverse-Wishart distribution with components,

$$\begin{aligned} \mu_i \mid \Sigma_i \quad &\sim \quad \mathcal{N}\left(\frac{\tau_0^2 \mu_0 + W_i^*}{\tau_0^2 + 1}, \; \frac{\Sigma_i}{\tau_0^2 + 1}\right), \\ \Sigma_i \quad &\sim \quad \text{InvWish}\left[\nu_0 + 1, \; \left\{S_0 + \frac{\tau_0^2}{\tau_0^2 + 1}(W_i^* - \mu_0)(W_i^* - \mu_0)^\top\right\}^{-1}\right]. \end{aligned}$$

Next, following West, Müller, and Escobar (1994), we derive the weights, q_0 and q_j, by computing the marginal (augmented-data) likelihood $p(W_i^* \mid \mu_i, \Sigma_i)$ and $p(W_i^* \mid \mu_j, \Sigma_j)$, respectively,

$$\begin{aligned} q_0 \quad &\propto \quad \alpha \, \frac{2\tau_0^2 \, \Gamma\left(\frac{\nu_0+1}{2}\right)}{(\tau_0^2 + 1)\, \Gamma\left(\frac{\nu_0 - 1}{2}\right)} |S_0|^{-1/2} \left\{1 + \frac{\tau_0^2}{\tau_0^2 + 1}(W_i^* - \mu_0)^\top S_0^{-1}(W_i^* - \mu_0)\right\}^{-(\nu_0+1)/2}, \\ q_j \quad &\propto \quad |\Sigma_j|^{-1/2} \exp\left\{\frac{1}{2}(W_i^* - \mu_j)^\top \Sigma_j^{-1}(W_i^* - \mu_j)\right\} \quad \text{for} \quad j = 1, \ldots, n, \quad \text{and} \quad j \neq i, \end{aligned}$$

where $\sum_{j=0, j\neq i}^n q_j = 1$. q_0 is proportional to the bivariate t density with $(\nu_0 - 1)$ degrees of freedom, the location parameter μ_0, and the scale matrix $\tau_0^2(\nu_0 - 1)S_0/(1 + \tau_0^2)$. q_j is proportional to the bivariate normal density with mean μ_j and variance Σ_j.

Given these weights, we can approximate $p(\mu, \Sigma \mid W^*)$ via a Gibbs sampler by sampling (μ_i, Σ_i) given $(\mu^{(i)}, \Sigma^{(i)}, W_i^*)$ for each i. This step creates clusters of units where some units share the same values of the population parameters. At a particular iteration, we have $J \leq n$ clusters each of which has n_j units with $\sum_{j=1}^{J} n_j = n$. Note that the number of clusters J can vary from one iteration to another. Bush and MacEachern (1996) recommend adding the 'remixing' step to prevent the Gibbs sampler from repeatedly sampling a small set of values. In our application, we update the new values of the parameters (μ_i, Σ_i) by using the newly configured cluster structure. That is, for each cluster j, we update the parameters with $(\tilde{\mu}_j, \widetilde{\Sigma}_j)$ by drawing them from the following conditional distribution,

$$\begin{aligned}
\tilde{\mu}_j \mid \widetilde{\Sigma}_j, \{W_i^* : i \in j\text{th cluster}\} &\sim \mathcal{N}\left(\frac{\tau_0^2 \mu_0 + n_j \overline{W}_j^*}{\tau_0^2 + n_j}, \ \frac{\widetilde{\Sigma}_j}{\tau_0^2 + n_j}\right), \\
\widetilde{\Sigma}_j \mid \{W_i^* : i \in j\text{th cluster}\} &\sim \text{InvWish}\left(\nu_0 + n_j, \ S_{n_j}^{-1}\right),
\end{aligned}$$

where $S_{n_j} = S_0 + \sum_{i \in j\text{th cluster}}^{n_j} (W_i^* - \overline{W}_j^*)(W_i^* - \overline{W}_j^*)^\top + \frac{\tau_0^2 n_j}{\tau_0^2 + n_j}(\overline{W}_j^* - \mu_0)(\overline{W}_j^* - \mu_0)^\top$, and $\overline{W}_j^* = \sum_{i \in j\text{th cluster}}^{n_j} W_i^*/n_j$. Given these new draws, we set $\mu_i = \mu_j^*$ and $\Sigma_i = \Sigma_j^*$ for each i that belongs to the jth cluster.

When small scale survey data, W_i^{survey}, are available for some counties, we can update the posterior draws of μ_i and Σ_i conditional on $\mu^{(i)}, \Sigma^{(i)}, W_i^*$ and W_i^{survey}. In this case, the posterior distribution will be conditional on two data points instead of just one. The corresponding components of the Dirichlet mixture can be modified accordingly. At the remixing step, one can include the survey data directly.

Finally, to update α, we use the algorithm developed by Escobar and West (1995). Namely, the conditional posterior distribution of α has the form of the following gamma mixture,

$$\alpha \mid \eta, J \ \sim \ \omega \, \mathcal{G}\left(a_0 + J, \ b_0 - \log \eta\right) \ + \ (1 - \omega) \, \mathcal{G}\left(a_0 + J - 1, b_0 - \log \eta\right),$$

where $\omega = (a_0 + J - 1)/\{n(b_0 - \log \eta)\}$, and η is a latent variable that follows a beta distribution, $\mathcal{B}(\alpha + 1, J)$. This completes one cycle of our Gibbs sampler.

References

Achen, C. H. and Shively, W. P. (1995). *Cross-Level Inference.* University of Chicago Press, Chicago.

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics* **2**, 1152–1174.

Brown, P. J. and Payne, C. D. (1986). Aggregate data, ecological regression, and voting transitions. *Journal of the American Statistical Association* **81**, 452–460.

Burden, B. C. and Kimball, D. C. (1998). A new approach to the study of ticket splitting. *American Political Science Review* **92**, 3, 533–544.

Bush, C. A. and MacEachern, S. N. (1996). A semiparametric Bayesian model for randomized block designs. *Biometrika* **83**, 275–285.

Cho, W. K. T. and Gaines, B. J. (2004). The limits of ecological inference: The case of split-ticket voting. *American Journal of Political Science* **48**, 1, 152–171.

Cleave, N., Brown, P. J., and Payne, C. D. (1995). Evaluation of methods for ecological inference. *Journal of the Royal Statistical Society, Series A, General* **158**, 55–72.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B, Methodological* **39**, 1–37.

Dey, D., Müller, P., and Sinha, D., eds. (1998). *Practical nonparametric and semiparametric Bayesian statistics.* Springer-Verlag Inc, New York.

Duncan, O. D. and Davis, B. (1953). An alternative to ecological correlation. *American Sociological Review* **18**, 665–666.

Escobar, M. D. (1994). Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association* **89**, 268–277.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90**, 577–588.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* **1**, 209–230.

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *The Annals of Statistics* **2**, 615–629.

Ferguson, T. S. (1983). *Recent Advances in Statistics: Papers in Honor of Herman Chernoff on His Sixtieth Birthday (eds., M. Haseeb Rizvi, Jagdish S. Rustagi, and David Siegmund)*, chap. Bayesian density estimation by mixtures of normal distributions, 287–302. Academic Press, New York.

Freedman, D. A., Klein, S. P., Sacks, J., Smyth, C. A., and Everett, C. G. (1991). Ecological regression and voting rights (with discussion). *Evaluation Review* **15**, 673–816.

Freedman, D. A., Ostland, M., Roberts, M. R., and Klein, S. P. (1998). "Review of 'A Solution to the Ecological Inference Problem' ". *Journal of the American Statistical Association* **93**, 1518–1522.

Gelman, A., Park, D. K., Ansolabehere, S., Price, P. N., and Minnite, L. C. (2001). Models, assumptions and model checking in ecological regressions. *Journal of the Royal Statistical Society, Series A* **164**, 101–118.

Goodman, L. (1953). Ecological regressions and the behavior of individuals. *American Sociological Review* **18**, 663–666.

Goodman, L. A. (1959). Some alternatives to ecological correlation. *The American Journal of Sociology* **64**, 610–624.

Greenland, S. and Robins, J. M. (1994). Ecologic studies: Biases, misconceptions, and counterexamples. *American Journal of Epidemiology* **139**, 747–760.

Grofman, B. (1991). Statistics without substance: A critique of Freedman et al. and Clark and Morrison. *Evaluation Review* **15**, 6, 746–769.

Imai, K. and King, G. (2004). Did illegal overseas absentee ballots decide the 2000 U.S. presidential election? *Perspectives on Politics* **2**, 3, 537–549.

King, G. (1997). *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data.* Princeton University Press, Princeton, NJ.

King, G. (1999). Comment on "Review of 'A Solution to the Ecological Inference Problem' ". *Journal of the American Statistical Association* **94**, 352–355.

King, G., Rosen, O., and Tanner, M. A. (1999). Binomial-beta hierarchical models for ecological inference. *Sociological Methods & Research* **28**, 61–90.

King, G., Rosen, O., and Tanner, M. A., eds. (2004). *Ecological Inferece: New Methodological Strategies.* Cambridge University Press.

Lange, K. (1999). *Numerical Analysis for Statisticians.* Springer Verlag, New York.

Lichtman, A. J. (1991). Passing the test: Ecological regression analysis in the Los Angeles county case and beyond. *Evaluation Review* **15**, 6, 770–799.

Meng, X.-L. and Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association* **86**, 899–909.

Mukhopadhyay, S. and Gelfand, A. E. (1997). Dirichlet process mixed generalized linear models. *Journal of the American Statistical Association* **92**, 633–639.

Orchard, T. and Woodbury, M. A. (1972). A missing information principle: Theory and applications. *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability* **1**, 697–715.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review* **15**, 351–357.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association* **82**, 528–550.

Wakefield, J. (2004a). Ecological inference for 2 × 2 tables (with discussion). *Journal of the Royal Statistical Society, Series A* **167**, 385–445.

Wakefield, J. (2004b). *Ecological Inference: New Methodological Strategies (eds. Gary King, Ori Rosen and Martin Tanner)*, chap. Prior and likelihood choices in the analysis of ecological data, forthcoming. Cambridge University Press, Cambridge.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing.* Chapman & Hall, London.

West, M., Müller, P., and Escobar, M. D. (1994). *Aspects of Uncertainty: A Tribute to D. V. Lindley (eds. A.F.M. Smith and P.R. Freedman)*, chap. Hierarchical priors and mixture models, with application in regression and density estimation, 363–386. John Wiley & Sons, London.