Bayesian Analysis (2013) **8**, Number 2, pp. 453–478

Multiple-Shrinkage Multinomial Probit Models with Applications to Simulating Geographies in Public Use Data

Lane F. Burgette* and Jerome P. Reiter†

Abstract. Multinomial outcomes with many levels can be challenging to model. Information typically accrues slowly with increasing sample size, yet the parameter space expands rapidly with additional covariates. Shrinking all regression parameters towards zero, as often done in models of continuous or binary response variables, is unsatisfactory, since setting parameters equal to zero in multinomial models does not necessarily imply "no effect." We propose an approach to modeling multinomial outcomes with many levels based on a Bayesian multinomial probit (MNP) model and a multiple shrinkage prior distribution for the regression parameters. The prior distribution encourages the MNP regression parameters to shrink toward a number of learned locations, thereby substantially reducing the dimension of the parameter space. Using simulated data, we compare the predictive performance of this model against two other recently-proposed methods for big multinomial models. The results suggest that the fully Bayesian, multiple shrinkage approach can outperform these other methods. We apply the multiple shrinkage MNP to simulating replacement values for areal identifiers, e.g., census tract indicators, in order to protect data confidentiality in public use datasets.

Keywords: Confidentiality, Dirichlet process, disclosure, spatial, synthetic

1 Introduction

In models of discrete choices, agents often choose from a large number of outcome categories. For example, a researcher may conceptualize immigrants to the U.S. as choosing to make one of several hundred metropolitan areas their new home. A marketer may be interested in understanding which car models—among the dozens available—are likely to interest a consumer with a given set of characteristics. Finally, as we motivate further in Section 4, a statistical agency may seek to model associations between people's demographic variables and home census tract identifier, with the intention of releasing simulated values of data subjects' locations for release in public use datasets. This could enable the agency to protect data subjects' confidentiality while releasing datasets with fine levels of areal geography.

Models of response variables with large numbers of outcome categories encounter several difficulties. Foremost is the rate at which the model dimensions expand when adding new covariates. If there are p categories, adding a covariate whose values are specific to the decision-maker (as opposed to an outcome category-specific covariate)

*RAND Corporation, Arlington, VA, burgette@rand.org
†Department of Statistical Science, Duke University, Durham, NC, jerry@stat.duke.edu

 DOI:10.1214/13-BA816

adds $p-1$ identifiable regression parameters to the model. On the other hand, each additional observation typically carries a small amount of information relative to standard models of continuous or ordered categorical outcomes. These issues combine to make regularization—either through Bayesian approaches or penalties for the likelihood— an essential aspect of modeling.

Regularization with unordered categorical outcomes introduces a distinct set of challenges from regularization with continuous or binary outcomes. First, in models of continuous or binary outcomes, regression parameters set equal to zero correspond to no effect; hence, shrinkage towards zero carries special importance in these models. This is not necessarily the case with unordered categorical outcomes: even when a regression parameter for a particular covariate and outcome category equals zero, changing that covariate's value can impact the probability of selecting the category. This is because other categories may have non-zero regression parameters for that covariate that cause their probabilities to change, which in turn can impact the probabilities of the categories with null regression parameters. Consequently, zeros in the vector of multinomial regression parameters do not have the same importance that they do in models of continuous or binary outcomes, and global shrinkage toward zero may not be a reasonable regularization strategy.

A second, related issue is that in standard formulations of the multinomial logit (MNL) and multinomial probit (MNP) models, the analyst chooses a base category to identify the model. The choice of a base category can interact with the prior distribution in unpredictable and undesirable ways (Lenk and Orme 2009). For example, Krishnapuram et al. (2005) and Sha et al. (2004) proposed multinomial models (each with a base category) that encourage regression parameters to equal zero. This can imply a strong dependence on the base category. Such a penalty should work well when there is a single main group of categories whose regression parameters are nearly equal and the base category is in that group. However, a log odds of zero may mean very different things when the base category is changed from one value to another.

In this manuscript, we propose a novel strategy for Bayesian multinomial regression modeling with large numbers of outcome levels. In particular, we break from the "shrink toward zero" approach that has dominated previous regularization strategies for multinomial models in favor of a strategy that shrinks toward multiple values; that is, we identify groups of regression parameters that are indistinguishable from one another. Arguably, with multinomial outcomes it is more important to identify such groups than to identify coefficients that are indistinguishable from zero. For example, in a model of immigrants' choices of location, we may find that those with high education levels are more likely than their less educated peers to select Seattle, San Francisco, or Raleigh and relatively less likely to move to Phoenix, Detroit, or Atlanta. However, within these groups of cities, the ratios of selection probabilities may be insensitive to education levels.

To implement this strategy, we use Bayesian MNP models with a modified version of the multiple-shrinkage prior distribution of MacLehose and Dunson (2010). This prior distribution was designed with a different purpose in mind—namely, strongly shrinking

parameters in binary logistic regression that appear to be unimportant, while minimizing shrinkage of larger effects—but we adapt it to achieve the within-group shrinkage for multinomial models. The prior distribution is constructed using a Dirichlet process (DP) (Ferguson 1973; Blackwell and MacQueen 1973), so that the coefficients cluster around a small number of learned locations without the analyst having to specify the number in advance. To avoid the base category problems, we use a symmetric multinomial model, which enforces a sum-to-zero identification rule for the latent utilities (Burgette and Hahn 2010). This treats each outcome category equitably in the prior distribution and removes the worry that the regularization properties depend on the choice of a base category.

We propose models based on both normal and Laplace distributions in the DP mixture. The normal kernel offers easier computation, but the Laplace kernel tends to result in tighter within-component distributions of coefficients near the component means, which we expect to better achieve the objective of regularization by grouping coefficients. We note that Laplace distributions sometimes are used for robustness because of their heavier-than-normal tails. That is not our motivation here, as we expect the mixture formulation of the prior distribution to supplant the robustness of using Laplace tails.

The DP has been employed previously in multinomial applications, but the focus has been on increasing flexibility of the model in applications with modest numbers of outcome categories, clustering over observations rather than outcome categories. For example, Kim et al. (2004) and Burda et al. (2008) suggest DP-based models that allow for household-level heterogeneity in regression parameters, though they apply their methods to analyses with four and five outcome categories, respectively. Shahbaba and Neal (2009) present a DP mixture of multinomial logit (MNL) models, which allows for nonlinear relationships. De Blasi et al. (2010) investigate consistency properties of nonparametric mixed MNL models, and consider a simulation with $p = 3$ outcome categories. In contrast to these applications, we use the DP as a means to regularize multinomial models with much larger p.

The closest work to our own is the L_1-penalized MNL model of Friedman et al. (2010; FHT), which corresponds to maximum *a posteriori* (MAP) estimates under a Laplace (or double exponential) prior on the regression parameters. This model avoids specifying a base category via the penalty, since for two parameter configurations that imply the same fitted probabilities, one will be preferred by the penalty. Cawley et al. (2007) also takes this general approach. Our framework differs from this work in two key ways. First, FHT focus on MAP or penalized maximum likelihood estimates, whereas our approach offers full Bayesian inference. Second, the FHT model shrinks only to zero rather than the multiple shrinkage we advocate.

Another closely related work is the model of Taddy (2012; MT), who describes an inverse regression approach to sentiment modeling that can, for example, be used to model diners' restaurant experiences based on text in written reviews. An MNL model for large covariate spaces is embedded in this work. Like FHT, MT uses Laplace-type regularization. However, instead of choosing a global tuning parameter via cross-

validation, MT introduces separate gamma-distributed hyperpriors that regulate the Laplace shrinkage for each of the non-intercept regression parameters. After specifying this hierarchical structure for the regression parameters, MT uses cyclic coordinate descent to produce MAP estimates.

In addition to these two approaches, several researchers have developed models for large multinomial outcomes in the field of discrete choice models. This includes the early work of McFadden (1978), which can be used to skirt the estimation of numerous nuisance parameters that arise in models of residential moves; see also Duncombe et al. (2001). Large multinomial responses also naturally occur in the field of topic modeling (Blei et al. 2003). These models include multinomial responses with large numbers of categories, but they are buried in the model. Since interest focuses on a much lower dimensional quantity, the modeling goals are distinct from those considered here. Thus, when evaluating the MNP models that we develop, we compare their performance to those of the FHT and MT models, which we consider to be the closest competitors.

The remainder of the article is arranged as follows. In Section 2, we formally describe the models and their estimation. In Section 3, we present results of simulation studies and compare our methods with the FHT and MT models. In Section 4, we describe a motivating application for the development of these models, which is to predict areal indicators of individuals' homes (i.e., census tracts) from their demographic characteristics for the purpose of releasing simulated indicators in public use datasets. To our knowledge, no one has proposed releasing simulated aggregated geography as a means of protecting confidentiality. In Section 5, we conclude with a discussion and suggested directions for future research.

2 The Multiple-Shrinkage Multinomial Probit

In contexts where the number of parameters grows with the sample size, Bayesian semiparametric and nonparametric approaches use the shrinking or regularizing properties of the prior distribution to make the model tractable. For the large multinomial outcome setting considered here, we seek to fit a model with a number of parameters that is both fixed and smaller than the sample size. Even so, each observation offers little information to estimate the model, so we use ideas from the literature on semiparametric Bayesian modeling to shrink strongly yet flexibly. In particular, we use a collection of truncated Dirichlet process (DP) priors and a modified version of the multiple-shrinkage prior distribution of MacLehose and Dunson (2010) to shrink the regression parameters toward a small number of learned locations. We begin our description of the model with a review of the DP and further discussion of the multiple-shrinkage prior distribution.

The DP is the workhorse of many Bayesian semiparametric models. Sethuraman (1994) demonstrated the "stick-breaking" formulation of the DP, which gives intuition

into its behavior. If $D \sim \text{DP}(\alpha, D_0)$, we have the following almost-sure representation:

$$\begin{aligned} v_j &\overset{\text{iid}}{\sim} \text{Beta}(1, \alpha) \\ \pi_j &= v_j \prod_{k=1}^{j-1}(1 - v_k) \\ z_j &\overset{\text{iid}}{\sim} D_0 \\ \Pr(D = d) &= \sum_{j=1}^{\infty} \pi_j \mathbf{1}(d = z_j). \end{aligned}$$

The term "stick-breaking" comes from the analogy of breaking a piece of size v_1 from a unit length stick, after which we set $\pi_1 = v_1$. Then, from the remaining stick with length $(1 - v_1)$, break off a proportion v_2, and set $\pi_2 = v_2(1 - v_1)$. Repeating this infinitely provides the weights for the DP. These discrete masses are at the locations z_j, which are assumed to have been drawn independently from the base measure D_0. In our model, the z_j are pairs of shrinkage locations and scales. A key feature of the distribution is the stochastic decay of the π_j, which means that $E(\pi_j - \pi_k) > 0$ for $j > k$. When the DP is used as a mixing distribution, this encourages many observations to be assigned to the same mixture components.

With a continuous or binary outcome variable, recent research has shown that it can be desirable to strongly shrink variables that have small estimated effects while minimally shrinking those that have stronger apparent effects; for example, see the "horseshoe" estimator of Carvalho et al. (2010). The multiple shrinkage prior distribution of MacLehose and Dunson (2010) does just this: effects that appear to correspond to noise variables are drawn toward a shrinkage location fixed at zero, and those with larger apparent effects should be drawn toward a non-zero shrinkage location. In the case of multinomial models, a different goal guides our selection of this prior distribution: we wish to find groups of parameters that appear to be nearly equal to each other, and shrink within the groupings.

The multiple-shrinkage prior distribution encourages the regression parameters to cluster; however, it does not demand it. We truncate the DPs at the pth term, i.e., we set each $v_p = 1$. This allows each of the p regression parameters for a particular covariate to be assigned to its own cluster. However, the prior distribution disfavors such allocations.

With this background in mind, we now define the MNP models formally. We work with a formulation of the MNP that assumes a latent vector of Gaussian utilities, $W_i = \{w_{ij} : j = 1, \ldots, p\}$, for every observation $i = 1, \ldots, n$. If there are q covariates including the intercept, $x_i = (1, x_{i1}, \ldots, x_{i,q-1})'$, that vary by decision-maker (rather than outcome category), let $X_i = (I_p, x_{i1}I_p, \ldots, x_{i,q-1}I_p)$. Let $\boldsymbol{\mu}_k$, $\boldsymbol{\tau}_k$, $\boldsymbol{\lambda}_k$ and $\boldsymbol{\beta}_k$ each be columns of $p \times q$ matrices, and let $\beta = (\boldsymbol{\beta}_0', \ldots, \boldsymbol{\beta}_{q-1}')'$. We propose two MNP models, one based on Laplace kernels (shown first) and another based on normal kernels. The

model based on Laplace kernels is

$$
\begin{aligned}
D_{0k} &\equiv \text{normal}(c_k, d_k) \times \text{gamma}(a_k, b_k) && (1)\\
(\mu_{jk}, \tau_{jk}) &\overset{\text{ind}}{\sim} \text{truncated-DP}(\alpha, D_{0k}; p) && (2)\\
\lambda_{jk} &\overset{\text{ind}}{\sim} \text{exponential}(2/\tau_{jk}) && (3)\\
\beta_{jk} &\overset{\text{ind}}{\sim} \text{normal}(\mu_{jk}, \lambda_{jk}) && (4)\\
p(W_i) &\propto \varphi(W_i; X_i\beta, I)\mathbf{1}\{\textstyle\sum_j w_{ij} = 0\} && (5)\\
y_i &= \arg\max_j w_{ij}. && (6)
\end{aligned}
$$

For the model based on normal kernels, we replace (3) with

$$\lambda_{jk} = 1/(16\tau_{jk}). \quad (7)$$

The above distributions are parametrized such that the expectation of a gamma(a, b) variate is ab; the expectation of an exponential$(2/\tau)$ variate is $2/\tau$; and, the normal is parametrized by its variance. We use φ to denote the normal density.

In these models, D_{0k} is the base measure related to the kth covariate, and (μ, τ) pairs are drawn from a truncated DP for each outcome category and each covariate. Using (3) mixes the variances over the exponential$(2/\tau)$ distribution, resulting in a Laplace distribution that has a MAP estimate corresponding to a lasso (or L_1-penalized) estimate (Park and Casella 2008; Hans 2009). However, since these Laplace distributions are not merely centered at zero, the prior distribution results in the type of multiple shrinkage described earlier. Forcing the W_i utilities to sum to zero allows us to define this model symmetrically with respect to the category labels, i.e., without a base category. Finally, assuming that the ith decision-maker chooses the category with the highest w_{ij} value completes the probit model specification. We note that the model can accommodate a single shrinkage location, which results in a model that is similar to a Bayesian version of the FHT model (although in the probit rather than logit framework).

For the Laplace version of the model, we specify default hyperparameters following the advice of MacLehose and Dunson (2010). They note that $a_k = b_k = 6.5$ results in unit width prior 95% credible regions — conditional on a shrinkage location — for the Laplace distributions, and they found that this provides meaningful shrinkage without requiring a proliferation of shrinkage locations. Our experience is in accord with this claim. We also specify $\alpha = 1$, $c_k = 0$, and $\sqrt{d_k} = 1.5$, which allows for the existence of strong effects without letting β estimates drift off to $-\infty$ if certain covariate/outcome patterns are not observed in the data. For the normal version of the model, we use the same prior distributions except with hyperparameters $a_k = 1/b_k = 15$, resulting in marginal kernels with approximate unit prior credible width (as is the case with the Laplace kernels) and nearly normal marginal distributions (t_{30}).

Since each $\boldsymbol{\beta}_k$ can take on at most p unique values, we truncate the underlying DPs at the pth component without making the model less general. Therefore, we are able to use the blocked Gibbs sampler described by Ishwaran and James (2002) that is

simpler than corresponding samplers for the full DP, while displaying favorable mixing properties. The details of the estimation algorithm are given in the Appendix.

We note that these models are not likelihood identified. For instance, if each $\boldsymbol{\beta}_k$ is identically equal to a constant C_k, then $\Pr(y_i) = p^{-1}$, regardless of the C_k values. In many cases predictions or fitted selection probabilities (rather than parameter estimates) are of primary importance, so we would argue that this is not particularly worrisome. It would be possible to identify the model by requiring that each $\boldsymbol{\beta}_k$ sums to zero (as in Burgette and Hahn (2010)), but this would seriously complicate the model estimation. Even so, this is still an example of a symmetric MNP model, as the prior is invariant to relabeling the values that y_i can take on. If marginal estimates of β parameters are of primary interest, one could consider post-processing the drawn values into an identifiable scale, in a style similar to that of the McCulloch and Rossi (1994) MNP model. However, we find in practice that the $\boldsymbol{\beta}_k$ blocks nearly do sum to zero merely by the requirement $\sum_j w_{ij} = 0$ for each i. This means that the marginal distributions of β parameters can be interpreted as though they were from a formally identified scale; we provide an example of this in Section 4.

3 Simulation Studies

In this section, we present two sets of simulation results. The first set demonstrates how the prior distributions used in Section 2 can engender multiple shrinkage and motivates potential advantages of using the Laplace versus normal kernels. The second set compares the performances of both the Laplace and normal kernel multiple shrinkage MNPs with two other approaches, as well as with each other.

3.1 Studies of the MNP prior distributions

We begin with a visualization of how the Laplace kernel hyperprior for β allows for multiple shrinkage. Figure 1 displays one simulated realization from this distribution using the default hyperparameters. The DP is truncated at 50 terms, though only three clear peaks are visible. The remaining 47 are close to zero and minimally impact the distribution. If the MNP parameters were drawn from this distribution, we could make the rough interpretation of there being three groups: low, medium-high, and high. As we increase the related covariate, probabilities of the categories that were drawn from the "low" mixture component would become less popular, with mass moving to categories whose parameters were chosen from the "medium-high" and especially the "high" mixture components. Within groups of categories whose parameters were drawn from a particular mixture component, changes to the covariate would result in small changes in relative probabilities.

Similar distributions can be generated from the normal kernel but with a notable difference. The Laplace density produces component distributions that are peaked tightly around their means. As a result, the mixture of Laplace kernels favors posterior distributions for β with many components that are tightly clustered relative to the posterior

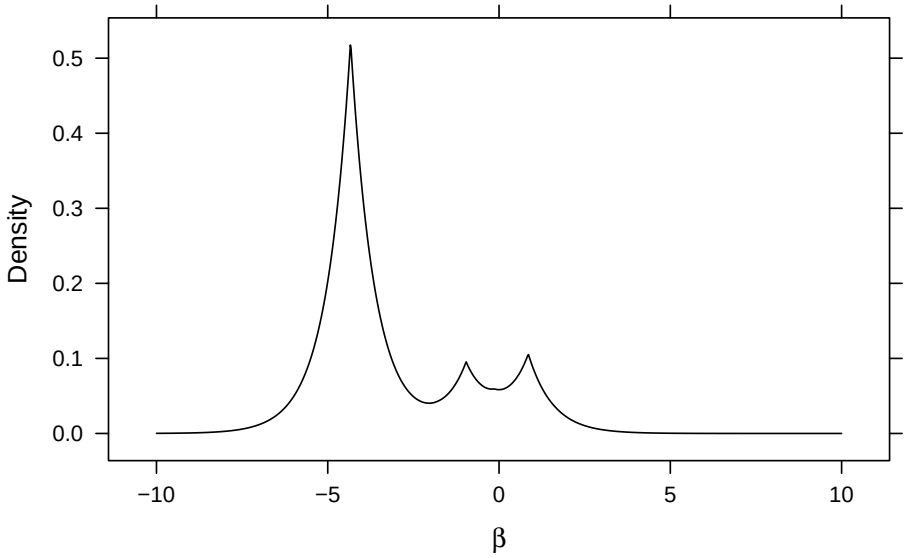

Figure 1: Realization of the prior for regression parameters in the multiple shrinkage MNP.

corresponding to the mixture of normals. To demonstrate this feature of the Laplace kernels, we consider a case with only one true component and where strong shrinkage is obviously desirable. We uniformly draw $n = 500$ outcome variables y from a set of $p = 25$ outcome categories, so that each category has probability .04. We then generate n independent draws of a standard normal covariate x, and regress y on x using a multinomial logit model. Since x is unrelated to y, the true model corresponds to regression coefficients for x, $\beta = \{\beta_j : j = 1, \ldots, 50\}$, equal to zero. Figure 2 displays the fitted probabilities at $x \in \{0, 1\}$ based on posterior means from (1) – (7), as well as those based on maximum likelihood (ML) estimates of β. The Laplace kernel tends to shrink the predicted probabilities closer to each other, and to .04, than the normal kernel does. We note that both methods offer greater shrinkage than the ML estimates, as expected.

3.2 Comparisons of methods

We next compare the performance of the Laplace and normal kernel multiple-shrinkage MNPs to the MNL models of FHT and MT via repeated sampling studies. For each repetition, we simulate $n = 2500$ records with $q - 1 = 2$ covariates, x_1 and x_2, drawn uniformly from $[0, 1]$ and one outcome, y, with $p = 50$ levels. This (n, p) is motivated by the dimensions of the application in Section 4. We generate each y_i, where $i =$

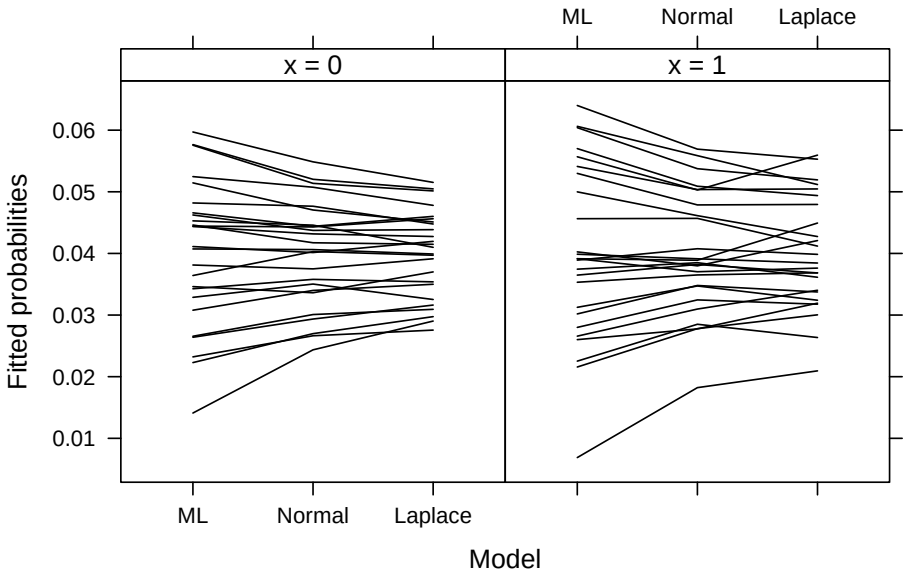

Figure 2: Fitted probabilities where the truth for all categories is 0.04. "ML" is maximum likelihood; "Normal" refers to the multiple shrinkage prior, except with a DP mixture of normals rather than Laplace distributions; "Laplace" is our preferred multiple-shrinkage prior, consisting of a DP mixture of Laplace distributions. Note the stronger shrinkage conferred by the mixture of Laplaces.

$1, \ldots, 2500$, using

$$y_i \overset{\text{ind}}{\sim} \text{multinomial}(1, q_1(x), \ldots, q_p(x)) \tag{8}$$

$$q_j(x) \propto \exp(\beta_{0j} + x_{i1}\beta_{1j} + x_{i2}\beta_{2j}), \text{ where } j = 1, \ldots, p. \tag{9}$$

We note that generating from multinomial logit likelihoods results in a mismatch with the MNP models. Any bias induced by the differing likelihoods will work against the relative performance of the MNP models.

We consider three scenarios for generating each $(\beta_{0j}, \beta_{1j}, \beta_{2j})$, the details of which are summarized in Box 1. Scenario 1 and Scenario 2 are designed so that neither β_{1j} nor β_{2j} are equal across j; thus, the data generators do not *a priori* favor setting groups of regression parameters equal to one another. In Scenario 1, we draw each (β_{1j}, β_{2j}) from homoscedastic Laplace distributions, for which the lasso-type estimates coincide with Bayesian MAP estimators. Hence, the lasso penalty in FHT and MT is, in a sense, the right one to use. In Scenario 2, we draw each (β_{1j}, β_{2j}) from asymmetric distributions, so that the flexibility of prior distributions based on mixture models is desirable. Scenario 3 is designed to favor procedures that set groups of regression parameters equal to one another. To implement this, we draw the regression parameters from distribu-

Box 1: Simulation specifications.

Simulation model:

$$\begin{aligned} y_i &\overset{\text{ind}}{\sim} \text{multinomial}(1, q_1(x), \ldots, q_p(x)) \\ q_j(x) &\propto \exp(\beta_{0j} + x_{i1}\beta_{1j} + x_{i2}\beta_{2j}), \text{ where } j = 1, \ldots, p. \end{aligned}$$

Scenario 1: Unequal coefficients generated from Laplace

$$\begin{aligned} \beta_{0j} &\overset{\text{iid}}{\sim} .2\,\text{normal}(0, 1) \\ \beta_{1j} &\overset{\text{iid}}{\sim} .4\,\text{Laplace}(0, 1) \\ \beta_{2j} &\overset{\text{iid}}{\sim} .4\,\text{Laplace}(0, 1) \end{aligned}$$

Scenario 2: Unequal coefficients generated asymmetrically

$$\begin{aligned} \beta_{0j} &= 0 \\ \beta_{1j} &\overset{\text{iid}}{\sim} 3\,\text{beta}(5, 1) \\ \beta_{2j} &\overset{\text{iid}}{\sim} 3\,\text{beta}(1, 5) \end{aligned}$$

Scenario 3: Mixtures of equal coefficients, varying importance

$$\begin{aligned} \beta_{0j} &\overset{\text{iid}}{\sim} .5\,\text{normal}(0, 1) \\ P(\beta_{1j} = C_1) &= .9, \;\; P(\beta_{1j} = 0) = .1 \\ P(\beta_{2j} = C_2) &= .1, \;\; P(\beta_{2j} = 0) = .9 \end{aligned}$$

For *Low* information, $C_1 = C_2 = 1$.
For *Medium* information, $C_1 = C_2 = 2$.
For *High* information, $C_1 = C_2 = 3$.
For *Mixed* information, $C_1 = 1$ and $C_2 = 3$.

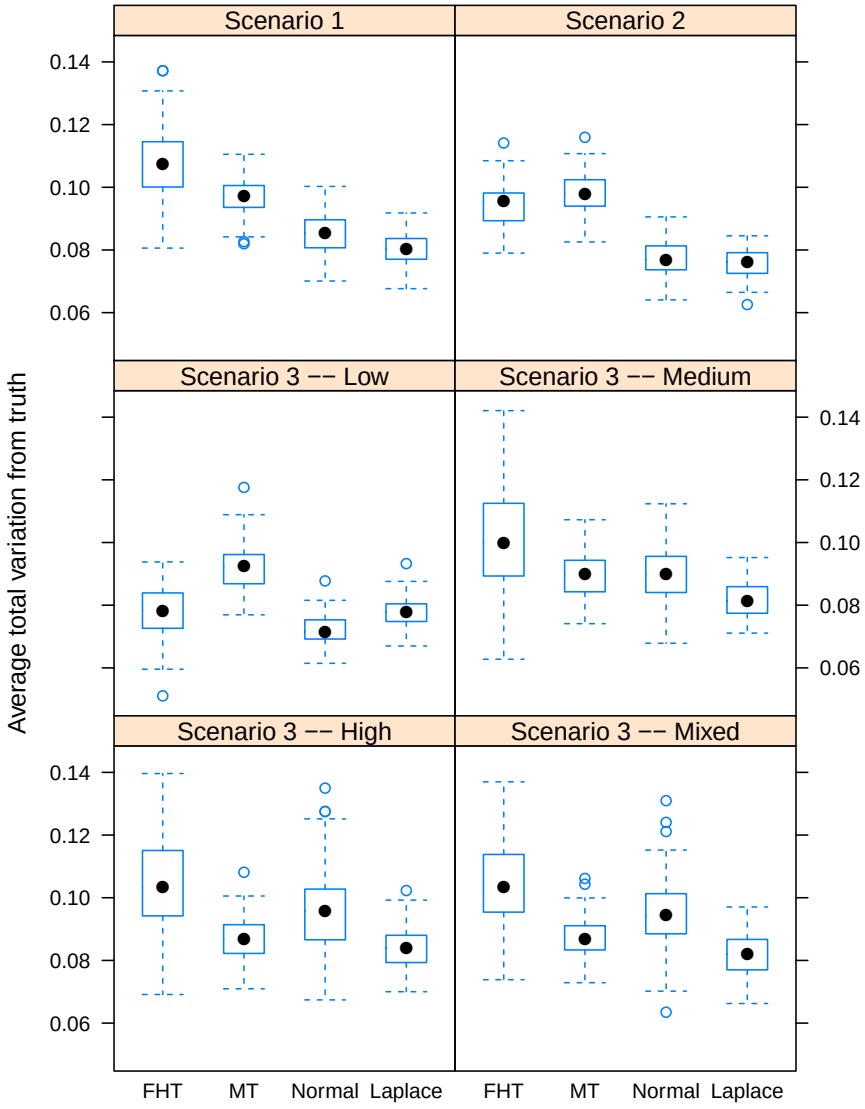

Figure 3: Simulation results. We display average percent total variation from true selection probabilities for FHT, MT and multiple-shrinkage MNP with Laplace kernels and normal kernels. See Box 1 for a description of the data generation. Results based on 100 simulated datasets for each scenario.

tions with two point masses, altering the distances in those masses to reflect differing amounts of predictive power in the covariates. We specify four distinct cases representing differing predictive power in the covariates. These include low, medium, and high signal strengths, and a mixed condition where one set of β parameters corresponds to the "low" signal strength, and the other corresponds to the "high" condition.

To assess performance, we compare the model-based fitted probabilities against the true multinomial probabilities for particular covariate arrangements. We make this comparison via the total variation norm, which is defined to be

$$\mathrm{TV}_{x_i}(P_T, P_E) = .5 \sum_{j=1}^{p} |P_T(y_i = j|x_i) - P_E(y_i = j|x_i)|, \tag{10}$$

where P_T and P_E are the true and estimated multinomial probabilities, respectively (Burgette and Nordheim 2012). This measure is equivalent to

$$\max_{A \in \mathcal{A}} |\Pr_T(y_i \in A|x_i) - \Pr_E(y_i \in A|x_i)|$$

where $\mathcal{A}$ is the set of all subsets of $\{1, \ldots, p\}$. We estimate this difference on a 5×5 grid over the covariate space, and report the average over the grid. We avoid performance metrics based directly on likelihoods or regression parameters because the likelihoods differ between the logit and probit specifications.

The FHT method requires the selection of a tuning parameter that sets the strength of the penalty for non-zero regression parameters. FHT suggest using ten-fold cross-validation to select the tuning parameter, as implemented in their `glmnet` package in R. We follow the default behavior of their software, which uses a deviance criterion in the cross-validation. Cross-validation via prediction error is also an option in their software. The need to choose this tuning parameter (rather than marginalizing over a prior distribution) is one of the major differences between FHT and the MT approach. The MT method is implemented in the `textir` package in R. We use the default settings of the `mnlm` function here.

Figure 3 displays the results from 100 simulation runs of each scenario. In Scenario 1, the two fully Bayesian MNP models perform favorably relative to FHT and MT, even though the data generation closely matches the assumptions of MT and FHT. Evidently, in these simulations the gains from the fully Bayesian analysis outweigh any penalty that might be incurred by assuming a more complex model (i.e., multiple shrinkage locations). In Scenario 2, the data-generating algorithm results in multi-modality when pooled across the length of β and asymmetry within $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. The flexibility of the multiple shrinkage models pays off in this situation. In fact, in Scenario 2, when we examine the estimated coefficients in the MNP methods, we find that allowing multiple shrinkage locations encourages coefficient estimates to be closer to their true values (which are not all zero) than when all are shrunk to the single value zero. We note that there is little performance difference between the two MNP models, though the Laplace kernels slightly outperform the normal kernels, especially in Scenario 1.

In Scenario 3, we discover an interesting set of tradeoffs. In the low information

setting, the normal kernel MNP performs best. The FHT model and Laplace kernel MNP have similar performance, and MT is least effective. In the high information setting, the ordering reverses with MT and the Laplace kernel MNP performing best. In the medium information setting, the Laplace kernel MNP performs best, with FHT as least effective. In the mixed setting — i.e., when one covariate belongs to the high information setting, and another corresponds to the low information setting — the Laplace kernel MNP again performs best, often significantly better than the normal kernel MNP and FHT.

We interpret these results as follows. In the low information setting, the cross-validated approach is quite aggressive in shrinking coefficients to zero, at times even choosing an intercepts-only model, which performs well in this case. The MT prior does not shrink coefficients as strongly to zero due to the use of separate scale parameters for each regression parameter, which results in underperformance in this case. In the scenarios with stronger signals, FHT shrinks some coefficients too far toward zero whereas MT does not, resulting in the relative performances of the methods. Turning to the two MNP models, across all scenarios they tend to allocate coefficients to small numbers of components with non-zero values and modest variances, which better approximates the true values of β and thus explains their strong overall performance.

Examining parameter estimates for the normal and Laplace kernel MNP models in Scenario 3, we find as in Section 3.1 that the Laplace kernel tends to result in tighter groupings of regression parameters within mixture groupings than the normal kernel does. This plays out in their relative performances. In the low information setting, the models tend not to recognize that there are two modes in the β distributions. With this being the case, the stronger within-mixture shrinkage results in poorer performance for the Laplace than the normal kernels. (We also investigated the low information data generation, except modified to have $\boldsymbol{\beta}_0 = 0$. The normal kernels still outperformed the Laplace kernels, indicating that the better performance of the normal kernels in the low information setting is not being driven by the fact that the intercept parameters were drawn from a normal distribution.) In the medium and high information settings, both the Laplace and normal kernels tend to put coefficients in their correct mixture components, but within components the Laplace kernel shrinks coefficients more strongly towards the corresponding component means and hence closer to the true values. In the case of signals of mixed strength, the Laplace kernel is more accurate than the normal kernel, reflecting the relative importance of accurately estimating the large effects.

Taken as a whole, the simulations suggest that the Laplace kernel MNP offers the most favorable performance. Across scenarios, its estimates are never beaten badly by other competing methods, and it often provides the highest predictive accuracy. This is not to say, however, that analysts should always prefer the Laplace (or normal) kernel MNP to the methods of FHT and MT. In particular, both of these methods are orders of magnitude faster than our proposed Gibbs sampler for the MNP models. For example, we ran the MCMC simulations for 6000 iterations (including 1000 discarded for burn-in), which took around 20 minutes on a standard laptop computer. For problems of this size (in terms of n, p and q), FHT fits are typically available in one minute (when using cross-validation, and less otherwise), and the MT method gives results in seconds. For

the extremely large problems considered by MT, the multiple-shrinkage MNP would be infeasible as we currently run it. In fact, MT reports that even the FHT software is unable to manage the large models considered in his paper.

Based on our experience, the multiple shrinkage MNP is most useful in the case of moderate p (say, $20 \leq p \leq 250$) where the sample size is moderate relative to p. When the sample size is large relative to p, the likelihood dominates the prior, minimizing the differences among the various methods, but the Gibbs sampler for the multiple shrinkage models can be slow to run. Since the multiple shrinkage is performed within covariates, we would not recommend the model for small p but large q. In practice, the computational speed is primarily a function of n and p, so — from a computational standpoint — analysts should not be restricted to small q.

When dealing with large multinomial response variables, a combination of approaches may be fruitful. For example, when a very large number of covariates are under consideration, one could use the FHT or MT method to explore many possible models. After having settled on one or a few models of ultimate interest, one could use the multiple-shrinkage MNP (if it is feasible) to form final model estimates, predictions, or inference.

4 Simulating Areal Identifiers via the Multiple-Shrinkage MNP

Government statistical agencies and other data stewards often collect data with areal geographies, such as county or census tract identifiers, that they seek to disseminate as public use files. However, sharing areal identifiers can result in high risks to data subjects' confidentiality, particularly when the data include demographic characteristics that are readily available in external databases. For example, there may be only one person of a certain age, sex, race, and marital status—which may be available to ill-intentioned users at low cost—in a particular county (but many in the state), so that releasing county level indicators carries too high risk of this person being identified in the data. To reduce risks, agencies typically release geographic information only at highly aggregated levels, if at all. For example, the U.S. Health Insurance Portability and Accountability Act mandates that released geographic units comprise at least 20,000 individuals; and, the U.S. Bureau of the Census does not release public use files with geographic identifiers of areas with fewer than 100,000 people (Wang and Reiter 2012). Such disclosure limitation requirements degrade the utility of data for legitimate users, especially for analyses that would benefit from finer spatial resolution. Further, aggregation can create or magnify ecological inference fallacies (Robinson 1950; Freedman 1999).

We propose an alternative to aggregation for releasing areal geographic information: release values of areal identifiers that are simulated from models designed to preserve spatial relationships among the attributes in the data. This is an example of what is known as partially synthetic data in the literature on statistical disclosure limitation

(Reiter 2003). To describe this approach more fully, we modify the scenario of Wang and Reiter (2012), who used tree-based models to simulate latitudes and longitudes of respondents' home addresses. Suppose that a statistical agency has collected data on a random sample of 10,000 heads of households in a state. The data comprise each person's census tract, age, sex, and education. Suppose that combining census tract, age, sex, and education uniquely determines a large percentage of records in the sample and the population, but that that age, sex, and education without census tract do not uniquely identify many people. Therefore, the agency wants to replace census tract for all people in the sample to disguise their identities. The agency fits an MNP model of census tract on age, sex, and education, and for each person generates a draw from the predictive distribution of census tract. These simulated values replace the actual census tracts, and the result is one partially synthetic dataset. The agency repeats this process say ten times, and these ten datasets are released to the public to enable inference via methods akin to multiple imputation combining rules (Reiter and Raghunathan 2007).

A related partially synthetic data approach is used to protect locations in the Census Bureau's OnTheMap project (Machanavajjhala et al. 2008). In that project, Machanavajjhala et al. (2008) synthesize the street blocks where people live conditional on the street blocks where they work and other block-level attributes. They use multinomial regressions to simulate home-block values, constraining the possible outcome space for each individual based on where they work. This constraint, which avoids the task of estimating large multinomial regressions, is somewhat particular to the setting of OnTheMap. For example, this constraint would not sensibly apply in the typical demographic survey with only one areal location per individual. The multiple shrinkage MNP model does not require these constraints. We also note that the approach of Wang and Reiter (2012) differs from the multiple shrinkage MNP approach, since they consider point-referenced data whereas we use data attached to areal identifiers.

To illustrate partial synthesis of areal geographies, we consider data that record the causes of all deaths for the year 2007 in Alamance, Durham, Orange, and Wake counties of North Carolina. These counties include the Raleigh, Durham, and Chapel Hill communities. These mortality data are in fact publicly available and so do not require disclosure protection. Nonetheless, they are ideal test data for methods that protect confidentiality of geographies since, unlike many datasets on human individuals, locations are available and can be revealed for comparisons. Similar data (but point-referenced) from 2002 were used by Wang and Reiter (2012).

In 2007, in these counties 7373 residents passed away. The deaths were spread among 200 census tracts. We seek to simulate new values of every person's census tract identifier, leaving other attributes at their original values. To do so, we use the Laplace kernel MNP model from Section 2 to estimate the probability that person i was from the jth census tract, where $j = 1, \ldots, 200$, as a function of several attributes on the file. These include indicators for age 18 or under, age greater than 65, race of black/non-black, and whether the cause of death was recorded as being cardiac-related or not. We expect the multiple shrinkage framework to be desirable for the race variable in particular, since the data exhibit racial clustering over tract-level geography. To fit the model, we run the MCMC for 50,000 iterations, storing every 10th draw.

Figure 4: Plot of probabilities of assignment to census tracts for a non-black respondent, aged greater than 65 years, who died of a non-cardiac cause. The colors correspond of deciles of the multinomial probabilities, with white corresponding to low probability, and black corresponding to high.

The results of the Laplace kernel MNP model indicate that there are strong spatial associations in the data. For example, Figures 4 and 5 display tract probabilities $\Pr(y_i = j)$ for people older than age 65 who died of a non-cardiac cause for black and non-black races, respectively. The eastern-most county in these plots is Wake; the city of Raleigh is at its center. The region of small census tracts to the north and west of Raleigh in the adjoining county is the city of Durham. Durham is characterized by a relatively high proportion of black residents, especially compared to the west and north portions of Raleigh. The fitted probabilities reflect this, with much of the mass shifting from Raleigh to Durham when we change race from non-black (Figure 4) to black (Figure 5).

To demonstrate the extent to which synthetic data generated from the MNP model preserve the associations of the observed variables, we create $m = 20$ partially synthetic datasets by sampling 20 times from the posterior predictive distributions of the census tracts. We then apply spatial simultaneous autoregressive lag models (e.g., Banerjee et al. 2004) to each of the synthetic spatial datasets, and combine the results according to the rules derived in Reiter (2003). In particular, the spatial regression models take on the form

$$Y = \rho VY + X\beta + \varepsilon.$$

Figure 5: Plot of probabilities of assignment to census tracts for a black respondent, aged greater than 65 years, who died of a non-cardiac cause. The colors correspond of deciles of the multinomial probabilities, with white corresponding to low probability, and black corresponding to high.

Here, Y is a p-vector with a single measure from each tract, and V is a right stochastic matrix defined as follows. Let $\mathcal{N}_j$ contain the tract identifiers of the regions that border tract j. The jth row of V has elements $1/|\mathcal{N}_j|$ in the columns corresponding to the elements in $\mathcal{N}_j$ and zeros elsewhere. Hence, the Y value in each cell is assumed to consist of a fraction of the average value from the neighboring tracts, a contribution from a linear regression, and a normal additive error. The scalar ρ therefore captures an extent of the spatial association net of the covariates. To find the maximum likelihood estimates of these models, we use the `spautolm` function in the `spdep` package in R (Bivand 2011).

We begin with a model of the tract-level rates of cardiac-related deaths. We model this as a function of tract-level rates of young (age $\leq$ 18), old (age $>$ 65), and black study subjects. In the spatial models, we drop the 12 tracts that had fewer than 10 records in the genuine data since the tract-level rates for these units are highly volatile and can degrade the estimates, though they were included in the model that produces the synthetic spatial identifiers themselves.

Table 1 summarizes the results. The cause-of-death variable does not have a strong

	Synthetic geography		True geography	
Parameter	Estimate	95% CI	Estimate	95% CI
Intercept	0.213	(0.073, 0.353)	0.193	(0.074, 0.311)
Young	-0.194	(-0.687, 0.298)	-0.103	(-0.485, 0.279)
Old	0.237	(0.099, 0.375)	0.221	(0.107, 0.335)
Black	-0.049	(-0.151, 0.053)	-0.007	(-0.063, 0.048)
ρ	0.041	(-0.183, 0.265)	0.105	(-0.106, 0.316)

Table 1: Parameter estimates from the spatial simultaneous autoregressive lag model of tract-level percent cardiac-related deaths. The "synthetic geography" aggregates the variables according to the synthetic spatial identifiers that result from the multiple-shrinkage MNP. The explanatory variables are the tract-level means of the indicated traits.

spatial pattern, and the imputations preserve this: the estimates of ρ from the synthetic and genuine data are not significantly different from zero. The imputed tract identifiers also do a good job of preserving the β parameter estimates. The 95% confidence intervals from the imputed sets cover the estimates from the true data. (Since both the response and covariate values are aggregated by tract in this model, variables on both sides of the regression equation are changing with each set of imputed geographic identifiers.)

We also switch the roles of race and cause-of-death in the spatial regression. Although such a model (i.e., one predicting race) is not of great substantive interest, it does offer a test of the synthesizer in an application with strong spatial patterns. We summarize the results in Table 2. Here again, the imputed geography preserves many of the key features of the data. The measure of spatial association ρ is estimated to be strongly significant in both sets of regressions: the corresponding confidence interval from the synthetic data covers the value from the true data. The synthetic geographic identifiers preserve (in-)significance of the β parameters. Although two of the corresponding confidence intervals do not cover the estimates from the true data, they barely miss doing so. Finally, if the insignificant regression parameters are dropped so that we model percent black as a function of percent old study subjects, all of the confidence intervals from the synthetic data cover the values estimated using the true data.

The process of imputing new spatial identifiers would not be worthwhile if we were preserving statistical relationships between the observed variables simply by preserving the true tract identifier values. To assess the extent to which the synthetic identifiers were changed from their original values, we examine the $m = 20$ sets of imputed tract identifiers that were used to perform the spatial regressions described above. For 6164 records — just shy of 85% — none of the 20 imputed identifiers was the true one. For 99.4% of the records, the true identifier was imputed zero or one times.

As further evidence that the MNP model moves census tracts around, consider the simplistic approach to breaking confidentiality of taking the records that have a single tract imputed several times and assuming that the most frequently-imputed value is the

	Synthetic geography		True geography	
Parameter	Estimate	95% CI	Estimate	95% CI
Intercept	0.454	(0.320, 0.587)	0.591	(0.419, 0.762)
Young	0.038	(-0.578, 0.654)	-0.625	(-1.386, 0.136)
Old	-0.436	(-0.602, -0.269)	-0.644	(-0.861, -0.427)
Cardiac	-0.085	(-0.265, 0.094)	-0.126	(-0.411, 0.159)
ρ	0.514	(0.360, 0.669)	0.663	(0.541, 0.785)

Table 2: Parameter estimates from the spatial simultaneous autoregressive lag model of tract-level percent black study subjects. The "synthetic geography" columns aggregate the variables according to the synthetic spatial identifiers that result from the multiple-shrinkage MNP. The explanatory variables are the tract-level means of the indicated traits.

true one. Among the 653 records that had the same identifier imputed three times, it was correct 127 times; 51 records had the same tract imputed four times, though none was correct; and, three records had the same identifier imputed five times, though in only one case was it correct. In short, if a potential data intruder took the most common identifier as the true one, more often than not he would be wrong. Although this is not a formal disclosure risk assessment—see Reiter and Mitra (2009) for formal risk assessment approaches for synthetic categorical data—it does suggest that the favorable preservations of the spatial associations shown in Tables 1 and 2 are not the result of inadequate shuffling of the true identifiers.

As a final note on this analysis, we return to the identifiability issue noted in Section 2. The Laplace kernel MNP model is not technically identified as it is described: for each added covariate, we can only identify $p-1$ parameters rather than the p parameters that enter into the model. However, the model is identified if we require that each group of p parameters sums to zero. This is the identifying restriction that corresponds to forcing the latent W_i to sum to zero, which we do enforce. In this application, we find that the loss of identification is minor, because the sampled $\boldsymbol{\beta}_k$ parameters (where $k = 1, \ldots, q$) in practice nearly *do* have mean zero, even though the model does not demand it. Figure 6 displays trace plots of the mean of each group of regression parameters (i.e., $\boldsymbol{\beta}_k$ for $k = 1, \ldots, q$). These numbers are centered around zero and small in magnitude relative to the estimated effects, so the under-identification is not important. Thus, the marginal distributions of the estimated parameters honestly reflect uncertainty. Heuristically, we expect the iteration-by-iteration average of the groups of parameters to be closest to zero when p is relatively large, but this property is easy to check from output of the MCMC.

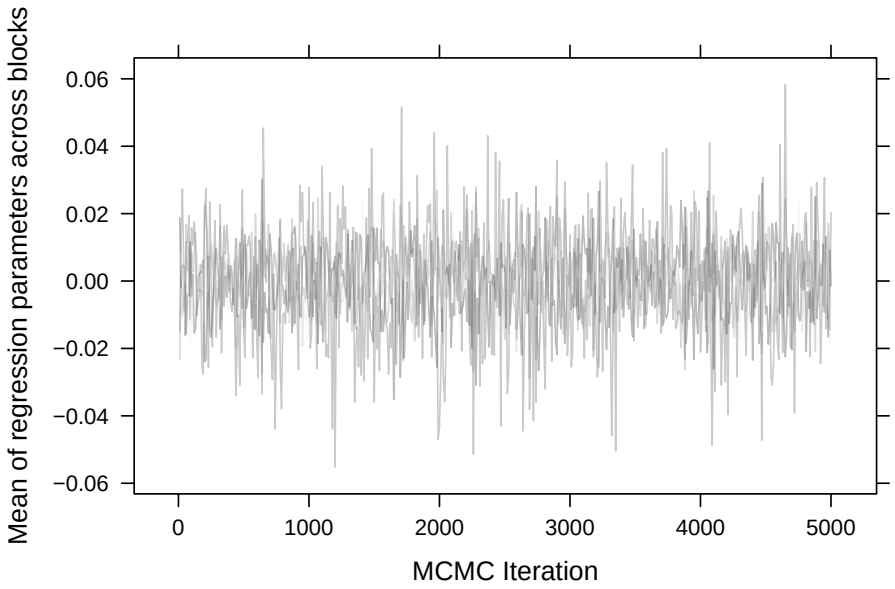

Figure 6: Trace plots of means of regression parameters across blocks of p parameters that relate to each covariate. If the model were fully identified these quantities would be identically zero. On average the magnitude of the deviations from zero is 0.008, which indicates that the under-identification is small. Results are thinned to every 10th stored draw.

5 Concluding Remarks

There are several ways in which one could extend the MNP model while working within the proposed multiple-shrinkage framework. For example, one could use the hierarchical DP of Teh et al. (2006) to encourage similar shrinkage patterns across some or all of the covariates. Further, this model is built on an i.i.d. normal error structure. One could consider more general substitution patterns, i.e., the way in which probabilities change if one outcome category is removed from consideration, by allowing for more general covariance structures. A good deal of care would have to be taken in doing this, since standard inverse-Wishart-based MNP models (e.g., McCulloch and Rossi 1994; Imai and van Dyk 2005; Burgette and Hahn 2010) often encounter numerical problems in the form of degenerate covariance matrices when p is even moderately large (say, $p = 10$).

When applying the MNP model to synthesize areal identifiers, it is important to recognize that the MNP model does not fully account for the spatial structure in the data. Areal adjacencies are not part of the synthesis so that, for example, individuals living in the same or adjacent tracts in the original data may be far away from one another in the simulated data. Further, areal adjacencies are not used to estimate

parameters in the MNP model. An alternative model might encourage each tract *a priori* to have associated regression parameters that are similar in some way to the regression parameters of its neighbors.

Synthesizing areal geographies may not suffice to protect confidentiality; it may be necessary to simulate values of non-geographic variables as well (e.g., age, race, marital status). One approach is to simulate from hierarchical, area-level spatial models (Banerjee et al. 2004), which can be challenging with large datasets. Another strategy is to mask attribute data using spatial smoothing techniques (Zhou et al. 2010). We note that applying either of these approaches alone, i.e., without simulating geography, leaves the original areal geographies on the file, which may result in too high disclosure risks. An open research question involves quantifying the trade offs in disclosure risk and data quality for different amounts of synthesis, e.g., simulating areal identifiers plus only age versus simulating age, race, and marital status.

In some contexts, n or p may be too large to estimate the MNP models efficiently with fully Bayesian approaches. Nonetheless, there are settings in which our methods apply directly. For example, many state-wide or national cancer registries publish counts of cancer incidence by subjects' sex, race, age (typically categorized), and cancer type. Doing so in moderately aggregated regions like census tracts may represent too high disclosure risks. Instead of suppressing the tract-level counts, the agency can use the MNP to synthesize these regions, perhaps after stratifying on larger aggregates like counties to facilitate computation and preserve counts within the larger aggregates.

Although challenges remain, we anticipate that the MNP model presented here will help researchers in a range of fields — economics, marketing, and sociology, among others — construct flexible and principled models of categorical variables with many categories.

References

Banerjee, S., Carlin, B., and Gelfand, A. (2004). *Hierarchical modeling and analysis for spatial data*, volume 101. Chapman & Hall/CRC. 468, 473

Bivand, R. (2011). *spdep: Spatial dependence: weighting schemes, statistics and models*. R package version 0.5-41.
URL http://CRAN.R-project.org/package=spdep 469

Blackwell, D. and MacQueen, J. (1973). "Ferguson distributions via Pólya urn schemes." *The Annals of Statistics*, 1(2): 353–355. 455

Blei, D., Ng, A., and Jordan, M. (2003). "Latent Dirichlet allocation." *The Journal of Machine Learning Research*, 3: 993–1022. 456

Burda, M., Harding, M., and Hausman, J. (2008). "A Bayesian mixed logit-probit model for multinomial choice." *Journal of Econometrics*, 147(2): 232–246. 455

Burgette, L. F. and Hahn, P. R. (2010). "Symmetric Bayesian multinomial probit models." *Duke University Statistical Science Technical Report*, 1–20. 455, 459, 472

Burgette, L. F. and Nordheim, E. V. (2012). "The trace restriction: An alternative identification strategy for the Bayesian multinomial probit model." *Journal of Business and Economic Statistics*, 30(3): 404–410. 464

Carvalho, C., Polson, N., and Scott, J. (2010). "The horseshoe estimator for sparse signals." *Biometrika*, 97(2): 465–480. 457

Cawley, G., Talbot, N., and Girolami, M. (2007). "Sparse multinomial logistic regression via Bayesian L1 regularisation." *Advances in neural information processing systems*, 19: 209–216. 455

De Blasi, P., James, L., and Lau, J. (2010). "Bayesian nonparametric estimation and consistency of mixed multinomial logit choice models." *Bernoulli*, 16(3): 679–704. 455

Duncombe, W., Robbins, M., and Wolf, D. (2001). "Retire to where? A discrete choice model of residential location." *International Journal of Population Geography*, 7(4): 281–293. 456

Ferguson, T. (1973). "A Bayesian analysis of some nonparametric problems." *The Annals of Statistics*, 1(2): 209–230. 455

Freedman, D. A. (1999). "Ecological Inference and the Ecological Fallacy." In Smelser, N. J. and Baltes, P. B. (eds.), *International Encyclopedia of the Social Sciences*, volume 6, 4027–4030. Elsevier. 466

Friedman, J., Hastie, T., and Tibshirani, R. (2010). "Regularization paths for generalized linear models via coordinate descent." *Journal of Statistical Software*, 33(1): 1–22. 455, 456, 458, 460, 463, 464, 465, 466

Hans, C. (2009). "Bayesian lasso regression." *Biometrika*, 96(4): 835–845. 458

Imai, K. and van Dyk, D. (2005). "A Bayesian analysis of the multinomial probit model using marginal data augmentation." *Journal of Econometrics*, 124(2): 311–334. 472

Ishwaran, H. and James, L. (2002). "Approximate Dirichlet Process computing in finite normal mixtures." *Journal of Computational and Graphical Statistics*, 11(3): 508–532. 458, 477

Kim, J., Menzefricke, U., and Feinberg, F. (2004). "Assessing heterogeneity in discrete choice models using a Dirichlet process prior." *Review of Marketing Science*, 2: 1–39. 455

Krishnapuram, B., Carin, L., Figueiredo, M. A. T., and Hartemink, A. J. (2005). "Sparse multinomial logistic regression: Fast algorithms and generalization bounds." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 957–968. 454

Lenk, P. and Orme, B. (2009). "The value of informative priors in Bayesian inference with sparse priors." *Journal of Marketing Research*, 46: 832–845. 454

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., and Vilhuber, L. (2008). "Privacy: Theory meets practice on the map." In *IEEE 24th International Conference on Data Engineering*, 277–286. 467

MacLehose, R. and Dunson, D. (2010). "Bayesian Semiparametric Multiple Shrinkage." *Biometrics*, 66(2): 455–462. 454, 456, 457, 458

McCulloch, R. and Rossi, P. (1994). "An exact likelihood analysis of the multinomial probit model." *Journal of Econometrics*, 64(1): 207–240. 459, 472

McFadden, D. (1978). "Modelling the choice of residential location." In Karlqvist, A., Lundqvist, L., Snickars, F., and Weibull, J. (eds.), *Spatial Interaction Theory and Planning Models*, 75–96. Amsterdam: North-Holland. 456

Park, T. and Casella, G. (2008). "The Bayesian lasso." *Journal of the American Statistical Association*, 103(482): 681–686. 458

Reiter, J. (2003). "Inference for partially synthetic, public use microdata sets." *Survey Methodology*, 29(2): 181–188. 467, 468

Reiter, J. P. and Mitra, R. (2009). "Estimating risks of identification disclosure in patrially synthetic data." *Journal of Privacy and Confidentiality*, 1: 99–110. 471

Reiter, J. P. and Raghunathan, T. E. (2007). "The multiple adaptations of multiple imputation." *Journal of the American Statistical Association*, 102(480): 1462–1471. 467

Robinson, W. S. (1950). "Ecological correlations and the behavior of individuals." *American Sociological Review*, 15: 351–357. 466

Sethuraman, J. (1994). "A constructive definition of Dirichlet priors." *Statistica Sinica*, 4: 639–650. 456

Sha, N., Vannucci, M., Tadesse, M., Brown, P., Dragoni, I., Davies, N., Roberts, T., Contestabile, A., Salmon, M., Buckley, C., et al. (2004). "Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage." *Biometrics*, 60(3): 812–819. 454

Shahbaba, B. and Neal, R. (2009). "Nonlinear models using Dirichlet process mixtures." *The Journal of Machine Learning Research*, 10: 1829–1850. 455

Taddy, M. (2012). "Multinomial inverse regression for text analysis." *Journal of the American Statistical Association*. Forthcoming. 455, 456, 460, 463, 464, 465, 466

Teh, Y., Jordan, M., Beal, M., and Blei, D. (2006). "Hierarchical Dirichlet processes." *Journal of the American Statistical Association*, 101(476): 1566–1581. 472

Wang, H. and Reiter, J. (2012). "Multiple imputation for sharing precise geographies in public use data." *Annals of Applied Statistics*, 6(1): 229–252. 466, 467

Zhou, Y., Dominici, F., and Louis, T. A. (2010). "A smoothing approach for masking spatial data." *Annals of Applied Statistics*, 4(3): 1451–1475. 473

Appendix: MCMC details

Updating utilities

We use the following lemma: If $(x', y')' \sim \text{normal}((\mu_x', \mu_y')', \Sigma)$, where we have the partitioning

$$\Sigma = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}$$

then $(y|x) \sim \text{normal}(\mu_{y|x}, \Sigma_{y|x})$ where

$$\Sigma_{y|x} = \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}$$

and

$$\mu_{y|x} = \mu_y + \Sigma_{yx}\Sigma_{xx}^{-1}(x - \mu_x).$$

We consider the distribution of $W_i^* \equiv (w_{i,1}, w_{i,2}, \ldots, w_{i,p-1}, \bar{w}_i)'$. If $y \sim \text{normal}(0, I)$, then $W_i^* \sim \text{normal}(0, TT')$ where

$$TT' = \begin{bmatrix} 1 & 0 & \cdots & 0 & p^{-1} \\ 0 & & & & \\ \vdots & & I_{p-2} & & \vdots \\ 0 & & & & \\ p^{-1} & & \cdots & & p^{-1} \end{bmatrix}.$$

For our sampler, we will be interested in the distribution of $w_{i,1}|W_{i,-1}^*$. Via a draw from this conditional distribution, we infer a draw from the conditional distribution of $(w_{i,1}, w_{i,p})$, taking the sum-to-zero restriction into account. Dropping the i subscripts, this conditional variance is given by

$$\Sigma_{w_1|W_{-1}^*} = 1 - [0, \ldots, 0, p^{-1}]\Sigma_{W_{-1}^*, W_{-1}^*}^{-1}[0, \ldots, 0, p^{-1}]'.$$

We calculate

$$\Sigma_{W_{-1}^*, W_{-1}^*}^{-1} = \begin{bmatrix} I_{p-2} + .5J_{p-2}J_{p-2}' & -.5pJ_{p-2} \\ -.5pJ_{p-2}' & .5p^2 \end{bmatrix}.$$

It follows that

$$\Sigma_{w_1|W_{-1}^*} = 1 - p^{-2}(.5p^2) = 0.5.$$

Similarly,

$$\mu_{w_1|W_{-1}^*} = \mu_{w_1} + [-.5J_{p-2}', .5p](W_{-1}^* - \mu_{W_{-1}^*}).$$

Once we have these conditional distributions, we only need to derive the truncations. For simplicity, we will jointly sample the jth and y_ith elements of W_i. In this case, we have

$$w_{ij} \leq \sum_{k \notin \{j, y_i\}} w_{ik} + \min_{k \notin \{j, y_i\}}(w_{ik}).$$

Other MCMC details

After updating the latent utilities, the algorithm proceeds as follows:

1. Update (w_{ij}, w_{iy_i}) for all i and $j \neq y_i$.

2. Update the regression coefficients:

$$\beta \sim \text{normal}(\hat{\beta}, \hat{\Sigma}_\beta)$$

where $\hat{\Sigma}_\beta = (X'X + \Lambda^{-1})^{-1}$ and $\hat{\beta} = \hat{\Sigma}_\beta(X'z + \Lambda^{-1}\mu_0)$.

3. Update the mixing parameters matrix Λ:

$$\lambda_j \overset{\text{ind}}{\sim} \text{inv-Gaussian}(a, b)$$

where $a = \sqrt{\tau_{k_j}}/|\beta_j - \mu_{k_j}|$ and $b = \tau_{k_j}$.

4. Update $\{(\mu_t, \tau_t)\}_{t=1}^p$ via

$$\begin{aligned} \mu_t &\overset{\text{ind}}{\sim} \text{normal}(b_t, B_t) \\ \tau_t &\overset{\text{ind}}{\sim} \text{gamma}(n_t + a_1, 1/(\sum_{j:k_j=t} \lambda_j/2 + 1/b_1)) \end{aligned}$$

where $B_t = (1/d + \sum_{j:k_j=t} 1/\lambda_j)^{-1}$ and $b_t = B_t(c/d + \sum_{j:k_j=t} \beta_j/\lambda_j)$.

5. Update p according to the truncated stick-breaking scheme outlined in Ishwaran and James (2002). Draw

$$v_k \overset{\text{ind}}{\sim} \text{beta}(1 + r_k, \alpha + \sum_{l=k+1}^{N} r_l)$$

and set

$$p_k = v_k \prod_{j=1}^{k-1}(1 - v_j)$$

where r_k counts the number of β components assigned to the kth mixture component.

6. Update the vector of coefficient configurations k. For each i and j, draw from

$$\Pr(k_{ij}) \propto p_{ij} N(\beta_{ij}|\mu, \lambda_j)\exp(\lambda_j|2/\tau_l).$$

Making predictions

We will be interested in making draws from the posterior predictive distribution of the model. To do so, we need to make draws from a multivariate normal that preserve the sum-to-zero property of the W_i vectors. As above, we operate on w_i^*, though this time conditioning only on its last element, which is defined to be $\bar{w}_i$. Doing so, we see that our draws should be from a normal distribution with mean $\{X_i\beta\}_{-p} - p^{-1}(J_p' X_i\beta)J_{p-1}$ and variance $I_{p-1} - p^{-1}J_{p-1}J_{p-1}'$. A single draw from this distribution can be used to impute the areal identifier; repeated draws give Monte Carlo estimates of the assignment probabilities.

Acknowledgments

The authors wish to thank the reviewers and editors for very helpful comments that improved the manuscript. This research was supported by National Institute of Health grant 1R21AG032458-01A1. The research was carried out when the first author was a postdoctoral research associate in the Department of Statistical Science at Duke University.