Biometrika (2013), **xx**, x, *pp.* 1–23

The mismatched Gaussians prior for structured model dependence

By Kristian Lum and James Johndrow

Network Dynamics and Simulation Science Laboratory,
Virginia Tech and
Duke University Department of Statistical Science

Summary

Here we develop a novel class of probability distributions on the unit interval, which we refer to as mismatched Gaussian (MG) distributions. In the univariate setting, MG distributions form a similar class to the Beta family. However, unlike the Beta, MG has a convenient Gaussian latent variable representation that facilitates generalization to the multivariate case and allows for efficient computation by MCMC. There are many applications where a computationally tractable multivariate distribution on probabilities is desirable, but we focus specifically on inclusion probabilities for Bayesian variable selection on *a priori* clustered data. Dependent inclusion probabilities are of particular interest in spatial and time series applications and in hierarchical models, where natural clusters are created by the spatio-temporal or hierarchical structure, and often the set of important predictors varies over clusters. In these cases, the number of observations in each cluster may be small, so it is important to borrow information across clusters in estimating inclusion probabilities, while retaining the tendency to concentrate prior mass around sparse coefficient vectors in any particular local model. Multivariate MG priors accomplish this while also retaining the automatic multiplicity correction that is a hallmark of fully Bayesian variable selection. We illustrate the method with several simulation examples and an application to estimation of the number of uncounted deaths during a violent conflict in Colombia.

Some key words: variable selection, dependent, spatial, time series, copula, data augmentation, MCMC, capture-recapture

1. Introduction

This paper introduces a class of distributions on the unit interval that have straightforward multivariate generalizations and give rise to easy computational algorithms. These distributions, which we refer to as mismatched Gaussian (MG) distributions, are of particular interest as multivariate priors on inclusion probabilities in Bayes variable selection (BVS) for regression models. The basic motivation for dependent inclusion probabilities is data that can be clustered *a priori* where the appropriate model varies by cluster. For example, one might wish to perform variable selection in GLMs estimated on data from a stratified survey, where the important covariates are expected to differ across strata, but where it is desirable to borrow strength in estimating variable inclusion probabilities. Another example is areal or point-referenced spatial data with multiple observations per

area or location, where one expects the set of important covariates to be more similar for nearby regions than distant ones. These types of data arise commonly in social science applications, yet there currently exists no method for flexibly smoothing inclusion probabilities across the component models. Our multivariate MG distributions easily handle this situation. While we focus on time-series, spatial, and survey settings, numerous extensions to other applications – including many unrelated to Bayes variable selection – are possible.

Bayesian variable selection typically involves the use of mixture priors on regression coefficients, e.g.:

$$Y = X\beta + \epsilon \tag{1}$$

$$p(\beta_j) = \pi\delta_0 + (1-\pi)N(0, \tau_0^2) \tag{2}$$

where δ_0 is the Dirac delta function and $\pi \in (0, 1)$. This topic has been extensively studied (see e.g. George & McCulloch (1997)), though only recently has it been shown that hyperpriors on π are necessary to achieve automatic multiplicity correction in these models (Scott & Berger (2010)). In general, treatments of BVS in which hyperpriors are employed have chosen a Beta(a, b) prior on π. BVS is one of many approaches to inducing sparsity in regression models, but is theoretically one of the most appealing. Asymptotically, BVS achieves near-optimal posterior contraction rates around sparse vectors Castillo & van der Vaart (2012). Recent work in the compressed sensing literature has shown that the (computationally intractable) maximum-likelihood analogue of BVS nearly achieves information theoretic lower bounds for recovery of sparse vectors in finite samples (Reeves & Gastpar (2012)). These strong theoretical results confirm the emphasis in the Bayesian literature on BVS as the gold-standard for sparsity inducing models.

Unfortunately, when employing BVS for p larger than about 30, it becomes impossible to explore the entire model space, even with modern cluster computing resources and despite the fact that with Beta priors on p, the posterior model probabilities are available in closed form. Some authors have focused on efficient algorithms for exploring high-probability regions of the sample space, making use of parallel and graphical computing to greatly expand the number of models that can be visited (see Hans et al. (2007)). Nonetheless, these algorithms can only explore a tiny fraction of the full model space, and thus others have focused on the development of shrinkage priors that approximate the sparsity-inducing properties of BVS at a much lower computational cost. A wide variety of priors exist, with recent work focusing on the local-global family of shrinkage priors. Within this family, the horseshoe (Carvalho et al. (2010)), which employs half-Cauchy priors on both the local and global scales, deserves special note for its consistently strong out of sample predictive performance. A notable recent paper proposes a new member of the local-global family with optimal posterior contraction rates in sparse problems (see Bhattacharya et al. (2012)).

The vast majority of recent work, including that cited above, has related to large p, small n cases, and thus the focus has been on approximating the full BVS posterior in these cases either through shrinkage priors or stochastic search algorithms. However, many modern applications do not fit this paradigm. Often, while p is small relative to n, we expect a substantial degree of structure in the data and wish to employ a hierarchical modeling framework in which data are clustered *a priori*. This is commonly encountered in all types of survey data and in spatial and temporal sampling. In these cases, the number of observations in each cluster is often not large relative to p. Moreover, we expect

substantial variability across the local models with regard to the set of variables that are important. The standard Bayesian approach to this problem is the use of hierarchical priors to flexibly borrow information and induce dependence across the clusters. This becomes quite challenging with Beta priors on inclusion probabilities. While numerous multivariate generalizations of the Beta distribution have been proposed that could in theory be used to model dependent inclusion probabilities (see e.g. Trippa et al. (2011)), they are practically quite difficult to work with and lead to challenging computation. This work is motivated by the need for alternative priors on π that allow for dependence in inclusion probabilities across constituent local models while maintaining the multiplicity correction and leading to computationally feasible MCMC algorithms for computation. The priors we propose here are by comparison extremely simple to implement, adding only an additional layer of hierarchical Gaussian latent variables to the model. Moreover, the latent Gaussian representation of the MG distribution allows access to a vast array of methods for multivariate normal distributions to be applied directly to the modeling of dependent inclusion probabilities. We illustrate numerous such applications in this paper.

We have found no direct precedent for this work. One method has been proposed for the special case of lattice data (Smith & Fahrmeir (2007)), which utilized an Ising prior to induce dependence in inclusion probabilities between neighboring grid points. However, the method does not extend naturally to the more general time series and spatial settings that we consider here, where time points/locations generally do not reside on a lattice, and is not applicable to other types of grouped data. A somewhat more apt comparison can be made to fused Lasso, which includes the usual L_1 penalty on regression coefficients as well as an additional penalty on successive differences between regression coefficients in time-series or other settings in which the data are naturally ordered. This approach tends to result in smoothing across time of the sparsity pattern in the lasso paths.

While existing methods may address certain niche applications, they do not extend easily to the general local model variable selection problem that we consider here. Moreover, the regularization-based approaches do not provide estimates of uncertainty in model parameters. The fully Bayesian approach that we propose here is much more general and can be used in virtually any setting where the data can be naturally grouped, ordered, or arranged spatially. In addition, our method allows for formal characterization of uncertainty in model parameters via the posterior distribution. Our method is also computationally straightforward and can be used in a variety of regression applications. By combining this representation of covariate inclusion indicators with any conditionally normal representation of a likelihood, we achieve a fully conjugate Bayesian model with very straight-forward computation. Because many GLM likelihoods can be represented as conditionally normal using data augmentation methods, our approach can easily be incorporated in many GLM settings without sacrificing full conjugacy. We propose a Markov Chain Monte Carlo (MCMC) algorithm for estimation, which also benefits from the full conjugacy of the model, as we are able to marginalize over the regression coefficients to obtain good mixing.

The rest of this article is organized as follows. Section 2 introduces our new class of priors for variable selection, referred to as Mismatched Gaussians (MG) priors, and derives their properties. Section 3 describes how to induce dependence in inclusion probabilities using a Gaussian copula and MG marginal distributions. Section 5 proposes MCMC algorithms for computation in a variety of local GLM settings. Section 6 provides an empirical evaluation of the proposed method via several simulation examples. Section 7

illustrates the use of the model in applications the number of victims of human rights abuses via multiple systems estimation with regional models. Section 8 concludes and offers suggestions for future directions.

2. Mismatched Gaussians Distributions

2·1. *Description*

Here we introduce the Mismatched Gaussians (MG) prior on inclusion probabilities and derive some of its properties. We focus initially on univariate properties; the dependent case will be described later. Consider the standard Bayesian variable selection prior on coefficients ($\boldsymbol{\beta}$) in a regression model:

$$p(\beta_j) = \pi\delta_0 + (1-\pi)N(0,\tau^2) \tag{3}$$

where β_j is the j^{th} component of the p-vector $\boldsymbol{\beta}$. In a fully Bayesian treatment, we also specify a prior on the model inclusion probability π. The default choice is $\text{Beta}(1,1) = U(0,1)$. An extensive treatment of the properties of Beta priors on π is given in Scott & Berger (2010). However, we need not restrict ourselves to Beta priors, and in fact any prior with support on the unit interval will provide automatic correction for multiplicity. We propose the following alternative prior on π:

$$\pi = \Phi(\mu) \qquad\qquad \mu \sim N(\mu_0, \sigma_0^2). \tag{4}$$

We refer to this as the mismatched Gaussian (MG) prior, owing to the fact that π is a standard Gaussian CDF transform of a generally non-standard Gaussian random variable, and write $\pi \sim MG(\mu_0, \sigma_0^2)$ to indicate that π has an MG distribution. Note that if $\mu_0 = 0$ and $\sigma_0^2 = 1$ then $\pi \sim U(0,1)$. While this case is uninteresting at this point, it becomes relevant later when we introduce dependent priors on π, since it allows us to create a multivariate distribution with $U(0,1)$ marginals (e.g. a Gaussian copula).

The density of π is given by

$$f(\pi) = \frac{\phi_0(\Phi^{-1}(\pi))}{\phi(\Phi^{-1}(\pi))}, \tag{5}$$

where ϕ_0 is the normal density function with mean μ_0 and variance σ_0^2. This form is easily derived by a standard transformation of random variables and noting that $\frac{d}{dx}f^{-1}(x) = \frac{1}{f'(f^{-1}(x))}$.

2·2. *Correspondence to Beta distribution*

The obvious question at this point is how does the MG family compare to the family of Beta distributions? We would like a similarly flexible class of priors that would allow us to reflect different prior beliefs or prior information about model size. Fortunately, the answer is that MG distributions form a very similar class to the Beta family. The $\text{Beta}(a,b)$ distribution can have five shapes:

1. If $a < 1$ and $b < 1$, the density has poles at 0 and 1
2. If $a > b = 1$, the density has a pole at 0
3. If $b > a = 1$, the density has a pole at 1
4. If $a, b > 1$, the density has a mode between zero and 1 and the density is bounded away from infinity

5. Finally, there is the special case of $a = b = 1$, which gives the $U(0, 1)$.

Reformulating equation 5 as

$$f(\pi) = \frac{1}{\sigma_0} \exp\{-\frac{1}{2\sigma_0^2}\left((\Phi^{-1}(\pi))^2(1-\sigma_0^2) - 2\mu_0\Phi^{-1}(\pi) + \mu_0^2\right)\} \tag{6}$$

sheds some light on the range of possible shapes attainable by the MG distribution. Note that unless $\sigma_0^2 = 1$, the term $-(\Phi^{-1}(\pi))^2(1-\sigma_0^2)/(2\sigma_0^2)$ dominates both limits of the density, as the log-likelihood is quadratic in $\Phi^{-1}(\pi)$. If $\sigma_0^2 = 1$, the quadratic term disappears, and the kernel is linear in $\Phi^{-1}(\pi)$. The shape of the MG distribution corresponds to that of the Beta distribution as follows:

1. *Poles at zero and one.* If $\sigma_0^2 > 1$, then $1 - \sigma_0^2 < 0$, so as $\pi \to 0$ or $\pi \to 1$, the term $(\Phi^{-1}(\pi))^2(1-\sigma_0^2) \to -\infty$, and $f(\pi) \to \infty$, giving us poles at 0 and 1.
2. *Pole at zero.* If $\sigma_0^2 = 1$ then the term $\mu_0\Phi^{-1}(\pi)/2$ controls the limiting behavior. If $\mu_0 < 0$, then $f(\pi) \to \infty$ as $\pi \to 0$, and $f(\pi) \to 0$ as $\pi \to 1$, giving the pole at zero type.
3. *Pole at one.* Clearly, $\mu_0 > 0$ will give us the opposite limiting behavior as for the previous case, leading to a pole at one.
4. *Unique mode, bounded density.* If $\sigma_0^2 < 1$, $f(\pi) \to 0$ as $\pi \to 0, 1$ and is nonzero elsewhere. In terms of the beta distribution, this is the unimodal type with density bounded away from infinity.
5. *Uniform Distribution.* If $\sigma_0^2 = 1$ and $\mu_0 = 0$, $f(\pi)$ reduces to the constant function 1, implying a uniform distribution on π.

Beyond its role in achieving the basic shapes, μ_0 also controls the symmetry of the distribution. This is already clear for the case of $\sigma_0^2 = 1$ by inspection of equation 6. To understand the more general case, consider the ratio of $f(\pi)$ to $f(1-\pi)$:

$$\frac{f(\pi)}{f(1-\pi)} = \exp\left\{\frac{2\Phi^{-1}(\pi)\mu_0}{\sigma_0^2}\right\},$$

a fact easily derived from the equality $\Phi^{-1}(\pi) = -\Phi^{-1}(1-\pi)$. If $\pi > 1/2$ and $\mu_0 > 0$, then $\Phi^{-1}(\pi) > 0$, and the term inside the exponential is positive, implying $f(\pi) > f(1-\pi)$ for all $\pi > \frac{1}{2}$. Therefore, $\int_0^{1/2} f(\pi)d\pi > \int_{1/2}^1 f(\pi)d\pi$, and the distribution is asymmetric and skewed to the left. For $\mu_0 < 0$, we have $f(\pi) < f(1-\pi)$ for $\pi > 1/2$, and the distribution is skewed to the right. The distribution is symmetric if $\mu_0 = 0$. For $\sigma_0 > 0$ ($\sigma_0 < 0$), the local minimum (global maximum) is given by $\Phi\left(\frac{\mu_0}{1-\sigma_0^2}\right)$.

One might now ask whether we can define an isomorphism between the MG family and the Beta family. This is clearly not the case, since the MG density goes to infinity at its singularities at an exponential rate in π, while the Beta grows as a power of π. In other words, the MG goes to infinity at its singularities at a faster rate than any Beta distribution, and thus there can be no Beta distribution that is identical to the members of the MG family that have singularities. However, from the point of view of reflecting a range of prior beliefs this is irrelevant, as it is rather impossible to imagine strong prior beliefs on whether the density should grow polynomially or exponentially near its boundary.

While we cannot define an exact correspondence between the families, we can provide examples of similar Betas for a given MG via moment matching. Examples are shown in Figure 1. The black line shows the density of a Beta distribution with the same mean and variance as the sample from $f(\pi)$, shown in gray. Here, we calculate the moments of the MG by simulation. The differing rate at which the distributions approach the poles is most evident in the bottom row. Otherwise, these distributions are remarkably similar.

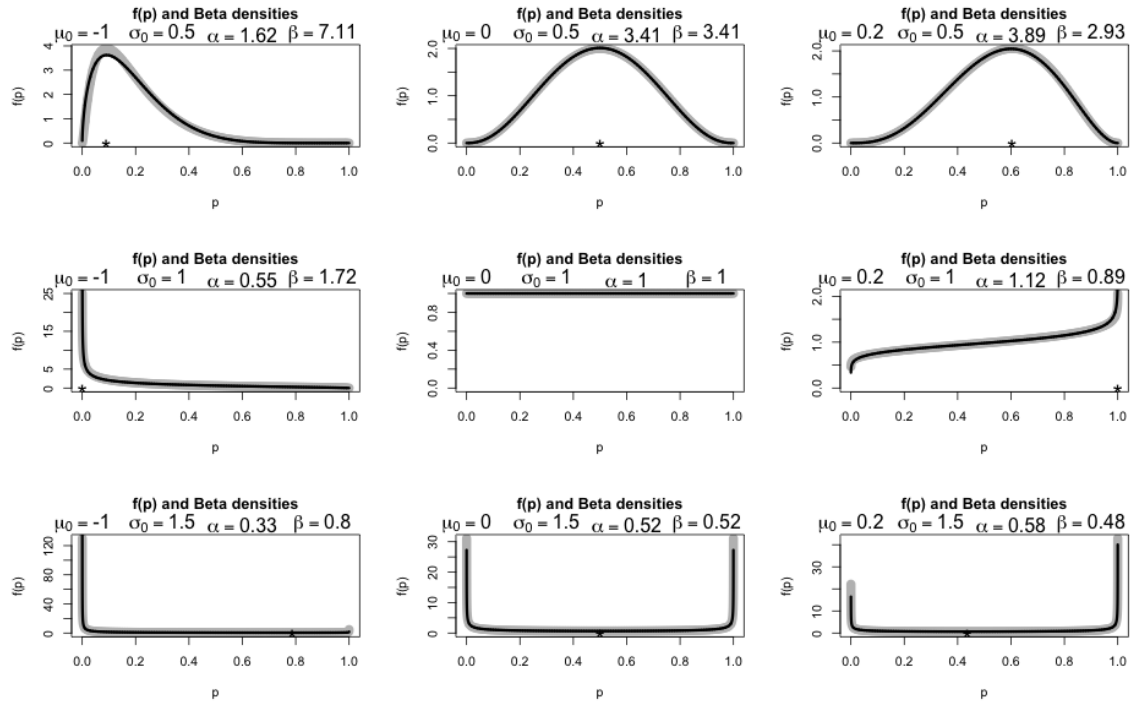


Fig. 1: Columns (left to right) show left, symmetric, and right skewed MG densities (gray) with the corresponding moment-matched Beta density overlaid (black). Rows (top to bottom) similarly show $\sigma_0 < 0$, $\sigma_0 = 0$ and $\sigma_0 > 0$ densities.

2·3. *Moments*

The raw moments of the MG distribution are defined by the integrals

$$E(\pi^k) = \int_{\mathbb{R}} \Phi(\mu)^k \phi_0(\mu) d\mu,$$

which are not available in closed form. However, the first moment can be approximated using a Taylor series expansion. Since $\Phi(\mu) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x}{\sqrt{2}}\right)\right]$, where erf is the error

function, we can re-write this as

$$\begin{aligned}\mathbb{E}(\pi) &= \int \left[\frac{1}{2} + \frac{1}{\sqrt{\pi}} \sum_{i=0}^{\infty} \frac{(-1)^i \mu^{2i+1}}{i! \sqrt{2}^{(2i+1)} (2i+1)}\right] \phi_0(\mu) d\mu \\ &= \frac{1}{2} + \frac{1}{\sqrt{\pi}} \sum_{i=0}^{\infty} \frac{(-1)^i}{i!(2i+1) 2^{(2i+1)/2}} \mathbb{E}_0(\mu^{2i+1}).\end{aligned}$$

where $\mathbb{E}_0$ is the expectation of μ with respect to the measure ϕ_0, and the exchange of summation and integration follows from Fubini and the fact that erf is an entire function. The quality of the approximation and the number of terms required to achieve negligible error depends on the hyper parameters μ_0 and σ_0 through the higher-order raw moments of the normal distribution $\mathbb{E}_0(\mu^{2i+1})$. In general, we need $\mathbb{E}_0(\mu^{2i+1})$ to be small relative to $i!$, which dominates the denominator of the summands. The critical terms in $\mathbb{E}_0(\mu^{2i+1})$ are $\sigma^{2i}\mu$ and μ^{2i+1}. When $\sigma < 1$ and $\mid \mu \mid < 1$, these terms are decreasing in i, and thus the remainder in the Taylor series is going to zero quickly. However, when $\sigma > 1$ or $\mid \mu \mid > 1$, these terms are increasing in i, and thus we need $i! >> \sigma^{2i}$ and $i! >> \mu^{2i+1}$ for the remainder to be small. For moderate values of μ or σ (e.g. 2), this may require 15-20 terms for the error to be negligible. This approach could of course be expanded to higher moments and correlations with significantly more mathematical effort. While the existence of an analytic functional representation of the moments is a noteworthy feature of MG distributions, we generally use simulation to approximate moments of the MG distribution rather than the Taylor series representation.

2·4. *Latent Gaussian hierarchical representation*

Lastly, we note that the MG prior can be expressed hierarchically using latent Gaussians in the usual way:

$$\begin{aligned}\gamma &= \mathbb{1}[Z > 0] \\ Z &\sim N(\mu, 1) \\ \mu &\sim N(\mu_0, \sigma_0^2).\end{aligned}$$

Under this specification, $\pi = \Pr(\gamma = 1) \sim MG(\mu_0, \sigma_0^2)$. We will make extensive use of the latent variable representation of the MG prior in constructing computational algorithms in section 5.

3. Dependent Inclusion Probabilities using Copulas

While MG forms a class of distributions on [0,1] that is as flexible as the Beta family, there is no compelling reason to use MG priors in place of Beta priors in a simple variable selection case. In fact, MG is somewhat less convenient, since unlike the Beta, MG priors do not lead to analytically tractable prior model probabilities. Rather, the strength of the MG family lies in the ease with which it can accommodate dependence in inclusion probabilities, which we now demonstrate.

Clearly, inducing dependent inclusion probabilities requires a multivariate distribution with margins supported on the unit interval. A p-variate generalization of the MG

distribution can be specified as:

$$\mu_1, \ldots, \mu_p \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \tag{7}$$
$$\pi_j = \Phi(\mu_j) \quad j = 1, \ldots, p. \tag{8}$$

Note that we can also write this as:

$$\begin{aligned} \xi_1, \ldots, \xi_p &\sim N(\mathbf{0}, \boldsymbol{\Psi}_0) \\ \mu_j &= \mu_{0j} + \sigma_{0j}\xi_j \\ \pi_j &= \Phi(\mu_j) \end{aligned}$$

where $\boldsymbol{\Psi}_0$ is the correlation matrix corresponding to $\boldsymbol{\Sigma}_0$, μ_{0j} is the j^{th} entry of $\boldsymbol{\mu}$, and σ_{0j}^2 is the j^{th} diagonal entry of $\boldsymbol{\Sigma}_0$. We can further re-express this as:

$$\begin{aligned} F(v_1, \ldots, v_p) &= \boldsymbol{\Phi}_{\boldsymbol{\Psi}_0}(\Phi^{-1}(v_1), \ldots, \Phi^{-1}(v_p)) \\ \mu_j &= \mu_{0j} + \sigma_{0j}\xi_j \qquad\qquad \xi_j = \Phi^{-1}(v_j) \\ \pi_j &= \Phi(\mu_j) \end{aligned}$$

where $v_1, \ldots, v_p \in [0,1]^p$, $\Phi_{\boldsymbol{\Psi}_0}$ is a p-variate normal CDF with mean vector $\mathbf{0}$ and correlation matrix $\boldsymbol{\Psi}_0$, and Φ^{-1} is the univariate probit function - i.e. $F(v_1, \ldots, v_p)$ is a Gaussian copula. A copula is a multivariate distribution on the unit hypercube with uniform marginals. The fundamental result for copulas is Sklar's theorem (1959), which states that any p-variate multivariate distribution can be represented by a copula and a collection of p marginal distributions. In the case where the marginals are MG and the copula is Gaussian, nothing is gained by representing the multivariate MG distribution in this way rather than using the more intuitive representation in equations 7 and 8. However, in much the same way that the MG distribution is a special case of a large family of distributions on $[0,1]$ that can be obtained via CDF transforms of location-scale random variables with support on the real line, our approach to generating multivariate MG distributions can be expanded beyond the Gaussian dependence structure. This is much clearer when the marginals are considered separately from the dependence, and therefore it is useful to recast the multivariate MG distribution as a copula and a collection of MG marginals. For instance, we could introduce a different dependence structure but maintain the MG marginals by writing

$$\begin{aligned} \pi_j &= \Phi(\mu_{0j} + \sigma_{0j}\Phi^{-1}(v_j)) \\ F(v_1, \ldots, v_p) &= G_{\nu, \boldsymbol{\Psi}_0}(\Phi^{-1}(v_1), \ldots, \Phi^{-1}(v_p)), \end{aligned}$$

where $G_{\nu,\boldsymbol{\Psi}_0}$ is a multivariate t CDF with degrees of freedom ν and scale matrix $\boldsymbol{\Psi}_0$. This would give a multivariate MG distribution with tail dependence. Given the interest in non-Gaussian tails and tail dependence in areas such as finance, this construction may be of practical use. For example, it could be applied to model default rates in asset pricing. Moreover, because the multivariate t can be represented as a scale mixture of normals, this particular multivariate MG distribution is computationally tractable. While we do not consider non-Gaussian dependence structures further in this paper, and we are primarily interested in dependent inclusion probabilities, it is important to keep in mind that our method could be applied anytime a multivariate distribution on the unit hypercube is desired.

4. Structured Model Dependence using the MG prior

In the following section we construct a series of multivariate MG priors designed for several broad classes of applications. To begin, we consider the data-augmented form of the Bayesian variable selection prior. Let there be r regions/groups/time points and p covariates. Let β_{ij} be the the coefficient of the jth variable in the ith local model, and introduce γ_{ij} such that:

$$p(\beta_{ij} \mid \gamma_{ij}) = \delta_0 \mathbb{1}(\gamma_{ij} = 0) + N(0, \tau^2)\mathbb{1}(\gamma_{ij} = 1)$$

Now write the MG distribution in its hierarchical form:

$$\begin{aligned} \gamma_{ij} &= \mathbb{1}[Z_{ij} > 0] \\ Z_{ij} &\sim N(\mu_j, \rho(\theta_j)) \\ \mu_j &\overset{iid}{\sim} N(\mu_0, \sigma_0^2), \end{aligned}$$

where $\rho(\theta)$ is a covariance function specifying the dependence structure of $\mathbf{Z}_j$. The hierarchical representation provides a simple framework for building Gaussian copula models for the inclusion probabilities through the Gaussian latent variables Z_{ij}. We focus on two settings: (1) spatial or time series applications with multiple observations per time point or spatial location; and (2) general hierarchical models as arise frequently in survey applications.

4·1. *Continuous Spatial and Temporal Data*

We first consider a point-referenced spatial or continuous-time context. We now have a regression model for each location. The general approach to inducing dependence is:

$$\begin{aligned} \gamma_{ij} &= \mathbb{1}[Z_{ij} > 0] \\ Z_{ij} &= \mu + \eta_{ij} + \epsilon_{ij} \\ \mu &\sim N(\mu_0, \sigma_0^2) \end{aligned}$$

where i indexes region, j covariate, and η is a spatial effect. If this were a model for point-referenced binary data, we could take $\epsilon_{ij} \overset{iid}{\sim} N(0, 1)$ and specify independent Gaussian Process (GP) priors on η_j for each $j \in \{1, \ldots, p\}$. While the variance of the Z's in a non-spatial binary model is not identified and is usually fixed at 1, the variance of the GP is identified in this model as long as the variance of the ϵ's is fixed, since it determines the spatial signal:noise ratio. However, in our case, we are using this as a prior on covariate inclusion probabilities, so somewhat more care is needed to ensure that the induced prior on π is still meaningful for a given choice of μ_0, σ_0^2.

The previous section outlined the range of prior beliefs about π that can be reflected by MG priors and how to choose μ_0 and σ_0^2 to achieve them. The default choice in Bayesian variable selection is to choose a $U(0, 1)$ prior on π, which in the MG case means choosing $\mu_0 = 0$ and $\sigma_0^2 = 1$, under the assumption that the marginal variance of the Z's is 1. However, in this case, the variance of the GP is not known a priori. An alternative specification that retains uniform marginals could be constructed as follows:

$$\begin{aligned} Z_{ij} &= \mu + \eta_{ij} + \epsilon_{ij} \\ \mu \mid \sigma^2 &\sim N(0, \sigma^2 + 1) \qquad\qquad \epsilon_{ij} \overset{iid}{\sim} N(0, 1) \\ \boldsymbol{\eta}_j \mid \sigma^2 &\overset{ind}{\sim} GP(\sigma^2, \phi_j, \nu_j). \end{aligned}$$

We refer to the induced prior on π as the continuous MG process. The time-series case is identical but with a one-dimensional GP.

Conditional on μ, the distribution of the vector $\boldsymbol{Z}_j$, which has one entry for each location, is:

$$\boldsymbol{Z}_j \sim N(\mu\mathbf{1}, \Sigma + I)$$

where Σ is the GP covariance matrix with variance σ^2. Let $\pi_{ij} = Pr(\gamma_{ij} = 1)$. Marginal of the process at all other locations, $\pi_{ij} = 1 - \Phi(\mu/(\sigma^2 + 1))$, which is analogous to the independent case but rescaled based on the marginal variance of the Z's. However, the γ_{ij}'s are dependent, with polychoric correlation determined by $\Sigma + I$. Thus, the continuous MG process can be viewed as a Gaussian copula model on any finite collection of point-referenced inclusion indicators $\boldsymbol{\gamma}_j$ with marginals $\mathbb{1}(\mu + \eta_{ij}/(1 + \sigma^2) > 0)$ and correlation matrix $\frac{1}{1+\sigma^2}(\Sigma + I)$. Alternatively, using the compact notation as before, we can define the prior on any finite collection of inclusion probabilities as a Gaussian copula on $\boldsymbol{\pi}_j$ with marginals $\Phi(\mu + \eta_{ij}/(1 + \sigma^2))$ and correlation matrix $\frac{1}{1+\sigma^2}(\Sigma + I)$.

Note that a single intercept term is shared across all locations and covariates, i.e. $\mu_j = \mu$. The model we propose here reflects the belief that regressions across regions should have about the same number of covariates, and that each covariate should have the same prior inclusion probability, but that the identity of important covariates should differ across space in a dependent way. One could consider alternative hierarchical specifications on μ, where μ varies across covariate or region, with dependence induced through a hierarchical prior. This would be appropriate in the case where one believes that more covariates could be important at some locations than at others.

4·2. *Discrete Spatial or Temporal Data*

The general approach we propose for discrete spatial and time series is to specify a Markov Random Field (MRF) structure for the covariance across time and location in the Z_j's. We illustrate this in the case of region-specific regressions in an areal model. The classical MRF structure for areal data is the conditionally autoregressive (CAR) model. Suppose we have m spatial regions and let W be a matrix such that $w_{ii'} = 1$ if i and i' share a border and $w_{ii'} = 0$ otherwise. Let D_W be the diagonal matrix whose nonzero entries are equal to $W\mathbf{1}$, the row-wise sums of W. If we specify $\boldsymbol{\eta}_j \sim CAR(\tau_j)$, then $\boldsymbol{\eta}_j \propto \exp\{-\frac{1}{2}\boldsymbol{\eta}_j^T(D_W - W)\boldsymbol{\eta}_j\}$. Unfortunately, the matrix $D_W - W$ is not invertible, and this distribution is improper. Because we must again pay attention to the marginal variance of the Z's in our model, we instead use the ρ-CAR model, which does have an invertible precision matrix and thus, an identified marginal variance.

The ρ-CAR model has precision matrix $D_W - \rho W$, which is invertible for any $\rho \in [1/\lambda_1, 1/\lambda_m]$, where $\lambda_1, \ldots, \lambda_m$ are the ordered eigenvalues of $D_W - W$. In practice, $\rho < 1$ will generally be sufficient to guarantee invertibility. Usually, a prior is placed on ρ in these models. However, in our case we will again embed this spatial process into a joint MG prior on areal inclusion probabilities for each predictor, and thus the ρ-CAR is just a stand-in for the CAR-model. As such, we fix $\rho = 0.99$, which approximates a CAR model while maintaining reasonable condition numbers for $D_W - \rho W$. A ρ-CAR model is then defined conditionally by $\eta_{ij} \mid \eta_{-ij} \sim N(\rho\frac{1}{n_i}\sum_{k\sim i}\eta_{kj}, \frac{\tau_j}{n_i})$, where n_i is the number of regions that are neighbors of region i. Alternatively, the joint distribution can be written as $\pi(\boldsymbol{\eta}_j) \propto \exp\{\boldsymbol{\eta}_j^T(D_W - \rho W)\boldsymbol{\eta}_j)$, so long as $\rho < 1$. [ADD SOME CITATIONS]

Ideally, we would specify a prior on μ to make the induced prior on π again $U(0, 1)$. However, in the ρ-CAR model, the variance of the spatial effects varies across regions in a

manner we cannot control– as a function of the number of neighbors. The problem is that because the marginal variance of the spatial effects in a ρ-CAR model will not in general be equal, any prior on μ will necessarily induce different priors on π for each region. There will undoubtedly be cases in which one wishes to continue with the conditional model despite these shortcomings, not least of which is the lack of any well-established alternative for discrete time/space data. In that case, we suggest introducing covariate-specific means μ_j with hierarchical priors:

$$\begin{aligned} \boldsymbol{Z}_j \mid \boldsymbol{\eta}_j &\sim N(\mu_j \mathbf{1} + \boldsymbol{\eta}_j, I) \\ \boldsymbol{\eta}_j &\sim \rho - CAR(\tau_j) \\ \mu_j \mid \tau^2 &\sim N(\mu_0, \tau_j^2) \\ \mu_0 &\sim N(0, \sigma_0^2) \end{aligned}$$

where μ_j are covariate-specific means. Note that we can write this marginally of $\boldsymbol{\eta}_j$ as $\boldsymbol{Z}_j \sim N(\mu_j \mathbf{1}, \tau_j C + I)$ where $C = (D_W - \rho W)^{-1}$. We select strong prior on τ_j^2 and σ_0^2 such that the majority of the mass in these priors favors values such that $\sigma_0^2 > \tau_j^2 \max_i c_{ii} + 1$. In this case, while the induced priors on π_{ij} will differ across regions and covariates, the covariates in all regions will all have the double pole shaped prior. This is somewhat reminiscent of the horseshoe shrinkage prior and will tend to favor relatively sparse models. This is usually desirable in high-dimensional settings.

4·3. *Hierarchical Models*

Consider modeling an outcome of interest given covariates where the data are collected using a stratified survey design. Often, two or more levels of hierarchy exist in surveys, and it is natural in these cases to borrow information across strata (higher-level clusters) and groups within strata (lower-level clusters). The natural approach in the setting without variable selection is to specify a hierarchical model on the regression coefficients:

$$\begin{aligned} y_{ij} &= x_{ij}\beta_j + \epsilon_{ij} \qquad\qquad \epsilon_{ij} \sim N(0, \sigma^2) \\ \beta_j &\overset{iid}{\sim} N(\beta, \tau^2 I_p) \end{aligned}$$

where j indexes strata and i group within stratum. In our case we are also interested in inferences on important covariates and how they vary across groups and strata. We would expect that different covariates may be important in different strata or groups, and that the subset of important covariates is more likely to be similar for groups within the same stratum than for groups in different strata. For dependent variable selection, we employ an MG prior on school-specific inclusion probabilities with a hierarchical prior on μ to borrow information across groups and strata:

$$\begin{aligned} Z_{ij} &\sim N(\mu_j, 1) \\ \mu_j &\overset{iid}{\sim} N(\mu_0, \zeta) \qquad\qquad \mu \sim N(0, 1-\zeta) \\ \zeta &\sim \text{Beta}(a_0, b_0) \end{aligned}$$

Note that $\text{Var}(\mu_j \mid \zeta)$ is always 1 in this model, and $E(\mu_j \mid \zeta) = 0$, so the induced prior on π_j, the group-level inclusion probability, is $U(0,1)$. This relatively simple hierarchical setup allows for borrowing of information across groups for Bayesian variable selection in group-specific linear models.

5. Computation

The MG prior is an alternative prior on inclusion probabilities in the well-studied Bayesian variable selection problem for linear models. The main computational challenge in this setting is efficiently exploring the model space in higher-dimensions.

5·1. *Exploring the Model Space*

For the special case of Gaussian linear models with a Beta prior on inclusion probabilities, marginal likelihoods are analytically tractable, so for small p, the problem reduces to enumerating the model space and calculating marginal likelihoods. In our case, the posterior is not analytically tractable, and therefore variable selection must always be performed within an MCMC algorithm for model computation. There is a significant literature on stochastic-search variable selection within MCMC for larger p linear models and GLMs (see e.g. Scott & Carvalho (2008)), with focus on reversible-jump algorithms that propose local moves in the neighborhood of the current model as follows:

1. *Birth step*: Add a predictor to the model by randomly selecting an index $k \in \{j : \gamma_j = 0\}$ and propose to set $\gamma_k = 1$
2. *Death step*: Remove a predictor from the model by randomly selecting $k \in \{j : \gamma_j = 1\}$ and propose to set $\gamma_k = 0$.
3. *Exchange step*: Select an index $k \in \{j : \gamma_j = 0\}$ and an index $k' \in \{j : \gamma_j = 1\}$ and propose to set $\gamma_k = 1$ and $\gamma_{k'} = 0$.

These methods have a tendency to become trapped in local modes, leading to more recent work that also adds the possibility of a global move which proposes an entirely new model. To improve the acceptance rates for these global moves, predictors for the new model are usually selected based on the current MCMC estimate of their marginal inclusion probabilities.

Efficiently exploring the model space is the most challenging aspect of computation in models with MG priors, and is complicated by the fact that we cannot obtain closed-form marginal likelihoods even in the Gaussian case. However in the current work, we have in mind applications with small to moderate p (at most 30), and thus we use a simpler approach to propose new models. At every MCMC iteration, with probability q, we propose to change the value of each γ_j sequentially conditional on γ_{-j}. The value of q is tuned to give metropolis acceptance rates of approximately 0.3. While this approach would not be practical with large p, in the small to moderate p applications in sections 6 and 7, this approach is computationally tractable. For large p applications, we suggest a more sophisticated approach, such as that in Hans et al. (2007).

5·2. *Computation for MG Latent Variables*

Computation for models with MG priors is quite straightforward, requiring only some additional Gibbs sampling steps for the MG latent variables. Since the rest of the computation will differ depending on the sampling model, and since in general Bayesian computation for regression models is at this point very standard, we describe only the steps related to the MG inclusion probabilities. We outline a general algorithm. Suppose $\mathbf{Y}_i \mid \boldsymbol{\beta}_i, \boldsymbol{\gamma}_i, \sim f(\mathbf{Y}_i \mid \boldsymbol{\beta}_i, \boldsymbol{\gamma}_i)$ and $\boldsymbol{\beta}_i \sim g(\boldsymbol{\beta}_i \mid \boldsymbol{\theta})$. Let $P(\mathbf{Y}_i \mid \boldsymbol{\gamma}_i) = \int f(\mathbf{Y}_i \mid \boldsymbol{\beta}_i, \boldsymbol{\gamma}_i) g(\boldsymbol{\beta}_i \mid \boldsymbol{\theta}) d\boldsymbol{\beta}_i$ be the integrated likelihood of the model that includes variables indexed by $\boldsymbol{\gamma}_i$. In practice, f may include more random variables than just $\boldsymbol{\beta}_i$– if possible, integrate them out; if not, condition on them. Let the conditional normal distribution of $Z_{ij} \mid Z_{-ij}$ be denoted

by $N(Z_{ij} \mid Z_{-ij}, \boldsymbol{\theta}_Z)$ where $\boldsymbol{\theta}_Z$ are the additional parameters necessary to specify the distribution. As usual, i references group and j references covariates. A general algorithm for computation proceeds as follows. For $i = 1...r$ and $j = 1...p$

1. Propose γ^*_{ij} from $\pi(\gamma^*_{ij})$ as described above and let $\boldsymbol{\gamma}^*_i$ be the proposal for $\boldsymbol{\gamma}_i$ such that the jth component reflects the proposed change, γ^*_{ij}. Calculate $p_{ij} = Pr(Z_{ij} > 0 \mid Z_{-ij})$ and $r = \frac{P(\mathbf{Y}_i|\boldsymbol{\gamma}^*_i)p_{ij}^{\gamma^*_{ij}}(1-p_{ij})^{\gamma^*_{ij}}\pi(\gamma_{ij})}{P(\mathbf{Y}_i|\boldsymbol{\gamma}_i)p_{ij}^{\gamma_{ij}}(1-p_{ij})^{\gamma_{ij}}\pi(\gamma^*_{ij})}$. Accept $\boldsymbol{\gamma}^*_i$ with probability $\min(1, r)$.
2. Sample $Z_{ij} \mid \boldsymbol{\gamma}_i, \sim N_{\gamma_{ij}}(Z_{ij} \mid Z_{-ij}, \boldsymbol{\theta}_Z)$, where N_0 is the conditional distribution of Z_{ij} truncated to be less than 0, and N_1 is truncated to be positive.

After $\boldsymbol{\gamma}$ and $\mathbf{Z}$ have been updated for all i and j, update $\boldsymbol{\theta}_Z$.

As a more concrete example, we outline the computation for discrete spatial models, as described in section 4·2. For $i = 1...r$ and $j = 1...p$

1. Propose γ^*_{ij} from $\pi(\gamma^*_{ij})$, typically $\pi(\gamma^*_{ij}) = Bern(p_{ij})$, where $p_{ij} = \Phi(\mu_j + \eta_{ij})$. Calculate $r = \frac{P(\mathbf{Y}_i|\boldsymbol{\gamma}^*_i)p_{ij}^{\gamma^*_{ij}}(1-p_{ij})^{\gamma^*_{ij}}\pi(\gamma_{ij})}{P(\mathbf{Y}_i|\boldsymbol{\gamma}_i)p_{ij}^{\gamma_{ij}}(1-p_{ij})^{\gamma_{ij}}\pi(\gamma^*_{ij})}$. Accept γ^*_{ij} with probability $\min(r, 1)$.
2. Sample $Z_{ij} \mid \gamma_{ij}, \eta_{ij}, \mu_j \sim N_{\gamma_{ij}}(\mu_j + \eta_{ij}, 1)$, where N_0 is the normal distribution truncated to be less than 0, and N_1 is the normal distribution truncated to be positive.
3. Sample $\eta_{ij} \mid Z_{ij}, \boldsymbol{\eta}_{-ij} \sim N(\rho\frac{1}{n_i}\sum_{k\sim i}\eta_{kj}, \frac{\tau_j}{n_i})$.

After $\boldsymbol{\gamma}$, $\mathbf{Z}$, and $\boldsymbol{\eta}$ are updated for all i and j, update $\boldsymbol{\mu}$ and $\boldsymbol{\tau}^2$ from Normal and Inverse-Gamma distributions respectively. Update μ_0 from a normal distribution.

We also note that efficient MCMC algorithms for dependent Bayes variable selection in regression with non-Gaussian likelihoods – including logistic regression, multinomial logit, and negative binomial regressions for count data – can be developed easily by combining our approach with the Polya-Gamma data augmentation scheme developed in Polson et al. (2012). We utilize this method in several of the data examples presented below.

6. Simulation Studies

6·1. *Hierarchical Model Simulation*

We conducted a simulation study to assess the performance of the model in recovering the true model and in out-of-sample prediction. Data were simulated in a manner consistent with the hierarchical structure in many surveys: we include ten strata and twenty groups within each stratum with 25 observations per group. The model for each group is determined as follows:

1. Sample $\mu_h \overset{iid}{\sim} N(0, 0.25)$
2. Sample $\mu_{hi} \mid \mu_h \overset{iid}{\sim} N(\mu_h, 0.75)$
3. Sample $Z_{hij} \mid \mu_{hi} \overset{iid}{\sim} N(\mu_{hi}, 1)$.
4. Set $\gamma_{hij} = \mathbb{1}(Z_{hij} > 0)$.

We also sample β_{hij} hierarchically as:

1. Sample $\beta_h \overset{iid}{\sim} N(0, 1)$

2. Sample $\beta_{hi} \overset{iid}{\sim} N(\beta_h, 1)$
3. Sample $\beta_{hij} \overset{iid}{\sim} N(\beta_{hi}, 1)$,

where h indexes strata, i indexes groups within strata, and j indexes covariates. We then sample $y_{hi} \sim N(X_{hi}\beta_{hi}, I_{n_{hi}})$ with n_{hi} the number of observations in each group. In this simulation study there are five covariates. We also sample 5 held out observations per group from the same model. We then fit either the hierarchical MG model with a hierarchical model on the β's, a hierarchical regression model without BVS, Lasso, or regression using stepwise selection by AIC. The Lasso and stepwise models were fit either to each group separately, to the data within each stratum pooled, or to all of the data pooled. We then assessed out-of-sample prediction on the held-out data by predictive mean square error (PMSE). These results are shown in table 1. Note that the model with BVS and MG priors outperforms competitors both in terms of PMSE and model selection accuracy.

	One Model	Stratum-level Models	Group-level Models
MG	NA	NA	1.17
Hierarchical β's	NA	NA	1.25
AIC	20.22	16.01	30.58
Lasso	20.67	16.74	36.86

Table 1: Predictive mean square error for MG priors compared to alternative approaches with data simulated from a model with structured dependence in inclusion probabilities

6·2. *Capture-recapture Simulation*

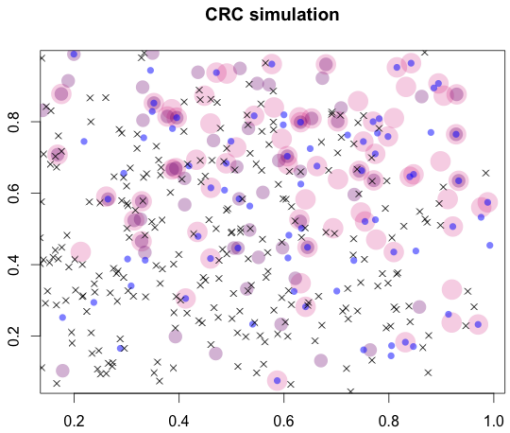
Capture-Recapture (CRC) seeks to estimate the total size of a population based upon multiple sub-samples of the population and the overlaps among them. Let there be $k = \{1, ..., K\}$ sub-samples (or lists) collected. Then, the data consists of counts of the number of individuals that appear on each list intersection. We define a list intersection as a binary string, $d_1...d_K$, where $d_k = 1$ for an intersection including list k and 0 otherwise. $Y_{d_1...d_K}$ is the number of individuals who appear in intersection $d_1...d_K$. The traditional method (see Fienberg (1972)) then fits a Poisson regression to the above counts, $Y_{d_1...d_K} \sim \text{Pois}(\mu_{d_1...d_K})$, where $\log(\mu_{d_1...d_K}) = \alpha + \beta_1 d_1 + ...\beta_K d_K$. The number of elements seen on no list ($Y_{0...0}$) is then estimated as e^{α}. Incorporating list dependence is as simple as including interaction terms to the log-linear regression. For example, to account for dependence between list one and two, one would also include $\beta_{12} d_1 d_2$ in the equation for $\log(\mu_{\mathbf{d}})$. CRC generally relies upon the assumption that all members of the population have equal probability of being sampled by each list. When this assumption breaks down, stratification is often done to subdivide the samples into sub-populations within which it is plausible that the probability of capture for each member of the sub-population is approximately equal by list (Bruno et al. (1994)). However, the degree to which stratification is possible is dependent upon the amount of data available. If one stratifies too much, each stratum is left with so little data pertaining to it that independently obtaining reasonable estimates for individual strata is impossible.

Estimation of these models is further complicated by the fact the there is often little information *a priori* about which model ought to be used, i.e. which interaction indicators

between lists should be included. However, it is logical that spatially proximate regions should have similar models, as collection agencies may operate similarly regionally. This makes capture-recapture a perfect candidate for using an MG prior for structured model dependence, which will allow for finer spatial stratification through borrowing of information across space, both in terms of the mean estimates and informing about variable inclusion probabilities.

We consider a case in which there are four lists (K=4) which sample individuals at spatial locations contained in $\mathcal{S} = [0,1]^2$. The intensity surface for the point distribution, $\lambda(\mathbf{s})$, be given by $\lambda(\mathbf{s}) = c\left(\frac{\sqrt{2}}{2} - ||\mathbf{s} - .5||\right)$. Our first simulated list can detect an event at location $\mathbf{s} = \{x, y\}$ with probability $Pr(L_1(x,y) = 1) = \frac{1}{8}(x+y)^3$. In order to induce list dependence in some spatially adjacent regions, list two samples an event with probability $Pr(L_2(x,y) = 1|L_1(x,y)) = \frac{3}{16}y^3 + \frac{1}{16} + \frac{1}{2}\mathbb{1}[L_1(x,y) = 1]\mathbb{1}[x < 0.4]$. We make a similar conditional statement for L_3; $Pr(L_3(x,y) = 1 \mid L_1(x,y)) = \frac{3}{16}x^2 + \frac{1}{16} + \frac{1}{2}\mathbb{1}[L_1(x,y) = 1]\mathbb{1}[x > 0.4]$. List four is generated independently of all other lists, $Pr(L_4(x,y) = 1) = \frac{3}{16}x^2 + \frac{1}{16}$. In summary, we sample locations according to a non-homogeneous Poisson process in continuous space. At each of these locations, each of the four lists detects individuals with spatially varying probabilities and spatially varying dependence. This structure does not precisely simulate from our model, though it does induce the properties that exist in real applications and which our model addresses. We discretize the space into a 5×5 grid of cells to create the areal units. An example simulation is shown in Figure 2.

Fig. 2: Black x's represent unsampled individuals. Those sampled by List 1 appear in a large, transparent pink circle. Those sampled by List 2 are indicated by a medium semi-transparent circle. List 3's observations appear as small, opaque circles. List 4 is omitted for clarity of presentation.



Because of its convenient latent variable representation using the Polya-Gamma latent variable scheme for fully conjugate computation derived in Polson et al. (2012), we fit a negative binomial regression model: $\mathbf{Y}_i \sim NegBin(\mathbf{p}_i)$ where $\mathbf{p}_i = \frac{\exp[\alpha_i + \mathbf{X}_i^T \boldsymbol{\beta}_i]}{1+\exp[\alpha_i + \mathbf{X}_i^T \boldsymbol{\beta}_i]}$, $\mathbf{X}_i$

are the intersection indicators as above and $\alpha_i \sim CAR(\tau_\alpha)$. Our prior for β is the ρ-CAR MG prior of section 4·2. Even with this added complexity, we retain fully conjugate updates for all parameters. We take a Bayesian model averaging approach to estimating the total number of un-sampled individuals. For region i, we use the "model average" estimate (Raftery et al. (1997)), $\hat{\alpha_i} = \frac{1}{M}\sum_{m=1}^{M} \alpha_i^{(m)}$, i.e. the posterior mean estimate over all iterations of α_i averaged over all of the sampled models.

To investigate what is gained by including both spatial smoothing of the response surface and spatial smoothing of the inclusion probabilities, we select models for comparison that include some of these features. We fit an equivalent spatial model as the above in the absence of model selection, i.e. $\gamma_{ij} = 1$, using R-INLA (Rue & Martino (2009)). Two other points of comparison are estimated on each region independently: Bayesian model averaging (BMA) and the AIC to determine a single model.

Table 2 shows the results of this experiment, where we allow the sampling scheme to be the same, but the total number of events, N, and the number of unsampled points (in parentheses) varies in the four simulations. We find that if we do not share information among all of the strata, there are many regions for which no reasonable estimate can be produced. This is demonstrated by the fact that in the methods that do not leverage the shared information, there are several estimates with infinite confidence intervals. In these intervals, some of estimates were on the order of 10^{13}. By comparing our method to the R-INLA implementation, we see the added value of allowing different models by region– our model has empirical coverage rate approximately equal to the nominal coverage rate of 0.95, it has the lowest root mean squared error as applied to the difference between the true number of unsampled individuals and the estimated number. Figure 3 highlights the ability of the MG prior to accurately recover the true model.

7. Estimating the number of killings in Casanare, Colombia

We analyze a data set of the number of individuals killed in Casanare, Colombia between 2001-2006[1], a time period during which the region was in the midst of a conflict involving paramilitaries, guerrillas, and the Colombian military (Guberek et al. (2010)). In prior analyses of this data, it was not possible to estimate the number of undocumented killings in each municipality and each year separately because of the sparsity of the data collected. In one analysis, Guzmán et al. (2007) aggregated all of the southern municipalities (Sabanalarga, Villanueva, Monterrey, Aguazul, Tauramena, Maní, Chameza, Recetor and Yopal) to get an estimate of the combined total number of killings in the south for two nested time periods, 2001-2004 and 1998-2005. The later analyses of Guberek et al. (2010) and Lum et al. (2010) used slightly less regional aggregation, grouping the area into four regions: Center (Yopal and Augazul); Piedemonte (Sacama, La Salina, Tamara, Recetor, Chameza, and Nunchia); South(Tauramena, Monterrey, Villanueva, Maní, and Sabanalarga); and the Plains (Hato Corozal, Paz de Ariporo, Poré, San Luis de Palenque, Trinidad, and Orocue). They were, however, able to temporally disaggregate the analysis to produce yearly estimates. The necessity of the spatial aggregation was due to the inability of the available models to reliably produce estimates in a sparse data setting.

[1] This data is made freely available by Benetech's Human Rights Data Analysis Group (HRDAG) at https://www.hrdag.org/about/CasanareSummaries.html

True Count	MG	INLA	BMA	AIC
N = 1,785 (1,046)				
Empirical Coverage	0.96	0.88	0.8	0.8
RMSE	18	21	7233	76230893497
RMSE - no outliers	18	21	31	31
Mean Int. Length	58	106	Inf	Inf
Mean Int. Length- no outliers	58	106	166	236
N = 2,312 (1,348)				
Empirical Coverage	0.92	0.96	0.84	0.84
RMSE	26	27	221069260534	221497269721
RMSE - no outliers	26	27	39	39
Mean Int. Length	66	71	Inf	Inf
Mean Int. Length- no outliers	66	71	155	200
N = 4,641 (2,714)				
Empirical Coverage	0.92	0.84	0.92	0.84
RMSE	31	47	59	75177676116
RMSE - no outliers	31	47	59	43
Mean Int. Length	134	104	Inf	Inf
Mean Int. Length- no outliers	134	104	323	174
N = 11,626 (6,805)				
Empirical Coverage	1	0.96	0.88	0.88
RMSE	53	67	146	225
RMSE - no outliers	53	67	146	168
Mean Int. Length	357	269	Inf	709
Mean Int. Length- no outliers	357	269	783	709

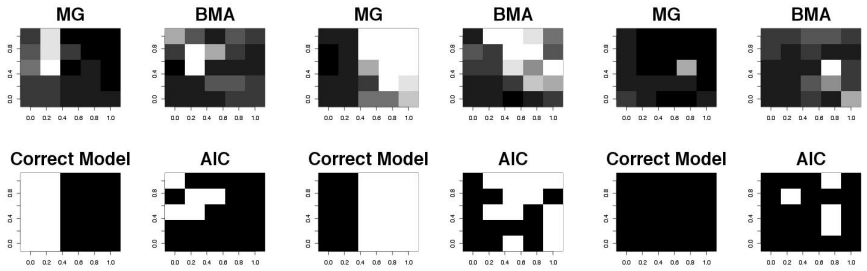
Table 2: Results of the CRC simulation example for four simulated datasets

We fit the model used in the simulation example with an additional temporal component: $\mathbf{Y}_{it} \sim NegBin(\mathbf{p}_{it})$, where $\mathbf{p}_{it} = \frac{\exp[\alpha_i+\zeta_{it}+\mathbf{X}_{it}^T\boldsymbol{\beta}_i]}{1+\exp[\alpha_i+\zeta_{it}+\mathbf{X}_{it}^T\boldsymbol{\beta}_i]}$ and t, which indexes time, ranges from 2001 to 2006. We specify $\zeta_{it} \sim AR(1)$. Notice that the region-specific model remains constant for all years within a region. Using our modified model, we obtain estimates of the number of unreported murders from 2000-2004 for each individual municipality and year.

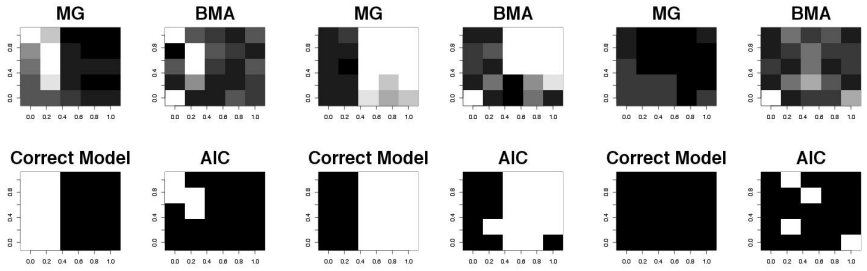
We include data from five administrative lists and include up to three-way interaction terms as potential covariates, resulting in 26 potential covariates for each region. We enforce the restriction that main effects and an intercept must be present, leaving 20 covariates in each region for the algorithm to include or exclude appropriately. The results of our analysis are shown in Table 3 and Figures 5 and 4. If we aggregate our results to the level of coarseness of previous analyses, we find that for the most part, our analyses are consistent with the others in that the previous confidence and credible intervals generally contain our estimates. So, despite permitting a much more discretized, detailed view of the pattern of undocumented killings, this analysis largely confirms the previous results obtained by large-scale spatial aggregation. However, our more detailed analysis has allowed other aspects of the dynamics to emerge. We find that for the majority of

the regions, the number of undocumented killings peaked in 2004. Nunchia and Recetor, however, see an earlier spike in 2003. Interestingly, these regions are neighbors. As our model did not include sharing of information in the time series component, this shouldn't be an artifact of the spatial smoothing. Our overall estimate of the total number of undocumented killings is larger than previous estimates– we estimate a total of 4,777 undocumented killings compared to the previous estimate of XXX.

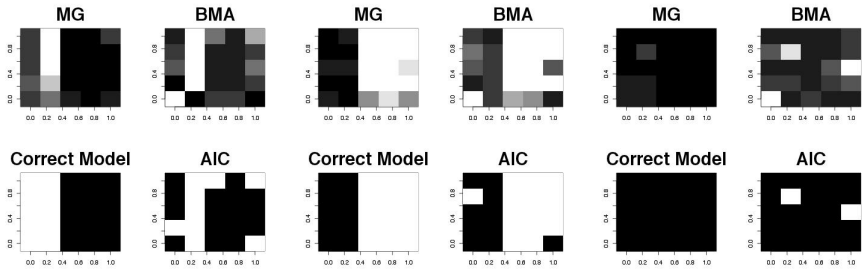
Fig. 3: Comparison of inclusion probabilities of β_{12}, β_{13}, and β_{23} for the $N = 5,317$ simulation



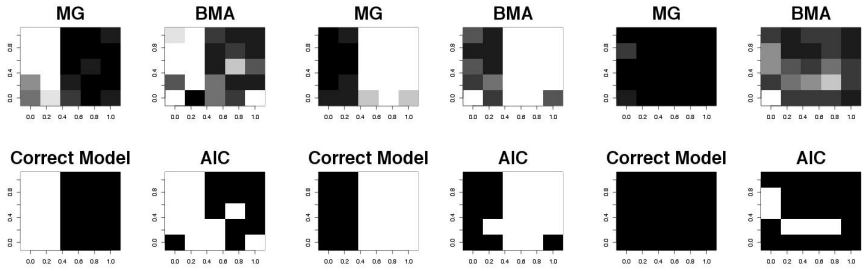
Comparison of inclusion probabilities of β_{12}, β_{13}, and β_{23} for the $N = 7,087$ simulation



Comparison of inclusion probabilities of β_{12}, β_{13}, and β_{23} for the $N = 14,165$ simulation



Comparison of inclusion probabilities of β_{12}, β_{13}, and β_{23} for the $N = 35,392$ simulation



	2001	2002	2003	2004	2005	2006
aguazul	58 [24 , 116]	74 [31 , 143]	109 [48 , 210]	296 [135 , 567]	174 [79 , 334]	56 [23 , 115]
chameza	2 [0 , 6]	4 [1 , 9]	7 [2 , 17]	3 [0 , 8]	2 [0 , 6]	2 [0 , 5]
hato corozal	1 [0 , 4]	2 [0 , 4]	3 [0 , 6]	3 [1 , 7]	3 [0 , 6]	3 [0 , 7]
la salina	1 [0 , 2]	1 [0 , 2]	1 [0 , 2]	2 [0 , 4]	1 [0 , 3]	1 [0 , 2]
mani	18 [7 , 35]	14 [5 , 28]	17 [6 , 33]	40 [16 , 76]	23 [9 , 45]	10 [3 , 22]
monterrey	18 [7 , 35]	26 [11 , 48]	25 [10 , 47]	38 [16 , 72]	33 [14 , 63]	12 [4 , 26]
nunchia	1 [0 , 3]	1 [0 , 3]	2 [0 , 5]	4 [1 , 9]	2 [0 , 5]	2 [0 , 5]
orocue	2 [0 , 5]	2 [0 , 5]	2 [0 , 6]	5 [1 , 12]	5 [1 , 10]	4 [1 , 8]
paz de ariporo	3 [0 , 7]	3 [0 , 7]	11 [5 , 21]	18 [8 , 34]	12 [5 , 24]	10 [4 , 19]
pore	5 [2 , 10]	6 [3 , 12]	5 [2 , 10]	7 [3 , 14]	6 [2 , 11]	4 [1 , 9]
recetor	2 [0 , 6]	2 [0 , 7]	11 [4 , 23]	3 [0 , 7]	3 [0 , 8]	2 [0 , 6]
sabanalarga	3 [0 , 6]	2 [0 , 6]	2 [0 , 6]	4 [1 , 9]	3 [0 , 6]	3 [0 , 6]
sacama	1 [0 , 3]	1 [0 , 3]	1 [0 , 3]	1 [0 , 4]	1 [0 , 3]	1 [0 , 3]
san luis de palenque	1 [0 , 4]	1 [0 , 4]	2 [0 , 6]	2 [0 , 5]	2 [0 , 5]	2 [0 , 4]
tamara	2 [0 , 4]	2 [0 , 4]	2 [0 , 5]	2 [0 , 5]	2 [0 , 5]	3 [1 , 7]
tauramena	23 [8 , 46]	28 [10 , 55]	19 [6 , 39]	35 [13 , 68]	34 [13 , 66]	22 [8 , 44]
trinidad	3 [1 , 7]	3 [1 , 7]	4 [1 , 9]	6 [2 , 12]	5 [2 , 11]	5 [2 , 11]
villanueva	15 [5 , 30]	24 [10 , 46]	43 [19 , 78]	101 [47 , 183]	29 [12 , 55]	17 [6 , 35]
yopal	408 [114 , 1158]	455 [130 , 1285]	788 [225 , 2209]	1327 [385 , 3741]	467 [133 , 1321]	178 [48 , 518]

Table 3: The posterior mean [credible interval] by year and municipality of the number of undocumented killings in Casanare, Colombia.

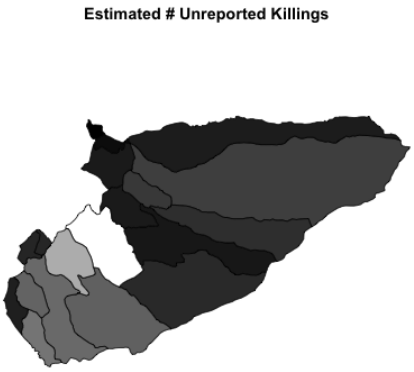


Fig. 4: Posterior mean estimate of the number of unreported killings by region (left) and the mean estimates as a percent of the total population (right).

8. Discussion and Future Directions

Recently, the variable selection and shrinkage estimation literature has focused mostly on high-dimensional cases where p is large, and perhaps $p > n$. The main threads in this area have been the development of shrinkage priors that approximate the optimal properties of Bayes variable selection, and efficient stochastic search algorithms for exploring the posterior model space when using BVS with large p. While many interesting applications fit this paradigm, there are also numerous settings in which variable selection is important where the challenge is not the number of covariates, but rather complex structure in the data that should be exploited to achieve good inference and prediction. Here we have focused on local modeling scenarios in which n is large relative to p, but the data are broken into many clusters *a priori*, and n is not large relative to p in each cluster. This situation arises often in social sciences, which rely heavily on structured surveys, and in spatial statistics and time series applications. The main obstacle to use of BVS in these situations is the need to borrow strength across the clusters in estimating variable inclusion probabilities. The difficulty of working with multivariate generalizations of the Beta distribution makes this difficult to achieve using existing approaches to modeling inclusion probabilities.

This work introduces a novel class of priors on variable inclusion probabilities that easily accommodates dependence and leads to simple Gibbs sampling algorithms. Our MG priors form a very similar class to the Beta family, making it possible to reflect a wide array of prior beliefs by specifying different prior hyperparameters. Our priors rely on probit transforms of normals with varying location and scale, creating a two-parameter family on the unit interval. Because MG priors have a Gaussian latent variable representation, it is simple to generate dependent inclusion probabilities with a Gaussian

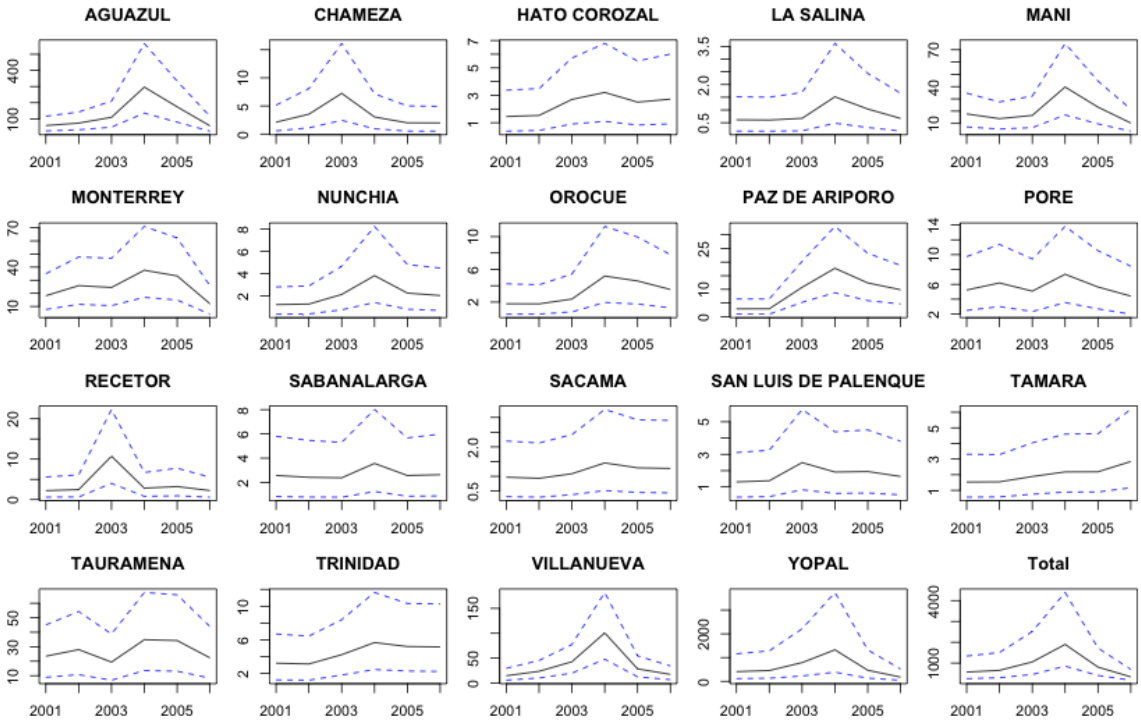


Fig. 5: The posterior mean (solid) and 95% posterior credible intervals (dashed) of the number of undocumented killings by region across the years 2001 to 2006.

copula. Our method can be used in virtually any survey, time-series, or spatial setting, a few of which we illustrate here with simulations and real data examples.

We anticipate that this method can be applied easily to numerous problems in the social sciences, allowing for inference on how important predictors vary across clusters within a larger sample in a manner not previously possible. Future methodological work could include development of more efficient stochastic search methods for updating variable inclusion probabilities that would allow for extension to larger p settings. One can also imagine an analogue of this approach that develops classes of shrinkage priors that easily accommodate structured dependence in shrinkage factors.

References

Bhattacharya, A., Pati, D., Pillai, N. S. & Dunson, D. B. (2012). Bayesian shrinkage. *arXiv preprint arXiv:1212.6088* .

Bruno, G., andE. Merletti, R. L., Biggieri, A., McCarty, D. & Pagano, C. (1994). National diabetes programs: Applications of capture-recapture to "count" diabetes. *Diabetes Care* 17 548–556.

Carvalho, C. M., Polson, N. G. & Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* 97 465–480.

Castillo, I. & van der Vaart, A. (2012). Needles and straws in a haystack: Posterior concentration for possibly sparse sequences. *Annals of Statistics, to appear* .

Fienberg, S. E. (1972). Multiple-recapture census for closed populations and incomplete contingency tables. *Biometrika* 59 591–603.

George, E. & McCulloch, R. (1997). Approaches for bayesian variable selection. *Statistica Sinica* 7 339–374.

Guberek, T., Guzmán, D., Price, M., Lum, K. & Ball, P. (2010). To count the uncounted: An estimation of lethal violence in casanare. Tech. rep., Benetech Human Rights Program.

Guzmán, D., Guberek, T., Hoover, A. & Ball, P. (2007). Missing people in casanare. Tech. rep., Benetech Human Rights Data Analysis Group, http://www.hrdag.org/about/colombia.shtml.

Hans, C., Dobra, A. & West, M. (2007). Shotgun stochastic search for "large p" regression. *Journal of the American Statistical Association* 102 507–516.

Lum, K., Price, M., Guberek, T. & Ball, P. (2010). Measuring elusive populations with bayesian model averaging for multiple systems estimation: A case study on lethal violations in casanare, 1998 - 2007. *Statistics, Politics, and Policy* 1.

Polson, N. G., Scott, J. G. & Windle, J. (2012). Bayesian inference for logistic models using polya-gamma latent variables. Tech. rep., University of Texas at Austin.

Raftery, A., Madigan, D. & Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* 92 179–191.

Reeves, G. & Gastpar, M. (2012). The sampling rate-distortion tradeoff for sparsity pattern recovery in compressed sensing. *Information Theory, IEEE Transactions on* 58 3065–3092.

Rue, H. & Martino, S. (2009). *INLA: Functions which allow to perform a full Bayesian analysis of structured additive models using Integrated Nested Laplace Approximaxion*. R package version 0.0.

Scott, J. & Berger, J. (2010). Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics* 38 2587–2619.

Scott, J. G. & Carvalho, C. M. (2008). Feature-inclusion stochastic search for gaussian graphical models. *Journal of Computational and Graphical Statistics* 17 790–808.

Smith, M. & Fahrmeir, L. (2007). Spatial bayesian variable selection with application to functional magnetic resonance imaging. *Journal of the American Statistical Association* 102 417–431.

Trippa, L., Müller, P. & Johnson, W. (2011). The multivariate beta process and an extension of the polya tree model. *Biometrika* 98 17–34.