Statistica Sinica Preprint No: SS-2015-0382R2	
Title:	A Bayesian generalized CAR model for correlated signal detection
Manuscript ID:	SS-2015-0382R2
URL:	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202015.0382
Complete List of Authors:	Andrew Brown Gauri Datta and Nicole Lazar
Corresponding Author:	Andrew Brown
E-mail:	ab7@g.clemson.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0382

A BAYESIAN GENERALIZED CAR MODEL FOR CORRELATED SIGNAL DETECTION

D. Andrew Brown, Gauri S. Datta and Nicole A. Lazar

Clemson University and University of Georgia

Abstract: Over the last decade, large-scale multiple testing has found itself at the forefront of modern data analysis. Often data are correlated, so that the observed test statistic used for detecting a non-null case, or signal, at each location in a dataset carries some information about the chances of a true signal at other locations. Brown, Lazar, Datta, Jang, and McDowell (2014) proposed, in the neuroimaging context, a Bayesian multiple testing model that accounts for the dependence of each volume element on the behavior of its neighbors through a conditional autoregressive (CAR) model. Here, we propose a generalized CAR model that allows for inclusion of points with no neighbors at all, something that is not possible under conventional CAR models. We consider also neighborhoods based on criteria other than physical location, such as genetic pathways in microarray determined from existing biological knowledge. This provides a unified framework for the simultaneous modeling of dependent and independent cases, resulting in stronger Bayesian learning in the posterior. We justify the selected prior distribution and prove that the resulting posterior distribution is proper. We illustrate the utility of our proposed model by using it to analyze both simulated and real microarray data in which the genes exhibit dependence that is determined by physical adjacency on a chromosome or predefined gene pathways.

Key words and phrases: conditional autoregressive model, enrichment, microarray, multiple testing, significance analysis of microarrays, spike-and-slab prior

1. Introduction

High throughput data analysis presents many challenges across a variety of disciplines. Many of the problems are ubiquitous in the sciences, and exacerbated when the datasets are massive in size. Often, the goal is to detect the presence of a signal over a very large number of cases, creating a massive multiple testing problem. Prior to the last two decades, most multiple testing procedures were constructed to control an overall error rate for a relatively small number of simultaneous tests (Efron (2010)). The advent of high throughput technology revealed that classical procedures can be inappropriate in the presence of thousands of simultaneous tests (Benjamini (2010)).

Suppose our data consist of J cases, each of which arises independently from a normal

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0382

distribution with case-specific mean, $y_j \sim N(\theta_j, \sigma^2),\ j = 1, \ldots, J$. For example, in genetics, y_j may be the test statistic quantifying the differential expression of gene j between cancerous and healthy tissue (Efron and Tibshirani (2002)). In functional neuroimaging, signals are observed over time at thousands of points in the brain and a test statistic y_j is calculated at each point j to summarize the observed difference in that area's signal between some stimulus condition versus a baseline (Friston, Holmes, Worsley, Poline, Frith, and Frackowiak (1995)). A common question of interest is whether or not $\theta_j = 0$ at each j, and a hypothesis test is conducted at each of thousands of locations to determine which of the θ_j are non-zero, indicating an interesting signal. The goal is to find a statistical procedure which corrects for multiplicity without sacrificing too much sensitivity.

A similar issue arises in variable selection, where one is interested in determining which variables contribute in a meaningful way to the observed response. A Bayesian approach is to assume the coefficient corresponding to each variable, β_j, belongs to either the null class in which $\beta_j \sim N(0, \sigma^2)$, or the non-null class, $\beta_j \sim N(\theta_j, \sigma^2),\ \theta_j \neq 0$ (e.g., Mitchell and Beauchamp (1988); George and McCulloch (1993); Scott and Berger (2006); Efron (2008)). Each β_j is assigned a prior probability p of belonging to the null class with the interpretation that $p \approx 1$ models very sparse signals. This mixing proportion, p, is usually unknown, but it can be assigned a prior distribution to reflect the researcher's beliefs about the level of sparsity in the data, or it can be estimated via empirical Bayes. The posterior inclusion probabilities are estimated from the posterior distribution. A Beta prior on p induces a multiplicity adjustment in that the model automatically penalizes for the number of tests in *a posteriori* probability statements. Scott and Berger (2010) discuss this issue and the conditions under which multiplicity correction can be induced.

Much of the work thus far developed assumes independent hypothesis tests. This assumption is untenable in many applications. Nontrivial dependence structures are known to exist in neuroimaging data (Lee, Jones, Caffo, and Bassett (2014)), syndromic surveillance (Banks, Datta, Karr, Lynch, Niemi, and Vera (2012)), gene microarray (Zhao, Kang, and

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0382

Yu (2014)), and RNA sequencing (RNA-seq; Love, Huber, and Anders (2014)). Correlation can cause the null distribution of the observed test statistics to be over- or under-dispersed relative to the theoretical null under independence. Consequently, either too few or too many test statistics may be declared significant. The deleterious impact correlation can have on empirical Bayes methods and false discovery rate control (FDR; Benjamini and Hochberg (1995)) was investigated in Qiu, Klebanov, and Yakovlev (2005). Efron (2007) focused on the effects of dependence on the distribution of test statistics. Work has been done on incorporating known dependence structure into Bayesian models for identification of interesting cases (e.g., Smith and Fahrmeir (2007); Li and Zhang (2010); Stingo, Chen, Tadesse, and Vannucci (2011); Lee, Jones, Caffo, and Bassett (2014); Zhao, Kang, and Yu (2014); Zhang, Guindani, Versace, and Vannucci (2014)), but it has been limited, particularly with respect to exploring the multiple testing adjustments incurred through data-dependent estimation of inclusion probabilities.

Many datasets include isolated observations. For example, genes in microarray data share common pathways, but many genes are in no pathway at all. It is important to include as many cases as possible when evaluating the posterior distribution. A standard CAR model assumes every observation has at least one neighbor, so that one is forced to either use an inappropriate neighborhood structure or exclude isolated points. In the current paper, we extend a model proposed by Brown, Lazar, Datta, Jang, and McDowell (2014) for Bayesian multiple testing and discuss more general neighborhood structures to provide a unified treatment of cases with at least one neighbor and with no neighbor. We use a less restrictive improper prior distribution for the variance components and establish the propriety of the posterior distribution of our model, ensuring that inferences are valid. In addition to allowing the newly proposed model to identify isolated non-null cases, we demonstrate that the inclusion of isolated points results in stronger Bayesian learning and improved estimation of the signal strengths of the selected cases.

We motivate a common model used in Bayesian signal detection in Section 2. This leads

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0382

to our proposed extension to accommodate local dependence. We prove the propriety of the posterior distribution of our proposed model and discuss computation. In Section 3, we report on simulated correlated microarray data used to study our model's performance against a prior assuming independence and assuming a conventional CAR structure. We also compare it against results obtained from the significance analysis for microarrays (SAM) procedure (Efron, Tibshirani, Storey, and Tusher (2001); Tusher, Tibshirani, and Chu (2001)). We apply our procedure to two gene microarray datasets in Section 4, one using dependence determined by adjacency on a chromosome, and the other with gene pathways defining the neighborhoods. Section 5 concludes with a discussion.

2. Methods

2.1 Mixture Priors for Multiplicity Adjustment

To facilitate a Bayesian multiple testing correction, we postulate a "two groups model" (Efron (2008)). We assume two possible cases for each of the observed test quantities, reflected through the prior on θ_j,

$$\pi(\theta_j \mid p,\ \tau^2) = p\delta_0(\theta_j) + (1-p)\varphi_{0,\tau^2}(\theta_j) \quad j = 1, \ldots, J, \tag{2.1}$$

where $\delta_0(\cdot)$ is the Dirac delta spike at zero and $\varphi_{0,\tau^2}(\cdot)$ is the Gaussian density function with mean zero and variance τ^2. So-called "spike-and-slab" or "spike-and-bell" priors of this form are standard in Bayesian variable selection, introduced by Mitchell and Beauchamp (1988) for variable selection in linear regression. The model was used by Geweke (1996), who provided a procedure for selecting models subject to order constraints among the variables included in each. Similarly, George and McCulloch (1993) treated each regression coefficient as arising from a mixture of two continuous distributions with different variances for stochastic search variable selection. Literature on Bayesian variable selection was reviewed in Clyde and George (2004) and O'Hara and Sillanpää (2009). Scott and Berger (2006) explored Model (2.1) and

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0382

ways in which it induces multiplicity correction, along with graphical displays and decision rules for inference.

By allowing p to be determined by the data, the joint posterior distributions obtained from spike-and-slab priors adapt to the number of tests, resulting in the posterior inclusion probabilities, $p_j := P(\theta_j \neq 0 \mid \mathbf{y})$, being penalized to account for the multiple tests (Scott and Berger (2006, 2010)). Specifically, Scott and Berger (2006) used a Beta$(\alpha, 1)$ prior density on p, where α is specified, and with $\pi(\tau^2,\ \sigma^2) = (\tau^2 + \sigma^2)^{-2}$ as a prior density for the variance components. Under this model, Scott and Berger (2006, Lemma 3) showed that

$$p_j = 1 - E\left[\left(1 + \frac{1-p}{p}\sqrt{\frac{\sigma^2}{\sigma^2+\tau^2}}\exp\left(\frac{y_j^2\tau^2}{2\sigma^2(\sigma^2+\tau^2)}\right)\right)^{-1}\right], \tag{2.2}$$

where the expectation is taken with respect to the joint posterior distribution of p, σ^2, and τ^2.

2.2 Incorporation of Local Dependence

Posterior inference can be sharpened if we exploit correlation among potential predictors in the search for interesting signals. Brown, Lazar, Datta, Jang, and McDowell (2014) proposed allowing the continuous component of (2.1) to share information across observations by writing θ_j as $\gamma_j\mu_j$, $\gamma_j \overset{\text{iid}}{\sim} \text{Bern}(1-p)$ and μ_j is Gaussian, so that $\mathbf{y} \sim N(\mathbf{\Gamma}\boldsymbol{\mu}, \sigma^2\mathbf{I})$, with $\mathbf{y} = (y_1, \ldots, y_J)^T$, $\mathbf{\Gamma} = \text{diag}\{\gamma_i,\ i = 1, \ldots, J\}$, and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_J)^T$. The lattice structure of datasets such as those arising from neuroimaging and gene microarray makes a conditional autoregressive model (CAR; Besag (1974)) a natural choice for incorporating local dependence into the prior on $\boldsymbol{\mu}$. Since the potential non-null signals are expected to be as much positive as negative, *a priori* it is reasonable to assume such signals have zero means. Thus, we consider prior distributions of the form $\mu_j \mid \boldsymbol{\mu}_{(-j)} \sim N\left(\sum_{i=1}^J c_{ji}\mu_i, \tau_j^2\right),\ j = 1, \ldots, J$, where $\boldsymbol{\mu}_{(-j)} = (\mu_1, \ldots, \mu_{j-1}, \mu_{j+1}, \ldots, \mu_J)^T$, $c_{jj} = 0$, and $c_{ji} = 0$ except when cases j and i are neighbors. The intrinsic autoregressive model (IAR; Besag, York, and Mollié (1991)) takes $c_{ji} = w_{ji}/w_{j\cdot}$ and $\tau_j^2 = \tau^2/w_{j\cdot}$, where $w_{ji} \neq 0$ if and only if sites j and i are neighbors

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0382

and $w_{j\cdot} = \sum_{i=1}^{J} w_{ji}$. Under an IAR model for $\boldsymbol{\mu}$, the prior density is

$$\pi(\boldsymbol{\mu} \mid \tau^2) \propto \exp\left(-\frac{1}{2\tau^2}\boldsymbol{\mu}^T(\mathbf{D}_w - \mathbf{W})\boldsymbol{\mu}\right), \tag{2.3}$$

where $\mathbf{D}_w = \text{diag}\{w_{j\cdot},\ j = 1, \ldots, J\}$ and $\mathbf{W} = \{w_{ji}\}_{j,i=1}^{J}$. Note that $(\mathbf{D}_w - \mathbf{W})\mathbf{1} = \mathbf{0}$ so that the IAR is improper. However, a "propriety parameter", ρ, can be used such that $\mu_j \mid \boldsymbol{\mu}_{(-j)} \sim N(\rho \sum_{i=1}^{J} w_{ji}\mu_i / w_{j\cdot},\ \tau^2/w_{j\cdot})$ with precision matrix $\tau^{-2}(\mathbf{D}_w - \rho\mathbf{W})$. This matrix is nonsingular if $\lambda_1^{-1} < \rho < \lambda_J^{-1}$, where $\lambda_1 < 0$ and $\lambda_J > 0$ are the smallest and largest eigenvalues of $\mathbf{D}_w^{-1/2}\mathbf{W}\mathbf{D}_w^{-1/2}$, respectively (Banerjee, Carlin, and Gelfand (2015)).

Any data having a lattice structure with known or suspected correlations occurring along predefined networks can be modeled with a CAR model. For instance, genes in microarray are known to express themselves in clusters along a chromosome (e.g., Xiao, Reilly, and Khodursky (2009)), or to behave in concert along specific gene pathways (Subramanian, Tamayo, Mootha, Mukherjee, Ebert, Gillette, Paulovich, Pomeroy, Golub, Lander, and Mesirov (2005)). Neighborhoods can be defined in terms of adjacency on a chromosome or based on predefined pathways determined from prior knowledge. Care should be taken in defining neighborhoods, though, as such datasets often include genes that are not members of any known pathway and thus are isolated. Including isolated points in the IAR induces zero rows in the precision matrix, a problem that cannot be fixed with a propriety parameter. In response, we adjust the neighborhood weights to allow for inclusion of the isolated points while avoiding a singular precision matrix.

We modify the usual IAR model by defining the neighborhood weights about μ_j to be $c_{ji} = w_{ji}/(d + w_{j\cdot})$ with conditional variance $\tau^2/(d + w_{j\cdot})$, where $d > 0$. The consequent precision matrix is $\tau^{-2}(\mathbf{D}_w + d\mathbf{I} - \mathbf{W})$. Then $\mathbf{x}^T(\mathbf{D}_w + d\mathbf{I} - \mathbf{W})\mathbf{x} = \sum_{i=1}^{J} dx_i^2 + (1/2)\sum_i \sum_j w_{ij}(x_i - x_j)^2 \geq 0$, with strict inequality for $\mathbf{x} \neq \mathbf{0}$. Thus, with $d > 0$, we are able to include isolated points in the model while maintaining the propriety of the distribution. If we take $d = 1$, then for any isolated point j', $w_{j'\cdot} = 0$ so that $E(\mu_{j'} \mid \boldsymbol{\mu}_{(-j')}) = 0$, $Var(\mu_{j'} \mid \boldsymbol{\mu}_{(-j')}) = \tau^2$,

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0382

and $\mu_{j'} \mid \tau^2 \sim N(0, \tau^2)$, independently of other points. Hence, we can facilitate conditional independence of μ_j while allowing all J points to share information about plausible values of the hypervariance through the prior distribution on the variance components. Taking this view, we can express the traditional IAR model as a special case in which every point in a dataset has at least one neighbor and $d = 0$.

Let the joint density of the data and parameters be given by $f(\mathbf{y}, \boldsymbol{\psi}, \boldsymbol{\mu}, p, \boldsymbol{\gamma})$, where $\boldsymbol{\psi}$ contains nuisance parameters modeled in the prior distribution. When $\gamma_j = 0$ for all j, $\boldsymbol{\mu}$ does not appear in the resulting likelihood and thus is Bayesianly unidentified (Gelfand and Sahu (1999)). This means that $\int_{\boldsymbol{\mu}} f(\mathbf{y}, \boldsymbol{\psi} \mid \boldsymbol{\gamma} = \mathbf{0}, \boldsymbol{\mu}) \pi(\boldsymbol{\mu}) d\boldsymbol{\mu} \equiv \int_{\boldsymbol{\mu}} f(\mathbf{y}, \boldsymbol{\psi} \mid \boldsymbol{\gamma} = \mathbf{0}) \pi(\boldsymbol{\mu}) d\boldsymbol{\mu} = f(\mathbf{y}, \boldsymbol{\psi} \mid \boldsymbol{\gamma} = \mathbf{0}) \int_{\boldsymbol{\mu}} \pi(\boldsymbol{\mu}) d\boldsymbol{\mu}$, so that we must have a proper prior on $\boldsymbol{\mu}$ for the posterior distribution to be proper, making the inclusion of ρ necessary when $d = 0$. See McLachlan and Peel (2000, Ch. 4) for further discussion of prior distributions in finite mixture models.

Usually, there is little direct information available about ρ, so estimating it may be difficult. Previous work has shown that appreciable interaction between adjacent points only occurs when ρ is close to its upper bound under the $d = 0$ model (Banerjee, Carlin, and Gelfand (2015)). To give the data more freedom in determining the spatial association without specific regard for interpretability, we consider the prior $\pi_\rho(\rho) \propto I(\lambda_1^{-1} < \rho < \lambda_J^{-1})$. It is important to note that inclusion of ρ is still possible when $d > 0$, provided that ρ is bounded between the reciprocals of the smallest and largest eigenvalues of $(\mathbf{D}_w + d\mathbf{I})^{-1/2}\mathbf{W}(\mathbf{D}_w + d\mathbf{I})^{-1/2}$.

An additional advantage of including ρ in the joint model for $\boldsymbol{\mu}$ is that, under positive spatial association, the posterior distribution becomes insensitive to the choice of d in the neighborhood weights. This is because as d grows, ρ is allowed to increase as well. In other words, if $d_1 < d_2$, then $\lambda_{J,1}^{-1} < \lambda_{J,2}^{-1}$, where $\lambda_{J,i} > 0$ is the maximum eigenvalue of $(\mathbf{D}_w + d_i\mathbf{I})^{-1/2}\mathbf{W}(\mathbf{D}_w + d_i\mathbf{I})^{-1/2},\ i = 1, 2$. A proof is in the Supplementary Material.

We also wish to avoid strong information about either the noise variance, σ^2, or the hypervariance, τ^2. Gelman (2006) suggested that the priors specified for the variance pa-

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0382

rameters in hierarchical models may have a disproportionate effect in that they can restrict posterior inference. Conversely, priors used for scale hyperparameters that are intended to be noninformative may, in fact, be *too* weak in placing considerable probability on unreasonable extreme values in the posterior. To address this, Gelman (2006) proposed the use of a weakly informative prior on the scale hyperparameter such as the folded-t distribution. Scott and Berger (2006) argued for the use of a joint prior on σ^2 and τ^2 with density $\pi_{(\tau^2,\sigma^2)}(\tau^2,\ \sigma^2) = (\tau^2+\sigma^2)^{-2} = (\sigma^2)^{-1}(1+\tau^2/\sigma^2)^{-2}(\sigma^2)^{-1} \equiv \pi_{\tau^2|\sigma^2}(\tau^2 \mid \sigma^2)\pi_{\sigma^2}(\sigma^2)$ so that a standard improper prior on σ^2 can be used while scaling τ^2 by σ^2. The prior on $\tau^2 \mid \sigma^2$ is similar to the prior on τ^2 which results from placing a folded-t_2 prior on τ with scale σ. Supplementary Figure 1 illustrates a slight difference between the two priors for small values of τ^2 so that the Scott-Berger (SB) prior on τ^2 is slightly less informative than a folded-t. However, the SB prior and the folded-t-based prior are tail equivalent in that the ratio of the two densities is $O(1)$ as $\tau^2 \to \infty$. We thus follow the precedent set by Scott and Berger (2006) and use the same joint prior distribution on τ^2 and σ^2.

This leads to the model

$$\begin{gathered}
y_j \mid \gamma_j,\ \mu_j,\ \sigma^2 \overset{\text{indep}}{\sim} N(\gamma_j\mu_j, \sigma^2); \qquad \gamma_j \mid p \overset{\text{iid}}{\sim} \text{Bern}(1-p), \quad j = 1,\ldots,J \\
\mu_j \mid \boldsymbol{\mu}_{(-j)},\ \tau^2,\ \rho \sim N\left(\sum_{i=1}^{J} \frac{\rho w_{ji}\mu_i}{d+w_{j\cdot}},\ \frac{\tau^2}{d+w_{j\cdot}}\right), \quad d \geq 0, \quad j = 1,\ldots,J \\
p \sim \text{Beta}(\alpha, 1), \quad \alpha \geq 1; \qquad \rho \sim \text{Unif}(\nu_1^{-1},\ \nu_J^{-1}) \\
\pi_{\tau^2|\sigma^2}(\tau^2 \mid \sigma^2) = \left(\frac{1}{\sigma^2}\right)\left(1+\frac{\tau^2}{\sigma^2}\right)^{-2}, \quad \tau^2 > 0; \qquad \pi_{\sigma^2}(\sigma^2) = \frac{1}{\sigma^2}, \quad \sigma^2 > 0,
\end{gathered} \tag{2.4}$$

where ν_1 and ν_J are the smallest and largest eigenvalues of $(\mathbf{D}_w + d\mathbf{I})^{-1/2}\mathbf{W}(\mathbf{D}_w + d\mathbf{I})^{-1/2}$, respectively. Since we are using an improper prior in (2.4), the posterior density is not guaranteed to be integrable. We provide a proof in the Appendix that the posterior distribution is indeed proper.

The practical effect of ρ having room to increase along with d in our proposed model can be seen through simulation. Consider a 20×20 array of observations arising from both

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0382

null and non-null distributions. The activation pattern was created by drawing from an Ising distribution (e.g., Higdon (1994)), $p(\mathbf{x}) \propto \exp(\beta \sum_{i\sim j} I(x_i = x_j)),\ \mathbf{x} \in \{0,1\}^{400}$, with interaction parameter $\beta = 0.35$. The null cases ($x_i = 0$) were drawn from $N(0,1)$ and the non-null cases were drawn from $N(3.5,1)$. The binary activation pattern and simulated data array are displayed in Supplementary Figures 2 and 3. We use these data to estimate the posterior distributions under model (2.4) with $d = 0,\ 1,$ and 5. See Subsection 2.3 for implementation details. A descriptive measure of spatial association is Moran's I, $I(\mathbf{y}) = n \sum_i \sum_j w_{ij}(y_i - \overline{y})(y_j - \overline{y}) / [(\sum_{i\neq j} w_{ij}) \sum_i (y_i - \overline{y})^2]$, where values away from zero are evidence of spatial association according to the predefined neighborhood structure (e.g., Banerjee, Carlin, and Gelfand (2014, Sec. 4.1)). Figure 1 displays smoothed histograms of realizations of $I(\mathbf{y}^*)$ from 2,000 replications each from the three respective posterior predictive distributions, $p(\mathbf{y}^* \mid \mathbf{y}) = \int_{\boldsymbol{\Theta}} f(\mathbf{y}^* \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid \boldsymbol{y}) d\boldsymbol{\theta}$, along with the approximate marginal posterior densities of ρ. For each value of d, the posterior of ρ tends to concentrate near its upper bound and the posterior predictive densities of I are nearly indistinguishable. Regardless of the value of d, ρ adjusts accordingly and the overall spatial association is consistent with the data.

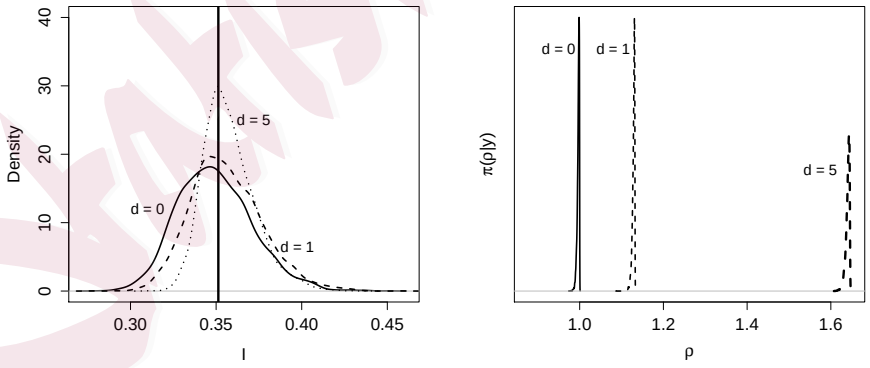

Figure 1: Smoothed histograms of 2,000 realizations of Moran's I from the corresponding posterior predictive distributions (left panel) and estimated marginal posterior distributions of ρ under model (2.4) (right panel). The dark vertical line in the left panel is at the observed value, $I(\mathbf{y})$.

Including as many cases as possible when evaluating the posterior distribution is important. It is often the case that the dataset to be analyzed contains subsets of correlated observations among many independent observations. A standard CAR structure assumes ev-

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0382

ery observation has at least one neighbor and this may be inappropriate, excluding isolated points for example. With our proposed approach, adjusting the weights with $d > 0$ in the denominator allows for the inclusion of all cases when evaluating the posterior distribution while simultaneously preserving the dependence among the cases sharing common networks as well as the independence of the isolated cases. In the sequel, we consider the performance of our model when $d = 0$ or $d = 1$. These are admittedly ad hoc values, and may not be appropriate for all situations.

2.3 Computational Implementation

We facilitate Gibbs sampling (Geman and Geman (1984)) by reparameterizing the model as $\tau^2 = \eta\sigma^2$. For ease of notation, take $\mathbf{D}_w^* := \mathbf{D}_w + d\mathbf{I}$. Then

$$\begin{aligned}\pi(\boldsymbol{\mu}, \boldsymbol{\gamma}, \sigma^2, \eta, \rho, p \mid \mathbf{y}) \quad \propto \quad & (\sigma^2)^{-J-1} \exp\left(-\frac{(\mathbf{y}-\boldsymbol{\Gamma}\boldsymbol{\mu})^T(\mathbf{y}-\boldsymbol{\Gamma}\boldsymbol{\mu})}{2\sigma^2}\right) \\ & \times|\eta(\mathbf{D}_w^* - \rho\mathbf{W})^{-1}|^{-1/2} \exp\left(-\frac{\boldsymbol{\mu}^T(\mathbf{D}_w^* - \rho\mathbf{W})\boldsymbol{\mu}}{2\eta\sigma^2}\right) \\ & \times p^{J-\sum_{i=1}^J \gamma_i + \alpha - 1}(1-p)^{\sum_{i=1}^J \gamma_i}(1+\eta)^{-2} I(\nu_1^{-1} < \rho < \nu_J^{-1}).\end{aligned}$$

The full conditional distribution of $\boldsymbol{\mu}$ is $N(\mathbf{R}\boldsymbol{\Gamma}\mathbf{y},\ \sigma^2\mathbf{R})$, where $\mathbf{R} = (\boldsymbol{\Gamma} + \eta^{-1}(\mathbf{D}_w^* - \rho\mathbf{W}))^{-1}$ (Carlin and Louis (2009)). To avoid matrix inversion with extremely large J, we update $\boldsymbol{\mu}$ element-wise. The full conditional distributions are given in the Supplementary Material.

We use rejection sampling to draw from the conditional distribution of η. However, the importance ratio determining the acceptance probability is $\eta^2/(1+\eta)^2 \to 0$ as $\eta \to 0$, meaning that iterations can slow down on this step with candidate densities that concentrate on extremely small values of η. Possible alternatives are an adaptive Metropolis algorithm (Carlin and Louis, 2009) or adaptive rejection sampling (Gilks and Wild, 1992).

We follow Carlin and Banerjee (2003) and use slice sampling (Neal (2003)) to draw from the full conditional distribution of ρ. Our experience is that the algorithm performs better with the "doubling" procedure outlined by Neal (2003) to adaptively determine good

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0382

proposal intervals. The determinant $|\mathbf{D}_w^* - \rho\mathbf{W}|^{1/2} \propto |\mathbf{I} - \rho(\mathbf{D}_w^*)^{-1/2}\mathbf{W}(\mathbf{D}_w^*)^{-1/2}|^{1/2}$ in the conditional density of ρ can be quickly computed using the eigenvalues of $(\mathbf{D}_w^*)^{-1/2}\mathbf{W}(\mathbf{D}_w^*)^{-1/2}$. Matrix computations can also be eased by calculating $\boldsymbol{\mu}^T\mathbf{D}_w^*\boldsymbol{\mu} = \sum_{j=1}^J (d + w_j.)\mu_j^2$ and $\boldsymbol{\mu}^T\mathbf{W}\boldsymbol{\mu}$ before searching for an acceptable update for ρ. These only need to be calculated once for each Gibbs iteration.

From (S1.1) in the Supplementary Material, we can see the strong dependence between $\boldsymbol{\gamma}$ and p in their conditional distributions. On the k^{th} iteration, if the sample draw $p^{(k)}$ is close to 1, then most of the draws $\gamma_i^{(k)},\ i = 1, \ldots, J$, will be zero. But then $\sum_i \gamma_i$ will be close to zero so that the conditional Beta density will concentrate close to 1, leading to another high value of p, and so on. Thus, in spite of the computationally convenient conditional conjugacy, an MCMC routine can get stuck in the region of the parameter space with most $\gamma_i = 0$, slowing convergence. The situation can be ameliorated by reparameterizing to eliminate boundary constraints on p and using Langevin-Hastings proposals to push the chain toward the posterior mode (Gilks and Roberts (1996)). Randomly mixing in ordinary Metropolis proposals in place of Langevin-Hastings proposals offers further improvements when the chain is far from the mode (Carlin and Louis (2009, Ch. 3)).

The quantities of interest are the marginal posterior inclusion probabilities for each signal j, $P(\gamma_j = 1 \mid \mathbf{y})$. To estimate this from the posterior sample draws, we recognize that in Model (2.4), $P(\gamma_j = 1 \mid \mathbf{y}) = E(p_j^*)$, where the expectation is taken with respect to $\pi(\boldsymbol{\mu}, \boldsymbol{\gamma}_{(-j)}, \sigma^2, \eta, \rho, p, \mid \mathbf{y})$, and $p_j^* := P(\gamma_j = 1 \mid \boldsymbol{\mu}, \boldsymbol{\gamma}_{(-j)}, \sigma^2, \eta, \rho, p, \mathbf{y})$ is given by

$$p_j^* = \frac{(1-p)\varphi_{0,\sigma^2}(y_j - \mu_j)}{(1-p)\varphi_{0,\sigma^2}(y_j - \mu_j) + p\varphi_{0,\sigma^2}(y_j)}. \tag{2.5}$$

This quantity can be estimated by $N^{-1}\sum_{i=1}^N \hat{p}_j^{*,(i)}$, where $\hat{p}_j^{*,(i)}$ is the plug-in estimate of p_j^* evaluated with the i^{th} draws $p^{(i)}, \mu_j^{(i)}, \sigma^{2,(i)}$ from the posterior, and N is the Monte Carlo sample size. Similarly, we use (2.2) to estimate the inclusion probabilities under the Scott-Berger model using $p^{(i)}, \sigma^{2,(i)}, \tau^{2,(i)}$ drawn from the appropriate posterior. Both of these estimators

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0382

are "Rao-Blackwellized" in the sense of Gelfand and Smith (1990) and thus have smaller Monte Carlo variance than other more naive estimators (Carlin and Louis (2009, Ch. 3)).

3. Simulation Studies

To evaluate performance, we simulated a dataset in a manner similar to Xiao, Reilly, and Khodursky (2009), resulting in a correlation structure as sometimes arises among genes on chromosomes. For the j^{th} gene on the i^{th} subject, $i = 1, \ldots, 10, j = 1, \ldots, 1000$, the observed expression level X_{ij} was drawn from $N(\mu_{ij}, 1)$. Five of the subjects were taken as controls with baseline (i.e., null case) expression levels over all 1,000 genes, so that $\mu_{ij} = 0$ for $i = 1, \ldots, 5,\ j = 1, \ldots, 1000$. The remaining five "treatment" subjects' data were simulated so that 100 genes were differentially expressed: For $i = 6, \ldots, 10,\ \mu_{ij} = 1.5,\ j = 1, \ldots, 20, 111, \ldots, 130, 211, \ldots, 230$ and $\mu_{ij} = -1.5,\ j = 311, \ldots, 330, 411, \ldots, 430$. For each control subject, we generated the gene expression levels by drawing the vector of observations $(X_{i,1}, \ldots, X_{i,1000})^T = \mathbf{X}_i \sim N(\mathbf{0}, \mathbf{I}),\ i = 1, \ldots, 5$. For the treatment group, the null cases were again simulated as i.i.d. standard normal. We modeled correlation among the differentially expressed cases by drawing each group of twenty test statistics as $\mathbf{X}_i^{(k)} \sim N(\boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}),\ i = 6, \ldots, 10$, where $\mathbf{X}_i^{(k)}, k = 1, \ldots, 5$, is the k^{th} cluster of non-null cases, (i.e., $\boldsymbol{\mu}^{(k)} = (1.5, 1.5, \ldots, 1.5)^T, k = 1, 2, 3,\ \boldsymbol{\mu}^{(k)} = (-1.5, -1.5, \ldots, -1.5)^T,\ k = 4, 5$), and $\boldsymbol{\Sigma} = \{0.9^{|i-j|}\}_{i,j=1}^{20}$. Pooled t statistics, $t_j,\ j = 1, \ldots, 1000$, were then calculated between the control and treatment conditions for each gene and subsequently normalized via probit transformation of the p-values (Efron (2010)), yielding test statistics $\mathbf{y} = (y_1, \ldots, y_{1000})^T$, where $y_j = \Phi^{-1}(F(t_j))$ with $F(\cdot)$ being the cdf of the t statistics.

We analyzed the simulated data using both our proposed model and the Scott-Berger (SB) model assuming independence. In our model, we considered the sharing of information across genes using three different neighborhood structures. These neighborhoods, displayed

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0382

graphically in Supplementary Figure 4, have adjacency matrices

$$\mathbf{W}_1 = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 & 0 \\ 1 & 0 & 1 & \dots & 0 & 0 \\ 0 & 1 & 0 & 1 & \dots & 0 \\ \vdots & & & \ddots & & \vdots \\ 0 & 0 & 0 & \dots & 1 & 0 \end{bmatrix}, \quad \mathbf{W}_2 = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & \dots & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & \dots & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & \dots & 0 & 0 & 0 \\ \vdots & & & & & \ddots & & \vdots & \\ 0 & 0 & 0 & 0 & 0 & \dots & 1 & 1 & 0 \end{bmatrix},$$

$$\mathbf{W}_3 = \begin{bmatrix} 0 & 1 & \frac{1}{2} & \frac{1}{3} & 0 & 0 & \dots & 0 & 0 & 0 \\ 1 & 0 & 1 & \frac{1}{2} & \frac{1}{3} & 0 & \dots & 0 & 0 & 0 \\ \frac{1}{2} & 1 & 0 & 1 & \frac{1}{2} & \frac{1}{3} & \dots & 0 & 0 & 0 \\ \vdots & & & & & & \ddots & & \vdots & \\ 0 & 0 & 0 & 0 & 0 & 0 & \dots & \frac{1}{2} & 1 & 0 \end{bmatrix}.$$

We implemented also the Significance Analysis for Microarrays (SAM) procedure as outlined in Efron (2010). SAM is a popular method for analyzing microarray data designed to approximately control the false discovery rate (FDR). For this procedure, we varied the FDR criterion between 0.05 and 0.15 to study performance across FDR levels commonly used in practice.

Each neighborhood was one in which every location had at least one neighbor so that, in (2.4), $w_{j\cdot} > 0$ for all j. We took $d = 0$, reducing to the IAR model considered in Brown, Lazar, Datta, Jang, and McDowell (2014). To enforce sparsity *a priori*, we took $\alpha = 150$ in the prior on p. In the SB model, we found that the best results were obtained with a uniform prior on p. The data generating mechanism was different from what is assumed under either our model or SB, allowing us to study robustness, as well.

We implemented the SB model using Gibbs sampling with nested rejection sampling steps for the non-standard distributions. To draw from the posterior of our proposed model, we used Gibbs sampling with nested rejection and slice sampling steps described in Subsec-

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0382

tion 2.3. The algorithms were coded entirely in R (R Core Team (2015)). For both models, a single chain used a burn-in period of 5,000 iterations followed by an additional 10,000 sampling iterations, thinning to every fifth draw for a final sample of size 2,000. We ran three chains in parallel and assessed convergence with trace plots and scale reduction factors for selected parameters (Gelman and Rubin (1992)). Upon attaining approximate convergence, the retained draws from each of the three chains were combined for a final Monte Carlo sample size of 6,000. The posterior inclusion probabilities were estimated using (2.2) for SB and (2.5) for our model. We thresholded the estimated posterior inclusion probabilities at 0.95.

Table 1 displays the empirical error rates for the SB model, the SAM procedure, and our model using $\mathbf{W}_1$, $\mathbf{W}_2$, and $\mathbf{W}_3$ as neighborhoods. The SB model results in the highest overall misclassification proportion, due to the false non-discoveries. In this case, the SB model is overcorrecting for multiplicity to the point that no cases are selected at all (hence the identically zero false discovery proportion). The SAM procedure performs better in terms of non-discoveries and overall misclassification proportion, but false positives account for 13% - 16% of all discoveries. Our proposed model performs better than SB or SAM, regardless of the selected neighborhood structure. The first-order neighborhood with unit weights ($\mathbf{W}_1$; top illustration in Supplementary Figure 4) performs best both in terms of non-discoveries and false discoveries, but all of the error rates are very close when compared to the other two approaches.

	SB	SAM(0.05)	SAM(0.10)	SAM(0.15)	CAR ($\mathbf{W}_1$)	CAR ($\mathbf{W}_2$)	CAR ($\mathbf{W}_3$)
FNP	0.100	0.094	0.088	0.086	0.055	0.058	0.062
FDP	0.000	0.125	0.133	0.158	0.000	0.022	0.024
MCP	0.100	0.094	0.089	0.087	0.052	0.056	0.060

Table 1: False non-discovery proportions (FNP), false discovery proportions (FDP), and misclassification proportions (MCP) for the simulated gene expression data

Figure 2 displays the empirical receiver operating characteristic (ROC) curves for the SB, SAM, and CAR($\mathbf{W}_1$) models. The ROC curves for the CAR($\mathbf{W}_2$) and CAR($\mathbf{W}_3$) models are virtually indistinguishable from CAR($\mathbf{W}_1$) and are not displayed. The approximate areas

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0382

under the curves for CAR($\mathbf{W}_1$), SB, and SAM are 0.897, 0.869, and 0.850, respectively. Hence, ours attains the best overall discriminatory power.

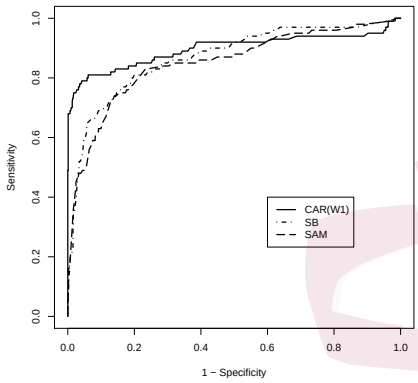

Figure 2: Empirical ROC curves for the simulated microarray data

Insight into the reasons for the discrepancies between our model and the SB model can be gained by examining the smoothed approximate posterior densities of p, σ^2, and $\tau^2 = \eta\sigma^2$, displayed in Figure 3. Incorporating the dependence results in much stronger Bayesian learning about these parameters. In this simulation, 100 out of 1,000 cases were non-null, so that we expected p to be large, though not exactly 0.9 since correlation among the cases reduces the effective sample size (Carlin and Louis (2009, Ch. 3)). That was indeed the case under our model. The SB model results in considerably lower estimates of p and weakly identified distributions of σ^2 and τ^2, contributing to the error rates observed in Table 1.

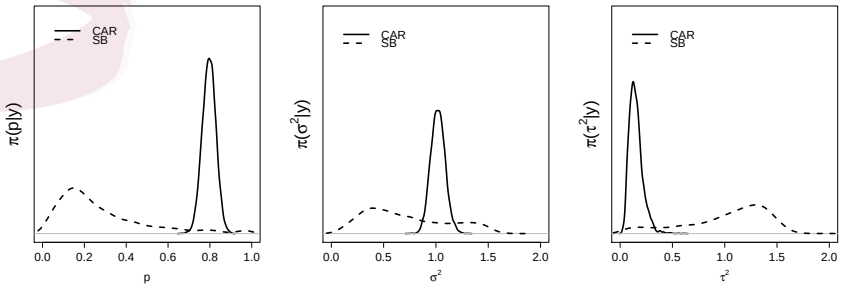

Figure 3: Smoothed posterior estimates of p, σ^2, and τ^2 for the simulated microarray data.

Supplementary Figure 5 plots the estimated inclusion probabilities versus the test statis-

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0382

tics. All of the statistics are assigned relatively high probabilities of being non-null under SB, but the lack of information about σ^2 and τ^2 prevents distinctions being drawn that are strong enough to pass a 0.95 threshold. Our approach, on the other hand, results in stronger statements about the likelihoods of cases being non-null. The 'jagged' quality of the curve corresponding to the CAR($\mathbf{W}_1$) model is due to the estimated inclusion probabilities being not a function of the y_j values alone, but also of their location with respect to nearby observed values. To see this, consider the circled point in Supplementary Figure 5, which corresponds to the circled case in the graphical depiction of the test statistics in Supplementary Figure 6. This relatively extreme observation is in the middle of uninteresting test statistics. This is a truly null case, so the incorporation of local dependence prevents a false discovery.

Instead of physical adjacency, it may be desired to facilitate the sharing of information within gene sets such as those used in enrichment analysis (Subramanian, Tamayo, Mootha, Mukherjee, Ebert, Gillette, Paulovich, Pomeroy, Golub, Lander, and Mesirov (2005)). To illustrate this, we simulated another dataset with expression characteristics similar to the simulation carried out in Efron and Tibshirani (2007). We again considered a collection of 1,000 genes, with genes 11-20, 111-130, 211-230, 311-330, 411-430 defining five different gene sets. The set consisting of genes 111-130 was simulated as differentially expressed by drawing them independently from $N(2.5, 1)$; similarly, the genes 411-430 were drawn independently from $N(-1.5, 1)$. The remaining genes, including those in the remaining gene sets, were all drawn from $N(0, 1)$. To distinguish dependence determined by pathway membership from dependence determined by physical adjacency, the genes were labeled and randomly permuted so that genes sharing common pathways were not adjacent in the physical sense. Many genes were members of no pathway and so were isolated.

Suppose a researcher were to mistakenly assume the dependence structure for these data followed the same pattern as the previous example in which every gene is correlated with its physical neighbors and the usual proper IAR with adjacency matrix $\mathbf{W}_1$ is used with $d = 0$ in (2.4). We compared the performance of this CAR structure to our proposed approach

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0382

with the adjacency matrix $\mathbf{W}$ determined by defining genes that are in the same set to be neighbors. To include all observations without reducing the rank of the precision matrix, we set $d = 1$ in (2.4) so that the marginal distributions of isolated μ_j were $N(0, \tau^2)$. We implemented both models using MCMC with the same burn-in and sampling settings as the previous simulation.

Supplementary Figure 7 displays the empirical ROC curves from our model under both neighborhood assumptions. It is apparent that making incorrect assumptions about the neighborhood structure severely inhibits the model's discriminatory power. In fact, thresholding the posterior inclusion probabilities at 0.95 as before results in no cases being identified at all under the physical adjacency assumption. The misspecified correlation results in the model overestimating the noise variability in the data, as is clear upon examination of Supplementary Figure 8. Superimposed on the histogram are two mean-zero normal densities with variances equal to the posterior means of σ^2 under both models. The overestimation of σ^2 results in poor identification of the non-null cases, indicated with the dark tick marks along the x-axis. Incorporating a more appropriate neighborhood structure results in improved estimation of the variance components and thus superior discrimination among cases.

Even with knowledge of an approximately correct dependence structure, one might want to use a standard CAR model by discarding the isolated cases. The isolated points provide information about the parameters common to both the null and non-null components of the data distribution and hence useful information would be discarded. Consider the smoothed marginal posterior densities of p, σ^2, and τ^2 displayed in Figure 4. Eliminating the isolated cases reduces J in (2.4), so we obtain considerably less posterior concentration about the error variance, which in turn affects the amount of information available to estimate the second variance component, τ^2. As is the case in any hypothesis testing scenario, the operating characteristics are directly affected by the amount of information we have about the error variability. This results in the "borderline" cases being misclassified as noise at a 0.95 inclusion probability threshold, thus increasing the false non-discovery proportion.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0382

(See Supplementary Figure 9.) The false discovery, false non-discovery, and misclassification proportions at the 0.95 threshold with and without isolated cases are given in Table 2.

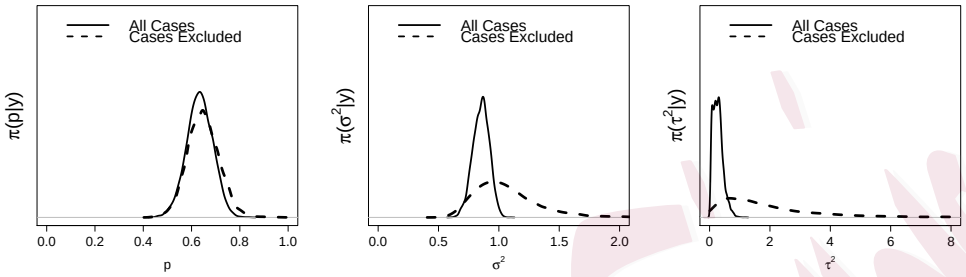

Figure 4: Smoothed posterior estimates of p, σ^2, and τ^2 with and without the isolated cases.

	With Isolated Cases	Without Isolated Cases
FNP	0.008	0.178
FDP	0.000	0.000
MCP	0.008	0.130

Table 2: Error rates for the simulated pathway example using pathway-based neighborhoods.

In addition to detection, there is often an interest in estimating the true signal strengths of the non-null cases. Figure 5 plots the smoothed approximate posterior densities and approximate 95% credible intervals about the signals, μ, for two typical non-null cases in the simulated gene pathway data. Again, the posteriors were evaluated with and without the isolated cases using neighborhoods defined by pathway membership. The true signal strengths for these two cases were $E(Z) \approx -2.78$ (top panel) and $E(Z) \approx 2.12$ (bottom panel), indicated in the plots by vertical lines. The Figure illustrates the reduction in posterior uncertainty that can be obtained by including all of the data points. For the top panel, the approximate 95% credible intervals with and without the isolated cases are $[-3.36, -2.40]$ and $[-3.77, -2.12]$, respectively. For the bottom panel, the intervals are $[1.70, 2.66]$ and $[1.24, 3.07]$ with and without the isolated cases, respectively. The average widths of the approximate 95% credible intervals over all of the cases in (non-null) Pathway 2 with and without isolated observations are 0.960 (0.014) and 1.71 (0.060), respectively. Likewise, the average widths over cases in Pathway 5 with and without isolated observations are 0.934 (0.011) and 1.70 (0.043), respectively. By including the isolated points, we attain an approx-

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0382

imate four-fold increase in the precisions of the signal estimates.

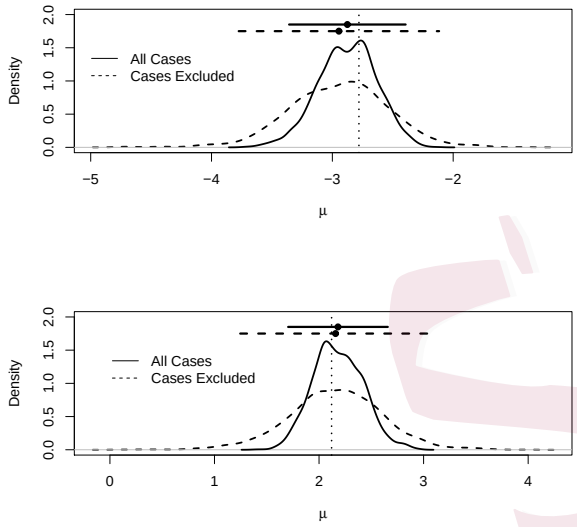

Figure 5: Smoothed approximate posterior densities of the signal strengths μ_j for two non-null cases in the simulated gene pathway example. The bars at the top are the approximate 95% posterior credible intervals. The vertical lines indicate the true means of the non-null distributions of the test statistics.

In practice, the most appropriate neighborhood structure to use in the CAR model may not be known. For gene microarray, only biologically meaningful dependence structures would usually be considered so that one would not be faced with completely unguided choices. Even among interpretable neighborhoods, we still might wish to compare competing neighborhood assumptions. Consider again the simulated pathway data, only we do not know whether the dependence is among biologically-determined pathways, or if it is a function of physical adjacency. In this case, we can gauge the spatial dependence by considering Moran's I statistic under different neighborhood assumptions, whence the competing models can be examined by looking at the strength of estimated spatial association according to each. For the simulated pathway data, we have $I(\mathbf{y}) = 0.0041$ for the (incorrect) adjacency assumption, and $I(\mathbf{y}) = 2.0747$ for the (correct) pathway assumption. The lack of association under the adjacency structure is indicative of the invalid assumption, since we would expect there to be some kind of association (otherwise there's no need for a CAR model at all). Further, we can investigate competing models' predictive capabilities through the use of root mean

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0382

square predictive error (RMSPE) and the Wantanabe-Akaike information criterion (WAIC; Wantanabe (2010)), the latter of which is asymptotically equivalent to Bayesian leave-one-out cross validation but averages over the posterior distribution instead of relying on point estimates, unlike AIC or DIC (Gelman, Carlin, Stern, Dunson, Vehtari, and Rubin (2014)). The RMSPEs for the adjacency assumption and pathway assumption are 1.58 and 1.50, respectively. Likewise, the WAICs for the adjacency and pathway assumption are 3016.50 and 2794.61, respectively. The correct neighborhood structure is favored according to both criteria.

The true correlation structure among gene expression data is complex. The implementation of our proposed model (and similar CAR models), however, requires neighborhood structures to be specified *a priori*, and this specification can potentially be incomplete or simplistic. To study the performance of our proposed model under partially incorrect neighborhood assumptions, we simulated 1,000 expression levels over ten subjects (five treatment, five control) with genes 11-20, 111-130, 211-230, 311-330, 411-430 defining five different gene sets as before. We again supposed that the second and fifth sets were enriched in the treatment group. In contrast to the previous simulation, we assumed for each treatment subject that the correlation among signals was induced according to a directed graph, as might occur in metabolic pathways in which a signal originates from a single gene and cascades via several networks to other genes downstream (Stingo, Chen, Tadesse, and Vannucci (2011)). A parent gene was selected at random from each of the two active gene sets, and the signals for each were simulated as $\mu_p \sim N(2.5, 0.75^2)$. Within each pathway, seven additional child nodes were selected at random and their signals were taken to be $\mu_{c_1}, \ldots, \mu_{c_7} \overset{\text{iid}}{\sim} N(0.92\mu_p, 0.75^2)$. The remaining signals were drawn independently from $N(0.92 \sum_{s=1}^{7} \mu_{c_s}, 0.75^2)$. The observed expression levels were taken to be $X_{ik} \overset{\text{indep.}}{\sim} N(\mu_k, 1),\ i = 1, \ldots, 10$, where k indexes over all genes, including the parent nodes and child nodes, and i indexes the subjects. It was taken that the assumed neighborhood structure omits two genes that are members of one of the active sets, and likewise for one of the inactive sets. Lastly, we randomly selected 30 isolated

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0382

genes and drew their expression levels from $N(\mathbf{0}, \mathbf{\Sigma})$ for all subjects, treatment and control, where $\mathbf{\Sigma} = \{0.9^{|i-j|}\}_{i,j=1}^{30}$. Our proposed model thus incorrectly assumes uniform correlation within each pathway, uses incomplete pathway definitions, and ignores correlation among a subset of genes with no known pathway membership.

Table 3 displays the empirical false nondiscovery proportions, false discovery proportions, and overall misclassification proportions for these simulated data under the SB model and our proposed pathway-determined CAR model with 0.95 posterior probability threshold, as well as SAM with three common thresholds. Despite the model misspecification, our model performs comparably to SB and the generally applicable SAM procedure. In fact, our approach still yields the smallest empirical misclassification proportion, though they are all close. Even though we have partially incorrect assumptions about the correlation structure, our approach is still superior to that of assuming independence in terms of predictive capability, as evident in the lower RMSPE and WAIC values, also displayed in Table 3. Further, the assumed neighborhood structure in our model predicts non-zero spatial correlations that are fairly consistent with those observed in the actual data under the same structure. This is evident in Supplementary Figure 10, which plots a histogram of realized $I(\mathbf{y})$ values from the posterior predictive distribution along with the observed value. Under the assumed neighborhood structure, $P(I(\mathbf{y}^*) \geq I(\mathbf{y}) \mid \mathbf{y}) \approx 0.1175$.

	SB	CAR	SAM(0.05)	SAM(0.10)	SAM(0.15)
FNP	0.012	0.009	0.009	0.009	0.009
FDP	0.000	0.031	0.061	0.088	0.114
MCP	0.012	0.010	0.011	0.012	0.013
RMSPE	1.326	0.759	–	–	–
WAIC	3470.804	2778.664	–	–	–

Table 3: Error rates and predictive capabilities of the independence Bayesian (SB) and the CAR models under incorrect correlation assumptions, along with error rates from the SAM procedure with three common thresholds. SAM is not used for prediction, so the RMSPE and WAIC are not applicable.

These simulation results indicate that overall performance of the mixture prior can be improved by using common information across local neighborhoods, when it is available. This improvement is due to the induced penalty on the inclusion probabilities of statistics sur-

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0382

rounded by uninteresting observations, and to improved estimation of the mixing proportion and variance components in the data. By choosing the neighborhood weights appropriately, we demonstrate how our model can accommodate isolated genes that have no neighbors. Our proposed approach is evidently superior to a conventional CAR model, even when the correct pathway information is used to define the neighborhoods but isolated points are discarded. Incorporating spatial dependence and isolated cases results in much sharper Bayesian learning in the posterior distribution, thereby reducing uncertainty. Simple diagnostics such as Moran's I under different assumed neighborhood structures can be helpful when competing neighborhoods are available, as well as considering predictive capabilities through measures such as WAIC. We demonstrate that even simplistic correlation assumptions still perform competitively with alternatives such as the SAM procedure while predicting dependence features that are consistent with the observed data. Judging from model fit criteria, partially incorrect correlation assumptions are better than ignoring them altogether.

4. Applications

4.1 E. Coli Data

For application of our model, we considered the microarray expression data from Xiao, Reilly, and Khodursky (2009). This dataset contains transcriptional activity on the *Escherichia coli* chromosome measured as log ratios of transcript abundances between a control and various chemical, physiological, and genetic perturbations comprising 53 experimental conditions. The observed gene expression levels are the average log ratios across conditions. We have 4 replicate measurements on 4,276 genes. Test statistics are calculated by dividing the mean difference by the standard error plus a small constant to reduce the effect of extreme observations, as done with SAM. The statistics are probit transformed to yield equivalent z statistics.

The *E. coli* chromosome has been shown to have correlated expression levels according to gene location, but with a circular structure so that the first and last genes on the chromosome

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0382

are considered neighbors (Xiao, Reilly, and Khodursky (2009)). This structure led us to use the adjacency matrix obtained by replacing the last elements of the first row and first column of $\mathbf{W}_1$ in Section 3 by 1. We took $d = 0$ since each point has two neighbors.

We simulated the posterior distributions of the SB and CAR models using the MCMC algorithms described in Sections 2 and 3. The resulting posterior activation probabilities were thresholded at 0.99 to select genes as being differentially expressed. To evaluate performance, we compared the selected genes to a list of 41 genes identified in Macnab (1992) as having a known or suspected function in the *E. coli* chromosome. This list serves as a reference with which we calculated approximate false discovery and false non-discovery proportions.

To illustrate the effect of the shape parameter in the prior for p in our model, we simulated the posterior distribution while varying α. For the SB model, we again took $p \sim \text{Unif}(0, 1)$. Table 4 gives the empirical error rates. For lower values of α, the sensitivity results in generally higher error rates compared to SB. However, for higher α, we attain uniformly better performance, with all three error rates outperforming the independence model. The effect of α on the marginal posterior distributions of p and ρ can be seen in Supplementary Figure 11. Higher α values result in higher estimated values of p, as expected, but they also result in sharper posterior inferences about both p and ρ. The false discovery proportions are all quite high. As this analysis is based on a dataset, there is no way of knowing which of these are true false discoveries. The high FDP could be due to Macnab (1992) listing the *most* interesting genes, not necessarily *all* interesting genes.

	SB	$\alpha = 1$	$\alpha = 500$	$\alpha = 1000$	$\alpha = 1775$
FNP	0.002	0.001	0.001	0.001	0.001
FDP	0.407	0.782	0.532	0.422	0.379
MCP	0.007	0.033	0.011	0.007	0.006

Table 4: Error rates for the E. Coli data under the independence (SB) model and the CAR model with selected values of α in $\pi_p(p)$. For the CAR model, $d = 0$. The SB model uses a uniform prior on p.

4.2 Male vs. Female Lymphoblastoid Cell Data

For an application with a different neighborhood structure, we considered mRNA ex-

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0382

pression profile data collected from lymphoblastoid cell lines derived from fifteen males and seventeen females. This dataset was analyzed in Subramanian, Tamayo, Mootha, Mukherjee, Ebert, Gillette, Paulovich, Pomeroy, Golub, Lander, and Mesirov (2005) with gene set enrichment analysis (GSEA), who sought to identify gene sets enriched in males (male > female) and enriched in females (female > male). Each cell line contains measurements on 22,283 genes. The existing catalog for this dataset includes 319 cytogenetic gene sets, 24 for each of the 24 human chromosomes, and 295 associated with cytogenetic bands. For our analysis, we calculated two-sample t-statistics and normalized.

We considered two variants of our proposed CAR testing approach, both of which use the 319 gene sets to define the neighborhoods, but with one model including the isolated points and the other excluding isolated points, allowing for a conventional CAR model. With the isolated cases included, we set $d = 1$ in (2.4); $d = 0$ when they were excluded. We took $\alpha = 1$ in the prior on p. The MCMC algorithm, identical under both models, used a burn in of 25,000 iterations, followed by 10,000 sampling iterations, of which every fifth draw was retained. The posterior inclusion probabilities were calculated using (2.5) and thresholded at 0.99 to identify differentially expressed genes.

The estimated posterior inclusion probabilities under both models were quite similar, though not exactly the same. This can be seen in Supplementary Figure 12, which plots the posterior inclusion probabilities versus the test statistics for both models. The differences result in a couple of disagreements on the selection of differentially expressed genes, listed in Table 5.

The disagreements between the two approaches can partly be explained by the reliability of the estimates themselves. Table 5 lists the effective sample sizes (ESS) of the MCMC draws of the signals for each identified gene, $\mu_j^{(1)}, \ldots, \mu_j^{(2000)}$ (Kass, Carlin, Gelman, and Neal (1998)), which approximate of the number of *independent* pieces of information about a parameter produced by an MCMC algorithm. Lower numbers reflect higher autocorrelation in the chain and hence slower convergence. The differences between the two approaches were

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0382

	ESS		Pathway					
Gene	gCAR	CAR	chrX	chrXq13	chrXp22	chrY	chrYp11	chrYq11
201028_s_at	1855.221	1164.651	•		•	•	•	
201909_at	2000.000	1012.763				×•	×•	
204409_s_at	2000.000	557.216				×•		×•
204410_at	2106.911	1221.072				×•		×•
205000_at	2000.000	694.079				×•		×•
205001_s_at	2000.000	2000.000				×•		×•
206624_at	2000.000	1561.820				×•		×•
206700_s_at	2148.696	615.271				×•		×•
214131_at	1594.909	1873.845				•		•
214218_s_at	2000.000	85.484	×•	×•				
214983_at	1778.600	2000.000				•		
221728_x_at	1936.754	131.264	×•	×•				
203974_at	2000.000				× (Isolated case)			

Table 5: Genes selected by the generalized CAR model including isolated cases (gCAR) or conventional CAR model with isolated cases excluded (CAR). The × indicates selection under the gCAR with isolated points; the • indicates selection when ignoring isolated points. Also listed are the effective sample sizes (ESS) of $\mu_j^{(1)}, \ldots, \mu_j^{(2000)}$.

substantial. With a couple of exceptions, the retained values from the Markov chain under the conventional CAR were more highly correlated than in the gCAR, thus reducing the amount of available information about these parameters. On the other hand, including the isolated cases in this instance generally results in the Markov chain samples being almost as good as an i.i.d. sample from the posterior.

Table 5 also indicates the pathway membership for each case. There are six gene sets in which the identified individual genes appear, and the clustering of genes along the X and Y chromosome is apparent. If one were to use the approach of simply declaring as enriched the pathways including the individual differentially expressed genes, our results would agree closely with those found in the GSEA. In particular, the GSEA identified the Y chromosome (`chrY`) and two Y bands with at least 15 genes (`chrYp11, chrYq11`) as being associated with higher expression levels in males. Two of the genes selected by both models appear in the `chrX` and `chrXq13` gene sets. These genes are associated with the set of X chromosome inactivation genes, which is expected to be enriched in females. Note that the gCAR identifies an isolated gene (`203974_at`) that does not appear in any of the predefined gene sets. This gene would have been missed entirely if we used a conventional CAR structure, since it would have been discarded from the analysis. We notice also a particular gene selected only by the conventional CAR that appears in both the X and Y chromosome. This curious case could

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0382

be a false positive, though, as the slower MCMC convergence under the conventional CAR model makes the posterior inference less reliable than its gCAR counterpart.

The applications presented here illustrate two different approaches to defining neighborhoods across which information may be shared when searching for non-null cases. While the results are sensitive to the choice of hyperparameters as well as the threshold, it is apparent that "good" choices can lead to desirable operating characteristics. Applying our proposed approach to these data yields results consistent with past analyses. These results demonstrate our model's ability to harness shared information between cases without sacrificing the possibility of identifying independent cases, something that would not be possible under the conventional CAR assumption. In this analysis, we found that including the isolated cases substantially reduced the autocorrelation, resulting in quicker convergence and much greater sampling efficiency than is obtained from the conventional CAR. This is an important consideration when performing MCMC in a large-scale setting, where the computational burden limits the feasible number of iterations. Our results suggest that there could be a positive effect on the sampling efficiency of an MCMC algorithm by including isolated, independent cases in our generalized CAR structure. This is possibly an interesting topic for future research that we do not pursue here.

5. Discussion

We present a unified approach to correlated Bayesian testing whereby isolated cases and neighboring cases can be analyzed simultaneously. This allows for improved estimation of the signal strengths, the possibility of identifying isolated cases, and sharper posterior inferences. We suggest simple diagnostics that can aid a researcher in determining the most appropriate neighborhood structure when choosing among plausible models. When little prior information is available concerning correlation structure, we note that there exist in the literature proposed techniques for discovering structure such as sparse factor modeling (West, 2003) and independent components analysis (Comon, 1994). We demonstrate the

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0382

robustness of our approach to model misspecification by applying it to simulated data with a complex correlation structure in which the assumptions are partially incorrect. It performs competitively with well-established procedures.

The results presented here are seen to be sensitive to the choice of the shape parameter in the prior for the mixing proportion, p, in our proposed model. This is in part because large α values result in both prior and posterior concentration of p about large values. Large values of p mean that the Gibbs sampler tends to visit sparser models more often, and thus parameters that only appear in non-null cases are updated less frequently. Certain applications necessitate the use of a prior that enforces known sparsity (e.g., Carvalho, Chang, Lucas, Nevins, Wang, and West (2008)). While the choice of shape hyperparameter does have a considerable effect on subsequent inferences, we demonstrate how finding a good value leads to desirable operating characteristics. In working with our model, we found that an acceptable value of the shape parameter seems to depend on the strength of the correlation across neighboring observations. The best way to choose this value or otherwise tune the prior to approximately match the true *a priori* non-null probability in the data is still an open problem worthy of further investigation.

A related point is the thresholding of the location-specific *a posteriori* inclusion probabilities. We use in this paper an informal 0.95 decision rule for selecting non-null cases. Decision rules in the Bayesian testing paradigm have been proposed through average risk optimization and use of the so-called Bayes false discovery rate (bFDR) (e.g., Efron and Tibshirani (2002); Tadesse, Ibrahim, Vannucci, and Gentleman (2005); Bogdan, Ghosh, and Tokdar (2008)). Most results concerning the relationship between bFDR and frequentist error measures are based on assumed independence in the data, which we are not considering. Performance also is determined in part by specification of the prior probabilities of the hypotheses. The approach of Scott and Berger (2006) on which our model is based enjoys the virtue of inducing an automatic multiplicity adjustment, even in the correlated case (Brown, Lazar, Datta, Jang, and McDowell (2014)). While expression (2.5) allows our calculations

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0382

to viewed as a fully Bayesian treatment of local false discovery rates (Efron (2010, Ch. 5)), much work remains to be done on establishing optimal decision rules under dependence.

This paper provides a glimpse at the possibility of facilitating more reliable inference by capturing (or at least approximating) the true dependence structures that are inherent in modern high-dimensional data. While this issue has garnered more interest in the recent literature (e.g., Smith and Fahrmeir (2007); Li and Zhang (2010); Stingo, Chen, Tadesse, and Vannucci (2011); Lee, Jones, Caffo, and Bassett (2014); Zhang, Guindani, Versace, and Vannucci (2014); Zhao, Kang, and Yu (2014)) relatively limited work has been done on modeling nontrivial dependence in Bayesian models for signal detection, particularly in the high-dimensional setting where classical multiple testing approaches are no longer appropriate. We propose using a Markov random field in the continuous component of the spike-and-slab mixture. An avenue of future work could be the exploration of other dependence structures. Work has been done on modeling complex measures of distance and covariance structures (e.g., Dryden, Koloydenko, and Zhou (2009)), but there is a need for much further research toward building a class of multiple testing models capable of dealing with a wide variety of dependence types.

Supplementary Materials

The online Supplementary Material for this paper contains the full conditional distributions derived from Model (2.4), additional facts including details of simplifications used in the proof of posterior propriety in the Appendix, and Supplementary Figures discussed in Sections 3 and 4.

Acknowledgements

The authors thank Sayan Mukherjee for providing the lymphoblastoid cell data. We are grateful to the Editor, an associate editor, and two anonymous referees for their constructive feedback.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0382

Appendix. Proof of Posterior Propriety of the Proposed Model

Let $\boldsymbol{\Theta} = (\boldsymbol{\mu}^T, \sigma^2, \tau^2, \rho)^T$ and let $f(\mathbf{y}, \boldsymbol{\Theta}, \boldsymbol{\gamma}, p)$ be the joint density of the data and the parameters. For any $\boldsymbol{\gamma} \in \{0,1\}^J$,

$$\begin{aligned} f(\mathbf{y}, \boldsymbol{\Theta}, \boldsymbol{\gamma}, p) &= (2\pi\sigma^2)^{-J/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{J}(y_j - \gamma_j\mu_j)^2\right) \\ &\quad \times (2\pi)^{-J/2}|\tau^2(\mathbf{D}_w^* - \rho\mathbf{W})^{-1}|^{-1/2} \exp\left(-\frac{1}{2\tau^2}\boldsymbol{\mu}^T(\mathbf{D}_w^* - \rho\mathbf{W})\boldsymbol{\mu}\right) \\ &\quad \times (\sigma^2+\tau^2)^{-2}\pi_\rho(\rho)\alpha p^{J-\sum_j\gamma_j+\alpha-1}(1-p)^{\sum_j\gamma_j} \\ &\equiv f(\mathbf{y}, \boldsymbol{\Theta} \mid \boldsymbol{\gamma})\pi_{(\boldsymbol{\gamma},p)}(\boldsymbol{\gamma}, p), \end{aligned}$$

where $\pi_{(\boldsymbol{\gamma},p)}(\boldsymbol{\gamma}, p) = \alpha p^{J-\sum_j\gamma_j+\alpha-1}(1-p)^{\sum_j\gamma_j}$. Hence,

$$\int_p\int_{\boldsymbol{\Theta}} f(\mathbf{y}, \boldsymbol{\Theta}, \boldsymbol{\gamma}, p)d\boldsymbol{\Theta}dp \;\propto\; \int_{\boldsymbol{\Theta}} f(\mathbf{y}, \boldsymbol{\Theta} \mid \boldsymbol{\gamma})d\boldsymbol{\Theta},$$

since $\int_0^1 p^{J-\sum_j\gamma_j+\alpha-1}(1-p)^{\sum_j\gamma_j}dp < \infty$. So, it suffices to establish that

$$\int_\rho\int_{\tau^2}\int_{\sigma^2}\int_{\boldsymbol{\mu}} f(\mathbf{y}, \boldsymbol{\mu}, \sigma^2, \tau^2, \rho \mid \boldsymbol{\gamma})d\boldsymbol{\mu}d\sigma^2d\tau^2d\rho < \infty, \quad \forall\boldsymbol{\gamma} \in \{0,1\}^J.$$

Write $f_{\mathbf{1}}(\mathbf{y} \mid \boldsymbol{\mu}) := f(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\gamma} = \mathbf{1})$ to simplify notation. We have that

$$f_{\mathbf{1}}(\mathbf{y} \mid \boldsymbol{\mu}) = \prod_{j=1}^{J} f(y_j \mid \mu_j, \gamma_j = 1) = \prod_{j=1}^{J} N(y_j \mid \mu_j, \sigma^2).$$

Since $(\mathbf{D}_w^* - \rho\mathbf{W})$ is positive definite, $\boldsymbol{\mu} \sim N_J(\mathbf{0},\ \tau^2(\mathbf{D}_w^* - \rho\mathbf{W})^{-1})$. Thus, the marginal density of $\mathbf{y}$ is that of a $N_J(\mathbf{0},\ \sigma^2\mathbf{I} + \tau^2(\mathbf{D}_w^* - \rho\mathbf{W})^{-1})$ distribution. Integration over $\boldsymbol{\mu}$ yields

$$f_{\mathbf{1}}(\mathbf{y}, \sigma^2, \tau^2, \rho) = \pi_{(\tau^2,\sigma^2)}(\tau^2, \sigma^2)\pi_\rho(\rho)N_J(\mathbf{y} \mid \mathbf{0}, \sigma^2\mathbf{I} + \tau^2(\mathbf{D}_2^* - \rho\mathbf{W})^{-1}).$$

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0382

Take $\eta := \tau^2/\sigma^2$ and integrate over σ^2 using the inverse gamma integral to obtain

$$\begin{aligned} f_{\mathbf{1}}(\mathbf{y},\eta,\rho) &\propto \pi_\rho(\rho)(1+\eta)^{-2}|\mathbf{I}+\eta(\mathbf{D}_w^*-\rho\mathbf{W})^{-1}|^{-1/2} \\ &\quad \times \int_0^\infty (\sigma^2)^{-(J/2)-1}\exp\left(-\frac{1}{2\sigma^2}\mathbf{y}^T(\mathbf{I}+\eta(\mathbf{D}_w^*-\rho\mathbf{W})^{-1})^{-1}\mathbf{y}\right)d\sigma^2 \\ &\propto \pi_\rho(\rho)(1+\eta)^{-2}\frac{|\mathbf{I}+\eta(\mathbf{D}_w^*-\rho\mathbf{W})^{-1}|^{-1/2}}{(\mathbf{y}^T(\mathbf{I}+\eta(\mathbf{D}_w^*-\rho\mathbf{W})^{-1})^{-1}\mathbf{y})^{J/2}}. \end{aligned}$$

But $\mathbf{y}^T(\mathbf{I}+\eta(\mathbf{D}_w^*-\rho\mathbf{W})^{-1})^{-1}\mathbf{y} = \mathbf{y}^T(\mathbf{D}_w^*)^{1/2}(\mathbf{D}_w^*+\eta(\mathbf{I}-\rho\mathbf{W}^*)^{-1})^{-1}(\mathbf{D}_w^*)^{1/2}\mathbf{y} = \mathbf{x}^T(\mathbf{D}_w^*+\eta(\mathbf{I}-\rho\mathbf{W}^*)^{-1})^{-1}\mathbf{x}$, where $\mathbf{W}^* = (\mathbf{D}_w^*)^{-1/2}\mathbf{W}(\mathbf{D}_w^*)^{-1/2}$ and $\mathbf{x} = (\mathbf{D}_w^*)^{1/2}\mathbf{y}$. Let $w_{(J)} = \max_{1\le j\le J} w_j$. Then, after substantial simplification (see Supplementary Material), it can be shown that, for all $\rho \in (\nu_1^{-1}, \nu_J^{-1})$,

$$(\mathbf{x}^T((w_{(J)}+d)\mathbf{I}+\eta(\mathbf{I}-\rho\mathbf{W}^*)^{-1})^{-1}\mathbf{x})^{-J/2} \le k'(w_{(J)}+d+\eta)^{J/2}, \tag{5.1}$$

where $0 < k' < \infty$ is constant.

Similarly, after substantial simplification (see Supplementary Material), we can show that

$$|\mathbf{I}+\eta(\mathbf{D}_w^*-\rho\mathbf{W}^*)^{-1}|^{-1/2} \le \frac{k'\prod_{j=1}^J \max\{(1-\rho\nu_j)^{1/2},1\}}{(w_{(1)}+d+\eta)^{J/2}}, \tag{5.2}$$

establishing that

$$f(\mathbf{y},\eta,\rho) \le C(w_{(J)}+d+\eta)^{J/2}\left(\frac{\prod_{j=1}^J \max\{(1-\rho\nu_j)^{1/2},1\}}{(w_{(1)}+d+\eta)^{J/2}}\right)(1+\eta)^{-2}\pi_\rho(\rho) \tag{5.3}$$

where $0 < C < \infty$. Also,

$$\int_0^\infty \Big(w_{(J)}+d+\eta\Big)^{J/2}\Big(w_{(1)}+d+\eta\Big)^{-J/2}(1+\eta)^{-2}d\eta < \infty.$$

It follows that $\int_{\nu_1^{-1}}^{\nu_J^{-1}}\int_0^\infty f_{\mathbf{1}}(\mathbf{y},\eta,\rho)d\eta d\rho < \infty$, showing that $\int_{\mathbf{\Theta}} f_{\mathbf{1}}(\mathbf{y},\mathbf{\Theta})d\mathbf{\Theta} < \infty$.

Now consider the case $|\{\gamma_j : \gamma_j = 1\}| = J_1 < J$, where $|\cdot|$ denotes cardinality. Let

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0382

$S_1 = \{j : \gamma_j = 1\} \subsetneq \{1, \ldots, J\}$, $\mathbf{y}_1 = \{y_j : j \in S_1\}$, and $\mathbf{y}_0 = \{y_j : j \in S_1^c\}$ so that we may partition $\mathbf{y}$ as $\mathbf{y} = (\mathbf{y}_0^T, \mathbf{y}_1^T)^T$. Then $\int_{\sigma^2}\int_{\tau^2}\int_{\rho}\int_{\boldsymbol{\mu}}\int_{\mathbf{y}_1} f(\mathbf{y}_0, \mathbf{y}_1, \tau^2, \sigma^2, \rho, \boldsymbol{\mu} \mid \boldsymbol{\gamma}^*) d\mathbf{y}_1 d\boldsymbol{\mu} d\rho d\tau^2 d\sigma^2$ is proportional to

$$\int_0^\infty (\sigma^2)^{-\frac{J-J_1}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{j\in S_1^c} y_j^2\right) \int_0^\infty (\sigma^2+\tau^2)^{-2} d\tau^2 d\sigma^2 < \infty, \quad \forall \mathbf{y}_0 \ (a.e.).$$

The proof is completed by considering the case $\gamma_j = 0$, for all j. The preceding argument still applies, though, with $S_1 = \emptyset$ and $J_1 = 0$. The result is therefore established.

References

Banerjee, S. and Carlin, B. P., and Gelfand, A. E. (2015). *Hierarchical Modeling and Analysis for Spatial Data*. 2nd ed. Chapman & Hall/CRC, Boca Raton.

Banks, D., Datta, G., Karr, A., Lynch. J., Niemi, J., and Vera, F. (2012). Bayesian CAR models for syndromic surveillance on multiple data streams: Theory and practice. *Info. Fus.* **13**, 105-116.

Benjamini, Y. (2010). Discovering the false discovery rate. *J. Roy. Stat. Soc. B.* **72**, 405-416.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B.* **57**, 289-300.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Stat. Soc. B.* **36**, 192-236.

Besag, J., York, J. C., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion). *Ann. Inst. Stat. Math.* **43**, 1-59.

Bogdan, M., Ghosh, J. K., and Tokdar, S. T. (2008). A comparison of the Benjamini-Hochgberg procedure with some Bayesian rules for multiple testing. In *Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen,* eds. Balakrishnan, N., Peña, E. A., and Silvapulle, M. J. Institute of Mathematical Statistics, Beachwood. pp. 211-230.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0382

Brown, D. A., Lazar, N. A., Datta, G. S., Jang, W., and McDowell, J. E. (2014). Incorporating spatial dependence into Bayesian multiple testing of statistical parametric maps in functional neuroimaging. *NeuroImage*. **84**, 97-112.

Carlin, B. and Banerjee, S. (2003). Hierarchical multivariate CAR models for spatiotemporally correlated survival data. In *Bayesian Statistics 7*, eds. Bernardo, J. M, Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M. Oxford University Press, Oxford. pp.45-63.

Carlin, B. P. and Louis, T. A. (2009). *Bayesian Methods for Data Analysis*. Chapman & Hall/CRC, Boca Raton.

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q., and West, M. (2008). High-dimensional sparse factor modeling: Applications in gene expression genomics. *J. Am. Stat. Assoc.*, **103**, 1438-1456.

Clyde, M. and George, E. I. (2004). Model uncertainty. *Stat. Sci.*, **19**, 81-94.

Comon, P. (1994). Independent component analysis: A new concept? *Signal Processing*, **36**, 287-314.

Dryden, I. L., Koloydenko, A., and Zhou, D. (2009). Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *Ann. Appl. Stat.* **3**, 1102-1123.

Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *J. Am. Stat. Assoc.*, **102**, 93-103.

Efron, B. (2008). Microarrays, empirical Bayes, and the two-groups model. *Stat. Sci.*, **23**, 1-22.

Efron, B. (2010). *Large Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press, New York.

Efron, B. and Tibshirani, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Gen. Epidem.*, **23**, 70-86.

Efron, B. and Tibshirani, R. (2007). On testing the significance of sets of genes. *Ann. Appl. Stat.*, **1**, 107-129.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0382

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.*, **96**, 1151-1160.

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D., and Frackowiak, R. S. J. (1995). Statistical parametric maps in functional imaging: A general linear approach. *Hum. Br. Map.*, **2**, 189-210.

Gelfand, A. E. and Sahu, S. K. (1999). Identifiability, improper priors, and Gibbs sampling for generalized linear models. *J. Am. Stat. Assoc.*, **94**, 247-253.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Am. Stat. Assoc.*, **85**, 398-409.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayes. Anal.*, **1**, 515-533.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian Data Analysis*. 3rd ed. Chapman & Hall/CRC, Boca Raton.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Stat. Sci.*, **7**, 457-511.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Patt. Anal. Mach. Int.*, **6**, 721-741.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *J. Am. Stat. Assoc.* **88**, 881-889.

Geweke, J. (1996). Variable selection and model comparison in regression. In *Bayesian Statistics 5*, eds. Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M. Oxford University Press, Oxford. pp. 609-620.

Gilks, W. R. and Roberts, G. O. (1996). Strategies for improving MCMC. In *Markov Chain Monte Carlo in Practice*, eds. Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. Chapman & Hall/CRC, Boca Raton. pp. 89-114.

Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *J. Roy. Stat. Soc. C.* **41**, 337-348.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0382

Higdon, D. (1994). Spatial applications of Markov chain Monte Carlo for Bayesian inference. Unpublished doctoral thesis. Univ. of Washington, Dept. of Statistics.

Kass, R. E., Carlin, B. P., Gelman, A., and Neal, R. (1998). Markov chain Monte Carlo in practice: A roundtable discussion. *Am. Stat.*, **52**, 93-100.

Lee, K.-J., Jones, G. L., Caffo, B. S., and Bassett, S. S. (2014). Spatial Bayesian variable selection models on functional magnetic resonance imaging time-series data. *Bayes. Anal.*, **9**, 699-732.

Li, F. and Zhang, N. R. (2010). Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *J. Am. Stat. Assoc.*, **105**, 1202-1214.

Love, M. I., Huber. W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. BioRxiv Preprint. DOI: `http://dx.doi.org/10.1101/002832`.

Macnab, R. M. (1992). Genetics and biogenesis of bacterial flagella. *Ann. Rev. Gen.*, **26**, 131-158.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. John Wiley and Sons, New York.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *J. Am. Stat. Assoc.*, **83**, 1023-1032.

Neal, R. M. (2003). Slice sampling. *Ann. Stat.*, **31**, 705-741.

O'Hara,, R. B. and Sillanpää, M. J. (2009). A review of Bayesian variable selection methods: What, how, and which. *Bayes. Anal.*, **4**, 85-118.

Qiu, X., Klebanov, L., and Yakolev, A. (2005). Correlation between gene expression levels and limitations of the empirical Bayes methodology for finding differentially expressed genes. *Stat. Appl. Gen. Mol. Biol.*, **4**, article 34.

R Core Team. (2015). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna.

Scott, J. G. and Berger, J. O. (2006). An exploration of aspects of Bayesian multiple testing. *J. Stat. Plan. Inf.*, **136**, 2144-2162.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0382

Scott, J. G. and Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Ann. Stat.*, **38**, 2587-2619.

Smith, M. and Fahrmeir, L. (2007). Spatial Bayesian variable selection with application to functional magnetic resonance imaging. *J. Am. Stat. Assoc.*, **102**, 417-431.

Stingo, F. C., Chen, Y. A., Tadesse, M. G., and Vannucci, M. (2011). Incorporating biological information in the linear models: A Bayesian approach to the selection of pathways and genes. *Ann. Appl. Stat.*, **5**, 1978-2002.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Nat. Acad. Sci.*, **102**, 15545-15550.

Tadesse, M., Ibrahim, J. G., Vannucci, M., and Gentleman, R. (2005). Wavelet thresholding with Bayesian false discovery rate control. *Biometrics*, **61**, 25-35.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Nat. Acad. Sci.*, **98**, 5116-5121.

Wantanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learning Res.*, **11**, 3571-3594.

West, M. (2003). Bayesian factor regression models in the 'large p, small n' paradigm. In *Bayesian Statistics 7*, eds. Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M. Oxford University Press, Oxford. pp. 723-732.

Xiao, G., Reilly, C., and Khodursky, A. B. (2009). Improved detection of differentially expressed genes through incorporation of gene locations. *Biometrics*, **65**, 805-814.

Zhang, L., Guindani, M., Versace, F., and Vannucci, M. (2014). A spatio-temporal nonparametric Bayesian variable selection model of fMRI data for clustering correlated time courses. *NeuroImage*, **95**, 162-175.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0382

Zhao, Y., Kang, J., and Yu, T. (2014). A Bayesian nonparametric mixture model for selecting genes and gene subnetworks. *Ann. Appl. Stat.* **8**, 999-1021.

Department of Mathematical Sciences, Clemson University, Clemson, SC 29634-0975, U.S.A.

E-mail: ab7@clemson.edu

Department of Statistics, University of Georgia, Athens, GA 30602, U.S.A.

E-mail: gauri@uga.edu

Department of Statistics, University of Georgia, Athens, GA 30602, U.S.A.

E-mail: nlazar@uga.edu

Statistica Sinica