This article was downloaded by: [NIH Library]
On: 26 December 2013, At: 06:55
Publisher: Taylor & Francis
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK

Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uasa20

Bayesian Selection and Clustering of Polymorphisms in Functionally Related Genes

David B Dunson[a], Amy H Herring[a] & Stephanie M Engel[a]

[a] David B. Dunson is Senior Investigator, Biostatistics Branch, National Institute of Environmental Health Sciences, Research Triangle Park, NC 27709 . Amy H. Herring is Associate Professor of Biostatistics, University of North Carolina, Chapel Hill, NC 27599. Stephanie M. Engel is Assistant Professor, Department of Community and Preventive Medicine, Mount Sinai Medical School, New York, NY 10029. This research was supported in part by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences. The PIN Study was supported by a grant from the National Institute of Child Health and Human Development, National Institutes of Health (HD28684); cooperative agreements from the Association of Schools of Public Health/ Centers for Disease Control and Prevention (S455-16/17, S0807-18/20, S1099-19/21); and funds from the Wake Area Health Education Center in Raleigh, NC; and supported in part by a grant (RR00046) from the General Clinical Research Centers Program of the Division of Research Resources, National Institutes of Health. The second author's work was supported in part by grants from the National Institutes of Health (National Institute of Child Health and Human Development 1R03HD04578001A1, National Institute of Environmental Health Sciences P30ES10126) and the Environmental Protection Agency (RD831843010).
Published online: 01 Jan 2012.

To cite this article: David B Dunson, Amy H Herring & Stephanie M Engel (2008) Bayesian Selection and Clustering of Polymorphisms in Functionally Related Genes, Journal of the American Statistical Association, 103:482, 534-546, DOI: 10.1198/016214507000000554

To link to this article: http://dx.doi.org/10.1198/016214507000000554

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the "Content") contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden. Terms & Conditions of access and use can be found at http://www.tandfonline.com/page/terms-and-conditions

Bayesian Selection and Clustering of Polymorphisms in Functionally Related Genes

David B. DUNSON, Amy H. HERRING, and Stephanie M. ENGEL

In epidemiologic studies, there is often interest in assessing the relationship between polymorphisms in functionally related genes and a health outcome. For each candidate gene, single nucleotide polymorphism (SNP) data are collected at a number of locations, resulting in a large number of possible genotypes. Because instabilities can result in analyses that include all the SNPs, dimensionality is typically reduced by conducting single SNP analyses or attempting to identify haplotypes. This article proposes an alternative Bayesian approach for reducing dimensionality. A multilevel Dirichlet process prior is used for the distribution of the SNP-specific regression coefficients within genes, incorporating a variable selection-type mixture structure to allow SNPs with no effect. This structure allows simultaneous selection of important SNPs and soft clustering of SNPs having similar impact on the health outcome. The methods are illustrated using data from a study of pro- and anti-inflammatory cytokine polymorphisms and spontaneous preterm birth.

KEY WORDS: Bayesian; Clustering; Dirichlet process; Genetic association; Hierarchical regression; Multiple testing; Nonparametric Bayes; Single nucleotide polymorphisms; Sparse regression.

Downloaded by [NIH Library] at 06:55 26 December 2013

1. INTRODUCTION

In epidemiologic research, there is commonly interest in the association between multiple, closely related *exposures* and a health outcome. Some examples include drinking water disinfection byproducts, agricultural chemicals, and single nucleotide polymorphisms (SNPs) in candidate genes. When the number of exposures is large (e.g., 30+) and the exposures are correlated (e.g., due to linkage disequilibrium between polymorphisms), it is well known that maximum likelihood estimation can result in unstable estimates and inferences. For this reason, analysts typically apply dimensionality reduction techniques, with the most common being (1) consider exposures one at a time in univariate analyses; (2) collapse exposures into class-specific summaries; and (3) run a model selection procedure, such as stepwise selection, to obtain a parsimonious model on which to base final inferences. There are clear problems with each of these approaches: (1) can produce misleading results by not adjusting for correlated exposures, (2) can discard valuable information on variability in the effect within a class, and (3) can result in overestimation of the regression coefficients due to selection bias.

For these reasons, many authors have proposed hierarchical regression procedures, which shrink the exposure-specific regression coefficients toward a common distribution, using empirical Bayes (Thomas et al. 1985), "semi-Bayes" (Greenland 1992, 1993), or fully Bayes approaches. Greenland (1993) provided a review and demonstrated improved performance of the empirical and semi-Bayes approaches relative to methods based on maximum likelihood estimation (MLE). Such hierarchical regression procedures have been used in numerous articles in the epidemiologic literature. For example, De Roos, Poole, Teschke, and Olshan (2001) considered applications to multiple paternal occupational exposures and neuroblastoma in the offspring, and Hung et al. (2004) considered applications to genetic association studies with multiple markers. For related methods for multiple outcomes, refer to Meng and Dempster (1987) and Coull, Hobert, Ryan, and Holmes (2001).

These methods are based on shrinking the exposure-specific regression coefficients toward a normal prior distribution, potentially with unknown mean and variance. Although this shrinkage certainly improves the stability of estimates, many epidemiologists would prefer to avoid the assumption that the regression coefficients for the different exposures follow a normal distribution. In addition, there is typically interest in grouping or clustering the different exposures based on their effects on the outcome. In particular, one wishes to identify exposures having similar effects, including those that are not associated with the outcome, in drawing mechanistic conclusions. Although grouping can be done subjectively based on examination of estimated regression coefficients, it would be appealing to have a formal clustering procedure.

This problem is somewhat related to subset selection in regression, which focuses on identifying predictors with nonzero coefficients from among a potentially high-dimensional set of candidates (refer to George and McCulloch 1997; Clyde and George 2004, for reviews of Bayesian approaches). However, following standard epidemiologic practice, we are at least as interested in estimating regression coefficients for the different exposures and in grouping exposures according to the magnitude of their effect, as we are in identifying exposures that are associated with the response. Hence, the problem is one of clustering the regression coefficients incorporating information on the exposure class.

The Bayesian approach provides a natural framework for clustering of the exposures in this manner. For a recent article on Bayesian variable selection and clustering in high-dimensional data, refer to Tadesse, Sha, and Vannucci (2005).

David B. Dunson is Senior Investigator, Biostatistics Branch, National Institute of Environmental Health Sciences, Research Triangle Park, NC 27709 (E-mail: *dunson1@niehs.nih.gov*). Amy H. Herring is Associate Professor of Biostatistics, University of North Carolina, Chapel Hill, NC 27599. Stephanie M. Engel is Assistant Professor, Department of Community and Preventive Medicine, Mount Sinai Medical School, New York, NY 10029. This research was supported in part by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences. The PIN Study was supported by a grant from the National Institute of Child Health and Human Development, National Institutes of Health (HD28684); cooperative agreements from the Association of Schools of Public Health/Centers for Disease Control and Prevention (S455-16/17, S0807-18/20, S1099-19/21); and funds from the Wake Area Health Education Center in Raleigh, NC; and supported in part by a grant (RR00046) from the General Clinical Research Centers Program of the Division of Research Resources, National Institutes of Health. The second author's work was supported in part by grants from the National Institutes of Health (National Institute of Child Health and Human Development 1R03HD045780-01A1, National Institute of Environmental Health Sciences P30ES10126) and the Environmental Protection Agency (RD831843010).

In the Public Domain
Journal of the American Statistical Association
June 2008, Vol. 103, No. 482, Applications and Case Studies
DOI 10.1198/016214507000000554

Downloaded by [NIH Library] at 06:55 26 December 2013

Their focus was on clustering samples of data into groups while simultaneously selecting discriminating variables. In contrast, our focus is on clustering not the data but the unknown exposure effects into groups, while allowing an unknown subset to have no association with the outcome. A related problem was considered by Gopalan and Berry (1998), who used a Dirichlet process (DP) prior (Ferguson 1973, 1974) to cluster treatment groups in a clinical trial in order to adjust for multiple comparisons. From a Bayesian perspective, the multiple-comparison problem can be considered as an issue of appropriately choosing a prior to account for dependency in multiple, related hypotheses (refer to Westfall, Johnson, and Utts 1997; Berry and Hochberg 1999; Gonen, Westfall, and Johnson 2003; Berry and Berry 2004; Dunson 2005).

Although the Gopalan–Berry (1998) approach could be directly modified to allow clustering of the regression coefficients for the different exposures, such an approach would not incorporate information on exposure class or allow identification of exposures having no effect. To perform simultaneous variable selection and clustering, both within and across exposure classes, this article proposes an alternative approach. In the one-class case, our approach relies on a mixture prior, with a point mass for exposures with no effect and a DP component to cluster the nonnull exposures. In the multiple-class case, a multilevel DP structure is chosen, allowing common clusters across exposure classes, while also introducing class-specific clusters.

Section 2 motivates the problem through application to selection and clustering of polymorphisms in functionally related genes. Section 3 describes the regression model and proposes the hierarchical clustering prior. Section 4 develops methods for posterior computation. Section 5 applies the method to data on cytokine polymorphisms and risk for spontaneous preterm birth, while also presenting results from simulation studies. Section 6 contains a discussion.

2. IDENTIFYING POLYMORPHISMS PREDICTING DISEASE

This article is motivated by the problem of selection and clustering of polymorphisms in functionally related genes. Using the nomenclature of Section 1, genes correspond to *classes* and *exposures* to single nucleotide polymorphisms (SNPs). For a given gene, SNPs can be collected within the coding region, which consists of the sequence of amino acids that codes directly for the protein product of the gene, within regulatory regions upstream of the coding region, or within intronic sequences. SNPs that occur within regulatory regions are thought to be much more likely to affect biologic function and gene expression.

For subject i, $i = 1, \ldots, n$, and gene c, $c = 1, \ldots, C$, the SNP data consist of the individual's genotype at p_c loci. Let $g_{icl} \in \{1, 2, 3\}$ denote a categorical variable indicating the genotype of individual i for locus l within gene c, with $g_{icl} = 1$ if homozygous for the more common allele, $g_{icl} = 2$ if heterozygous, and $g_{icl} = 3$ if homozygous for the minor allele. Including the genotype data for each loci within each of C functionally related genes, we obtain the vector $\mathbf{g}_i = (\mathbf{g}'_{i1}, \ldots, \mathbf{g}'_{iC})'$, with $\mathbf{g}_{ic} = (g_{ic1}, \ldots, g_{ic,p_c})'$ for $c = 1, \ldots, C$.

Scientific interest focuses on assessing the relationship between $\mathbf{g}_i$ and a health outcome, y_i, adjusting for potential confounders, $\mathbf{z}_i = (z_{i1}, \ldots, z_{iq})'$. For example, we are interested in relating SNPs in cytokine gene regulatory regions to risk for spontaneous preterm birth using data from the Pregnancy, Infection, and Nutrition (PIN) study (Savitz et al. 1999), which enrolled women between 24 and 29 weeks of gestation, collecting blood at the intake visit. As shown in Table 1, there are 12 cytokines (soluble proteins that mediate and regulate immunity and inflammation) of interest, including IL1α, IL1β, IL2–IL6, IL10, IL13, LTA, TGFβ1, and TNF. The number of sites within regulatory regions at which SNP data are collected ranges from one to three per cytokine (22 total), with three genotypes per site. The genotypes varied in frequency, with $\sum_{i=1}^{n} 1(g_{icl} = h)$, $h = 1, 2, 3$, ranging from 8 to 221, with $n = 447$ (excluding women with missing genotype data).

Results for one-SNP-at-a-time logistic regression analyses, with spontaneous preterm birth ($y_i = 1$ preterm, $y_i = 0$ full term) as the outcome, are provided in Table 1. The genotype at a single locus is included in the model through the use of two indicator values, with the $g_{icl} = 1$ category used as the reference. Results are stratified on ethnicity (white, African American), because African American women have higher rates of spontaneous preterm birth and potentially different genetic factors. An association between a marker and disease may occur if the marker is in linkage disequilibrium with a disease gene or because of population substructure (admixture). Such admixture is thought to be particularly apparent in African Americans, potentially leading to false positives (Ziv and Burchard 2003). Although all three genotypes were represented in the study for each ethnic group, some categories had no women with spontaneous preterm births, so certain genotype-specific odds ratios could not be obtained. In addition, we were unable to obtain convergence for the full model with all SNPs included simultaneously.

Focusing on a smaller group of common proinflammatory cytokines and two genotype categories per SNP, Mulherin Engel et al. (2005a) reported an association with spontaneous preterm birth based on a semi-Bayes analysis (refer also to Mulherin Engel et al. 2005b). Our interest here is in identifying the specific SNPs predictive of spontaneous preterm birth, while also clustering SNPs within and across cytokines that have a similar risk of spontaneous preterm birth. In addition, we want to address the curse-of-dimensionality problem that occurs as additional loci are added to the model by using a Bayesian approach, with a prior that favors sparse regression models. In particular, it is expected that genetic variation at most loci will be unimportant in predicting risk, but that there is one or perhaps a few important variants hidden in a sea of nulls. Sparse regression models have been proposed by several authors as a biologically motivated approach for analyzing massive genomics datasets. For example, Carvalho et al. (2005) and Lucas et al. (2006) proposed sparse modeling approaches motivated by gene expression data. Our approach is conceptually different, being based on a combination of variable selection and clustering.

Table 1. Summary of one-site-at-a-time logistic regression analyses with spontaneous preterm birth as the outcome variable

			Odds ratio for spontaneous preterm birth	
Cytokine	Site	Genotype	White	African American
IL1α	+4845	GG	1.0	1.0
		GT	1.1 (.6, 1.9)	.8 (.5, 1.5)
		TT	1.8 (.6, 5.2)	*
	IVS5-109	AA	1.0	1.0
		AC	.7 (.4, 1.3)	1.1 (.6, 1.9)
		CC	.4 (.1, 1.7)	1.1 (.4, 3.2)
IL1β	1061	CC	1.0	1.0
		TC	1.8 (1.0, 3.3)	.7 (.3, 1.6)
		TT	1.1 (.4, 3.0)	1.1 (.5, 2.5)
	+3594	CC	1.0	1.0
		CT	1.2 (.7, 2.1)	.9 (.5, 1.7)
		TT	*	*
	−581	TT	1.0	1.0
		TC	2.0 (1.1, 3.8)	.6 (.3, 1.4)
		CC	1.0 (.4, 2.8)	1.1 (.5, 2.4)
IL2	−385	TT	1.0	1.0
		TG	1.6 (.9, 2.9)	.7 (.3, 1.6)
		GG	1.1 (.4, 2.9)	*
IL4	−589	CC	1.0	1.0
		CT	.8 (.4, 1.6)	.7 (.3, 1.7)
		TT	28.4 (3.3, 241.5)	.9 (.4, 2.0)
	−1099	TT	1.0	1.0
		GT	.7 (.3, 1.9)	1.8 (1.0, 3.3)
		GG	*	.5 (.1, 3.9)
	−33	CC	1.0	1.0
		TC	.8 (.4, 1.7)	1.9 (1.0, 3.7)
		TT	14.9 (2.9, 76.5)	1.7 (.7, 4.0)
IL5	−746	TT	1.0	1.0
		TC	1.4 (.4, 5.0)	.9 (.5, 1.6)
		CC	2.5 (.7, 8.7)	.7 (.2, 3.6)
IL6	−174	GG	1.0	1.0
		CG	1.1 (.6, 2.1)	1.1 (.5, 2.6)
		CC	1.2 (.5, 2.7)	*

Downloaded by [NIH Library] at 06:55 26 December 2013

3. HIERARCHICAL CLUSTERING OF GENETIC POLYMORPHISMS

3.1 Model and Background

For simplicity, we will focus on the binary response regression model:

$$\Pr(y_i = 1|\mathbf{g}_i, \mathbf{z}_i) = \mathcal{L}\left(\mathbf{z}_i'\boldsymbol{\kappa} + \sum_{c=1}^{C}\sum_{l=1}^{p_c}\sum_{h=1}^{3} 1(g_{icl} = h)\beta_{clh}\right), \tag{1}$$

where $\mathcal{L}: \Re \to (0, 1)$ is a monotone link function, $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_q)'$ are unknown coefficients including the intercept and slopes for the confounders, and β_{clh} is a coefficient for genotype h within locus l of gene c. As a convention to ensure frequentist identifiability, one could let $\beta_{cl1} = 0$ for all c, l, so that individuals who are homozygous for the more common allele are placed in a baseline category. Even with this constraint, there are $2\sum_{c=1}^{C} p_c$ coefficients for the genotype effects. Hence, when the number of loci is moderate to large or $\sum_{c,l,h} 1(g_{icl} = h)$ is small for some c, l, h (as is typically the case), maximum likelihood estimation of model (1) runs into difficulties. In particular, the estimated regression coefficients, $\widehat{\boldsymbol{\beta}}$, can be unstable, taking values known to be unreasonable a priori, and a unique MLE may not exist.

A natural solution to this problem is to borrow information across SNPs for functionally related genes, while favoring a sparser form of regression model (1) having many of the β_{clh} coefficients equal to 0 and only a few unique values of the remaining coefficients. Allowing β_{clh} coefficients to equal 0 corresponds to our prior expectation that most of the genetic variants will have essentially no impact on disease risk. Borrowing of information across functionally related genes is also well motivated biologically. Consider the cytokine application and suppose that one finds that an SNP in IL1α increases the log odds of preterm birth by .5. One would then be more optimistic about finding a similar magnitude of effect for genetic variants in related cytokines.

Cytokines with a similar mechanism and an SNP that causes the same magnitude of change in gene expression might have a very similar log odds ratio. In the absence of additional data that would allow reliable grouping of cytokines a priori, we treat them as exchangeable, while allowing flexible clustering. Even if one does not believe exact clustering of the coefficients is bio-

Table 1. (Continued.)

Cytokine	Site	Genotype	Odds ratio for spontaneous preterm birth	
			White	African American
IL10	−854	CC	1.0	1.0
		TC	.8 (.5, 1.5)	.9 (.5, 1.7)
		TT	1.0 (.4, 2.9)	.5 (.2, 1.1)
	−627	CC	1.0	1.0
			.8 (.5, 1.5)	1.0 (.5, 1.8)
			.7 (.2, 2.2)	.5 (.2, 1.3)
	−1082	AA	1.0	1.0
		AG	1.5 (.7, 3.1)	1.1 (.6, 2.0)
		GG	1.5 (.7, 3.5)	1.2 (.5, 3.1)
IL13	+2034	GG	1.0	1.0
		AG	1.3 (.7, 2.3)	.8 (.5, 1.5)
		AA	5.4 (1.8, 16.3)	.4 (.0, 3.0)
	−1112	CC	1.0	1.0
		TC	1.6 (.9, 2.9)	2.2 (1.2, 4.2)
		TT	2.6 (.9, 8.1)	1.1 (.4, 3.0)
	IVS3-24	CC	1.0	1.0
		TC	1.3 (.7, 2.3)	1.1 (.4, 2.6)
		TT	4.7 (1.6, 13.9)	.9 (.4, 2.2)
LTA	IVS1+90	AA	1.0	1.0
		AG	1.3 (.7, 2.3)	1.4 (.7, 2.8)
		GG	1.8 (.7, 4.4)	.9 (.4, 2.1)
	IVS1-82	GG	1.0	1.0
		CG	1.7 (.9, 3.2)	1.3 (.7, 2.3)
		CC	1.3 (.5, 3.2)	.9 (.2, 3.5)
TGFβ1	L10P	TT	1.0	1.0
		TC	.8 (.4, 1.4)	1.1 (.6, 2.0)
		CC	.5 (.2, 1.3)	1.2 (.5, 2.7)
	−1347	CC	1.0	1.0
		CT	.8 (.4, 1.5)	1.2 (.7, 2.1)
		TT	.5 (.1, 1.7)	1.4 (.3, 5.5)
TNF	−308	GG	1.0	1.0
		GA	1.6 (.9, 2.9)	1.0 (.5, 2.0)
		AA	3.3 (.9, 11.9)	.5 (.1, 3.9)

*No women with spontaneous preterm births in this category.

Downloaded by [NIH Library] at 06:55 26 December 2013

logically realistic, the approach is justified as adaptively grouping the SNPs into categories based on the magnitude of their effects to stabilize estimation in high dimensions. In addition, the clustering is *soft*, so the estimated SNP-specific coefficients will differ even for SNPs in the same cluster with greater than 50% posterior probability.

Removing the $\beta_{cl1} = 0$ constraint, we propose the following nonparametric prior for the SNP-specific coefficients:

$$\beta_{clh} \sim G_c \quad \text{for } h = 1, 2, 3,\ l = 1, \ldots, p_c,\ c = 1, \ldots, C. \qquad (2)$$

Here, G_c is the unknown distribution of the regression coefficients for SNPs in gene c, and $\mathcal{G} = \{G_c, c = 1, \ldots, C\}$ is the collection of unknown distributions for the SNPs in the different genes. Within a gene, dependence in the regression coefficients is accommodated by assuming that the SNP-specific regression coefficients are drawn from a common distribution. Between genes, dependence is accommodated by assuming that the different distributions in the collection $\mathcal{G}$ have similar features.

3.2 Simultaneous Variable Selection and Clustering

We first consider the case in which all the SNPs under study relate to a single gene, so that $C = 1$. In this case, repressing the c subscript, we let $\beta_{lh} \overset{\text{iid}}{\sim} G$ for $l = 1, \ldots, p$ and $h = 1, 2, 3$. Then, to allow for uncertainty in G, while clustering the SNPs having identical regression coefficients, we choose the following mixture prior:

$$G = \pi_0 \delta_0 + (1 - \pi_0) G^*, \qquad G^* \sim \mathrm{DP}(\alpha_0 G_0), \qquad (3)$$

where $0 \le \pi_0 \le 1$ is a point mass probability, δ_θ denotes the Dirac measure concentrated at θ, and G^* is a random probability measure assigned a DP prior, with precision α and base probability measure G_0. Prior (3) falls within the framework described in MacEachern, Kottas, and Gelfand (2001).

Note that (3) is related to mixture priors used routinely in variable selection applications (George and McCulloch 1997; Clyde and George 2004), with the important distinction that G^* is replaced with a unimodal parametric density in most of the previous literature in this area. In particular, the most common choice corresponds to a normal density centered on 0, although heavier tailed distributions have also been considered.

Without restrictions, model (1) is overparameterized, because we include both an intercept, κ_1, and coefficients for

each of the genotype categories. This overparameterization allows the genotypes to be treated as exchangeable, avoiding the need to specify a reference genotype at each locus in advance. Instead, we adaptively choose the baseline category as corresponding to those genotypes having $\beta_{lh} = 0$. If $\beta_{lh} = 0$ for $h = 1, 2, 3$, then polymorphisms at locus l are not predictive of disease risk, and the locus effectively drops out of model (1). To avoid ambiguity, the baseline category is defined to correspond to the low-risk group by restricting G^* to have positive support on $\Re^+$.

Because G^* is almost surely discrete under the DP prior, nonzero coefficients will also be clustered together with positive probability. This allows cases, such as $\beta_{l1} = 0, \beta_{l2} = \beta_{l3} = .5$, which implies that having one or two copies of the minor allele at locus l conveys the same increase in the risk of disease. Such clustering can also occur across locations, for example, $\beta_{lh} = \beta_{l'h'}$.

To obtain more insight into the variable selection and clustering process, we consider some theoretical properties implied by the DP prior for G^*. First, following Sethuraman's (1994) stick-breaking representation:

$$G = \pi_0\delta_0 + (1-\pi_0)\sum_{h=1}^{\infty}\pi_h\delta_{\theta_h},$$
$$\frac{\pi_h}{\prod_{l=1}^{h-1}(1-\pi_l)} \overset{\text{iid}}{\sim} \text{beta}(1,\alpha_0), \qquad \theta_h \overset{\text{iid}}{\sim} G_0, \tag{4}$$

where $\{\pi_h, h = 1, \ldots, \infty\}$ is an infinite sequence of random weights, with $\pi_h = V_h\prod_{l=1}^{h-1}(1-V_l)$ and $V_h \overset{\text{iid}}{\sim} \text{beta}(1,\alpha_0)$ for $h = 1, \ldots, \infty$, and $\boldsymbol{\Theta} = \{\theta_h, h = 1, \ldots, \infty\}$ is an infinite sequence of random atoms, with $\theta_h \overset{\text{iid}}{\sim} G_0$. Letting $\theta_0 = 0$, $\widetilde{\pi}_0 = \pi_0$, and $\widetilde{\pi}_h = (1-\pi_0)\pi_h$, $h = 1, 2, \ldots, \infty$, we have $G = \sum_{h=0}^{\infty}\widetilde{\pi}_h\delta_{\theta_h}$. Hence, G is an infinite mixture of atoms at random locations, with one fixed atom at 0. The other locations are generated from G_0, which we assume corresponds to $\text{N}_+(\mu_0, \sigma_0^2)$, which denotes a $\text{N}(\mu_0, \sigma_0^2)$ density truncated below by 0.

To study the clustering behavior, it is of interest to extend the DP results of Antoniak (1974) to prior (3), which mixes in a point mass at 0. Let $p_+ = 3p = p_{0+} + p_{1+}$ denote the total number of genotype categories across the p loci, with $p_{0+} = \sum_{l=1}^{p}\sum_{h=1}^{3} 1(\beta_{lh} = 0)$ denoting the number of SNPs having zero regression coefficients. Then, for $\beta_{lh} \sim G$, it follows directly from (3) that

$$\Pr(p_{0+} = h|\pi_0, \alpha_0, p_+) = \binom{p_+}{h}\pi_0^h(1-\pi_0)^{p-h}, \qquad h = 1, \ldots, p_+, \tag{5}$$

so that, conditional on π_0, the prior distribution for the number of SNPs having zero coefficients is binomial. We refer to the group of SNPs having zero coefficients as the *null cluster*.

Theorem 1. Assume $\beta_{lh} \overset{\text{iid}}{\sim} G$ for $l = 1, \ldots, p$, $h = 1, 2, 3$, with G defined by expressions (3) and (4). Then, letting k^* denote the number of unique nonzero elements of $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_p')$, with $\boldsymbol{\beta}_l = (\beta_{l1}, \beta_{l2}, \beta_{l3})'$, the prior distribution for k^* is

$$\Pr(k^* = k|\pi_0, \alpha_0, p_+) = \sum_{h=0}^{p_+}\binom{p_+}{h}(1-\pi_0)^h\pi_0^{p_+-h}\frac{a_h(k)\alpha_0^k}{\alpha_0^{(h)}}, \tag{6}$$

where $a_h(k)$ are the absolute values of Stirling numbers of the first kind (refer to Abramowitz and Stegun 1964, p. 833), and $\alpha^{(h)} = \alpha(\alpha+1)\cdots(\alpha+h-1)$.

The proof is straightforward using expressions (3) and (4) and the result of Antoniak (1974), page 1161.

From Theorem 1, it is clear that the number of nonnull clusters, which are defined as groups of SNPs having identical nonzero regression coefficients, increases stochastically with α_0 and decreases with π_0. Hence, α_0 and π_0 are key hyperparameters controlling the clustering of SNPs into null and nonnull groups. Note also that the number of clusters increases automatically as the number of SNPs under consideration, p_+, increases. It is also apparent that the approach performs simultaneous variable selection and clustering, classifying a subset of SNPs as having no effect while clustering the remaining SNPs.

To obtain additional insight into the clustering process, we derive prior probabilities of the coefficients $\boldsymbol{\beta}$ falling into different $\mathcal{C}$ classes, where $\boldsymbol{\beta}$ belongs to class $\mathcal{C}(m_0, m_1, \ldots, m_{p_+})$ if there are m_0 elements of $\boldsymbol{\beta}$ equal to 0, m_1 nonzero elements of $\boldsymbol{\beta}$ that occur once, m_2 nonzero elements that occur twice, up to m_{p_+} nonzero elements that occur p_+ times. It follows that $k = 1(m_0 > 0) + \sum_{h=1}^{p_+} m_h = 1(m_0 > 0) + k^*$ is the number of unique elements of $\boldsymbol{\beta}$, denoted $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$.

Theorem 2. Assume $\beta_{lh} \overset{\text{iid}}{\sim} G$ for $l = 1, \ldots, p$, $h = 1, 2, 3$, with G defined by expression (3). Then, the prior probability that $\boldsymbol{\beta}$ belongs to class $C(m_0, m_1, \ldots, m_{p+})$ is

$$\Pr\{\boldsymbol{\beta} \in \mathcal{C}(m_0, m_1, \ldots, m_{p_+})\} = \left[\sum_{m_0=0}^{p_+}\binom{p_+}{m_0}\pi_0^{m_0}(1-\pi_0)^{p_+-m_0} \times \frac{(p_+-m_0)!}{\prod_{h=1}^{p_+-m_0}h^{m_h}(m_h!)}\frac{\alpha_0^{\sum_{h=1}^{p_+-m_0}m_h}}{\alpha_0^{(p_+-m_0)}}\right]. \tag{7}$$

This theorem follows from proposition 3 of Antoniak (1974) after appropriate modification to mix in the fixed point mass.

Theorem 2 can be used to derive the prior probabilities corresponding to a number of interesting special cases. For example, the probability that none of the SNPs has an effect is simply $\Pr\{\boldsymbol{\beta} \in \mathcal{C}(p_+, 0, \ldots, 0)\} = \pi_0^{p_+}$. The probability that all the SNPs have an equivalent nonnull effect is

$$\Pr\{\boldsymbol{\beta} \in \mathcal{C}(0, \ldots, 0, 1)\} = (1-\pi_0)^{p_+}\frac{\alpha_0(p_+-1)!}{\prod_{h=1}^{p_+}(\alpha_0+h-1)}.$$

Other class probabilities corresponding to different numbers of null and nonnull SNPs, and various clustering in the nonnull SNPs, can be calculated easily. Potentially, π_0 and α_0 can be chosen subjectively based on back-calculating from these probabilities.

Downloaded by [NIH Library] at 06:55 26 December 2013

Downloaded by [NIH Library] at 06:55 26 December 2013

3.3 Semiparametric Hierarchical Clustering

We now consider the case in which SNPs occur within different functionally related genes, and interest focuses on variable selection and clustering within and across genes. In particular, to borrow information across genes, it is appealing to develop a method that allows SNPs for different genes to be assigned to the same cluster, while also allowing clusters to be gene specific. One approach would be to extend the prior of Section 2.2 to incorporate a c subscript on G and then account for dependency in the elements of $\mathcal{G} = \{G_c : c = 1, \ldots, C\}$ by applying the dependent Dirichlet process (DDP) of MacEachern (1999, 2000; see also De Iorio, Müller, Rosner, and MacEachern 2004). This could be accomplished by defining parallel stick-breaking formulations for each G_c and modeling dependence through a stochastic process for the atoms.

This DDP approach allows for dependence in the coefficients between genes but does not allow clustering of SNPs in different genes. Another possibility for defining dependence in $\mathcal{G}$ is to use the mixture approach of Müller, Quintana, and Rosner (2004), letting $G_c = \pi F_c + (1-\pi)F_0$, with $\{F_0, F_1, \ldots, F_C\}$ assigned independent DP priors. This formulation allows atoms to be shared across genes through the incorporation of the global component, F_0. Alternatively, one could use the hierarchical DP of Teh, Jordan, Beal, and Blei (2006), which would let $G_c \sim \mathrm{DP}(G)$, with $G \sim \mathrm{DP}(\alpha G_0)$. Either of these approaches could potentially be generalized to incorporate a point mass at 0.

We propose a simpler additive formulation more closely related to traditional multilevel regression models:

$$\begin{aligned}
&\beta_{clh} = \psi_c + \gamma_{clh}, \\
&\qquad h = 1, 2, 3,\ l = 1, \ldots, p_c,\ c = 1, \ldots, C, \\
&\gamma_{clh} \overset{\text{iid}}{\sim} F, \qquad F \sim \delta_0 \mathrm{DP}(\pi_0, \alpha_0 F_0), \\
&\qquad F_0 = \mathrm{N}_+(\mu_0, \sigma_0^2), \qquad \forall c, h, l, \\
&\psi_c \overset{\text{iid}}{\sim} H, \qquad H \sim \delta_0 \mathrm{DP}(\pi_1, \alpha_1 H_0), \\
&\qquad H_0 = \mathrm{N}_+(\mu_1, \sigma_1^2), \qquad \forall c,
\end{aligned} \tag{8}$$

where ψ_c is a gene-specific factor, γ_{clh} is an SNP-specific factor, and $\delta_0 \mathrm{DP}(\pi_0, \alpha_0 F_0)$ is shorthand for the prior described in expression (3), consisting of a point mass at zero with probability π_0 and a $\mathrm{DP}(\alpha F_0)$ prior with probability $1 - \pi_0$.

This multilevel formulation allows clustering of SNPs both within and across genes. For the sake of parsimony, we ignore possible within-locus dependency in $\{\beta_{cl1}, \beta_{cl2}, \beta_{cl3}\}$, which may arise when one or two copies of a disease allele convey a similar risk. Our motivation for this simplifying assumption is that we typically have no a priori knowledge of the dependency structure, and a priori independence does not imply a posteriori independence.

There are $k_\Gamma \le p_+ = 3\sum_{c=1}^{C} p_c$ unique values $\boldsymbol{\Gamma} = (\Gamma_1, \ldots, \Gamma_{k_\Gamma})'$ of the SNP-specific factor $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1', \ldots, \boldsymbol{\gamma}_C')$, with $\boldsymbol{\gamma}_c = (\boldsymbol{\gamma}_{c1}', \ldots, \boldsymbol{\gamma}_{c,p_c}')'$ for $c = 1, \ldots, C$ and $\boldsymbol{\gamma}_{cl} = (\gamma_{cl1}, \gamma_{cl2}, \gamma_{cl3})'$ for $l = 1, \ldots, p_c$. There are also $k_\Psi \le C$ unique values $\boldsymbol{\Psi} = (\Psi_1, \ldots, \Psi_{k_\Psi})'$ of the gene-specific factor $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_C)'$. Because of the masses at 0, one of the unique values of each of these vectors will typically correspond to 0, so we let $\Gamma_1 = 0$ and $\Psi_1 = 0$ without loss of generality. From this characterization, SNPs from the same gene (say SNPs c, l, h and c, l', h') belong to the same cluster if $\gamma_{clh} = \gamma_{cl'h'}$, whereas SNPs from different genes (say SNPs c, l, h and c', l', h') belong to the same cluster if $\gamma_{clh} = \gamma_{c'l'h'}$ and $\psi_c = \psi_{c'}$. In particular, given that the SNPs are not null, the probabilities of belonging to the same cluster are

$$\begin{aligned}
&\Pr(\beta_{clh} = \beta_{cl'h'} | \beta_{clh} \ne 0, \beta_{cl'h'} \ne 0) \\
&\quad = \pi_0^2(1-\pi_1) + \frac{(1-\pi_0)^2}{\alpha_0 + 1}, \\
&\Pr(\beta_{clh} = \beta_{c'l'h'} | \beta_{clh} \ne 0, \beta_{c'l'h'} \ne 0) \\
&\quad = \frac{\pi_1^2(1-\pi_0)^2}{\alpha_0+1} + \frac{(1-\pi_1)^2\pi_0^2}{\alpha_1+1} + \frac{(1-\pi_1)^2(1-\pi_0)^2}{(\alpha_1+1)(\alpha_0+1)}.
\end{aligned}$$

After some algebra, it follows that SNPs from the same gene are clustered together with higher probability than SNPs from different genes. Such within-gene dependence is often biologically reasonable. A priori, SNPs in different regulatory regions for the same cytokine should have a higher chance of belonging to the same cluster than SNPs for different cytokines, because the biological action of cytokines can vary.

To obtain additional insight into the clustering properties, we focus on the null cluster containing SNPs with zero coefficients. For greater flexibility, we choose hyperprior densities for π_0 and π_1 as follows:

$$\pi_0 \sim \mathrm{beta}(a_0, b_0) \qquad \text{and} \qquad \pi_1 \sim \mathrm{beta}(a_1, b_1), \tag{9}$$

where $\mathbf{a} = (a_1, a_2)'$ and $\mathbf{b} = (b_1, b_2)'$ are prespecified hyperparameters. Integrating out π_0 and π_1, the probability that SNP c, l, h is null ($\beta_{clh} = 0$) is

$$\begin{aligned}
\Pr(\beta_{clh} = 0 | \mathbf{a}, \mathbf{b}) &= \int\int \pi_0\pi_1 \frac{\pi_0^{a_0-1}(1-\pi_0)^{b_0-1}}{B(a_0, b_0)} \\
&\qquad \times \frac{\pi_1^{a_1-1}(1-\pi_1)^{b_1-1}}{B(a_1, b_1)}\, d\pi_0\, d\pi_1 \\
&= \frac{a_0}{a_0+b_0}\frac{a_1}{a_1+b_1}.
\end{aligned} \tag{10}$$

Similarly, the probability that all the SNPs in the cth gene belong to the null cluster is

$$\Pr(\boldsymbol{\beta}_c = \mathbf{0} | \mathbf{a}, \mathbf{b}) = \frac{a_0^{(3p_c)}}{(a_0+b_0)^{(3p_c)}}\frac{a_1}{a_1+b_1}, \tag{11}$$

and the probability that all SNPs in all genes belong to the null cluster is

$$\Pr(\boldsymbol{\beta} = \mathbf{0} | \mathbf{a}, \mathbf{b}) = \frac{a_0^{(p_+)}}{(a_0+b_0)^{(p_+)}}\frac{a_1^{(C)}}{(a_1+b_1)^{(C)}}. \tag{12}$$

To illustrate the dependence structure, note that the probability that SNP l, h in gene c is null given another SNP l', h' in gene c is null is

$$\begin{aligned}
\Pr(\beta_{clh} = 0 | \beta_{cl'h'} = 0, \mathbf{a}, \mathbf{b}) &= \frac{\Pr(\beta_{clh} = \beta_{cl'h'} = 0 | \mathbf{a}, \mathbf{b})}{\Pr(\beta_{cl'h'} = 0 | \mathbf{a}, \mathbf{b})} \\
&= \frac{a_0+1}{a_0+b_0+1},
\end{aligned} \tag{13}$$

which is always higher than $\Pr(\beta_{clh} = 0|\mathbf{a}, \mathbf{b})$. If we instead condition on knowledge that an SNP in a different gene is null, we obtain

$$\begin{aligned}\Pr(\beta_{clh} &= 0|\beta_{c'l'h'} = 0, \mathbf{a}, \mathbf{b}) \\ &= \frac{\Pr(\beta_{clh} = \beta_{c'l'h'} = 0|\mathbf{a}, \mathbf{b})}{\Pr(\beta_{c'l'h'} = 0|\mathbf{a}, \mathbf{b})} \\ &= \frac{a_0 + 1}{a_0 + b_0 + 1}\frac{a_1 + 1}{a_1 + b_1 + 1}, \end{aligned} \tag{14}$$

which is also higher than $\Pr(\beta_{clh} = 0|\mathbf{a}, \mathbf{b})$ [shown in expression (10)] but lower than the probability in expression (13). Thus, the dependence between SNPs in a gene is higher than the dependence between SNPs in different genes, with the magnitude of the difference depending on the hyperparameters $\mathbf{a}$ and $\mathbf{b}$. In the limit as $a_0 + b_0 \to \infty$ and $a_1 + b_1 \to \infty$, holding $a_0/(a_0 + b_0)$ and $a_1/(a_1 + b_1)$ constant, expressions (10), (13), and (14) are equivalent and there is no borrowing of information across SNPs about the probability of membership in the null cluster.

3.4 Prior Elicitation

Motivated by the cytokine application, we illustrate a strategy for prior elicitation. In particular, in choosing a_0, b_0, a_1, b_1, we recommend back-calculating from prior probabilities corresponding to different global and local hypotheses. For example, one can specify (i) the prior probability that none of the SNPs is associated with spontaneous preterm birth, $\Pr(\boldsymbol{\beta} = 0|\mathbf{a}, \mathbf{b})$; (ii) the probability that a randomly selected SNP is null, $\Pr(\beta_{clh} = 0|\mathbf{a}, \mathbf{b})$; (iii) the probability that two SNPs within a gene are null, $\Pr(\beta_{clh} = \beta_{cl'h'} = 0|\mathbf{a}, \mathbf{b})$; and (iv) the probability that two SNPs within different genes are null, $\Pr(\beta_{clh} = \beta_{c'l'h'} = 0|\mathbf{a}, \mathbf{b})$. The hierarchical structure implies that the probabilities are ordered (i) < (iv) < (iii) < (ii), so one should choose values consistent with this constraint. Because (i)–(iv) are simple analytic functions of a_0, b_0, a_1, b_1, it is straightforward to solve the system of nonlinear equations using numerical methods.

In the cytokine application, we let $\Pr(\boldsymbol{\beta} = 0|\mathbf{a}, \mathbf{b}) = .5$ to set the probability of the global null hypothesis equal to .5, corresponding to a 50% chance that any of the SNPs are predictive of spontaneous preterm birth. This represents a Bayesian approach to limit false positives that arise in multiple testing. We then let $\Pr(\beta_{clh} = 0|\mathbf{a}, \mathbf{b}) = .8$, noting that the Bayesian Bonferroni approach (Westfall et al. 1997), which treats local hypotheses as independent, ignoring correlation, would instead choose $.5^{1/66} = .990$. A 1% chance that an SNP is important is unrealistically low, given that we are studying promising candidate SNPs. In addition, such a low prior probability would result in a very conservative procedure, requiring very large sample sizes to detect a health effect of the magnitude that would be expected in this study (e.g., odds ratio between .5 and 2). As plausible values for probabilities (iii) and (iv), we choose .75 and .73, respectively. These values are chosen to be slightly lower than .8, with a modest degree of within-gene dependence. In simulation studies, we have found a high degree of robustness to the specific values chosen.

The precision parameters α_0 and α_1 are assigned gamma hyperprior distributions:

$$\begin{aligned}\alpha_0 &\sim \text{gamma}(a_{\alpha_0}, b_{\alpha_0}) \qquad \text{and} \\ \alpha_1 &\sim \text{gamma}(a_{\alpha_1}, b_{\alpha_1}),\end{aligned} \tag{15}$$

where $a_{\alpha_0}, b_{\alpha_0}, a_{\alpha_1}, b_{\alpha_1}$ are prespecified hyperparameters. In choosing these values, we recommend letting $a_{\alpha_0} = b_{\alpha_0} = a_{\alpha_1} = b_{\alpha_1} = 1$ as a somewhat vague prior for the number of global and local clusters, which favors smaller numbers of clusters.

For the hyperparameters characterizing the base distributions, $\mu_0, \sigma_0^2, \mu_1, \sigma_1^2$, we recommend letting $\mu_0 = \mu_1 = 0$ and $\sigma_0^2 = \sigma_1^2 = 1$. The resulting F_0 and H_0 correspond to truncated normal distributions assigning high probability to a wide range of plausible values for the increase in risk of spontaneous preterm birth attributable to an adverse genotype. Potentially, hyperprior densities could be chosen for the means and variances for greater flexibility. This may be a useful strategy in cases in which there are very large numbers of candidate genes and SNPs, and less is known about scientifically plausible values for the regression coefficients.

4. POSTERIOR COMPUTATION

For ease in posterior computation, we focus on the case in which the link function $\mathcal{L}(\cdot)$ corresponds to the cdf for the t distribution with ν degrees of freedom:

$$\begin{aligned}\Pr(y_i = 1|\mathbf{g}_i, \mathbf{z}_i) = \int_{\Re+}\int_{\Re+} &\mathrm{N}(y_i^*; \mathbf{z}_i'\boldsymbol{\kappa} + \mathbf{x}_i'\boldsymbol{\beta}, \phi_i^{-1}\sigma^2) \\ &\times \text{gamma}(\phi_i; \nu/2, \nu/2)\, d\phi_i\, dy_i^*,\end{aligned} \tag{16}$$

where $\mathbf{x}_i$ is a vector of 0/1 indicators of genotype category within each locus of each gene. This form facilitates the use of an efficient data augmentation algorithm, while also allowing one to obtain a highly accurate approximation to a logistic regression model by letting $\sigma^2 = \pi(\nu - 2)/3\nu$ and $\nu = 7.3$ (O'Brien and Dunson 2004, and the references therein).

Let $y_i = 1(y_i^* > 0)$, where $y_i^* = \mathbf{z}_i'\boldsymbol{\kappa} + \mathbf{x}_i'\boldsymbol{\beta} + \phi_i^{-1/2}\epsilon_i$, with $\phi_i \sim \text{gamma}(\nu/2, \nu/2)$ and $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$. Then, the algorithm alternates between (1) sampling y_i^* and ϕ_i from their respective truncated normal and gamma full conditional posterior distributions for $i = 1, \ldots, n$ and (2) sampling unknowns related to $\boldsymbol{\beta}$ jointly with $\boldsymbol{\kappa}$, assuming a $\mathrm{N}(\boldsymbol{\kappa}_0, \boldsymbol{\Sigma}_\kappa)$ prior for $\boldsymbol{\kappa}$. The first step is straightforward, so we focus our attention on step 2.

Letting the (clh) superscript denote a quantity obtained excluding element c, l, h, the conditional prior distribution of γ_{clh} given $\boldsymbol{\gamma}^{(clh)}$ is

$$\begin{aligned}&\frac{\alpha_0(1-\pi_0)}{\alpha_0 + p - p_{\Gamma_1}^{(clh)} - 1}\mathrm{N}_+(\mu_0, \sigma_0^2) + \pi_0\delta_0 \\ &\qquad + \sum_{l=2}^{k_\Gamma^{(clh)}} \frac{p_{\Gamma_l}^{(clh)}(1-\pi_0)}{\alpha_0 + p - p_{\Gamma_1}^{(clh)} - 1}\delta_{\Gamma_l^{(clh)}},\end{aligned} \tag{17}$$

where $p_{\Gamma_l}^{(clh)}$ is the number of elements of $\boldsymbol{\gamma}^{(clh)}$ equal to $\Gamma_l^{(clh)}$, $\boldsymbol{\Gamma}^{(clh)} = (\Gamma_l^{(clh)}, l = 1, \ldots, k_\Gamma^{(clh)})'$, $\Gamma_1^{(clh)} = 0$, $\Gamma_l^{(clh)}$ for $l = 2, \ldots, k_\Gamma^{(clh)}$ denotes the unique nonzero values of $\boldsymbol{\gamma}^{(clh)}$, and $k_\Gamma^{(clh)}$ is the number of atoms in (17).

Downloaded by [NIH Library] at 06:55 26 December 2013

Downloaded by [NIH Library] at 06:55 26 December 2013

The conditional distribution of ψ_c given $\boldsymbol{\psi}^{(c)} = (\psi_1, \ldots, \psi_{c-1}, \psi_{c+1}, \ldots, \psi_C)'$ is

$$\frac{\alpha_1(1-\pi_1)}{\alpha_1 + C - p_{\Psi_1}^{(c)} - 1}\mathrm{N}_+(\mu_1, \sigma_1^2) + \pi_1\delta_0 + \sum_{l=2}^{k_\Psi^{(c)}} \frac{p_{\Psi_l}^{(c)}(1-\pi_1)}{\alpha_1 + C - p_{\Psi_1}^{(c)} - 1}\delta_{\Psi_l^{(c)}}, \quad (18)$$

where $p_{\Psi_l}^{(c)}$ is the number of elements of $\boldsymbol{\psi}^{(c)}$ equal to $\Psi_l^{(c)}$, $\boldsymbol{\Psi}^{(c)} = (\Psi_l^{(c)}, l = 1, \ldots, k_\Psi^{(c)})'$, $\Psi_1 = 0$, Ψ_l for $l = 2, \ldots, k_\Psi^{(c)}$ denotes the unique nonzero values of $\boldsymbol{\psi}^{(c)}$, and $k_\Psi^{(c)}$ is the number of atoms in (18).

As shorthand, let $\mathbf{u}^{(clh)} = (u_0^{(clh)}, u_1^{(clh)}, \ldots, u_{k_\Gamma^{(clh)}}^{(clh)})'$ and $\mathbf{w}^{(c)} = (w_0^{(c)}, w_1^{(c)}, \ldots, w_{k_\Psi^{(c)}}^{(c)})'$ denote the probability weights on the respective mixture components in expressions (17) and (18). Updating conditional priors (17) and (18) using information in the data, we obtain the following full conditional posterior distributions:

$$(\gamma_{clh}|-) = U_0^{(clh)}\mathrm{N}_+\big(\gamma_{clh}; E_\gamma^{(clh)}, V_\gamma^{(clh)}\big) + \sum_{l=1}^{k_\Gamma^{(clh)}} U_l^{(clh)}\delta_{\Gamma_l^{(clh)}}, \quad (19)$$

$$(\psi_c|-) = W_0^{(c)}\mathrm{N}_+\big(\psi_c; E_\psi^{(c)}, V_\psi^{(c)}\big) + \sum_{l=1}^{k_\Psi^{(c)}} W_l^{(c)}\delta_{\Psi_l^{(c)}}, \quad (20)$$

where $\tilde{y}_i^{(c)} = y_i^* - \mathbf{z}_i'\boldsymbol{\kappa} - \mathbf{x}_i\boldsymbol{\gamma} - \sum_{c' \neq c} x_{ic'}\psi_{c'}$, $x_{ic} = \sum_{l=1}^{p_c}\sum_{h=1}^{3} x_{iclh}$, $\tilde{y}_i^{(clh)} = y_i^* - \mathbf{z}_i'\boldsymbol{\kappa} - \mathbf{x}_i^{(clh)'}\boldsymbol{\beta}^{(clh)} - x_{iclh}\psi_c$, the posterior means and variances in the $\mathrm{N}_+(\cdot)$ components are

$$V_\gamma^{(clh)} = \left(\sigma_0^{-2} + \sum_{i=1}^n \sigma_i^{-2}x_{iclh}^2\right)^{-1},$$

$$E_\gamma^{(clh)} = V_\gamma^{(clh)}\left(\sigma_0^{-2}\mu_0 + \sum_{i=1}^n \sigma_i^{-2}x_{iclh}\tilde{y}_i^{(clh)}\right),$$

$$V_\psi^{(c)} = \left(\sigma_1^{-2} + \sum_{i=1}^n \sigma_i^{-2}x_{ic}^2\right)^{-1},$$

$$E_\psi^{(c)} = V_\psi^{(c)}\left(\sigma_1^{-2}\mu_1 + \sum_{i=1}^n \sigma_i^{-2}x_{ic}\tilde{y}_i^{(c)}\right),$$

and the updated mixture weights are defined as follows:

$$U_0^{(clh)} = c_u \times \frac{u_0^{(clh)}\mathrm{N}_+(0; \mu_0, \sigma_0^2)\prod_{i=1}^n \mathrm{N}(\tilde{y}_i^{(clh)}; 0, \sigma_i^2)}{\mathrm{N}_+(0; E_\gamma^{(clh)}, V_\gamma^{(clh)})},$$

$$U_l^{(clh)} = c_u \times u_l^{(clh)}\prod_{i=1}^n \mathrm{N}\big(\tilde{y}_i^{(clh)}; x_{iclh}\Gamma_l^{(clh)}, \sigma_i^2\big),$$

$$W_0^{(c)} = c_w \times \frac{w_0^{(c)}\mathrm{N}_+(0; \mu_1, \sigma_1^2)\prod_{i=1}^n \mathrm{N}(\tilde{y}_i^{(c)}; 0, \sigma_i^2)}{\mathrm{N}_+(0; E_\psi^{(c)}, V_\psi^{(c)})},$$

$$W_l^{(c)} = c_w \times w_l^{(c)}\prod_{i=1}^n \mathrm{N}\big(\tilde{y}_i^{(c)}; x_{ic}\Psi_l^{(c)}, \sigma_i^2\big),$$

where c_u and c_w are normalizing constants and $\sigma_i^2 = \phi_i^{-1}\sigma^2$, for $i = 1, \ldots, n$.

We follow West, Müler, and Escobar (1994) and MacEachern (1994) in alternating between updating (i) the cluster allocation and (ii) the cluster-specific parameters. Let $\mathcal{S}_{clh} = l$ if $\gamma_{clh} = \Gamma_l^{(clh)}$ for $l = 1, \ldots, k_\Gamma^{(clh)}$ and $\mathcal{T}_c = l$ if $\psi_c = \Psi_l^{(c)}$ for $l = 1, \ldots, k_\Psi^{(c)}$ index the allocation of γ_{clh} and ψ_c to clusters. In addition, let $\mathcal{S}_{clh} = 0$ denote that $\gamma_{clh} \notin \boldsymbol{\Gamma}^{(clh)}$, so that SNP c, l, h cannot be grouped with the other SNPs and a new cluster needs to be introduced. Also, $\mathcal{T}_c = 0$ denotes that $\psi_c \notin \boldsymbol{\Psi}^{(c)}$, so that a new cluster is introduced for gene c. The conditional posterior distributions of $\mathcal{S}_{clh}$ and $\mathcal{T}_c$ are, respectively,

$$\big(\mathcal{S}_{clh}|\mathbf{S}^{(clh)}, \mathbf{T}, \boldsymbol{\Gamma}^{(clh)}, \boldsymbol{\Psi}, \text{data}\big) = \text{Multinomial}\big(0, 1, \ldots, k_\Gamma^{(clh)}; U_0^{(clh)}, U_1^{(clh)}, \ldots, U_{k_\Gamma^{(clh)}}^{(clh)}\big), \quad (21)$$

$$\big(\mathcal{T}_c|\mathbf{S}, \mathbf{T}^{(c)}, \boldsymbol{\Gamma}, \boldsymbol{\Psi}^{(c)}, \text{data}\big) = \text{Multinomial}\big(0, 1, \ldots, k_\Psi^{(c)}; W_0^{(c)}, W_1^{(c)}, \ldots, W_{k_\Psi^{(c)}}^{(c)}\big). \quad (22)$$

In step 2(i), we sample from these multinomial distributions. When $\mathcal{S}_{clh} = 0$, a new value for γ_{clh} is drawn from $\mathrm{N}_+(E_\gamma^{(clh)}, V_\gamma^{(clh)})$, whereas, when $\mathcal{T}_c = 0$, a new value for ψ_c is drawn from $\mathrm{N}_+(E_\psi^{(c)}, V_\psi^{(c)})$.

Then, in step 2(ii), we update $\boldsymbol{\kappa}$ and the cluster-specific parameters $\boldsymbol{\Gamma}^* = (\Gamma_2, \ldots, \Gamma_{k_\Gamma})'$ and $\boldsymbol{\Psi}^* = (\Psi_2, \ldots, \Psi_{k_\Psi})'$ by alternately sampling from their conditional posterior distributions. Given the cluster allocation indicators, $\mathbf{S}, \mathbf{T}$, and number of clusters, k_Γ, k_Ψ, these conditional distributions have normal and truncated normal forms, so details are excluded. Finally, the full conditional posterior distributions of π_0 and π_1 are, respectively

$$(\pi_0|\mathbf{S}, \boldsymbol{\Gamma}, \mathbf{T}, \boldsymbol{\Psi}, \boldsymbol{\kappa}, \mathbf{y}^*) = \text{beta}\left(a_0 + \sum_{c,l,h} 1(\mathcal{S}_{clh} = 1), b_0 + p - \sum_{c,l,h} 1(\mathcal{S}_{clh} = 1)\right), \quad (23)$$

$$(\pi_1|\mathbf{S}, \boldsymbol{\Gamma}, \mathbf{T}, \boldsymbol{\Psi}, \boldsymbol{\kappa}, \mathbf{y}^*) = \text{beta}\left(a_1 + \sum_{c=1}^C 1(\mathcal{T}_c = 1), b_1 + C - \sum_{c=1}^C 1(\mathcal{T}_c = 1)\right). \quad (24)$$

In step 2(iii), we sample from these conditionals. Finally, in step 2(iv), we update α_0 and α_1 by applying the approach of Escobar and West (1995).

5. CYTOKINES AND PRETERM BIRTH APPLICATION

5.1 Real Data Results

We applied the approach to data from the cytokines and spontaneous preterm birth application introduced in Section 2. Using the approach to prior elicitation proposed in Section 3.4, we obtained $a_0 = .48, b_0 = .089, a_1 = .78, b_1 = .046$, which implies $\Pr(\beta_{ch} = 0) = .8$ and $\Pr(\boldsymbol{\beta} = 0) = .5$. We ran analyses separately for African Americans ($n = 195$) and whites

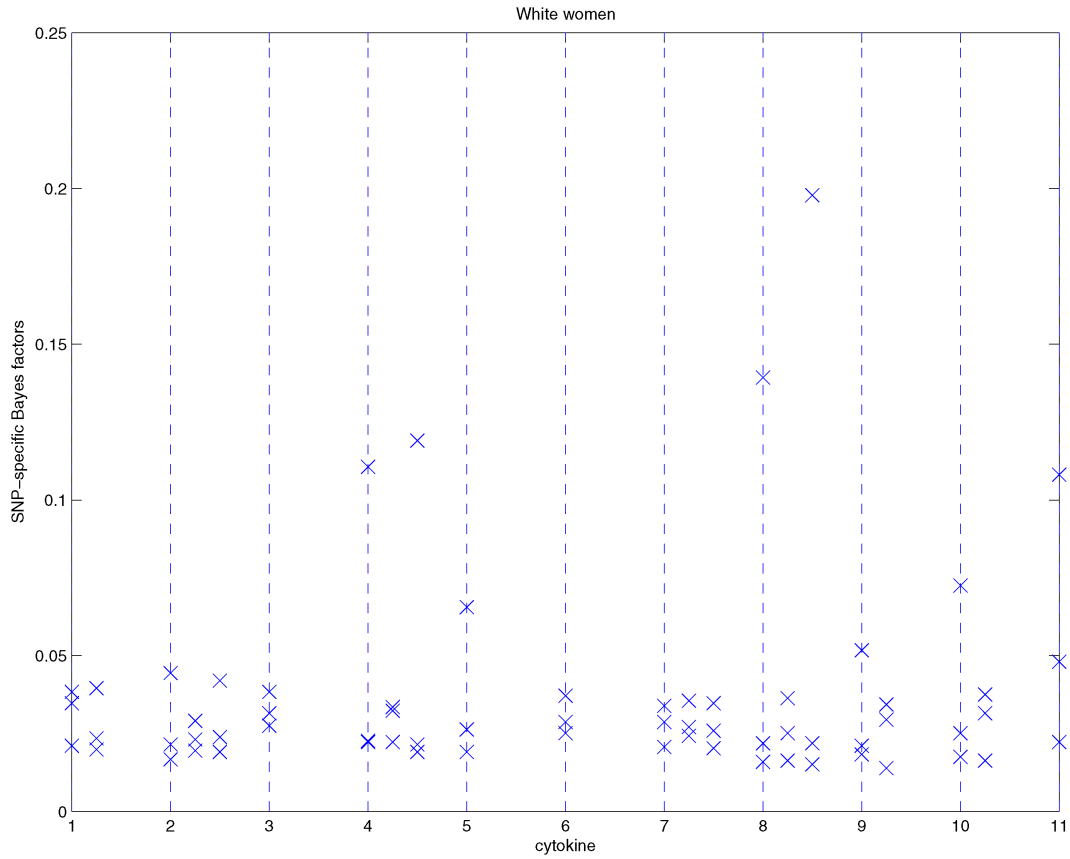

Figure 1. Estimated Bayes factors for the different genotype × locus categories of each cytokine (white women).

Downloaded by [NIH Library] at 06:55 26 December 2013

($n = 252$), following standard epidemiologic practice in this area. In each case, the Markov chain Monte Carlo (MCMC) algorithm was run for 100,000 iterations, discarding a burn-in of 5,000 iterations and collecting every 10th sample to thin the chain. Samples converged quickly to a stationary distribution, and mixing was rapid, suggesting that the proposed algorithm is efficient.

For whites, the posterior probability of the global alternative hypothesis that any of the SNPs were predictive of spontaneous preterm birth was $\Pr(H_1|\text{data}) = .170$, with the corresponding Bayes factor being BF = .204. For African Americans, the values were $\Pr(H_1|\text{data}) = .110$ and BF = .123. Hence, the data support the null hypothesis that cytokine polymorphisms are not predictive of spontaneous preterm birth. These results are robust to moderate changes in the prior, and we repeated the analysis with $\Pr(\boldsymbol{\beta} = 0) = .2$ and $\Pr(\beta_{ch} = 0) = .5$ without change in the conclusion. Conclusions were also unchanged running the analysis for African Americans and whites combined. SNP-specific BFs are provided in Figures 1 and 2 for whites and African Americans. In each case, the SNP-specific BFs were well below 1. In general, estimated coefficients given inclusion in the model tend to parallel the results shown in Table 1, with the extreme estimates occurring at similar locations. However, the model-averaged estimates are all approximately 0, reflecting the low inclusion probabilities.

5.2 Simulation Study

A potential concern is that the approach may be overly conservative, particularly in cases in which there are effects only for one or two of the SNPs. In such cases, borrowing of information across the SNPs regarding the probability of inclusion in the model can conceivably cause SNP-specific effects to be obscured. To assess whether the approach is capable of detecting SNP-specific effects, we ran a small simulation study. In particular, using the sample size and genotype data for the women in the PIN study but randomly permuting the assignment of SNPs to subjects, we simulated the spontaneous preterm birth outcome variable under four different scenarios: (i) null model ($\beta_{clh} = 0$ for all c, l, h); (ii) one SNP (cytokine IL13, location −1,112, genotype TC) positive ($\beta_{822} = 1$, $\beta_{clh} = 0$ for all $c, l, h \neq 8, 2, 2$); (iii) two SNPs positive ($\beta_{822} = \beta_{912} = 1$); and (iv) two nonnull clusters at $\beta = 1$ [SNPs $\{(2, 3, 2), (3, 1, 2), (5, 1, 2), (7, 1, 2), (9, 2, 2)\}$] and $\beta = 1.5$ [SNPs $\{(4, 2, 2), (6, 1, 2), (8, 2, 2), (10, 2, 2)\}$].

Under each scenario, we simulated 25 datasets, implementing the MCMC algorithm as for the real data example in each case, but with the algorithm run for 10,000 iterations with a 1,000 burn-in. Summaries of the results for cases (i)–(iii) are presented in Table 2. In case (i), 0/25 datasets had estimated Bayes factors greater than 1 for either the global or the local alternative, suggesting that the approach does not tend to produce many false positives. Motivated by these results, we recommend a threshold of BF > 1 for identifying interesting SNPs for further study. Borrowing of information across the many null SNPs pulls down the BF through a Bayesian adjustment for multiple comparisons, and a higher threshold would result in very low power. In case (ii), 22/25 of the datasets had

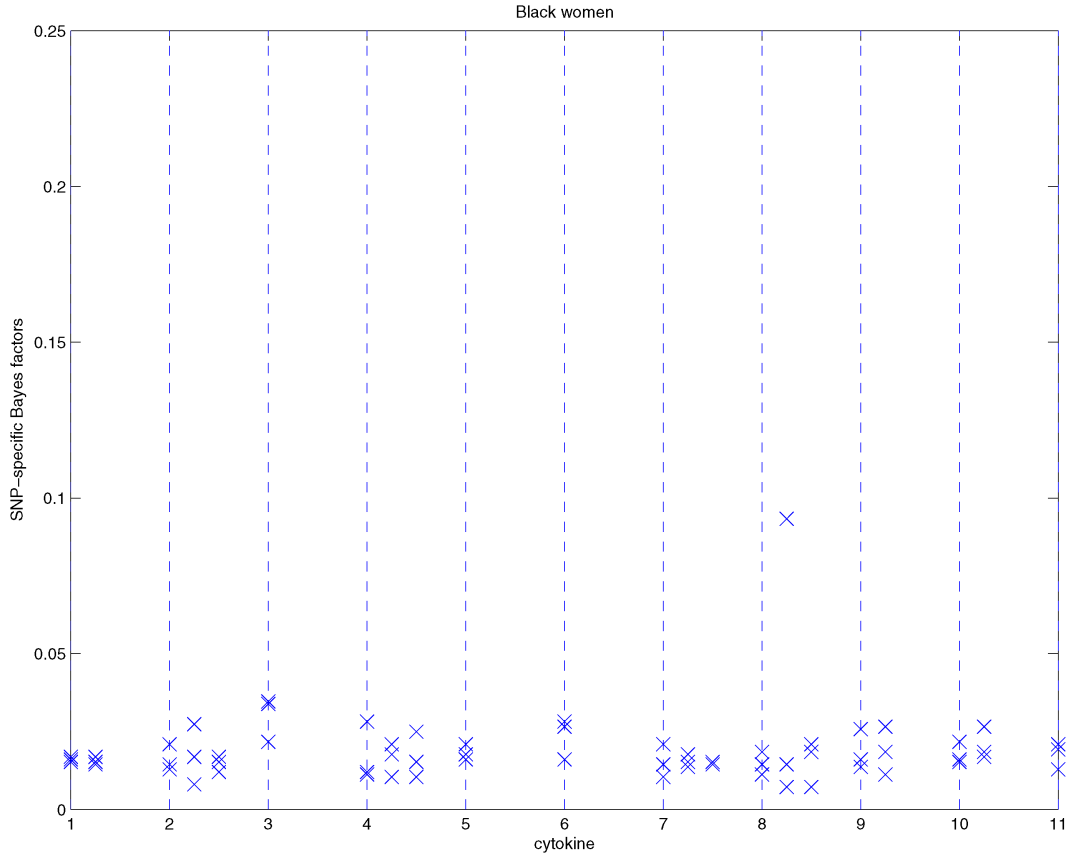

Figure 2. Estimated Bayes factors for the different genotype × locus categories of each cytokine (African American women).

$BF_{8,2,2} > 1$ and 18/25 datasets had a global BF > 1, suggesting that the approach has good power to detect single SNP effects. In addition, the preferred model posterior means for the SNP-specific coefficients were close to the true value, with mild attenuation toward 0. Results were even better in case (iii), with higher power and less attenuation.

For the more complex case (iv), results are presented in Figures 3 and 4. Figure 3 shows the estimated marginal inclusion probabilities for each of the SNPs, ordered so that 1–57 are in the null cluster, 58–62 are in the $\beta = 1$ cluster, and 63–66 are in the $\beta = 1.5$ cluster. Values for each of the 25 simulated datasets are shown, with the horizontal line representing the average in-

Table 2. Simulation results

		Summary across simulations*			
Case	Quantity	Mean	Median	[25th, 75th]	Proportion > 0
(i)	$\log \mathrm{BF}$	−2.31	−2.40	[−2.56, −2.08]	.00
	$\log \overline{\mathrm{BF}}_-$	−3.89	−4.04	[−4.18, −3.65]	.00
	$\overline{\beta}_-$	.0013	.001	[.0007, .0016]	1.00
(ii)	$\log \mathrm{BF}$	2.26	2.75	[−.77, 4.62]	.72
	$\log \mathrm{BF}_{8,3}$	3.35	4.06	[.40, 5.84]	.88
	$\log \overline{\mathrm{BF}}_-$	−2.76	−2.61	[−3.37, −2.32]	.00
	$\beta_{8,3}$	.72*	.92	[.00, 1.07]	.72
	$\overline{\beta}_-$	.0005	.0000	[.0000, .0000]	.04
(iii)	$\log \mathrm{BF}$	>10	6.11	[3.64, >10]	1.00
	$\log \mathrm{BF}_{8,3}$	>10	4.84	[3.13, 7.47]	.96
	$\log \mathrm{BF}_{9,1}$	>10	5.20	[1.98, 6.96]	1.00
	$\log \mathrm{BF}_-$	−2.12	−2.14	[−2.41, −1.78]	.00
	$\beta_{8,3}$	.82	.87	[.80, 1.02]	.88
	$\beta_{9,1}$	.82	.89	[.69, .97]	.92
	β_-	.0000	.0000	[.0000, .0000]	.0000

NOTE: BF, global Bayes factor in favor of H_1; $\mathrm{BF}_{c,h}$, local Bayes factor in favor of $H_{1,ch}$, $(\overline{\beta}_-, \overline{\mathrm{BF}}_-)$, average coefficients and BFs for negative SNPs.

*Preferred model estimator of coefficients (0 if posterior probability inclusion is less then .5).

Downloaded by [NIH Library] at 06:55 26 December 2013

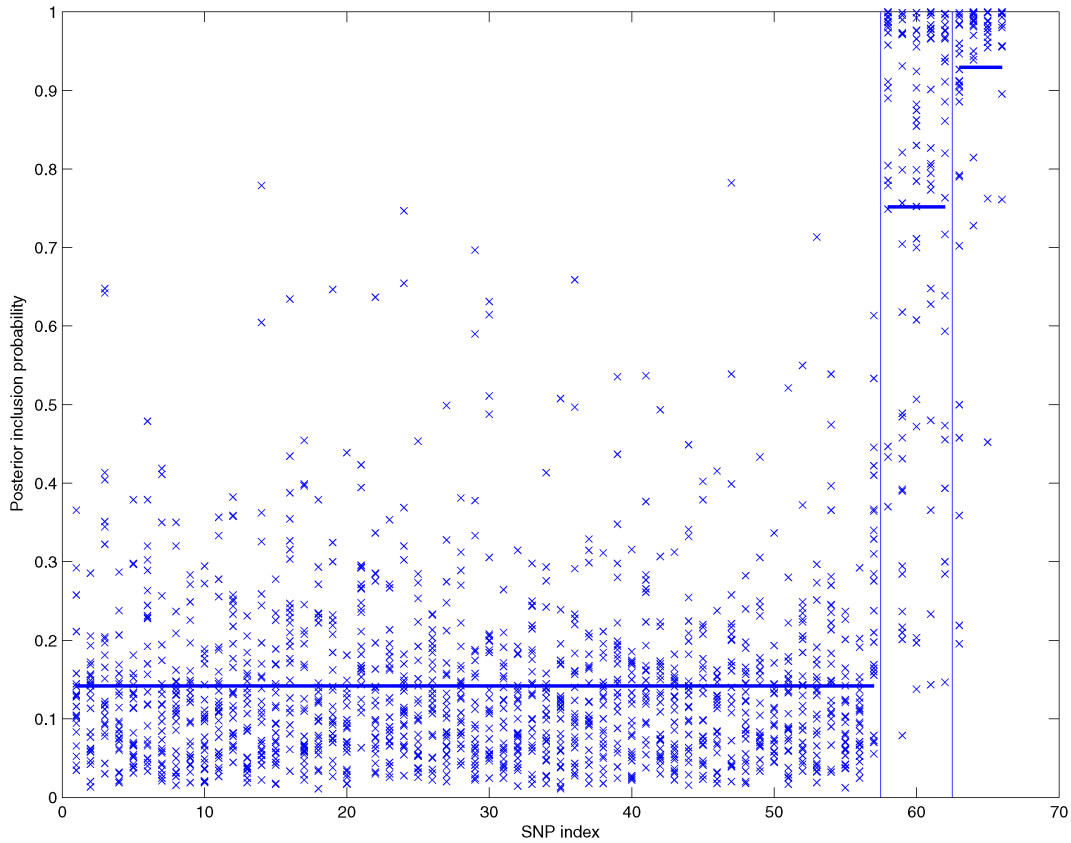

Figure 3. Posterior inclusion probabilities for each SNP in the simulation case (iv). SNPs 1–57 are null, SNPs 58–62 have a coefficient of 1, and SNPs 63–66 have a coefficient of 1.5.

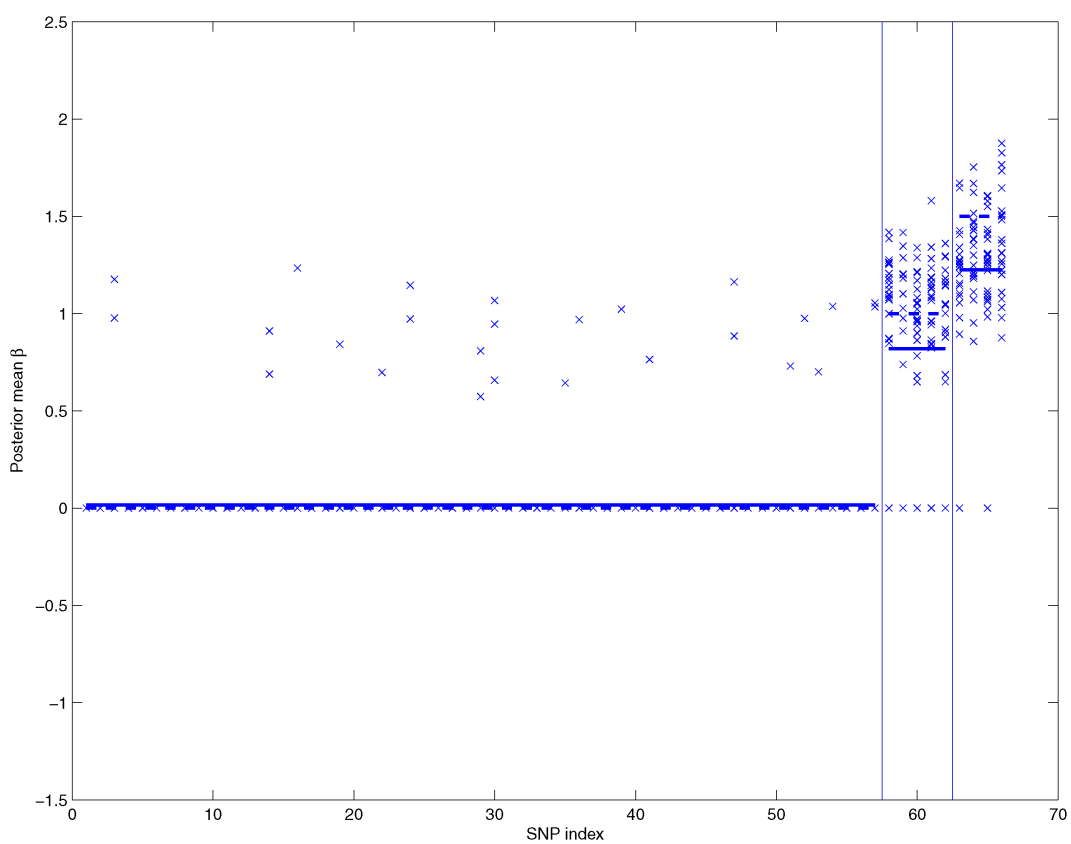

Figure 4. Preferred model posterior means (solid lines) and true values (dashed lines) of the coefficients within each cluster in simulation case (iv).

Downloaded by [NIH Library] at 06:55 26 December 2013

Downloaded by [NIH Library] at 06:55 26 December 2013

clusion probability across SNPs in a cluster and across datasets. Clearly, the approach tends to assign substantially higher inclusion probabilities to the SNPs in the two nonnull clusters. Figure 4 shows the preferred model posterior means for β_{clh} for all c, l, h across the different simulated datasets. On average, the null SNPs have estimated coefficients close to 0, while the nonnull SNPs have estimated coefficients close to the true value.

6. DISCUSSION

This article has proposed a semiparametric Bayesian approach for simultaneous variable selection and clustering in applications involving many related predictors. There is a rapidly expanding literature on methods for identifying important predictors from an extremely high-dimensional set of candidates, primarily motivated by gene expression data (Efron, Tibshirani, Storey, and Tusher 2001; Newton, Kendziorski, Richmond, Blattner, and Tsui 2001; Ibrahim, Chen, and Gray 2002; among many others). Our motivation is somewhat different in that we are interested in more focused genetic studies that collect genotype data at a moderate number of locations (e.g., 30+), corresponding to regulatory or coding regions for functionally related genes. Such studies are potentially conducted as a second stage after preliminary identification of promising candidate genes through gene expression studies. Because we have a more modest number of predictors, we can be more ambitious in attempting to address questions about overall significance and clustering of effect sizes. Our method should also be useful in epidemiologic studies collecting information for environmental exposures, such as pesticides or nutrients, that can be grouped into prespecified classes.

Our motivation was genetic epidemiology studies in which investigators preselect SNPs based on presumed functionality judged from the literature. This extremely common strategy tends to limit the number of SNPs that need to be genotyped, but can potentially misrepresent variation within a gene. Technological advances now allow one to use a dense collection of *tagSNPs* that may have no function in themselves but are instead markers of variability within a gene. TagSNPs can be selected to be approximately evenly spaced across a gene, or they can be selected on the basis of estimates of linkage disequilibrium (LD), which will result in a denser set of markers covering areas of low LD. Our proposed method is promising as an approach for identifying regions of a gene that may contain a functional SNP(s) from a field of anonymous tagSNPs. Otherwise, by relying on preselection of a small number of SNPs, there is always the possibility that important variability exists at other locations. Hence, inferences are necessarily limited by the SNPs chosen, and one cannot make general conclusions about the importance of a particular gene in predicting a health outcome.

We have proposed a particular strategy of prior elicitation that treats the different SNPs as exchangeable within a gene, while also treating the genes as exchangeable. Although this is a reasonable default strategy for many studies, in certain cases there may be information available to suggest that certain genes and SNPs are particularly promising candidates, while little or no information is available for others. In such cases, as noted by Wacholder, Chanock, Garcia-Closas, El Ghormli, and Rothman (2004), the exchangeability assumption is implausible. Fortunately, it is straightforward to modify our procedure to allow the prior probabilities of inclusion to vary for the different genes and SNPs under study. In future work, it will be interesting to extend the methodology to accommodate missing genotype data.

[Received August 2005. Revised January 2007.]

REFERENCES

Abramowitz, M., and Stegun, I. A. (1964), *Handbook of Mathematical Functions*, Washington: National Bureau of Standards.

Antoniak, C. E. (1974), "Mixtures of Dirichlet Processes With Application to Nonparametric Problems," *The Annals of Statistics*, 2, 1152–1174.

Berry, D. A., and Hochberg, Y. (1999), "Bayesian Perspectives on Multiple Comparisons," *Journal of Statistical Planning and Inference*, 82, 215–227.

Berry, S. M., and Berry, D. A. (2004), "Accounting for Multiplicities in Assessing Drug Safety: A Three Level Hierarchical Mixture Model," *Biometrics*, 60, 418–426.

Carvalho, C., Lucas, J., Wang, Q., Chang, J., Nevins, J., and West, M. (2005), "High-Dimensional Sparse Factor Models and Latent Factor Regression," ISDS Discussion Paper 2005-14, Duke University.

Clyde, M., and George, E. I. (2004), "Model Uncertainty," *Statistical Science*, 19, 81–94.

Coull, B. A., Hobert, J. P., Ryan, L. M., and Holmes, L. B. (2001), "Crossed Random Effect Models for Multiple Outcomes in a Study of Teratogenesis," *Journal of the American Statistical Association*, 96, 1194–1204.

De Iorio, M., Müller, P., Rosner, G. L., and MacEachern, S. N. (2004), "An ANOVA Model for Dependent Random Measures," *Journal of the American Statistical Association*, 99, 205–215.

De Roos, A. J., Poole, C., Teschke, K., and Olshan, A. F. (2001), "An Application of Hierarchical Regression in the Investigation of Multiple Paternal Occupational Exposures and Neuroblastoma in Offspring," *American Journal of Industrial Medicine*, 39, 477–486.

Dunson, D. B. (2005), "Bayesian Semiparametric Isotonic Regression for Count Data," *Journal of the American Statistical Association*, 100, 618–627.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. G. (2001), "Empirical Bayes Analysis of a Microarray Experiment," *Journal of the American Statistical Association*, 96, 1151–1160.

Escobar, M. D., and West, M. (1995), "Bayesian Density-Estimation and Inference Using Mixtures," *Journal of the American Statistical Association*, 90, 577–588.

Ferguson, T. S. (1973), "A Bayesian Analysis of Some Nonparametric Problems," *The Annals of Statistics*, 1, 209–230.

——— (1974), "Prior Distributions on Spaces of Probability Measures," *The Annals of Statistics*, 2, 615–629.

George, E. I., and McCulloch, R. E. (1997), "Approaches for Bayesian Variable Selection," *Statistica Sinica*, 7, 339–373.

Gonen, M., Westfall, P. H., and Johnson, W. O. (2003), "Bayesian Multiple Testing for Two-Sample Multivariate Endpoints," *Biometrics*, 59, 76–82.

Gopalan, R., and Berry, D. A. (1998), "Bayesian Multiple Comparisons Using Dirichlet Process Priors," *Journal of the American Statistical Association*, 93, 1130–1139.

Greenland, S. (1992), "A Semi-Bayes Approach to the Analysis of Correlated Multiple Associations, With an Application to an Occupational Cancer-Mortality Study," *Statistics in Medicine*, 11, 219–230.

——— (1993), "Methods for Epidemiologic Analyses of Multiple Exposures—A Review and Comparative-Study of Maximum-Likelihood, Preliminary-Testing, and Empirical-Bayes Regression," *Statistics in Medicine*, 12, 717–736.

Hung, R. J., Brennan, P., Malaveille, C., Porru, S., Donato, F., Boffetta, P., and Witte, J. S. (2004), "Using Hierarchical Modeling in Genetic Association Studies With Multiple Markers: Application to a Case-Control Study of Bladder Cancer," *Cancer Epidemiology Biomarkers & Prevention*, 13, 1013–1021.

Ibrahim, J. G., Chen, M.-H., and Gray, R. J. (2002), "Bayesian Models for Gene Expression With DNA Microarray Data," *Journal of the American Statistical Association*, 97, 88–99.

Lucas, J., Carvalho, C., Wang, Q., Bild, A., Nevins, J. R., and West, M. (2006), "Sparse Statistical Modelling in Gene Expression Genomics," in *Bayesian Inference for Gene Expression and Proteomics*, eds. K.-A. Do, P. Müller, and M. Vannucci, Cambridge University Press, pp. 155–176.

MacEachern, S. N. (1994), "Estimating Normal Means With a Conjugate Style Dirichlet Process Prior," *Communications in Statistics: Simulation and Computation*, 23, 727–741.

——— (1999), "Dependent Nonparametric Processes," in *ASA Proceedings of the Section on Bayesian Statistical Science*, Alexandria, VA: American Statistical Association.

——— (2000), "Dependent Dirichlet Processes," unpublished manuscript, Ohio State University, Dept. of Statistics.

MacEachern, S. N., Kottas, A., and Gelfand, A. E. (2001), "Spatial Nonparametric Bayesian Models," in *ASA Proceedings of the Joint Statistical Meetings*.

Meng, C. Y. K., and Dempster, A. P. (1987), "A Bayesian Approach to the Multiplicity Problem for Significance Testing With Binomial Data," *Biometrics*, 43, 301–311.

Mulherin Engel, S. A., Ericksen, H. C., Savitz, D. A., Thorp, J., Chanock, S. J., and Olshan, A. F. (2005a), "Risk of Spontaneous Preterm Birth Is Associated With Common Proinflammatory Cytokine Polymorphisms," *Epidemiology*, 16, 469–477.

Mulherin Engel, S. A., Olshan, A. F., Savitz, D. A., Thorp, J., Ericksen, H. C., and Chanock, S. J. (2005b), "Risk of Small-for-Gestational Age Is Associated With Common Anti-Inflammatory Cytokine Polymorphisms," *Epidemiology*, 16, 478–486.

Müller, P., Quintana, F., and Rosner, G. (2004), "A Method for Combining Inference Across Related Nonparametric Bayesian Models," *Journal of the Royal Statistical Society*, Ser. B, 66, 735–749.

Newton, M. A., Kendziorski, C. M., Richmond, C. C., Blattner, F. R., and Tsui, K. W. (2001), "On Differential Variability of Expression Ratios: Improving Statistical Inference About Gene Expression Changes From Microarray Data," *Journal of Computational Biology*, 8, 37–52.

O'Brien, S. M., and Dunson, D. B. (2004), "Bayesian Multivariate Logistic Regression," *Biometrics*, 60, 739–746.

Savitz, D. A., Dole, N., Williams, J., et al. (1999), "Determinants of Participation in an Epidemiological Study of Preterm Delivery," *Paediatric and Perinatal Epidemiology*, 13, 114–125.

Sethuraman, J. (1994), "A Constructive Definition of the Dirichlet Process Prior," *Statistica Sinica*, 2, 639–650.

Tadesse, M. G., Sha, N., and Vannucci, M. (2005), "Bayesian Variable Selection in Clustering High-Dimensional Data," *Journal of the American Statistical Association*, 100, 602–617.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006), "Hierarchical Dirichlet Processes," *Journal of the American Statistical Association*, 101, 1566–1581.

Thomas, D. C., Siemiatycki, J., Dewar, R., Robins, J., Goldberg, M., and Armstrong, B. G. (1985), "The Problem of Multiple Inference in Studies Designed to Generate Hypotheses," *American Journal of Epidemiology*, 122, 1080–1095.

Wacholder, S., Chanock, S., Garcia-Closas, M., El Ghormli, L., and Rothman, N. (2004), "Assessing the Probability That a Positive Report Is False: An Approach for Molecular Epidemiology Studies," *Journal of the National Cancer Institute*, 96, 434–442.

West, M., Müller, P., and Escobar, M. D. (1994), "Hierarchical Priors and Mixture Models, With Applications in Regression and Density Estimation," in *A Tribute to D. V. Lindley*, eds. A. F. M. Smith and P. R. Freeman, New York: Wiley.

Westfall, P. H., Johnson, W. O., and Utts, J. M. (1997), "A Bayesian Perspective on the Bonferroni Adjustment," *Biometrika*, 84, 419–427.

Ziv, E., and Burchard, E. G. (2003), "Human Population Structure and Genetic Association Studies," *Pharmacogenomics*, 4, 431–441.

Downloaded by [NIH Library] at 06:55 26 December 2013