The Annals of Applied Statistics
2011, Vol. 5, No. 3, 1780–1815
DOI: 10.1214/11-AOAS455

BAYESIAN VARIABLE SELECTION REGRESSION FOR GENOME-WIDE ASSOCIATION STUDIES AND OTHER LARGE-SCALE PROBLEMS

BY YONGTAO GUAN AND MATTHEW STEPHENS[1]

University of Chicago

We consider applying Bayesian Variable Selection Regression, or BVSR, to genome-wide association studies and similar large-scale regression problems. Currently, typical genome-wide association studies measure hundreds of thousands, or millions, of genetic variants (SNPs), in thousands or tens of thousands of individuals, and attempt to identify regions harboring SNPs that affect some phenotype or outcome of interest. This goal can naturally be cast as a variable selection regression problem, with the SNPs as the covariates in the regression. Characteristic features of genome-wide association studies include the following: (i) a focus primarily on identifying relevant variables, rather than on prediction; and (ii) many relevant covariates may have tiny effects, making it effectively impossible to confidently identify the complete "correct" subset of variables. Taken together, these factors put a premium on having interpretable measures of confidence for individual covariates being included in the model, which we argue is a strength of BVSR compared with alternatives such as penalized regression methods. Here we focus primarily on analysis of quantitative phenotypes, and on appropriate prior specification for BVSR in this setting, emphasizing the idea of considering what the priors imply about the total proportion of variance in outcome explained by relevant covariates. We also emphasize the potential for BVSR to estimate this proportion of variance explained, and hence shed light on the issue of "missing heritability" in genome-wide association studies. More generally, we demonstrate that, despite the apparent computational challenges, BVSR can provide useful inferences in these large-scale problems, and in our simulations produces better power and predictive performance compared with standard single-SNP analyses and the penalized regression method LASSO. Methods described here are implemented in a software package, pi-MASS, available from the Guan Lab website http://bcm.edu/cnrc/mcmcmc/pimass.

1. Introduction. The problem of identifying relevant covariates in a regression model, sometimes known as variable selection, arises frequently in many fields. As computational and data-collection technologies have developed, the number of covariates typically measured in these kinds of problems has steadily increased, and it is now not unusual to come across data sets involving many thou-

Received December 2009; revised December 2010.
[1]Supported by NIH Grants HG02585 and HL084689.
Key words and phrases. Bayesian regression, variable selection, shrinkage, genome-wide, association study, multi-SNP analysis, heritability.

sands or millions of covariates. Here we consider one particular setting where data sets of this size are common: genome-wide association studies (GWAS).

Current typical GWAS [e.g., Wellcome Trust Case Control Consortium (2007)] measure hundreds of thousands, or millions, of genetic variants (typically Single Nucleotide Polymorphisms, or SNPs), in hundreds, thousands, or tens of thousands of individuals, with the primary goal being to identify which regions of the genome harbor SNPs that affect some phenotype or outcome of interest. While many GWAS are case-control studies, here we focus primarily on the computationally-simpler setting where a continuous phenotype has been measured on population-based samples, before briefly considering the challenges of extending these methods to binary outcomes.

Most existing GWAS analyses are "single-SNP" analyses, which simply test each SNP, one at a time, for association with the phenotype. Strong associations between a SNP and the phenotype are interpreted as indicating that SNP, or a nearby correlated SNP, likely affects phenotype. The primary rationale for GWAS is the idea that, by examining these SNPs in more detail—for example, examining which genes they are in or near—we may glean important insights into the biology of the phenotype under study.

In this paper we examine the potential to apply Bayesian Variable Selection Regression (BVSR) to GWAS (or other similar large-scale problems). Variable selection regression provides a very natural approach to analyzing GWAS: the phenotype is treated as the regression response, SNPs become regression covariates, and the goal of identifying genomic regions likely to harbor SNPs affecting phenotype is accomplished by examining the genomic locations of SNPs deemed likely to have nonzero regression coefficients. However, BVSR requires the use of computationally-intensive Markov chain Monte Carlo (MCMC) algorithms, and, prior to performing this work, it was unclear to us whether such algorithms could produce reliable results in a practical time-frame for problems as large as a typical GWAS. One important contribution of this paper is to show that, even using relatively simple MCMC algorithms, BVSR can indeed produce useful inferences in problems of this size. Another important contribution is to discuss *how* BVSR should be used for GWAS analysis, with particular focus on choice of appropriate prior distribution. Further, and perhaps most importantly, we give reasons *why* one might want to use BVSR to analyze GWAS—rather than less computationally-demanding approaches such as single-SNP analyses, or penalized regression approaches such as LASSO [Tibshirani (1996)]—by emphasizing qualitative advantages of BVSR in this context. In particular, we emphasize that, unlike penalized regression approaches, BVSR naturally produces easily-interpretable measures of confidence—specifically, posterior probabilities—that individual covariates have nonzero regression coefficients. This is a particularly important advantage in GWAS because the primary goal of the analysis is to identify such covariates, and to use these identifications to learn about underlying biology (in contrast to other settings where prediction may be the primary goal).

Although our work is motivated by GWAS, many of the ideas and results should be of more general interest. In brief, the key elements are as follows:

- We demonstrate that BVSR can be practical for large problems involving hundreds of thousands of covariates and thousands of observations.
- We introduce some new ideas for prior specification in BVSR. In particular, we emphasize the benefits of focusing on what the priors imply about the total proportion of variance in response explained by relevant covariates (henceforth abbreviated as PVE). We note that standard approaches to prior specification in BVSR, which put the same priors on the regression coefficients irrespective of how many covariates are included in the model, imply that models with many relevant covariates are likely to have much larger PVE than models with few relevant covariates. We propose a simple alternative prior that does not make this potentially undesirable assumption, and has the intuitively appealing property that it applies stronger shrinkage in more complex models (i.e., models with more relevant covariates).
- We emphasize the potential for BVSR to estimate the total amount of signal in a data set, specifically the PVE, even when there is insufficient information to reliably identify all relevant covariates. As a result, BVSR has the potential to shed light on the so-called "missing heritability" observed in many GWAS [Maher (2008); Yang et al. (2010)].
- We compare and contrast BVSR with a penalized-regression approach, the LASSO [Tibshirani (1996)]. Despite the considerable literature on both BVSR and penalized regression, there exist few comparisons (either qualitative or quantitative) of these two approaches. We chose the LASSO as a representative of penalized regression approaches both because of its popularity and because previous papers have applied it to the specific context of GWAS [e.g., Hoggart et al. (2008); Wu et al. (2009)]. In our limited simulation study BVSR outperforms LASSO in terms of predictive performance. In addition, we emphasize the qualitative advantage of BVSR over LASSO, and other penalized regression methods, that it produces posterior probabilities for each covariate having a nonzero regression coefficients. This qualitative advantage seems more fundamental, since predictive performance of different methods may vary depending on the underlying assumptions.

The remainder of the paper is organized as follows. In Section 2 we describe BVSR and our choice of priors. In Section 3 we discuss computation and inference, including Markov chain Monte Carlo algorithms used, and a Rao–Blackwellization approach to estimating the marginal posterior inclusion probability for each covariate. Section 4 reviews our main goals in applying BVSR to GWAS. In Section 5 we examine, through simulations, the effectiveness of BVSR for various tasks, including estimating the PVE, prediction, and identifying relevant covariates. For some of these tasks we compare BVSR with LASSO and single-SNP analyses. We also illustrate BVSR on a GWAS for C-reactive protein.

In Section 6 we briefly consider the challenges of extending our methods to deal with binary phenotypes. Finally, in Section 7 we discuss some limitations and pitfalls of BVSR as we have applied it in this context, and potential future directions.

2. Models and priors. This section introduces notation and specifies the details of our BVSR model and priors used. Our formulation up to Section 2.1 is in the same vein as much previous work on BVSR, but with particular emphasis on putting priors on hyperparameters that are often considered fixed and known. Key relevant references include Mitchell and Beauchamp (1988), George and McCulloch (1993), Smith and Kohn (1996), Raftery, Madigan and Hoeting (1997) and Brown, Vannucci and Fearn (2002); see also Miller (2002) and O'Hara and Sillanpää (2009) for more background and references.

We consider the standard normal linear regression

$$\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\beta}, X, \tau \sim N_n(\boldsymbol{\mu} + X\boldsymbol{\beta}, \tau^{-1} I_n), \tag{2.1}$$

relating a response variable $\mathbf{y}$ to covariates X. Here $\mathbf{y}$ is an n-vector of observations on n individuals, $\boldsymbol{\mu}$ is an n-vector with components all equal to the same scalar μ, X is an n by p matrix of covariates, $\boldsymbol{\beta}$ is a p-vector of regression coefficients, τ denotes the inverse variance of the residual errors, $N_n(\cdot, \cdot)$ denotes the n-dimensional multivariate normal distribution and I_n the n by n identity matrix. The variables $\mathbf{y}$ and X are observed, whereas $\boldsymbol{\mu}$, $\boldsymbol{\beta}$, and τ are parameters to be inferred. In more detail, $\mathbf{y} = (y_1, \ldots, y_n)$, where y_i is the measured response on individual i, and $X = (\mathbf{x}_{\cdot 1}, \ldots, \mathbf{x}_{\cdot p})$, where $\mathbf{x}_{\cdot j} = (x_{1j}, \ldots, x_{nj})^T$ is a column vector containing the observed values of the jth covariate. For example, in the context of a GWAS, y_i is the measured phenotype of interest in individual i, and x_{ij} is the genotype of individual i at SNP j, typically coded as 0, 1 or 2 copies of a particular reference allele. [By coding the genotypes as 0, 1, or 2, we are assuming an additive genetic model. It would be straightforward to include dominant and recessive effects by adding another covariate for each SNP, as in Servin and Stephens (2007), e.g., although this would increase computational cost.]

In many contexts, including GWAS, the number of covariates is very large—and, in particular, $p \gg n$—but only a small subset of the covariates are expected to be associated with the response (i.e., have nonzero β_j). Indeed, the main goal of GWAS is to identify these relevant covariates. To this end, we define a vector of binary indicators $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p) \in \{0, 1\}^p$ that indicate which elements of β are nonzero. Thus,

$$\mathbf{y}|\boldsymbol{\gamma}, \mu, \tau, \boldsymbol{\beta}, X \sim N_n(\boldsymbol{\mu} + X_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \tau^{-1} I_n), \tag{2.2}$$

where $X_{\boldsymbol{\gamma}}$ denotes the design matrix X restricted to those columns j for which $\gamma_j = 1$, and $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ denotes a corresponding vector of regression coefficients. In general, for observational studies one would be reluctant to conclude any causal interpretation for $\boldsymbol{\gamma}$, but in the context of GWAS, it is usually reasonable to interpret

$\gamma_j = 1$ as indicating that SNP j, or an unmeasured SNP correlated with SNP j, has a causal (functional) affect on $\mathbf{y}$. This is because in GWAS reverse causation is generally implausible (phenotypes cannot causally affect genotype, since genotype comes first temporally), and there are few potential unmeasured confounders that could affect both genotype and $\mathbf{y}$ [Smith and Ebrahim (2003)]. A well-documented exception to this is population structure; here we assume that this has been corrected for prior to analysis, for example, by letting $\mathbf{y}$ be the residuals from regressing the observed phenotype values against measures of population structure, obtained, for example, by model-based clustering [Pritchard et al. (2000)] or principal components analysis [Price et al. (2006)].

Taking a Bayesian approach to inference, we put priors on the parameters:

$$\tau \sim \text{Gamma}(\lambda/2, \kappa/2), \tag{2.3}$$

$$\mu|\tau \sim N(0, \sigma_\mu^2/\tau), \tag{2.4}$$

$$\gamma_j \sim \text{Bernoulli}(\pi), \tag{2.5}$$

$$\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\tau, \boldsymbol{\gamma} \sim N_{|\boldsymbol{\gamma}|}\big(0, (\sigma_a^2/\tau) I_{|\boldsymbol{\gamma}|}\big), \tag{2.6}$$

$$\boldsymbol{\beta}_{-\boldsymbol{\gamma}}|\boldsymbol{\gamma} \sim \delta_0, \tag{2.7}$$

where $|\boldsymbol{\gamma}| := \sum_j \gamma_j$, $\boldsymbol{\beta}_{-\boldsymbol{\gamma}}$ denotes the vector of $\boldsymbol{\beta}$ coefficients for which $\gamma_j = 0$, and δ_0 denotes a point mass on 0. Here $\pi, \sigma_a, \lambda, \kappa$, and σ_μ are hyperparameters. The hyperparameters π and σ_a have important roles, with π reflecting the sparsity of the model, and σ_a reflecting the typical size of the nonzero regression coefficients. Rather than setting these hyperparameters to prespecified values, we place priors on them, hence allowing their values to be informed by the data; the priors used are detailed below. (Later we will argue that this ability to infer π and σ_a from the data is an important advantage of analyzing all SNPs simultaneously, rather than one at a time.) The remaining hyperparameters are less critical, and, in practice, we consider the posterior distributions for which $\sigma_\mu^2 \to \infty$ and $\nu, \kappa \to 0$, which has the attractive property that the resulting relative marginal likelihoods for $\boldsymbol{\gamma}$ are invariant to shifting or scaling of $\mathbf{y}$. Thus, for example, inference of which genetic variants are associated with height would be unaffected by whether height is measured in meters or inches. [Taking these limits is effectively equivalent to using the improper prior $p(\mu, \tau) \propto 1/\tau$, but we prefer to formulate proper priors and take limits in their posteriors, to verify sensible limiting behavior.]

The parameter π controls the sparsity of the model, and where the appropriate level of sparsity is uncertain a priori, as is typically the case, it seems important to specify a prior for π rather than fixing it to an arbitrary value. In GWAS, and probably in many other settings with extreme sparsity, uncertainty in π may span orders of magnitude: for example, there could be just a few relevant covariates or hundreds. In this case a uniform prior on π seems inappropriate, since this would inevitably place most of the prior mass on larger numbers of covariates

(e.g., uniform on 10^{-5} to 10^{-3} puts about 90% probability on $>10^{-4}$). Instead, we put a uniform prior on $\log \pi$:

$$\log \pi \sim U(a, b), \tag{2.8}$$

where $a = \log(1/p)$ and $b = \log(M/p)$, so the lower and upper limits on π correspond, respectively, to an expectation of 1 and M covariates in the model. In applications here we used $M = 400$, with this arbitrary limit being imposed partly due to computational considerations (larger M can increase computing time considerably). The assumption of a uniform distribution is, of course, somewhat artificial but has the merit of being easily interpretable. An alternative, which may be preferable in some settings, would be a normal prior on $\log(\pi/(1-\pi))$.

The above formulation, with the exception of our slightly nonstandard prior on π, follows previous work. However, since many formulations of BVSR differ slightly from one another, we now comment on some of the choices we made:

(1) We chose, in (2.6), to put independent priors on the elements of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$. An alternative common choice is Zellner's g-prior [Zellner (1986); Agliari and Parisetti (1988)], which assumes correlations among the regression coefficients mimicking the correlations among covariates,

$$\boldsymbol{\beta}_{\boldsymbol{\gamma}} \sim N_{|\boldsymbol{\gamma}|}\left(0, \frac{g}{\tau} X_{\boldsymbol{\gamma}}^t X_{\boldsymbol{\gamma}}\right).$$

For GWAS we prefer the independent priors because we view the $\boldsymbol{\beta}$'s as reflecting causal effects of X on $\mathbf{y}$, and there seems to be no good reason to believe that the correlation structure of causal effects will follow that of the SNPs.

(2) Some authors center each of the vectors $\mathbf{y}$ and $\mathbf{x}_{\cdot 1}, \ldots, \mathbf{x}_{\cdot p}$ to have mean 0, and set $\mu = 0$. This approach yields the same posterior on $\boldsymbol{\gamma}$ as our limiting prior on μ (derivation omitted), and simplifies calculations, and so we use it henceforth.

(3) It is common in variable selection problems to scale the covariates $\mathbf{x}_{\cdot 1}, \ldots, \mathbf{x}_{\cdot p}$ to each have unit variance, to avoid problems due to different variables being measured on different scales. In GWAS these covariates are measured on the same scale, being counts of the reference allele, and so we do not scale the covariates in this way in our examples. However, one could so scale them, which would correspond to a prior assumption that SNPs with less variable genotypes (i.e., those with a lower minor allele frequency) have larger effect sizes; see Wakefield (2009) for relevant discussion.

(4) The priors assume that the β_j are exchangeable, and, in particular, that all covariates are, a priori, equally plausible candidates to affect outcome $\mathbf{y}$. In the context of a GWAS, this assumption means we are ignoring information that might make some SNPs better candidates for affecting outcome than others. Our priors also ignore the fact that functional SNPs may tend to cluster near one another in the genome. These choices were made purely for simplicity; one attractive feature of BVSR is that one could modify the priors to incorporate these types of information, but we leave this to future work.

(5) Some formulations of BVSR use a similar "sparse" prior, where the marginal prior on β_j is a mixture of a point mass at 0 and a normal distribution, whereas others [e.g., George and McCulloch (1993)] instead use a mixture of two normal distributions, one with a substantially larger variance than the other. The sparse formulation seems computationally advantageous in large problems because sparsity facilitates certain operations (e.g., integrating out $\boldsymbol{\beta}$ given $\boldsymbol{\gamma}$).

2.1. *Novel prior on σ_a^2.* While the above formulation is essentially standard and widely used, there is considerable variability in how different authors treat the hyperparameter σ_a. Some fix it to an essentially arbitrary value, while others put a prior on this parameter. Several different priors have been suggested, and the lack of consensus among different authors may reflect the fact that most of them seem not to have been given a compelling motivation or interpretation. Here we suggest a way of thinking about this prior that we believe aids interpretation, and hence appropriate prior specification. Specifically, we suggest focusing on what the prior implies about the proportion of variance in $\mathbf{y}$ explained by $X_{\boldsymbol{\gamma}}$ (the PVE). For example, almost all priors we have seen previously in this context assume independence of π and σ_a, which implies independence of $\boldsymbol{\gamma}$ and σ_a. While this assumption may seem natural initially, it implies that more complex models are expected to have substantially higher PVE. In our application this assumption does not capture our prior beliefs. For example, it seems quite plausible a priori that there could be either a large number of relevant covariates with small PVE, or a small number of covariates with large PVE.

Here we suggest specifying a prior on σ_a^2 given $\boldsymbol{\gamma}$ by considering the induced prior on the PVE, and, in particular, by making this induced prior relatively flat in the range of $(0, 1)$. To formalize this, let $V(\boldsymbol{\beta}, \tau)$ denote the empirical variance of $X\boldsymbol{\beta}$ relative to the residual variance τ^{-1}:

$$V(\boldsymbol{\beta}, \tau) := \frac{1}{n} \sum_{i=1}^{n} [(X\boldsymbol{\beta})_i]^2 \tau, \tag{2.9}$$

where this expression for the variance assumes that the covariates have been centered, and so $X\boldsymbol{\beta}$ has mean 0. Then the total proportion of variance in $\mathbf{y}$ explained by X if the true values of the regression coefficients are $\boldsymbol{\beta}$ is given by

$$\mathrm{PVE}(\boldsymbol{\beta}, \tau) := V(\boldsymbol{\beta}, \tau)/\big(1 + V(\boldsymbol{\beta}, \tau)\big). \tag{2.10}$$

Our aim is to choose a prior on $\boldsymbol{\beta}$ given τ so that the induced prior on $\mathrm{PVE}(\boldsymbol{\beta}, \tau)$ is approximately uniform. To do this, we exploit the fact that the expected value of $V(\boldsymbol{\beta}, \tau)$ (with expectation being taken over $\boldsymbol{\beta}|\tau$) depends in a simple way on σ_a:

$$v(\boldsymbol{\gamma}, \sigma_a) := E[V(\boldsymbol{\beta}, \tau)|\boldsymbol{\gamma}, \sigma_a, \tau] = \sigma_a^2 \sum_{j\,:\,\gamma_j = 1} s_j, \tag{2.11}$$

where $s_j = \frac{1}{n}\sum_{i=1}^n x_{ij}^2$ is the variance of covariate j. Define

$$h(\boldsymbol{\gamma}, \sigma_a) = v(\boldsymbol{\gamma}, \sigma_a)/\big(1 + v(\boldsymbol{\gamma}, \sigma_a)\big). \tag{2.12}$$

Intuitively, h gives a rough guide to the expectation of PVE for a given value of $\boldsymbol{\gamma}$ and σ_a. (It is not precisely the expectation since it is the ratio of expectations, rather than the expectation of the ratio.) To accomplish our goal of putting approximately uniform prior on PVE, we specify a uniform prior on h, independent of $\boldsymbol{\gamma}$, which induces a prior on σ_a given $\boldsymbol{\gamma}$ via the relationship

$$\sigma_a^2(h, \gamma) = \frac{h}{1-h}\frac{1}{\sum_{j:\gamma_j=1} s_j}. \tag{2.13}$$

In all our MCMC computations, we parameterize our model in terms of $(h, \boldsymbol{\gamma})$, rather than $(\sigma_a, \boldsymbol{\gamma})$. Note that the induced prior on σ_a^2 is diffuse: if $Z = h/(1-h)$, and $h \sim U(0,1)$, then Z has a probability density function $f(z) = 1/(1+z)^2$, which is heavy tailed.

Our prior on σ_a has interesting connections with the prior suggested by Liang et al. (2008). While Liang et al. (2008) use a g prior, if we consider the case where the covariates are orthogonal with variances $s_j = 1$, then their parameter g is effectively equivalent to our $n\sigma_a^2$. They suggest putting a Beta$(1, a/2 - 1)$ prior on $g/(1+g)$, with $a = 3$ or 4; the case $a = 4$ is uniform on $g/(1+g)$, or in our notation uniform on $n\sigma_a^2/(1+n\sigma_a^2)$. In contrast, our prior is uniform on $|\boldsymbol{\gamma}|\sigma_a^2/(1+|\boldsymbol{\gamma}|\sigma_a^2)$. Thus, our σ_a is effectively $n/|\boldsymbol{\gamma}|$ times the value of σ_a from Liang et al. (2008), and so our σ_a is larger than theirs (implying less shrinkage), provided that the number of relevant covariates $|\boldsymbol{\gamma}|$ is less than n. Qualitatively, perhaps the main difference between the priors is that our prior applies less shrinkage (larger σ_a) in simpler models, which seems intuitively appealing.

Of course, choice of appropriate prior distributions may vary according to context, and we do not argue that the prior used here is universally superior to other choices. However, we do believe the priors outlined above are suitable for general use in most GWAS applications. In addition, we emphasize that these priors incorporate two principles that we believe should be helpful more generally: first, it seems preferable to place prior distributions on the hyperparameters π and σ_a, rather than fixing them to specific values, as this provides the potential to learn about them from the data; second, when comparing priors for σ_a, it is helpful to consider what the priors imply about PVE.

3. Computation and inference. We use Markov chain Monte Carlo to obtain samples from the posterior distribution of $(h, \pi, \boldsymbol{\gamma})$ on the product space $(0,1) \times (0,1) \times \{0,1\}^p$, which is given by

$$\mathsf{p}(h, \pi, \boldsymbol{\gamma}|\mathbf{y}) \propto \mathsf{p}(\mathbf{y}|h, \boldsymbol{\gamma})\mathsf{p}(h)\mathsf{p}(\boldsymbol{\gamma}|\pi)\mathsf{p}(\pi). \tag{3.1}$$

Here we are exploiting the fact that the parameters $\boldsymbol{\beta}$ and τ can be integrated out analytically to compute the marginal likelihood $\mathsf{p}(\mathbf{y}|h, \boldsymbol{\gamma})$. For each sampled value

of $h, \boldsymbol{\gamma}$ from this posterior, we also obtain samples from the posterior distributions of $\boldsymbol{\beta}$ and τ by sampling from their conditional distributions given $\mathbf{y}, \boldsymbol{\gamma}, h$.

Our Markov chain Monte Carlo algorithm for sampling $h, \pi, \boldsymbol{\gamma}$ is detailed in Appendix A. In brief, it is a Metropolis–Hastings algorithm [Metropolis et al. (1953); Hastings (1970)], using a simple local proposal to jointly update $h, \pi, \boldsymbol{\gamma}$. In particular, it explores the space of covariates included in the model, $\boldsymbol{\gamma}$, by proposing to add, remove, or switch single covariates in and out of the model. To improve computational performance, we use three strategies. First, in addition to the local proposal moves, we sometimes make a longer-range proposal by compounding randomly many local moves. This technique, named "small-world proposal," improves the theoretical convergence rate of the MCMC scheme [Guan and Krone (2007)]. Second, and perhaps more importantly, when proposing new values for $\boldsymbol{\gamma}$, and specifically when proposing to add a variable to the model, we focus more attention on those covariates with the strongest *marginal* associations. This idea is related to the sure independence screen [Fan and Lv (2008)] (SIS), which uses marginal associations as an initial screening step. However, it is a "softer" use of these marginal associations than the SIS, because every variable continues to have positive probability of being proposed. Simulations (not shown) show that taking account of the marginal associations in this way dramatically increases the acceptance rate compared to a proposal that ignores the marginal associations. Finally, when estimating quantities of interest, we make use where possible of Rao–Blackwellization techniques [Casella and Robert (1996)], detailed below, to reduce Monte Carlo variance.

We note that our computational scheme is relatively simple, and one can create data sets where it will perform poorly, for example, multiple correlated covariates that are far apart along a chromosome, where an efficient algorithm would require careful joint updates of the γ_i values for those correlated covariates. (In our current implementation, swap proposals only apply to SNPs that are close to one another in the genome, which is motivated by the fact that correlations decay quickly with respect to distance between SNPs.) However, our main focus in this paper is not on producing a computational scheme that will deal with difficult situations that might arise, but rather on prior specification, and to provide an initial assessment of the potential for BVSR to be applied to large-scale problems. Indeed, we hope that our results stimulate more research on the challenging computational problems that can occur in applying BVSR to GWAS and similar settings.

3.1. *Posterior inclusion probabilities via Rao–Blackwellization.* In the context of GWAS, a key inferential question is which covariates have a high probability of being included in the model. That is, we wish to compute the *posterior inclusion probability* (PIP) of the jth covariate, $\Pr(\gamma_j = 1|\mathbf{y})$. Although one could obtain a simple Monte Carlo estimate of this probability by simply counting the proportion of MCMC samples for which $\gamma_j = 1$, this estimator may have high

sampling variance. To improve precision, we instead use the Rao–Blackwellized estimate,

$$\Pr(\gamma_j = 1|\mathbf{y}) \approx (1/M)\sum_{i=1}^{M} \Pr\bigl(\gamma_j = 1|\mathbf{y}, \boldsymbol{\gamma}_{-j}^{(i)}, \boldsymbol{\beta}_{-j}^{(i)}, \tau^{(i)}, h^{(i)}, \pi^{(i)}\bigr), \tag{3.2}$$

where $\boldsymbol{\gamma}^{(i)}, \boldsymbol{\beta}^{(i)}, \tau^{(i)}, h^{(i)}, \pi^{(i)}$ denote the ith MCMC sample from the posterior distribution of these parameters given $\mathbf{y}$, and $\boldsymbol{\gamma}_{-j}$ and $\boldsymbol{\beta}_{-j}$ denote the vectors $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ excluding the jth coordinate. The probabilities that are being averaged here essentially involve simple univariate regressions of residuals against covariate j, and so are fast to compute for all j even when p is very large. Details are given in Appendix B.

3.2. *Estimating proportion of variance explained.* To perform inference on the total proportion of variance in $\mathbf{y}$ explained by measured covariates, we use samples from the posterior distribution of $\mathrm{PVE}(\boldsymbol{\beta}, \tau)$, which is defined at equation (2.10). These posterior samples are obtained by simply computing $\mathrm{PVE}(\boldsymbol{\beta}^{(i)}, \tau^{(i)})$ for each sampled value of $\boldsymbol{\beta}, \tau$ from our MCMC scheme.

3.3. *Predicting future exchangeable observations.* Given observed covariates x_{n+1} for a future individual, we can predict a value of y_{n+1} for that individual by

$$E(y_{n+1}|\mathbf{y}) = x_{n+1}E(\boldsymbol{\beta}|\mathbf{y}). \tag{3.3}$$

To estimate $E(\boldsymbol{\beta}|\mathbf{y})$, we use the Rao–Blackwellized estimates

$$E(\beta_j|\mathbf{y}) \approx (1/M)\sum_{i=1}^{M} E\bigl(\beta_j|\gamma_j = 1, \mathbf{y}, \theta_{-j}^{(i)}\bigr)\Pr\bigl(\gamma_j = 1|\mathbf{y}, \theta_{-j}^{(i)}\bigr). \tag{3.4}$$

Expressions for the two terms in this sum are given in Appendix B.

3.4. *Assessing predictive performance.* Suppose that we estimate $\boldsymbol{\beta}$ to be $\hat{\boldsymbol{\beta}}$. One way to assess the overall quality of this estimate is to ask how well it would predict future observations, on average. Motivated by this, we define the mean squared prediction error (MSPE):

$$\mathrm{MSPE}(\hat{\boldsymbol{\beta}}; \boldsymbol{\beta}, \tau) = E(X\hat{\boldsymbol{\beta}} - \mathbf{y})^2 = \sum_{j=1}^{m} s_j(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^2 + 1/\tau, \tag{3.5}$$

where $\boldsymbol{\beta}$ is the true value of the parameter, and s_j is the variance of $\mathbf{x}_{\cdot j}$, defined at (2.11).

The MSPE has the disadvantage that its scale depends on the units of measurement of $\mathbf{y}$. Hence, we define a *relative prediction gain*, RPG, which contrasts the MSPE from an estimated $\boldsymbol{\beta}$ with the prediction loss from simply predicting the

mean of $\mathbf{y}$ for each future observation (MSPE_0) and to the prediction error attained by the true value of $\boldsymbol{\beta}$ ($\text{MSPE}_{\text{opt}} = 1/\tau$):

$$\text{RPG} = \frac{\text{MSPE}_0 - \text{MSPE}(\hat{\boldsymbol{\beta}})}{\text{MSPE}_0 - \text{MSPE}_{\text{opt}}}. \tag{3.6}$$

The RPG does not depend on τ or on the scale of measurement of $\mathbf{y}$, and indicates what proportion of the extractable signal we are successfully obtaining from the data. For example, if the total proportion of variance in $\mathbf{y}$ explained by $X_{\boldsymbol{\gamma}}$ is 0.2, then an RPG of 0.75 indicates that we are effectively able to extract three-quarters of this signal, leaving approximately 0.05 of the variance in $\mathbf{y}$ "unexplained." Note that $\text{RPG} = 0$ if the prediction performs as well as the mean, and $\text{RPG} = 1$ if the prediction performs as well as the true value of $\boldsymbol{\beta}$. If $\text{RPG} < 0$, then the prediction is worse than simply using the mean, effectively indicating a problem with "overfitting."

4. Goals and expectations. At this point it seems helpful to review what we are attempting to achieve, and why it might be achievable despite the apparent severity of the computational burden. In brief, our primary goal is to extract more information from signals that exist in these data, particularly marginal associations, than do standard single-SNP analyses that test each SNP, one at a time, for association with phenotype. (The vast majority of GWAS published so far restrict themselves to such single-SNP analyses.) One of the main difficulties in single-SNP analysis is to decide how confident one should be that individual SNPs are truly associated with the phenotype. This difficulty stems from the fact that confidence should depend on the unknown values of π and σ_a. In a single-SNP analysis one must make assumptions, either implicitly or explicitly, about these parameters. An important aim of our approach is to instead estimate these parameters from the data, and hence provide more data-driven estimates of confidence for each SNP being associated with phenotype. To get intuition into why the data are informative about π and σ_a, consider the following examples. First, suppose that in a GWAS involving 300,000 SNPs, there are 10 SNPs (in different genomic regions) that show very strong marginal associations. Then, effectively, we immediately learn that π is likely at least of the order of $10/300{,}000$ (and, of course, it may be considerably higher). Further, the estimated size of the effects at these 10 SNPs also immediately gives some idea of plausible values for σ_a (or, more precisely, for σ_a/τ). Conversely, suppose that in a different GWAS none of the 300,000 SNPs show even modest marginal associations. This immediately suggests that either π or σ_a/τ (or both) must be "small" (because if both were large, then there would be many strong effects, and we would have seen some of them). More generally, we note that the strength of the effect size at the most strongly associated SNPs immediately puts an upper bound on what kinds of effect size are possible, and hence an upper bound on plausible values for σ_a/τ. In essence, BVSR provides

a model-based approach to quantifying these qualitative ideas, taking account of relevant factors (e.g., sample size) that affect the amount of information in the data.

Another limitation of single-SNP analyses, at least as conventionally applied, is that once some SNPs are confidently identified to be associated with outcome, they are not controlled for, as they should be, in analysis of subsequent SNPs. Controlling for SNPs that truly affect phenotype should help in identifying further such SNPs, and so a second key aim of our approach is to accomplish this. To see why our approach should attain this goal, note that our Rao–Blackwellization procedure for estimating marginal posterior inclusion probabilities is effectively a conventional single-SNP analysis that controls for the SNPs currently in the model. Thus, for example, if we start our MCMC algorithm from a point that includes the strongest marginal associations, it effectively immediately accomplishes this second goal.

We note two things we are *not* attempting to do. First, we are not attempting to identify a single best model (i.e., combination of SNPs), or to estimate posterior probabilities for specific models. In this context—and, we would argue, many other contexts where BVSR may be appropriate—these goals are of no interest, because the combination of small effect sizes and $p \gg n$ mean that the posterior probability on any particular model is going to be very small, and the chance of identifying the "correct" model is effectively zero. Neither are we attempting to identify combinations of SNPs that interact in a nonadditive way to affect the phenotype—SNPs that have little marginal signal, but whose effect is only revealed when they are considered together in combination with others. While such combinations of SNPs may exist, and identifying them would be of considerable interest, this seems considerably more challenging, both statistically and computationally, than our more modest goals here.

Finally, we note a particular feature of GWAS studies that may make it easier to obtain useful results from BVSR than in other contexts. Specifically, correlations among SNPs tend to be highly "local": each SNP is typically correlated with only a relatively small number of other SNPs that are near to it (linearly along the DNA sequence), and any two randomly chosen SNPs are typically uncorrelated with one another. Put another way, the matrix $X'X$ tends to have a highly banded structure, with large values clustering near the diagonal. To understand why this is helpful, note that one of the main potential pitfalls in applying MCMC to BVSR is that the MCMC scheme may get stuck in a "local mode" where a particular covariate (A, say) is included in the model, whereas in fact a different correlated covariate (B, say) should have been included. To help avoid getting stuck like this, the MCMC scheme could include specific steps that propose to interchange correlated covariates (e.g., remove A from the model and add B to the model), and the local correlation structure among SNPs in a GWAS means that this is easily implemented by simply proposing to interchange nearby SNPs. Furthermore, and perhaps more importantly, the local correlation structure means that getting stuck in such local modes may not matter very much, because if A and B are correlated,

then they are also almost certainly close to one another in the genome, and hence implicate a similar set of genes, and correctly identifying a set of implicated genes is the ultimate goal of most GWAS analyses.

5. Simulations and comparisons with other methods. We now present a variety of simulation results to illustrate features of our method, and assess its performance. Because our priors and methodology were primarily motivated by GWAS, these simulations are designed to mimic certain features of a typical GWAS. These include particularly that $p \gg n$ (in our simulations $p \approx$ 10,000–300,000 and $n \approx$ 1,000), extreme sparsity (in most of our simulations $\sim$30 covariates affect response), and small effect sizes (most relevant covariates individually explain $<1\%$ of the variance of $\mathbf{y}$).

5.1. *Simulation details*. We performed simulations based on three different genotype data, including both simulated and real genotypes. The first is simulated 10,000 independent SNPs (henceforth 10K), the second is real genotypes at $\sim$317,000 SNPs (henceforth 317K), and the third is real genotypes at $\sim$550,000 SNPs (henceforth 550K). Both 317K and 550K genotypes closely mimic real GWAS, and comparison between them can illustrate the scalability of our method. The 10K data set is helpful for several reasons: smaller simulations run faster; they allow us to assess methods in a simpler setting where computational problems are less of an issue; and the independence of the covariates avoids problems with deciding what is meant by a "true association" when covariates are correlated with one another.

For the 10K data, we simulated genotypes as follows. At each SNP $j = 1, \ldots, 10{,}000$ the minor allele frequency f_j is drawn from a uniform distribution on $[0.05, 0.5]$, and then genotypes x_{ij} $(i = 1, \ldots, n)$ are drawn independently from a Binomial$(2, f_j)$ distribution. We use $n = 1{,}000$ and 6,000.

Both 317K and 550K data sets come from an association study performed by the Pharmacogenomics and Risk of Cardiovascular Disease (PARC) consortium [Reiner et al. (2008); Barber et al. (2010)]. The 317K genotypes come from the Illumina 317K BeadChip SNP arrays for 980 individuals and the 550K genotypes come from the Illumina 610K SNP chip plus a custom 13,680 SNP Illumina i-Select chip in 988 individuals (550K SNPs remain after QC). We replaced missing genotypes with their posterior mean given the observed genotypes, which we computed under a Hidden Markov Model [Scheet and Stephens (2006)] implemented in the software package BIMBAM [Guan and Stephens (2008)].

For both the simulated and real genotypes we simulated sets of phenotypes in the following way. First, we specified a value of PVE, the total proportion of variance in $\mathbf{y}$ explained by the relevant SNPs, that we wanted to achieve in the simulated data. Then we randomly selected a set of 30 "causal" SNPs, C, and simulated effect sizes β_j for each of these SNPs independently from an effect size distribution $\mathcal{E}(\cdot)$ (discussed below). Next we computed the value of τ that gives the desired

value for PVE($\boldsymbol{\beta}, \tau$) in equation (2.10). Finally, we simulated phenotypes for each individual using $y_i = \sum_{j \in C} \beta_j x_{ij} + N(0, \tau^{-1})$.

Unless otherwise stated, for the 10K SNP data sets we run BVSR for 1 million iterations, and for the 317K and 550K SNP data sets we use 2 million iterations. Run times for each data set varied from a few minutes to about one day on a single MAC Pro with 3 GHz processor. (Note that the running time per iteration depends primarily on the inferred values for $|\boldsymbol{\gamma}|$, not the total number of SNPs.)

5.2. *Other methods.* In results presented below we compare our method with two other methods: simple single-SNP analysis that tests each SNP one at a time for association with phenotype, and the penalized regression method LASSO [Tibshirani (1996)].

For the single SNP analyses we ranked SNPs by their single-SNP Bayes factors, computed using equation (A.2), with $\boldsymbol{\gamma}$ in the numerator being the vector with jth component 1 and all other components 0, and averaging over $\sigma_a = 0.4, 0.2$, and 0.1 as in Servin and Stephens (2007). (Using standard single-SNP p values instead of Bayes Factors gives very similar performance in terms of ranking SNPs.)

The LASSO procedure [Tibshirani (1996)] estimates $\boldsymbol{\beta}$ by minimizing the penalized residual sum of squares:

$$\operatorname*{argmin}_{\boldsymbol{\beta}} (\mathbf{y} - X\boldsymbol{\beta})^t (\mathbf{y} - X\boldsymbol{\beta}) + \lambda \sum_j |\beta_j|. \tag{5.1}$$

For sufficiently large penalties, λ, LASSO produces sparse estimates $\hat{\boldsymbol{\beta}}$. Its main practical advantage over BVSR appears to be computational: for example, one can efficiently find the global optimal solution path for $\boldsymbol{\beta}$ as λ varies. To apply the LASSO procedure, we used the `lars` package (v. 0.9-7) in R [Efron et al. (2004)].

5.3. *Inference of PVE, and its relationship to heritability.* The total proportion of variance in $\mathbf{y}$ explained by the relevant covariates $X_{\boldsymbol{\gamma}}$, or PVE, is commonly used to summarize the results of a linear regression. In GWAS the PVE is, conceptually, closely related to the "heritability" of the trait, which is widely used, for better or worse, as a summary of how "genetic" the phenotype is. The key difference between the PVE and heritability is that the PVE reflects the optimal predictive accuracy that could be achieved for a linear combination of the *measured* genetic variants, whereas heritability reflects the accuracy that could be achieved by *all* genetic variants. In recent GWAS, it has been generally observed, across a range of different diseases and clinical traits, that the proportion of phenotypic variance explained by "significant" genetic variants is much lower than previous estimates of heritability from family-based studies [Maher (2008)]. There are several possible explanations for this "missing heritability": for example, it may be that previous estimates of heritability are inflated for some reason. However, two explanations have received particular attention: some of the missing heritability could reflect

genetic variants that were measured but simply did not reach stringent levels of "significance" in standard analyses, while other parts of the missing heritability could reflect genetic variants that were unmeasured (and not strongly correlated with measured variants). Because the measured genetic variants in current GWAS studies are predominantly "common" genetic variants (those with a population frequency exceeding a few percent), the relative contribution of these two factors is connected to the contentious topic of the relative contributions of common vs rare variants to phenotypic variation and disease risk [Pritchard (2001)]. Comparing the PVE with heritability should provide some insights into the relative contributions of these two factors. For example, at the simplest level, if the PVE is almost as big as the heritability, then this suggests that most phenotypic variation is due to variation at SNPs that are highly correlated with measured genetic variants, and perhaps that rare genetic variants, which are usually not strongly correlated with measured common variants, contribute little to phenotypic variation.

An important feature of BVSR that allows it to estimate the PVE, together with measures of confidence, is its use of *Bayesian model averaging* (BMA) to average over uncertainty in which covariates are relevant. This is very different from single SNP analyses and standard penalized regression approaches, which typically result in identification of a single set of potentially-relevant covariates, and so do not naturally provide estimates of the PVE that take account of the fact that this set may be missing some relevant covariates and include some irrelevant covariates. Since, as far as we are aware, the ability of BVSR to estimate PVE has not been examined previously, we performed simulation studies to assess its potential.

For both real and simulated genotype data (described above), we simulated 50 independent sets of phenotype data, each containing 30 randomly-chosen "causal" SNPs affecting phenotype, varying PVE from 0.01 to 0.5 in steps of 0.01. Our Bayesian model assumes, through the prior on $\boldsymbol{\beta}$, that the effect size distribution $\mathcal{E}$ is normal. To check for robustness to deviations from this assumption, we simulated phenotype data using both $\mathcal{E} = N(0, 1)$ (as effectively assumed by our model) and $\mathcal{E} = \mathrm{DE}(1)$, where DE denotes the double exponential distribution. The results from these two different distributions were qualitatively similar, and so we show only the results for $\mathcal{E} = \mathrm{DE}(1)$.

Figure 1 shows estimates of PVE obtained by our method against the true values. For both simulated and real SNP data there is a generally good correspondence between the true and inferred values, and 90% credible intervals (CI) for PVE covered the true value in 85% of cases. As might be expected, the uncertainty in PVE is greater when there is a larger number of SNPs, presumably due to the increased difficulty in reliably identifying relevant variants. In addition, the uncertainty in PVE tends to be greater when the true PVE is smaller. Our intuition is that when the data contain no SNPs with strong individual effects, it remains difficult to rule out the possibility that many SNPs may have very small effects that combine to produce an appreciable PVE. Nonetheless, even when the true PVE is small, the inferred posterior interval for PVE does exclude large values, illustrating that even in this case our method is able to extract information from the data.

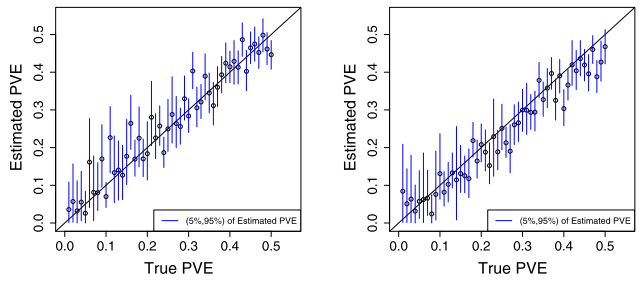


FIG. 1. *Comparison of true and inferred values for the proportion of variance in* $\mathbf{y}$ *explained by relevant covariates* (PVE). *Panel* (a) *shows results for* 1,000 *individuals with* 10,000 *independent simulated SNPs*; *Panel* (b) *shows results for* 980 *individuals with* 317*K real SNP genotypes*. *Circles indicate posterior mean for* PVE; *vertical bars indicate the symmetric* 90% *credible interval*.

5.4. *Many causal SNPs with tiny effects*. The simulations above involve 30 causal SNPs explaining *in total* between 0.01 and 0.5 of the total variance in $\mathbf{y}$. We note that this is a relatively subtle level of signal: in the following sections we will see that, for PVE $= 0.30$, and the sample sizes we used, it is typically not possible to confidently identify the majority of causal SNPs, nor to achieve the predictive performance that is similar to one would obtain if one knew the causal variants. Thus, to estimate the PVE, BVSR must not only identify variants that are confidently associated with $\mathbf{y}$, but also estimate how many additional variants of small effects it might be missing and what their effect sizes might be. Clearly, there must be some limit to its ability to accomplish all these tasks: in particular, if there were very many variants of minuscule effects, then it would be difficult to distinguish this from the null model in which no variants have any effect. To try to test these limits, we ran more challenging simulations involving many more SNPs with tiny individual effects, but a nontrivial overall PVE. Specifically, we considered two cases: (i) 300 causal SNPs out of the 10K simulated SNPs in 1,000 individuals; and (ii) 1,000 causal SNPs out of the 317K real SNPs in 980 individuals. In each case we simulate the effect sizes using a normal distribution. We simulated 10 independent sets of phenotypes with PVE $= 0.3$ in each case. For comparison in each case we also simulated 10 independent sets of phenotypes under a "null" model with no causal SNPs (PVE $= 0$).

For these data sets, to give BVSR some chance to identify the large number of causal SNPs, we increased M, the upper limit on the expected number of nonzero regression coefficients in our prior on π, to $M = 1{,}000$. Plots of 99% and 95% credible intervals for PVE in each simulation are shown in Figure 2.

Somewhat surprisingly, for the first set of simulations, with 300 causal SNPs out of 10K and PVE $= 0.3$, BVSR remains able to provide reasonable estimates of PVE: for example, for 5 of the 10 simulations the interquartile range of the posterior on PVE spans the true value of PVE $= 0.3$, and in 7 simulations the 90%

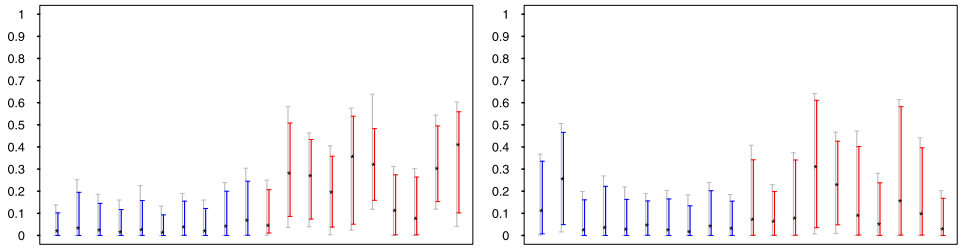


FIG. 2. *Plots showing estimation of* PVE *for simulations with large numbers of causal variants of very small effect. Panels* (a) *and* (b) *are for* 10*K and* 317*K data sets, respectively. The grey lines denote* 99% *CI and colored lines denote* 95% *CI. The blue color indicates null simulation* (PVE = 0), *red indicates alternative simulations* (PVE = 0.3). *The * denotes the median.*

symmetric CI includes PVE = 0.3. Further, there is a clear qualitative difference between the results of PVE = 0.3 and PVE = 0. Less surprisingly, for the extremely challenging case of 1,000 causal SNPs out of 317K, the estimates of PVE are considerably less precise. However, even here, these admittedly limited simulations appear to show systematic differences between PVE = 0.3 and PVE = 0. For example, for PVE = 0.3, 8 of the 90% CIs cover PVE = 0.2 and 6 CIs cover PVE = 0.3; whereas for PVE = 0, only 2 of the 90% CIs cover 0.2 and 1 CI covers 0.3.

5.5. *Identifying the causal SNPs.* In existing GWAS the vast majority of studies published so far restrict their analysis to the simplest possible approach of testing each SNP, one at a time, for association with phenotype. One possible advantage of a multi-SNP analysis like ours is to improve power compared with this simple single-SNP approach. However, since each SNP is typically correlated with only a small number of other (nearby) SNPs, and so any two randomly chosen SNPs will be typically uncorrelated, the gain in power might be expected to be small (at least in the absence of interactions among SNPs). Further, one might be concerned that if our MCMC scheme does not mix adequately, then the results of the multi-SNP approach could actually be worse than those from a simpler analysis.

We performed two types of simulations to investigate these issues, the first using the 10K data set (independent SNPs), and the second using the chromosome 22 of the 550K data set (9,041 correlated SNPs). In each case we simulated 100 phenotype data sets as described above, with 30 causal SNPs and PVE = 0.25.

For the 10K simulations we compared BVSR, single-SNP analyses, and LASSO in their ability to identify the causal SNPs as follows. For BVSR and single-SNP analyses we first computed, for each SNP, a measure of the evidence for association with phenotype. For BVSR we used the PIPs [equation (3.2)]; for single-SNP analysis we used the univariate Bayes Factor as described in Section 5.2. We then

consider thresholding this measure of evidence: for any given threshold, we consider all causal SNPs exceeding the threshold to be true positives, and all other SNPs exceeding the cutoff to be false positives. We compare methods by constructing curves showing the trade-off between true positives and false positives as the threshold is varied. For LASSO, we first computed the solution path as λ varies. Then, for each solution on this path we defined all causal SNPs with nonzero regression coefficients to be true positives, and all other SNPs with nonzero regression coefficients to be false positives. We then constructed curves showing the trade-off between true positives and false positives as λ is varied.

For the real (correlated) SNPs we performed a similar comparison, but assessed the methods in their ability to identify the correct genomic *regions* rather than individual SNPs. This is because the three methods differ qualitatively in the way they identify SNP associations when SNPs are correlated with one another: single-SNP analyses tend to identify significant associations at any SNP that is strongly correlated with a causal SNP; LASSO tends instead to select just one or a few correlated SNPs; and BVSR tends to spread the association signal (the PIPs) out among correlated SNPs. While it may be important to be aware of these qualitative differences when interpreting results from the methods, they are not our main interest here, and we assess the methods at the level of regions in an attempt to reduce the influence of these qualitative differences. (Further, it could be argued that identifying regions of interest is the primary goal of GWAS.) To describe the approach in more detail, we partitioned chromosome 22 into 200 kb nonoverlapping regions (different choices of region size that we tried produced qualitatively similar results). We then used each method to assign each region a "region statistic" indicating the strength of the evidence for an association in that region. For single SNP analysis we used the maximum single SNP Bayes factor within each region; for BVSR we used the sum of the PIP for SNPs in the region; and for LASSO we used the penalty λ at which any SNP in that region is included in the model. Similar to the SNP-level comparisons, we plot how true and false positive regions vary as the threshold on the region statistic is varied. (We averaged results over two different starting positions for the first window, 0, and 100 kb.)

Figure 3 shows curves of the trade-off between true and false positives for each method in the two different simulations. Each point on the curve shows the total true vs false positives across the hundred simulated data sets, using a common threshold across data sets. (An alternative way to combine data sets is to use a different threshold in each data set, vary the thresholds in such a way as to produce the same number of positive findings in each data set; the two different ways to combine data sets give similar results.)

For a given number of false positives, the multi-SNP approaches (BVSR and LASSO) always yield as many or more true positives than the single-SNP analysis. For the 10K simulated SNPs BVSR and LASSO perform similarly, whereas for the real genotypes BVSR is better. (The reasons for this difference are unclear to us.) The results demonstrate that, even in the case where single-SNP tests

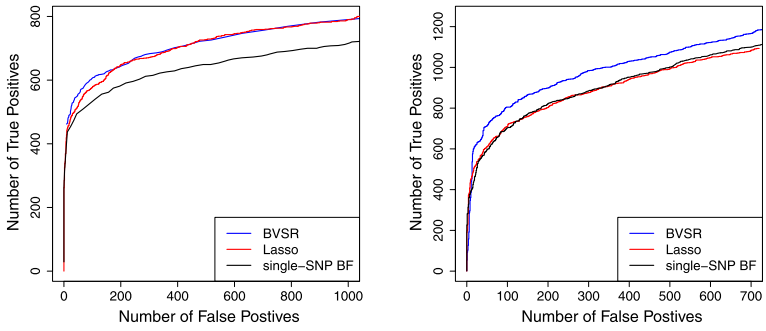


FIG. 3. *Graphs showing the trade-off between true positive and false positive SNP identifications for the different methods*: *BVSR* (*blue*), *LASSO* (*red*), *and single SNP analyses* (*black*). *Both plots show results that are summed across* 100 *data sets* (*see text for further explanation*). *Panel* (a) *is for independent simulated SNPs*; *Panel* (b) *is based on real genotype data for chromosome* 22.

might be expected to perform extremely well—that is, independent SNPs with no interactions—it is still possible to gain slightly in power by performing multi-SNP analyses. Our intuitive explanation for the gain in power of the multi-SNP approaches is that, once one identifies a causal variant, controlling for it will improve power to detect subsequent causal variants. Because the SNPs are independent, this gain is expected to be small: indeed, if the SNPs were exactly orthogonal, then one would expect no gain by controlling for identified variants. However, our results show that even in the case of independent SNPs the gain is measurable because the finite sample size produces nonzero sample correlations between "independent" SNPs.

We note that, at least in these simulations, most of the gain from the multi-SNP methods occurs when the number of false positives is small but nontrivial: that is, the multi-SNP methods promote some of the moderately-difficult-to-detect causal SNPs slightly higher in the SNP rankings, but not so far as to put them at the very top. This suggests that multi-SNP analysis may be most useful when used in combination with other types of data or analysis that attempt to distinguish true and false positives among the SNPs near the top of the association rankings [as in Raychaudhuri et al. (2009), e.g., where information on gene similarities taken from PubMed abstracts are used in this way].

5.6. *Prediction performance.* We used the same simulated data as in the previous section to compare predictive performance of BVSR and LASSO. To measure predictive accuracy, we use the relative prediction gain, defined at (3.6). For our method we compute $\mathrm{RPG}(\bar{\boldsymbol{\beta}})$ where $\bar{\boldsymbol{\beta}}$ is the posterior mean for $\boldsymbol{\beta}$. For LASSO we compute the RPG in two ways, which we will refer to as RPG_1 and RPG_2. For RPG_1 we first compute $\mathrm{RPG}(\boldsymbol{\beta}^{(i)})$ for each $\boldsymbol{\beta}^{(i)}$ in the solution path for $\boldsymbol{\beta}$ output by

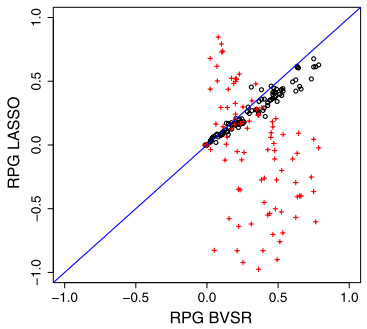


FIG. 4. *Comparison of the relative prediction gain* (RPG) *for BVSR* (*x-axis*) *and LASSO* (*y-axis*). *Black circles are results from the optimal LASSO solution without refitting* (RPG_1); *Red crosses are corresponding results with refitting* (RPG_2), *described in the main text*.

the `lars` package, and take the minimum of these relative prediction errors. Note that by taking the minimum over λ in this way we are effectively assuming that an oracle has given us the optimal value for λ; in practice, one would need to obtain λ through other means, such as cross-validation, which would result in worse accuracy than RPG_1. For RPG_2 we take a two-stage approach to prediction. First, we use LASSO to select the SNPs that should have nonzero coefficients (using the λ used for RPG_1), and then we estimate the regression coefficients of these SNPs using ordinary least squares ($\boldsymbol{\beta}_{\text{OLS}}$), and compute $\text{RPG}_2 := \text{RPG}(\boldsymbol{\beta}_{\text{OLS}})$. The motivation for this procedure is that if LASSO is able to reliably identify the correct coefficients, then the refitting procedure will improve predictive performance by avoiding the known tendency for LASSO to overshrink nonzero regression coefficients; however, as we shall see below, the refitting can be counterproductive when the correct coefficients are not reliably identified.

Figure 4 compares the RPG obtained from the three methods on 100 simulated data sets. The RPG from our Bayesian approach is higher than that obtained directly from the optimal LASSO solution (RPG_1) in 82 of the 100 data sets, and mean RPG is higher (0.315 vs 0.261). The refitting procedure has a substantial effect on predictive accuracy, and, in particular, it substantially increases the variance of the performance: for some data sets the refitting procedure improves predictive performance, but for the majority of data sets it results in much worse RPG. Indeed, RPG_2 is often negative, indicating that predictive performance after refitting is substantially worse than simply using the mean phenotype value, which is the symptom of "overfitting." This behavior makes intuitive sense: in cases when the optimal LASSO solution does a good job of precisely identifying many of the relevant covariates, and no irrelevant ones, the refitting step improves predictive performance, but when the first stage includes several false positives the refitting procedure is counter-productive.

Although our Bayesian model is sparse, our estimated $\bar{\boldsymbol{\beta}}$ is not sparse due to the averaging in (3.4). In some contexts one might want to obtain a sparse predictor,

so, to examine how this might impact predictive accuracy, we computed the RPG for each data set using only the P covariates with highest posterior inclusion probabilities (setting other coordinates of $\bar{\boldsymbol{\beta}}$ to 0), where $P = 10, 30, 100$. The average RPG for these sparse estimates of $\boldsymbol{\beta}$ were essentially unchanged from using the nonsparse estimate $\bar{\boldsymbol{\beta}}$ (RPG $= 0.313, 0.315,$ and 0.315, resp.).

We also examined the benefits of using Bayesian model averaging (BMA) to perform prediction, by computing the RPG obtained using only those covariates with a posterior inclusion probability $>t$ where $t = 0.2, 0.5, 0.8$. [When $t = 0.5$ this is the "median probability model" of Barbieri and Berger (2004).] Specifically, we computed the RPG for $\hat{\boldsymbol{\beta}}_j = I(\hat{\Pr}(\gamma_j = 1) > t)\hat{E}(\beta_j|\gamma_j = 1)$, where the two quantities on the right-hand side are estimated from (3.2) and (3.4). These estimates have some shrinkage because $E(\beta_j|\gamma_j = 1)$ is a shrinkage estimate of β_j (due to the normal prior on $\boldsymbol{\beta}$), but they do not have the additional shrinkage term $\Pr(\gamma_j = 1)$ that BMA provides to further shrink variables that are not confidently included in the model. The average RPG's for these non-BMA estimates were notably worse than for the BMA-based estimates: $0.244, 0.291,$ and 0.272, respectively, compared with 0.315 for BMA.

Taken together, these results suggest that BMA is responsible for a moderate amount of the gain in predictive performance of BVSR compared with RPG_1, with some of the remainder being due to LASSO's tendency to over-shrink estimates of the nonzero regression coefficients. One way to think of this is that LASSO has only a single parameter, λ, that controls both shrinkage and sparsity. In this setting the true solution is very sparse, so λ needs to be big enough to keep the solution sufficiently sparse; but having λ this big also creates an overly strong shrinkage effect. In contrast, BVSR effectively avoids this problem by having two parameters, σ_a controlling shrinkage, and π controlling sparsity. As we have seen, in this context the strategy of refitting the β coefficients at the LASSO solution fails to improve average predictive performance. Other possible ways around this problem include using a more flexible penalized regression model (e.g., the Elastic Net [Zou and Hastie (2005)] has two parameters, rather than one), or using a procedure that does not overshrink large effect sizes, for example, SCAD [Fan and Li (2001)]. Comparisons of these methods with BVSR would be an interesting area for future work.

5.7. *Calibration of the posterior inclusion probabilities.* One of the main advantages of BVSR compared with Bayesian single-SNP analysis methods is that BVSR allows the hyperparameters π and σ_a to be estimated from the data, and thus provides data-driven estimates of the posterior inclusion probabilities (PIPs). One hope is that estimating these parameters from the data will lead to better-calibrated estimates of the PIPs than the single-SNP approach which effectively requires one to supply educated guesses for these parameters. To assess this, Figure 5(a) shows the calibration of the PIPs from BVSR, for the simulations used in

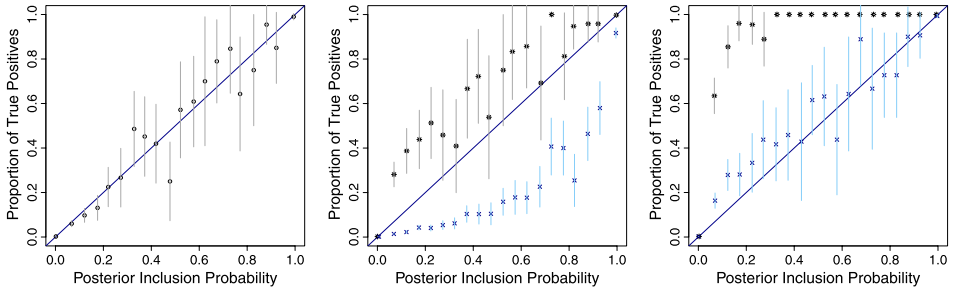


FIG. 5. *Calibration of the posterior inclusion probabilities* (*PIPs*) *from BVSR. The graph was obtained by binning the PIPs obtained from BVSR in* 20 *bins of width* 0.05. *Each point on the graph represents a single bin, with the* x *coordinate being the mean of the PIPs within that bin, and the* y *coordinate being the proportion of SNPs in that bin that were true positives* (*i.e., causal SNPs in our simulations*). *Vertical bars show* ± 2 *standard errors of the proportions, computed from a binomial distribution. Panel* (a) *is the result of BVSR, using the priors described here. The fact that the points lie near the line* $y = x$ *indicates that the PIPs are reasonably well calibrated, and thus provide a reliable assessment of the confidence that each SNP should be included in the regression. Panel* (b) *is the result from BVSR fixing* π *to be either* $5\times$ *smaller* (*black star*) *or* $5\times$ *larger* (*blue cross*) *than the true value* (σ_a *fixed to true value*). *Panel* (c) *is the result of fixing* σ_a *to be either* $5\times$ *smaller* (*black star*) *or* $5\times$ *larger* (*blue cross*) *than the true value* (π *fixed to true value*).

the estimation of PVE above (fifty data sets with PVE $= 0.01$–0.5 for both normal and exponential effect size distributions). The figure shows that the PIPs are reasonably well calibrated. In particular, SNPs with high PIP have a high probability of being causal variants in the simulations.

To illustrate the potential benefits of using moderately-diffuse prior distributions on π and σ_a, allowing their values to be informed by the data, rather than fixing them to specific values, we also applied BVSR with either π or σ_a fixed to an "incorrect" value (approximately 5 times larger or smaller than the values used in the simulations). Figure 5(b) and (c) show how, as might be expected, this can result in poorly-calibrated estimates of the PIP (of course, if one were lucky enough to fix both π and σ_a to their "correct" values, then calibration of PIPs will be good, but, in practice, the correct values are not known). We note that fixing σ_a to be five-fold too large seems to have only a limited detrimental effect on calibration, which is consistent with the fact that in single-SNP analyses, with moderate sample sizes, BFs are relatively insensitive to choice of σ_a provided it is not too small [e.g., Stephens and Balding (2009), Figure 1]. This suggests that, in specifying priors on σ_a, it may be prudent to err on the side of using a distribution with too long a tail rather than too short a tail. Note that, as in Bayesian single-SNP analyses, although the numerical value of the PIP is sensitive to choice of π, the ranking of SNPs is relatively insensitive to choice of π (and, indeed, σ_a). Consequently, in contrast to the calibration plot, power plots of the kind shown in Figure 3 are not sensitive to choice of prior on either π or σ_a (results not shown).

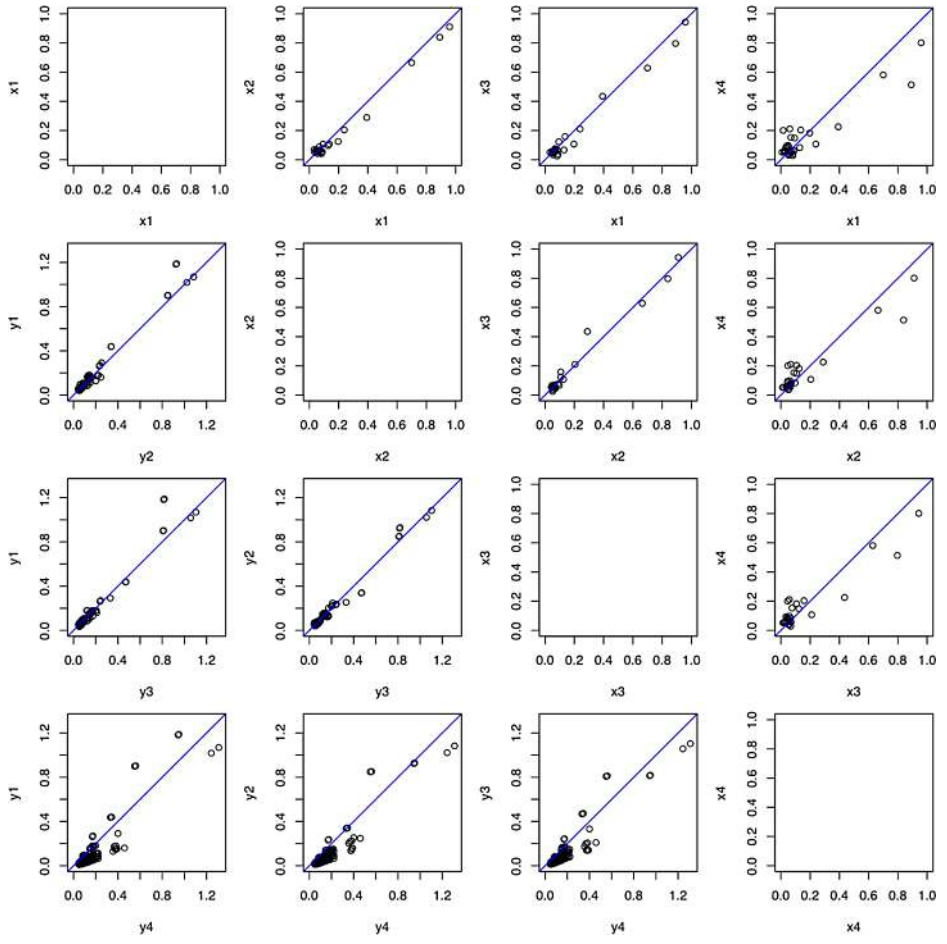
5.8. *Real data analysis*: *PARC GWAS for C-reactive protein*. We applied BVSR to analyze a GWAS study to identify genetic variants associated with plasma C-reactive protein (CRP) concentration. CRP is a protein found in the blood that is associated with inflammation, and is predictive of future cardiovascular disease [Ridker et al. (2002)]. The data come from the Pharmocogenetics and Risk of Cardiovascular Disease (PARC) study [Reiner et al. (2008) and references therein].

The available genotype data consisted of 1968 individuals genotyped on either the Illumina 317K chip (980 individuals) or the Illumina 610K SNP chip plus a custom 13,680 SNP Illumina i-Select chip (988 individuals). These genotype data had undergone basic quality control filters (e.g., removing SNPs with very high proportions of missing data, or showing strong departures from Hardy–Weinberg equilibrium) prior to our analysis. To merge the two data sets, we used genotype imputation [Servin and Stephens (2007); Marchini et al. (2007)], using the software package BIMBAM [Guan and Stephens (2008)] to replace missing or unmeasured genotypes with their posterior mean given the observed genotype data [see Guan and Stephens (2008) for discussion of this strategy]. After imputing missing genotypes, we removed SNPs with (estimated) minor allele frequency <0.01, leaving a total of 530,691 SNPs.

The phenotype data consisted of plasma concentrations of CRP, measured multiple times for each individual, both before and after exposure to statin drugs. These multiple measures were adjusted for covariates (age, sex, smoking status, and body mass index), quantile normalized to a standard normal distribution, and averaged to produce a single summary measure of CRP concentration for each individual (relative to other individuals in the same study), as described in Reiner et al. (2008).

After removing individuals with missing phenotypes, we had phenotype and genotype data on a total of 1,682 individuals. We performed four independent MCMC runs, two with 2 million iterations, and two using 4 million iterations. These longer runs took approximately 60 and 90 CPU hours on a Mac Pro 3 GHz desktop. Comparing results among runs, we found three of the runs gave very good agreement in all aspects we examined, whereas the fourth run showed mild but noticeably greater departure from the others, suggesting possible convergence or mixing issues. For example, Figure 6(a) compares the estimated PIPs for each pair of runs, and Figure 6(b) compares the estimated posterior distribution of PVE among runs. The remainder of the results in this section are based on pooling the results from all four runs.

The usual way to summarize single-SNP analyses is to report the SNPs with the strongest marginal evidence for association. Thus, it might seem natural in a multi-SNP analysis to focus on the SNPs with the largest posterior inclusion probabilities (PIPs). However, this can be misleading. For example, if there are many SNPs in a region that are highly correlated with one another, and all approximately equally associated with the phenotype, then it may be that the correct conclusion is that at least one of these SNPs should be included in the model, but there might be

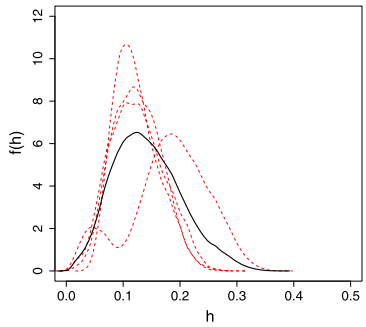


(a)

FIG. 6. *Illustration of the consistency of results across four different runs of the MCMC algorithm for the CRP data. In panel* (a) *the* (i, j)*th plot compares results for runs* i *and* j*. Plots in the upper triangle* $(j > i)$ *compare estimated posterior inclusion probabilities* (*PIPs*) *for each SNP. Plots in the lower triangle compare estimated posterior expected number of SNPs in* 1 *Mb regions* (*so each point corresponds to a single region*)*. The line* $y = x$ *is marked in blue. Panel* (b) *shows posterior distributions of* PVE *from the four MCMC runs.*

considerable uncertainty about which one. In this case, even though the posterior probability of at least one SNP being included in the model would be high (near 1), none of the individual PIPs may be very big, and concentrating on the PIPs alone would risk missing this signal in the data. To avoid this problem, we prefer to initially summarize results at the level of *regions*, as we now illustrate.

We divided the genome into overlapping regions, each 1 Megabase (10^6 bases) in length, with the overlap between adjacent regions being 0.5 Megabases. For each



(b)

FIG. 6. (*Continued*).

region we computed two quantities: (i) an estimate, E, of the posterior expected number of SNPs included in the model, being the sum of the estimated PIPs for all SNPs in the region; (ii) an estimate of the probabilities, P, that the region contains (a) 1 SNP, (b) 2 SNPs, or (c) more than 2 SNPs included in the model. The latter quantities (ii) are perhaps the most natural summary of the evidence that the region harbors genetic variants affecting phenotype, but (i) has the advantage that it can be easily approximated using Rao–Blackwellization, resulting in lower Monte Carlo error. Thus, in practice, we suggest examining both quantities, and placing more trust in (i) where the two disagree.

The results are summarized in Figure 7, which also shows results for a single permutation of the phenotypes for comparison. The plot clearly identifies two regions with very strong evidence for an association with CRP in both plots (e.g., $E > 0.95$), and a third region with moderately strong evidence (e.g., $E > 0.75$). Multiple other regions show modest signals ($E = 0.1$ to 0.5), that might generally be considered worthy of follow-up in larger samples, although at this level of signal the majority are, of course, unlikely to be truly associated with CRP.

The three regions with the strongest association signals contain the genes *CRP*, *HNF1A*, and *APOE/APOC*, all of which have shown robustly-replicated SNP associations with C-reactive protein levels in several other GWAS using single-SNP analyses [e.g., Reiner et al. (2008); Ridker et al. (2008)]. In addition, in these data, these three regions all contain single SNPs showing strong associations: the largest single-SNP Bayes factors in each of these regions are $10^{6.2}$, $10^{5.5}$, and $10^{4.9}$, respectively. Thus, in this case the identification of regions of interest from BVSR is largely concordant with what one would have obtained from a single-SNP analysis. However, we highlight two advantages of the BVSR analysis. First, the estimated posterior probabilities obtained for each region are easier to interpret than the single-SNP Bayes Factors. For example, the estimated posterior probability that the HNF1A region contains at least one SNP included in the model is 0.96, and this seems much more helpful than knowing that the largest single-SNP Bayes

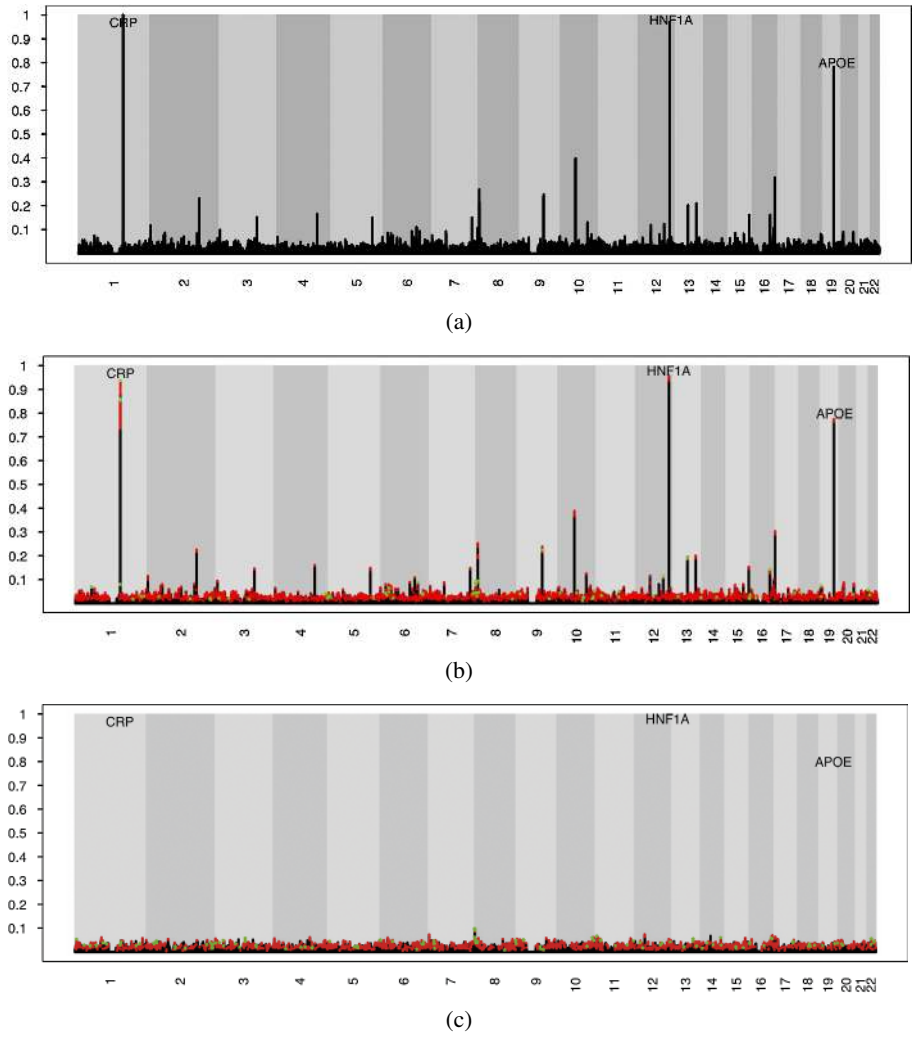


FIG. 7. *For each* 1 *Mb region we show an estimate from BVSR that the region contains* 1 (*black*), 2 (*red*), *or more than* 2 (*green with* ⋆) *SNPs in the regression. The* 1 *Mb regions overlap by* 0.5 *Mb, and so any SNP with a large PIP would cause a signal to occur in* 2 *adjacent regions on the plot. Panel* (a) *shows sum of PIP in each* 1 *Mb region* (*truncated at* 1). *Panel* (b) *shows estimated probabilities that each genomic region harbors variants associated with CRP levels. In panel* (c) *we permute phenotype once and produce the same plot as a comparison.*

Factor in the region is $10^{4.9}$. Similarly, for the next most associated region, which is on chromosome 10 near the gene FAM13C1, the posterior probability of 0.42 is simpler to interpret than the fact that the largest single-SNP Bayes Factor is $10^{3.9}$. And while these single-SNP BFs are easily converted to posterior probabilities of association by specifying a prior probability of association (effectively π in

our model), the multi-SNP analysis reduces the risk of specifying an inappropriate value for π by learning about π from the data.

A second advantage of BVSR is its ability to estimate the PVE. To illustrate this, we first consider a typical single-SNP analysis in this context, which estimates the PVE for "significant" SNPs by performing ordinary least-squares regression on those SNPs. Applying this approach to these data, using a relatively liberal (by GWAS standards) threshold for significance (single-SNP BF $>10^4$), we find that significant SNPs explain approximately 6% of the overall variance in CRP after controlling for covariates. Comparing this with some previous estimates of heritability of CRP in the range 0.35–0.4 [Pankow et al. (2001); Lange et al. (2006)] suggests that a substantial amount of genetic variation influencing CRP remains to be identified, a feature that has become known as "missing heritability." One question of interest is to what extent this shortfall might be explained by measured genetic variants that simply failed to pass the significance threshold, vs being explained by unmeasured genetic variants or other factors. To assess this, we examine PVE obtained from applying BVSR on measured SNPs. The posterior distribution for PVE [Figure 6(b)] has mean 0.14, with a symmetric 90% CI of $[0.05, 0.25]$. Note that, as one might expect, the lower part of this CI is similar to the estimated PVE of "significant" SNPs. Because most of the posterior distribution lies above 0.06, we infer that larger studies of *the same set of SNPs* might be expected to uncover considerably more signal than this study. (Consistent with this, a larger study involving 6,345 women typed at a subset of the SNPs considered here identified four additional genome regions containing SNPs associated with CRP levels [Ridker et al. (2008)].) Conversely, the fact that the upper part of the CI (0.25) remains well short of previous estimates of heritability suggests that not all of the missing heritability is likely to be explained by simply conducting larger studies of the same SNPs, and that some alternative factors (e.g., unmeasured rare variants) may also contribute.

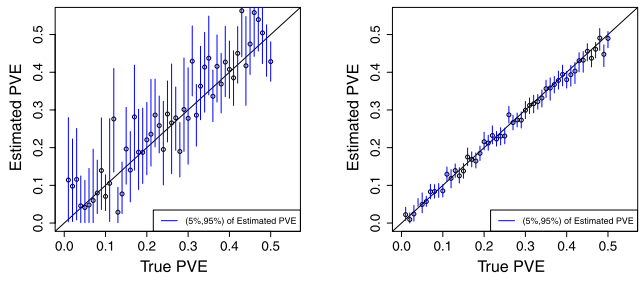
6. Extension to binary phenotypes. Although we have focused here on quantitative traits, BVSR is also potentially applicable to binary phenotypes, and this is important for GWAS applications because they often involve binary phenotypes. In this section we briefly summarize our attempts to extend BVSR in this way.

A standard approach to applying BVSR to binary phenotypes is to use a probit link function. In practice, this is usually accomplished by introducing latent variables $\mathbf{z}$ which are assumed to follow the standard linear regression (2.2), and to be related to the observed outcomes $\mathbf{y}$ by $y_i = 1$ $(z_i > 0)$ [Albert and Chib (1993)]. Posterior inference is performed by integrating out $\mathbf{z}$ using Markov chain Monte Carlo, which requires implementation of only one additional update compared with the quantitative trait (an update of the $\mathbf{z}$ variables).

A nice feature of this probit-based approach using latent Gaussian variables is that it would allow us to use the same priors as for quantitative outcomes, except

that these priors now relate to the unobserved latent (Gaussian) variables, rather than the observed (binary) outcomes. Furthermore, we can continue to summarize the overall signal by estimating the PVE of the latent variables. However, the way we have set things up, with an improper prior on τ, this would lead to improper posteriors on τ and $\mathbf{z}$ [because the likelihood $p(\mathbf{y}|\mathbf{z})$ is unchanged by multiplying $\mathbf{z}$ by any positive constant]. This could be rectified in a number of ways. For example, we could fix τ [e.g., to 1, as in Albert and Chib (1993)]. Here we instead choose to impose an identifiability constraint directly on the elements of $\mathbf{z}$, by constraining them to have (empirical) variance 1, because this allows us to re-use exactly the same computer code as for the quantitative phenotypes (whereas fixing τ would necessitate some changes). In addition, in an attempt to improve mixing, we make the approximation that the marginal distribution of the elements of $\mathbf{z}$ will be normal, which should be a reasonable approximation under the linear regression model (2.2) provided that there are no very large values for β. Specifically, we restrict $z_1, \ldots, z_n|\mathbf{y}$ to take a fixed set of values, being the n equally-spaced quantiles of a standard normal $N(0, 1)$ distribution, with the values corresponding to the n_0 individuals with $y_i = 0$ being constrained to be the first n_0 of these quantiles. The intuitive motivation for this constraint is that it can reduce the potential to fall into poor local optima by ruling out implausible configurations of $\mathbf{z}$ that correspond to some SNPs having very large effects. (Of course, this may not be a good idea in settings where very large effects are more plausible.) With this constraint in place, local Metropolis–Hastings proposals for $\mathbf{z}$ simply involve randomly picking a pair of individuals (i, j) with the same (binary) phenotype value and proposing to swap the values of z_i and z_j. (For long-range proposals, we simply compound this local proposal randomly many times.)

To provide a brief illustration of the potential for this approach, we applied the method to some simple simulated data sets. The genotypes were simulated in the same way described in Section 5.1, using 10,000 independent SNPs genotyped in $n = 1{,}000$ and 6,000 individuals. We simulated latent normal phenotypes by randomly selecting 30 causal SNPs and simulating a quantitative phenotype $\mathbf{z}$ with prespecified PVE as in Section 5.1. We then converted these n quantitative phenotypes to n binary phenotypes by mapping the largest $n/2$ z values to $y = 1$ and the remainder to $y = 0$. Figure 8 illustrates how reliably we are able to infer the PVE of the latent variables from the binary data. More generally, we find that provided we limit analyses to thousands of SNPs, we are able to obtain generally reliable results for binary traits (e.g., results from multiple independent runs largely agree with one another). Thus, for example, we should be able to obtain reliable results for small genomic regions, such as individual genes, which can itself be of considerable interest [Servin and Stephens (2007)]. However, our experience with larger real data sets involving hundreds of thousands of SNPs indicates that mixing is, as one might expect, harder for binary traits than for quantitative traits, and that to obtain reliable results in practice for GWAS may require longer MCMC runs and/or further methodological innovation.



(a) (b)

FIG. 8. *Comparison of true and inferred values of* PVE *for binary phenotypes. The estimated* PVE *is on the y-axis and the true* PVE *on the x-axis. Panels* (a) *and* (b) *are for* $n = 1{,}000$ *and* $n = 6{,}000$ *individuals, respectively. Circles indicate posterior mean for* PVE; *vertical bars indicate the symmetric* 90% *credible interval.*

7. Discussion. In this paper we have demonstrated that BVSR can be successfully applied to large problems, with a particular focus on genome-wide association studies. We have argued that BVSR has several potential benefits compared with standard single-SNP analyses, among them the ability to obtain data-driven estimates of hyperparameters that must otherwise be specified more subjectively by the user, and the ability to estimate the overall signal (PVE) that might be accounted for by relevant covariates, even when confidently identifying the relevant covariates is not possible. We have also introduced a novel, more interpretable, approach to prior specification in BVSR, and shown that BVSR can provide a competitive alternative to the penalized regression procedure LASSO.

However, despite our generally upbeat assessment, there are a number of potential limitations of the methods we have described here, which present both pitfalls to be aware of in practice, as well as challenges and opportunities for future work.

One important aspect of analysis of any GWAS is the potential for data quality to adversely impact results. For example, although modern genotyping technologies provided very high quality genotypes on average, some SNPs are much harder to genotype accurately than others, and genotyping error can occur at some SNPs at an appreciable rate. This can cause false positive associations if genotyping error is correlated with phenotype (which it can be, particularly in case-control studies if the DNA quality differs appreciably between cases and controls [Clayton et al. (2005)]). While quality control is vital to any study, it is of potentially even greater import in multi-SNP analyses than in single-SNP analyses, because in multi-SNP analyses the association results at one SNP affect the results at other SNPs, and so low quality data at a few SNPs may impact estimated associations at other SNPs. Thus, it seems particularly important to attempt to impose stringent data quality filters before embarking on a computationally-intensive multi-SNP analysis.

One limitation of the methods we present here is the assumption that effect sizes are normally distributed. Although our simulations with exponential effect sizes

suggest a certain amount of robustness to this assumption, it is important to note that there are some phenotypes where the normality assumption is clearly wrong. For example, in type 1 diabetes, one region of the genome, the MHC, contains genetic variants whose effect on phenotype may be substantially greater than any other region. When such regions of unusually large effect are known, it would be prudent to run methods like ours both including and excluding data at these loci, to check for robustness of conclusions. More generally, the robustness of our BVSR could be improved by replacing the assumption of normally-distributed effects with a heavy-tailed distribution such as a t with small or moderate degree of freedom, or indeed with a prior on the degrees of freedom.

Another related issue is that we assume the residual distribution of the phenotypes to be normal. To improve robustness to this assumption, we typically normal quantile transform the observed phenotypes to have a normal distribution before analysis (which while not strictly ensuring that the residuals are normal, does in our experience limit problems that might otherwise be caused by deviations from normality, such as occasional outlying values). Again, the use of a t distribution for the residuals might be preferable.

We view the work presented here as just the very start of what could be done with BVSR in GWAS. One important extension would be to incorporate additional information into the prior distribution on which variables are included in the regression ($\boldsymbol{\gamma}$ in our notation). Here we have assumed that variables are included in the model, independently, with common probability π. This independence assumption ignores likely local spatial dependence of $\boldsymbol{\gamma}$. In particular, it would be unsurprising to see multiple functional variants occurring in a single gene, and, indeed, analyses of genetic data in the CRP gene have suggested that it contains multiple SNPs affecting CRP levels [Verzilli et al. (2008); Stephens and Balding (2009)]. The independence assumption in the prior we use here makes it overly skeptical about this possibility. Another important possibility is that one could allow the prior probability of each SNP being included in the regression to depend on annotations of the SNP, such as where it lies relative to a gene, or whether it lies in a genomic region that is conserved across several species (a sign that the region may be functional). Of course it is not generally known a priori how much such annotations should affect the prior inclusion probabilities. However, with BVSR one could *estimate* hyperparameters that affect the prior inclusion probabilities from the data [Veyrieras et al. (2008)].

Finally, despite our focus on GWAS, many of the issues we have discussed here have broad relevance. In particular, while the computational challenges of BVSR remain considerably greater than penalized regression methods, we believe that the qualitative advantages of BVSR make it worth investing effort into designing more efficient inference algorithms for BVSR, to be able to better deal with the very large-scale applications that are becoming increasingly common.

APPENDIX A: DETAILS OF MCMC SCHEME

We use Markov chain Monte Carlo to obtain samples from the posterior distribution of $(h, \pi, \boldsymbol{\gamma})$ on the product space of $(0,1) \times (0,1) \times \{0,1\}^p$, which is given by

$$\mathsf{p}(h, \pi, \boldsymbol{\gamma}|\mathbf{y}) \propto \mathsf{p}(\mathbf{y}|h, \boldsymbol{\gamma})\mathsf{p}(h)\mathsf{p}(\boldsymbol{\gamma}|\pi)\mathsf{p}(\pi). \tag{A.1}$$

Here we are exploiting the fact that the parameters $\boldsymbol{\beta}$ and τ can be integrated out analytically to compute the marginal likelihood $\mathsf{p}(\mathbf{y}|h, \boldsymbol{\gamma})$. Indeed, in the limit for the hyperparameters $\lambda, \kappa \to 0$ and $\sigma_\mu \to \infty$ that we use here, we have

$$\frac{\mathsf{p}(\mathbf{y}|h, \boldsymbol{\gamma})}{\mathsf{p}(\mathbf{y}|h, \boldsymbol{\gamma} = \mathbf{0})} = n^{1/2}|\Omega|^{1/2}\frac{1}{\sigma_a(h, \boldsymbol{\gamma})^{|\boldsymbol{\gamma}|}}\left(\frac{\mathbf{y}^t\mathbf{y} - \mathbf{y}^t X_{\boldsymbol{\gamma}}\Omega X_{\boldsymbol{\gamma}}^t\mathbf{y}}{\mathbf{y}^t\mathbf{y} - n\bar{\mathbf{y}}^2}\right)^{-n/2}, \tag{A.2}$$

where $\Omega := (\sigma_a(h, \boldsymbol{\gamma})^{-2} I_{|\boldsymbol{\gamma}|} + X_{\boldsymbol{\gamma}}^t X_{\boldsymbol{\gamma}})^{-1}$ and $\mathbf{0}$ denotes the p-vector of all 0s. [For derivation, see Servin and Stephens (2007), Protocol S1 equation (13).] Note that here $\sigma_a(h, \boldsymbol{\gamma})$ is given by equation (2.13).

For each sampled values of $h, \boldsymbol{\gamma}$ from this posterior, we obtain samples from the posterior distributions of $\boldsymbol{\beta}$ and τ by sampling from their conditional distributions given $\mathbf{y}, \boldsymbol{\gamma}, h$:

$$\begin{aligned} \tau|\mathbf{y}, h, \boldsymbol{\gamma} &\sim \Gamma\big(n/2, 2/(\mathbf{y}^t\mathbf{y} - \mathbf{y}^t X_{\boldsymbol{\gamma}}\Omega X_{\boldsymbol{\gamma}}^t\mathbf{y})\big), \\ \boldsymbol{\beta}_{\boldsymbol{\gamma}}|\tau, \mathbf{y}, h, \boldsymbol{\gamma} &\sim N\big(\Omega X_{\boldsymbol{\gamma}}^t\mathbf{y}, (1/\tau)\Omega\big), \\ \boldsymbol{\beta}_{-\boldsymbol{\gamma}}|\tau, \mathbf{y}, h, \boldsymbol{\gamma} &\sim \delta_0. \end{aligned} \tag{A.3}$$

Our Markov chain Monte Carlo algorithm for sampling $h, \pi, \boldsymbol{\gamma}$ is based on a Metropolis–Hastings algorithm [Metropolis et al. (1953); Hastings (1970)], using a simple local proposal to jointly update $h, \pi, \boldsymbol{\gamma}$. In outline, the local proposal proceeds as follows. First a new proposed value of $\boldsymbol{\gamma}$, $\boldsymbol{\gamma}'$, is obtained by small modification of the current value (see below for more details); then a new value of π is proposed from a Beta$(|\boldsymbol{\gamma}'|, p - |\boldsymbol{\gamma}'| + 1)$ distribution; finally a proposed new value for h is obtained by adding a $U(-0.1, 0.1)$ random variable to the current value (reflecting proposed values that lie outside $[0, 1)$ about the boundary). The proposal distribution for π is proportional to its full conditional distribution given $\boldsymbol{\gamma}'$ inside the finite range of the prior on π [given by (2.8)]; on the infrequent occasions that the proposed value for π lies outside this range, it is of course rejected.

In addition to the local proposal described above, we sometimes (with probability 0.3 each iteration) make a longer-range proposal by compounding randomly-many local moves [the number being uniform on $(2, \ldots, 20)$]. This technique, named "small-world proposal," improves the theoretical convergence rate of the MCMC scheme [Guan and Krone (2007)].

We now give details on our update proposal for $\boldsymbol{\gamma}$. When adding a covariate into the model we use a *rank based proposal* that focuses more attention on covariates that are more likely to be included in the model. To do this, we first rank

the covariates based on their association with phenotype $\mathbf{y}$ (specifically we rank them by the Bayes factor for the model including only that covariate vs the null model containing no covariates, evaluated at $\sigma_a = 1$). Let Q_t be a distribution on $(0, \ldots, t-1)$ which has decreasing probability. Here we choose Q_t to be a mixture $Q_t = 0.3U_t + 0.7G_t$, where U_t is a uniform distribution on $\{0, \ldots, t-1\}$ and G_t is a geometric distribution truncated to $\{0, \ldots, t-1\}$, with its parameter chosen to give a mean of 2,000.

Now let $\boldsymbol{\gamma}^+$ denote the set of covariates that are currently in the model, $\boldsymbol{\gamma}^+ = \{i : \gamma_i = 1\}$. Let $\boldsymbol{\gamma}^-$ denote the complimentary set. We define three different types of moves, namely, add a covariate, remove a covariate, and exchange a pair of covariates in and out of the current model. Each move starts by setting $\boldsymbol{\gamma}' = \boldsymbol{\gamma}$. Then we randomly choose among the following:

- Add covariate: Generate $r \sim Q_{p-k}$, and find the covariate $i \in \boldsymbol{\gamma}^-$ that has rank r (among covariates in $\boldsymbol{\gamma}^-$). Set $\boldsymbol{\gamma}'_i = 1$.
- Remove covariate uniformly: Uniformly pick $i \in \boldsymbol{\gamma}^+$, and set $\boldsymbol{\gamma}'_i = 0$.
- Add a covariate and remove another: Pick i uniformly from $\boldsymbol{\gamma}^+$ and j uniformly from $\boldsymbol{\gamma}^-$, and set $\boldsymbol{\gamma}'_i = 0$; $\boldsymbol{\gamma}'_j = 1$.

In our current implementation, at each update we randomly select among these moves with probabilities 0.45, 0.45, and 0.1.

APPENDIX B: CALCULATIONS FOR RAO–BLACKWELLIZED ESTIMATES

In this appendix we derive the calculations need to compute the terms in equation (3.2).

Let θ_{-j} denote the parameters $(\gamma_{-j}, \boldsymbol{\beta}_{-j}, \tau, h, \pi)$. Note that

$$\Pr(\gamma_j = 1 | \mathbf{y}, \theta_{-j}) = \frac{\lambda}{1+\lambda}, \tag{B.1}$$

where

$$\begin{aligned}
\lambda &:= \frac{p(\gamma_j = 1 | \mathbf{y}, \theta_{-j})}{p(\gamma_j = 0 | \mathbf{y}, \theta_{-j})} \\
&= \frac{p(\mathbf{y} | \gamma_j = 1, \theta_{-j})}{p(\mathbf{y} | \gamma_j = 0, \theta_{-j})} \frac{p(\boldsymbol{\beta}_{-j} | \gamma_j = 1, \gamma_{-j}, \tau, h, \pi)}{p(\boldsymbol{\beta}_{-j} | \gamma_j = 0, \gamma_{-j}, \tau, h, \pi)} \frac{p(\gamma_j = 1 | \gamma_{-j}, \tau, h, \pi)}{p(\gamma_j = 0 | \gamma_{-j}, \tau, h, \pi)} \qquad \text{(B.2)} \\
&= \frac{p(\mathbf{y} | \gamma_j = 1, \theta_{-j})}{p(\mathbf{y} | \gamma_j = 0, \theta_{-j})} \frac{p(\boldsymbol{\beta}_{-j} | \gamma_j = 1, \gamma_{-j}, \tau, h)}{p(\boldsymbol{\beta}_{-j} | \gamma_j = 0, \gamma_{-j}, \tau, h)} \frac{\pi}{1-\pi}.
\end{aligned}$$

The second term here arises because in our parameterization $\boldsymbol{\beta}_{-j}$ is not independent of γ_j (because its prior variance, σ_a, is a function of $h, \boldsymbol{\gamma}$). This term is easily computed from the fact that $\boldsymbol{\beta}_{-j} | \boldsymbol{\gamma}, \tau, h$ are i.i.d. $\sim N(0, \sigma^2(h, \boldsymbol{\gamma})/\tau)$.

To compute the numerator of the first term note that

$$\mathbf{y}|\gamma_j = 1, \qquad \theta_{-j} \sim N(X_{\boldsymbol{\gamma}-j}\beta_{\boldsymbol{\gamma}-j} + \mu + X_j\beta_j, 1/\tau I), \tag{B.3}$$

with the priors on μ, β_j [from (2.3)] being

$$\begin{aligned} \mu|\tau &\sim N(0, \sigma_\mu^2/\tau), \\ \beta_j|\tau &\sim N(0, \sigma_a^2/\tau). \end{aligned} \tag{B.4}$$

Integrating out μ, β_j gives

$$p(\mathbf{y}|\gamma_j = 1, \tau) = (2\pi)^{-n/2}\tau^{n/2}\frac{|\Omega|^{1/2}}{\sigma_\mu\sigma_a}\exp\left(-\frac{1}{2}(R^tR - R^tX\Omega X^tR)\tau\right), \tag{B.5}$$

where $R = \mathbf{y} - X_{\boldsymbol{\gamma}-j}\beta_{\boldsymbol{\gamma}-j}$, $\boldsymbol{\gamma} - j$ denotes the vector obtained by taking $\boldsymbol{\gamma}$ and setting the jth coordinate to 0, $\Omega = (X^tX + \nu^{-1})^{-1}$, $\nu = \begin{pmatrix} \sigma_\mu^2 & 0 \\ 0 & \sigma_a^2 \end{pmatrix}$, and $X = (1, X_j)$ is an $n \times 2$ design matrix whose first column is all 1s. [See equation (8) from Protocol S1 in Servin and Stephens (2007).] The posterior distribution on β_j is given by

$$\beta_j|\mathbf{y}, \qquad \theta_{-j} \sim N(\Omega X^tR, \Omega). \tag{B.6}$$

Similarly, to compute the denominator of the first term, we use

$$\mathbf{y}|\gamma_j = 0, \qquad \theta_{-j} \sim N\big(X_{\boldsymbol{\gamma}-j}\beta_{\boldsymbol{\gamma}-j} + \mu, (1/\tau)I\big), \tag{B.7}$$

with priors on $\mu|\tau \sim N(0, \sigma_\mu^2/\tau)$. Integrate out μ to get

$$p(\mathbf{y}|\gamma_j = 0, \tau) = (2\pi)^{-n/2}\tau^{n/2}\frac{\Omega_0^{1/2}}{\sigma_\mu}\exp\left(-\frac{1}{2}(R^tR - \Omega_0 n^2\bar{R}^2)\tau\right), \tag{B.8}$$

where $\Omega_0 = (\sigma_\mu^{-2} + n)^{-1}$ and $\bar{R} = \frac{1}{n}\sum R_i$.

From this we obtain

$$\frac{p(\mathbf{y}|\gamma_j = 1, \theta_{-j})}{p(\mathbf{y}|\gamma_j = 0, \theta_{-j})} = \frac{|\Omega|^{1/2}}{\Omega_0^{1/2}}\frac{1}{\sigma_a}\exp\left(\frac{\tau}{2}(R^tX\Omega X^tR - \Omega_0 n^2\bar{R}^2)\right). \tag{B.9}$$

In the limit $\sigma_\mu \to \infty$ we have $\Omega_0 \to n$ and $\nu \to \begin{pmatrix} 0 & 0 \\ 0 & \sigma_a^2 \end{pmatrix}$ and the above expression becomes

$$\frac{p(\mathbf{y}|\gamma_j = 1, \theta_{-j})}{p(\mathbf{y}|\gamma_j = 0, \theta_{-j})} = |\Omega|^{1/2}\frac{n^{1/2}}{\sigma_a}\exp\left(\frac{\tau}{2}(R^tX\Omega X^tR - n\bar{R}^2)\right). \tag{B.10}$$

Note that this calculation effectively involves a univariate regression of the residuals R against covariate j. Furthermore, all covariates $j \notin \boldsymbol{\gamma}^+$ use the same residuals: only for $j \in \boldsymbol{\gamma}^+$ do the residuals need to be recomputed.

Acronyms used in the paper.

- BMA: Bayesian model averaging
- BVSR: Bayesian variable selection regression
- GWAS: genome wide association studies
- LASSO: least absolute shrinkage and selection operator, a popular variable selection method
- MCMC: Markov chain Monte Carlo
- PIP: posterior inclusion probability
- PVE: proportion of variance explained
- RPG: relative prediction gain
- SNP: single nucleotide polymorphism
- SIS: sure independence screen, a two-stage variable selection procedure.

Acknowledgments. We thank two anonymous referees, and the editor and associate editor for helpful comments on the initial submission. We thank P. Carbonetto for useful discussions.

REFERENCES

AGLIARI, A. and PARISETTI, C. C. (1988). A-g reference informative prior: A note on Zellner's *g* prior. *J. Roy. Statist. Soc. Ser. D* **37** 271–275.

ALBERT, J. H. and CHIB, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88** 669–679. MR1224394

BARBER, M. J., MANGRAVITE, L. M., HYDE, C. L., CHASMAN, D. I., SMITH, J. D., MCCARTY, C. A., LI, X., WILKE, R. A., RIEDER, M. J., WILLIAMS, P. T., RIDKER, P. M., CHATTERJEE, A., ROTTER, J. I., NICKERSON, D. A., STEPHENS, M. and KRAUSS, R. M. (2010). Genome-wide association of lipid-lowering response to statins in combined study populations. *PLoS ONE* **5** e9763.

BARBIERI, M. M. and BERGER, J. O. (2004). Optimal predictive model selection. *Ann. Statist.* **32** 870–897. MR2065192

BROWN, P. J., VANNUCCI, M. and FEARN, T. (2002). Bayes model averaging with selection of regressors. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 519–536. MR1924304

CASELLA, G. and ROBERT, C. P. (1996). Rao–Blackwellisation of sampling schemes. *Biometrika* **83** 81–94. MR1399157

CLAYTON, D. G., WALKER, N. M., SMYTH, D. J., PASK, R., COOPER, J. D., MAIER, L. M., SMINK, L. J., LAM, A. C., OVINGTON, N. R., STEVENS, H. E., NUTLAND, S., HOWSON, J. M. M., FAHAM, M., MOORHEAD, M., JONES, H. B., FALKOWSKI, M., HARDENBOL, P., WILLIS, T. D. and TODD, J. A. (2005). Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat. Genet.* **37** 1243–1246.

EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression (with discussion). *Ann. Statist.* **32** 407–499. MR2060166

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70** 849–911. MR2530322

GEORGE, E. I. and MCCULLOCH, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88** 881–889.

GUAN, Y. and KRONE, S. M. (2007). Small world MCMC and convergence to multi-modal distributions: From slow mixing to fast mixing. *Ann. Appl. Probab.* **17** 284–304. MR2292588

GUAN, Y. and STEPHENS, M. (2008). Practical issues in imputation-based association mapping. *PLoS Genet*. **4** e1000279.

HASTINGS, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57** 97–109.

HOGGART, C. J., WHITTAKER, J. C., DE IORIO, M. and BALDING, D. J. (2008). Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genet*. **4** e1000130.

LANGE, L. A., BURDON, K., LANGEFELD, C. D., LIU, Y., BECK, S. R., RICH, S. S., FREEDMAN, B. I., BROSNIHAN, K. B., HERRINGTON, D. M., WAGENKNECHT, L. E. and BOWDEN, D. W. (2006). Heritability and expression of c-reactive protein in type 2 diabetes in the diabetes heart study. *Ann. Hum. Genet*. **70** 717–725.

LIANG, F., PAULO, R., MOLINA, G., CLYDE, M. A. and BERGER, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *J. Amer. Statist. Assoc*. **103** 410–423. MR2420243

MAHER, B. (2008). Personal genomes: The case of the missing heritability. *Nature* **456** 18–21.

MARCHINI, J., HOWIE, B., MYERS, S., MCVEAN, G. and DONNELLY, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet*. **39** 906–913.

METROPOLIS, N., ROSENBLUTH, A., ROSENBLUTH, M., TELLER, A. and TELLER, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys*. **21** 1087–1092.

MILLER, A. (2002). *Subset Selection in Regression*, 2nd ed. *Monographs on Statistics and Applied Probability* **95**. Chapman & Hall/CRC, Boca Raton, FL. MR2001193

MITCHELL, T. J. and BEAUCHAMP, J. J. (1988). Bayesian variable selection in linear regression. *J. Amer. Statist. Assoc*. **83** 1023–1036. MR0997578

O'HARA, R. B. and SILLANPÄÄ, M. J. (2009). A review of Bayesian variable selection methods: What, how and which. *Bayesian Anal*. **4** 85–117. MR2486240

PANKOW, J. S., FOLSOM, A. R., CUSHMAN, M., BORECKI, I. B., HOPKINS, P. N., ECKFELDT, J. H. and TRACY, R. P. (2001). Familial and genetic determinants of systemic markers of inflammation: The NHLBI family heart study. *Atherosclerosis* **154** 681–689.

PRICE, A. L., PATTERSON, N. J., PLENGE, R. M., WEINBLATT, M. E., SHADICK, N. A. and REICH, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet*. **38** 904–909.

PRITCHARD, J. K. (2001). Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet*. **69** 124–137.

PRITCHARD, J. K., STEPHENS, M., ROSENBERG, N. A. and DONNELLY, P. (2000). Association mapping in structured populations. *Am. J. Hum. Genet*. **67** 170–181.

RAFTERY, A. E., MADIGAN, D. and HOETING, J. A. (1997). Bayesian model averaging for linear regression models. *J. Amer. Statist. Assoc*. **92** 179–191. MR1436107

RAYCHAUDHURI, S., PLENGE, R. M., ROSSIN, E. J., NG, A. C. Y., PURCELL, S. M., SKLAR, P., SCOLNICK, E. M., XAVIER, R. J., ALTSHULER, D., DALY, M. J. and CONSORTIUM, I. S. (2009). Identifying relationships among genomic disease regions: Predicting genes at pathogenic snp associations and rare deletions. *PLoS Genet*. **5** e1000534.

REINER, A. P., BARBER, M. J., GUAN, Y., RIDKER, P. M., LANGE, L. A., CHASMAN, D. I., WALSTON, J. D., COOPER, G. M., JENNY, N. S., RIEDER, M. J., DURDA, J. P., SMITH, J. D., NOVEMBRE, J., TRACY, R. P., ROTTER, J. I., STEPHENS, M., NICKERSON, D. A. and KRAUSS, R. M. (2008). Polymorphisms of the HNF1A gene encoding hepatocyte nuclear factor-1 alpha are associated with C-reactive protein. *Am. J. Hum. Genet*. **82** 1193–1201.

RIDKER, P. M., RIFAI, N., ROSE, L., BURING, J. E. and COOK, N. R. (2002). Comparison of C-reactive protein and low-density lipoprotein cholesterol levels in the prediction of first cardiovascular events. *N. Engl. J. Med*. **347** 1557–1565.

RIDKER, P. M., PARE, G., PARKER, A., ZEE, R. Y., DANIK, J. S., BURING, J. E., KWIATKOWSKI, D., COOK, N. R., MILETICH, J. P. and CHASMAN, D. I. (2008). Loci related to metabolic-syndrome pathways including LEPR, HNF1A, IL6R, and GCKR associate with plasma c-reactive protein: The women's genome health study. *Am. J. Hum. Genet.* **82** 1185–1192.

SCHEET, P. and STEPHENS, M. (2006). A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* **78** 629–644.

SERVIN, B. and STEPHENS, M. (2007). Efficient multipoint analysis of association studies: Candidate regions and quantitative traits. *PLoS Genet.* **3** e114.

SMITH, G. D. and EBRAHIM, S. (2003). Mendelian randomization: Can genetic epidemiology contribute to understanding environmental determinants of disease? *Internat. J. Epidemiology* **32** 1–22.

SMITH, M. and KOHN, R. (1996). Nonparametric regression using Bayesian variable selection. *J. Econometrics* **75** 317–343.

STEPHENS, M. and BALDING, D. J. (2009). Bayesian statistical methods for genetic association studies. *Nat. Rev. Genet.* **10** 681–690.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

VERZILLI, C., SHAH, T., CASAS, J. P., CHAPMAN, J., SANDHU, M., DEBENHAM, S. L., BOEKHOLDT, M. S., KHAW, K. T. T., WAREHAM, N. J., JUDSON, R., BENJAMIN, E. J., KATHIRESAN, S., LARSON, M. G., RONG, J., SOFAT, R., HUMPHRIES, S. E., SMEETH, L., CAVALLERI, G., WHITTAKER, J. C. and HINGORANI, A. D. (2008). Bayesian meta-analysis of genetic association studies with different sets of markers. *Am. J. Hum. Genet.* **82** 859–872.

VEYRIERAS, J.-B., KUDARAVALLI, S., KIM, S. Y., DERMITZAKIS, E. T., GILAD, Y., STEPHENS, M. and PRITCHARD, J. K. (2008). High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet.* **4** e1000214.

WAKEFIELD, J. (2009). Bayes factors for genome-wide association studies: Comparison with P-values. *Genet. Epidemiol.* **33** 79–86.

WELLCOME TRUST CASE CONTROL CONSORTIUM (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447** 661–678.

WU, T. T., CHEN, Y. F., HASTIE, T., SOBEL, E. and LANGE, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* **25** 714–721.

YANG, J., BENYAMIN, B., MCEVOY, B. P., GORDON, S., HENDERS, A. K., NYHOLT, D. R., MADDEN, P. A., HEATH, A. C., MARTIN, N. G., MONTGOMERY, G. W., GODDARD, M. E. and VISSCHER, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42** 565–569.

ZELLNER, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques* (P. K. Goel and A. Zellner, eds.) *Stud. Bayesian Econometrics Statist.* **6** 233–243. North-Holland, Amsterdam. MR0881437

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** 301–320. MR2137327

DEPARTMENTS OF PEDIATRICS
AND MOLECULAR AND HUMAN GENETICS
BAYLOR COLLEGE OF MEDICINE
ONE BAYLOR PLAZA
USDA CHILDREN'S NUTRITION RESEARCH CENTER
1100 BATES ST., STE. 2070
HOUSTON, TEXAS 77030
USA
E-MAIL: yongtaog@bcm.edu

DEPARTMENT OF STATISTICS
UNIVERSITY OF CHICAGO
ECKHART HALL ROOM 126
5734 S. UNIVERSITY AVENUE
CHICAGO, ILLINOIS 60637
USA
E-MAIL: mstephens@uchicago.edu