Bayesian Variable Selection in Structured High-Dimensional Covariate Spaces With Applications in Genomics

Fan LI and Nancy R. ZHANG

We consider the problem of variable selection in regression modeling in high-dimensional spaces where there is known structure among the covariates. This is an unconventional variable selection problem for two reasons: (1) The dimension of the covariate space is comparable, and often much larger, than the number of subjects in the study, and (2) the covariate space is highly structured, and in some cases it is desirable to incorporate this structural information in to the model building process. We approach this problem through the Bayesian variable selection framework, where we assume that the covariates lie on an undirected graph and formulate an Ising prior on the model space for incorporating structural information. Certain computational and statistical problems arise that are unique to such high-dimensional, structured settings, the most interesting being the phenomenon of phase transitions. We propose theoretical and computational schemes to mitigate these problems. We illustrate our methods on two different graph structures: the linear chain and the regular graph of degree k. Finally, we use our methods to study a specific application in genomics: the modeling of transcription factor binding sites in DNA sequences.

KEY WORDS: Ising model; Markov chain Monte Carlo; Motif analysis; Phase transition; Undirected graph.

1. INTRODUCTION

Consider the standard multiple regression problem

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

where $\mathbf{Y}$ is $n \times 1$ variable response, $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)$ is a $n \times p$ matrix of covariates, and $\boldsymbol{\epsilon}$ is a $n \times 1$ error term and $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I)$. In this paper, we focus on variable selection for this model with (a) a very large number of covariates, possibly much larger than the sample size (i.e., the "large p" paradigm, West 2003), and (b) information about substantial structure among covariates which can help us in the model building process.

This scenario of variable selection in a high-dimensional structured covariate space appears often in modern applied statistics. Here we list a few motivating examples:

1. In cancer genomics, mutations and DNA copy number aberrations can now be detected in high throughput fashion along the genomic sequence. A common goal is to link certain features of the genomic profile (X) to clinical phenotypes (Y). Regression models, if employed for this task, would face thousands of covariates (the mutations along the genome sequence) with possibly only hundreds of patient samples. However, the fact that these noisy mutation measurements are spaced linearly along the genome sequence provides location information that should be considered in the model building process. It is often reasonable, for example, to assume that adjacent measurements on the chromosome are both assaying the same underlying genetic defect, and thus should be grouped when added to the model.

2. In functional MRI (fMRI) studies of the brain, fMRI images are collected while subjects are assessed in the performance of tasks (Y). Then, the two-dimensional and three-dimensional images are scanned for regions of the brain that are associated with task performance. The images are often very large, containing more than thousands of voxels. The covariates in this case are voxel intensities, and in variable selection, our goal is to select voxels that are associated with Y. Since true signals usually represent connected regions in the brain, the smoothness of the signal in space should be incorporated into the variable selection process.

3. Gene expression can now be quantified at the genomic scale using technologies such as microarrays. With this data and available genomic sequence data, there has been much effort in the statistical modeling of the dependence of gene expression on promoter sequence composition. Linear regression models have been applied to this problem, with the response being gene expression, and the covariates being the counts of certain word patterns in the upstream promoter sequence of the gene. The words that are selected in the model may be binding sites for transcription factors. If we let the set of potential covariates be all L length words, then $p = 4^L$, which, for example, would be 16,384 for $L = 7$. Usually, n would be a subset of all of the genes in the genome, that is comparable to p. In this problem, we are aided by the fact that, due to the degeneracy of transcription fact binding sites, true motifs can be represented by words that are clustered by Hamming distance. Similar words often have similar effects on expression. It is this information that we would like to incorporate into the model building process.

In all of the above examples, the known structure among the large number of covariates can be represented by an undirected graph: the structure in the DNA copy number data can be represented by a one-dimensional linear chain; that in the fMRI data by a two-dimensional or three-dimensional lattice; and that in the motif data by a regular graph of degree L (more

detailed discussion is given in Section 5). Bayesian paradigm is a natural choice to incorporate such prior graphical structure. For example, Bayesian multivariate sparse latent factor model (West 2003) provides a flexible platform for introducing prior design-dependent covariate structure in feature selection in high-dimensional settings. Our focus is to identify important covariates instead of latent factors in this paper, and thus we adopt the Bayesian "spike and slab" approaches to variable selection (e.g., George and McCulloch 1993, 1997; Brown, Vannucci, and Fearn 1998; Ishwaran and Rao 2003, 2005a; Clyde and George 2004; and reference therein). The basic idea behind this framework is to define latent variables $\gamma = (\gamma_i : 1 \le i \le p)$, where γ_i is the indicator of whether covariate i is included in the model. Then, Markov chain Monte Carlo (MCMC) methods are used to stochastically approximate the posterior distribution of γ given the data. For a detailed comparison of Bayesian and frequentist penalized regression approaches, see Ishwaran and Rao (2005a). These MCMC based procedures involve extensive computing and have been traditionally applied to regression problems where p is not too large, although recently they have been applied with some success to high-dimensional problems (Ibrahim, Chen, and Gray 2002; Ishwaran and Rao 2003, 2005b; Tadesse, Sha, and Vannucci 2005). The small sample size and the high dimensionality in these problems render the variable selection problem difficult. In this paper, we introduce dependence in the γ's, with the effect of guiding the Markov chain to effectively search over a smaller set of configurations in the γ's—configurations that are smooth with respect to an underlying graph. Thus, instead of the set of 2^p possible models, the search is biased for a much smaller subset, depending on the graph structure. The main thrust of this paper is to use a class of Ising priors for the latent variables γ to flexibly incorporate the covariate space structure and improve stochastic model selection, and to provide guidance on how to avoid some of the consequent complications when p is large.

Graphical models have been extensively used in Bayesian methodology for other types of problems, such as segmentation and smoothing. For example, hidden Markov models assume a linear graph, and are very useful for segmentation of one-dimensional data. Two-dimensional to three-dimensional lattices have been used for the smoothing of fMRI data (Smith and Fahrmeir 2007). Informative priors for related covariates (e.g., interactions, grouped covariates), which can be viewed as overlaying on undirected acyclic graphs, were also discussed before by Chipman (1996). However, formal methods for graphical representation of substantive structural information among covariates in Bayesian variable selection, especially in high-dimensional settings, have since received relatively little attention. When p becomes large, that is, in the thousands, many new theoretical and computational issues arise for certain graph structures, the most interesting and problematic of which is the phenomenon of phase transitions: Certain global characteristics of the distribution of γ, such as the model size $\gamma_1 + \cdots + \gamma_p$, undergo a dramatic change given an infinitesimal change in the hyperparameters. For both efficiency and computational tractability, it is critically important to understand the phase transition behavior of the distribution of γ, and to avoid it. Such phase transition behavior in Ising models has been explored at great length in statistical physics. To our knowledge, this issue

has not been previously studied in the context of Bayesian variable selection. In Section 3, we give guidance for choosing the hyperparameters to avoid the phase transition behavior in high dimensions when the prior distribution on γ is exchangeable. This method can be applied to problems where there is underlying symmetry in the covariate space, such as the three examples listed at the beginning of this section. Exchangeability in prior covariate structure is desirable if a priori we do not want to bias our procedure towards the inclusion of any particular covariate.

As one may expect, in high-dimensional settings one of the most important determining factors in the practicality of a Monte Carlo algorithm is its computational efficiency. In this paper, we adopt the Gibbs sampling algorithms, as first suggested by George and McCulloch (1993). We discuss the computational challenges that arise in this method, and implement an efficient algorithm which we use to analyze a high-dimensional dataset where $p > 8000$ in Section 5.

The rest of the paper is organized as follows. Section 2 describes the formulation of the general Ising prior. Section 3 discusses the issue of hyperparameter selection, with emphasis on phase transition behavior. Section 4 presents simulation studies under a linear chain prior. Section 5 presents a real application to the modeling of transcription factor binding sites in DNA sequences. Section 6 concludes with a discussion.

2. FORMULATION OF GENERAL MODEL

2.1 Ising Prior for Covariate Spaces

Let the observed data be $\mathbf{X}$ and $\mathbf{Y}$ for which we assume the simple linear model (1) as described in the Introduction. As mentioned before, the Bayesian variable selection method relies on introducing a latent variable $\gamma_i \in \{0, 1\}$ for each covariate that indicates whether this covariate is included in the model. The prior distribution for the regression parameters $\boldsymbol{\beta}$ is assumed to depend on $\gamma = (\gamma_1, \ldots, \gamma_p)'$ as follows: given γ, β_i are independent with conjugate Gaussian mixture priors

$$\beta_i | \gamma_i \sim (1 - \gamma_i)I_0 + \gamma_i N(0, \sigma^2 v^2), \qquad (2)$$

where I_0 is a point mass at 0. For the residual variance σ^2, the inverse gamma (IG) conjugate prior is often assumed

$$\sigma^2 | \gamma \sim IG(\nu/2, \nu\lambda/2).$$

When $\nu = 0$, the IG prior reduces to a flat prior, which is adopted in this paper. With certain prior being further assumed for γ, the variable selection is then based on a stochastic search in the posterior covariate spaces $\gamma | \mathbf{Y} \in \{0, 1\}^p$ given the data. The prior for γ is traditionally assumed to be iid Bernoulli, which is equivalent to assuming the covariates are independent a priori. In other words, the prior information of structure in $\mathbf{X}$ is not incorporated. Intuitively, proper incorporation of such information would improve stochastic search of the covariate spaces. In this paper, we propose a general Ising prior for γ and investigate its consequences under the high-dimensional scenario.

We assume that the covariates $i = 1, \ldots, p$ lie in an undirected graph which can be represented by an edge set $\mathcal{E} = \{(i, j) : 1 \le i \ne j \le p\}$. Given this graph, let $\mathbf{a} = (a_1, \ldots, a_p)'$ be

a vector and $\mathbf{B} = (b_{i,j})_{p \times p}$ be a symmetric matrix of real numbers where $b_{i,j} = 0$ for all $(i, j) \notin \mathcal{E}$. Then, we assume the Ising prior distribution for $\boldsymbol{\gamma}$:

$$P(\boldsymbol{\gamma}) = e^{\mathbf{a}'\boldsymbol{\gamma} + \boldsymbol{\gamma}'\mathbf{B}\boldsymbol{\gamma} - \psi(\mathbf{a},\mathbf{B})}, \qquad (3)$$

where $\psi(\mathbf{a}, \mathbf{B})$ is the normalizing constant:

$$\psi(\mathbf{a}, \mathbf{B}) = \log\left(\sum_{\boldsymbol{\gamma} \in \{0,1\}^p} e^{\mathbf{a}'\boldsymbol{\gamma} + \boldsymbol{\gamma}'\mathbf{B}\boldsymbol{\gamma}}\right).$$

The constant $\psi(\mathbf{a}, \mathbf{B})$ is referred to as the partition function in statistical physics. Without loss of generality we assume that $a_i < 0$. If $\mathbf{B}$ were 0, then $\psi(\mathbf{a}, \mathbf{0}) = \sum_{i=1}^{p} \log(1 + e^{a_i})$, but in general there is no closed form for ψ.

In the Ising prior (3), the hyperparameters $\mathbf{a}$ control the sparsity of $\boldsymbol{\gamma}$ and the entries in $\mathbf{B}$ control the smoothness of $\boldsymbol{\gamma}$ over $\mathcal{E}$. Often, there is underlying symmetry in the covariate space such that the prior distribution on $\boldsymbol{\gamma}$ should be exchangeable, that is, for any permutation π of $\{1, \ldots, p\}$, the law of $\boldsymbol{\gamma}$ is equal to the law of $\boldsymbol{\gamma}(\pi) = (\gamma_{\pi_1}, \ldots, \gamma_{\pi_p})$. Under this setting, we do not favor a priori the inclusion of any covariate into the model. Thus, the graph must be regular, that is, each vertex has the same degree, and $\mathbf{a} = a(1, 1, \ldots, 1)$. The hyperparameters $\{b_{ij}\}$ represent the prior belief on the strength of coupling between the pairs of neighbors (i, j). Larger b_{ij} means tighter coupling. When $\mathbf{B} = 0$, the prior is back to iid Bernoulli. Further restrictions on b_{ij} are often placed to reduce the number of hyperparameters. For example, with lack of specific prior information on the strength of connection between each pair of neighbors, it is natural to assume b_{ij}'s to be constant. Then $(\mathbf{a}, \mathbf{B})$ reduce to two hyperparameters (a, b). In many problems, b_{ij} is not constant, but exchangeability implies that $\sum_j b_{ij}$ is constant across vertices. An example of nonconstant b_{ij} is given in Section 5.

The utility of this general Ising model owes to the fact that it is easily adaptable to a wide variety of problems. We will illustrate this by presenting two examples with different graph (covariate) structure in Sections 4 and 5.

2.2 Gibbs Sampling of $f(\boldsymbol{\gamma}|\mathbf{Y})$

To sample from $f(\boldsymbol{\gamma}|\mathbf{Y})$, we adopt the Gibbs sampling scheme that samples directly from the ergodic Markov chain: $\boldsymbol{\gamma}^0, \boldsymbol{\gamma}^1, \boldsymbol{\gamma}^2, \ldots$. When the average model size is sparse, each update sweep of $\boldsymbol{\gamma}$ in this scheme can be accomplished in linear time.

Let $\gamma_{(-i)} = \{\gamma_j : j \neq i\}$; $I_{(-i)}$ be the set of indices $\{\gamma_j = 1 : j \neq i\}$; $I_i = I_{(-i)} \cup \{i\}$; $p_i = |I_i|$ and $p_{(-i)} = |I_{(-i)}|$. For the prior distribution (3), there is a simple form for the conditional distribution

$$P(\gamma_i | \gamma_{(-i)}) = \frac{e^{\gamma_i(a + b\sum_{j \in I_{(-i)}} \gamma_j)}}{1 + e^{a + b\sum_{j \in I_{(-i)}} \gamma_j}}.$$

The posterior distribution of $\boldsymbol{\gamma}$ given the data can be decomposed by Bayes formula,

$$P(\gamma_i = 1 | \gamma_{(-i)}, \mathbf{Y})$$

$$= \frac{P(\gamma_i = 1 | \gamma_{(-i)})}{P(\gamma_i = 1 | \gamma_{(-i)}) + F(i | \gamma_{(-i)})^{-1} \cdot P(\gamma_i = 0 | \gamma_{(-i)})}, \qquad (4)$$

where $F(i | \gamma_{(-i)}) = \frac{P(\mathbf{Y} | \gamma_i = 1, \gamma_{(-i)})}{P(\mathbf{Y} | \gamma_i = 0, \gamma_{(-i)})}$ is the Bayes factor and can be explicitly computed for the linear regression model under the

priors $\boldsymbol{\beta}$ and σ specified in the previous section. Specifically, integrating out $\boldsymbol{\beta}$ and σ, we have

$$F(i | \gamma_{(-i)})$$

$$= v^{-1} \cdot \frac{|A_{(-i)}|^{1/2}}{|A_i|^{1/2}} \cdot \left(\frac{Y'Y - Y'X_{I_{(-i)}} A_{(-i)}^{-1} X'_{I_{(-i)}} Y}{Y'Y - Y'X_{I_i} A_i^{-1} X'_{I_i} Y}\right)^{n/2}, \qquad (5)$$

where $A_i = X'_{I_i} X_{I_i} + v^{-2} I_{p_i}$ and $A_{(-i)} = X'_{I_{(-i)}} X_{I_{(-i)}} + v^{-2} I_{p_{(-i)}}$.

Hence, one can sample directly from the posterior distribution of $\boldsymbol{\gamma}$ by constructing a Markov chain on $\{0, 1\}^p$ where at each iteration, an index is picked, say i, and γ_i is sampled from $P(\gamma_i | \gamma_{(-i)}, \mathbf{Y})$ using Equation (4). The index i can either be picked in a fixed order, or randomly.

Evaluating $F(i | \gamma_{(-i)})$ in (5) is the computationally intensive step during each iteration, because it involves inverting and calculating the determinant of the p_i by p_i matrix A_i. Note that one of the matrices $A_{(-i)}^{-1}$ and A_i^{-1} is in fact always available from the last iteration, and that A_i^{-1} can be obtained from $A_{(-i)}^{-1}$ by a low-rank update, which is an $O(p_i^2)$ operation. Then, each sweep through all of the γ_i's would be $O(pp_i^2)$ operation. This underlies the importance of limiting the size of the model during the sampling of $\boldsymbol{\gamma}$: even though the Bayesian formulation does not limit the model size in each iteration, it is desirable in the interest of computation for the model to be sparse. The model size is greatly affected by the choice of the hyperparameters, which will be discussed intensively in the next section. Various low-rank update algorithms can be developed using the numerical methods such as the Cholesky or LU decomposition of matrix. Details of the algorithm we used is given in the Appendix.

3. HYPERPARAMETER SELECTION

Hyperparameter selection is an important part of any type of Bayesian inference. In particular, for regression problems when p is large, the selection of hyperparameters need to be based not only on prior beliefs but also on considerations of computational efficiency. In this section, we focus on exploring two aspects of hyperparameter selection for the general model: (1) phase transition of the Ising prior, which induces critical slow down of the MCMC and dramatic change in model behavior; and (2) the influence of hyperparameter choice on the average size of the selected model, which dictates the computation time of Gibbs sampler.

3.1 Phase Transition of Ising Model

Under the general models (2) and (3), three hyperparameters need to be specified: the shrinkage v, the sparsity a and the smoothness b. In this paper, we focus our attention on the setting of the sparsity a and smoothness b of the underlying Ising model. The hyperparameter v is the prior variance of β_i given that $\gamma_i = 1$, and should be set based on expectations on the magnitude of β_i if covariate i were indeed a true predictor. Usually this information is not available, but we find that the following procedure yields satisfying results in practice: For every covariate, perform a single linear regression $Y \sim X_i$ to obtain a naive estimate of the coefficient $\hat{\beta}_i$, and then choose v based on the variance of the $\hat{\beta}_i$'s. Alternatively, one can adopt the approach of Ishwaran and Rao (2005a), which assumes a

hierarchical model where v itself follows a bimodal distribution with a spike at 0 and a continuous right tail. This gives a more adaptive model that removes subjectivity in the choice of v, and is shown by Ishwaran and Rao (2005a) to have a desired "selective shrinkage" property when used in combination with a rescaled spike and slab model. The hierarchical model for v should give better results in practice, but in this paper we choose the simpler prior which allows a more transparent study of the effects of the hyperparameters a and b.

The choice of hyperparameters (a, B) must consider any possible phase transition points in the Ising prior. It is widely known, for example, that Ising models on lattices of dimension ≥ 2 undergo transition between an ordered and a disordered underlying state at or near the phase transition boundary in terms of (a, B), leading to various dramatic consequences such as critical slow down of the MCMC. One immediate consequence in Bayesian variable selection for large p is the drastic change in the proportion of $\gamma_i = 1$, for example, from <1% to >90% near the phase transition boundary. This is illustrated by a simple example in Figure 1, which assumes that $b_{ij} = b$ and shows the expected proportion of $\gamma_i = 1$ versus b of an Ising model defined on a 6 degree regular graph with $p = 8192$ vertices and a fixed at 4. As one can see, the model size increases gradually from 150 to 200 as b increases until b reaches 1.35, where the model size suddenly jumps to over 8000 with small change in b. After b passes 1.4, the model size becomes stable again at around 8000. That is, the Ising model undergoes phase transition near the pair of hyperpameters $(a, b) = (-4, 1.35)$. Since the computational cost of sampling from the posterior of γ is of quadratic order of the model size, (a, b) must be chosen to avoid the phase transition point and guarantee a small average model size.

The main difficulty in analyzing a high-dimensional Ising model lies in the analytical intractability of the partition function $\psi(a, B)$, due to the many combinatorial interaction terms when summing over all states, that is, $\sum_{(i,j) \in \mathcal{E}} \gamma_i \gamma_j$. Nevertheless, the behavior of an Ising model on a wide class of regular graphs can be approximated by mean field theory (for a nice overview, see Yedidia 2001), which are useful in providing ballpark estimates of certain quantities, such as model size, clumping behavior (i.e., $E[\sum_{(i,j) \in \mathcal{E}} \gamma_i \gamma_j]$), and phase transition point. The main idea of mean field theory is to replace all interactions

Phase transition of a hypercube prior (p=8192, k=6, a=−4)

Figure 1. Phase transition (in terms of model size) of Ising model.

to any γ_i with an average interaction, which becomes exact as the dimension of the graph goes to infinity. This yields useful approximations for the partition function, which, as the normalizing factor of an exponential family, encodes many properties of the joint distribution of γ. For example, the derivative of the partition function with respect to a gives the mean of γ. As shown in Appendix A.2, the phase transition boundary of the Ising model with exchangeable distribution can be studied by examining the set of mean field approximations to the partition function $\psi(a, B)$, which can be expressed as $\min_t \phi(t)$, where

$$\phi(t) = \log(1 - t) - \left(a + \log\frac{1-t}{t}\right)t - kbt^2, \qquad 0 < t < 1,$$
(6)

where $k = \sum_j b_{ij}$, which, due to exchangeability, does not rely on i. To minimize $\phi(t)$, we look for solutions $\hat{t}$ to

$$\frac{d\phi}{dt} = -\log\left(\frac{1-t}{t}\right) - a - 2kbt = 0,$$
(7)

that satisfy $\frac{d^2\phi}{dt^2} = \frac{1}{t(1-t)} - 2kb > 0$. These solutions can be easily found numerically. To study them qualitatively, the left panel of Figure 2 shows the two sides of Equation (7) for varying kb. The intersection of the lines and the logit function are possible solutions $\hat{t}$ for given values of (a, kb). The nature of the solutions can be described as follows:

1. When $a > -2$: there is one minima of $\phi(t)$.
2. When $a = -2$: there is one inflection point (i.e., $\frac{d^2\phi}{dt^2} = 0$), $t^* = \frac{1}{2}$.
3. When $a < -2$: let the two solutions to equation $\text{logit}(t) = a + \frac{1}{1-t}$ be t_1^* $(> 1/2)$ and t_2^* $(< 1/2)$. Then when $\frac{1}{t_2^*(1-t_2^*)} < 2kb < \frac{1}{t_1^*(1-t_1^*)}$, there are two minima and one maxima of $\phi(t)$; when $2kb = \frac{1}{t_2^*(1-t_2^*)}$ or $\frac{1}{t_1^*(1-t_1^*)}$, one minima and one inflection point; when $2kb < \frac{1}{t_2^*(1-t_2^*)}$ or $2kb > \frac{1}{t_1^*(1-t_1^*)}$, one minima.

Therefore, for any given $a < -2$, the mean field approximate $\phi(t)$ transits between unimodal and multimodal states at $b_i^* = \frac{1}{2kt_i^*(1-t_i^*)}$ $(i = 1, 2)$, which are the phase transition points. The right panel of Figure 2 shows these regions in the $(a, 2kb^*)$ plane. In theory, for any given a, any b that is above the solid line $[>b_1^*(a)]$ or below the dashed line $[>b_1^*(a)]$ avoids phase transition in the Ising model. However, the model is only sparse for b below the dashed line. Thus, because of our a priori belief in a sparse model, and to limit the model size for computational efficiency, we always choose b that is below the dashed line in applications.

We have derived a ballpark estimate of the phase transition boundary for exchangeable Ising prior defined on regular graphs using mean field approximation. There is no analytical solution for the phase transition points for the posterior distribution. In the next section, we derive some heuristic guidelines for choosing hyperparameters to avoid the phase transition point when sampling from the posterior distribution in the scenario where the sample size is large and the true model is sparse.
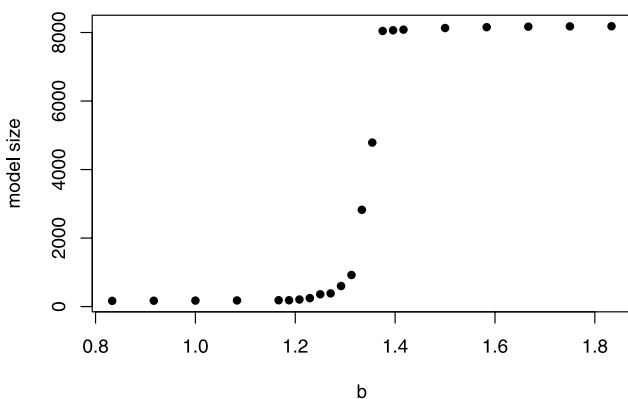
Graphic illustration of equation logit(t)=a+2kbt **Phase transition boundary (a,b*(a)) of a hypercube model**
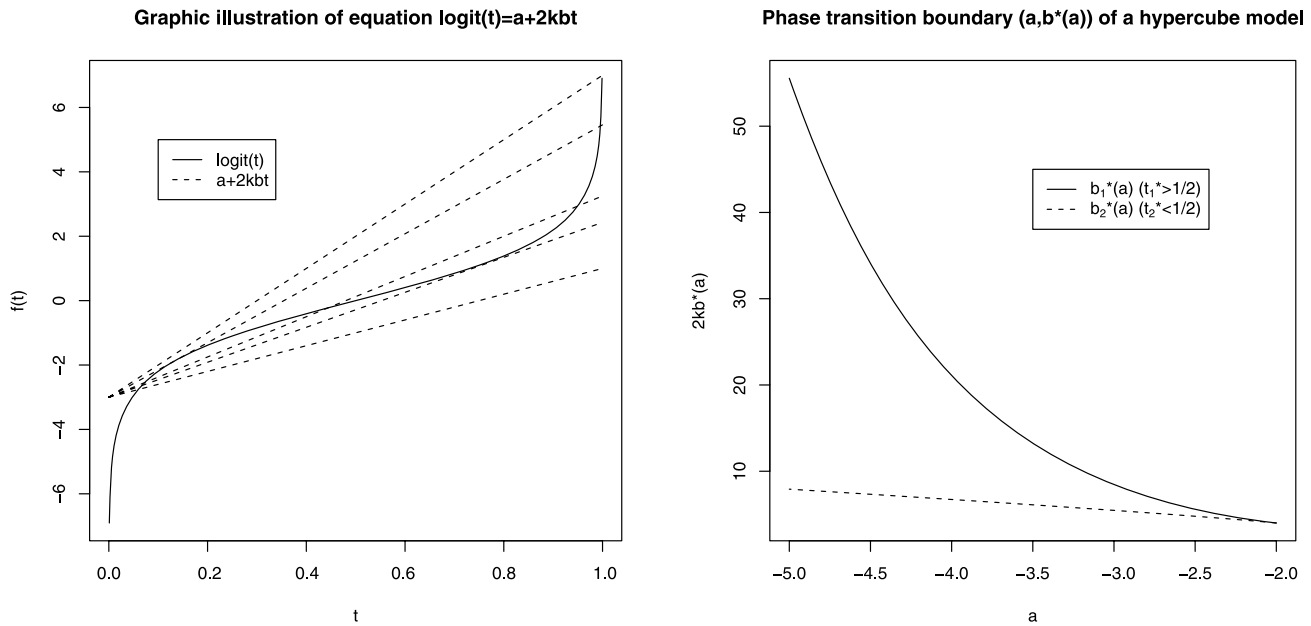
Figure 2. Phase transition boundary of Ising model.

3.2 Posterior Model Size

We now examine the influence of hyperparameter choice on the posterior model size, which depends not only on a and b but also on the hyperparameter v, the number of data points n, and the correlation structure within $\mathbf{X}, \mathbf{Y}$. Taking the log of the Bayes factor (5), we have

$$\log F\big(i|\gamma_{(-i)}\big) = -\log v + \frac{1}{2}\log\big(|A_{(-i)}|/|A_i|\big)$$
$$+ \frac{n}{2}\log(1 + \Delta/n\hat{\sigma}^2),$$

where $\Delta = Y'(X_{I_i}A_{(i)}^{-1}X_{I_i}' - X_{I_{(-i)}}A_{-i}^{-1}X_{I_{(-i)}}')Y$ is the difference in sum of squared error between the posterior mean fit of the smaller model and that of the larger model, and $\hat{\sigma}^2 = Y'(I - X_{I_i}A_i^{-1}X_{I_i})Y$ is an estimate of the variance σ^2. The second term $\log(|A_{(-i)}|/|A_i|) = \log n + C_v(\mathbf{X})$, where $C_v(\mathbf{X}) = O(1)$ depends on the correlation structure within $\mathbf{X}$. For large n, and assuming that the true $\beta_i = 0$, the third term $\Delta/\hat{\sigma}^2$ is approximately chi-square distributed, giving us the approximation

$$\log \frac{P(\gamma_i = 1|\gamma_{(-i)}, \mathbf{Y})}{P(\gamma_i = 0|\gamma_{(-i)}, \mathbf{Y})} \approx a + b \sum_{(i,j)\in\mathcal{E}} \gamma_j - \log v - \log n$$
$$+ C_v(\mathbf{X}) + Z^2(\mathbf{X}, \mathbf{Y})/2, \quad (8)$$

where $Z(\mathbf{X}, \mathbf{Y}) \sim \mathrm{N}(0, 1)$. The terms $C_v(\mathbf{X})$ and $Z(\mathbf{X}, \mathbf{Y})$ introduce higher than second-order interactions among γ_i, making it difficult to analyze the phase transition behavior in the posterior distribution of $\boldsymbol{\gamma}$. However, (8) still gives useful insights, the most important of which is the following: Let $n \to \infty$, and a, b, v remain fixed. If the true model contains $c \ll p$ predictors, then phase transition will not occur. This is because in the right-hand side of (8), $a - \log v - \log n \to -\infty$ while the interaction terms remain bounded for all but c of the predictors. However, when sample size n is moderate and p is large, $\log n$ is often not large enough to preclude phase transition behavior, in which case we have found the following heuristics to be useful:

1. The posterior model size decreases with increasing v, with a C-fold increase in v equivalent to a $\log C$ decrease in a.
2. The posterior model size decreases with increasing sample size, with a C-fold increase in sample size equivalent to a $\log C$ decrease in a.

When the number of covariates is large, we assume that the bulk of them follow the null model. Such sparse models are easier to interpret, and make Gibbs sampling in spike and slab model selection procedures computationally feasible for high-dimensional data. Thus, the above approximation provides useful guidelines in quantifying the effect of v and n, relative to (a, b), on the posterior distribution of $\boldsymbol{\gamma}$. We found the following to be a good strategy: First choose v based on the expected signal magnitude β, then choose b based on desired smoothness. Finally, based on v, b, and n, choose a based on (1–2) above and the mean field approximations in Section 3.1 to avoid phase transition and obtain the desired posterior model size.

4. SIMULATIONS: LINEAR CHAIN PRIOR

The linear chain prior, where $P(\gamma_i|\gamma_{(-i)}) = P(\gamma_i|\gamma_{i-1})$, is a simple example of the general model (3). Smoothness of models along the linear chain prior can be easily visualized by plotting the posterior marginal distribution $P(\gamma_i = 1|\mathbf{Y})$ versus the linear ordering i. It is well known that phase transition does not occur under this setting (see, e.g., Brusch 1967). Also, closed form formulas are available for marginal probabilities on γ_i's. Due to its simplicity and convenience for visualization, we start with simulations under the linear chain prior assumption to examine the basic question: When and how does graph-based smoothing improve the accuracy of variable selection in regression models?

We will simulate the data $(\mathbf{X}, \mathbf{Y})$ from two different models, summarized in Figure 3. Under the first model, X has a direct effect on Y, with the effect being smooth along the underlying graph. We will see that, not surprisingly, our method produces
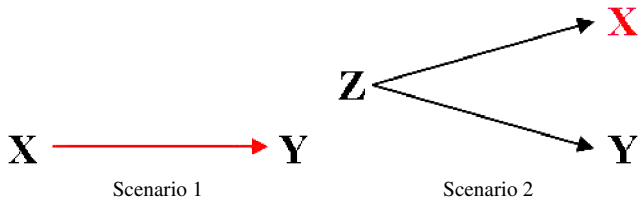
Figure 3. Design of simulation studies. Scenario 1: X_i has a direct effect on Y, with the effect being smooth in i. Scenario 2: X and Y related through Z. X is smooth in i. The online version of this figure is in color.

more accurate model estimates than the independent prior assumption. In the second simulation, X does not have a direct effect on Y, but the two are related through a latent variable Z. X itself, rather than the relationship between X and Y, is smooth. We will see that under this more subtle scenario, the Ising prior improves accuracy if the smoothness in X is strong compared to the strength of the effect of Z on Y.

4.1 The Linear Chain Prior

First, we quantify the effects of the hyperparameters in the linear chain prior in more detail. In a linear chain, each vertex γ_i has two neighbors γ_{i-1} and γ_i. To make this model exchangeable, we circularize the chain by adding an edge between γ_1 and γ_{p+1}. To reflect the linear ordering of the covariates, we assume that $\boldsymbol{\gamma}$ is Markov with transition matrix

$$Q = \begin{pmatrix} q_0 & 1-q_0 \\ 1-q_1 & q_1 \end{pmatrix},$$

and that $\gamma_1 \sim \pi$, where $\pi = (\frac{1-q_1}{2-q_0-q_1}, \frac{1-q_0}{2-q_0-q_1})$ is the stationary distribution with regards to Q. The above formulation is equivalent to the following 1D Ising model

$$P(\gamma_i = 1 | \gamma_{i-1}, \gamma_{i+1}) = \frac{e^{a+b(\gamma_{i-1}+\gamma_{i+1})}}{1 + e^{a+b(\gamma_{i-1}+\gamma_{i+1})}}, \quad (9)$$

where $a = \log(r/w_0^2)$, $b = \log(w_1 w_0)$, and

$$r = \frac{1-q_0}{1-q_1} = \frac{\pi_1}{\pi_0}, \qquad w_0 = \frac{q_0}{1-q_1}, \qquad w_1 = \frac{q_1}{1-q_0}. \quad (10)$$

This parameterization has an intuitive interpretation: r is the prior odds of $\gamma_i = 1$, w_0 reflects the increase in probability of $\gamma_i = 0$ if we knew that $\gamma_{i-1} = 0$, and w_1 is the increase in probability of $\gamma_i = 1$ if we knew that $\gamma_{i-1} = 1$. Note that if $w_1 = 1$, then the γ_i's would be iid. The pair (r, w_1) completely specifies the model. With r kept fixed, the expected size of the model [i.e., $pP(\gamma_i = 1)$] remains fixed. In our simulations, we compare between models with fixed expected size and varying smoothness, and thus use (r, w) for parameterization instead of (a, b).

4.2 Simulation Model 1: Smooth in γ

First consider the following simulation model:

$$Y_k = \sum_i X_{k,i} \beta \gamma_i + \epsilon_{k,i}, \qquad i = 1, \dots, p; k = 1, \dots, n; \quad (11)$$

where $\epsilon_i \sim N(0, 1)$. We let $p = 1000$ and $n = 100$, and set γ to be the piecewise constant vector $\gamma_i = I(i \in [245, 260] \cup [745, 760])$. The true β is between $(0.1, 1)$ and can vary within a block. In particular, we explore the mixture of strong and

weak signals: the signals at even indices (β_1) are strong and at odd indices (β_2) are weak. For covariates $\mathbf{X}$, we assume $X_i \sim N(0, 1)$ and study two correlation structures: (1) independent $\mathbf{X}$: X_i are iid; (2) correlated $\mathbf{X}$: in the blocks [241, 265] and [741, 765], let $\text{cor}(X_i, X_j) = 0.75 - 0.03|i-j|$, that is, the piecewise correlation between two covariates is negatively proportional to their distance (maximum 0.75). To add noise, we also let $\mathbf{X}$ be correlated as $\text{cor}(X_i, X_j) = 0.4 - 0.02|i-j|$, in two blocks that do not contain true signal: [41, 60] and [941, 960]. We varied v, r, and w while keeping the stationary distribution π fixed. For each setting of hyperparameters, we ran the Gibbs sampler 10 times with random start in $\boldsymbol{\gamma}$. Each run has 2000 iterations with the first 1000 iterations as burn-in. It takes 2 minutes to run 2000 iterations with average posterior model size of 40 on a Sun Unix V880 with 1200 Mhz CPU. In all of our experiments, the 10 simulations lead to highly similar posterior summary statistics.

For high-dimensional covariate spaces, the traditional posterior summary statistics of counting the occurrence of each particular posterior model is infeasible because any model is most likely to be sampled only once in a MCMC with workable length. So here we focus on the posterior marginal probabilities $P(\gamma_i = 1|\mathbf{Y})$, an approach used by Smith and Kohn (1996) and Ibrahim, Chen, and Gray (2002) among others. These posterior marginals are obtained by dividing the number of iterations where $\gamma_i = 1$ over the total number of iterations excluding the burn-in period. This choice is motivated by its simplicity of interpretation and the fact that the posterior of γ is a natural by-product of our Gibbs sampler, which marginalizes over β. Even though there is no rigorous way to threshold the posterior marginals, we choose to select the variables corresponding to the top M (e.g., 100) marginals in our application, a common procedure that appears to give satisfying results.

With 2000 iterations, the marginal posterior probabilities in our simulation are very stable over random restarts, implying convergence. To better visualize and summarize the comparison between models, we further compute the ROC curve as follows: only those covariates i with $P(\gamma_i = 1|\mathbf{Y})$ greater than a threshold are deemed positives, and those below the threshold are deemed negatives, then the ROC curve reflects the pair of (true positive rate, false positive rate) achieved by varying the calling threshold. The bigger area under the ROC curve (maximum 1), the better the discriminating power of the model.

The posterior marginal probability of $\boldsymbol{\gamma}$ of a representative simulation under the independent $\mathbf{X}$ model [true signal $(\beta_1, \beta_2) = (0.8, 0.4)$] is shown in Figure 4, where the true $\gamma_i = 1$ is labeled by lines. The corresponding ROC curves are shown in left panel of Figure 5 (only ROC curves are presented hereafter). The hyperparameters are fixed $v = 1$, $r = \pi_1/\pi_0 = 0.03$, and varying $w_1 = 1, 5, 7$, where $w_1 = 1$ corresponds to the independent Bernoulli prior. It is clear that in the simple independent X case the assumed Markov chain prior indeed yields significantly better results. The improvement becomes even more pronounced for harder tasks with weaker signal (smaller β). This pattern is consistently observed in each of our simulations under various settings of hyperparameters and signals. Under the correlated X model, performance of the Markov chain prior is similar to that of the independent prior for
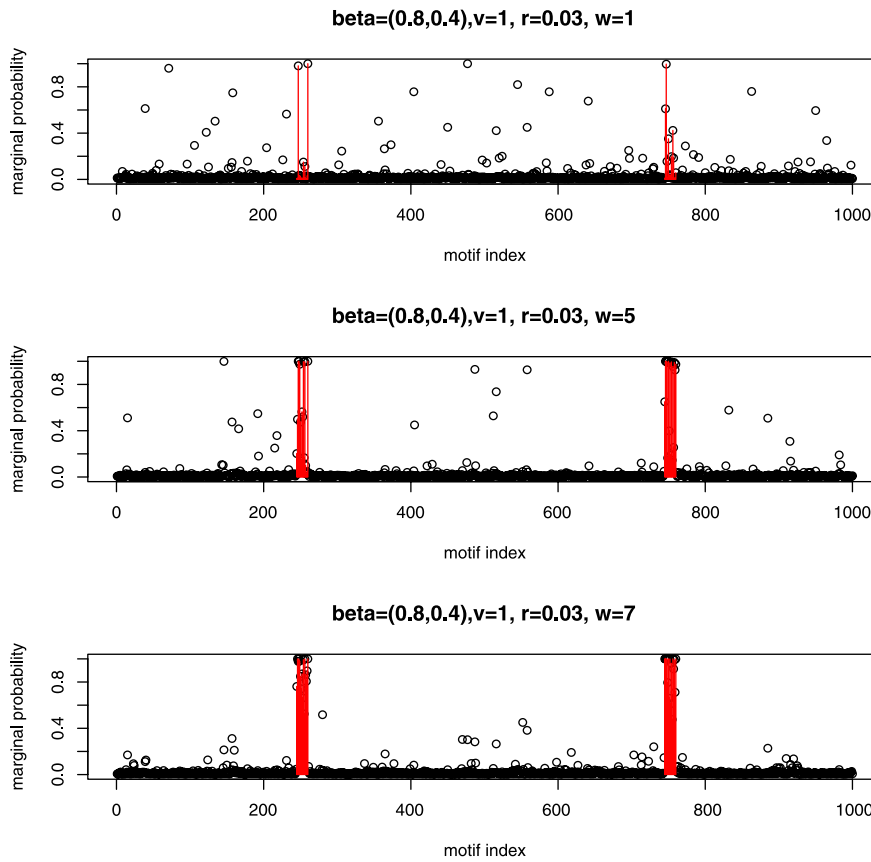
Figure 4. Marginal probability of $\boldsymbol{\gamma}$ under simulation model (11) and independent X. The online version of this figure is in color.

moderate to strong signal [$\max(\beta_1, \beta_2) > 0.3$], but consistently better for weak signal [$\max(\beta_1, \beta_2) < 0.3$]. The ROC curves of a representative simulation under the correlated $\mathbf{X}$ model with $(\beta_1, \beta_2) = (0.2, 0.1)$ are shown in the right panel of Figure 5, where the gain from the Markov chain prior is evident. We also experimented with other patterns of signals, for example, all β's in the same block are the same. The results are similar to what is represented above.
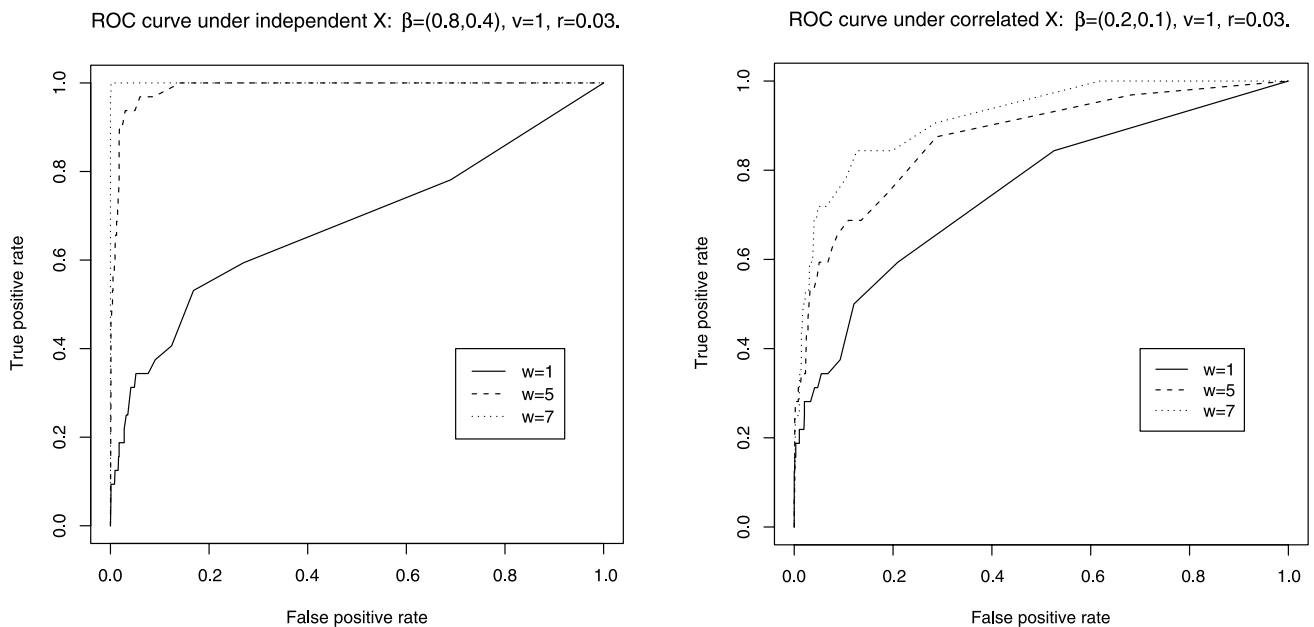

Figure 5. ROC curves under simulation model (11): independent X (left) and correlated X (right).

4.3 Simulation Model 2: Smooth in X

It is intuitively obvious that in simulation model (11), a smoothed model fit performs better: The truth agrees with the model! We now study a more complicated scenario where the relationship between consecutive covariates is more subtle. We let $\mathbf{X}_k = (X_{k,1}, \ldots, X_{k,p})$ be piecewise continuous:

$$\mathbf{X}_{k,i} = \delta Z_k I(i \in [i^* - L_{k,1}, i^* + L_{k,2}]) + \xi_{k,i}, \quad (12)$$

where $\xi_{k,i} \sim N(0, 1)$, $Z_k \sim \text{Bernoulli}(1/2)$, and $L_{k,1}$ and $L_{k,2}$ are independent Poisson random variables with mean μ_L. Thus, with probability $1/2$, $\mathbf{X}_k$ has a jump of magnitude δ centered at location i^*. The length of the jump in X_k is a Poisson random variable. Then, let the response Y depend only on whether a jump occurred at i^*:

$$Y_k \sim \beta Z_k + \epsilon_k.$$

Hence, $\mathbf{Y}$ is related to $\mathbf{X}$ only through the latent variable $\mathbf{Z}$, the indicator for a jump centered at i^*. The goal is to locate i^* by regressing Y and X.

Model (12) poses a much harder variable selection task than Model (11) because the effect is indirect (goes through Z). This means that a small underlying effect size (β) usually leads to poor performance of the Baysian variable selection procedure with any w. However, our simulations show that setting $w > 1$ consistently improves performance over $w = 1$. Figure 6 shows the covariate matrix X and its correlation structure (heatmap) for a typical simulation run.

We present the results under model (12) with $\delta = 0.35$, $\beta = 3.5$, and 10 jump locations in X, $i^* = (50, 150, \ldots, 950)$. Figure 7 shows the ROC curves of the posterior marginal probability of $\boldsymbol{\gamma}$ with fixed $v = 1$, $r = 0.02$ and varying $w = 1, 5, 7$. We can see that the smoothed prior for $\boldsymbol{\gamma}$ always outperforms the iid prior. The jump size at $\delta = 0.35$ is small, and pooling information across neighboring covariates in this case can help significantly in identifying the location of i^*. Here larger w
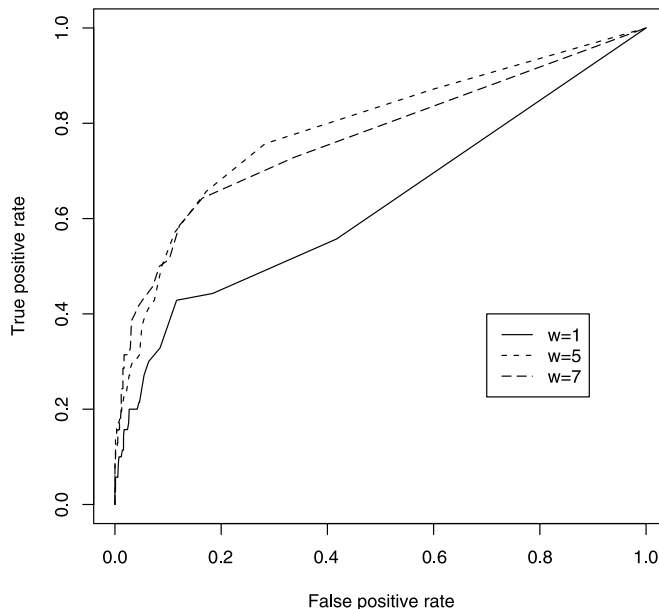

Figure 7. ROC curves under simulation model (12).

($w > 1$) does not necessarily result in better performance, which is not surprising because the extra signal gained from pooling information over a large neighborhood is countered by the extra noise introduced into the model.

5. APPLICATION TO DNA MOTIF FINDING: HYPERCUBE PRIOR

5.1 Background and Motivation

Transcription factors are proteins that regulate gene expression by binding to its surrounding sequence in the genome.
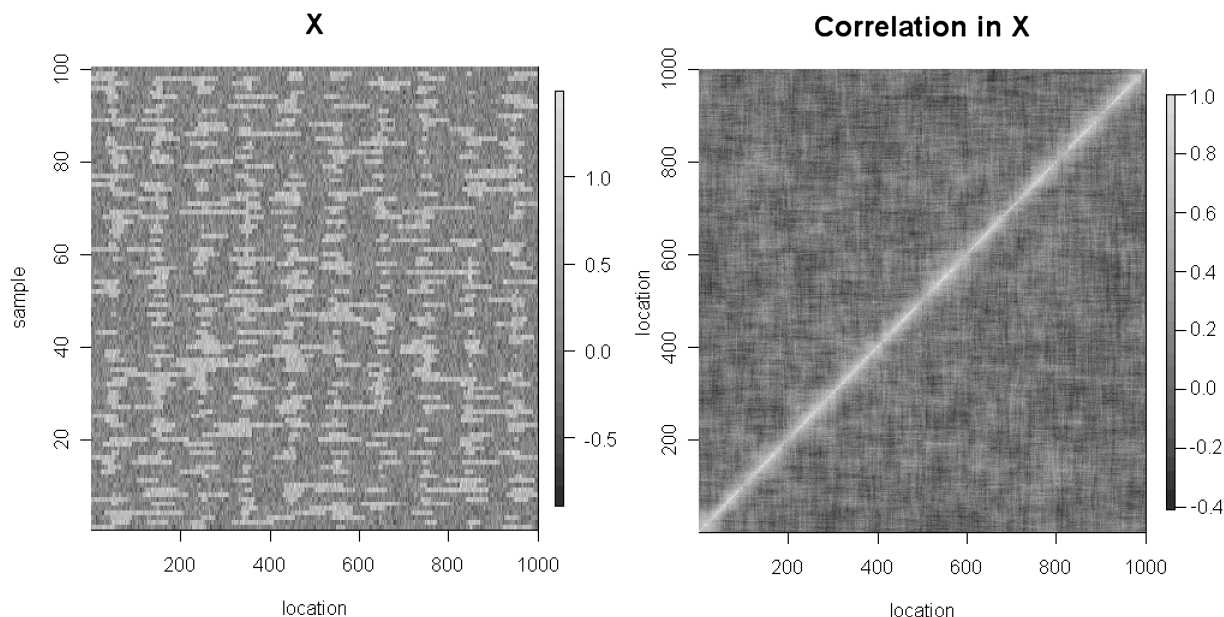

Figure 6. The heatmap on the left displays, in rows, the values of covariates X for simulation model 2 ($n = 100, p = 1000$). Notice the smoothness in X along the rows. The heatmap on the right displays the correlation structure in X.

Transcription factor binding sites (TFBS) usually contain low-entropy patterns called motifs. An important problem in biology is the modeling of the relationship between expression level of genes and the repertoire of motifs in their promoter sequences. Regression models have been applied to this problem in studies such as Bussemaker, Li, and Siggia (2001), Conlon et al. (2003), Tadesse, Vannucci, and Liò (2004), and Zhang, Wildermuth, and Speed (2008).

Transcription factors are usually degenerate, in the sense that words which are close together in Hamming distance are more likely to be alternative binding sites for the same transcription factor. The degeneracy of transcription factor binding sites have been modeled in a variety of ways, such as using position specific scoring matrices (PSSMs) and consensus sequences. Usually, a binding site is composed of one or multiple core sequences, which can tolerate very little variation, and flanking sequences which can take on different values. The strength of attraction of the transcription factor to the binding site depends on the flanking sequence. An example is the MCB motif, which regulates gene expression at the start of the S-phase in the yeast cell cycle. Its most common form is ACGCGT. The core sequence is the four bases in the center, CGCG, which cannot be changed. However, the flanking bases are allowed to wobble, with variants of MCB including TCGCGA and CCGCGT. Even though different transcription factor binding sites have different position specific base patterns, existing studies have shown that they share position-specific entropy patterns (Schneider et al. 1986; Mirny and Gelfand 2002; Moses et al. 2003). That is, if each position in the motif is modeled as an independent multinomial distribution over the alphabet $\{A, C, G, T\}$, then the entropy of this distribution is low in the middle 3–4 positions and high in the flanking sequence. This is due to the fact that each turn of the DNA helix encompasses 3.6 bases, and transcription factors usually contact DNA in its major or minor groove, which limits the size of the core sequence. Work by Kechris et al. (2004) have incorporated such prior knowledge on position-specific entropy to raise the sensitivity in algorithms for motif identification.

We will use linear regression models to find words whose presence in the promoter sequence is associated with various gene expression patterns. The response variable is a measurement of the strength of the expression pattern of each gene. The covariates are the counts of all words of length L in the promoter sequence of that gene. Therefore, the number of predictors are on the order of 4^L, and the genes used in the analysis usually number in the thousands. To reflect the fact that motifs should be clustered in Hamming distance, we model the words as vertices on a L-dimensional hypercube, with the edge weights b_{ij} chosen based on position-specific entropy obtained from previous studies. Below we give a detailed description of the model.

5.2 Model Description

Let $\mathcal{A} = \{A, C, G, T\}$ be the DNA alphabet, and let L be a fixed word length. We denote by $\mathcal{W} = \mathcal{W}_L = \mathcal{A}^L$ the set of all words of length L on $\mathcal{A}$. For any pair of words $w, w' \in \mathcal{W}$, let $d(w, w')$ be their Hamming distance, that is,

$$d(w, w') = \sum_{i=1}^{L} I(w_i \neq w_i').$$

Based on the studies of Schneider et al. (1986) and Kechris et al. (2004), we formulate the matrix B in the Ising prior based on the Hamming distance and the location of the mismatches between the pairs of words

$$B_{w,w'} = \begin{cases} 0, & d(w, w') > D \\ b \sum_{i=1}^{L} g_i I(w_i \neq w_i'), & d(w, w') \leq D, \end{cases} \quad (13)$$

where $g_i > 0$ is a weight corresponding to the ith position. The above model defines a L-d hypercube on vertices $V = \mathcal{W}_L$, where there is an edge between two words if they are within D of each other in hamming distance. If the two words are connected by an edge, then the weight on that edge depends on the position(s) of mismatch. As the studies Schneider et al. (1986), Kechris et al. (2004), Mirney and Gelfand (2002), and Moses et al. (2003) show, g_i should be small in the middle of the motif, and large in the flanking regions. The parameter b controls the strength of the clustering effect.

In the example below we let $L = 7$, which is long enough to cover the core region (3–4 bases) and a few flanking bases, but still allow computational tractability. We let $D = 1$ and

$$g_i = \begin{cases} 1, & i \in L_1 \\ 0, & i \in L_2, \end{cases} \quad (14)$$

where $L_1 = \{1, 2, 6, 7\}$ are the "flanking positions" and $L_2 = \{3, 4, 5\}$ are the "core positions." Thus, no mismatch is allowed in the core positions, and only 1 mismatch is allowed in the flanking positions. We chose this model because it is the simplest model that distinguishes between core and flanking regions, and we show in the next section that the simple structural information already substantially improves detection accuracy over the independent prior.

5.3 Analysis of Spellman et al. (1988) Data

As an illustration, we analyze the α-arrest yeast sporulation experiment of Spellman et al. (1998) to find motifs that are related to the cell cycle. This is a classic dataset that has been analyzed previously by many motif finding methods (Bussemaker et al. 2001; Tadesse, Vannucci, and Liò 2004; Zhang, Wildermuth, and Speed 2008). Previous regression based approaches have used as covariates either nondegenerate words, degenerate words on the IUPAC alphabet, or a known set of precurated PSSMs. A reliable list of precurated PSSMs is not always available, and the set of degenerate words using the IUPAC alphabet is too large (the IUPAC alphabet consists of 17 letters, thus the set of all words of length 7 on the IUPAC alphabet is $17^7 = 410,338,673$ instead of $4^7 = 16,384$). Thus, we find the approach of starting with nondegenerate words and using a graphical model to borrow strength between "neighboring" words to be more attractive.

This dataset consists of samples taken at 18 timepoints spanning two cell cycles. Using any single timepoint as the response variable in the regression is not sufficient in capturing the complexity of the experiment. We follow the approach suggested in Zhang, Wildermuth, and Speed (2008) and use the scores of the first principal component of the data, the loadings of which are plotted versus time in Figure 8. We used a list of 1600 genes in the regression, which includes the original 800 "cell cycle
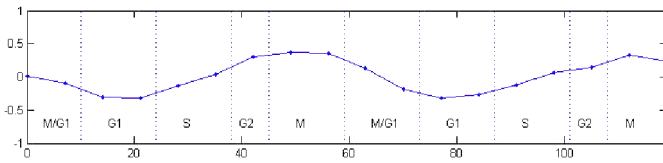
Figure 8. Loadings of the first principal component for yeast cell cycle dataset. The online version of this figure is in color.

genes" identified by Spellman et al. (1998) as well as 800 control genes that are not differentially expresses across time. A minor technical detail is that in yeast, a word and its reverse complement should be considered the same motif. Thus, there are 8192, instead of $4^7 = 16,384$, covariates, with each being the pair of words $\{w, w^{RC}\}$ where w^{RC}_{7-i} is the complement base of w_i for $i = 1, \ldots, 7$. It is straightforward to show that for length 7 words, these 8192 covariates still lie on a hypercube with degree $k = 6$ (a regular graph). Thus, its phase transition boundary can be directly obtained from the general results in Section 3.1. Specifically, we use the model in (14) with $a = -5$, $2kb = 10$ (i.e., $b = 0.83$), and $v = 1$. By (8), this is equivalent to a sparsity parameter of $a' = -\log n - \log v - 5 = -8.2$ for the posterior model, which gives a posterior model size of $\sim 47 \pm 5$. These values lie within the phase transition boundary.

Although yeast is one of the most well-studied organisms in terms of transcription regulation, much is still unknown about the possible forms of cell cycle motifs. Unless otherwise noted, we use as a gold standard the set of experimentally validated motifs in the Sachromyces cerevisiae Promoter Database (Zhu and Zhang 1999).

Figure 9 shows the histogram of the marginal probabilities $\log_{10} P(\gamma_i = 1 | \mathbf{Y})$. Due to the large size of the covariate space, and the sparsity of our model, most of the motifs (including some that are known to be biologically relevant to the cell cycle) have very low $\log_{10} P(\gamma_i = 1 | \mathbf{Y})$. However, many known cell cycle related motifs are ranked high in the list. Thus, as for the previous example, we find that it is more meaningful to filter motifs based on ranking or relative (rather than absolute) posterior marginal probability. For example, in the top $M = 100$ motifs, 29 have a neighboring motif in the hypercube that is also selected. We call such clusters of more than one selected motif that are connected in the hypercube graph islands. There are 12 islands in the top 100 motifs, listed in Table 1. Almost
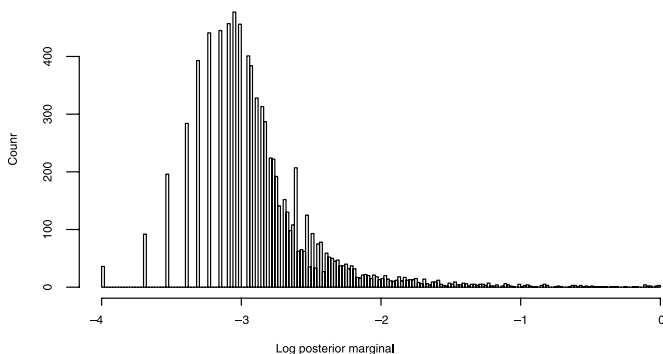
all known cell cycle regulatory motifs are part of an island, including MCB (ACGCGT), SCB (TTTTCGTG), SFF (TTGTTT), and SWI5 (GCTGG). The words that are grouped together in the same island are also known variants of the same TRBS. For example, it is known that TTTCGTG and TTTCGCG are the two most common alternative forms of the SCB motif, and that the first 'A' in the MCB motif ACGCGT can be replaced by other letters, such as a 'T.' Other than the known motifs, a few interesting candidates also appear in Table 1. The island of 4 motifs comprising GCCCGTT, GCCCGAT, GTCCGAT, GTC-CGCT are a putative MCM1 domains (Zhang, Wildermuth, and Speed 2008). MCM1 is an important regulator in the cell cycle, but due to the high degeneracy of its binding sites it is often missed by existing motif finding algorithms. For example, Bussemaker, Li, and Siggia (2001), which is the first paper on regression based modeling of this problem, can only detect this motif by considering motif pairs rather than singletons. However, due to the hypercube graphical structure, this cluster has quite a strong signal. Another interesting cluster is GAGAACG, GCGAACG, which contains the ABF/BAF1 site. BAF1 is known to be a regulator of genes involved in the cell cycle, including CDC19.

It is meaningful to compare the results obtained from the hypercube model to results obtained from the model that assumes prior independence of $\boldsymbol{\gamma}$. Out of the top 100 motifs in the independent model, there are eight islands comprising 19 different motifs, which are also listed in Table 1. The fact that these islands appear in the independent model, and that they include many of the known motifs of the cell cycle (MCB, SCB, SFF, and SWI5), is independent evidence that the graphical model based on Hamming distance is appropriate for analysis of motif data. However, without the underlying graphical model, weaker signals, such as the MCM1 cluster and the ABF/BAF1 site, are lost. The effect of the hypercube model can also be seen in the relative magnitude of the marginal probabilities. Known motifs, such as TACGCGT (MCB), TTTCGTG (SCB), TTGTTTA (SFF), TTGGTCG (MCM1) have a large increase in marginal probability under the hypercube model, the set of motifs that have a decrease in marginal probability are not enriched with known cell-cycle regulatory motifs.

6. DISCUSSION

Model building in high-dimensional covariate spaces with a priori known structure is a frequently met problem in modern statistics. In this paper, we have explored the use of Ising priors on the latent indicator variables $\boldsymbol{\gamma}$ under the framework of Bayesian variable selection. We proposed a general framework that can flexibly adapt to a large variety of problems. As illustration, we studied two scenarios in Sections 4 and 5. In both scenarios the assumed structure on the covariate space can be encoded into graphs, but the different nature of the graphs called for different approaches to hyperparameter selection. In the first example, the graph is a linear chain, which allows easy plotting and closed-form analysis. Of particular interest is the second example involving the hypercube prior, where the selection of hyperparameters need to take into consideration the phase transition behavior induced by the graph. We have found that mean field approximations are useful in this context. Avoiding phase transition and controlling the posterior model size is crucial for

Figure 9. Histogram of $\log_{10} P(\gamma_i = 1 | \mathbf{Y})$ for Spellman et al. yeast cell cycle dataset.

Table 1. Islands in top 100 motifs ranked by $P(\gamma_i = 1|\mathbf{Y})$ from hypercube model

	Independent model			Hypercube model	
	$P(\gamma_i = 1|\mathbf{Y})$	Name		$P(\gamma_i = 1|\mathbf{Y})$	Name
Island 1, 5 words:			**Island 1, 5 words:**		
GACGCGT	1	MCB	GACGCGT	1	MCB
TACGCGT	0.7876	MCB	TACGCGT	0.9262	MCB
GGCGCGT	0.711		GGCGCGT	0.7691	
TTCGCGT	0.1529		TTCGCGT	0.2284	
TTCGCGA	0.0982		TTCGCGA	0.1554	
Island 2, 2 words:			**Island 2, 2 words:**		
GCTGGTT	0.9418	Swi5	GCTGGTT	0.9589	Swi5
GCTGGAT	0.0916		GCTGGAT	0.2477	
Island 3, 2 words:			**Island 3, 4 words:**		
TTTCGCG	0.8678	SCB	**GCCCGTT**	**0.9547**	**MCM1**
TTTCGTG	0.6117	SCB	**GCCCGAT**	**0.1062**	
			GTCCGAT	**0.0633**	**MCM1**
			GTCCGCT	**0.097**	
Island 4, 2 words:			**Island 4, 2 words:**		
CTGCGCT	0.3865		TGTTTGT	0.8589	
CTGCGTT	0.0962	RME1	TGTTTTT	0.1202	STE12
Island 5, 2 words:			**Island 5, 2 words:**		
TCGCGTC	0.2053		TTTCGCG	0.8318	SCB
GCGCGTC	0.2017		TTTCGTG	0.79	SCB
Island 6, 2 words:			**Island 6, 2 words:**		
TTGGTCG	0.1029		CTGCGCT	0.4159	
TCGGTCG	0.0742	MCM1	CTGCGTT	0.1423	RME1
Island 7, 2 words:			**Island 7, 2 words:**		
GCCGACT	0.0992	BAS1	TAGCCAG	0.3352	
GCCGACG	0.0541	BAS1	TAGCCGG	0.1142	
Island 8, 2 words:			**Island 8, 2 words:**		
TTGTTTA	0.0941	SFF, ROX1	TCGCGTC	0.2332	
TTGTTTT	0.064	ROX1	GCGCGTC	0.1932	
			Island 9, 2 words:		
			GAGAACG	**0.1483**	
			GCGAACG	**0.063**	**ABF1, BAF1**
			Island 10, 2 words:		
			TTGTTTA	0.1409	SFF, ROX1
			TTGTTTT	0.0861	ROX1
			Island 11, 2 words:		
			TTGGTCG	0.1394	
			TCGGTCG	0.0958	MCM1
			Island 12, 2 words:		
			GCCGACT	0.1135	BAS1
			GCCGACG	0.0743	BAS1

computational feasibility of Bayesian variable selection algorithms in high dimensions, which is a main concern dictating the methods in this paper.

The inference in this paper is based on the latent variables via thresholding the posterior inclusion probabilities $P(\gamma_i = 1|\mathbf{Y})$, where the coefficients β is integrated out, an approach advocated first by Smith and Kohn (1996). Barbieri and Berger (2004) proposed to use instead the posterior median model [the model consisting of those variables with $P(\gamma_i = 1|\mathbf{Y}) \geq 50\%$], which they showed is predictively optimal. Under the spike and slab regression setting, the median model is equivalent to the posterior model (Barbieri and Berger 2004). Alternatively, Ishwaran and Rao (2005a) and Dey, Ishwaran, and Rao

(2008) proposed a procedure based on rescaling the responses Y, which is shown to have better finite sample performance. Furthermore, approaches based on thresholding posterior values of β's (which are not integrated out) (e.g., Ishwaran and Rao 2003, 2005a, 2005b), has also been shown to be very useful in high-dimensional settings.

Introducing the smoothing parameter b in the prior distribution for γ also increases the stickiness of the Markov chain, and thus causes slower mixing rate. However, in both the simulation and the real data example that we explored, the effect on mixing rate was not significant even for very large values of the smoothing parameter. Block-wise updating schemes, or modifications of the Swendsen–Wang algorithm proposed by Nott and

Green (2004) for variable selection, can be applied and may be useful when mixing rate becomes a concern.

L_1 penalized regression methods such as the fused Lasso (Tibshirani et al. 2005) and the group Lasso (Yuan and Lin 2006), as well as Markov random field models on the regression parameters β (Wei and Li 2007; Wei and Pan 2009) have been proposed for structured variable selection in high-dimensional settings. However, the underlying model assumptions of these methods are very different than those proposed in this paper: Their methods assume smoothness in β while ours assumes dependency in γ. Intuitively, these L_1 penalty methods might not work well in the simulation setting of Section 4.2, where the true β's are not smooth. This easily dismissed but not-too-subtle distinction might be important in some applications. For example, in the application to transcription factor binding site prediction described in Section 5.3, there is no reason to expect that the true values for β are piece-wise constant.

Ising priors do not enforce directionality on the underlying graph. Prior information often comes in the form of constraints, such as X_i must be selected if X_j is in the model, where one can set $P(\gamma_j = 1 | \gamma_i = 0) = 0$. These constraints can be easily modeled via a directed acyclic graph (DAG). The class of priors proposed in Chipman (1996) can be viewed as a special case of general DAG priors. Another applicable situation is when there is prior information for causal relationships among the covariates. Computation under the DAG prior in high-dimensional regression settings is a challenging but exciting area of future research.

We focus on linear regression for continuous outcomes in our discussion. The methods can be readily extended, with care taken in computational efficiency, to nonlinear regression for binary and categorical outcomes, and accelerated failure time models for survival outcomes.

The R and Fortran code are available at by request from the authors.

APPENDIX

A.1 Fast Updating of A_i^{-1} From $A_{(-i)}^{-1}$

To simplify discussion, consider the case first where $D = 0$, so $A_i = X_{I_i}' X_{I_i}$. The case where $D \neq 0$ is analogous. Define $A_{(-i)} = X_{I_{(-i)}}' X_{I_{(-i)}}$, $\Sigma_{I_{(-i)},i} = X_{I_{(-i)}}' X_i$, $\sigma_{ii} = X_i' X_i$. The matrix A_i can be expressed in the following partitioned forms:

$$A_i = \begin{pmatrix} A_{(-i)} & \Sigma_{I_{(-i)},i} \\ \Sigma_{I_{(-i)},i}' & \sigma_{ii} \end{pmatrix}.$$

Then, the matrix A_i^{-1} can be computed as:

$$A_i^{-1} = \begin{pmatrix} A^{11} & A^{12} \\ A^{21} & A^{22} \end{pmatrix}, \qquad (A.1)$$

where

$$A^{11} = \left(A_{(-i)} - \Sigma_{I_{(-i)},i}\sigma_{ii}^{-1}\Sigma_{I_{(-i)},i}'\right)^{-1} \overset{\text{def}}{=} (A_{11\cdot2})^{-1},$$

$$A^{12} = -(A_{11\cdot2})^{-1}\Sigma_{I_{(-i)},i}\sigma_{ii}^{-1},$$

$$A^{21} = -\sigma_{ii}^{-1}\Sigma_{I_{(-i)},i}'(A_{11\cdot2})^{-1},$$

$$A^{22} = \sigma_{ii}^{-1} + \sigma_{ii}^{-1}\Sigma_{I_{(-i)},i}'(A_{11\cdot2})^{-1}\Sigma_{I_{(-i)},i}\sigma_{ii}^{-1}.$$

Of the four quantities above, the computation of A^{12}, A^{21}, A^{22} are $O(p_{(-i)}^2)$. The explicit form of A^{11} is

$$A^{11} = A_{(-i)}^{-1} + \frac{1}{\sigma_{ii}(1 - \Sigma_{I_{(-i)},i}'A_{(-i)}^{-1}\Sigma_{I_{(-i)},i}/\sigma_{ii})}$$
$$\times A_{(-i)}^{-1}\Sigma_{I_{(-i)},i}\left(A_{(-i)}^{-1}\Sigma_{I_{(-i)},i}\right)'. \quad (A.2)$$

Thus the computation of A^{11} can be done via a low-rank update of $A_{(-i)}^{-1}$, available from the previous iteration, and thus would also be $O(p_{(-i)}^2)$.

Calculating the determinant of a matrix is computationally equivalent to obtaining its Cholesky factor. So now we describe the fast updating of the Cholesky factor of A_i^{-1}. Let $A^{11} = \widetilde{L}_{(-i)}\widetilde{L}_{(-i)}'$, $A_{(-i)}^{-1} = L_{(-i)}L_{(-i)}'$, and $A_i^{-1} = L_iL_i'$. Notice the right-hand side of Equation (A.2) is also of the form $A + vv'$, the computation of $\widetilde{L}_{(-i)}$ thus can be done via a low rank update of the Cholesky factor of $A_{(-i)}^{-1}$, $L_{(-i)}$. The lower triangular matrix L_i has the following partitioned form

$$L_i = \begin{pmatrix} \widetilde{L}_{(-i)} & \mathbf{0} \\ L_{(-i),i} & l_{ii} \end{pmatrix},$$

where $L_{(-i),i}$ is $1 \times p_{(-i)}$, and $\mathbf{0} = (0, \ldots, 0)_{p_{(-i)}}'$. This implies

$$A_i^{-1} = \begin{pmatrix} A^{11} & \widetilde{L}_{(-i)}L_{(-i),i}' \\ L_{(-i),i}\widetilde{L}_{(-i)}' & L_{(-i),i}L_{(-i),i}' + l_{ii}^2 \end{pmatrix}. \qquad (A.3)$$

Comparing expressions (A.1) and (A.3), we have $A^{12} = \widetilde{L}_{(-i)}L_{(-i),i}'$, and $A^{22} = L_{(-i),i}L_{(-i),i}' + l_{ii}^2$. The vector $L_{(-i),i}$ thus can be obtained from solving an upper triangular linear system, the computation of which is $O(p_{(-i)}^2)$.

A.2 Mean Field Approximation for Exchangeable Ising Models

For a general Ising model on γ, let $E(\gamma)$ be the *energy function*, defined as $E(\gamma) = -(\sum_i a_i\gamma_i + \sum_{ij} b_{ij}\gamma_i\gamma_j)$, and let

$$\psi(\lambda) = -\log\left[\sum_\gamma e^{-E_0(\gamma) - \lambda(E(\gamma) - E_0(\gamma))}\right],$$

where E_0 is a "simple" energy function which we will define later. Then, $\psi(a, b) = \psi(1)$. One can verify that $\psi(\lambda)$ is concave in λ, which gives us the inequality $\psi = \psi(1) \leq \psi(0) + \dot\psi(0)$, and thus

$$\psi(1) \leq -\log\left[\sum_\gamma e^{-E_0(\gamma)}\right] + \mathbb{E}_0[E(\gamma) - E_0(\gamma)].$$

By $\mathbb{E}_0$, $\mathbb{V}\text{ar}_0$, or $\mathbb{P}_0$, we mean expectation, variance, and probability under the density $p(\gamma) = e^{-E_0(\gamma)}/\sum_\gamma e^{-E_0(\gamma)}$. The above inequality is true for every energy function E_0, and hence it is still true when we optimize over E_0:

$$\psi(1) \leq \min_{E_0 \in \mathcal{F}}\left\{-\log\left[\sum_\gamma e^{-E_0(\gamma)}\right] + \mathbb{E}_0[E(\gamma) - E_0(\gamma)]\right\}. \quad (A.4)$$

The idea in mean field approximations is to choose a class of energy functions $\mathcal{F}$ simple enough so that the minimization in (A.4) is analytically tractable. Often, the choice is the class of linearly additive energy functions:

$$E_0(\gamma) = -\sum_i h_i\gamma_i, \qquad (A.5)$$

with h_i being freely varying parameters. With this parameterization, optimization over $\mathcal{F}$ is equivalent to optimization over $\mathbf{h} = (h_1, \ldots, h_m)$.

Let $\phi(\mathbf{h})$ be the function being minimized in (A.4) for $\mathcal{F}$ defined as in (A.5):

$$\phi(\mathbf{h}) = -\log\left[\sum_{\gamma} e^{\sum_i h_i \gamma_i}\right] - \sum_i (a_i - h_i)\mathbb{E}_0(\gamma_i) - \sum_{ij} b_{i,j}\mathbb{E}_0(\gamma_i\gamma_j).$$

Since $\mathbb{E}_0(\gamma_i) = P_0(\gamma_i = 1) = \frac{e^{h_i}}{(1+e^{h_i})}$, and $\mathbb{E}_0(\gamma_i\gamma_j) = \frac{e^{h_i+h_j}}{(1+e^{h_i})(1+e^{h_j})}$, we have

$$\phi(\mathbf{h}) = -\sum_i \log(1+e^{h_i}) - \sum_i (a_i - h_i)\frac{1}{1+e^{-h_i}}$$
$$- \sum_{ij} b_{i,j}\frac{1}{(1+e^{-h_i})(1+e^{-h_j})}.$$

Since we assume that the vertices are exchangeable, the optimizing $\mathbf{h}$ must have $h_i = h$, and hence, we have a one-dimensional optimization problem:

$$\phi(h) = -n\log(1+e^h) - n(a-h)(1+e^{-h})^{-1} - Nb(1+e^{-h})^{-2},$$

where N is the total number of edges. We let $N = kn$, where $k = \sum_j b_{ij}$ is the sum of weights for edges coming out of each vertex in the graph, and to make things simpler we reparameterize $t = (1+e^{-h})^{-1}$. With a slight abuse of notation, this gives us

$$\frac{\phi(h)}{n} = \phi(t) = \log(1-t) - \left(a + \log\frac{1-t}{t}\right)t - kbt^2. \quad (A.6)$$

For any given a, the phase transition points are the b^*'s that introduces a change in the nature of the minimizer t of Equation (A.6), as discussed in Section 3.1.

[Received April 2008. Revised January 2010.]

REFERENCES

Barbieri, M. M., and Berger, J. O. (2004), "Optimal Predictive Model Selection," *The Annals of Statistics*, 32, 870–897. [1212]

Brown, P. J., Vannucci, M., and Fearn, T. (1998), "Multivariate Bayesian Variable Selection and Prediction," *Journal of the Royal Statistical Society, Ser. B*, 60 (3), 627–641. [1203]

Brush, S. G. (1967), "History of the Lenz–Ising Model," *Reviews of Modern Physics*, 39, 883–893. [1206]

Bussemaker, H. J., Li, H., and Siggia, E. D. (2001), "Regulatory Element Detection Using Correlation With Expression," *Nature Genetics*, 27 (2), 167–171. [1210,1211]

Chipman, H. (1996), "Bayesian Variable Selection With Related Predictors," *Canadian Journal of Statistics*, 24, 17–36. [1203,1213]

Clyde, M., and George, E. (2004), "Model Uncertainty," *Statistical Science*, 19 (1), 81–94. [1203]

Conlon, E. M., Liu, X. S., Lieb, J. D., and Liu, J. S. (2003), "Integrating Regulatory Motif Discovery and Genome-Wide Expression Analysis," *Proceedings of the National Academy of Science of the United States of America*, 100 (6), 3339–3344. [1210]

Dey, T., Ishwaran, H., and Rao, J. S. (2008), "An In-Depth Look at Highest Posterior Model Selection," *Econometric Theory*, 24, 377–403. [1212]

George, E., and McCulloch, R. E. (1993), "Variable Selection via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881–889. [1203]

——— (1997), "Approaches for Bayesian Variable Selection," *Statistica Sinica*, 7, 339–373. [1203]

Ibrahim, J. G., Chen, M. H., and Gray, R. J. (2002), "Bayesian Models for Gene Expression With DNA Microarray Data," *Journal of the American Statistical Association*, 97, 88–99. [1203,1207]

Ishwaran, H., and Rao, J. S. (2003), "Detecting Differentially Expressed Genes in Microarrays Using Bayesian Model Selection," *Journal of the American Statistical Association*, 98, 438–455. [1203,1212]

——— (2005a), "Spike and Slab Variable Selection: Frequentist and Bayesian Strategies," *The Annals of Statistics*, 33, 730–773. [1203-1205,1212]

——— (2005b), "Spike and Slab Gene Selection for Multigroup Microarray Data," *Journal of the American Statistical Association*, 100, 764–780. [1203,1212]

Kechris, K., van Zwet, E., Bickel, P., and Eisen, M. B. (2004), "Detecting DNA Regulatory Motifs by Incorporating Positional Trends in Information Content," *Genome Biology*, 5, R50. [1210]

Mirny, L., and Gelfand, M. (2002), "Structural Analysis of Conserved Base Pairs in Protein-DNA Complexes," *Nucleic Acids Research*, 30, 1704–1711. [1210]

Moses, A. M., Chiang, D. Y., Kellis, M., Lander, E. S., and Eisen, M. B. (2003), "Position Specific Variation in the Rate of Evolution in Transcription Factor Binding Sites," *BMC Evolutionary Biology*, 3, 19. [1210]

Nott, D., and Green, P. J. (2004), "Bayesian Variable Selection and the Swendsen–Wang Algorithm," *Journal of Computational and Graphical Statistics*, 13, 141–157. [1213]

Schneider, T., Stormo, G. D., Gold, L., and Ehrenfeucht, A. (1986), "Information Content of Binding Sites on Nucleotide Sequences," *Journal of Molecular Biology*, 188, 415–431. [1210]

Smith, M., and Fahrmeir, L. (2007), "Spatial Bayesian Variable Selection With Application to Functional Magnetic Resonance Imaging," *Journal of the American Statistical Association*, 102, 417–431. [1203]

Smith, M., and Kohn, R. (1996), "Nonparametric Regression Using Bayesian Variable Selection," *Journal of Econometrics*, 75, 317–343. [1207,1212]

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998), "Comprehensive Identification of Cell Cycle-Regulated Genes of the Yeast Saccharomyces Cerevisiae by Microarray Hybridization," *Molecular Biology of the Cell*, 9 (12), 3273–3297. [1210,1211]

Tadesse, M. G., Sha, N., and Vannucci, M. (2005), "Bayesian Variable Selection in Clustering High-Dimensional Data," *Journal of the American Statistical Association*, 100, 602–617. [1203]

Tadesse, M. G., Vannucci, M., and Liò, P. (2004), "Identification of DNA Regulatory Motifs Using Bayesian Variable Selection," *Bioinformatics*, 20, 2553–2561. [1210]

Tibshirani, R., Saunders, M., Rosset, R., Zhu, J., and Knight, K. (2005), "Sparsity and Smoothness via the Fused Lasso," *Journal of the Royal Statistical Society, Ser. B*, 67 (1), 91–108. [1213]

Wei, P., and Pan, W. (2009), "Network-Based Genomic Discovery: Application and Comparison of Markov Random Field Models," technical report, University of Minnesota, Dept. of Biostatistics. [1213]

Wei, Z., and Li, H. (2007), "A Markov Random Field Model for Network-Based Analysis of Genomic Data," *Bioinformatics*, 23, 1537–1544. [1213]

West, M. (2003), "Bayesian Factor Regression Models in the 'Large p, Small n' Paradigm," in *Bayesian Statistics 7*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford University Press, pp. 723–732. [1202,1203]

Yedidia, J. S. (2001), "An Idiosyncratic Journey Beyond Mean Field Theory," in *Advanced Mean Field Methods: Theory and Practice*, Cambridge, MA: MIT Press, pp. 21–36. [1205]

Yuan, M., and Lin, Y. (2006), "Model Selection and Estimation in Regression With Grouped Variables," *Journal of the Royal Statistical Society, Ser. B*, 68 (1), 49–67. [1213]

Zhang, N. R., Wildermuth, M. C., and Speed, T. P. (2008), "Transcription Factor Binding Site Prediction With Multivariate Gene Expression Data," *The Annals of Applied Statistics*, 2, 332–365. [1210,1211]

Zhu, J., and Zhang, M. Q. (1999), "SCPD: A Promoter Database of Yeast Saccharomyces Cerevisiae," *Bioinformatics*, 15, 607–611. [1211]