2010

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

Fan Li

Nancy R. Zhang
University of Pennsylvania

Recommended Citation

Li, F., & Zhang, N. R. (2010). Bayesian Variable Selection in Structured High-Dimensional Covariate Spaces With Applications in Genomics. *Journal of the American Statistical Association, 105* (491), 1202-1214. http://dx.doi.org/10.1198/jasa.2010.tm08177

At the time of publication, author Nancy R. Zhang was affiliated with Stanford University. Currently, she is a faculty member at the Statistics Department at the University of Pennsylvania.

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

Abstract

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.

Keywords

Ising model, Markov chain Monte Carlo, motif analysis, phase transition, undirected graph

Disciplines

Statistics and Probability

Comments

At the time of publication, author Nancy R. Zhang was affiliated with Stanford University. Currently, she is a faculty member at the Statistics Department at the University of Pennsylvania.

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

Fan Li
Department of Health Care Policy
Harvard Medical School
Boston, MA 02115-2899, USA
li@hcp.med.harvard.edu

Nancy R. Zhang
Department of Statistics
Stanford University
Stanford, CA 94305-4065, USA
nzhang@stat.stanford.edu

November 18, 2007

SUMMARY

We consider the problem of regression modeling in high dimensional spaces where there is known structure among the covariates. Such problems are becoming increasingly relevant as high-throughput data collection schemes become increasingly common. A fundamental goal in such problems is to find a small set of covariates that are associated with a response variable, which we discuss from the perspective of statistical variable selection. However, 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 many cases it is desirable to incorporate this structural information in the model building process.

We approach this problem through the Bayesian variable selection framework, where we formulate a general Ising prior on the model space for incorporating structural information. However, 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. As examples we discuss two specific applications in genomics, one arising from DNA copy number analysis, and the other arising from the modeling of transcription factor binding sites in DNA sequences.

Key words and phrases: Bayesian variable selection, DNA copy number data, Ising model, Markov chain Monte Carlo, motif analysis, phase transition

1 Introduction

Consider the standard multiple regression problem

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1.1}$$

where $\mathbf{Y}$ is $n \times 1$ variable response, $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_m)$ is a $n \times m$ matrix of covariates, and $\boldsymbol{\epsilon}$ is an $n \times 1$ error term. We employ the standard assumption that $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I)$. In this paper, we study the problem of variable selection for this model when m is large, possibly much larger than n, and when there is known structure among the 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. In Section 3, we will discuss in detail two problems of this kind that arise in high-throughput genomics. The first problem concerns the analysis of array-based comparative genomic hybridization (array-CGH) data, where the covariates are measurements of DNA quantity at m locations in the genome, collected for n patients. Array-CGH data detects gains and losses of DNA copy number, a common phenomenon in cancer. The response variable in this case may be an observed phenotype, or a clinical outcome, and we would like to find genome locations that, when gained or lost, predicts the response. Since the copy number measurements are located linearly along the genome, we would like to use this ordering information in the model building process. The second problem that we consider arises in the study of transcription regulation, where the response is the expression level of n genes, and the covariates are the counts of appearances of all L-length words in the upstream region of that gene. Through a model such as (1.1) we would like to find words whose appearance in the upstream region has an effect on the gene's transcription level, because such words are likely to be binding sites for transcription factors. In this case, the covariates, the L-length words, can be viewed as vertices on a hypercube. Due to the degeneracy of transcription factor binding sites (TFBS) neighboring words on the hypercube often have similar effects on expression, and it is this information that we would like to incorporate in building the model.

Greedy stepwise or exhaustive enumeration approaches for searching for the best model are clearly impractical for such high dimensional problems. Penalized regression schemes such as the LASSO (Tibshirani (1996)) have gained in popularity, and recently there has been increasing work in L_1-type penalties for incorporating covariate space structure, for example, the fused LASSO (Tibshirani et al. (2005)) and the group LASSO (Yuan and Lin (2006)). However, in this paper we consider the Bayesian approach to variable selection, where we find the incorporation of covariate space structure to be more natural. The basic idea behind Bayesian variable selection (George and McCulloch (1993,1997), Brown et al. (1998,2002)) is to define latent variables $(\gamma_i : 1 \leq i \leq m)$, where γ_i is the indicator of whether covariate i is included in the model. Then, Markov chain monte carlo methods are used to explore the model space $\{\boldsymbol{\gamma} : \gamma_i \in \{0, 1\}\}$ and approximate the posterior distribution of $\boldsymbol{\gamma}$ given the data. The covariate space structure is used to aid the search for the best model by assuming that $\boldsymbol{\gamma}$ lies on a graph and that the prior distribution for $\boldsymbol{\gamma}$ is Markov with respect to this graph. In the first problem involving array-CGH data, the graph is a linear chain, whereas in the second problem involving TFBS modeling, the graph is a hypercube.

Non-independent priors for $\boldsymbol{\gamma}$ have been employed previously in smaller scaled problems, where $m \ll n$. When m becomes large, i.e. in the thousands, many new theoretical and computational issues arise. The most interesting, and problematic, of which is the phenonmenon of phase transitions: Certain global characteristics of the prior distribution of $\boldsymbol{\gamma}$, such as the model size $\gamma_1 + \cdots + \gamma_m$, undergo a dramatic change given an infinitesimal change in the hyperparameters. Since the computational efficiency of the MCMC algorithm depends heavily on the model size, it is critically important to understand the phase transition behavior of the distribution of $\boldsymbol{\gamma}$, and to be able 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 considered in the context of Bayesian variable selection algorithms.

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. Bayesian approaches to variable selection has previously been applied in high dimensions, e.g., by Tadesse et al. (2005), who used Metropolis-Hastings based approaches. In this paper, we explore the performance of Gibbs sampling algorithms, as first suggested by George and McCulloch (1993). We discuss the computational challenges that arise in this method, and give an efficient algorithm which we use to analyze two high-dimensional data sets in Section 3.

2 Notations and Formulation of General Model

2.1 Model

Let the observed data be $\mathbf{X}$ and $\mathbf{Y}$ for which we assume the simple linear model (1.1) as described in the introduction. Following George and McCulloch (1993,1997), we assume that the prior distribution for the regression parameters $\boldsymbol{\beta}$ depends on latent variables $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_m)'$, with $\gamma_i \in \{0, 1\}$. Given $\boldsymbol{\gamma}$, β_i are independent with conjugate Gaussian mixture priors

$$\beta_i|\gamma_i \sim (1 - \gamma_i)N(0, \sigma^2 v_0^2) + \gamma_i N(0, \sigma^2 v_1^2), \tag{2.1}$$

which, in matrix form, is

$$\boldsymbol{\beta}|\boldsymbol{\gamma} \sim N_p(0, \sigma^2 D_{\boldsymbol{\gamma}}^2),$$

where $D_{\boldsymbol{\gamma}} = \text{diag}((1 - \gamma_i)v_0 + \gamma_i v_1 : 1 \leq i \leq m)$. It is assumed that $v_1 > v_0 \geq 0$, so that β_i has a larger prior variance if i is included in the model.

A special case of the prior (2.1) that plays a crucial role in computations is

$$\beta_i|\gamma_i \sim (1 - \gamma_i)I_0 + \gamma_i N(0, \sigma^2 v^2), \tag{2.2}$$

where I_0 is a point mass at 0. The prior (2.2) can be obtained from (2.1) by letting $v_0 \to 0$, $v_1 = v$.

We use the inverse gamma conjugate prior for the variance σ^2,

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

which is equivalent to the assumption $\nu\lambda/\sigma^2 \sim \chi^2_\nu$.

To formulate the prior for $\boldsymbol{\gamma}$, we assume that the covariates $i = 1, \ldots, m$ lie in an undirected graph which can be represented by an edge set $\mathcal{E} = \{(i, j) : 1 \leq i \neq j \leq m\}$. Given this graph, let $\boldsymbol{a} = (a_1, \ldots, a_m)'$ be a real vector and $\boldsymbol{B} = (b_{i,j})_{m \times m}$ be a matrix of real numbers where $b_{i,j} = 0$ for all $(i, j) \notin \mathcal{E}$. Then, we assume the following exponential form for the prior distribution of $\boldsymbol{\gamma}$:

$$P(\boldsymbol{\gamma}) = e^{\boldsymbol{a}'\boldsymbol{\gamma} + \boldsymbol{\gamma}'\boldsymbol{B}\boldsymbol{\gamma} - \psi(\boldsymbol{a}, \boldsymbol{B})}, \tag{2.3}$$

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

$$\psi(\boldsymbol{a}, \boldsymbol{B}) = \sum_{\gamma \in \{0,1\}^m} e^{\boldsymbol{a}'\boldsymbol{\gamma} + \boldsymbol{\gamma}'\boldsymbol{B}\boldsymbol{\gamma}}.$$

This exponential form is called the Ising model in physics and Monte Carlo literature, where $\psi(\boldsymbol{a}, \boldsymbol{B})$ is referred to as the *partition function*. Without loss of generality we assume that $a_i < 0$. If $\boldsymbol{B}$ were 0, then $\psi(\boldsymbol{a}, \boldsymbol{0}) = \sum_{i=1}^{m} \log(1 + e^{a_i})$, but in general there is no closed form for ψ.

Often, we do not want to favor a priori the inclusion of any covariate into the model. If the graph $\mathcal{E}$ were regular, then this can be achieved by letting $\boldsymbol{a} = aI_m$, where $I_m = (1, 1, \ldots, 1) \in \Re^m$, and $\boldsymbol{B}$ be chosen so that it is symmetric in $i = 1, \ldots, m$. We will illustrate this concretely with the motif example in Section 3.2. In general, the hyperparameter a controls the sparsity of $\boldsymbol{\gamma}$ and the entries in $\boldsymbol{B}$ control the smoothness of $\boldsymbol{\gamma}$ over $\mathcal{E}$.

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

To sample from the posterior distribution $f(\boldsymbol{\gamma}|\mathbf{Y})$, we adopt the Gibbs sampling scheme that sample directly from the ergodic Markov chain

$$\boldsymbol{\gamma}^0, \boldsymbol{\gamma}^1, \boldsymbol{\gamma}^2, \ldots. \tag{2.4}$$

This scheme is of particular interest because, when the average model size is sparse, each update sweep of $\boldsymbol{\gamma}$ can be accomplished in linear time.

Let $\gamma_{(-i)} = \{\gamma_j : j \neq i\}$; $I_{(-i)}$ be the set of indices $\{\gamma_j = 1 : j \neq i\}$; $I_i = I_{(-i)} \bigcup \{i\}$; $m_i = |I_i|$ and $m_{(-i)} = |I_{(-i)}|$. For the prior distribution (2.3), there is a simple form for the conditional distribution

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

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

$$P(\gamma_i = 1|\gamma_{(-i)}, \mathbf{Y}) \;=\; \frac{P(\gamma_i = 1|\gamma_{(-i)})}{P(\gamma_i = 1|\gamma_{(-i)}) + F(i|\gamma_{(-i)})^{-1} \cdot P(\gamma_i = 0|\gamma_{(-i)})} \tag{2.5}$$

where $F(i|\gamma_{(-i)})$ is the Bayes factor, that is,

$$F(i|\gamma_{(-i)}) = \frac{P(\mathbf{Y}|\gamma_i = 1, \gamma_{(-i)})}{P(\mathbf{Y}|\gamma_i = 0, \gamma_{(-i)})}.$$

The Bayes factor can be explicitly computed for the linear regression model. To compute the term $P(Y|\gamma_i = 1, \gamma_{(-i)})$, first integrate out $\boldsymbol{\beta}$ under the special conjugate prior (2.2),

$$P(Y|\gamma_i = 1, \gamma_{(-i)}, \sigma^2) \;=\; e^{-\frac{Y'Y - Y'X_{I_i}A_i^{-1}X'_{I_i}Y}{2\sigma^2}} \sigma^{-n}|A_i|^{-\frac{1}{2}}|D_{I_i}|^{-\frac{1}{2}}, \tag{2.6}$$

where $A_i = X'_{I_i}X_{I_i} + D_{I_i}^{-2}$. Then, integrating out σ from (2.6), we have

$$P(Y|\gamma_i = 1, \gamma_{(-i)}) \;\propto\; |A_i|^{-\frac{1}{2}}|D_{I_i}|^{-\frac{1}{2}} \left(\frac{Y'Y - Y'X_{I_i}A_i^{-1}X'_{I_i}Y + \nu\lambda}{2}\right)^{-\frac{n+\nu}{2}}.$$

$P(Y|\gamma_i = 0, \gamma_{(-i)})$ can be obtained similarly, with I_i replaced by $I_{(-i)}$. Therefore,

$$F(i|\gamma_{(-i)}) = v^{-1} \cdot \frac{|A_{(-i)}|^{\frac{1}{2}}}{|A_i|^{\frac{1}{2}}} \cdot \left(\frac{Y'Y - Y'X_{I_{(-i)}}A_{(-i)}^{-1}X'_{I_{(-i)}}Y + \nu\lambda}{Y'Y - Y'X_{I_i}A_i^{-1}X'_{I_i}Y + \nu\lambda}\right)^{\frac{n+\nu}{2}}. \tag{2.7}$$

Hence, one can sample directly from the posterior distribution of $\boldsymbol{\gamma}$ by constructing a Markov chain on $\{0,1\}^m$ 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 (2.5). The index i can either be picked in a fixed order, or randomly.

2.3 Computational Issues

Evaluating the Bayes Factor $F(i|\gamma_{(-i)})$ in (2.7) is the computationally intensive step during each iteration, because it involves inverting and calculating the determinant of the m_i by m_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 $O(m_{(-i)}^2)$ operation. Then, each sweep through all of the γ_i's (assuming the γ_i's are sampled in fixed order) would be $O(mm_{(-i)}^2)$. Various fast update algorithms can be developed using numerical methods to obtain inverse and determinant of matrix, e.g., by Cholesky or LU decomposition. Details of the algorithm we used are given in the Appendix 6.1.

This shows the importance of limiting the size of the model during the sampling of $\boldsymbol{\gamma}$: even though the Bayesian formulation allows the model size in each iteration to be larger than n, 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 later sections. This also explains why we choose the special prior (2.2) over the general prior (2.1). Even with fast update algorithms, the latter would lead to a computational task of quadratic order $O(m^2)$, which is impractical when m is very large.

3 Examples

In Section 2, we proposed a general Ising prior (2.3) on γ for incorporating the structure in the covariate space, and gave a general formula (2.5) for the Gibbs sampler to sample from $P(\boldsymbol{\gamma}|Y)$. We believe that the utility of this model owes to the fact that it is easily adaptable to

a wide variety of problems. We present here two examples with different covariate structure. Through these examples, we will discuss the selection of hyperparameters, which is paramount to both the quality of the results as well as the efficiency of the computation. In particular, when m is large, the selection of hyperparameters need to be based not only on prior beliefs but also on considerations of computational efficiency.

In the first example, the underlying graph is a linear Markov chain. It is a well known fact that for this simplest of graphs closed form formulas are available for marginal probabilities on the γ_i's, which can be used to guide hyperparameter selection. Another convenient fact about the linear Markov chain is that it does not exhibit phase transition behavior as $m \to \infty$ (see, e.g., Brush (1967)). This is not true in the second example, where the underlying graph is a hypercube. Hence, of primary concern in the second example is the selection of hyperparameters to avoid phase transition behavior.

3.1 DNA copy number analysis: linear Markov chain prior

3.1.1 Background of Application

High throughput platforms for DNA copy number analysis has generated massive data sets to catalogue this specific type of genetic variation. During the past decade, several different technologies have been developed to measure DNA copy number at a fine scale at thousands to hundreds of thousands of locations in the genome. We let $X_{k,i}$ be the copy number measurement in sample k at location i. A value for $X_{k,i}$ that is lower than baseline indicates a possible loss of that region of the genome, and a value that is higher than baseline indicates a possible gain. We are interested in finding regions of the genome that may be associated with an observed trait $\mathbf{Y}$, which may be clinical outcome, response to treatment or the measurement of another biomarker. Since neighboring measurements on a chromosome are noisy surrogates for the underlying copy number of contiguous locations on the chromosome, it is desirable for the model to pool evidence neighboring clones in finding regions of the genome that are associated with the response.

3.1.2 Model Description

To reflect the linear ordering of the measurements along the chromosome, we assume that $\boldsymbol{\gamma}$ is Markov with transition matrix

$$P = \begin{pmatrix} p & 1-p \\ 1-q & q \end{pmatrix},$$

and that $\gamma_1 \sim \pi$, where

$$\pi = \left(\frac{1-q}{2-p-q}, \frac{1-p}{2-p-q} \right)$$

is the stationary distribution with regards to P. An equivalent parameterization of this Markov chain is

$$P(\gamma_i = 1 | \gamma_{i-1}, \gamma_{i+1}) = \frac{e^{a+b(\gamma_{i-1}+\gamma_{i+1})}}{1 + e^{a+b(\gamma_{i-1}+\gamma_{i+1})}}, \tag{3.1}$$

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

$$r = \frac{1-p}{1-q} = \frac{\pi_1}{\pi_0}, \quad w_0 = \frac{p}{1-q}, \quad w_1 = \frac{q}{1-p}. \tag{3.2}$$

The above 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 i.i.d.. The pair (r, w_1) completely specifies the model, and is more interpretable than (a, b). Thus, we will refer to r as the sparsity parameter and $w = w_1$ as the smoothness parameter. Also note that this parameterization is symmetric in the γ_i's, which means that a priori, every covariate has equal chance of being in the model.

3.1.3 Hyperparameter Selection

For simplicity, we assume a flat prior on σ^2 (i.e. $\nu = 0$, λ irrelevant), and focus on the selection of v, r, and w. The hyperparameter v is the prior variance of β_i given that $\gamma_i = 0$, 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 from our experience v only needs to have the correct order for the method to perform well. The selection of v based on the expected signal for low dimensional problems, and its interpretation, has been explored in George and McCulloch (1993, 1997), and Mitchell and Beauchamp (1988). Their discussion carries over to high dimensional settings, and we refer the reader to these papers for details on the selection of v.

We would like to explore further the influence of hyperparameter choice on model size, which is an important concern since the computation time for each sweep of the Gibbs sampler is on the order of the model size squared times m. The prior expectation of model size is $mP(\gamma_i = 1) = m\pi_1 = mr/(1-r)$, relying directly on the sparsity parameter r. However, the posterior model size is a complex function of r, w_1, v, as well as the number and strength of true predictors. As a rough heuristic, from the Laplace approximation of the Bayes factor (2.7) we have

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

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. Consider the simple case where $\mathbf{X}$ and $\mathbf{Y}$ are unrelated. If $v \to \infty$, then for large sample sizes $\Delta/\hat{\sigma}^2$ is approximately χ^2 distributed, and $\log|A_{(-i)}| - \log|A_i| = \log n + O(1)$. Hence, for v and n large, we have the approximation

$$P(\gamma_i = 1|\gamma_{(-i)}, \mathbf{Y}) \approx \frac{e^{a+b\sum_{j\in I_{(-i)}}\gamma_j - \log v - \log n + Z^2/2}}{1 + e^{a+b\sum_{j\in I_{(-i)}}\gamma_j - \log v - \log n + Z^2/2}}, \tag{3.3}$$

where $Z \sim N(0,1)$. This implies that for the case of $\mathbf{X}$ and $\mathbf{Y}$ unrelated (when, ideally, the posterior model should be the empty set): we have the following relationships:

1. The posterior model size is smaller for larger v, with

$$\lim_{v\to\infty} P(\gamma_i = 1|\gamma_{(-i)}, \mathbf{Y}) = 0.$$

2. The posterior model size decreases with increasing sample size, with

$$\lim_{n\to\infty} P(\gamma_i = 1|\gamma_{(-i)}, \mathbf{Y}) = 0.$$

These observations are intuitive. Larger v means less shrinkage on β, and thus each addition of a predictor to the model should be penalized more heavily. Also, as sample size increases the posterior model should be consistent, as verified by the second observation. When the number of covariates is large, we expect the bulk of them to follow the null model, and thus the above approximation is a good heuristic in relating the model size to v, n, and (a, b).

Hence, in choosing hyperparameters to achieve a certain modelsize, one needs to take into consideration not only the sparsity parameter r, but also the sample size n and the prior variance v. We found the following to be a good strategy: First choose v based on the expected signal magnitude of b, and then, based on n and w, choose r based on the heuristic in equation (3.3) and the desired running time and number of iterations of the Markov chain.

3.1.4 Simulation Studies

Scenario 1: Smooth in $\boldsymbol{\gamma}$. First consider the following simulation model:

$$Y_k = X_{k,i}\beta\gamma_i^* + \epsilon_{k,i}, \quad i = 1, \ldots, m; \quad k = 1, \ldots, n; \tag{3.4}$$

where $X_i \sim N(0, 1)$ and $\epsilon_i \sim N(0, 1)$. We let $m = 1000$ and $n = 100$, and set γ to be the piecewise constant vector $\gamma_i = I(i \in [245, 260] \cup [745, 760])$. This is a simple model of additive effects over two blocks of consecutive covariates. The true β used to generate the data is allowed to vary over $\{0.5, 1, 2\}$, and we experimented with Bayesian variable selection with varying levels of v, r, and w under the same stationary distribution π. For each setting of hyperparameters, we ran the Gibbs sampler 10 times with random start in $\boldsymbol{\gamma}$. Each run has 100,000 iterations with the first 50,000 iterations as burnin. For 100,000 iterations with average posterior model size of 40, the complete procedure takes about 2.5 hours to run on a Sun Fire Unix V880 with 1200Mhz ultraSprac III CPU. In all of our experiments, the 10 simulations lead to highly similar posterior summary statistics, indicating convergence of the MCMC.

For high dimensional covariate spaces (m in thousands or more), 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, as observed in our simulations. A natural alternative is to instead calculate the posterior marginal distribution of γ_i, $P(\gamma_i = 1|\mathbf{Y})$, by dividing the number of iterations where $\gamma_i = 1$ over the total number of iterations excluding the burnin period. To compare between models, we can 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)

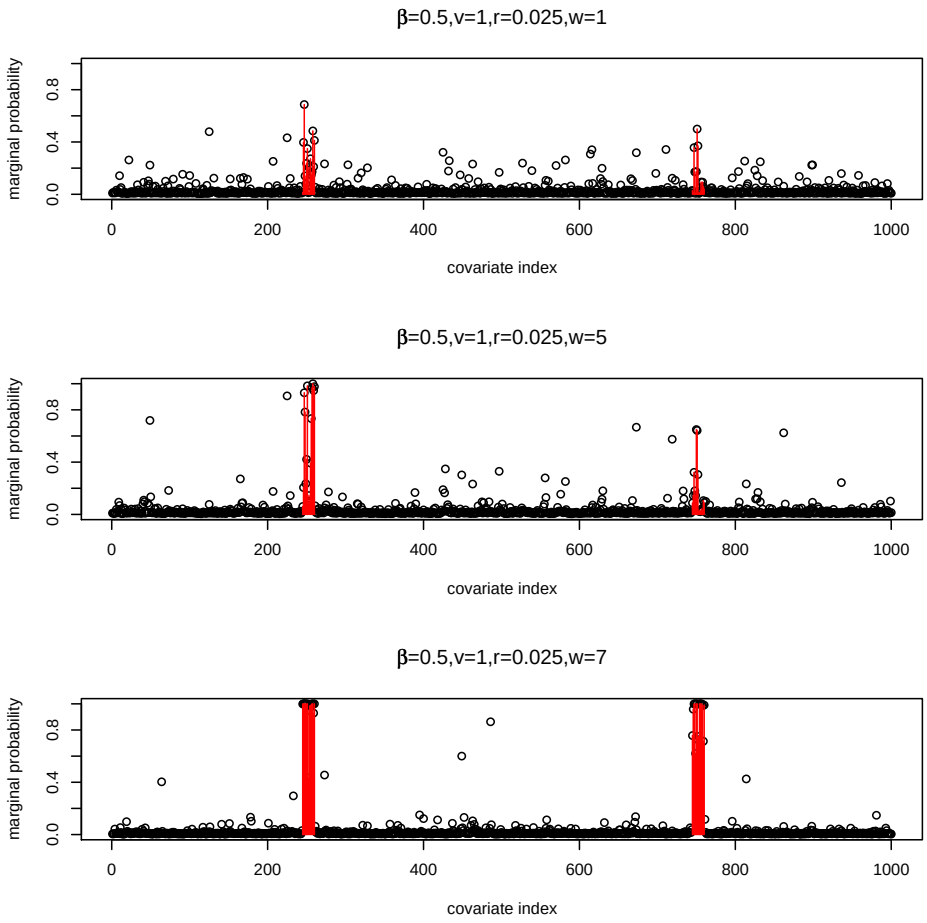

Figure 1. Marginal probability of $\boldsymbol{\gamma}$ under simulation model (3.4) (smooth in $\boldsymbol{\gamma}$)

achieved by varying the calling threshold. The bigger area under the ROC curve (maximum 1), the better the discriminating power of the model.

Here we present the results where the signal is weak ($\beta = 0.5$). Figure 1 shows the posterior marginal probability of $\boldsymbol{\gamma}$ with fixed $v = 1$, $r = \pi_1/\pi_0 = 0.025$, and varying $w_1 = 1, 5, 7$, where the indices of true $\gamma_i = 1$ are labeled by red lines. Figure 2 shows the corresponding ROC curves. Note that $w_1 = 1$ corresponds to the case of γ_i's i.i.d.. It is quite clear from the results that in this simple additive model the assumed Markov chain prior indeed yields significantly better results. For the easier tasks where the underlying models have stronger signal (larger β), the improvement becomes even more pronounced. This pattern is consistently observed in each of our simulations.

Scenario 2: Smooth in X. It is intuitively obvious that in simulation model (3.4), 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 $\mathbf{X}_k = (X_{k,1}, \ldots, X_{k,m})$ be piecewise continuous:

$$\mathbf{X}_{k,i} = \delta Z_k I(i \in [i^* - L_{k,1}, i^* + L_{k,2}]) + \xi_{k,i}, \tag{3.5}$$

where $\xi_{k,i} \sim N(0,1)$, $Z_k \sim Bernoulli(1/2)$, 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 and length

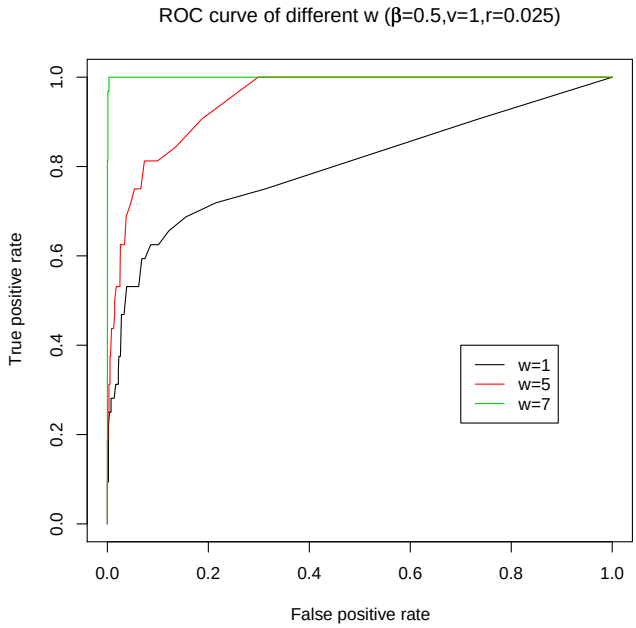

Figure 2. ROC curve under simulation model (3.4) (smooth in γ)

$L_{k,1} + L_{k,2}$ covering location i^*. Then, let the response depend only on whether there is a jump at i^*:

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

Hence, $\mathbf{Y}$ is related to $\mathbf{X}$ *only* through the latent variable $\boldsymbol{Z}$. X may or may not contain a jump near i^*, and since the jump involves the neighboring covariates as well, pooling information across adjacent covariates might aid in the determination of the value of Z_k, and thus also in the prediction of the value of Y.

It is quite clear that model (3.5) pose a much harder variable selection task than model (3.4) because of the extra noise introduced in X. This means a small underlying effect size (β) usually leads to poor performance of the Baysian variable selection procedure with any w. However, as β increases, the models differ in performance.

Here we present the results under model (3.5) with $\delta = 0.35$, $\beta = 3.5$, and 10 spikes in X, $i^* = (50, 150, \cdots, 950)$. Figure 3 shows the posterior marginal probability of $\boldsymbol{\gamma}$ with fixed $v = 1$, $r = 0.02$ and varying $w = 1, 5, 7$. Figure 4 shows the corresponding ROC curves. The gain over a smoothed prior for $\boldsymbol{\gamma}$ is understandably less than that for the first simulation model. However, it is clear that when the jump size δ is small, pooling information across neighboring covariates can help significantly in identifying the location of i^*. An interesting feature shown in Figure 4 is that above certain value (> 1), larger w does not necessarily result in larger area under ROC curve. This is not surprising because the extra signal from pooling information over a large neighborhood under overly large w tends to be outpassed by the extra noise introduced at the same time.

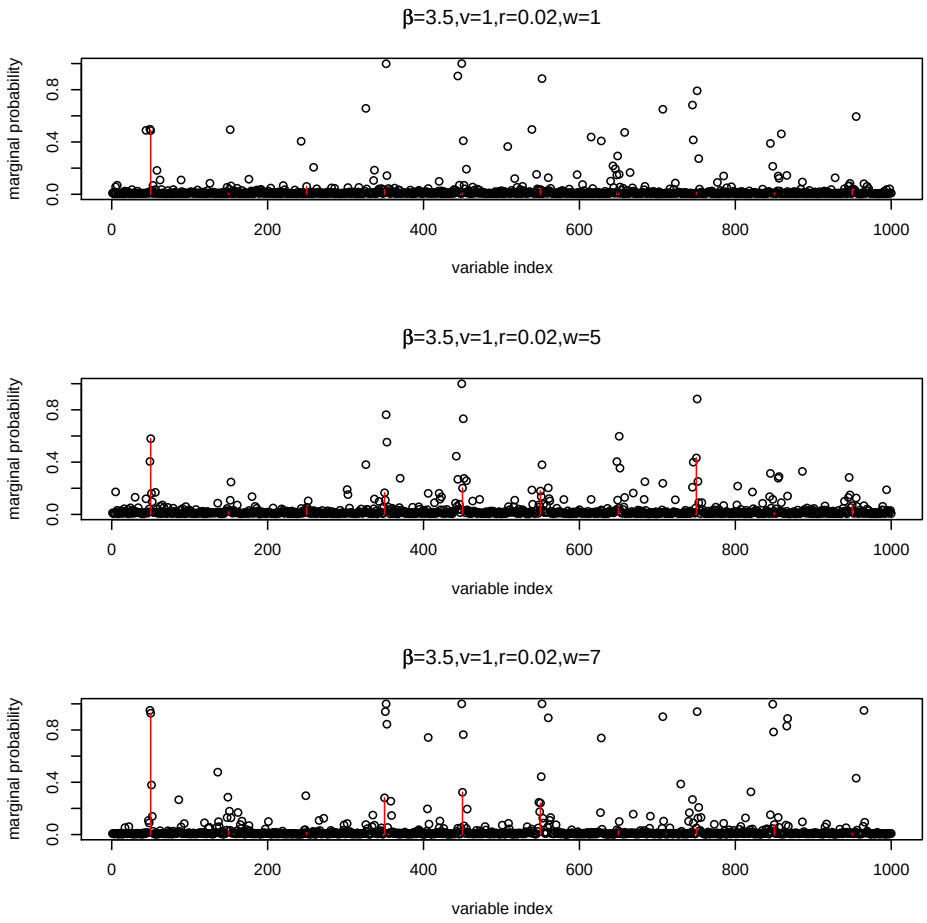

Figure 3. Marginal probability of γ under simulation model (3.5) (smooth in X)

3.1.5 Results on a Colorectal Cancer Data Set

The two simulation models in the last section reflect two different hypotheses for the relationship between DNA copy number and cancer outcome. The first model, which we will call the "multiple-genes model", reflects the hypothesis that the dosage level of multiple genes in an aberrant region in the genome contribute collectively to the cancer outcome. This type of model applies, for example, to the well-known effects of trisomy and contiguous gene deletion syndromes. Recently it has been hypothesized (Mitelman et al. (1997), Duesberg et al. (2005)) that the dosage effect of whole sets of genes also play an important role in cancer. Alternatively, the second "one-gene" model applies to the case where the aberrant region is caused by the selection for a single oncogene, with the other genes in the region having no or little effect on the outcome. This type of model has been proposed to explain many cases of recurrent escalating amplifications in neoplasms such as the ERBB2 region in breast cancer. As we have shown using our simulation study, both models could potentially benefit from the linear Markov prior on $\boldsymbol{\gamma}$. However, the size of the improvement depends both on the error structure in the data as well as the strength of the hypothesized effect.

As an example, we analyze the BAC array-CGH data from colorectal liver metastasis resected from 50 patients, taken from Mehta et al. (2005). For this data set, clinical variables

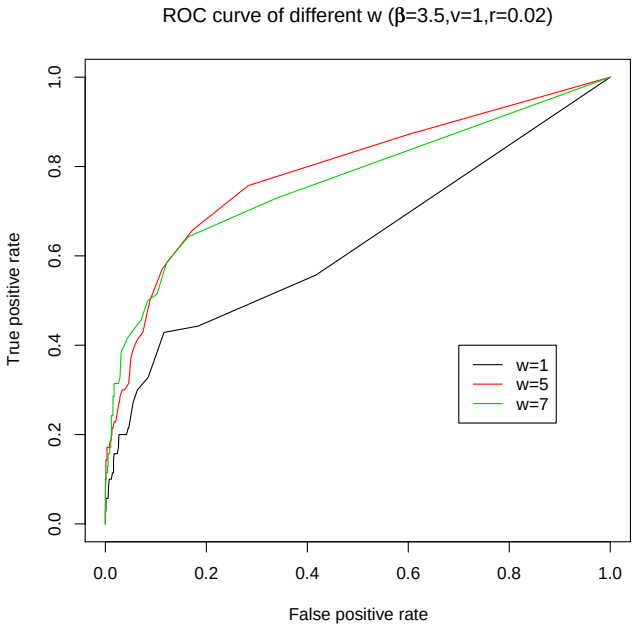

Figure 4. ROC curve under simulation model (3.5) (smooth in X)

such as the overall survival time of the patient are available. The covariates are the measurements of DNA copy number at 2153 (m) locations along the whole genome. Our goal is to identify regions of the genome that have prognostic value in predicting overall survival. The survival time for this data set is fully observed (no censoring), thus we model it by the Gaussian distribution. We applied a square root transform to the survival time to stabilize variance.

We use the linear Markov chain prior with $w_1 \in \{0, 20\}$ to analyze this data. The MCMC chain ran for 150000 iterations, of which the first 100000 iterations were used as burn-in. The results from 10 random restarts confirmed the convergence of the chain. Figure 5 shows the posterior marginal probabilities for γ_i plotted against location in the genome. For this data set, the most prominent spike in the posterior marginal probabilities has a height ≈ 0.4, indicating that there is no single genomic location which has a strong correlation with survival. This is consistent with the conclusions of Mehta et al. (2005), who found that, although the total fraction of genome altered is a significant independent predictor of survival, no single clone has a significant independent effect.

However, quite a few regions have marginal posterior probabilities that rise above the bulk. This is especially noticeable when one zooms in on the marginal probability plots for each chromosome separately (Figure 6). The mean posterior model size, to the nearest integer, is 53 for both values of w. The mean of $P(\gamma_i|\mathbf{Y})$ is 0.0245 and 0.0246 respectively for $w_1 = 0, 20$, which are both roughly equal to $53/2153 = 0.0246$, the marginal posterior probability for each γ_i assuming that all covariates are a posteriori equally likely to be in the model. Table 1 lists the clones whose posterior marginal probability are three fold above the mean.

For colorectal carcinoma, several regions of the genome have already been confirmed by numerous studies to have good prognostic value in predicting survival. The two regions that have received the most attention are chromosome arm 18q and 20q, both of which figure

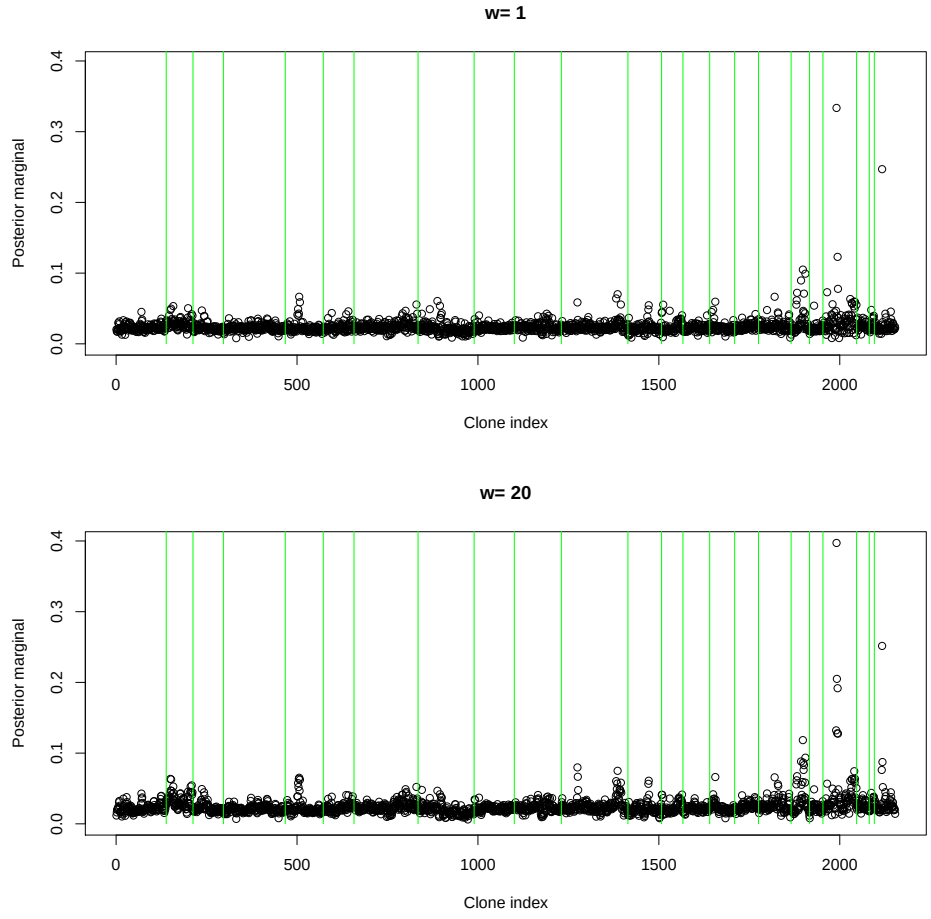

Figure 5. Marginal probability of γ for Mehta et al. colorectal cancer data.

prominently in Table 1. First, consider chromosome arm 18q. Several retrospective studies have identified correlations between loss of heterozygosity events in this region and reduced survival for patients with colorectal carcinoma. The effect is not always strong, as other studies have failed to identify this correlation. The evidence is strongest in the 18q21 region, which contains several cancer related genes, including DCC (deleted in colorectal carcinoma gene) and SMAD2 and SMAD4 (mothers against decapentaplegic homologue 2 and 4). However, it has been hypothesized by (Ji et al. 2007) that other candidate colorectal cancer genes may reside in this area which also provide good prognosis value.

Next, consider chromosome arm 20q13, which has been identified in breast and ovarian cancer with speculations about prognostic significance. In the case of colorectal carcinoma, several studies have reported amplifications in the 20q11-13 region and have found correlation between amplification in this region with worse outcomes.

For any value of $w_1 \in (1, 20)$, the chromosomal regions 18q21 and 20q13 contain the clones with the highest posterior marginal probabilities. The posterior marginal probability increases slightly but steadily with increase in w_1. Also worth noting is chromosome 11, which has also been linked with poor prognosis in Tagawa et al. (1997). This region is noticeably more separated from the baseline probability of 0.0254 for $w_1 = 20$.

For this data set, the difference between the results for different values of w_1 is not striking.

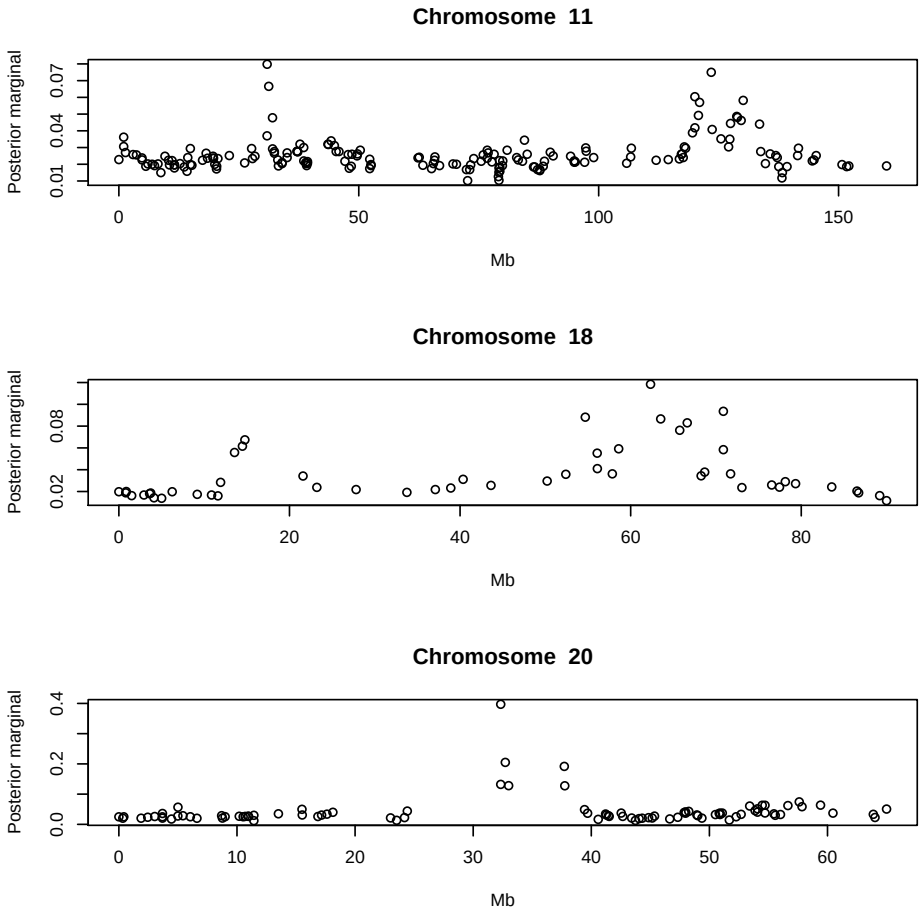

Figure 6. Chromosomes 11, 18, and 20 which contain regions of higher than 3 fold increase in posterior marginal probability than average.

Note that the simulation studies in the previous section show that the gain obtained from increased w is larger under the multiple-gene hypothesis than under the one-gene hypothesis, and that under the one-gene hypothesis the gain is larger if the separation between states (δ/σ_X) is small. Hence, the similarity of results between $w_1 = 0$ and $w_1 = 20$ may be due to the high signal/noise ratio of the array-CGH data, the small effect size, or both.

3.2 DNA motif finding: hypercube prior

3.2.1 Background of Application

Transcription factors are proteins that regulate gene expression by binding to its surrounding sequence in the genome. 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 et al. (2001), Conlon et al. (2003), Zhang et al. (2007).

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

Chrom	Mb	Posterior
11	30.898	0.0798
11	123.46	0.075
18	54.679	0.0882
18	62.332	0.1184
18	63.523	0.0866
18	65.752	0.0762
18	66.63	0.083
18	70.867	0.0936
20	32.33	0.1322
20	32.33	0.3972
20	32.718	0.205
20	33	0.128
20	37.715	0.1918
20	37.753	0.1276
20	57.607	0.0744
23	48.84	0.0762
23	50.202	0.2516
23	54.599	0.0874

Table 1. Clones with greater than three-fold increase of $P(\gamma_i = 1|\mathbf{Y})$ over mean value for $w_1 = 20$.

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 do not tolerate 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 can not be changed. However, the flanking bases are allowed to wobble, with variants of MCB including `TCGCGA` and `CCGCGT`. Even though different motifs have different position specific bases, studies have shown that they share position-specific entropy patterns (Mirny and Gelfand (2002), Schneider et al. (1986), 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 to model the dependence of gene expression on the count of various motifs in its promoter sequence. The response variable is the expression of each gene. In de novo motif detection, 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.

3.2.2 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, i.e.

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

We then let $m_i > 0$ be a weight corresponding to the i-th position,

$$B_{w,w'} = \begin{cases} 0, & d(w, w') > D \\ b \sum_{i=1}^{L} m_i I(w_i \neq w'_i), & d(w, w') \leq D. \end{cases} \tag{3.6}$$

The above model defines a a hypercube on vertices $V = \mathcal{W}_L$, where there is an edge between two words if they are within D of each other a hamming distance. If the two words are connected by an edge, then the weight on that edge depends on the position(s) of mismatch. We let m_i be small for i in the middle of the motif, and large for i in the flanking regions. The parameter b controls the strength of the clustering effect. We enforce $m_1 + \cdots + m_L = 1$.

In the example below we will let $D = 1$, $L = 7$, and

$$m_i = \begin{cases} 1, & i \in L_1; \\ 0, & i \in L_2. \end{cases}, \tag{3.7}$$

where $L_1 = \{1, 2, L-1, L\}$ are the "flanking regions" and $L_2 = \{3, \ldots, L-2\}$ are the "core regions" where no mismatch is allowed.

3.2.3 Hyperparameter Selection

We discuss mainly the choice of the sparsity parameter a and the smoothness parameter b for the hypercube graph. The choice of v follows similar considerations as in the linear Markov model. However, unlike in the linear Markov case, there are no analytic formulas for the prior model size $\sum_{i=1}^{m} P(\gamma_i = 1)$, and no direct interpretation of the parameter b as for the parameters w_0 and w_1 in terms of the prior odds in (3.2). More importantly, unlike the previous 1-dimensional (linear Markov chain) case, the Ising model on most graphs of dimension two or higher (including the hypercube) exhibits phase transition behavior.

Generally speaking, the Ising model undergoes transition between an ordered and a disordered underlying state at or near the phase transition point, leading to various dramatic consequences such as the critical slow down of the MCMC. But the most relevant consequence to our application is the drastic change in the proportion of $\gamma_i = 1$ (e.g., from $< 1\%$ to $> 90\%$) in the prior and consequently the posterior distribution with a tiny change in b at or near the phase transition point. Since the computational cost of sampling from the posterior of

$\boldsymbol{\gamma}$ is of quadratic order of the model size, (a, b) must be chosen to avoid the phase transition point and guarantee a small model size on average. We recommend simulating first from the prior of $\boldsymbol{\gamma}$ to aid in the choice of (a, b). Furthermore, the behavior of an Ising model on a wide class of regular graphs can be approximated by mean field equations (see, for example, 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 major difficulty in analyzing a high dimensional Ising model lies in the analytical intractability of the partition function $\psi(a, b)$, due to the complicated combinatorics generated by the interaction terms when summing over all states, i.e., $\sum_{(i,j)\in\mathcal{E}} \gamma_i\gamma_j$. The main idea of mean field theory is to replace all interactions to any one vertex with an average or effective interaction, which, for some graphs such as the hypercube, becomes exact as the dimension goes to infinity. A brief derivation of the mean field equations of the hypercube model is given in Appendix 6.2.

As shown in Appendix 6.2, the key to track the phase transition of the Ising model is to study the nature of the minimizer of the mean field approximate $\phi(p)$,

$$\phi(p) = \log(1-p) - \left(a + \log\frac{1-p}{p}\right)p - kbp^2, \tag{3.8}$$

where k is the degree of the hypercube, and $0 < p < 1$. To minimize $\phi(p)$, we look for solutions $\hat{p}$ to

$$\frac{d\phi}{dp} = -\log\left(\frac{1-p}{p}\right) - a - 2kbp = 0, \tag{3.9}$$

that satisfy

$$\frac{d^2\phi}{dp^2} = \frac{1}{p(1-p)} - 2kb > 0.$$

The solutions can be found numerically. To study it qualitatively, the left panel of Figure 7 shows the two sides of equation (3.9) for varying kb. The intersection of the line and the logit function are possible solutions $\hat{p}$ for fixed (a, kb). The nature of the solutions are can be described as follows:

1. When $a > -2$: there is one minima of $\phi(p)$.
2. When $a = -2$: there is one inflation point (i.e., $\frac{d^2\phi}{dp^2} = 0$), $p^* = \frac{1}{2}$.
3. When $a < -2$: let the two solutions to equation $\text{logit}(p) = a + \frac{1}{1-p}$ be $p_1^*(> 1/2)$ and $p_2^*(< 1/2)$. Then when $\frac{1}{p_2^*(1-p_2^*)} < 2kb < \frac{1}{p_1^*(1-p_1^*)}$, there are two minima and one maxima of $\phi(p)$; when $b = \frac{1}{p_2^*(1-p_2^*)}$ or $2kb = \frac{1}{p_1^*(1-p_1^*)}$, one minima and one inflation point; when $2kb < \frac{1}{p_2^*(1-p_2^*)}$ or $2kb > \frac{1}{p_1^*(1-p_1^*)}$, one minima.

Therefore, for any given $a < -2$, the mean field approximate $\phi(p)$ transits between uni-mode and multi-mode states at $b_i^* = \frac{1}{2kp_i^*(1-p_i^*)}, (i = 1, 2)$, which are the phase transition points. The right panel of Figure 7 shows these points $(a, 2kb^*)$. In theory, for given hyperparameter a, any b that is well above the solid line $(> b_1^*(a))$ or below the dashed line $(> b_1^*(a))$ should avoid phase transition in the Ising model. But for the computational efficiency (limiting model size), we choose b that is below the dashed line in our application.

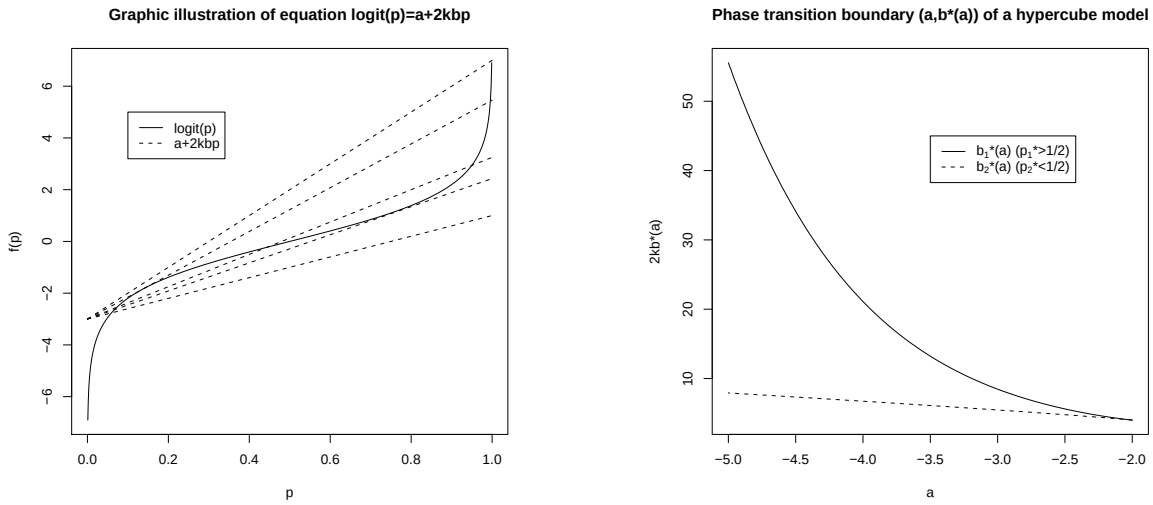

Figure 7. Phase transition boundary of Ising model

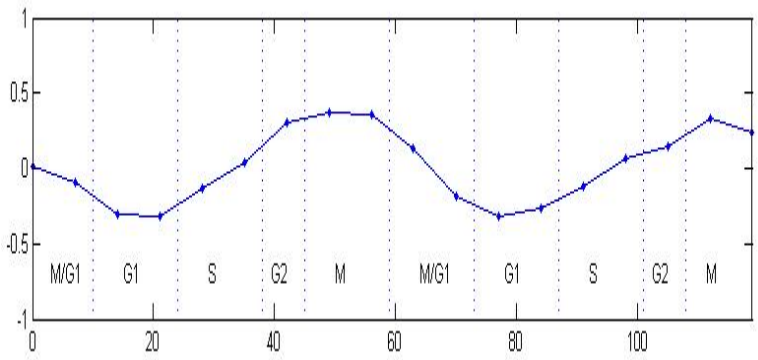

Figure 8. Loadings of the first principal component for yeast cell cycle data set.

3.2.4 Analysis of Spellman et al. (1998) 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 data set that has been analyzed previously by many motif finding methods (Bussemaker et al., (2001), Zhang et al. (2007)). Previous regression based approaches have used as covariates either nondegenerate words, degenerate words on the IUPAC alphabet, or a known set of pre-curated PSSMs. A reliable list of pre-curated 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 $4^7 = 16384$). 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 data set 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 et al. (2007) and use the scores of the first principal component of the data, the loadings of which are plotted versus time in Figure 8. A minor technical detail is that in yeast, a word and its reverse complement

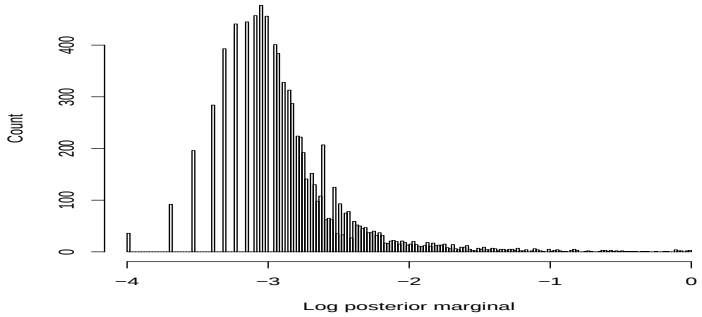

Figure 9. Histogram of $\log_{10} P(\gamma_i = 1|\mathbf{Y})$ for Spellman et al. yeast cell cycle data set.

should be considered the same motif. Thus, there are 8192, instead of $4^7 = 16384$, 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 not hard to show that for length 7 words, these 8192 covariates still lie on a hypercube with degree $k = 6$. We use the model in (3.7) with $a = -5$, $2kb = 14$.

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 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 2. 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 2. The island of 4 motifs comprising `GCCCGTT, GCCCGAT, GTCCGAT, GTCCGCT` are a putative MCM1 domains (Zhang et al., 2007). 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 et al. (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

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	
Island 4, 2 words:			GTCCGAT	0.0633	MCM1
CTGCGCT	0.3865		GTCCGCT	0.097	
CTGCGTT	0.0962	RME1	Island 4, 2 words:		
Island 5, 2 words:			TGTTTGT	0.8589	
TCGCGTC	0.2053		TGTTTTT	0.1202	STE12
GCGCGTC	0.2017		Island 5, 2 words:		
Island 6, 2 words:			TTTCGCG	0.8318	SCB
TTGGTCG	0.1029		TTTCGTG	0.79	SCB
TCGGTCG	0.0742	MCM1	Island 6, 2 words:		
Island 7, 2 words:			CTGCGCT	0.4159	
GCCGACT	0.0992	BAS1	CTGCGTT	0.1423	RME1
GCCGACG	0.0541	BAS1	Island 7, 2 words:		
Island 8, 2 words:			TAGCCAG	0.3352	
TTGTTTA	0.0941	SFF, ROX1	TAGCCGG	0.1142	
TTGTTTT	0.064	ROX1	Island 8, 2 words:		
			TCGCGTC	0.2332	
			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

Table 2. Islands in top 100 motifs ranked by $P(\gamma_i = 1|\mathbf{Y})$ from hypercube model.

obtained from the model that assumes prior independence of $\boldsymbol{\gamma}$. Out of the top 100 motifs in the hypercube model, there are 8 islands comprising 19 different motifs, which are also listed in Table 2. 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.

Detailed lists of the top 100 motifs found by each model, and their marginal probabilities, are given in Supplementary table 1.

4 Discussion

Model building in high dimensional covariate spaces with a priori known structure is an important 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. Ising priors has been applied to smoothing-type problems such as segmentation of MRI images in Smith and Fahrmeir (2007), which specifically looked at two and three dimensional lattices. To our knowledge, its utility in guiding model selection has not been explored. We proposed a general framework that can flexibly adapt to a large variety of problems. As illustration, we studied two problems in genomics in Section 3. In both problems the a priori known structure in the covariate space can be encoded into regular graphs, but the different nature of the graphs called for different approaches to hyperparameter selection. 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 simulating from the prior, guided by mean field approximations, is useful in this context.

When the covariate space is large, computational efficiency is a main concern dictating both the distributional form of the prior for $\boldsymbol{\gamma}, \boldsymbol{\beta}$ as well as the choice of hyperparameters. In the application of Gibbs sampling to variable selection, George and McCullough (1993) first proposed sampling from the auxiliary Markov chain

$$\boldsymbol{\beta}^{(1)}, \sigma^{(1)}, \boldsymbol{\gamma}^{(1)}, \boldsymbol{\beta}^{(2)}, \sigma^{(2)}, \boldsymbol{\gamma}^{(2)}, \ldots,$$

which has been a popular alternative to the direct sampling scheme (2.4). However, it is important to note that the auxiliary sampling strategy does not allow for sparse models, because the distribution on β_i conditional on $\gamma_i = 0$ must be non-degenerate to ensure ergodicity of the Markov chain. Assuming a point mass at 0 for the distribution of β_i when $\gamma_i = 0$ allows for $O(md^2)$ computation time for each sweep of $\boldsymbol{\gamma}$, where d is the model size. This is why the direct sampling strategy, coupled with the prior (2.2), is especially computationally attractive in high dimensions.

Introducing the smoothing parameter b in the prior distribution for $\boldsymbol{\gamma}$ also increases the stickiness of the Markov chain, and thus causes slower mixing rate. However, in both the simulation and real data examples that we explored, the mixing rate was very fast 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 and the group Lasso have been proposed for structured variable selection in high dimensional settings. However, the underlying model assumptions for these methods are very different than those proposed in this paper: The former enforces smoothness in $\boldsymbol{\beta}$ while the latter assumes smoothness in $\boldsymbol{\gamma}$. This easily dismissed but not-too-subtle distinction might be important in some applications.

The methods described in this paper can be extended to nonlinear regression for binary and categorical outcomes, and accelerated failure time models for survival outcomes.

R and Fortran code is available at

`http://www-stat.stanford.edu/ nzhang/BVS/.`

5 Acknowledgements

We thank Alan Zaslavsky for constructive comments and general support. We also thank Andrea Montanari for helpful discussions.

6 Appendix

6.1 Fast updating of A_i^{-1}

Here we describe the 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 form:

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

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

$$A_i^{-1} = \begin{pmatrix} A^{11} & A^{12} \\ A^{21} & A^{22} \end{pmatrix}, \tag{6.1}$$

where

$$\begin{cases} A^{11} &= (A_{(-i)} - \Sigma_{I_{(-i)},i}\sigma_{ii}^{-1}\Sigma'_{I_{(-i)},i})^{-1} \stackrel{\text{def}}{=} (A_{11\cdot 2})^{-1} \\ A^{12} &= -(A_{11\cdot 2})^{-1}\Sigma_{I_{(-i)},i}\sigma_{ii}^{-1} \\ A^{21} &= -\sigma_{ii}^{-1}\Sigma'_{I_{(-i)},i}(A_{11\cdot 2})^{-1} \\ A^{22} &= \sigma_{ii}^{-1} + \sigma_{ii}^{-1}\Sigma'_{I_{(-i)},i}(A_{11\cdot 2})^{-1}\Sigma_{I_{(-i)},i}\sigma_{ii}^{-1} \end{cases}.$$

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

$$A^{11} = A_{(-i)}^{-1} + \frac{1}{\sigma_{ii}(1 - \Sigma'_{I_{(-i)},i} A_{(-i)}^{-1} \Sigma_{I_{(-i)},i}/\sigma_{ii})} A_{(-i)}^{-1} \Sigma_{I_{(-i)},i} (A_{(-i)}^{-1} \Sigma_{I_{(-i)},i})'. \tag{6.2}$$

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 $O(m_{(-i)}^2)$.

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 $A_i^{-1} = L_i L_i'$. Notice the right side of equation (6.2) 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 $1 \times m_{(-i)}$, and $\mathbf{0} = (0, ..., 0)'_{m_{(-i)}}$. This implies

$$A_i^{-1} = \begin{pmatrix} A^{11} & \widetilde{L}_{(-i)} L'_{(-i),i} \\ L_{(-i),i} \widetilde{L}'_{(-i)} & L_{(-i),i} L'_{(-i),i} + l_{ii}^2 \end{pmatrix}. \tag{6.3}$$

Comparing expressions (6.1) and (6.3), 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 $O(m_{(-i)}^2)$.

The efficiency of the updating can be further improved by the systematic relation between the two consecutive A matrices in a sweep (say, sweep j), as shown below

γ_{i-1}^j	γ_i^{j-1}	$A_{(-i)}^j$	A_i^j
0	0	$A_{-(i-1)}^j$	add a row/column from $A_{-(i-1)}^j$
0	1	delete a row/column from $A_{-(i-1)}^j$	$A_{-(i-1)}^j$
1	0	A_{i-1}^j	add a row/column from A_{i-1}^j
1	1	delete a row/column from A_{i-1}^j	A_{i-1}^j

Since we are interested in sparse models, most calculation will follow the first case (i.e.,$\gamma_{i-1}^j = 0, \gamma_i^{j-1} = 0$) in the above table, which is relatively simple.

6.2 Mean field approximation of the hypercube model

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

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

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

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

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

$$\psi(1) \leq \min_{E_0 \in \mathcal{F}} \left\{ -\log\left[\sum_{\boldsymbol{\gamma}} e^{-E_0(\boldsymbol{\gamma})}\right] + \mathbb{E}_0[E(\boldsymbol{\gamma}) - E_0(\boldsymbol{\gamma})] \right\}. \tag{6.4}$$

The idea in mean field approximations is to choose a class of energy functions $\mathcal{F}$ simple enough so that the minimization in (6.4) is analytically tractable. Often, the choice is the class of linearly additive energy functions:

$$E_0(\boldsymbol{\gamma}) = -\sum_i h_i \gamma_i, \tag{6.5}$$

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

Let $\phi(\boldsymbol{h})$ be the function being minimized in (6.4) for $\mathcal{F}$ defined as in (6.5):

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

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:

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

Now, in the hypercube model, we assume that all edges have the same weight $b_{ij} = b$ (this is the smoothing parameter), and that $a_i = a$ (this is the *external field*). Thus, due to the symmetry in the model, the optimizing $\boldsymbol{h}$ 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 is twice the degree of each node in the hypercube, and to make things simpler we reparameterize $p = (1+e^{-h})^{-1}$. With a slight abuse of notation, this gives us:

$$\frac{\phi(h)}{n} = \phi(p) = \log(1-p) - \left(a + \log\frac{1-p}{p}\right)p - kbp^2. \tag{6.6}$$

For any given a, the phase transition points are the b^*'s that introduces a change in the nature of the minimizer p of equation (6.6), as discussed in Section 3.2.3.

References

[1] Bentz, M., Dohner, H., Huck, K., Schutz, B., Ganser, A., Joos, S., du Manoir, S., and Lichter, P. (1995). Comparative genomic hybridization in the investigation of myeloid leukemias. *Genes Chromosomes Cancer* **12**, 193-200.

[2] Brown, P.J., Vannucci, M. and Fearn, T. (1998). Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society, Series B* **60(3)**, 627-641.

[3] Brown, P.J., Vannucci, M. and Fearn, T. (2002). Bayes model averaging with selection of regressors. *Journal of the Royal Statistical Society, Series B* **64(3)**, 519-536.

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

[5] Bussemaker, H.J., Li, H. and Siggia, E.D. (2001). Regulatory element detection using correlation with expression. *Nature Genetics* **27(2)**, 167-171.

[6] 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 National Academy of Science USA* **100(6)**, 3339-3344.

[7] George, E. and McCulloch, R.E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88**, 881-889.

[8] George, E. and McCulloch, R.E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica* **7**, 339-373.

[9] Jen, J., Kim, H., Piantadosi, S., Liu, Z.F., Levitt, R.C., Sistonen, P., Kinzler, K.W., Vogelstein, B. and Hamilton, S.R. (1994). Allelic Loss of Chromosome 18q and Prognosis in Colorectal Cancer. *The New England Journal of Medicine* **331**, 213-221.

[10] Ji, H., Kumm, J., Zhang, M., Farnam, K., Salari, K., Faham, M., Ford, J.M., and Davis, R.W. (2007). Molecular inversion probe analysis of gene copy alterations reveals distinct categories of colorectal carcinoma. *Cancer Research* **66(16)**, 7910-7919.

[11] 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.

[12] Mehta, K.R., Nakao, K., Zuraek, M.B., Ruan, D.T., Bergsland, E.K., Venook, A.P., Moore, D.H., Tokuyasu, T.A., Jain, A.N., Warren, R.S., Terdiman, J.P. and Waldman FM. (2005). Fractional genomic alteration detected by array-based comparative genomic hybridization independently predicts survival after hepatic resection for metastatic colorectal cancer. *Clinical Cancer Research* **11(5)** 1791-7.

[13] Matsuyama, H., Oba, K., Matsuda, K., Yoshihiro, S., Tsukamoto, M., Kinjo, M., Sagiyama, K., Takei, M., Yamaguchi, A., Sasaki, K. and Naito, K. (2007). Haploinsufficiency of 8p22 may influence cancer-specific survival in prostate cancer. *Cancer Genetics and Cytogenetics* **174(1)**, 24-34.

[14] Mirny, L. and Gelfand, M. (2002). Structural analysis of conserved base pairs in protein-DNA complexes. *Nucleic Acids Research* **30**, 1704-1711.

[15] Mitchell, T. J. and Beauchamp, J.J. Bayesian Variable Selection in Linear Regression. (1988). *Journal of the American Statistical Association* **83**, 1023-1032.

[16] Mitelman, F., Mertens, F. and Johanson, B. (1997). A breakpoint map of recurrent chromosomal rearrangements in human neoplasia. *Nature Genetics* **15**, 417-474.

[17] 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.

[18] Nott, D. and Green, P.J. (2004). Bayesian variable selection and the Swendsen-Wang algorithm. *Journal of Computational and Graphical Statistics* **13**, 141 - 157.

[19] 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.

[20] 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.

[21] Spellman, P.T., Sherlock, G., 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.

[22] 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.

[23] Tagawa, Y., Yasutake, T., Sawai, T., Nanashima, A., Jibiki, M., Morinaga, M., Akama, F., Nakagoe, T. and Ayabe, H. (1997). Clinical and pathological significance of numerical aberrations of chromosomes 11 and 17 in colorectal neoplasms. *Clinical Cancer Research* **3(9)**, 1587-1592.

[24] Thijs G, Marchal K, Lescot M, Rombauts S, Moor BD, Rouz P, Moreau Y. (2002). A Gibbs sampling method to detect overrepresented motifs in upstream regions of coexpressed genes. *Journal of Computational Biology* **9**:447-464.

[25] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58(1)**, 267-288.

[26] 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, Series B* **67(1)**, 91-108.

[27] Yedidia, J.S. (2001). An Idiosyncratic Journey Beyond Mean Field Theory. *Advanced Mean Field Methods, Theory and Practice* 21-36. The MIT Press.

[28] Yuan, M. and Lin, Y. (2006). Model Selection and Estimation in Regression with Grouped Variables. *Journal of the Royal Statistical Society, Series B* **68(1)**, 49-67

[29] Zhang, N.R., Wildermuth, M.C., Speed, T.P. Transcription Factor Binding Site Prediction with Multivariate Gene Expression Data. *Annals of Applied Statistics*, in press.

[30] Zhu, J. and Zhang, M.Q. (1999). SCPD: A Promoter Database of Yeast Saccharomyces cerevisiae. *Bioinformatics* **15**, 607-611