Bayesian Variable Selection in Regression with Networked Predictors *

Feng Tai † Wei Pan ‡ Xiaotong Shen §

Abstract

We consider Bayesian variable selection in linear regression when the relationships among a possibly large number of predictors are described by a network given *a priori*. A class of motivating examples is to predict some clinical outcomes with high-dimensional gene expression profiles and a gene network, for which it is assumed that the genes neighboring to each other in the network are more likely to participate together in relevant biological processes and thus more likely to be simultaneously included in (or excluded from) the regression model. To account for spatial correlations induced by a predictor network, rather than using an independent (and identical) prior distribution for each predictor's being included in the model as implemented in the standard approach of stochastic search variable selection (SSVS), we propose a Gaussian Markov random field (MRF) and a binary MRF as priors. We evaluate and compare the performance of the new methods against the standard SSVS using both simulated and real data.

Keywords: Binary Markov random field (BMRF); Gaussian Markov random field (GMRF); gene network; high dimension; Markov chain Monte Carlo (MCMC); microarray data; stochastic search variable selection (SSVS).

1 Introduction

We consider linear regression with "large p, small n" data as arising in genomics, in which we would like to predict soem clinical outcome using high-dimensional gene expression profiles. In such

*The authors were partially supported by NIH grants HL65462 and GM081535.

†Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN 55455, USA

‡Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN 55455, USA, Email: weip@biostat.umn.edu

§School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA, Email: xshen@stat.umn.edu

an application, variable (or gene) selection is crucial for predictive performance and ellucidating underlying biological processes. Most exsiting methods for variable selection are generic, ignoring subject-matter prior knowledge on predictors. For example, a popular Bayesian variable selection method is the Stochastic Search Variable Selection (SSVS) proposed by George and McCulloch (1993, 1997). SSVS introduces a latent binary vector $\boldsymbol{\gamma}$ to indicate whether a predictor or variable should be included in the model or not, and uses a Bayesian hierarchical model to estimate $\boldsymbol{\gamma}$ for variable selection. The regression coefficient β_i follows a normal mixture distribution, $\pi(\beta_i|\gamma) = (1-\gamma_i)N(0, v_0) + \gamma_i N(0, v_1)$. Lee *et al.* (2002) applied SSVS to microarray data in the context of classification, using a mixture of a normal and a point mass instead, $\pi(\beta_i|\gamma) = (1-\gamma_i)I_0 + \gamma_i N(0, v_1)$, treating all the genes equally apriori by giving independent and identical priors for the probability of a gene being in the final model; i.e. $\pi(\gamma) \equiv 1/2^p$. In Bae and Mallick (2004), instead of using indicator vector γ for variable selection, they modeled β by assuming $\beta|\Lambda \sim N(0, \Lambda), \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p)$ and put three different priors on Λ. Variable (gene) selection was based on applying a threshold on the posterior of λ_i to eliminate genes with small variances λ_i.

On the other hand, there has been rapidly accumulating biological knowledge in the form of various gene networks. A gene network can be expressed as an undirected graph with nodes representing genes and edges representing interactions between genes, which provides a natural neighborhood structure for any gene. The importance of incorporating biological knowledge into genomic data analysis has been increasingly recognized. For instance, in a different context of detecting differentially expressed genes, Wei and Li (2007) proposed a binary Markov random field (BMRF) model to account for the local dependency of the genes in a network, while Wei and Pan (2008) proposed a Gaussian markov random field (GMRF) model for the same purpose. In the context of linear regression, Li and Li (2008) and Pan et al (2009) proposed network-based penalty functions for variable selection in the framework of penalized regression, in which some smoothness assumption on the regression coefficients is imposed. Here we would like to take a Bayesian approach, which differs from the above penalized regression methods in that we have a less stringent smoothness assumption: we only assume the smoothness of the prior probabilities of the predictors' being selected, rather than of their effect sizes (i.e. regression coefficients). Specifically, we investigate three different spatial priors in the framework of SSVS, targeting applications to regression analysis for high-dimensional microarray data. Instead of treating all the genes independently and identically *a priori*, we assign dependent priors to reflect the relationships among the genes over a gene network. We introduce three different priors to model the potential spatial correlations among the genes based on their network structure. Specifically, we assume the probability of a gene's being informative

depends on that of its direct neighbors in the network. In other words, we assume the spatial dependency among γs as induced by the network.

Markov random field models for binary spatially correlated variables have been widely used in image analysis and spatial statistics to account for local dependencies. The basic autologistic model was developed by Besag (1972, 1974) with a broad range of applications, as shown by Heikkinen and Högmander (1994) and Hoeting *et al.* (2000). Weir and Pettitt (2000) proposed a hidden conditional autoregressive Gaussian process to model binary spatially correlated responses. Wei and Pan (2008) used a GMRF to model the prior probabilities of the binary statuses of some binary variables, and Wei and Pan (2009) compared the performance of the independence, GMRF and BMRF priors in the same context. Smith and Smith (2006) compared three binary Markov random fields, which are popular Bayesian priors for spatial smoothing. Smith and Fahrmeir (2007) extended Bayesian variable selection to a series of spatially linked regressions, incorporating the spatial correlation among the indicators $\boldsymbol{\gamma}$ by specifying a binary markov random field prior. It is very similar to, but not exactly the same as, our method. They placed an Ising prior on some binary indicator variables across multiple regressions. A difference between their method and ours is that they had repeated measures of each covariates from multiple sites, resulting in a matrix of binary indicators $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_N)$ from locations (1,. . . ,N). They modeled spatial correlations across different sites within each covariate. Specifically, for a N-dimension binary vector of covariate j, $\boldsymbol{\gamma}_j = (\gamma_{j1}, \ldots, \gamma_{jN})'$, all elements in $\boldsymbol{\gamma}_j$ are assumed to be spatially correlated, but for all p covariates, $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_p$ are assumed to be independent (i.e. $p(\boldsymbol{\gamma}) = \prod_{j=1}^{p} \boldsymbol{\gamma}_j$). However, in our method, we only have one "site" ($N = 1$), and we consider the spatial correlation between covariates instead of within covariates. All elements of a p-vector $\boldsymbol{\gamma}_N = (\gamma_{1N}, \ldots, \gamma_{pN})$ are spatially correlated based on a given network. During the preparation of this manuscript, we learned the recent work of Li and Zhang (2008), who proposed an Ising model to introduce a spatial prior for γ; Monni and Li (2009) proposed a different network-based prior for γ and considered both linear models for continuous responses and probit models for binary responses. In addition to some differences from theirs in implementations and applications, here we also study a GMRF model and a scaled BMRF (SBMRF) model.

The rest of this paper is organized as follows. We first review SSVS, then propose our new methods with three Markov random field (MRF) models as priors: GMRF, BMRF and SBMRF. After describing some details on the posterior distributions and sampling schemes, we apply our methods to both simulated and real data, followed by a short discussion.

2 Statistical Models

2.1 Review of SSVS

SSVS (George and McCulloch 1993, 1997) starts from the standard linear model

$$f(Y|\beta,\sigma) = N_n(X\beta,\sigma^2 I),$$

where Y is a $n \times 1$ vector of the response variable and $X = (X_1,\ldots,X_p)$ is an $n \times p$ matrix of predictors. The regression coefficient β is a $p \times 1$ unknown vector and σ is an unknown positive scalar.

In order to conduct variable selection, we define a vector

$$\gamma = (\gamma_1,\ldots,\gamma_p)',$$

where $\gamma_i = 1$ or 0 indicates whether predictor i should be included in or excluded from the model respectively. We model the uncertainty underlying variable selection by a mixture prior $\pi(\beta,\sigma,\gamma) = \pi(\beta|\sigma,\gamma)\pi(\sigma|\gamma)\pi(\gamma)$, which can be conditionally specified as follows,

$$\pi(\beta|\sigma,\gamma) = N_p(0, D_\gamma R_\gamma D_\gamma),$$

where R_γ is a correlation matrix and D_γ is a diagonal matrix with its ith diagonal element denoted by

$$(D_\gamma^2)_{ii} = \begin{cases} v_{0_\gamma} \text{ if } \gamma_i = 0, \\ v_{1_\gamma} \text{ if } \gamma_i = 1. \end{cases}$$

With this prior, each component of β is modeled as coming from a mixture of two normals

$$\pi(\beta_i|\sigma,\gamma) = (1-\gamma_i)N(0, v_{0_{\gamma(i)}}) + \gamma_i N(0, v_{1_{\gamma(i)}}).$$

The idea of variable selection is that, by setting $v_{0_{\gamma(i)}}$ and $v_{1_{\gamma(i)}}$ "small" and "large" respectively, if the data supports $\gamma_i = 0$ over $\gamma_i = 1$, then β_i should be small enough so that the corresponding predictor X_i plays a negligible role and thus should be excluded from the model. A simple choice for R_γ is $R_\gamma = I$. The residual variance σ^2 is conveniently modeled by an inverse gamma distribution,

$$\pi(\sigma^2|\gamma) = IG(\nu,\lambda).$$

The prior for γ has the form

$$\pi(\gamma) = \prod w_i^{\gamma_i}(1-w_i)^{1-\gamma_i}.$$

For simplicity, usually $\pi(\gamma) \equiv 1/2^p$ is used to substantially reduce computational cost. We interpret $w_i = P(\gamma_i = 1)$ as the prior probability that β_i is large enough to have X_i included in the model.

Based on data Y, the posterior $\pi(\gamma|Y)$ updates the prior probabilities on each of the 2^p possible values of γ. The γs with higher posterior probabilities $\pi(\gamma|Y)$ identify the more promising models that are more strongly supported by the data and the prior distribution. MCMC is usually used to explore the posteriors of β, σ and γ.

2.2 Spatial priors for γ

For the standard SSVS, $\pi(\gamma) = \prod w_i^{\gamma_i}(1-w_i)^{1-\gamma_i}$, which implies the components of γ are *a priori* independent. In other words, the genes are treated independently apriori, and are further assumed to have the same prior probabilities to be included in the model by specifying $w_i \equiv w_0$ for all i, where w_0 is a pre-specified constant. In order to account for the dependency among the genes over a gene network, we propose to incorporate biological knowledge of the gene network by specifying a spatial prior for γ over the gene network. A gene network can be expressed as an undirected graph with nodes for genes and edges for interactions between genes, which provides a natural neighborhood structure for a Markov Random Field (MRF). Here, we consider two different MRF models as priors.

1 Gaussian Markov Random Field (GMRF)

We define θ_i as a logit transformation of $w_i = Pr(\gamma_i = 1)$

$$\theta_i = \log\left(\frac{w_i}{1-w_i}\right),$$

and model θ_i by an Intrinsic Gaussian Conditional Autoregression (ICAR) model (Besag and Kooperberg 1995):

$$\theta_i|\theta_{(-i)} \sim N\left(\frac{1}{m_i}\sum_{j\in\delta_i}\theta_j, \frac{\tau^2}{m_i}\right),$$

where $\theta_{(-i)} = \{\theta_j : j \neq i\}$, δ_i is a set of indices of direct neighbors of gene i, and $m_i = |\delta_i|$ is the size of δ_i as determined by a given gene network. The ICAR model accounts for spatial correlations and smoothness among the prior probabilities of the genes' being included in the model. The same idea can be found in Wei and Pan (2008), but in a different context.

2 Binary Markov Random Field (BMRF)

Instead of specifying a full conditional distribution of θ_is as in the ICAR model, a BMRF specifies a full conditional distribution of γ directly,

$$\pi(\gamma_i|\gamma_{(-i)}) \propto \exp(\alpha_0 + \alpha_1 k_i),$$

where $\gamma_{(-i)} = \{\gamma_j : j \neq i\}$, m_{i0} and m_{i1} are the numbers of $\gamma_j = 0$ and $\gamma_j = 1$ for $j \in \delta_i$ respectively, and $k_i = m_{i1} - m_{i0}$. This model is also called an autologistic model. The joint distribution of γ involves a normalizing factor $Z(\alpha)$, which depends on $\alpha = (\alpha_0, \alpha_1)'$ and is analytically intractable. A simple alternative to estimate α is to use a pseudo-likelihood approximation:

$$\text{pl}(\alpha) = \prod_i \pi(\gamma_i|\gamma_{(-i)}).$$

Using the pseudo-likelihood is equivalent to regressing θ_i on k_i,

$$\theta_i = \log\left(\frac{w_i}{1-w_i}\right) = \alpha_0 + \alpha_1 k_i.$$

Notice that α_0 is closely related to the marginal probability $Pr(\gamma_i = 1|\theta_i)$ for all $k_i = 0$. In practice, we specify α_0 to control the overall number of the genes (or variables) to be selected a priori. $\alpha_1 > 0$ is usually assumed, indicating that γ_i has a higher probability to be 1 than 0 if the number of 1's is greater than the number of 0's in its neighborhood. Another alternative is to replace k_i by a scaled $k_i^* = k_i/m_i$, where $m_i = m_{i1} + m_{i0}$ is the neighborhood size for gene i (Wei and Li 2008); we call it the scaled Binary Markov Random Field (SBMRF)

3 Estimation

3.1 Gibbs sampling

We use the Gibbs sampling to simulate posterior distributions. The full conditional posterior distribution for β is a multivariate normal distribution

$$Pr(\beta|\sigma, \gamma, Y) = N(\Lambda X'Y, \sigma^2\Lambda),$$

where $\Lambda = (X'X + \sigma^2(D_\gamma R_\gamma D_\gamma)^{-1})^{-1}$, and we choose $R_\gamma = I$ for simplicity. σ^2 follows an inverse gamma distribution

$$Pr(\sigma|\beta, Y) = IG\left(\frac{n}{2} + \nu, \frac{1}{2}||Y - X\beta||_2 + \lambda\right).$$

1 GMRF

For GMRF, we have

$$Pr(\gamma_i|\beta,\theta) = Ber\left(\frac{a_i}{a_i+b_i}\right),$$
$$a_i = f(\beta_i|\gamma_i=1)\cdot\frac{\exp(\theta_i)}{1+\exp(\theta_i)},\ b_i = f(\beta_i|\gamma_i=0)\cdot\frac{1}{1+\exp(\theta_i)}. \tag{3.1}$$

The joint distribution of θ given all other parameters under the ICAR specification is

$$Pr(\theta|\gamma,\tau^2) \propto \left(\prod_i \frac{\exp(\gamma_i\theta_i)}{1+\exp(\theta_i)}\right)\ \exp\left(-\frac{1}{2\tau^2}\sum_{i\neq j} w_{ij}(\theta_i-\theta_j)^2\right)$$

and using an inverse gamma as the prior of τ^2 leads to

$$Pr(\tau^2|\theta) = IG\left(\frac{p-1}{2}+0.5, \frac{1}{2}\sum_{i\neq j} w_{ij}(\theta_i-\theta_j)^2+0.005\right).$$

Rather than drawing θ as a vector, it is better to draw it component-wise from

$$Pr(\theta_i|\gamma,\tau^2,\theta_{j\neq i}) \propto \left(\frac{\exp(\gamma_i\theta_i)}{1+\exp(\theta_i)}\right)\ \exp\left(-\frac{m_i}{2\tau^2}(\theta_i-\frac{1}{m_i}\sum_{j\in\delta_i}\theta_j)^2\right).$$

Due to the log-concavity of $Pr(\theta_i|\gamma,\tau^2,\theta_{j\neq i})$, an adaptive rejection sampling can be directly applied. Under the ICAR specification, the mean of θ_i's is undetermined. Hence, we put a constraint that $\sum_i \theta_i = \theta_0$, where θ_0 is a fixed number to reflect the prior belief of the proportion of the variables to be selected in the model. In practice, we found that sampling τ^2 and θs at the same time might cause some convergence problems. Thus, we fixed $\tau^2 = 0.49$ in the following simulations and real data examples.

2 BMRF

For BMRF or SBMRF, we have

$$Pr(\gamma_i|\beta,\theta) = Ber\left(\frac{a_i}{a_i+b_i}\right),$$
$$a_i = f(\beta_i|\gamma_i=1)\cdot\frac{\exp(\alpha_0+\alpha_1 k_i)}{1+\exp(\alpha_0+\alpha_1 k_i)},\ b_i = f(\beta_i|\gamma_i=0)\cdot\frac{1}{1+\exp(\alpha_0+\alpha_1 k_i)}. \tag{3.2}$$

And

$$Pr(\alpha|\gamma) = \left(\prod_i \frac{\exp(\gamma_i(\alpha_0+\alpha_1 k_i))}{1+\exp(\alpha_0+\alpha_1 k_i)}\right)\pi(\alpha_0)\pi(\alpha_1).$$

To ensure $\alpha_1 > 0$, we use a gamma prior $G(\lambda,\nu)$ for α_1 and have $\pi(\alpha_1) = \alpha_1^{\lambda-1}\exp(-\nu\alpha_1)$. In this way, $Pr(\alpha_1|\gamma)$ is log-concave for $\lambda \geq 1$. Thus, an adaptive rejection sampling can be directly applied. In our applications, we used $G(\lambda=3,\nu=0.5)$ as the prior, with most of its mass between 0 and 15, which was used by Hoeting *et al.* (2000).

3.2 Computation

To avoid a potential bias in parameter estimation, we updated the θ_i and γ_i in random orders. In MCMC sampling, the most costly step is to generate β from a multivariate normal distribution, which requires recomputing the inverse of a large covariance matrix. This step consumes almost the whole computing time due to the high dimensionality of the data. Thus in practice, the computing times are about the same for all four priors, even though the MRF priors have more parameters to estimate. For $p = 1329$ as in a real data example, the time of sampling 100 MCMC samples for all priors differed within 1 second.

3.3 Variable selection and response prediction

Variable selection is based on the marginal frequencies of the variables appearing in the posterior samples, i.e., the posterior mean of γ_is, reflecting the importance of each gene.

We predict a response by $\hat{y}$ based on each MCMC samples:

$$\hat{y} = \frac{1}{B}\sum_t X\hat{\beta}_t,$$

where B is the number of MCMC samples and $\hat{\beta}_t$ is the value of β in the tth MCMC sample. Thus the predictive model is not just only built on those genes with larger $\hat{\gamma}$, but possibly based on other genes.We also tried

$$\hat{y} = \frac{1}{B}\sum_t X\hat{\beta}_t\hat{\gamma}_t,$$

which produced similar results.

4 Results

To evaluate the performance of our proposed network-based SSVS, we conducted both simulations and real data studies with the four SSVS methods : the standard SSVS with an independent prior (SSVS+IND), SSVS with a GMRF prior (SSVS+GMRF), SSVS with a BMRF prior (SSVS+BMRF) and SSVS with a scaled BMRF prior (SSVS+SBMRF).

4.1 Simulations

Simulated data were generated from a linear regression model

$$Y = X\beta + \epsilon.$$

Two simple networks were considered.

1) A simple random network (RanN) that consisted of $p = 100$ variables. First, we randomly divided 100 variables into 10 groups, and generated a graph containing 10 subgraphs, each corresponding to one of the 10 groups of variables. Each subgraph was completely connected and there was no edges between any two subgraphs. Then we randomly deleted 300 edges ending up with a graph having 100 nodes and a total of 271 edges. Next, we randomly added some edges to connect the 10 subgraph together. One of the 10 groups was selected to be informative (with variables numbered from 20 to 34), which contained 15 variables and 50 edges as shown in Fig 4.1. Those informative βs were simulated from $N(0, 2^2)$ and remaining βs were set to 0. Lastly, we simulated X from a multivariate normal distribution, $X \sim MVN(0, \mathbf{I})$.
2) A simple regulatory network (RegN) as used by Li and Li (2008). We had 10 transcription factors (TFs), each of which formed a subnetwork with its 10 regulated genes; there was no connection between any two subnetworks, or between any two regulated genes. The resulting network consisted of 110 genes and 100 edges. We assumed two TFs and their regulated genes were informative with non-zero regression coefficients, while the regression coefficients for the other genes were zero:

$$\beta = (5, \underbrace{\frac{5}{\sqrt{10}}, \cdots, \frac{5}{\sqrt{10}}}_{10}, -3, \underbrace{\frac{-3}{\sqrt{10}}, \cdots, \frac{-3}{\sqrt{10}}}_{10}, 0, \cdots, 0)'.$$

The expression levels of TFs were drawn independently from standard normal, $X_{TF_j} \sim N(0, 1)$, and the expression levels of the genes that TF_j regulated followed $N(0.7X_{TF_j}, 0.51)$.

In both simulation set-ups, the random error ϵ was iid from $N(0, \sigma^2)$, where $\sigma^2 = \sum \beta_j^2/r$. We chose the signal-to-noise ratio (SNR) $r = 2$ or 4. For the random network, we specified $w_i = Pr(\gamma_i = 1) = 15/100 = 0.15$ for the independence (IND) prior, and the constraints $\theta_0 = \text{logit}(0.15)$ for the GMRF model and $\alpha_0 = \text{logit}(0.15)$ for the BMRF and SBMRF models. A similar set-up was used for the regulatory network, except $w_i = 22/110 = 0.2$ and $\theta_0 = \alpha_0 = \text{logit}(0.2)$. For each simulation run, we generated 50 training samples and 100 test samples; the simulation was repeated 100 times. In each run, 10,000 MCMC samples were generated with the first 8000 as the burn-in period. For the GMRF prior, we fixed $\tau^2 = 0.49$, finding that it worked well in practice. The starting values of θs were randomly generated from $N(\theta_0, 1)$. We randomly picked one simulation sample and applied three different random initial θs; the results were very stable, indicating convergence. The results shown in Table 1 were based on only one initial value of θ. The prediction mean-squared error (PMSE) was calculated for each test data set. In Table 1, column *ninfo* shows the number of true

informative genes in the top 15 (for RanN) or top 22 (for RegN) most frequently selected genes by each model. SSVS+GMRF had a smaller PMSE than SSVS+IND in all situations, but selected a smaller proportion of informative genes for the regulatory network. SSVS+SBMRF had a smaller PMSE than SSVS+IND and included more informative genes in all situations. SSVS+BMRF had smaller PMSE than SSVS+IND only for the regulatory network when SNR=4, and included more informative genes except for the random network when SNR=2. If we pool γs from 200 runs (100 each from SNR=2 or 4) for the random network, we found all models performed well in terms of gene selection. Histograms of γ are shown in Fig 4.2. A dot line indicates the cut-off point for distinguishing signal from noise genes. The cut-off points for all models as shown in Fig 4.2 completely separated signal and noise genes. In general, the MRF priors better separated signal and noise genes than the independence prior.

Table 1: Simulation results.

Network	SNR	prior	ninfo	pmse
RanN (15)	2	IND	6.66 (0.15)	62.15 (2.43)
		GMRF	12.96 (0.24)	53.11 (2.07)
		BMRF	4.55 (0.60)	71.19 (3.06)
		SBMRF	9.38 (0.31)	60.11 (2.50)
	4	IND	7.81 (0.13)	34.72 (1.33)
		GMRF	14.63 (0.08)	26.98 (1.07)
		BMRF	9.83 (0.62)	35.31 (2.11)
		SBMRF	11.77 (0.29)	32.78 (1.54)
RegN (22)	2	IND	16.08 (0.18)	55.52 (1.22)
		GMRF	13.36 (0.73)	55.11 (2.62)
		BMRF	21.19 (0.14)	57.99 (1.77)
		SBMRF	21.63 (0.11)	52.66 (1.30)
	4	IND	17.31 (0.15)	33.47 (0.78)
		GMRF	13.27 (0.69)	30.52 (2.10)
		BMRF	21.79 (0.09)	30.23 (1.20)
		SBMRF	21.98 (0.01)	28.86 (0.83)

4.2 Two Real Data Examples

1 Glioblastoma Data

We applied our proposed methods to a microarray gene expression data set of glioblastoma studied by Horvath *et al.* (2006). Glioblastoma is the most common primary malignant brain tumor of adults and one of the most lethal of all cancers. Patients with this disease have a median survival of 15 months from the time of diagnosis despite surgery, radiation and chemotherapy. Gene expression data from two independent sets of clinical tumor samples ($n = 55$ and $n = 65$) were obtained using Affymetrix HG U133A genechips. The RMA normalization method (Irizarry *et al.*, 2003) was applied to the gene expression data. Here we aimed to build a predictive model for log survival time and to identify biologically important genes. Nine patients that were still alive by the end of study were excluded from analysis, leading to 50 and 61 samples for two data sets respectively. We combined two data sets together and deleted two outliers, whose survival times were extremely short, resulting in a total of 109 subjects. We randomly split the data into two parts with 72 samples in the training and 37 in the test data. The gene network we used was a protein-protein interaction (PPI) network (Chuang *et al.* 2007). We mapped the genes in the microarray data to the PPI network and selected the largest subnetwork, which included 1329 genes. The prior probability for a gene being included in the model, w_i, was set to 0.05 for the independence prior in the standard SSVS, and the constraint θ_0 for the ICAR prior was set to logit(0.05). No intercept ($\alpha_0 = 0$) was fitted for the BMRF and SBMRF priors. We ran a total of 10000 MCMC iterations with a burn-in period of 8000 iterations, and the analysis was based on the last 2000 MCMC samples. The PMSEs for the methods are shown in Table 2.

Table 2: PMSEs for the glioblastoma data.

	IND	GMRF	BMRF	SBMRF
PMSE	0.54	0.55	0.64	0.53

In summary, for this example, the four priors for γ performed pretty similarly to each other in terms of prediction, though SSVS+BMRF performed slightly worse than others with a larger PMSE. For gene selection, as shown in Fig 4.3, $\hat{\gamma}$s for the independence prior and GMRF were roughly normally distributed around the specified prior at 0.05, and for the BMRF prior it was also normally distributed around 0.02. The BMRF prior seemed to better separate the informative and non-informative genes, however, it also included much more genes. Since our prior was set to reflect the belief of 5% of informative genes in a total of 1329 genes, we plotted the top 66 selected genes

for all priors except BMRF (not shown); the network structures looked very similar, though most of the selected genes did not overlap. For this dataset, the similar performance of the methods in terms of PMSE (Table 2) and their widely varied genes being selected can be presumably explained by the fact that the genes were barely informative in predicting the survival outcome, as shown by Binder and Schumacher (2008).

2 NCI-60 Dataset

The NCI-60 cell line data set was generated from a drug discovery project at the National Cancer Institute (NCI). The 60 cell lines from 9 different tissues of origin were exposed to thousands of compounds. Growth inhibitory effects of each compound were measured for each cell line and reported as GI50, the concentration that inhibits growth by 50%. The data set was originally analyzed by Staunton et al (2001) to predict a dichotomized chemosensitivity. Compounds that had a relatively broad and balanced range of effects across the 60 cell lines had been used for analysis. Here, the response variable used was normalized $\log_{10}(GI50)$ values across all cell lines for each compound and there were a total of 232 compounds. Gene expression data were derived using high density Hu6800 Affymetrix microarrays containing 7129 probe sets. The original data were provided as average difference values between perfect match and mismatch scores. The gene expression data used here was pre-processed and contained only 1517 probe sets (Staunton et al., 2001), for which the minimum change in gene expression across all 60 cell lines was greater than 500 average difference units. Data were logged (base 2) and median centered.

The 1517 probe sets corresponded to 1408 unique genes according to their ENTREZ IDs. For probe sets with the same ENTREZ ID, we took the average of their measurements as the expression level for that gene. Mapping to the PPI network, we found that 996 genes formed a connected subnetwork with 7310 edges. The average number of direct neighbors was 14.7, ranging from 1 to 120. The response variable was GI50 for one compound with a relatively high predictive accuracy according to Staunton et al (2001). Data were randomly split into a training set and a test set with sample sizes 40 and 20 respectively. We applied all four methods to the training set for model building and to the test set for prediction. The results are shown in Table 3. For SSVS+GMRF, we set $\theta_0 = \text{logit}(0.1)$ and $\alpha_0 = \text{logit}(0.1)$ for SSVS+BMRF and SSVS+sBMRF. However, the model size selected by SSVS+BMRF and SSVS+SBMRF was sensitive to α_0, as pointed out by Li and Zhang (2008).

The methods SSVS+GMRF and SSVS+SBMRF yielded smaller PMSEs than SSVS+IND, while SSVS+BMRF had the largest PMSE. The frequencies of the selected genes by the four methods

Table 3: PMSEs for the NCI-60 data.

	IND	GMRF	BMRF	SBMRF
PMSE	0.77	0.56	1.29	0.62

are shown in Fig 4.4. Again the method SSVS+BMRF selected most genes. Fig 4.5 shows the top 20 most frequently selected genes and their associated edges. None of the genes selected by SSVS+IND were connected to each other, while several genes selected by SSVS+GMRF or SSVS+SBMRF were connected. In contrast, most of the genes selected by SSVS+BMRF were highly connected to each other, suggesting that the method SSVS+BMRF seemed to favor the selection of the genes with large degrees, in consistent with its use of k_i, rather than its scaled version k_i^* as used in SSVS+SBMRF.

We searched the Cancer Genes database (Higgins et al 2006) and identified that, among the top 20 genes selected by the four methods, there were respectively 6, 5, 4 and 6 genes related to the cancer gene pathways or functional groups: genes CXCR4, PRKAR1A, PROS1, RBBP8, SOD1 and YWHAB for SSVS+IND; HMGA2, PRKACG, PTP4A2, RPN1 and TYMS for SSVS+GMRF; FUS, RPS13, RPSA and SNRPD2 for SSVS+BMRF; and HADHB, LASP1, PSMC1, SNRPD2, TPM4 and YWHAB for SSVS+SBMRF.

5 Discussion

In this paper, we have investigated four different models for the prior probabilities of the predictors' being included in a linear regression model in the framework of Stochastic Search Variable Selection (SSVS). Compared to the standard independence prior that treats the predictors as independent a priori, the three Markov random field priors aim to capture spatial correlations among the predictors as suggested by a given predictor network. The same idea can be found in Wei and Li (2007) and Wei and Pan (2008), but in a simpler non-regression context. In the simulation study, we have demonstrated that the proposed MRF priors performed better than the independence prior in terms of both prediction and variable selection, even though there did not appear to exist a unanimous winner. For the real data, although some of the new methods still performed better, the difference was smaller.

Although the MRF priors introduce additional parameters into the SSVS model, the increase of computational demand is negligible as compared to the independence prior. Considering the potential gain in prediction and gene selection, but without significant increase in computing time,

the MRF priors provide a good means to incorporate network structures to improve statistical efficiency. In particular, it is easier to specify some prior parameter to control the final model size with the GMRF prior, while it is more difficult for the BMRF and SBMRF due to the latter two's dependence on several parameters. We also explored putting a zero point mass on non-informative βs and using conjugate priors for β as mentioned in George (1997), $\beta_\gamma \sim N(0, c\sigma^2(X'_\gamma X_\gamma)^{-1})$. This set up requires $(X'_\gamma X_\gamma)$ to be positive definite, thus can only choose a number of genes no more than the sample size, which may be a shortcoming for the high-dimensional and low-sample-sized setting. Our simulation results (not shown) indicated that it had similar performance in identifying informative genes to the methods presented here, but worse in predictive performance.

Here we have introduced some MRF priors to smooth the prior probabilities of the predictors' being selected over a given predictor network. For the same purpose of incorporating network information into linear regression, Li and Li (2008) and Pan et al (2009) derived network-constrained penalties to induce smoothness in (weighted) regression coefficients β_i's or $|\beta_i|$'s in the framework of penalized regression, in which the smoothness assumption is much stronger than assumed here. Nevertheless, our methods can be extended to smooth β_i's directly, e.g. by imposing a GMRF prior on β_i's. Alternatively, a penalized approach to smoothing the prior probabilities as done here might be more efficient computationally than the MCMC implementations proposed here. In addition, although we have only applied the proposed methods to linear regression, it is conceptually straightforward to extend them to classification (Mallick et al 2005) and nonlinear regression with generalized linear models (Monni and Li 2009) or the Cox proportional hazards model (Gui and Li 2005). More studies are needed.

Acknowledgement

WP thanks helpful discussions with Hongzhe Li and Peng Wei.

References

[1] K. Bae and B. K. Mallick. Gene selection using a two-level hierarchical Bayesian model. Bioinformatics 20 (2004), 3423-3430.

[2] J. Besag. Nearest-neighbor systems and the auto-logistic model for binary data. JRSS-B 34 (1972), 75-83.

[3] J. Besag. Spatial interaction and the statistical analysis of lattice systems. JRSS-B 36 (1974), 192-236.

[4] J. Besag and C. Kooperberg. On conditional and intrinsic autoregressions. Biometrika 82 (1995),

733-746.

[5] H. Binder and M. Schumacher. Comment on 'Network-constrained regularization and variable selection for analysis of genomic data'. Bioinformatics 24 (2008), 2566-2568.

[6] H. Y. Chuang *et al.* Network-based classification of breast cancer metastasis. Molecular System Biology 3 (2007), 140-149.

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

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

[9] W. R. Gilks and P. Wild. Adaptive rejection sampling for gibbs sampling. Applied Statiststics 41 (1992), 337-348.

[10] J. Gui and H. Li. Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. Bioinformatics 21 (2005), 3001-3008.

[11] J. Heikkinen and H. Harri*ö*gmande. Fully bayesian approach to image restoration with an application in biogeography. Applied Statistics 43 (1994), 569-582.

[12] M. E. Higgins, M. Claremont, J. E. Major, C. Sander and A. E. Lash. CancerGenes: a gene selection resource for cancer genome projects. Nucleic Acids Research 35 (2006), D721-D726.

[13] J. A. Hoeting, M. Leecaster and D. Bowden. An Improved Model for Spatially Correlated Binary Responses. Journal of Agricultural, Biological and Environmental Statistics 5 (2000), 102-114.

[14] K. E. Lee, N. Sha, E. R. Dougherty, M. Vannucci and B. K. Mallick. Gene selection: a Bayesian variable selection approach. Bioinformatics 19 (2003), 90-97.

[15] C. Li and H. Li. Network-constrained regularization and variable selection for analysis of genomic data. Bioinformatics 24 (2008), 1175-1182.

[16] F. Li and N. R. Zhang. Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. Manuscript (2008).

[17] B. K. Mallick, D. Ghosh and M. Ghosh. Bayesian classification of tumors by using gene expression data. J. R. Statist. Soc. B 67 (2005), 219-234.

[18] S. Monni and H. Li. Bayesian variable selection for graph-structured covariates with applications in genomics. Manuscript (2009).

[19] W. Pan, B. Xie and X. Shen. Incorporating predictor network in penalized regression with application to microarray data. To appear in Biometrics (2009).

Available at `http:` `http://www.biostat.umn.edu./rrs.php` as Research Report 2009-001, Di-

vision of Biostatistics, University of Minnesota.

[20] D. Smith and M. Smith. Estimation of binary markov random fields using markov chain monte carlo. Journal of Computational and Graphical Statistics 15 (2006), 207-227.

[21] M. Smith and L. Fahrmeir. Spatial bayesian variable selection with application to functional magnetic resonance imaging. Journal of the American Statistical Association, 102 (2007), 417-431.

[22] P. Wei and W. Pan. Incorporating gene networks into statistical tests for genomic data via a spatially correlated mixture model. Bioinformatics 24 (2008), 404-411.

[23] P. Wei and W. Pan. Network-based genomic discovery: application and comparison of Markov random field models. To appear in Applied Statistics (2009).
Available at `http: http://www.biostat.umn.edu./rrs.php` as Research Report 2009-009, Division of Biostatistics, University of Minnesota.

[24] Z. Wei and H. Li. A markov random field model for network-based analysis of genomic data. Bioinformatics 23 (2007), 1537-1544.

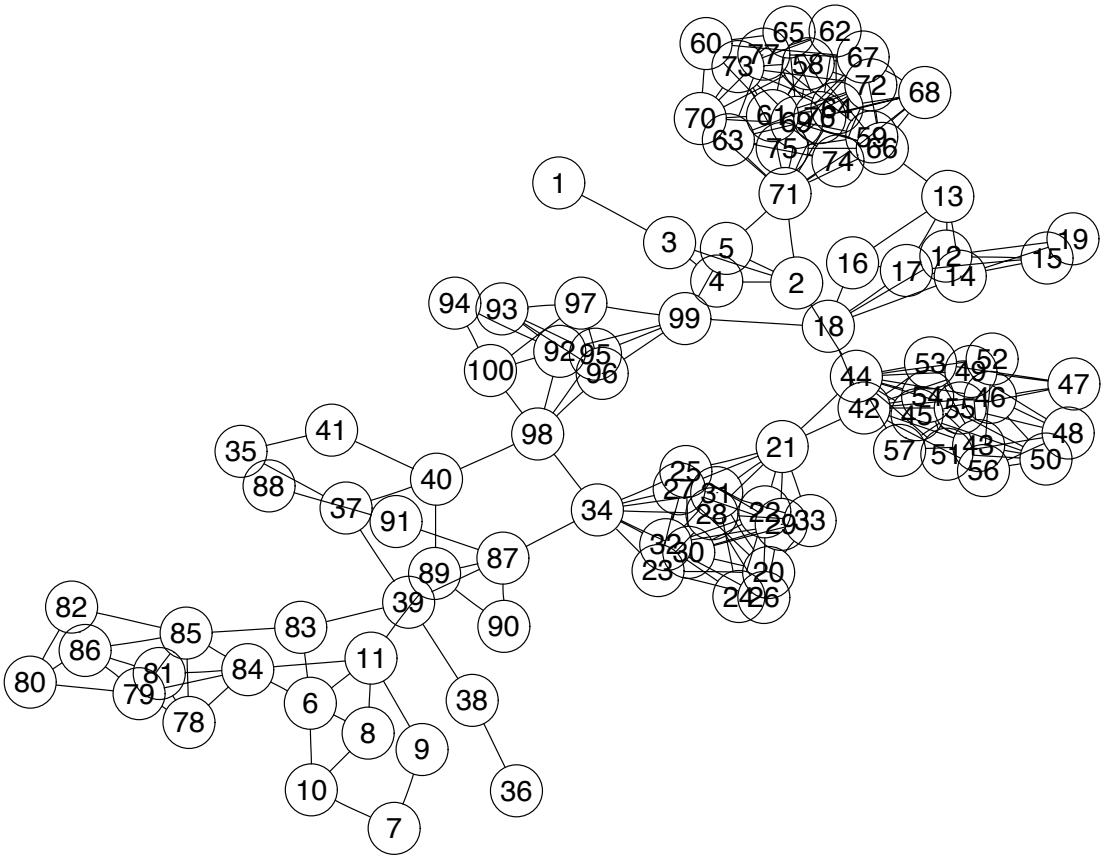

Figure 4.1: A random network used in simulation.

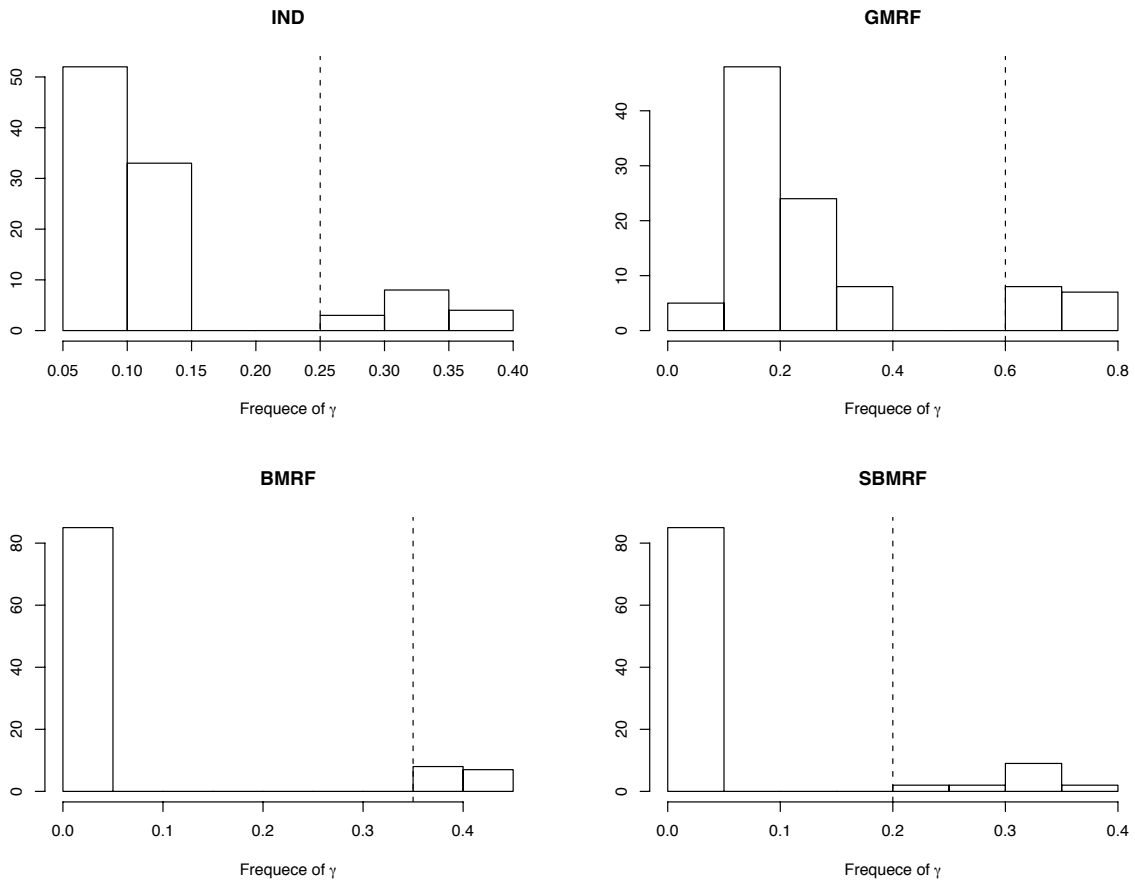

Figure 4.2: Histograms of $\hat{\gamma}_i$ in simulations for the random network.

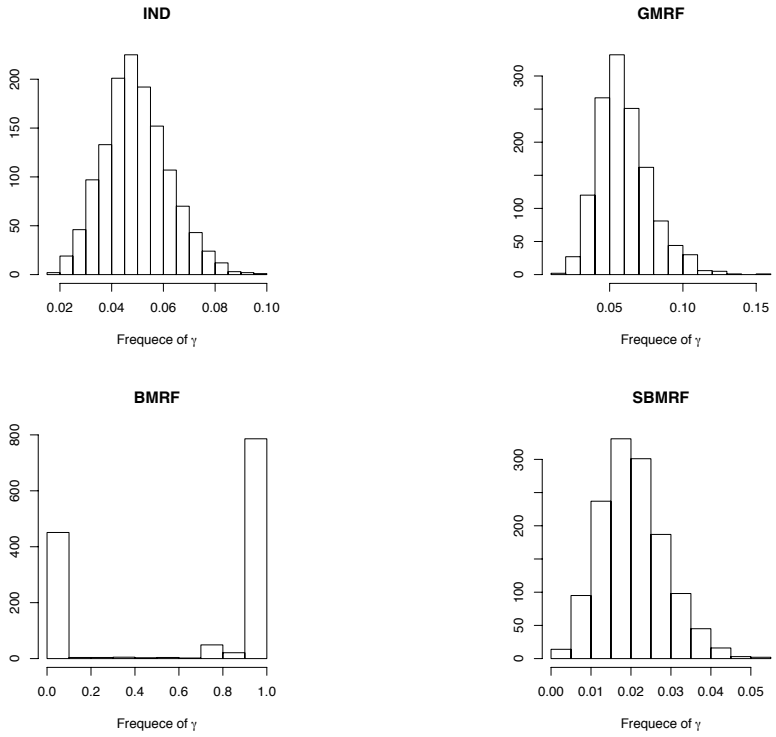

Figure 4.3: Frequencies of the genes being selected for the glioblastoma data.

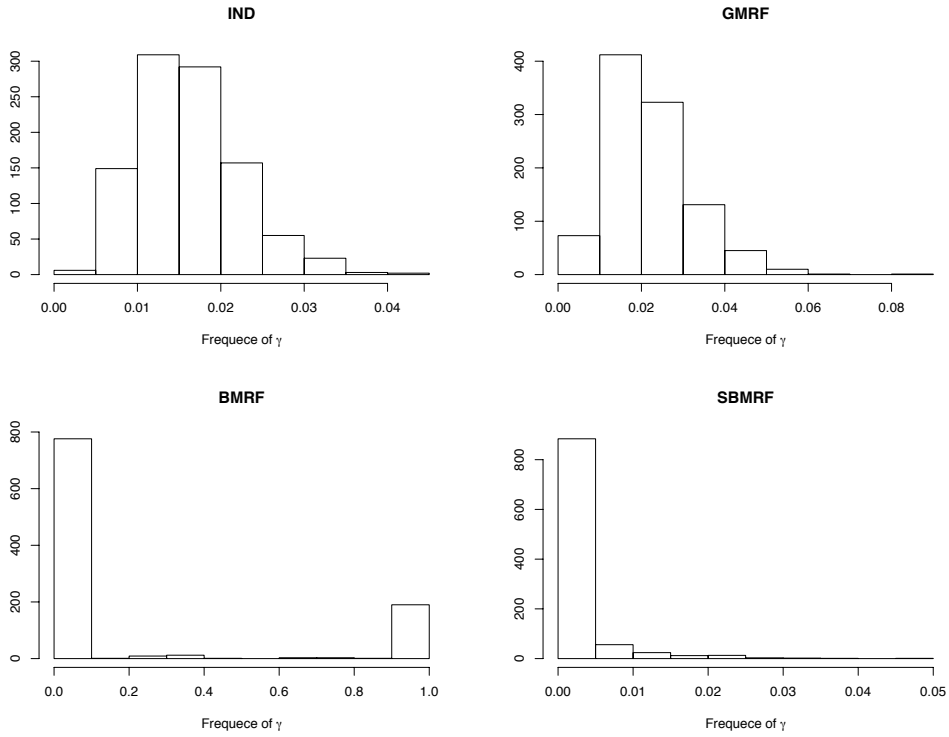

Figure 4.4: Frequencies of the genes being selected for the NCI-60 data.

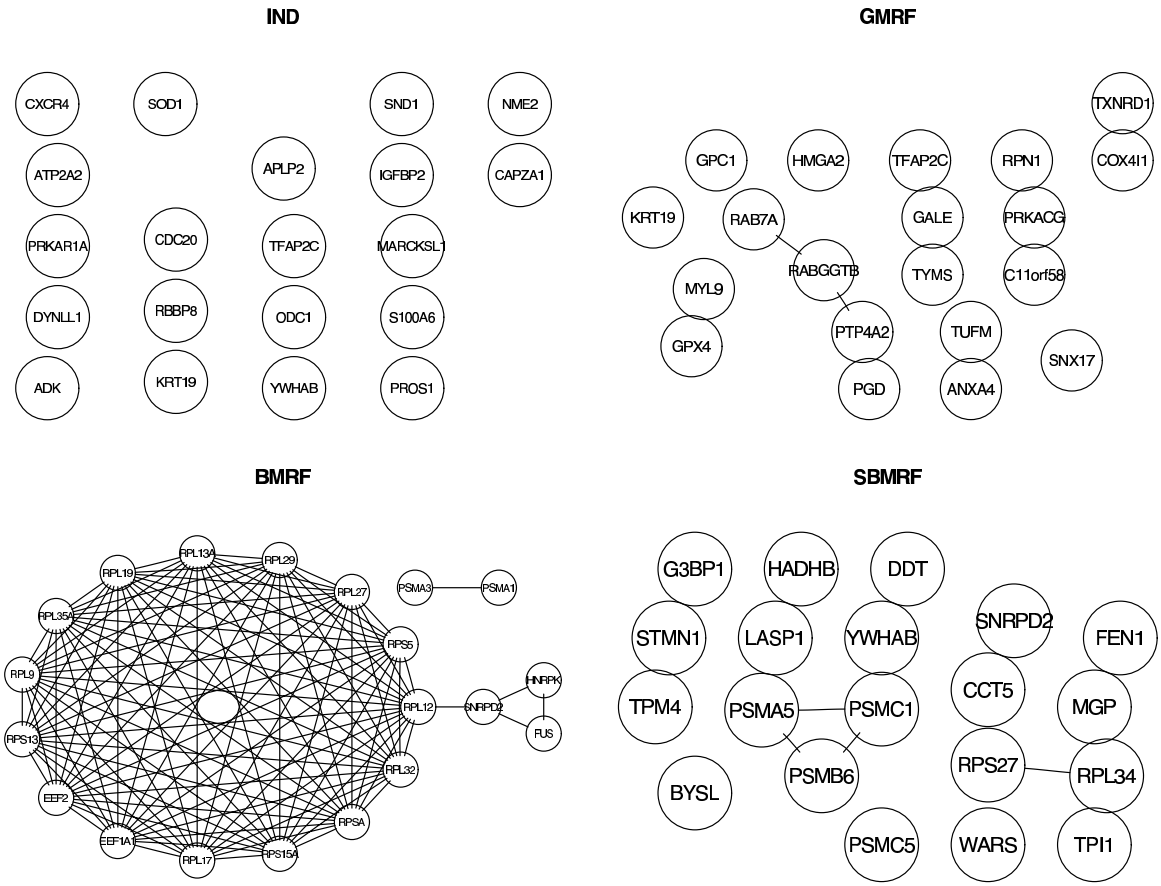

Figure 4.5: Subnetworks of the top 20 selected genes by the methods for the NCI-60 data.