BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 16 2013, pages 1987–1996
doi:10.1093/bioinformatics/btt335

Gene expression

Advance Access publication June 8, 2013

Joint network and node selection for pathway-based genomic data analysis

Shandian Zhe[1], Syed A. Z. Naqvi[1], Yifan Yang[2] and Yuan Qi[1,3,*]

[1]Department of Computer Science, [2]Department of Biology, and [3]Department of Statistics, Purdue University, West Lafayette, IN 47907, USA

Associate Editor: Martin Bishop

ABSTRACT

Motivation: By capturing various biochemical interactions, biological pathways provide insight into underlying biological processes. Given high-dimensional microarray or RNA-sequencing data, a critical challenge is how to integrate them with rich information from pathway databases to jointly select relevant pathways and genes for phenotype prediction or disease prognosis. Addressing this challenge can help us deepen biological understanding of phenotypes and diseases from a systems perspective.

Results: In this article, we propose a novel sparse Bayesian model for joint network and node selection. This model integrates information from networks (e.g. pathways) and nodes (e.g. genes) by a hybrid of conditional and generative components. For the conditional component, we propose a sparse prior based on graph Laplacian matrices, each of which encodes detailed correlation structures between network nodes. For the generative component, we use a spike and slab prior over network nodes. The integration of these two components, coupled with efficient variational inference, enables the selection of networks as well as correlated network nodes in the selected networks.

Simulation results demonstrate improved predictive performance and selection accuracy of our method over alternative methods. Based on three expression datasets for cancer study and the KEGG pathway database, we selected relevant genes and pathways, many of which are supported by biological literature. In addition to pathway analysis, our method is expected to have a wide range of applications in selecting relevant groups of correlated high-dimensional biomarkers.

Availability: The code can be downloaded at www.cs.purdue.edu/homes/szhe/software.html.

Contact: alanqi@purdue.edu

Received on February 9, 2013; revised on June 5, 2013; accepted on June 6, 2013

1 INTRODUCTION

With the popularity of high-throughput biological data such as microarray and RNA-sequencing data, many variable selection methods—such as lasso (Tibshirani, 1996) and elastic net (Zou and Hastie, 2005)—have been proposed and applied to select relevant genes for disease diagnosis or prognosis. Nevertheless, these approaches ignore invaluable biological pathway information accumulated over decades of research; hence, their selection results can be difficult to interpret biologically and their predictive performance can be limited by a small sample size of expression profiles. To overcome these limitations, a promising direction is to integrate expression profiles with rich biological knowledge in pathway databases. Because pathways organize genes into biologically functional groups and model their interactions that capture *correlation* between genes, this information integration can improve not only the predictive performance but also interpretability of the selection results. Thus, a critical need is to integrate pathway information with expression profiles for joint selection of pathways and genes associated with a phenotype or disease.

Despite their success in many applications, previous sparse learning methods are limited by several factors for the integration of pathway information with expression profiles. For example, group lasso (Yuan and Lin, 2007) can be used to utilize memberships of genes in pathways via a $l_{1/2}$ norm to select groups of genes, but they ignore pathway structural information. An excellent work by Li and Li (2008) overcomes this limitation by incorporating pathway structures in a Laplacian matrix of a global graph to guide the selection of relevant genes. In addition to graph Laplacians, binary Markov random field priors can be used to represent pathway information to influence gene selection (Li and Zhang, 2010; Stingo and Vannucci, 2010; Wei and Li, 2007, 2008). These network-regularized approaches do not explicitly select pathways. However, not all pathways are relevant, and pathway selection can yield insight into underlying biological processes. A pioneering approach to joint pathway and gene selection by Stingo *et al.* (2011) uses binary Markov random field priors and couples gene and pathway selection by hard constraints—for example, if a gene is selected, all the pathways it belongs to will be selected. However, this consistency constraint might be too rigid from a biological perspective: an active gene for cancer progression does not necessarily imply that *all* the pathways it belongs to are active. Given the Markov random field priors and the nonlinear constraints, posterior distributions are inferred by a Markov Chain Monte Carlo (MCMC) method (Stingo *et al.*, 2011). But the convergence of MCMC for high-dimensional problems is known to take a long time.

To overcome these limitations, we propose a new sparse Bayesian approach, called Network and NOde Selection (NaNOS), for joint pathway and gene selection. NaNOS is a sparse hybrid Bayesian model that integrates conditional and generative components in a principled Bayesian framework

*To whom correspondence should be addressed.

(Lasserre *et al*., 2006). For the conditional component, we use a graph Laplacian matrix to encode information of each network (e.g. a pathway) and incorporate it into a sparse prior distribution to select individual networks. For the generative component, we use a spike and slab prior distribution to choose relevant nodes (e.g. genes) in selected networks. For this hybrid model, we do not impose the hard consistency constraints used by Stingo *et al*. (2011). Furthermore, the prior distribution of our model does not contain intractable partition functions. This enables us to give a full Bayesian treatment over model parameters and develop an efficient variational inference algorithm to obtain approximate posterior distributions for Bayesian estimation. As described in Section 3, our inference algorithm is designed to handle both continuous and discrete outcomes.

Simulation results in Section 4 demonstrate superior performance of our method over alternative methods for predicting continuous or binary responses, as well as comparable or improved performance for selecting relevant genes and pathways. Furthermore, on real expression data for diffuse large B cell lymphoma (DLBCL), pancreatic ductal adenocarcinoma (PDAC) and colorectal cancer (CRC), our results yield meaningful biological interpretations supported by biological literature.

2 MODEL

In this section, we present the hybrid Bayesian model, NaNOS, for network and node selection. First, let us start from the classical variable selection problem. Suppose we have N independent and identically distributed samples $\mathcal{D} = \{(\mathbf{x}_1, t_1), \ldots, (\mathbf{x}_N, t_N)\}$, where $\mathbf{x}_i$ and t_i are the explanatory variables and the response of the i-th sample, respectively. The explanatory variables can be various biomarkers, such as gene expression levels or single-nucleotide polymorphisms. Following the tradition in variable selection, we normalize the values of each variable so that its mean and standard deviation are 0 and 1, respectively. The response can be certain phenotype or disease status. We aim to predict the response vector $\mathbf{t} = [t_1, \ldots, t_N]^\top$ based on the explanatory variables $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^\mathrm{T}$ and to select a small number of variables relevant for the prediction. Because the number of variables (e.g. genes) is often much bigger than the number of samples, the prediction and selection tasks are statistically challenging.

To reduce the difficulty of variable selection, we can use valuable information from networks, each of which contains certain variables as nodes and represents their interactions. For example, biological pathways cluster genes into functional groups, revealing various gene interactions. Based on M networks, we organize the explanatory variables $\mathbf{x}_i$ into M subvectors, each of which comprises the values of explanatory variables in its corresponding network. If a variable (i.e. a gene) appears in multiple networks (i.e. pathways), we duplicate its value in these networks. Note that networks here are exchangeable with graphs; we can use them to represent not only biological pathways but also linkage disequilibrium structures for genetic variation analysis.

Our model is a Bayesian hybrid of conditional and generative models based on a general framework proposed by (Lasserre *et al*., 2006). The conditional component selects individual networks via 'discriminative' training, the generative component chooses relevant nodes in the selected networks and the two models are glued together through a joint prior distribution, so that the selected networks can guide node selection and, in return, the selected nodes can influence network selection.

Specifically, for the conditional model, we use a Gaussian data likelihood function for the continuous response

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \tau) = \prod_{i=1}^{N} \mathcal{N}(t_i|\mathbf{x}_i^\mathrm{T}\mathbf{w}, \tau^{-1}) \tag{1}$$

where $\mathbf{w}$ are regression weights, each of which represents the contribution of the corresponding node to the response, and τ is the precision parameter. For the unknown variance τ, we assign an uninformative diffuse Gamma prior, $\mathrm{Gam}(\tau|g, h)$ with $g = h = 10^{-6}$.

For the binary response, we use a logistic likelihood

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}) = \prod_{i=1}^{N} \sigma(\mathbf{x}_i^\mathrm{T}\mathbf{w})^{t_i}[1 - \sigma(\mathbf{x}_i^\mathrm{T}\mathbf{w})]^{1-t_i} \tag{2}$$

where $t_i \in \{0, 1\}$, $\mathbf{w}$ are classifier weights and $\sigma(\cdot)$ is the logistic function [i.e. $\sigma(y) = (1 + \exp(-y))^{-1}$]. Based on the M networks, we partition $\mathbf{w}$ into M groups, so that $\mathbf{w} = [\mathbf{w}_1, \ldots, \mathbf{w}_M]^\top$ where $\mathbf{w}_k$ are the weights for the explanatory variables in the k-th network.

To incorporate the topological information of a network, we use its normalized Laplacian matrix representation. Specifically, given an adjacent matrix $\mathbf{G}_k$ that represents the edges (i.e. interactions) between nodes in the k-th network, the normalized Laplacian matrix $\mathbf{L}_k$ is defined as

$$\mathbf{L}_k(i,j) = \begin{cases} 1 & i = j \text{ and } \deg(i) \neq 0 \\ -\frac{1}{\sqrt{\deg(i)\deg(j)}} & i \neq j \text{ and } \mathbf{G}_k(i,j) \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

where $\deg(i) = \sum_j \mathbf{G}_k(i,j)$ is the degree of the i-th node in the k-th network.

Based on the graph Laplacian matrices, we design the following mixture prior over $\mathbf{w}_k$ to select relevant networks:

$$p(\mathbf{w}_k|\alpha_k) = \mathcal{N}(\mathbf{w}_k|\mathbf{0}, s_1\mathbf{L}_k^{-1})^{\alpha_k}\mathcal{N}(\mathbf{w}_k|\mathbf{0}, s_2\mathbf{I}_k)^{1-\alpha_k} \tag{3}$$

where α_k is a binary variable indicating whether the k-th network is selected, $s_1 > s_2$, $s_2 \approx 0$ and $\mathbf{I}_k$ is an identity matrix. We set the hyperparameters s_1 and s_2 based on cross-validation (CV) in our experiments. To make sure $\mathbf{L}_k$ is strictly positive-definite, we add a diagonal matrix $10^{-6}\mathbf{I}_k$ to $\mathbf{L}_k$. In (3), $\mathbf{L}_k$ captures the correlation information between nodes in the k-th network. Note that if we replace $\mathbf{L}_k$ by $\mathbf{I}_k$ in the slab component, the prior (3) becomes a simple generalization of the classical spike and slab prior (George and McCulloch, 1997) for group selection. When $\alpha_k = 1$, the k-th network is selected and the elements of $\mathbf{w}_k$ are encouraged to be similar to each other due to the Laplacian matrix $\mathbf{L}_k$; when $\alpha_k = 0$, because s_2 is close to zero, the corresponding Gaussian prior prunes $\mathbf{w}_k$. We use a Bernoulli prior distribution to reflect the uncertainty in α_k, $p(\alpha_k) = (u_k)^{\alpha_k}(1 - u_k)^{1-\alpha_k}$ where $u_k \in [0, 1]$ is the selection probability. Without any prior preference over selecting or pruning the k-th network, we assign a uniform prior over u_k: $p(u_k) = 1$ [i.e. $p(u_k) = \mathrm{Beta}(u_k; a, b)$ where $a = b = 1$].

To identify relevant nodes, we introduce a latent vector $\widetilde{\mathbf{w}}_k$ in the generative model for each network k, which is tightly linked to $\mathbf{w}_k$ as explained later. We use a spike and slab prior:

$$\begin{aligned} p(\widetilde{\mathbf{w}}_k|\beta_k) &= \prod_{j=1}^{p_k} \mathcal{N}(\tilde{w}_{kj}|0,r_1)^{\beta_{kj}} \mathcal{N}(\tilde{w}_{kj}|0,r_2)^{1-\beta_{kj}} \\ &= \prod_{j=1}^{p_k} \mathcal{N}(0|\tilde{w}_{kj},r_1)^{\beta_{kj}} \mathcal{N}(0|\tilde{w}_{kj},r_2)^{1-\beta_{kj}} \\ &= p(\mathbf{0}|\widetilde{\mathbf{w}}_k,\beta_k) \end{aligned} \tag{4}$$

where p_k is the number of nodes in the k-th network, $r_2 \approx 0$ and β_{kj} is a binary variable indicating whether to select the j-th node in the k-th network. We give β_{kj} a Bernoulli prior, $p(\beta_{kj}) = (v_{kj})^{\beta_{kj}}(1-v_{kj})^{1-\beta_{kj}}$, and a uniform prior over v_{kj}: $p(v_{kj}) = 1$ (i.e. $p(v_{kj}) = \text{Beta}(v_{kj}|c,d)$ where $c = d = 1$). As shown above, the spike and slab prior $p(\widetilde{\mathbf{w}}_k|\beta_k)$ has the same form as $p(\mathbf{0}|\widetilde{\mathbf{w}}_k,\beta_k)$, which can be viewed as a generative model—in other words, the observation $\mathbf{0}$ is sampled from $\widetilde{\mathbf{w}}_k$. This view enables us to combine the sparse conditional model for network selection with the sparse generative model for node selection via a principled hybrid Bayesian model.

Specifically, to link the conditional and generative models together, we introduce a prior on $\widetilde{\mathbf{w}}_k$:

$$p(\widetilde{\mathbf{w}}_k|\mathbf{w}_k) = \mathcal{N}(\widetilde{\mathbf{w}}_k|\mathbf{w}_k,\lambda\mathbf{I}) \tag{5}$$

where the variance λ controls how similar $\widetilde{\mathbf{w}}_k$ and $\mathbf{w}_k$ are in our joint model. For simplicity, we set $\lambda = 0$ so that $p(\widetilde{\mathbf{w}}_k|\mathbf{w}_k) = \delta(\widetilde{\mathbf{w}}_k - \mathbf{w}_k)$ where $\delta(f) = 1$ if $f = 0$ and $\delta(f) = 0$ otherwise. The graphical model representation of the joint model is given in Figure 1.

The network and node selections are consistent with each other in a probabilistic sense. If a network is pruned, all its node are removed. Because $\mathbf{w}_k = \widetilde{\mathbf{w}}_k$ is enforced by the prior $\delta(\widetilde{\mathbf{w}}_k - \mathbf{w}_k)$, when $\alpha_k = 0$, $\mathbf{w}_k = \mathbf{0}$ implies $\widetilde{\mathbf{w}}_k = \mathbf{0}$. As a result, the spike component in (4) will be selected for all the nodes in the k-th network (i.e. $\beta_{kj} = 0$ for $j = 1, \ldots, p_k$) with a higher probability than the slab component. On the other hand, it is easy to see that if one or multiple nodes in a network are selected, then this network will be selected too. Note that if a node appears in multiple networks and is selected, our model will not force *all* the networks that contain this node to be chosen. The reason is that we duplicate the value of this node in the networks and treat their corresponding regression or classification weights as separate model parameters.

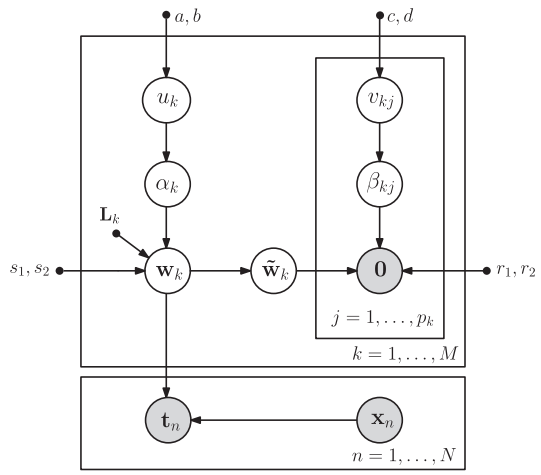

Fig. 1. The graphical model representation of NaNOS

3 ALGORITHM

In this section, we present the variational Bayesian algorithm for model estimation. Specifically, we develop the variational updates to efficiently approximate the posterior distribution of weights $\mathbf{w}$, the network-selection indicators α, the node-selection indicators β, the network- and node-selection probabilities $\mathbf{u}$ and $\mathbf{v}$ and the precision parameter τ for regression. Based on the posteriors of α and β, we can decide which networks and nodes are selected.

For regression, based on the model specification in Section 2, the posterior distribution of our model is

$$\begin{aligned} &p(\mathbf{w},\widetilde{\mathbf{w}},\alpha,\beta,\mathbf{u},\mathbf{v},\tau|\mathbf{t},\mathbf{X}) \\ &= \frac{1}{Z}\mathcal{N}(\mathbf{t}|\mathbf{X}\mathbf{w},\tau^{-1}\mathbf{I})\text{Gamma}(\tau)\cdot \\ &\quad \prod_k p(\mathbf{w}_k|\alpha_k)p(\widetilde{\mathbf{w}}_k|\mathbf{w}_k)p(\mathbf{0}|\widetilde{\mathbf{w}}_k,\beta_k)\ \text{Bern}(\alpha_k|u_k)\ \text{Beta}(u_k)\cdot \\ &\quad \prod_j \text{Bern}(\beta_{kj}|v_{kj})\text{Beta}(v_{kj}) \end{aligned} \tag{6}$$

where $p(\mathbf{w}_k|\alpha_k)$ and $p(\mathbf{0}|\widetilde{\mathbf{w}}_k,\beta_k)$ are defined in (3) and (4), $p(\widetilde{\mathbf{w}}_k|\mathbf{w}_k) = \delta(\widetilde{\mathbf{w}}_k - \mathbf{w}_k)$ and Z is the normalization constant. For classification, the posterior distribution is similar to (6), except that we replace the Gaussian likelihood (1) by the logistic function (2) and remove the precision parameter τ and its prior for regression in (6).

Classical Markov chain Monte Carlo methods can be applied to approximate the posterior distribution. However, given the high dimensionality of the parameters (e.g. $\mathbf{w}$ and α), it would take a long time for a sampler to converge. In practice, it is even difficult to judge the sampler's convergence. Thus, we resort to a computationally efficient variational approximation to (6).

Specifically, we approximate the exact posterior distribution in (6) by a factorized distribution: $Q(\theta) = Q(\mathbf{w})Q(\alpha)Q(\beta)Q(\mathbf{u})Q(\mathbf{v})Q(\tau)$, where θ denotes all the latent variables. Note that, for classification, we do not have $Q_\tau(\tau)$. Because we set $p(\widetilde{\mathbf{w}}|\mathbf{w}) = \delta(\widetilde{\mathbf{w}} - \mathbf{w})$, we do not need a separate distribution $Q(\widetilde{\mathbf{w}})$. To solve $Q(\theta)$, we minimize the Kullback-Leibler (KL) divergence between the exact and approximate posterior distributions of θ:

$$\text{KL}(Q(\theta)||p(\theta|\mathbf{t},\mathbf{X})) = \int Q(\theta)\ln\frac{Q(\theta)}{p(\theta|\mathbf{t},\mathbf{X})}\text{d}\theta \tag{7}$$

Applying coordinate descent for the minimization of (7), we obtain efficient updates for the variational distributions as described in the following sections. The updates are iterative: we update one of the variational distributions at a time while having all the other variational distributions fixed, and iterate these updates until convergence. Because these updates monotonically decrease the value of the KL divergence (7), which is lower bounded by zero, they are guaranteed to converge in terms of the KL value (Bishop, 2006).

3.1 Regression

The variational distributions for regression have the following forms:

$$Q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}, \Sigma) \tag{8}$$

$$Q(\alpha) = \prod_k \gamma_k^{\alpha_k}(1-\gamma_k)^{1-\alpha_k} \tag{9}$$

$$Q(\beta) = \prod_k \prod_j (\eta_{kj})^{\beta_{kj}}(1-\eta_{kj})^{1-\beta_{kj}} \tag{10}$$

$$Q(\mathbf{u}) \propto \prod_k (u_k)^{\tilde{a}_k-1}(1-u_k)^{\tilde{b}_k-1} \tag{11}$$

$$Q(\mathbf{v}) \propto \prod_k \prod_j (v_{kj})^{\tilde{c}_{kj}-1}(1-v_{kj})^{\tilde{d}_{kj}-1} \tag{12}$$

$$Q(\tau) = \Gamma(\tau|\tilde{g}, \tilde{h}) \tag{13}$$

Their parameters are iteratively updated as follows:

$$\Sigma = (\mathbf{A} + \langle\tau\rangle\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} \qquad \mathbf{m} = \langle\tau\rangle\Sigma\mathbf{X}^{\mathrm{T}}\mathbf{t} \tag{14}$$

$$\tilde{a}_k = \gamma_k + a \qquad \tilde{b}_k = 1 - \gamma_k + b \tag{15}$$

$$\tilde{c}_{kj} = \eta_{kj} + c \qquad \tilde{d}_{kj} = 1 - \eta_{kj} + d \tag{16}$$

$$\gamma_k = 1/(1+\exp(\langle\ln(1-u_k)\rangle - \langle\ln u_k\rangle + \frac{p_k}{2}\ln\frac{s_1}{s_2} - \frac{1}{2}\ln|\mathbf{L}_k| + \frac{1}{2}\mathrm{tr}\Big(\langle\mathbf{w}_k\mathbf{w}_k^{\mathrm{T}}\rangle\Big(\frac{1}{s_1}\mathbf{L}_k - \frac{1}{s_2}\mathbf{I}_k\Big)\Big)) \tag{17}$$

$$\eta_{kj} = 1/\Big(1+\exp\Big(\langle\ln(1-v_{kj})\rangle - \langle\ln v_{kj}\rangle + \frac{1}{2}\ln\frac{r_1}{r_2} + \frac{1}{2}\langle(w_{kj})^2\rangle\Big(\frac{1}{r_1} - \frac{1}{r_2}\Big)\Big)\Big) \tag{18}$$

$$\tilde{h} = h + \frac{1}{2}\mathbf{t}^{\mathrm{T}}\mathbf{t} - \mathbf{m}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{t} + \frac{1}{2}\sum\nolimits_i \mathbf{x}_i^{\mathrm{T}}\langle\mathbf{w}\mathbf{w}^{\mathrm{T}}\rangle\mathbf{x}_i \tag{19}$$

$$\tilde{g} = g + \frac{N}{2} \tag{20}$$

where $\mathbf{A} = \frac{1}{s_1}\mathrm{diag}(\{\gamma_k\mathbf{L}_k\}_k) + \frac{1}{s_2}\mathrm{diag}(\{(1-\gamma_k)\mathbf{I}_k\}_k) + \frac{1}{r_1}\mathrm{diag}(\eta) + \frac{1}{r_2}\mathrm{diag}(1-\eta)$ [note that $\mathrm{diag}(\{\gamma_k\mathbf{L}_k\}_k)$ is a block-diagonal matrix], $\langle\cdot\rangle$ means expectation over the corresponding variational distribution, and the required moments in the above equations are

$$\langle\mathbf{w}\mathbf{w}^{\mathrm{T}}\rangle = \Sigma + \mathbf{m}\mathbf{m}^{\mathrm{T}} \qquad \langle\tau\rangle = \tilde{g}/\tilde{h}$$
$$\langle\ln u_k\rangle = \psi(\tilde{a}_k) - \psi(\tilde{e}_k) \quad \langle\ln(1-u_k)\rangle = \psi(\tilde{b}_k) - \psi(\tilde{e}_k)$$
$$\langle\ln v_{kj}\rangle = \psi(\tilde{c}_{kj}) - \psi(\tilde{f}_{kj}) \quad \langle\ln(1-v_{kj})\rangle = \psi(\tilde{d}_{kj}) - \psi(\tilde{f}_{kj})$$

where $\psi(x) = \frac{\mathrm{d}}{\mathrm{dx}}\ln\Gamma(x)$, $\tilde{e}_k = \tilde{a}_k + \tilde{b}_k$ and $\tilde{f}_{kj} = \tilde{c}_{kj} + \tilde{d}_{kj}$.

3.2 Classification

Compared with regression, the classification task is more challenging. Because of the logistic function (2), we cannot directly solve the variational distribution $Q(\mathbf{w})$. Therefore, we use a lower bound proposed by (Jaakkola and Jordan, 2000) to replace the logistic function in the joint distribution:

$$\sigma(y)^t(1-\sigma(y))^{1-t} \geq \sigma(\xi)\exp\Big(\frac{(2t-1)y - \xi}{2} - f(\xi)((2t-1)^2y^2 - \xi^2)\Big) \tag{21}$$

where $f(\mathbf{x}) = \frac{1}{4\xi}\tanh(\xi/2)$, and ξ is a variational parameter. Note that the equality is achieved when $\xi = (2t-1)y$. Because the logarithm of the lower bound (21) is quadratic in y, it essentially converts the logistic function into a Gaussian form so that the variational inference becomes tractable.

Combining the maximization of the lower bound (21) with the minimization of the KL divergence (7), we obtain the variational updates for classification. They are the same as those for the regression task, except for that $Q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}, \Sigma)$, now we have

$$\Sigma = \Big(\mathbf{A} + 2\sum\nolimits_i f(\xi_i)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\Big)^{-1} \quad \mathbf{m} = \frac{1}{2}\Sigma\mathbf{X}^{\mathrm{T}}(2\mathbf{t} - \mathbf{1}) \tag{22}$$

where $\mathbf{A}$ is the same as in the regression.

In addition, maximization of the lower bound of the logistic function gives the update for the variational parameter ξ_i:

$$\xi_i^2 = \mathbf{x}_i^{\mathrm{T}}\langle\mathbf{w}\mathbf{w}^{\mathrm{T}}\rangle\mathbf{x}_i. \tag{23}$$

3.3 Computational cost

The computational cost of the proposed algorithm is dominated by (14) for regression and (22) for classification. For both cases, it takes $O(p^3)$ for matrix inversion to obtain Σ and $O(Np + p^2)$ to obtain $\mathbf{m}$ for each iteration. Thus, the total cost is $O(p^3 + Np)$ and, for most applications where $p > N$, it simplifies to $O(p^3)$.

4 EXPERIMENTS

In this section, we apply NaNOS to synthetic and real gene expression data to select pathways (i.e. networks) and genes (i.e. nodes), and provide biological analysis of our results. We also compare NaNOS with alternative methods, including lasso (Tibshirani, 1996), elastic net (Zou and Hastie, 2005), group lasso (Jacob *et al.*, 2009; Yuan and Lin, 2007), the network-constrained regularization approach [Li and Li (2008), henceforth 'LL'] and the sparse Bayesian model with the classical spike and slab prior (George and McCulloch, 1997). For lasso and elastic net, we used the Glmnet software package (www-stat.stanford.edu/~tibs/glmnet-matlab/). For group lasso, we treat each pathway as a group. To handle genes appearing in multiple pathways (i.e. groups), we first duplicated their expression levels for each group—as suggested by (Jacob *et al.*, 2009)—and then used the SLEP software package (www.public.asu.edu/~jye02/Software/SLEP/) for group lasso estimation. For the spike and slab model, we implemented variational inference similar to our updates in Section 3. Just as NaNOS, all these software packages use the Gaussian likelihood for regression and the logistic likelihood for classification. We used the default configuration of these software packages for the maximum number of iterations, initial values and the threshold for convergence. To tune regularization weights in lasso, group lasso and the LL approach, we conducted thorough 10-fold CV on training data (i.e. not using the test data) using a large computer cluster. The CV grids on the

free parameters are summarized here: for lasso, $\alpha = [0 : 0.01 : 1]$; for elastic net, $\alpha = [0 : 0.01 : 1]$ and $\beta = [0 : 0.01 : 1]$; for group lasso (both regression and logistic regression), $\alpha = [0 : 0.01 : 1]$; and for the LL approach, $\lambda_1 = [1 : 25 : 300]$ and $\lambda_2 = [1 : 25 : 300]$ (we also did a second-level CV after we pruned the range of λ_1 and λ_2 values based on the first-level CV). Finally, for NaNOS, the CV grids are $s_1 = r_1 = [0.1, 1, 3]$ and $s_2 = r_2 = [10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}]$.

On the synthetic data for which we knew the true relevant pathways, we also compared NaNOS with a popular tool for gene set enrichment analysis (GSEA) (Mootha *et al*., 2003; Subramanian *et al*., 2005). We treated each pathway as a set, used GSEA's default configuration and applied its suggested criterion false discovery rate (FDR) $<25\%$ to discover enriched pathways. We then identified all the genes in these enriched pathways as target genes. Because GSEA cannot provide predictions on responses **t**, we did not include it for comparison on the real data.

4.1 Simulation studies

We first compare all the methods on synthetic data in the following three experiments.

Experiment 1. We followed the first and second data generation models used by Li and Li (2008). Specifically, we simulated expression levels of 200 transcription factors (TFs), each controlling 10 genes in a simple tree-structured regulatory network, and assumed that four pathways—including *all* of their genes—have effect on the response **t**. We sampled the expression levels of each TF from a standard normal distribution, $x_{TF} \sim \mathcal{N}(0, 1)$ and the expression level of each gene that this TF regulates from $\mathcal{N}(0.7x_{TF}, 0.51)$. This implies a correlation of 0.7 between the TF and its target genes.

For the first model with the continuous response, we designed a weight vector for each pathway, $\rho = [1, \frac{1}{\sqrt{10}}, \ldots, \frac{1}{\sqrt{10}}]$, corresponding to the TF and 10 genes it regulates, and then sampled **t** as follows:

$$\mathbf{w} = [5\rho, -5\rho, 3\rho, -3\rho, \mathbf{0}^\top]^\top$$
$$\mathbf{t} = \mathbf{X}\mathbf{w} + \epsilon$$

where $\epsilon \sim \mathcal{N}(0, \sigma_e^2)$ and **0** is a vector of all zeros.

The second model is the same as the first one, except that the genes regulated by the same TF can have either positive or negative effect on the response **t**. Specifically, we set

$$\rho = \left[1, \frac{-1}{\sqrt{10}}, \frac{-1}{\sqrt{10}}, \frac{-1}{\sqrt{10}}, \underbrace{\frac{1}{\sqrt{10}}, \ldots, \frac{1}{\sqrt{10}}}_{7}\right].$$

For the first and second models, the noise variance was set to be $\sigma_e^2 = (\Sigma_j w_j^2)/4$ so that the signal-to-noise ratio was 12.85 and 7.54, respectively.

For the binary response, we followed the same procedure as for the continuous response to generate expression profiles **X** and the parameters **w**. Then we sampled **t** from (2).

For each of the settings, we simulated 100 samples for training and 100 samples for test. We repeated the simulation 50 times. To evaluate the predictive performance, we calculated the prediction mean-squared error for regression and the error rate for classification. To examine the accuracy of gene and pathway selection, we also computed sensitivity and specificity and summarized them in the F_1 score, $F_1 = 2>$ (sensitivity $\times$ specificity)/(sensitivity + specificity). The bigger the F_1 score, the higher the selection accuracy.

All the results are summarized in Figure 2, in which the error bars represent the standard errors. For all the settings, NaNOS gives smaller errors and higher F_1 scores for gene selection than the other methods, except that, for classification of the samples from the second data model, NaNOS and group lasso obtain the comparable F_1 scores. All the improvements are significant under the two-sample t-test ($P<0.05$). We also show the accuracy of group lasso, GSEA and NaNOS for pathway selection in Figure 5. Again, NaNOS achieves significantly higher selection accuracy. Because the LL approach was developed for regression, we did not have its classification results. While the LL approach uses the topological information of all the pathways, they are merged together into a *global* network for regularization. In contrast, using a sparse prior over individual pathways, NaNOS can explicitly select pathways relevant to the response, guiding the gene selection. This may contribute to its improved performance.

Experiment 2. For the second experiment, we did not require all genes in relevant pathways to have effect on the response. Specifically, we simulated expression levels of 100 TFs, each regulating 21 genes in a simple regulatory network. We sampled the expression levels of the TFs, the regulated genes and their response in the same way as in Experiment 1, except that we set

$$\rho = \left[1, \underbrace{\frac{1}{\sqrt{21}}, \ldots, \frac{1}{\sqrt{21}}}_{10}, \underbrace{0, \ldots, 0}_{11}\right]$$

for the first data generation model and

$$\rho = \left[1, \frac{-1}{\sqrt{21}}, \frac{-1}{\sqrt{21}}, \frac{-1}{\sqrt{21}}, \underbrace{\frac{1}{\sqrt{21}}, \ldots, \frac{1}{\sqrt{21}}}_{7}, \underbrace{0, \ldots, 0}_{11}\right] \quad (24)$$

for the second data generation model. Note that the last 11 zero elements in ρ indicate that the corresponding genes have no effect on the response **t**, even in the four relevant pathways.

The results for both the continuous and binary responses are summarized in Figures 3 and 5. For regression based on the first data model, NaNOS and LL obtain the comparable F_1 scores; for all the other cases, NaNOS significantly outperforms the alternative methods in terms of both prediction and selection accuracy ($P<0.05$).

Experiment 3. Finally, we simulated the data as in Experiment 2, except that we replaced $\sqrt{21}$ in the denominators in (24) with 21, to obtain a weaker regulatory effect of the TF. Again, as shown in Figures 4 and 5, NaNOS outperforms the competing methods significantly.

4.2 Application to expression data

Now we demonstrate the proposed method by analyzing gene expression datasets for the cancer studies of DLBCL (Rosenwald *et al*., 2002), CRC (Ancona *et al*., 2006) and

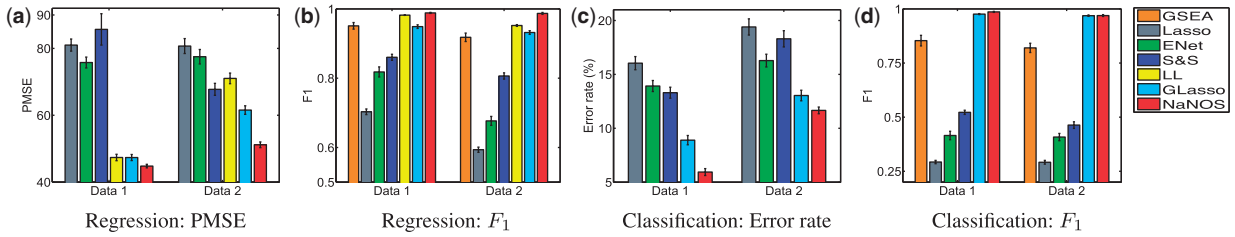

Fig. 2. Prediction errors and F_1 scores for gene selection in Experiment 1. ENet, S&S and GLasso stand for elastic net, the spike and slab model and group lasso, respectively; Data 1 and 2 indicate the first and second data generation models

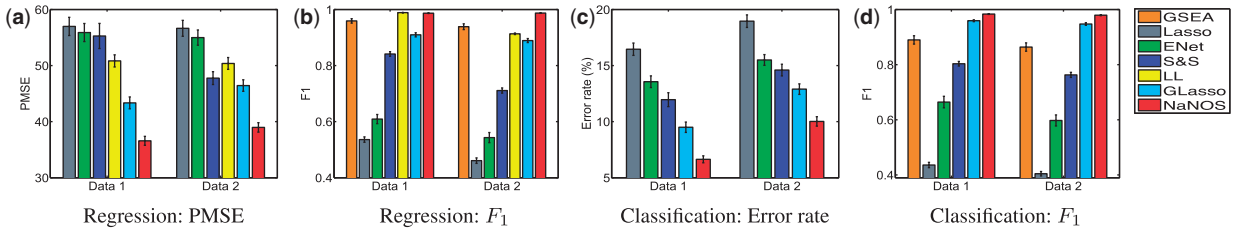

Fig. 3. Prediction errors and F_1 scores for gene selection in Experiment 2

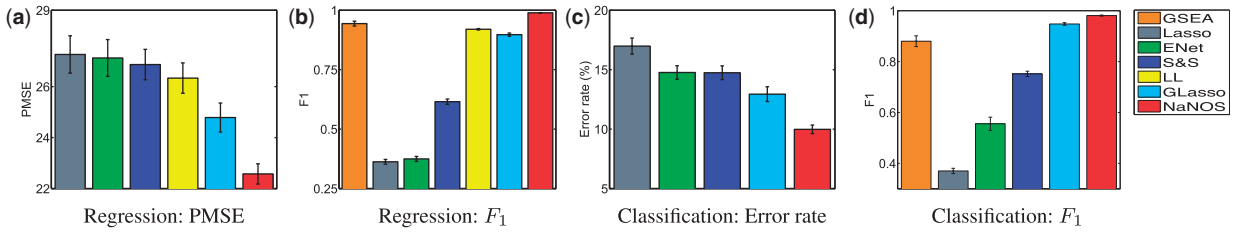

Fig. 4. Prediction errors and F_1 scores for gene selection in Experiment 3

PDAC (Badea *et al.*, 2008). We used the probeset-to-gene mapping provided in these studies. For the CRC and PDAC datasets in which multiple probes were mapped to the same genes, we took the average expression level of these probes. We used the pathway information from the KEGG pathway database (www.genome.jp/kegg/pathway.html) by mapping genes from the cancer studies into the database, particularly in the categories of Environmental Information Processing, Cellular Processes and Organismal Systems.

4.2.1 Diffuse large B cell lymphoma We used gene expression profiles of 240 DLBCL patients from an uncensored study in the Lymphoma and Leukemia Molecular Profiling Project (Rosenwald *et al.*, 2002). From 7399 probes, we found 752 genes and 46 pathways in the KEGG dataset. The median survival time of the patients is 2.8 years after diagnosis and chemotherapy. We used the logarithm of survival times of patients as the response variable in our analysis.

We randomly split the dataset into 120 training and 120 test samples 100 times and ran all the competing methods on each partition. The test performance is visualized in Figure 6a. NaNOS significantly outperforms lasso, elastic net and group lasso. Although the results of the LL approach can contain connected subnetworks, these subnetworks do not necessarily correspond to (part of) a biological pathway. For instance, they may

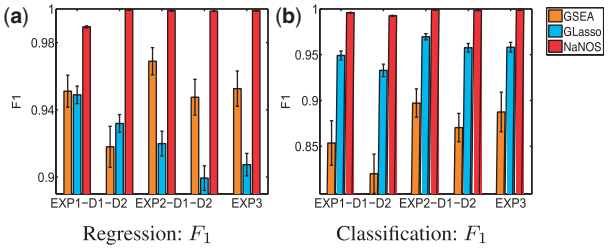

Fig. 5. F_1 scores for pathway selection. 'EXP' stands for 'Experiment' and 'D' stands for 'Data model'

consist of components from multiple overlapped pathways. In contrast, NaNOS explicitly selects relevant pathways. Four pathways had the selection posterior probabilities larger than 0.95 and they were consistently chosen in all the 100 splits. Two of these pathways are discussed below.

First, NaNOS selected the antigen processing and presentation pathway. The part of this pathway containing selected genes is visualized in Figure 7a. A selected regulator CIITA was shown to regulate two classes of antigens MHC I and II in DLBCL (Cycon *et al.*, 2009). The loss of MHC II on lymphoma cells—including the selected HLA-DMB, -DQB1, -DMA, -DRA, -DRB1, -DPA1, -DPB1 and -DQA1—was shown to be related to poor prognosis and reduced survival in DLBCL patients (Rosenwald *et al.*, 2002).

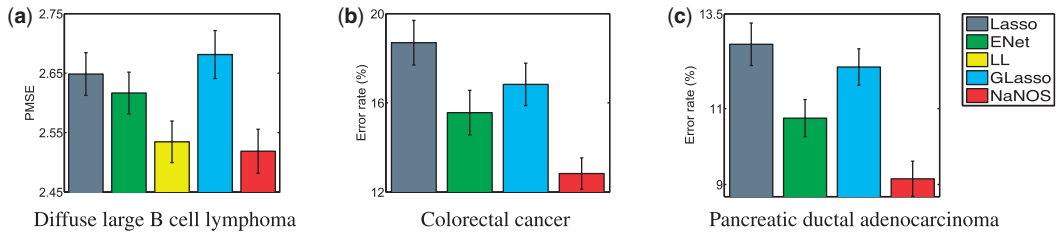

Fig. 6. Predictive performance on three gene expression studies of cancer

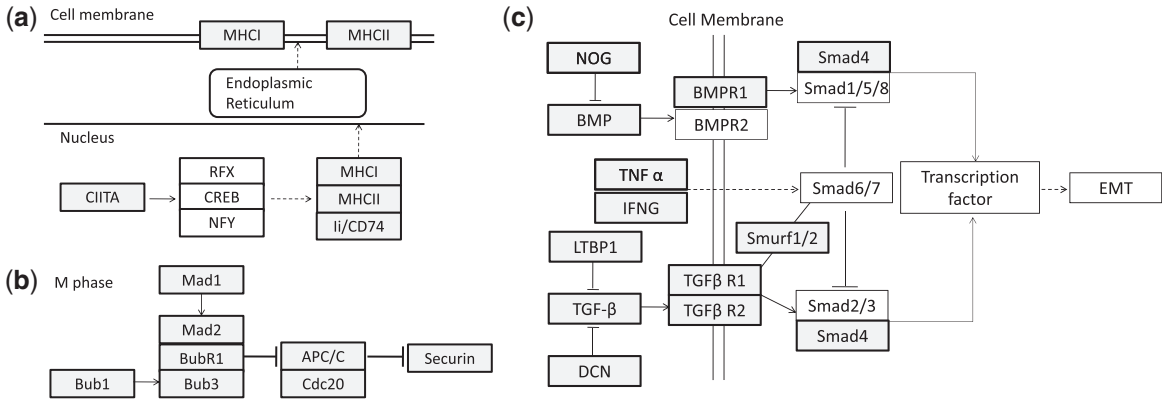

Fig. 7. Examples of part of identified pathways. (**a**) The antigen processing and presentation pathway for DLBCL; (**b**) the cell cycle pathway for CRC; (**c**) the TGF-β signaling pathway for PDAC. Shaded and unshaded boxes indicate selected and not selected genes, respectively

The selected MHC I (e.g. HLA-A,-B,-C,-G) was reported to be absent from the cell surface, allowing the escape from immunosurveillance of lymphoma (Amiot *et al.*, 1998). And the selected Ii/CD74 and HLA-DRB were proposed to be monoclonal antibody targets for DLBCL drug design (Dupire and Coiffier, 2010).

Second, NaNOS chose cell adhesion molecules (CAMs). Adhesive interactions between lymphocytes and the extracellular matrix (ECM) are essential for lymphocytes' migration and homing. For example, the selected CD99 is known to be overexpressed in DLBCL and correlated with survival times (Lee *et al.*, 2011), and LFA-1 (ITGB2/ITGAL) can bind to ICAM on the cell surface and further lead to the invasion of lymphoma cells into hepatocytes (Terol *et al.*, 1999).

4.2.2 Colorectal cancer We applied our model to a CRC dataset (Ancona *et al.*, 2006). It contains gene expression profiles from 22 normal and 25 tumor tissues. We mapped 2455 genes from 22 283 probes into 67 KEGG pathways. The goal was to predict whether a tissue has the CRC or not and select relevant pathways and genes.

We randomly split the dataset into 23 training and 24 test samples 50 times and ran all the methods on each partition. The test performance is visualized in Figure 6b. Again, based on a two-sample *t*-test, NaNOS outperforms the alternatives significantly ($P<0.05$). Three out of the four pathways with the selection posterior probabilities larger than 0.95 are discussed below. They were selected 20, 50 and 50 times in the 50 splits.

First, NaNOS selected the cell cycle pathway. This selection is consistent with the original result by Ancona *et al.* (2006). As shown in Figure 7b, NaNOS selected mitotic spindle assembly related genes. Specifically, Bub1 and Mad1 may regulate the checkpoint complex (MCC) containing Mad2, BubR1 and Bub3. The upregulated MCC may in turn inhibit ability of APC/C to ubiquitinate securin and further lead to mitotic event extension in CRC (Menssen *et al.*, 2007). NaNOS also chose cyclin/CDK complexes, among which CycD/CDK4 overexpression is found in mouse colon tumor and CDK1, CDK2, CycE are increased in human CRC (Vermeulen *et al.*, 2003; Wang *et al.*, 1998). NaNOS further identified the minichromosome maintenance (MCM) complex—including MCM2 and MCM5—which are biomarkers for the CRC stage identification (Giaginis *et al.*, 2009). Moreover, the selected TP53 and c-Myc are known to be closely related to CRC (Menssen *et al.*, 2007).

Second, NaNOS chose the intestinal immune network for IgA production. A greatly increased level of IgA—as a result of long-term intestinal inflammation—can increase the chance of CRC (Rizzo *et al.*, 2011) and serve as an effective biomarker for early diagnosis of CRC (Chalkias *et al.*, 2011). Also, selected chemkines in this pathway, such as CXCR4 and CXCL12, may contribute to CRC progression (Sakai *et al.*, 2012).

Third, NaNOS selected the cytokine–cytokine receptor interaction pathway as well as several well-known CRC-related molecules in this pathway. For instance, CXCL13 is a biomarker for stage II CRC prognosis (Agesen *et al.*, 2012), CXCL10 dramatically increases with CRC progression (Toiyama *et al.*, 2012) and IL10 secreted by CRC cells can accelerate tumor proliferation and be used for the prognosis of CRC progression (Toiyama *et al.*, 2010).

4.2.3 Pancreatic ductal adenocarcinoma This cancer dataset includes expression profiles from 39 PDAC and 39 normal subjects (Badea *et al.*, 2008). By mapping 2781 genes from 54 677 probes

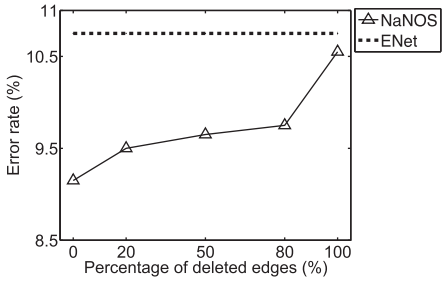

Fig. 8. The predictive performance of NaNOS when the pathway structures are inaccurate. When more edges are randomly selected and removed from each pathway, the performance of NaNOS degrades smoothly, but still better than the competing methods

into KEGG pathways, we obtained 67 pathways. Our goal was to predict whether a subject has the pancreatic cancer and select relevant pathways and genes. We randomly split the dataset into 39 training and 39 test samples 50 times and ran all the methods on each partition. The test performance is visualized in Figure 6c. Based on a two-sample *t*-test, NaNOS significantly outperforms lasso, elastic net and group lasso.

To investigate the sensitivity of NaNOS to the structural noise in the pathway database, we randomly chose 20, 50, 80 and 100% edges in each pathway and removed them. We tested NaNOS for each case and reported the average test error rate in the new Figure 8. As expected, the error rate of NaNOS gradually increases with more edges being removed because less topological information in pathways is available. But NaNOS still consistently outperformed all the alternative methods such as elastic net, the second best method on this dataset. This experiment demonstrates (i) that by exploiting subtle correlation information embedded in the pathway topology, NaNOS can boost its modeling power and predictive performance, and (ii) that NaNOS is robust to small perturbation in pathway topology.

We also examined the impact of the important prior distributions on pathway and gene selection probabilities u_k and v_{kj}. As described in Section 2, we used the uniform priors [i.e. the Beta(1,1) prior] over u_k and v_{kj}, indicating no prior preference over selecting a pathway or gene or not. The average test error based on the uninformative priors is 9.15 ± 0.5, as visualized in Figure 6c. If we change the prior to an informative one, Beta(1,10) (mean 0.09 and standard deviation 0.083) that strongly prefers sparsity, then the average test error increases slightly to 10.0 ± 0.4. This *minor* increase in error may stem from the oversparification caused by the sparsity prior that are overconfident (suggested by a small variance). Now if we use another informative prior Beta(10,1) (mean 0.91 and standard deviation 0.083) that strongly prefers dense—instead of sparse—estimation, then the average test error increases to 11.2 ± 0.5. This relatively larger error increase is exactly what we expected because now the *wrong* dense prior aims to select most pathways and genes. What is important is that, no matter which of these two informative priors we chose, NaNOS consistently outperformed lasso and group lasso in Figure 6c. Between these two extreme cases, if we use an uninformative or weak sparse prior [e.g. Beta(0.5,0.5)], we find that similar prediction error rates were obtained for NaNOS as in Figure 6c. The above analysis indicates that NaNOS is robust to the prior choice.

In addition to using the even splitting strategy with the same number of training and test samples, we also tested the performance of all the algorithms in another setting with more training samples—specifically, 62 training and 16 test samples. We repeated the random partitioning 50 times. The average error rates for NaNOS, elastic net, lasso and group lasso are 8.00 ± 0.89, 9.90 ± 1.00, 12.0 ± 1.0 and 11.0 ± 0.14, respectively. Again, the two-sample *t*-test indicates that NaNOS outperforms the alternative methods significantly ($P < 0.05$).

Three out of the five pathways with the selection posterior probabilities larger than 0.95 are discussed below. They were selected 35, 50 and 50 times in the 50 splits.

The first selected pathway was the transforming growth factor beta (TGF-β) signaling pathway. It is essential in epithelial-mesenchymal transition (EMT)—a critical component for developmental and cancer processes—and related to PDAC (Krantz *et al*., 2012). The selected part of this pathway is visualized in Figure 7c. It shows that IFNG, TNF-α, LTBP1, DCN, TGF-β and its receptor TGF-β R1 were selected. The TGF-β ligand—via its receptor—propagates the signal through phosphorylation of Smads including the selected Smad 4, which in turn translocate into the nucleus and interact with Snail TFs to regulate EMT (Krantz *et al*., 2012). The selected BMP ligand (i.e. BMP2) is bound to BMP R1 and R2 receptors to activate Smad1, which is in a protein complex including Smad4. Gordon *et al*. (2009) showed that in PANC-1 cell line, this protein complex mediates EMT partially by increasing the activity of MMP-2.

The second identified pathway was ECM–receptor interaction. It is associated with desmoplastic reaction, a hallmark in PDAC (Shields *et al*., 2012). In this pathway, NaNOS selected the integrin receptors—including ITGB1, ITGA2, ITGA3, ITGA5, ITGA6—and the ECM proteins—collagens including COL1A1 and COL1A2, and laminins including LAMC2 and LAMB3. Important interactions among them were revealed in a previous study by Weinel *et al*. (1992).

The third chosen pathway was CAMs. CAMs are pivotal in pancreatic cancer invasion by mediating cell–cell signal transduction and cell–matrix communication (Keleg *et al*., 2003). In this pathway, the selected molecules include calcium-dependent cadherin family molecules (CDH2, CDH3) and neural-related molecules (MAG); both of them have shown to be related to PDAC (Kameda *et al*., 1999; Keleg *et al*., 2003).

5 DISCUSSION

As shown in the previous section, the new Bayesian approach, NaNOS, outperformed the alternative sparse learning methods on both simulation and real data by a large margin. Now we discuss three factors that may contribute to the improved performance of NaNOS.

First, the spike and slab prior (3) and its generalization (4) in NaNOS separate weight regularization from the selection of variables (pathways or genes). Both the (generalized) spike and slab prior and elastic net can be viewed as mixture models in which one component encourages the selection of variables and the other helps remove irrelevant ones. However, unlike the elastic net where the weights over l_1 and l_2 penalty functions are fixed, the spike and slab prior has the selection indicators over these two components estimated from data. When a variable is

selected, the model has a Gaussian prior over its value (i.e. weight) that is equivalent to a l_2 regularizer (as in ridge regression) and does not shrink the value of the selected variable as l_1 penalty would do. By contrast, lasso or elastic net, with a fixed mixture weight, has sparsity penalty over both pruned and selected variables, which can greatly shrink the values of selected variables and hurt predictive performance.

Second, NaNOS incorporates correlation structures encoded in pathways for variable selection. Specifically, it uses pathway structures into the extended spike and slab prior distribution to explicitly model the detailed relationships between correlated genes. In contrast, lasso and elastic net do not use this valuable correlation information in their models. By comparing prediction accuracies of NaNOS when 0 and 100% edges are removed from pathways (Fig. 8), we can see that the detailed correlation information captured by the pathway topology can greatly improve modeling quality.

Third, NaNOS has the capability of selecting both relevant pathways and genes due to its two-layer sparse structure. By contrast, with l_1/l_2 penalty, group lasso encourages the selection of all the genes in chosen pathways, leading to *dense* estimation. This may be undesirable in practice and deteriorate the predictive performance of group lasso. NaNOS enhances the *flexibility* of group lasso by conducting sparse estimation at both the pathway (or group) and gene levels. Meanwhile, our Bayesian estimation effectively avoids overfitting, a problem often plaguing flexible models.

NaNOS has been applied to joint pathway and gene selection in this article. Inspired by the seminal works in (Chuang *et al*., 2007; Fröhlich *et al*., 2006; Srivastava *et al*., 2008; Zycinski *et al*., 2013), we can use NaNOS in a variety of biomedical applications where there are abundant high-dimensional biomarkers of individual samples and other information sources—for example, the gene ontology (GO) and protein–protein interaction networks information—that capture correlation in the high-dimensional space. Here we discuss two approaches to apply NaNOS when we have only GO or other group information without network topology. The first approach is to compute some distance or similarity scores between genes based on the GO information [e.g. following the approach by Srivastava *et al*. (2008)] and then estimate the network topology based on a network learning method, for example, graphical lasso (Friedman *et al*., 2008). With the estimated network topology, we can compute the graph Laplacian matrices and apply NaNOS to select genes and groups of genes. The second approach is to directly use the group membership information in NaNOS by replacing the graph Laplacian matrices with identity matrices. This approach becomes useful when we even do not have any information available to learn the network topology. As shown in Figure 8, even when all the edges were removed and we had only group information, NaNOS still outperformed the second best method, elastic net, in terms of prediction accuracy.

Funding: This work was supported by NSF IIS-0916443, NSF CAREER Award IIS-1054903, and the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-0939370.

Conflict of Interest: none declared.

REFERENCES

Agesen,T. *et al.* (2012) ColoGuideEx: a robust gene classifier specific for stage II colorectal cancer prognosis. *Gut*., **61**, 1560–1567.

Amiot,L. *et al.* (1998) Loss of HLA molecules in B lymphomas is associated with an aggressive clinical course. *Br. J. Haematol*., **100**, 655–663.

Ancona,N. *et al.* (2006) On the statistical assessment of classifiers using DNA microarray data. *BMC Bioinformatics*, **7**, 387.

Badea,L. *et al.* (2008) Combined gene expression analysis of whole-tissue and microdissected pancreatic ductal adenocarcinoma identifies genes specifically overexpressed in tumor epithelia. *Hepatogastroenterology*, **55**, 2016–2027.

Bishop,C.M. (2006) *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ.

Chalkias,A. *et al.* (2011) Patients with colorectal cancer are characterized by increased concentration of fecal hb-hp complex, myeloperoxidase, and secretory IgA. *Am. J. Clin. Oncol*., **34**, 561–566.

Chuang,H. *et al.* (2007) Network-based classification of breast cancer metastasis. *Mol. Syst. Biol*., **3**, 140.

Cycon,K. *et al.* (2009) Alterations in CIITA constitute a common mechanism accounting for downregulation of MHC class II expression in diffuse large B-cell lymphoma (DLBCL). *Exp. Hematol*., **37**, 184–194.

Dupire,S. and Coiffier,B. (2010) Targeted treatment and new agents in diffuse large B cell lymphoma. *Int. J. Hematol*., **92**, 12–24.

Friedman,J. *et al.* (2008) Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**, 432–441.

Fröhlich,H. *et al.* (2006) Kernel based functional gene grouping. In: *International Joint Conference on Neural Networks*. IEEE Computer Society, Los Alamitos, CA, USA, pp. 3580–3585.

George,E.I. and McCulloch,R.E. (1997) Approaches for bayesian variable selection. *Statistica Sinica*, **7**, 339–373.

Giaginis,C. *et al.* (2009) Clinical significance of MCM-2 and MCM-5 expression in colon cancer: association with clinicopathological parameters and tumor proliferative capacity. *Dig. Dis. Sci*., **54**, 282–291.

Gordon,K. *et al.* (2009) Bone morphogenetic proteins induce pancreatic cancer cell invasiveness through a Smad1-dependent mechanism that involves matrix metalloproteinase-2. *Carcinogenesis*, **30**, 238–248.

Jaakkola,T.S. and Jordan,M.I. (2000) Bayesian parameter estimation through variational methods. *Stat. Comput*., **10**, 25–37.

Jacob,L. *et al.* (2009) Group lasso with overlap and graph lasso. In: *Proceedings of the 26th International Conference on Machine Learning*. New York, pp. 433–440.

Kameda,K. *et al.* (1999) Expression of highly polysialylated neural cell adhesion molecule in pancreatic cancer neural invasive lesion. *Cancer Lett*., **137**, 201–207.

Keleg,S. *et al.* (2003) Invasion and metastasis in pancreatic cancer. *Mol. Cancer*, **2**, 14.

Krantz,S. *et al.* (2012) Contribution of epithelial-to-mesenchymal transition and cancer stem cells to pancreatic cancer progression. *J. Surg. Res*., **173**, 105–112.

Lasserre,J. *et al.* (2006) Principled hybrids of generative and discriminative models. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Vol. 1, IEEE Computer Society, Washington, DC, USA, pp. 87–94.

Lee,S. *et al.* (2011) Clinicopathologic characteristics of CD99-positive diffuse large B-cell lymphoma. *Acta. Haematol*., **125**, 167–174.

Li,C. and Li,H. (2008) Network-constrained regularization and variable selection for analysis of genomics data. *Bioinformatics*, **24**, 1175–1182.

Li,F. and Zhang,N. (2010) Bayesian variable selection in structured high-dimensional covariate space with applications in genomics. *J. Am. Stat. Assoc*., **105**, 1202–1214.

Menssen,A. *et al.* (2007) c-MYC delays prometaphase by direct transactivation of MAD2 and BubR1: identification of mechanisms underlying c-MYC-induced DNA damage and chromosomal instability. *Cell Cycle*, **6**, 339–352.

Mootha,V.K. *et al.* (2003) PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet*., **34**, 267–273.

Rizzo,A. *et al.* (2011) Intestinal inflammation and colorectal cancer: a double-edged sword? *World J. Gastroenterol*., **17**, 3092–3100.

Rosenwald,A. *et al.* (2002) The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N. Engl. J. Med*., **346**, 1937–1947.

Sakai,N. *et al.* (2012) CXCR4/CXCL12 expression profile is associated with tumor microenvironment and clinical outcome of liver metastases of colorectal cancer. *Clin. Exp. Metastasis*, **29**, 101–110.

Shields,M. *et al.* (2012) Biochemical role of the collagen-rich tumour microenvironment in pancreatic cancer progression. *Biochem. J.*, **441**, 541–552.

Srivastava,S. *et al.* (2008) A novel method incorporating gene ontology information for unsupervised clustering and feature selection. *PLoS One*, **3**, 12.

Stingo,F.C. and Vannucci,M. (2010) Variable selection for discriminant analysis with Markov random field priors for the analysis of microarray data. *Bioinformatics*, **27**, 495–501.

Stingo,F.C. *et al.* (2011) Incorporating biological information into linear models: A Bayesian approach to the selection of pathways and genes. *Ann. Appl. Stat.*, **5**, 1978–2002.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, **102**, 15545–15550.

Terol,M. *et al.* (1999) Expression of beta-integrin adhesion molecules in non-Hodgkin's lymphoma: correlation with clinical and evolutive features. *J. Clin. Oncol.*, **17**, 1869–1875.

Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. R Stat. Soc., B*, **58**, 267–288.

Toiyama,Y. *et al.* (2010) Loss of tissue expression of interleukin-10 promotes the disease progression of colorectal carcinoma. *Surg. Today*, **40**, 46–53.

Toiyama,Y. *et al.* (2012) Evaluation of CXCL10 as a novel serum marker for predicting liver metastasis and prognosis in colorectal cancer. *Int. J. Oncol.*, **40**, 560–566.

Vermeulen,K. *et al.* (2003) The cell cycle: a review of regulation, deregulation and therapeutic targets in cancer. *Cell Prolif.*, **36**, 131–149.

Wang,Q. *et al.* (1998) Altered expression of cyclin D1 and cyclin-dependent kinase 4 in azoxymethane-induced mouse colon tumorigenesis. *Carcinogenesis*, **19**, 2001–2006.

Wei,Z. and Li,H. (2007) A Markov random field model for network-based analysis of genomic data. *Bioinformatics*, **23**, 1537–1544.

Wei,Z. and Li,H. (2008) A hidden spatial-temporal Markov random field model for network-based analysis of time course gene expression data. *Ann. Appl. Stat.*, **2**, 408–429.

Weinel,R. *et al.* (1992) Expression and function of VLA-α_2, -α_3, -α_5 and –$alpha_6$-integrin receptors in pancreatic carcinoma. *Int. J. Cancer*, **52**, 827–833.

Yuan,M. and Lin,Y. (2007) Model selection and estimation in regression with grouped variables. *J. R Stat. Soc., B*, **68**, 49–67.

Zou,H. and Hastie,T. (2005) Regularization and variable selection via the elastic net. *J. R Stat. Soc., B*, **67**, 301–320.

Zycinski,G. *et al.* (2013) Knowledge Driven Variable Selection (KDVS) a new approach to enrichment analysis of gene signatures obtained from high-throughput data. *Source Code Biol. Med.*, **8**, 2.