Web-based Supplementary Materials for Bayesian Nonlinear Model Selection for Gene Regulatory Networks by Yang Ni, Francesco C. Stingo and Veerabhadran Baladandayuthapani

December 11, 2014

Web Appendix A: Inverted Pareto Distribution

Here we discuss the inverted Pareto distribution as it has been defined as the prior distribution for the smoothing parameter τ. Compared to the standard Gamma prior, the inverted Pareto distribution achieves a degree of flexibility in skewness. The inverted Pareto distribution $Ip(a, b)$ has pdf,

$$\pi(\tau|a, b) = \frac{a}{b}(\frac{\tau}{b})^{a-1}, \text{ for } a > 0, \ 0 < \tau < b.$$

The skewness of inverted Pareto distribution can be adjusted through parameter a: it is uniform distribution on $(0, b)$ when $a = 1$, puts its mass on small values when $a < 1$ and puts its mass on large values when $a > 1$. See Web Figure 1 for the density with several different values of a.

Web Appendix B: Specification of Mixture Prior

We elicit the prior by intersecting the Gamma distribution and the inverted Pareto distribution at a pre-specified cut-point. This cut-point provides the best separation of the non-linear and linear components defined by small and large values of τ respectively; and was chosen through simulation studies on a variety of test functions. In order to allow for uncertainty of this choice, we allow for overlapping between the two distributions such that the MCMC sampler has the ability to sample (with positive probability) from one distribution or the other.

Two datasets were generated for this purpose: one is completely linear and the other is completely non-linear. Then we fit our nDAG model to both datasets and plot the kernel density estimation of the MCMC samples of smoothing parameter τ in Web Figure 2 from which we observe the cut-point is around 30.

For the parameter b of the inverted Pareto distribution which controls the range of τ, we choose it to be large enough so that the fitted curve is practically linear for any τ greater than b (Morrissey et al., 2011).

The resulting mixture prior is presented in Web Figure 3.

Web Appendix C: Posterior Inference

Our major interest is in the network parameters, γ_{gj}, and the smoothing parameters, τ_{gj}, that are in the posterior distribution, $\pi(\tau_{\mathbf{g}}, \gamma_{\mathbf{g}}|\mathbf{y_g})$. We therefore integrate out the spline coefficient, β_{gj}, the constant term, μ_g and the error precision, λ_g, from the full likelihood to obtain the marginal log-likelihood:

$$p(\mathbf{y_g}|\tau_{\mathbf{g}}, \gamma_{\mathbf{g}}) \propto |\mathbf{P}_{\gamma_{\mathbf{g}}}|^{1/2}|\mathbf{A_g}|^{-1/2}(\mathbf{1_n}'\mathbf{B_g}\mathbf{1_n} + \kappa_\mu)^{-1/2}(a_g + b_\lambda)^{-(a_\lambda + n/2),}$$

where $\mathbf{1_n}$ is a column vector of ones, $\gamma_\mathbf{g} = (\gamma_{g1}, \ldots, \gamma_{gG})'$, $\mathbf{A_g} = \mathbf{X_{\gamma_g}}'\mathbf{X_{\gamma_g}} + \mathbf{P_{\gamma_g}}$, $\mathbf{B_g} = \mathbf{I_n} - \mathbf{X_{\gamma_g}}\mathbf{A_g}^{-1}\mathbf{X_{\gamma_g}}'$, $\mathbf{I_n}$ is the identity matrix, $a_g = \frac{1}{2}\left\{\mathbf{y_g}'\mathbf{B_g}\mathbf{y_g} - \frac{(\mathbf{y_g}'\mathbf{B_g}\mathbf{1_n})^2}{\mathbf{1_n}'\mathbf{B_g}\mathbf{1_n}+\kappa_\mu}\right\}$, the block diagonal matrix $\mathbf{P_g} = diag(\tau_{g1}\mathbf{K}, \tau_{g2}\mathbf{K}, \ldots, \tau_{gG}\mathbf{K})$, $\mathbf{X_{\gamma_g}}$ is the submatrix of $\mathbf{X}$ corresponding to the columns j, for which $\gamma_{gj} = 1$, and $\mathbf{P_{\gamma_g}}$ is the submatrix of $\mathbf{P_g}$ with the jth block, for which $\gamma_{gj} = 1$. The ordering of the nodes in the graph implies $\gamma_{gj} = 0$ for all $j \geq g$. Then the posterior distribution $\pi(\tau_\mathbf{g}, \gamma_\mathbf{g}|\mathbf{y_g})$ is expressed as

$$\pi(\tau_\mathbf{g}, \gamma_\mathbf{g}|\mathbf{y_g}) \propto p(\mathbf{y_g}|\tau_\mathbf{g}, \gamma_\mathbf{g}) \prod_{j=1}^{g-1} \pi(\gamma_{gj}|\rho)\pi(\tau_{gj}|\phi_{gj}, \gamma_{gj})\pi(\phi_{gj}|\omega, \gamma_{gj}).$$

Below we briefly describe the MCMC algorithm implemented to sample from the posterior distribution $\pi(\tau_\mathbf{g}, \gamma_\mathbf{g}|\mathbf{y_g})$.

SAMPLING SCHEME

Since the posterior distribution $\pi(\tau_\mathbf{g}, \gamma_\mathbf{g}|\mathbf{y_g})$ is analytically intractable, we construct an MCMC sampler to obtain a posterior sample of the parameters τ_{gj} and γ_{gj}. Moreover, since the parameter space of the binary variables γ_{gj} is enormous, it is impractical to compute the explicit posterior probabilities for all possible subsets. Instead, we use stochastic search variable selection (SSVS) and Gibbs sampling (George and McCulloch, 1993) to explore the posterior distribution of the models.

The MCMC consists of two parts: (I) A Metropolis-Hastings (M-H) step for network parameter $\gamma_\mathbf{g} = \{\gamma_{gj}\}_{j=1}^G$, smoothing parameter $\tau_\mathbf{g} = \{\tau_{gj}\}_{j=1}^G$ and mixture indicator $\phi_\mathbf{g} = \{\phi_{gj}\}_{j=1}^G$ and (II) a Gibbs sampler for parameters ρ and ω. To obtain a good mixing of the chain, we accomplish the M-H step in three intermediate steps: (A) sampling $\gamma_\mathbf{g}$, $\tau_\mathbf{g}$ and $\phi_\mathbf{g}$; (B) sampling $\tau_\mathbf{g}$ and $\phi_\mathbf{g}$; and (C) sampling $\tau_\mathbf{g}$.

For $g = 1, \ldots, G$, at each iteration:

(A) Update τ, γ and ϕ jointly (between-model move). This involves sampling the indicator variable γ. For that, we randomly choose j from set $[g-] = \{1, \ldots, g-1\}$ and change

the value of γ_{gj} to γ^*_{gj}, either from 0 to 1 or 1 to 0. If $\gamma^*_{gj} = 1$, then we propose τ^*_{gj} and ϕ^*_{gj} from the log-normal and Bernoulli distributions, respectively. We accept $(\gamma^*_{gj}, \tau^*_{gj}, \phi^*_{gj})$ according to the M-H acceptance ratio,

$$\alpha_{\gamma,\tau} = \frac{p(\mathbf{y_g}|\tau^*_{\mathbf{g}}, \gamma^*_{\mathbf{g}})\pi(\gamma^*_{gj}|\rho)\pi(\tau^*_{gj}|\phi^*_{gj}, \gamma^*_{gj})\pi(\phi^*_{gj}|\omega, \gamma^*_{gj})q(\tau_{gj}|\gamma_{gj})q(\phi_{gj}|\gamma_{gj})q(\gamma_{gj}|\gamma^*_{gj})}{p(\mathbf{y_g}|\tau_{\mathbf{g}}, \gamma_{\mathbf{g}})\pi(\gamma_{gj}|\rho)\pi(\tau_{gj}|\phi_{gj}, \gamma_{gj})\pi(\phi_{gj}|\omega, \gamma_{gj})q(\tau^*_{gj}|\gamma^*_{gj})q(\phi^*_{gj}|\gamma^*_{gj})q(\gamma^*_{gj}|\gamma_{gj})},$$

where $q(\cdot|\cdot)$ is a generic notation for the proposed densities.

(B) Update τ and ϕ. Here we want τ to be updated more often as it is continuous and therefore has a much bigger parameter space than that of a discrete parameter. So in the next two steps, we perform within-model moves for the parameter τ. The first step is to update τ and ϕ together, since τ has a mixture prior that depends on the parameter ϕ. Similarly to what we have done for γ, we randomly choose j from $\{j : \gamma_{gj} \neq 0\}$ and switch ϕ_{gj} to ϕ^*_{gj}, either from 0 to 1 or 1 to 0. For the parameter τ_{gj}, we propose its candidate τ^*_{gj} from the log-normal distribution with the mean equal to τ_{gj}. We accept $(\tau^*_{gj}, \phi^*_{gj})$ according to the ratio

$$\alpha_{\tau} = \frac{p(\mathbf{y_g}|\tau^*_{\mathbf{g}}, \gamma_{\mathbf{g}})\pi(\tau^*_{gj}|\phi^*_{gj}, \gamma_{gj})\pi(\phi^*_{gj}|\omega, \gamma_{gj})q(\tau_{gj}|\tau^*_{gj}, \gamma_{gj})q(\phi_{gj}|\phi^*_{gj}, \gamma_{gj})}{p(\mathbf{y_g}|\tau_{\mathbf{g}}, \gamma_{\mathbf{g}})\pi(\tau_{gj}|\phi_{gj}, \gamma_{gj})\pi(\phi_{gj}|\omega, \gamma_{gj})q(\tau^*_{gj}|\tau_{gj}, \gamma_{gj})q(\phi^*_{gj}|\phi_{gj}, \gamma_{gj})}.$$

(C) Update τ. Here we update τ alone, using the random walk Metropolis-within-Gibbs algorithm. We propose τ^*_{gj} from the log-normal distribution with mean τ_{gj} and accept it with

$$\alpha_{\tau} = \frac{p(\mathbf{y_g}|\tau^*_{\mathbf{g}}, \gamma_{\mathbf{g}})\pi(\tau^*_{gj}|\phi_{gj}, \gamma_{gj})q(\tau_{gj}|\tau^*_{gj}, \gamma_{gj})}{p(\mathbf{y_g}|\tau_{\mathbf{g}}, \gamma_{\mathbf{g}})\pi(\tau_{gj}|\phi_{gj}, \gamma_{gj})q(\tau^*_{gj}|\tau_{gj}, \gamma_{gj})},$$

for each $j \in \{j : \gamma_{gj} \neq 0\}$ sequentially.

(D) Update ρ. We implement a simple Gibbs step to draw ρ^* from $Beta($
$\sum_{j<g} \gamma_{gj} + a_\rho, \sum_{j<g}(1 - \gamma_{gj}) + b_\rho)$.

(E) Update ω. Likewise, we apply a simple Gibbs step to draw ω^* from $Beta(\sum_{j<g} I(\phi_{gj} = 1) + a_\omega, \sum_{j<g} I(\phi_{gj} = 0) + b_\omega)$.

Upon completion, the sampling scheme results in a list of visited models based on the indicator variables. A simple approach for model selection is to select edges whose posterior inclusion probability (i.e., the marginal posterior probability of $\gamma_{gj} = 1$) is higher than a pre-specified threshold. A popular threshold is 0.5, which results in the so-called *median probability model* (Barbieri and Berger, 2004). Marginal posterior probabilities can be approximated by the fraction of time each γ_{gj} is visited by the Markov chain.

An alternative approach, which we use in this paper, is to pick the model with the highest (joint) posterior probability based on the MCMC samples. This approach is not commonly used since, in several applications, a stochastic search results in several models with very similar posterior probabilities. By contrast, in our applications we have observed that the highest probability model is clearly identified, i.e., its posterior probability is much higher than that of the second model in the ranking. We speculate that this desirable feature is due to the greater flexibility of our approach. The two approaches would coincide in many cases, for example, when there is a model with a posterior probability considerably higher than those of the other models (a model with a posterior probability higher than 0.5 is a special case). In our case, we observe a very good agreement between these two approaches.

ESTIMATION AND PREDICTION

Given $\tau_{\mathbf{g}}$ and $\gamma_{\mathbf{g}}$, we derive the conditional posterior distribution of the parameters not sampled in the MCMC algorithm, namely the spline coefficients β_{gj}, the intercept μ_g, and the precision λ_g. The precision parameter has the following posterior distribution,

$$\lambda_g|\mathbf{y_g}, \tau_{\mathbf{g}}, \gamma_{\mathbf{g}} \sim Gamma(a_\lambda + \frac{n}{2}, b_\lambda + b_g),$$

where

$$b_g \;=\; \frac{1}{2}\mathbf{y_g'Dy_g} + \frac{1}{4}\mathbf{y_g'DX}_{\gamma_\mathbf{g}}(-\frac{1}{2}\mathbf{X}'_{\gamma_\mathbf{g}}\mathbf{DX}_{\gamma_\mathbf{g}} - \frac{1}{2}\mathbf{P}_{\gamma_\mathbf{g}})^{-1}\mathbf{X}'_{\gamma_\mathbf{g}}\mathbf{Dy_g}, \quad \mathbf{D} = \mathbf{I_n} - \frac{\mathbf{1_n 1'_n}}{n + \kappa_\mu}.$$

The constant term and spline coefficients are multivariate normal:

$$\alpha_{\mathbf{g}}|\mathbf{y_g}, \lambda_g, \tau_{\mathbf{g}}, \gamma_{\mathbf{g}} \sim N(\mu_{\alpha_{\mathbf{g}}}, \mathbf{T}^{-\mathbf{1}}_{\alpha_{\mathbf{g}}}),$$

where

$$\alpha_{\mathbf{g}} = \begin{pmatrix} \mu_g \\ \beta_{\mathbf{g}} \end{pmatrix}, \ \mathbf{Z}_{\gamma_{\mathbf{g}}} = \begin{pmatrix} \mathbf{1_n} & \mathbf{X}_{\gamma_{\mathbf{g}}} \end{pmatrix}, \ \mathbf{C_g} = \begin{pmatrix} \kappa_\mu & 0 \\ 0 & \mathbf{P}_{\gamma_{\mathbf{g}}} \end{pmatrix},$$

$$\mu_{\alpha_{\mathbf{g}}} = (\mathbf{Z}'_{\gamma_{\mathbf{g}}}\mathbf{Z}_{\gamma_{\mathbf{g}}} + \mathbf{C_g})^{-1}\mathbf{Z}'_{\gamma_{\mathbf{g}}}\mathbf{y_g}, \qquad \mathbf{T}_{\alpha_{\mathbf{g}}} = \lambda_g(\mathbf{Z}'_{\gamma_{\mathbf{g}}}\mathbf{Z}_{\gamma_{\mathbf{g}}} + \mathbf{C_g}).$$

Posterior inference on $\alpha_{\mathbf{g}}$, e.g. estimation of $\alpha_{\mathbf{g}}$ under a squared error loss, can be performed through Monte Carlo integration. The posterior expected value of $\alpha_{\mathbf{g}}$ is

$$E[\alpha_{\mathbf{g}}|\mathbf{y_g}] \approx \frac{1}{N}\sum_{i=1}^{N} \mu_{\alpha_{\mathbf{g}}}({\tau_{\mathbf{g}}}^{(i)}, {\gamma_{\mathbf{g}}}^{(i)}),$$

where ${\tau_{\mathbf{g}}}^{(i)}, {\gamma_{\mathbf{g}}}^{(i)}$ are N posterior samples drawn from the MCMC procedure.

Suppose we observe a new set of data $\mathbf{y^*_1}, \ldots, \mathbf{y^*_{g-1}}$ with sample size m. Let $\mathbf{X^*_g}$ be the spline design matrix of $\mathbf{y^*_1}, \ldots, \mathbf{y^*_{g-1}}$ and $\mathbf{X}^*_{\gamma_{\mathbf{g}}}$ be the same matrix given $\gamma_{\mathbf{g}}$ and let $\mathbf{Z}^*_{\gamma_{\mathbf{g}}} = (\mathbf{1_m}\ \mathbf{X}^*_{\gamma_{\mathbf{g}}})$. Then the predictive distribution of new observations $\mathbf{y^*_g}$ given $\tau_{\mathbf{g}}, \gamma_{\mathbf{g}}$ is a multivariate student's t-distribution with mean

$$\mu^*_{\mathbf{g}} = \{\mathbf{I_m} - \mathbf{Z}^*_{\gamma_{\mathbf{g}}}(\mathbf{Z}^{*\prime}_{\gamma_{\mathbf{g}}}\mathbf{Z}^*_{\gamma_{\mathbf{g}}} + \mathbf{Z}'_{\gamma_{\mathbf{g}}}\mathbf{Z}_{\gamma_{\mathbf{g}}} + \mathbf{C_g})^{-1}\mathbf{Z}^{*\prime}_{\gamma_{\mathbf{g}}}\}^{-\mathbf{1}}\mathbf{Z}^*_{\gamma_{\mathbf{g}}}(\mathbf{Z}^{*\prime}_{\gamma_{\mathbf{g}}}\mathbf{Z}^*_{\gamma_{\mathbf{g}}} + \mathbf{Z}'_{\gamma_{\mathbf{g}}}\mathbf{Z}_{\gamma_{\mathbf{g}}} + \mathbf{C_g})^{-1}\mathbf{Z}'_{\gamma_{\mathbf{g}}}\mathbf{y_g};$$

hence, by Bayesian model averaging over the high-probability model (Raftery et al., 1997; Hoeting et al., 1999), the mean of the predictive distribution can be approximated by

$$E[\mathbf{y^*_g}|\mathbf{y_g}, \mathbf{X^*_g}] \approx \frac{1}{N}\sum_{i=1}^{N} \mu^*_{\mathbf{g}}({\tau_{\mathbf{g}}}^{(i)}, {\gamma_{\mathbf{g}}}^{(i)}).$$

Web Appendix D: Simulated Examples of Single Regression

In addition to the full network analysis in section 5, we conduct another simulation study for variable selection under single nonlinear regression setting. In this example, we generate 49 predictor variables independently from the standard uniform $[0, 1]$ distribution. After standardizing each predictor, a response variable then has the following functional form:

$$\mathbf{y} = 2 + 3\mathbf{x_1} - 4\mathbf{x_2^2} - \exp(\mathbf{x_3}) + 4\cos\left(\frac{4\pi\mathbf{x_4}}{7}\right) + 2\sin\left(\frac{8\pi\mathbf{x_5}}{7}\right) + \epsilon, \quad (1)$$

where $\epsilon \sim N_n(\mathbf{0}, \sigma^2\mathbf{I_n})$. Hence, the true model contains only the first five variables.

The hyperparameters setting is the same as in section 5. We run an MCMC algorithm with $20,000$ iterations, in which the first $2,000$ iterations are considered as a "burn-in" period for both methods. We conduct 50 simulations for each method.

In the simulated true model, we have one linear function and four nonlinear functions: quadratic, exponential, cosine (with approximately one period within the data range), and sine (with approximately two periods within the data range) functions. We randomly choose one simulation and plot the reconstructions for the nMixDAG after adjusting for the mean. This plot appears in Web Figure 4, with the solid curves representing the true curves and the dashed curves the estimated curves. The 95% credible bands are also plotted as dotted lines. The performance of the reconstruction is quite reasonable, as the true curves and the estimated curves pretty much lie on top of each other and the true curves are contained in the credible bands. The quality of the reconstructions of the nDAG is almost identical to that of the nMixDAG (plot not shown).

We vary the sample sizes ($n = 150, 200, 250$) and error variances ($\sigma^2 = 4, 16$) to explore the performance of SSVS, nDAG, nMixDAG, spikeSlabGAM and SpAM under different signal-to-noise settings. We tabulate the results in Web Table 1 where the true positive

rate (TPR), the false discovery rate (FDR), the false negative rate (FNR=1-TPR) and mean squared prediction error (MSPE) are reported. The linear method SSVS is apparently not competitive with all the nonlinear approaches in terms of both prediction and variable selection and nMixDAG is always slightly better than nDAG. Although the difference is not substantial, spikeSlabGAM is the best in prediction. For instance, when $n = 200, \sigma^2 = 16$, the average MSPE of nMixDAG, spikeSlabGAM and SpAM are 21.161, 19.994 and 26.079, respectively. In terms of variable selection, generally, SpAM has the highest TPR yet at the price of unreasonable FDR (around 70%). On the other hand, although nMixDAG has lower TPR than spikeSlabGAM and SpAM, it has 0 FDR for all settings. For example, when $n = 250, \sigma^2 = 16$, nMixDAG has an average 0.804 TPR and 0 FDR while spikeSlabGAM has 0.980 TPR and 0.279 FDR, and SpAM has 1.000 TPR and 0.733 FDR. The trade-off between TPR and FDR shows that our method nMixDAG is more parsimonious than spikeSlabGAM and SpAM.

Web Appendix E: GBM Data Analysis

In section 6, we analyzed the TCGA-based GBM gene expression data using our proposed model nMixDAG. TCGA provides microarray-based gene expression data for a large cohort of hundreds of GBM tumor specimens (241 in our case study) (TCGA, 2008). Newly-diagnosed glioblastomas are selected retrospectively from biospecimen repositories and further reviewed and processed through TCGA Biospecimen Core Resource to ensure less than 50% necrosis and more than 80% tumor nuclei for RNA extraction. Each qualified biospecimen is assayed using three platforms: Affymetrix U133A, Affymetrix Exon 1.0 ST and custom Agilent 244K, from which messenger RNA (mRNA) expression profiles are generated. Subsequently, the mRNA expression profiles are integrated into a single estimate of relative gene expression for each gene and for each sample. The heat map of marginal posterior inclusion probability for the edge between each pair of genes is shown in Web Figure 5 from which the sparsity of

our model is evident. Also, the marginal probabilities appear separated, i.e. either close to 0 or 1, which suggests the highest probability model in our application is clearly identifiable.

In addition, as kindly suggested by the anonymous associate editor, we include more genes in our real data analysis. We focus on the full RTK/PI3K signaling pathway[1] instead of the frequently mutated genes from the three core pathways, which consists of 195 genes. Despite the dimension being much higher and having run the same length of chains, we found strong evidence of convergence for all parameters: the PSRFs for τ (95% of τ values ranging from 1.000 to 1.0966), ρ (=1.000) and ω (=1.000) are close to 1 and the correlations of posterior probabilities for γ and ϕ are 0.911 and 0.971, respectively. Totally, we found 463 connections and 91 of them are nonlinear. While some connections are well studied: e.g. Ras activates PI3K family (Yan et al., 1998), some have not been fully understood, especially the nonlinear regulations listed In Web Tables 5, 6, 7. They are ordered by the degree of nonlinearity. We find 8 hub genes are: PTK2, ITGA3, KDR, SYK, PRLR, SOS2, PDPK1 and HSP90AB1. Two of them, KDR and PDPK1, were previously detected as driver genes for GBM in the literatures (TCGA, 2008; Cerami et al., 2010). The other 6 newly found hub genes might be potential GBM driver genes which need to be validated through biological experiment. We also apply the spikeSlabGAM to this dataset for comparison. The difference in WAIC between nMixDAG and spikeSlabGAM is substantial (103559 vs 211571), which again indicates nMixDAG has higher prediction power.

Web Appendix F: Markov Equivalence Class

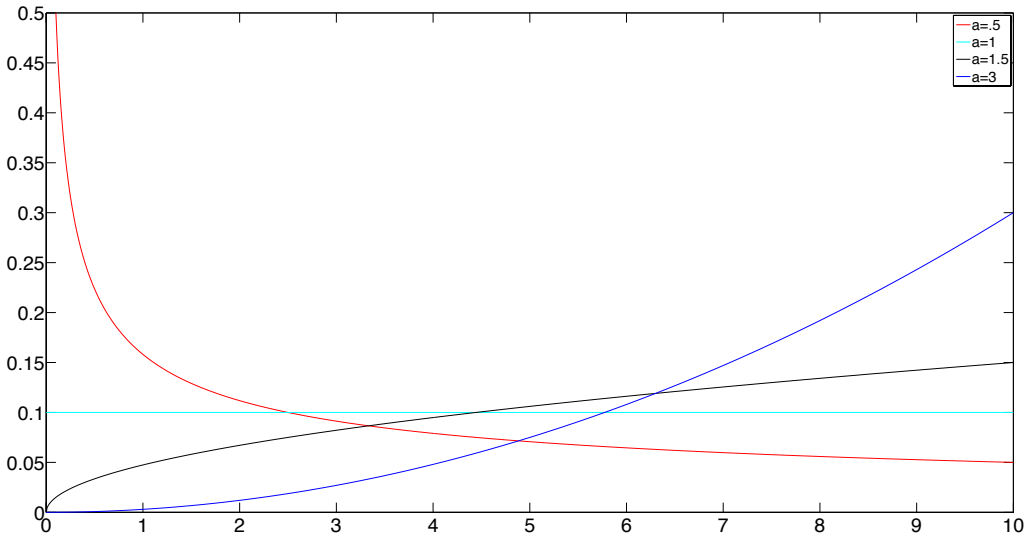
Our model is defined based on a prior ordering of the nodes. Without a prior ordering, we cannot distinguish two DAGs within the same Markov equivalence class (MEC) in which all DAGs have the same conditional independence assertions. If we ignore this aspect, we would define a computational inefficient approach that will never discriminate two Markov equivalent DAGs. More importantly, for (linear) Gaussian DAG, if the parameter priors are

[1] http://www.genome.jp/kegg/

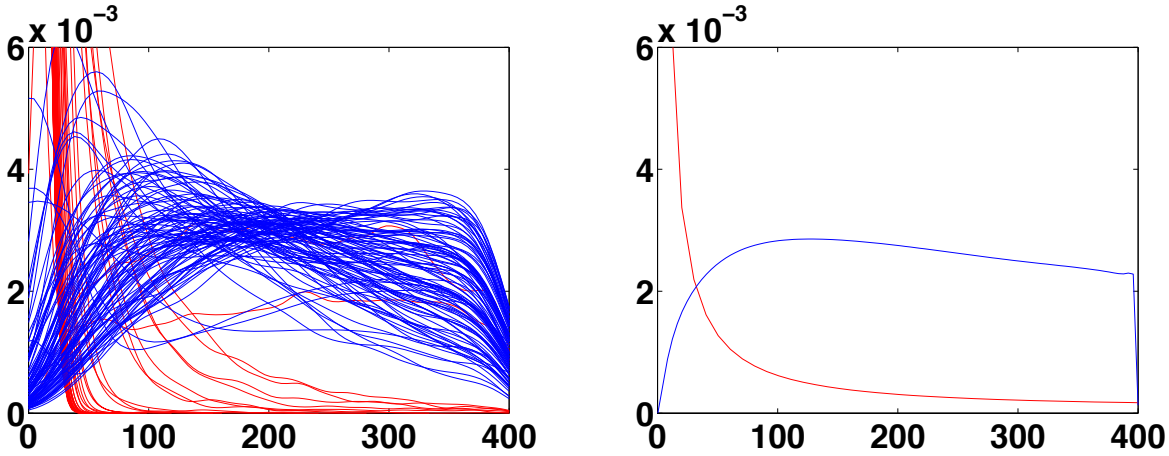
carefully chosen, the marginal likelihood is equivalent for Markov equivalent DAGs (Geiger and Heckerman, 2002). However, since our proposed DAG models are nonlinear, two Markov equivalent DAGs can represent different sets of distributions. In our work the ordering of the nodes is naturally obtained for the pathway information. Here we discuss the amount of prior information needed to discriminate two Markov equivalent DAGs when the prior ordering of the nodes is not available.

One possible way is to discriminate Markov equivalent graphs by adopting an informative prior based on the number of common edges between the proposed network and the reference network. However, it is not as straightforward as it appears. Consider the following example in Web Figure 6. Suppose the graph on the left panel is the reference network. The two graphs in middle and on the right are Markov equivalent and they have the same number of common edges with the reference graph. Hence the proposed prior would fail to distinguish between these two. In practice, the amount of prior biological information needed to discriminate two Markov equivalent DAGs is commensurate to the amount of the information needed to define an ordering of the model.

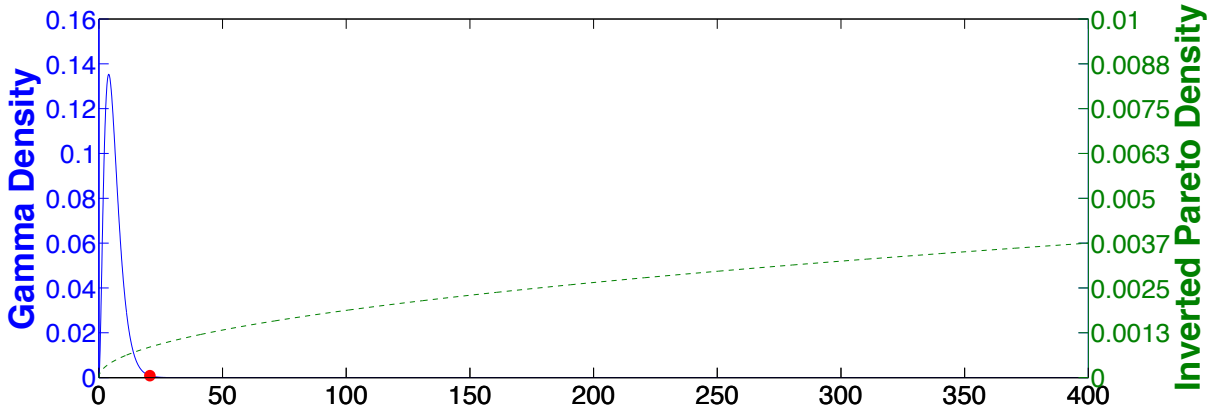
Another possible way to avoid repeatedly analyzing Markov equivalent DAGs is inspired by Andersson et al. (1997). They suggest to treat each MEC as a single model through *essential graph*, the union of all the graphs within the same MEC. There is a one-to-one relationship between MEC and its essential graph. Therefore, working with the MEC space could be a feasible and effective approach (Chickering, 2002). However, this approach does not help solve the issue of discriminating Markov equivalent DAGs either.



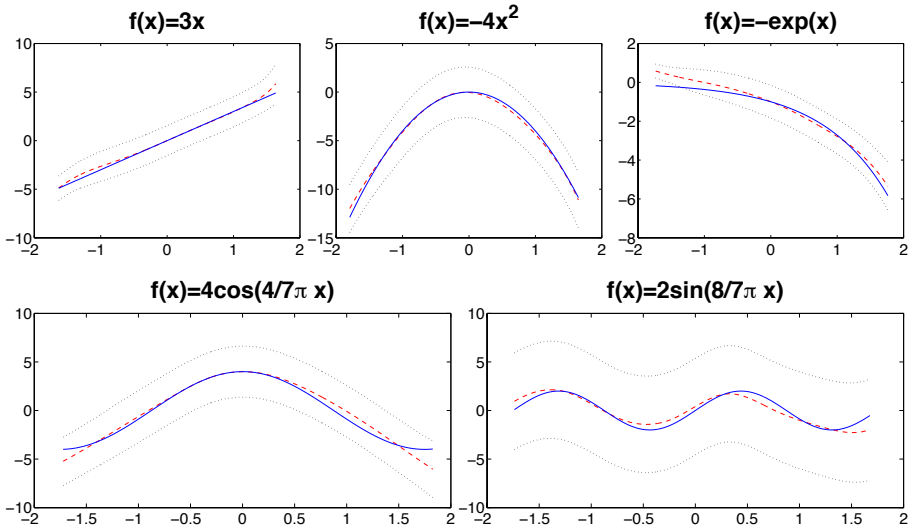
Web Figure 1: The inverted Pareto distribution with $b = 10$ and $a = 0.5$ (red), 1 (cyan), 1.5 (black), 3 (blue).



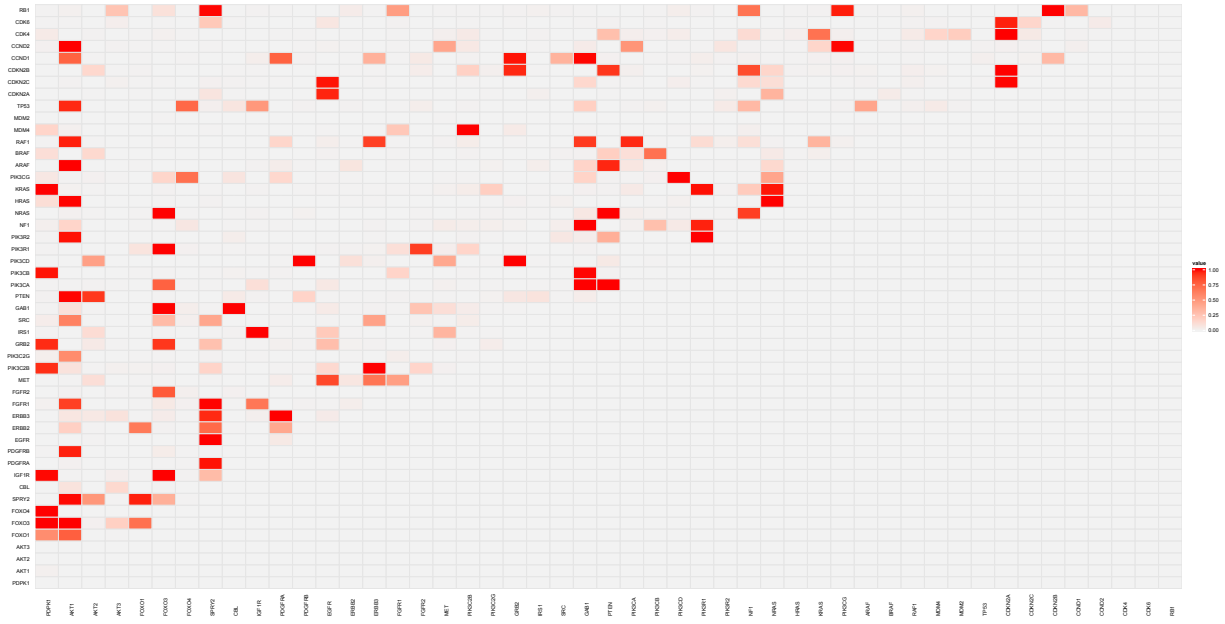
Web Figure 2: Specification of mixture prior. Kernel density estimation of the smoothing parameter τ. The blue curves correspond to linear fit while the red curves correspond to non-linear fit.



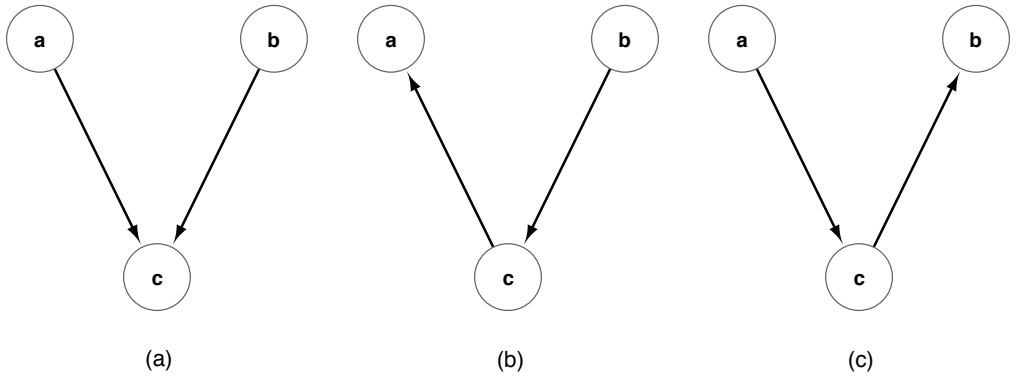
Web Figure 3: Gamma and Pareto distributions are plotted on different scales with their densities defined by the y-axis on the left and right sides, respectively. The solid curve is the Gamma density; the dashed curve is the inverted Pareto density. The red dot is the intersection of the two densities if they were plotted on the left y-axis scale.



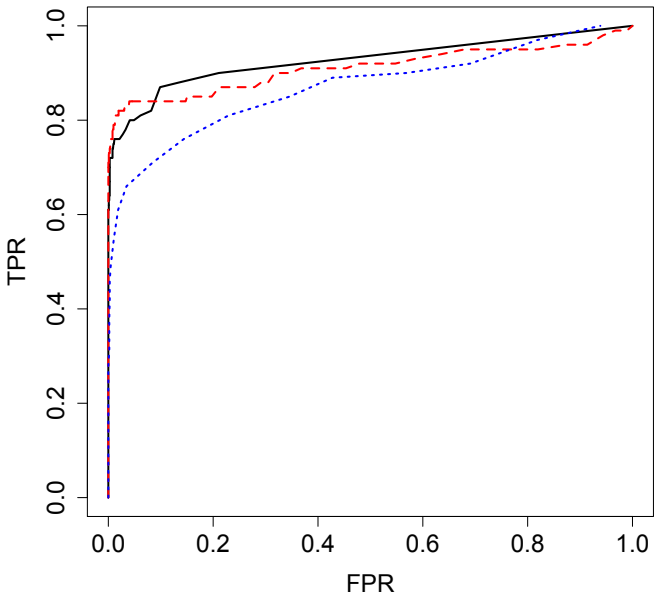
Web Figure 4: Single regression. Reconstruction of the five functions used to simulate the data. The solid lines are the true functions, the dashed lines are the estimated functions, with credible bands shown as dotted lines.



Web Figure 5: GBM data. The heat map of marginal posterior inclusion probability for each pair of genes. The darker the color, the higher the probability.



Web Figure 6: Markov equivalence class. Example of indistinguishable Markov equivalent DAGs. (a) The reference network; (b) and (c) Markov equivalent networks.



Web Figure 7: Receiver operating characteristic curves for one simulation from scenario 3. The solid curve represents nMixDAG; the dashed curve represents spikeSlabGAM; and the dotted curve represents SpAM.

Web Table 1: Single regression. Operating characteristics for SSVS, nDAG, nMixDAG, spikeSlabGAM and SpAM are calculated under different sample sizes and error variances. The bold numbers indicate the best performance. The standard error for each statistic is given in the parentheses.

Scenario		SSVS	nDAG	nMixDAG	spikeSlabGAM	SpAM
$n = 250$	TPR	0.452(0.097)	0.988(0.048)	**1.000**(**0.000**)	**1.000**(**0.000**)	**1.000**(**0.000**)
$\sigma^2 = 4$	FDR	0.073(0.133)	**0.000**(**0.000**)	**0.000**(**0.000**)	0.133(0.135)	0.735(0.096)
	FNR	0.548(0.097)	0.012(0.048)	**0.000**(**0.000**)	**0.000**(**0.000**)	**0.000**(**0.000**)
	MSPE	27.533(5.263)	6.089(2.312)	5.800(2.667)	**4.691**(**0.904**)	6.860(1.592)
$n = 250$	TPR	0.412(0.110)	0.800(0.057)	0.804(0.086)	0.980(0.061)	**1.000**(**0.000**)
$\sigma^2 = 16$	FDR	0.077(0.159)	**0.000**(**0.000**)	**0.000**(**0.000**)	0.279(0.162)	0.733(0.090)
	FNR	0.588(0.110)	0.200(0.057)	0.196(0.086)	0.020(0.061)	**0.000**(**0.000**)
	MSPE	39.278(6.952)	20.090(3.875)	19.882(3.805)	**18.833**(**3.446**)	23.193(4.824)
$n = 200$	TPR	0.352(0.095)	0.740(0.093)	0.732(0.096)	0.940(0.093)	**0.980**(**0.061**)
$\sigma^2 = 16$	FDR	0.128(0.187)	**0.000**(**0.000**)	**0.000**(**0.000**)	0.271(0.198)	0.692(0.112)
	FNR	0.648(0.095)	0.260(0.093)	0.268(0.096)	0.060(0.093)	**0.020**(**0.061**)
	MSPE	40.952(7.829)	21.320(4.566)	21.161(4.522)	**19.994**(**3.855**)	26.079(5.464)
$n = 150$	TPR	0.292(0.101)	0.644(0.123)	0.652(0.113)	0.880(0.114)	**0.912**(**0.108**)
$\sigma^2 = 16$	FDR	0.142(0.226)	**0.000**(**0.000**)	**0.000**(**0.000**)	0.366(0.166)	0.681(0.097)
	FNR	0.708(0.101)	0.356(0.123)	0.348(0.113)	0.120(0.886)	**0.088**(**0.108**)
	MSPE	41.446(8.593)	24.057(9.382)	23.152(8.470)	**20.977**(**3.682**)	26.939(4.916)

Web Table 2: Simulated networks. The average operating characteristics for nMixDAG with dimension $G = 50, 100, 150, 200$. The standard error for each statistic is given in the parentheses.

			Methods			
Dimension	AUC5%	AUC	TPR	TPR_linear	TPR_nonlinear	FDR
50	0.718(0.036)	0.879(0.019)	0.618(0.027)	0.730(0.029)	0.514(0.040)	0.023(0.021)
100	0.721(0.025)	0.867(0.012)	0.591(0.023)	0.670(0.031)	0.518(0.037)	0.005(0.008)
150	0.697(0.022)	0.854(0.011)	0.586(0.023)	0.686(0.039)	0.493(0.033)	0.001(0.003)
200	0.669(0.024)	0.840(0.012)	0.582(0.019)	0.715(0.023)	0.459(0.033)	0.004(0.009)

Web Table 3: Simulated networks. Sensitivity analysis on the hyper-prior of the parameter ρ. Three different values of parameters (a_ρ, b_ρ) are tested.

Prior	$(a_\rho, b_\rho) = (2, 2)$	$(a_\rho, b_\rho) = (0.5, 0.5)$	$(a_\rho, b_\rho) = (2, 3)$	$(a_\rho, b_\rho) = (3, 2)$
AUC5%	0.718(0.036)	0.703(0.033)	0.715(0.039)	0.721(0.032)
AUC	0.879(0.019)	0.870(0.017)	0.878(0.019)	0.881(0.019)
TPR	0.618(0.027)	0.610(0.028)	0.618(0.028)	0.625(0.027)
TPR_linear	0.730(0.029)	0.725(0.029)	0.728(0.034)	0.743(0.029)
TPR_nonlinear	0.514(0.040)	0.503(0.043)	0.516(0.040)	0.517(0.039)
FDR	0.023(0.021)	0.034(0.029)	0.038(0.026)	0.036(0.026)

Web Table 4: Simulated networks. Sensitivity analysis of the hyper-prior on the smoothing parameter τ. Four different sets of parameters (a_τ, b_τ) and (k_τ, θ_τ) are tested.

Prior	$(a_\tau, b_\tau) = (1.5, 400)$ $(k_\tau, \theta_\tau) = (3, 2)$	$(a_\tau, b_\tau) = (3, 600)$ $(k_\tau, \theta_\tau) = (3, 5)$	$(a_\tau, b_\tau) = (2, 500)$ $(k_\tau, \theta_\tau) = (3, 4)$	$(a_\tau, b_\tau) = (1.5, 300)$ $(k_\tau, \theta_\tau) = (1, 5)$	$(a_\tau, b_\tau) = (1.5, 400)$ $(k_\tau, \theta_\tau) = (3, 1)$
AUC5%	0.718(0.036)	0.718(0.034)	0.720(0.034)	0.709(0.036)	0.701(0.033)
AUC	0.879(0.019)	0.878(0.019)	0.881(0.019)	0.875(0.018)	0.870(0.018)
TPR	0.618(0.027)	0.620(0.027)	0.623(0.030)	0.619(0.028)	0.611(0.029)
TPR_linear	0.730(0.029)	0.735(0.030)	0.740(0.033)	0.736(0.033)	0.731(0.033)
TPR_nonlinear	0.514(0.040)	0.514(0.042)	0.515(0.044)	0.510(0.041)	0.500(0.042)
FDR	0.023(0.021)	0.034(0.024)	0.031(0.024)	0.038(0.027)	0.039(0.025)

Web Table 5: GBM data analysis. The top 30 of 91 Nonlinear regulations (from column Source to column Target) identified by nMixDAG. They are ordered by the nonlinearity measure defined as the posterior probability $p(\phi = 1|\mathbf{Y})$, which is shown in the last column.

Target	Source	Nonlinearity
ITGB7	IL7R	0.9987
GNB1	JAK1	0.9980
IL2RG	IL2RB	0.9976
PPP2R5E	PPP2R5C	0.9944
CDK4	MDM2	0.9930
MAPK1	PPP2R3A	0.9926
TLR2	CSF1R	0.9898
MAP2K2	PKN1	0.9875
IL2RB	SYK	0.9861
GYS2	PRLR	0.9860
KRAS	PTK2	0.9825
RAF1	PTK2	0.9776
G6PC3	FOXO3	0.9724
PCK1	PTEN	0.9714
YWHAH	YWHAZ	0.9669
CASP9	ITGA3	0.9657
SGK3	PTK2	0.9652
HSP90B1	PDPK1	0.9623
YWHAB	RAC1	0.9622
CSF3R	SYK	0.9591
PIK3R2	PIK3R1	0.9500
PPP2CB	RAC1	0.9489
PPP2R5C	HSP90AA1	0.9392
FOXO3	JAK1	0.9303
RPS6KB2	JAK1	0.9222
SGK3	IFNAR1	0.9164
HSP90AA1	RAC1	0.9163
ITGA8	FLT1	0.9101
CHUK	EIF4E	0.9073
MAP2K1	RPS6	0.9063

Web Table 6: GBM data analysis. The middle 30 of 91 Nonlinear regulations (from column Source to column Target) identified by nMixDAG. They are ordered by the nonlinearity measure defined as the posterior probability $p(\phi = 1|\mathbf{Y})$, which is shown in the last column.

Target	Source	Nonlinearity
PIK3CB	GNGT1	0.8784
IL2RA	CSF1R	0.8742
ITGB6	PRLR	0.8728
CCNE2	PKN1	0.8726
PPP2R3C	NRAS	0.8716
YWHAZ	PPP2CA	0.8598
GNB2	GNB1	0.8593
SYK	CSF1R	0.8592
FASLG	GYS2	0.8448
MAPK1	YWHAH	0.8410
PPP2CB	SYK	0.8395
SYK	TLR2	0.8267
EIF4E2	EIF4E	0.8001
PPP2R5D	HSP90AB1	0.7987
CREB3L2	GSK3B	0.7985
EIF4B	INSR	0.7961
MAPK3	EIF4B	0.7923
MYB	CCNE2	0.7686
EIF4E2	EIF4EBP1	0.7576
JAK1	INSR	0.7507
PPP2R1A	HSP90AB1	0.7466
BRCA1	RPS6KB1	0.7463
PKN1	AKT1	0.7462
CDC37	HSP90AB1	0.7434
CREB1	HSP90AB1	0.7415
PIK3CG	SYK	0.7351
PDPK1	IGF1R	0.7337
IL4R	PDGFRB	0.7302
PPP2CA	HSP90AB1	0.7232
ITGA5	KIT	0.7111

Web Table 7: GBM data analysis. The last 31 of 91 Nonlinear regulations (from column Source to column Target) identified by nMixDAG. They are ordered by the nonlinearity measure defined as the posterior probability $p(\phi = 1|\mathbf{Y})$, which is shown in the last column.

Target	Source	Nonlinearity
PRKACA	AKT1	0.6960
GNB5	IL4R	0.6911
MAP2K2	PPP2R1A	0.6859
CCNE2	GNG4	0.6813
MCL1	RPS6KB1	0.6636
MAP2K1	YWHAH	0.6598
CCNE1	CDK2	0.6477
IKBKB	PTK2	0.6465
GSK3B	PDPK1	0.6320
PIK3R3	EDG4	0.6167
BCL2L11	CASP9	0.6111
CCNE2	CDK2	0.6105
PPP2R2B	IL4R	0.5877
RBL2	PDPK1	0.5800
CASP9	MAPK3	0.5749
PPP2R2A	PTK2	0.5692
IL4R	CSF1R	0.5689
EIF4E	NRAS	0.5682
PIK3CG	P2RY5	0.5672
EIF4EBP1	NRAS	0.5592
HRAS	GNB2	0.5589
HSP90AB1	GNB1	0.5584
PPP2R5C	PPP2R3C	0.5549
GSK3B	RAF1	0.5486
GNG4	GNB5	0.5460
PPP2R5B	BCR	0.5400
PDPK1	RAC1	0.5317
HSP90AB1	PIK3CA	0.5293
HSP90AA1	SOS2	0.5287
IL4R	TLR2	0.5262
BCL2L11	GNG13	0.5152

References

Andersson, S. A., Madigan, D., and Perlman, M. D. (1997). A characterization of markov equivalence classes for acyclic digraphs. *The Annals of Statistics* **25,** 505–541.

Barbieri, M. M. and Berger, J. O. (2004). Optimal predictive model selection. *The Annals of Statistics* **32,** 870–897.

Cerami, E., Demir, E., Schultz, N., Taylor, B. S., and Sander, C. (2010). Automated network analysis identifies core pathways in glioblastoma. *PLoS ONE* **5,** e8918.

Chickering, D. M. (2002). Learning equivalence classes of bayesian network structures. *The Journal of Machine Learning Research* **2,** 445–498.

Geiger, D. and Heckerman, D. (2002). Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *The Annals of Statistics* **30,** 14121440.

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

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical science* pages 382–401.

Morrissey, E. R., Juarez, M. A., Denby, K. J., and Burroughs, N. J. (2011). Inferring the time-invariant topology of a nonlinear sparse gene regulatory network using fully Bayesian spline autoregression. *Biostatistics* **12,** 682–694.

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* **92,** 179–191.

TCGA (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455,** 1061–1068.

Yan, J., Roy, S., Apolloni, A., Lane, A., and Hancock, J. F. (1998). Ras isoforms vary in their ability to activate raf-1 and phosphoinositide 3-kinase. *Journal of Biological Chemistry* **273,** 24052–24056.