Harvard University
Harvard University Biostatistics Working Paper Series

Year 2010 *Paper* 114

Modeling Dependent Gene Expression

Donatello Telesca* Peter Muller†
Giovanni Parmigiani‡ Ralph S. Freedman**

*UCLA School of Public Health, donatello.telesca@gmail.com
†University of Texas M.D. Anderson Cancer Center, pm@odin.mdacc.tmc.edu
‡Harvard School of Public Health and Dana Farber Cancer Institute, gp@jimmy.harvard.edu
**The University of Texas M.D. Anderson Cancer Center

http://biostats.bepress.com/harvardbiostat/paper114

Modeling Dependent Gene Expression

DONATELLO TELESCA[1,5], PETER MÜLLER[2],
GIOVANNI PARMIGIANI[3], RALPH S. FREEDMAN[4]

Author's Footnote

[1] UCLA School of Public Health, Department of Biostatistics.

[2] The University of Texas M.D. Anderson Cancer Center, Department of Biostatistics.

[3] Dana Farber Cancer Institute, Department of Biostatistics and Computational Biology and Harvard School of Public Health, Department of Biostatistics.

[4] The University of Texas M.D. Anderson Cancer Center, Department of Gynecologic Oncology.

February 8, 2010

[5]For Correspondence

Donatello Telesca, Ph.D.

Department of Biostatistics

UCLA School of Public Health

Los Angeles, CA 90095–1772

phone: 310 825 6402

e-mail: donatello.telesca@gmail.com

Abstract

In this paper we propose a Bayesian modeling approach for inference about dependence of high throughput gene expression. Our goals are to use prior knowledge about pathways to anchor inference about dependence among genes; to account for this dependence while making inferences about differences in mean expression across phenotypes; and to explore differences in the dependence itself across phenotypes. Useful features of the proposed approach are a model-based parsimonious representation of expression as an ordinal outcome, a novel and flexible representation of prior information on the nature of dependencies, and the use of a coherent probability model over both the structure and strength of the dependencies of interest. We evaluate our approach through simulations and in the analysis of data on expression of genes in the Complement and Coagulation Cascade pathway in ovarian cancer.

Keywords: Conditional Independence, Microarray Data, Probability Of Expression, Probit Models, Reciprocal Graphs, Reversible Jumps MCMC.

1 INTRODUCTION

Inferring patterns of dependence from high throughput geneomic data poses significant challenges. Statistically, the problem is one of learning about dependence structures in high dimension, with relatively low signal. A promising direction for strenghtening this inference is the explicit consideration of information from known 'pathways' — biochemical processes described in terms of a series of relationships among genes and their products.

In this paper we take this perspective, and propose a Bayesian approach to achieve three related goals in the context of gene expression analysis: to use prior knowledge about pathways to anchor inference about dependence among genes; to account for this dependence while making inferences about differences in mean expression across phenotypes; and to explore differences in the dependence itself across phenotypes. The proposed model builds on the POE model (Parmigiani et al. 2002) and integrates inference about probability of differential expression with inference about dependence between genes through the formulation of a coherent probability model. Our proposed inferences are local in the sense that the model is centered around a specific pathway. Formally, variable selection is used to remove and add structure relative to the centering pathway. This is in contrast to approaches aimed at learning dependence structures de novo from expression data, without guidance by a prior pathway structure.

Some of the existing approaches for probabilistic modeling of dependence structures attempt to explore the space of all possible graphical models, often restricted to Directed Acyclic Graphs (DAGs) or Bayesian networks (BN) (Lauritzen 1996) and decomposable models (Dawid and Lauritzen 1993). A comprehensive review of statistical methodology for network data is provided in Kolaczyk (2009). Recent literature includes the application of BN and dynamic BN to microarray data (Murphy and Mian 1999, Friedman et al. 2000),

with applications and extensions of this methodology reported in Ong et al. (2002) and Beal et al. (2005) among others. Although appealing, these techniques have computational and methodological limitations related to modeling conditional independence under the "large p, small n" paradigm and the difficult specification of consistent prior models across dimensions (Dobra et al. 2004). Other authors (Scott and Carvalho 2008, Jones et al. 2005) have reported difficulties with the performance of standard trans-dimensional MCMC methods (Giudici and Green 1999) in the exploration of the model space, and suggested alternative stochastic search schemes. For a decision theoretic perspective on graphical model selection see Sebastiani (2005).

To overcome these problems, we focus on variations of a baseline model that represents known dependence structures. The centering anchors the model space to a prior path diagram elicited from sets of bimolecular interactions derived by previous experiments.

Our idea is similar to the modeling approaches described in Wei and Li (2007) and Wei and Li (2008), who introduced conditional independence between genes, via a Markov random field (mrf) defined over binary hidden states of differential expression. These authors propose to consider a fixed mrf mirroring exactly the topology of a prior pathway and ignoring the directionality of the edges. We contrast this approach in three fundamental ways. First, we provide an alternative interpretation of the connections encoded into a prior pathway. We develop a prior model for the dependence structure that is based on the reciprocal graphs (Koster 1996). This class of graphical models takes full account of the directionality of the edges included in the pathway and allows for the Markovian characterization of cycles, which often arise in biological depictions of genetic interactions. Also, recognizing that a known pathway is often summarizing results obtained under different experimental conditions, we allow for significant deviations from the prior dependence structure. This requires explicit consideration of a model determination strategy, but enables inference on the model

parameters as well as inference on the dependence structure between genes. Finally, our focus is on identifying significant interactions between genes in a prior pathway, as opposed to identifying differentially expressed genes in a given pathway.

The rest of this article is organized as follows. In Section 2 we introduce the proposed model. Section 3 discusses estimation and inferential details associated with the proposed model. We validate our approach with a simulation study in Section 4. Section 5 employs the model for the analysis of epithelial ovarian cancer expression data, to derive inference about active genetic interactions. In the example, a well known molecular pathway provides prior information on the dependence structure. A final section concludes with a critical discussion of limitations and possible extensions.

2 DEPENDENT PROBABILITY OF EXPRESSION

In Section 2.1, we discuss graphical models and notation, and in Section 2.2, we review the POE (Probability of Expression) model Parmigiani et al. (2002), which defines biological events via latent three-way indicators of relevant biological states. The original POE model assumes independence across genes, conditional on hyperparameters. We extend the original model by formalizing more complex relationships among variables via a cascade of conditional dependences, guided by a predefined interaction map.

2.1 **Modeling Dependence: Background on Graphical Models**

Networks of relationships among expression levels can be represented as graphs that describe how genes influence each other (for an example in Ovarian Cancer see Wang et al. (2005)). More formally, a graph is often characterized as an algebraic structure $\mathcal{G} = \{V, E\}$, composed of a set of nodes V, in our case genes, and a set of edges $E \in \{V\} \times \{V\}$. A graph $\mathcal{G}$ defines the Markov properties of a statistical model in a graphical fashion, via the specification of a

set of conditional dependencies.

Biochemical networks often include the presence of cycles and feed-back relationships. This requires some care when trying to characterize a coherent probabilistic model that accurately portrays prior biochemical knowledge. For this purpose, we focus on a class of graphical models known as reciprocal graphs (Koster 1996). Reciprocal graph are defined as a natural generalization of other well known classes, including directed acyclic graphs (DAG) and Markov random Fields (mrf), among others. Reciprocal graphs are defined through the coherent probabilistic interpretation of directed $a(\rightarrow b)$, indirected $(a - b)$ and reciprocal edges $(a \rightleftarrows b)$. Here, for simplicity, we consider a subset of the reciprocal graph family excluding undirected edges.

The proposed model and inference is based on the directed graph $\mathcal{G}$. But sometimes it is of interest to describe the implied conditional independence structure, i.e. the Markov properties. When desired, the Markov properties of our model are defined in terms of an undirected graph $\mathcal{G}^m = \{V, E^m\}$ elicited via moralization (Koster 1996, Lauritzen 1996) of a graph $\mathcal{G}$. In substance, the moralization procedure consists in adding an undirected edge between parents of a common child and replacing the remaining directed edges with undirected ones. In $\mathcal{G}^m$, standard Markov field properties hold, in the sense that two genes i and j are disconnected when they are conditionally independent, given the rest of the genes (Besag 1974). For example, consider the reciprocal graph $\mathcal{G}$ represented in Figure 1. The class of Markov equivalent models spanned by $\mathcal{G}$, may be represented with the moral (undirected) graph $\mathcal{G}^m$, for which the following Markov property holds: $1 \perp 2 \mid 3, 4$, that is $P(1, 3 \mid 2, 4) = P(1 \mid 2, 4)P(3 \mid 2, 4)$. The correspondence between $\mathcal{G}$ and $\mathcal{G}^m$ is not 1–to–1 as $\mathcal{G}^m$ could arise from the moralization of an entire class of Markov equivalent reciprocal graphs. Further details about moralization in reciprocal graphs and covariance equivalence are discussed in Koster (1996) and Spirtes et al. (1998). Here, our inference will be based

on $\mathcal{G}$ only, and the directionality will be based on prior knowledge. The undirected graph $\mathcal{G}^m$ provides a convenient summary of the conditional independence structure if desired.

2.2 Dependent Gene Expression and Hidden Systems of Simultaneous Equations

Following Parmigiani et al. (2002), we consider data in the form of an $(p \times n)$ expression matrix $\mathbf{Y}$, with the generic element y_{ij} denoting the observed gene expression for gene i in sample j, $i = 1, \ldots, p$ and $j = 1, \ldots n$. We introduce latent variables $e_{ij} \in \{-1, 0, 1\}$ indexing three possible expression categories for each entry in $\mathbf{Y}$. For example, if $\mathbf{Y}$ represents ratios of expression level relative to a normal reference, they can be interpreted as high, normal and low. Given e_{ij}, for each gene i and each sample j we assume a mixture parameterized with $\theta = (\alpha_j, \mu_i, \kappa_i^-, \kappa_i^+)$ as follows:

$$p(y_{ij} - (\alpha_j + \mu_i)|e_{ij}) = f_{ij}(y_{ij}) \text{ with } \begin{cases} f_{-1i} &= U(-\kappa_i^-,\ 0) \\ f_{0i} &= N(0,\ \sigma_i^2) \\ f_{1i} &= U(0,\ \kappa_i^+). \end{cases} \quad (1)$$

In words, we assume that the observed expressions arise from a mixture of a Gaussian distribution and two uniform distributions designed to capture a broad range of departures relative to the Gaussian. The interpretation of the Gaussian component varies depending on the experimental design and sampling scheme, and can be trained in a supervised way if data are available (Garrett and Parmigiani 2004). When the technology used for measuring expression has an internal reference, as in Section 5, the high (low) class can be interpreted as over– (under) expression compared to the reference. By approximately specifying probabilities, one can collapse this model to one with binary indicators. In that case our strategy results in a boolean network defined on the latent e_{ij}'s.

In (1), α_j is a sample–specific effect, included to adjust for systematic variation across

samples; μ_i is a gene–specific effect, modeling the overall abundance of each gene, and κ_i^- and κ_i^+ parameterize the limits of variation in the tails. Finally σ_i^2 is the variance of the normal component of the distribution of gene i. We follow Parmigiani et al. (2002) in defining a conditionally conjugate prior for μ_i, σ_i^2 and κ_i^- and κ_i^+. Let $\mathcal{G}a(a,b)$ denote a Gamma distribution with expectation a/b:

$$\begin{aligned} p(\mu_i \,|\, m_\mu, \tau_\mu) &= N(m_\mu, \tau_\mu), & p(1/\sigma_i^2 \,|\, \gamma_\sigma, \lambda_\sigma) &= \mathcal{G}a(\gamma_\sigma, \lambda_\sigma), \\ p(1/\kappa_i^- \,|\, \gamma_\kappa^-, \lambda_\kappa^-) &= \mathcal{G}a(\gamma_\kappa^-, \lambda_\kappa^-), & p(1/\kappa_i^+ \,|\, \gamma_\kappa^+, \lambda_\kappa^+) &= \mathcal{G}a(\gamma_\kappa^+, \lambda_\kappa^+); \end{aligned}$$

where $min(\kappa_i^+, \kappa_i^-) > \kappa_0 \sigma_i$ and $\kappa_0 = 5$. The restriction on κ_i^- and κ_i^+ ensures that the gene-specific mixture distribution has heavier tails than its normal component, preserving interpretability of the three-way latent classes. For the sample–specific effect α_j, we impose an identifiability constraint $\alpha_j \sim N(0, \tau_\alpha^2)$ with $\sum_{j=0}^{n} \alpha_t = 0$.

Specifying a prior model for e_{ij} we deviate from Parmigiani et al. (2002), defining the model in terms of latent normal variables (Albert and Chib 1993). For each gene and sample we introduce a latent Gaussian variable z_{ij}, and define:

$$e_{ij} = \begin{cases} 1 & \text{if } z_{ij} > 1 & \text{high epression} \\ 0 & \text{if } -1 < z_{ij} \leq 1 & \text{normal expression} \\ -1 & \text{if } z_{ij} \leq -1 & \text{low expression} \end{cases} \tag{2}$$

where the distribution of z_{ij} is defined by the following system of simultaneous equation model (SEM):

$$z_{ij} = m_{ij} + \sum_{k \neq i} \beta_{ik}(z_{kj} - m_{kj}) + \epsilon_{ij}, \quad i = 1, ..., p, \quad j = 1, ..., n; \tag{3}$$

with $\epsilon_{ij} \sim N(0, s_i^2)$. Let $\mathbf{Z}_j = z_{1j}, ..., z_{pj}$, denote the p–dimensional vector of latent probit scores, associated with sample j. Also, let $\mathbf{B}$ be the $(p \times p)$ matrix whose diagonal elements are unity and whose off-diagonal (i,k) components is $-\beta_{ik}$. Provided $\mathbf{B}$ is non–singular, the process above defines a proper joint probability density function (Besag 1974). More

precisely, defining the marginal precision matrix $\mathbf{H}_z = diag(1/s_1, ..., 1/s_p)$ and $\mathbf{\Omega} = \mathbf{B}'\mathbf{H_z}\mathbf{B}$, we have

$$P(\mathbf{Z}_j \mid \mathbf{m}_j, \mathbf{\Omega}) = \frac{|\mathbf{\Omega}|^{1/2}}{(2\pi)^{1/2}} \exp\{-\frac{1}{2}(\mathbf{Z}_j - \mathbf{m}_j)'\mathbf{\Omega}(\mathbf{Z}_j - \mathbf{m}_j)\} \tag{4}$$

where $\mathbf{m}_j = (m_{1j}, ..., m_{pj})'$.

If $\mathbf{e}_j = (e_{1j}, ..., e_{pj})'$, the implied probabilities for the indicators e_{ij} are:

$$P(\mathbf{e}_j \mid \mathbf{m}_j, \mathbf{\Omega}) = \int_{A_{pj}} \cdots \int_{A_{1j}} P(\mathbf{Z}_j \mid \mathbf{m}_j, \mathbf{\Omega})\, d\mathbf{Z}_j, \tag{5}$$

where A_{ij} is the interval $(-\infty, -1]$ if $e_{ij} = -1$, $(-1, 1]$ if $e_{ij} = 0$ and $(1, \infty)$ if $e_{ij} = 1$. We use notations $\pi^+_{ij} = p(z_{ij} > 1 \mid y)$, $\pi^-_{ij} = p(z_{ij} < -1 \mid y)$ and $p^\star_{ij} = \pi^+_{ij} - \pi^-_{ij}$.

In this context, we propose to use a reciprocal graph, $\mathcal{G} = \{V, E\}$, to describe a dependence structure among the three-way indicators e_{ij} that reflects a priori knowledge about a pathway. Relationships between genes are captured via a set of conditional independences over the joint distribution of the classes $\mathbf{e}_j = (e_{ij},\ i = 1, ..., p)$. This is implemented by structuring the matrix $\mathbf{B}$ so that the off–diagonal element (i, k) is null ($\beta_{ik} = 0$), if and only if the edge $k \to i$ is not in $\{E\}$, (Spirtes et al. 1998). The resulting concentration matrix $\mathbf{\Omega} = \mathbf{B}'\mathbf{H_z}\mathbf{B}$, will have zero off diagonal elements ($\omega_{ik} = 0$), structured compatibly with the Markov properties encoded in the moral graph $\mathcal{G}^m = \{V, E^m\}$, (Koster 1996).

For each gene and sample, the mean m_{ij} may be modeled as a linear function ($m_{ij} = \mathbf{x}'_j b_i$) of, say, a design vector $\mathbf{x}_j$. This allows for comparisons across groups. For example, if $\mathbf{x}_j = 1$ and -1 for samples under two different biologic conditions, then the posterior distribution for b_i formalizes difference on the differential expression of gene i under the two conditions, adjusting for the dependence among the genes.

Finally, the autoregressive scheme in (3), implicitly assumes that genetic interactions are invariant across all the cross-sample biological variation represented in the study. Relaxing

this assumption is important and can be achieved by including an interaction term relating the covariate or phenotype information in $\mathbf{x}_t$ with the neighboring probit scores $\mathbf{z_{kj}}$ in (3).

In summary, we assume a mixture model for the observed gene expressions y_{ij}. The noisy data y_{ij} is reduced to latent trinary indicators which are used to define the dependence structure. Because of the nonlinear shrinkage induced by the mixture model, the y_{ij} do not come from a multivariate normal, and the patterns of dependence could be more complex.

2.3 Priors over graphical structures and dependence parameters

We define a prior probability model for the dependence structure $\mathcal{G}$. In words, the prior is based on a pathway diagram, that summarizes substantive prior information about the pathway of interest. We interpret the pathway as a reciprocal graph $\mathcal{G}_0 = \{V, E_0\}$, (See example in Section 2.1). The prior on $\mathcal{G}$ is defined on the set of all graphs that can be obtained by deleting edges from $\mathcal{G}_0$. More formally, we define the model space generated by $\mathcal{G}_0$ as $M(\mathcal{G}_0) = \{\mathcal{G} = (V, E) : E \subset E_0\}$. If E_0 comprises a total number of K edges, then $M(\mathcal{G}_0)$ includes $D = 2^K$ possible models.

The definition of the the prior $p(\mathcal{G})$ can be seen as stating joint probabilities for the multiple hypothesis testing problem implicitely defined by inclusion versus exclusion of all possible edges. Following the standard Bayesian variable selection scheme (George and McCulloch 1993, Brown and Vannucci 1998, Dobra et al. 2004), we can consider edge inclusions as exchangeable Bernoulli trials with common inclusion probability φ. If $|E_0|$ is the total number of possible edges and $k_{\mathcal{G}}$ is the number of edges included in $\mathcal{G}$, it follows that $P(\mathcal{G} \mid \varphi) = \varphi^{k_{\mathcal{G}}}(1-\varphi)^{|E_0|-k_{\mathcal{G}}}$. When the inclusion probability φ comes from the Beta family ($\varphi \sim \mathcal{B}(a_\varphi, b_\varphi)$), Scott and Berger (2006) and Carvalho and Scott (2009) show that this class of prior model probabilities yield a strong control over the number of "false" edges included in $\mathcal{G}$.

A key feature of the proposed prior is the restriction to subsets of $\mathcal{G}_0$. Inference under the proposed model populates existing pathways with probabilistic information associated with a biological system at a temporal cross section of its dynamic. The restriction to $M(\mathcal{G}_0)$ is important to keep MCMC posterior simulation across the model space practicable. For global searches, without restriction to a focused set of models, trans–dimensional MCMC becomes impracticable. Local focus does not preclude some extensions beyond $M(\mathcal{G}_0)$ to facilitate discovery of previously unknown interactions. For example, consider an arbitrary graph $\mathcal{G}$, without restriction to $M(\mathcal{G}_0)$, and let $m_{\mathcal{G}}$ denote the number of deleted *and* added edges relative to $\mathcal{G}_0$. One could replace $k_{\mathcal{G}}$ in the prior by $m_{\mathcal{G}}$ and allow for graphs beyond $\mathcal{G}_0$. Little would change in the proposed inference. But centering on models close to $\mathcal{G}_0$ is important. See also related comments in Section 6.

Our model is completed defining priors over the non-zero parameters $\beta_{ij} \sim N(0, \sigma_\beta^2)$ $(i, j = 1, ..., p)$. This defines a conjugate prior for the normal model (3). This formulation is derived as a natural characterization of the SEM in (3).

3 ESTIMATION AND INFERENCE

3.1 **Model Determination via RJ–MCMC**

We implement posterior inference for $(\boldsymbol{\theta}, \mathbf{B}, \mathcal{G})$ by setting up posterior MCMC simulation. We define the current state $x = (\boldsymbol{\theta}, \mathbf{B}, \mathcal{G})$ as the complete set of unknowns and write $\pi(dx)$ short for the target posterior distribution $p(\boldsymbol{\theta}, \mathbf{B}, \mathcal{G} \mid \mathbf{Y})$.

The MCMC is defined by the following transition probabilities: (a) Update the parameter vector $(\boldsymbol{\theta}, \mathbf{B})$; and (b) Update $\mathcal{G}$ ensuring that the proposed graph $\mathcal{G}'$ is in the set $M\{\mathcal{G}_0\}$. This move usually involves changes to $\mathbf{B}$ as well.

The updates in (a) follow the usual M-H scheme. More care is needed for the updates in

(b) as they involve adding or deleting an edge in $\mathcal{G}$, therefore changing the dimension of the parameter space. We implement a reversible jump MCMC (RJ), Green (1995).

i) Draw an edge $(k \to i)$ at random from E_0. If in the current state $\mathcal{G}$, $(k \to i) \notin E$ propose the birth of the new edge $k \to i$. If $(k \to i) \in E$ propose the death of $k \to i$.

ii) If we propose the birth $k \to i$, the structural matrix $\mathbf{B}$ gets populated with a new element $\beta'_{ik} = u$, where $u \sim q(u)$. If we propose a the death of the edge $k \to i$, we simply set $\beta'_{ik} = 0$.

Steps i) and ii) generate a candidate $x' = (\mathbf{B}', \mathcal{G}')$. Let $m =$ index the move proposed in step i), and let m' index the reverse move. The acceptance probability is (Green, 1995)

$$R(x, x') = \min\left\{1, \frac{\pi(dx')}{\pi(dx)} \frac{q(m' \mid x')}{q(m \mid x)\, q(u)}\right\}, \tag{6}$$

where $q(m \mid x)$ is the probability of proposing move m when the chain is in state x, $q(u)$ is the density function of u. In general $R(x, x')$ might incluce an additional factor involving the Jacobian of a possible (deterministic) transformation of (x, u) to define x'. The described RJ involves no such transformation. The move m is generated in step i) by a uniform draw from E_0, implying $q(m \mid x) = q(m' \mid x')$. Finally, $q(u)$ is the proposal p.d.f. The acceptance probability of a birth R_b is then defined as:

$$R_b = min\left\{1; \frac{p(x' \mid \mathbf{Y})}{p(x \mid \mathbf{Y})} q(u)^{-1}\right\}.$$

If the proposed element β'_{ik} of $\mathbf{B}'$ defines a a singular matrix, $\mathbf{\Omega}$ is not positive definite and we reject move m' setting R_b to zero. Given this sampling scheme, the probability of a deletion is simply defined as $R_d = 1/R_b$.

3.2 Graphical Model Selection

The posterior probability $p(\mathcal{G}, \mathbf{B} \mid \mathbf{Y})$ and the corresponding MCMC posterior simulation characterize our knowledge about the pathway in the light of the data. Based on this posterior probability, we may be interested in selecting a representative graph $\mathcal{G}$. The posterior only summarizes the evidence for each $\mathcal{G}$. It does not yet tell us which $\mathcal{G}$'s we should finally report.

This model selection problem has been discussed by different authors. Drton and Perlman (2007) discuss graphical model selection from the frequentist perspective, under the assumption that $n \geq p+1$, while Jones et al. (2005) or Meinshausen and Bühlmann (2006), describe selection techniques for problems where the sample size n is small when compared to the number of variables p. From a Bayesian perspective, Carvalho and Scott (2009) provide a comprehensive discussion of Objective Bayesian model selection in Gaussian Graphical Models.

In the context of the model described in Section 2, graphical model selection can be defined by removing elements $(k \rightarrow i) \in E_0$ specified by the prior graph $\mathcal{G}_0 = \{V, E_0\}$. This is equivalent to the vanishing of the structural parameters β_{ik} in the matrix $\mathbf{B}$, characterizing the joint distribution of latent probit scores $\mathbf{Z}$ (Ronning and Kukuk 1996). If the edge set E_0 has size $|E_0| = Q$, graphical model selection involves testing Q hypothesis

$$H_q^0 : \beta_{(q)} = 0, \quad \text{vs.} \quad H_q^1 : \beta_{(q)} \neq 0, \qquad \text{for } q = 1, ..., Q.$$

When testing a large number of hypotheses it is important to address possible multiplicity problems by controlling some pre–defined error rate. A popular choice is to control the False Discovery Rate (FDR) (Benjamini and Hochberg 1995). Several authors (Carvalho and Scott 2009, Scott and Carvalho 2008, Scott and Berger 2006) suggest considering the shrinkage prior defined in Section 3.1 and report how including edges with inclusion probability $P(\beta_{ik} \neq$

$0) > 0.5$ (median model), yields strong control over the number of false positives.

4 SIMULATION STUDY

We validate and illustrate the proposed method with a simulation study with $p = 50$ genes form $n = 30$ samples. We define $\mathbf{Y}$ as the $(p \times n)$ matrix of simulated mRNA intensities and consider a balanced design where 15 columns of $\mathbf{Y}$ are form "normal" samples and 15 columns of $\mathbf{Y}$ are associated with "tumor" samples. Thus $\mathbf{x_{ij}} = (1, 0)'$ if y_{ij} is a normal sample and $\mathbf{x_{ij}} = (1, 1)'$ if y_{ij} is a tumor sample.

We generate simulated data $\mathbf{Y}$ as follows. Given a set of latent scores $\mathbf{W} \sim \mathcal{MN}(\mathbf{0}, \mathbf{\Sigma_z}, \mathbf{I_T})$, were $\mathbf{\Omega}_z = \mathbf{\Sigma}_z^{-1}$ encodes a known conditional dependence structure, and covariate effects $\mathbf{b_i} \sim N_2(\mathbf{m_i}, \sigma_\mathbf{b}^\mathbf{2} I_2)$, we define $z_{ij} = w_{ij} + \mathbf{x_{ij}}' \mathbf{b_i}$. We then generate the intensity matrix $\mathbf{Y}$ from a three-way mixture of Gaussian distributions:

$$\begin{aligned} y_{ij} \mid z_{ij} \leq -1 &\sim N(-4, 2^2), \\ y_{ij} \mid z_{ij} > 3 &\sim N(4, 2^2), \\ y_{ij} \mid -1 < z_{ij} \leq 3 &\sim N(0, 1). \end{aligned} \tag{7}$$

The precision matrix $\mathbf{\Omega}_z$ is defined as follows. First we obtain the $p \times p$ matrix $\mathbf{B}$ by defining $\gamma_{ij} =_d Ga(2, 1)$, $c_{ij} = \{-1, 1\}$ with $P(c_{ij} = 1) = 0.5$ and δ_0 a Dirac mass at 0, so that $\mathbf{B}_{ii} = 1$ (for $i = 1, ..., p$), and the off diagonal elements $\mathbf{B}_{ij} = \pi_0\, \delta_0 + (1 - \pi_0)\, c_{ij}\gamma_{ij}$. The simulation truth is deliberately chosen different from the assumed analysis model (1).

We then generate $\mathbf{\Omega}_z$ by rescaling $\mathbf{B}'\mathbf{B}$ to a correlation matrix. The simulation model (7) is deliberately different from the assumed analysis model, but still includes a meaningful notion of true dependence structure and strength.

We use a prior Graph $\mathcal{G}_0 = \{V, E_0\}$ spanned by the set of edges $E = E^* \cup \tilde{E}$, with E^* spanning the simulation truth of non–zero elements in $\mathbf{B}_{ij}$ (in our example $|E^*| = 50$) and

$\tilde{E}$ serving as a random mispecification set including false edges (in our example $|\tilde{E}| = 87$).

In Figure 3, we display the classification results for the expression measurements generated under the dependence schemes just described. We calculated posterior probabilities of over– and under–expression from 50,000 posterior samples (thinned by 10), obtained after conservatively discarding 50,000 iterations. Figure 3 *(left panel)* shows the simulation truths as indicators (e_{ij}) of over–(white), normal–(grey) and under–expression(black). The right panel reports a unidimensional summary of the probabilities of over– or under–expression ($p^*_{ij} = \pi^+_{ij} - \pi^-_{ij}$). The elements p^*_{ij} are defined in the $[-1, 1]$ scale and may be compared directly with the three-way indicators e_{gt}. We note that the p^* scale provides improved resolution over genes with signal and recovers well the generating truth.

Posterior inference includes a posterior distribution on the dependence structure. In Figure 4 *(left panel)* we report the number of edges included in the model by MCMC iteration, for two chains starting at opposite sides of the model saturation spectrum. Despite the size of the mispecification set $\tilde{E}$, the trans-dimensional Markov chains converge fairly rapidly towards models of size comparable to $|E^*| = 50$. In the same figure, marginalizing over all possible graphs $M\{\mathcal{G}_0\}$ we report the posterior expected SEM coefficients $E(\beta_{ik} \mid \mathbf{Y})$ and the edge inclusion probabilities $P(\beta_{ik} \neq 0 \mid \mathbf{Y})$ (*right panel*). In this plot, we report the false edges as solid circles. Most solid circles lie in the area below an inclusion probability of 0.5. This shows how the adopted probability scheme, not only penalizes for model complexity, but effectively controls the number of false discoveries, allowing for a genuine recovery of the generating conditional dependence structure.

5 CASE STUDY

Wang et al. (2005) report a study of epithelial ovarian cancer (EOC). The goal of the study is to characterize the role of the tumor microenvironment in favoring the intra–peritoneal

spread of EOC. To this end the investigators collected tissue samples form patients with benign (b) and malignant (m) ovarian pathology. Specimens were collected, among other sites, from peritoneum adjacent to the primary tumor. RNA was co-hybridized with reference RNA to a custom made cDNA microarray including combination of the Research genetics RG_HsKG_031901 8k clone set and 9,000 clones selected from RG_Hs_seq_ver_070700. A complete list of genes is available at `http://nciarray.nci.nih.gov/gal_files/index.shtml`, 'custom printings'. See the array labeled Hs_CCDTM–17.5k–1px.

In the following discussion we focus on the comparison of 10 peritoneal samples from patients with benign ovarian pathology (bPT) versus 14 samples from patients with malignant ovarian pathology (mPT). The raw data was processed using BRB ArrayTool (`http://linus.nci.nih.gov/BRB-ArrayTools.html`). In particular, spots with minimum intensity less than 300 in both fluorescence channels were excluded from further analysis. See Wang et al. (2005) for a detailed description.

One subset of genes reported on the NIH custom microarray are 61 genes in the coagulation and complement pathway from KEGG (`http://www.genome.ad.jp`), shown in Figure 2. Genes on this pathway are of interest for their role in the inflammatory process. The arches in the pathway are interpreted as prior judgement about (approximate) conditional dependence (Section 2.1). However, recognizing that the pathway represents a protein system rather than gene expression, we allow for significant deviation from this structure, explicitly including model determination in our analysis.

We fit the model presented in Section 2 to this set of 61 genes. The prior set of conditional dependences between genes is represented as a reciprocal graph in Figure 2 and includes a set of 148 possible edges. Reported inference is based on 50,000 MCMC samples, thinned by 10, after discarding 50,000 observation for burn–in.

Recording the number of times the sampler visits a particular edge we calculate the

posterior probability $v_{ik} = P(\beta_{ik} \mid \mathbf{Y})$, for each edge $(k \longrightarrow i)$ in the prior graph $\mathcal{G}_0$. In Figure 5, we show the set of selected genetic interactions when we consider edges with inclusion probabilities greater than 0.5 (median model). Edge directionality is inherited from $\mathcal{G}_0$ (Figure 2).

The posterior distribution on $\mathbf{e_g}$ provides inference on differential expression, appropriately adjusted for dependence. Starting from the Complement and Coagulation Cascade pathway, we identify a set of 24 genes exhibiting patterns of dependence in their differential expression profiles across healty and tumor tissues. In order to give an interpretation to our findings, we searched the scientific literature using the Information Hyperlinked Over Protein (IHOP) tool implemented by Hoffman (Hoffman and Valencia 2004), available at : `www.ihop-net.org`.

For example, our study confirms the centrality of the peptide IL8 (Intelukin-8) in the regulation of the chemokine (CXC and CC motifs) genes. The protein encoded by this gene has been reported by several authors to play an important role in the response to inflammatory stimuli, resistance to apoptosis and tumoral angiogenesis. See Terranova and Rice (1997)) or Brat et al. (2005), for comprehensive discussions on IL8 and its receptors.

One other example is the finding of dependent expression profiles associated with the Thrombine pathway (F2$\longrightarrow$F2R and F2$\longrightarrow$THBD). This patwhay plays a central role in the coagulation cascade and has been reported as a potential mediator of cellular function in the ovarian follicle (Roach et al. 2002).

6 DISCUSSION

We propose a probability model for the analysis of dependent gene expression data. Dependence between genes is modeled via the explicit consideration of prior information from pathways representing known biochemical processes. We characterize a biochemical pathway

as a reciprocal graph depicting a coherent set of conditional dependence relationships between three-way classes of gene under-, normal- and over–expression. Modeling dependence between latent indicators of class membership, is likely to represent a more sensible approach, for this kind of data, when compared with methods that model correlations between observables directly. Acknowledging that a known pathway represents only prior information, we seek posterior inference for the model parameters as well as for the pathway itself via an RJ-MCMC scheme. We showed, through simulation studies, that our model allows for the recovery of the true dependence structure, even under a misspecified prior pathway.

Our model of mRNA abundance relies on the Probability of Expression (POE) Model of Parmigiani et al. (2002), and assumes that the variability of expression across tissue samples can be fully characterized by heavy tailed mixtures of Normal and Uniform random variables. While this is a simplification of reality, it contributes to denoising data and is likely to provide useful summaries, allowing for the investigation of the many aspects associated with expression data analysis, from data normalization, to DE analysis, to the characterization of molecular profiles. The general framework presented in this article is also adaptable to other models of gene expression analysis.

In the construction of the dependent probability model, it is important to acknowledge the limitations of the information provided in a biochemical network. In fact, a pathway may not necessarily describe relations among transcript levels, although it carries some information about it. On a related note, the proposed methodology is currently restricted to known biochemical pathways. Our model could be extended to discover novel genetic interactions, by allowing adding new edges between nodes in the prior graph $\mathcal{G}_0$. This, however, would come at a substantial computational cost and would require a challenging reformulation of the prior over graphs $p(\mathcal{G})$, to penalize for model complexity and, at the same time, to favor models closer to the structure of the prior pathway $\mathcal{G}_0$. Initial progress in this direction was

reported by Braun et al. (2008) and, in the context of Bayesian Networks, by Mukherjee and Speed (2008).

In this article we model dependence between three-way variables as dependence between latent Gaussian quantities. This is only a convenient restriction on the possible shapes of dependence characterizing a matrix of ordinal random variables. Extensions of our model considering a richer class of dependence structures are, in principle, appealing. However these would require a higher level of complexity and limitations on the clique size contributing to the joint distribution of the three-way indicators.

References

Albert, J. H. and S. Chib (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association 88*, 669–679.

Beal, M., F. Falciani, Z. Ghahramani, C. Rangel, and D. Wild (2005). A Bayesian approach to reconstructing genetic regulatory networks with hidden factors. *Bioinformatics 21*, 349–356.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B 57*, 289–300.

Besag, J. (1974). Spatial interections and the statistical analysis of lattice systems. *JRSS B*, 302–339.

Brat, D. J., A. C. Bellail, and G. V. M. Erwin (2005). The role of interlukin-8 and its receptors in gliomagenesis and tumoral angiogenesis. *Neuro-Oncology 7*, 122–133.

Braun, R., L. Cope, and G. Parmigiani (2008). Identigying differential correlation in gene/pathway combinations. *BMC Bioinformatics 9*, 488.

Brown, P. J. and M. Vannucci (1998). Multivariate Bayesian model selection and prediction. *Journal of the Royal Statistical Society, Series B 60*(3), 627–641.

Carvalho, C. M. and J. G. Scott (2009). Objective Bayesian model selection in gaussian graphical models. *Biometrika 96*(3), 497.

Dawid, A. P. and S. L. Lauritzen (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *Annals of Statistics 3*, 1272–1317.

Dobra, A., C. Hans, B. Jones, G. Yao, and M. West (2004). Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis 90*, 196–212.

Drton, M. and M. D. Perlman (2007). Multiple testing and error control in Gaussian graphical model selection. *Statistical Science 22*, 430 – 449.

Friedman, N., M. Linial, I. Nachman, and D. Pe'er (2000). Using Bayesian networks to analyze expression data. *Journal of Computational Biology 7*, 601–620.

Garrett, E. S. and G. Parmigiani (2004). A nested unsupervised approach to identifying novel molecular subtypes. *Bernoulli 10*(6), 951–969.

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

Giudici, P. and P. J. Green (1999). Decomposable graphical Gaussian model determination. *Biometrika 86 (4)*, 785–801.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika 82 (4)*, 711–732.

Hoffman, R. and A. Valencia (2004). A gene network for navigating the literature. *Nature Genetics 36*, 664–664.

Jones, B., C. Carvalho, A. Dobra, C. Hans, and M. West (2005). Experiments in stochastic computation for high–dimensional graphical models. *Statistical Science 20*, 388–400.

Kolaczyk, E. D. (2009). *Statistical Analysis of Network Data: Methods and Models*. New York: Springer.

Koster, J. T. A. (1996). Markov properties of non recursive causal models. *Annals of Statistics 24*, 2148–2177.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford: Clarendon.

Meinshausen, N. and P. Bühlmann (2006). High dimensional graphs and variable selection with the lasso. *Annals of Statistics 34*, 1436–1462.

Mukherjee, S. and T. P. Speed (2008). Netrwork inference using informative priors. *PNAS 105*(38), 14133–14318.

Murphy, K. and S. Mian (1999). Modeling gene expression data using dynamic Bayesian networksayesian networks. *Technical Report, Computer Science Division, UC Berkley*.

Ong, I., J. Glasner, and D. Page (2002). Modelling regulatory pathways in e.coli from time series expression profiles. *Bioinformatics 18*, S241–S248.

Parmigiani, G., E. S. Garrett, R. Anbazhagan, , and E. Gabrielson (2002). A statistical framework for expression–based molecular classification in cancer (with discussion). *JRSS B 64*, 717–736.

Roach, L. E., J. J. Petrik, L. Plante, and J. LaMarre (2002). Thrombin generation and presence of thrombin in ovarian follicles. *Biology of reproduction 66*, 1350–1358.

Ronning, G. and M. Kukuk (1996). Efficient estimation of ordered probit models. *Journal of The American Statistical Association 97*, 1122–1140.

Scott, J. and C. M. Carvalho (2008). Feature-inclusion stochastic search for Gaussian graphical models. *Journal of Computational and Graphical Statistics 17*, 790–808.

Scott, J. G. and J. O. Berger (2006). An exploration of aspects of Bayesian multiple testing. *Journal of Computational and Graphical Statistics 136*(7), 2144:2162.

Sebastiani, P. (2005). Normative selection of Bayesian networks. *Journal of Multivariate Analysis 93*, 340–357.

Spirtes, P., T. S. Richardson, C. Meek, R. Scheines, and C. Glymour (1998). Using path diagrams as a structural equation modeling tool. *Sociological Methods & Research 27*(2), 182–225.

Terranova, P. F. and V. M. Rice (1997). Review: cytokine involvement in ovarian processes. *American Journal of Reproductive Immunology 37*, 50–63.

Wang, X., E. Wang, and J. Kavanagh (2005). Ovarian cancer, the coagulation pathway, and inflammation. *Journal of Translational Medicine*, 3–25.

Wei, Z. and H. Li (2007). A Markov random field model for network–based analysis of genomic data. *Bioinformatics 23*, 1357–1544.

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

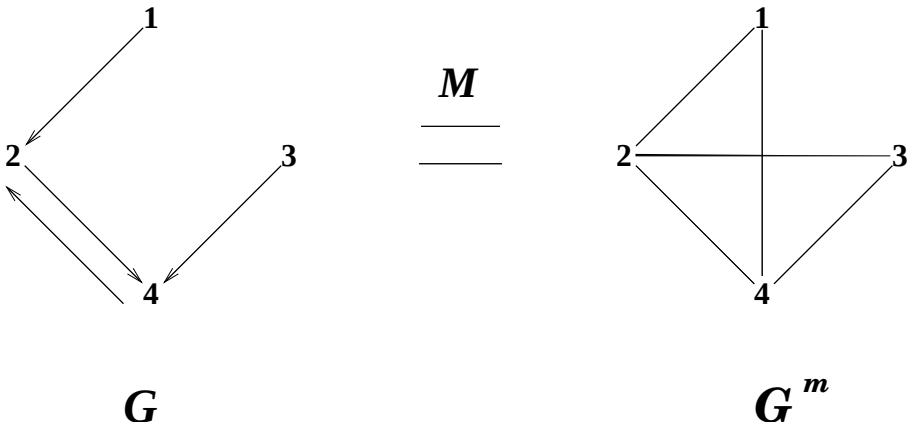

Figure 1: (Example) moralization of a reciprocal graph.

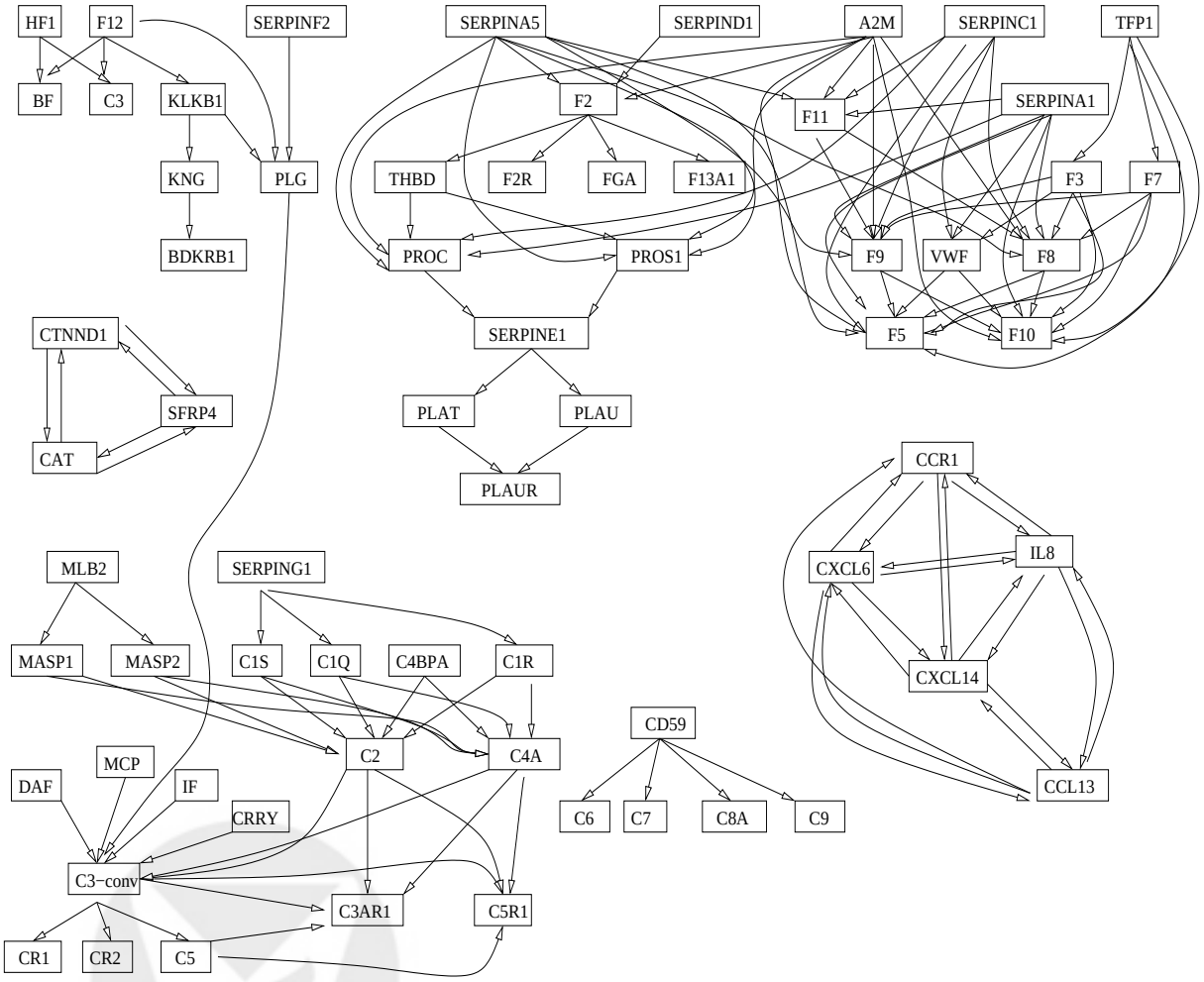

Figure 2: Complement and coagulation cascades pathway (Wang et al. 2005).

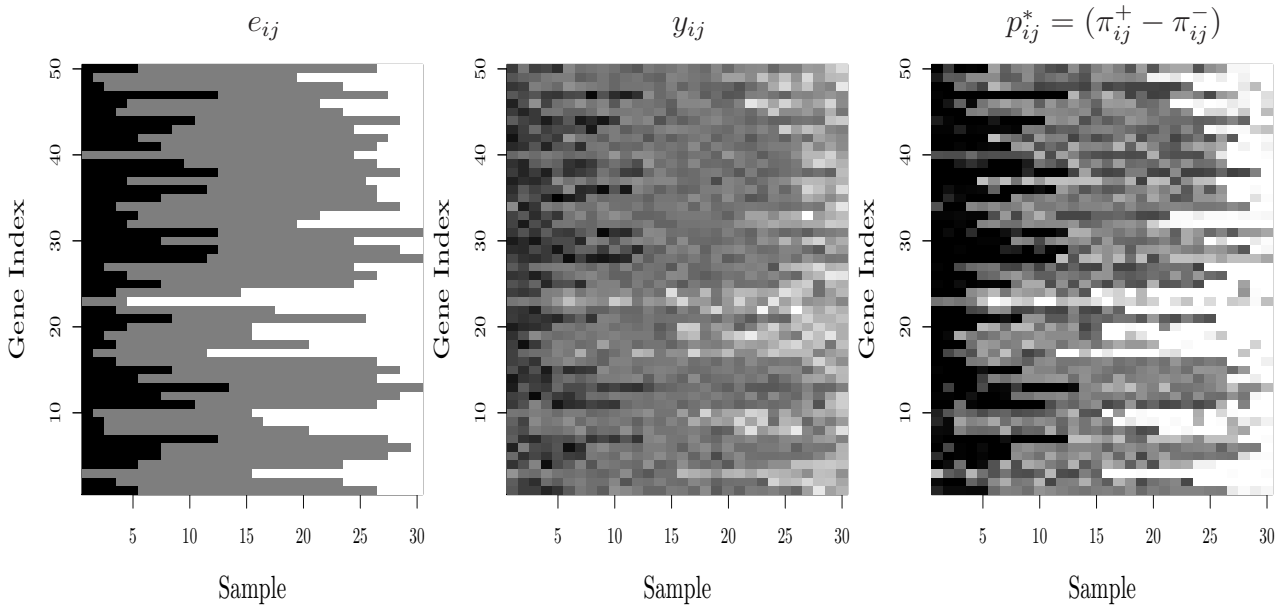

Figure 3: Simulation Study: *(Left Panel)* Simulation signal e_{ij}. *(Central Panel)* Simulated mRNA abundance y_{ij}. *(Right Panel)* DepPOE estimate of p^*_{ij}.

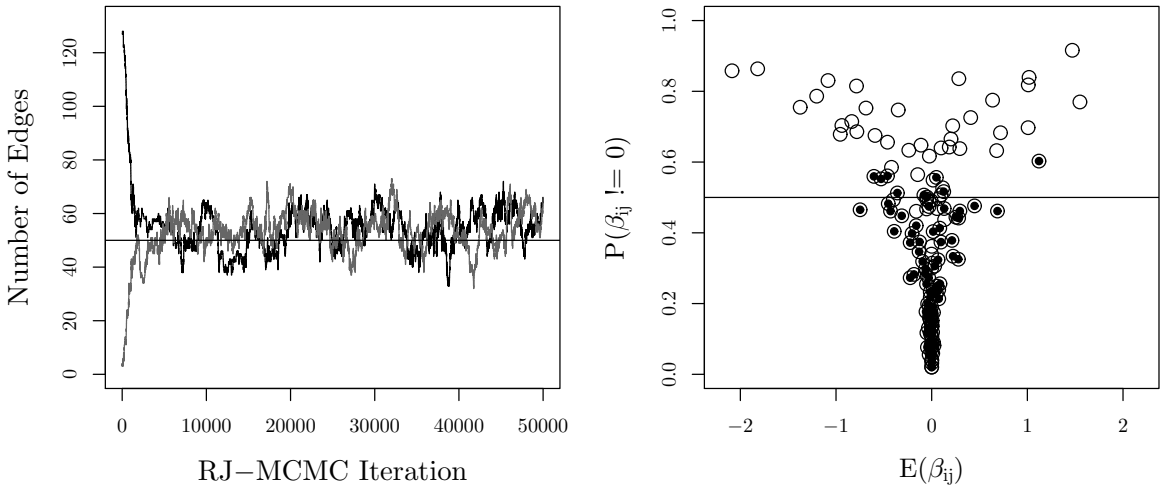

Figure 4: Simulation Study: (Left Panel) Number of edges included in the model by MCMC iteration, for two chains with starting points at the two extremes of the saturation spectrum. (Right Panel) Posterior expected SEM coefficients $E(\beta_{ik} \mid \mathbf{Y})$ Vs. posterior inclusion probabilities $P(\beta_{ik} \neq 0)$. False edges are represented with a solid circle.

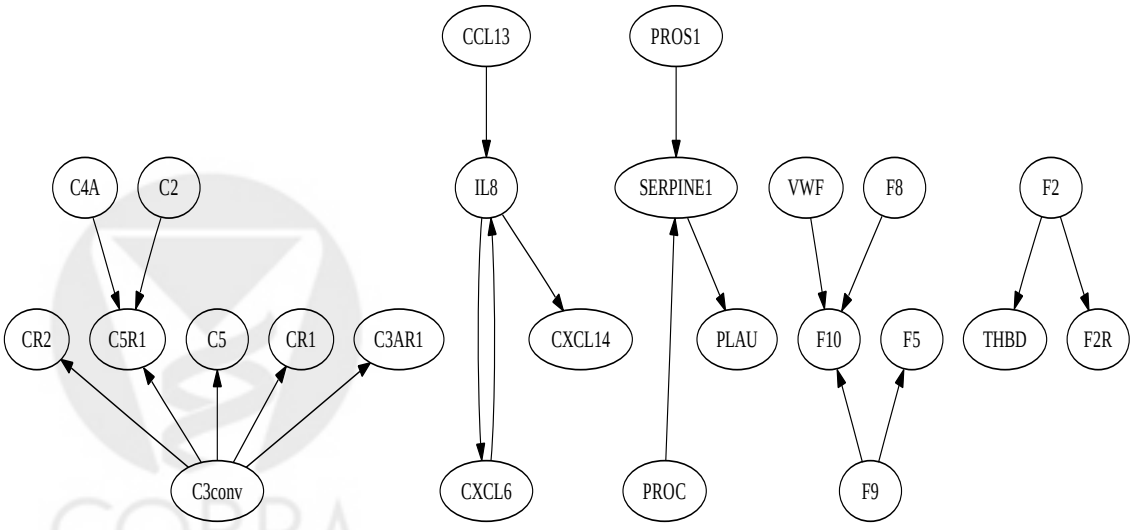

Figure 5: Case Study. Posterior pathway obtained selecting edges with inclusion probabilities greater than 0.5 (Median model).