Reciprocal Graphical Models for Integrative Gene Regulatory Network Analysis

Yang Ni[*], Yuan Ji[†], and Peter Müller[‡]

Abstract. Constructing gene regulatory networks is a fundamental task in systems biology. We introduce a Gaussian reciprocal graphical model for inference about gene regulatory relationships by integrating messenger ribonucleic acid (mRNA) gene expression and deoxyribonucleic acid (DNA) level information including copy number and methylation. Data integration allows for inference on the directionality of certain regulatory relationships, which would be otherwise indistinguishable due to Markov equivalence. Efficient inference is developed based on simultaneous equation models. Bayesian model selection techniques are adopted to estimate the graph structure. We illustrate our approach by simulations and application in colon adenocarcinoma pathway analysis.

Keywords: simultaneous equation models, Markov equivalence, directed cycles, feedback loop, multimodal genomic data.

1 Introduction

In this paper, we develop a reciprocal graphical model (RGM) to infer gene regulatory relationships and gene networks. This includes in particular directed edges without time course or interventional data. RGMs allow for undirected edges, directed edges and directed cycles and therefore are ideally suited for modeling regulatory relationships including feedback loops. Exploiting genomic data from multiple modalities/platforms, we are able to determine the directionality of certain regulatory relationships, which would be otherwise indistinguishable due to Markov equivalence. Such inference about directionality becomes possible because basic biology fixes the directionality for some edges, for example, between DNA methylation and gene expression of the same gene. Conditioning on such known directionality enables us to infer directionality for other edges. Statistically, the class of probability models determined by RGMs is strictly larger than the class of probability models determined by directed acyclic graphs (DAGs) and Markov random fields (MRFs). Computationally, the connection of RGMs with simultaneous equation models (SEMs) facilitates computation-efficient implementation of full posterior inference.

Most recent graphical model approaches in biostatistics and bioinformatics are restricted to DAGs (Stingo et al., 2010; Yajima et al., 2015; Ni et al., 2015) and MRFs (Wang and West, 2009; Dobra et al., 2012; Green and Thomas, 2013; Mitra et al., 2013; Wang et al., 2013). These approaches use the conditional independence struc-

[*]Department of Statistics and Data Sciences, The University of Texas at Austin, yangni87@gmail.com

[†]Program for Computational Genomics & Medicine, NorthShore University HealthSystem, Department of Public Health Sciences, The University of Chicago, koaeraser@gmail.com

[‡]Department of Mathematics, The University of Texas at Austin, pmueller@math.utexas.edu

ture represented by the graphical models. The popularity of DAGs and MRFs is due to their mathematical tractability and easy computation, despite some inherent limitations. MRFs model undirected relationships between genes and do not account for the directionality of edges. However, biological interactions between genes are often asymmetric. For example, it is only possible for a regulator to regulate its targets, not vice versa. DAGs allow for directed edges but only as arbitrary factorization of a joint probability model. Also, DAGs explicitly prohibit directed cycles. However, feedback loops are quite common motifs and have key functional roles in many cellular processes such as regulating gene expressions and acting as bistable switches (Shin et al., 2010).

RGMs were first proposed by Koster (1996) but remain curiously under-used in the biostatistics and bioinformatics literature, with few exceptions. Zhang et al. (2005) developed a hierarchical RGM to discover the relationship between cholesterol levels and pulmonary function in a longitudinal study. The graph structure is fixed and a generalized EM algorithm was developed to estimate the parameters. Telesca et al. (2012a,b) use RGMs for latent variables that represent active/inactive proteins and differential vs. non-differential gene expression, respectively. However, the use of the RGM is restricted to representation and for convenient summaries. The actual inference model is based on the implied conditional independence structure only (after moralization). Also, they use RGMs with only directed edges, excluding possible undirected edges. By contrast, our use of RGMs is directly on the observed gene expressions and we use the RGM to build the probability model in a way that allows us to infer direction for directed edges.

Because different RGMs may determine equivalent conditional independence structure, inference algorithms that use only the implied conditional independence structure can not possibly distinguish such equivalent graphs based on observational data. In particular, the directionality of some regulatory relationships cannot be determined (examples are given in Section 2.2). In this paper, we integrate different sources of genomic information including DNA copy number and DNA methylation in a way that allows us to determine the direction of some edges from first biological principles. This allows us then to identify the gene regulatory networks that best fit the data, including inference on edge directions in some cases. That is, the additional genetic and epigenetic information together with fundamental biological knowledge can inform the directionality of the regulatory relationships between genes. Integrating multimodal data in constructing networks has been explored by other authors as well (Cai et al., 2013; Zhang and Kim, 2014; Oates et al., 2016). In addition, covariate-adjusted Gaussian graphical models (Rothman et al., 2010; Cai et al., 2012; Bhadra and Mallick, 2013; Chen et al., 2016; Kundu and Kang, 2016) also allow for integration of multimodal data. The relationships between different modalities are modeled by a multivariate regression with errors following a Gaussian graphical model. By contrast, in the upcoming discussion, all modalities are modeled within the same RGM framework.

The proposed approach is motivated by a genomic study of colon adenocarcinoma (COAD). Our goal is to reconstruct a gene regulatory network with genes from the RAS-MAPK pathway which is critical in the initiation and progression of COAD (TCGA, 2012). Integrating copy number and methylation information, we find biologically meaningful gene interactions as well as novel regulatory relationships that need to be validated by further biological experiment.

The rest of the article is organized as follows. We develop the proposed model in Section 2. We present simulation studies in Section 3 and the application in Section 4. Section 5 provides our closing discussion.

2 Model

2.1 Notation

A graph $\mathcal{G} = (V, E)$ consists of a set of *vertices* $V = \{1, \ldots, p\}$ and a set of *edges* E connecting these vertices. We consider both *directed* and *undirected* edges $E = E^d \cup E^u$ with $E^d \subseteq \{(i, j) \mid i, j \in V\}$ and $E^u \subseteq \{\{i, j\} \mid i, j \in V\}$ where the ordered pair (i, j) denotes a directed edge from vertex i to vertex j and $\{i, j\}$ denotes an undirected edge. If $\{(i, j), (j, i)\} \subseteq E^d$, two directed edges are drawn between nodes i and j, which differs from some graphical model literature such as Frydenberg (1990) where $\{(i, j), (j, i)\} \subseteq E^d$ implies an undirected edge $\{i, j\} \in E^u$. This distinction is important as $\{(i, j), (j, i)\}$ and $\{i, j\}$ have different Markov properties and interpretations in RGMs. We use the vertices to index a set of random variables, $\boldsymbol{Y} = \boldsymbol{Y}_V = (Y_1, \ldots, Y_p)^T$. For example, in our application Y_j represents gene expressions for gene j.

A *path* of length K is an ordered sequence $(i_0, \ldots, i_K)$ of distinct vertices except possibly $i_0 = i_K$ such that $\{i_{k-1}, i_k\} \in E$ or $(i_{k-1}, i_k) \in E$ for $k = 1, \ldots, K$. A path is called undirected if $\{i_{k-1}, i_k\} \in E$ for all $k = 1, \ldots, K$. A *path component* is a set of vertices that are all connected by an undirected path. A *reciprocal graph* (RG) is a graph $\mathcal{G} = (V, E)$ such that there are no directed edges between vertices in the same path component (Koster, 1996). A *chain component*, a similar notion in chain graphs, is a special case of a path component in RGs with the distinction that the edges between two chain components in a chain graph must point in the same direction. Some examples of RG with four vertices are given in Figure 1a-1c. A violation of the definition of RG can be found in $1 \rightleftarrows 2$ where nodes 1 and 2 form a path component but there is a directed edges from 1 to 2. The *boundary* of a vertex i is $\mathrm{bd}(i) = \{j \mid \{j, i\} \in E \text{ or } (j, i) \in E\}$ and the *boundary* of a subset $V_0 \subseteq V$ is $\mathrm{bd}(V_0) = \bigcup_{i \in V_0} \mathrm{bd}(i) \backslash V_0$. An anterior set is a subset $V_0 \subseteq V$ such that $\mathrm{bd}(V_0) = \emptyset$ and the smallest anterior set containing V_0 is denoted by $\mathrm{an}(V_0)$. For example, $an(\{1\}) = \{1, 2, 3, 4\}$ in Figure 1a because $bd(\{1\}) = \{2, 3\}$, $bd(\{1, 2, 3\}) = \{4\}$ and $bd(\{1, 2, 3, 4\}) = \emptyset$. and $an(\{1\}) = \{1, 3\}$ in Figure 1b because $bd(\{1\}) = \{3\}$ and $bd(\{1, 3\}) = \emptyset$.

The Markov property (i.e. conditional independence relationships of $\boldsymbol{Y}$) of an RG relies on the notion of *moralization*. To moralize a graph $\mathcal{G}$, we connect all vertices in the boundary of each path component of $\mathcal{G}$ by undirected edges and then replace all directed edges by undirected edges. The resulting moral graph is an undirected graph and is denoted by $\mathcal{G}^m$. For example, the moral graphs of RGs in 1a, 1b and 1c are shown in Figure 1g, 1h and 1i. Later we will use graph *separation* to introduce a global Markov property. In an undirected graph, two sets V_1 and V_2 are said to be separated by a third set V_3 if every path between V_1 and V_2 intersects V_3. For instance, in Figure 1g, nodes 3 and 4 are separated by 1 and 2.
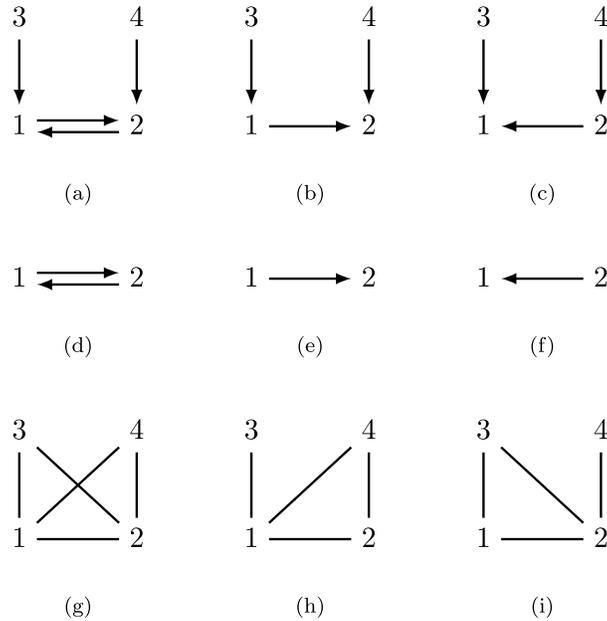
Figure 1: Graphs for illustration. Graphs in (a) – (c) are reciprocal graphs and imply different conditional independence. Graphs in (d) – (f) are Markov equivalent as they imply the same Markov properties. Graphs in (g) – (i) are moral graphs of (a) – (c). In our applications, we will use nodes 1 and 2 to represent gene expression of two genes and nodes 3 and 4 to represent copy number (or methylation) of the same genes.

2.2 Reciprocal graphical models, Markov equivalence and integrative genomics

A graphical model is a mapping between a family of distribution and an underlying graph. In this paper, we focus on the class of RGs which by definition strictly contains MRFs and DAGs as subclasses.

The probability distribution of $\boldsymbol{Y}$ is said to be (global) *Markov* with respect to an RG $\mathcal{G}$ if $\boldsymbol{Y}_{V_1} \perp\!\!\!\perp \boldsymbol{Y}_{V_2} \mid \boldsymbol{Y}_{V_3}$ whenever V_3 separates V_1 and V_2 in $\mathcal{G}^m_{an(V_1 \cup V_2 \cup V_3)}$. RGMs can represent global Markov properties beyond the conditional independence structure that is encoded in MRFs and DAGs. RGMs are a strictly larger class of probability model than MRFs and DAGs. For example, in Figure 1a, the only two conditional independence relationships are $3 \perp\!\!\!\perp 4$ and $3 \perp\!\!\!\perp 4 \mid 1, 2$, that is, the distribution of (Y_1, Y_2, Y_3, Y_4) can be factorized in two ways,

$$p(Y_1, Y_2, Y_3, Y_4) = p(Y_3|Y_1, Y_2)p(Y_4|Y_1, Y_2)p(Y_1, Y_2) = p(Y_1, Y_2|Y_3, Y_4)p(Y_3)p(Y_4).$$

There is no MRF or DAG that encodes the same conditional independence relationships. More importantly for our application, RGMs are particularly useful for the construction of genomic networks because of the ability to model feedback loops which are not allowed

by MRFs or DAGs. For example, in Figure 1a, gene 1 may regulate the expression of gene 2 while the status of gene 2 may reciprocally affect gene 1 through a feedback regulatory mechanism.

Two graphs are said to be *Markov equivalent* if they have the same Markov properties. For example, graphs 1d, 1e and 1f are Markov equivalent as they have the same Markov property, namely, no conditional independence assertion. Markov equivalence is indeed an equivalence relation and hence induces Markov equivalence class. Graphs within the same Markov equivalence class are usually nonidentifiable from observational data. In particular, observational data do not allow for inference on the direction of the edges in Figures 1d–1f.

However, with prior knowledge, sometimes we are able to distinguish the relationships between variables. In this paper, we introduce such prior knowledge by deliberately considering edges with (biologically) known direction if included. This allows for inference on other edges. We develop this approach based on integrating genomic information across different platforms and exploiting the central dogma of molecular biology that mRNA is produced by transcription from segments of DNA on which the copy number and methylation are measured, but the reverse processes are rare and biologically uninterpretable. For illustration, let vertices 1 and 2 in Figure 1 represent mRNA gene expressions and vertices 3 and 4 represent copy numbers of the corresponding genes and assume there exists dependence between copy number and gene expression for the same gene and hence by the central dogma, there are directed edges from 3 to 1 ($3 \rightarrow 1$) and from 4 to 2 ($4 \rightarrow 2$). Now we are able to fully identify the relationship between mRNAs 1 and 2 because each of graphs 1a, 1b and 1c defines a distinct set of conditional independence relationships. For example, in terms of pairwise Markov property, graph 1a implies $3 \perp\!\!\!\perp 4$ and $3 \perp\!\!\!\perp 4 \mid 1, 2$, graph 1b implies $1 \perp\!\!\!\perp 4$, $3 \perp\!\!\!\perp 4$, $2 \perp\!\!\!\perp 3 \mid 1$, $2 \perp\!\!\!\perp 3 \mid 1, 4$, $3 \perp\!\!\!\perp 4 \mid 1$, $3 \perp\!\!\!\perp 4 \mid 1, 2$ and $1 \perp\!\!\!\perp 4 \mid 3$ and graph 1c implies $2 \perp\!\!\!\perp 3$, $3 \perp\!\!\!\perp 4$, $1 \perp\!\!\!\perp 4 \mid 2$, $1 \perp\!\!\!\perp 4 \mid 2, 3$, $3 \perp\!\!\!\perp 4 \mid 2$, $3 \perp\!\!\!\perp 4 \mid 1, 2$ and $2 \perp\!\!\!\perp 3 \mid 4$. Since a network can be equivalently described by pairwise gene regulations, the argument above can be applied to the case with more than two genes. We remark that although we fix the direction of edges between DNA and mRNA, we do not force the inclusion of such edges. Therefore, if the data do not support the association between DNA and mRNA, the strategy described above is not sufficient to distinguish Markov equivalent graphs. In our application (Section 4), however, we do observe strong association between DNA copy number and mRNA gene expression.

2.3 Simultaneous equation models

For a formal description of our approach, we still need the mapping from the RGM to a family of probability models for the observational data. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_p)^T$ denote the mRNA gene expressions for genes $1, \ldots, p$. Let $\boldsymbol{X} = (X_1, \ldots, X_{2p})^T$ be the set of DNA level measurements for genes $1, \ldots, p$ with X_{2i-1} and X_{2i} being the copy number and the methylation for gene i, respectively. We first state the simultaneous equation model (SEM) and will then introduce the mapping. An SEM for $\boldsymbol{Y}$ and $\boldsymbol{X}$ is given by

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{Y} + \boldsymbol{B}\boldsymbol{X} + \boldsymbol{E}, \tag{1}$$

where $\boldsymbol{A} = (a_{ij}) \in \mathbb{R}^{p \times p}$ with zeros on the diagonal, $\boldsymbol{B} = (b_{ik}) \in \mathbb{R}^{p \times 2p}$, $\boldsymbol{E} = (\epsilon_1, \ldots, \epsilon_p)^T \sim \mathrm{N}_p(0, \boldsymbol{\Sigma})$ and $\boldsymbol{E}$ and $\boldsymbol{X}$ are independent. Let $\boldsymbol{I}_p$ denote a $p \times p$ identity matrix, model (1) can be equivalently expressed as

$$\boldsymbol{Y} \mid \boldsymbol{X} \quad \sim \quad \mathrm{N}_p \left\{ (\boldsymbol{I}_p - \boldsymbol{A})^{-1} \boldsymbol{B} \boldsymbol{X}, (\boldsymbol{I}_p - \boldsymbol{A})^{-1} \boldsymbol{\Sigma} (\boldsymbol{I}_p - \boldsymbol{A})^{-T} \right\}, \tag{2}$$

provided $\boldsymbol{I}_p - \boldsymbol{A}$ is invertible, which is practically always the case by random matrix theory (Rudelson, 2008; Rudelson and Vershynin, 2008). To link an SEM to an RGM, we draw a path diagram $\mathcal{G} = (V, E)$ of SEM by the following rules:

(i) define vertices $V = \{1, \ldots, p, p+1, \ldots, 3p\}$ which represent $(\boldsymbol{Y}, \boldsymbol{X}) = (Y_1, \ldots, Y_p, X_1, \ldots, X_{2p})$;

(ii) draw directed edges $E^d = \{(j, i) \mid a_{ij} \neq 0 \text{ or } b_{i,j-p} \neq 0\}$; and

(iii) draw undirected edges $E^u = \{\{i, j\} \mid i, j = p+1, \ldots, 3p\}$.

In words, (i) we introduce a node for each variable in $(\boldsymbol{Y}, \boldsymbol{X})$ with nodes $j = 1, \ldots, p$ corresponding to Y_j and $p + k$, $k = 1, \ldots, 2p$ corresponding to X_k; (ii) nodes $i = 1, \ldots, p$ (i.e. Y_i nodes) become targets of directed edges from node j if the corresponding $a_{ij} \neq 0$ or $b_{i,j-p} \neq 0$ and only directed edges are allowed between genes; (iii) we introduce undirected edges between every pair X_k and X_l (i.e. nodes $k+p$ and $l+p$). If needed, the inclusion of undirected edges between x-vertices could be explicitly modeled. However, we do not pursue this direction since the focus of this paper is on learning gene regulations. Let $A = \mathrm{diag}(A_1, A_2)$ denote a block diagonal matrix with diagonal blocks A_1 and A_2. Figure 2 shows an example of an RGM with $p = 2$,

$$\boldsymbol{A} = \left[\begin{array}{cc} 0 & * \\ * & 0 \end{array} \right] \text{ and } \boldsymbol{B} = \left[\begin{array}{cccc} * & * & 0 & 0 \\ 0 & 0 & * & 0 \end{array} \right],$$

with $*$ indicating non-zero elements. The path diagram of an SEM following rules (i)–(iii) is Markov with respect to the path diagram $\mathcal{G}$ (Spirtes, 1995; Koster, 1996) if $\boldsymbol{\Sigma}$ is diagonal $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_p)$, i.e. the residuals are independent; the formal theorem of the connection between RGMs and SEMs from Koster (1996) is provided in Supplementary Material A (Ni et al., 2017) for completeness. Since the main inferential goal of this paper is to investigate the regulatory relationship between genes, we will not model the marginal distribution of $\boldsymbol{X}$ and will only focus on the conditional distribution of $\boldsymbol{Y} \mid \boldsymbol{X}$. We fix b_{ik} to 0 for $k \neq 2i - 1$ or $2i$ because copy number and methylation of gene i, in principle, only directly affect the expression of gene i. The associations between copy number and methylation of one gene and gene expression of another gene are hard to interpret and likely to be indirect. For example, the intergenic association of copy number and expression could be explained by the intragenic association between the copy number and expression of the gene, and the intergenic association between expression of both genes. Graphically, in Figure 2, imagine that nodes 3 and 1 are the copy number and expression of one gene, respectively, and node 2 is the expression of another gene. Then the fact that node 3 is connected with node 1 and node 1 connected with node 2 can lead to an association between node 3 and node 1. However, the association is

indirect and confounds with the presented path diagram in Figure 2. Moreover, without this assumption, RGMs that belong to the same Markov equivalence class cannot be differentiated. For example, if we draw directed edges $3 \to 2$ and $4 \to 1$ in Figures 1a, 1b and 1c, then they all become Markov equivalent and their moral graphs are the same as depicted in Figure 1g.

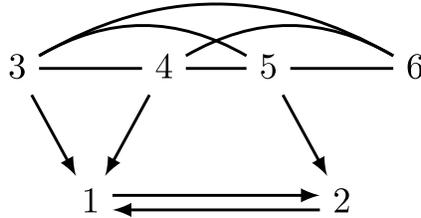Figure 2: Path diagram for simultaneous equations.

2.4 Priors and structural learning

The structural zeros in $\boldsymbol{A}$ and $\boldsymbol{B}$ correspond to missing edges in the RG. Therefore, learning the graph structure is equivalent to finding sparse estimators for $\boldsymbol{A}$ and $\boldsymbol{B}$. Towards this end, we define a non-local prior for $\boldsymbol{A}$ and $\boldsymbol{B}$, which is constructed as follows. We put a thresholded prior on each element of $\boldsymbol{A}$ and $\boldsymbol{B}$. We first write a_{ij} and b_{ik} as

$$a_{ij} = \tilde{a}_{ij}\mathrm{I}(|\,\tilde{a}_{ij}\,| > t_i) \ \text{ and } \ b_{ik} = \tilde{b}_{ik}\mathrm{I}(|\,\tilde{b}_{ik}\,| > t_i) \ \text{ with } \ t_i \sim p(t_i),$$

for $i = 1, \ldots, p$, $j \neq i$ and $k = 2i - 1, 2i$. The threshold parameter t_i controls the minimum effect sizes of a_{ij} and b_{ik}. We do not fix t_i but instead learn it from the data by assigning a prior $p(t_i) = \mathrm{Uniform}(0, t_0)$. We impose normal priors for $\tilde{a}_{ij}$ and $\tilde{b}_{ik}$, $\tilde{a}_{ij} \sim \mathrm{N}(0, \tau_{ij})$ and $\tilde{b}_{ik} \sim \mathrm{N}(0, \nu_{ik})$ and conjugate hyperpriors $\tau_{ij} \sim \mathrm{IG}(\alpha_\tau, \beta_\tau)$ and $\nu_{ik} \sim \mathrm{IG}(\alpha_\nu, \beta_\nu)$. Ni et al. (2018) show that marginally, after integrating with respect to t_i, the induced marginal prior for a_{ij} or b_{ik} is a mixture of a point mass at 0 and a non-local prior (Johnson and Rossell, 2010). Non-local priors shrink small effects to zero, which is desirable in our setting where we are only interested in edges with moderate to strong effects. We complete our prior specifications with a conjugate prior $\sigma_i \sim \mathrm{IG}(\alpha_\sigma, \beta_\sigma)$.

Posterior inference is straightforward by Markov chain Monte Carlo (MCMC); a detailed algorithm is presented in Supplementary Material B. We want to point out that we do not need to invert $\boldsymbol{I} - \boldsymbol{A}$ in (2) because the likelihood can be evaluated by

$$
\begin{aligned}
p(\boldsymbol{Y} \mid \boldsymbol{X}) &= \mathrm{N}_p\left\{\boldsymbol{Y}|(\boldsymbol{I}_p - \boldsymbol{A})^{-1}\boldsymbol{B}\boldsymbol{X}, (\boldsymbol{I}_p - \boldsymbol{A})^{-1}\boldsymbol{\Sigma}(\boldsymbol{I}_p - \boldsymbol{A})^{-T}\right\} \\
&= \mathrm{N}_p\left\{(\boldsymbol{I}_p - \boldsymbol{A})\boldsymbol{Y} - \boldsymbol{B}\boldsymbol{X}|\boldsymbol{0}, \boldsymbol{\Sigma}\right\}|\det(\boldsymbol{I}_p - \boldsymbol{A})|.
\end{aligned}
$$

Let $\boldsymbol{\Gamma}_A = (\gamma_{A,ij})$ and $\boldsymbol{\Gamma}_B = (\gamma_{B,ik})$ denote matrices indicating non-zero elements in $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively. An estimated graph can be reported by selecting edges for which the marginal posterior probability $p(\gamma_{A,ij} = 1 \mid \mathrm{Data})$ or $p(\gamma_{B,ik} = 1 \mid \mathrm{Data})$ exceeds a certain cutoff. This is the median probability model (MPM) when the cutoff is fixed at $1/2$. Instead of being fixed at an arbitrary value, the cutoff can be chosen to control the

posterior expected false discovery rate (FDR) at a desired level α (Newton et al., 2004; Müller et al., 2006). Alternatively we report the highest posterior probability model (HPM) by maximizing $p(\boldsymbol{\Gamma}_A, \boldsymbol{\Gamma}_B \mid \text{Data})$. However, this is only feasible when the model space is small (i.e. with very small p).

3   Simulations

We carry out a simulation study to validate the model's ability to recover a true graph with sample sizes similar to the later application. We generate data by mimicking the actual data from the later application, using a sample size of $n = 276$ and $p = 10$ genes. For each gene, we generate two hypothetical DNA level measurements $\boldsymbol{X} \sim N_{2p}(\boldsymbol{0}, \boldsymbol{I}_{2p})$ (corresponding to copy number and methylation in the real data).

We consider five scenarios. In scenario 1, we randomly set 4/5th entries in $\boldsymbol{A}$ and 1/3rd potentially nonzero entries in $\boldsymbol{B}$ to zero. To ensure that the true model is identifiable, we restrict the simulation truth to include at least one edge from DNA level measurements to each gene (this is relaxed in scenario 2). We draw nonzero elements of $\boldsymbol{A}$ and $\boldsymbol{B}$ from $\{-0.5, 0.5\}$ with equal probability and let $\boldsymbol{\Sigma} = 0.25\boldsymbol{I}_{10}$ in the simulation truth. $\boldsymbol{Y}$ is then generated from (2).

In scenario 2, we reduce the signal by drawing nonzero elements of $\boldsymbol{A}$ and $\boldsymbol{B}$ from $\{-0.4, 0.4\}$ and increase the noise to $\boldsymbol{\Sigma} = \boldsymbol{I}_{10}$. Moreover, we do not force every gene to be connected to at least one of its DNA level measurements. As a result one gene is independent of both its DNA level measurements.

Scenario 3 further reduces the signal by generating nonzero entries of $\boldsymbol{A}$ and $\boldsymbol{B}$ from $\frac{1}{2}U(-0.4, -0.2) + \frac{1}{2}U(0.2, 0.4)$. The diagonal elements σ_i of $\boldsymbol{\Sigma}$ are drawn from $U(0.5, 1.5)$.

Scenario 4 is the same as scenario 1, except that the conditional distribution of $\boldsymbol{Y} \mid \boldsymbol{X}$ is misspecified and generated from a p-dimensional multivariate t-distribution $\mathrm{T}_p(\boldsymbol{\mu}, \boldsymbol{\Theta}, \delta)$ with location $\boldsymbol{\mu} = (\boldsymbol{I}_p - \boldsymbol{A})^{-1}\boldsymbol{B}\boldsymbol{X}$, scale matrix $\boldsymbol{\Theta} = (\boldsymbol{I}_p - \boldsymbol{A})^{-1}\boldsymbol{\Sigma}(\boldsymbol{I}_p - \boldsymbol{A})^{-T}$ and degrees of freedom $\delta = 3$.

In scenario 5, we consider another misspecified model as simulation truth. Instead of fixing b_{ik} to 0 for $k \neq 2i - 1$ or $2i$, we randomly set 5 of the b_{ik} for $k = 2i - 3, 2i - 2, 2i + 1, 2i + 2$ to nonzero which makes certain edges nonidentifiable (due to Markov equivalence). All other simulation specifications are the same as scenario 1.

The hyperparameters are specified as $\alpha_\sigma = \beta_\sigma = \alpha_\tau = \beta_\tau = \alpha_\nu = \beta_\nu = 0.01$ and $t_0 = 1$. We run 50,000 MCMC iterations, discard the first 25,000 iterations as burn-in and retain every 5th sample. The graph is estimated as an MPM.

In Table 1, we report the true positive rate (TPR), FDR, Matthews correlation coefficient (MCC) and the area under the ROC curve (AUC). The performance in scenario 1 is nearly perfect. As to be expected with the decreased signal-to-noise ratio in scenario 2, the TPR drops and the FDR rises relative to scenario 1. Similarly, due to model misspecification we observe less favorable summaries under scenarios 4 and 5 compared to simulations under scenario 1.

Scenario	MCC	TPR	FDR	AUC
1	0.99 (0.02)	1.00 (0.00)	0.01 (0.02)	1.00 (0.00)
2	0.89 (0.07)	0.97 (0.03)	0.11 (0.08)	0.99 (0.02)
3	0.54 (0.10)	0.71 (0.07)	0.30 (0.10)	0.84 (0.05)
4	0.87 (0.08)	0.97 (0.04)	0.14 (0.08)	0.99 (0.02)
5	0.67 (0.03)	0.86 (0.01)	0.24 (0.03)	0.94 (0.01)

Table 1: Simulation results across 5 scenarios. The average operating characteristics are calculated on the basis of 50 simulations; standard deviations are given within parentheses.

Method	MCC	TPR	FDR	$\|\|\Delta\|\|_F$	Time
RGM	0.84 (0.09)	0.95 (0.05)	0.17 (0.10)	1.35 (0.72)	8.86 (0.40)
CAPME	0.00 (0.00)	1.00 (0.00)	0.60 (0.00)	2.26 (0.47)	5.86 (0.10)
ANTAC	0.74 (0.07)	0.79 (0.05)	0.11 (0.07)	8.21 (6.48)	0.01 (0.00)

Table 2: Comparison with CAPME and ANTAC. For fair comparison, MCC, TPR and FDR are reported based on the network of Y. The column labeled by $\|\|\Delta\|\|_F$ is the measure of error in estimating the conditional covariance matrix. The last column is the computational time in minutes on a 2.6 GHz Xeon E5-2690 v3 CPU. The data are generated using settings in scenario 2. The average operating characteristics are calculated on the basis of 50 simulations; standard deviations are given within parentheses.

Comparison with CAPME and ANTAC. We benchmark inference under the RGM against two recent covariate adjusted Gaussian graphical model approaches, CAPME (Cai et al., 2012) and ANTAC (Chen et al., 2016). We use the setting of scenario 2 and compare estimation of the graph structure and the estimation of the conditional covariance matrix $\Theta = (I_p - A)^{-1}\Sigma(I_p - A)^{-T}$ of $Y|X$. Let $\Delta = \widehat{\Theta} - \Theta$ be the difference between the estimated $\widehat{\Theta} = (I_p - \widehat{A})^{-1}\widehat{\Sigma}(I_p - \widehat{A})^{-T}$ and the true Θ. We then compute the Frobenius norm of the difference $\|\|\Delta\|\|_F$. The tuning parameters of CAPME are chosen through 5-fold cross validation on a 100×100 grid and those of ANTAC are set to default theoretical values. For a fair comparison, the graph structure learning performance is evaluated based on the network of Y for all methods and only the skeleton (i.e. undirected version) of the graph is used for CAPME and ANTAC since they do not infer directionality of edges. The performance as well as computation times are summarized in Table 2. RGM compares favorably against both methods in terms of network structure learning and covariance matrix estimation. The computational times of RGM and CAPME are of the same order whereas ANTAC is two orders of magnitude faster as it does not require tuning and Monte Carlo sampling.

Sensitivity analysis. The parameter t_i, a key component in our model, controls the minimum effect size and induces network sparsity. The model includes a prior on t_i, indexed with a hyperparameter t_0. We perform sensitivity analysis with respect to this hyperparemter t_0. We choose scenario 2 and use different values of $t_0 = 0.5, 0.75, 1.0, 1.25, 1.5$. The results are summarized in Table 3. In summary, the method appears to be robust with respect to t_0.

t_0	MCC	TPR	FDR	AUC
0.5	0.88 (0.07)	0.98 (0.03)	0.12 (0.07)	0.99 (0.01)
0.75	0.89 (0.06)	0.98 (0.03)	0.11 (0.06)	0.99 (0.01)
1	0.89 (0.07)	0.97 (0.03)	0.11 (0.08)	0.99 (0.02)
1.25	0.87 (0.08)	0.95 (0.05)	0.11 (0.08)	0.98 (0.03)
1.25	0.86 (0.10)	0.93 (0.07)	0.11 (0.07)	0.98 (0.02)

Table 3: Sensitivity analysis of hyperparameter $t_0 \in \{0.5, 0.75, 1.0, 1.25, 1.5\}$. The data are generated using settings in scenario 2. The average operating characteristics are calculated on the basis of 50 simulations; standard deviations are given within parentheses.

$(n, p+q)$	MCC	TPR	FDR	AUC
(200, 40)	0.87 (0.11)	0.90 (0.05)	0.11 (0.15)	0.98 (0.02)
(300, 40)	0.95 (0.03)	0.95 (0.03)	0.03 (0.03)	1.00 (0.01)
(200, 80)	0.73 (0.10)	0.61 (0.06)	0.10 (0.17)	0.91 (0.03)
(300, 80)	0.85 (0.04)	0.78 (0.06)	0.06 (0.03)	0.95 (0.02)
(200, 120)	0.61 (0.09)	0.43 (0.07)	0.09 (0.18)	0.81 (0.03)
(300, 120)	0.74 (0.06)	0.62 (0.05)	0.09 (0.09)	0.86 (0.02)

Table 4: Scalability of RGM. The average operating characteristics are calculated on the basis of 50 simulations; standard deviations are given within parentheses.

Scalability. The reported simulations and later case studies include networks with 30 vertices ($p = 10$ and $q = 20$). To investigate the scalability of RGM, we extend the simulations in scenario 2 by considering six different combinations of sample sizes (n) and number of nodes ($p + q$): $(n, p + q) \in \{200, 300\} \times \{40, 80, 120\}$ with $p = q$. The network sparsity is kept approximately the same as in the other scenarios. The results are summarized in Table 4 and show no evidence of significant deterioration with increasing number of vertices.

4   Analyzing colon adenocarcinoma data

We use TCGA-Assembler (Zhu et al., 2014) to retrieve mRNA gene expression (GE), DNA copy number (CN) and DNA methylation (ME) data for colon adenocarcinoma (COAD) from the Cancer Genome Atlas (TCGA). We focus on genes that are mapped to the RAS-MAPK pathway, which is critical for initiation of carcinogenesis in COAD (Colussi et al., 2013). The RAS-MAPK pathway includes $p = 10$ core genes. Restricting to samples with available GE, CN and ME data, the sample size is $n = 276$. We run two parallel MCMC simulations, each with 50,000 iterations, discard the first 50% as burn-in and thin the chains to every 5th sample. MCMC diagnostics show no evidence for lack of practical convergence (for details see Supplementary Material C). We summarize the posterior distribution on the unknown graph by controlling the posterior expected FDR < 10%.

We find that all genes are associated with their respective copy number and NRAS, MAP2K1, MAPK1 and MAPK3 are associated with methylation as well. Therefore, the

(a) Full network

(b) Feed-forward loop

(c) Negative feedback loop

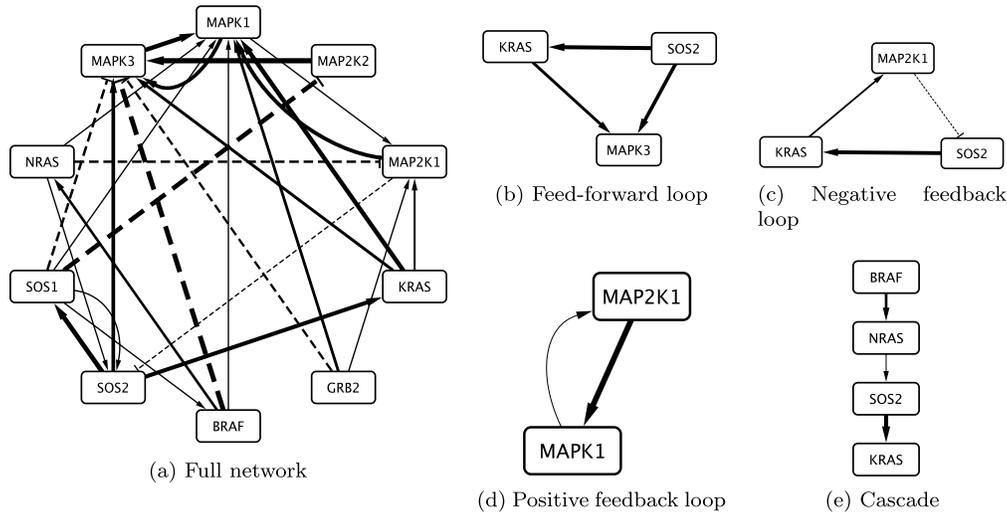(d) Positive feedback loop

(e) Cascade

Figure 3: Estimated gene network for colon adenocarcinoma. Solid lines with arrowheads represent stimulatory interactions, whereas dashed lines with horizontal bars denote inhibitory regulation. Line width is proportional to its posterior probability. Panel (a) full network; Panels (b)–(e) network motifs.

gene interactions are fully identifiable by the argument in the last paragraph of Section 2.2. The full network of gene interactions is shown in Figure 3a. The solid lines with arrowheads represent stimulatory interactions and the dashed lines with perpendicular bars denote inhibitory regulations. Uncertainties about the selected edges are quantified as posterior inclusion probabilities which are displayed as edge width in Figure 3. In total, we find 21 stimulatory ($a_{ij} > 0$) and 7 inhibitory ($a_{ij} < 0$) regulatory relationships.

Gene networks are often made up of a small set of recurrent regulation patterns, called network motifs, which can be thought of as fundamental building blocks for the network and are expected to occur more often in gene networks than in random networks. In Figures (3b)–(3e), we display four motifs identified by our method that are commonly observed in gene networks (Alon, 2007). Figure (3b) shows a feed-forward loop among SOS2, KRAS and MAPK3. Part of this feed-forward loop is well studied in COAD (Zenonos and Kyprianou, 2013). SOS2 binds KRAS and removes guanosine diphosphate (GDP) molecules from KRAS and thus allows guanosine triphosphate (GTP) molecules to bind and activate it. The active KRAS would eventually activates MAPK3 through the kinase cascade. Another important motif is feedback loop. We present a negative and a positive feedback loop in Figures 3c and 3d, respectively. The regulatory relationships in the negative feedback loop of SOS2, KRAS and MAP2K1 have been studied extensively. SOS2→KRAS→MAP2K1 is part of the well-known MAP kinases cascade (Plotnikov et al., 2011; Zenonos and Kyprianou, 2013) while MAP2K1 can phosphorylate and inhibit SOS2 and thereby reduces MAP2K1 activation (Holt et al., 1996; Mendoza et al., 2011). For the positive feedback loop (Figure 3d), MAP2K1
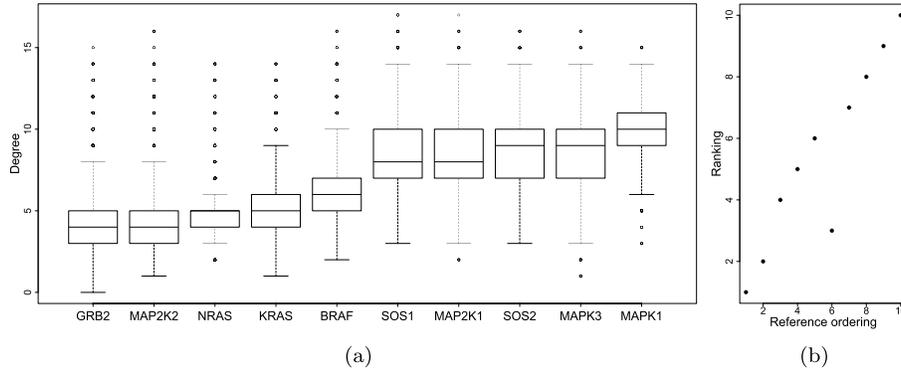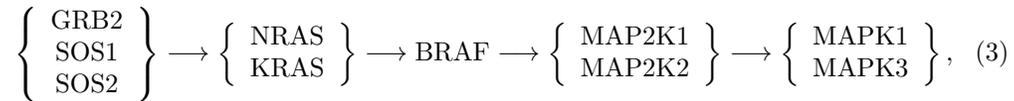
Figure 4: Panel (a): box plots for posterior distribution of degree of each gene; Panel (b): scatter plot for ranking vs reference ordering. Ranking of the genes is based on the Score$_j$ defined by (4). The reference ordering is based on (3) where the genes with the same bracket are assigned arbitrary ordering.

activating MAPK1 is again part of the MAP kinases cascade but the reversed activation is less explored in the literature. Similarly, another network motif (Figure 3e), regulatory cascade, also need to be validated by further biological experiment.

On the gene level, we calculate the posterior distribution of degree of each gene, that is, the number of edges connected to the gene. We visualize these posterior distributions by box plots in Figure 4a. Highly connected genes are often called hub genes which are usually more involved in multiple regulatory activities than non-hub genes. MAPK1 and MAPK3 appear to be the two most highly connected genes. Several studies have shown that overexpression of MAPK1/3 plays an critical role in the progression of COAD (Fang and Richardson, 2005) and is responsible for proliferation, differentiation, survival, migration and angiogenesis of tumors in many cancers (Dhillon et al., 2007).

For another quantitative exploration of the known MAPK signalling cascade we consider the following inference. The following signaling pathway/cascade has been extensively studied and validated in biological literature (Plotnikov et al., 2011; Zenonos and Kyprianou, 2013),

$$\left\{ \begin{array}{c} \text{GRB2} \\ \text{SOS1} \\ \text{SOS2} \end{array} \right\} \longrightarrow \left\{ \begin{array}{c} \text{NRAS} \\ \text{KRAS} \end{array} \right\} \longrightarrow \text{BRAF} \longrightarrow \left\{ \begin{array}{c} \text{MAP2K1} \\ \text{MAP2K2} \end{array} \right\} \longrightarrow \left\{ \begin{array}{c} \text{MAPK1} \\ \text{MAPK3} \end{array} \right\}, \quad (3)$$

where genes within each curly brace belong to the same gene family (except for GRB2 for which the protein often binds to SOS to form a protein complex) and play similar roles in the pathway. We label each gene in cascade (3) from left to right with integer $j = 1, \ldots, p$ as our reference ordering. The labels within each curly brace are arbitrary. To compare our findings with this well established cascade, we score each gene j by

$$\text{Score}_j = \text{indegree}_j - \text{outdegree}_j, \tag{4}$$

where indegree$_j$ is the number of edges pointing towards gene j and outdegree$_j$ is the number of edges pointing away from gene j. Intuitively, gene j is likely to be on the left of gene k in cascade (3) if Score$_j <$ Score$_k$. We then rank the scores in an increasing order and plot the rank against the reference ordering in Figure 4b. We also calculate the normalized Kendall's tau distance, which is the ratio of the number of discordant pairs over the total number of pairs. The normalized Kendall's tau distance lies between 0 and 1, with 0 indicating a perfect agreement of the two orderings and 1 indicating a perfect disagreement. The resulting Kendall's tau distance is 0.07. Our findings appear to be consistent with the biologically validated pathway since both Figure 4b and the Kendall's tau distance indicate a good concordance between our rankings and the reference ordering.

For comparison, we apply ANTAC to this dataset and the results are provided in Supplementary Material D.

5 Discussion

In this article, we have introduced a Gaussian RGM to model gene regulatory relationships from genomic data. RGMs are statistically more general and biologically more interpretable than MRFs and DAGs. By integrating DNA level information, we are able to differentiate between RGMs that belong to the same Markov equivalence class. We exploit the connection between RGMs and SEMs for efficient inference. We constructed a prior probability model for the unknown graph using a thresholded model which marginally defines a mixture of non-local prior and a point mass. We use simulation studies to illustrate the performance of our method in terms of graph structure learning. Our method is applied to a colon cancer pathway analysis. Some of our findings are consistent with the literature, while others need to be validated by biological experiments. Although our applications focus on gene regulatory networks, the proposed approach is general and can be potentially applied to other scientific settings such as climate sciences and macroeconomics. The approach works for any network where some edges have known direction if included (inclusion itself need not be fixed).

The link between RGMs and SEMs is based on the assumption that the gene expressions are multivariate Gaussian, which could be thought of as one limitation of our model. We have empirically demonstrated the robustness of the proposed model against slight model misspecification. More general sampling models might need an additional hierarchical layer of latent variables, such as latent probit scores for binary outcomes. Another limitation of this work is that we only consider *cis-regulations*. That is, DNA copy number or methylation of a gene only affects the mRNA gene expression of the same gene. Although *trans-regulations* are also important, including them in our model will introduce Markov equivalent graphs that are nonidentifiable. For example, adding edges $3 \to 2$ and $4 \to 1$ in Figures 1b and 1c will lead to two Markov equivalent graphs.

In the simulations and application, the sample size n is always larger than the number p of genes. Although with proper priors the posterior distribution is always proper even when $p > n$, the performance of the proposed model will deteriorate as p grows as we have seen in Table 4. The graph space is super-exponential in p. Hence, in high dimensional problems the results should be interpreted with caution.

Supplementary Material

Supplementary Material for "Reciprocal Graphical Models for Integrative Gene Regulatory Network Analysis" (DOI: 10.1214/17-BA1087SUPP; .pdf).

References

Alon, U. (2007). "Network motifs: theory and experimental approaches." *Nature Reviews Genetics*, 8(6): 450–461. 1105

Bhadra, A. and Mallick, B. K. (2013). "Joint high-dimensional Bayesian variable and covariance selection with an application to eQTL analysis." *Biometrics*, 69(2): 447–457. MR3071063. doi: https://doi.org/10.1111/biom.12021. 1096

Cai, T. T., Li, H., Liu, W., and Xie, J. (2012). "Covariate-adjusted precision matrix estimation with an application in genetical genomics." *Biometrika*, 100(1): 139–156. MR3034329. doi: https://doi.org/10.1093/biomet/ass058. 1096, 1103

Cai, X., Bazerque, J. A., and Giannakis, G. B. (2013). "Inference of gene regulatory networks with sparse structural equation models exploiting genetic perturbations." *PLoS Computational Biology*, 9(5): e1003068. 1096

Chen, M., Ren, Z., Zhao, H., and Zhou, H. (2016). "Asymptotically normal and efficient estimation of covariate-adjusted Gaussian graphical model." *Journal of the American Statistical Association*, 111(513): 394–406. 1096, 1103

Colussi, D., Brandi, G., Bazzoli, F., and Ricciardiello, L. (2013). "Molecular pathways involved in colorectal cancer: implications for disease behavior and prevention." *International Journal of Molecular Sciences*, 14(8): 16365–16385. 1104

Dhillon, A. S., Hagan, S., Rath, O., and Kolch, W. (2007). "MAP kinase signalling pathways in cancer." *Oncogene*, 26(22): 3279–3290. 1106

Dobra, A., Lenkoski, A., and Rodriguez, A. (2012). "Bayesian inference for general Gaussian graphical models with application to multivariate lattice data." *Journal of the American Statistical Association*. 1095

Fang, J. Y. and Richardson, B. C. (2005). "The MAPK signalling pathways and colorectal cancer." *The Lancet Oncology*, 6(5): 322–327. 1106

Frydenberg, M. (1990). "The chain graph Markov property." *Scandinavian Journal of Statistics*, 333–353. 1097

Green, P. J. and Thomas, A. (2013). "Sampling decomposable graphs using a Markov chain on junction trees." *Biometrika*, 100(1): 91–110. MR3034326. doi: https://doi.org/10.1093/biomet/ass052. 1095

Holt, K. H., Kasson, B. G., and Pessin, J. E. (1996). "Insulin stimulation of a MEK-dependent but ERK-independent SOS protein kinase." *Molecular and Cellular Biology*, 16(2): 577–583. 1105

Johnson, V. E. and Rossell, D. (2010). "On the use of non-local prior densities in Bayesian hypothesis tests." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(2): 143–170. 1101

Koster, J. T. (1996). "Markov properties of nonrecursive causal models." *The Annals of Statistics*, 24(5): 2148–2177. MR1421166. doi: https://doi.org/10.1214/aos/1069362315. 1096, 1097, 1100

Kundu, S. and Kang, J. (2016). "Semiparametric Bayes conditional graphical models for imaging genetics applications." *Stat*, 5(1): 322–337. 1096

Mendoza, M. C., Er, E. E., and Blenis, J. (2011). "The Ras-ERK and PI3K-mTOR pathways: cross-talk and compensation." *Trends in Biochemical Sciences*, 36(6): 320–328. 1105

Mitra, R., Müller, P., Liang, S., Yue, L., and Ji, Y. (2013). "A Bayesian graphical model for ChIP-seq data on histone modifications." *Journal of the American Statistical Association*, 108(501): 69–80. 1095

Müller, P., Parmigiani, G., and Rice, K. (2006). "FDR and Bayesian multiple comparisons rules." 1102

Newton, M. A., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004). "Detecting differential gene expression with a semiparametric hierarchical mixture method." *Biostatistics*, 5(2): 155–176. 1102

Ni, Y., Ji, Y., and Müller, P. (2017). "Supplementary Material for "Reciprocal Graphical Models for Integrative Gene Regulatory Network Analysis"." *Bayesian Analysis*. doi: https://doi.org/10.1214/17-BA1087SUPP. 1100

Ni, Y., Stingo, F., and Baladandayuthapani, V. (2018). "Bayesian graphical regression." *Journal of the American Statistical Association*, just accepted. 1101

Ni, Y., Stingo, F. C., and Baladandayuthapani, V. (2015). "Bayesian nonlinear model selection for gene regulatory networks." *Biometrics*, 71(3): 585–595. 1095

Oates, C. J., Smith, J. Q., and Mukherjee, S. (2016). "Estimating causal structure using conditional DAG models." *Journal of Machine Learning Research*, 17(54): 1–23. 1096

Plotnikov, A., Zehorai, E., Procaccia, S., and Seger, R. (2011). "The MAPK cascades: signaling components, nuclear roles and mechanisms of nuclear translocation." *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1813(9): 1619–1633. 1105, 1106

Rothman, A. J., Levina, E., and Zhu, J. (2010). "Sparse multivariate regression with covariance estimation." *Journal of Computational and Graphical Statistics*, 19(4): 947–962. 1096

Rudelson, M. (2008). "Invertibility of random matrices: norm of the inverse." *Annals of Mathematics*, 575–600. MR2434885. doi: https://doi.org/10.4007/annals.2008.168.575. 1100

Rudelson, M. and Vershynin, R. (2008). "The Littlewood–Offord problem and invertibility of random matrices." *Advances in Mathematics*, 218(2): 600–633. MR2407948. doi: https://doi.org/10.1016/j.aim.2008.01.010.   1100

Shin, S.-Y., Rath, O., Zebisch, A., Choo, S.-M., Kolch, W., and Cho, K.-H. (2010). "Functional roles of multiple feedback loops in ERK and Wnt signaling pathways that regulate epithelial-mesenchymal transition." *Cancer Research*, 70(17): 6715.   1096

Spirtes, P. (1995). "Directed cyclic graphical representations of feedback models." In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, 491–498. Morgan Kaufmann Publishers Inc.   1100

Stingo, F. C., Chen, Y. A., Vannucci, M., Barrier, M., and Mirkes, P. E. (2010). "A Bayesian graphical modeling approach to microRNA regulatory network inference." *The Annals of Applied Statistics*, 4(4): 2024.   1095

TCGA (2012). "Comprehensive molecular characterization of human colon and rectal cancer." *Nature*, 487(7407): 330–337.   1096

Telesca, D., Müller, P., Kornblau, S. M., Suchard, M. A., and Ji, Y. (2012a). "Modeling protein expression and protein signaling pathways." *Journal of the American Statistical Association*, 107(500): 1372–1384.   1096

Telesca, D., Müller, P., Parmigiani, G., and Freedman, R. S. (2012b). "Modeling dependent gene expression." *The Annals of Applied Statistics*, 6(2): 542–560.   1096

Wang, H. and West, M. (2009). "Bayesian analysis of matrix normal graphical models." *Biometrika*, 96(4): 821–834.   1095

Wang, W., Baladandayuthapani, V., Holmes, C. C., and Do, K.-A. (2013). "Integrative network-based Bayesian analysis of diverse genomics data." *BMC Bioinformatics*, 14(Suppl 13): S8.   1095

Yajima, M., Telesca, D., Ji, Y., and Müller, P. (2015). "Detecting differential patterns of interaction in molecular pathways." *Biostatistics*, 16(2): 240–251. MR3365426. doi: https://doi.org/10.1093/biostatistics/kxu054.   1095

Zenonos, K. and Kyprianou, K. (2013). "RAS signaling pathways, mutations and their role in colorectal cancer." *World Journal of Gastrointestinal Oncology*, 5(5): 97–101. 1105, 1106

Zhang, D., Wells, M. T., Turnbull, B. W., Sparrow, D., and Cassano, P. A. (2005). "Hierarchical graphical models: An application to pulmonary function and cholesterol levels in the normative aging study." *Journal of the American Statistical Association*, 100(471): 719–727.   1096

Zhang, L. and Kim, S. (2014). "Learning gene networks under SNP perturbations using eQTL datasets." *PLoS Computational Biology*, 10(2): e1003420.   1096

Zhu, Y., Qiu, P., and Ji, Y. (2014). "TCGA-assembler: open-source software for retrieving and processing TCGA data." *Nature Methods*, 11(6): 599–600.   1104