Bayesian Analysis (2016) **11**, Number 1, pp. 99–124

Bayesian Graphical Models for Differential Pathways

Riten Mitra[*], Peter Müller[†], and Yuan Ji[‡]

Abstract. Graphical models can be used to characterize the dependence structure for a set of random variables. In some applications, the form of dependence varies across different subgroups. This situation arises, for example, when protein activation on a certain pathway is recorded, and a subgroup of patients is characterized by a pathological disruption of that pathway. A similar situation arises when one subgroup of patients is treated with a drug that targets that same pathway. In both cases, understanding changes in the joint distribution and dependence structure across the two subgroups is key to the desired inference. Fitting a single model for the entire data could mask the differences. Separate independent analyses, on the other hand, could reduce the effective sample size and ignore the common features. In this paper, we develop a Bayesian graphical model that addresses heterogeneity and implements borrowing of strength across the two subgroups by simultaneously centering the prior towards a global network. The key feature is a hierarchical prior for graphs that borrows strength across edges, resulting in a comparison of pathways across subpopulations (differential pathways) under a unified model-based framework. We apply the proposed model to data sets from two very different studies: histone modifications from ChIP-seq experiments, and protein measurements based on tissue microarrays.

Keywords: autologistic regression, histone modifications, Markov random fields, networks, reverse phase protein arrays.

1 Introduction

1.1 Background

We discuss inference for the comparison of graphical models. The discussion is motivated by two biomedical inference problems. In the first application, we study how protein pathways change across different disease subpopulations. In the second motivating application, we compare dependence structure of histone modification (HM) counts in promoter regions across genes with high versus low expression. Both applications address important questions in biomedical research that cannot easily be addressed by existing methods. Details of the applications are reported later. In both cases, dependence structure is formalized as a graph with the nodes representing proteins or HMs. Both applications require coherent inference on how conditional dependence structure changes for the same set of nodes across different biological conditions. To achieve such

[*] Department of Biostatistics, University of Louisville, r0mitr01@louisville.edu

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

[‡] Center for Clinical and Research Informatics, Northshore University HealthSystem, jiyuan@uchicago.edu

DOI: 10.1214/14-BA931

inference, we introduce graphical models that impose a hierarchical prior on *a pair of graphs*. The main contribution of this paper is an approach for inference for comparing two graphs.

Graphical models can be used to characterize the dependence structure of a set of random variables. We focus on Markov random field (MRF) models. An MRF can be represented as an undirected graph $G = (V, E)$, where $V = \{1, \ldots, m\}$ are the nodes and edges are pairs of nodes, $E = \{\{i, j\},\ i \neq j \in V\}$. The nodes index random variables. In our case studies, the variables are protein expression or HM counts. The edges represent conditional dependence between the corresponding random variables. The absence of an edge between any two nodes indicates conditional independence of the corresponding random variables. In general, an undirected graph is a more general concept than an MRF. For example, the graph G itself might be the observed data, without any notion of modeling conditional independence structure. However, for the purpose of this paper, we shall assume that an MRF is indistinguishable from any undirected graphical model.

Some commonly used examples of MRF models are Gaussian graphical models (GGM). A GGM uses a multivariate normal distribution to describe the joint distribution of the nodes. The absence of an edge, i.e., conditional independence, corresponds to a zero entry in the precision matrix of the Gaussian distribution. The G-Wishart distribution is a conjugate prior for the inverse covariance matrix under the constraint to these zero entries. This feature has, in recent years, motivated the development of a range of computational techniques for sampling from this distribution. However, inference in these high dimensional graphical models remains computationally challenging. Several authors therefore propose sparsity control as one means of mitigating the computational challenge. Related proposals are mostly based on lasso and penalized likelihood techniques (Yuan and Lin, 2007). Recently developed lasso methods that deal with conditional likelihoods through neighborhood selection include Chen et al. (2013); Ravikumar et al. (2010); Meinshausen and Bühlmann (2006); Yang et al. (2012) and Yang et al. (2013). A common theme of these approaches is the maximization of some penalized versions of conditional likelihood per node. Node-specific inference is then merged to reconstruct the entire graph. Some of these methods have been shown to be consistent under some sparsity constraints of the true graph.

In contrast, the Bayesian approach relies on carefully specified priors. The role of prior specifications has been discussed, in great detail among many others, in Dobra et al. (2004) Jones et al. (2004), Scott and Carvalho (2008) and Carvalho and Scott (2009). For decomposable graphs, Carvalho et al. (2007) proposed a direct and efficient method based on the perfect ordering of cliques. For general graphs, Piccioni (2000) proposed a block Gibbs sampler using Bayesian iterative proportional scaling. However, this technique relies on clique enumeration and is computationally expensive. To address these limitations and increase computational efficiency, Mitsakakis et al. (2011), Dobra and Lenkoski (2011) and Wang and Carvalho (2010) proposed several approaches based on novel reversible jump and Metropolis Hasting steps. These improvements followed the theory for non-decomposable graphs developed in Atay-Kayis and Massam (2005).

Graphical models are increasingly used for inference in biomedical research problems. The first applications to biological networks date back to as early as Wright (1934).

More recently, Lauritzen and Sheehan (2003) demonstrated the use of these models in detecting allele networks and analyzing pedigrees. Zhang (2012) proposed a novel Bayesian graphical model for multilocus disease association in genome-wide case control studies. Another example is the work of Stingo et al. (2010) who propose a Bayesian graphical modeling approach to infer a miRNA regulatory network. A key feature of their Bayesian approach is the use of priors to include important covariates like sequence information. Applications like these are often characterized by a large number of nodes (regulatory elements) and small sample sizes. This has motivated the use of sparsity to reduce dimensionality in network inference. For example, Dobra et al. (2004) impose sparsity constraints in GGMs for modeling gene–gene interactions in high dimensional data. Sparsity is also sometimes desirable as a biologically meaningful constraint, e.g., in the application of Stingo et al. (2010) where small number of microRNAs regulate a large number of genes. Instead of a GGM, Mitra et al. (2013) propose a graphical model for binary indicators based on an autologistic model. In particular, we carried out inference with multivariate count data, by using the binary indicators at the nodes of the graph as latent variables with an additional sampling model for the observed data given the latent indicators.

1.2 Multiple Graphs

Biomedical research problems related to unknown networks often naturally lead to joint inference for multiple related graphs. Biological networks across related disease subcategories, related genes, protein-pathways targeted by the same drug, are natually modeled to share some common characteristics. Danaher et al. (2013) and Guo et al. (2011) introduced the idea in frequentist models. They estimate multiple related GGMs for observations belonging to distinct classes. Their method borrows strength across the classes through an appropriate convex penalty functions. Some other examples of joint graphical modeling using penalized likelihood appear in Chiquet et al. (2011); Hara and Washio (2013); Yang et al. (2012); Mohan et al. (2012) and Mohan et al. (2013). For example, Mohan et al. (2012, 2013) used the perturbed-node joint graphical lasso which introduces a convex optimization that is based upon the use of a row-column overlap norm penalty.

Our approach adds a novel perspective to this problem through hierarchical Bayesian priors. We use priors to borrow strength across multiple graphs and sharpen inference for datasets with small sample sizes. To our knowledge, this is the first Bayesian formulation of joint graphical inference, except for Peterson et al. (2014) who discuss the special case of GGMs and focus a slightly different problem by considering the case of several related graphs, with an MRF over graphs. Instead, we focus here on the comparison of two graphs with a general sampling model. We build on the models of Mitra et al. (2013) to achieve this. The approach is motivated by problems where a heterogeneous population of samples gives rise to two subgroups. Assuming a common network for both subgroups could bias inference to a large extent. On the other hand, treating them as independent samples would result in a loss of efficiency. A typical example are protein measurements for cancer patients. Different disease subtypes give rise to patient subgroups, which, while sharing many common characteristics, possess unique features that are specific to each subtype. Typically, the unique features and differences across subgroups are the focus of interest. The proposed hierarchical graphical models borrow strength across subgroups and allow inference on the differences between the two groups. The proposed approach can be characterized as a hierarchical model across graphs.

The rest of the article is organized as follows. In the next section, we state the proposed model. In Section 3, we discuss model choice between the proposed differential graph model and a model based on two separate independent graphs. Section 4 describes a posterior simulation scheme to implement posterior inference. In Section 5, we report some simulation experiments to validate the proposed graphical model. In Section 6, we describe the application of our model to two case studies. Finally, we conclude with a brief discussion on the importance of these results and their relevance to the current state of biological research.

2 Model

2.1 A Differential Prior Model for Graphs

We propose a prior model for a pair of graphs that represent conditional independence structure. We denote the two unknown networks sharing the same set of nodes by $G_1 = (V, E^1)$ and $G_2 = (V, E^2)$, respectively. For later reference we first summarize the overall model structure. We define a joint probability model for G_1, G_2, π, $\boldsymbol{y}$, $\boldsymbol{\theta}$, and $\boldsymbol{\beta}$. Here $G_k, k = 1, 2$, denote the two graphs, π denotes a hyperparameter that is interpreted as the overall probability of edges matching across G_1 and G_2, $\boldsymbol{y}$ is the observed data, and $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ are parameters that index the sampling model for $\boldsymbol{y}$, with $\boldsymbol{\beta}$ describing the strength of dependence for those outcomes that are not constrained to be conditionally independent by G_k. Also, $\boldsymbol{y} = (\boldsymbol{y}^1, \boldsymbol{y}^2)$ are the data arranged by group, and similarly for $\boldsymbol{\beta}^k$ and $\boldsymbol{\theta}^k$. In the following sections, we define a joint probability model,

$$p(\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\beta}, G_1, G_2) \propto p(G_1, G_2 \mid \pi)\, p(\pi) \prod_{k=1}^{2} p(\boldsymbol{y}^k \mid \boldsymbol{\theta}^k, \boldsymbol{\beta}^k, G_k)\, p(\boldsymbol{\beta}^k \mid G_k)\, p(\boldsymbol{\theta}^k). \tag{1}$$

The first two factors are the key elements of the proposed model. It is the joint prior over the two related graphs which helps achieve meaningful inference about the comparison of G_1 and G_2. The third factor is the sampling model for the observed data $\boldsymbol{y}$. For the moment we only need to assume that it can be chosen to respect the conditional independence structure that is specified in G_k. Details of the sampling model, as well as the prior on $\boldsymbol{\beta}^k$ and $\boldsymbol{\theta}^k$, are discussed below, in Section 2.2. Let $G^k_{ij} = I(\{i, j\} \in E^k)$ denote an indicator for the presence of an edge in G_k. We define $\delta_{ij} = |G^2_{ij} - G^1_{ij}|$ as a latent indicator for a difference between the two graphs at the edge $\{i, j\}$.

We assume that prior expert information can be summarized as a prior guess $G_0 = (V, E_0)$. Let $U(G_0)$ denote a uniform distribution on the space of all subgraphs of G_0. We start the model construction with a uniform prior for G_1,

$$G_1 \sim U(G_0). \tag{2}$$

In words, $p(G_1)$ has support only on the edges of the prior graph G_0. No prior mass will be placed on any edges outside G_0. Each edge in G_0 is included in G_1 independently with probability 0.5. We complete the prior $p(G_1, G_2 \mid \pi)$ with independent priors for the differences δ_{ij}

$$\delta_{ij} \sim \text{Ber}(\pi),\ i < j,\ \text{and}\ \pi \sim Beta(a, b). \tag{3}$$

Together G_1 and δ_{ij} implicitly define G_2 by $G^2_{ij} = G^1_{ij}(1 - \delta_{ij}) + (1 - G^1_{ij})\delta_{ij}$ for all edges $\{i, j\},\ i < j$. We refer to (3) as the *differential graph model*, and refer to π as the global probability of similarity.

Conditional on π and G_1, the prior on G_2 places more mass on structures closer (when $\pi < 0.5$) or distant ($\pi > 0.5$) to G_1. Fixing $\pi = 0.5$ recovers the special case of $G_k \sim U(G_0)$, $k = 1, 2$, independently. In any case, the marginals are identical, $p(G_1) = p(G_2)$. This does not remain true under general, alternative priors $p(G_1)$. In view of this asymmetry, it is natural to think of G_1 as a reference graph.

In some applications, informative prior information might not be readily available, and we assume instead $G_1 \sim U_m(V)$, where $U_m(V)$ denotes a uniform distribution over all graphs $G = (V, E)$ of size $|V| = m$.

In the applications and in the description of posterior inference, we fix the hyperparameters at $a = b = 1$. However, $p(G_2 \mid G_1)$ diverges from $p(G_2)$ after marginalizing out π. The conditionals posterior distributions $p(\pi \mid G_1, G_2, data)$ and $p(\delta \mid \pi, data)$ form the basis of inference in this differential model. The inclusion probability π plays a significant role in the differential graph model. Note that π is an unknown parameter to be estimated from the available data. It informs us about the global similarity between two networks. When the data suggests network similarity, π is closer to 0. This, in turn, enforces similarity between the two networks. In the absence of any information, we could use a uniform hyperprior for π. However, depending on the context availability of expert knowledge, we can construct an informative prior around π. In general, we recommend using different values of a and b if more specific prior information were available. These hyperparameters would then reflect prior beliefs on the commonality across graphs. Overall, posterior learning about π is essential for borrowing strength.

In addition, several interesting alternatives could arise out of the general framework described above. For example, instead of the uniform prior on G_1, we could center G_1 at a prior guess, say by $p(G_1) \propto \rho^{d(G_1, G_0)}$ where d is a distance between the two graphs and ρ is a pre-specified concentration parameter.

2.2 The Sampling Model

Conditional on G_k, $k = 1, 2$, we assume a sampling model for the observed data. Let y_{kti}, $i = 1, \ldots, m$, $t = 1, \ldots, n_k$, $k = 1, 2$, denote the observed data for experimental unit t in group k, and i indexes the coordinates of the m-dimensional response vector. Let $\boldsymbol{y}^k$ denote all data for the kth group. We assume a sampling model

$$p(\boldsymbol{y}^k \mid \boldsymbol{\beta}^k, \boldsymbol{\theta}^k, G_k). \tag{4}$$

The sampling model is indexed by parameters $\boldsymbol{\beta}^k$ and $\boldsymbol{\theta}^k$, and it is defined to respect the conditional independence structure given by G_k. Details of the sampling model depend on the application. In both case studies that we discuss later, we set up a hierarchical model

$$p(\boldsymbol{y}^k \mid \beta^k, \boldsymbol{\theta}^k, G_k) = \int \prod_{i,t} p(y_{kti} \mid v_{kti}, \boldsymbol{\theta}^k)\; p(\boldsymbol{v}^k \mid \boldsymbol{\beta}^k, G_k)\, d\boldsymbol{v}^k.$$

The model uses a set of latent binary indicators $\boldsymbol{v}_{tk} = (v_{1tk}, \ldots, v_{mtk})$, $v_{kti} \in \{0, 1\}$ and $\boldsymbol{v}^k = (\boldsymbol{v}_{tk},\ t = 1, \ldots, n_k)$. The indicators are interpreted as activation of a protein or as presence of a histone modification, respectively. In the first case study with RPPA data, as well as in the upcoming simulation study, we use a normal sampling model,

$$p(y_{kti} \mid v_{kti}, \boldsymbol{\theta}^k) \propto \begin{cases} N(\mu_{1ik}, \sigma^2_{1ik}) & \text{if } v_{kti} = 0, \\ N(\mu_{2ik}, \sigma^2_{2ik}) & \text{if } v_{kti} = 1. \end{cases} \tag{5}$$

Let $\boldsymbol{\theta}^k = (\mu_{1ki}, \mu_{2ki}, \sigma_{1ki}, \sigma_{2ki},\ i = 1, \ldots, m, k = 0, 1)$ denote the parameters that index the sampling model.

We continue the construction of the hierarchical sampling model with a prior on $\boldsymbol{v}$. This is where we impose the conditional independence structure described by G_k. We use an autologistic model (Besag, 1974),

$$p(\boldsymbol{v}_{kt} \mid \boldsymbol{\beta}^k, G_k) \propto \exp\left\{ \sum_i \beta^k_i v_{kti} + \sum_{j:\ G^k_{ij}=1} \beta^k_{ij}(v_{kti} - m^k_i)(v_{jtk} - m^k_j) \right\}, \tag{6}$$

where $m^k_i = 1/\{1+\exp(-\beta^k_i)\}$. The interaction coefficients β^k_{ij} is zero whenever $G^k_{ij} = 0$. The inclusion of G_k in the conditioning subset highlights the dependence of the autologistic model on G_k.

We complete the sampling model with independent priors for the non-zero elements

$$\beta^k_{ij} \sim \mathrm{N}(0, \sigma^2_\beta),\ \{i, j\} \in E^k, \quad k = 1, 2. \tag{7}$$

Finally, we will discuss priors for $\boldsymbol{\theta}^k$ in the context of the case studies.

3 Graphical Model Choice

We argue for the differential graph model (3) and (2) over the default alternative of independent models with independent priors for G_1 and G_2, i.e., $p(G_k) = U(G_0)$, $k = 1, 2$, independently. The two main reasons for preferring the differential graph model over the independent models are multiplicity control and better modeling of the experimental setup. We discuss both issues in some more detail below and show that the latter matters.

Multiplicity Control for Comparing Dependence Structure Inference on comparing two graphs G_1 and G_2 can be described as a massive multiple comparison problem. For each possible edge $\{i, j\}$ we decide whether to report the edge as different across the two subgroups. Recall that δ_{ij} denotes the truth about comparing edge $\{i, j\}$ across the two graphs. Let $p = m(m-1)/2$ denote the number of comparisons, that is, the number of edges.

Posterior inference under (3) automatically adjusts for multiplicities, in the following sense. Scott and Berger (2006, 2010) consider inference for a family of hypotheses $\delta_i = 0$ versus $\delta_i = 1$, $i = 1, \ldots, p$. They consider the special case of variable selection in a linear model, but the argument is more general. Assume that the model includes a hyperparameter π that can be interpreted as the overall rate of true comparisons, for example as $p(\delta_i = 1 \mid \pi) = \pi$, independently, $i = 1, \ldots, p$. Including π in the parameter vector and adding a hyperprior $p(\pi)$ allows us to learn about the overall level of noise and thus on all δ_i. If the data suggests that many comparisons are negative, then π is estimated to be closer to zero, and thus posterior inference for any particular δ_i is shrunk towards zero. On the other hand, if there is evidence that many comparisons are likely to be positive, then $E(\delta_i \mid \boldsymbol{y})$ is shrunk towards larger values. Compared with inference that fixes π or inference that does not include learning about the overall level of observed differences, posterior inference under the hierarchical Bayesian model can be said to adjust for multiplicities. In other words, a full Bayesian model can define a more general prior over the space of joint hypotheses $\boldsymbol{\delta} = (\delta_i,\ i = 1, \ldots, p)$ than many other approaches. The discussion in Scott and Berger (2010) is about hypotheses or model selection related to the mean of the observed outcomes. In contrast, inference about $\boldsymbol{\delta} = (\delta_{ij},\ i < j)$ in the differential graph model is related to inference about the dependence structure and comparing dependence structure across two conditions. However, the argument remains valid and explains how (3) adjusts for multiplicity.

Posterior Inference under the Differential Versus the Independent Graph Model The other important justification for proposing the differential graph model (3) over two independent graphs is that it better reflects biological reasoning and assumptions. In the motivating applications, the graphs correspond to conditional independence structure of protein activation under different biologic conditions. It is assumed a priori that the two graphs are different. In fact, we are interested in inference about how they differ. However, the differences are not expected to be many, making a hierarchical model that allows most edges to be unchanged across G_1 and G_2 more appropriate than two independent models. And most importantly, these differences between the two model choices matter. They can lead to very different posterior inference, as we show next.

Consider a stylized multiple comparison problem, with inference for a vector $\boldsymbol{\delta} = (\delta_i,\ i = 1, \ldots, p)$ of comparisons. Let $p_\mu(\boldsymbol{\delta})$ denote the joint prior under a hierarchical model with $p_\mu(\delta_i = 1 \mid \pi) = \pi$ and $p_\mu(\pi) = U(0, 1)$. The model p_μ is a stylized proxy for the differential graph model. Let $p_\nu(\boldsymbol{\delta})$ denote a model with $p(\delta_i = 1) = p_0$ for fixed p_0. The model is a stylized proxy for the independent graph model which fixes the probability of matching edges by implication of the independent prior on the graphs G_k. Under both models we assume the same sampling model $p(\boldsymbol{y} \mid \boldsymbol{\delta})$. Finally, let $p_\mu(\boldsymbol{\delta} \mid \boldsymbol{y})$

and $p_\mu(\boldsymbol{y})$ denote the posterior distribution and marginal model under μ, and similar for $p_\nu(\boldsymbol{\delta} \mid \boldsymbol{y})$ and $p_\nu(\boldsymbol{y})$. For any two probability models P and Q, let

$$\mathrm{KL}(P, Q) = \int \log \frac{P(x)}{Q(x)}\, dP(x)$$

denote the Kullback–Leibler (KL) divergence between P and Q. Posterior inference in the multiple comparison problem under p_μ and p_ν differs substantially. The KL divergence of the two posterior models diverges as the size of the comparison increases.

Theorem 1. *If $p_\nu(\boldsymbol{\delta} \mid \boldsymbol{y})/p_\nu(\boldsymbol{\delta}) < M_1$ and $p_\mu(\boldsymbol{\delta} \mid \boldsymbol{y})/p_\mu(\boldsymbol{\delta}) > M_0 > 0$ are bounded from above and from below, respectively, then*

$$KL\left[p_\mu(\cdot \mid \boldsymbol{y}),\ p_\nu(\cdot \mid \boldsymbol{y})\right] \to \infty$$

as $p \to \infty$, almost surely. Recall that in the application to inference for the edges in a graph, $p = m(m-1)/2$ is the number of edges. The proof is given in the appendix. The result is independent of the actual sampling model, as long as the stated condition holds. In words, the bounds on $p(\boldsymbol{\delta} \mid \boldsymbol{y})/p(\boldsymbol{\delta})$ say that the data must not be "unlimited informative". In other words, we require that the likelihood should not dominate the prior. These conditions are easily met for fixed and moderate sample sizes when the sampling model is Gaussian or Bernoulli. The theorem states that, for large enough networks, the posterior probabilities under the two models differ in KL divergence by arbitrary amounts.

4 Posterior Inference

4.1 Posterior MCMC Simulation

Posterior inference for model (1) is implemented as posterior Markov chain Monte Carlo (MCMC) simulation. We use $[x \mid y, z]$ to generically indicate a transition probability that changes x while conditioning on the currently imputed values of y and z. In writing the transition probabilities, we include only those conditioning parameters which appear in the kernel. Recall that $\boldsymbol{\delta}$ is a function of G_1 and G_2. MCMC posterior simulation proceeds by iterating over the following transition probabilities: $[\boldsymbol{\theta}^k \mid \beta^k, G_k, \boldsymbol{y}^k]$, $[\boldsymbol{\beta}^k \mid \boldsymbol{\theta}^k, G_k, \boldsymbol{y}^k]$ for $k = 1, 2$, $[\pi \mid \boldsymbol{\delta}]$, $[\boldsymbol{\delta}, \boldsymbol{\beta}^2 \mid \boldsymbol{\theta}^2, \boldsymbol{\beta}^1, G_1, \pi, \boldsymbol{y}^2]$, $[G_1, \boldsymbol{\beta}^1 \mid \boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \boldsymbol{\beta}^2, G_2, \pi, \boldsymbol{y}^1]$.

The transition probabilities $[\boldsymbol{\theta}^k \mid \ldots]$ update the parameters $\boldsymbol{\theta}^k$ in the sampling model. In both examples later, these transition probabilities are straightforward standard MCMC implementations. The transition probability $[\pi \mid \ldots]$ updates the global similarity parameter π. We use the complete conditional posterior probability. Let $m_1 = \sum_{i<j} \delta_{ij}$ and $m_0 = \sum_{i<j}(1 - \delta_{ij})$ denote the number of mismatches and matches between edges of the two graphs G_1 and G_2. We have $p(\pi \mid \delta) \propto \pi^{m_1+a-1}(1-\pi)^{m_0+b-1}$. We recognize this as the kernel of $Beta(m_1 + a, m_0 + b)$.

This above step, though computationally simple, is critical to the posterior inference scheme. Without this updating, π would be fixed. When π is fixed to 0.5, posterior inference would proceed just as in an independent model. Fixing π to a known value would make the networks dependent. However, the degree of borrowing would then be entirely dictated by the a priori information and ignore available data. In contrast, the proposed posterior updating step produces a more general inference by allowing a flexible interplay between the data and priors.

The remaining steps are to update $\boldsymbol{\beta}^k$ and G_k. Here we run across a critical computational hurdle in the form of normalizing constants. Let $c(\boldsymbol{\beta}^k, G_k)$ denote the normalization constant in (6). The complete conditional posterior distributions for $\boldsymbol{\beta}^k$ and G_k involve evaluation of this constant.

In general, these constants emerge as a challenging problem in many graphical models. Some approaches to solve this problem for GGMs using G-Wishart priors required approximation techniques like the Monte Carlo integration of Atay-Kayis and Massam (2005) and the Laplace approximation of Lenkoski and Dobra (2011). Alternative methods sample over the joint space of graphs and precision matrices through reversible jump procedures (Dobra et al., 2011; Giudici and Green, 1999; Lenkoski and Dobra, 2011). For these methods, suitable choice of tuning parameters remains a major challenge. Wang and Li (2012), alternatively, devised a completely novel MCMC exploiting the partial analytic structure (PAS) of G-Wishart distributions which resulted in automated proposal choices for the RJ steps. Moreover, computation of prior normalizing constants is avoided via the implementation of an exchange algorithm. A detailed review of these methods and the associated challenges for G-wishart priors appears in Wang and Li (2012). For Bayesian GGMs not based on G-Wishart priors, one can cleverly exploit a reparametrization of the precision matrix Ω and then impose a prior on the new parameters. Some good examples of this strategy can be found in Wong et al. (2003) and Wang (2012). They replace the problem of imposing priors on G by the specification of priors on partial correlations, which can be specified independently, without any constraints. Moreover, normalizing constants are required for each possible graph size rather than for every individual graph. The latter substantially reduces the scale of the problem. A parsimonious GGM proposed by Wang (2012) entirely does away with normalizing constants by avoiding model selection on the space of zero-constrained Gs. Instead, they induce a weaker version of parsimony through shrinkage priors on the individual elements of the inverse covariance matrix.

For autologistic models like (6), the same problem arises, that is, the evaluation of an analytically intractable normalization constant. Here, the constant is expressed as a sum over all possible m-dimensional binary vectors $\boldsymbol{v}_t \in \{0,1\}^m$. This is computationally intractable for the massively repeated evaluation that is needed in MCMC simulation. Though introduced decades earlier in Besag (1974), the scope of application of autologistic models has been limited due to these constants. Several techniques to approximate these constants have been suggested (Atchade et al., 2008; Moeller et al., 2006). We used an importance sampling technique that is described in Mitra et al. (2013) where we used the same second-order autologistic model for inference with a single graph. We briefly summarize the strategy and refer to Mitra et al. (2013) for details.

We first describe the implementation of a transition probability to change $\boldsymbol{\beta}^k$. For simplicity we drop the super-index ^{k} in the following argument. We implement an importance sampling estimate to approximate the ratio of the normalizing constants $c(\boldsymbol{\beta}, G)$

that are required for the evaluation of the acceptance probabilities in the Metropolis–Hastings transition probabilities to update $\boldsymbol{\beta}$. The use of importance sampling estimates to evaluate ratios of normalizing constants is discussed in Chen and Shao (1997), and more recently reviewed in Chen, Shao and Ibrahim (2000, Chapter 5). Let $p_v(\boldsymbol{v};\ \boldsymbol{\beta}, G)$ denote the autologistic probability (6), and let $K(\boldsymbol{v};\ \boldsymbol{\beta}, G)$ denote the un-normalized expression on the right-hand side of (6). We generate a proposal for a new $\tilde{\boldsymbol{\beta}}$ by drawing $\tilde{\beta}_i \sim q(\tilde{\beta}_i;\ \beta_i) = N(\beta_i, c)$. Next we sample M binary vectors $\boldsymbol{v}_i \sim p_v(\boldsymbol{v}_i;\ \boldsymbol{\beta}, G)$. By the law of large numbers, the sample average $R = \frac{1}{M}\sum_{i=1}^{M} K(\boldsymbol{v}_i;\ \tilde{\boldsymbol{\beta}}, G)/K(\boldsymbol{v}_i;\ \boldsymbol{\beta}, G)$ converges to $c(\tilde{\boldsymbol{\beta}}, G)/c(\boldsymbol{\beta}, G)$. We use R to approximate the ratio of the normalization constants $c(\tilde{\boldsymbol{\beta}}, G)/c(\boldsymbol{\beta}, G)$ that appears in the Metropolis–Hastings acceptance ratio for the proposal $\tilde{\boldsymbol{\beta}}$. In our experience, the importance sampling is fast and sufficiently accurate with an importance sampling size of $M = 5{,}000$.

We similarly construct another transition probability to update G_k in a Metropolis–Hastings step. An added complication is that a change in G_k requires to add or remove coefficients in the autologistic model (6). Implementation requires a transdimensional MCMC (Green, 1995). We construct a candidate $\widetilde{G}_k$ by adding or deleting an edge from G_k. We first describe the transition probabilities to update G_2.

Updating G_2 conditional on G_1 is equivalent to updating $\boldsymbol{\delta}$ since G_2 is a deterministic function of G_1 and $\boldsymbol{\delta}$. We update $(\delta_{ij}, \beta^2_{ij})$ one edge at a time. The transition probability implies a possible change in dimension of $\boldsymbol{\beta}^2$ when it involves a change of δ_{ij}. We use a reversible jump (RJ) MCMC implementation and jointly propose a candidate $(\tilde{\delta}_{ij}, \tilde{\beta}^2_{ij})$. Without loss of generality, assume that the currently imputed state is $\delta_{ij} = 0$ and we propose $\tilde{\delta}_{ij} = 1$. Thus the parameter vector under the proposed new state is incremented by an additional coefficient β^2_{ij}. We generate $\tilde{\beta}^2_{ij}$ from a normal proposal distribution $q(\tilde{\beta}^2_{ij})$. The joint proposal $(\tilde{\delta}_{ij}, \tilde{\beta}^2_{ij})$ implicitly defines a proposal $\tilde{G}_2$. Let $\widetilde{\boldsymbol{\beta}}^2$ denote $\boldsymbol{\beta}^2$ with β^2_{ij} replaced by $\tilde{\beta}^2_{ij}$. The acceptance probability becomes $\min\{1, A\}$ with

$$A = \frac{p(\boldsymbol{v}^2 \mid \tilde{G}_2, \widetilde{\boldsymbol{\beta}}^2)}{p(\boldsymbol{v}^2 \mid G_2, \boldsymbol{\beta}^2)} \frac{p(\tilde{\beta}^2_{ij})}{q(\tilde{\beta}^2_{ij})}. \tag{8}$$

The Jacobian is 1 since the proposal involves no deterministic transformation. Evaluation of A again requires the normalization constant $c(\widetilde{\boldsymbol{\beta}}^2, \widetilde{G}^2)$. We proceed as before with the importance sampling method. The RJ acceptance probability can now easily be evaluated. Updating G_1 and $\boldsymbol{\beta}^1$ conditional on currently imputed values of G_2 proceeds similar to the above step.

4.2 Posterior Summaries

One of the desired inference summaries are estimates for the graphs G_1 and G_2. We report graphs $\bar{G}_1$ and $\bar{G}_2$ based on marginal inclusion probabilities. Let $\bar{P}^k_{ij} = p(\{i, j\} \in E_k \mid \boldsymbol{y})$ denote the posterior probability of edge $\{i, j\}$ being included in graph G_k. We report graphs $\bar{G}_k$ including all edges with $\bar{P}^k_{ij} > \lambda$ for some threshold λ. That is, we report estimates

$$\bar{G}^k_{ij} = I(\bar{P}^k_{ij} > \lambda). \tag{9}$$

The threshold is set to control the posterior expected false discovery rate (FDR) for edge inclusion,

$$\mathrm{FDR}_\lambda = \frac{\sum_{i\ j}\left[(1-\bar{P}_{ij})I(\bar{P}_{ij} > \lambda)\right]}{\sum_{i,j} I(\bar{P}_{ij} > \lambda) + \epsilon},$$

adding ϵ to avoid division by zero.

The main inference targets are the δ_{ij}. We summarize posterior inference on δ_{ij} by a similar argument as $\bar{P}^k_{ij}$. We use a rule that reports an edge as different across graphs when

$$\bar{\delta}_{ij} = I\left[p(\delta_{ij} = 1 \mid \boldsymbol{y}) > \kappa\right]. \tag{10}$$

This is similar to the rule $\{\bar{P}^k_{ij} > \lambda\}$ that we use to report estimated graphs. Only now the rule is deciding the report of edges as different across conditions. Again, κ can be chosen to control the FDR for the multiple comparison problem.

In summary, posterior inference is implemented with reversible jump MCMC simulation. Computation is not trivial, but does not involve any non-standard techniques. Inference is computation intensive. However, conditional on imputed values of $\boldsymbol{\beta}^k$ and $\boldsymbol{\theta}^k$, the implementation of the transition probabilities to update π and the two graphs are linear in p (recall that p is the number of edges). In other words, the joint model adds an $O(p)$ term to the computational cost, compared to inference for a single graph.

5 Simulations

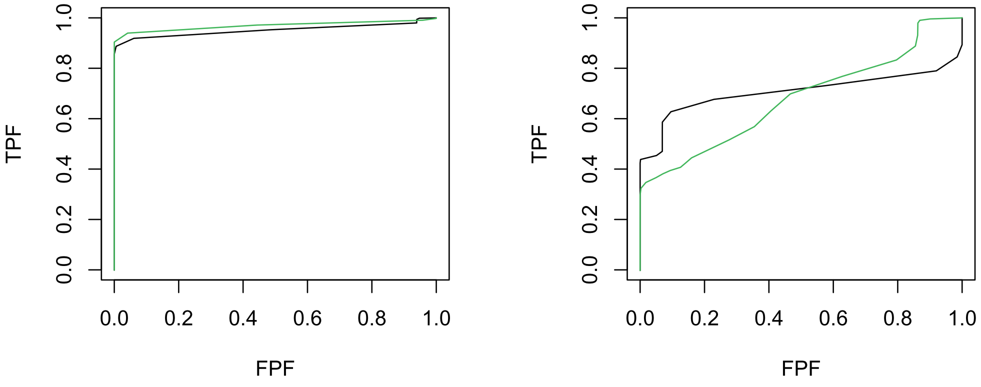
We set up a simulation experiment to validate the proposed model. For each simulated data set we carried out inference under four models: (1) the proposed model (*differential graph model*); (2) a model with two independent priors for G_1 and G_2, identical to $p(G_1)$ in (3) (*independent graph model*); (3) joint graphical inference by Guo et al. (2011); (4) joint graphical inference by Danaher et al. (2013); and (5) Independent graphical lasso.

The Beta hyper-parameters of the proposed differential model (3) were set as $a = 11$ and $b = 1$. We fixed the number of observations for subgroup 1 at 330 and subgroup 2 at 48. The graph G_1 was generated by setting up vertices for $m = 7$ nodes. For each pair of nodes $\{i, j\}$ we included an edge between them with probability $p = 0.5$. For each imputed edge $\{i, j\}$ we generated values of β^1_{ij} using a discrete uniform prior over three possible values, $\beta^1_{ij} \sim \mathrm{Unif}(\{\log(2), \log(4), -\log(2)\}$. These values were chosen arbitrarily. Next, we used π to generate G_2 from the conditional prior distribution $p(G_2 \mid G_1, \pi)$. In the simulation truth, we used several choices of π. Values are indicated in the upcoming tables of results.

Hypothetical data $\boldsymbol{y}^k$ was generated using the sampling model (5) with latent binary indicators generated as in (6). We fixed the model parameters $\boldsymbol{\theta}^k$ with $\mu_{1ik} \sim N(4, 0.2)$, $\mu_{2ik} \sim N(1, 0.2)$ and $\sigma_{1ik} = \sigma_{2ik} = 0.1$. We generated 20 hypothetical datasets under this assumed sampling model. We then evaluated inference under the proposed differential graph model with sampling model (5) and (6).

To assess model performance, recall rule (10) that reports an edge as positive when $\bar{\delta}_{ij} = I[p(\delta_{ij} = 1 \mid \boldsymbol{y}) > \kappa]$. As we vary the threshold κ, we generate a family of decision rules. The receiver operating characteristic (ROC) curve for each model plots sensitivity versus the false positive rate as we vary κ. The area under the ROC curve (AUC) along with misclassification rates is often used as a summary to compare competing classification rules. For each dataset we computed AUC for the five models (1) through (5). The frequentist lasso methods (3)–(5) required specifying the glasso penalization parameter ρ which we set to 0.03. Inference under models (3) and (4) was implemented using the *glasso* package and *jgl* in R, while (5) was executed with code obtained from Guo et al. (2011). For each method we recorded 4 measures of model performance, including the AUC for estimating G_1 (AUC1 in Table 1); AUC for estimating G_2 (AUC2); the average across these two (AUC12); and ER_κ, as the error rate for detecting differential edges obtained under a given posterior probability threshold κ in (10). The latter is evaluated as the proportion $ER_\kappa = \frac{1}{p}\sum_{i;ji<j}[\bar{G}_{ij}^k \neq \delta_{ij}^o]$ where δ_{ij}^o is the simulation truth and $p = m(m-1)/2$ is the number of possible edges. The last two summaries, AUC12 and ER_κ, provide a combined summary of how the model jointly estimates the pair of networks. AUC12 focuses on average performance over individual graphs while ER_κ addresses the detection of differential edges. We fixed $\kappa = 0.75$ throughout. The ROC curves for the frequentist methods are obtained by thresholding the values of the estimated inverse covariance matrix at different cutoff values. Each cutoff yields a binary matrix of estimated differences, which is then used to compute the corresponding sensitivity and specificity. The average AUC values for all methods along with their standard errors (across repeat simulation) are summarized in Table 1. We observe that the differential prior compares consistently favorably with the independent prior in terms of combined accuracy and the estimation of G_2.

Figure 1 shows smoothed ROC curves under the two models for estimating both graphs for one of the sample data sets. The smoothed curves are based on kernel density



(a) Group 1 (b) Group 2

Figure 1: ROC curves for a simulated data set. The green and black curves represent the operating characteristics of the differential graph model (solid) and the independent graph model (dotted), respectively.

estimates of the distribution of $\bar{\bar{\delta}}_{ij}$ for both, true positives and true negatives. For more details we refer the readers to Lloyd (1998).

We next varied the penalization parameters of the frequentist methods and found that performance was very sensitive to these values. Specifying an optimal value of λ remains a challenge. In interpreting the overall comparison, one should, however, keep in mind that these approaches were never specifically intended for a comparison of edges in two graphs and are more focussed on the shrinkage of graph coefficients.

Performance	Diff-Bayes	Ind-Bayes	Glasso	JGL	Guo
AUC1	0.95 (0.06)	0.97 (0.04)	0.97 (0.04)	0.97 (0.04)	0.89 (0.11)
AUC2	0.84 (0.10)	0.68 (0.13)	0.68 (0.12)	0.71 (0.11)	0.65 (0.13)
AUC12	0.90 (0.08)	0.83 (0.07)	0.83 (0.03)	0.84 (0.03)	0.77 (0.03)
ER_κ	0.118 (0.04)	0.82 (0.06)	0.119 (0.06)	0.119 (0.07)	0.12 (0.10)

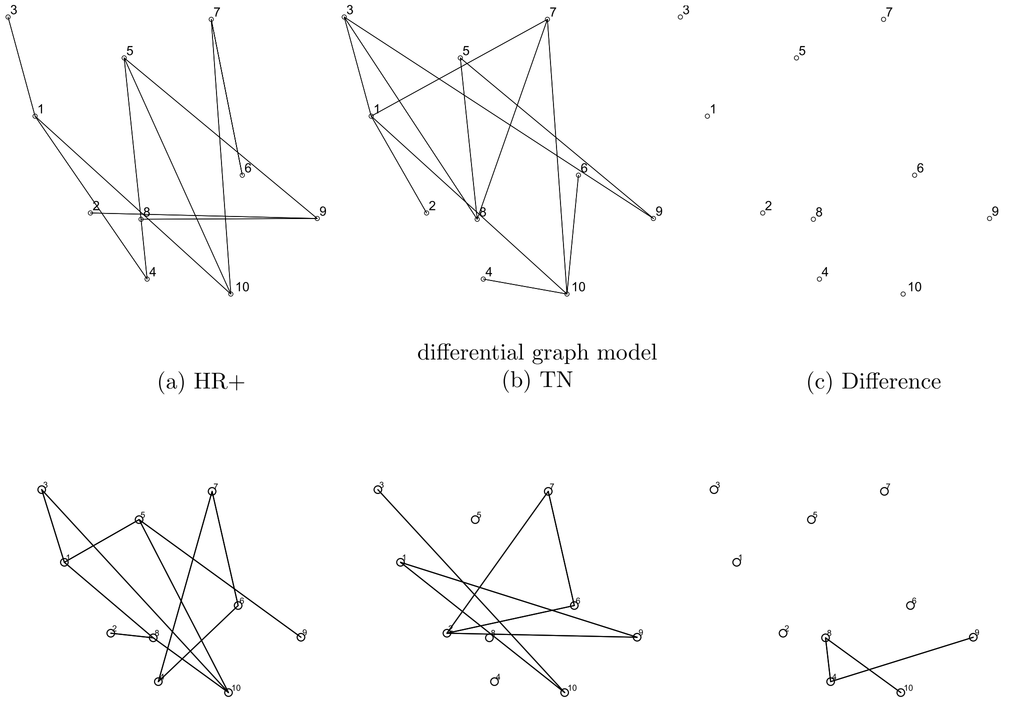
Table 1: Comparing differential prior model against independent priors and other frequentist alternatives. In parentheses, standard deviations (..) over repeat simulations.

Besides the comparison of Table 1, several other considerations lead us to favor the proposed approach when the primary goal is a comparison of dependence structure across two conditions. First, the Bayesian paradigm allows the incorporation of prior expert knowledge, when available. Second, we model the differential structure directly on the space of latent graphs, rather than relying on features of an assumed sampling model. This makes the approach very flexible. For example, the sampling model could be replaced by any alternative sampling model without substantially changing the implementation of posterior simulation. Lastly, the Bayesian approach includes a full probabilistic description of uncertainties as the posterior distribution $p(\boldsymbol{\delta} \mid \boldsymbol{y})$. Overall, the joint estimation of differential pathways in the differential model allows improved inference on differences across the two graphs. The relative advantage over independent analyses decreases when sample sizes increase (simulations not shown). However, inference under the differential prior provides substantial gains in AUC (and a significantly lower error rate) under unequal and lower sample sizes. Asymptotically, as both sample sizes increase, and the data essentially reveals the true graphs, both models achieve an AUC of 100%.

6 Case Studies

6.1 RPPA Data

Reverse Phase Protein Arrays (RPPAs) is a recently developed high-throughput technology that is designed to measure protein activation for many samples simultaneously. A typical RPPA experiment starts out with a mixture of cultured cells from patient samples. These cells are treated with therapeutic agents. Proteins extracted from these cells are then fixed onto a slide. A typical slide usually consists of thousands of individual patient samples. Investigators wanting to study a particular pathway design an RPPA experiment with each array on the slide hybridized against an antibody that



differential graph model

(a) HR+ (b) TN (c) Difference

independent model

(d) HR+ (e) TN (f) Difference

Figure 2: RPPA data. Posterior graphs and estimated differences δ_{ij} under the differential graph model (a,b,c) and the independent graph model (d,e,f). Inference under the differential prior reports no uncommon edges between the two graphs while the independent prior selects 3 edges.

binds to a targeted protein of interest. Therefore, each RPPA produces a data set of measurements for one protein across multiple samples.

We analyzed data from an RPPA experiment on the mitogen-activated protein kinase (MAPK) pathway from breast cancer cell lines. The dataset consists of measurements of 10 proteins selected from the MAPK pathway for 255 patient samples. Patients were classified into three clinical groups based on the activation status of three biomarkers HR+, Triple Negative(TN+) and HER2+. The sample sizes for the three subgroups samples were 139, 63 and 53 patients, respectively. The goal was to estimate differences in the protein networks for each pair of these three subgroups by combining prior knowledge of protein interaction and RPPA measurements.

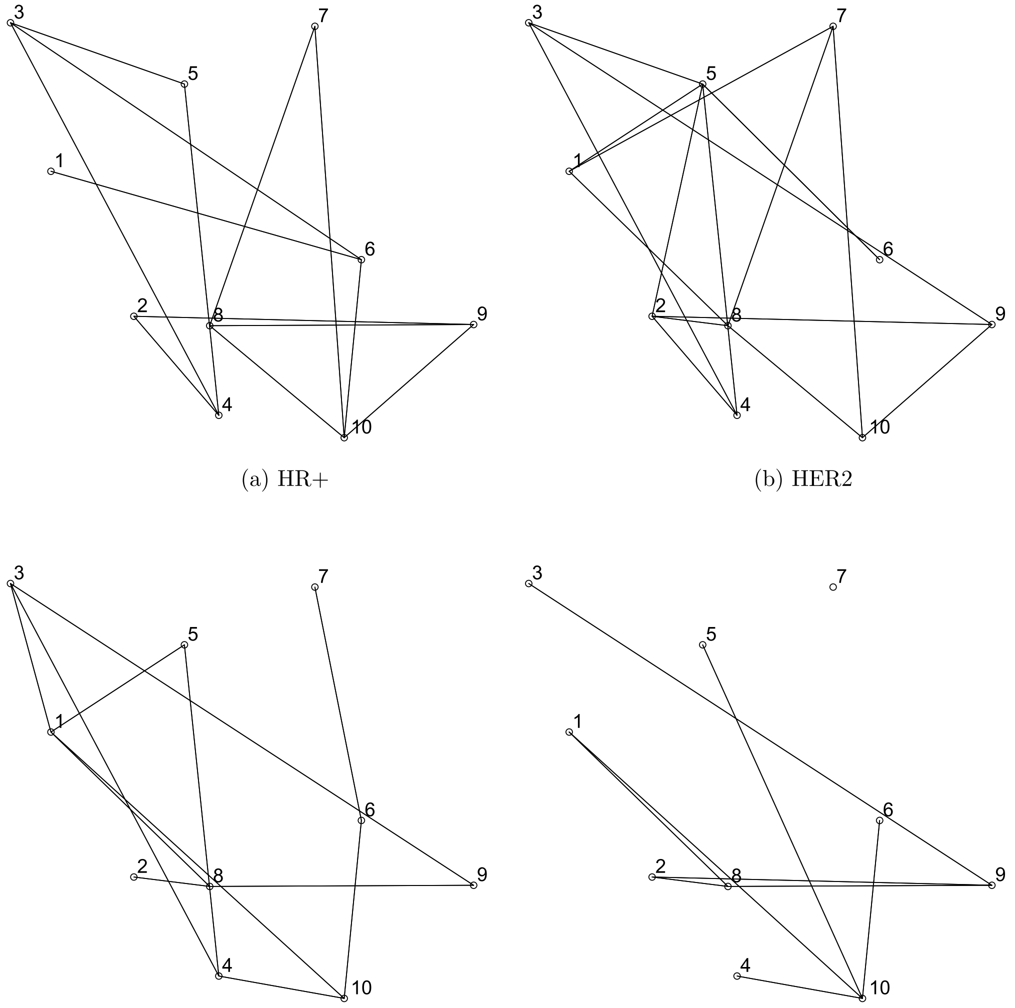
We first focus on comparing TN versus HR+. We carried out MCMC posterior simulation, using 16,000 iterations in total, and keeping the last 8,000 iterations to evaluate posterior summaries. The total time for running these simulations was 23 minutes on a Dell Optiplex 980 desktop computer. Figure 2 summarizes posterior inference for the comparison of HR+ versus TN. Figures 2(c,f) include all edges with $\bar{\delta}_{ij} = 1$ based on a FDR criterion of 0.01. Figure 2(c) shows the reported differences under the differential graph model for (G_1, G_2). Panel 2(f) shows the same inference under two independent models for G_1 and G_2. The additional edges that are reported by the independent model are due to the un-adjusted high posterior probabilities $\bar{P}_{ij} = p(\delta_{ij} = 1 \mid \boldsymbol{y})$. In contrast, the differential path model learns that most edges remain unchanged across conditions and shrinks inference on δ_{ij} towards 0. For reference, Panels 2(a,b,d,e) show the estimated graphs $\bar{G}^k_{ij}$.

Next, Figure 3 shows the posterior estimated networks $\bar{G}_k$ under the remaining two comparisons, that is HR+ versus HER2, and TN versus HER2. The cutoff in $\bar{G}^k$ was chosen to achieve a posterior expected FDR of 0.01. From inspection of the two estimated graphs in Figure 3 alone, it is not clear which differences should be reported. And it is impossible to attach probabilities to any such report. In these two particular comparisons, inference under the differential graph model finds no edges to be significantly different (again under FDR control at 0.01). The strength of the proposed model is that it facilitates such inference as straightforward summaries of the posterior distribution $p(\boldsymbol{\delta} \mid \boldsymbol{y})$.

6.2 ChIP-Seq Data of Histone Modifications

Histones are proteins that wrap short segments of DNA (about 140 base pairs) around small spherical structures called nucleosomes. A nucleosome consists of an octamer of four core proteins (two sets of H2A, H2B, H3, and H4). Post-translational modifications of these proteins by methyl, acetyl and phosphoryl groups significantly influence important biochemical processes such as gene activation, nucleosome assembly and higher-order chromatin packing. The correlation of these histone modifications (HMs) with translational activity and occurrence of promoters has been well documented in Barski et al. (2007), among many others. By influencing gene expression, HMs can overwrite the inscribed genetic code. Therefore, HMs can be said to be epigenetic markers that share the importance of DNA in explaining heredity. Recent research has been increasingly suggestive of a fundamental association between HMs and the pathology of some major diseases. For example, patterns of HMs have been found to be important predictors of cancer prognosis and subsets of HMs are used as potential informants of clinical decisions (Kurdistani, 2007, 2011). Both, global patterns in HMs and their cellular heterogeneity have been considered in building therapeutic regimens. A comprehensive list of HMs appears at `http://bioinfo.hrbmu.edu.cn/hhmd` (Zhang et al., 2010).

Despite their obvious importance, knowledge about the association of HMs with translational processes and disease markers still remains restricted to marginal association, in the sense that most reported associations are for individual HMs. The mechanism of co-localization of multiple HMs and its relation to transcription remains largely unknown. The famously hypothesized *histone code* (Strahl and Allis, 2000) suggests that the presence or absence of these HM co-localizations regulates gene transcription combinatorially. Since the emergence of this hypothesis, several experimental results have provided strong evidence for a cross talk mechanism between HMs. Many plausible mechanisms of co-localization have also been uncovered, e.g., the concurrent activity



(a) HR+

(b) HER2

(c) TN

(d) HER2

Figure 3: RPPA data. Posterior estimated graphs $\bar{G}_1$ and $\bar{G}_2$ for the comparison of HR+ versus HER2 (a,b) and for TN versus HER2 (c,d). $\bar{G}_k$ shows the estimated conditional independence structure for protein activation in the MAPK pathway for the respective comparisons.

of several enzymes on different parts of the histone tails. It is now known that the distributions of HMs vary with genomic locations, and there are attempts to demarcate functional domains over the genome by signatures of histone patterns (Liu et al., 2005). We go a step further and aim to demonstrate that functionally diverse genomic regions (e.g., regions of high and low gene expression) are characterized by different cross-talk mechanisms. Our model-based approach formalizes this dependence through difference graphs between two networks.

Data for HMs was obtained from a ChIP-seq experiment for CD4 positive T lymphocytes (Wang et al., 2008; Barski et al., 2007). The data reported counts for $m = 39$ types of HMs including 18 acetylations, 20 methylations, and one special histone modification H2A.Z at a given set of genomic locations. A high count indicated an enrichment of the HM at the corresponding genomic location. Here, genomic location refers to a segment of DNA of around 2,000 base pairs. In Mitra et al. (2013) we discussed inference for the same data using a model for a single population, using the marginal model for $\boldsymbol{y}^1$ that is implied by (1). The upcoming results under the proposed differential graph model extended this analysis to inference across two subpopulations, as we explored differences in dependence structure of HMs across regions corresponding to high and low transcription. To achieve this goal we restricted data to the HM counts in known promoter regions that were related to specific genes. Then we separated the genomic region belonging to promoters into protein coding and non-coding regions. This defined the two groups. In each group we had count data for all 39 HMs.

For the HM data we used a slight modification of the sampling model (5), replacing the normal sampling model with a mixture of log normal model. Let $LN(\mu, \sigma)$ denote a log-normal distribution with location parameter μ and scale parameter σ. We assume

$$p(y_{kti} \mid \boldsymbol{\theta}^k, v_{kti}) \propto \begin{cases} \mathrm{LN}(\mu_{1ik}, \sigma^2_{1ik}) & v_{kti} = 0, \\ \mathrm{LN}(\mu_{2ik}, \sigma^2_{2ik}) & v_{kti} = 1. \end{cases} \tag{11}$$

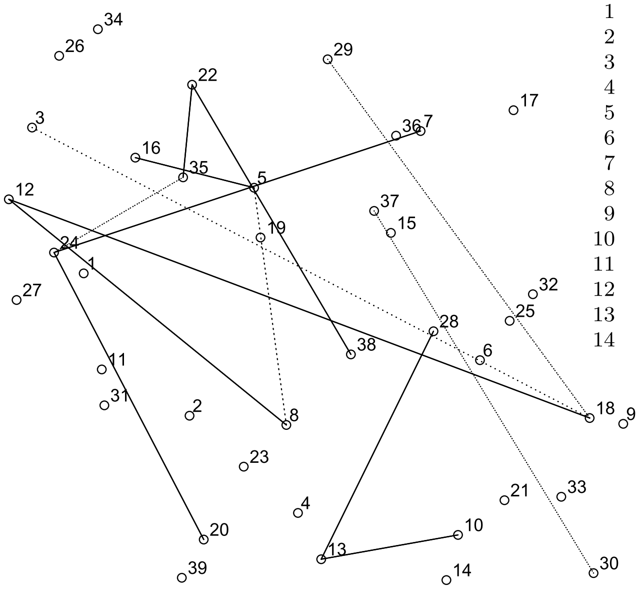
The autologistic model (6) remains unchanged. In the context of the application to HM data, the indicators v_{kti} are interpreted as indicators for the presence of histone modification i in sample t under condition k. The primary aim of the study is inference about the difference across conditions $k = 1, 2$ of the dependence structure of v_{kti}, $i = 1, \ldots, m$.

We implemented inference under the proposed differential graph model by simulating a total of 8000 iterations of the earlier described MCMC posterior simulation. This took approximately 8.5 hours on a Dell Optiplex 980 desktop computer. Figure 4 shows the estimated differences δ_{ij} in the dependence structure across the high versus low expression regions. In the figure, we indicate the different HMs names with running indices 1 through 39, as indicated.

To select edges to be reported in the difference graph of Figure 4 we used a variation of the rule in (10). First we noted that $\bar{\delta}_{ij}$ in (10) can be justified as a Bayes rule under the loss function

$$L(\boldsymbol{\delta}, \mathbf{d}) = \sum_i \sum_j [d_{ij}(1 - \delta_{ij}) + c(1 - d_{ij})\delta_{ij}], \tag{12}$$

where $\mathbf{d} = \{d_{ij}\}$ is a set of decision rules in which $\{d_{ij} = 1\}$ denotes the decision to declare that there exists a difference between the two graphs at the edge $\{i, j\}$. Under $L(\boldsymbol{\delta}, \mathbf{d})$, $d^*_{ij} = \bar{\delta}_{ij}$ is the Bayes rule. The tradeoff c determines the threshold κ in (10). See, for example, Müller et al. (2007) for a discussion of this interpretation



1	H2AK5ac	15	H4K16AC
2	H2AK9ac	16	H4K5AC
3	H2BK120ac	17	H4K8AC
4	H2BK12ac	18	H4K91AC
5	H2BK20ac	19	H2AZ
6	H2BK5ac	20	H3BK5ME1
7	H3K14AC	21	H3K27ME1
8	H3K18AC	22	H3K27ME2
9	K23AC	23	H3K27ME3
10	H3K27AC	24	H3K36ME1
11	H3K36AC	25	H3K36ME3
12	H3K4AC	26	H3K4ME1
13	H3K9AC	27	H3K4ME2
14	H4K12AC	28	H3K4ME3
		29	H3K79ME1
		30	H3K79ME2
		31	H3K79ME3
		32	H3K9ME1
		33	H3K9ME2
		34	H3K9ME3
		35	H4R2ME1
		36	H4R2ME2
		37	H4K20ME1
		38	H4K20ME3
		39	H4R3ME2

Figure 4: HM data. Estimated differences between graphs G_1 and G_2. The solid edges are present in the high expression network G_1 and not in low expression network G_2. The dashed edges are present in low expression network and not in the high expression network. The HM names corresponding to nodes 1 through 39 are listed.

of thresholding rules. Understanding the underlying loss function (12), one could now criticize the rule (10). In particular, the loss function penalizes all false negatives (second term in $L(\boldsymbol{\delta}, \bar{\delta})$) equally with a penalty c. This interpretation leads us to adopt an alternative rule. Let $m_{ij} = \delta_{ij}|\beta^1_{ij} - \beta^2_{ij}|$ denote the (true) extent of differential strength in edge $\{i, j\}$. We use m_{ij} to weigh false negatives differently, in the loss function

$$L(\boldsymbol{\delta}, \mathbf{d}, \boldsymbol{\beta}) = \sum_i \sum_j [-d_{ij}m_{ij} + c_1(1 - d_{ij})m_{ij} + c_2 d_{ij}].$$

Here c_1 and c_2 are two fixed constants. In words, false negatives $(1 - d_{ij})\delta_{ij}$ and true positives $d_{ij}\delta_{ij}$ are weighted by the size of the difference m_{ij}. The relative weights of true positives, false negatives and sampling cost are $-1, c_1$ and c_2, respectively. It is easy to show that the optimal rule d^*_{ij} is given by

$$\widehat{d}_{ij} = I\left[E(m_{ij} \mid \boldsymbol{y}) > c_2/(1 + c_1)\right].$$

In other words, we include edges in the difference graph by thresholding the posterior probability of the differences δ_{ij} weighted by the strengths β_{ij} of the edges. The rule $\widehat{d}$ is a variation of the earlier introduced rule (10), replacing the posterior mean of δ_{ij} by a weighted quantity $m_{ij} = \delta_{ij}|\beta^1_{ij} - \beta^2_{ij}|$. In our implementation, we use a cutoff

$c_2/(1+c_1) = 0.6$. Figure 4 shows $\widehat{d}_{ij}$. For ease of display, we used a numerical index for each HM (instead of its full name) in plotting the graph.

In summary, most edges occur with high posterior probabilities in both networks. Some of the common edges are between the variants of the same type of HM. For example, edges (H3K27me3, H3K27me2) and (H4R2me1, H4R2me2) are common to both networks. Apart from these, the list of significant common edges include edges between H3K4me and H3K9me methylation groups. The following edges correspond to the solid lines in Figure 4: (H3K36ac, H3K18ac), (H4K8ac, H3K4ac), (H3K4ac, H3K27ac), (H4K16ac, H2BK20ac), (H3K36ac, H4K91ac), and (H2BK12ac, H4K5ac). The following is a list of some positive edges in the low expression network that do not appear in the high expression network. These correspond to the (black) dotted lines in Figure 4: (H4K91ac, H3K4me3), (H3K9me3, H3K36me1), (H3R2me2, H3K79me2), (H4K8ac, H3K79me1), (H3K79me1, H4K20me1), and (H2BK12ac, H3K18ac). Finally, we report HMs with high posterior probabilities of high connectivity, defined by 8 or more edges. In both the high and low expression networks, the top connecting HMs include H3K4ac and H3K9ac. Both of these HMs are known to be associated with transcriptional activity in promoter regions. The activating mark H3K4me1 connects to a large number of nodes in the high expression network. The top connecting HMs in the low expression network include H4K8ac and H3K27me3.

7 Conclusion

We propose a model for joint inference on dependence structure G_1 and G_2 in two related subgroups. The main goal is inference about relative differences of the two dependence structures. In the application to protein activation, such inference formalizes the notion of pathway activation and disruption in one subpopulation relative to the other. In the application to histone modification counts, inference on relative differences between G_1 and G_2 gets us closer to understanding what is known as the histone code, i.e., epigenetics marks of HMs that reveal the regulatory mechanisms of HMs on gene expressions.

Among the limitations of the proposed model are the restriction to low and moderate size graphs, the restriction to comparing two graphs, the computation-intensive estimation and the lack of informative priors in the current implementation.

First, the proposed inference is only suitable (and intended) for problems with a moderate number of nodes, say, $m \leq 40$. The joint graphical prior only mitigates the challenge of inference for high-dimensional problems, but does not entirely solve it. Second, we restrict discussion to two related subpopulations, simply because this is all we need for the two motivating case studies. Some straightforward extensions to multiple subpopulations are possible. For example, assume that subpopulation 1 is the reference population. Sets of parameters $(\boldsymbol{\delta}^k, \boldsymbol{\beta}^k)$, $k = 2, \ldots, K$, could be used to augment the model to $(G_1, \ldots, G_k)$. Third, implementation requires MCMC simulation, with trans-dimensional transition probabilities to allow for the addition and deletion of edges. We use a reversible jump MCMC scheme. Such MCMC schemes are easy to describe, but notoriously difficult to implement. An efficient implementation requires attention to many housekeeping details. Finally, we described the model with generic priors $p(G_k)$ that did not make use of prior expert opinion beyond centering the graph around a prior guess G_0. However, nothing changes in the remaining discussion if $p(G_k)$ were replaced by more informative models.

A natural next step would be to embed such prior models in a larger framework where the subgroups are unknown a priori. The goal would be then to estimate the subgroup categories and the graphs simultaneously. The work of Rodriguez et al. (2011) is a significant step in that direction. They used a DP mixture of GGMs for this purpose. However, unlike us, they did not focus on differential edges between the graphs conditioned on subgroups.

The model can also be extended in several directions to try to locate finer differences and similarities between two graph topologies. One modification would be to center the δ_{ij} not around a global mean, but around a mean specific to a local neighborhood. A local neighborhood of an edge, or a pair of vertices, could be defined as subset of vertices connected to both i and j, or some other criterion. The newly defined means for the subgraphs (could be overlapping) could be independent or could themselves be centered.

Appendix

Proof of Theorem 1. We first introduce some notation. We make use of an approximation of the sampling model under p_ν by a normal distribution, after appropriate standardization. For any π let Y_π^p denote the standardized version of a binomial variable $Y \sim \text{Bin}(p, \pi)$, with mean $p\pi$. Let q_p be the p.m.f of Y_π^p. By the Central Limit Theorem, $q_p \to \text{N}(0, 1)$ weakly and also in probability.

In the following argument, we need a slightly different mode of convergence. We say p_ν is LR-convergent if $\log(q_p/\phi)$ converges to 0 uniformly, where ϕ is the density of the standard normal distribution. Let $s_p = \sqrt{\pi(1-\pi)/p}$. The maximum and minimum of the support of q_p is $a_p = (1-\pi)/s_p$ and $b_p = -\pi/s_p$. The distance between two consecutive point masses in the support of q_ν is $d_p = s_p$. In the proof, we will use that a_p and b_p are $O(\sqrt{p})$ and d_p is $O(\sqrt{1/p})$.

We begin with a general form of KL divergence between two posteriors. Since the space of possible graphs or binary configurations is discrete, the integral in the formula of KL divergence reduces to a summation:

$$\sum_\delta p_\mu(\delta \mid \boldsymbol{y}) \log\left\{\frac{p_\mu(\delta \mid \boldsymbol{y})}{p_\nu(\delta \mid \boldsymbol{y})}\right\} = \sum_\delta \frac{p_\mu(\delta \mid \boldsymbol{y})}{p_\mu(\delta)} p_\mu(\delta) \log\left\{\frac{p_\mu(\delta)}{p_\nu(\delta)}\frac{p_\nu(\boldsymbol{y})}{p_\mu(\boldsymbol{y})}\right\} =$$
$$= \sum_\delta \left[\frac{p_\mu(\delta \mid \boldsymbol{y})}{p_\mu(\delta)} p_\mu(\delta) \log\left\{\frac{p_\mu(\delta)}{p_\nu(\delta)}\right\}\right] - \log\{B(\boldsymbol{y}, \mu, \nu)\}$$

Here $B(\boldsymbol{y}, \mu, \nu)$ is the Bayes factor of μ with respect to ν. Since$\frac{p_\mu(\delta|\boldsymbol{y})}{p_\mu(\delta)} > M_0$, the first term is greater than M_0 KL(p_μ, p_ν).

We now produce a lower bound for the second term, $-\log\{B(\boldsymbol{y}, \mu, \nu)\}$. To see this, write the Bayes factor as $B(\boldsymbol{y}, \mu, \nu) = \sum_\delta p_\mu(\delta)\, p_\nu(\delta \mid \boldsymbol{y})/p_\nu(\delta)$. Using $\sum_\delta p_\mu(\delta) = 1$ and the upper bound M_1 for $p_\nu(\delta \mid \boldsymbol{y})/p_\nu(\delta)$, we conclude $-\log\{B(\boldsymbol{y}, \mu, \nu)\} > -\log(M_1)$.

Thus the divergence of the posteriors would primarily be driven by the divergence of the priors under the given assumptions. We need to show only lower bounds since our aim is to show divergence to positive infinity.

The KL divergence for the priors can be written as

$$\mathrm{KL}(p_\mu(\boldsymbol{\delta}), p_\nu(\boldsymbol{\delta})) = \sum_{\boldsymbol{\delta}} p_\mu(\boldsymbol{\delta}) \log \frac{p_\mu(\boldsymbol{\delta})}{p_\nu(\boldsymbol{\delta})}.$$

For each $\boldsymbol{\delta}$ let $k_\delta = \sum_{ij} \delta_{ij}$ denote the number of non-zero elements, and recall that m is the number of edges in the graph. Then

$$p_\mu(\boldsymbol{\delta}) = \int_0^1 \lambda^{k_\delta}(1-\lambda)^{p-k_\delta}\, dp = \frac{k_\delta!(p-k_\delta)!}{(p+1)!}.$$

This distribution induces a uniform prior for the number of non-zero entries k_δ

$$p_\mu(k_\delta = k) = \frac{p!}{k_\delta!(p-k_\delta)!}\frac{k_\delta!(p-k_\delta)!}{(p+1)!} = \frac{1}{p+1}.$$

Now $p_\nu(\boldsymbol{\delta}) = p_0^{k_\delta}(1-p_0)^{p-k_\delta}$ and thus $p_\nu(k_\delta = k) = C_{p,k}\, p_0^k(1-p_0)^{n-k}$ where $C_{p,k} = p!/\{k!(p-k)!\}$ is the binomial coefficient or the number of size k subsets of a set of p items. Partitioning the sum in the KL expression into subsets which have the same value of k_δ we have

$$\mathrm{KL} = \sum_{k=0}^{p} \frac{1}{p+1}\ \log\left\{\frac{1/(p+1)}{C_{p,k}\, p_0^k(1-p_0)^{n-k}}\right\}.$$

Thus the divergence is equal to $\mathrm{KL}(X, Y)$ where X is a discrete uniform random variable on $\{0, 1, \ldots, p+1\}$ and $Y \sim \mathrm{Bin}(p, p_0)$. The priors μ and ν induce a uniform and a binomial distribution, respectively, on the same support. Since a one–one transformation preserves KL, we can further write this as $\mathrm{KL}\{T(X), T(Y)\}$ where $T(x) = \sqrt{p}\{\frac{\frac{x}{p} - p_0}{\sqrt{p_0(1-p_0)}}\}$ Thus, we can rewrite $\mathrm{KL}(p_\mu(\cdot), p_\nu(\cdot))$ as $\mathrm{KL}(X', Y^p_{p_0})$ where $Y^p_{p_0}$ is a standardized version of binomial variable with mean p_0 and X' is a discrete uniform variable on the same support. Using the earlier defined notation of q_p for the p.m.f. of a standardized binomial variable, we have

$$\mathrm{KL} = \frac{1}{p+1}\sum_{i=0}^{p} \log\left\{\frac{(1/q_i)}{p+1}\right\} = -\log(p+1) - \frac{1}{p+1}\sum_{i=0}^{p}\log(q_i).$$

From Lemma 1 below, the second term is $O(p)$. It is precisely $-(K+z_p)\{\frac{p^2}{1+p}\}+c_p$ where K is a constant and the sequences z_p and c_p go to 0. From Lemma 1, the second term is $O(p)$. It is precisely $-(K+z_p)\{\frac{p^2}{1+p}\}+c_p$ where K is a constant and the sequences z_p and c_p go to 0. Now $\log(p+1)/p \to 0$ as $p \to \infty$. In other words, p diverges faster than $\log(p+1)$. Therefore, $\mathrm{KL} = O(p)$, which completes the proof. □

Lemma 1. *Consider q_p as defined above. Then the average log-probability mass function of q_p defined by $Q = \frac{1}{p+1}\sum_{i=0}^{p}\{\log(q_i)\}$ is $O(p)$.*

Proof. Let ϕ_i denote the normal density evaluated at the ith point mass in the support of $Y_{p_0}^{p}$. Then the average probability mass function can be written as

$$Q = \frac{1}{p+1}\sum_{i=0}^{p}\log\left\{\frac{q_i}{\phi_i}(\phi_i)\right\} = \frac{1}{p+1}\sum_{i=0}^{p}\log\frac{q_i}{\phi_i} + \frac{\frac{1}{p+1}}{d_h}\sum_{a_p}^{b_p} d_h\log(\phi_i).$$

From the assumption on LR convergence the first term tends to 0. The second term can be written as

$$\frac{1/(p+1)}{d_h}\left[\sum_{0}^{p} d_h\log(\phi_i) - \int_{a_p}^{b_p}\log(\phi)\right] + \frac{1/(p+1)}{d_h}\int_{a_p}^{b_p}\log(\phi).$$

Now, $\{\frac{\frac{1}{p+1}}{d_h}\}$ is equal to $\frac{\sqrt{p}\sqrt{p_0(1-p_0)}}{p+1}$ which converges to 0. Also, since a_p and b_p go to infinity and d_h goes to 0, the term in the third bracket goes to 0 by the definition of Riemann integral.

The integrand in the second term is $\log(C) - \frac{1}{2}x^2$, implying the second term to be $-\{\frac{\frac{1}{p+1}}{d_h}\}(a_p^3 - b_p^3)/6$. Here C is $\sqrt{\frac{1}{2\pi}}$.

From the definitions of a_p and b_p which are $O(\sqrt{p})$ and noting that $\frac{\frac{1}{p+1}}{d_h} = \frac{\sqrt{p}\sqrt{p_0(1-p_0)}}{p+1}$ the last term is is equal to $-\frac{p^2}{1+p}(K + z_p)$ where K is a constant and z_p goes to 0. Thus the expression is $O(p)$. □

References

Atay-Kayis, A. and Massam, H. (2005). "A Monte Carlo method for computing the marginal likelihood in nondecomposable Gaussian graphical models." *Biometrika*, 92(2): 317–335. MR2201362. doi: http://dx.doi.org/10.1093/biomet/92.2.317. 100, 107

Atchade, Y., Lartillot, N., and Robert, C. (2008). "Bayesian computation for statistical models with intractable normalizing constants." *Technical report, University of Michigan, Department of Statistics*. 107

Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). "High-resolution profiling of histone methylations in the human genome." *Cell*, 129: 823–837. 113, 115

Besag, J. (1974). "Spatial Interaction and the Statistical Analysis of Lattice Systems." *Journal of Royal Statistical Society Series B*, 135: 192–236. MR0373208. 104, 107

Carvalho, C. and Scott, J. (2009). "Objective Bayesian model selection in Gaussian graphical models." *Biometrika*, 96(3): 497–512. MR2538753. doi: http://dx.doi.org/10.1093/biomet/asp017. 100

Carvalho, C. M., Massam, H., and West, M. (2007). "Simulation of Hyper-inverse Wishart Distributions in Graphical Models." *Biometrika*, 94(3): 647–659. MR2410014. doi: http://dx.doi.org/10.1093/biomet/asm056. 100

Chen, M.-H. and Shao, Q.-M. (1997). "On Monte Carlo methods for estimating ratios of normalizing constants." *The Annals of Statistics*, 25: 1563–1594. MR1463565. doi: http://dx.doi.org/10.1214/aos/1031594732. 108

Chen, M.-H., Shao, Q.-M., and Ibrahim, J. (2000). *Monte Carlo Methods in Bayesian Computation*. Springer Verlag, New York. MR1742311. doi: http://dx.doi.org/10.1007/978-1-4612-1276-8. 108

Chen, S., Witten, D., and Shojaie, A. (2013). "Selection and Estimation for Mixed Graphical Models." arXiv:1311.0085. 100

Chiquet, J., Grandvalet, Y., and Ambroise, C. (2011). "Inferring multiple graphical structures." *Statistics and Computing*, 21(4): 537–553. MR2826691. doi: http://dx.doi.org/10.1007/s11222-010-9191-2. 101

Danaher, P., Wang, P., and Witten, D. M. (2013). "The joint graphical lasso for inverse covariance estimation across multiple classes." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2): 373–397. MR3164871. doi: http://dx.doi.org/10.1111/rssb.12033. 101, 109

Dobra, A., Hans, C., Jones, B., Nevins, J. R., and West, M. (2004). "Sparse graphical models for exploring gene expression data." *Journal of Multivariate Analysis*, 90: 196–212. MR2064941. doi: http://dx.doi.org/10.1016/j.jmva.2004.02.009. 100, 101

Dobra, A. and Lenkoski, A. (2011). "Copula Gaussian Graphical Models and Their Application to Modeling Functional Disability Data." *The Annals of Applied Statistics*, 5(2A): 969–993. MR2840183. doi: http://dx.doi.org/10.1214/10-AOAS397. 100

Dobra, A., Lenkoski, A., and Rodriguez, A. (2011). "Bayesian inference for general Gaussian graphical models with application to multivariate lattice data." *Journal of the American Statistical Association*, 106(496): 1418–1433. MR2896846. doi: http://dx.doi.org/10.1198/jasa.2011.tm10465. 107

Giudici, P. and Green, P. (1999). "Decomposable graphical Gaussian model determination." *Biometrika*, 86(4): 785–801. MR1741977. doi: http://dx.doi.org/10.1093/biomet/86.4.785. 107

Green, R. (1995). "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination." *Biometrika*, 82. MR1380810. doi: http://dx.doi.org/10.1093/biomet/82.4.711. 108

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2011). "Histone modifications as markers of cancer prognosis: a cellular view." *Biometrika*, 98: 1–15. MR2804206. doi: http://dx.doi.org/10.1093/biomet/asq060. 101, 109, 110

Hara, S. and Washio, T. (2013). "Learning a common substructure of multiple graphical Gaussian models." *Neural Networks*, 38: 23–38. doi: http://dx.doi.org/10.1016/j.neunet.2012.11.004. 101

Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C., and West, M. (2004). "Experiments in Stochastic Computation for High-Dimensional Graphical Models." *Statistical Science*, 20: 388–400. MR2210226. doi: http://dx.doi.org/10.1214/088342305000000304. 100

Kurdistani, S. (2007). "Histone modifications as markers of cancer prognosis: a cellular view." *British Journal of Cancer*, 97: 1–5. doi: http://dx.doi.org/10.1038/sj.bjc.6603844. 113

— (2011). "Histone modifications in cancer biology and prognosis." *Epigenetics and Disease*, 67: 91–106. doi: http://dx.doi.org/10.1007/978-3-7643-8989-5_5. 113

Lauritzen, S. L. and Sheehan, N. A. (2003). "Graphical Models for Genetic Analyses." *Statistical Science*, 18(4): 489–514. MR2059327. doi: http://dx.doi.org/10.1214/ss/1081443232. 101

Lenkoski, A. and Dobra, A. (2011). "Computational aspects related to inference in Gaussian graphical models with the G-Wishart prior." *Journal of Computational and Graphical Statistics*, 20(1): 140–157. MR2816542. doi: http://dx.doi.org/10.1198/jcgs.2010.08181. 107

Liu, C. L., Kaplan, T., Kim, M., Buratowski, S., Schreiber, S. L., Friedman, N., and Rando, O. J. (2005). "Single-nucleosome mapping of histone modifications in S. cerevisiae." *PLoS Biology*, 3: e328. 114

Lloyd, C. J. (1998). "Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems." *Journal of the American Statistical Association*, 93(444): 1356–1364. doi: http://dx.doi.org/10.1080/01621459.1998.10473797. 111

Meinshausen, N. and Bühlmann, P. (2006). "High-dimensional graphs and variable selection with the lasso." *The Annals of Statistics*, 34(3): 1436–1462. MR2278363. doi: http://dx.doi.org/10.1214/009053606000000281. 100

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 American Statistical Association*, 108: 69–90. MR3174603. doi: http://dx.doi.org/10.1080/01621459.2012.746058. 101, 107, 115

Mitsakakis, N., Massam, H., and Escobar, M. D. (2011). "A Metropolis-Hastings Based Method for Sampling from the G-Wishart Distribution in Gaussian Graphical Models." *Electronic Journal of Statistics*, 5: 18–30. MR2763796. doi: http://dx.doi.org/10.1214/11-EJS594. 100

Moeller, J., Pettitt, A. N., Berthelsen, K. K., and Reeves, R. W. (2006). "An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants." *Biometrika*, 93(2): 451–458. MR2278096. doi: http://dx.doi.org/10.1093/biomet/93.2.451. 107

Mohan, K., Chung, M. J.-Y., Han, S., Witten, D. M., Lee, S.-I., and Fazel, M. (2012). "Structured Learning of Gaussian Graphical Models." In: *NIPS*, 629–637. 101

Mohan, K., London, P., Fazel, M., Lee, S.-I., and Witten, D. (2013). "Node-based learning of multiple gaussian graphical models." arXiv:1303.5145. MR3190845. 101

Müller, P., Parmigiani, G., and Rice, K. (2007). "FDR and Bayesian Multiple Comparisons Rules." In: *Bayesian Statistics 8*. Oxford University Press. MR2433200. 115

Peterson, C., Stingo, F., and Vannucci, M. (2014). "Bayesian Inference of Multiple Gaussian Graphical Models." *Journal of the American Statistical Association*. doi: http://dx.doi.org/10.1080/01621459.2014.896806. 101

Piccioni, M. (2000). "Independence structure of natural conjugate densities to exponential families and the Gibbs' sampler." *Scandinavian journal of statistics*, 27(1): 111–127. MR1774047. doi: http://dx.doi.org/10.1111/1467-9469.00182. 100

Ravikumar, P., Wainwright, M. J., and Lafferty, J. D. (2010). "High-dimensional Ising model selection using 1-regularized logistic regression." *The Annals of Statistics*, 38(3): 1287–1319. MR2662343. doi: http://dx.doi.org/10.1214/09-AOS691. 100

Rodriguez, A., Lenkoski, A., and Dobra, A. (2011). "Sparse covariance estimation in heterogeneous samples." *Electronic Journal of Statistics*, 5: 981–1014. MR2836767. doi: http://dx.doi.org/10.1214/11-EJS634. 118

Scott, J. and Carvalho, C. (2008). "Feature-inclusion stochastic search for Gaussian graphical models." *Journal of Computational and Graphical Statistics*, 17(4): 790–808. MR2649067. doi: http://dx.doi.org/10.1198/106186008X382683. 100

Scott, J. G. and Berger, J. O. (2006). "An exploration of aspects of Bayesian multiple testing." *Journal of Statistical Planning and Inference*, 136(7): 2144–2162. MR2235051. doi: http://dx.doi.org/10.1016/j.jspi.2005.08.031. 105

Scott, J. G. and Berger, J. O. (2010). "Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem." *Annals of Statistics*, 38(5): 2587–2619. MR2722450. doi: http://dx.doi.org/10.1214/10-AOS792. 105

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): 2024–2048. MR2829945. doi: http://dx.doi.org/10.1214/10-AOAS360. 101

Strahl, B. D. and Allis, C. D. (2000). "The language of covalent histone modifications." *Nature*, 403: 41–45. 113

Wang, H. (2012). "Bayesian graphical lasso models and efficient posterior computation." *Bayesian Analysis*, 7(4): 867–886. MR3000017. doi: http://dx.doi.org/10.1214/12-BA729. 107

Wang, H. and Carvalho, C. M. (2010). "Simulation of hyper-inverse Wishart distributions for non-decomposable graphs." *Electronic Journal of Statistics*, 4: 1470–1475. MR2741209. doi: http://dx.doi.org/10.1214/10-EJS591. 100

Wang, H. and Li, S. Z. (2012). "Efficient Gaussian graphical model determination under G-Wishart prior distributions." *Electronic Journal of Statistics*, 6: 168–198. MR2879676. doi: http://dx.doi.org/10.1214/12-EJS669. 107

Wang, Z., Zang, C., Rosenfeld, J. A., Schones, D. E., Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Peng, W., Zhang, M. Q., and Zhao, K. (2008). "Combinatorial patterns of histone acetylations and methylations in the human genome." *Nature Genetics*, 40: 897–903. 115

Wong, F., Carter, C. K., and Kohn, R. (2003). "Efficient estimation of covariance selection models." *Biometrika*, 90(4): 809–830. MR2024759. doi: http://dx.doi.org/10.1093/biomet/90.4.809. 107

Wright, S. (1934). "The method of path coefficients." *Annals of Mathematical Statistics*, 5(3): 161–215. doi: http://dx.doi.org/10.1214/aoms/1177732676. 100

Yang, E., Ravikumar, P., Allen, G. I., and Liu, Z. (2013). "On graphical models via univariate exponential family distributions." arXiv:1301.4183. 100

Yang, E., Ravikumar, P. D., Allen, G. I., and Liu, Z. (2012). "Graphical Models via Generalized Linear Models." In: *NIPS*, volume 25, 1367–1375. 100, 101

Yuan, M. and Lin, Y. (2007). "Model selection and estimation in the Gaussian graphical model." *Biometrika*, 94(1): 19–35. MR2367824. doi: http://dx.doi.org/10.1093/biomet/asm018. 100

Zhang, Y. (2012). "A novel Bayesian graphical model for genome-wide multi-SNP association mapping." *Genetic epidemiology*, 36(1): 36–47. doi: http://dx.doi.org/10.1002/gepi.20661. 101

Zhang, Y., Lv, J., Liu, H., Zhu, J., Su, J., Wu, Q., Qi, Y., Wang, F., and Li, X. (2010). "HHMD: the human histone modification database." *Nucleic Acids Research*, 38: D149–154. 113