Mixed Graphical Models via Exponential Families

Eunho Yang* **Yulia Baker**† **Pradeep Ravikumar*** **Genevera I. Allen**† **Zhandong Liu**‡

*University of Texas, Austin , †Rice University , ‡Baylor College of Medicine

Abstract

Markov Random Fields, or undirected graphical models are widely used to model high-dimensional multivariate data. Classical instances of these models, such as Gaussian Graphical and Ising Models, as well as recent extensions (Yang et al., 2012) to graphical models specified by univariate exponential families, assume all variables arise from the same distribution. Complex data from high-throughput genomics and social networking for example, often contain discrete, count, and continuous variables measured on the same set of samples. To model such heterogeneous data, we develop a *novel class* of mixed graphical models by specifying that each node-conditional distribution is a member of a possibly different univariate exponential family. We study several instances of our model, and propose scalable M-estimators for recovering the underlying network structure. Simulations as well as an application to learning mixed genomic networks from next generation sequencing and mutation data demonstrate the versatility of our methods.

1 Introduction

Markov Networks, or undirected graphical models, are a popular tool for modeling, visualization, inference, and exploratory analysis of multivariate data with wide-ranging applications. The Gaussian Graphical Model, for continuous (Gaussian) variables, and the Ising / Potts model, for binary / categorical variables, are two widely used classes of Markov Networks. Recently, Yang et al. (2012) extending Besag (1974) introduce a more general class of graphical models constructed by assuming the node-conditional distributions arise from a univariate exponential family distribution. While this work permits graphical modeling for varied types of variables such as count data (e.g. Poisson graphical models) or left-skewed data (e.g. exponential graphical models), the models assume that all variables belong to the same type. There are many big-data examples, however, in economics, marketing, and advertising, among others, where observations are collected on a set of mixed variables, or variables of many different types. Consider high-throughput genomics, for example, where for a given biological sample, technologies can measure gene expression (continuous variables from microarrays or counts from RNA-sequencing), point mutations (binary variables from SNP-arrays), copy number variation (categorical variables after processing CGH-arrays), and epigenetic data (continuous variables from methylation arrays). Scientists are interested in studying relationships both *between* and *within* these different types of genomic markers to better understand the genetic basis of disease. To this end, new classes of mixed graphical models are needed that construct Markov Networks for sets of heterogeneous variables.

Existing models for mixed graphs are limited to one particular case: a Gaussian and Ising mixed model. This model was initially proposed by Lauritzen and Wermuth (1989) (and further studied in (Frydenberg and Lauritzen, 1989; Lauritzen, 1992; Lauritzen et al., 1989; Lauritzen, 1996)), where they formulated a Markov Network over nodes with a subset of continuous variables and a subset of discrete categorical or binary variables. The construction of this model is simple and assumes that the continuous variables conditioned on all possible configurations of the discrete vector are distributed as multivariate Gaussian. This model specification however scales exponentially with the number of discrete variables, and accordingly several others have proposed specializations of this Gaussian-Ising mixed graphical model. Lee and Hastie (2012) considered a specialization involving only pairwise interactions between any two variables, while Cheng et al. (2013) further allowed for three-way inter-

Appearing in Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, Reykjavik, Iceland. JMLR: W&CP volume 33. Copyright 2014 by the authors.

actions between two binary and one continuous variable. In addition to these specializations, these recent Gaussian-Ising models are limited to allowing variables to one of two specific types (binary/Ising, and continuous/Gaussian).

In this paper, we propose a general class of mixed graphical models that permits each variable to belong to a potentially different type. Our construction is a natural extension of that of the Gaussian-Ising model and the class of exponential family MRFs (Yang et al., 2012). Suppose the conditional distribution of each variable conditioned on other variables belongs to an arbitrary and *potentially different* univariate exponential family distribution. Two key model specification questions arise: (1) Do these node-conditional distributions jointly form a proper density over the nodes and if so, what is the form of this density? and, (2) What is the natural parameter space of these models, or in other words, under what conditions are mixed graphs normalizable? We carefully study both of these questions, showing that there indeed exists a proper joint distribution over variables of mixed node types; we call this the class of *mixed exponential MRFs*. We also show that there exist general conditions under which certain classes of these mixed graphical models are normalizable. Thus, for the first time, our work provides a general class of mixed graphical models, beyond the Gaussian-Ising instance, to encompass varied types of heterogeneous variables.

While our construction of general mixed graphical models is a natural extension of that of Markov Random Fields for variables of one type, there are possibly other ways of jointly modeling variables of mixed types. First, there has been much recent interest in non-parametric extensions of graphical models using things like copula transforms (Dobra and Lenkoski, 2011; Liu et al., 2012) or robust estimators of relationships between variables such as with Spearman's or Kendall's Tao rank-correlation (Xue and Zou, 2012). While such approaches could be employed for mixed types of variables, non-parametric approaches in general might not adequately account for differing domains of mixed variables and likely have less statistical power than parametric methods for recovering graph structure in high-dimensional settings. Second, our construction is closely related to that of conditional random field (CRF) models (Lafferty, 2001), and particularly CRFs constructed via node-conditional exponential families as recently investigated by Yang et al. (2013a). Deriving a mixed MRF from such CRFs by taking a product of a conditional CRF distributions and marginal MRF distributions however, has a key disadvantage in that the resulting distribution ends up with much more complicated terms. (A discussion of such formulations is given in the appendix). Third, relationships between variables of different types could be approached via types of multi-response regression models (Cai et al., 2013); these are particularly popular approaches for eQTL mapping of point mutations to gene expression, for example (Lee et al., 2010). While these approaches may be effective at finding connections between two sets of variables, they cannot model relationships within sets of variables, are limited to only two types of variables, and do not correspond to a coherent joint probabilistic model.

In this paper, we make several major contributions including: (1) Construction of a general class of mixed graphical models that permits each node to belong to a potentially different variable type, thus broadly generalizing the applicability of mixed statistical models; (2) Careful discussion of the conditions on the natural parameters under which these graphical model exist for paired sets of variables; (3) Development of an M-estimator to learn the structure of mixed graphical models via neighborhood selection. We demonstrate the applicability of our models through both simulation studies as well as a real high-throughput genomics example jointly learning a breast cancer genetic network of mutations (binary) and gene expression (counts via RNA-sequencing).

2 Mixed Graphical Models

Suppose we have a p-variate random response vector $X = (X_1, \ldots, X_p)$, with each response variable X_r taking values in a set $\mathcal{X}_r$. Suppose also that $G = (V, E)$ is an undirected graph over p nodes corresponding to the p response variables. Given the underlying graph G, and the set of cliques (fully-connected sub-graphs) $\mathcal{C}$ of the graph G, the corresponding Markov random field (MRF) is a set of distributions over the random vector X, that satisfy *Markov independence assumptions* with respect to the graph G. By the Hammersley-Clifford Theorem (Lauritzen, 1996), any such distribution, when positive also has the following specific factored form. Let $\{\phi_c(X_c)\}_{c \in \mathcal{C}}$ denote a set of clique-wise sufficient statistics, so that ϕ_c only depends on variables X_c corresponding to the clique $c \in \mathcal{C}$. Then any strictly positive distribution over the random vector X within the Markov random field family takes the following factored form:

$$P(X) \propto \exp\big\{ \sum_{c \in \mathcal{C}} \phi_c(X_c) \big\}. \tag{1}$$

The sufficient statistics $\{\phi_c(X_c)\}_{c \in \mathcal{C}}$ are typically specified according to the type and properties of the random vector; the Ising model for instance corresponds to linear and quadratic sufficient statistics over binary data, while the Gaussian MRF corresponds to linear and quadratic sufficient statistics over continuous real-valued data. *Mixed* Markov random fields (MRFs) would correspond to the setting where the random variables belong to heterogeneous domain sets, so that the sets $\{\mathcal{X}_r\}_{r\in V}$ are potentially all distinct. There is however a lack of any substantial model specification and statistical theory (beyond the references noted in the introduction) for such mixed MRFs. A key reason for this is the difficulty in setting appropriate sufficient statistics $\{\phi_c(X_c)\}_{c\in\mathcal{C}}$ specifying the MRFs over such heterogeneous random variables, some of which could be discrete, and others could be continuous with disparate properties.

Interestingly, when considering the heterogeneous random variables individually, we do have considerable understanding in specifying *univariate* distributions over these varied types of variables, discrete as well as continuous. A popular class of univariate family of distributions for instance is the exponential family class of distributions: $P(Z) = \exp(\theta\, B(Z) + C(Z) - D(\theta))$, with sufficient statistics $B(Z)$, base measure $C(Z)$, and log-normalization constant $D(\theta)$. Such exponential family distributions include a wide variety of commonly used distributions over varied continuous and discrete data, such as Gaussian, Bernoulli, multinomial, Poisson, exponential, gamma, chi-squared, beta, any of which can be instantiated with particular choices of the functions $B(\cdot)$, and $C(\cdot)$. Such univariate exponential family distributions are thus popularly used to model a heterogeneous variety of data types including skewed continuous data and count data. Additionally, through generalized linear models, they are used to model the response of various data types conditional on a set of covariates. The key question is whether we can combine multiple such univariate exponential family distributions into a single *mixed* MRF distribution over heterogeneous multivariate data.

We consider the following generalization of the construction in Besag (1974); Yang et al. (2012). Note that the conditional distribution of a variable conditioned on the rest of the variables can be specified by a univariate distribution. Accordingly, suppose that the node-conditional distributions of variables X_r conditioned on the rest of the response variables, $X_{V\backslash r}$ is given by an arbitrary univariate exponential family that depends on the node $r \in V$:

$$\begin{aligned} &P(X_r|X_{V\backslash r}) \qquad (2)\\ &= \exp\Big\{E_r(X_{V\backslash r})\, B_r(X_r) + C_r(X_r) - \bar{D}_r(X_{V\backslash r})\Big\}. \end{aligned}$$

Here, the functions $B_r(\cdot), C_r(\cdot)$ are specified by the choice of a univariate exponential family, and the parameter $E_r(X_{V\backslash r})$ is an *arbitrary function* of the all variables except X_r. Note that the exponential family for each variable X_r could be distinct.

Consider the joint MRF distribution as in (1) with arbitrary sufficient statistics $\{\phi_c(X_c)\}_{c\in\mathcal{C}}$. Would the node-conditional distributions as specified in (2) be consistent with a joint MRF distribution, possibly under some restriction over the choice of the exponential families, over the functions $\{E_r(\cdot)\}_{r\in V}$ specifying the node-conditional distributions?

Theorem 1. *Consider a p-dimensional random vector $X = (X_1, X_2, \ldots, X_p)$, with each variable X_r taking values in a potentially distinct set $\mathcal{X}_r$. Consider the node-conditional distributions, of each variable X_r conditioned on the rest of random variables, as specified in* (2) *by heterogeneous univariate exponential family distributions. These are* consistent *with a joint MRF distribution over the random vector X, as in* (1), *that is Markov with respect to a graph $G = (V, E)$ with clique-set $\mathcal{C}$ of size at most k, if and only if the functions $\{E_r(\cdot)\}_{r\in V}$ specifying the node-conditional distributions have the form:*

$$\theta_r + \sum_{t\in N(r)} \theta_{rt}\, B_t(X_t) + \ldots + \sum_{t_2,\ldots,t_k\in N(r)} \theta_{r\,t_2\ldots t_k}(X) \prod_{j=2}^{k} B_{t_j}(X_{t_j}),$$

where $\theta_{r\cdot} := \{\theta_r, \theta_{rt}, \ldots, \theta_{r\,t_2\ldots t_k}\}$ is a set of parameters, and $N(r)$ is the set of neighbors of node r according to an undirected graph $G = (V, E)$. Moreover, the corresponding consistent joint MRF distribution in turn has the following form:

$$\begin{aligned} P(X;\theta) = \exp\Big\{ &\sum_{r\in V} \theta_r B_r(X_r) + \sum_{r\in V}\sum_{t\in N(r)} \theta_{rt}\, B_r(X_r) B_t(X_t) + \\ &\ldots + \sum_{(t_1,\ldots,t_k)\in\mathcal{C}} \theta_{t_1\ldots t_k} \prod_{j=1}^{k} B_{t_j}(X_{t_j}) + \sum_{r\in V} C_r(X_r) - A\big(\theta\big)\Big\}, \quad (3) \end{aligned}$$

where $A\big(\theta\big)$ is the log-normalization constant.

Theorem 1 states that one could choose arbitrary and potentially different exponential families for each of the node-conditional distributions, and yet obtain a valid consistent joint MRF distribution if and only if the functions $E_r(X_{V\backslash r})$ specifying the canonical parameter in the univariate exponential family distributions (2) have a specific form. Then, not only is there is a corresponding consistent joint MRF distribution, it has the specific form as specified in (3). We term this class of MRFs specified in Theorem 1 as the class of *mixed exponential MRFs.*

This class of mixed exponential MRFs allows us, for the first time, to specify joint distributions over varied heterogeneous random variables. We note that it recovers the exponential MRF family of (Yang et al., 2012) over *homogeneous* multivariate data, by setting all the exponential families to be the same. Theorem 1 can thus be understood as an extension of their framework to the heterogeneous setting where the exponential distributions comprising the node-conditional distributions could all be *distinct*, thus being able to simultaneously model variables from disparate domains with disparate characteristics, such as discrete, continuous, skewed-continuous, etc.

An important special case of the mixed exponential MRF family is when the joint distribution has clique factors of size at most two.The joint distribution in (3) would take the form:

$$P(X;\theta) = \exp\Big\{ \sum_{r\in V} \theta_r B_r(X_r) + \sum_{(r,t)\in E} \theta_{rt}\, B_r(X_r)\, B_t(X_t) + \sum_{r\in V} C_r(X_r) - A\big(\theta\big) \Big\}, \tag{4}$$

with the log-normalization term given by $A\big(\theta\big) := \log \int_{\mathcal{X}^p} \exp\big\{ \sum_{r\in V} \theta_r B_r(X_r) + \sum_{(r,t)\in E} \theta_{rt}\, B_r(X_r)\, B_t(X_t) + \sum_{r\in V} C_r(X_r) \big\}$.

In order to build the joint distribution under this framework (3), we only need to specify $B_r(\cdot)$ and $C_r(\cdot)$ for each random variable $r \in V$. As an example, consider the pairwise MRF with three types of random variables: Gaussian, Ising and Poisson. Let (V_G, E_G) be the sub-graph corresponding only to Gaussian variables. (V_I, E_I) and (V_P, E_P) are defined similarly. We also have sets of cross-edges; denote E_{GI} as the set of edges between Gaussian and Ising variables, and so on. Then, the mixed MRF distribution in (3) takes the form:

$$\begin{aligned} P(X) \propto \exp\Big\{ & \sum_{r\in V_G} \frac{\theta^g_r}{\sigma_r} X_r + \sum_{r'\in V_I} \theta^i_r X_r + \sum_{r''\in V_P} \theta^p_r X_r \\ & + \sum_{(r,t)\in E_G} \frac{\theta^{gg}_{rt}}{\sigma_r\sigma_t} X_r X_t + \sum_{(r',t')\in E_I} \theta^{ii}_{r't'} X_{r'} X_{t'} + \sum_{(r'',t'')\in E_P} \theta^{pp}_{r''t''} X_{r''} X_{t''} \\ & + \sum_{(r,r')\in E_{GI}} \frac{\theta^{gi}_{rr'}}{\sigma_r} X_r X_{r'} + \sum_{(r,r'')\in E_{GP}} \frac{\theta^{gp}_{rr''}}{\sigma_r} X_r X_{r''} + \sum_{(r',r'')\in E_{IP}} \theta^{ip}_{r'r''} X_{r'} X_{r''} \\ & - \sum_{r\in V_G} \frac{X_r^2}{2\sigma_r^2} - \sum_{r''\in V_P} \log(X_{r''}!) \Big\}. \end{aligned}$$

An important question that arises, particularly given interactions between heterogeneous types, is under what constraints on the parameters θ is the mixed MRF distribution in (3) well-defined, so that $A\big(\theta\big) < \infty$, and the distribution is normalizable. We will study this further in the next section.

3 Manichean Graphical Models

An important subclass of our mixed exponential MRF family in (4) is when the random variables belong to just one of two types. Specifically, suppose the set of random variables $\{X_1, \ldots, X_p\}$ is partitioned into two groups: $\{Y_1, \ldots, Y_{p_Y}\}$ of variables Y_r taking values in a set $\mathcal{Y}$; and $\{Z_1, \ldots, Z_{p_Z}\}$ of variables Z_r taking values in a set $\mathcal{Z}$ where $p = p_Y + p_Z$. Collating these groups into random vectors $Y := (Y_1, \ldots, Y_{p_Y})$ and $Z := (Z_1, \ldots, Z_{p_Z})$, and $X := (Y, Z)$, consider the mixed MRF family over $X = (Y, Z)$ from (4) where the exponential families for the node-conditional distributions of variables in $\{Y_1, \ldots, Y_{p_Y}\}$ are specified by the sufficient statistic $B_Y(\cdot)$ and base-measure $C_Y(\cdot)$, and those for the variables in $\{Z_1, \ldots, Z_{p_Z}\}$ are specified by the sufficient statistic $B_Z(\cdot)$ and base-measure $C_Z(\cdot)$. With these choices of the univariate exponential families, we then obtain the following sub-class of mixed MRF distributions:

$$\begin{aligned} P(Y,Z;\theta) \propto \exp\Big\{ & \sum_{r\in V_Y} \theta^y_r B_Y(Y_r) + \sum_{r'\in V_Z} \theta^z_{r'} B_Z(Z_{r'}) + \\ & \sum_{(r,t)\in E_Y} \theta^{yy}_{rt}\, B_Y(Y_r)\, B_Y(Y_t) + \sum_{(r',t')\in E_Z} \theta^{zz}_{r't'}\, B_Z(Z_{r'})\, B_Z(Z_{t'}) + \\ & \sum_{(r,r')\in E_{YZ}} \theta^{yz}_{rr'}\, B_Y(Y_r)\, B_Z(Z_{r'}) + \sum_{r\in V_Y} C_Y(Y_r) + \sum_{r'\in V_Z} C_Z(Z_{r'}) \Big\} \end{aligned} \tag{5}$$

where V_Y, V_Z are the sets of nodes corresponding to variables in Y and Z respectively; and E_Y are the set of edges restricted to nodes in V_Y, E_Z are the set of edges restricted to nodes in V_Z, and E_{YZ} is the set of "heterogeneous" edges between nodes in V_Y and V_Z. We term the subclass of joint distributions in (5) as Manichean mixed exponential MRFs, or Manichean MRFs in short (after the philosophy that loosely, places elements into one of two types).

Past work (Lauritzen, 1996; Lee and Hastie, 2012; Cheng et al., 2013) in modeling joint MRFs over heterogenous random variables has been restricted to this dual-type setting, where the random variables belong to one of two types. Indeed, it has been specifically focused on the setting where one set of variables is binary or discrete, and the other set of variables is continuous; and the resulting mixed MRFs were either a Gaussian-Ising or Gaussian-discrete MRF, which can be seen as special cases of our Manichean exponential MRF family. In illustrating our class of mixed MRFs via examples, we thus largely focus on this Manichean setting. We interleave our discussions with an analysis of the *normalizability conditions* over the model parameters for such mixed MRFs. We delineate two settings of such Manichean MRFs: one where at least one of the domains $\mathcal{Y}$ or $\mathcal{Z}$ is finite, and the other where both the domains are infinite.

3.1 When one of the domains is finite

We first focus on the case when at least one of the domains $\mathcal{Y}$ or $\mathcal{Z}$ is finite. Let us assume without loss of generality that $\mathcal{Z}$ is finite, and that both the maximum and minimum values in $\mathcal{Z}$ are finite; $\max\{z : z \in \mathcal{Z}\} < \infty$ and $\min\{z : z \in \mathcal{Z}\} > -\infty$. As we will show, such a setting allows an easier specification of the normalizability conditions for (5).

We first note that given the *pairwise* joint distribution in (5), the conditional distribution $P(Y|Z)$ can be derived as follows:

$$P(Y|Z) \propto \exp\Big\{ \sum_{r\in V_Y} \theta_r^y B_Y(Y_r) + \sum_{(r,t)\in E_Y} \theta_{rt}^{yy}\, B_Y(Y_r) B_Y(Y_t) + \sum_{(r,r')\in E_{YZ}} \theta_{rr'}^{yz} B_Y(Y_r) B_Z(Z_{r'}) + \sum_{r\in V_Y} C_Y(Y_r) \Big\}. \qquad (6)$$

The distribution (6) can be understood to belong to an exponential family with node-wise sufficient statistics $B_Y(Y_r)$ for $r \in V_Y$, and pairwise sufficient statistics $B_Y(Y_r)\, B_Y(Y_t)$ for $(r,t) \in E_Y$. The canonical parameters $\bar{\theta}_r^y(Z)$ corresponding to the node-wise sufficient statistics are linear functions of the conditioned random vector Z, given by $\bar{\theta}_r^y(Z) = \theta_r^y + \sum_{r'\in N(r)} \theta_{rr'}^{yz} B_Z(Z_{r'})$. The canonical parameters $\bar{\theta}_{rt}^{yy}(Z)$ for the pair-wise sufficient statistics are simply constants independent of Z: $\bar{\theta}_{rt}^{yy}(Z) := \theta_{rt}^{yy}$. The log-partition function (denote $A_{Y|Z}(\cdot)$) of (6) is some function of $\bar{\theta}^y(Z) := \{\bar{\theta}_r^y(Z)\}_{r\in V_Y}$ and $\bar{\theta}^{yy}(Z)$ (see Yang et al. (2013a) for further details of such exponential conditional graphical models, also known as conditional random fields (CRFs)). Note that we can obtain higher-order interaction terms by considering higher-order interactions in the corresponding joint distribution beyond the pairwise terms in (5).

The following theorem shows that the normalizability condition for the joint distribution in (5) can be expressed in terms of the *conditional* log-partition function $A_{Y|Z}(\cdot)$:

Theorem 2. *The Manichean MRF joint distribution in* (5) *is normalizable iff*

$$E_{Z'}\Big[\exp\big\{A_{Y|Z'}(\bar{\theta}^y(Z'), \bar{\theta}^{yy})\big\}\Big] < \infty,$$

where $E_{Z'}[\cdot]$ *is the expectation with respect to a random vector* Z' *that follows the pairwise MRF distribution:*

$$P(Z') \propto \exp\Big\{ \sum_{r'\in V_Z} \theta_{r'}^z B_Z(Z'_{r'}) + \sum_{r'\in V_Z} C_Z(Z'_{r'}) + \sum_{(r',t')\in E_Z} \theta_{r't'}^{zz}\, B_Z(Z'_{r'})\, B_Z(Z'_{t'}) \Big\},$$

where the parameters $\theta_{r'}^z, \theta_{r't'}^{zz}$*, the sufficient statistics* $B_Z(\cdot)$*, the base measure* $C_Z(\cdot)$ *and the node/edge sets* (V_Z, E_Z) *are as specified in the Manichean MRF joint distribution in* (5).

The following corollary of Theorem 2 then addresses the normalizability of the joint distribution in (5) for the case where one of the domains is finite.

Corollary 1. *Suppose that the domain* $\mathcal{Z}$ *is finite, with* $\max\{z : z \in \mathcal{Z}\} < \infty$ *and* $\min\{z : z \in \mathcal{Z}\} > -\infty$. *Suppose also that the conditional distribution* (6) *given* Z *is well-defined (i.e. normalizable) for all* $Z \in \mathcal{Z}$. *Then, the log-partition function is finite, and the joint in* (5) *is well-defined and normalizable as well.*

Example: Gaussian - Ising The Gaussian - Ising mixed graphical model (Lee and Hastie, 2012; Cheng et al., 2013) can be seen to be a special case of the mixed MRF family in (5) with univariate Gaussian and Bernoulli distributions as the exponential family distributions. Specifically, suppose that the domain of the variables $\{Y_r\}$ is $\mathcal{Y} = \mathbb{R}$, and that their corresponding choice of a univariate exponential family is the univariate Gaussian distribution with known σ^2, so that the sufficient statistics and base measure are given by $B_Y(Y_r) = \frac{Y_r}{\sigma_r}$, and $C_Y(Y_r) = -\frac{Y_r^2}{2\sigma_r^2}$, respectively. Suppose also that the univariate exponential family corresponding to variables $\{Z_l\}$ is the Bernoulli distribution, so that $\mathcal{Z} = \{-1, 1\}$, and the sufficient statistics and base measure are given by $B_Z(Z_{r'}) = Z_{r'}$, $C_Z(Z_{r'}) = 0$ for all $r' \in V_Z$. Substituting these in (5), we obtain the following mixed MRF distribution:

$$P(Y,Z) \propto \exp\Big\{ \sum_{r\in V_Y} \frac{\theta_r^y}{\sigma_r} Y_r + \sum_{r'\in V_Z} \theta_{r'}^z Z_{r'} + \sum_{(r,t)\in E_Y} \frac{\theta_{rt}^{yy}}{\sigma_r\sigma_t} Y_r Y_t + \sum_{(r',t')\in E_Z} \theta_{r't'}^{zz}\, Z_{r'}\, Z_{t'} + \sum_{(r,r')\in E_{YZ}} \frac{\theta_{rr'}^{yz}}{\sigma_r}\, Y_r\, Z_{r'} - \sum_{r\in V_Y} \frac{Y_r^2}{2\sigma_r^2} \Big\}.$$

The conditional Gaussian distribution given Z, $P(Y|Z)$ is well defined as long as $\Theta \prec 0$ where Θ is a matrix defined as $[\Theta]_{rt} = \begin{cases} -\frac{1}{\sigma_r^2} & \text{if } r = t \\ \frac{\theta_{rt}^{yy}}{\sigma_r\sigma_t} & \text{otherwise.} \end{cases}$ Thus, from Corollary 1, the joint Gaussian - Ising distribution is normalizable so long as $\Theta \prec 0$.

Example: Poisson - Ising We can define a mixed Poisson - Ising model as follows. Suppose that the domain of the variables $\{Y_r\}$ is $\mathcal{Y}_r = \{0, 1, 2, \ldots\}$. The natural choice of an exponential family distribution for such over count-valued domain is the univariate Poisson distribution, with sufficient statistic and base measure are given by $B_Y(Y_r) = Y_r$, and $C_Y(Y_r) = -\log(Y_r!)$, respectively. Suppose that the remaining variables $\{Z_r\}$ are binary with $\mathcal{Z}_{r'} = \{0, 1\}$, so that the corresponding natural univariate exponential family is the Bernoulli distribution, with sufficient statistics and base measure are given by $B_Z(Z_{r'}) = Z_{r'}$, $C_Z(Z_{r'}) = 0$. Substituting these in (5), we obtain the following mixed MRF distribution:

$$P(Y,Z) \propto \exp\Big\{ \sum_{r\in V_Y} \theta^y_r Y_r + \sum_{r'\in V_Z} \theta^z_{r'} Z_{r'} + \sum_{(r,t)\in E_Y} \theta^{yy}_{rt} Y_r Y_t + \sum_{(r',t')\in E_Z} \theta^{zz}_{r't'}\, Z_{r'}\, Z_{t'} + \sum_{(r,r')\in E_{YZ}} \theta^{yz}_{rr'}\, Y_r\, Z_{r'} - \sum_{r\in V_Y} \log(Y_r!) \Big\}. \quad (7)$$

As a specialization of Corollary 1 to this setting, we obtain the following corollary for the normalizability of the Poisson - Ising distribution:

Corollary 2. *The Poisson-Ising distribution* (7) *is well-defined iff $\theta_{rt} \le 0$ all $(r,t) \in E_Y$ (i.e., for the pairwise terms over count-valued variables).*

The condition specified in the corollary is the one required for normalizability of any conditional distribution $P(Y|Z)$ of the count-valued variables conditioned on the binary variables (Yang et al., 2013a). The corollary then follows from an application of Corollary 1.

3.2 When both domains $\mathcal{Y}$ and $\mathcal{Z}$ are infinite

Now consider the setting where both domains $\mathcal{Y}$ and $\mathcal{Z}$ are infinite, with $\sup\{z : z \in \mathcal{Z}\} = \infty$ or $\inf\{z : z \in \mathcal{Z}\} = -\infty$; and with the same for $\mathcal{Y}$. Instances of variables with such infinite domains include real-valued variables (e.g. the Gaussian exponential family) or count-valued variables (e.g. the Poisson exponential family). The following proposition provides a necessary condition when the joint mixed MRF distribution is normalizable under such a setting:

Proposition 1. *Suppose that both domains $\mathcal{Y}$ and $\mathcal{Z}$ are infinite. Then, if the mixed MRF distribution* (5) *is normalizable, it necessarily holds that unnormalized mass in* (5) *should converge to zero. That is, letting $X := (Y,Z)$, for all $r,t \in V := V_Y \cup V_Z$, there exists $(x_0, x_1) \in \mathcal{X}_r \times \mathcal{X}_t$, so that it necessarily holds that*

$$\theta_r B_r(X_r) + \theta_t B_t(X_t) + \theta_{rt} B_r(X_r) B_t(X_t) + C_r(X_r) + C_t(X_t) < 0,$$

for all values $(X_r, X_t) \in \mathcal{X}_r \times \mathcal{X}_t$ s.t. $|X_r| \ge |x_0|$ and $|X_t| \ge |x_1|$.

Note that the node indices r and t range over any variable in $X := (Y,Z)$, and thus over any of the two types of variables in Y or Z.

In the sequel, we focus on a popular exponential family setting of linear sufficient statistics $B_r(X_r) = X_r$ (which includes popular instances such as Gaussian, Poisson, Bernoulli, exponential, etc.). In the following theorem, we derive *necessary* conditions in order for the normalizability condition in Proposition 1 to be satisfied.

Theorem 3. *Suppose that both domains $\mathcal{Y}$ and $\mathcal{Z}$ are infinite, and that $B_r(X_r) = X_r$, $r \in V$, for $X := (Y,Z)$ and $V := V_Y \cup V_Z$. Then the mixed MRF distribution expression over $X := (Y,Z)$ in* (5) *is not normalizable if neither of the following conditions are satisfied for all $r,t \in V$ with non-zero θ_{rt}:*

(a) *both $\mathcal{X}_r$ and $\mathcal{X}_t$ are infinite only from one side, so that $\sup\{x : x \in \mathcal{X}_r\} < \infty$ or $\inf\{x : x \in \mathcal{X}_r\} > -\infty$, with the same for $\mathcal{X}_t$.*

(b) *for $\forall \alpha, \beta > 0$ such that $-C_r(X_r) = O(X_r^\alpha)$ and $-C_t(X_t) = O(X_t^\beta)$, it holds that $(\alpha-1)(\beta-1) \ge 1$.*

3.2.1 Example: Gaussian - Poisson

Suppose that the domain of the variables $\{Y_r\}$ is $\mathcal{Y} = \mathbb{R}$, and that their corresponding choice of a univariate exponential family is the univariate Gaussian distribution with known σ^2 as discussed earlier, with sufficient statistics and base measure given by $B_Y(Y_r) = \frac{Y_r}{\sigma_r}$ and $C_Y(Y_r) = -\frac{Y_r^2}{2\sigma_r^2}$ respectively. Also suppose that the remaining random variables $\{Z_r\}$ are count-valued so that $\mathcal{Z} = \{0,1,2,\ldots\}$, and that the corresponding choice of the univariate exponential family is the Poisson distribution, with sufficient statistic and base measure are given by $B_Z(Z_{r'}) = Z_{r'}$, $C_Z(Z_{r'}) = -\log(Z_{r'}!)$ respectively. Substituting these in (5), we obtain the following mixed MRF:

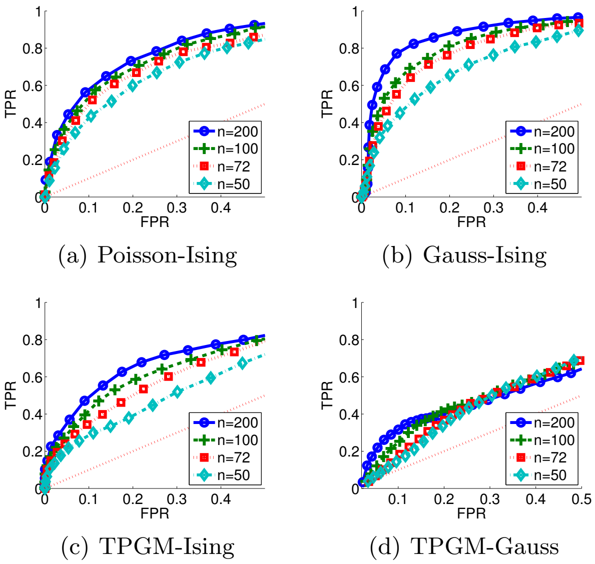
$$P(Y,Z) \propto \exp\Big\{ \sum_{r\in V_Y} \frac{\theta^y_r}{\sigma_r} Y_r + \sum_{r'\in V_Z} \theta^z_{r'} Z_{r'} + \sum_{(r,t)\in E_Y} \frac{\theta^{yy}_{rt}}{\sigma_r \sigma_t} Y_r Y_t + \sum_{(r',t')\in E_Z} \theta^{zz}_{r't'}\, Z_{r'}\, Z_{t'} + \sum_{(r,r')\in E_{YZ}} \frac{\theta^{yz}_{rr'}}{\sigma_r} Y_r\, Z_{r'} - \sum_{r\in V_Y} \frac{Y_r^2}{2\sigma_r^2} - \sum_{r'\in V_Z} \log(Z_{r'}!) \Big\}. \quad (8)$$

As we show in the following corollary however, there can be *no interactions* between the continuous (corresponding to Gaussian) and count-valued (corresponding to Poisson) variables for the distribution to be normalizable!

Corollary 3. *The Gaussian-Poisson distribution* (8) *is not normalizable unless $\theta_{rt} = 0$ for all $(r,t) \in E_{YZ}$.*

Corollary 3 follows from an application of Theorem 3. The Gaussian random variables are infinite in both directions, so that they do not satisfy the first condition. Moreover, $-C_Y(X_r) = O(X_r^2)$, so that $\alpha = 2$, while $\log(Z_{r'}!)$ is no faster than $Z_{r'}^{1.5}$ asymptotically, so that $\beta = 1.5$, with $(2-1)(1.5-1) < 1$, so that the second condition is also violated. Gaussian-exponential mixed graphical models can also be analyzed along similar lines.

How can we then model mixed MRFs over continuous and count-valued variables? A useful distribution towards addressing this is provided by the univariate Truncated Poisson distribution introduced by



(a) Poisson-Ising

(b) Gauss-Ising

(c) TPGM-Ising

(d) TPGM-Gauss

Figure 1: ROC curves for different types of Manichean graphical models when $p_Y = 36$, $p_Z = 36$.

Yang et al. (2013b), which is a finite-domain distribution (over a finite set of non-negative integers) that has the shape of the Poisson distribution over its finite domain. Consider the mixed MRF where the variables $Z_{r'}$ of the random sub-vector $Z := (Z_1, \dots, Z_l)$ follows the univariate Truncated Poisson distribution with values in the set $\mathcal{Z} = \{0, 1, 2, \dots, R\}$, where R is a fixed positive constant, a truncating parameter. The sufficient statistics and base measure are given by $B_Z(Z_{r'}) = Z_{r'}$, and $C_Z(Z_{r'}) = -\log(Z_{r'}!)$, respectively. Then, appealing to results in the previous Section 3.1 where one of the domains is finite, we would conclude that the Gaussian-TPGM mixed models are normalizable under the condition that $\Theta \prec 0$.

4 Learning Mixed Graphical Models

We now consider the task of learning a mixed exponential MRF distribution given i.i.d. observations. We focus on the two type Manichean case, with variables (Y, Z) distributed as in (5) with unknown parameters θ^*; given n i.i.d. samples $\mathcal{D} := \left\{Y^{(i)}, Z^{(i)}\right\}_{i=1}^n$ drawn from this unknown mixed MRF, the task is to recover the parameters, and in particular the underlying MRF graph structure. While regularized MLE estimators would be a natural choice, these involve the log-partition function of the mixed MRF (5), which is typically intractable due to the summation or integration over the domains of varied types of variables. Accordingly, we follow the node-neighborhood estimation based approach of (Meinshausen and Bühlmann, 2006; Ravikumar et al., 2010; Yang et al., 2012, 2013a): instead of maximizing the joint likelihood, we separately learn node-wise *conditional distributions* at every node, which would yield estimates of parameter-sets $\{\theta^*_{rt}\}_{t \in N(r)}$, as well as node-neighborhoods $N(r)$ separately; these can be stitched together to obtain the overall graph structure, as in Meinshausen and Bühlmann (2006); Ravikumar et al. (2010); Yang et al. (2012, 2013a).

By Theorem 1, the node-wise conditional distribution of Y_r given the rest of nodes $Y_{V_Y \backslash r}$ and Z, has the form: $P(Y_r | \, Y_{V_Y \backslash r}, Z; \theta^*) = \exp\Big\{ B_Y(Y_r) \eta(Y_{V_Y \backslash r}, Z; \theta^*) + C_Y(Y_r) - \bar{D}_r\big(\eta(Y_{V_Y \backslash r}, Z; \theta^*)\big)\Big\}$, which can be seen to be a *univariate* exponential distribution with canonical parameter $\eta(Y_{V_Y \backslash r}, Z; \theta^*) = \theta^{*y}_{r} + \sum_{t \in V_Y \backslash r} \theta^{*yy}_{rt} B_Y(Y_t) + \sum_{r' \in V_Z} \theta^{*yz}_{rr'} B_Z(Z_{r'})$.

Hence, given i.i.d. samples from the joint mixed MRF distribution $P(Y, Z; \theta^*)$, the corresponding negative log-conditional-likelihood can be written as $\ell(\theta; \mathcal{D}) := -\frac{1}{n} \log \prod_{i=1}^n P(Y_r^{(i)} | Y^{(i)}_{V_Y \backslash r}, Z^{(i)}; \theta)$. The corresponding ℓ_1 regularized M-estimator of the conditional distribution parameters is then given as:

$$\min_{\theta \in \mathbb{R}^{1 + (p_Y - 1) + p_Z}} \ell(\theta; \mathcal{D}) + \lambda_{Y,n} \|\theta^{yy}\|_1 + \lambda_{Z,n} \|\theta^{yz}\|_1, \quad (9)$$

where $\lambda_{Y,n}$, $\lambda_{Z,n}$ are the regularization constants, and could have different values. Given this M-estimator, we can recover the homogeneous-neighborhood of Y_r (i.e. interactions with nodes in $\{Y_{\backslash r}\}$) as $N_Y(r) = \{t \in V_Y \setminus r \mid \theta^{yy}_{rt} \neq 0\}$, and its heterogeneous-neighborhood (i.e. interactions with nodes in $\{Z_{r'}\}$) as $N_Z(r) = \{r' \in V_Z \mid \theta^{yz}_{rr'} \neq 0\}$. The node-conditional distribution for $Z_{r'}$ given the rest of nodes, $P(Z_{r'} | Y, Z_{V_Z \backslash r'}; \theta^*)$, and its corresponding M-estimator can also be defined similarly as above.

For the statistical analysis of the M-estimator above, in particular its sparsistency, or consistent recovery of neighborhood sets, we can directly appeal to results in Yang et al. (2013a) where they analyzed the sparsistency of an M-estimator for a conditional random field, and which has a similar form as the M-estimator above; from which it can be shown that the M-estimator in (9) recovers the neighborhood sets $N_Y(r)$ and $N_Z(r)$ exactly for all $r \in V$ with high probability under standard incoherence conditions.

5 Numerical Experiments

Simulation Studies To study the performance of our M-estimator for learning the mixed network structures, we generate data via a Gibbs sampler from four instances of our Manichean graphical model: Gaussian-Ising, Gaussian-Truncated Poisson (TPGM), Poisson-Ising, and Truncated Poisson-Ising. A lattice graph connecting nearest neighbors on a two-

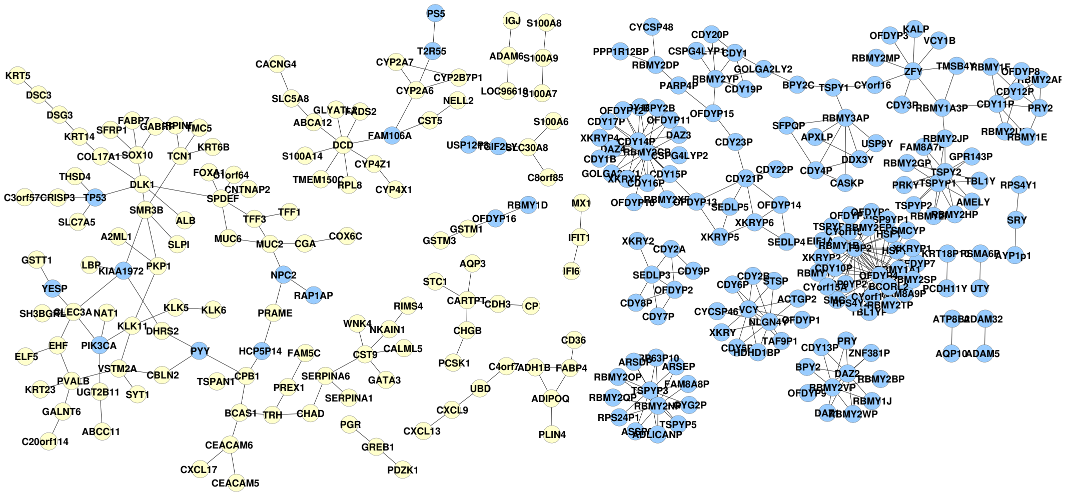


Figure 2: Connected components of a breast cancer genetic network estimated by the Truncated Poisson and Ising mixed graphical model for gene expression via RNA-sequencing (yellow nodes) and genomic mutations including both point and copy number aberrations (blue nodes) measured on the same set of 697 breast cancer subjects. Key highly mutated cancer biomarkers such as TP53 and PIK3C are found to have many inter-connections to gene expression variables that are consistent with the cancer genomics literature.

dimensional grid with $p = 72$ variables, $p_Y = 36$ and $p_Z = 36$, is employed. In Figure 1, we report receiver-operators characteristic (ROC) curves for recovering the true edge structure of the graph by varying the sparsity parameters, λ_Y and λ_Z which are held at a constant ratio, for four different sample sizes, n =50, 72, 100 and 200.

Results indicate that we are able to recover the graph structure via neighborhood selection for instances beyond the existing Gaussian-Ising graph. Mixed graph recovery, even in $p > n$ cases, is indeed possible for all pairs except the Truncated Poisson (TPGM) - Gaussian model. In this instance, as the truncation level increases, the strengths of connections is weakened because the TPGM tends towards the PGM for which the paired Poisson-Gaussian model does not allow heterogeneous interactions as we had shown in the previous section. (For properties of the truncated Poisson distribution, see Yang et al. (2013b)).

Genomics Example We employ our Manichean graphical model on a high-throughput genomic example to identify connections both between and within gene expression biomarkers and genomic mutation biomarkers for invasive breast carcinoma. Level III RNA-sequencing data for 806 patients was downloaded from The Cancer Genome Atlas (TCGA) (Cancer Genome Atlas Research Network, 2012) and processed according to techniques described in Allen and Liu (2013). Level II non-silent somatic mutation and level III copy number variation data was downloaded from TCGA for 951 patients. Copy number data was segmented using standard methods described in (Zhang) and merged with the mutation data to form an indicator matrix of whether a point mutation or copy number aberration occurs in each gene biomarker. There are $n = 697$ patients common to both data sets, and our analysis considers the top 2% of genes filtered by expression variance across samples ($p_Y = 329$) and gene aberrations that occurred in at least 15% of patients ($p_Z = 177$). As RNA-sequencing data is count-valued and the mutation status is binary, we fit our Truncated Poisson - Ising Manichean graphical model to this data. Stability selection (Liu et al., 2010) was used to determine the optimal level of regularization. Results visualized in Figure 2 show highly connected modules exhibiting *within* connections and identified several *between* connections that are consistent with the cancer genomics literature. For example, TP53 is known to be highly mutated in breast cancer and a regulator of gene expression. Two such genes that have been experimentally validated as influenced by TP53 mutations, DLK1 and THSD4 (Lin et al., 2010; Wu et al., 2010), were identified as inter-connected neighbors to TP53 in our graph. Overall, the formulation of mixed graphical models via node-conditional exponential family distributions permits us to learn connections between heterogeneous variables such as genomic cancer biomarkers.

Acknowledgements

E.Y. and P.R. acknowledge support from NSF via IIS-1149803 and DMS-1264033, and ARO via W911NF-12-1-0390. Y.B. and G.A. acknowledge support from NSF DMS 1209017 and 1264058 Z.L. acknowledges support from NSF DMS-1263932 and NIH-1DP5OD009134.

References

G. I. Allen and Z. Liu. A local poisson graphical model for inferring networks from sequencing data. *IEEE Trans. NanoBioscience*, 12(1):1–10, 2013.

J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):192–236, 1974.

T. T. Cai, H. Li, W. Liu, and J. Xie. Covariate-adjusted precision matrix estimation with an application in genetical genomics. *Biometrika*, 100(1): 139–156, 2013.

Cancer Genome Atlas Research Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.

J. Cheng, E. Levina, and J. Zhu. High-dimensional mixed graphical models. *arXiv preprint arXiv:1304.2810*, 2013.

A. Dobra and A. Lenkoski. Copula gaussian graphical models and their application to modeling functional disability data. *The Annals of Applied Statistics*, 5 (2A):969–993, 2011.

M. Frydenberg and S. L. Lauritzen. Decomposition of maximum likelihood in mixed graphical interaction models. *Biometrika*, 76(3):539–555, 1989.

J Lafferty. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning (ICML-2001)*, 2001.

S. L. Lauritzen. Propagation of probabilities, means, and variances in mixed graphical association models. *Journal of the American Statistical Association*, 87 (420):1098–1108, 1992.

S. L. Lauritzen. *Graphical models*. Oxford University Press, 1996.

S. L. Lauritzen and N. Wermuth. Graphical models for associations between variables, some of which are qualitative and some quantitative. *The Annals of Statistics*, pages 31–57, 1989.

S. L. Lauritzen, A. H. Andersen, D. Edwards, K. G. Jöreskog, and S. Johansen. Mixed graphical association models [with discussion and reply]. *Scandinavian Journal of Statistics*, pages 273–306, 1989.

J. D. Lee and T. J. Hastie. Learning mixed graphical models. *arXiv preprint arXiv:1205.5012*, 2012.

S. Lee, J. Zhu, and E. P. Xing. Adaptive multi-task lasso: with application to eqtl detection. In *Advances in neural information processing systems*, pages 1306–1314, 2010.

A. Lin, R. T. Wang, S. Ahn, C. C. Park, and D. J. Smith. A genome-wide map of human genetic interactions inferred from radiation hybrid genotypes. *Genome research*, 20(8):1122–1132, August 2010.

H. Liu, K. Roeder, and L. Wasserman. Stability approach to regularization selection (stars) for high dimensional graphical models. *Arxiv preprint arXiv:1006.3316*, 2010.

H. Liu, F. Han, M. Yuan, J. Lafferty, and L. Wasserman. High-dimensional semiparametric gaussian copula graphical models. *The Annals of Statistics*, 40(4):2293–2326, 2012.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34:1436–1462, 2006.

P. Ravikumar, M. J. Wainwright, and J. Lafferty. High-dimensional ising model selection using ℓ_1-regularized logistic regression. *Annals of Statistics*, 38(3):1287–1319, 2010.

G. Wu, X. Feng, and L. Stein. A human functional protein interaction network and its application to cancer data analysis. *Genome biology*, 11(5):R53, 2010.

L. Xue and H. Zou. Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *The Annals of Statistics*, 40(5):2541–2571, 2012.

E. Yang, P. Ravikumar, G. I. Allen, and Z. Liu. Graphical models via generalized linear models. In *Neur. Info. Proc. Sys.*, 25, 2012.

E. Yang, P. Ravikumar, G. I. Allen, and Z. Liu. Conditional random fields via univariate exponential families. In *Neur. Info. Proc. Sys.*, 26, 2013a.

E. Yang, P. Ravikumar, G. I. Allen, and Z. Liu. On poisson graphical models. In *Neur. Info. Proc. Sys.*, 26, 2013b.

J. Zhang. Convert segment data into a region by sample matrix to allow for other high level computational analyses. Bioconductor Package.