Square Root Graphical Models: Multivariate Generalizations of Univariate Exponential Families that Permit Positive Dependencies, by David Inouye, Pradeep Ravikumar, and Inderjit Dhillon

Samson Koelle

Department of Statistics, University of Washington Seattle, WA, 98195, USA

1 Abstract

Understanding relationships between different categorical or quantitative random variables is a central problem in clustering, network reconstruction, and regression. This paper proposes a multivariate distribution based on a Markov Random Field which encapsulates multivariate Gaussian and Ising graphical models. The parameter space is expanded using a square root transformation of sufficient statistics to enable new multivariate dependencies.

2 Introduction

Markov Random Fields (MRFs) are widely used to model multivariate distributions in genomics, natural language processing, and image analysis. In these models, random variables, which we consider as taking values in a phase space, and conditional probability dependencies are represented as nodes and edges in an undirected graph, respectively. A typical statistical task in this setting is analysis of latent factors leading to observed correlations between variables. For example, some latent variable such as cell type or document topic may manifest itself as a tendency for certain genes or words to co-occur in the same cell or document. Another common task is causal inference, e.g. we wish to learn that the promotion or inhibition of expression of a certain gene may itself depend on the expression level of another gene. Importantly, both of these tasks may be accomplished through consideration of conditional probability dependencies - the altered distribution of a certain variable given the phase of another, even when conditioned on the phases of all variables other than the two in question, and much work has been done to these ends using multivariate distributions with a MRF representation, including the Ising/Potts Model for categorical data, and the multivariate Gaussian distribution (**?**) (**?**).

Unfortunately, these models are ill-suited for continuous variables which are not normally distributed, such as airport delay times, or take potentially unbounded integer values, such as gene counts from next-generation sequencing or word counts from natural-language processing. Given the complex ways in which variables with different phase spaces interact, and the increasing resolution with which these variables can be observed, the general theory of multivariate distributions is lacking. For example, the multivariate Poisson distribution does not allow for negative correlations, and the multivariate exponential distribution distribution has a singular component. Partly as a response to these difficulties, multivariate copulas have become popular for modeling multivariate conditional probability dependencies non-parametrically, or using a amenable distribution such as a multivariate Gaussian. However, the additional layer of complexity assumed by parametric copulas may not be desirable, and non-parametric models often have inferior statistical power and interpretability. A different approach is simply to Gaussianize univariate data using a power or quantile transformation, but this approach is often insensitive to meaningful outliers, and ignores potentially important-to-understand processes which generate the data.

The weaknesses of these multivariate approaches are evident in contrast with the clear framework for univariate modelling provided by generalized linear models (GLMs). GLMs allow a variable to have a wide range of conditional probability dependencies, with only the weak assumption that the conditional univariate distributions in question are from an exponential family. Exponential family distributions include the Normal, Poisson, Exponential, and Bernoulli distributions, and many others, so this condition is often met. In this situation, conditional probability dependencies can in general be optimized using methods such as gradient descent, even in high-dimensions. Often, a convex or information theoretic sparsity penalty is applied to improve interpretability and prediction. Interestingly, the repeated application of this approach to all nodes of a graph and smoothing of the results asymptotically recovers the conditional dependence structure of high-dimensional multivariate Gaussian and Ising distributions. A major line of research in statistics over the past ten years has been extending this joint-regression strategy either through some intrinsic mathematical improvement or targeting of a particular problem.

This paper widens these lines of reasoning by formulating a useful new distribution upon which they are applicable. While the normalizing constant of the multivariate Gaussian distribution can easily be calculated from its parameters, this is in general not the case, and, more problematically, the joint probability normalizing constants resulting from previous multivariate extension of univariate conditional dependencies infinite in important situations, such as when each univariate conditional distribution is exponential or Poisson, and variables have a tendency to both positively co-occur. The proposed distribution is normalizable in these situations, and though the normalizing constant is somewhat difficult to approximate, this normalizability both enables model-comparison tasks such as selection of a sparsity-inducing tuning parameter, and opens a door for analytic improvement of stochastic integration techniques. The normalizability of the proposed distribution is enabled by a nuance of previous GLM-based approaches to creating multivariate exponential family distributions whose conditional probability dependencies are equivalent to Markov Random Field, as well as, somewhat independently, a nuance of the Poisson distribution. The combination of these two factors gives the proposed multivariate Poisson distribution even less restriction on conditional probability dependencies than the multivariate Normal distribution, whose dependencies must form a positive-definite matrix. This new distribution has conditional probabilities which form a MRF, and is optimizable using well-characterized algorithms.

3 Methods

In this section, I describe the derivation of the SRGM distribution, the properties which lead to its normalizability, its optimization given data using a proximal gradient descent algorithm, and the stochastic approximation of its normalizing constant. We are given data $\mathbb{X}$ composed of a set of n i.i.d. observations of p variables $\boldsymbol{x} = \{x_1, \ldots, x_p\}$. Generally, vectors will be bolded, and matrices capitalized.

3.1 Graphical Models

A graph G consists of a set of nodes or vertices $\boldsymbol{V} = \{V_1, \ldots, V_p\}$ and a set of edges $\boldsymbol{E}$ which link the nodes in a pairwise manner. Given a multivariate vector of univariate random variables $\boldsymbol{x} = \{x_1, \ldots, x_p\}$, create a Markov Random Field by initializing p nodes, uniquely assigning univariate random variables $x_1, \ldots, x_p$ to these nodes, and constructing an edge set E which satisfies the following properties.

1. $P(x_i|\boldsymbol{x}_{-i}) = P(x_i|\boldsymbol{x}_{-i,-j})$ implies $e_{ij} \notin E$, where $\boldsymbol{x}_{-i}$ and $\boldsymbol{x}_{-i,-j}$ are the multivariate random vector $\boldsymbol{x}$ with the i-th, and i-th and j-th entries removed and e_{ij} is an edge linking the nodes corresponding to x_i and x_j.

2. $P(x_i|\boldsymbol{x}_{-i}) = P(x_i|\boldsymbol{N}(x_i))$, where $\boldsymbol{N}(x_i)$ denotes the set of univariate random variables whose corresponding nodes are connected to the node of x_i via an edge.

3. $P(\boldsymbol{x}_A|\boldsymbol{x}_{B\cap C}) = P(\boldsymbol{x}_A|\boldsymbol{x}_C)$, where A, B, and C are non-intersecting sets of nodes, and every from A to B in $\boldsymbol{E}$ passes through C.

In practice, nodes can correspond to observables such as gene expression, or latent factors which effect observables, such as chromatin state. Edges may be of quantitatively different strengths, and we often assume that they are sparse. While it may be easy to estimate conditional probabilities $P(x_i|x_j)$, ensuring the above conditions are met, i.e. that $\boldsymbol{E}$ is minimal, is a more challenging problem that will be discussed in section **??**.

The Hammersley-Clifford Theorem implies that the density of a multivariate probability measure of a Markov Random Field such as the one described in the previous paragraph factorizes as

$$P(\boldsymbol{x}) = \frac{1}{Z(\Phi)} \exp \sum_{c \in \mathcal{C}} \phi_c(\boldsymbol{x})$$

where $\mathcal{C}$ are the cliques of the graph, ϕ_c are compatibility functions which depend only on the variables whose nodes are in the clique, and $Z(\Phi)$ is a normalizing constant which depends on both the compatibility functions and the domain of $\boldsymbol{x}$. This means that when we assign variables to nodes and conditional probability dependencies to edges, the choice of compatibility functions from conditional probability simplexes is crucial for constructing a multivariate distribution. This choice is constrained however by the necessary condition that

$$Z(\Phi) = \int_{\mathcal{D}} \exp \sum_{c \in \mathcal{C}} \phi_c(\boldsymbol{x}) d\boldsymbol{x} < \infty,$$

where $\mathcal{D}$ is the domain of P. The next section introduces some background on univariate distributions which will inform our choice of compatibility functions.

3.2 Exponential Family Distributions

The univariate exponential family of distributions of a random variable x includes all probability laws of the form

$$P(x|\theta) = \exp(\eta(\theta)^T T(x) + B(x) - A(\theta)),$$

where $\eta(\theta)$ are the *natural parameters* of the exponential family, $T(x)$ is a set of *sufficient statistics*, $B(x)$ is the *base measure* of the distribution, and

$$A(\theta) = \log \int_{\mathcal{D}} \exp(\eta(\theta)^T T(x) + B(x)) dx$$

is the *normalizing constant*. This class includes many commonly encountered discrete and continuous distributions, several of which are summarized in the table below.

Name	Standard PDF	$\eta(\theta)$	$T(x)$	$B(x)$	$A(\theta)$
Exponential	$\lambda \mathrm{e}^{-\lambda x}$	$-\lambda$	x	0	$-\log \lambda$
Poisson	$\frac{\lambda^x \mathrm{e}^{-\lambda}}{x!}$	$-\log \lambda$	x	$-\log x!$	λ
Normal	$\frac{1}{\sigma\sqrt{2\pi}} \exp \frac{-(x-\mu)^2}{2\sigma^2}$	$\{\frac{\mu}{\sigma^2}, \frac{-1}{2\sigma^2}\}$	$\{x, x^2\}$	$-\frac{1}{2}\log(2\pi)$	$\frac{\mu^2}{2\sigma^2} + \log \sigma$
Multivariate Normal	$\frac{1}{\sqrt{(2\pi)^p \lvert\Sigma\rvert}} \exp \frac{-(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}{2}$	$\{\Sigma^{-1}\mu, \frac{-1}{2}\Sigma^{-1}\}$	$\{\boldsymbol{x}, \boldsymbol{x}^T\boldsymbol{x}\}$	$\frac{-p}{2}\log(2\pi)$	$\frac{\boldsymbol{\mu}^T \Sigma^{-1} \boldsymbol{\mu}}{2} + \frac{\log \lvert\Sigma\rvert}{2}$

Table 1: Exponential family parameterizations of some common distributions.

The prevalence of exponential family distributions in univariate statistics is based upon a number of useful properties, including that $T(\{x_1, \dots x_n\}) = \sum_{i=1}^n T(x_i)$, a set of fixed size, is sufficient for optimization of the model given arbitrarily large numbers of observations.

This follows immediately from the product law $P(\{x_1, \ldots x_n\}) = \prod_{i=1}^{n} P(x_i)$. Furthermore, exponential family distributions are maximum entropy distributions given given certain constraints on moments. For example, the univariate normal distribution is the maximum entropy real-valued distribution with specified first and second moments. For this reason, exponential family distributions both appear in physical and information theoretic systems, and also are the distributions which, given data, minimize the amount of prior information built into a model.

3.3 Exponential Family Markov Random Fields

For all of the above reasons, we often wish to model a particular variable with a univariate exponential family distribution. In particular, let us assume that the natural parameter η of that distribution is contingent on a set of effector variables, i.e.

$$P(x_s|\boldsymbol{x}_{-s}) = \exp(\eta(\boldsymbol{x_{-s}})T(x_s) + B(x_s) - A_s), \tag{1}$$

where A_s is the normalizing constant of the conditional distribution. As shown in (**?**), specifying a joint distribution of the form

$$P(\boldsymbol{x}) = \exp(\sum_{s=1}^{p}(\eta(\boldsymbol{x_{-s}})T(x_s) + B(x_s)) - A),$$

whose conditionals are given by (1), necessitates that the functions $\eta(x_s)$ must be of the form

$$\eta(\boldsymbol{x}_{-s}) = \sum_{c \in \mathcal{C}_s}(\theta_c \prod_{i \in c \setminus s} T(x_i)),$$

where $\mathcal{C}_s$ is the set of cliques containing site s, $B(x_s)$ is determined by the univariate exponential family of node s, A_s is a normalizing constant, and $\Theta_{\mathcal{C}_s}$ is the set of clique parameters θ_c for site s. That is, in order for a MRF to have node-conditional distributions from an exponential family, its compatibility functions must be equivalent to a scalar clique weights multiplied by the sufficient statistics of the variables in the clique. The joint distribution specified by these conditionals is therefore

$$P(\boldsymbol{x}|\Theta) = \exp\,(\sum_{s=1}^{p}(\sum_{c\in\mathcal{C}_s}(\theta_c \prod_{i\in c} T(x_i)) + B(x_s)) - A(\Theta)).$$

We will at this point restrict ourselves to considering pairwise graphical models, whose compatibility functions factor into functions of only two variables. The resultant probability law is

$$\begin{aligned} P(\boldsymbol{x}|\Theta,\Phi) &= \exp(\Theta^T T(x) + T(x)^T \Phi T(x) + \sum_{s=1}^{p} B(x_s) - A(\Theta,\Phi)) && (2) \\ &= \exp(\Theta^T T(x) + T(x)^T \Phi T(x) + \boldsymbol{B}(\boldsymbol{x}) - A(\Theta,\Phi)) && (3) \end{aligned}$$

where $\boldsymbol{B}(\boldsymbol{x}) := \sum_{s=1}^{p} B(x_s)$. Here, Θ and Φ are arrays of clique-specific parameters of orders 1 and 2 respectively, i.e. $\Theta \in \mathbb{R}^p$ and $\Phi \in \mathbb{R}^{p\times p}$. These arrays completely specify the underlying compatibility functions. In particular, the diagonal terms of Φ must be equal to zero in order for node conditionals to be of the form

$$\begin{aligned} P(x_s|\boldsymbol{x}_{-s}) &= \exp(\Theta_s T(x_s) + 2T(\boldsymbol{x_{-s}})^T \Phi_{-s} T(x_s) + B(x_s) - A_s) && (4) \\ &= \exp(\Theta_s T(x_s) + \sum_{eN(s)} \Phi_e T(x_s) + B(x_s) - A_s). && (5) \end{aligned}$$

Note that in the case $T(x) = x$, the node conditionals of this MRF are generalized linear models.

This class clearly includes the Ising model via the parameterization $\{\Theta = \mathbf{0}$, $x_s \in \{-1, 1\}$, $\Phi_{ij} \in \mathbb{R}\}$ and appropriate normalizing constant. It also includes the multivariate Gaussian and distribution via the parameterization $\Theta = \Sigma^{-1}\mu$ and $\Phi = -\frac{1}{2}\Sigma^{-1}$. Since Σ^{-1} has non-zero diagonal elements, (**?**) proposes that in order for the diagonal terms of Φ to be 0,

$$\boldsymbol{B}(\boldsymbol{x}) = \frac{-p}{2}\log(2\pi) - \sum_{s=1}^{p} \frac{{x_s}^2}{2\Phi_{ss}^2}.$$

This violates the condition that $B(x)$ does not depend on parameters, unless we make the additional assumption that variance is known. As an alternative approach, (**?**) adds self-

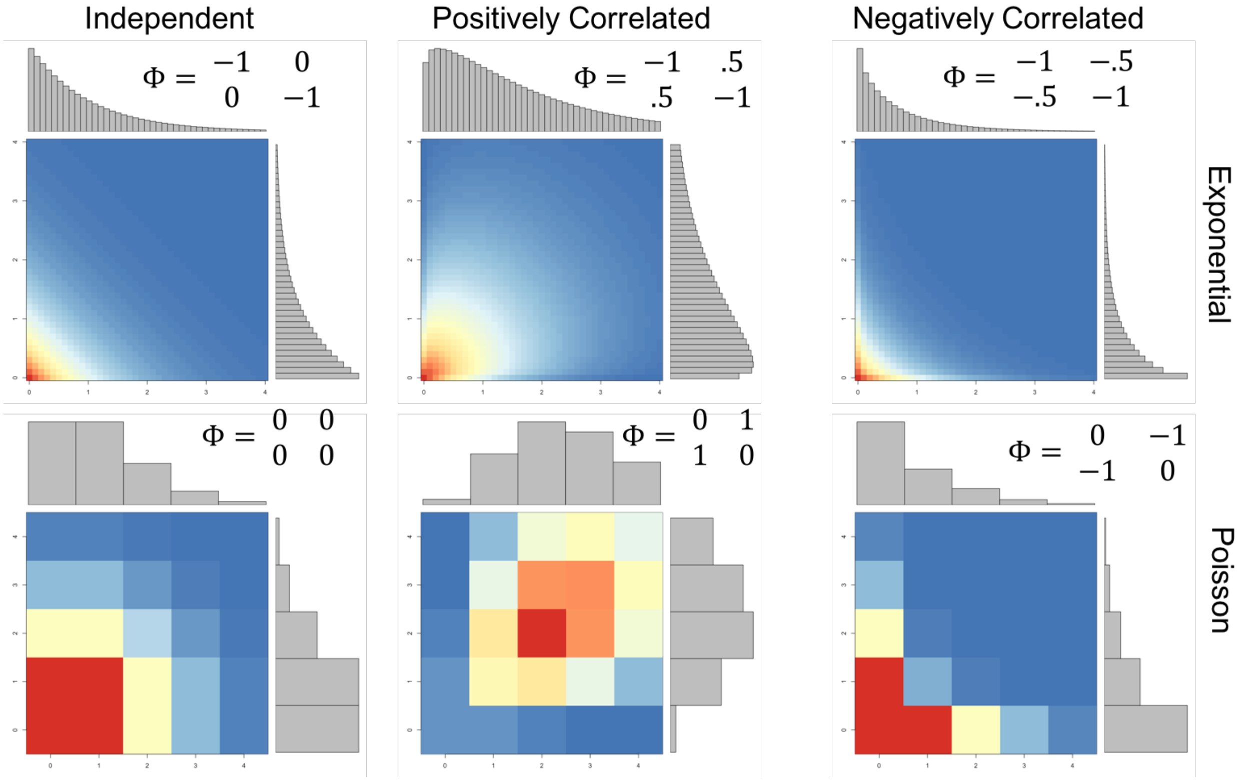


Figure 1: Selected square root graphical model probability densities with $\Theta = \mathbf{0}$. The independent parameterization of the Poisson follows from $\eta(\theta) = \log\lambda$, as in Table 1.

edges to the edge set $\boldsymbol{E}$. Thus, each node is in its own neighborhood, and non-zero diagonal terms of Φ are allowed. Note that self-edges satisfy the Markov property, and, by their role in the multivariate Gaussian graphical model, are necessary to add to G. This nuanced difference is partly why (**?**) reported unsatisfactory multivariate generalizations of the Poisson and exponential distributions. Since $T(\boldsymbol{x})^T\Phi T(\boldsymbol{x}) \in O(x^2)$ while $T(\boldsymbol{x})^T\Theta \in O(x)$, Φ must be positive-definite. In these cases, zeroes on the diagonals of Φ render positive values elsewhere in Φ equivalent to non-normalizability. This situation is not entirely analogous to the Gaussian case, however, since the Gaussian distribution has a quadratic power term, while the exponential and Poisson distributions do not.

3.4 Square Root Graphical Models

The proposed *Square Root Graphical Model* (SRM) distribution

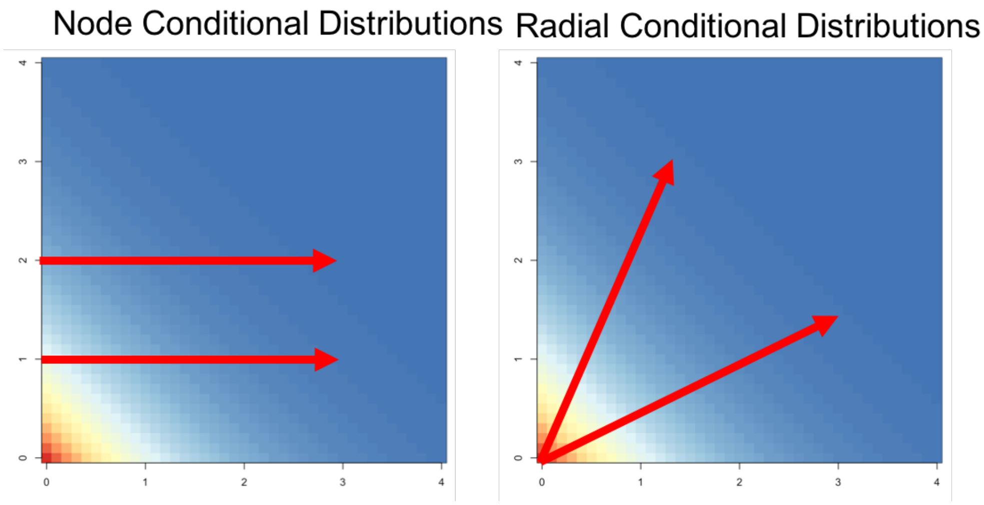
$$P(\boldsymbol{x}|\Theta, \Phi) = \exp(\Theta^T \sqrt{T(\boldsymbol{x})} + \sqrt{T(\boldsymbol{x})}^T \Phi \sqrt{T(\boldsymbol{x})} + \sum_{s=1}^{p} B(x_s) - A(\Theta, \Phi))$$

ameliorates this concern by using square roots of sufficient statistics. The diagonal terms of Φ are therefore non-zero and in $O(x)$ when composed with the square root sufficient statistics, which can be considered as simply a redefinition of the sufficient statistic $\widetilde{T(x)} := \sqrt{T(x)}$. These diagonal terms are the scalar which completely characterize the compatibility functions associated with self-edges. To recover the multivariate Gaussian, we set entry-wise $T(x) = x^2$ and replace $\sqrt{T(\boldsymbol{x})}$ with sgn $(\boldsymbol{x})\sqrt{T(\boldsymbol{x})}$ in the above expression, while otherwise use the same parameterization as in the previous section. Note that the choice $T'(x) = T(x)^2$ in the SRM distribution generically recovers the non-square root distribution described in the previous section. Some example Poisson and exponential SRM distributions are displayed in **??**.

The node conditionals of the SRM distribution are straightforward to observe by considering that Θ_s and the $s - th$ row and column of Φ are the only factors which effect the distribution of each random variable x_s. Separating terms into those which are multiplied by $\sqrt{x_s}$ and x_s, we see that

$$P(x_s|x_-s, \eta_{1s}, \eta_{2s}) = \exp\left(\eta_{1s}^{node}\sqrt{x_s} + \eta_{2s}^{node} x_s + B(x_s) - A(\eta_{1s}^{node}, \eta_{2s}^{node})\right)$$

where $\eta_{1s}^{node} = \Theta_s + 2\Phi_{-s}\sqrt{x_{-si}}$ and $\eta_{2s}^{node} = \Phi_{ss}$. This distributions illustrates the useful difference of the SRM compared to the previous model: Node-conditionals are identical to their univariate exponential families in the case that $\eta_1 = 0$, and because the dominating term in the node-conditional distribution is of the same order as the base exponential family, the node-conditionals have tail behavior which corresponds to the base distribution, alleviating a concern of previous models (**?**). Note that this is not the case in the non-square root model for any distribution whose sufficient statistics are in $\emptyset(x)$, and that keeping the term of largest order in the sufficient statistics, in this case Φ_{ss}, at the same order as the natural parameter of the univariate node-conditional distribution exponential family distribution


Node Conditional Distributions

Radial Conditional Distributions

Figure 2: Example node and radial conditional distributions

ensures normalizability. In this way, square root sufficient statistics are often a natural choice pairwise graphical models.

3.5 Normalizability of Square Root Graphical Models

The SRM has a natural parameter space

$$\mathcal{N} = \{(\Theta, \Phi) | \int A(\Theta, \Phi) < \infty\},$$

where

$$A(\Phi, \Theta) = \log \int_{\mathcal{D}} \exp\left(\Theta^T \sqrt{T(\boldsymbol{x})} + \sqrt{T(\boldsymbol{x})}\Phi\sqrt{T(\boldsymbol{x})} + \sum_{s=1}^{p} B(x_s)\right) d\mu(\boldsymbol{x}). \tag{6}$$

The domain of the natural parameter of the SRM distribution depends only on the sufficient statistics and base measures and base measures of conditional probability distributions of the sites which make up the base graph, as well as the natural parameter spaces of the node-conditionals. Observe the extended normalizability of the SRM compared with the previous model by radially factorizing $A(\Theta, \Phi)$ along $\mathcal{D}$:

$$A(\Theta, \Phi) = \log \int_{\mathcal{V}} \int_{\mathcal{Z}(\mathcal{V})} \exp\left(\Theta^T \sqrt{T(z\boldsymbol{v})} + \sqrt{T(z\boldsymbol{v})}\Phi\sqrt{T(z\boldsymbol{v})} + \sum_{s=1}^{p} B(zv_s)\right) d\mu(z) d\boldsymbol{v}, \tag{7}$$

where $\mathcal{D}$ is the domain of the probability measure, $\mathcal{V} = \{\frac{|\boldsymbol{v}|}{\|\boldsymbol{v}\|_1} = 1, \boldsymbol{v} \in \mathcal{D}\}$, and $\mathcal{Z}(\mathcal{V}) = \{z \in \mathbb{R}_+ | z\boldsymbol{v} \in \mathcal{D}\}$. Since $\mathcal{V}$ is bounded and exp and log are continuous, A is finite as long as

$$\begin{aligned} A^{rad}(\Theta, \Phi) :&= \int_{\mathcal{Z}(\mathcal{V})} \Theta^T \sqrt{T(z\boldsymbol{v})} + \sqrt{T(z\boldsymbol{v})}\Phi\sqrt{T(z\boldsymbol{v})} + \sum_{s=1}^{p} B(zv_s) d\mu(z) \\ &= \int_{\mathcal{Z}(\mathcal{V})} (\Theta^T \sqrt{\boldsymbol{v}})\sqrt{z} + (\sqrt{\boldsymbol{v}^T}\Phi\sqrt{\boldsymbol{v}})z + \sum_{s=1}^{p} B(zv_s) d\mu(z) \\ &= \int_{\mathcal{Z}(\mathcal{V})} \eta_1^{rad}\sqrt{z} + \eta_2^{rad} z + \sum_{s=1}^{p} B(x_s) d\mu(z) < \infty, \end{aligned}$$

where $\eta_1^{rad} = \Theta^T\sqrt{\boldsymbol{v}}$, $\eta_2^{rad} = \sqrt{\boldsymbol{v}^T}\Phi\sqrt{\boldsymbol{v}}$. This is the case if if $\eta_2 < 0$ or $\eta_2 = 0$, and $\eta_1 \leq 0$. Here we have assumed that $T(x) = x$, as is the case in the Exponential and Poisson distributions. Note that this general condition is weaker than negative definiteness, but nevertheless requires that the diagonals of $\Phi \leq 0$. However, this is not required for distributions such as the SRM Poisson whose base measure is decreasing. In particular, $\lim_{x\to\infty} \log \frac{1}{x!} < x^{-1}$.

3.6 Selection of Model Parameters

The consistency of M-estimators for estimating model parameters was studied by (**?**) in the case of the general exponential family MRF, and is assumed to be applicable throughout. Given a data matrix $\mathbb{X} \in \mathbb{R}^{n\times p}$ containing n i.i.d. observations of p variables, the authors seek to estimate the model parameters $\Phi \in \mathbb{R}^{p\times p}$ and $\Theta \in \mathbb{R}^p$ that maximize the joint likelihood

$$L(\Theta, \Phi | \mathbb{X}) = \prod_{i=1}^{n} \exp(\Theta^T \sqrt{T(x)} + \sqrt{T(x)}^T \Phi \sqrt{T(x)} - A(\Theta, \Phi)).$$

After adding a sparsity inducing penalty on the off-diagonal terms of Φ, this is approximated by minimizing the sum of the penalized node conditional negative log likelihood joint objective functions, which is

$$O(\Theta, \Phi|\mathbb{X}) = \sum_{s=1}^{p}(-\frac{1}{n}\sum_{i=n}^{n}(\eta_{1si}\sqrt{x_{si}} + \eta_{2si}x_{si} - A_{node}(\eta_{1si}, \eta_{2si})) + \lambda\|\Phi_{-s}\|_1).$$

Here, $\eta_{1si} = \Theta_s + 2\Phi_{-s}\sqrt{x_{-si}}$ and $\eta_{2si} = \Phi_{ss}$. Note that this is an exact approximation of the joint likelihood except for the convolution of the normalizing constants.

Optimization via proximal gradient descent necessitates that the gradients

$$\begin{aligned}
\frac{dO}{d\phi_{ss}} &= \frac{-1}{n}\sum_{i=1}^{n} x_{si} - \frac{dA_{node}}{d\eta_2}, \\
\frac{dO}{d\phi_{-s}} &= \frac{-2}{n}\sqrt{\mathbb{X}_{-s}}^T\sqrt{\mathbb{X}_s} - 2\sqrt{\mathbb{X}_{-s}}\frac{dA_{node}}{d\eta_1} \qquad \text{and} \\
\frac{dO}{d\theta_s} &= \frac{-1}{n}\sum_{i=1}^{n}\sqrt{x_{si}} - \frac{dA_{node}}{d\eta_1}.
\end{aligned}$$

be computed. Therefore, we compute

$$\frac{dA_{node}}{d\eta 1} = \frac{(A_{node}((\eta_1 + \epsilon), \eta_2) - A_{node}((\eta_1, \eta_2)))}{\epsilon}$$

$$\frac{dA_{node}}{d\eta_2} = \frac{(A_{node}((\eta_1), \eta_2 + \epsilon) - A_{node}((\eta_1, \eta_2)))}{\epsilon}$$

using an approximation of A_{node}. This is easy to compute since A_{node} is an integral over one dimension, but has an analytic solution

$$A_{node}(\eta_1, \eta_2) = \sqrt{\pi}\eta_1\frac{\exp\left(\frac{\eta_1^2}{-4\eta_2}\right)}{(2(-\eta_2)^{1.5})}(1 - \operatorname{erf}\left(\frac{-\eta_1}{2\sqrt{-\eta_2}}\right)) + \frac{1}{-\eta_2}$$

in the case that the node-conditional is exponential. The suggested value of ϵ is 0.0001.

After gradients are computed, the optimization algorithm proposes new values of Φ and Θ which move increase the likelihood. After the off-diagonal values of each proposed Φ are soft-thresholded using the proximal operator $ST(\Phi_{off}) = \operatorname{sgn}(\Phi_{s-s})(\Phi_{s-s} - \lambda)_+$, the penalized node-conditional probability of the move is computed, and significant improvements are accepted. After each computation of the gradient, large jumps are proposed, and if these are rejected, smaller jumps are considered.

3.7 Estimation of Normalization Constant

Given model parameters Θ and Φ, we use Annealed Importance Sampling to estimate the normalizing constant $A(\Theta, \Phi)$. AIS is an instance of importance sampling in the sense that the magnitude of a certain distribution is compared to another at certain points, thereby enabling comparison of their normalizing constants. In our case, this comparison is made over a sequence of intermediate densities

$$\hat{P}_i(\boldsymbol{x}|\Theta_i, \Phi_i) = \exp(\Theta_i^T \sqrt{T(\boldsymbol{x})} + \sqrt{T(\boldsymbol{x})}^T \Phi_i \sqrt{T(\boldsymbol{x})} + \sum_{s=1}^{p} B(x_s))$$

where $\Phi_i = \Phi_{diag} + \frac{i}{n}(\Phi - \Phi_{diag})$, and $\Theta_i = \frac{i}{n}\Theta$. P_0 is simply a probability law composed of the product of independent exponential family distributions, and therefore has a computable normalizing constant upon which we can iteratively compare.

The usefulness of AIS stems from the slight differences between distributions which are being compared. Subsequent distributions can be sampled from efficiently using MCMC based upon the previous distribution. For each particle path $x = \{x_0, \ldots x_n\}$, we compute the weight $w(x) = \prod_{i=1}^{n} \frac{f_{i-1}(x_{i-1})}{f_i(x_{i-1})}$. The average of the sample weights converges to the ratio of the normalizing constants of P_0 and P_n. Since, the normalizing constant of P_0 is calculable by summing the normalizing constants of the independent univariate exponential family distributions with parameters given by the diagonals of Φ, we can solve for $A(\Theta, \Phi)$.

4 Results

5 Discussion

As it turns out, there is a semigroup of parametric distributions, including the multivariate Normal, for which the inverse covariance matrix, also known as the precision matrix, of the variables corresponds to the edge structure of the graph. However, this is definitely not the case when the number of parameters p is greater than the number of observations n, and also fails in the examples that we consider from non-Gaussian multivariate distributions. Dempster.

Note the importance of the normalizing constant in parametric model selection. While a probability law which approximates data can be learned non-parametrically, a parametric distribution is often preferable when prior knowledge is assumed. Suppose we wish to estimates cell-specific networks of genes from the count numbers of these genes in the cytoplasma. Not Suppose we wish to estimate the developmental trajectory of possibly reproducing labelled sets of objects. This task can be accomplished parametrically or non-parametrically. can be useful for model selection, orAn example of learning a parametric distribution from a non-parametric one is creating a mixture model from non-parametric clusters. The crucial goal of research on multivariate distributions in the high-dimensional setting is estimability of the model given data, for which a parallel inference approach may be valuable. The normalizable multivariate distributions presented here node-conditionals which are which are In many cases, a normalization free inference mechanism such as score-matching, or a parallel regression strategy which does not consider normalizability may be used. In the cases where a Interpreting self-edges in a conditional probability model is not obvious. Variables have a trivial conditional probability dependence on themselves which can be extended into a n-simplex of probability relationships using n-roots of sufficient statistics, meaning that it seems like the sufficient statistics of any variable to any integer power are available through trivial self-cliques and a valid MRF distribution is maintained. By Yang's theorem, the component sufficient statistics must be n-roots of the usual ones, and multivariate dependencies must necessarily be modeled at this level.

The GLM-based approach to is rooted in (?) and (?), who proved a necessary tensor-factorization of conditional probability dependencies when these dependencies are modelling using a Markov Random Field.

are modeled through clique-specific "compatibility functions", which depend only on subsets of variables of each clique, i.e. simplicial combination of non-zero conditional probability dependencies, of the base graph.

Counting cliques is a hard problem.

When the sufficient statistic is a linear function $T(X) = X$, then the order two multivariate exponential family is in fact a generalized linear model, and this linear effect node-conditional model can be extended to graphical models in which each site in the graph may have a probability density corresponding to different exponential families, such as Bernoulli and Normal. The arumIn order to ensure that the resulting multivariate distribution is normalizable, the natural parameter space can be restricted to negative or positive definite matrices, but in cases where the base measure decreases with the sufficient statistic, this condition may be relaxed because the decrease in base measure cancels out increases in sufficient statistics. The Poisson distribution has the quickly decreasing base measure $\frac{1}{x!}$, which enables . In particular,

On the other hand, if univariate marginal distributions are not exponential families, and therefore (by the theorem of Pitman-Koopman-Darmois) do not have bounded sets of node-specific sufficient statistics, then the number of potential interactions between which will explode since the dimensionalities of the sufficient statistics are each themselves growing. Dependencies between parameters are often interesting in their own right or as part of the regression problem of optimizing θ in $Y \approx \mathbb{X}\theta$ where $\mathbb{X} = \{X_1, \dots X_n\}$. (VM's comment: **You are operating with mathematical notation without defining it; very hard to understand what you are trying to say**) Sufficiency is a poorly understood notion in the multivariate case. Note that inferring dependencies locally is also a regression problem $X_a \approx \mathbb{X}_{-a}\theta_a$, where $\mathbb{X}_{-a}$ are adjacent sites of parameter a and θ_a are their edge weights (**?**). [[Score matching These models may be useful for statistical tasks such as clustering, network reconstruction, regression, as well as for inference of latent factors.

Recently, there has been work extending these models to combining distributions, both in specific mixed data genomic settings, for example in the case when both chromatin binding states and mRNA expression levels of thousands of genes can be assayed simultaneously. Other attempts to create a multivariate Poisson distribution have been unsuccessful for alternate reasons such computational complexity, restrictions to positive-interactions, or unnatural transforms applied to sufficient statistics (Inouye 2013, 11-16).

Cross-validation

References

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, July 2008.