Understanding the Development of Mixed Graphical Models

Landon Buechner
Texas A&M University

Abstract

Two of the most well studied Markov random fields are the Gaussian and Ising graphical models. Despite being a powerful framework within which to model the covariance structure of multivariate data, these models are limited by the constraint that all variables must come from a single distribution. Note that datasets found in the real world typically contain a mix of both discrete and continuous data. Further still, there is no guarantee that any of the variables come from the same named distribution. In this paper the class of mixed exponential MRFs (Yang et al. 2014) is introduced followed by a derivation of the joint density for the Gaussian-Potts model. Various parameter estimations schemes are discussed in order to motivate a structure learning algorithm for the Gaussian-Potts model (Lee et al. 2014). The goal of this review is to serve as a starting point for further investigations of undirected graphical models.

1 Introduction

Probabilistic graphical models is a sub-field of statistics that is incredibly vast. In fact, most of modern machine learning models have some form of graphical representation. The field can be subdivided into two subsets: directed and undirected graphical models. The latter, also known as Markov Random Fields (MRFs), is the main subject reviewed in this survey. The focus will be on a class of parametric graphical models constructed from the exponential family of distributions.

1.1 Undirected Graphical Models

It can be seen that the well known Ising, Potts, and Gaussian graphical models (GGM) are characterized by the property that the node-wise marginal distribution are members of the exponential family; namely the Gaussian, Bernoulli, and multinomial distributions respectively. Formalized by (Yang et al. 2012), models of this type are called exponential family graphical models. They are a powerful framework within which to model the conditional relationships between random variables and are used in a variety of fields to better understand multivariate data. For example, neuroscientists use GGMs to analyse the functional connectivity between regions in the brain. It is interesting to note that in the statistical literature 'omics' (genomics, proteomics, etc) serves as a popular test bed for statistical research because of the wide range of complex biological interactions that lend themselves to graphical modelling.

In order to contextualize the ideas discussed in this review, consider the following modelling problem. For a given collection of airports in the United States it can be seen that, under a necessary transformation, daily departure delays are normally distributed. Airline companies as well government organizations such as the FAA need to understand how delays affect traffic on the tarmac in order to optimize operations. It is reasonable to assume they may ask a question such as, what is the expected departure delay for flights out of IAH given that a ATL and ORD are experiencing delays? Questions of this type can be asked for every airport and all together can be understood as a network of airports connected through departure delays. A Gaussian graphical model would be appropriate for this situation but it would be limited for the following reason. Weather is arguably the leading cause for departure delays, yet the mentioned model does not take into account this information. Imagine that for each airport we observe both departure delays as well as a categorical weather state. In the language of conditional dependence this allows use to ask, what is the expected departure delay at IAH given that is is raining at ATL and snowing at ORD? In general, graphical

Proceedings of the 24$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s).

models are useful when studying how departure delays propagate from airport to airport. Despite this, it is not clear how to incorporate categorical weather information into a Gaussian graphical model. This is the main limitation of exponential family graphical models because most real world datasets contain a mix of both discrete and continuous types.

To resolve this shortcoming, Yang and his collaborators extended their original class of exponential family MRFs by introducing *mixed* exponential family MRFs (Yang et al. 2014). Together these two papers outline how to construct an incredibly descriptive and flexible class of graphical models and are the main subject of interest in this survey. The Gaussian-Potts model is derived and used to review the current state of parameter estimation research for undirected graphical models. Before moving forward it is of interest to discuss a landmark paper by Julian Besag which was fundamental to the development of exponential family graphical models.

The keystone paper by (Besag 1974) showed, via the Hammersley-Clifford Theorem, that a Markov random field can be constructed by specifying conditional distributions belonging to the exponential family. His work is of importance because it extended the research of Hammersley-Clifford which connected Gibbs and Markov random fields. At the time Besag was interested in spatial lattice models and considered applications in ecology. For example, consider analysing how an infection spreads through crop sites laid out in a rectangular grid.

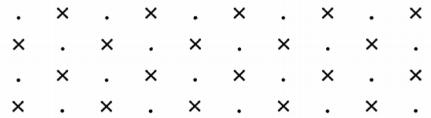


FIG. 1. Coding pattern for a first-order scheme.

$$\prod p_{i,j}(x_{i,j};\, x_{i-1,j}, x_{i+1,j}, x_{i,j-1}, x_{i,j+1}),$$

It is important to note that this model has a predefined graph structure, namely it assumes that there is no interactions between crops beyond neighboring sites. Besag considered more general spatial interaction models where, given a set of conditional distributions of the form $P(X_i|X_1, ..., X_{i-1}, X_{i+1}, ..., X_n)$, it is of interest to construct a joint probability distribution.

Prior to his work models such as the Ising model in statistical physics and the Gaussian graphical model were already thoroughly researched. His work is significant because it established a general class of Markov random fields within which the aforementioned models are members of. As we will see, the development of increasingly more general families of graphical models is a reoccurring theme in the research covered in this survey.

2 Exponential Family MRF's

Define $G = (V, E)$ as the undirected graph over random vector $X = (X_1, \ldots, X_p)$ with node set $V = \{1, \ldots, p\}$ and edge set $E = V \times V$. Note that X is a p-dimensional random vector with each variable taking values in a set $\mathcal{X}_r$.

Theorem 1 (Hammserley-Clifford + Besag + Yang). *If X has marginal distributions arising from the exponential family and satisfies the Markov independence properties*

- *Global:* $X_A \perp\!\!\!\perp X_B \,|\, X_C$ *is satisfied for any* $A, B, C \subset V$ *where C is a separating sub-graph.*
- *Local:* $X_r \perp\!\!\!\perp X_{V \setminus N(r)} \,|\, X_{N(r)}$ *for every node r with neighbors* $N(r) \subset V$.
- *Pairwise:* $X_r \perp\!\!\!\perp X_t \,|\, X_{V\setminus\{r,t\}}$ *when* $t \notin N(r) \subset V$

then the joint distribution can be factorized over the set maximal cliques $\mathcal{C}$ which encode the conditional independence structure of the graph G.

Let $\phi_c(X_c)$ be a sufficient statistic dependent on the collection of random variables $\{X_r\}$ that correspond to the clique $c \in \mathcal{C}$. The exponential family of Markov random fields then have joint densities of the form

$$P(X) \propto \prod_{c \in \mathcal{C}} \exp\left\{\phi_c(X_c)\right\}$$

Note that it is not specified whether or not the sufficient statistics must come from a single or a collection of distributions in the exponential family. This is because Theorem 1 holds true for both the *homogeneous* and *heterogeneous* exponential family Markov random fields whose definitions will be made precise below.

Before preceding, I quickly review the defining characteristics of the exponential family. We will see that these distributions serve as building blocks for a large class of undirected graphical models. A univariate random variable Z with distribution in the exponential family $\mathcal{F}$ has a density of the form

$$P(Z) = \exp\left\{\eta^T B(Z) + C(Z) - A(\eta)\right\}$$

Some of the most well known distributions such as the Gaussian, Poisson, beta, and multinomial are members of $\mathcal{F}$. The canonical parameter η, sufficient statistic $B(Z)$, base measure $C(Z)$, and log-partition function $A(\eta)$ stated above are specified by the choice of distribution. It is important to note that these densities are only defined when the canonical parameter is in the natural parameter space $\mathcal{N} = \{\eta : D(\eta) < \infty\}$. When considering distributions in the exponential family we can write $\eta = \eta(\theta)$ where $\eta : \Theta \to \mathcal{N}$ is a bijection and Θ is given a parameter set.

2.1 Homogeneous MRFs

The work by (Yang et al. 2012) investigated the class of homogeneous Markov random fields characterized by the property that for each node $r \in V$ the marginal distribution of X_r arises from the *same* distribution in $\mathcal{F}$. A readily available example is the Gaussian graphical model. Building upon the work of (Besag 1974), they showed that a Markov random field can be constructed given a set of node-neighborhood conditional distributions. In particular, the distribution of each X_r conditioned on the remaining nodes $X_{V\setminus r} = (X_1, ..., X_{r-1}, X_{r+1}, ..., X_p)$. Note that canonical parameter is a function of the form $\eta_r \equiv \eta_r(X_{V\ \setminus r}, \theta)$.

$$P(X_r|X_{V\ \setminus r}) = \exp\left\{\eta_r B_r(X_r) + C_r(X_r) - D_r(\eta_r)\right\}$$

This is precisely the form of a generalized linear model and for any given distribution in $\mathcal{F}$ the natural parameter space $\mathcal{N}$ is well known. The properties that make the exponential family nice to work with combined with the results of (Besag 1974) leads to a natural approach to estimating the conditional independence structure of X.

Recall that the class of homogeneous exponential family MRFs is limited to modelling data of of a single type (continuous or discrete) and consequently a single distribution. If the homogeneity condition were to be relaxed it is not apparent whether or not a valid MRF could be constructed. Additionally, if such a class of graphical models did exist, what would be the natural parameter space? The answers to these questions are answered in (Yang et al. 2014).

2.2 Heterogeneous MRFs

In the formulation of both the homogeneous and heterogeneous exponential family MRFs, Yang and his collaborators consider potential higher order interactions between random variables. Instead, this survey considers pairwise interaction models for a number of reasons. Pairwise graphical models are usually sufficient to describe any given multivariate data found in practice. Furthermore, it is quite hard to find datasets that express high order interactions while still being in the wheelhouse of parametric statistics. This narrowing of focus is not a serious detriment to the depth of this review because higher order interactions can be accounted for by simply adding a few additional terms to the joint density.

Theorem 2 (Yang et al. 2014)**.** *Let $X = (X_1, \ldots, X_p)$ be a p-dimensional random vector. The collection of conditional distributions $P(X_r|X_{V\setminus r}) \in \mathcal{F}$ are sufficient to construct a pairwise MRF over the random vector X that is Markov with respect to a graph $G = (V, E)$ if and only if the functions $\{\eta_r(\cdot)\}_{r\in V}$ that specify the conditional distributions have the form*

$$\eta_r(\theta, X_{V\setminus r}) = \theta_r + \sum_{(r,t)\in E} \theta_{rt} B_t(X_t)$$

These conditional distributions result in the density

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

with log-partition function

$$A(\theta) = \log \int_X \exp\left\{\sum_{r\ \in V} \theta_r B_r(X_r) + \sum_{(r,t)\in E} \theta_{rt} B_r(X_r) B_t(X_t) + \sum_{r\ \in V} C_r(X_r)\right\} dX$$

In other words, for a given node $r \in V$ the functions η_r must be linear combinations of the sufficient statistics of the remaining $p - 1$ random variables. It is important to note that Theorem 2 outlines how to construct *mixed* exponential family MRFs. Thus the sufficient statistic $B_t(X_t)$ for each neighbor X_t of node r can be potentially different. Also it is not immediately clear that $P(X)$ contains terms corresponding to distinct distributions in $\mathcal{F}$. Assume X_1 is Gaussian, then $B_1(X_1)$ contains both a linear and quadratic term. If X_2 is Poisson then $B_2(X_2)$ is just a linear term. Within the theoretical constraints which are beyond the scope of this survey, an ambitious statistician could hypothetically count out every distribution in $\mathcal{F}$ and construct a valid mixed graphical model. Conclusively, $P(X)$ does in fact provide a construction of an MRF for a mixture of distributions, it just a compact representation. See that if all of the node-wise distributions were Gaussians then the mixed MRF would reduce to a Gaussian graphical model. It is easy to see that collection of homogeneous graphical models is contained in this larger family. The class of heterogeneous MRFs is incredibly rich.

Alluding to the philosophy of Manichaeism, a school of thought based on a dualistic interpretation of the universe, (Yang et al. 2014) call the subset of pairwise models *Manichean* MRFs. Researchers across disciplines would benefit from using mixed pairwise MRFs because data in the real world typically takes on a combination of both continuous/discrete and left-skewed/right-skewed. It is pertinent to mention an example in order to showcase the relevance of mixed MRFs. The application discussed in (Yang et al. 2014) shows how connections between genetic biomarkers and genomic mutation biomarkers in a genetic breast cancer network can be identified via a Poisson-Ising model. The observed RNA-sequence counts from each patient follow a Poisson distribution and the mutation state for each biomarker is Bernoulli. With this data they fit a Poisson-Ising model and found that the estimated graph structure accurately reflects experimental evidence stated in the cancer genomics literature.

To take things full circle consider the previously discussed airline delay network. We now know how to account for the categorical weather states in light of Theorem 2. This particular heterogeneous MRF is called the Gaussian-Multinomial or Gaussian-Potts model and will be the object of interest for the remainder of this work.

3 Gaussian-Potts Model

In (Yang et al. 2014) the Gaussian-Potts model is not fully discussed and the following derivation of the MRF is my own. As prescribed, we first need to find the sufficient statistics and base measure for the Gaussian and multinomial distributions. Note that in their work they provide the form of the Gaussian-Ising model and state the theoretical properties which ensure the MRF is well defined. I leave that particular derivation to future investigations since Gaussian-Potts model is known to exist.

Define $G = (V, E)$ as the graph that encodes the conditional independence structure of the partitioned random vector $X = (Y, Z)$. Let (V_Y, E_Y), (V_Z, E_Z), and (V_{YZ}, E_{YZ}) be the two homogeneous sub-graphs and the heterogeneous sub-graph respectively and see that the first two correspond to Gaussian and Potts graphical models. The innovation of mixed MRFs is that we can now model cross-type interactions. Let $\{Y_r\}$ be the set of univariate Gaussians with domain $\mathcal{Y}_r = \mathbb{R}$ and known variances σ_r^2. Let $\{Z_r\}$ be the set of multinomial random variables with each node $r' \in V_Z \subset V$ taking on values in a potentially unique finite set $\mathcal{Z}_{r'} = \{0, 1, 2, \ldots\}$. Yang and his collaborators provide the Gaussian terms and I supplement the multinomial sufficient statistic and base measure.

$\mathcal{F}$	$B(\cdot)$	$C(\cdot)$
Gaussian	$\frac{Y_r}{\sigma_r}$	$-\frac{Y_r^2}{2\sigma_r^2}$
Multinomial	$\mathbb{I}[\ Z_r = j\ ]$	0

The joint density in Theorem 2 is expressed in terms of the parameters $\{\theta_r, \theta_{rt}\}$ which correspond to the node-wise and cross-node potentials. Extending this, let α denote the Gaussian potentials, θ to multinomial, and λ to the cross-type potentials. Define the aggregate set of parameters as $\boldsymbol{\Theta} = \{\theta, \alpha, \lambda\}$. This can be represented in matrix form and is visualized below. Furthermore, the identity functions for the multinomial sufficient statistics are written as $\mathcal{I}_{r'}^{j} = \mathbb{I}[\ Z_{r'} = j\ ]$ and $\mathcal{I}_{(r',t')}^{jk} = \mathbb{I}[\ Z_{r'} = j\ ,\ Z_{t'} = k\ ]$. The subscripts $r \in V_Y$ and $r' \in V_Z$ enumerate nodes in V_Y and V_Z respectively. Finally, the superscripts make clear which multinomial sufficient statistic is being scaled. All together the Gaussian-Potts joint density is proportional to

$$\begin{aligned} P(Y,Z) \propto \exp\Bigg\{ & \sum_{\substack{r' \in V_Z \\ j \in \mathcal{Z}_{r'}}} \theta_{r'}^{j}\, \mathcal{I}_{r'}^{j} + \sum_{\substack{(r',t') \in E_Z \\ j,\, k \in\ \mathcal{Z}_{r'},\ \mathcal{Z}_{t'}}} \theta_{r't'}^{jk}\, \mathcal{I}_{(r',t')}^{jk} \\ & + \sum_{\substack{(r,t') \in E_{YZ} \\ j \in \mathcal{Z}_{t'}}} \lambda_{rt'}^{j} Y_r\, \mathcal{I}_{t'}^{j} + \sum_{r \in V_Y} \alpha_r \frac{Y_r}{\sigma_r} \\ & + \sum_{(r,t) \in E_Y} \frac{\alpha_{rt}}{\sigma_r \sigma_t} Y_r Y_t - \sum_{r \in V_Y} \frac{Y_r^2}{2\sigma_r^2} \Bigg\} \end{aligned}$$

It is straight forward to construct any pairwise mixed MRFs given two distributions in $\mathcal{F}$. Note that the existence of any constructed mixed MRF is directly related to the normalizability of $P(\cdot)$ given parameter set $\boldsymbol{\Theta}$. (Yang et al. 2014) provides such conditions for the Gaussian-Ising, Gaussian-Poisson, and the Poisson-Ising models. Further research is required to establish the constraints beyond these three. Some other models to consider are the Gaussian-Beta, Exponential-Binomial, Poisson-Gamma, etc. Looking ahead, notice that the node-wise conditional distributions can be found by combining what is known about the form of η_r as seen in Theorem 2 with the conditional distribution stated in Section 2.1.

4 Parameter Estimation

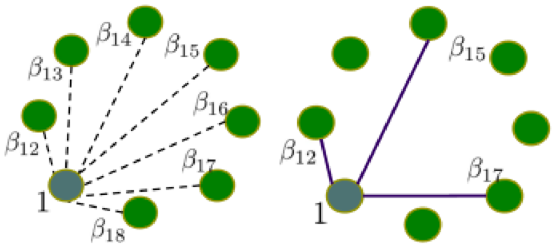
The notation for the Gaussian-Potts graphical model is maintained in this section. I briefly review the recent research devoted to structure learning for homogeneous models before discussing a special parameter estimation scheme (Lee et al. 2014) tailored to the Gaussian-Potts model.

4.1 Learning Homogeneous MRFs

Gaussian graphical models arguably have received the most attention in the literature compared to the other models mentioned thus far. Two main paradigms exist for estimating the paramaters of GGMs. The first can be attributed to (Meinshausen and Bühlmann 2006) where separate node-wise regressions are considered. With Lagrange multiplier ω, the regression for each node is of the form.

$$\tilde{\alpha}_r = \underset{\alpha_r \in \mathbb{R}^{p-1}}{\arg\min} \left(Y_r - \sum_{t \neq r}^{p} \alpha_{rt} Y_t\right)^2 + \omega \sum_{t \neq r}^{p} |\alpha_{rt}|$$

If the distributions for each node Y_r conditioned on the remaining $Y_{V \setminus r}$ are known, then the joint density for the Markov random field is immediately specified (Besag 1974). Within this framework, two nodes (Y_r, Y_t) are said to be conditionally independent whenever the regression coefficient $\alpha_{rt} = 0$. Since all of the nodes are regressed on one another for every pair of nodes $r, t \in V$ there are two parameters α_{rt} and α_{tr} coming from the two corresponding conditional density functions. The pair of parameters can be such that $\alpha_{rt} \neq \alpha_{tr}$ because the spans of $Y_{V \setminus r}$ and $Y_{V \setminus t}$ are not guaranteed to be identical. One approach to reconcile this is to let $\alpha^*_{rt} = \max\{\alpha_{rt}, \alpha_{tr}\}$. In doing so, a symmetric matrix of parameters that encodes the conditional independence structure of Y is created. A single node-neighborhood regression is visualized below to build intuition.



Credit: CMU Lecture notes

The multivariate Gaussian distribution is uniquely characterized by the property that it's inverse covariance matrix $\Omega = \Sigma^{-1}$ completely encodes the conditional independencies of the random variables. Namely, for a pair of nodes $r, t \in V$ it is such that $Y_r \perp\!\!\!\perp Y_t \,|\, Y_{r,t} \iff \Omega_{rt} = 0$. With moderate effort it can be seen that the entries of Ω are linear functions of the node-wise regression coefficients. Whenever $\alpha_{rt} = 0$ we can claim $\Omega_{rt} = 0$ implying conditional independence. Consequently, the graph structure of a Gaussian graphical model is completely encoded in the precision matrix. The Graphical Lasso (Friedman et al. 2008) takes advantage of this equivalent representation by estimating Ω via the regularized negative log-likelihood $\ell(\Omega) + \lambda\|\Omega\|$. Contrasting the separate node-wise regression scheme of (Meinshausen and Bühlmann 2006), the graphical lasso has p^2 parameters to estimate instead of $2(p-1)^2$ which makes it computationally efficient. Parameter estimation for graphical models is all about finding a balance between accuracy and computational complexity.

Regularized node-neighborhood regression is the standard approach in the case of Ising models (Ravikumar et al. 2010). Logistic regressions are performed as the marginal distributions in the Ising model are Bernoulli. Note that Poisson regression would be performed for a Poisson MRF etc. It is safe to deduce that parameter estimation of homogeneous MRFs is done through generalized linear models. In (Yang et al. 2012), they take note of (Meinshausen and Bühlmann 2006) and recover the graph structure through the ℓ_1 regularized conditional log-likelihood.

Multinomial MRFs are somewhat trickier. In line with the notation in the preceding section recall that the sub-graph $(V_Z, E_Z) \subset G$ corresponds to the Potts model and is parameterized by θ. For example, let node Z_1 takes on values $\{1, 2, 3\}$ and Z_2 take values in {a,b,c}. You may intuitively, and correctly, guess that parameter estimation in this sub-graph is done by multi-class logistic regression. Unlike the Gaussian graphical model in which a single parameter α_{rt} determines conditional independence, in this particular example $|\mathcal{Z}_1 \times \mathcal{Z}_2| = 9$ parameters are needed. Multiclass logistic regressions could be performed in an attempt to recover the graph structure but this would be naive. All nine $\{\theta_{11}, ..., \theta_{33}\}$ must equal 0 as a *group* in order for Z_1 and Z_2 to be conditionally independent. In other words there is additional structure in the parameter space that needs to taken into account. One approach is to use a group penalized lasso (Jalali et al. 2011). The same problem arises in the Gaussian-Potts model (Lee et al. 2014) and is the subject of interest in the following sections.

4.2 Learning Heterogeneous MRFs

Let $X = (Y, X)$ be a $p+q$-dimensional random vector with p Gaussian and q multinomial random variables with corresponding graph $G = (V, E)$. We know it is possible to estimate the conditional independence structure of an exponential family MRF via node-wise conditional distributions. Simply plug in the relevant sufficient statistics and base measures into the conditional distribution as prescribed by the exponential family and perform $p + q$ separate node wise regressions. Per (Yang et al. 2012, 2014) this holds true for both homogeneous and heterogeneous cases. With the Gaussian-Potts model in mind see that the distribution of the Gaussian nodes conditioned on the multinomial nodes $P(Y|Z)$ takes the form

$$\exp\left\{\sum_{r\in V}\theta_r B_Y(Y_r) + \sum_{(r,t)\in E_Y}\theta_{rt}B_Y(Y_r)B_Y(Y_t) + \sum_{(r,t')\in E_{YZ}}\theta_{rt'}B_Y(Y_r)B_Z(Z_{t'}) + \sum_{r\in V_Y}C_Y(Y_r)\right\}$$

where B_Y and B_Z are the relevant sufficient statistics.

I now make precise the conditions that characterize the conditional independence structure for this particular mixed MRF. Recall that the Gaussian-Potts model has parameters $\boldsymbol{\Theta} = \{\theta, \alpha, \lambda\}$. A pair of Gaussian random variables (Y_r, Y_t) are said to be conditionally independent whenever

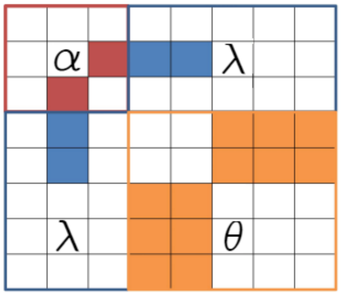
$$\alpha_{rt} = 0 \iff Y_r \perp\!\!\!\perp Y_t \,|\, (Y_{V\setminus\{r,t\}}, Z)$$

A pair of multinomial random variables $(Z_{r'}, Z_{t'})$ are conditionally independent when the *collection* of parameters that enumerate all possible combination of states are all zero. See that this translates to a zero-valued sub-matrix in $\boldsymbol{\Theta}$. Symbolically, $\forall j \in \mathcal{Z}_{r'}$ and $\forall k \in \mathcal{Z}_{t'}$

$$\theta^{jk}_{r't'} = 0 \iff Z_{r'} \perp\!\!\!\perp Z_{t'} \,|\, (Z_{V\setminus\{r',t\}}, Y)$$

Finally for the cross-type interactions, $(Y_r, Z_{t'})$ are conditionally independent whenever there is a corresponding zero-valued sub-vector in $\boldsymbol{\Theta}$. Like the previous case, $\forall j \in \mathcal{Z}_{r'}$

$$\lambda^{j}_{rt'} = 0 \iff Y_r \perp\!\!\!\perp Z_{t'} \,|\, (Y_{V\setminus r}, Z_{V\setminus t'})$$



Symmetric matrix represents the parameters Θ of the model.

Credit: Lee et al. 2014

4.3 Pseudolikelihood & Group Lasso

In light of the node-neighborhood scheme and the graphical lasso, there is actually a third framework for learning the structure of an exponential family MRF. Note that the success of the graphical lasso is reliant on the nice properties of the multivariate Gaussian. Maximum likelihood estimation for complicated distributions such as the Gaussian-Potts model is comparatively harder because of the computational demands that come with computing the high-dimensional integral in the log-partition function $A(\theta)$. The pseudolikelihood (Besag 1975) is an alternative approach that bypasses this difficulty at the cost of estimate accuracy.

$$\hat{\ell}(\boldsymbol{\Theta}) = -\sum_{r}^{p} \log P(Y_r \,|\, Y_{V\setminus r},\, Z\,; \boldsymbol{\Theta}) - \sum_{r'}^{q} \log P(Z_{r'} \,|\, Z_{V\setminus r'},\, Y\,; \boldsymbol{\Theta})$$

Quoting (Lee et al. 2014), the pseudolikelihood is a consistent estimator of the maximum likelihood and can be interpreted as node-wise regressions that enforce symmetry. Compare this to the asymmetry that occurs in the Meinshausen and Bühlmann method. Additionally, the lost accuracy is acceptable because the zeros in $\boldsymbol{\Theta}$ are a priority as they reveal which random variables are *not* conditionally dependent. Sparsity induced by regularization makes this quite easy and further motivates the pursuit of zeros.

It is of interest to note that work on structure learning for the Gaussian-Potts model (Lee et al. 2014) was published prior to the formalization of mixed Markov random fields (Yang et al. 2014). The trio of papers from Yang and Lee collectively are actually self referential so it is safe to assume there was some degree of communication between them.

The estimator combines all of the topics discussed thus far. First, via Hammersley-Clifford and (Besag 1974), we know that a valid MRF can be constructed by specifying the node-wise conditional distributions in $\mathcal{F}$. Then through Yang, the aforementioned conditionals are no longer restricted to a single named distribution. Next the pseudolikelihood (Besag 1975) provides a consistent and computationally efficient estimate of the conditional independence structure. Finally, the the necessity of the group lasso to account for additional structure in $\boldsymbol{\Theta}$ (Jalali et al. 2011) makes the following estimator of (Lee et al. 2014) seem entirely appropriate

$$\min_{\boldsymbol{\Theta}} \ell_\omega(\boldsymbol{\Theta}) = \ell(\boldsymbol{\Theta}) + \omega\left(\sum_{r=1}^{p}\sum_{t=1}^{r-1}|\alpha_{rt}| + \sum_{r=1}^{p}\sum_{t'=1}^{q}\|\lambda_{rt'}\|_2 + \sum_{r'=1}^{q}\sum_{t'=1}^{r'-1}\|\theta_{r't'}\|_F\right)$$

See that each scalar, vector, and sub-matrix corresponding the various potential parameters in $\boldsymbol{\Theta}$ are forced to zero as groups. They went further still and proposed a weighting scheme to penalize each group as a function of size which was validated on simulated data.

References

J. Besag. Statistical analysis of non-lattice data. *The statistician*, 1975.

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

E. Yang, P. Ravikumar, G.I. Allen, and Z. Liu. On graphical models via univariate exponential family distributions, 2014.

E. Yang, Y. Baker, P. Ravikumar, G.I. Allen, and Z. Liu. Mixed Graphical Models via Exponential Families. *arXiv preprint arXiv:1301.4183*, 2013.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 2008.

N. Meinshausen and P. B¨uhlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 2006.

P. Ravikumar, M.J. Wainwright, and J.D. Lafferty. High-dimensional ising model selection using l1-regularized logistic regression. *The Annals of Statistics*, 2010.

A. Jalali, P. Ravikumar, V. Vasuki, S. Sanghavi, UT ECE, and UT CS. On learning discrete graphical models using group-sparse regularization. In Proceedings of the *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

M. Wainwright and M. Jordan. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends in Machine Learning*, 2008.