Journal of Machine Learning Research 21 (2020) 1-21 Submitted 5/19; Published 12/20

Nonparametric graphical model for counts

Arkaprava Roy ARK007@UFL.EDU
Department of Biostatistics
University of Florida
Gainesville, FL 32611, USA

David B Dunson DUNSON@DUKE.EDU
Department of Statistics
Duke University
Durham, NC 27708-0251, USA

Editor: Robert McCulloch

Abstract

Although multivariate count data are routinely collected in many application areas, there is surprisingly little work developing flexible models for characterizing their dependence structure. This is particularly true when interest focuses on inferring the conditional independence graph. In this article, we propose a new class of pairwise Markov random field-type models for the joint distribution of a multivariate count vector. By employing a novel type of transformation, we avoid restricting to non-negative dependence structures or inducing other restrictions through truncations. Taking a Bayesian approach to inference, we choose a Dirichlet process prior for the distribution of a random effect to induce great flexibility in the specification. An efficient Markov chain Monte Carlo (MCMC) algorithm is developed for posterior computation. We prove various theoretical properties, including posterior consistency, and show that our COunt Nonparametric Graphical Analysis (CONGA) approach has good performance relative to competitors in simulation studies. The methods are motivated by an application to neuron spike count data in mice.

Keywords: Conditional independence, Dirichlet process, Graphical model, Markov random field, Multivariate count data

1. Introduction

Graphical models provide an appealing framework to characterize dependence in multivariate data $X_i = (X_{i1}, \ldots, X_{iP})$ in an intuitive way. This article focuses on undirected graphical models or Markov random fields (MRFs). In this approach, each random variable is assigned as a node of a graph G which is characterized by the pair (V, E). Here V and E denote the set of nodes and set of connected edges of the graph G, with $V = \{1, \ldots, P\}$ and $E \subseteq V \times V$. The graph G encodes conditional independence relationships in the data. We say X_l and X_k are conditionally independent if $P(X_l, X_k | X_{-(l,k)}) = P(X_l | X_{-(l,k)}) P(X_k | X_{-(l,k)})$, with $X_{-(l,k)}$ denoting all random variables excluding X_l and X_k. Conditional independence between two random variables is equivalent to the absence of an edge between those two corresponding nodes in the graph. Thus the conditional independence of X_l and X_k would imply that the edge (k, l) is not present i.e. $(k, l) \notin E$.

Although there is a rich literature on graphical models, most of the focus has been specifically on Gaussian graphical models. For bounded discrete data, Ising (Ravikumar et al., 2010; Kolar et al., 2010) and multinomial graphical models (Jalali et al., 2011) have been studied. However, for unbounded count-valued data, the existing literature is limited. Multivariate count data are routinely collected in genomics, sports, imaging analysis, and text mining among many other areas, but most of the focus has been on latent factor and covariance structure models (Wedel et al., 2003; Zhou et al., 2012). The goal of this article is to address this gap and provide a flexible framework for statistical inference in count graphical models.

Besag first introduced pair-wise graphical models, deemed 'auto-models' in his seminal paper on MRFs (Besag, 1974). To define a joint distribution on a spatial lattice, he started with an exponential family representation of the marginal distributions and then added first-order interaction terms. In the special case of count data, he introduced the Poisson auto-model. In this approach, the random variable at the *i-th* location X_i follows a conditional Poisson distribution with mean μ_i, dependent on the neighboring sites. Then μ_i is given the form $\mu_i = \exp(\alpha_i + \sum_j \beta_{ij} X_j)$. It can be shown that this conditional density model admits a joint density if and only if $\beta_{ij} \leq 0$ for all pairs of (i, j). Hence, only non-negative dependence can be accommodated. Gamma and exponential auto-models also have the same restriction due to the non-negativity of the random variables.

Yang et al. (2013) truncated the count support within the Poisson auto-model to allow both positive and negative dependence, effectively treating the data as ordered categorical. Allen and Liu (2012) fit the Poisson graphical model locally in a manner that allows both positive and negative dependence, but this approach does not address the problem of global inference on G. Chiquet et al. (2018) let $X_{ij} \sim \text{Poi}(\exp(\mu_j + Z_{ij}))$ for $1 \leq i \leq n, 1 \leq j \leq V$ and $Z_i \sim \text{MVN}(0, \Sigma)$. The graph is inferred through sparse estimation of Σ^{-1}. Hadiji et al. (2015) proposed a non-parametric count model, with the conditional mean of each node an unknown function of the other nodes. Yang et al. (2015) defined a pairwise graphical model for count data that only allows negative dependence. Inouye et al. (2016a,b, 2017) models multivariate count data under the assumption that the square root or more generally the *j-th* root, of the data is in an exponential family. This model allows for positive and negative dependence but under strong distributional assumptions.

In the literature on spatial data analysis, many count-valued spatial processes have been proposed, but much of the focus has been on including spatial random effects instead of an explicit graphical structure. De Oliveira (2013) considered a random field on the mean function of a Poisson model to incorporate spatial dependence. However, conditional independence or dependence structure in the mean does not necessarily represent that of the data. The Poisson-Log normal distribution, introduced by Aitchison and Ho (1989), is popular for analyzing spatial count data (Chan and Ledolter, 1995; Diggle et al., 1998; Chib and Winkelmann, 2001; Hay and Pettitt, 2001). Here also the graph structure of the mean does not necessarily represent that of the given data. Hence, these models cannot be regarded as graphical models for count data. To study areal data, conditional autoregressive models (CAR) have been proposed (Gelfand and Vounatsou, 2003; De Oliveira, 2012; Wang and Kockelman, 2013). Although these models have an MRF-type structure, they assume the graph G is known based on the spatial adjacency structure, while our focus is on inferring unknown G.

Gaussian copula models are popular for multivariate non-normal data (Xue-Kun Song, 2000; Murray et al., 2013). Mohammadi et al. (2017) developed a computational algorithm to build graphical models based on Gaussian copulas using methods developed by Dobra et al. (2011). However, it is difficult to model multivariate counts with zero-inflated or multimodal marginals using a Gaussian copula.

Within a semiparametric framework, Liu et al. (2009) proposed a nonparanormal graphical model in which an unknown monotone function of the observed data follows a multivariate normal model with unknown mean and precision matrix subject to identifiability restrictions. This model has been popular for continuous data, providing a type of Gaussian copula. For discrete data, the model is not directly appropriate, as mapping discrete to continuous data is problematic. To the best of our knowledge, there has been no work on nonparanormal graphical models for counts. In general, conditional independence cannot be ensured if the function of the random variable is not continuous. For example if f is not monotone continuous, then conditional independence of X and Y does not ensure conditional independence of $f(X)$ and $f(Y)$.

In addition to proposing a flexible graphical model for counts, we aim to develop efficient Bayesian computation algorithms. Bayesian computation for Gaussian graphical models (GGMs) is somewhat well-developed (Dobra and Lenkoski, 2011; Wang, 2012, 2015; Mohammadi et al., 2015). Unfortunately, outside of GGMs, the likelihood-based inference is often problematic due to intractable normalizing constants. For example, the normalizing constant in the Ising model is too expensive to compute except for very small P. There are approaches related to surrogate likelihood (Kolar and Xing, 2008) by relaxation of the log-partition function (Banerjee et al., 2008). Kolar et al. (2010) use conditional likelihood. Besag (1975) chose a product of conditional likelihoods as a pseudo-likelihood to estimate MRFs. For exponential family random graphs, Van Duijn et al. (2009) compared maximum likelihood and maximum pseudo-likelihood estimates in terms of bias, standard errors, coverage, and efficiency. Zhou and Schmidler (2009) numerically compared the estimates from a pseudo-posterior with exact likelihood-based estimates and found they are very similar in small samples for Ising and Potts models. Also for pseudo-likelihood based methods asymptotic unbiasedness and consistency have been studied (Comets, 1992; Jensen and Künsch, 1994; Mase, 2000; Baddeley and Turner, 2000). Pensar et al. (2017) showed consistency of marginal pseudo-likelihood for discrete-valued MRFs in a Bayesian framework.

Recently Dobra et al. (2018) used pseudo-likelihood for estimation of their Gaussian copula graphical model. Although pseudo-likelihood is popular in the frequentist domain for count data (Inouye et al., 2014; Ravikumar et al., 2010; Yang et al., 2013), its usage is still the nonstandard in Bayesian estimation for count MRFs. This is mainly because calculating conditional densities is expensive for count data due to unbounded support, making posterior computations hard to conduct. We implement an efficient Markov Chain Monte Carlo (MCMC) sampler for our model using pseudo-likelihood and pseudo-posterior formulations. Our approach relies on a provably accurate approximation to the normalizing constant in the conditional likelihood. We also provide a bound for the approximation error due to the evaluation of the normalizing constant numerically.

In Section 2, we introduce our novel graphical model. In Section 3, some desirable theoretical results are presented. Then we discuss computational strategies in Section 4

and present simulation results in Section 5. We apply our method to neuron spike data in mice in Section 6. We end with some concluding remarks in Section 7.

2. Modeling

Before introducing the model, we define some of the Markov properties related to the conditional independence of an undirected graph. A clique of a graph is the set of nodes where every two distinct nodes are adjacent; that is, connected by an edge. Let us define $\mathcal{N}(j) = \{l : (j,l) \in E\}$. For three disjoint sets A, B and C of V, A is said to be separated from B by C if every path from A to B goes through C. A path is an ordered sequence of nodes $i_0, i_1, \ldots, i_m$ such that $(i_{k-1}, i_k) \in E$. The joint distribution is locally Markov if $X_j \perp V \setminus \{X_j, N(j)\}|N(j)$. If for three disjoint sets A, B and C of V, X_A and X_B are independent given X_C whenever A and B are separated by C, the distribution is called globally Markov. The joint density is pair-wise Markov if for any $i, j \in V$ such that $(i,j) \notin E$, X_i and X_j are conditionally independent.

We consider here a pair-wise MRF (Wainwright et al., 2007; Chen et al., 2014) which implies the following joint probability mass function (pmf) for the P dimensional random variable X,

$$\Pr(X_1, \ldots, X_P) \propto \exp\left\{ \sum_{i=1}^{P} f(X_i) + \sum_{l=2}^{P} \sum_{j<l} f(X_j, X_l) \right\}, \tag{1}$$

where $f(X_i)$ is called a node potential function, $f(X_j, X_l)$ an edge potential function and we have $f(X_j, X_l) = 0$ if there is no edge (j,l). Thus this distribution is pair-wise Markov by construction. Then (1) satisfies the Hammersley-Clifford theorem (Hammersley and Clifford, 1971), which states that a probability distribution having a strictly positive density satisfies a Markov property with respect to the undirected graph G if and only if its density can be factorized over the cliques of the graph. Since our pair-wise MRF is pair-wise Markov, we can represent the joint probability mass function as a product of mass functions of the cliques of graph G. The existence of such a factorization implies that this distribution has both global and local Markov properties.

Completing a specification of the MRF in (1) requires an explicit choice of the potential functions $f(X_j)$ and $f(X_j, X_l)$. In the Gaussian case, one lets $f(X_j) = -\alpha_j X_j^2$ and $f(X_j, X_l) = -\beta_{jl} X_j X_l$, where α_j and β_{jl} correspond to the diagonal and off-diagonal elements of the precision matrix $\Sigma^{-1} = \mathrm{cov}(X)^{-1}$. In general, the node potential functions can be chosen to target specific univariate marginal densities. If the marginal distribution is Poisson, the appropriate node potential function is $f(X_j) = \alpha_j X_j - \log(X_j!)$. One can then choose the edge potential functions to avoid overly restrictive constraints on the dependence structure, such as only allowing non-negative correlations. Yang et al. (2013) identify edge potential functions with these properties for count data by truncating the support; for example, to the range observed in the sample. This reduces the ability to generalize results, and in practice, estimates are sensitive to the truncation level. We propose an alternative construction of the edge potentials that avoids truncation.

2.1 Model

We propose the following modified pmf for P-dimensional count-valued data X,

$$\Pr(X_1, \ldots, X_P) \propto \exp\Big(\sum_{j=1}^{P} [\alpha_j X_j - \log(X_j!)] - \sum_{l=2}^{P} \sum_{j<l} \beta_{jl} F(X_j) F(X_l) \Big),$$

where $F(\cdot)$ is a monotone increasing bounded function with support $[0, \infty)$, $f(X_j) = \alpha_j X_j - \log(X_j!)$ and $f(X_j, X_l) = -\beta_{jl} F(X_j) F(X_l)$ using the notation of (1).

Lemma 1 *Let $F(\cdot)$ be uniformly bounded by U, then the normalizing constant, say $A(\alpha, \beta)$, can be bounded as,*

$$\exp\Big(\sum_{j=1}^{P} \exp(\alpha_j) - U^2 \sum_{l=2}^{P} \sum_{j<l} |\beta_{jl}| \Big) \leq A(\alpha, \beta) \leq \exp\Big(\sum_{j=1}^{P} \exp(\alpha_j) + U^2 \sum_{l=2}^{P} \sum_{j<l} |\beta_{jl}| \Big).$$

These bounds can be obtained by elementary calculations. The constant $A(\alpha, \beta)$ is the sum of the above pmf over the support of X. The sum reduces to a product of P many exponential series sums after replacing the function $F(\cdot)$ by its maximum.

Thus by modifying the edge potential function in this way using a bounded function of X, we can allow unrestricted support for all the parameters, allowing one to estimate both positive and negative dependence. Under the monotonicity restrictions on $F(\cdot)$, inference on the conditional independence structure tends to be robust to the specific form chosen. We let $F(\cdot) = (\tan^{-1}(\cdot))^\theta$ for some positive $\theta \in \mathbb{R}^+$ to define a flexible class of monotone increasing bounded functions. The exponent θ provides additional flexibility, including impacting the range of $F(X)$, $\big(0, (\frac{\pi}{2})^\theta\big)$. The parameter θ can be estimated along with the other parameters, including the baseline parameters α controlling the marginal count distributions and the coefficients β_{jl} controlling the graphical dependence structure. For simplicity and interpretability, we propose to estimate θ to minimize the difference in covariance between $F(X)$ and X. Figure 1 illustrates how θ controls the range and shape of $F(\cdot)$. Figure 2 shows how the difference between covariances of $F(X)$ and X vary for different values of θ in sparse and non-sparse data cases. In both cases, the difference function has a unique minimizer. Although the same strategy could be used to tune the truncation parameter in the Yang et al. (2013) approach, issues arise in estimating the support of the data based on a finite sample, as new data may fall outside of the estimated support. Besides, their approach is less flexible in relying on parametric assumptions, while we use a mixture model for the αs to induce a nonparametric structure.

Letting X_t denote the t^{th} independent realization of X, for $t = 1, \ldots, n$, the pmf is

$$\Pr(X_{t1}, \ldots, X_{tP}) \propto \exp\Big(\sum_{j=1}^{P} [\alpha_{tj} X_{tj} - \log(X_{tj}!)] - \sum_{l=2}^{P} \sum_{j<l} \beta_{jl} (\tan^{-1}(X_{tj}))^\theta (\tan^{-1}(X_{tl}))^\theta \Big), \quad (2)$$

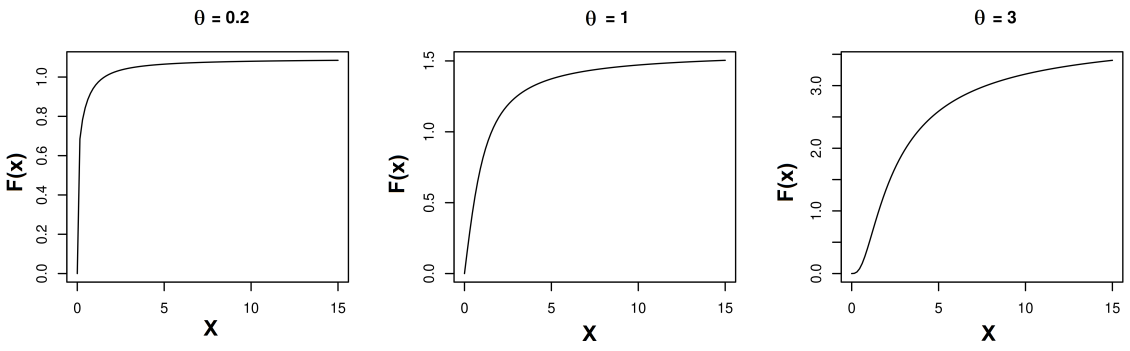

Figure 1: $F(\cdot) = (\tan^{-1})^{\theta}(\cdot)$ for different values of θ. The parameter θ controls both shape and range of $F(\cdot)$.

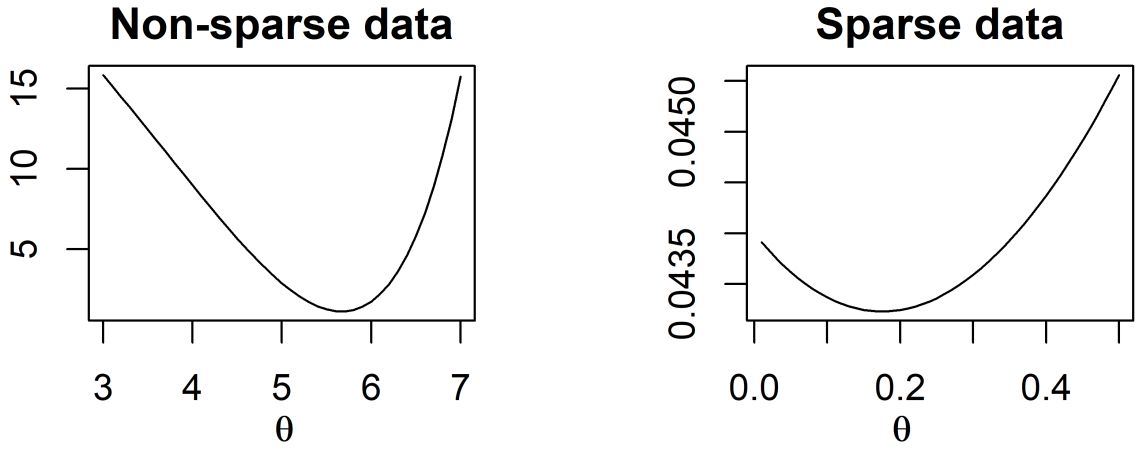

Figure 2: $\|cov(\tan^{-1}(X)^{\theta}) - cov(X)\|_F$ for different values of θ. $\|\|_F$ stands for the Frobenius norm.

where α_{tj}'s are coefficients of different node potential functions and β_{jl}'s are coefficients of the edge potential functions as before. We vary α_{tj} with t to allow more flexibility in modeling the marginal densities. If $\beta_{jl} = 0$, then X_{tj} and X_{tl} are conditionally independent for all t. We call our proposed method COunt Nonparametric Graphical Analysis (CONGA).

Now we reparametrize (2) using $\log(\lambda_{tj}) = \alpha_{tj}$ and rewrite the model as,

$$\Pr(X_{t1}, \ldots, X_{tP}) \propto \prod_{j=1}^{P} \frac{\lambda_{tj}^{X_{tj}}}{X_{tj}!} \exp\Big(- \sum_{l=2}^{P} \sum_{j<l} \beta_{jl} (\tan^{-1}(X_{tj}))^{\theta} (\tan^{-1}(X_{tl}))^{\theta} \Big). \qquad (3)$$

This reparametrizated model is more intuitive to understand. Due to the Poisson type marginal in (3), this model is suitable for data with over-dispersed marginals with respect to the Poisson at each node. Over-dispersion is typical in broad applications. We consider this reparametrized model in the rest of the paper.

2.2 Prior structure

To proceed with Bayesian computation, we put priors on the parameters. We have two sets of parameters in (3), β and λ. For the β_{jl} parameters, we choose simple iid Gaussian priors. It is straightforward to consider more elaborate shrinkage or variable selection priors for the β_{jl}'s, but we find usual Gaussian priors have good performance in small to moderate-dimensional applications

The parameter λ_{tj}'s represent random effects; these parameters are not individually identifiable and are given random effects distributions $\lambda_{tj} \sim D_j$. The distribution D_j controls over-dispersion and the shape of the marginal count distribution for the j^{th} node. To allow these marginals to be flexibly determined by the data, we take a Bayesian nonparametric approach using Dirichlet process priors $D_j \sim \text{DP}(M_j D_0)$, with D_0 a Gamma base measure and M_j a precision parameter, having $M_j \sim \text{Ga}(c, d)$ for increased data adaptivity.

3. Theoretical properties

We explore some of the theoretical properties of our proposed CONGA method.

Theorem 2 *If we have $\beta_{jl} = 0$, then X_{tj} and X_{tl} are conditionally independent for all t under* (3).

This result is easy to verify by simply calculating the conditional probabilities. The details of the proof are in the Appendix.

We study posterior consistency under a fixed P and increasing n regime, assuming the prior of Section 2.2 with prespecified θ. Let G_j be the density on α_{tj}, induced by $\lambda_{tj} \sim D_j$. Let the parameter space for G_j be $\mathcal{G}_j$ and that for β be $\mathbb{R}^q$, where $q = P(P-1)/2$. Thus the complete parameter space for $\kappa = \{\beta, G_1, \ldots, G_p\}$ is $\Psi = \mathbb{R}^q \times \mathcal{G}_1 \times \cdots \times \mathcal{G}_P$. We consider the prior $\tilde{\Gamma}_j$ on G_j and χ on β.

Let κ^0 be the truth for κ. We make the following assumptions.

Assumptions

1. For some large $T > 0$, let $\mathcal{G}\dashv = \{G : G([-T, T]) = 1\}$. Then $G_j^0 \in \mathcal{G}$ and G_j^0 is in the support of $\tilde{\Gamma}_j$.

2. For some large $C > 0$, let $\mathcal{Q} = \{\beta : \|\beta\|_\infty < C\}$, where $\|\cdot\|_\infty$ stands for the infinity norm. Then $\beta^0 \in \mathcal{Q}$ and β^0 is in the support of χ.

3. $E(X_{tj}) < \infty$ for all pairs of (t, j)

Theorem 3 *Under the assumptions, 1-3, the posterior for κ is consistent at κ^0.*

We show that the truth belongs to the Kullback-Leibler support of the prior. Thus the posterior probability of any neighborhood around the true p.m.f converges to one in $P_{\kappa_0}^{(n)}$-probability as n goes to ∞ as a consequence of Schwartz (1965). Here $P_\kappa^{(n)}$ is the distribution of a sample of n observations with parameter κ. Hence, the posterior is weakly consistent. The posterior is said to be strongly consistent if the posterior probability of any neighborhood around the true p.m.f convergences to one almost-surely. Support of the data is a countable space. The weak and strong topologies on countable spaces are equivalent by Scheffe's theorem. In particular, weak topology and total variation topology are equivalent for discrete data. Weak consistency implies strong consistency. Thus the posterior for κ is also strongly consistent at κ^0. A detailed proof is in the Appendix.

Instead of assuming bounded support on the true distribution of random effects, one can also assume it to have sub-Gaussian tails. The posterior consistency result still holds with minor modifications in the current proof. Establishing graph selection consistency of the proposed method is an interesting area of future research when p is growing with n and λ_{tj}'s are fixed effects. Since we are interested in a non-parametric graphical model, we do not explore that in this paper.

4. Computation

As motivated in Section 2.1, we estimate θ to minimize the differences in the sample covariance of X and $F(X)$. In particular, the criteria is to minimize $\|cov(\tan^{-1}(X)^\theta) - cov(X)\|_F$. This is a simple one dimensional optimization problem, which is easily solved numerically.

To update the other parameters, we use an MCMC algorithm, building on the approach of Roy et al. (2018). We generate proposals for Metropolis-Hastings (MH) using a Gibbs sampler derived under an approximated model. To avoid calculation of the global normalizing constant in the complete likelihood, we consider a pseudo-likelihood corresponding to a product of conditional likelihoods at each node. This requires calculations of P local normalizing constants which is computationally tractable.

The conditional likelihood at the *j-th* node is,

$$P(X_{tj}|X_{t,-j}) = \frac{\exp\left[\{\log(\lambda_{tj})X_{tj} - \log(X_{tj}!)\} - \sum_{j\neq l}\beta_{jl}\{\tan^{-1}(X_{tj})\}^\theta\{\tan^{-1}(X_{tl})\}^\theta\right]}{\sum_{X_{tj}=0}^{\infty}\exp\left[\{\log(\lambda_{tj})X_{tj} - \log(X_{tj}!)\} - \sum_{j\neq l}\beta_{jl}\{\tan^{-1}(X_{tj})\}^\theta\{\tan^{-1}(X_{tl})\}^\theta\right]} \quad (4)$$

The normalizing constant is

$$\sum_{X_{tj}=0}^{\infty}\exp\left[\{\log(\lambda_{tj})X_{tj} - \log(X_{tj}!)\} - \sum_{j\neq l}\beta_{jl}\{\tan^{-1}(X_{tj})\}^\theta\{\tan^{-1}(X_{tl})\}^\theta\right].$$

We truncate this sum at a sufficiently large value B for the purpose of evaluating the conditional likelihood. The error in this approximation can be bounded by

$$\exp(\lambda_{tj})(1 - CP(B+1, \lambda_{tj})) \exp\left\{ - \sum_{j \neq l : \beta_{jl} < 0} \beta_{jl} (\pi/2)^{\theta} (\tan^{-1}(X_{tl}))^{\theta} \right\},$$

where $CP(x, l)$ is the cumulative distribution function of the Poisson distribution with mean l evaluated at x. The above bound can in turn be bounded by a similar expression with $(\tan^{-1}(X_{tl}))^{\theta}$ replaced by $(\pi/2)^{\theta}$. One can tune B based on the resulting bound on the approximation error. In our simulation setting, even $B = 70$ makes the above bound numerically zero. We use $B = 100$ as a default choice for all of our computations.

We update $\lambda_{\cdot j}$ using the MCMC sampling scheme described in Chapter 5 of Ghosal and Van der Vaart (2017) for the Dirichlet process mixture prior of λ_{ij} based on the above conditional likelihood. For clarity this algorithm is described below:

(i) Calculate the probability vector Q_j for each j such that $Q_j(k) = \text{Pois}(X_{ij}, \lambda_{kj})$ and $Q_j(i) = M_j \text{Ga}(\lambda_{i,j}, a + X_{i,j}, b + 1)$.

(ii) Sample an index l from $1 : T$ with probability $Q_j / \sum_k Q_j(k)$.

(iii) If $l \neq i$, $\lambda_{ij} = \lambda_{lj}$. Otherwise sample a new value as described below.

(iv) M_j is sampled from $\text{Gamma}(c + U, d - \log(\delta))$, where U is the number of unique elements in $\lambda_{\cdot j}$, δ is sampled from $\text{Beta}(M_j, T)$, and $M_j \sim \text{Ga}(c, d)$ a priori.

When we have to generate a new value for λ_{tj} in step (iii), we consider the following scheme.

(i) Generate a candidate λ^c_{tj} from $\text{Gamma}(a + X_{tj}, b + 1)$.

(ii) Adjust the update $\lambda^c_{tj} = \lambda^0_{tj} + K_1(\lambda^c_{tj} - \lambda^0_{tj})$, where λ^0_{tj} is the current value and $K_1 < 1$ is a tuning parameter, adjusted with respect to the acceptance rate of the resulting Metropolis-Hastings (MH) step.

(iii) We use the pseudo-likelihood based on the conditional likelihoods in (4) to calculate the MH acceptance probability.

To generate β, we consider a new likelihood that the standardized $(\tan^{-1}(X_{tl}))^{\theta}$ follows a multivariate Gaussian distribution with precision matrix Ω such that $\Omega_{pq} = \Omega_{qp} = \beta_{pq}$ with $p < q$ and $\Omega_{pp} = (Var((\tan^{-1}(X_{tl}))^{\theta})^{-1})_{pp}$. Thus diagonal entries do not change over iterations. We update $\Omega_{l,-l} = \{\Omega_{l,i} : i \neq l\}$ successively. We also define $\Omega_{-l,-l}$ as the submatrix by removing l-th row and column. Let $s = (F(x) - \bar{F}(X))^T (F(x) - \bar{F}(X))$. Thus s is the $P \times P$ gram matrix of $(\tan^{-1} X)^{\theta}$, standardized over columns.

(i) Generate an update for $\Omega_{l,-l}$ using the posterior distribution as in Wang (2012). Thus a candidate $\Omega^c_{l,-l}$ is generated from $\text{MVN}(-C s_{l,-l}, C)$, where $C = ((s_{22} + \gamma)\Omega^{-1}_{-l,-l} + D_l^{-1})^{-1}$, where D_l is the prior variance corresponding to $\Omega_{l,-l}$

(ii) Adjust the update $\Omega^c_{l,-l} = \Omega^0_{l,-l} + K_2 \frac{(\Omega^c_{l,-l} - \Omega^0_{l,-l})}{\|(\Omega^c_{l,-l} - \Omega^0_{l,-l})\|_2}$, where $\Omega^0_{l,-l}$ is the current value and K_2 is a tuning parameter, adjusted with respect to the acceptance rate of the following MH step. Also K_2 should always be less than $\|(\Omega^c_{l,-l} - \Omega^0_{l,-l})\|_2$.

(iii) Use the pseudo-likelihood based on the conditional likelihoods in (4), multiplying over t to calculate the MH acceptance probability. $\pi(\theta^0|\theta^c) = \tilde{\pi}(\theta_G)$ and $\pi(\theta^c|\theta^0) = \tilde{\pi}(\theta'_G)$, where θ_G is the original Gibbs update.

5. Simulation

We consider four different techniques for generating multivariate count data. One approach is based on a Gaussian copula type setting. The other three are based on competing methods. We compare the methods based on false positive and false negative proportions. We include an edge in the graph between the j^{th} and l^{th} nodes if the 95% credible interval for β_{jl} does not include zero. There is a decision-theoretic proof to justify such an approach in Thulin (2014). We compare our method CONGA with TPGM, SPGM, LPGM, huge, BDgraph, and ssgraph. The first three are available in R package `XMRF` and the last two are in R packages `BDgraph` and `ssgraph` respectively. The function huge is from R package `huge` which fits a nonparanormal graphical model. The last two methods fit graphical models using Gaussian copulas and `ssgraph` uses spike and slab priors in estimating the edges.

To simulate data under the first scheme, we follow the steps given below.

(i) Generate n many multivariate normals of length c from $\text{MVN}(0_c, \Omega^{-1}_{c\times c})$, where 0_c is the vector of zeros of length c. This produces a matrix X of dimension $n \times c$.

(ii) We calculate the matrix $P_{n\times c}$, which is $P_{ij} = \Phi(X_{ij})$, where Φ is the cumulative density function of the standard normal.

(iii) The Poisson random variable $Y_{n\times c}$ is $Y_{ij} = QP(P_{ij}, \lambda)$ for a given mean parameter λ with QP the quantile function of Poisson(λ).

Let $X_{:,l}$ denote the *l-th* column of X. In the above data generation setup, $\Omega_{pq} = 0$ implies that $Y_{:,p}$ and $Y_{:,q}$ are conditionally independent due to Lemma 3 of Liu et al. (2009). The marginals are allowed to be multimodal at some of the nodes, which is not possible under the other simulation schemes.

Apart from the above approach, we also generate the data using R package `XMRF` from the models Sub-Linear Poisson Graphical Model (SPGM), Truncated Poisson graphical Model (TPGM) (Yang et al., 2013), and Local Poisson Graphical Model (LPGM) (Allen and Liu, 2012).

We choose $\nu_3 = 100$, which is the prior variance of the normal prior of β_{jl} for all j, l. The choice $\nu_3 = 100$ makes the prior weakly informative. The parameter γ is chosen to be 5 as given in Wang (2012). For the gamma distribution, we consider $a = b = 1$. For the Dirichlet process mixture, we take $c = d = 10$. We consider $n = 100$ and $P = 10, 30, 50$. We collect 5000 post burn MCMC samples after burning in 5000 MCMC samples.

We compare the methods based on two quantities p_1 and p_2. We define these as p_1 = Proportion of falsely connected edges in the estimated graph (false positive) and p_2 =

Proportion of falsely not connected edges in the estimated graph (false negative). We show the pair (p_1, p_2) in Tables 1 to 3 for number of nodes 10, 30 and 50. All of these results are based on 50 replications. To evaluate the performance of CONGA, we calculate the proportion of replications where zero is included in the corresponding 95% credible region, constructed from the MCMC samples for each replication. For the other methods, the results are based on the default regularization as given in the R package `XMRF`. Our proposed method overwhelmingly outperforms the other methods when the data are generated using a Gaussian copula type setting instead of generating from TPGM, SPGM, or LPGM. For other cases, our method performs similarly to competing methods when the number of nodes is large. In these cases, the competing methods TPGM, SPGM, or LPGM are levering on modeling assumptions that CONGA avoids. CONGA beats BDgraph and ssgraph in almost all the scenarios in terms of false-positive proportions. The false-negative proportions are comparable. The function 'huge' performs similarly to CONGA when the data are generated using TPGM, SPGM, and LPGM. But CONGA is better than all other methods when the data are generated using the Gaussian copula type setting. This is likely because the other cases correspond to simulating data from one of the competitor's models.

Table 1: Performance of the competing methods against our proposed method with 10 nodes. Top row indicates the method used to estimate and the first column indicates the method used to generate the data. p_1 and p_2 stand for false positive and false negative proportions.

	CONGA		TPGM		SPGM		LPGM		bdgraph		ssgraph		huge	
Data generation method	p_1	p_2	p_1	p_2	p_1	p_2	p_1	p_2	p_1	p_2	p_1	p_2	p_1	p_2
Multi-Poisson	0.08	0	0.22	0.29	0.21	0.34	0.22	0.29	0	0.90	0.27	0.07	0.16	0.20
TPGM	0.04	0.25	0.10	0.02	0.07	0.03	0.10	0.03	0	0.93	0.30	0.15	0.12	0.13
SPGM	0.06	0.23	0.09	0.04	0.07	0.03	0.09	0.04	0	0.95	0.28	0.14	0.12	0.12
LPGM	0.05	0.24	0.07	0.06	0.11	0.07	0.07	0.07	0	0.92	0.31	0.15	0.10	0.09

Table 2: Performance of the competing methods against our proposed method with 30 nodes. Top row indicates the method used to estimate and the first column indicates the method used to generate the data. p_1 and p_2 stand for false positive and false negative proportions.

	CONGA		TPGM		SPGM		LPGM		bdgraph		ssgraph		huge	
Data generation method	p_1	p_2	p_1	p_2	p_1	p_2	p_1	p_2	p_1	p_2	p_1	p_2	p_1	p_2
Multi-Poisson	0	0	0.08	0.57	0.04	0.76	0.08	0.57	0.43	0.25	0.42	0.25	0.13	0.25
TPGM	0.06	0.23	0.05	0.23	0.06	0.23	0.06	0.23	0.41	0.20	0.37	0.21	0.09	0.19
SPGM	0.07	0.22	0.06	0.23	0.06	0.22	0.06	0.23	0.40	0.21	0.38	0.21	0.08	0.18
LPGM	0.07	0.23	0.06	0.22	0.06	0.22	0.06	0.21	0.39	0.19	0.40	0.22	0.08	0.19

Table 3: Performance of the competing methods against our proposed method with 50 nodes. Top row indicates the method used to estimate and the first column indicates the method used to generate the data. p_1 and p_2 stand for false positive and false negative proportions.

	CONGA		TPGM		SPGM		LPGM		bdgraph		ssgraph		huge	
Data generation method	p_1	p_2	p_1	p_2	p_1	p_2	p_1	p_2	p_1	p_2	p_1	p_2	p_1	p_2
Multi-Poisson	0	0	0.01	0.88	0.02	0.76	0.02	0.75	0.46	0.22	0.44	0.25	0.15	0.26
TPGM	0.11	0.23	0.03	0.29	0.03	0.33	0.03	0.33	0.42	0.23	0.43	0.25	0.07	0.21
SPGM	0.11	0.25	0.03	0.33	0.03	0.31	0.03	0.33	0.43	0.21	0.41	0.26	0.08	0.22
LPGM	0.12	0.23	0.03	0.32	0.03	0.34	0.03	0.31	0.43	0.23	0.44	0.26	0.08	0.21

6. Neuron spike count application

The dataset records neuron spike counts in mice across 37 neurons in the brain under the influence of three different external stimuli, 2-D sinusoids with vertical gradient, horizontal gradient, and the sum. These neurons are from the same depth of the visual cortex of a mouse. The data are collected for around 400-time points. In Figure 3, we plot the marginal densities of the spike counts of four neurons under the influence of stimuli 0. We see that there are many variations in the marginal densities, and the densities are multi-modal for some of the cases. Marginally at each node, we also have that the variance is more than the corresponding mean for each of the three stimuli.

6.1 Estimation

We apply exactly the same computational approach as used in the simulation studies. To additionally obtain a summary of the weight of evidence of an edge between nodes j and l, we calculate $S_{jl} = \big(|0.5 - P(\beta_{jl} > 0)|\big)/0.5$, with $P(\beta_{jl} > 0)$ the posterior probability estimated from the MCMC samples. We plot the estimated graph with edge thickness proportional to the values of S_{jl}. Thus thicker edges suggest greater evidence of an edge in Figures 4 to 6. To test for similarity in the graph across stimuli, we estimate 95% credible regions for $\Delta_{jl}^{s,s'} = \beta_{jl}^{s} - \beta_{jl}^{s'}$, denoting the difference in the (j,l) edge weight parameter under stimuli s and s', respectively. We flag those edges (j,l) having 95% credible intervals for $\Delta_{jl}^{s,s'}$ not including zero as significantly different across stimuli.

6.2 Inference

We find that there are 129, 199, and 110 connected edges respectively for stimuli 0, 1, and 2. Among these edges, 38 are common for stimulus 0 and 1. The number is 15 for stimulus 0 and 2, and 28 for stimulus 1 and 2. There are 6 edges that are common for all of the stimuli. These are (13,16), (8,27), (5,8), (33,35), (3,4) and (9, 14). Each node has at least one edge with another node. We plot the estimated network in Figures 4 to 6. We calculate the number of connected edges for each node and list the 5 most connected nodes in Table 4. We also list the most significant 10 edges for each stimulus in Table 5. We find that node 27 is present in all of them. This node seems to have significant interconnections with other nodes for all of the stimuli. We also test the similarity in the estimated weighted network

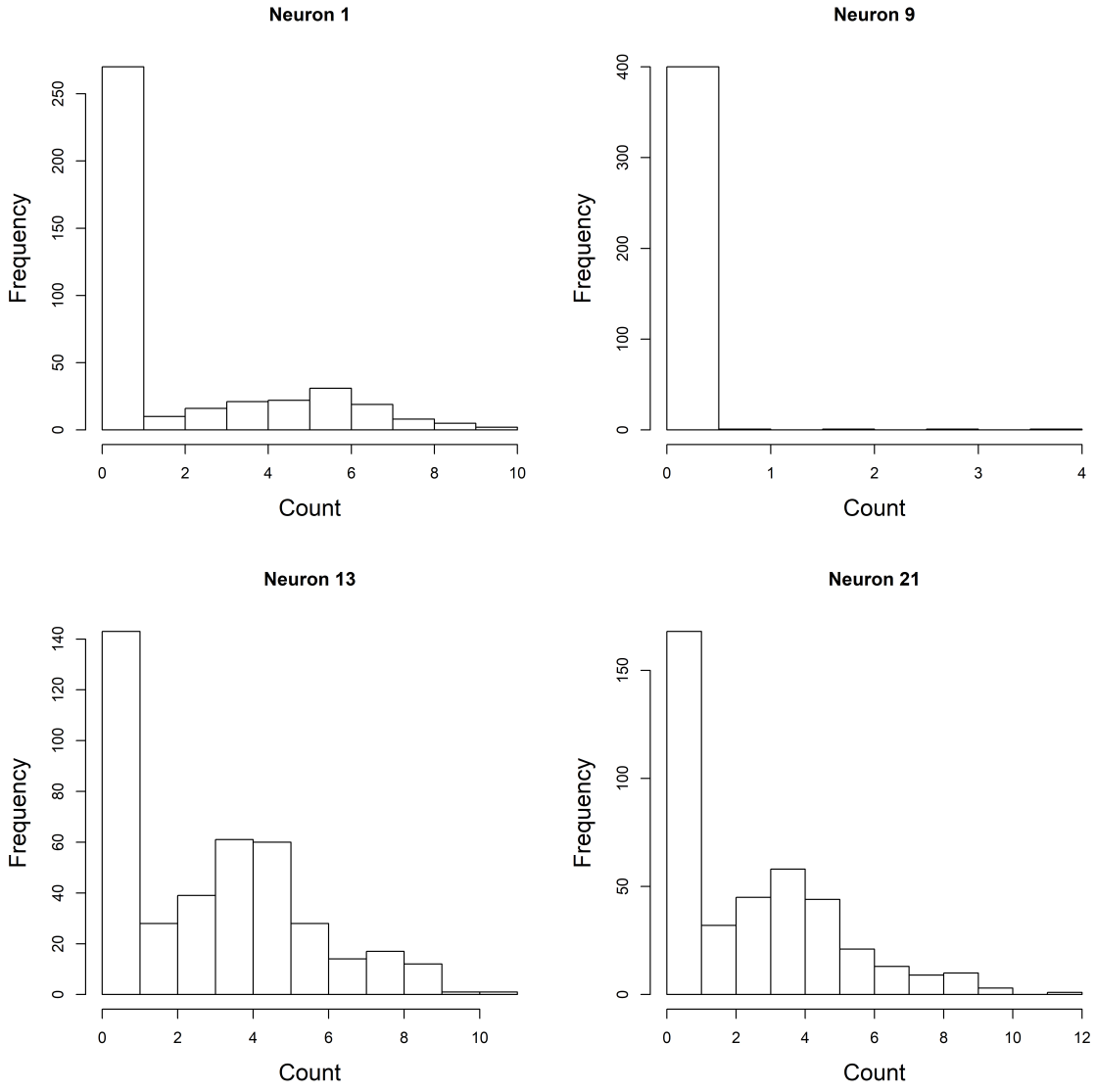

Figure 3: Marginal densities of spike count of the four selected neurons under the influence of stimuli 0.

across stimuli. Here we find 82.13% similarity between the estimated weighted networks under the influence of stimulus 0 and 1. It is 84.98% for the pair 0 and 2. For 1 and 2, it is 79.43%. Stimulus 0 is a combination of stimuli 1 and 2. This could be the reason that the estimated graph under influence of stimulus 0 has the highest similarity with the other estimated graphs.

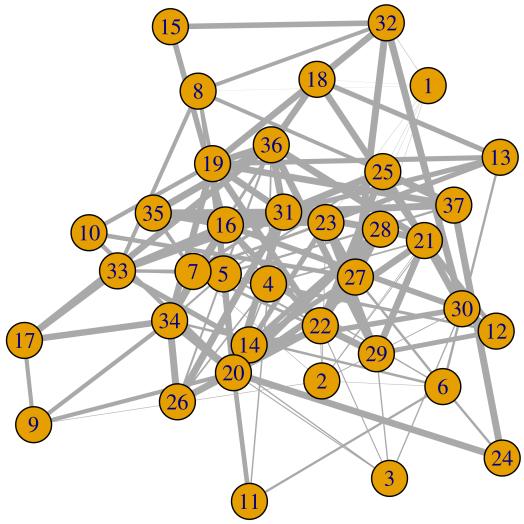

Figure 4: Estimated weighted network under the influence of stimuli 0. The edge width is proportional to the degree of significance.

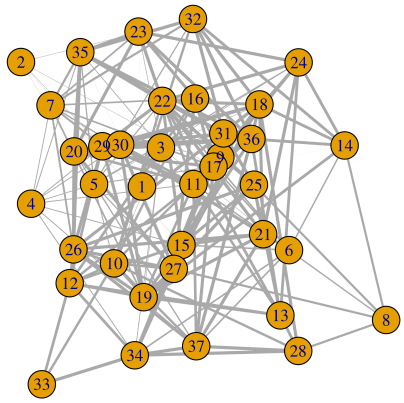

Figure 5: Estimated weighted network under the influence of stimuli 1. The edge width is proportional to the degree of significance.

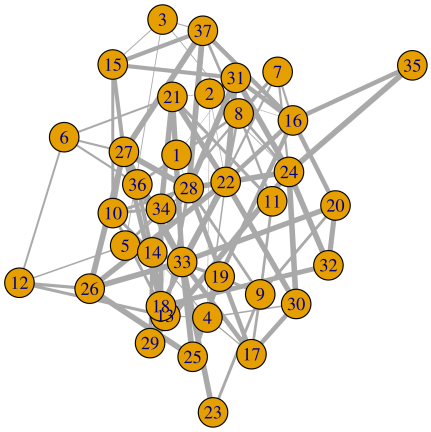

Figure 6: Estimated weighted network under the influence of stimuli 2. The edge width is proportional to the degree of significance.

Table 4: Top 5 nodes with maximum number of connected edges under the influence of stimuli 0, 1 and 2 are listed below.

Stimuli 0		Stimuli 1		Stimuli 2	
Node number	Number of connected edges	Node number	Number of connected edges	Node number	Number of connected edges
37	12	27	16	32	11
6	11	3	15	23	10
9	11	5	14	3	9
25	11	8	14	18	9
27	11	23	14	27	9

Table 5: Top 10 most significant edges under the influence of stimulus 0, 1 and 2 with 1 as the estimated measure of significance are listed below.

Stimuli 0		Stimuli 1		Stimuli 2	
Neuron 1	Neuron 2	Neuron 1	Neuron 2	Neuron 1	Neuron 2
24	35	24	28	14	30
26	30	24	30	16	35
26	37	24	35	21	35
28	37	24	37	21	36
29	33	26	28	24	28
30	32	26	31	24	29
30	35	28	37	24	37
31	33	34	37	25	26
35	36	35	36	26	36
35	37	36	37	31	36

7. Discussion

Our count nonparametric graphical analysis (CONGA) method is useful for multivariate count data, and represents a starting point for more elaborate models and other research directions. One important direction is to time series data. In this respect, a simple extension is to define an autoregressive process for the baseline parameters α_{tj}, inducing correlation in $\alpha_{t-1,j}$ and α_{tj}, while leaving the graph as fixed in time. A more elaborate extension would instead allow the graph to evolve dynamically by replacing the β_{jl} parameters with β_{tjl}, again defining an appropriate autoregressive process.

In this paper, we proposed to tune θ by minimizing the difference $\|cov((\tan^{-1}(X))^{\theta}) - cov(X)\|_F$. However, we could have easily placed a prior on θ and updated it within our posterior sampling algorithm. As the gradient of the pseudo-likelihood with respect to θ is easy to compute, it is possible to develop efficient gradient-based updating algorithms. When λ_{tj}'s are fixed effects, an interesting area of research is to establish graph selection consistency. Such a theory would likely give us more insight regarding the role of θ. Graph selection is expected to suffer both for too small and too large θ.

An additional interesting direction is flexible graphical modeling of continuous positive-valued multivariate data. Such a modification is conceptually straightforward by changing the term $\log(X_{tj}!)$ to the corresponding term in the gamma distribution. All the required functions to fit the CONGA algorithm along with a supplementary R code with an example usage are provided at `https://github.com/royarkaprava/CONGA`.

Acknowledgments

This research was partially supported by grant R01-ES027498-01A1 from the National Institute of Environmental Health Sciences (NIEHS) of the National Institutes of Health (NIH) and grant R01-MH118927 from the National Institute of Mental Health (NIMH) of the NIH.

Appendix

Proof of Theorem 2

The conditional probability is given by,

$$P(X_{tj}, X_{tl}|X_{t,-(j,l)}) = \frac{\exp\big(\sum_{h\in(j,l)}(\alpha_{th}X_{th} - log(X_{th}!) - \sum_{g\neq h}\beta_{gh}(\tan^{-1}(X_{tg}))^{\theta}(\tan^{-1}(X_{th}))^{\theta}\big)}{\sum_{X_{tj}=0}^{\infty}\sum_{X_{tl}=0}^{\infty}\exp\big(\sum_{h\in(j,l)}(\alpha_{th}X_{th} - log(X_{th}!) - \sum_{g\neq h}\beta_{gh}(\tan^{-1}(X_{tg}))^{\theta}(\tan^{-1}(X_{th}))^{\theta}\big)},$$

where $X_{t,-(j,l)} = \{X_{ti} : i \neq (j,l)\}$ and $\log(\lambda_{th}) = \alpha_{th}$. Since $\beta_{jl} = 0$, we can break the exponentiated terms into two such that X_{tj} and X_{jl} are separated out. That would immediately give us, $P(X_{tj}, X_{tl}|X_{t,-(j,l)}) = P(X_{tj}|X_{t,-(j,l)})P(X_{tl}|X_{t,-(j,l)})$.

Proof of Theorem 3

For $q, q^* \in$ the space of probability measure $\mathcal{P}$, let the Kullback-Leibler divergences be given by

$$K(q^*, q) = \int q^* \log\frac{q^*}{q} \qquad V(q^*, q) = \int q^* \log^2\frac{q^*}{q}.$$

Let us denote $p_{i,\alpha,\beta}(X_i)$ as the probability distribution of the data given below,

$$\frac{1}{A(\alpha_i,\beta)}\exp\bigg(\sum_{j=1}^{P}[\alpha_{ij}X_{ij} - \log(X_{ij}!)] + \sum_{l=2}^{P}\sum_{j<l}\beta_{ijl}(\tan^{-1}(X_{ij}))^{\theta}(\tan^{-1}(X_{il}))^{\theta}\bigg),$$

where $A(\alpha_i, \beta)$ is the normalizing constant and $\alpha_i = \{\alpha_{i1}, \ldots, \alpha_{iP}\}, \beta = \{\beta_{jl} : 1 \leq j < l \leq P\}$. Let $E(X_{ij}) = Q$. We have,

$$\frac{\partial\log(A(\alpha_i,\beta))}{\partial\alpha_{ij}} = \frac{A(\alpha_{ij},\beta)}{A(\alpha_i,\beta))}E(X_{ij}) \leq Q, \quad \text{as,} \quad \frac{A(\alpha_{ij},\beta)}{A(\alpha_i,\beta))} \leq 1,$$

and

$$\frac{\partial(A(\alpha_i,\beta))}{\partial\beta_{jl}} \leq (\frac{\pi}{2})^{2\theta}(A(\alpha_i,\beta))$$

Thus we have, $\frac{\partial\log(A(\alpha_i,\beta))}{\partial\alpha_{ik}} \leq Q$ for all $1 \leq k \leq P$ and $\frac{\partial\log(A(\alpha_i,\beta))}{\partial\beta_{jl}} \leq (\frac{\pi}{2})^{2\theta}$.
This implies,

$$-\sum_{j=1}^{P} v_{ij} \leq \log\frac{p_{i,\kappa^0}(X_i)}{p_{i,\kappa}(X_i)} \leq \sum_{j=1}^{P} v_{ij},$$

where $v_{tj} = (TQX_{tj} + C\sum_{l=2}^{P}\sum_{j<l}(\tan^{-1}(X_{tj})) + (\frac{\pi}{2})^{2\theta}(T + qC)$. We have $E(v_{tj}) < \infty$ due to the last assumption. From the dominated convergence theorem as $n \to \infty$, we have κ converges to κ^0. Thus Kullback-Leibler divergences go to zero.

Thus the posterior is weakly consistent. The weak and strong topologies on countable spaces are equivalent by Scheffe's theorem. Thus the posterior for κ is also strongly consistent at κ^0.

References

John Aitchison and CH Ho. The multivariate Poisson-log normal distribution. *Biometrika*, 76(4):643–653, 1989.

Genevera I Allen and Zhandong Liu. A log-linear graphical model for inferring genetic networks from high-throughput sequencing data. In *Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on*, pages 1–6. IEEE, 2012.

Adrian Baddeley and Rolf Turner. Practical maximum pseudolikelihood for spatial point patterns: (with discussion). *Australian & New Zealand Journal of Statistics*, 42(3):283–322, 2000.

Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate aussian or binary data. *Journal of Machine Learning Research*, 9(Mar):485–516, 2008.

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

Julian Besag. Statistical analysis of non-lattice data. *The Statistician*, pages 179–195, 1975.

KS Chan and Johannes Ledolter. Monte carlo em estimation for time series models involving counts. *Journal of the American Statistical Association*, 90(429):242–252, 1995.

Shizhe Chen, Daniela M Witten, and Ali Shojaie. Selection and estimation for mixed graphical models. *Biometrika*, 102(1):47–64, 2014.

Siddhartha Chib and Rainer Winkelmann. Markov chain monte carlo analysis of correlated count data. *Journal of Business & Economic Statistics*, 19(4):428–435, 2001.

Julien Chiquet, Mahendra Mariadassou, and Stéphane Robin. Variational inference for sparse network reconstruction from count data. *arXiv preprint arXiv:1806.03120*, 2018.

Francis Comets. On consistency of a class of estimators for exponential families of Markov random fields on the lattice. *The Annals of Statistics*, pages 455–468, 1992.

Victor De Oliveira. Bayesian analysis of conditional autoregressive models. *Annals of the Institute of Statistical Mathematics*, 64(1):107–133, 2012.

Victor De Oliveira. Hierarchical Poisson models for spatial count data. *Journal of Multivariate Analysis*, 122:393–408, 2013.

Peter J Diggle, JA Tawn, and RA Moyeed. Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(3):299–350, 1998.

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

Adrian Dobra, Alex Lenkoski, and Abel Rodriguez. Bayesian inference for general Gaussian graphical models with application to multivariate lattice data. *Journal of the American Statistical Association*, 106(496):1418–1433, 2011.

Adrian Dobra, Reza Mohammadi, et al. Loglinear model selection and human mobility. *The Annals of Applied Statistics*, 12(2):815–845, 2018.

Alan E Gelfand and Penelope Vounatsou. Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics*, 4(1):11–15, 2003.

Subhashis Ghosal and Aad Van der Vaart. *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press, 2017.

Fabian Hadiji, Alejandro Molina, Sriraam Natarajan, and Kristian Kersting. Poisson dependency networks: Gradient boosted models for multivariate count data. *Machine Learning*, 100(2-3):477–507, 2015.

John M Hammersley and Peter Clifford. Markov fields on finite graphs and lattices. *Unpublished manuscript*, 1971.

John L Hay and Anthony N Pettitt. Bayesian analysis of a time series of counts with covariates: an application to the control of an infectious disease. *Biostatistics*, 2(4): 433–444, 2001.

David Inouye, Pradeep Ravikumar, and Inderjit Dhillon. Admixture of Poisson mrfs: A topic model with word dependencies. In *International Conference on Machine Learning*, pages 683–691, 2014.

David I Inouye, Pradeep Ravikumar, and Inderjit S Dhillon. Generalized root models: beyond pairwise graphical models for univariate exponential families. *arXiv preprint arXiv:1606.00813*, 2016a.

David I Inouye, Pradeep Ravikumar, and Inderjit S Dhillon. Square root graphical models: Multivariate generalizations of univariate exponential families that permit positive dependencies. *arXiv preprint arXiv:1603.03629*, 2016b.

David I Inouye, Eunho Yang, Genevera I Allen, and Pradeep Ravikumar. A review of multivariate distributions for count data derived from the Poisson distribution. *Wiley Interdisciplinary Reviews: Computational Statistics*, 9(3):e1398, 2017.

Ali Jalali, Pradeep Ravikumar, Vishvas Vasuki, and Sujay Sanghavi. On learning discrete graphical models using group-sparse regularization. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 378–387, 2011.

Jens Ledet Jensen and Hans R Künsch. On asymptotic normality of pseudo likelihood estimates for pairwise interaction processes. *Annals of the Institute of Statistical Mathematics*, 46(3):475–486, 1994.

Mladen Kolar and Eric P Xing. Improved estimation of high-dimensional Ising models. *arXiv preprint arXiv:0811.1239*, 2008.

Mladen Kolar, Le Song, Amr Ahmed, Eric P Xing, et al. Estimating time-varying networks. *The Annals of Applied Statistics*, 4(1):94–123, 2010.

Han Liu, John Lafferty, and Larry Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10(Oct):2295–2328, 2009.

Shigeru Mase. Marked gibbs processes and asymptotic normality of maximum pseudo-likelihood estimators. *Mathematische Nachrichten*, 209(1):151–169, 2000.

Abdolreza Mohammadi, Ernst C Wit, et al. Bayesian structure learning in sparse Gaussian graphical models. *Bayesian Analysis*, 10(1):109–138, 2015.

Abdolreza Mohammadi, Fentaw Abegaz, Edwin van den Heuvel, and Ernst C Wit. Bayesian modelling of dupuytren disease by using Gaussian copula graphical models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66(3):629–645, 2017.

Jared S Murray, David B Dunson, Lawrence Carin, and Joseph E Lucas. Bayesian Gaussian copula factor models for mixed data. *Journal of the American Statistical Association*, 108(502):656–665, 2013.

Johan Pensar, Henrik Nyman, Juha Niiranen, Jukka Corander, et al. Marginal pseudo-likelihood learning of discrete Markov network structures. *Bayesian analysis*, 12(4):1195–1215, 2017.

Pradeep Ravikumar, Martin J Wainwright, and John D Lafferty. High-dimensional Ising model selection using ℓ1-regularized logistic regression. *The Annals of Statistics*, 38(3): 1287–1319, 2010.

Arkaprava Roy, Brian J Reich, Joseph Guinness, Russell T Shinohara, and Ana-Maria Staicu. Spatial shrinkage via the product independent Gaussian process prior. *arXiv preprint arXiv:1805.03240*, 2018.

Lorraine Schwartz. On bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 4(1):10–26, 1965.

Måns Thulin. Decision-theoretic justifications for Bayesian hypothesis testing using credible sets. *Journal of Statistical Planning and Inference*, 146:133–138, 2014.

Marijtje AJ Van Duijn, Krista J Gile, and Mark S Handcock. A framework for the comparison of maximum pseudo-likelihood and maximum likelihood estimation of exponential family random graph models. *Social Networks*, 31(1):52–62, 2009.

Martin J Wainwright, John D Lafferty, and Pradeep K Ravikumar. High-dimensional graphical model selection using ℓ_1 regularized logistic regression. In *Advances in neural information processing systems*, pages 1465–1472, 2007.

Hao Wang. Bayesian graphical lasso models and efficient posterior computation. *Bayesian Analysis*, 7(4):867–886, 2012.

Hao Wang. Scaling it up: Stochastic search structure learning in graphical models. *Bayesian Analysis*, 10(2):351–377, 2015.

Yiyi Wang and Kara M Kockelman. A Poisson-lognormal conditional-autoregressive model for multivariate spatial analysis of pedestrian crash counts across neighborhoods. *Accident Analysis & Prevention*, 60:71–84, 2013.

Michel Wedel, Ulf Böckenholt, and Wagner A Kamakura. Factor models for multivariate count data. *Journal of Multivariate Analysis*, 87(2):356–369, 2003.

Peter Xue-Kun Song. Multivariate dispersion models generated from Gaussian copula. *Scandinavian Journal of Statistics*, 27(2):305–320, 2000.

Eunho Yang, Pradeep K Ravikumar, Genevera I Allen, and Zhandong Liu. On Poisson graphical models. In *Advances in Neural Information Processing Systems*, pages 1718–1726, 2013.

Eunho Yang, Pradeep Ravikumar, Genevera I Allen, and Zhandong Liu. Graphical models via univariate exponential family distributions. *The Journal of Machine Learning Research*, 16(1):3813–3847, 2015.

Mingyuan Zhou, Lauren A Hannah, David B Dunson, and Lawrence Carin. Beta-negative binomial process and Poisson factor analysis. *In AISTATS*, pages 1462–1471, 2012.

Xiang Zhou and Scott C Schmidler. Bayesian parameter estimation in Ising and Potts models: A comparative study with applications to protein modeling. Technical report, Duke University, 2009.