Bayesian Analysis (2019) **14**, Number 4, pp. 1271–1301

Hierarchical Normalized Completely Random Measures for Robust Graphical Modeling

Andrea Cremaschi[*,†], Raffaele Argiento[‡,§], Katherine Shoemaker[¶,‖], Christine Peterson[**] and Marina Vannucci[††]

Abstract. Gaussian graphical models are useful tools for exploring network structures in multivariate normal data. In this paper we are interested in situations where data show departures from Gaussianity, therefore requiring alternative modeling distributions. The multivariate t-distribution, obtained by dividing each component of the data vector by a gamma random variable, is a straightforward generalization to accommodate deviations from normality such as heavy tails. Since different groups of variables may be contaminated to a different extent, Finegold and Drton (2014) introduced the Dirichlet t-distribution, where the divisors are clustered using a Dirichlet process. In this work, we consider a more general class of nonparametric distributions as the prior on the divisor terms, namely the class of normalized completely random measures (NormCRMs). To improve the effectiveness of the clustering, we propose modeling the dependence among the divisors through a nonparametric hierarchical structure, which allows for the sharing of parameters across the samples in the data set. This desirable feature enables us to cluster together different components of multivariate data in a parsimonious way. We demonstrate through simulations that this approach provides accurate graphical model inference, and apply it to a case study examining the dependence structure in radiomics data derived from The Cancer Imaging Atlas.

Keywords: graphical models, Bayesian nonparametrics, normalized completely random measures, hierarchical models, radiomics data, t-distribution.

1 Introduction

Graphical models describe the conditional dependence relationships among a set of random variables. A graph $G = (V, E)$ specifies a set of vertices $V = \{1, 2, \ldots, p\}$ and a set of edges $E \subset V \times V$. In a directed graph, edges are denoted by ordered pairs $(i, j) \in E$. In an undirected graph, $(i, j) \in E$ if and only if $(j, i) \in E$ (Lauritzen, 1996). Here we focus on undirected graphical models, also known as Markov random fields. In

[*]Department of Cancer Immunology, Institute of Cancer Research, Oslo University Hospital, Oslo, Norway
[†]Oslo Centre for Biostatistics and Epidemiology (OCBE), University of Oslo, Oslo, Norway
[‡]ESOMAS Department, University of Torino, Torino, Italy
[§]Collegio Carlo Alberto, Torino, Italy
[¶]Department of Statistics, Rice University, Houston, TX, USA
[‖]Department of Biostatistics, University of Texas MD Anderson Cancer Center, Houston, Texas, USA
[**]Department of Biostatistics, University of Texas MD Anderson Cancer Center, Houston, Texas, USA
[††]Department of Statistics, Rice University, Houston, TX, USA

 https://doi.org/10.1214/19-BA1153

this class of models, the absence of an edge between two vertices means that the two corresponding variables are conditionally independent given the remaining variables, while an edge is included whenever the two variables are conditionally dependent.

In the context of multivariate normal data, graphical models are known as Gaussian graphical models (GGMs) or covariance selection models (Dempster, 1972). In this setting, the graph structure G implies constraints on the precision matrix (the inverse of the covariance matrix). Specifically, a zero entry in the precision matrix corresponds to the absence of an edge in the graph, meaning that the corresponding nodes (variables) are conditionally independent. Since graphical model estimation corresponds to estimation of a sparse matrix, regularization methods are a natural approach. In particular, the graphical lasso (Meinshausen and Bühlmann, 2006; Yuan and Lin, 2007; Friedman et al., 2008), which imposes an L_1 penalty on the sum of the absolute values of the entries of the precision matrix, is a popular method for achieving the desired sparsity. Among Bayesian approaches, the Bayesian graphical lasso, proposed as the Bayesian analogue to the graphical lasso, places double exponential priors on the off-diagonal entries of the precision matrix (Wang, 2012; Peterson et al., 2013), while approaches which enforce exact zeros in the precision matrix have been proposed by Roverato (2002), Jones et al. (2005), and Dobra et al. (2011). Gaussian graphical models have been widely applied in genomics and proteomics to infer various types of networks, including co-expression, gene regulatory, and protein interaction networks (Friedman, 2004; Dobra et al., 2004; Mukherjee and Speed, 2008; Stingo et al., 2010; Telesca et al., 2012; Peterson et al., 2016).

Some extensions of standard Gaussian graphical models exist in the literature for the analysis of data that show departures from normality. Among others, Pitt et al. (2006) used copula models and Bhadra et al. (2018) used Gaussian scale mixtures. Here, we build upon the approach of Finegold and Drton (2011, 2014), who introduced a vector of positive latent contamination parameters (*divisors*) regulating the departure from Gaussianity and then modeled those as a sample from a nonparametric distribution, specifically a Dirichlet process. Their model, however, does not allow the exchange of information among the vectors of observed data, since independent Dirichlet process priors are used for each of the n samples. We propose to use a more flexible class of nonparametric prior distributions, known as normalized completely random measures (NormCRMs), and consider a hierarchical construction where the nonparametric priors for the divisors are conditionally independent, given their centering measure, which is itself a completely random measure. NormCRMs were first introduced by Regazzini et al. (2003) with the name of Normalized Random Measures with Independent increments (NRMI), and subsequently studied by several researchers in statistics and machine learning (James et al., 2009; Lijoi and Prünster, 2010; Caron and Fox, 2017). One of the most commonly used measures in this class is the Normalized Generalized Gamma (NGG) process (Lijoi et al., 2007). For illustrations of the use of this prior in mixture models, see Argiento et al. (2010), Barrios et al. (2013), and Argiento et al. (2016). Theoretical and clustering properties of hierarchical CRMs were first investigated by Camerlenghi et al. (2019) (see also Camerlenghi et al., 2017, 2018). Subsequently, Argiento et al. (2019) have focused on clustering and computational issues arising under mixture models built upon this class of priors. In this paper, we exploit the clustering characterization of

these constructions to induce sharing of information. More specifically, we focus our attention on the normalized generalized gamma process, which has been shown to yield a quite flexible clustering structure. Furthermore, we devise a suitable MCMC algorithm for posterior sampling.

We are motivated by an application to radiomics data derived from magnetic resonance imaging (MRI) of glioblastoma patients collected as part of The Cancer Imaging Atlas. In the development of personalized cancer treatment, there is great interest in using information from tumor imaging data to better characterize a patient's disease, as these medical images are collected as a routine part of diagnosis. There have been a large number of different numerical summaries proposed, but the interpretation of these features is not immediate. It is hypothesized that clinically relevant features may be capturing related aspects of the underlying disease. Statistical modeling of the dependencies in radiomics data poses challenges, however, as the features exhibit outliers and overdispersion due to heterogeneity of the tumor presentation across patients.

The paper is organized as follows: we begin in Section 2 with a review of graphical models. In Section 3, we lay out the proposed model and summarize computational methods for inference. We then illustrate the application of the method to both simulated and a publicly available radiomics data set in Section 4. Finally, we conclude with a discussion on the current model as well as future directions in Section 5.

2 Background

2.1 Gaussian Graphical Models

Let $\mathbf{X}_i \in \mathbb{R}^p$ be a random vector, with $i = 1, \ldots, n$. In GGMs, the conditional independence relationships between pairs of nodes encoded by a graph G correspond to constraints on the precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ of the multivariate normal distribution

$$\mathbf{X}_i \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad i = 1, \ldots, n, \tag{1}$$

with $\boldsymbol{\mu} \in \mathbb{R}^p$ the mean vector and $\boldsymbol{\Sigma} \in \mathbb{R}^p \times \mathbb{R}^p$ a positive definite symmetric matrix. Specifically, the precision matrix $\boldsymbol{\Omega}$ is constrained to the cone of symmetric positive definite matrices with off-diagonal entry ω_{ij} equal to zero if there is no edge in G between nodes i and j.

In Bayesian analysis, the standard conjugate prior for the precision matrix $\boldsymbol{\Omega}$ is the Wishart distribution. Given the constraints of a graph among the variables, Roverato (2002) proposed the G-Wishart distribution as the conjugate prior. The G-Wishart is the Wishart distribution restricted to the space of precision matrices with zeros specified by a graph G. The G-Wishart density $W_G(b, D)$ can be written as

$$p(\boldsymbol{\Omega}|G, b, D) = I_G(b, D)^{-1} |\boldsymbol{\Omega}|^{(b-2)/2} \exp\Big\{ -\frac{1}{2}\mathrm{tr}(\boldsymbol{\Omega} D)\Big\}, \quad \boldsymbol{\Omega} \in P_G$$

where $b > 2$ is the degrees of freedom parameter, D is a $p \times p$ positive definite symmetric matrix, I_G is the normalizing constant, and P_G is the set of all $p \times p$ positive definite

symmetric matrices with $\omega_{ij} = 0$ if and only if $(i, j) \notin E$. Even when the graph structure is known, sampling from this distribution poses computational difficulties since both the prior and posterior normalizing constants are intractable. Dobra et al. (2011) proposed a reversible jump algorithm to sample over the joint space of graphs and precision matrices that does not scale well to large graphs. Wang and Li (2012) and Lenkoski (2013) proposed sampler methods that do not require proposal tuning and circumvent computation of the prior normalizing constant through the use of the exchange algorithm, improving both the accuracy and efficiency of the computations. Mohammadi and Wit (2015) proposed a sampling methodology based on birth-death processes for the appearance or removal of an edge in the graph. Their algorithm, implemented in the R package BDgraph, can be used with the approximation of the normalizing constant of the G-Wishart prior calculated either via the Monte Carlo method of Atay-Kayis and Massam (2005) or the Laplace approximation of Lenkoski and Dobra (2011).

To sum up, we can write the standard Gaussian graphical model in the Bayesian setting as:

$$\begin{aligned} &\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n | \boldsymbol{\mu}, \boldsymbol{\Omega}, \overset{iid}{\sim} \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Omega}), \\ &\boldsymbol{\mu} \sim \mathcal{N}_p(\boldsymbol{\mu}_0, \mathbb{I}_p/\sigma^2_{\boldsymbol{\mu}}) \\ &\boldsymbol{\Omega} | G \sim G\text{-Wishart}(G, b, D), \\ &G \sim \pi(G), \end{aligned} \tag{2}$$

with the symbol $\mathbb{I}_p$ indicating the identity matrix of dimension p. The last ingredient to fully specify the model is the prior for the graph G. When prior knowledge is not available, a uniform prior is often used (see Lenkoski and Dobra, 2011). However, it is well known that this prior is not optimal for sparsity as it favors graphs with a moderately large number of edges. To overcome this issue, Dobra et al. (2004) and Jones et al. (2005) suggested assigning a small data-dependent inclusion probability to each edge, i.e., $\pi(G) \propto d^{|E|}(1-d)^{\binom{p}{2}-|E|}$, with $d = 2/(p-1)$. This prior, adopted also in this paper, is called the Erdős-Rényi prior, and it reduces to the uniform prior when $d = 0.5$.

2.2 Robust Graphical Models

Assume $\boldsymbol{Y}_i \in \mathbb{R}^p$ is a vector of observed data on p variables for subject i, with $i = 1, \ldots, n$. When data show departures from normality, robust models are needed. In particular, as noted by Finegold and Drton (2011, 2014), t-distributions are well-suited to accommodate heavy tails, and result in minimal loss of efficiency when the data are in fact normal. They propose introducing the normal variables $\boldsymbol{X}_i$ in model (2) as latent quantities, and modeling the observed data as:

$$Y_{ij} = \mu_j + \frac{X_{ij}}{\sqrt{\theta_{ij}}} \quad j = 1, \ldots, p, \tag{3}$$

where $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{ip})$, for $i = 1, \ldots, n$, are data and variable specific perturbation parameters (divisors), taking into account the deviation from normality of the observations. Using the invariance under linear transformation property of the Gaussian

distribution, we can express the sampling model as:

$$\boldsymbol{Y}_i|\boldsymbol{\mu},\boldsymbol{\Omega},\boldsymbol{\theta}_i \overset{iid}{\sim} \mathcal{N}_p(\boldsymbol{\mu}, diag(\sqrt{\boldsymbol{\theta}_i})\ \boldsymbol{\Omega}\ diag(\sqrt{\boldsymbol{\theta}_i})), \quad i = 1\ldots, n. \tag{4}$$

Different distributions of the vector $\boldsymbol{\theta}_i$ yield different models. Let P_0 denote a gamma$(\nu/2, \nu/2)$ distribution with mean 1 and variance $2/\nu$. If $\theta_{i1} = \theta_{i2} = \cdots = \theta_{ip}$ and $\theta_{i1} \sim P_0$ (i.e., just one common divisor for all the components), then a multivariate t-distribution is assumed for the observations. We refer to this model as $\boldsymbol{Y}_i \sim t_{p,\nu}(\boldsymbol{\mu}, \boldsymbol{\Omega})$. On the other hand, if $\theta_{i1}, \ldots, \theta_{ip} \overset{\text{iid}}{\sim} P_0$ (i.e., p different divisors, one for each component of the data), then $\boldsymbol{Y}_i$ is distributed according to an alternative t-distribution, as introduced by Finegold and Drton (2011), and denoted by $\boldsymbol{Y}_i \sim t^*_{p,\nu}(\boldsymbol{\mu}, \boldsymbol{\Omega})$. As an intermediate case, Finegold and Drton (2014) consider $\theta_{i1}, \ldots, \theta_{ip}|P_i \overset{\text{iid}}{\sim} P_i$, $P_i \sim DP(\kappa, P_0)$, where $P_i \sim DP(\kappa, P_0)$ is the Dirichlet process with mass parameter κ and centering measure P_0. We refer to this model as $\boldsymbol{Y} \sim t^{\kappa}_{p,\nu}(\boldsymbol{\mu}, \boldsymbol{\Omega})$. A realization from the Dirichlet process P_i is almost surely a discrete random probability measure. To give an illustration, let $p = 2$. Therefore, if $(\theta_{i1}, \theta_{i2})|P_i \overset{\text{iid}}{\sim} P_i$, then with probability $\mathbb{P}(\theta_{i1} = \theta_{i2}) = \frac{1}{\kappa+1}$, and $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}) \sim t_{2,\nu}(\boldsymbol{\mu}, \boldsymbol{\Omega})$. On the other hand, with probability $\frac{\kappa}{\kappa+1}$ we have the alternative t case. Indeed, the two are limiting cases of the Dirichlet t-distribution when $\kappa \to 0$ or $\kappa \to +\infty$, respectively. Even though the Dirichlet process has proven to perform well in several contexts, it is well known that the clustering it induces is often inaccurate as it is affected by the so-called *rich-gets-richer* effect. In the next section, we propose a more flexible approach to mitigate this behavior and to allow for a more flexible clustering structure.

3 Proposed Method

3.1 Robust Graphical Modeling via Hierarchical Normalized Completely Random Measures

We propose an extension of the Dirichlet t model that uses a more flexible class of nonparametric distributions, namely the class of hierarchical normalized completely random measures (NormCRM). Through the use of these measures, we are able to address some of the limitations of the Dirichlet process. First, the tendency towards a highly skewed distribution of cluster sizes can be mitigated by the use of the more flexible NormCRM. In addition, we show how exploiting a hierarchical construction facilitates the *sharing of information* across cluster components in the dataset.

Let Θ be the Euclidean space and let us consider the class of almost surely discrete random probability measures that can be written as:

$$\tilde{P}(\cdot) = \sum_{l\geq 1} \frac{J_l}{\mathcal{T}} \delta_{\tau_l}(\cdot) = \sum_{l\geq 1} w_l \delta_{\tau_l}(\cdot), \tag{5}$$

where $\mathcal{T} = \sum_{l\geq 1} J_l$. We assume $\tilde{P}$ to be a homogeneous NormCRM, whose law is characterized by a *Lèvy intensity measure* ν that factorizes into $\nu(ds, d\tau) = \alpha(s)P(d\tau)ds$,

where α is the density of a nonnegative measure, absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^+$ and regular enough to guarantee that $0 < \mathcal{T} < \infty$ almost surely, and P is a probability measure over $(\Theta, \mathcal{B})$. In general, such a factorization does not hold true, but by adopting it we ensure that the random jumps $\{J_l\}_{l\geq 1}$ and the random locations $\{\tau_l\}_{l\geq 1}$ are independent sequences. The random locations $\tau_1, \tau_2, \ldots$ are independent and identically distributed according to the base distribution P, while the unnormalized random masses $J_1, J_2, \ldots$ are distributed according to a Poisson random measure with intensity α. The Dirichlet process is encompassed by this class when $\alpha(s) = \kappa s^{-1}\mathrm{e}^{-s}$, for $\kappa > 0$ and $s > 0$. Even though our approach can be implemented with a general NormCRM, in what follows we consider the specific case of the normalized generalized gamma (NGG) process (Lijoi et al., 2007), which is obtained by choosing $\alpha(s) = \frac{\kappa}{\Gamma(1-\sigma)} s^{-1-\sigma}\mathrm{e}^{-s}$, for $0 \leq \sigma < 1$. This nonparametric prior has been shown to be very effective for model-based clustering (Lijoi et al., 2007). We also refer readers to Argiento et al. (2015), for an application in biostatistics. Note that, when $\sigma = 0$, the Dirichlet process is recovered.

A sample $\theta_1, \ldots, \theta_p | \tilde{P} \overset{\text{iid}}{\sim} \tilde{P} = \sum_{l\geq 1} w_l \delta_{\tau_l}$ can be represented via the set of variables $\boldsymbol{l} = (l_1, \ldots, l_p) | \{w_l\}_{l\geq 1} \overset{\text{iid}}{\sim} \text{Discrete}(\{w_l\}_{l\geq 1})$, by letting $\theta_j = \tau_{l_j}$, for $j = 1, \ldots, p$. Let $\boldsymbol{l}^* = (l_1^*, \ldots, l_K^*)$ be the set of the K unique values in $\boldsymbol{l}$. A partition $\rho = \{C_1, \ldots, C_K\}$ of the indices $\{1, \ldots, p\}$ can be defined by letting $C_h = \{j : l_j = l_h^*\}$, for $h = 1, \ldots, K$. The partition ρ is called l-clustering. Let now $\theta_h^* = \tau_{l_h^*}$. When the centering measure P is diffuse, the θ_h^*s coincide with the unique values in $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ and the l-clustering coincides with the *natural clustering*, i.e., we can also write $C_h = \{j : \theta_j = \theta_h^*\}$, for $h = 1, \ldots, K$. The l-clustering and the natural clustering can be different when the centering measure P is discrete (see Argiento et al., 2019). In particular, in the discrete case, each θ_h^* is the value shared by all the indices in the so-called l-cluster C_h, for $h = 1, \ldots, K$. Furthermore, extending the results in Pitman (1996), Ishwaran and James (2003), and Argiento et al. (2019), one can show that, for P either atomic or diffuse, the following characterization holds:

$$\mathcal{L}(\rho, d\theta_1^*, \ldots, d\theta_K^*) = \mathcal{L}(\rho)\mathcal{L}(d\theta_1^*, \ldots, d\theta_K^* | K) = \text{eppf}(\boldsymbol{e}; \kappa, \sigma) \prod_{h=1}^{K} P(d\theta_h^*), \tag{6}$$

with $\boldsymbol{e} = (e_1, \ldots, e_K)$ the vector of l-cluster sizes in the partition ρ, such that $e_h = \#C_h$, for each $h = 1, \ldots, K$, and with the notation $\text{eppf}(\boldsymbol{e}; \kappa, \sigma)$ indicating the exchangeable partition probability function (eppf) of the NGG process, a symmetric function of the cluster sizes $\boldsymbol{e}$, as introduced by Pitman (2003). The explicit analytical form of the eppf of a generic (homogeneous) NormCRM can be derived (see formulas (36)-(37) in Pitman (2003)) and enables the construction of a Gibbs sampler based on the Chinese restaurant process representation. In the Dirichlet process case, De Blasi et al. (2015) pointed out that the predictive distribution induced by (6), i.e., the probability that θ_p belongs to a new l-cluster given $(\theta_1, \ldots, \theta_{p-1})$, depends only on the dimension p, while in the NGG process case this probability depends on both p and K, leading to a more flexible prior. Furthermore, the probability that θ_p belongs to a previously observed l-cluster C_h is proportional to $e_h - \sigma$, for $h = 1, \ldots, K$. These two properties mitigate the rich-gets-richer behavior arising when considering the Dirichlet case.

The first step towards our proposed robust graphical modeling construction is to replace the Dirichlet prior on the divisors with an NGG process, yielding the following robust graphical model:

$$\begin{aligned}
&\boldsymbol{Y}_i|\boldsymbol{\mu},\boldsymbol{\Omega},\boldsymbol{\theta}_i \overset{iid}{\sim} \mathcal{N}_p(\boldsymbol{\mu}, diag(\sqrt{\boldsymbol{\theta}_i})\ \boldsymbol{\Omega}\ diag(\sqrt{\boldsymbol{\theta}_i})), \quad i = 1\ldots,n,\\
&\theta_{i1},\ldots,\theta_{ip}|P_i \overset{\text{iid}}{\sim} P_i, \quad i = 1,\ldots,n,\\
&P_1,\ldots,P_n|\kappa,\sigma \overset{\text{iid}}{\sim} NGG(\kappa,\sigma,P),
\end{aligned} \tag{7}$$

where P is the distribution on the space of the divisors. Suitable prior distributions can be assigned to κ, σ, $\boldsymbol{\mu}$, and $\boldsymbol{\Omega}$. We denote by $\rho_i = \{C_{i1},\ldots,C_{iK_i}\}$ the l-clustering induced by P_i in each data vector, for $i = 1,\ldots,n$, as described earlier in this section.

When P is diffuse, model (7) does not allow for sharing of information across the data vectors. This can be seen by using characterization (6) to marginalize model (7) with respect to the infinite-dimensional parameters $P_1,\ldots,P_n$, and rewriting the last two lines of (7) as:

$$\begin{aligned}
&\rho_i|\kappa,\sigma \overset{\text{ind}}{\sim} \text{eppf}(\boldsymbol{e}_i;\kappa,\sigma), \quad i = 1,\ldots,n,\\
&\theta^*_{i1},\ldots,\theta^*_{iK_i}|K_i \overset{\text{iid}}{\sim} P, \quad i = 1,\ldots,n,
\end{aligned}$$

where $(\rho_i,\boldsymbol{\theta}^*_i)$ represent the partition and the vector of unique values induced by P_i on the data components, and $\boldsymbol{e}_i = (e_{i1},\ldots,e_{iK_i})$ is the vector of l-cluster sizes in the partition ρ_i, such that $e_{ih} = \#C_{ih}$, for each $h = 1,\ldots,K_i$. By this re-writing, it is clear that the sharing of information among the different clustering structures is achieved only via the conditional dependence of ρ_i given κ and σ. In particular, we cannot have shared divisors across data vectors, but only across components of the same data vector, since $\boldsymbol{\theta}^*$'s are all i.i.d. from the diffuse distribution P. We overcome this limitation by considering a more flexible hierarchical model formulation that allows for additional sharing of information across the samples. Specifically, we assume P to be a random probability measure, namely an NGG process centered on a diffuse measure P_0. In formulas, the proposed model can be written as follows:

$$\begin{aligned}
&\boldsymbol{Y}_i|\boldsymbol{\mu},\boldsymbol{\Omega},\boldsymbol{\theta}_i \overset{iid}{\sim} \mathcal{N}_p(\boldsymbol{\mu}, diag(\sqrt{\boldsymbol{\theta}_i})\ \boldsymbol{\Omega}\ diag(\sqrt{\boldsymbol{\theta}_i})), \quad i = 1\ldots,n,\\
&\theta_{i1},\ldots,\theta_{ip}|P_i \overset{\text{iid}}{\sim} P_i, \quad i = 1,\ldots,n,\\
&P_1,\ldots,P_n|\kappa,\sigma,P \overset{\text{iid}}{\sim} NGG(\kappa,\sigma,P),\\
&P|\kappa_0,\sigma_0 \sim NGG(\kappa_0,\sigma_0,P_0).
\end{aligned} \tag{8}$$

The law of $(P_1,\ldots,P_n)$, as given by the last two lines of (8), is called the hierarchical NGG (HNGG) process. For ease of notation, we will refer to the mixture model (8) as t-HNGG. Theoretical and clustering properties of hierarchical normalized completely random measures have been investigated first by Camerlenghi et al. (2019), and a detailed study of the clustering induced by these measures in the context of mixture models has been conducted in Argiento et al. (2019). An attractive feature

of this construction is that it induces a two-layered hierarchical clustering structure that allows components of different observed data vectors to be clustered together. This two-layered structure consists of an l-clustering ρ_i within each group i, and a clustering η that merges elements of $\rho_1, \ldots, \rho_n$. In order to define η more precisely, let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_n^\top)^\top$ be the matrix where the row $\boldsymbol{\theta}_i$ is the vector of all divisors of the i-th observation, and let $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_M)$ be the vector of unique values found in the matrix $\boldsymbol{\theta}$. Let $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_n)$ indicate the l-clustering partitions in each data vector, and $\boldsymbol{\theta}^* = (\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_n^*)$ the corresponding multidimensional array of parameter values. We define a clustering of the indices of the multidimensional array $\boldsymbol{\theta}^*$ by letting $\eta = \{D_1, \ldots, D_M\}$ where $D_m = \{(i,h) : \theta_{ih}^* = \psi_m, h = 1, \ldots, K_i, i = 1, \ldots, n\}$, with $m = 1, \ldots, M$. We also let $\boldsymbol{d} = (d_1, \ldots, d_M)$, with $d_m = \#D_m$. Then, the law of the matrix $\boldsymbol{\theta}$ of divisors, given in the last three lines of (8), can be characterized in terms of $\boldsymbol{\rho}$, η and $\boldsymbol{\psi}$ as:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\rho}, \eta, d\boldsymbol{\psi}) &= \mathcal{L}(\eta|\boldsymbol{\rho}) \prod_{i=1}^{n} \mathcal{L}(\rho_i) \prod_{m=1}^{M} P_0(d\psi_m) \\ &= \text{eppf}(\boldsymbol{d}; \kappa_0, \sigma_0) \prod_{i=1}^{n} \text{eppf}(\boldsymbol{e}_i; \kappa, \sigma) \prod_{m=1}^{M} P_0(d\psi_m). \end{aligned} \tag{9}$$

Full details on the derivation of formula (9) can be found in Argiento et al. (2019). We also note that the partially exchangeable partition probability function of Camerlenghi et al. (2019) can be obtained from (9) by marginalizing with respect to $(\eta, \boldsymbol{\psi})$.

We call the *natural clustering* induced by $\boldsymbol{\theta}$ the partition of indices $\mathcal{I} = \{\mathcal{I}_1, \ldots, \mathcal{I}_M\}$ such that $(i,j) \in \mathcal{I}_m$ iff $\theta_{ij} = \psi_m$. Since the sets of indices $\mathcal{I}_m$, for $m = 1, \ldots, M$, can be recovered from $(\boldsymbol{\rho}, \eta)$, formula (9) characterizes the law of the natural clustering. The relationship between $\mathcal{I}$ and $(\boldsymbol{\rho}, \eta)$ is clarified in formulas:

$$\begin{aligned} \mathcal{I}_m^{(\rho_i, \eta)} &:= \bigcup_{h=1}^{K_i} \{(i,j) : j \in C_{ih}, (i,h) \in D_m\}, \qquad m = 1, \ldots, M, \\ \mathcal{I}_m &:= \mathcal{I}_m^{(\boldsymbol{\rho}, \eta)} = \bigcup_{i=1}^{n} \mathcal{I}_m^{(\rho_i, \eta)}, \qquad\qquad m = 1, \ldots, M. \end{aligned} \tag{10}$$

Formula (9) can be described in terms of a Chinese restaurant franchise process. In our context, each observation represents a different restaurant in the franchise, each serving p customers, one for each component of the data vector. Customers entering the i-th restaurant are allocated to the tables according to $\text{eppf}(\boldsymbol{e}_i; \kappa, \sigma)$, independently from the other restaurants in the franchise, and generate the partition $\rho_i = (C_{i1}, \ldots, C_{iK_i})$, for $i = 1, \ldots, n$. In this metaphor, the elements of ρ_i represent the tables of the i-th restaurant. Conditionally on $T = \sum_{i=1}^{n} K_i$, the tables of the franchise are grouped according to the law described by $\text{eppf}(\boldsymbol{d}; \kappa_0, \sigma_0)$, thus obtaining a partition of tables. Hence, the elements of η can be interpreted as clusters of tables. In addition, all tables in the same cluster D_m share the same dish ψ_m, for $m = 1, \ldots, M$. Moreover, $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_M)$ is and i.i.d. sample from P_0. Under this metaphor, e_{ih}, for $h = 1, \ldots, K_i$

and $i = 1, \ldots, n$, represents the number of customers seated at the h-th table in the i-th restaurant, while d_m, for $m = 1, \ldots, M$, is the number of tables where the m-th dish is served across the franchise. Finally, in this metaphor, the natural clustering induced by the corresponding $\boldsymbol{\theta}$ is formed of clusters of customers that share the same dish across the franchise, and not only in the same restaurant.

3.2 Predictive Structure of the hierarchical NGG

In this paper, we make use of a marginal MCMC algorithm for simulating the nonparametric quantities involved in model (8). This algorithm is based on integrating out the infinite-dimensional parameters $P_1, \ldots, P_n, P$ and on the characterization of the generalized Chinese restaurant franchise process via formula (9). To drastically reduce the computational complexity, it is convenient to consider the predictive structure induced by the hierarchical NGG process by using a standard augmentation trick (see James et al., 2009; Lijoi and Prünster, 2010). More specifically, we introduce $n + 1$ auxiliary random variables $\boldsymbol{U} = (U_1, \ldots, U_n, U_0)$, referring to the n clustering structures in each data vector, and to the one existing across the whole dataset, respectively. For the NGG process, each partition ρ_i has the following law, jointly with U_i:

$$\text{eppf}(\boldsymbol{e}_i, u_i; \kappa, \sigma) = \frac{u_i^{p-1}}{\Gamma(p)} e^{-\frac{\kappa}{\sigma}((u_i+1)^\sigma - 1)} \prod_{h=1}^{K_i} \frac{\kappa}{(u_i+1)^{e_{ih}-\sigma}} \frac{\Gamma(e_{ih}-\sigma)}{\Gamma(1-\sigma)}, \tag{11}$$

where K_i is the number of clusters and $\boldsymbol{e}_i = (e_{i1}, \ldots, e_{iK_i})$ is the vector of cluster sizes in ρ_i. The joint law of (η, U_0) has an analogous expression.

Suppose now to have a new variable in the i-th group, whose index is $p + 1$. We will use an abuse on notation by indicating with $(p + 1) \in C_{ih}$, for $h = 1, \ldots, K_i$, the event that the new variable is allocated to the h-th cluster in the i-th group, and with $(p + 1) \in C_{iK_i+1}$ the event that the new variable is assigned to a new cluster. It can be shown that the allocation probabilities of the new variable are the following, for $h = 1, \ldots, K_i$:

$$\begin{aligned} P_{ih}^{(to)} &= \mathbb{P}((p+1) \in C_{ih} | \rho_i, U_i) \propto \frac{\text{eppf}(e_{i1}, \ldots, e_{ih}+1, \ldots, e_{iK_i}; \kappa, \sigma, u_i)}{\text{eppf}(e_{i1}, \ldots, e_{iK_i}; \kappa, \sigma, u_i)} = e_{ih} - \sigma, \\ P_i^{(tn)} &= \mathbb{P}((p+1) \in C_{iK_i+1} | \rho_i, U_i) \propto \frac{\text{eppf}(e_{i1}, \ldots, e_{iK_i}, 1; \kappa, \sigma, u_i)}{\text{eppf}(e_{i1}, \ldots, e_{iK_i}; \kappa, \sigma, u_i)} = \kappa(u_i+1)^\sigma, \end{aligned} \tag{12}$$

corresponding to the allocation probabilities of a new customer entering the i-th restaurant, and sitting at an existing or at a new table in the generalized Chinese restaurant metaphor. In case a new cluster arises, the partition η needs to be updated. The allocation probabilities of the new element $T + 1$ are, for $m = 1, \ldots, M$:

$$\begin{aligned} P_m^{(do)} &= \mathbb{P}((T+1) \in D_m | \eta, U_0) \propto \frac{\text{eppf}(d_1, \ldots, d_m+1, \ldots, d_M; \kappa_0, \sigma_0, u_0)}{\text{eppf}(d_1, \ldots, d_M; \kappa_0, \sigma_0, u_0)} = d_m - \sigma_0, \\ P^{(dn)} &= \mathbb{P}((T+1) \in D_{M+1} | \eta, U_0) \propto \frac{\text{eppf}(d_1, \ldots, d_M, 1; \kappa_0, \sigma_0, u_0)}{\text{eppf}(d_1, \ldots, d_M; \kappa_0, \sigma_0, u_0)} = \kappa_0(u_0+1)^{\sigma_0}, \end{aligned} \tag{13}$$

corresponding to the allocation probabilities that a newly generated table will join a new or an existing cluster of tables. Additional details on how to derive (12) and (13) can be found in the Supplementary Materials (Cremaschi et al., 2019).

To complete the generalized Chinese restaurant franchise process metaphor, not only does a new customer have to select a table, but also a dish from the franchise menu. Suppose the new customer enters the i-th restaurant, and let θ_{ip+1} be the label of the selected dish. The table is picked according to the predictive rules (12) of the i-th restaurant. The customer can choose between joining an existing table with label $h = 1, \ldots, K_i$, or occupying the (K_i+1)-th new one. The first choice leads to sharing the dish on the h-th table in the i-th restaurant, i.e. $\theta_{i(p+1)} = \theta^*_{ih}$. On the other hand, if a new table is chosen, the customer can select a dish from the menu of dishes according to (13). This menu contains dishes that are already served in other tables across the franchise, as well as infinitely many new ones, since the centering measure P_0 is diffuse. Following Argiento et al. (2019), the full-conditional allocation probability, for $i = 1, \ldots, n$ is

$$\begin{aligned}
&\mathbb{P}((p+1) \in C_{ih}, (i, p+1) \in \mathcal{I}_m^{(\rho_i,\eta)} | \boldsymbol{\rho}, \eta, \boldsymbol{U}) \\
&\qquad = \mathbb{P}((i, p+1) \in \mathcal{I}_m^{(\rho_i,\eta)} | (p+1) \in C_{ih}, \boldsymbol{\rho}, \eta, \boldsymbol{U}) \mathbb{P}((p+1) \in C_{ih} | \boldsymbol{\rho}, \eta, \boldsymbol{U}) \\
&\qquad \propto \begin{cases} P_{ih}^{(to)} & h = 1, \ldots, K_i, \quad m = m_h, \\ P_m^{(do)} P_i^{(tn)} & h = K_i + 1, \quad m = 1, \ldots, M \\ P^{(dn)} P_i^{(tn)} & h = K_i + 1, \quad m = M + 1, \\ 0 & \text{otherwise}, \end{cases}
\end{aligned} \tag{14}$$

where m_h is such that $(i, h) \in D_{m_h}$, and the sets $\mathcal{I}_m^{(\rho_i,\eta)}$, for $m = 1, \ldots, M$, are defined in (10). These equations are the main blocks to compute the full-conditional allocation probabilities needed for posterior sampling, as presented in the next section. Such probabilities can be calculated in closed form using equation (9). However, in the next section we make use of auxiliary variables as these simplify the sampling algorithm. The conditional predictive probabilities hereby specified characterize the prior clustering induced by our nonparametric modeling. We refer to Argiento et al. (2019) for results on the distribution of relevant quantities, such as the prior distribution of the number of different dishes or the dependence induced by our model across observations (e.g. correlation, coskewness).

3.3 MCMC Algorithm

In this section, we describe the MCMC algorithm for posterior inference from model (8), embedded within the graphical modeling part described in (2). The state space of the Gibbs sampler is given by $(\boldsymbol{\mu}, \boldsymbol{\Omega}, G, \boldsymbol{\rho}, \eta, \boldsymbol{\psi})$. We describe the parameter updates by splitting them into two blocks: the *graphical model* block, which comprises the full conditionals of $(\boldsymbol{\mu}, \boldsymbol{\Omega}, G)$, and the *generalized Chinese restaurant franchise* block, which includes those for $(\boldsymbol{\rho}, \eta, \boldsymbol{\psi})$. For simplicity, we will remove the indexing of the Gibbs sampler iteration.

- **Graphical model updates:** In the following, we will consider the law of $(\boldsymbol{\mu}, \boldsymbol{\Omega}, G)$ conditionally upon the variables $(\boldsymbol{\rho}, \eta, \boldsymbol{\psi})$.

– For the update of $(\boldsymbol{\Omega}, G)$, we resort to the birth-death algorithm of Mohammadi and Wit (2015) available in the R package `BDgraph`, and suitable for non-decomposable graphs. The algorithm proceeds by first adding/removing an edge of the graph, and then updating the precision matrix $\boldsymbol{\Omega}$ using the algorithm presented in Lenkoski (2013). These moves have probabilities

$$\mathbb{P}((i,j) \in E | \boldsymbol{\mu}, \boldsymbol{\Omega}, G, \boldsymbol{Y}, \boldsymbol{\theta}) \propto \beta_{ij}^b(\boldsymbol{\mu}, \boldsymbol{\Omega}, G, \boldsymbol{Y}, \boldsymbol{\theta}), \quad (i,j) \notin E,$$
$$\mathbb{P}((i,j) \notin E | \boldsymbol{\mu}, \boldsymbol{\Omega}, G, \boldsymbol{Y}, \boldsymbol{\theta}) \propto \beta_{ij}^d(\boldsymbol{\mu}, \boldsymbol{\Omega}, G, \boldsymbol{Y}, \boldsymbol{\theta}), \quad (i,j) \in E,$$

with β_{ij}^b and β_{ij}^d the birth and death rates of edge (i,j), respectively, computed in such a way that the stationary distribution of the Markov process is the joint full-conditional of $(\boldsymbol{\Omega}, G)$, given $(\boldsymbol{Y}, \boldsymbol{\rho}, \eta, \boldsymbol{\psi})$ (see Theorem 3.1 in Mohammadi and Wit, 2015). This algorithm is particularly efficient since the Markov process specification ensures that the birth/death moves are always accepted, contrarily to the reversible jump algorithm of Giudici and Green (1999), also implemented in the package `BDgraph`.

– **Updating $\boldsymbol{\mu}$:** This full-conditional is conjugate. A-priori $\boldsymbol{\mu} \sim \mathcal{N}_p(\boldsymbol{\mu}_0, \mathbb{I}_p/\sigma_{\boldsymbol{\mu}}^2)$, hence:

$$\boldsymbol{\mu} | G, \boldsymbol{\Omega}, \boldsymbol{\theta}, \boldsymbol{Y} \sim \mathcal{N}_p(\boldsymbol{m}_{\boldsymbol{\mu}}, \boldsymbol{S}_{\boldsymbol{\mu}}),$$
$$\boldsymbol{S}_{\boldsymbol{\mu}} = \mathbb{I}_p/\sigma_{\boldsymbol{\mu}}^2 + \sum_{i=1}^{n} \left(diag(\sqrt{\boldsymbol{\theta}_i}) \; \boldsymbol{\Omega} \; diag(\sqrt{\boldsymbol{\theta}_i}) \right),$$
$$\boldsymbol{m}_{\boldsymbol{\mu}} = \boldsymbol{S}_{\boldsymbol{\mu}} \left[\boldsymbol{\mu}_0/\sigma_{\boldsymbol{\mu}}^2 + \sum_{i=1}^{n} \left(diag(\sqrt{\boldsymbol{\theta}_i}) \; \boldsymbol{\Omega} \; diag(\sqrt{\boldsymbol{\theta}_i}) \right) \boldsymbol{Y}_i^\top \right].$$

- **Generalized Chinese restaurant franchise process updates:** We refer to the notation of Sections 3.1 and 3.2. Conditionally to the vector of auxiliary variables $\boldsymbol{U}$ and the graphical model parameters $(\boldsymbol{\mu}, \boldsymbol{\Omega}, G)$, the joint law of (8) is:

$$\mathcal{L}(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n | \boldsymbol{\rho}, \eta, \boldsymbol{\theta}, \boldsymbol{U}, \boldsymbol{\mu}, \boldsymbol{\Omega}, G) \mathcal{L}(\rho_1, \ldots, \rho_n | U_1, \ldots, U_n) \mathcal{L}(\eta | U_0) \prod_{m=1}^{M} P_0(d\psi_m)$$
$$= \prod_{i=1}^{n} f(\boldsymbol{y}_i | \boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\theta}_i) \prod_{i=1}^{n} \text{eppf}(\boldsymbol{e}_i; \kappa, \sigma, P_0, u_i) \text{eppf}(\boldsymbol{d}; \kappa_0, \sigma_0, P_0, u_0) \prod_{m=1}^{M} P_0(d\psi_m),$$

with f representing the multivariate Gaussian density introduced in model (8). It is important to point out that, under model (8), the observations y_{ij} are now components of the vector $\boldsymbol{y}_i$ and are no longer conditionally independent. Thus, it is useful to introduce the following conditional likelihood for a subset of indices $t \subset \{1, \ldots, p\}$:

$$f(\boldsymbol{y}_{it} | \boldsymbol{y}_{i\setminus t}, \boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\theta}_i) = \mathcal{N}\left(\boldsymbol{y}_{it} \middle| \boldsymbol{\mu}_c, \left[diag(\sqrt{\boldsymbol{\theta}_i}) \; \boldsymbol{\Omega} \; diag(\sqrt{\boldsymbol{\theta}_i}) \right]_{tt} \right),$$
$$\boldsymbol{\mu}_c = \boldsymbol{\mu}_t - \boldsymbol{\Omega}_{tt}^{-1} \boldsymbol{\Omega}_{t\setminus t} (\boldsymbol{y}_{i\setminus t} - \boldsymbol{\mu}_{\setminus t}) \sqrt{\boldsymbol{\theta}_{i\setminus t}}, \tag{15}$$

with $\backslash t = \{1, \dots, p\} \cap \bar{t}$, and $\bar{t}$ indicating the complementary set of t. This allows us to write the following updates:

– **Update of $\boldsymbol{U}$ and $\boldsymbol{\psi}$:** using the expression given in (11) of eppf$(\cdot, u_0; \kappa_0, \sigma_0, P_0)$ and eppf$(\cdot, u_i; \kappa, \sigma, P_0)$, and the centering measure P_0, we have:

$$\begin{aligned} p(U_i|\rho_i, \kappa, \sigma) &\propto u_i^{p-1} e^{-\frac{\kappa}{\sigma}((u_i+1)^\sigma - 1)} \\ &\quad \times \prod_{h=1}^{K_i} \left(\frac{\kappa}{(u_i+1)^{e_{ih}-\sigma}} \frac{\Gamma(e_{ih}-\sigma)}{\Gamma(1-\sigma)} \right), i = 1, \dots, n, \\ p(U_0|\eta, \kappa_0, \sigma_0, T) &\propto u_0^{T-1} e^{-\frac{\kappa_0}{\sigma_0}((u_0+1)^{\sigma_0} - 1)} \\ &\quad \times \prod_{m=1}^{M} \left(\frac{\kappa_0}{(u_0+1)^{d_m-\sigma_0}} \frac{\Gamma(d_m-\sigma_0)}{\Gamma(1-\sigma_0)} \right), \\ p(\psi_m|\boldsymbol{Y}, \boldsymbol{\rho}, \eta) &\propto \prod_{i=1}^{n} f(\boldsymbol{y}_{i\mathcal{I}_m^{(\boldsymbol{\rho}_i,\eta)}}|\boldsymbol{y}_{i\backslash\mathcal{I}_m^{(\boldsymbol{\rho}_i,\eta)}}, \boldsymbol{\mu}, \boldsymbol{\Omega}, \psi_m) P_0(d\psi_m), \\ &\quad m = 1, \dots, M. \end{aligned} \tag{16}$$

These quantities are often known up to a normalizing constant, making necessary to implement a series of Metropolis-Hastings (MH) steps. Specifically, we use an adaptive MH scheme for the random variables $\boldsymbol{U}$, following the guidelines of Griffin and Stephens (2013). The sampling of the unique values $\boldsymbol{\psi}$ is achieved by performing M independent standard MH steps. This approach is necessary since the full-conditional distribution of $\boldsymbol{\psi}$ presents an intractable normalizing constant, and does not allow the use of a direct sampler (Finegold and Drton, 2011).

– **Update of $(\boldsymbol{\rho}, \eta)$:** We report now the full-conditional distributions for the clustering variables $(\boldsymbol{\rho}, \eta)$. The updating takes advantage of the augmented predictive representation given in Section 3.2, inspired by (Favaro and Teh, 2013) and by the popular Algorithm 8 of (Neal, 2000). Indeed, due to the non-conjugate setting of our model, we augment the sample space to include a set of N_c auxiliary variables $\boldsymbol{\psi}^c = (\psi_1^c, \dots, \psi_{N_c}^c) \overset{\text{iid}}{\sim} P_0$. Let the superscript $(-ij)$ denote the conditioning on the random variables modified after the removal of the j-th observation of the i-th restaurant, for $j = 1, \dots, p$ and $i = 1, \dots, n$. Then, conditionally upon $\boldsymbol{Y}$ and $(\boldsymbol{\rho}^{-ij}, \eta^{-ij})$, and resorting to (14), the probability of assigning the j-th customer to the h-th table of the i-th restaurant, where the m-th dish is served, is:

$$\begin{aligned} &\mathbb{P}(j \in C_{ih}^{-ij}, \theta_{ij} = \psi_m|\boldsymbol{Y}, \boldsymbol{\mu}, \boldsymbol{\Omega}, G, \boldsymbol{\rho}^{-ij}, \eta^{-ij}, \boldsymbol{\theta}^{-ij}, \boldsymbol{\psi}^c, \boldsymbol{U}) \\ &\propto \begin{cases} P_{ih}^{(to)} f(\boldsymbol{y}_{ij}|\boldsymbol{y}_{i\backslash j}, \boldsymbol{\mu}, \boldsymbol{\Omega}, \theta_{ih}^*), & h = 1, \dots, K_i^{-ij} \\ P_i^{(tn)} P_m^{(do)} f(\boldsymbol{y}_{ij}|\boldsymbol{y}_{i\backslash j}, \boldsymbol{\mu}, \boldsymbol{\Omega}, \psi_m), & h = K_i^{-ij}+1, m = 1, \dots, M^{-ij} \\ P_i^{(tn)} P^{(dn)} f(\boldsymbol{y}_{ij}|\boldsymbol{y}_{i\backslash j}, \boldsymbol{\mu}, \boldsymbol{\Omega}, \psi_{n_c}^c)/N_c, & \\ \quad h = K_i^{-ij}+1, m = M^{-ij}+1, n_c = 1, \dots, N_c, & \end{cases} \end{aligned} \tag{17}$$

where $K_i^{-ij}+1$ and $M^{-ij}+1$ are the new table and dish labels, respectively. The updating process continues by re-allocating C_{ih} to a cluster of tables. To this end, we have to assign C_{ih} to a D_m, for $h = 1, \ldots, K_i$ and $m = 1, \ldots, M$. More formally, let the superscript $(-ih)$ indicate the conditioning on the variables after the removal of all the observations in C_{ih}. Conditionally on $\boldsymbol{Y}$ and $(\boldsymbol{\rho}^{-ih}, \eta^{-ih})$, and using again (14), the probability of assigning the h-th table of the i-th restaurant to the m-th cluster is:

$$\mathbb{P}((i,h) \in D_m^{-ih}, \theta_{ih}^* = \psi_m | \boldsymbol{Y}, \boldsymbol{\rho}^{-ih}, \eta^{-ih}, \boldsymbol{\psi}^{-ih}, \boldsymbol{\psi}^c, \boldsymbol{U}) \tag{18}$$
$$\propto \begin{cases} P_m^{(do)} f(\boldsymbol{y}_{iC_{ih}} | \boldsymbol{y}_{i\setminus C_{ih}}, \boldsymbol{\mu}, \boldsymbol{\Omega}, \psi_m), & m = 1, \ldots, M^{-ih}, \\ P^{(dn)} f(\boldsymbol{y}_{iC_{ih}} | \boldsymbol{y}_{i\setminus C_{ih}}, \boldsymbol{\mu}, \boldsymbol{\Omega}, \psi_{n_c}^c)/N_c, & m = M^{-ih}+1, \quad n_c = 1, \ldots, N_c. \end{cases}$$

where $M^{-ih}+1$ indicates the new dish labels.

Given the output from the MCMC chain, one can estimate the graph structure by considering the median graph (Barbieri et al., 2004) as the graph represented by those edges $(i,j) \in E$ for which the posterior edge inclusion probability $\mathbb{P}((i,j) \in E|\boldsymbol{Y})$ is greater than 0.5. Additionally, we can estimate the precision matrix of the sampling model (4) by considering the contribution of the divisors $\boldsymbol{\theta}$, as

$$\boldsymbol{\Omega}_{\boldsymbol{\theta}} = \frac{1}{n} \sum_{i=1}^{n} diag(\sqrt{\boldsymbol{\theta}_i})\ \boldsymbol{\Omega}\ diag(\sqrt{\boldsymbol{\theta}_i}).$$

We obtain an analogous estimate $\hat{\boldsymbol{\Omega}}_{\boldsymbol{\theta}}$ by averaging over the MCMC samples.

One important feature of the nonparametric prior distributions imposed in models (7) and (8) is the ability to cluster the data via the unique values of the divisors $\boldsymbol{\theta}$. In the applications below, we illustrate the properties of the random partitions imposed on $\boldsymbol{\theta}$ by reporting the posterior mean of the number of clusters in each data vector for the independent model (7), and the posterior distribution of the number of clusters among all the data vectors for the hierarchical model (8). Both quantities are computed by using the saved iterations of the posterior chains of the random partitions $\boldsymbol{\rho}$ and η.

4 Applications

4.1 Simulation Study

In this section we illustrate the performance of the proposed method via simulation studies. In particular, we employ two simulated scenarios inspired by the work of Finegold and Drton (2014), for ease of comparison. Analogously, an edge $(i,j) \in E$ is considered "positive" if $\mathbb{P}((i,j) \in E|\boldsymbol{Y}) > \epsilon$, for a range of values of $\epsilon \in (0,1)$. We compare results across different models in terms of the receiver operating characteristic (ROC) curves, by calculating true and false positive rates for each of 50 replicated datasets and then computing the ROC curves by averaging over the 50 replicates.

AR(1) graph, $n = p = 25$

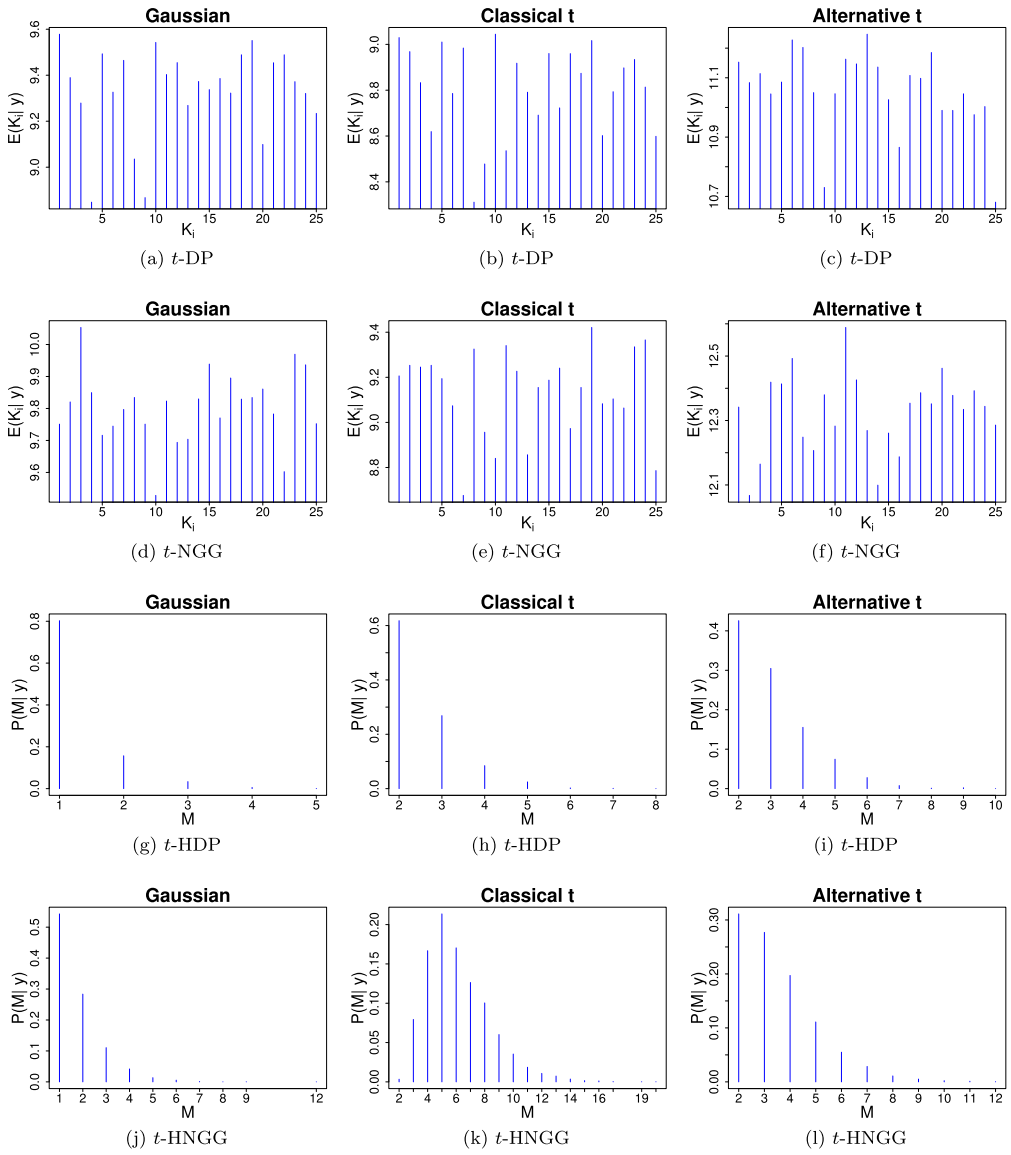
In this simulation setting, $n = 25$ data vectors are simulated from model (4) with an AR(1) graph structure on G, induced by a tri-diagonal precision matrix $\boldsymbol{\Omega}$ where the off-diagonal non-zero elements are set to -1, and the diagonal ones are set to 3 apart from the two extremes that are set to 2. The mean vector $\boldsymbol{\mu}$ is simulated as p independent standard normal random variables. The divisors $\boldsymbol{\theta}$ are set to recover different distribution structures, namely the multivariate Gaussian ($\theta_{ij} = 1$ for $i = 1, \ldots, n$, $j = 1, \ldots, p$), the classical multivariate t-Student ($\theta_{i1} = \cdots = \theta_{ip} \overset{\text{iid}}{\sim} \text{gamma}(\nu/2, \nu/2)$, $i = 1, \ldots, n$), and the alternative multivariate t-Student ($\theta_{11}, \ldots, \theta_{pp} \overset{\text{iid}}{\sim} \text{gamma}(\nu/2, \nu/2)$). Where required, $\nu = 3$.

Here, we investigate performance of four different models: an independent Dirichlet model (t-DP) obtained from (7) with $\kappa \sim \text{gamma}(1, 1)$ and $\sigma = 0$ ($\mathbb{E}(M) = n4.31$, $\text{sd}(M) = \sqrt{n}1.56$); an independent t-NGG model obtained from (7) with $\kappa \sim \text{gamma}(1, 1)$ and by setting $\sigma = 0.1$ ($\mathbb{E}(M) = n4.40$, $\text{sd}(M) = \sqrt{n}1.70$); a t-HDP model in the form of equation (8) with $\kappa, \kappa_0 \sim \text{gamma}(1, 1)$ and $\sigma = \sigma_0 = 0$ ($\mathbb{E}(M) = 4.50$, $\text{sd}(M) = 1.56$); and a t-HNGG model obtained from (8) with $\kappa, \kappa_0 \sim \text{gamma}(1, 1)$ and by setting $(\sigma, \sigma_0) = (0.5, 0.1)$ ($\mathbb{E}(M) = 7.67$, $\text{sd}(M) = 2.41$). Alternatively, Beta hyperpriors can be imposed on σ, σ_0 (see our second simulation setting below and Argiento et al. (2019) for a full sensitivity analysis on the parameters $(\kappa, \kappa_0, \sigma, \sigma_0)$).

Furthermore, in all models, we set the prior distribution for G to be uniform with edge probability $d = 0.05$. We also set $b = p$ and $D = \mathbb{I}_p$ for the prior distribution of $\boldsymbol{\Omega}$, and $\nu = 3$. For each replicated dataset, we ran an MCMC chain with 50,000 iterations, of which the first 40,000 are discarded as burn-in period, and 5,000 are saved from the remaining ones, after thinning, for estimation purposes.

In order to elucidate the properties of the clustering structure of the divisors induced by our model, in Figure 1 we show the posterior distributions of the number of clusters for each of the four fitted models, on one of the replicated datasets for each of the three different scenarios. As expected, both the t-HDP and the t-HNGG model induce a lower posterior mean number of clusters (in the natural clustering sense), when compared to the number of clusters in each data vector induced by the independent t-DP and t-NGG models. This is possible thanks to the ability of the hierarchical models to exploit the sharing of information across data vectors. This effect is particularly clear when looking at the Gaussian scenario, where the proposed model is able to effectively cluster the data into one cluster with high posterior probability. On the other hand, in the alternative multivariate t case it is clear how the tuning of the hyperparameters plays a crucial role in the resulting partition structure. The classical t case shows that neither hierarchical model accurately captures the original number of clusters, equal to 25. However, the t-HNGG model outperforms the t-HDP model, allowing for higher posterior probability on partitions characterized by a larger number of clusters.

Figure 2 shows the comparison of the ROC curves for the four different models, computed by averaging over the 50 replicates, for each of the three simulation settings. We can observe an agreement in the results for the Gaussian case, while the proposed model performs better in the other scenarios, due to the presence of non-unitary divisors



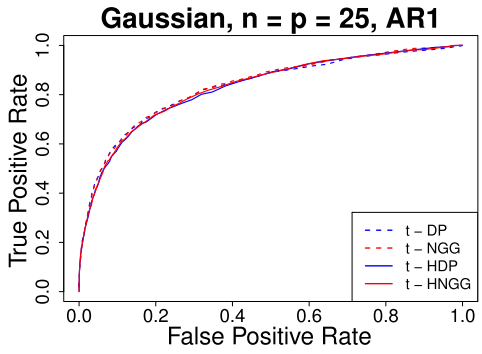
(a) t-DP (b) t-DP (c) t-DP

(d) t-NGG (e) t-NGG (f) t-NGG

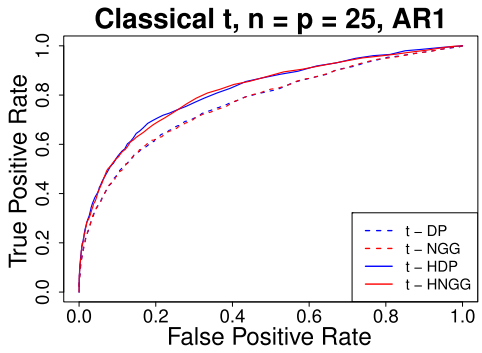
(g) t-HDP (h) t-HDP (i) t-HDP

(j) t-HNGG (k) t-HNGG (l) t-HNGG

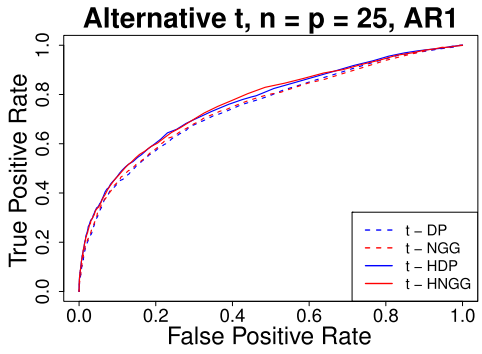
Figure 1: Simulation study, AR(1) graph: Posterior number of clusters, comparing the four models under study (independent t-DP, independent t-NGG, t-HDP, and t-HNGG), for data generated from a Gaussian distribution (first column), a Classical t distribution (second column) and an Alternative t distribution (third column). Posterior means for each data vector are reported for the independent t-DP and t-NGG models (first and second rows, respectively), while posterior distributions are shown for the t-HDP and the t-HNGG models (third and fourth rows).



(a)



(b)



(c)

Figure 2: Simulation study, AR(1) graph: ROC curves comparing the independent t-DP and t-NGG models with the t-HDP and the t-HNGG models, for data generated from a Gaussian distribution, a Classical t distribution and an Alternative t distribution.

that can be captured by the flexible nonparametric structure. The t-HDP and t-HNGG models show comparable performance. Furthermore, in Table 1 we report the L^1, L^2,

	Gaussian			
	t-DP	t-NGG	t-HDP	t-HNGG
L^1	15.5948	15.3678	3.3862	3.4717
L^2	26.4608	26.2770	6.8677	6.8875
Max	9.2539	9.1041	1.1490	1.2120

	Classical t			
	t-DP	t-NGG	t-HDP	t-HNGG
L^1	12.7382	13.0816	4.3448	4.3848
L^2	16.6510	17.2454	9.0376	8.7959
Max	6.1404	6.3945	1.5965	1.6277

	Alternative t			
	t-DP	t-NGG	t-HDP	t-HNGG
L^1	8.5563	9.2981	5.3283	5.2390
L^2	11.3243	12.0485	9.0608	8.8924
Max	4.7448	5.3204	2.8309	2.7630

Table 1: Simulation study, AR(1) graph: Average distances between $\hat{\boldsymbol{\Omega}}_{\boldsymbol{\theta}}$ and $\boldsymbol{\Omega}_{\boldsymbol{\theta}}$.

and maximum modulus distances between the estimated and the simulated precision matrices, averaged over the 50 replicates, for each of the three simulated scenarios. For two matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{p\times p}$, these measures are defined as:

$$\begin{aligned} d_{L^1}(\boldsymbol{A},\boldsymbol{B}) &= \max_{1\le j\le p} \sum_{i=1}^{p} |a_{ij}-b_{ij}|, \\ d_{L^2}(\boldsymbol{A},\boldsymbol{B}) &= \sqrt{\sum_{i=1}^{p}\sum_{j=1}^{p}(a_{ij}-b_{ij})^2}, \\ d_{max}(\boldsymbol{A},\boldsymbol{B}) &= \max_{ij} |a_{ij}-b_{ij}|. \end{aligned} \tag{19}$$

The proposed model clearly outperforms the independent ones in all simulated scenarios. Once again, the two hierarchical models yield comparable results. Additional details on this analysis are reported in the Supplementary Materials, where the posterior estimates of the precision and covariance matrices are compared for the different models and simulation scenarios.

Contaminated data, $n = 100$ and $p = 30$

Next, we illustrate the behavior of our model on a more complex simulated data structure. In particular, we simulate $n = 100$ p-dimensional random vectors, with $p = 30$. In this set of simulations, the graph structure G is produced by splitting the p-dimensional graph into three random graphs of size 10 each, while the elements of the related precision matrix $\boldsymbol{\Omega}$ are set to 3 on the diagonal (2 at the extremes), and to -1 for the off-diagonal non-zero elements. Then, the values are multiplied by a constant factor, yielding to a minimum eigenvalue of $\boldsymbol{\Omega}$ bigger than 0.5. The divisor matrix $\boldsymbol{\theta}$ is produced by working on its vectorized version, $vec(\boldsymbol{\theta})$. We sample $n_r, n_c \sim$ Poisson(10), and

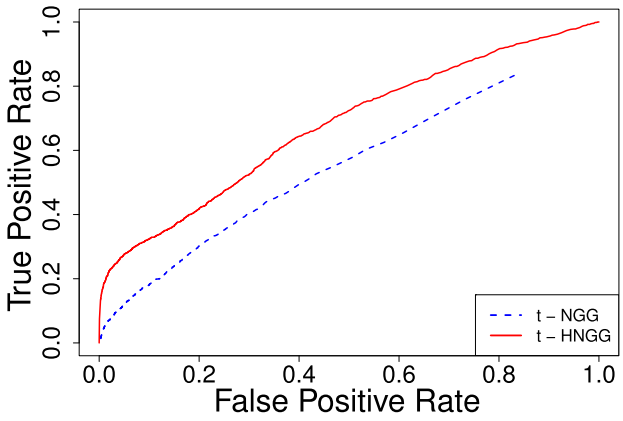


Figure 3: Simulation study, contaminated data: ROC curves, t-NGG vs t-HNGG models.

associate to $n_r n_c$ randomly selected elements of $vec(\boldsymbol{\theta})$ a divisor $\psi_m \sim \text{Unif}[0.01, 0.2]$. We repeat this process without replacement to produce 4 divisors, and set all the other elements of $vec(\boldsymbol{\theta})$ to 1. For this example, we fit the t-NGG and t-HNGG models with hyperpriors $\sigma, \sigma_0 \sim \text{Beta}(2, 18)$, where the prior expectation of the Beta distribution is equal to 0.1. The rest of the setting is unchanged from the previous simulation study.

Figure 3 shows the comparison of the ROC curves for our t-HNGG model vs the independent counterpart, the t-NGG model. Curves were computed by averaging over 50 replicated datasets. We observe a clear improvement in the t-HNGG model fitting. Figure 4 reports a summary of the posterior number of clusters as well as posterior distributions of σ and σ_0, obtained under the two models. In particular, in the independent model we show the posterior mean of the number of clusters in each data vector, while the posterior distribution of the number of clusters M is reported for model (8). As expected, the number of clusters in each data vector obtained under the t-NGG is higher than in the t-HNGG case, due to the lack of sharing of information. Furthermore, the posterior mode of the number of clusters in the hierarchical model is very close to the number of unique divisors used to simulate the data (i.e., 5 different divisors including 1). We also notice that the posterior distributions of σ and σ_0 show a clear departure from the Dirichlet process case (achieved when $\sigma = \sigma_0 = 0$), supporting the choice of the use of the NGG process as a building block for our model.

Finally, Table 2 reports the L^1, L^2, and maximum modulus distances between estimated and true precision matrices, averaged over the 50 replicates, and calculated using formulas (19). We also add comparisons with methodologies available in the literature, i.e., the Graphical-Lasso (Meinshausen and Bühlmann, 2006) and the Bayesian Graphical-Lasso (Wang, 2012). The proposed model outperforms both the standard methods and the independent t-NGG model.

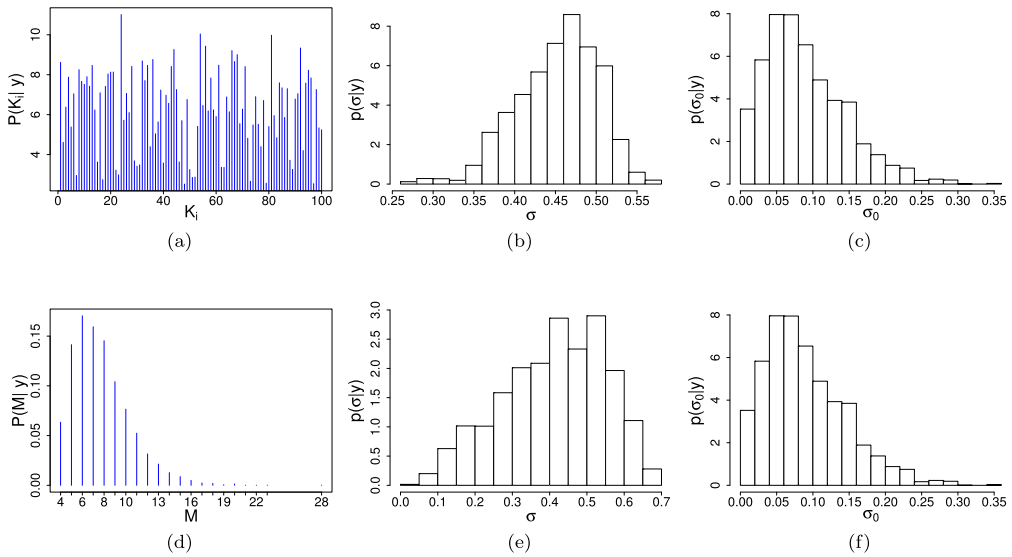


Figure 4: Simulation study, contaminated data: t-NGG vs t-HNGG models. (a) Posterior mean of the number of clusters in each data vector (t-NGG); (b,c) posterior distribution of σ and σ_0 (t-NGG); (d) posterior distribution of the number of clusters M (t-HNGG); (e,f) posterior distribution of σ and σ_0 (t-HNGG).

	G-Lasso	Bayes G-Lasso	t-NGG	t-HNGG
L^1	6.9213	6.9089	6.6516	5.6287
L^2	17.2031	18.3216	16.2151	8.6058
Max	3.2770	3.4513	3.1293	1.6143

Table 2: Simulation study, contaminated data: Average distances between $\hat{\boldsymbol{\Omega}}_{\boldsymbol{\theta}}$ and $\boldsymbol{\Omega}_{\boldsymbol{\theta}}$.

4.2 Case Study on Radiomics Features

Radiomics is the study of numerical features extracted from radiographic image data, which can be used to quantitatively summarize tumor phenotypes (Lambin et al., 2012; Gillies et al., 2016). Cellular diagnostic techniques such as biopsies are not only invasive, but they also do not allow for a thorough or complete investigation of the entire tumor environment, while manual review of images by radiologists is expensive, time-consuming, and not always consistent across raters. Quantitative imaging features mined with radiomics techniques can be used to get a more comprehensive picture of the entire lesion environment without having to take multiple biopsies or depend on qualitative visual assessments. It has been hypothesized that trends in radiomic features are reflective of complementary tumor characteristics at the molecular, cellular, and genetic levels (Aerts et al., 2014).

Although the development of novel radiomic features is an active area of research (Shoemaker et al., 2018), in this work, we focus on the so-called first and second order

features, as these are the most commonly used in practice (Gillies et al., 2016). First order features consider the collection of intensity values across all voxels in the image, without regard for their spatial orientation, and may be referred to as histogram-based or non-spatial. Examples of first order features include volume, intensity, mean, median, entropy, kurtosis. Second order features account for voxel position in addition to intensity, and are also called spatial features. Examples of second order features include eccentricity, solidity, and texture features. These features are often computed on multiple combinations of angle, distance, and number of grey levels, leading to a large set of features that can be used in model building and data analysis. However, there are challenges in the use of radiomics data for statistical modeling. Firstly, features often exhibit departures from normality due to the heterogeneity of the tumor images across patients. Secondly, they are often highly correlated. Such dependence is partially structural in nature, as the features are all calculated on the same voxel data. To date, most efforts at predictive modeling begin with filtering of the features by selecting a single representative for each cluster of highly correlated features (Gillies et al., 2016) or applying rank-based filtering methods across all features (Parmar et al., 2015) or within each class of features (Aerts et al., 2014). The screened features are then used as input to machine learning algorithms for prediction or classification such as random forests, support vector machines, or regularized regression. There is a push in the field, however, away from "black box" modeling. For example, there is an interest in establishing the genetic basis of the features (known as "radiogenomics", see Gevaert et al., 2014) and, more generally, in enhancing the interpretability of the features, models, and results obtained (Morin et al., 2018). Investigation of the relationships between features supports the search for links between radiomic features, genotypes, phenotypes, and clinical outcomes in more complex statistical models (Stingo et al., 2013) aimed at not only using imaging features for prediction, but understanding their interdependence and the genomic and clinical factors that shape them.

In this case study, we focus on glioblastoma data collected as part of The Cancer Imaging Atlas (TCIA), which provides imaging data on the same set of subjects whose clinical and genomic data are available through The Cancer Genome Atlas (TCGA). Specifically, we obtained radiomic features extracted from magnetic resonance imaging (MRI) images by Bakas et al. (2017), which made a standard set of features publicly available with the goal of providing reproducible and accessible data. This data set includes more than 700 radiomic features for 102 subjects diagnosed with glioblastoma (GBM). The features provided include intensity, volumetric, morphologic, histogram-based, and textural features, as well as spatial information and parameters extracted from a glioma growth model (Hogea et al., 2008). Each subject has scans in the MRI modalities of T1-weighted pre-contrast (T1), T1-weighted post-contrast (T1-Gd), T2, and T2-Fluid-Attenuated Inversion Recovery (FLAIR). The MRI images were segmented into the following regions: the enhancing part of the tumor core (ET), the non-enhancing part of the tumor core (NET), and the peritumoral edema (ED), and these segmentations were manually checked and approved by a neurologist.

To obtain a usable feature set for the proposed robust graphical model, we first applied a log transformation to improve symmetry and reduce the impact of outlying large values in the untransformed data. To account for negative values and to handle

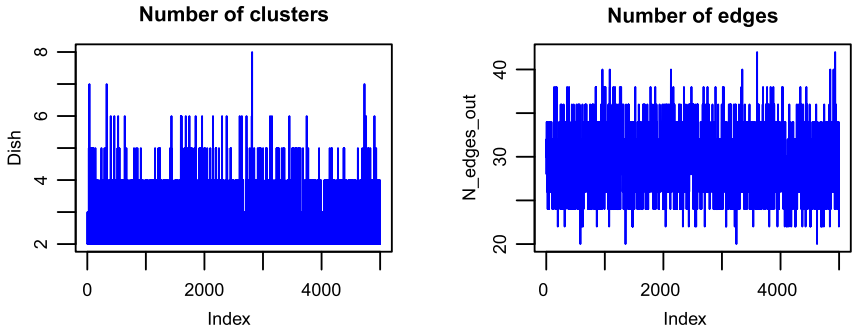


Figure 5: Case study on radiomics data: Trace plots of the parameters for the numbers of edges and the number of clusters for the t-HNGG model.

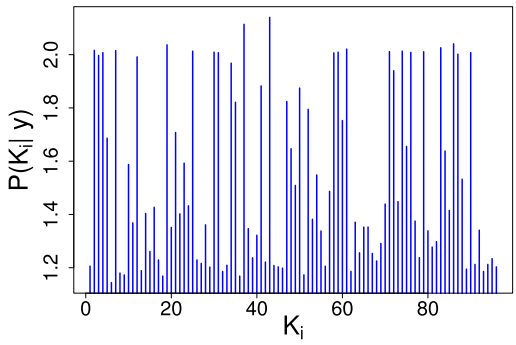
the presence of zeros, the features with negative values were shifted up by the minimum value, and 1 was added to each observation for all features before the log transformation was applied. We then assessed the pairwise correlation between all the log-transformed features. If a pair had absolute correlation greater than 0.8, we removed the feature with a higher mean absolute correlation to all other features. In order to focus on features with potential clinical importance, we obtained survival information from the TCGA database, and filtered the features to include only those with p-value ≤ 0.05 in a univariate Cox proportion hazard model for overall survival. This resulted in a set of 26 features for downstream analysis. The features that remain are fairly representative of the different types of features provided in Bakas et al. (2017), as well as from the different regions of the brain and MRI modalities. See Supplementary Materials for detailed information on these features.

Analysis

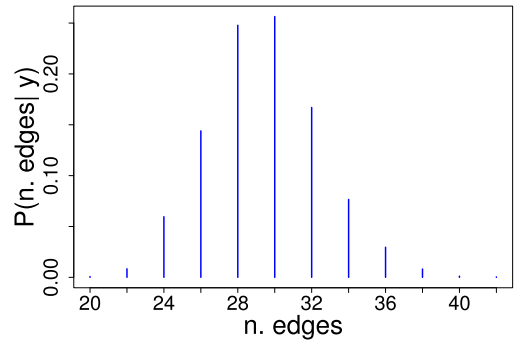
The t-HNGG model was applied to the screened features. As in the first simulation study, we set $\kappa, \kappa_0 \sim \text{gamma}(1,1)$ and $(\sigma, \sigma_0) = (0.5, 0.1)$, yielding $\mathbb{E}(M) = 8.83$ and $\text{sd}(M) = 2.66$. We ran an MCMC chain with 30,000 iterations, with 20,000 burn-in iterations and thinned by 2. The edge inclusion was determined by thresholding the posterior probability of inclusion (PPI) at 0.5, as in the median model of Barbieri et al. (2004). Following Peterson et al. (2015), we computed the Bayesian false discovery rate (FDR) for the selected model; the resulting value of 0.053 suggests that our edge selection procedure is reasonable.

To assess convergence, we applied the Geweke diagnostic criteria (Geweke et al., 1991) on four parameters: κ, κ_0, the number of clusters, and the number of edges. The test gave non-significant p-values for each of the parameters, indicating that the chains converged. The trace plots for the number of clusters and the number of edges are given in Figure 5. Summaries of the posterior means of the number of clusters in each data vector, the number of edges, and the number of clusters are given in Figure 6.

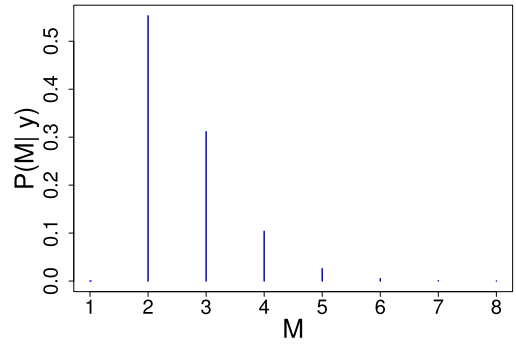
For comparison, we applied the graphical lasso (GLasso) and the Bayesian graphical



(a)



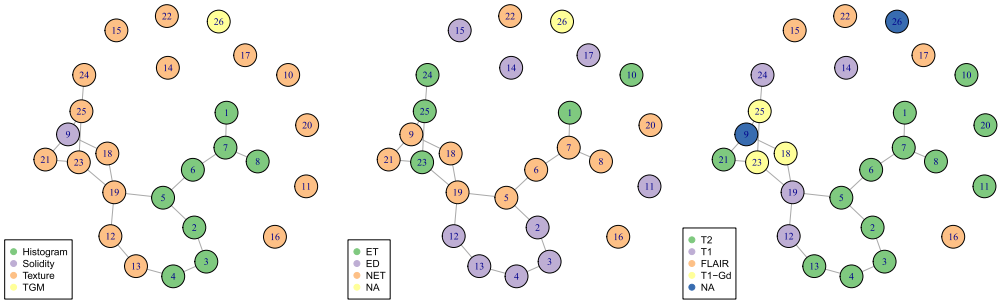
(b)



(c)

Figure 6: Case study on radiomics data: (a) Posterior mean of the number of clusters in each data vector; (b) Posterior distribution of the number of edges; (c) Posterior distribution of the number of clusters.

lasso (BGLasso) methods. The regularization parameter for the GLasso was chosen as 0.45 by minimizing the Bayesian information criterion (BIC), and the gamma prior for the regularization parameter of the BGLasso was set such that the prior mean was also 0.45, with shape = 4.5 and scale = 1/10. The BGLasso was run for 13,000 total



(a) Type (b) Region (c) Modality

Figure 7: Case study on radiomics data: The resulting graph from the t-HNGG model is depicted in plots (a)-(c). In each plot, colors are used to indicate class membership of the graph nodes, according to different characteristics of the features, i.e, (a) feature type, (b) feature region, and (c) imaging modality.

iterations, with a burn in of 3,000. The sampled precision matrices for this method are not sparse, so a threshold of 0.1 on the absolute value of the entries in the posterior mean of the precision matrix was chosen to create the adjacency matrix.

Results

The graph inferred by the proposed t-HNGG method is presented in Figure 7, with three different color schemes to indicate class membership of the nodes by feature type, feature region, and imaging modality. In this illustration, we see that features from the same type and modality are more likely to be identified as connected, while there are fewer links dictated by region: this could imply that it is more critical to have features divided over separate regions of the tumor than it is to have a large number of features or to have scans in multiple modalities, as the former is more likely to give independent information from the different features.

Regarding the comparison to other methods, GLasso and BGLasso produced very similar graphs, as is to be expected, and these had fewer edges overall than the graph inferred via the t-HNGG model, although there were a couple of connections selected under the lasso methods that were not identified in the t-HNGG model. Table 3 reports edge similarities between the three methods. In all three graphs, edges are captured that we expect to see, such as adjacent bins in various histograms, e.g., there is a connection between bin 1 and bin 2 of the histogram for the T2 modality of the NET region. The busyness features over three different modalities are connected in all three models. However, the two GLasso models only select couplets and triplets, and none of these are particularly surprising, linking together similar features that could be considered adjacent in a qualitative sense. As one would expect, reducing the PPI threshold in the t-HNGG model increases the number of selected edges, while increasing this threshold

	GL	BGL	t-HNGG
GL	7	7	5
BGL		8	6
t-HNGG			19

Table 3: Case study on radiomics data: Number of edges in each of the graphs inferred by GLasso, BGLasso and the t-HNGG model, and number of shared edges between pairs of graphs.

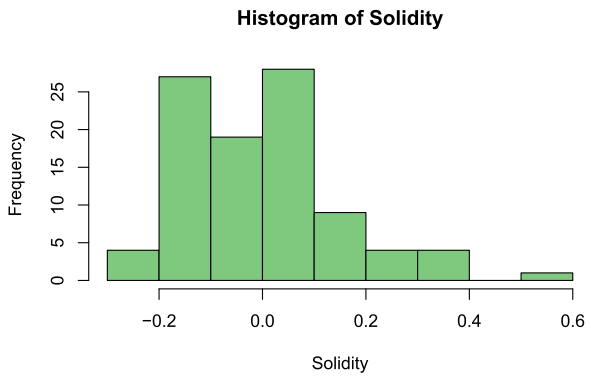


Figure 8: Case study on radiomics data: A histogram of the data for the 9th feature, solidity of the NET region.

reduces the number of selections. We found, however, that the overlapping edges between the t-HNGG and the graphs inferred with GLasso and BGLasso remained consistent across the range of PPI thresholds between 0.2 and 0.9.

An interesting edge captured by the t-HNGG model that is not captured by the other models is one between a histogram feature and a busyness texture feature. Histogram features display only the first-order information about the pixels and are not often used to infer any information about the adjacency or texture of the images. However, this particular histogram feature is of the first bin of the histogram, so this could suggest that heavier tailed pixel distributions are harbingers of busyness. The end bin of the histogram was also found to be a significant feature for glioma classification by Cho and Park (2017). Further, there are no edges in the graphs inferred by the LASSO-type models that connect the solidity feature to any other feature, unlike in the t-HNGG graph. Failure to recover edges might be attributed to the non-normal distribution shown by this feature, as it can be seen in Figure 8, showing once again the power of the t-HNGG model to handle outliers. Edges and dependencies, and lack thereof, can be used to inform more complex models for classification and characterization, to inform radiologists and clinicians as they begin to utilize radiomics, and to enhance the general interpretation as statisticians move away from the "black box models" often used on these complicated feature sets.

5 Conclusion

In this paper, we have proposed a class of robust Bayesian graphical models based on a nonparametric hierarchical prior construction that allows for flexible deviations from Gaussianity in the distribution of the data. The proposed model is an extension of the t-Dirichlet model presented in Finegold and Drton (2014), where departure from Gaussianity is accounted for by including suitable latent variables (divisors) in the sampling model, to allow for skewness. In our proposed construction, the law of the divisors is described by a hierarchical normalised completely random measure. In particular, in this paper we have focused on a hierarchical NGG process, yielding to what we have called a t-HNGG model. The advantage of this choice is twofold: on one side, by extending the characterization to the NGG process, we induce a more flexible clustering structure when compared to the Dirichlet process case; on the other side, by allowing for an additional level of hierarchy in the nonparametric prior setting, we achieve sharing of information across the data sample. For posterior inference, we have implemented a suitable MCMC algorithm, which is built upon the generalized Chinese restaurant franchise metaphor to exploit dependency among the components of each data vector (i.e,. customers seated in the same restaurant).

We have illustrated performances of our proposed methodology on simulated data and on a case study on numerical features extracted from radiographic image data which can be used to quantitatively summarize tumor phenotypes, and that are known to show non-Gaussian characteristics. On simulated data, we have shown good recovery of the main features of the data, such as the graph structure and the precision matrix. Additionally, a comparison with existing methodologies such as the GLasso and the Bayesian GLasso has shown how these methods are outperformed by our proposed model in the presence of non-Gaussian data. On the real data, our model has resulted in a less sparse graph than those inferred by GLasso and Bayesian GLasso. Furthermore, the inferred relationships highlighted by our estimated graph have revealed interesting interpretation in terms of important characteristics of the data. These relationships and dependencies, and lack thereof, can provide valuable information for follow-up classifications and characterization of radiomics data.

Acknowledgment

Andrea Cremaschi was supported by the University of Oslo (UiO). Marina Vannucci and Christine B. Peterson were partially supported by NSF/DMS 1811568 and NSF/DMS 1811445. Katherine Shoemaker was partially supported by NIH grant T32 - CA09652.

Supplementary Material

Supplementary Material for "Hierarchical Normalized Completely Random Measures for Robust Graphical Modeling"
(DOI: 10.1214/19-BA1153SUPP; .pdf). We include in this file additional theoretical justifications, details of the MCMC updates, as well as some additional results from the

applications presented in the paper. Details on the features analysed in the radiomics case study are reported.

References

Aerts, H. J., Velazquez, E. R., Leijenaar, R. T., Parmar, C., Grossmann, P., Cavalho, S., Bussink, J., Monshouwer, R., Haibe-Kains, B., Rietveld, D., et al. (2014). "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach." *Nature Communications*, 5. 1289, 1290

Argiento, R., Bianchini, I., and Guglielmi, A. (2016). "A blocked Gibbs sampler for NGG-mixture models via a priori truncation." *Statistics and Computing*, 26(3): 641–661. MR3489862. doi: https://doi.org/10.1007/s11222-015-9549-6. 1272

Argiento, R., Cremaschi, A., and Vannucci, M. (2019). "Hierarchical Normalized Completely Random Measures to Cluster Grouped Data." *Journal of the American Statistical Association*. doi: https://doi.org/10.1080/01621459.2019.1594833. 1272, 1276, 1277, 1278, 1280, 1284

Argiento, R., Guglielmi, A., Hsiao, C. K., Ruggeri, F., and Wang, C. (2015). "Modeling the association between clusters of SNPs and disease responses." In *Nonparametric Bayesian Inference in Biostatistics*, 115–134. Springer. MR3382181. 1276

Argiento, R., Guglielmi, A., and Pievatolo, A. (2010). "Bayesian density estimation and model selection using nonparametric hierarchical mixtures." *Computational Statistics and data Analysis*, 54: 816–832. MR2580918. doi: https://doi.org/10.1016/j.csda.2009.11.002. 1272

Atay-Kayis, A. and Massam, H. (2005). "A Monte Carlo method for computing the marginal likelihood in nondecomposable Gaussian graphical models." *Biometrika*, 92(2): 317–335. MR2201362. doi: https://doi.org/10.1093/biomet/92.2.317. 1274

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., Freymann, J. B., Farahani, K., and Davatzikos, C. (2017). "Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features." *Scientific Data*, 4: 170117.
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5685212/ 1290, 1291

Barbieri, M. M., Berger, J. O., et al. (2004). "Optimal predictive model selection." *The Annals of Statistics*, 32(3): 870–897. MR2065192. doi: https://doi.org/10.1214/009053604000000238. 1283, 1291

Barrios, E., Lijoi, A., Nieto-Barajas, L. E., and Prünster, I. (2013). "Modeling with normalized random measure mixture models." *Statistical Science*, 28: 313–334. MR3135535. doi: https://doi.org/10.1214/13-STS416. 1272

Bhadra, A., Rao, A., and Baladandayuthapani, V. (2018). "Inferring network structure in non-normal and mixed discrete-continuous genomic data." *Biometrics*, 74(1): 185–195. MR3777939. doi: https://doi.org/10.1111/biom.12711. 1272

Camerlenghi, F., Lijoi, A., Orbanz, P., and Prünster, I. (2019). "Distribution theory for hierarchical processes." *Annals of Statistics*, 47(1): 67–92. MR3909927. doi: https://doi.org/10.1214/17-AOS1678. 1272, 1277, 1278

Camerlenghi, F., Lijoi, A., and Prünster, I. (2017). "Bayesian prediction with multiple-samples information." *Journal of Multivariate Analysis*, 156: 18–28. MR3624682. doi: https://doi.org/10.1016/j.jmva.2017.01.010. 1272

Camerlenghi, F., Lijoi, A., and Prünster, I. (2018). "Bayesian nonparametric inference beyond the Gibbs-type framework." *Scandinavian Journal of Statistics*, 45(4): 1062–1091. MR3884900. 1272

Caron, F. and Fox, E. B. (2017). "Sparse graphs using exchangeable random measures." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(5): 1295–1366. MR3731666. doi: https://doi.org/10.1111/rssb.12233. 1272

Cho, H. and Park, H. (2017). "Classification of low-grade and high-grade glioma using multi-modal image radiomics features." In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 3081–3084. 1294

Cremaschi, A., Argiento, R., Shoemaker, K., Peterson, C., Vannucci, M. (2019). "Supplementary Material for "Hierarchical Normalized Completely Random Measures for Robust Graphical Modeling"." *Bayesian Analysis*. doi: https://doi.org/10.1214/19-BA1153. 1280

De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I., and Ruggiero, M. (2015). "Are Gibbs-type priors the most natural generalization of the Dirichlet process?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2): 212–229. 1276

Dempster, A. (1972). "Covariance selection." *Biometrics*, 28: 157–175. 1272

Dobra, A., Hans, C., Jones, B., Nevins, J. R., Yao, G., and West, M. (2004). "Sparse graphical models for exploring gene expression data." *Journal of Multivariate Analysis*, 90(1): 196–212. MR2064941. doi: https://doi.org/10.1016/j.jmva.2004.02.009. 1272, 1274

Dobra, A., Lenkoski, A., and Rodriguez, A. (2011). "Bayesian Inference for General Gaussian Graphical Models With Application to Multivariate Lattice Data." *Journal of the American Statistical Association*, 106(496): 1418–1433. MR2896846. doi: https://doi.org/10.1198/jasa.2011.tm10465. 1272, 1274

Favaro, S. and Teh, Y. (2013). "MCMC for Normalized Random Measure Mixture Models." *Statistical Science*, 28(3): 335–359. MR3135536. doi: https://doi.org/10.1214/13-STS422. 1282

Finegold, M. and Drton, M. (2011). "Robust graphical modeling of gene networks using classical and alternative t-distributions." *The Annals of Applied Statistics*, 1057–1080. MR2840186. doi: https://doi.org/10.1214/10-AOAS410. 1272, 1274, 1275, 1282

Finegold, M. and Drton, M. (2014). "Robust Bayesian Graphical Modeling Using Dirichlet t-Distributions." *Bayesian Analysis*, 9(3): 521–550. MR3256052. doi: https://doi.org/10.1214/13-BA856. 1271, 1272, 1274, 1275, 1283, 1295

Friedman, J., Hastie, T., and Tibshirani, R. (2008). "Sparse inverse covariance estimation with the graphical lasso." *Biostatistics*, 9(3): 432–441. 1272

Friedman, N. (2004). "Inferring cellular networks using probabilistic graphical models." *Science*, 303(5659): 799–805. 1272

Gevaert, O., Mitchell, L. A., Achrol, A. S., Xu, J., Echegaray, S., Steinberg, G. K., Cheshier, S. H., Napel, S., Zaharchuk, G., and Plevritis, S. K. (2014). "Glioblastoma multiforme: exploratory radiogenomic analysis by using quantitative image features." *Radiology*, 273(1): 168–174. 1290

Geweke, J. et al. (1991). *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*, volume 196. Federal Reserve Bank of Minneapolis, Research Department Minneapolis, MN, USA. 1291

Gillies, R. J., Kinahan, P. E., and Hricak, H. (2016). "Radiomics: Images Are More than Pictures, They Are Data." *Radiology*, 278(2): 563–577. 1289, 1290

Giudici, P. and Green, P. J. (1999). "Decomposable graphical Gaussian model determination." *Biometrika*, 86(4): 785–801. MR1741977. doi: https://doi.org/10.1093/biomet/86.4.785. 1281

Griffin, J. E. and Stephens, D. A. (2013). "Advances in Markov chain Monte Carlo." *Bayesian Theory and Applications*, 104–144. MR3221161. 1282

Hogea, C., Davatzikos, C., and Biros, G. (2008). "An image-driven parameter estimation problem for a reaction-diffusion glioma growth model with mass effects." *Journal of Mathematical Biology*, 56(6): 793–825. MR2385684. doi: https://doi.org/10.1007/s00285-007-0139-x. 1290

Ishwaran, H. and James, L. F. (2003). "Generalized weighted Chinese restaurant processes for species sampling mixture models." *Statistica Sinica*, 1211–1235. MR2026070. 1276

James, L., Lijoi, A., and Prünster, I. (2009). "Posterior analysis for normalized random measures with independent increments." *Scandinavian Journal of Statistics*, 36: 76–97. MR2508332. doi: https://doi.org/10.1111/j.1467-9469.2008.00609.x. 1272, 1279

Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C., and West, M. (2005). "Experiments in stochastic computation for high-dimensional graphical models." *Statistical Science*, 388–400. MR2210226. doi: https://doi.org/10.1214/088342305000000304. 1272, 1274

Lambin, P., Rios Velazquez, E., Leijenaar, R., Carvalho, S., van Stiphout, R. G., et al. (2012). "Radiomics: Extracting more information from medical images using advanced feature analysis." *European Journal of Cancer*, 48(4): 441–446. 1289

Lauritzen, S. (1996). *Graphical Models*. Clarendon Press (Oxford and New York). MR1419991. 1271

Lenkoski, A. (2013). "A direct sampler for G-Wishart variates." *Stat*, 2: 119–128. 1274, 1281

Lenkoski, A. and Dobra, A. (2011). "Computational aspects related to inference in Gaussian graphical models with the G-Wishart prior." *Journal of Computational and Graphical Statistics*, 20(1): 140–157. MR2816542. doi: https://doi.org/10.1198/jcgs.2010.08181. 1274

Lijoi, A., Mena, R. H., and Prünster, I. (2007). "Controlling the reinforcement in Bayesian nonparametric mixture models." *Journal of the Royal Statistical Society B*, 69: 715–740. MR2370077. doi: https://doi.org/10.1111/j.1467-9868.2007.00609.x. 1272, 1276

Lijoi, A. and Prünster, I. (2010). "Models beyond the Dirichlet process." In Hjort, N., Holmes, C., Müller, P., and Walker (eds.), *In Bayesian Nonparametrics*, 80–136. Cambridge University Press. MR2730661. 1272, 1279

Meinshausen, N. and Bühlmann, P. (2006). "High-dimensional graphs and variable selection with the lasso." *Annals of Statistics*, 34(3): 1436–1462. MR2278363. doi: https://doi.org/10.1214/009053606000000281. 1272, 1288

Mohammadi, A. and Wit, E. C. (2015). "Bayesian structure learning in sparse Gaussian graphical models." *Bayesian Analysis*, 10(1): 109–138. MR3420899. doi: https://doi.org/10.1214/14-BA889. 1274, 1281

Morin, O., Vallières, M., Jochems, A., Woodruff, H. C., Valdes, G., Braunstein, S. E., Wildberger, J. E., Villanueva-Meyer, J. E., Kearney, V., Yom, S. S., Solberg, T. D., and Lambin, P. (2018). "A Deep Look into the Future of Quantitative Imaging in Oncology: A Statement of Working Principles and Proposal for Change." *International Journal of Radiation Oncology*Biology*Physics*. 1290

Mukherjee, S. and Speed, T. (2008). "Network inference using informative priors." *Proceedings of the National Academy of Sciences of the United States of America*, 105(38): 14313–14318. 1272

Neal, R. (2000). "Markov Chain sampling Methods for Dirichlet process mixture models." *Journal of Computational and Graphical Statistics*, 9: 249–265. MR1823804. doi: https://doi.org/10.2307/1390653. 1282

Parmar, P., C.and Grossmann, Bussink, J., Lambin, P., and Aerts, H. J. (2015). "Machine learning methods for quantitative radiomic biomarkers." *Scientific Reports*, 5: 13087. 1290

Peterson, C., Stingo, F., and Vannucci, M. (2016). "Joint Bayesian variable and graph selection for regression models with network-structured predictors." *Statistics in Medicine*, 35(7): 1017–1031. MR3476525. doi: https://doi.org/10.1002/sim.6792. 1272

Peterson, C., Stingo, F. C., and Vannucci, M. (2015). "Bayesian Inference of Multi-

ple Gaussian Graphical Models." *Journal of the American Statistical Association*, 110(509): 159–174. PMID: 26078481. MR3338494. doi: https://doi.org/10.1080/01621459.2014.896806. 1291

Peterson, C., Vannucci, M., Karakas, C., Choi, W., Ma, L., and Maletić-Savatić, M. (2013). "Inferring metabolic networks using the Bayesian adaptive graphical lasso with informative priors." *Statistics and Its Interface*, 6(4): 547–558. MR3164658. doi: https://doi.org/10.4310/SII.2013.v6.n4.a12. 1272

Pitman, J. (1996). "Some developments of the Blackwell-MacQueen urn scheme." *Lecture Notes-Monograph Series*, 245–267. MR1481784. doi: https://doi.org/10.1214/lnms/1215453576. 1276

Pitman, J. (2003). "Poisson-Kingman Partitions." In *Science and Statistics: a Festschrift for Terry Speed*, volume 40 of *IMS Lecture Notes-Monograph Series*, 1–34. Hayward (USA): Institute of Mathematical Statistics. MR2004330. doi: https://doi.org/10.1214/lnms/1215091133. 1276

Pitt, M., Chan, D., and Kohn, R. (2006). "Efficient Bayesian inference for Gaussian copula regression models." *Biometrika*, 93(3): 537–554. MR2261441. doi: https://doi.org/10.1093/biomet/93.3.537. 1272

Regazzini, E., Lijoi, A., and Prünster, I. (2003). "Distributional results for means of random measures with independent increments." *The Annals of Statistics*, 31: 560–585. MR1983542. doi: https://doi.org/10.1214/aos/1051027881. 1272

Roverato, A. (2002). "Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models." *Scandinavian Journal of Statistics*, 29(3): 391–411. MR1925566. doi: https://doi.org/10.1111/1467-9469.00297. 1272, 1273

Shoemaker, K., Hobbs, B. P., Bharath, K., Ng, C. S., and Baladandayuthapani, V. (2018). "Tree-based methods for characterizing tumor density heterogeneity." In *Pacific Symposium on Biocomputing*, volume 23, 216–227. World Scientific. 1289

Stingo, F., Chen, Y., Vannucci, M., Barrier, M., and Mirkes, P. (2010). "A Bayesian graphical modeling approach to microRNA regulatory network inference." *Annals of Applied Statistics*, 4(4): 2024–2048. MR2829945. doi: https://doi.org/10.1214/10-AOAS360. 1272

Stingo, F. C., Guindani, M., Vannucci, M., and Calhoun, V. D. (2013). "An integrative Bayesian modeling approach to imaging genetics." *Journal of the American Statistical Association*, 108(503): 876–891. MR3174670. doi: https://doi.org/10.1080/01621459.2013.804409. 1290

Telesca, D., Müller, P., Kornblau, S., Suchard, M., and Ji, Y. (2012). "Modeling protein expression and protein signaling pathways." *Journal of the American Statistical Association*, 107(500): 1372–1384. MR3036401. doi: https://doi.org/10.1080/01621459.2012.706121. 1272

Wang, H. (2012). "Bayesian graphical lasso models and efficient posterior computa-

tion." *Bayesian Analysis*, 7(2): 771–790. MR3000017. doi: https://doi.org/10.1214/12-BA729. 1272, 1288

Wang, H. and Li, S. (2012). "Efficient Gaussian graphical model determination under G-Wishart prior distributions." *Electronic Journal of Statistics*, 6: 168–198. MR2879676. doi: https://doi.org/10.1214/12-EJS669. 1274

Yuan, M. and Lin, Y. (2007). "Model selection and estimation in the Gaussian graphical model." *Biometrika*, 94(1): 19–35. MR2367824. doi: https://doi.org/10.1093/biomet/asm018. 1272