A hierarchical nonparametric approach for robust graphical modelling

Raffaele Argiento[1] & Andrea Cremaschi[2]

[1] *School of Mathematics, Statistics and Actuarial Sciences Cornwallis Building, University of Kent, Canterbury, CT2 7NF (UK) and Istituto di Matematica Applicata e Tecnologie Informatiche "E. Magenes" (CNR-IMATI), 10133 Milan (IT)*
E-mail raffaele@mi.imati.cnr.it

[2] *School of Mathematics, Statistics and Actuarial Sciences Cornwallis Building, University of Kent, Canterbury, CT2 7NF (UK)*
E-mail ac429@kent.ac.uk

Résumé. Les modéles gaussiens graphiques sont des outils utiles pour explorer les structures des rèseaux multivariées. Cependant, des modèles alternatifs sont nécessaires quand les données sont fortement non gaussiennes. La t-distribution, obtenue en divisant chaque composante du vecteur des données par une variable aléatoire gamma, est une généralisation simple pour gérer une telle situation. La t-distribution Dirichlet, introduite par Finegold et Drton (2013), est obtenue quand la loi des diviseurs est un processus Dirichlet. Dans cette dernière, sous la condition d'un paramètre de masse partagé, un processus de Dirichlet est introduit pour chaque observation, de telle sorte que l'on peut regrouper les composantes de chaque donnée par rapport à leur déviation de la distribution normale (regroupement de données aberrantes).

Dans ce travail nous considérons une classe plus générale de distributions non paramétriques, la classe des mesures complétement alèatoires normalisées (NCRM), qui permet un regroupement des composantes plus flexible. De plus, pour emprunter plus d'informations parmi les données, on modélise la dépendance parmi les NCRM via une structure hiérarchique non-paramétrique. Au niveau des données chaque NCRM est centrée sur la même mesure de base, qui est elle-même une NCRM. La nature discréte de la mesure de base partagée implique que les processus au niveau des données partagent les mêmes atomes. Cette caractéristique souhaitée permet de regrouper ensemble des composantes de différentes données.

En guise d'illustration, nous décrirons des applications de notre modèle pour simuler des donnés multivariées. De plus, pour estimer la dépendance entre les mesures d'un indice acoustiques prises à différents niveaux de fréquence, nous utiliserons des données provenant de l'étude de l'isolation sonore d'une façade.

Mots-clés. Modéles graphiques, inférence bayésienne non paramétrique, modéles, t-distribution.

Abstract. Useful tools for exploring multivariate network structures are Gaussian graphical models. However, alternative models are needed when data are strongly non-Gaussian. The t-distribution, obtained by dividing each component of the data vector

by a gamma random variable, is the straightforward generalisation to accommodate such issue. The Dirichlet t-distribution, introduced by Finegold and Drton (2014), is obtained when the law of the divisors is the Dirichlet process. In the latter, conditionally to a shared mass parameter, a Dirichlet process is introduced for every observation, so that one can cluster the components of each data point according to their deviation from the Normal distribution (outlier clustering).

In this work, we consider a more general class of nonparametric distributions, namely the class of normalised completely random measures (NCRM), which yields a more flexible component clustering. Moreover, in order to borrow more information across the data, we model the dependence among the NCRM through a nonparametric hierarchical structure. At data level each NCRM is centered on the same base measure, which is a NCRM itself. The discreteness of the shared base measure implies that the processes at data level share the same atoms. This desired feature allows to cluster together components of different data.

For illustrative purposes, we will describe applications of our model to simulated multivariate data. Moreover, in order to estimate the dependence between measurements of an acoustic index taken at different levels of frequency, we will use data from a façade sound insulation study.

Keywords. Graphical models, Bayesian nonparametric inference, inference, hierarchical models, t-distribution.

1 Graphical Models

Graphical models provide useful tools to study multivariate network structures. In the following, first we define mathematically undirected graphs, then we mention in which sense they identify the dependency structure in a multivariate Gaussian vector.

An undirected graph is a pair $G = (V, E)$, where $V = \{1, \dots, p\}$ denotes the set of vertices and $E \subset V \times V$ is the set of edges (also called adjacency matrix). The pair of vertices (i, j) belongs to the graph G if $E_{ij} = 1$, and $E_{ij} = 0$ otherwise. In an undirected graph $E_{ij} = 0$, for $i = j$.

In this work, for computational reasons, we restrict the analysis to the class of decomposable undirected graphs (see Lauritzen, 1996).

Let $\boldsymbol{Y} \in \mathbb{R}^p$ be the population variable and assume it being normally distributed with mean vector $\boldsymbol{\mu}$ and non-singular symmetric positive-definite covariance matrix Σ, i.e., $Y \sim \mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$. It is well known that the precision matrix $\Sigma^{-1} = \Omega = (\omega_{ij})_{i,j=1,\dots,p}$ is such that $\omega_{ij} = 0$ if and only if the two components Y_i and Y_j of $\boldsymbol{Y}$ are, conditionally to the rest of the data vector, independent. Dempster (1972) firstly introduced a covariance selection model where the adjacency matrix E of an undirected graph G is used to describe the structure of the precision matrix Ω of a multivariate normal variate $\boldsymbol{Y}$. In a nutshell, the undirected graph G identifies the network dependency structure so that ω_{ij} is equal to

zero for all pairs $(i, j) \notin E$.

The precision matrix Ω belongs to $M(G)$, the set of all symmetric positive-definite matrices satisfying the graph structure in E. Dawid and Lauritzen (1993) firstly attempted a Bayesian analysis of the covariance selection problem, by introducing a suitable prior for the covariance matrix Σ such that $\Omega \in M(G)$. They named such distribution *Hyper Inverse Wishart*, in short:

$$\Sigma \sim HIW_G(b, D),$$

where b and D represent the degrees of freedom and the location matrix, respectively.

We are now ready to write down the standard graphical model in Bayesian parametric setting (see Giudici and Green, 1999):

$$\begin{aligned} &\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n | \boldsymbol{\mu}, \Sigma, G \overset{iid}{\sim} \mathcal{N}_p(\boldsymbol{\mu}, \Sigma) \\ &\boldsymbol{\mu} | \Sigma, G \sim \mathcal{N}_p(\boldsymbol{\mu}_0, \Sigma / n_0) \\ &\Sigma | G \sim HIW_G(b, D) \\ &G \sim \pi(G). \end{aligned} \tag{1}$$

The last ingredient to fully specify the model is the prior for the undirected decomposable graph G, $\pi(G)$. Giudici and Green (1999) proposed a uniform prior over the set of decomposable graphs of dimension p, and firstly introduced a reversible jump MCMC algorithm to perform posterior inference under the parametric model 1. For a review of possible prior choices for G, we refer to Armstrong et al. (2009).

2 Robust Bayesian Graphical Modeling

In their recent works, Finegold and Drton (2013, 2014), consider data that are strongly non-normal, and propose an alternative approach for network inference in both Bayesian and classical setting. Shortly, their model for the population variable $\boldsymbol{Y} \in \mathbb{R}^p$ can be described as follows:

$$Y_j = \mu_j + \frac{X_j}{\sqrt{\tau_j}} \quad j = 1, \ldots, p; \quad \boldsymbol{X} = (X_1, \ldots, X_p) \sim N(0, \Sigma).$$

Where three cases are considered. Let $P_0 = gamma(\nu/2, \nu/2)$

1. $\tau_1 = \tau_2 = \cdots = \tau_p$, $\tau_1 \sim P_0$. $\boldsymbol{Y} \sim t_{p,\nu}(\mu, \Sigma)$ *multivariate t-distribution*;

2. $\tau_1, \tau_2, \ldots, \tau_p \overset{\text{i.i.d.}}{\sim} P_0$. $\boldsymbol{Y} \sim t^*_{p,\nu}(\mu, \Sigma)$ *alternative t-distribution*;

3. $\tau_1, \tau_2, \ldots, \tau_p | P \overset{\text{i.i.d.}}{\sim} P$, $P \sim \text{DP}(\kappa, P_0)$. $\boldsymbol{Y} \sim t^{\kappa}_{p,\nu}(\mu, \Sigma)$ *Dirichlet t-model.*

Here, DP stands for the Dirichlet Process with mass parameter κ and centering measure P_0 (Ferguson, 1973).

At date, the literature of Bayesian nonparametric graphical models is not abundant. Here we mention the work of Rodriguez et al. (2011) , where they address the situation of data coming from heterogeneous populations. Interestingly, in the work of Finegold and Drton (2014), a nonparametric approach is used to explore models lying between the two extreme cases of the multivariate and alternative t-distributions.

3 Background on normalized completely random measures

The class of normalized completely random measures (NCRM) is a wide class of almost sure discrete random probability measures recently introduced in the literature by Regazzini et al. (2003), with the name of normalized random measures with independent increments. This class has been extensively studied in recent years and it has proven to be an effective building block in nonparametric mixture models (see for instance Argiento et al. 2015). It is well known that a NCRM P can be represented as

$$P = \sum_{i=1}^{+\infty} \xi_i \delta_{\tau_i} = \sum_{i=1}^{+\infty} \frac{J_i}{T} \delta_{\tau_i} \tag{2}$$

where $\xi_i := J_i/T$, $(J_i)_i$ are the points of a Poisson process on $\mathbb{R}^+$ with mean intensity $\rho(s)ds$ and $T = \sum_i J_i$. The random variables τ_i are independent from $\{J_i\}$, and τ_i's are i.i.d. from P_0. The function $\rho(\cdot)$ is called *Levy intensity* of the process, and it characterizes its distribution and must satisfy the regularity conditions

$$\int_0^{+\infty} \min\{1, s\}\rho(s)ds < \infty \quad \text{and} \quad \int_0^{+\infty} \rho(s)ds = +\infty, \tag{3}$$

so that the normalization is well defined, since $\mathbb{P}(0 < T < +\infty) = 1$. A remarkable family of NCRM is the class of normalized generalized gamma processes, obtained when

$$\rho(s) = (\kappa/\Gamma(1-\sigma)s^{-1-\sigma}e^{-s}\mathbb{I}_{(0,+\infty)}(s)ds.$$

We write $P \sim NGG(\sigma, \kappa, P_0)$, where (σ, κ, P_0) are the parameters of the NGG-process (see Argiento et al. 2010) for more details. This class encompasses the Dirichlet processes when $\sigma = 0$ and $\kappa > 0$. On the other hand, when $\sigma = 1/2$, P reduces to the normalized inverse-Gaussian process (see Argiento et al. 2009).

One of the main arguments in favour of NCRM's and in particular of the NGG processes, when compared with DP's, is a higher flexibility in clustering. For instance, when considering a sample of size n from a NGG process, the distribution of the number K_n of

distinct values in the sample has a further degree of freedom, σ, which tunes its variance, unlike the DP case where the distribution of K_n can be highly peaked. The parameter σ also drives a richer reinforcement mechanism in the predictive distributions of the sample. Moreover, NGG processes are of Gibbs-type, a class of random probabilities which stands out for their mathematical tractability, see De Blasi et al. (2015).

4 A Hierarchical NGG Model

In this work we are going to consider the following hierarchical model:

$$
\begin{aligned}
&\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n | (\boldsymbol{\tau}_i), \boldsymbol{\mu}, \Sigma, G \overset{ind.}{\sim} \mathcal{N}_p \left(\boldsymbol{\mu}, \operatorname{diag}(1/\sqrt{\boldsymbol{\tau}_i}) \cdot \Sigma \cdot \operatorname{diag}(1/\sqrt{\boldsymbol{\tau}_i}) \right) \\
&(\tau_{ij} | P_i) \overset{\text{i.i.d.}}{\sim} P_i, \;\; j = 1, \ldots, p \\
&(P_i) | P \overset{\text{i.i.d.}}{\sim} P, \qquad P \sim NGG(\sigma, \kappa, P) \\
&P \sim NGG(\sigma_0, \kappa_0, P_0), \quad P_0 = gamma(\nu/2, \nu/2) \\
&\boldsymbol{\mu} \sim N_p(\mathbf{0}, \sigma_\mu^2 I_p), \quad \Sigma \sim HIW_G(b, D), \quad G \sim \pi(G).
\end{aligned}
$$

This is a generalization of the *Dirichlet t-distribution* in two directions. Firstly, for each data point (i.e., for each $i = 1, \ldots, n$) we model the divisors $\tau_{ij}, \; i = 1, \ldots, p$ as a random sample from a NGG process, so that clustering between the component of the jth sample is more flexible. Secondly, in order to borrow more information across the data, we model the dependence among the NGG's through a nonparametric hierarchical structure. At data level each NGG is centered on the same base measure, which is a NGG itself. The discreteness of the shared base measure implies that the processes at data level share the same atoms.

5 Application

The data considered in this work consist of forty measurements of an acoustic index, measuring the sound insulation of a residential building. Each measurement is a vector of observed values at 21 one-third/octave frequency bands. It is well known that measurements are more variable at low frequencies. Nevertheless, previous studies have shown that correlation between higher and lower frequencies exists. Our model is capable of identifying different frequencies according to the divisors τ_i, while at the same time recovering the network structure via the estimation of the graph G.

Bibliographie

[1] Argiento, R., Guglielmi, A., Hsiao, C.K., Ruggeri, F., Wang C. (2015). *Modelling the association between clusters of SNPs and disease responses.* in Nonparametric Bayesian Methods in Biostatistics and Bioinformatics (R. Mitra, P. Mueller Eds.), Springer.
[2] Argiento, R., Guglielmi, A., Pievatolo, A. (2010) *Bayesian density estimation and model selection using nonparametric hierarchical mixtures.* Computational Statistics & Data Analysis, Volume 54, Issue 4, pp. 816-832.
[3] Argiento, R., Guglielmi, A., Pievatolo, A. (2009) *A comparison of nonparametric priors in hierarchical mixture modelling of lifetime data.* Journal of Statistical Planning and Inference (2009) Volume 139, Issue 12, pp. 3989-4005.
[4] Armstrong, H., Carter, C. K., Wong, K. F. K. et Kohn, R. (2009), *Bayesian covariance matrix estimation using a mixture of decomposable graphical models*, Statistics and Computing, 19(3), 303–316.
[5] Dawid, A. P. et Lauritzen, L. S. (1993), *Hyper Markov Laws in the Statistical Analysis of Decomposable Graphical Models*, The Annals of Statistics, 21(3), 1272–1317.
[6] De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prunster, I., & Ruggiero, M. (2015). *Are Gibbs-type priors the most natural generalization of the Dirichlet process?.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, 37(2), 212-229.
[8] Dempster, A. P. (1972), *Covariance selection*, Biometrics, 28(1), 157–175.
[9] Ferguson, T. S. (1973), *A Bayesian analysis of some nonparametric problems*, The annals of statistics, 1(2), 209–230.
[10] Finegold, M. et Drton, M. (2011), *Robust Graphical Modeling of Gene Networks using Classical and Alternative t-Distributions*, The Annals of Applied Statistics, 5(2), 1057–1080.
[11] Finegold, M. et Drton, M. (2014), *Robust Bayesian Graphical Modeling Using Dirichlet t-Distributions*, Bayesian Analysis, 9(3), 521–550.
[12] Giudici, P. et Green, A. P. (1999), *Decomposable graphical Gaussian model determination*, Biometrika, 86(4), 785–801.
[13] Lauritzen, S. L. (1996), *Graphical models*, Clarendon Press.
[14] Regazzini, E., Lijoi, A.,& Prnster, I. (2003). *Distributional results for means of normalized random measures with independent increments.* Annals of Statistics, 560-585.
[15] Rodriguez, A., Lenkoski, A., & Dobra, A. (2011). *Sparse covariance estimation in heterogeneous samples.* Electronic Journal of Statistics, 5, 981–1014.