Developments in Bayesian Nonparametrics: Discussion

Federico Camerlenghi

April 21, 2021

University of Milano – Bicocca & Collegio Carlo Alberto

OUTLINE

SPARSE SPATIAL RANDOM GRAPHS

SUMMARY: SPARSE SPATIAL RANDOM GRAPHS

Network models are tailored for different applied problems:

- social networks: to describe friendship between individuals;
- email communication;
- biological networks: interactions between proteins.

A network may be represented as a set of nodes (individuals) and edges (interactions).

EXCHANGEABLE RANDOM MEASURES

According to (Caron & Fox; 2017), a graph is represented as a point process

$$Z := \sum_{i \geq 1} \sum_{j \geq 1} z_{i,j} \delta_{(\theta_i, \theta_j)}$$

- $\theta_i \in \mathbb{R}_+$: is the label of node i, i.e., time of appearance of the node;
- $z_{i,j}$: edge between node i and j, with $z_{i,j} = 1$ if the nodes are connected, 0 otherwise.

Z is typically assumed to be joint exchangeable, i.e.,

$$Z(A_i \times A_j) \stackrel{d}{=} Z(A_{\pi(i)} \times A_{\pi(j)}), \quad i, j \in \mathbb{N}$$

for any permutation π of $\mathbb{N}$ and any interval $A_i = [h(i-1), hi]$, $h > 0$.

Panero, Caron & Rousseau (2021) would like to:

- include covariate variables (age, job, gender) in the network to better describe the interactions;
- describe both dense and sparse graphs.

Spatial random graphs

Panero, Caron & Rousseau (2021) deal with the following:

$$Z := \sum_{i\geq 1}\sum_{j\geq 1} z_{i,j}\delta_{(\theta_i,\theta_j,x_i,x_j)}$$

- $\theta_i, z_{i,j}$ represent again nodes and edges;
- the probability of connection between nodes depends on $\vartheta_i = (x_i, w_i)$, where w_i is a sociability parameter.

Z has the following two properties:

- joint exchangeable with respect to the label coordinates θ_i;
- isometric invariant with respect to the space coordinates.

Prior distribution

The relevant quantities $\{(\theta_i, x_i, w_i)\}_{i\geq 1}$ are collected in a completely random measure

$$\tilde{\mu} := \sum_{i\geq 1} w_i \delta_{(x_i,w_i)}$$

having Lévy intensity given by $\rho(\mathrm{d}w)\mathrm{d}\theta\mathrm{d}x$.
The interaction between nodes is specified as follows:

$$z_{i,j}|\{(\theta_i, x_i, w_i)\}_{i\geq 1} \sim \mathrm{Bernoulli}\left(1 - \exp\left\{-\frac{2w_iw_j}{1+|x_i - x_j|^\beta}\right\}\right)$$

The authors are able to:

- choose a regularly varying Lévy intensity ρ to accommodate for real data problems;
- perform simulations and posterior inference;
- provide asymptotic properties in times

Comments and discussion

Benefits of the approach:

- to introduce covariate to each node and to accommodate for real data problems;
- theoretical guarantees of the proposed approach, i.e., asymptotic results on the number of nodes, edges, etc.;
- computational approach which reduces the computational complexity.

Open problems and questions:

- differentially private sparse spatial random graphs, as in (Borgs, Chayes & Smith; 2015)?
- dependent spatial random graphs based on dependent completely random measures?
- how to face prediction with spatial random graphs in presence of new nodes?

How to face prediction?

In many problems, one is interested to face prediction:

- predict new connections between nodes;
- out of sample prediction, allowing the possibility of observing new nodes.

These problems are relevant in biological frameworks, e.g., to predict protein interactions. Some other proposals are available in the literature:

- (Williamson; 2016)
- (Zhou; 2015).

Questions:

- is it possible to face prediction when the graph is represented as a point process?
- in general, how to define spatial random graphs to face prediction problems?

Dependent spatial random graphs

Consider a multiple-sample framework for random graphs:

$$Z_1, \ldots, Z_d$$

where

$$Z_\ell = \sum_{i \geq 1} \sum_{j \geq 1} z_{i,j,\ell} \delta_{(\theta_{i,\ell}, \theta_{j,\ell}, x_{i,\ell}, x_{j,\ell})}, \quad \ell = 1, \ldots, d.$$

Here one needs to exploit dependent random measures to model the prior opinion:

$$\tilde{\mu}_\ell = \sum_{i \geq 1} w_{i,\ell} \delta_{(x_{i,\ell}, w_{i,\ell})}, \quad \text{as } \ell = 1, \ldots, d.$$

Questions:

- is it possible to induce dependence across these spatial random graphs?
- are compound random measures (Griffin & Leisen; 2017) useful in this context?

References

- Borgs C., Chayes J. and Smith A. (2015). Private Graphon Estimation for Sparse Graphs. *Advances in Neural Information Processing Systems* (NIPS), **28**, 1369–1377.
- Caron F. and Fox E. (2017). Sparse graphs using exchangeable random measures. *J. Royal Stat. Soc. Ser. B*, **79**, 1295–1366.
- Griffin J. and Leisen F. (2017). Compound random measures and their use in Bayesian nonparametrics. *J. Royal Stat. Soc. Ser. B*, **79**, 525–545.
- Panero F., Caron F. and Rousseau J. (2021+). Sparse spatial random graphs. *In preparation*.
- Williamson S.A. (2016). Nonparametric Network Models for Link Prediction. *Journal of Machine Learning Research*, **17**, 1–21.
- Zhou M. (2015). Infinite edge partition models for overlapping community detection and link prediction. *Proc. Mach. Learn. Res.*, **38**, 1135–1143.

Measuring dependence in the Wasserstein distance for Bayesian nonparametric models

Summary: dependence in BNP models

Partial exchangeability: a probabilistic dependence across heterogeneous populations.

Partial exchangeability ($k = 2$)

The two sequences $\{(X_{i,j})_{j \geq 1} : i = 1, 2\}$ are partially exchangeable iff

$$(X_{1,1}, \cdots, X_{1,n_1}, X_{2,1}, \cdots, X_{2,n_2}) \overset{d}{=} (X_{1,\sigma(1)}, \cdots, X_{1,\sigma(n_1)}, X_{2,\pi(1)}, \cdots, X_{2,\pi(n_2)})$$

for every $n_1, n_2 \geq 1$ and every permutation σ and π of $\{1, \ldots, n_1\}$ and $\{1, \ldots, n_2\}$.

By de Finetti's representation theorem $\{(X_{i,j})_{j \geq 1} : i = 1, 2\}$ are partially exchangeable iff there exists a vector of dependent random probability measures $(\tilde{p}_1, \tilde{p}_2)$ such that:

$$(X_{1,j_1}, X_{2,j_2}) | \tilde{p}_1, \tilde{p}_2 \overset{\text{iid}}{\sim} \tilde{p}_1 \times \tilde{p}_2,$$
$$(\tilde{p}_1, \tilde{p}_2) \sim Q,$$

where Q is called the de Finetti measure of the sequence.

- $\tilde{p}_1 = \tilde{p}_2$: corresponds to exchangeablity, i.e., homogeneity across data;
- $\tilde{p}_1 \neq \tilde{p}_2$: corresponds to a general situation of dependence across data.

Main problems addressed by (Catalano, Lijoi & Prünster; 2021):

- how to quantify dependence of $(\tilde{p}_1, \tilde{p}_2)$ measuring how close we are to exchangeability;
- define a distance to measure the dependence, based on the Wasserstein metric.

Several Bayesian nonparametric models to accommodate for heterogeneity are based on transformations of vectors of random measures $(\tilde{\mu}_1, \tilde{\mu}_2)$:

- additive structures: (Müller, Quintana & Rosner; 2004), (Lijoi, Nipoti & Prünster; 2014);
- hierarchical structures: (Teh, Jordan, Beal & Blei; 2006) , (C, Lijoi, Orbanz & Prünster; 2019), (Griffin & Leisen; 2017);
- nested structures: (Rodriguez, Dunson & Gelfand; 2008).

Assumptions

Assume that $(\tilde{p}_1, \tilde{p}_2)$ is obtained as a suitable transformation of a random vector $(\tilde{\mu}_1, \tilde{\mu}_2)$

- $\tilde{\boldsymbol{\mu}} := (\tilde{\mu}_1, \tilde{\mu}_2)$: is a completely random vector with the same marginals;
- $\tilde{\boldsymbol{\mu}}^{co} := (\tilde{\mu}_1^{co}, \tilde{\mu}_2^{co})$: the comonotonic vector, where $\tilde{\mu}_1^{co} = \tilde{\mu}_2^{co}$ almost surely.

The authors define the following distance

$$d_{\mathcal{W}}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\mu}}^{co}) := \sup_A \mathcal{W}\left(\begin{pmatrix} \tilde{\mu}_1(A) \\ \tilde{\mu}_2(A) \end{pmatrix}, \begin{pmatrix} \tilde{\mu}_1^{co}(A) \\ \tilde{\mu}_2^{co}(A) \end{pmatrix} \right)$$

where $\mathcal{W}$ denotes the Wasserstein metric. They provide suitable bounds for different Bayesian nonparametric models:

- GM-dependent completely random measures (Lijoi, Nipoti & Prünster; 2014);
- compound random measures (Griffin & Leisen; 2017);
- GM-dependent random hazard rates (Lijoi & Nipoti; 2014).

Comments and discussion

Benefits of the approach:

- introduction of a new distance suitable for spaces of measures;
- the measure of dependence takes into account infinite dimensionality of the problem, overcoming the correlation;
- it provides us with a guide in the selection of the hyperparameters.

Open problems and questions:

- how close the posterior distribution of $(\tilde{\mu}_1, \tilde{\mu}_2)$ is to exchangeability?
- difference between $d_{\mathcal{W}}(\tilde{\mu}_1, \tilde{\mu}_2)$ and $d_{\mathcal{W}}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\mu}}^{ex})$?
- tightness of the upper bounds? Is it possible to determine lower bounds?

Bounds on posterior quantities

Remind Compound random measures of (Griffin & Leisen; 2017):

$$\tilde{\mu}_i | \tilde{\eta} = \sum_{j \geq 1} m_{i,j} J_j \delta_{\tilde{x}_j}$$

where

- $(m_{1,j}, m_{2,j}) \overset{\text{iid}}{\sim} h$, where h is a score distribution;
- $\tilde{\eta} = \sum_{j \geq 1} J_j \delta_{\tilde{x}_j}$ is a a completely random measure with Lévy measure ν^*.

BNP model

Consider the following Bayesian nonparametric model for $\mathbb{X}$-valued observations:

$$(X_{1,j_1}, X_{2,j_2}) | \tilde{p}_1, \tilde{p}_2 \overset{\text{iid}}{\sim} \tilde{p}_1 \times \tilde{p}_2$$

$$(\tilde{p}_1, \tilde{p}_2) = \left(\frac{\tilde{\mu}_1}{\tilde{\mu}_1(\mathbb{X})}, \frac{\tilde{\mu}_2}{\tilde{\mu}_2(\mathbb{X})} \right)$$

Posterior representation

Let $\boldsymbol{X}_i := (X_{i,1}, \ldots, X_{i,n_i})$, as $i = 1, 2$, be a sample from the model, then:

$$(\tilde{\mu}_1, \tilde{\mu}_2) | (\boldsymbol{X}_i, u_i)_{i=1}^2 \stackrel{d}{=} (\tilde{\mu}_1', \tilde{\mu}_2') + \sum_{\ell=1}^{k} (T_{1,\ell}, T_{2,\ell}) \sigma_\ell \delta_{x_\ell^*} \qquad (\star)$$

Questions:

- is it possible to measure how far the random vector $(\star)$ is from the exchangeable case?
- is it possible to do the same for other Bayesian nonparamteric models, in which a posterior representation is available?

Informative model-based clustering via Centered Partition Processes

References

- Camerlenghi F., Lijoi A., Orbanz P., and Prünster I. (2019). Distribution theory for hierarchical processes. *Ann. Statist.*, **47**, 67–92.
- Catalano M., Lijoi A. & Prünster I. (2021). Measuring dependence in the Wasserstein distance for Bayesian nonparametric models. *Ann. Statist.*, in press.
- Griffin J. and Leisen F. (2017). Compound random measures and their use in Bayesian nonparametrics. *J. Royal Stat. Soc. Ser. B*, **79**, 525–545.
- Lijoi A. and Nipoti B. (2014). A class of hazard rate mixtures for combining survival data from different experiments. *J. Amer. Statist. Assoc.*, **109**, 802–814.
- Lijoi A., Nipoti B. and Prünster I. (2014). Bayesian inference with dependent normalized completely random measures. *Bernoulli* **20**, 1260–1291.
- Müller P., Quintana F., and Rosner G. (2004). A method for combining inference across related nonparametric Bayesian models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **66**, 735–749.
- Regazzini E., Lijoi A. and Prünster I. (2003). Distributional results for means of random measures with independent increments. *Ann. Statist.* **31**, 560–585.
- Rodriguez A., Dunson D.B. and Gelfand A.E. (2008). The nested Dirichlet process. *J. Amer. Statist. Assoc.*, **103,** 1131–1154.
- Teh Y. W., Jordan M. I., Beal M. J. and Blei D. M. (2006). Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc.* **101**, 1566–1581.

Summary: Centered partition processes

Clustering issue

We are provided with

- $[N] := \{1, \ldots, N\}$: first N natural numbers, representing N different objects, e.g., birth defects;
- an initial partition $\boldsymbol{c}_0$ of the N objects.

How can we define a prior distribution on the space of partitions to include our prior guess $\boldsymbol{c}_0$?

A partition $\boldsymbol{c}$ of $[N]$ may be conveniently described using:

- K: number of blocks in the partition;
- $\{B_1, \ldots, B_K\}$: blocks of the partition, where B_k contains the cluster points in the kth cluster, and $|B_k| = \lambda_k$, with the constraint

$$\sum_{k=1}^{K} \lambda_k = N.$$

In the existing literature, different proposals to define a prior on the space of partitions are available:

- exchangeable models: the prior probability of $\boldsymbol{c}$ depends only on the cluster sizes $\lambda_1, \ldots, \lambda_k$ (Gnedin & Pitman; 2006);
- different proposals to relax exchangeablity, see (MacEachern; 1999).

Exchangeable Partition Probability Function

Under the exchangeable framework, the prior distribution on the space of partitions is called Exchangeable Partition Probability Function (EPPF).
A large class of EPPFs is the one induced by Gibbs-type priors:

$$p_0(\boldsymbol{c}) = \Pi_K^{(N)}(\lambda_1, \ldots, \lambda_K) = V_{N,K} \prod_{k=1}^{K} (1-\sigma)_{\lambda_k - 1}$$

where $(a)_b = \Gamma(a+b)/\Gamma(a)$, for $a, b > 0$. The parameter $\sigma < 1$ and the non–negative weights $\{V_{N,K} : \ N \geq 1,\ 1 \leq K \leq N\}$ must satisfy a recurrence relation.
See (Gnedin & Pitman; 2006) and (De Blasi et al.; 2015).

Idea of (Paganin et al.; 2021): include the prior guess $\boldsymbol{c}_0$ by a suitable penalization of the EPPF.

Centered Partition Processes

Ingredients:

- p_0 is a baseline EPPF;
- d is a distance in the space of partitions;
- ψ is a penalization term.

The proposed centred partition process is associated with the following prior on the space of partitions

$$p(\boldsymbol{c}|\boldsymbol{c}_0, \psi) \propto p_0(\boldsymbol{c}) e^{-\psi d(\boldsymbol{c}, \boldsymbol{c}_0)}$$

We have two limiting situations:

- $\psi \to 0$: baseline EPPF;
- $\psi \to +\infty$: $\boldsymbol{c} = \boldsymbol{c}_0$ with probability one.

Comments and Discussion

Benefits of the approach:

- include the prior guess $\boldsymbol{c}_0$ in the model;
- prior calibration of the parameter ψ;
- it allows to measure the uncertainty of the partition.

Open problems and questions:

- predictive and posterior properties of the model: is this tractable from a mathematical stand point?
- how can you improve the performance of the algorithm for prior calibration of ψ?
- extension to the case of feature allocation models: is this interesting?

Feature Allocation Models

Feature allocations are combinatorial structures:

- generalize the notion of partition, see (Broderick, Jordan & Pitman; 2013).
- $i \in [N]$ represents an individual which may display multiple features.

Denote by $x_1, \ldots, x_K$ the K distinct features out of the N individuals and

$$B_k := \{i \in [N] : \ i \text{ displays feature } k\}, \quad m_k := |B_k|.$$

- An index i may belong to more then one set B_k, this means that it displays more then one feature;
- m_k is the number of individuals displaying feature x_k.

Exchangeable Feature Allocations

- A random feature allocation $\boldsymbol{f}$ is termed exchangeable when its distribution depends only on $m_1, \ldots, m_K$, and not on the features' labels.
- Exchangeable feature allocation probability function (EFPF): is the distribution of the random feature allocation.

Examples of EFPF are the following:

- the EFPF induced by the three-parameters Indian Buffet Process (Teh, Görür, Ghahramani; 2009)

$$p(\boldsymbol{f}) = \frac{1}{K!}\left(\frac{\alpha}{(c+1)_{N-1}}\right)^K \exp\left\{-\alpha\sum_{i=1}^{N}\frac{(\sigma+c)_{i-1}}{(1+c)_{i-1}}\right\}\prod_{k=1}^{K}(1-\sigma)_{m_k-1}(c+\sigma)_{N-m_k}$$

- possible generalization, e.g., Gibbs-type Indian Buffet Processes by (Heaukulani & Roy; 2020).

Questions:

- is it possible to define centred feature allocation models?
- is this an interesting extension of the model?
- is there any application, where one is provided with a prior guess $\boldsymbol{f}_0$ for a feature allocation?

References

- Broderick T., Jordan M.I. and Pitman J. (2013). Cluster and feature modeling from combinatorial stochastic processes. *Statistical Science*, **28**, 289–312.
- 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**, 212–229.
- Gnedin A. and Pitman J. (2006). Exchangeable Gibbs partitions and Stirling triangles. *Journal of Mathematical Sciences*, **138**, 5674–5685.
- Heaukulani C. and Roy, D.M. (2020). Gibbs-type Indian Buffet Processes. *Bayesian Analysis*, **15**, 683–710.
- MacEachern S. N. (1999). Dependent nonparametric processes. In *Proceedings of the Bayesian Section.*, 50–55. Alexandria, VA: American Statistical Association.
- Paganin S., Herring A.H., Olshan A.F. and Dunson D.B. (2021). Centered Partition Processes: Informative Priors for Clustering. *Bayesian Analysis*, in press.
- Teh Y.W., Görür D. and Ghahramani Z. (2009). Indian Buffet Processes with Power-law Behavior. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*.