Submitted to the Annals of Applied Statistics

A SAMPLE DOCUMENT*

BY FIRST AUTHOR[†,‡,¶], SECOND AUTHOR[§,¶] AND THIRD AUTHOR[†,¶]

University of California Davis[¶]

The abstract should summarize the contents of the paper. It should be clear, descriptive, self-explanatory and not longer than 200 words.

1. Introduction.

2. Model.

2.1. *Stochastic block model.* A set of n nodes $\Gamma = \{v_1, \cdots, v_n\}$ is partitioned into k clusters $\Gamma_1, \cdots, \Gamma_k$. For simplicity, we use the notation $i \in \Gamma_l$ to denote that $v_i \in \Gamma_l$. The cluster of node v_i is represented by $z_i \in \{1, \cdots, k\}$, and the vector of clusters is $\mathbf{z} = (z_i)_{i=1}^n$. Define the adjacency matrix $\mathbf{A} \in \{0,1\}^{n\times n}$ where $\mathbf{A}_{i,j} = 1$ if an edge is observed between v_i and v_j and $\mathbf{A}_{i,j} = 0$ otherwise. We set $\mathbf{A}_{i,i} \equiv 0$ for any $i = 1, \cdots, n$, and assume that $\mathbf{A}_{i,j}$'s are conditionally independent given the cluster vector $\mathbf{z}$:

$$\mathbf{A}_{i,j}|z_i = q, z_j = l \overset{ind}{\sim} \text{Bernoulli}(\mathbf{C}_{q,l}), \qquad i \neq j,$$

where $\mathbf{C} \in [0,1]^{k\times k}$ denote the connecting probability matrix.

2.2. *Dynamic generalization of the stochastic block model.* Consider a growing dynamic network where edges appear over time. Assume that the edges will not disappear once constructed, and that the observed point processes $N_{i,j}(\cdot) \in \{0,1\}$ are independent realizations of intensity functions

$$\lambda_{i,j}(t) = f_{z_i,z_j}(t - \tau_{i,j}) \cdot g(d_{i,j}), \qquad t \in [0,T], \quad i,j = 1, \cdots, n,$$

where $[0,T]$ is overall time period, $\tau_{i,j}$ is the time lag of the edge from v_i to v_j, $d_{i,j}$ is the spatial distance between v_i and v_j, $f_{z_i,z_j}(\cdot)$ is the connecting intensity function from cluster z_i to z_j, and $g(\cdot)$ is a decreasing function

*Footnote to the title with the "thankstext" command.

†Some comment

‡First supporter of the project

§Second supporter of the project

MSC 2010 subject classifications: Primary 60K35, 60K35; secondary 60K35

Keywords and phrases: sample, LaTeX 2ε

that accounts for the decay of connection as the distance between any pair of nodes increases. Similar to the stochastic block model, we set $\lambda_{i,i}(\cdot) \equiv 0$ for $i = 1, \cdots, n$.
The integrated point process $N_{i,\cdot}(\cdot) := \sum_{j \neq i} N_{i,j}(\cdot)$ can be viewed as a realization of the intensity function $\lambda_{i,\cdot}(\cdot) = \sum_{j \neq i} \lambda_{i,j}(\cdot)$. For convenience, we denote $N_{i,\cdot}(\cdot)$ by $N_i(\cdot)$, and $\lambda_{i,\cdot}(\cdot)$ by $\lambda_{N_i}(\cdot)$.

In many applications, the edge from v_i to v_j is determined only by the activity of v_i. So we may have the following assumption.

ASSUMPTION 1. *$\tau_{i,j}$ only depends on v_i, that is, $\tau_{i,j} = \tau_i$ for all $j \neq i, i = 1, \cdots, n$.*

With Assumption 1, $\lambda_{N_i}(\cdot)$ can be written as

$$\begin{aligned}\lambda_{N_i}(t) &= \sum_{l=1}^{k} \left(f_{z_i,l}(t - \tau_i) \cdot \sum_{j \in \Gamma_l, j \neq i} g(d_{i,j}) \right) \\ &=: \sum_{l=1}^{k} f_{z_i,l}(t - \tau_i) \cdot w_{i,l}.\end{aligned}$$

Here $w_{i,l}$ measures the overall distance between node v_i and cluster Γ_l. For example, if the nodes represent neurons and are clustered by cell types, $w_{i,l}$ represents the distance between v_i and its neighbor cells which belong to cell types l. We assume that the cell types are distributed uniformly in the sense that $w_{i,l}$ and $w_{j,l}$ are identically distributed for any i, j such that $z_i = z_j$. More formally, we have the following assumption.

ASSUMPTION 2. *$w_{i,l} = \bar{w}_{z_i,l} + \epsilon_{i,l}$ where $\{\epsilon_{i,l}\}_{i \in \Gamma_l, l = 1, \cdots, k}$ are i.i.d. random variables with mean 0 and variance $\sigma^2 < \infty$.*

By Assumption 1 and 2, $\lambda_{N_i}(t) = \lambda_{z_i}(t - \tau_i) + \sum_{l=1}^{k} f_{z_i,l}(t - \tau_i)\epsilon_{i,l}$ for $t \in [0, T]$, where $\lambda_{z_i}(t) := \sum_{k=1}^{l} f_{z_i,l}(t) \cdot \bar{w}_{z_i,l}$.

Minor comments. Hoeffding's inequality might be useful later. (If $\epsilon_1, \cdots, \epsilon_d \overset{i.i.d}{\sim}$ sub-G(τ_0) and $\mathbb{E}\epsilon_i = 0$, then $\mathbb{P}(\langle a, \epsilon \rangle \geq t) \leq \exp\left\{-t^2/(2\|a\|_2^2 \tau_0^2)\right\}$ for any $a \in \mathbb{R}^d$.)

Let $\mathbf{F}_{k \times k} = [f_{q,l}(\cdot)]_{q,l \in \{1, \cdots, k\}}$, $\mathbf{W}_{n \times k} = [w_{i,l}]_{i \in \{1, \cdots, n\}, l \in \{1, \cdots, k\}}$, $\mathbf{Z} \in \{0, 1\}^{n \times k}$ with $\mathbf{Z}_{i,l} = 1$ if $z_i = l$ and 0 otherwise. Then

$$\begin{bmatrix} \lambda_{N_1}(\cdot + \tau_1) \\ \vdots \\ \lambda_{N_n}(\cdot + \tau_n) \end{bmatrix} = \operatorname{diag}\left(\mathbf{ZFW}^\top\right).$$

3. Method.

3.1. *k-means objective function.* In what follows, we review the k-means objective function in the Euclidean space, and introduce the objective function in our case where the samples are realizations of point processes.

k-means in $\mathbb{R}^d$. Let $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^d$ be an i.i.d. sample from distribution function F. Denote by F_n the empirical distribution function. The k-means problem is to solve

$$A_n := \operatorname*{arg\,min}_{A: A\subset\mathbb{R}^d, |A|=k} W_n(A, F_n) = \operatorname*{arg\,min}_{A: A\subset\mathbb{R}^d, |A|=k} \int \min_{a\in A} \|\mathbf{x}_i - a\|^2 \mathrm{d}F_n.$$

There are some theoretical guarantees of k-means. Pollard [13] shows that for a given k, $A_n \to \bar{A}$ almost surely, where $\bar{A} = \arg\min_A W(A, F)$ denotes the optimal population cluster centers. Bachem et al. [1] provides a uniform bound with a rate of $\mathcal{O}\left(n^{-1/2}\right)$ for the deviation between the empirical loss and the expected loss. The bound is uniform in the sense that it holds for any set of k cluster centers.

Note that $\bar{A}$ is a biased estimator of the true cluster centers (when they are well-defined). For example, if $k = 2$ and $F(x) = \frac{1}{2}\Phi(x; \mu_1, \sigma_1^2) + \frac{1}{2}\Phi(x; \mu_2, \sigma_2^2)$ is a mixture Gaussian distribution, and denote $X_1 \sim N(\mu_1, \sigma_1^2)$, then $\bar{A} = \{a_1, a_2\}$ where $a_1 = \mathbb{E}[X_1 \mathbf{1}_{X_1 \le (\mu_1+\mu_2)/2}] \neq \mu_1$ (and a similar expression for a_2).

This problem can be re-formulated as following

$$\min_{\{\Gamma_l\}_{l=1}^k} \frac{1}{n} \sum_{l=1}^{k} \sum_{i\in\Gamma_l} \|\mathbf{x}_i - \mathbf{c}_l\|^2, \tag{3.1}$$

where $\{\Gamma_l\}_{l=1}^k$ represent the clusters and form a partition of $\Gamma = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$, $\mathbf{c}_l = 1/|\Gamma_l| \cdot \sum_{\mathbf{x}_i\in\Gamma_l} \mathbf{x}_i$ is the sample center of the l-th cluster. For simplicity, we denote $\mathbf{x}_i \in \Gamma_l$ by $i \in \Gamma_l$ henceforth. The notation Γ_l is introduced in 2.1 but for a different set of nodes. What is the appropriate way to re-define this notation? We now extend this objective function to the context of point process.

k-means in point processes. Recall that by Assumption 1 and 2, $\lambda_{N_i}(t) = \lambda_{z_i}(t - \tau_i) + \sum_{l=1}^k f_{z_i,l}(t - \tau_i)\epsilon_{i,l}$, hence $\lambda_{N_i}(t + \tau_i) \overset{d}{=} \lambda_{N_j}(t + \tau_j)$ for any i, j satisfying $z_i = z_j$. So the k-means problem is to solve

$$\min_{\{\Gamma_l\}_{l=1}^k} \frac{1}{n} \sum_{l=1}^{k} \left(\min_{\{\tau_i\}_{i\in\Gamma_l}, \lambda_l} \sum_{i\in\Gamma_l} \|\hat{\lambda}_{N_i}(\cdot + \tau_i) - \lambda_l(\cdot)\|_2^2 \right), \tag{3.2}$$

or, equivalently,

$$\min_{\{\Gamma_l\}_{l=1}^k} \frac{1}{n} \sum_{l=1}^{k} \left(\min_{\{\tau_i\}_{i\in\Gamma_l}, F_l} \sum_{i\in\Gamma_l} \|\hat{F}_{N_i}(\cdot+\tau_i) - F_l(\cdot)\|_2^2 \right), \tag{3.3}$$

where $\hat{\lambda}_{N_i}(\cdot)$ is the intensity function estimated from $N_i(\cdot)$ using some smooth method, $\hat{F}_{N_i}(t) := 1/N_i[0,T] \cdot \sum_{j=1}^{N_i[0,T]} \mathbf{1}_{\{t_{N_i,j}\le t\}}$ is the empirical distribution of time of edges, $t_{N_i,j}$'s are the time of edges of $N_i(\cdot)$.

Likelihood can also be used as a measure of similarity between a point process and an intensify function, which yields the objective function

$$\min_{\{\Gamma_l\}_{l=1}^k} \frac{1}{n} \sum_{l=1}^{k} \left(\min_{\{\tau_i\}_{i\in\Gamma_l}, \lambda_l} \sum_{i\in\Gamma_l} d\big(N_i(\cdot+\tau_i), \lambda_l(\cdot)\big) \right), \tag{3.4}$$

where the distance is defined as the negative log-likelihood

$$d(N,\lambda) := -\log L(N;\lambda) = \int_0^T \lambda(t)\mathrm{d}t - \sum_{j=1}^{N(0,T]} \log\left(\lambda(t_{N,j})\right).$$

Other possible distance. The squared error distance is defined as

$$d(N,\lambda) := \int_0^T \lambda^2(t)\mathrm{d}t - 2\int_0^T \lambda(t)\mathrm{d}N(t) + ?$$

3.2. *Algorithm.* For (3.2) and (3.3), following the idea of Lloyd's algorithm [10], we iterate between two steps until convergence:

- Update time lags τ_i's and mean intensity functions λ_l's (or F_l's) based on current clustering using the method of shape invariant model.
- Update clustering by re-assigning each node ($\hat{\lambda}_{N_i}$ or $\hat{F}_{N_i}$) to its nearest center.

Shape invariant model has been analyzed in many literatures [3, 2, 14, 5, 15, 4, 6, 16]. The model aims to estimate a mean function (of time) from a set of similar functions where the variability consists of the time shifts and additive noise.

Following the method in [2], one can estimate the time lags based on Fourier coefficients, then take the mean of aligned curves as the estimation

of the mean curve. To be specific, the estimators for (3.2) are given by

$$\{\hat{\tau}_i\}_{i\in\Gamma_l} = \arg\min_{\tau_i, i\in\Gamma_l} \frac{1}{|\Gamma_l|}\sum_{i\in\Gamma_l}\left\|\hat{\lambda}_{N_i}(t+\tau_i) - \frac{1}{|\Gamma_l|}\sum_{j\in\Gamma_l}\hat{\lambda}_{N_j}(t+\tau_j)\right\|_2^2,$$
$$\hat{\lambda}_l(t) = \frac{1}{|\Gamma_l|}\sum_{i\in\Gamma_l}\hat{\lambda}_{N_i}(t+\hat{\tau}_i).$$

The estimation procedure for (3.3) is similar.

We might be able to prove similar theorems as those in Bigot and Gendre [2]. For example, given a cluster Γ_l, the convergence rate of

(3.5)
$$\arg\min_{\{\tau_i\}_{i\in\Gamma_l},\lambda_l}\sum_{i\in\Gamma_l}\|\hat{\lambda}_{N_i}(\cdot+\tau_i)-\lambda_l(\cdot)\|_2^2 \to \arg\min_{\{\tau_i\}_{i\in\Gamma_l},\lambda_l}\sum_{i\in\Gamma_l}\|\lambda_{N_i}(\cdot+\tau_i)-\lambda_l(\cdot)\|_2^2$$

and

(3.6)
$$\arg\min_{\{\tau_i\}_{i\in\Gamma_l},F_l}\sum_{i\in\Gamma_l}\|\hat{F}_{N_i}(\cdot+\tau_i)-F_l(\cdot)\|_2^2 \to \arg\min_{\{\tau_i\}_{i\in\Gamma_l},F_l}\sum_{i\in\Gamma_l}\|F_{N_i}(\cdot+\tau_i)-F_l(\cdot)\|_2^2.$$

The derivation of theory for (3.6) may be easier than (3.5) because $\hat{F}_{N_i}(t)\to F_{N_i}(t)$ as $n\to\infty$, but $\hat{\lambda}_{N_i}(t)$ depends on the choice of bandwidth.

How to derive the convergence of solution of (3.2) and (3.3)?

Should we take into account the error in $t_{N_i,j}$?

Another direction is to use negative log-likelihood as the distance. For reference see [15, 4, 14, 5]. The maximum likelihood estimator of the mean curve proposed in Gervini and Gasser [5] is showed to be $\sqrt{n}$-consistent and asymptotically normal. But the log-likelihood includes unknown $f_{l,l'}$

$$L(\lambda_l,\tau_i;N_i) = \int \mathbb{P}\left(N_i(t)\middle|\lambda_l(t-\tau_i)+\sum_{l'=1}^{k} f_{l,l'}(t-\tau_i)\epsilon_{i,l'}\right) f(\epsilon_{i,1})\cdots f(\epsilon_{i,k})\,\mathrm{d}\epsilon_{i,1}\cdots\mathrm{d}\epsilon_{i,k}$$

3.3. *Convex relaxation of k-means type clustering.*

Semidefinite programming relaxation. We briefly introduce a semidefinite programming relaxation (Peng-Wei relaxation) of k-means proposed by

Peng and Wei [12]. The k-means objective function in (3.1) can be re-written as

$$\sum_{l=1}^{k}\sum_{i\in\Gamma_l}\|\mathbf{x}_i-\mathbf{c}_l\|^2=\frac{1}{2}\sum_{l=1}^{k}\frac{1}{|\Gamma_l|}\sum_{i,j\in\Gamma_l}\|\mathbf{x}_i-\mathbf{x}_j\|^2$$
$$=\frac{1}{2}\sum_{l=1}^{k}\frac{1}{|\Gamma_l|}\langle\mathbf{1}_{\Gamma_l}\mathbf{1}_{\Gamma_l}^{\top},\mathbf{D}\rangle$$

where $\mathbf{D}\in\mathbb{R}^{n\times n}$ with entries $\mathbf{D}_{ij}=\|\mathbf{x}_i-\mathbf{x}_j\|^2$.
Hence (3.1) can be relaxed to

$$\min_{\mathbf{Z}}\ \langle\mathbf{Z},\mathbf{D}\rangle$$
$$\text{s.t.}\quad \mathbf{Z}\succeq 0,\quad \mathbf{Z}\geq 0,\quad \mathbf{Z}1_n=\mathbf{1}_n,\quad \mathrm{Tr}(\mathbf{Z})=k.$$

Proximity conditions are discussed in 4.3.

In our case, (3.2) is equivalent to

$$\min_{\mathbf{Z},\{\tau_i\}_{i=1}^n}\ \langle\mathbf{Z},\mathbf{D}(\tau_1,\cdots,\tau_n)\rangle \tag{3.7}$$

where $\mathbf{Z}=1/2\cdot\sum_{l=1}^{k}(1/|\Gamma_l|\cdot\mathbf{1}_{\Gamma_l}\mathbf{1}_{\Gamma_l}^{\top})$, $\mathbf{D}_{i,j}=\|\hat{\lambda}_{N_i}(t+\tau_i)-\hat{\lambda}_{N_j}(t+\tau_j)\|_2^2$.
Using Fourier transformation,

$$\mathbf{D}_{i,j}=\int_0^T\left(\int_0^\infty\hat{h}_{N_i}(\xi)e^{i2\pi\xi(t+\tau_i)}-\hat{h}_{N_j}(\xi)e^{i2\pi\xi(t+\tau_j)}\mathrm{d}\xi\right)^2\mathrm{d}t$$

where $\hat{h}_{N_i}(\xi)$ is the Fourier transform of $\hat{\lambda}_{N_i}(t)$. How to convexify?

4. Main result.

4.1. *Convergence of empirical loss.* For the case in (3.2), denote $\theta=\left(\{\Gamma_l\}_{l=1}^k,\{\tau_i\}_{i=1}^n,\{\lambda_l\}_{l=1}^k\right)$, show the convergence rate of

$$\hat{\theta}_n:=\underset{\{\Gamma_l\}_{l=1}^k,\{\tau_i\}_{i=1}^n,\{\lambda_l\}_{l=1}^k}{\arg\min}\frac{1}{n}\sum_{l=1}^{k}\sum_{i\in\Gamma_l}\|\hat{\lambda}_{N_i}(\cdot+\tau_i)-\lambda_l(\cdot)\|_2^2$$

to

$$\theta_n:=\underset{\{\Gamma_l\}_{l=1}^k,\{\tau_i\}_{i=1}^n,\{\lambda_l\}_{l=1}^k}{\arg\min}\frac{1}{n}\sum_{l=1}^{k}\sum_{i\in\Gamma_l}\|\lambda_{N_i}(\cdot+\tau_i)-\lambda_l(\cdot)\|_2^2.$$

Papers about shape invariant model might be useful.

4.2. *Convergence of Lloyd's algorithm using shape invariant model.*

Guarantees of Lloyd's algorithm. Lu and Zhou [11] provides a weak initialization condition under which Lloyd's algorithm converges to the optimal label estimators of sub-Gaussian mixture model. Also, see the reference therein.

4.3. *Proximity condition of the convex relaxation.* See Ling and Strohmer [9], Li et al. [8], Peng and Wei [12], Zhao, Levina and Zhu [17] for theory and proof. Ling and Strohmer [9] proposed a proximity condition under which the convex relaxation of RatioCuts is exactly the global optimal to the original ratio cut problem. The theorem is then applied to spectral clustering (as a special case of graph cuts) to obtain the theoretical guarantees for spectral clustering
Li, Chen and Xu [7] surveys recent theoretical advances in convex optimization approaches for community detection. Li et al. [8] compare different convex relaxations by relating them to corresponding proximity conditions. They present an improved proximity condition under which the relaxation proposed by Peng and Wei [12] recovers the underlying clusters exactly. The proximity condition states that for any $a \neq b$, the following holds

$$\min_{i\in\Gamma_a}\left\langle \mathbf{x}_i - \frac{\mathbf{c}_a+\mathbf{c}_b}{2}, \mathbf{w}_{b,a}\right\rangle > \frac{1}{2}\sqrt{\left(\sum_{l=1}^{k}\left\|\overline{\mathbf{X}}_l\right\|^2\right)\left(\frac{1}{n_a}+\frac{1}{n_b}\right)}$$

where $\mathbf{w}_{b,a} = \frac{\mathbf{c}_a-\mathbf{c}_b}{\|\mathbf{c}_a-\mathbf{c}_b\|}$ is the unit vector pointing from $\mathbf{c}_b$ to $\mathbf{c}_a$, $\overline{\mathbf{X}}_l$ is the centered data matrix of the l-th cluster, $\|\overline{\mathbf{X}}_l\|$ is the operator norm of $\overline{\mathbf{X}}_l$, and $n_a = |\Gamma_a|, n_b = |\Gamma_b|$.

Translate into point processes if (3.7) can be convexified.

5. Simulation.

6. Conclusion.

APPENDIX A: APPENDIX SECTION

Some words.

A.1. Appendix subsection. See Appendix A.

ACKNOWLEDGEMENTS

See Supplement A for the supplementary material example.

SUPPLEMENTARY MATERIAL

Supplement A: Title of the Supplement A
(http://www.e-publications.org/ims/support/dowload/imsart-ims.zip). Dum esset rex in accubitu suo, nardus mea dedit odorem suavitatis. Quoniam confortavit seras portarum tuarum, benedixit filiis tuis in te. Qui posuit fines tuos

REFERENCES

[1] Bachem, O., Lucic, M., Hamed Hassani, S. and Krause, A. (2017). Uniform Deviation Bounds for k-Means Clustering Technical Report.

[2] Bigot, J. and Gendre, X. (2013). Minimax properties of Fréchet means of discretely sampled curves. *Ann. Stat.* **41** 923–956.

[3] Bontemps, D. and Gadat, S. (2014). Bayesian methods for the Shape Invariant Model. *Electron. J. Stat.* **8** 1522–1568.

[4] Gamboa, F., Loubes, J.-M. and Maza, E. (2007). Semi-parametric estimation of shifts. *Electron. J. Stat.* **1** 616–640.

[5] Gervini, D. and Gasser, T. (2005). Nonparametric maximum likelihood estimation of the structural mean of a sample of curves. *Biometrika* **92** 801–820.

[6] Jérémie Bigot, B. and Gadat, S. (2010). A DECONVOLUTION APPROACH TO ESTIMATION OF A COMMON SHAPE IN A SHIFTED CURVES MODEL. *Ann. Stat.* **38** 2422–2464.

[7] Li, X., Chen, Y. and Xu, J. (2018). Convex Relaxation Methods for Community Detection.

[8] Li, X., Li, Y., Ling, S., Strohmer, T. and Wei, K. (2017). When Do Birds of a Feather Flock Together? k-Means, Proximity, and Conic Programming.

[9] Ling, S. and Strohmer, T. (2019). Certifying Global Optimality of Graph Cuts via Semidefinite Relaxation: A Performance Guarantee for Spectral Clustering Technical Report.

[10] Lloyd, S. P. (1982). Least Squares Quantization in PCM. *IEEE Trans. Inf. Theory* **28** 129–137.

[11] Lu, Y. and Zhou, H. H. Statistical and Computational Guarantees of Lloyd's Algorithm and its Variants Technical Report.

[12] Peng, J. and Wei, Y. (2005). Approximating K-means-type clustering via semidefinite programming Technical Report.

[13] Pollard, D. (1981). Strong Consistency of K-Means Clustering. *Source Ann. Stat.* **9** 135–140.

[14] Rønn, B. B. and Skovgaard, I. M. (2009). Nonparametric maximum likelihood estimation of randomly time-transformed curves Technical Report No. 1.

[15] Vimond, M. (2010). EFFICIENT ESTIMATION FOR A SUBCLASS OF SHAPE INVARIANT MODELS. *Ann. Stat.* **38** 1885–1912.

[16] Wang, K. and Gasser, T. (1997). ALIGNMENT OF CURVES BY DYNAMIC TIME WARPING Technical Report No. 3.

[17] Zhao, Y., Levina, E. and Zhu, J. I. (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *Ann. Stat.* **40** 2266–2292.

Address of the First and Second authors
Usually a few lines long
E-mail: first@somewhere.com
second@somewhere.com

Address of the Third author
Usually a few lines long
Usually a few lines long
E-mail: third@somewhere.com
URL: http://www.foo.com