Submitted to the Annals of Applied Statistics

A SAMPLE DOCUMENT[*]

By First Author[†,‡,¶], Second Author[§,¶,‖] and Third Author[†,‖]

Some University[¶] *and Another University*[‖]

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 $\Gamma_1, \cdots, \Gamma_k$. The cluster of node v_i is represented by $z_i \in \{1, \cdots, k\}$, and the clusters are obtained in the vector $\mathbf{z} = (z_i)_{i=1}^n$. Define the adjacency matrix $\mathbf{A} \in \{0,1\}^{n \times n}$ where (for $i < j$) $\mathbf{A}_{i,j} = 1$ if an edge is observed between v_i and v_j and $\mathbf{A}_{i,j} = 0$ otherwise. We set $\mathbf{A}_{i,i} \equiv 0$ and $\mathbf{A}_{i,j} = \mathbf{A}_{j,i}$ for any $i, j = 1, \cdots, n$, and assume that $\mathbf{A}_{i,j}$'s are conditionally independent given the cluster vector $\mathbf{z}$:

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

where $\mathbf{C} \in [0,1]^{k \times k}$ denote the (symmetric) 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 developed, and that the observed point processes $N_{i,j}(\cdot) \in \{0,1\}$ are independent realizations of intensity functions

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

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

[*]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, $\LaTeX\,2\varepsilon$

accounts for the decay of connection as the distance between any pair of nodes increases. Similar as stochastic block model, we set $\lambda_{i,i}(\cdot) \equiv 0$ and $\lambda_{i,j}(\cdot) = \lambda_{j,i}(\cdot)$.

The integrated point process is defined as following. Let $\lambda_{i,\cdot}(t) = \sum_{j\neq i} \lambda_{i,j}(t)$ for $t \in [0, T]$, then $N_{i,\cdot}(t) = \sum_{j\neq i} N_{i,j}(t)$ is a realization of $\lambda_{i,\cdot}(\cdot)$. For convenience, we abbreviate $\lambda_{i,\cdot}(\cdot)$ to $\lambda_i(\cdot)$, and $N_{i,\cdot}(\cdot)$ to $N_i(\cdot)$.

ASSUMPTION 1. *Need to fill the gap.*

By assumption 1, $N_i(t + \tau_i) \overset{d}{=} N_j(t + \tau_j)$ for any i, j such that $z_i = z_j$.

3. Method.

3.1. *k-means objective function.*

We will introduce the objective function of k-means in the Euclidean space as well as in our case where the samples are realizations of point process.

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 minimize

$$W_n(A, F_n) = \int \min_{a\in A} \|\mathbf{x}_i - a\|^2 dF_n$$

over all possible choices of the set A containing k points in $\mathbb{R}^d$. Denote by $\bar{A} = \arg\min_A W(A, F)$ the optimal population cluster centers, and $A_n = \arg\min_A W_n(A, F_n)$ the optimal sample cluster centers. Pollard [13] showed that for a given k,

$$A_n \to \bar{A} \text{ a.s.}$$

What does $\bar{A}$ represent? 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 I expect $\bar{A} = \{a_1, a_2\}$ where $a_1 = \mathbb{E}[X_1 \mathbf{1}_{X_1 \leq (\mu_1+\mu_2)/2}] \neq \mu_1$ (and a similar expression for a_2).

This problem can be reformulated as solving

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

where $\{\Gamma_l\}_{l=1}^k$ represent the clusters and form a partition of $\Gamma = \{1, 2, \cdots, n\}$, $\mathbf{c}_l = \frac{1}{|\Gamma_l|} \sum_{i\in\Gamma_l} \mathbf{x}_i$ is the sample center of the l-th cluster. We now extend this objective function to the context of point process.

k-means in point process. Denote by $d(N(\cdot), \lambda(\cdot))$ the distance between a point process and an intensity function. The problem of selecting distance function is discussed later. If we can assume that $N_i(t+\tau_i) \stackrel{d}{=} N_j(t+\tau_j)$ for any i, j such that $z_i = z_j$, the k-means problem is

$$\min_{\{\Gamma_l\}_{l=1}^k} \frac{1}{n} \sum_{l=1}^{k} \left(\min_{\{\tau_i\}_{i\in\gamma_l}, \lambda_l} \sum_{i\in\gamma_l} d(N_i(t+\tau_i), \lambda_l) \right). \tag{3.2}$$

I expect the solution of (3.2) to converge, as $n \to \infty$, to the minimizer of the population version of this objective function. But what does that mean? Read more paper about consistency of k-means problem.

Distance between a point process and an intensity function. For a given Poisson process $N(\cdot)$ and an intensity function $\lambda(\cdot)$, the distance can be defined as the negative log-likelihood

$$d(N, \lambda) \equiv -l(N; \lambda) = \int_0^T \lambda(t)dt - \sum_{j=1}^{N(0,T]} \log\left(\lambda(t_j^{(N)})\right),$$

where $t_j^{(N)}$'s are the time of events of $N(\cdot)$. Justify why it is reasonable to use Poisson process. See Daley and Springer [3] for details.

The squared error distance is defined as

$$d(N, \lambda) \equiv \int_0^T \lambda^2(t)dt - 2\int_0^T \lambda(t)dN(t)$$

It seems obvious but I do not know how to derive this.
Analyze its pros and cons and compare with the log-likelihood metric.

3.2. *Algorithm.* To solve the k-means problem in $\mathbb{R}^d$, Lloyd's algorithm [10] is a standard choice. In order to apply the Lloyd's algorithm, a good estimation of $\{\tau_i\}_{i=1}^n$ and $\{\lambda_l\}_{l=1}^k$, given clusters $\{\Gamma_l\}_{l=1}^k$, is needed.

Intensity estimation: shape invariant model. Shape invariant model (SIM) is analyzed in [2, 1, 14, 5, 15, 4, 6, 16]. The model is

$$Y_{s,j} = f(t_s - \theta_j^*) + \epsilon_{s,j}, \qquad s = 1, \cdots, n \text{ and } j = 1, \cdots, J,$$

where j is the index of curves, s is the index of observed points, $\epsilon_{s,j}$'s are i.i.d. Gaussian variables with zero expectation and variance σ^2, and f is T-periodic.

Depending on the choice of distance, we have two directions. If the squared error distance is adopted, one can estimate the time lags based on Fourier coefficients, then take the mean (in the Euclidean space) of aligned curves as the estimation of the mean curve. Relevant papers include [1, 6] and the reference therein. Bigot and Gendre [1] minimizes the sum of squared distance over time lags, then take the mean of curves aligned by estimated time lags. A minimax rate is derived and the proposed estimator is proved to achieve this minimax rate when both the number of curves and the number of sampling points go to infinity.
To be more specific, fixing a cluster Γ_l, we can model the smoothed intensity function

$$\hat{\lambda}_{N_i}(t) = \lambda_l(t - \tau_i) + \epsilon_i(t), \qquad i \in \Gamma_l, l = 1, \cdots, k, \tag{3.3}$$

or the empirical distribution of edge emerging time

$$\begin{aligned} \hat{F}_{N_i}(t) &\equiv \frac{1}{N_i(0,T]} \sum_{j=1}^{N_i(0,T]} \mathbf{1}_{\left\{t_j^{(N_i)} \le t\right\}} \\ &= F_l(t - \tau_i) + \epsilon_i(t), \qquad i \in \Gamma_l, l = 1, \cdots, k. \end{aligned} \tag{3.4}$$

Having the above model, one can then solve for time lags via

$$\underset{\{\tau_i\}_{i\in\Gamma_l}}{\arg\min} \sum_{i\in\Gamma_l} d\left(\hat{\lambda}_{N_i}(t+\tau_i),\ \frac{1}{|\Gamma_l|}\sum_{j\in\Gamma_l}\hat{\lambda}_{N_j}(t+\tau_j)\right)$$

or

$$\underset{\{\tau_i\}_{i\in\Gamma_l}}{\arg\min} \sum_{i\in\Gamma_l} d\left(\hat{F}_{N_i}(t+\tau_i),\ \frac{1}{|\Gamma_l|}\sum_{j\in\Gamma_l}\hat{F}_{N_j}(t+\tau_j)\right)$$

where the distance can be ℓ_2-distance or squared ℓ_2-distance (or K-S distance?).

Note that in (3.3), $\epsilon_i(\cdot) = (\hat{\lambda}_{N_i}(\cdot) - \lambda_{N_i}(\cdot)) + (\lambda_{N_i}(\cdot) - \lambda_l(\cdot))$ includes (i) the error between $\hat{\lambda}_{N_i}$ and λ_{N_i} and (ii) the error between λ_{N_i} and λ_l. Similar decomposition holds for (3.4). So in order to control these two error terms, we need convergence theory of smooth method (or empirical distribution function), as well as some assumptions about the distribution of $\{\lambda_{N_i}\}_{i\in\Gamma_l}$ (which is required by the theorems of SIM).
Find papers analyzing the behavior of Lloyd's algorithm with error in centers.

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.
Each point process is treated as its (expected) intensity function plus an error term (how to control?)

$$N_i(t) = \lambda_l(t-\tau_i) + \epsilon_i(t), \qquad i \in \Gamma_l, l = 1, \cdots, k, \tag{3.5}$$

One can then solve for the time lags and the mean intensity function base on the log-likelihood

$$\arg\max_{\lambda_l, \{\tau_i\}_{i\in\Gamma_l}} \sum_{i\in\Gamma_l} l(N_i(t), \lambda_l(t-\tau_i)).$$

Note that the likelihood implicitly includes the distribution of $\{\lambda_{N_i}\}_{i\in\Gamma_l}$.
How to formulate the distribution of $\{\lambda_{N_i}\}_{i\in\Gamma_l}$?
The convergence of MLE are proved for Gaussian model. Need some modification for Poisson process model.

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

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

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

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

Proximity conditions are discussed in 4.3.
But this relaxation replies on $\sum_{i\in\Gamma_l}(\mathbf{x}_i-\mathbf{c}_i) = 0$, how to generalize it in (3.2)?
Read other convex relaxation method and their proximity conditions.

3.4. *Other thoughts.*

- Non-asymptotic theory?
- If we want to estimate pairwise dissimilarity while incorporating time lag, we can estimate the time lag such that the sum of squared pairwise distance is maximized.
- What about using Fourier transformation and then k-means? What about functional principal component analysis?

4. Main result.

4.1. *Consistency of k-means type method.* Pollard [13] showed the almost sure convergence, as the sample size increases, of the set of means of the k clusters. The result can be generalized to any metric space for which all the closed balls are compact would do.

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.

Consistency of the shape invariant model. Consistent in terms of both number of curves and number of observed points go to infinity.

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 $n_a = |\Gamma_a|, n_b = |\Gamma_b|$.
It seems to require the clusters are well-separated?

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] BIGOT, J. and GENDRE, X. (2013). Minimax properties of Fréchet means of discretely sampled curves. *Ann. Stat.* **41** 923–956.
[2] BONTEMPS, D. and GADAT, S. (2014). Bayesian methods for the Shape Invariant Model. *Electron. J. Stat.* **8** 1522–1568.
[3] DALEY, D. J. and SPRINGER, V.-J. *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods, Second Edition.*
[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