Clustering Random Walk Time Series

Gautier Marti[1,2], Frank Nielsen[2], Philippe Very[1], and Philippe Donnat[1]

[1] Hellebore Capital Management
[2] Ecole Polytechnique

Abstract. We present in this paper a novel non-parametric approach useful for clustering independent identically distributed stochastic processes. We introduce a pre-processing step consisting in mapping multivariate independent and identically distributed samples from random variables to a generic non-parametric representation which factorizes dependency and marginal distribution apart without losing any information. An associated metric is defined where the balance between random variables dependency and distribution information is controlled by a single parameter. This mixing parameter can be learned or played with by a practitioner, such use is illustrated on the case of clustering financial time series. Experiments, implementation and results obtained on public financial time series are online on a web portal `http://www.datagrapple.com`.

1 Introduction

Random walks are sometimes used to perform data clustering [13] or can be a point of view on spectral clustering [21, 27]. In this paper, we consider the original converse problem: clustering random walks. These stochastic processes are an important mathematical formalization used to model, for instance, the path of molecules travelling in gas, or financial market prices as stated in the random walk hypothesis [3] and the efficient-market hypothesis [12].

1.1 Clustering time series

Partition-based clustering is the task of grouping a set of objects in such a way that objects in the same group (cluster) are more similar to each other than those in different groups. This task leverages a representation of the dataset and a distance between objects. In practice, such semantic representation and distance are unknown and the ones used are motivated by some heuristics.

When working with time series, researchers have considered, for instance, L^p metrics or Dynamic Time Warping (DTW) [7] for comparing them, and wavelets [23] or SAX [19] as means of representation. These approaches were found useful to detect anomalies in time series [16] with a strong focus on pattern recognition [15].

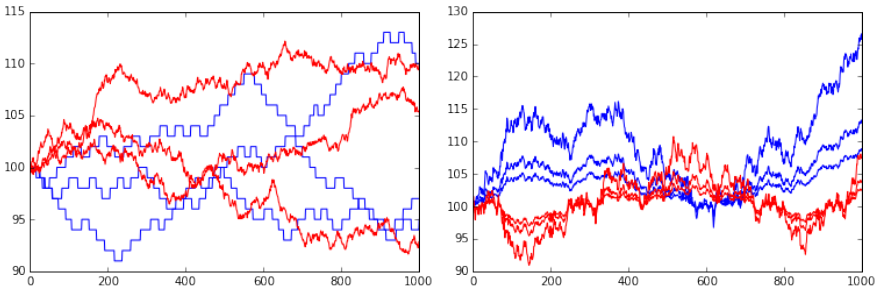

Fig. 1. Different criteria (apparently signal shape and homothetic scaling) are used for grouping these random walks in the two examples

1.2 Shortcomings of a standard approach

To understand why a standard approach fails to properly cluster random walks, we have to give a close look at the definition: a random walk is the sum $\sum_i X_i$ of a series of independent and identically distributed (i.i.d.) random variables X_i. So, there are no temporal patterns and thus approaches looking for them such as using a distance DTW and compressing time series using patterns as a way of representation are useless here. Note also that all information is carried by the increments X_i, it is therefore the underlying time series to study. By using a L^p metric between the increment time series, we may capture similarity in co-movements but, informally, we observe that we lose information of the random walk "shape", a criterion to take into account to cluster random walks as we can see it in the left panel of Figure 1. Moreover, since increments are independent and identically distributed, time does not matter in these time series and we actually consider equivalence classes of random walks consisting in all the permutations of the X_i. To cluster this special kind of time series, one cannot simply use the standard machinery of machine learning on time series. Common normalizations do not make sense either. So, this work is a first step to study the problem of clustering random walks with application to financial time series in mind [20].

To alleviate the shortcomings of a standard approach, this paper propounds in Section 2.1 a proper random walk representation capturing all information which is leveraged by a relevant distance. In Section 2.3, the approach is validated on synthetic datasets. In-depth results using the presented workflow on real and public financial time series from the credit default swap market, and implementations for reproducible research are available online (`http://www.datagrapple.com`).

2 Generic Non-Parametric Representation

We explain in this section our approach to represent and then cluster N random walks using a pre-processing we dubbed TS-GNPR for Generic Non-Parametric Representation of random walk Time Series.

2.1 Representation and distance

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. Ω is the sample space, $\mathcal{F}$ is the σ-algebra of events, and $\mathbf{P}$ is the probability measure. Let $\mathcal{V}$ be the space of all continuous real valued random variables defined on $(\Omega, \mathcal{F}, \mathbf{P})$. Let $\mathcal{U}$ be the space of random variables following a uniform distribution on $[0, 1]$ and $\mathcal{G}$ be the space of absolutely continuous cumulative distribution functions (cdf). We define the following representation of random vectors, that actually splits the joint behaviours of the marginal variables from their distributional information:

Let $\mathcal{T}$ be a mapping which transforms a random vector $X = (X_1, \ldots, X_N)$ into its TS-GNPR, an element of $\mathcal{U}^N \times \mathcal{G}^N$ representing X, defined as

$$\begin{aligned} \mathcal{T} : \mathcal{V}^N &\to \mathcal{U}^N \times \mathcal{G}^N \\ X &\mapsto (G_X(X), G_X) \end{aligned} \tag{1}$$

where $G_X = (G_{X_1}, \ldots, G_{X_N})$, and G_{X_i} being the cumulative distribution function of X_i. $\mathcal{T}$ is a bijection, and thus preserves the whole information. Note that it replicates Sklar's theorem [26], seminal result of copula theory.

Statistical distances (or non-metric divergences) have been intensively studied [4] for data processing. One important class of divergences is f-divergences that ensures the property of information monotonicity [1]. Informally, information monotonicity guarantees that the divergence between coarse-binned histograms is less than fine-binned histograms as some information are lost due to the binning process.

In our setting, which actually does not require the copula theory framework, using the generic *non-parametric* representation, we introduce artificially a separable divergence as follows: we leverage TS-GNPR by defining a distance d_θ between random variables taking into account both distributional forms and joint behaviours.

Let $(X, Y) \in \mathcal{V}^2$. Let G_X, G_Y be vectors of marginal cdf.

We define the following distance depending on $\theta \in [0, 1]$:

$$d_\theta^2(X, Y) = \theta d_1^2(G_X(X), G_Y(Y)) + (1 - \theta) d_0^2(G_X, G_Y),$$

where

$$d_1^2(G_X(X), G_Y(Y)) = 3\mathbf{E}[|G_X(X) - G_Y(Y)|^2], \tag{2}$$

and

$$d_0^2(G_X, G_Y) = \frac{1}{2} \int_{\mathbf{R}} \left(\sqrt{\frac{dG_X}{d\lambda}} - \sqrt{\frac{dG_Y}{d\lambda}} \right)^2 \mathrm{d}\lambda. \tag{3}$$

As particular cases, d_0 is the Hellinger distance, a particular f-divergence, quantifying the similarity between two probability distributions, and the distance $d_1 = \sqrt{(1-\rho_S)/2}$ is a distance correlation measuring statistical dependence between two random variables, where ρ_S is the Spearman's correlation between X and Y.

We can notice that for $\theta \in [0,1]$, $0 \leq d_\theta \leq 1$ and for $0 < \theta < 1$, d_θ is a metric. For $\theta = 0$ or $\theta = 1$, the separation axiom of metrics does not hold. This distance d_θ is invariant under diffeomorphism, i.e. let $h : \mathcal{V} \to \mathcal{V}$ be a diffeomorphism, let $(X,Y) \in \mathcal{V}^2$, we have $d_\theta(h(X), h(Y)) = d_\theta(X,Y)$. It is a desirable property as it ensures to be insensitive to scaling (e.g. choice of units) or measurement scheme (e.g. device, mathematical modelling) of the underlying phenomenon.

To apply the proposed distance on sampled data, we define a statistical estimate of d_θ: distance d_1 working with continuous uniform distributions can be approximated by rank statistics yielding to discrete uniform distributions, in fact coordinates of the multivariate empirical copula [9]; distance d_0 can be approximated using its discrete form working with estimates of marginal densities obtained from a basic kernel density estimator. For computing d_1, we need a bijective ranking function and since we consider application to time series, it is natural to choose arrival order to break ties. Let $(X_i)_{i=1}^M$ be M realizations of $X \in \mathcal{V}$. Let S_M be the permutation group of $\{1,\dots,M\}$ and let $\sigma \in S_M$ be any fixed permutation, say $\sigma = Id_{\{1,\dots,M\}}$. A bijective ranking function for $(X_i)_{i=1}^M$ can be defined as a function

$$\begin{aligned} \mathrm{rk}^X : \{1,\dots,M\} &\to \{1,\dots,M\} \\ i &\mapsto \#\{k \in \{1,\dots,M\} \mid \mathcal{P}_\sigma\} \end{aligned} \tag{4}$$

where $\mathcal{P}_\sigma \equiv (X_k < X_i) \vee (X_k = X_i \wedge \sigma(k) \leq \sigma(i))$.

Let $(X_i)_{i=1}^M$ and $(Y_i)_{i=1}^M$ be M realizations of random variables $X, Y \in \mathcal{V}$. An empirical distance between realizations of random variables can be defined by

$$\tilde{d}_\theta^2\left((X_i)_{i=1}^M, (Y_i)_{i=1}^M\right) \overset{a.s.}{=} \theta \tilde{d}_1^2 + (1-\theta)\tilde{d}_0^2, \tag{5}$$

where

$$\tilde{d}_1^2 = \frac{3}{M^2(M-1)} \sum_{i=1}^{M} \left(\mathrm{rk}^X(i) - \mathrm{rk}^Y(i)\right)^2 \tag{6}$$

and

$$\tilde{d}_0^2 = \frac{1}{2} \sum_{k=-\infty}^{+\infty} \left(\sqrt{g_X^h(hk)} - \sqrt{g_Y^h(hk)}\right)^2, \tag{7}$$

h being a suitable bandwidth, and $g_X^h(x) = \frac{1}{M}\sum_{i=1}^M \mathbf{1}\{\lfloor \frac{x}{h} \rfloor h \leq X_i < (\lfloor \frac{x}{h} \rfloor + 1)h\}$ being a density histogram estimating the probability density function g_X from $(X_i)_{i=1}^M$, M realizations of random variable $X \in \mathcal{V}$.

2.2 Parameter selection using clustering stability

To effectively use d_θ it boils down to select a particular value for θ. For instance, this value can be chosen by an expert who intends to give more weight on joint behaviours rather than distribution information, or the converse if one focuses on marginals. To aggregate both information in a balanced data-driven manner, we suggest using stability principles. Several researchers [6, 18, 25, 8] advocate that stability of some kind is a desirable property of clustering, i.e. partitions obtained should be similar while data undergo small perturbations, yet some critics have arose [17, 5] warning about the pitfalls of using stability as a method for clustering validation and model selection. In [24], authors conclude that stability is still a relevant criterion over finite samples.

Similarity between partitions To measure clustering stability, we first have to define a similarity measure between clusters, and then partitions. We consider a correlation-flavoured similarity which can be seen as the scalar product of representation vectors [10]. Given two clusters C_1 and C_2, their similarity s_C is defined by

$$s_C(C_1, C_2) = \frac{\#(C_1 \cap C_2)}{\sqrt{\#(C_1)\#(C_2)}} \tag{8}$$

where $\#(C)$ is the number of elements in a cluster C. Given two partitions $\mathcal{P}_1$ and $\mathcal{P}_2$, with the same size K, of a dataset $\mathcal{X}$, i.e. $\mathcal{P}_i = \{C_i^k\}_{k=1}^K$ for $i \in \{1, 2\}$, and $\mathcal{X} = \biguplus_{k=1}^K C_i^k$, we define a similarity s_P between $\mathcal{P}_1$ and $\mathcal{P}_2$ by averaging the pairwise similarities between clusters from $\mathcal{P}_1$ and $\mathcal{P}_2$, where each cluster in $\mathcal{P}_1$ is optimally assigned to a cluster in $\mathcal{P}_2$ with respect to maximizing the average cluster similarity, i.e.

$$s_P(P_1, P_2) = \max_{\sigma \in S_K} \frac{1}{K} \sum_{k=1}^{K} s_C(C_1^k, C_2^{\sigma(k)}) \tag{9}$$

Hungarian algorithm [22] is used to find the best assignment σ between the clusters from $\mathcal{P}_1$ and $\mathcal{P}_2$.

Time stability of a clustering Many different kind of data perturbations can be considered, a popular one being the bootstrap [11] as it preserves the statistical properties of the initial sample. In the context of time series context, it seems more natural to consider perturbations due to a time-sliding window. In a steady regime, practitioners want their model stable with respect to passing time. Since increments of the random walks are i.i.d. this perturbation also preserves the data statistical properties.

To define time stability, we suggest to apply a clustering algorithm at different periods and compute the partition similarities between the resulting clusterings. More precisely, we propose to apply the same clustering algorithm to a sliding window, compute all the similarities between partitions of two successive windows and finally average all of them.

Let $X = (x_1, \ldots, x_N)^\top$ be a matrix describing N time series, where each x_i is a vector in $\mathbf{R}^T$ and T is the time horizon under focus. Given a window of width H, we note $\mathcal{P}_H^K(t)$ the partition computed by a given clustering algorithm on the window $]t-H, t]$. Given a number of cluster K, a window width H, and a time step δt, the stability index is defined by

$$S_I(X, K, H, \delta t) = \frac{1}{W} \sum_{t=H}^{T} s_{\mathcal{P}}(\mathcal{P}_H^K(t), \mathcal{P}_H^K(t+\delta t)) \tag{10}$$

where $W = \left\lfloor \frac{T-H}{\delta t} \right\rfloor + 1$ is the number of slidings.

Stability index for model selection We present a simple example where time series are aggregated using a one level factorial model:

$\forall i \in [1, N]$, $\forall t \in [1, T]$,

$$x_i(t) = \sqrt{\rho_m} m(t) + \sqrt{\rho_k} f_{k(i)}(t) + \sqrt{\rho_s} \epsilon_i(t) \tag{11}$$

where m, $(f_k)_{k=1}^K$ and $(\epsilon_i)_{i=1}^N$ are multivariate uncorrelated Gaussian noises, $\rho_m, \rho_k \geq 0$, such that $\rho_m + \rho_k \leq 1$, and $\rho_s = 1 - \rho_m - \rho_k$.

In economical terms, m is a systemic factor that correlates all the x_i together whereas $(f_k)_{k=1}^K$ are sectoral factors that lead to the grouping of the time series in K clusters. Finally, $(\epsilon_i)_{i=1}^N$ are residual noises that decrease pairwise correlations. Two series x_i and x_j belong to the same clusters if they share the same sectoral factor, that is if $k(i) = k(j)$.

Here we choose $K = 10$ clusters among $N = 100$ time series, for an horizon $T = 500$. Time series are evenly distributed among the factors, forming clusters of size $\frac{N}{K} = 10$. We choose $\rho_m = 40\%$ and $\rho_k = 30\%$. We compute our stability index with a window of size $H = \frac{T}{2} = 250$ and a time step $\delta t = 5$ and obtain the results shown in Figure 2. We see that the stability index is equal to 1 for degenerated cases $K = 1$ and $K = N$ but also for the ground truth $K = 10$ clusters. This stability index usefulness depends on the signal-to-noise ratio $\sqrt{\rho_k/(1-\rho_k)}$, usually small in applications, and the length of the time series, usually finite horizon in applications, to obtain a good estimate. The mentioned bias which is obvious for $K = 1$ or $K = N$ exists for all values of K. We look for an estimate of this bias by computing the stability score on purely Gaussian noise and obtain the following stability curve plotted on Figure 3.

We thus propose the following adjusted stability index by subtracting this estimate. Given a set of time series X and a multivariate Gaussian noise $\mathcal{N}$, we define the adjusted stability index by

$$AS_I(X, K, H, \delta t) = \frac{S_I(X, K, H, \delta t) - S_I(\mathcal{N}, K, H, \delta t)}{1 - S_I(N, K, H, \delta t)} \tag{12}$$

$\theta^\star$ can be estimated similarly with this stability index.

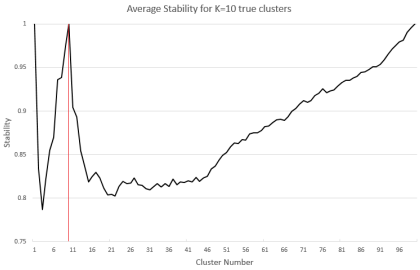

Fig. 2. Using time stability we accurately detect 10 clusters

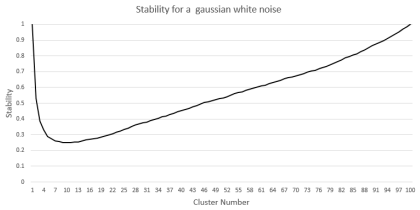

Fig. 3. Estimate of the stability index bias obtained on purely Gaussian noise

2.3 Validation and experiments

To benchmark our approach, we use the following generative model that generalizes the one presented in Section 2.2. Let $S \in \mathbf{N}$. Let $(K_1, \ldots, K_S) \in \mathbf{N}^S$. Let $(Y_k^s)_{k=1}^{K_s}$, $1 \leq s \leq S$, be i.i.d. random variables following a standard normal distribution. Let $p, K \in \mathbf{N}$. Let $N = pK \prod_{s=1}^{S} K_s$. Let $(Z_k^i)_{k=1}^{K}$, $1 \leq i \leq N$, be independent random variables. For $1 \leq i \leq N$, we define

$$X_i = \sum_{s=1}^{S} \sum_{k=1}^{K_s} \beta_{k,i}^s Y_k^s + \sum_{k=1}^{K} \alpha_{k,i} Z_k^i, \tag{13}$$

where $\alpha_{k,i} = 1$, if $i \equiv k - 1 \pmod{K}$, 0 otherwise; $\beta_k^s \in [0, 1]$, $\beta_{k,i}^s = \beta_k^s$, if $\lceil iK_s/N \rceil = k$, 0 otherwise. $(X_i)_{i=1}^N$ are partitioned into $Q = K \prod_{s=1}^{S} K_s$ clusters of p random variables each. Playing with the model parameters, we define in Table 2 some interesting test case datasets to study distribution clustering, dependence clustering or a mix of both. For clarity, we set $S \leq 2$ and $K \leq 4$, and use the following notations as a shorthand: $\mathcal{N} := \mathcal{N}(0, 1)$; $\mathcal{J} := \sum_{n \geq 0} (-1)^n \mathbf{1}_{\{t=T_n\}}$, with $T_n = \sum_{i=1}^{n} X_i$ and $X_i \sim \text{Pois}(\lambda)$ are i.i.d., with $\lambda = 5$; $\mathcal{L} := \text{Laplace}(0, 1/\sqrt{2})$; $\mathcal{S} := \text{t-distribution}(3)/\sqrt{3}$.

Our approach is essentially not algorithm dependent as can be seen in Table 1 where k-means++ [2] and Ward, a hierarchical clustering, algorithms have the same behaviour on datasets A, B and C which are described in Table 2. As expected algorithms working on standard representation, and TS-GNPR with $\theta = 1$ (working only on rank correlations) cannot retrieve distribution information which is the only information present in dataset A, whereas TS-GNPR with $\theta = 0$ (working only on distributions) or estimated $\theta^\star$ (working on an optimal mix of co-movements and distributions) can. On dataset B containing only co-movements information, all approaches but expectedly TS-GNPR with $\theta = 0$ perform accurately. Nonetheless, when distribution and dependence information are mixed (dataset C), only TS-GNPR with $\theta^\star$ can recover the ground truth. Notice that TS-GNPR with $\theta = 1$ achieves a much better Adjusted Rand Index (ARI) [14] than the standard representations (0.72 against 0.45) which shows

that working on a proper representation, even if only a part of the total information is available, is a better practice than working directly on the time series where heavy-tailed distributions can obfuscate the dependence relations between them.

Table 1. Comparative results for test case datasets

		Adjusted Rand Index		
Algo.	Representation	A	B	C
Ward	X	0	**0.94**	0.42
	$(X - \mu_X)/\sigma_X$	0	**0.94**	0.42
	$(X - \min)/(\max - \min)$	0	0.48	0.45
	TS-GNPR $\theta = 0$	**1**	0	0.47
	TS-GNPR $\theta = 1$	0	**0.91**	0.72
	TS-GNPR $\theta^\star$	**1**	**0.92**	**1**
k-m++	X	0	**0.90**	0.44
	$(X - \mu_X)/\sigma_X$	0	**0.91**	0.45
	$(X - \min)/(\max - \min)$	0.11	0.55	0.47
	TS-GNPR $\theta = 0$	**1**	0	0.53
	TS-GNPR $\theta = 1$	0.06	**0.99**	0.80
	TS-GNPR $\theta^\star$	**1**	**0.99**	**1**

Table 2. Model parameters for some interesting test case datasets

Dataset	Clustering	N	M	Q	K_1	β_k^1	K_2	β_k^2	Z_1^i	Z_2^i	Z_3^i	Z_4^i
A	Distribution	400	10000	4	0	0	0	0	$\mathcal{N}$	$\mathcal{J}$	$\mathcal{L}$	$\mathcal{S}$
B	Dependence	300	500	30	3	0.1	10	0.1	$\mathcal{N}$	$\mathcal{N}$	$\mathcal{N}$	$\mathcal{N}$
C	Mix	100	1000	20	0	0	10	0.1	$\mathcal{N}$	$\mathcal{N}$	$\mathcal{J}$	$\mathcal{J}$

3 Discussion

The aim over the long term is to design a full framework for the study of random walks in finite samples which will tackle multivariate inference and outlier detection based on clustering dynamics. The presented work was but a first step toward this goal by allowing us to have a proper data representation with a model selection based on a criterion that is dear to practitioners in finance, i.e. time stability. To complete this work, it remains to show that clustering using TS-GNPR could achieve consistency in simple factorial models where correlation matrices are slightly perturbed. We might also wish to improve the distance working on the TS-GNPR representation as we may want to compare distributions differently by taking into account, for instance, tail dependence.

References

[1] Amari, S.I., Cichocki, A.: Information geometry of divergence functions. Bulletin of the Polish Academy of Sciences: Technical Sciences 58(1), 183–195 (2010)

[2] Arthur, D., Vassilvitskii, S.: k-means++: The advantages of careful seeding. In: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. pp. 1027–1035. Society for Industrial and Applied Mathematics (2007)

[3] Bachelier, L.: Théorie de la spéculation. Gauthier-Villars (1900)

[4] Basseville, M.: Divergence measures for statistical data processing. Signal Processing 93(4), 621–633 (2013)

[5] Ben-David, S., Von Luxburg, U., Pál, D.: A sober look at clustering stability. In: Learning theory, pp. 5–19. Springer (2006)

[6] Ben-Hur, A., Elisseeff, A., Guyon, I.: A stability based method for discovering structure in clustered data. In: Pacific symposium on biocomputing. vol. 7, pp. 6–17 (2001)

[7] Berndt, D.J., Clifford, J.: Using dynamic time warping to find patterns in time series. In: KDD workshop. vol. 10, pp. 359–370. Seattle, WA (1994)

[8] Carlsson, G., Mémoli, F.: Characterization, stability and convergence of hierarchical clustering methods. The Journal of Machine Learning Research 11, 1425–1470 (2010)

[9] Deheuvels, P.: La fonction de dépendance empirique et ses propriétés. Un test non paramétrique d'indépendance. Acad. Roy. Belg. Bull. Cl. Sci.(5) 65(6), 274–292 (1979)

[10] Ding, C., He, X.: K-means clustering via principal component analysis. In: Proceedings of the twenty-first international conference on Machine learning. p. 29. ACM (2004)

[11] Efron, B.: Bootstrap methods: another look at the jackknife. The annals of Statistics pp. 1–26 (1979)

[12] Fama, E.F.: The behavior of stock-market prices. Journal of business pp. 34–105 (1965)

[13] Harel, D., Koren, Y.: On clustering using random walks. In: FST TCS 2001: Foundations of Software Technology and Theoretical Computer Science, pp. 18–41. Springer (2001)

[14] Hubert, L., Arabie, P.: Comparing partitions. Journal of classification 2(1), 193–218 (1985)

[15] Ivanov, P.C., Rosenblum, M.G., Peng, C., Mietus, J., Havlin, S., Stanley, H., Goldberger, A.L.: Scaling behaviour of heartbeat intervals obtained by wavelet-based time-series analysis. Nature 383(6598), 323–327 (1996)

[16] Keogh, E., Lin, J., Fu, A.: Hot sax: Efficiently finding the most unusual time series subsequence. In: Data mining, fifth IEEE international conference on. pp. 8–pp. IEEE (2005)

[17] Krieger, A.M., Green, P.E.: A cautionary note on using internal cross validation to select the number of clusters. Psychometrika 64(3), 341–353 (1999)

[18] Lange, T., Roth, V., Braun, M.L., Buhmann, J.M.: Stability-based validation of clustering solutions. Neural computation 16(6), 1299–1323 (2004)

[19] Lin, J., Keogh, E., Lonardi, S., Chiu, B.: A symbolic representation of time series, with implications for streaming algorithms. In: Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery. pp. 2–11. ACM (2003)

[20] Marti, G., Very, P., Donnat, P.: Toward a generic representation of random variables for machine learning. arXiv preprint arXiv:1506.00976 (2015)
[21] Meila, M., Shi, J.: A random walks view of spectral segmentation. In: AI and STATISTICS (AISTATS) (2001)
[22] Munkres, J.: Algorithms for the assignment and transportation problems. Journal of the Society for Industrial & Applied Mathematics 5(1), 32–38 (1957)
[23] Percival, D.B., Walden, A.T.: Wavelet methods for time series analysis, vol. 4. Cambridge University Press (2006)
[24] Shamir, O., Tishby, N.: Cluster stability for finite samples. In: NIPS (2007)
[25] Shamir, O., Tishby, N.: Model selection and stability in k-means clustering. In: Learning theory (2008)
[26] Sklar, M.: Fonctions de répartition à n dimensions et leurs marges. Université Paris 8 (1959)
[27] Von Luxburg, U.: A tutorial on spectral clustering. Statistics and computing 17(4), 395–416 (2007)