

# Time Series Clustering based on Factor Modeling

Xiang Zhu

*The University of Chicago, Chicago, USA.*

**Summary.** [TBA]

*Keywords:* [TBA]

## 1. Introduction

[Factor modelling:] In this paper we deal the clustering of their variates based on the factor modelling of the multiple time series. Factor modelling multivariate time series arises from a dimension-reduction viewpoint. Differently from the factor analysis for independent samples, we search for the factors which drive the serial dependence of the original time series. The factor model used in this paper can be traced back at least to that of Peña and Box (1987). We decompose a multivariate time series into two parts: a dynamic part driven by a lower-dimensional factor time series and a static part which is a vector white noise.

[High-dimensional time series:] High-dimensional time series data are often encountered nowadays in many fields including finance, economics, environmental and medical studies. For instance, understanding the dynamics of the returns of large number of assets is the key for asset pricing, portfolio allocation and risk management. Environmental time series are often of high-dimension due to a large number of indices monitored across many different locations.

[Clustering:] Clustering is an unsupervised learning procedure to identify hidden structure in the data set without ground truth labels. Data are organized into homogeneous classes where the within-group similarity and the between-group dissimilarity are maximized. When the data are time series, clustering would be useful to detect a few representative temporal patterns, quantify the affinity through the time or make forecasting. There are several features of time series which increase the difficulty of clustering. First of all, time series data are dynamic in essence and there are natural orders inside, so the temporal evolution should be considered in the analysis of similarities between series. In addition, time series databases are usually formed by large amounts of records and most of the standard clustering algorithms do not work efficiently on high-dimensional data. Time series clustering has become a very active research area nowadays. Refer the detailed and extensive review on clustering of time series by Liao (2005).

[Structure:] The rest of the paper is organized as follows. Section 2 first summarizes the factor modelling method for multiple time series (Lam and Yao, 2012) and then introduce the clustering approach based on the estimated factor loading matrix. Section 3 presents the performance comparisons between our method and several other existing methods both simulated dataset and classic test datasets for time series clustering. Two real data applications are discussed in section 4

## 2. Method

### 2.1. Clustering based on factor model

Suppose that there is some linear dynamic structure in the original  $p$ -variate time series  $\{\mathbf{y}_t\}_{t \in [T]}$ ,  $[T] = \{1, \dots, T\}$ . Conceptually,  $\mathbf{y}_t$  can be decomposed into two parts, a static part (i.e. white noise) and a dynamic part driven by a low-dimensional process

$$\mathbf{y}_t = \mathbf{A}\mathbf{x}_t + \boldsymbol{\epsilon}_t, \quad (1)$$

where  $\mathbf{x}_t$  is an  $r \times 1$  latent process with (unknown)  $r \leq p$ ,  $\mathbf{A}$  is a  $p \times r$  unknown constant matrix, and  $\boldsymbol{\epsilon}_t \sim \text{WN}(\boldsymbol{\mu}_\epsilon, \boldsymbol{\Sigma}_\epsilon)$  is a  $p$ -dimensional white-noise process. When  $r$  is much smaller than  $p$ , the serial dependence of  $\mathbf{y}_t$  is described by a lower-dimensional process  $\mathbf{x}_t$ , which is called a factor process. An effective dimension reduction is thus achieved in this way. The factor model (1) above can be traced back at least to Peña and Box (1987). For each time  $t \in [T]$  and each variate  $i \in [p]$ ,

$$y_{ti} = a_{i1}x_{t1} + a_{i2}x_{t2} + \dots + a_{ir}x_{tr},$$

where  $y_{ti}$  and  $x_{tj}$  are the  $i$ th and  $j$ th components of  $\mathbf{y}_t$  and  $\mathbf{x}_t$  respectively. Factor modelling of  $\{\mathbf{y}_t\}_{t \in [T]}$  implies that the temporal dynamics of its  $i$ th variate is totally determined by the  $i$ th row  $A_i$  of  $\mathbf{A}$ . In this work, the similarity between time series in different variates is calibrated by their hidden linear dynamic structures. More precisely, variates  $m$  and  $n$  in  $\{\mathbf{y}_t\}$  are similar if the factor loadings  $A_m$  and  $A_n$  are similar. Hence, clustering of  $p$  time series is simplified as clustering  $p$  row vectors in the loading matrix  $\mathbf{A}$ . The key for the clustering  $p$ -dimensional time series  $\mathbf{y}_t$  is to estimate the loading matrix  $\mathbf{A}$ . In a sense, the loading matrix conveys all useful information about the stochastic structure of the process since [WHAT?]. Then a measure of structural diversity between two variates  $i$  and  $j$  in multiple time series  $\mathbf{y}_t$  can be obtain by comparing the respective loadings  $A_i$  and  $A_j$  and a metric of similarity between these two variates are defined as

$$d(i, j) = \sqrt{\sum_{k=1}^r (a_{ik} - a_{jk})^2}.$$

[Check the properties of this metric!]. Since the white noise exhibits no serial correlations, the decomposition in the factor modelling is unique in the sense that both the number of factors (the dimension of the factor process) and the factor loading space are identifiable. (Rotation difference?) Thus,  $d(i, j)$  always exists and satisfies the classical properties of a distance, i.e. non-negativity, symmetry and triangular inequality.

### 2.2. Factor modelling for time series

Lam and Yao (2012), Lam et al. (2011).

To make the factor model (1) identifiable, several assumptions are required.

- 1 No linear combinations of  $\mathbf{x}_t$  are white noise: otherwise such components can be absorbed into  $\boldsymbol{\epsilon}_t$ .
- 2 The rank of  $\mathbf{A}$  is  $r$ : otherwise the model can be expressed equivalently in terms of a lower-dimensional factor.

3 Columns of  $\mathbf{A} := (\mathbf{a}_1, \dots, \mathbf{a}_r)$  are orthonormal: even though  $\mathbf{A}$  and  $\mathbf{x}_t$  are not uniquely determined but the *factor loading space*  $\mathcal{M}(\mathbf{A})$ , that is, the  $r$ -dimensional linear space spanned by the columns of  $\mathbf{A}$ , is uniquely defined.

The key for the inference for factor model is to determine the number of factors  $r$  and to estimate the  $p \times r$  factor loading matrix  $\mathbf{A}$ , or more precisely the factor loading space  $\mathcal{M}(\mathbf{A})$ . This paper assumes that  $\mathbf{x}_t$  is weakly stationary and the future white-noise components are uncorrelated with the factors up to the present. (Note that  $\mathbf{x}_t$  and  $\epsilon_s$  are assumed to be uncorrelated for any  $t$  and  $s$  in most factor modelling literature). Put

$$\Sigma_y(k) = \text{Cov}(\mathbf{y}_{t+k}, \mathbf{y}_t), \quad \Sigma_x(k) = \text{Cov}(\mathbf{x}_{t+k}, \mathbf{x}_t), \quad \Sigma_{x\epsilon} = \text{Cov}(\mathbf{x}_{t+k}, \epsilon_t).$$

For prescribed integer  $k_0 \geq 1$ , define

$$\mathbf{M} = \sum_{k=1}^{k_0} \Sigma_y(k) \Sigma_y^\top(k).$$

Consider  $p \times (p-r)$  matrix  $\mathbf{B} = (\mathbf{b}_1, \dots, \mathbf{b}_{p-r})$  associated with the factor loading matrix  $\mathbf{A}$  for which  $(\mathbf{A}, \mathbf{B})$  forms a  $p \times p$  orthogonal matrix:  $\mathbf{B}^\top \mathbf{A} = \mathbf{0}$  and  $\mathbf{B}^\top \mathbf{B} = \mathbf{I}_{p-r}$ . Simple algebra yields that  $\mathbf{M}\mathbf{B} = \mathbf{0}$ , implying that the columns of  $\mathbf{B}$  are the eigenvectors of  $\mathbf{M}$  corresponding to zero-eigenvalues. Thus, *the factor loading space  $\mathcal{M}(\mathbf{A})$  is spanned by the eigenvectors of  $\mathbf{M}$  corresponding to its non-zero eigenvalues and the number of the nonzero eigenvalues is  $r$ .* The sum-quantity  $\mathbf{M}$  accumulates the information from different time lags and this is useful when the sample size  $n$  is small. Nonnegative definite matrix  $\Sigma_y(k) \Sigma_y^\top(k)$  is used to avoid cancellation of information from different time lags. Small values of  $k_0$  are favoured, since the autocorrelation is often at its strongest at the small time lags. Estimation for  $\Sigma_y(k)$  with larger  $k$  is less accurate. To estimate  $r$  and  $\mathcal{M}(\mathbf{A})$ , just perform an eigenanalysis on

$$\hat{\mathbf{M}} = \sum_{k=1}^{k_0} \hat{\Sigma}_y(k) \hat{\Sigma}_y^\top(k),$$

where  $\hat{\Sigma}_y(k)$  denotes the sample covariance matrix of  $\mathbf{y}_t$  at time lag  $k$

$$\hat{\Sigma}_y(k) = \frac{1}{T-k} \sum_{t=1}^{T-k} (\mathbf{y}_{t+k} - \bar{\mathbf{y}}) \cdot (\mathbf{y}_t - \bar{\mathbf{y}})^\top, \quad \bar{\mathbf{y}} = \frac{1}{T} \sum_{t=1}^T \mathbf{y}_t.$$

The estimation of factor model is run as follows.

(a) A ratio-based estimator for the number of factors  $r$ :

$$\hat{r} = \arg \min_{1 \leq i \leq R} \frac{\hat{\lambda}_{i+1}}{\hat{\lambda}_i},$$

where  $\hat{\lambda}_1 \geq \dots \hat{\lambda}_p$  are the eigenvalues of  $\hat{\mathbf{M}}$  and  $r \leq R \leq p$  is a constant. (In practice take  $R = p/2$ .)

(b) The columns of the estimated factor loading matrix  $\hat{\mathbf{A}}$  are the  $\hat{r}$ th orthonormal eigenvectors of  $\hat{\mathbf{M}}$  corresponding to its  $\hat{r}$  largest eigenvalues.

(c) Estimated factor process and resulting residuals are  $\hat{\mathbf{x}}_t = \hat{\mathbf{A}}^\top \mathbf{y}_t$ ,  $\hat{\epsilon}_t = (\mathbf{I}_p - \hat{\mathbf{A}} \hat{\mathbf{A}}^\top) \mathbf{y}_t$ .

(d) Dynamic modelling of  $\hat{\mathbf{y}}_t$  is  $\hat{\mathbf{y}}_t = \hat{\mathbf{A}} \hat{\mathbf{x}}_t$ .

### 3. Numeric Studies

#### 3.1. Clustering quality evaluation criteria

Evaluating clustering systems is not a trivial task because clustering is an unsupervised learning process in the absence of the information of the actual partitions. We used classified datasets and compared how good the clustered results fit with the data labels. As is in Zhang et al. (2005), five objective clustering evaluation criteria were used in all of the numerical experiments: Jaccard score, Rand index, Folkes and Mallow index (Halkidi et al., 2001), clustering similarity measure (Gavrilov et al., 2000) and normalized mutual information (Strehl and Ghosh, 2002).

Consider  $\mathcal{G} = \{G_1, G_2, \dots, G_M\}$  as the clusters from a supervised dataset and  $\mathcal{A} = \{A_1, A_2, \dots, A_M\}$  as that obtained by a clustering algorithm under evaluations. Denote  $D$  as a dataset of original time series. For all pairs of the series, count the following quantities:

- a: the number of pairs belonging to the same cluster in both  $\mathcal{G}$  and  $\mathcal{A}$ ;
- b: the number of pairs belonging to the same cluster in  $\mathcal{G}$  but not in  $\mathcal{A}$ ;
- c: the number of pairs belonging to the same cluster in  $\mathcal{A}$  but not in  $\mathcal{G}$ ;
- d: the number of pairs belonging to different clusters in both  $\mathcal{G}$  and  $\mathcal{A}$ .

The used clustering evaluation criteria are defined as follows

- 1 Jaccard score

$$\text{Jaccard}(\mathcal{G}, \mathcal{A}) = \frac{a}{a + b + c};$$

- 2 Rand index

$$\text{Rand}(\mathcal{G}, \mathcal{A}) = \frac{a + d}{a + b + c + d};$$

- 3 Folkes and Mallow index

$$\text{FMI}(\mathcal{G}, \mathcal{A}) = \sqrt{\frac{a}{a + b} \cdot \frac{a}{a + c}};$$

- 4 Cluster similarity measure

$$\text{CSM}(\mathcal{G}, \mathcal{A}) = \frac{1}{M} \sum_{i=1}^M \max_{1 \leq j \leq M} \text{Sim}(G_i, A_j), \text{ where } \text{Sim}(G_i, A_j) = \frac{2|G_i \cap A_j|}{|G_i| + |A_j|};$$

- 5 Normalized Mutual Information

$$\text{NMI}(\mathcal{G}, \mathcal{A}) = \frac{\sum_{i=1}^M \sum_{j=1}^M N_{i,j} \log \left( \frac{N \cdot N_{i,j}}{|G_i| \cdot |A_j|} \right)}{\left( \sum_{i=1}^M |G_i| \log \left( \frac{|G_i|}{N} \right) \right)^{1/2} \cdot \left( \sum_{j=1}^M |A_j| \log \left( \frac{|A_j|}{N} \right) \right)^{1/2}}$$

where  $N$  is the total number of time series in the dataset and  $N_{i,j} = |G_i \cap A_j|$ .

All the used clustering evaluation criteria have value ranging from 0 to 1, where 1 corresponds to the case when  $\mathcal{G}$  and  $\mathcal{A}$  are identical. The larger the values of criteria, the more similar between  $\mathcal{A}$  and  $\mathcal{G}$ .

### 3.2. Simulation Experiment

For each replication of the experiment, the dataset  $S$  subjected to clustering was formed by one partial realization of length  $T = 200$  simulated from each of the autoregressive models enumerated in Table ???. At each trial of the experiment, one series of length  $T = 200$  were generated from each of models M1 to M6. These models were considered in (Vilar et al., 2010) and by other authors in previous works. The dimension of the factor process is six. The algorithm of Hartigan and Wong (1979) is used in the implementation of  $k$ -means clustering. Since the clustering results of  $k$ -means clustering depend on the initial clustering centers that shoue be randomly initialized in each run, we run  $k$ -means 100 times with random initialized centres every time. For each iteration, the number of random start for  $k$ -means algorithm is one. For  $AR(\infty)$  in this part of experiment, the maximum order of model to fit is set as the minimum of  $T - 1$  and  $10 \times \text{round}(\log_{10} T)$  and the estimation method is Yule-Walker.

M1	AR	$X_t = 0.6X_{t-1} + \epsilon_t$
M2	Bilinear	$X_t = (0.3 - 0.2\epsilon_{t-1}) \cdot X_{t-1} + 1 + \epsilon_t$
M3	EXPAR	$X_t = (0.9 \exp\{-X_{t-1}^2\} - 0.6)X_{t-1} + 1 + \epsilon_t$
M4	SETAR	$X_t = (0.3X_{t-1} + 1) \cdot \text{sgn}(X_{t-1} - 0.2) + \epsilon_t$
M5	NLAR	$X_t = 0.7 X_{t-1}  \cdot (2 +  X_{t-1} )^{-1} + \epsilon_t$
M6	STAR	$X_t = 0.8X_{t-1} - 0.8X_{t-1} \cdot (1 + \exp\{-10X_{t-1}\})^{-1} + \epsilon_t$

There are four types of dynamics in the loading matrix  $\mathbf{A}$ , that is, each row of  $\mathbf{A}$  equals to one the four types in Table ???. There are  $k$  rows for each type in the loading matrix  $\mathbf{A}$ . Thus,  $p = 4 \times k$ . The number of classes in this simulated dataset is 4 and the ground-truth labels are known.

Type I	(1,1,1,1,1,1)	Type II	(1,0,0,0,1,0)
Type III	(0,1,1,0,0,0)	Type IV	(0,0,0,1,0,1)

### 3.3. Evaluation of classical datasets

Ground truth of classification are available in these classical datasets. In this part, the clustering procedure based on factor modelling (FM) is evaluated on XXX datasets. Six classified datasets in UCR Time Series Classification/Clustering Page (Keogh et al., 2011). The reason why these six datasets are chosen is that .....**Explain it!**

[UCR Dataset description!]

Euclidean distance is used for both  $k$ -means and hierarchical clustering algorithms. The  $k$ -means clustering was conducted 100 times with random initialized centres for each experiment. As each run of hierarchical clustering for the same dataset always gets the same result, multiple runs are not needed in hierarchical clustering procedures. We use ward linkage (Ward, 1963) for the hierarchical clustering algorithm in our experiment. Table XXX lists the mean of evaluation criteria values of 100 runs for  $k$ -means with the estimated factor loading matrix (FM), the original data (OR), the extracted features using wavelet transformation (DW) and the matrix of  $AR(\infty)$  coefficients (AP) for each datasets. Table XXX presents the evaluation criteria values produced by these four methods for each datasets. **Methods compared here:**

DW An unsupervised feature extraction was performed for time series clustering using orthogonal wavelet transform (Zhang et al., 2005). The original time series could be well approximated

Table 1. Simulation Experiment Results																					
$T$	$p$	Rand				Jaccard				FMI				CSM				NMI			
		FM	OR	DW	AP	FM	OR	DW	AP	FM	OR	DW	AP	FM	OR	DW	AP	FM	OR	DW	AP
200	0.2T	.896	.895	.908	.886	.682	.724	.744	.718	.793	.829	.844	.824	.850	.856	.872	.850	.814	.872	.886	.863
	0.5T	.902	.905	.894	.901	.708	.741	.712	.750	.811	.842	.825	.846	.854	.859	.843	.861	.806	.880	.867	.878
	2T	.930	.894	.913	.886	.774	.729	.768	.730	.860	.834	.859	.831	.885	.842	.872	.838	.859	.867	.890	.859
	5T	.955	.903	.911	.908	.858	.749	.762	.760	.910	.846	.855	.853	.921	.853	.861	.859	.907	.878	.887	.884
400	0.2T	.937	.899	.916	.893	.806	.745	.771	.732	.875	.842	.860	.834	.908	.859	.876	.850	.873	.876	.895	.869
	0.5T	.923	.902	.905	.885	.773	.750	.744	.736	.854	.846	.844	.834	.881	.858	.855	.842	.849	.878	.880	.860
	2T	.963	.886	.892	.895	.880	.717	.712	.741	.926	.825	.825	.840	.937	.832	.834	.844	.925	.858	.864	.869
	5T	.948	.912	.914	.887	.831	.768	.771	.721	.895	.858	.860	.827	.912	.863	.866	.831	.894	.888	.891	.859
800	0.2T	.921	.893	.905	.899	.754	.722	.743	.754	.843	.830	.843	.847	.872	.843	.857	.858	.840	.867	.880	.875
	0.5T	.933	.902	.894	.883	.785	.743	.716	.706	.868	.843	.827	.819	.894	.853	.838	.828	.866	.877	.866	.855
	2T	.955	.892	.899	.912	.859	.726	.729	.774	.910	.832	.835	.861	.922	.837	.842	.866	.907	.865	.871	.889
	5T	.949	.910	.903	.917	.835	.764	.742	.787	.897	.856	.843	.869	.907	.860	.847	.872	.896	.886	.877	.895
1600	0.2T	.955	.900	.903	.870	.852	.753	.741	.700	.909	.847	.842	.812	.927	.855	.854	.818	.907	.876	.878	.841
	0.5T	.980	.915	.912	.880	.937	.777	.762	.719	.961	.864	.855	.824	.966	.871	.862	.828	.960	.893	.887	.852
	2T	.950	.888	.902	.892	.838	.719	.737	.726	.899	.827	.840	.832	.910	.832	.845	.836	.899	.860	.875	.865
	5T	.949	.894	.909	.892	.836	.730	.758	.733	.898	.834	.853	.835	.910	.838	.857	.837	.897	.867	.885	.865
3200	0.2T	.928	.894	.902	.871	.766	.730	.737	.719	.856	.834	.840	.822	.883	.843	.849	.824	.853	.867	.876	.844
	0.5T	.928	.882	.895	.880	.771	.701	.721	.711	.856	.816	.830	.821	.871	.822	.837	.824	.852	.853	.867	.852
	2T	.975	.890	.914	.892	.919	.729	.771	.726	.949	.833	.860	.831	.957	.836	.865	.834	.949	.863	.891	.865
	5T	xx	xx	xx	xx	xx	xx	xx	xx	xx	xx	xx	xx	xx	xx	xx	xx	xx	xx	xx	xx

by their wavelet coefficients in some proper dimension. The desired dimension of features is determined by minimizing the dimensionality and sum of squared errors between features and the original time series simultaneously. Time efficiency of this feature extraction method is guaranteed when Haar wavelet is used.

DP For clustering a set of dynamic structures on the class of autoregressive integrated moving-average invertible models, Piccolo (1990) used the Euclidean distance of their  $AR(\infty)$  operator coefficients to measure the structural diversity between two time series. In the experiments, the order parameter of the autoregressive model is chosen by Akaike Information Criterion and the method used to fit the model is

## 4. Data Analysis

[TBA]

## 5. Discussion

- Time series clustering  $\longrightarrow$  Ordinary clustering problem; Difficulty for clustering time series can be avoided. **Enrich this argument!**
- Assume that the dynamic part of univariate time series  $\{y_{kt}\}$  for each variate  $k$ ,  $k \in [p]$  can be described by latent factor process  $\{\mathbf{x}_t\}$  and its loading vector  $A_k$ . Find some effective test statistics to test the null hypothesis that there is no difference between the dynamic part of these two series, that is,  $H_0 : A_i = A_j$ . Two series are clustered together if the associated  $p$ -value is greater than the pre-specified significance level. **Try to find such test statistic; even if proposed in the next paper.**
- Co-clustering: consider jointly clustering. (Heard et al., 2005)

## 6. Conclusion

The general philosophy of this paper can be summarized as follows. The high-dimensional multivariate time series are decompose into two parts: a dynamic part driven by a lower-dimensional factor process and static part which is a vector white noise. In this factor model, the rows of loading matrix convey all the useful information about the stochastic dynamic structure of the original multiple process. Differences between variates in the raw data can be well approximated by those between rows in the loading matrix.

## References

- Gavrilov, M., D. Anguelov, P. Indyk, and R. Motwani (2000). Mining the stock market (extended abstract): which measure is best ? In *Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, USA, pp. 487–496. ACM.
- Halkidi, M., Y. Batistakis, and V. M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems* 17, 107–145.

**Table 2.** UCR time series data set

Dataset	k-means(100 trials)						hierarchical(ward)					
	Rand	Jaccard	FMI	CSM	NMI	Rand	Jaccard	FMI	CSM	NMI		
Trace	FM	0.7506	0.3599	0.5312	0.5585	0.5184	0.7506	0.4050	0.5844	0.5886	0.5421	
	OR	0.7501	0.3651	0.5375	0.5555	0.5198	0.7490	0.3315	0.4979	0.5367	0.5030	
	DW	0.7500	0.3609	0.5325	0.5537	0.5177	0.7490	0.3315	0.4979	0.5367	0.5030	
FaceFour	DP	0.8109	0.5242	0.6982	0.7477	0.7132	0.8394	0.5709	0.7367	0.7519	0.7841	
	FM	0.7271	0.2972	0.4583	0.5955	0.3529	0.6214	0.2132	0.3553	0.4820	0.2051	
	OR	0.7455	0.3562	0.5275	0.6471	0.4607	0.7261	0.3721	0.5500	0.6727	0.4850	
AP	DW	0.7409	0.3533	0.5247	0.6472	0.4534	0.7196	0.3487	0.5223	0.6516	0.4419	
	AP	0.6696	0.2171	0.3570	0.4761	0.1981	0.6645	0.2083	0.3450	0.4612	0.1906	
Yoga	FM	0.5002	0.3547	0.5242	0.5664	0.0019	0.5009	0.3950	0.5709	0.6062	0.0000	
	OR	0.5000	0.3692	0.5411	0.5735	0.0008	0.5013	0.3508	0.5197	0.5590	0.0043	
	DW	0.5000	0.3692	0.5411	0.5735	0.0008	0.5013	0.3508	0.5197	0.5590	0.0043	
uWGLX	AP	0.5013	0.4463	0.6339	0.6388	0.0000	0.5039	0.4180	0.5981	0.6203	0.0028	
	FM	0.8334	0.2143	0.3532	0.4675	0.3706	0.8395	0.2350	0.3809	0.4991	0.3878	
	OR	0.8557	0.2741	0.4303	0.5543	0.4404	0.8673	0.3185	0.4833	0.6103	0.5021	
uWGLY	DW	0.8562	0.2766	0.4333	0.5586	0.4416	0.8585	0.2917	0.4519	0.5376	0.4680	
	AP	0.7379	0.1132	0.2098	0.2772	0.1252	0.7771	0.1113	0.2014	0.2934	0.1282	
	FM	0.8322	0.2050	0.3403	0.4598	0.3628	0.8346	0.2469	0.3976	0.4998	0.3820	
AP	OR	0.8460	0.2620	0.4156	0.5283	0.4348	0.8387	0.2390	0.3863	0.5215	0.4137	
	DW	0.8466	0.2638	0.4180	0.5282	0.4355	0.8491	0.2899	0.4513	0.5390	0.4522	
	AP	0.7554	0.1437	0.2586	0.3304	0.2049	0.7725	0.1397	0.2489	0.3053	0.1967	



- Hartigan, J. A. and M. A. Wong (1979). Algorithm as 136: a k-means clustering algorithm. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* 28(1), 100–108.
- Heard, N. A., C. C. Holmes, D. A. Stephens, D. A. Hand, and G. Dimopoulos (2005). Bayesian coclustering of anopheles gene expression time series: study of immune defense response to multiple experimental challenges. *Proceedings of the National Academy of Sciences* 102(47), 16939–16944.
- Keogh, E., Q. Zhu, B. Hu, H. Y., X. Xi, L. Wei, and C. A. Ratanamahatana (2011). The ucr time series classification/clustering homepage.
- Lam, C. and Q. Yao (2012). Factor modelling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics* 40(2), 694–726.
- Lam, C., Q. Yao, and N. Bathia (2011). Estimation of latent factors for high-dimensional time series. *Biometrika* 98(4), 901–918.
- Liao, T. W. (2005). Clustering of time series data—a survey. *Pattern Recognition* 38, 1857–1874.
- Peña, D. and G. E. P. Box (1987). Identifying a simplifying structure in time series. *Journal of the American Statistical Association* 82, 836–843.
- Piccolo, D. (1990). A distance measure for classifying arima models. *Journal of Time Series Analysis* 11(2), 153–164.
- Strehl, A. and J. Ghosh (2002). Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research* 3(3), 583–617.
- Vilar, J. A., A. M. Alonso, and J. M. Vilar (2010). Non-linear time series clustering based on non-parametric forecast densities. *Computational Statistics and Data Analysis* 54, 2850–2865.
- Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* 58(301), 236–244.
- Zhang, H., T. B. Ho, Y. Zhang, and M. Lin (2005). Unsupervised feature extraction for time series clustering using orthogonal wavelet transform. *Informatica* 30, 305–319.