Time series clustering based on factor model

Xiang Zhu
Department of Statistics
The University of Chicago
Chicago, IL 60637
xiangzhu@uchicago.edu

1 Overview

The basic idea of this work is summarized as follows. The high-dimensional multivariate time series are decomposed into two parts: a dynamic part driven by a lower-dimensional factor process and a static part which is a vector of white noise. In the factor process, the rows of loading matrix capture most of the useful information about the stochastic dynamic structure of the original time series. Differences between variates of time series data can be approximated by the "distance" between the corresponding rows of the loading matrix. Hence, clustering the variates of time series data is achieved by clustering the rows of factor loading matrix.

2 Method

2.1 Clustering based on factor model

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

$$\mathbf{y}_t = \mathbf{A}\mathbf{x}_t + \boldsymbol{\epsilon}_t, \tag{1}$$

where $\mathbf{x}_t$ is an $r \times 1$ latent process with unknown dimension $r \leq p$, $\mathbf{A}$ is a $p \times r$ unknown constant matrix, and $\boldsymbol{\epsilon}_t$ is a p-dimensional white-noise process. The serial dependence of $\mathbf{y}_t$ is determined by the lower-dimensional process $\mathbf{x}_t$, which is called a factor process, and 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]. The factor model (1) maps the ith variate of $\mathbf{y}_t$ to the ith row of A. Specifically, consider two variates i and j at each time $t \in [T]$,

$$\begin{aligned} y_{ti} &= a_{i1}x_{t1} + a_{i2}x_{t2} + \ldots + a_{ir}x_{tr} + \epsilon_{ti} \\ y_{tj} &= a_{j1}x_{t1} + a_{j2}x_{t2} + \ldots + a_{jr}x_{tr} + \epsilon_{tj} \end{aligned}$$

where y_{ti} and x_{tk} are the ith and kth components of $\mathbf{y}_t$ and $\mathbf{x}_t$ respectively, $i \in [p]$, $k \in [r]$. We can see that, given the factor process $\mathbf{x}_t$, the temporal dynamics of the ith variate in $\{\mathbf{y}_t\}_{t\in[T]}$ is determined by its factor loading $A_i := (a_{i1}, \ldots, a_{ir})$ (the ith row of $\mathbf{A}$). Moreover, because they share the same factor process $\mathbf{x}_t$, the difference between variates i and j primarily depends on the difference between their factor loadings A_i and A_j. Based on these observations, the clustering of p variates in the time series data $\{\mathbf{y}_t\}$ can be simplified as the clustering of p row vectors in the loading matrix $\mathbf{A}$.

In our implementation, we use the Euclidean (L^2) norm as a metric of difference between the ith and jth variate:

$$d(i,j) := ||A_i - A_j||_2 = \sqrt{\sum_{k=1}^{r}(a_{ik} - a_{jk})^2}.$$

To cluster the rows of the loading matrix A, we can simply use the off-the-shelf clustering methods such as k-means and hierarchical clustering.

2.2 Factor modelling for time series

The factor modelling approach taken here is based on Lam and Yao [2012] and Lam et al. [2011], and we refer it to Lam-Yao procedure here.

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 $\mathbf{A} := (\mathbf{a_1}, \ldots, \mathbf{a_r})$ 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.
4. $\mathbf{x}_t$ is weakly stationary and the future white-noise components are uncorrelated with the factors up to the present. (Note that $\mathbf{x}_t$ and $\boldsymbol{\epsilon}_s$ are assumed to be uncorrelated for any t and s in most factor modelling literature).

For any prescribed integer $k_0 \geq 1$, define

$$\mathbf{M} = \sum_{k=1}^{k_0} \boldsymbol{\Sigma}_y(k)\boldsymbol{\Sigma}_y^{\mathsf{T}}(k), \tag{2}$$

where $\boldsymbol{\Sigma}_y(k) := \mathrm{Cov}(\mathbf{y}_{t+k}, \mathbf{y}_t)$ is the covariance matrix of $\mathbf{y}_t$ at time lag k. Consider the $p \times (p-r)$ matrix $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_{p-r})$ associated with the factor loading matrix $\mathbf{A}$ for which $(\mathbf{A}, \mathbf{B})$ forms a $p \times p$ orthogonal matrix: $\mathbf{B}^{\mathsf{T}}\mathbf{A} = \mathbf{0}$ and $\mathbf{B}^{\mathsf{T}}\mathbf{B} = \mathbf{I}_{p-r}$. It thus follows that $\mathbf{MB} = \mathbf{0}$, implying that the columns of $\mathbf{B}$ are the eigenvectors of $\mathbf{M}$ corresponding to zero-eigenvalues. The factor loading space $\mathcal{M}(\mathbf{A})$ is therefore spanned by the eigenvectors of $\mathbf{M}$ corresponding to its non-zero eigenvalues and the number of the non-zero eigenvalues is r. The sum-quantity $\mathbf{M}$ accumulates the information from different time lags and this is useful when the sample size is small. Non-negative definite matrix $\boldsymbol{\Sigma}_y(k)\boldsymbol{\Sigma}_y^{\mathsf{T}}(k)$ 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, and estimation for $\boldsymbol{\Sigma}_y(k)$ with larger k is often less accurate.

To estimate the number of factors r and the factor loading space $\mathcal{M}(\mathbf{A})$, just perform an eigen-analysis on

$$\widehat{\mathbf{M}} := \sum_{k=1}^{k_0} \widehat{\boldsymbol{\Sigma}}_y(k)\widehat{\boldsymbol{\Sigma}}_y^{\mathsf{T}}(k),$$

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

$$\widehat{\boldsymbol{\Sigma}}_y(k) := \frac{1}{T-k}\sum_{t=1}^{T-k}(\mathbf{y}_{t+k} - \overline{\mathbf{y}}) \cdot (\mathbf{y}_t - \overline{\mathbf{y}})^{\mathsf{T}}, \;\; \overline{\mathbf{y}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{y}_t.$$

The Lam-Yao procedure is summarized as follows.

1. A ratio-based estimator for the number of factors r:

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

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

2. The columns of the estimated factor loading matrix $\widehat{\mathbf{A}}$ are the $\hat{r}$th orthonormal eigenvectors of $\hat{\mathbf{M}}$ corresponding to its $\hat{r}$ largest eigenvalues.
3. The estimated factor process is $\hat{\mathbf{x}}_t = \hat{\mathbf{A}}^\top \mathbf{y}_t$, the resulting residuals are $\hat{\boldsymbol{\epsilon}}_t = (\mathbf{I}_p - \hat{\mathbf{A}}\hat{\mathbf{A}}^\top)\hat{\mathbf{y}}_t$, and the estimated dynamic component $\mathbf{y}_t$ is $\hat{\mathbf{y}}_t = \hat{\mathbf{A}}\hat{\mathbf{x}}_t$.

3 Results

Evaluating clustering methods is non-trivial because clustering is an unsupervised learning process where the information about the actual partitions is absent. We thus use pre-classified datasets and compare the clustering results with the known labels.

3.1 Evaluation criteria

Following Zhang et al. [2005], we used five objective clustering evaluation criteria to assess the clustering methods: 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].
Let $\mathcal{G} = \{G_1, G_2, \ldots, G_M\}$ denote the true clusters from a supervised dataset and $\mathcal{A} = \{A_1, A_2, \ldots, A_M\}$ denote the clusters obtained from a clustering algorithm. For all pairs of the series, count the following quantities $\{a, b, c, d\}$:

$a :=$ 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 $\mathcal{A}$.

Five clustering evaluation criteria are defined as follows

1. Jaccard score
$$\mathsf{Jacard}(\mathcal{G}, \mathcal{A}) := \frac{a}{a+b+c};$$
2. Rand index
$$\mathsf{Rand}(\mathcal{G}, \mathcal{A}) := \frac{a+d}{a+b+c+d};$$
3. Folkes and Mallow index
$$\mathsf{FMI}(\mathcal{G}, \mathcal{A}) := \sqrt{\frac{a}{a+b} \cdot \frac{a}{a+c}};$$
4. Cluster similarity measure
$$\mathsf{CSM}(\mathcal{G}, \mathcal{A}) := \frac{1}{M} \sum_{i=1}^{M} \max_{1 \le j \le M} \mathsf{Sim}(G_i, A_j), \text{ where } \mathsf{Sim}(G_i, A_j) := \frac{2|G_i \cap A_j|}{|G_i| + |A_j|};$$
5. Normalized mutual information
$$\mathsf{NMI}(\mathcal{G}, \mathcal{A}) := \frac{\sum_{i=1}^{M} \sum_{j=1}^{M} N_{i,j} \log\left(\frac{N \cdot N_{i,j}}{|G_i| \cdot |A_j|}\right)}{\left(\sum_{i=1}^{M} |G_i| \log\left(\frac{|G_i|}{N}\right)\right)^{1/2} \cdot \left(\sum_{j=1}^{M} |A_j| \log\left(\frac{|A_j|}{N}\right)\right)^{1/2}}$$
where N is the total number of time series in the dataset and $N_{i,j} = |G_i \cap A_j|$.

The five clustering evaluation criteria above range from 0 to 1, where 1 corresponds to the case when $\mathcal{G}$ and $\mathcal{A}$ are identical. Larger the values of criteria indicates higher level of similarity between $\mathcal{A}$ and $\mathcal{G}$.

3.2 Clustering methods

We considered four clustering methods here. Specifically, we applied the off-the-shelve methods (e.g. k-means or hierarchical clustering) on four types of "input" data.

1. FM: the factor loadings estimated by Lam-Yao procedure [Lam and Yao, 2012];
2. OR: the original time series data matrix;
3. DW: the Haar wavelet coefficients of the original time series [Zhang et al., 2005];
4. AP: the AR(∞) operator coefficients of the original time series [Piccolo, 1990].

3.3 Simulations

We simulated p-dimensional time series $\{\mathbf{y}_t\}$ of length T from the factor model (1), where the factor process $\{\mathbf{x}_t\}$ and loading matrix A are specified as follows. The factor process $\mathbf{x}_t$ has 6 variates, and the ith variate of $\mathbf{x}_t$ is given by the ith model in Table 1, $i \in [6]$. (These models were considered by Vilar et al. [2010] and other authors in previous work.) The loading matrix A has 4 types of dynamics, that is, each row of A is the same as one of the four types below. (Each type has the same number of rows in A.)

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)

For the simulation data, we know the number of clusters (4) and the ground-truth label (Type I-IV) of each variate in the original time series. The off-the-shelve clustering method used here is the k-means clustering [Hartigan and Wong, 1979]. For each setting of (T, p), the number of replications is 100. The results of simulations are summarized in Table 2.

Table 1: Models that are used to define the factor process in the simulation data.

Model	Name	Form
1	AR	$X_t = 0.6X_{t-1} + \epsilon_t$
2	Bilinear	$X_t = (0.3 - 0.2\epsilon_{t-1}) \cdot X_{t-1} + 1 + \epsilon_t$
3	EXPAR	$X_t = (0.9\exp\{-X_{t-1}^2\} - 0.6)X_{t-1} + 1 + \epsilon_t$
4	SETAR	$X_t = (0.3X_{t-1} + 1) \cdot \text{sign}(X_{t-1} - 0.2) + \epsilon_t$
5	NLAR	$X_t = 0.7\lvert X_{t-1}\rvert \cdot (2 + \lvert X_{t-1}\rvert)^{-1} + \epsilon_t$
6	STAR	$X_t = 0.8X_{t-1} - 0.8X_{t-1} \cdot (1 + \exp\{-10X_{t-1}\})^{-1} + \epsilon_t$

Table 2: The averaged clustering quality scores over 100 replications.

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

3.4 Real data

We also compared the clustering procedure based on factor modelling with other methods on real data. Five classified datasets were retrieved from the UCR Time Series Classification/Clustering Page [Keogh et al., 2011]. Results of this section is summarized in Table 3. The Euclidean distance was used for both k-means and hierarchical clustering algorithms. Ward linkage [Ward, 1963] was used in the hierarchical clustering algorithm. Results based on the k-means clustering were averaged over 100 trials with randomly initialized centers.

Table 3: Comparison of four clustering methods on five UCR time series datasets

		k-means (average 100 trials)					hierarchical (ward linkage)				
Dataset		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
	AP	0.8109	0.5242	0.6982	0.7477	0.7132	0.8394	0.5709	0.7367	0.7519	0.7841
FaceFour	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
	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
	AP	0.5013	0.4463	0.6339	0.6388	0.0000	0.5039	0.4180	0.5981	0.6203	0.0028
uWGLX	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
	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
uWGLY	FM	0.8322	0.2050	0.3403	0.4598	0.3628	0.8346	0.2469	0.3976	0.4998	0.3820
	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

References

M. Gavrilov, D. Anguelov, P. Indyk, and R. Motwani. 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*, pages 487–496, New York, USA, 2000. ACM.

M. Halkidi, Y. Batistakis, and V. M. On clustering validation techniques. *Journal of Intelligent Information Systems*, 17:107–145, 2001.

J. A. Hartigan and M. A. Wong. Algorithm as 136: a k-means clustering algorithm. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 28(1):100–108, 1979.

E. Keogh, Q. Zhu, B. Hu, H. Y., X. Xi, L. Wei, and C. A. Ratanamahatana. The ucr time series classification/clustering homepage, 2011. URL `www.cs.ucr.edu/~eamonn/time_series_data/`.

C. Lam and Q. Yao. Factor modelling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics*, 40(2):694–726, 2012.

C. Lam, Q. Yao, and N. Bathia. Estimation of latent factors for high-dimensional time series. *Biometrika*, 98(4):901–918, 2011.

D. Peña and G. E. P. Box. Identifying a simplifying structure in time series. *Journal of the American Statistical Association*, 82:836–843, 1987.

D. Piccolo. A distance measure for classifying arima models. *Journal of Time Series Analysis*, 11(2):153–164, 1990.

A. Strehl and J. Ghosh. Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3(3):583–617, 2002.

J. A. Vilar, A. M. Alonso, and J. M. Vilar. Non-linear time series clustering based on non-parametric forecast densities. *Computational Statistics and Data Analysis*, 54: 2850–2865, 2010.

J. H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

H. Zhang, T. B. Ho, Y. Zhang, and M. Lin. Unsupervised feature extraction for time series clustering using orthogonal wavelet transform. *Informatica*, 30:305–319, 2005.