Proceedings 1st International Workshop on Advanced Analytics and Learning on Temporal Data
AALTD 2015

An Exploratory Analysis of Multiple Multivariate Time Series

Lynne Billard[1], Ahlame Douzal-Chouakria[2], and Seyed Yaser Samadi[3]

[1] Department of Statistics, University of Georgia
[2] Université Grenoble Alpes, CNRS - LIG/AMA, France
[3] Department of Mathematics, Southern Illinois University

Abstract. Our aim is to extend standard principal component analysis for non-time series data to explore and highlight the main structure of multiple sets of multivariate time series. To this end, standard variance-covariance matrices are generalized to lagged cross-autocorrelation matrices. The methodology produces principal component time series, which can be analysed in the usual way on a principal component plot, except that the plot also includes time as an additional dimension.

1 Introduction

Time series data are ubiquitous, arising throughout economics, meteorology, medicine, the basic sciences, even in some genetic microarrays, to name a few of the myriad fields of application. Multivariate time series are likewise prevalent. Our aim is to use principal components methods as an exploratory technique to find clusters of time series in a set of S multivariate time series. For example, in a collection of stock market time series, interest may center on whether some stocks, such as mining stocks, behave alike but differently from other stocks, such as pharmaceutical stocks.

A seminal paper in univariate time series clustering is that of Košmelj and Batagelj (1990), based on a dissimilarity measure. Since then several researchers have proposed other approaches (e.g. Caiado et al (2015), D'Urso and Maharaj (2009)). A comprehensive summary of clustering for univariate time series is in Liao (2005). Liao (2007) introduced a two-step procedure for multivariate series which transformed the observations into a single multivariate series. Most of these methods use dissimilarity functions or variations thereof. A summary of Liao (2005, 2007) along with more recent proposed methods is in Billard et al. (2015). Though a few authors specify a particular model structure, by and large, the dependence information inherent to time series observations is not used.

Dependencies in time series are measured through the autocorrelation (or, equivalently, the autocovariance) functions. In this work, we illustrate how these dependencies can be used in a principal component analysis. This produces principal component time series, which in turn allows the projection of the original time series observations onto three dimensional principal component by time space. The basic methodology is outlined in Section2, and illustrated in Section 3.

2 Methodology

2.1 Cross-Autocorrelation functions for $S > 1$ series and $p > 1$ dimensions

Let $\mathbf{X}_{st} = \{(X_{stj}), j = 1, \dots, p\}$, $t = 1, \dots, N_s$, $s = 1, \dots, S$, be a p-dimensional time series of length N_s, for each series s. For notational simplicity, assume $N_s = N$ for all s. Let us also assume the observations have been suitably differenced/transformed so that the data are stationary.

For a standard single univariate series time series where $S = 1$ and $p = 1$, it is well-known that the sample autocovariance function at lag k is (dropping the $s = S = 1$ and $j = p = 1$ subscripts)

$$\hat{\gamma}(k) = \frac{1}{N} \sum_{t=1}^{N-k} (X_t - \bar{X})(X_{t+k} - \bar{X}), \quad k = 0, 1, \dots, \quad \bar{X} = \frac{1}{N} \sum_{t=1}^{N} X_t, \tag{2.1}$$

and the sample autocorrelation function at lag k is $\hat{\rho}(k) = \hat{\gamma}(k)/\hat{\gamma}(0)$, $k = 0, 1, \dots$.

These autocorrelation functions provide a measure of the time dependence between observations changes as their distance apart, lag k. They are used to identify the type of model and also to estimate model parameters. See, many of the basic texts on time series, e.g., Box et al. (2011); Brockwell and Davis (1991); Cryer and Chan (2008). Note that the divisor in Eq.(2.1) is N, rather than $N - k$. This ensures that the sample autocovariance matrix is non-negative definite.

For a single multivariate time series where $S = 1$ and $p \geq 1$, the cross-autocovariance function between variables (j, j') at lag k is the $p \times p$ matrix $\boldsymbol{\Gamma}(k)$ with elements estimated by

$$\hat{\gamma}_{jj'}(k) = \frac{1}{T} \sum_{t=1}^{T-k} (X_{tj} - \bar{X}_j)(X_{t+k,j'} - \bar{X}_{j'}), \quad k = 0, 1, \text{with } \bar{X}_j = \frac{1}{N} \sum_{t=1}^{N} X_{tj}, \tag{2.2}$$

and the cross-autocorrelation function between variables (j, j') at lag k is the $p \times p$ matrix, $\boldsymbol{\rho}(k)$, with elements $\{\rho_{jj'}(k), j, j' = 1, \ldots, p\}$ estimated by

$$\hat{\rho}_{jj'}(k) = \hat{\gamma}_{jj'}(k)/\{\hat{\gamma}_{jj}(0)\hat{\gamma}_{j'j'}(0)\}^{1/2}, \quad k = 0, 1, \ldots. \tag{2.3}$$

Unlike the autocorrelation function obtained from Eq.(2.1) with its single value at each lag k, Eq.(2.3) produces a $p \times p$ matrix at each lag k. The function Eq.(2.2) was first given by Whittle (1963) and shown to be nonsymmetric by Jones (1964). In general, $\rho_{jj'}(k) \neq \rho_{j'j}(k)$ for variables $j \neq j'$, except for $k = 0$, but $\boldsymbol{\rho}(k) = \boldsymbol{\rho}'(-k)$; see, e.g., Brockwell and Davis (1991).

When there are $S \geq 1$ series and $p \geq 1$ variables, the definition of Eqs.(2.2)-(2.3) can be extended to give a $p \times p$ sample cross-autocovariance function matrix between variables (j, j') at lag k, $\hat{\boldsymbol{\Gamma}}\,(k)$, with elements given by, for $j, j' = 1, \ldots, p$,

$$\hat{\gamma}_{jj'}(k) = \frac{1}{NS}\sum_{s=1}^{S}\sum_{t=1}^{N-k}(X_{stj} - \bar{X}_j)(X_{s,t+k,j'} - \bar{X}_{j'}),\ k = 0, 1, \tag{2.4}$$

$$\text{with } \bar{X}_j = \frac{1}{NS}\sum_{s=1}^{S}\sum_{t=1}^{N}X_{stj};$$

and the $p \times p$ sample cross-autocorrelation matrix at lag k, $\hat{\boldsymbol{\rho}}^{(1)}(k)$, has elements $\hat{\rho}_{jj'}(k)$, $j, j' = 1, \ldots, p$, obtained by substituting Eq.(2.4) into Eq.(2.3). This cross-autocovariance function in Eq.(2.4) is a measure of time dependence between observations k units apart for a given variable pair (j, j'), calculated across all S series. Notice, the sample means $\bar{X}_j$ in Eq.(2.4) are calculated across all NS observations.

An alternative approach is to calculate these sample means by series. In this case, the cross-autocovariance matrix $\hat{\boldsymbol{\Gamma}}\,(k)$ has elements estimated by, for $j, j' = 1, \ldots, p$, $s = 1, \ldots, S$,

$$\hat{\gamma}_{jj'}(k) = \frac{1}{NS}\sum_{s=1}^{S}\sum_{t=1}^{N-k}(X_{stj} - \bar{X}_{sj})(X_{s,t+k,j} - \bar{X}_{sj'}),\ k = 0, 1, \tag{2.5}$$

$$\text{with } \bar{X}_{sj} = \frac{1}{N}\sum_{t=1}^{N}X_{stj};$$

and the corresponding $p \times p$ cross-autocorrelation function matrix $\hat{\boldsymbol{\rho}}^{(2)}(k)$ has elements $\hat{\rho}_{jj'}(k)$ found by substituting Eq.(2.5) into Eq.(2.3).

Other model structures can be considered, which would provide other options for obtaining the relevant sample means. These include class structures, lag k structures, weighted series and/or weighted variable structures, and the like; see Billard et al. (2015).

2.2 Principal Components for Time Series

In a standard classical principal component analysis on a set of p-dimensional multivariate observations $\mathbf{X} = \{X_{ij}, i = 1, \dots n,\ j = 1, \dots, p\}$, each observation is projected into a corresponding ν^{th} order principal component, $PC_\nu(i)$, through the linear combination of the observation's variables,

$$PC_\nu(i) = w_{\nu 1}X_{i1} + \cdots + w_{\nu p}X_{ip}, \quad \nu = 1, \dots, p, \tag{2.6}$$

where $\mathbf{w}_\nu = (w_{\nu 1}, \dots, w_{\nu p})$ is the ν^{th} eigenvector of the correlation matrix $\boldsymbol{\rho}$ (or, equivalently for non-standardized observations, the variance-covariance matrix $\boldsymbol{\Sigma}$). The eigenvalues satisfy $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$, and $\sum_{\nu=1}^{p} \lambda_\nu = p$ (or, σ^2 for non-standardized data). A detailed description of this methodology for standard data can be found in any of the numerous texts on multivariate analysis, e.g., Joliffe (1986) and Johnson and Wichern (2007) for an applied approach, and Anderson (1984) for theoretical details.

For time series data, the correlation matrix $\boldsymbol{\rho}$ is replaced by the cross-autocorrelation matrix $\boldsymbol{\rho}(k)$, for a specific lag $k = 1, 2, \dots$, and the ν^{th} order principal component of Eq.(2.6) becomes

$$PC_\nu(s,t) = w_{\nu 1}X_{s1t} + \cdots + w_{\nu p}X_{spt}, \quad \nu = 1, \dots, p, \quad t = 1, \dots, N,\ s = 1, \dots, S. \tag{2.7}$$

The elements of $\boldsymbol{\rho}(k)$ can be estimated by $\hat{\rho}_{jj'}(k)$ from Eq.(2.4) or from Eq.(2.5) (or from other choices of model structure). The problem of non-positive definiteness, for lag $k > 0$, for the cross-autocorrelation matrix has been studied by Rousseeuw and Molenberghs (1993) and Jäckel (2002), with the recommendation that negative eigenvalues be re-set at zero.

3 Illustration

To illustrate, take a data set (<http://dss.ucar.edu/datasets/ds578.5>) where the observations are time series of monthly temperatures at $S = 14$ cities (weather stations) in China over the years 1923-88. In the present analysis, each month is taken to be a single variable corresponding to the twelve months (January, ..., December, respectively); hence, $p = 12$. Clearly, these variables are dependent as reflected in the cross-autocovariances when $j \neq j'$.

Let us limit the discussion to using the cross-autocorrelation functions at lag $k = 1$, evaluated from Eq.(2.4) and Eq.(2.3), and shown in Table 1. We obtain the corresponding eigenvalues and eigenvectors, and hence we calculate the principal components PC_ν, $\nu = 1, \dots, p$, from Eq.(2.6). A plot of $PC_1 \times PC_2 \times$ time is

displayed in Figure 1, and that for $PC_1 \times PC_3\times$ time is given in Figure 2. An interesting feature of these data highlighted by the methodology is that it is the $PC_1 \times PC_3$ pair that distinguishes more readily the city groupings. Figure 3 displays the $PC_1 \times PC_3$ values for all series and all times without tracking time (i.e., the 3-dimensional $PC_1 \times PC_3 \times$ time values are projected onto the $PC_1 \times PC_3$ plane). Hence, we are able to discriminate between cities.
Thus, we observe that cities 1-4 (Hailaer, HaErBin, MuDanJiang and ChangChun, respectively), color coded in black (and indicated by the symbol black ∘ and full lines ('lty=1')) have similar temperatures and are located in the north-eastern region of China. Cities 5-7 (TaiYuan, BeiJing, TianJin), identified by red (△ and lines $- \cdot -$ ('lty=4')), are in the north, and have similar but different temperature trends than do those in the north-eastern region. Two (BeiJing and TianJin) are located close to sea-level, while the third (TaiYuan) is further south (and so might be expected to have higher temperatures) but its elevation is very high so decreasing its temperature patterns to be more in line with BeiJing and TianJin. Cities 8-11 (ChengDu, WuHan, ChangSha, HangZhou), green ($*$) with lines $\cdots$ ('lty=3'), are located in central regions with ChengDu further west but elevated. Finally, cities 12-14 (FuZhou, XiaMen, GuangZhou), blue (□) with lines $- - -$ ('lty=8'), are in the southeast part of the country.

Pearson correlations between the variables X_j, $j = 1, \ldots, 12$, and the principal components PC_ν, $\nu = 1, \ldots, 12$, sand correlation circles (not shown) show that all months have an impact on PC_1 with the months of June, July and August having a slightly negative influence on PC_2. Plots for other $k \neq 1$ values give comparable results. Likewise, analyses using the cross-autocorrelations of Eq.(2.5) also produce similar conclusions.

4 Conclusion

The methodology has successfully identified cities with similar temperature trends, which trends *a priori* could not have been foreshadowed, but which do conform with other geophysical information thus confirming the usefulness of the methodology. The cross-autocorrelation functions for a p-dimensional multivariate time series have been extended to the case where there are $S \geq 1$ multivariate time series. These replaced the standard variance-covariance matrices for use in a principal component analysis, thus retaining measures of the time dependencies inherent to time series data. The methodology produces principal component time series, which can be compared in the usual way on a principal component plot, except that the plot also includes time as an additional plot dimension.

References

Anderson, T.W. (1984): *An Introduction to Multivariate Statistical Analysis* (2nd ed), John Wiley, New York.

Billard, L., Douzal-Chouakria, D. and Samadi, S. Y. (2015). Toward Autocorrelation Functions: A Non-Parametric Approach to Exploratory Analysis of Multiple Multivariate Time Series. Manuscript.

Box, G. E. P., Jenkins, G. M. and Reinsel, G. C. (2011): *Time Series Analysis: Forecasting and Control (4th. ed.).* John Wiley, New York.

Brockwell, P.J. and Davis, R.A. (1991): *Time Series: Theory and Methods.* Springer-Verlag, New York.

Caiado, J., Maharaj, E. A., D'Urso, P. Time series clustering, in Handbook of Cluster Analysis, Chapman & Hall, C. Hennig, M. Meila, F. Murtagh, R. Rocci (eds.), in press.

Cryer, J.D. and Chan, K.-S. (2008): *Time Series Analysis.* Springer-Verlag, New York.

D'Urso, P., Maharaj, E. A. (2009) Autocorrelation-based Fuzzy Clustering of Time Series, Fuzzy Sets and Systems, 160, 35653589. DOI: 10.1016/j.fss.2009.04.013.

Jäckel, P. (2002): *Monte Carlo Methods in Finance.* John Wiley, New York.

Johnson, R.A. and Wichern, D.W. (2007): *Applied Multivariate Statistical Analysis* (7th ed.), Prentice Hall, New Jersey.

Joliffe, I.T. (1986): *Principal Component Analysis*, Springer-Verlag, New York.

Jones, R.H. (1964): Prediction of multivariate time series. *Journal of Applied Meteorology*, 3, 285-289.

Košmelj, K. and Batagelj, V. (1990): Cross-sectional approach for clustering time varying data. *Journal of Classification* 7, 99-109.

Liao, T.W. (2005): Clustering of time series - a survey. *Pattern Recognition* 38, 1857-1874.

Liao, T.W. (2007): A clustering procedure for exploratory mining of vector time series. *Pattern Recognition* 40, 2550-2562.

Rousseeuw, P. and Molenberghs, G. (1993): Transformation of non positive semidefnite correlation matrices. *Communications in Statistics - Theory and Methods* 22, 965-984.

Whittle, P. (1963): On the fitting of multivariate autoregressions, and the approximate canonical factorization of a spectral density matrix. *Biometrika* 50, 129-134.

Table 1 - Sample Cross-Autocorrelations - $\hat{\boldsymbol{\rho}}(k),\ k = 1$

	Sample Cross-Autocorrelations $\hat{\rho}_{jj'}(1)$											
X_j	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
X_1	0.965	0.963	0.947	0.938	0.924	0.883	0.851	0.888	0.942	0.959	0.961	0.964
X_2	0.960	0.959	0.954	0.942	0.926	0.882	0.850	0.887	0.935	0.950	0.958	0.957
X_3	0.952	0.952	0.948	0.937	0.925	0.876	0.840	0.882	0.929	0.940	0.947	0.948
X_4	0.943	0.945	0.941	0.936	0.929	0.883	0.846	0.877	0.923	0.932	0.935	0.940
X_5	0.921	0.923	0.922	0.924	0.926	0.894	0.841	0.870	0.916	0.918	0.915	0.915
X_6	0.886	0.888	0.890	0.891	0.897	0.882	0.852	0.871	0.895	0.889	0.877	0.878
X_7	0.849	0.845	0.849	0.847	0.850	0.855	0.894	0.912	0.887	0.865	0.857	0.848
X_8	0.890	0.883	0.877	0.879	0.877	0.870	0.906	0.927	0.922	0.904	0.899	0.891
X_9	0.943	0.938	0.922	0.921	0.915	0.895	0.892	0.923	0.960	0.958	0.950	0.946
X_{10}	0.960	0.953	0.938	0.931	0.921	0.891	0.869	0.906	0.956	0.964	0.963	0.958
X_{11}	0.970	0.960	0.947	0.936	0.921	0.879	0.862	0.897	0.952	0.961	0.962	0.963
X_{12}	0.969	0.960	0.948	0.933	0.920	0.878	0.849	0.889	0.946	0.959	0.962	0.961

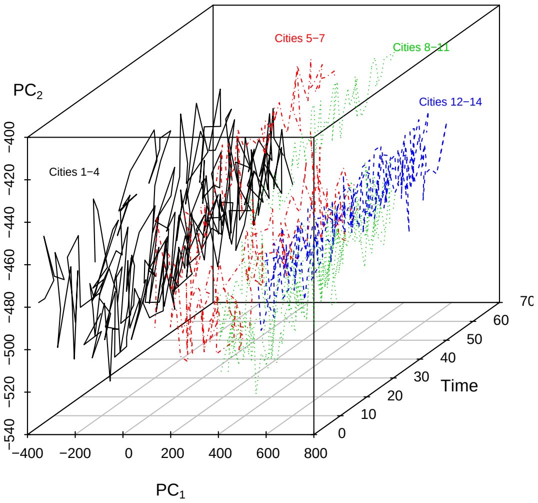

Figure 1 - Temperature Data: $PC_1 \times PC_2$ over Time – All Cities, $k = 1$

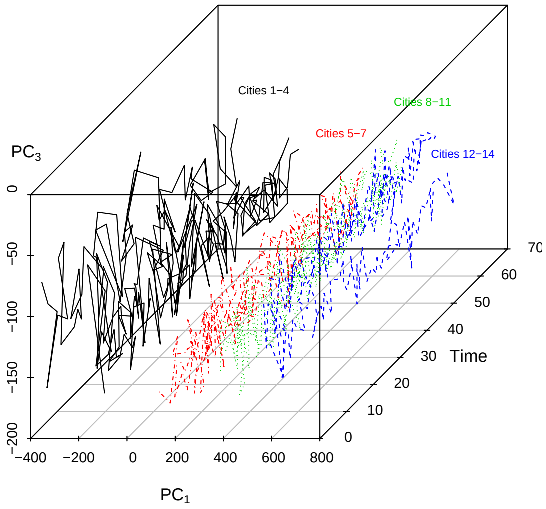

Figure 2 - Temperature Data: $PC_1 \times PC_3$ over Time − All Cities, $k = 1$

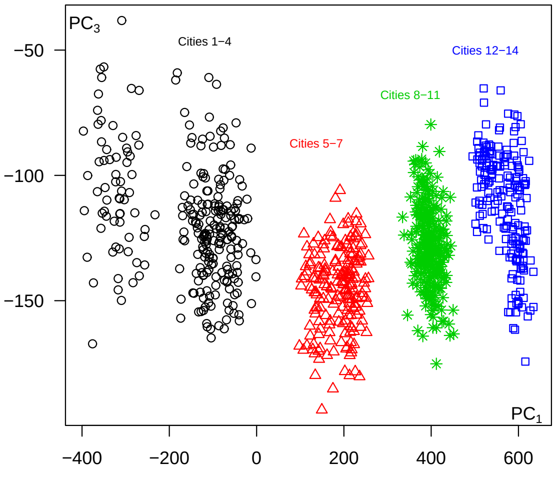

Figure 3 - Temperature Data: $PC_1 \times PC_3$ − All Cities, All Times, $k = 1$