Combining Graphical Models and PCA for Statistical Process Control

Roland Fried[1], Ursula Gather[1], Michael Imhoff[2], Melanie Keller[1] and Vivian Lanius[1]

[1] Department of Statistics, University of Dortmund, D-44221 Dortmund, Germany
[2] Surgical Department, Community Hospital Dortmund, D-44137 Dortmund, Germany

Abstract. Principal component analysis (PCA) is frequently used for detection of common structures in multivariate data, e.g. in statistical process control. Critical issues are the choice of the number of principal components and their interpretation. These tasks become even more difficult when dynamic PCA (Brillinger, 1981) is applied to incorporate dependencies within time series data. We use the information obtained from graphical models to improve pattern detection based on PCA.

Keywords. Time series analysis, dimension reduction, online monitoring, pattern detection

1 Introduction

Principal components analysis (PCA) is a frequently used tool for reducing the dimension of high-dimensional data. It is often applied e.g. in statistical process control when many variables have to be monitored simultaneously. The high sampling frequencies, that are possible nowadays for measuring process data, cause both autocorrelations and cross-correlations in the data. Brillinger (1981) suggests a dynamic version of PCA in the frequency domain that takes such correlations at various time-lags into account.

An important problem is the choice of the number of principal components (PC's). Versions of criteria like the eigenvalue criterion, the scree graph and the percentage of explained variability are difficult to use as we have to consider all frequencies. Further problems arise from the interpretation of the PC's. When using methods like Varimax or Procrustes rotation of the eigenvectors we also have to consider all frequencies. In conclusion, this method is computationally expensive when applied to high-dimensional time series.

These problems can be facilitated by partitioning the variables into subsets of closely related variables and extracting PC's for these subsets (MacGregor et al., 1994, Casin, 2001). The subsets need to be specified by background knowledge, that can be derived from prior analysis of historic data. We relate PCA to graphical models (Cox and Wermuth, 1996, Dahlhaus, 2000) and use the latter to enhance the results of pattern recognition based on PCA.

2 Dynamic principal component analysis

Principal component analysis searches those directions in a multivariate data space that capture the largest percentage of variability within the measured data. In the static case, where independent observations of a d-variate random variable $\mathbf{X}$ are observed, the solution is given by an analysis of the eigenvalues and the eigenvectors of the (estimated) covariance matrix $\boldsymbol{\Gamma}$: The j-th principal component $\chi_j = \mathbf{v}_j'\mathbf{X}$, which is defined to have maximum variance given the constraints that $||\mathbf{v}_j|| = 1$ and $\mathbf{v}_j \in \mathbb{R}^d$ is orthogonal to $\mathbf{v}_1, \ldots, \mathbf{v}_{j-1}$, is determined by an eigenvector $\mathbf{v}_j$ for the j-th largest eigenvalue λ_j of $\boldsymbol{\Gamma}$, that is the variance of χ_j. For dimension reduction one uses the first k PC's $\boldsymbol{\chi} = (\chi_1, \ldots, \chi_k)$ instead of the observed data with a suitable $k \leq d$.

PCA as defined above captures relationships only between simultaneous measurements. Brillinger (1981, chapter 9) suggests a dynamic version of PCA for a multivariate time series $\mathbf{x}(1), \ldots, \mathbf{x}(n)$ that takes into account correlations between the components at various time lags assuming that the underlying process $\{\mathbf{X}(t) : t \in \mathbb{Z}\}$ is stationary. For simplicity we further assume that the series has mean zero. While the PC's found by classical ("static") PCA are linear combinations of simultaneous measurements only, the PC's suggested by Brillinger are k-dimensional filtered series

$$\boldsymbol{\chi}(t) = \sum_h \boldsymbol{b}(h)\mathbf{x}(t-h), t \in \mathbb{Z},$$

i.e. linear combinations of past, present and future measurements, where $\{\boldsymbol{b}(h), h \in \mathbb{Z}\}$ is a linear filter of dimension $k \times d$. We get an approximation of the observed time series using another filter of dimension $d \times k$ as

$$\mathbf{x}^\star(t) = \sum_h \boldsymbol{c}(h)\boldsymbol{\chi}(t-h)\,, \quad t \in \mathbb{Z}\,.$$

The filters minimizing the mean square approximation error are given by

$$\boldsymbol{b}(h) = \frac{1}{2\pi}\int_0^{2\pi} \boldsymbol{B}(\alpha)\exp(ih\alpha)d\alpha,\;\; \boldsymbol{c}(h) = \frac{1}{2\pi}\int_0^{2\pi} \boldsymbol{C}(\alpha)\exp(ih\alpha)d\alpha$$

$$\boldsymbol{B}(\alpha) = \begin{bmatrix} \overline{\mathbf{v}_1(\alpha)'} \\ \vdots \\ \overline{\mathbf{v}_m(\alpha)'} \end{bmatrix},\;\; \boldsymbol{C}(\alpha) = \overline{\boldsymbol{B}(\alpha)}'\,,$$

where $\mathbf{v}_j(\alpha)$ is an eigenvector for the j-th largest eigenvalue $\lambda_j(\alpha)$ of the spectral density matrix $\boldsymbol{f}(\alpha), \alpha \in [0, 2\pi]$, and $\lambda_j(\alpha)$ is the spectral density of the j-th principal component series. The minimal mean square error is

$$E\left([\mathbf{X}(t) - \mathbf{X}^\star(t)]'\,[\mathbf{X}(t) - \mathbf{X}^\star(t)]\right) = \int_0^{2\pi} \sum_{j=k+1}^{d} \lambda_j(\alpha)d\alpha\,.$$

The principal component series are uncorrelated at all time lags. Thus, for dynamic PCA we need to perform a classical PCA for each frequency. For the estimation of the spectral density matrix we smooth the periodogram using a Daniell-window.

PCA is popular e.g. in statistical process control since in multivariate data often "redundant" information is contained as there are (approximate) relationships between the variables. A possibility to remove redundancy is to extract those PC's from the data, that contain the relevant information. Typically, one monitors the PC's jointly by Hotelling's T^2 statistic, that is

$\boldsymbol{\chi}(t)'\boldsymbol{\chi}(t)$ as the PC's are orthogonal, and additionally monitors the sum of the squared residuals $Q(t) = (\mathbf{x}(t) - \mathbf{x}^\star(t))'(\mathbf{x}(t) - \mathbf{x}^\star(t))$. Ideally, significant changes in the former represent a possible change within the same (still valid) relationships, while the latter signal a deviation from these relationships. For a closer inspection one needs to inspect the observed variables then as using T^2 and Q only means a large loss of information. In principle, one could also monitor the PC's individually as they are mutually orthogonal. However, interpretation of the PC's is usually not possible as they are linear combinations of all variables. Sometimes meaningful PC's can be obtained by applying a suitable transformation like Varimax or Procrustes rotation. While the former works automatically but does not always give easily interpretable results, for the latter we have to specify directions using background knowledge. If we were able to derive meaningful PC's in some way we could often specify possible sources of a disturbance directly without looking at the observed variables and avoiding possible problems like T^2 being significant without any of the observed variables showing an unusual pattern.

In a comparative study, Keller (2000) analyses physiological variables measured in intensive care using static PCA, dynamic PCA and multivariate singular spectrum analysis (MSSA, Read, 1993). She finds clinically relevant patterns like outliers and level shifts to be captured best by dynamic PCA, that also needs fewer PC's than the other methods. The essential information contained in 11 physiological variables could usually be compressed into five dynamic PC's. However, problems are noted with respect to a criterion based choice of the number of PC's as dynamic versions of the scree graph or the percentage of explained variability point at fewer PC's in most of the cases. Using too few PC's is not desirable since then the systematic (as opposed to singular or random) variability is partitioned incorrectly into the PC's and the residuals. In consequence we might miss relevant patterns.

This risk is further increased by the fact that PCA has a couple of optimality properties but none of them considers the variables individually (e.g. McCabe, 1984). Particularly, PCA does not guarantee that every variable has a large multiple correlation with the principal components, or, equivalently, that the remaining variability is small for all variables. However, this is interesting in statistical process control. Beale, Kendall and Mann (1967) suggest to base variable selection on this criterion in multivariate analysis. It should guide dimension reduction in the context of online-monitoring, at least if all measured variables are important.

Hence, statistical process control based on any form of PCA may result in unsatisfactory results as the PC's may be difficult to interpret and do not necessarily describe the course of every important variable well. Methods like Procrustes or Varimax rotation do not really solve these problems and are difficult to apply for dynamic PCA since we need to perform the rotation at all frequencies.

3 Graphical models

A *graph* $G = (V, E)$ consists of a finite set of *vertices* V and a set of *edges* $E \subseteq V \times V$, that are ordered pairs of vertices. It can be visualized by drawing a circle for each vertex and connecting each pair a, b of vertices whenever $(a, b) \in E$ or $(b, a) \in E$ by an edge. We focus exclusively on *undirected graphs* where $(a, b) \in E$ implies $(b, a) \in E$. In these graphs all edges represent symmetric associations and are symbolized by lines.

Dahlhaus (2000) introduces *partial correlation graphs* for multivariate time series, which generalize *conditional independence graphs* for independent observations. In these graphs, the vertices represent the measured variables, i.e. the components of the time series. Partial correlation graphs are defined by the *pairwise Markov property*, that states that two components a and b are partially uncorrelated at all time lags given the linear effects of all remaining components if they are not connected by an edge, i.e. $(a, b), (b, a) \notin E$. For independent measurements, the partial correlations given all remaining variables can easily be calculated from the *concentration matrix* $(g_{ab})_{1 \leq a,b \leq d}$, which is the inverse of the covariance matrix if the latter is regular, as they equal the negative scaled entries $g_{ab}(g_{aa}g_{bb})^{-1/2}$. For dynamic data, the partial spectral coherences, which are transforms of the partial correlations between the component processes at all time lags into the frequency domain, can analogously be derived from the inverse of the spectral density matrix.

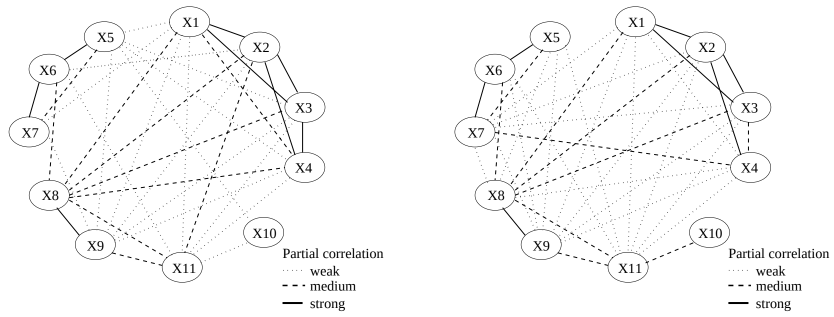
Partial correlation graphs reveal the essential linear relations between multiple variables. There is some (linear) relation between two variables a and b (possibly mediated by other variables) if they are *connected* by a *path*, i.e. if vertices $a = a_0, \ldots, a_l = b$, $l \geq 1$, exist such that there is an edge between each pair of successive vertices. A *connectivity component* of an undirected graph is a maximal subset of pairwise connected variables.

Although graphical models are based on pairwise relations given all the remaining variables, the *global Markov property* allows statements for situations, where we do condition on a separating subset only. Hence, this property is interesting for dimension reduction. It is equivalent to the pairwise Markov property if the spectral density matrix is regular at all frequencies. More precisely, the global Markov property implies that all partial correlations between variables in a set A and variables in a set B are zero given the linear effects of the variables in a set S separating A and B in G. Moreover, if S separates A and B in G then the course of the variables in B does not provide information (in terms of linear regression) on the variables in A given the variables in S and vice versa. This means, for reducing the conditional variance of variables in A (B) we should consider linear combinations of variables in $A \cup S$ ($B \cup S$). Therefore, separations in the partial correlation graph provide information on suitable directions for rotations and on subsets of variables which may be analysed individually.

For exemplification we study a 11-variate time series simulated from a VAR(1)-model. The model parameters have been fitted to a physiologic time series of heart rate, blood pressures etc. as analysed in Keller (2000). Figure 1 shows the partial correlation graph derived by estimating the partial spectral coherences from 600 observations using the program Spectrum (Dahlhaus and Eichler, 2000). Following a suggestion by Brillinger (1996) we measure the strength of the partial correlations via the area under the partial spectral coherences and use gradually distinct edges for illustration (Gather, Imhoff and Fried, 2002). There are three groups of strongly related variables, while the other partial correlations are rather small. Furthermore, variables 10 and 11 do not have any strong partial correlations. Hence, there may be between 3 and 6 strong relationships and accordingly we need between 5 and 8 PC's to capture the systematic variability. As this is a rule of thumb only we should additionally consider the percentage of explained variability. Experimenting with other processes as well we found about 500 observations to be necessary to detect the important associations in 11-variate time series. The "true" partial correlation graph for the underlying stochastic process estimated from 5000 observations is also shown in Figure 1.

In order to get meaningful components we use the partition $\{X_1, X_2, X_3, X_4\}$,

Fig. 1. Partial correlation graph estimated from a simulated time series of length 600 (left) and true partial correlation graph (right).

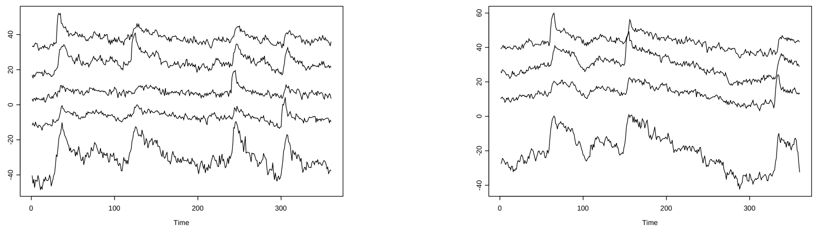


$\{X_5, X_6, X_7\}$, $\{X_8, X_9\}$, $\{X_{10}\}$, $\{X_{11}\}$ and then extract PC's for the former subsets individually treating X_{10} and X_{11} separately. This agrees with existing medical knowledge as these subsets represent intrathoracic pressures, arterial pressures, heart rate and pulse, temperature and pulsoxymetry respectively in the physiological time series. Figure 2 depicts the component processes corresponding to the subsets $\{X_1, X_2, X_3, X_4\}$ and $\{X_5, X_6, X_7\}$ for another time series simulated from the same model with several patches of five subsequent innovative outliers inserted. Innovative outliers are outlying "common" shocks in the VAR(1)-process which influence all variables according to the relationships within the process. Here, using one dynamic PC for each subset we find that these PC's represent the important patterns, but there are smoothing effects as we apply temporal filtering. These PC's are weakly correlated as this is true for the corresponding subsets. Alternatively, one could first calculate a PC for $\{X_8, X_9\}$ and then dynamically regress the other variables on it, before we extract further PC's from the residuals. In this way, we could further reduce the remaining variability and reduce the correlations among the components exploiting the moderate correlations between $\{X_8, X_9\}$ and the other variables. However, this is computationally more expensive and the results would be more difficult to interpret for an operator.

4 Conclusion

Partial correlation graphs with different edges representing different strengths of associations reveal the essential relationships within multivariate time series. Graph separations can be used to identify subsets of closely related variables as well as "isolated" variables. This provides information on the number of PC's needed to extract the systematic variability from the data. Moreover, we may find suitable partitions of the variables and apply methods like blockwise PCA to derive interpretable PC's. In our experience the percentage of explained variability is typically not much smaller than the percentage obtained using the same number of "ordinary" PC's. On the other hand, partitioning the variables reduces computational costs, affords fewer observations and allows more meaningful results. Dynamic PCA allows to incorporate temporal dependencies and captures the systematic variability within multivariate time series better than static PCA. However, for dy-

Fig. 2. Simulated time series (360 observations) partitioned into subsets of strongly correlated variables with inserted patterns. Left: Variables 1 to 4 (top, solid) and corresponding PC (bottom, bold solid). Right: Same for variables 5 to 7.



namic PCA more observations are needed to get reliable results than for static PCA. Therefore, dynamic PCA in its current form is mainly useful for retrospective pattern detection in multivariate time series, that is a difficult problem (Tsay, Peña and Pankratz, 2000). Often low order VAR-models are sufficient to capture the essential temporal dependencies between the variables. Dimension reduction techniques based on these model structures could be very useful for real-time process control.

References

Beale, E.M.L., Kendall, M.G. and Mann, D.W. (1967). The discarding of variables in multivariate analysis. *Biometrika*, **54**, 357-366.

Brillinger, D.R. (1981). *Time Series. Data Analysis and Theory.* San Francisco: Holden Day.

Brillinger, D.R. (1996). Remarks concerning graphical models for time series and point processes. *Revista de Econometria*, **16**, 1-23.

Casin, Ph. (2001). A generalization of principal component analysis to K sets of variables. *Computational Statistics & Data Analysis*, **35**, 417-428.

Cox, D.R. and Wermuth, N. (1996). *Multivariate Dependencies.* London: Chapman & Hall.

Dahlhaus, R. (2000). Graphical Interaction Models for Multivariate Time Series. *Metrika*, **51**, 157-172.

Dahlhaus, R. and Eichler, M. (2000). *Spectrum.* Program is available at http://www.statlab.uni–heidelberg.de/projects/graphical.models/

Gather, U., Imhoff, M. and Fried, R. (2002). Graphical models for multivariate time series from intensive care monitoring. *Statistics in Medicine*, to appear.

Keller, M. (2000). *Hauptkomponentenanalyse für intensivmedizinische Zeitreihen* (in German). Unpublished Diploma Thesis, Department of Statistics, University of Dortmund, Germany.

MacGregor, J.F., Jaeckle, C., Kiparissides, C. and Koutoudi, M. (1994). Process Monitoring and Diagnosis by Multiblock PLS Methods. *AIChE J.*, **40**, 826-838.

McCabe, G.P. (1984). Principal variables. *Technometrics*, **26**, 137-144.

Read, P.L. (1993). Phase portrait reconstruction using multivariate singular systems analysis. *Physica D*, **69**, 353-365.

Tsay, R.S., Peña, D., Pankratz, A.E. (2000). Outliers in multivariate time series. *Biometrika*, **87**, 789-804.