Partial Correlation Graphs and Dynamic Latent Variables for Physiological Time Series

Roland Fried[1], Vanessa Didelez[2], and Vivian Lanius[1]

[1] Fachbereich Statistik,
Universität Dortmund, D-44221 Dortmund, Germany
[2] Department of Statistical Science,
University College London, London, WC1E 6BT, U.K.

Abstract. Latent variable techniques are helpful to reduce high-dimensional time series to a few relevant variables that are easier to model and analyze. An inherent problem is the identifiability of the model and the interpretation of the latent variables. We apply graphical models to find the essential relations in the data and to deduce suitable assumptions leading to meaningful latent variables.

1 Introduction

In high-dimensional time series we may find strong correlations among the observed variables at several time lags. Statistical modelling should appropriately reflect these dependencies. However, complex models involve numerous parameters and require many observations to enable reliable inference. Thus, suitable strategies for dimension reduction provide a useful preliminary step.

A standard approach is to select a subset of the variables and to ignore the others. It is then important to know which and how much information we neglect. Alternatively, techniques like factor and principal component analysis (PCA) allow to extract latent variables describing the correlations among the observed variables and capturing more of their variability than a simple variable selection. However, the extracted variables are typically not easy to interpret although it is often important that they are meaningful.

In order to overcome these difficulties we propose to use partial correlation graphs to learn about the essential relations among the variables. These relations are visualized by a graph, where the variables are represented as vertices and the dependencies among them are shown as edges. Separations in the graph provide information about direct and indirect relations. This can be used to deduce suitable assumptions when applying factor analytic methods.

In this paper, we compare dimension reduction by variable selection, by straight-forward latent variable analysis and by latent variable analysis with restrictions derived from partial correlation graphs. We illustrate these approaches by analyzing physiological time series describing the human hemodynamic system and show that we can extract latent variables that explain

more of the observed variability than a simple variable selection but are still meaningful.

2 Partial correlation graphs and factor analysis

Graphical models visualize and clarify the dependencies among a set of variables (Whittaker (1990), Lauritzen (1996)). 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 restrict attention to undirected graphs where $(a, b) \in E$ implies $(b, a) \in E$ shown as undirected edge (a simple line) between a and b.

A *path* is a finite sequence of vertices $a_0, \ldots, a_n$, such that there is an edge connecting each pair of subsequent vertices. Subsets $A, B \subset V$ are *separated* by a subset $S \subset V$ if every path from a vertex in a to a vertex in B necessarily includes a vertex in S. A subset $C \subset V$ is called *complete* if all possible edges between pairs of variables in C exist.

Brillinger (1996) and Dahlhaus (2000) introduce partial correlation graphs for multivariate time series. These models focus on the essential linear, possibly time-lagged relations between pairs of component series which persist after eliminating all linear effects of the other variables. Here and in the following we assume that $Y_V = \{Y_V(t), t \in \mathbb{Z}\}$, $V = \{1, \ldots, d\}$, is a vector-valued weakly stationary time series with absolutely summable covariance function

$$\gamma_{ab}(h) = Cov(Y_a(t+h), Y_b(t)),\ h \in \mathbb{Z}.$$

For $A \subset V$ we denote the subprocess of all variables $a \in A$ by Y_A, and for $a \in V$ we denote the corresponding component process by Y_a.

The *cross-spectrum* between the time series Y_a and Y_b is the Fourier-transform of their covariance function,

$$f_{ab}(\lambda) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \gamma_{ab}(h) \exp(-i\lambda h),\ \lambda \in [-\pi, \pi]\,.$$

This defines a decomposition of γ_{ab} into periodic functions of frequencies λ. The variables Y_a and Y_b are uncorrelated at all time lags h iff $f_{ab}(\lambda)$ equals zero for all frequencies.

In order to distinguish between direct and induced linear relationships between two series Y_a and Y_b, the linear effects of the remaining variables on Y_a and Y_b have to be eliminated. The *partial cross-spectrum* between Y_a and Y_b is defined as the cross-spectrum between the series ϵ_a and ϵ_b,

$$f_{ab \cdot V \setminus \{a,b\}}(\lambda) = f_{\epsilon_a \epsilon_b}(\lambda)\,,$$

where $\epsilon_a(t)$ and $\epsilon_b(t)$ are the residual series obtained by subtracting all linear influences of $Y_{V \setminus \{a,b\}}$ from $Y_a(t)$ and $Y_b(t)$ respectively (Brillinger (1981)).

Similarly, the (partial) cross-spectrum between two vector time series can be defined. The *partial spectral coherency* is a standardization of the partial cross-spectrum

$$R_{ab\cdot V\setminus\{a,b\}}(\lambda) = \frac{f_{ab\cdot V\setminus\{a,b\}}(\lambda)}{\left[f_{aa\cdot V\setminus\{a,b\}}(\lambda) f_{bb\cdot V\setminus\{a,b\}}(\lambda)\right]^{1/2}} . \tag{1}$$

A *partial correlation graph* for a multivariate time series is an undirected graph $G = (V, E)$ with a vertex for each of the components $a \in V$ of the time series, where two vertices a and b are connected by an edge whenever their partial spectral coherency $R_{ab\cdot V\setminus\{a,b\}}(\lambda)$ is not identical to zero for all frequencies λ. A missing edge between a and b indicates that the linear relation between these two variables given the remaining ones is zero, which is denoted by $a \perp b | V \setminus \{a, b\}$. This is known as the *pairwise Markov property*. Under the assumption that the spectral density matrix is regular for all frequencies, Dahlhaus (2000) proves that the pairwise Markov property implies the *global Markov property*, which is a stronger property in general. It states that $A \perp B|S$ for all subsets $A, B, S \subset V$ such that S separates A and B in G.

Dynamic factor analysis allows to model a multivariate time series using a lower dimensional process of latent, i.e. unobserved variables called factors. A general dynamic factor model for an observed multivariate time series Y_V is given by a dynamic regression of Y_V

$$Y_V(t) = \sum_{u\in\mathbb{Z}} \Lambda(u) X(t-u) + e(t) \tag{2}$$

on an unobserved lower dimensional factor process X with an error process e. Here, $\Lambda(u)$ are matrices of unknown parameters called loadings. In this very general form the model is not identifiable, but we nevertheless use it as a starting point, in order to understand the assumptions that can be deduced from partial correlation graphs obtained from empirical data analysis. Brillinger's (1981) dynamic PCA in the frequency domain can be used for fitting model (2) with uncorrelated factors (Forni et al. (2000)). However, the factors extracted in this way are mixtures of all variables because all loadings are distinct from zero. Automatic rotations for improving interpretation, as in the non-dynamic case, are difficult to apply since we need to perform the rotation at each frequency individually. Problems inherent to dynamic PCA are discussed in more detail by Lanius and Gather (2003).

Under the assumption that the spectral density matrix of Y_V is regular at all frequencies, an algorithm has been derived by Fried and Didelez (2003b) to construct the partial correlation graph of Y_V given model (2) with Y_V and e both following a vector autoregressive model (Reinsel (1997)). In case of uncorrelated factors and uncorrelated error processes a pair of observed variables is connected by an edge if and only if both variables have nonzero

loadings for one of the factors. Thus, the resulting graph provides an assistance in identifying the number and types of factors. A complete subset in a partial correlation graph of Y_V can be regarded as generated by a latent factor. However, the identification of such common factors can be obscured by dependencies within the error process or the factors as such dependencies may cause additional edges in the partial correlation graph. Therefore, it seems reasonable to attribute only strong relationships to the factors while the weaker ones are ascribed to errors.

3 Analysis of physiological time series

In the following we analyze multivariate time series from 25 consecutive critically ill patients (9 female, 16 male, mean age 66 years) with extended hemodynamic monitoring requiring pulmonary artery catheterization, acquired on the surgical intensive care unit of the Klinikum Dortmund, a tertiary referral center. The hemodynamic variables heart rate HR, pulse PULS, arterial systolic pressure APS, arterial mean pressure APM, arterial diastolic pressure APD, pulmonary artery systolic pressure PAPS, pulmonary artery mean pressure PAPM, pulmonary artery diastolic pressure PAPD, central venous pressure CVP and blood temperature Temp were stored for each patient in one minute intervals with a standard clinical information system. Hence, 25 ten-variate time series with an average length of about 5200 time points were available for the following analysis.

When using methods for dimension reduction we want to explain as much of the clinically relevant variability in the data as possible, by a reduced set of variables, but not irrelevant artifacts and short-term fluctuations. Therefore we removed outliers for each variable individually using a robust filtering procedure based on the repeated median, which allows to preserve trends as well as systematic shifts in the data (Davies et al. (2003), Fried (2003)).

In order to get a general impression about the relationships between the physiological variables we constructed a partial correlation graph for each patient. We used the program "Spectrum" (Dahlhaus and Eichler (2000)) which estimates the cross-spectra by a nonparametric kernel estimator and allows simultaneous testing of all partial spectral coherencies being zero or not by constructing a sample-size dependent confidence bound. For improving the results we applied a stepwise search strategy based on graph separations described by Fried and Didelez (2003a). This strategy allows to overcome masking of weaker relations by stronger associations, which may occur when estimating all partial linear relations jointly.

We found the essential linear relations revealed by the final partial correlation graphs to match the physiological relations expected by physicians. A typical example of such a graph is shown in Figure 1. Different edge types are used to indicate different strengths of relations as measured by the area below the partial spectral coherencies. For all patients we identified strong

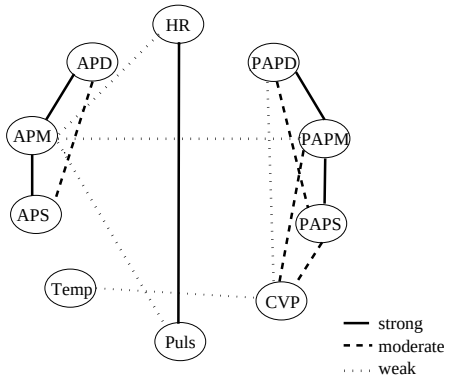

Fig. 1. Typical partial correlation graph for the hemodynamic variables of a patient.

partial correlations among the arterial pressures (APS, APM, APD), among the pulmonary artery pressures (PAPS, PAPM, PAPD) and between heart rate and pulse. The strength of the relation between the systolic and the diastolic pressure was always smaller than between each of these and the corresponding mean pressure. CVP was most strongly related to the pulmonary artery pressures, while the temperature did not show strong relations to any of the other variables. Hence, we can identify the following groups of strongly related variables from the partial correlation graphs: (APS, APM, APD), (PAPS, PAPM, PAPD,CVP), (HR, Puls).

The partitioning of the variables into strongly related subgroups is now used for variable selection. Due to the global Markov property the absence of edges between two groups of variables means that the variables in one of these groups do not contain information on the variables in the other group given the measurements of the separating variables. A variable can be regarded as very informative if it has strong relations to several other variables. Selecting e.g. APM from the strongly related subgroup of arterial pressures and neglecting APD and APSYS for clinical monitoring is therefore meaningful from both, a clinical and a statistical point of view. Applying these principles leads us to select PAPM, APM, HR and Temp.

An alternative approach for dimension reduction is to extract latent variables from the observed time series capturing as much of the total variability as possible. We scale the time series to unit variance and perform a dynamic PCA based on correlations as described by Brillinger (1981). We use four components, which is the minimal number of latent variables suggested by the partial correlation graphs.

For obtaining meaningful latent variables we can extract one component from each group of closely related variables applying dynamic PCA separately to each group. This corresponds to extracting factors as in model (2) with the $\Lambda(u)$ being restricted to be block-matrices. For heart rate and pulse, instead

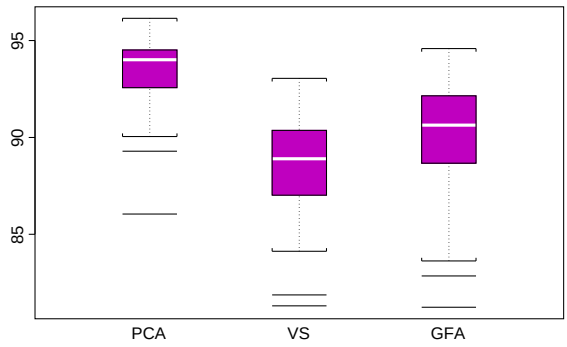

Fig. 2. Boxplots of the total explained variance (in percent): Dynamic PCA (left), variable selection (middle), grouped factor analysis (right).

of extracting a latent variable, we select the heart rate as its measurement is more reliable.

In the following we compare the percentage of variability explained by variable selection, dynamic PCA and grouped factor analysis. This is done via dynamic regression (Brillinger (1981)) of the observed variables on the selected variables and on the extracted components, respectively. Then we investigate the total residual variance as well as the individual residual variance for each variable.

Figure 2 shows that choosing the variables PAPM, APM, HR and Temp indeed explains a large part of the total variability, but less than a dynamic PCA with the same number of components, of course. Performing a grouped factor analysis allows to regain some of this loss while still providing meaningful latent variables. The variable selection explains more than 89% of the total variability for half of the patients whereas the extracted factors do so for about 75% of the patients.

Table 1 shows 5-point summaries of the explained variability for each variable. The factors derived from the groups describe the variables included in the selection very well. The explained variability increases substantially for the variables not captured well by the variable selection, see CVP and APS. When performing a standard dynamic PCA, the percentage of explained variability is at least 87% for 75% of the patients and each of the variables, which is rather high. However, these components are not meaningful to the physician. Thus, extracting latent variables from groups of closely related variables means a compromise between variable selection and factor analysis as we capture more of the total variability and of the variables neglected in the selection still working with interpretable variables.

	Min	25%	50%	75%	Max	Min	25%	50%	75%	Max
	PAPS					PAPM				
PCA	78.3	88.9	92.4	93.8	96.8	92.7	96.3	97.5	98.0	98.5
VS	57.7	72.2	83.1	88.8	94.7	100.0	100.0	100.0	100.0	100.0
GFA	61.5	75.5	84.1	87.8	93.4	85.9	93.9	95.4	96.6	97.9
	PAPD					CVP				
PCA	86.4	92.0	93.7	95.4	98.3	66.0	87.2	89.5	92.8	97.2
VS	70.5	79.5	84.8	89.7	95.2	15.3	46.9	68.0	76.4	92.5
GFA	75.6	83.3	87.5	90.7	96.4	23.6	62.1	80.4	85.2	95.4
	HR					PULS				
PCA	88.8	94.6	95.3	97.1	98.7	88.2	95.1	96.6	97.7	98.8
VS	100.0	100.0	100.0	100.0	100.0	67.3	94.5	97.3	98.7	99.8
GFA	100.0	100.0	100.0	100.0	100.0	68.7	94.4	97.3	98.7	99.8
	APS					APM				
PCA	77.9	89.4	92.7	94.8	96.8	91.6	96.3	97.3	97.7	99.2
VS	55.8	72.8	82.2	86.3	94.6	100.0	100.0	100.0	100.0	100.0
GFA	69.6	78.1	86.4	90.6	95.7	85.1	96.4	97.2	98.1	99.1
	APD					Temp				
PCA	74.1	90.8	93.8	94.7	96.7	38.9	88.4	92.3	95.8	98.6
VS	69.2	84.7	85.9	89.6	94.3	100.0	100.0	100.0	100.0	100.0
GFA	74.7	85.8	89.7	92.8	96.0	100.0	100.0	100.0	100.0	100.0

Table 1. Percentage of variability explained by PCA, by a variable selection (VS) and by a grouped factor analysis (GFA) for each variable.

4 Conclusion

Methods for dimension reduction aim at condensing the information provided by a high-dimensional time series into a few essential variables. Partial correlation graphs are a suitable tool to explore the relations among the observable variables. This information allows an advanced application of dimension reduction techniques. One possibility is to select suitable subsets of important variables from the graphs. Alternatively, we can enhance latent variable techniques. Deducing restrictions on the loading matrices from a graphical model combines variable selection and PCA as the percentage of explained variability is substantially higher than for a variable selection and we obtain meaningful latent variables. In our study the groups of closely related variables obtained from the data analysis agree with the groups anticipated from medical expertise. Therefore, we expect to gain reliable insights also in the relations among other variables, for which we have less background knowledge.

ACKNOWLEDGEMENTS

The financial support of the Deutsche Forschungsgemeinschaft (SFB 475, "Reduction of complexity in multivariate data structures") and the helpful comments by two referees are gratefully acknowledged.

References

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

BRILLINGER, D.R. (1996): Remarks Concerning Graphical Models For Time Series And Point Processes. *Revista de Econometria, 16, 1–23.*

DAHLHAUS, R. (2000): Graphical Interaction Models for Multivariate Time Series. *Metrika, 51, 157–172.*

DAHLHAUS, R. and EICHLER, M. (2000): SPECTRUM. A C program to calculate and test partial spectral coherences. Available via http://www.statlab.uni-heidelberg.de/projects/.

DAVIES, P.L, FRIED, R. and GATHER, U. (2003): Robust Signal Extraction for On-line Monitoring Data. *Journal of Statistical Planning and Inference, to appear.*

FORNI, M., HALLIN, M., LIPPI, M. and REICHLIN, L. (2000): The Generalized Dynamic Factor Model : Identification and Estimation. *The Review of Economics and Statistics, 82, 540–554.*

FRIED, R. (2003): Robust Filtering of Time Series with Trends. Technical Report 30/2003, SFB 475, University of Dortmund, Germany.

FRIED, R. and DIDELEZ, V. (2003a): Decomposability and Selection of Graphical Models for Multivariate Time Series. *Biometrika 90, 251–267.*

FRIED, R. and DIDELEZ, V. (2003b): Latent Variable Analysis and Partial Correlation Graphs for Multivariate Time Series. Technical Report 6/2003, SFB 475, University of Dortmund, Germany.

LAURITZEN, S.L. (1996): Graphical Models. Clarendon Press, Oxford.

LANIUS, V. and GATHER, U. (2003): Dimension Reduction for Time Series from Intensive Care. Technical Report 2/2003, SFB 475, University of Dortmund, Germany.

REINSEL, G.C. (1997): *Elements of Multivariate Time Series Analysis.* Second edition. Springer, New York.

WHITTAKER, J. (1990): *Graphical Models in Applied Multivariate Statistics.* Wiley, Chichester.