Analysis of High Dimensional Data from Intensive Care Medicine

Marcus Bauer[1], Ursula Gather[1] and Michael Imhoff[2]

[1] Department of Statistics, University of Dortmund, D-44221 Dortmund, Germany

[2] Surgical Department, Community Hospital, Beurhausstrasse 40, D-44137 Dortmund, Germany

Abstract. As high dimensional data occur as a rule rather than an exception in critical care today, it is of utmost importance to improve acquisition, storage, modelling, and analysis of medical data, which appears feasable only with the help of bedside computers. The use of clinical information systems offers new perspectives of data recording and also causes a new challenge for statistical methodology. A graphical approach for analysing patterns in statistical time series from online monitoring systems in intensive care is proposed here as an example of a simple univariate method, which contains the possibility of a multivariate extension and which can be combined with procedures for dimension reduction.

Keywords. Clinical information systems, decision support, high dimensional time series, online monitoring, phase space reconstruction

1 Introduction

Increasing technical possibilities in online recording of complex data structures produce manifold challenges for statistical methods. For instance, the use of clinical information systems (CIS) in intensive care medicine makes it possible to report online, simultaneously, and automatically up to 2000 physiological variables, laboratory data, device parameters etc. Even senior physicians may not be able to develop a systematic response to any problem involving more than seven variables (Miller, 1956). To allow for a more differentiated approach to therapy and computer aided clinical decision making, it seems necessary to develop tools for a suitable bedside decision support.

The patient data are multivariate time series. Thus modelling and analysis of the underlying dynamic is a central task, which should be solved to obtain tools for a decision support. In recent years much progress has been made in multivariate time series analysis, but for really high dimensional data, such as intensive care data, traditional multivariate time series methods fail because of the so-called "curse of dimensionality" (Friedmann, 1994).

Therefore, we see the necessity of new methodological approaches, which are able to extract the important information from high dimensional data and which work automatically and with fast algorithms. One should not underestimate or disregard that the results of a statistical online analysis have to be readable in an easy manner, such that physicians and nurses are able to recognize ad hoc the extracted information on the state of the patient.

In Section 2 we describe the data acquisition and storage with a CIS. The resulting demands for statistical methodology are formulated in Section 3 and a simple procedure for a graphical analysis of time series data as a tool for an ad hoc decision support is proposed in Section 4.

2 Data acquisition and storage

As the patient record is one of the most important tools in intensive care therapy, one has to give attention to acquisition and storage of these records. More and more devices with integrated microprocessors are in use at surgical intensive care units for monitoring patients and for therapeutic interventions. The use of CIS is unavoidable for processing the enormous data floods. In a clinical evaluation carried out at a major German surgical intensive care unit, a clinical information system was run for six

years. Experience has shown that a well configured and well maintained CIS improves dramatically the quality of therapy and care (Imhoff, 1992, 1995).
The CIS is based on a network of autonomous Unix workstations, one for each bed. Bedside devices (such as monitors, ventilators, etc.) are connected locally via serial interfaces. All patient data are stored on the local hard disk at the bedside and simultaneously mirrored onto a second workstation within the network. An administrative data server is used for administration of the network and the CIS, and may serve as a communication hub with central data services like the Hospital Information System. For undisturbed data analysis the patient record is transferred into a secondary SQL (Sybase SQL server) and exported into standard statistical software (SPSS, SAS).
Thus, it is guaranteed that most of the data relevant for patient monitoring is recorded in a regular, reliable, and correct way and hence the technical requirements for using statistical methods or tools for bedside decision support are fulfilled.

3 Challenges for statistical methods

The necessity of systematic research on statistical analysis of complex data structures has been repeatedly pointed out during the past two decades (see for example Tukey, 1977; Michie, 1994). Modern statistical methods and new areas like neural networks, statistical assistant systems, projection pursuit, "data mining" etc. are concerned with modelling, analysing and visualizing complex data structures. The need for appropriate methods and the existence of some of them is partly due to the increasing speed and capacity of computers. But in situations with really high dimensional data, which we often find in the lifesciences, the usefulness of the above-mentioned procedures is limited because the computational effort exceeds any possible computational power (Huber, 1993). Another fundamental problem is that in order to fill the highdimensional sample space one needs very large sample sizes, which are seldom given in praxis.
This is especially true in the context of intensive care medicine. Here the challenge for statistical methods is to develop new types of methods for data analysis, covering the

following features:

- ability to deal with multivariate / high dimensional time series
- allowing for individual patient monitoring
- designed for online-monitoring data
- ability of pattern identification
- implementation with fast algorithms
- allowing for simple interpretation.

To derive such methods, it is necessary to combine parsimonious model building with (automatic) procedures for dimension reduction on the existing computational basis. Another aspect, which has to be considered is robustness. (In intensive care medicine this is especially important for pattern identification.)

4 A new approach to pattern recognition for data from online monitoring systems

To fix ideas, let us be concerned in the following with the special task of analysing univariate intensive care online-monitoring data. We give an example of a simple univariate method, which satisfies the principles of parsimony, robustness and an ad hoc visualization and interpretation of the results.

One basic purpose of clinical monitoring of patients is to develop tools for automatic detection of qualitative patterns like outliers, level changes and trends in physiologic data. Different mathematical approaches exist, in particular in the framework of statistical time series analysis, and have been implemented in experimental and commercial software packages today. Mainly two approaches are pursued, i.e. procedures based on ARIMA-models (batch processing), see for example Imhoff *et al.* (1997), and state

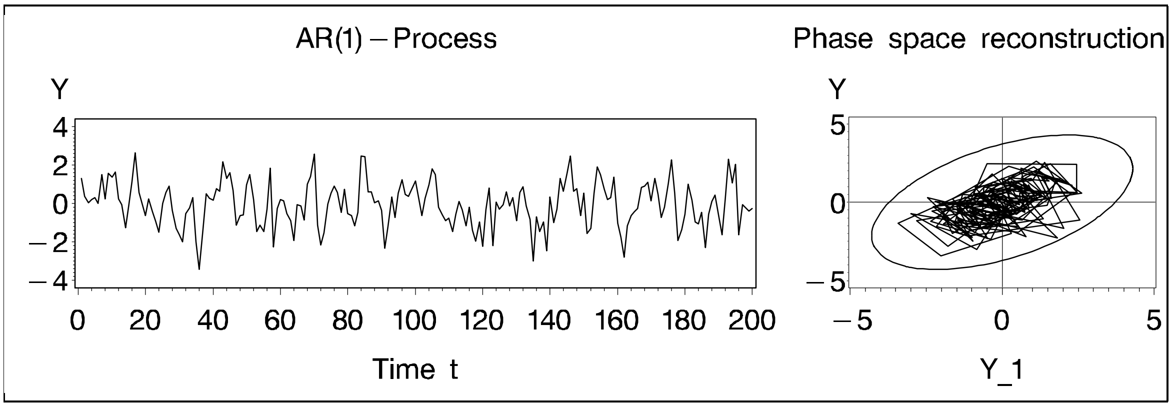

Figure 1: Realization of an AR(1)-process $Y_t = 0.5Y_{t-1} + \epsilon_t,\ \epsilon_t \sim N(0,1),\ t \in I\!N$ and its phase space reconstruction; the phase space vectors form an elliptic cloud

space models (sequential processing) following Smith & West (1983) as well as Daumer *et al.* (1996). It is not worked out yet how we can profit from these approaches to get a basis for future bedside online multivariate time series analysis. The reason for this is that existing methods have actually been constructed for monitoring univariate variables and their multivariate extensions fail because of the "curse of dimensionality". Further, the possibilities for online-monitoring are partly restricted. Here, we discuss a new graphical approach for pattern recognition in univariate time series based on phase space reconstruction. Let $\{y_t\}_{t\in\{1,\ldots,N\}}$ be a time series. Takens (1981) considered the set of m-dimensional vectors, the components of which are the time delayed observations of this time series:

$$\vec{\mathbf{y}}_t := (y_{t+(m-1)T}, \ldots, y_{t+2T}, y_{t+T}, y_t)', \quad \vec{\mathbf{y}}_t \in I\!R^m$$

with $T, m \in I\!N \backslash \{0\}$, and $t = 1, \ldots, N-(m-1)T$. The time delay is denoted by T and m is called the embedding dimension. Thus, the univariate time series is transformed into an m-dimensional space, the so-called phase space. The set $\{\vec{\mathbf{y}}_t \mid t = 1, \ldots, N-(m-1)T\}$ forms the phase space reconstruction. The phase space retains the properties of the state space, the axis of which are all variables, which characterize the dynamic. A mathematical justification of this approach is given in Takens (1981).

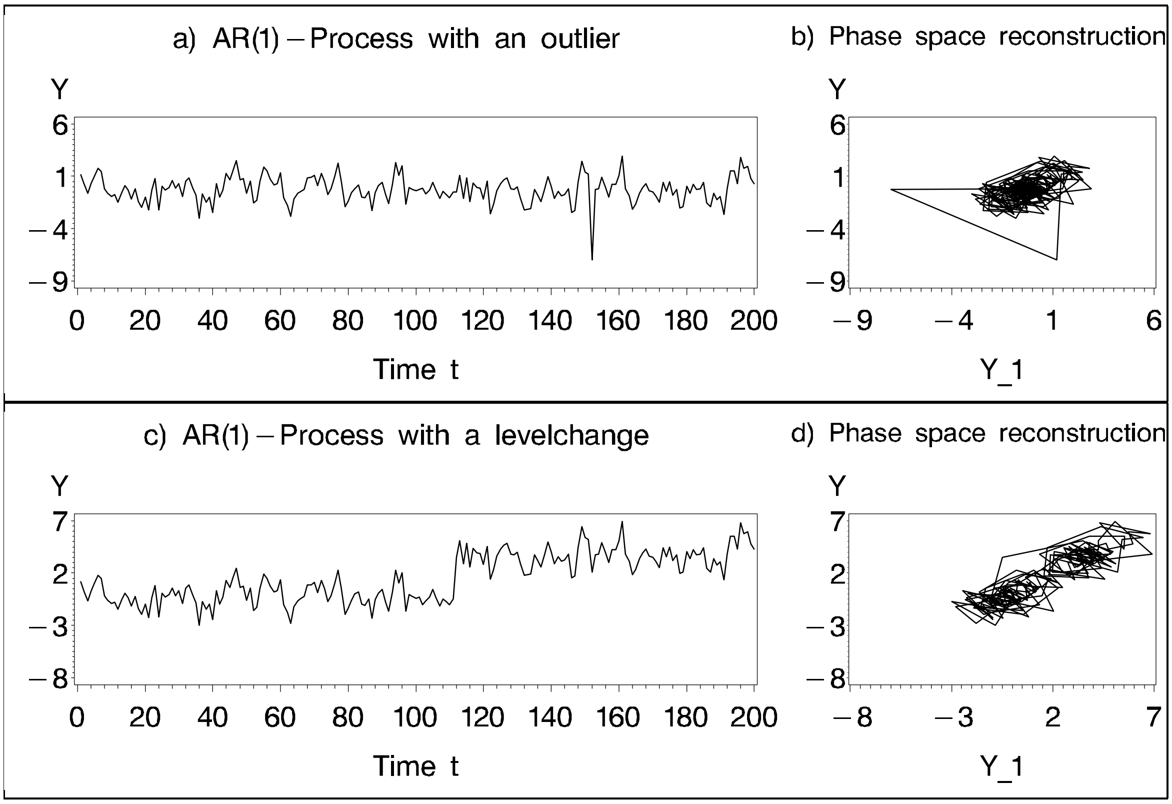

Figure 2: Simulated time series and phase space reconstruction of an AR(1)-process with outlier and with level shift

Analytical methods based on phase space reconstruction, which have been developed in theoretical physics to find properties of nonlinear dynamics, assume large sample sizes. Furthermore the data come from carefully controlled physical experiments. In the analysis of biological and ecological systems, we often have small sample sizes and random errors and moreover empirical data rather than data from controlled experiments. Thus the exact topological results of Takens (1981) are not longer valid for stochastic systems.

Nevertheless the concept of phase space reconstruction can be used for stochastic processes. In Figure 1 an observation of an AR(1)-process with Gaussian error terms and its 2-dimensional phase space reconstruction ($m = 2$) is depicted for $T = 1$. The chronological observations are combined in order to show the movement through space. The dependence structure can be clearly recognized by the elliptic form of the vector

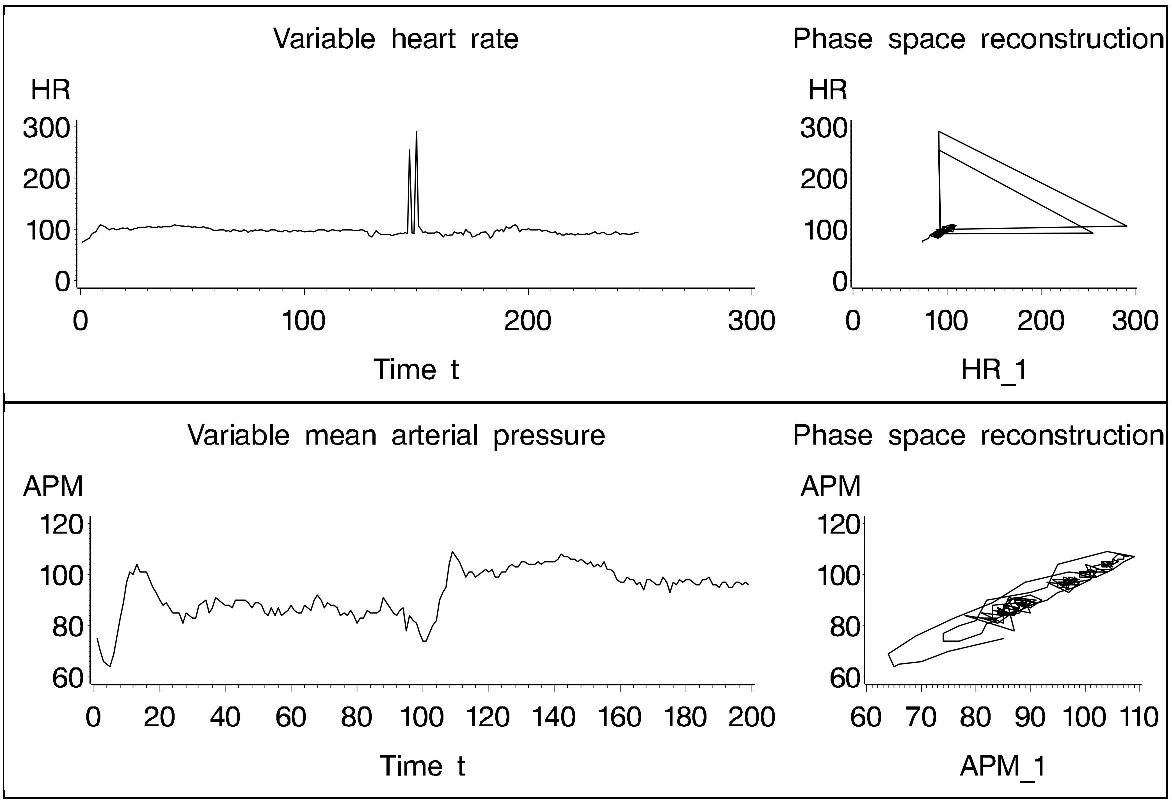

Figure 3: Typical time series and phase space reconstructions of physiological data, variables 'heart rate' and 'mean arterial pressure'

cloud. Typical disturbances of a time series like outliers, level shifts and trends can be visualized by phase space reconstructions, too. In Figure 2a an outlier is inserted in a simulated AR(1)-process and in Figure 2c a level change at time point 112 is added. The outlier at time point 152 arises in the phase space vectors $\vec{\mathbf{y}}_t$ and $\vec{\mathbf{y}}_{t+1}$, such that these vectors extrude from the regular observations (Figure 2b). Similar, all observations occuring after the level shift lie outside the original ellipse and form a new one (Figure 2d). Such features are often found in variables of intensive care data, for typical examples see Figure 3.

First attempts to model and monitor linear physiologic time series with phase space reconstructions are done by Bauer (1997), proposing a robust and automatic procedure with low computational effort for identification of special patterns in time series. Although, up to now, this pattern identification procedure has been succesfully applied to controlled clinical studies, the general use in automatic online monitoring remains for future developments. A generalization to multivariate time series and an extension in connection with procedures for dimension reduction in the situation of high dimensional data is under current research. The combination of this method with procedures for dimension reduction is necessary. A detailed analysis of the dependencies between the observed variables is a further fundamental task when modelling high dimensional data. Here the use of graphical models for multivariate time series (Dahlhaus, 1996) or the combination of methods from statistics and machine learning (Morik *et al.*, 1994) seem promising.

References

Bauer, M. (1997). *Identification of outliers and interventions in online monitoring data*. Ph.D. (in German), University of Dortmund.

Daumer, M., Falk, M. & Beyer, U. (1996). Online monitoring using multi-process Kalman filtering. Discussion paper 54 of SFB 386, TU Munich.

Dahlhaus, R. (1996). *Graphical interaction models for multivariate time series*. Technical Report, University of Heidelberg.

Friedmann, J.H. (1994). An overview of predictive learning and function approximation. In: *From Statistics to Neural Networks* (ed. V. Cherkassky, J.H. Friedmann & H. Wechsler), 1-61. Berlin: Springer.

Huber, P.J. (1993). Projection pursuit and robustness. In: *New Directions in Statistical Data Analysis and Robustness* (ed. S. Morgenthaler, E.M. Ronchetti & W.A. Stahel), 139-146. Basel: Birkhäuser.

Imhoff, M. (1992). Acquisition of ICU data: concepts and demands. *International Journal of Clinical Monitoring and Computing*, **9**, 229-237.

Imhoff, M. (1995). A clinical information system (CIS) in intensive care: how to make it work? *8th European Congress on Intensive Care Medicine*, Athens.

Imhoff, M., Bauer, M., Gather, U. & Löhlein, D. (1997). Statistical pattern detection in univariate time series of intensive care online-monitoring data. *10th European Congress on Intensive Care Medicine*, Paris, September 1997.

Michie, D., Spiegelhalter, D.J. & Taylor, C.C. (1994). *Machine Learning, Neural and Statistical Classification, Artificial Intelligence*. New York: Ellis Horwood.

Miller, G. (1956). The marginal number seven, plus or minus two: Some limits to our capacity for processing information. *Psychol. Rev.*, **63**, 81-97.

Morik, K., Potamias, G., Moustakis, V.S. & Charissis, G. (1994). Knowledgeable learning using MOBAL: a medical case study. *Applied Artificial Intelligence*, **8** (4), 579-592.

Smith, A.F.M. & West, M. (1983). Monitoring renal transplants: an application of the multi-process Kalman filter. *Biometrics*, **39**, 867-878.

Takens, F. (1981). Detecting strange attractors in turbulence. In: *Dynamic Systems and Turbulence* (ed. D.A. Rand & L.S. Young), 366-381. New York: Springer-Verlag.

Tukey, J.W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.