Robust online signal extraction from multivariate time series

Vivian Lanius, Ursula Gather [a,b,1]

[a] *Statistician, Berlin*

[b] *Department of Statistics, Technische Universität Dortmund, 44221 Dortmund*

Abstract

We introduce robust regression-based online filters for multivariate time series and discuss their performance in real time signal extraction settings. We focus on methods that can deal with time series exhibiting patterns such as trends, level changes, outliers and a high level of noise as well as periods of a rather steady state. In particular, the data may be measured on a discrete scale which often occurs in practice. Our new filter is based on a robust two-step online procedure. We investigate its relevant properties and its performance by means of simulations and a medical application.

Key words: Multivariate time series, signal extraction, robust regression, online methods

1 Introduction

In industrial and medical process control but also in economics often multivariate variables are recorded over time, where the univariate components may be dynamically dependent. Such time series are often non-stationary and they might exhibit patterns such as trends, level changes, spikes and periods of steadiness. Furthermore, the measurements may be overlaid with a high level of noise.
A typical example is a time series of online observations of vital parameters, i.e.

Email address: `vivian.lanius@uni-dortmund.de`, `gather@statistik.uni-dortmund.de` (Vivian Lanius, Ursula Gather).

[1] **Acknowledgement:** This research was initiated and conducted while the authors collaborated on the project ”Analysis of time series for online-monitoring data from intensive care medicine” of the Collaborative Research Center 475 supported by the DFG. The authors thank Michael Imhoff for recording, processing and providing the intensive care data.

15 December 2007

physiological variables, such as blood pressures and heart rate, measured by a clinical information system for critically ill patients and stored at least every minute. Reliable automatic monitoring of the hemodynamic system in real time is important in order to to support decision making at the bedside in time critical situations and thus for intensive care therapy.
The challenge is to develop methods that allow a fast and reliable filtering of such multivariate time series. Structural patterns of relevance are to be preserved, and noise and irrelevant artifacts should be removed.

For this purpose we derive regression-based filters for robust online extraction of signals from noisy and contaminated multivariate time series. In section 2 we describe the task more formally and review recent work on robust online signal extraction for univariate time series.
Section 3 deals with robust regression-based online filters for multivariate time series. Just generalizing robust univariate regression techniques to the multivariate setting leads to methods that are not affine equivariant. This yields a loss of efficiency if the error terms of the variables are highly correlated. We will therefore discuss multivariate regression methods in Subsection 3.1. Affine equivariant methods possibly do not possess a high breakdown point if the data is not in general position, i.e. when a number of datapoints larger than the dimension of the dataspace is located on lower dimensional hyperplanes. Intensive care data and financial time series, that exhibit the so-called compass rose pattern, (Crack and Ledoit, 1996) are often not in general position, especially not within short time windows. This is due to the fact that the data are often measured on a discrete scale. A compass rose pattern appears if (1) the changes in the time series from one timepoint to the next are small relative to the level of the time series, (2) these changes come in discrete jumps of a small number of measurement units and (3) the time series varies over a relatively wide range of values. Financial time series and also time series from intensive care often exhibit such a compass rose pattern (Lanius, 2005); an increasing amount of observations lies on the same hyperplane and thus the data is not in general position. There is thus a need for fast and robust online methods that can deal with such data. This will be further discussed in Subsection 3.2.
The aim of this paper is to extract multivariate signals in real time. The method which we will present in Subsection 3.3 is not affine equivariant but it has good efficiency properties and can also be applied to data that is not in general position. By means of a simulation study we will compare the relative efficiencies of this and some further regression-based multivariate online filters in Subsection 3.4.
The new method will be applied to a time series from intensive care in Subsection 3.5. We will summarize our findings and give some conclusions in Section 4.

2 Univariate Online Filters

Denote by $\boldsymbol{y}(t) = (y_1(t), \ldots, y_k(t))^\mathsf{T} \in \mathbb{R}^k$, $t \in \mathbb{Z}$, observations of a multivariate time series. As working model we choose the following simple additive model

$$\boldsymbol{y}(t) = \boldsymbol{\mu}(t) + \boldsymbol{\varepsilon}(t) + \boldsymbol{\eta}(t), \quad t \in \mathbb{Z}, \tag{1}$$

where $\boldsymbol{\mu}(t) = (\mu_1(t), \ldots, \mu_k(t))^\mathsf{T} \in \mathbb{R}^k$ represents the time-varying level of the components of $\boldsymbol{y}(t)$, while $\boldsymbol{\varepsilon}(t) = (\varepsilon_1(t), \ldots, \varepsilon_k(t))^\mathsf{T} \in \mathbb{R}^k$ is observational noise with $\mathsf{E}(\boldsymbol{\varepsilon}(t)) = \mathbf{0}$ and smoothly varying covariance matrix $\mathsf{Var}(\boldsymbol{\varepsilon}(t)) = \boldsymbol{\Sigma}(t)$. An outlier generating mechanism described by $\boldsymbol{\eta}(t) \in \mathbb{R}^k$ produces impulsive noise, which can affect more than one component at different time points. The signals $\mu_j(t)$, $j = 1, \ldots, k$, $t \in \mathbb{Z}$, are assumed to be smooth with some trends and few abrupt level shifts. To allow for correlations between the error components $\varepsilon_i(t)$ and $\varepsilon_j(t)$, $i \neq j$, the local covariance matrices $\boldsymbol{\Sigma}(t)$ can be non-diagonal. The actual observations may be on a discrete scale.

Online extraction of the k-dimensional signal vector $\boldsymbol{\mu}(t)$ can be achieved, e.g. by recursive filtering or moving window techniques. In the following, the focus is on filtering methods that rely on moving a time window $\{\boldsymbol{y}(t-w), \ldots, \boldsymbol{y}(t), \ldots \boldsymbol{y}(t+w)\}$ of width $N = 2w + 1$ through the series. For each time sequence the signal values $\boldsymbol{\mu}(t)$ in the center of the time window are approximated. A short time delay is achieved by choosing w small, though there is a trade-off with smoothness, which comes with longer time windows.

Online procedures for univariate robust signal extraction have been derived and discussed in the literature (see e.g. Gather *et al.*, 2006a,b; Davies *et al.*, 2004, among others). Assuming that the level of a univariate time series $y(t) \in \mathbb{R}$ is almost constant within each time window, location based filters have been suggested to approximate the signal. Efficient denoising is achieved by means of moving averages or other linear filters. However, such methods are easily affected by outliers. Robust methods, like running medians (Tukey, 1977) resist outliers, but have shortcomings in trend periods. A compromise are modified trimmed means (MTM; Lee and Kassam, 1985). Compared to running medians, they are more efficient under Gaussian noise and they also resist a few spikes successfully. In trend periods, however, the location model is not appropriate. To overcome the drawback of location based filters, Davies *et al.* (2004) propose to fit a linear trend $y(t + s) = \mu(t) + \beta(t)s$, $s = -w, \ldots, w$, within each time window. Here, regression estimators with high breakdown point can be applied, such as L_1 regression (Edgeworth, 1887), the least median of squares (Hampel, 1975; Rousseeuw, 1984), the least-trimmed squares (Rousseeuw, 1983), the repeated median (RM; Siegel, 1982) or deepest regression (Rousseeuw and Hubert, 1999). Davies *et al.* (2004) and Gather *et al.* (2006b) compare these methods with respect to properties such as computation time, robustness and the ability to preserve trends and level changes in the presence of outliers. Based on the results of various simulation studies therein as well as for computational reasons, for univariate online signal extraction the RM has been recommended as best compromise procedure.

Similar to location based MTM filters the RM estimator under the trend model has been modified by trimming observations with large regression residuals. Bernholt *et al.* (2006), Fried (2004) and Gather and Fried (2004) apply linear regression repeatedly within cascading windows. In the following this procedure is referred to as TRM regression. For identical inner and outer windows this filter is location- and scale equivariant, trend invariant (Fried *et al.*, 2006), slightly more efficient, and almost as robust as the RM-filter with a finite sample breakdown point of $\lfloor N/2 \rfloor / N$ (Bernholt *et al.*, 2006).

If one is interested in the level of the time series at the most recent time point $t+w$, one can derive online estimates $\mu^{online}(t+w) = \tilde{\mu}^{RM}(t) + \tilde{\beta}^{RM}(t)w$ (Gather *et al.*, 2006b).

3 Multivariate Online Filters

Assuming the simple model (1) for k-variate time series $\boldsymbol{y}(t) \in \mathbb{R}^k$, $k \geq 1$, the goal is to robustly extract the k-dimensional signal in real time. Under the location model a multivariate running mean is equivalent to univariate running means applied to each component of the response vector $\boldsymbol{y}(t)$. Robustification of location based filters in higher dimensions is not straightforward. A componentwise univariate signal extraction for $\boldsymbol{y}(t)$ could neglect correlations between the error components. There is no canonical extension of the univariate median to the multivariate case which is affine equivariant. Eg., a highly robust and affine equivariant location functional is based on the minimum covariance determinant estimator (MCD; Rousseeuw, 1984, 1985), and Koivunen (1996) constructs a robust MCD-based location filter in higher dimensions.
In this paper we deal with time series data for which the location model is not adequate. Instead, we will assume that the time series can be approximated locally by k linear trends, that is

$$\boldsymbol{y}(t+s) - \boldsymbol{\varepsilon}(t+s) = \boldsymbol{\mu}(t) + \boldsymbol{\beta}(t)s, \quad s = -w, \ldots, w. \tag{2}$$

Thus, in each time window $\{t-w, \ldots, t, \ldots, t+w\}$ a multivariate regression problem with a k-variate response and a univariate regressor ($m = 1$), namely the equidistant time points, has to be solved. Compared to k univariate regression models for each component, model (2) accounts for correlations between the errors of the k response variables. In the following we review some procedures for robust parameter estimation in multivariate linear regression models.

3.1 Robust multivariate linear regression

Let a multivariate, multiple regression model with response $\boldsymbol{y} \in \mathbb{R}^k$ and regressor $\boldsymbol{x} \in \mathbb{R}^m$ be given by

$$\boldsymbol{y} = \boldsymbol{\alpha} + \boldsymbol{\mathcal{B}}^{\mathsf{T}}\boldsymbol{x} + \boldsymbol{\varepsilon}, \tag{3}$$

where the errors $\boldsymbol{\varepsilon} \in \mathbb{R}^k$ are i.i.d. with mean vector $\mathbf{0}$ and positive definite covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}} \in \mathbb{R}^{k\times k}$; $\boldsymbol{\alpha} \in \mathbb{R}^k$ and $\boldsymbol{\mathcal{B}} \in \mathbb{R}^{m\times k}$ are the unknown intercept and slope parameters. Denoting the joint location of the random variables $\boldsymbol{z} = (\boldsymbol{x}^{\mathsf{T}}, \boldsymbol{y}^{\mathsf{T}})^{\mathsf{T}}$ by $\boldsymbol{\mu}$ and and their scatter matrix by $\boldsymbol{\Sigma}$, we write

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy} \end{pmatrix}.$$

Then, the least squares estimator $T_{LS} = (\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\mathcal{B}}}^{\mathsf{T}})^{\mathsf{T}}$ of $(\boldsymbol{\alpha}, \boldsymbol{\mathcal{B}})$ can be expressed as

$$\tilde{\boldsymbol{\mathcal{B}}} = \tilde{\boldsymbol{\Sigma}}_{xx}^{-1}\tilde{\boldsymbol{\Sigma}}_{xy}, \quad \text{and} \quad \tilde{\boldsymbol{\alpha}} = \tilde{\boldsymbol{\mu}}_y - \tilde{\boldsymbol{\mathcal{B}}}^{\mathsf{T}}\tilde{\boldsymbol{\mu}}_x, \tag{4}$$

where the corresponding components of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are estimated by the subvectors and submatrices of the sample mean vector $\tilde{\boldsymbol{\mu}} = \bar{\boldsymbol{z}}$ and the empirical covariance matrix $\tilde{\boldsymbol{\Sigma}} = \boldsymbol{S}_{zz}$. The estimate given by the multivariate LS functional T_{LS} is identical to the matrix resulting from k corresponding marginal univariate LS estimate vectors. It is affine equivariant and optimal under a multinormal error distribution. However, by possessing a breakdown point of zero, it is very sensitive to outliers, which yields a need for robust alternatives.

In most cases applying robust univariate regression methods to each response component leads to estimates that are not affine equivariant. This drawback occurs e.g. for generalizations of the univariate LMS or RM functional, as well as for other robust regression techniques discussed in the literature including L$_1$ regression (Rao, 1988; Bai *et al.*, 1990) and M-estimation (Koenker and Portnoy, 1990). Giving up affine equivariance implies a loss of efficiency if the error components are correlated (Chakraborty, 1999; Bickel, 1964). In the following, we firstly consider multivariate regression techniques that are both affine equivariant as well as highly robust. Estimators based on either quantile regression or multivariate ranks have been proposed, e.g. by Chakraborty and Chaudhuri (1997); Chakraborty (1999, 2003) and Ollila *et al.* (2003). Rousseeuw *et al.* (2004) have suggested the use of a regression functional $T_{reg} = (\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\mathcal{B}}}^{\mathsf{T}})^{\mathsf{T}}$ based on (4).

Regression-, x- and y-equivariance of T_{reg} are gained if $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are replaced by affine equivariant estimators. Also, the regression functional T_{reg} inherits the minimal breakdown point of the corresponding functionals for multivariate location and scatter. Many highly robust estimators of these quantities have been discussed in the literature, including, for example, the Stahel-Donoho estimator (Stahel, 1981; Donoho, 1982), the MCD estimator (Rousseeuw, 1983, 1984), S-estimators (Davies, 1987; Lopuhaä, 1989), and CM estimators (Kent and Tyler, 1996). Rousseeuw *et al.* (2004) suggest to estimate location and scatter of the random vector $(\boldsymbol{x}^\mathsf{T}, \boldsymbol{y}^\mathsf{T})^\mathsf{T}$ by means of the MCD. They call the resulting regression method *MCD regression.*

A different approach has been pursued by Agulló *et al.* (2006), who construct regression estimates based on the covariance matrix of the residuals. Denote for any matrix $(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\mathcal{B}}}^\mathsf{T})^\mathsf{T} = \boldsymbol{B} \in \mathbb{R}^{(m+1)\times k}$ of regression coefficients the corresponding residuals by $\boldsymbol{r}_i(\boldsymbol{B}) = \boldsymbol{y}_i - \boldsymbol{B}^\mathsf{T}\boldsymbol{x}_i$ and for each positive definite and symmetric matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{k\times k}$ the squared distances of the residuals with respect to $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$ by $d_i^2(\boldsymbol{B}, \boldsymbol{\Sigma}) = \boldsymbol{r}_i(\boldsymbol{B})^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{r}_i(\boldsymbol{B})$. The multivariate least-trimmed squares (MLTS) estimator is then defined as

$$\underset{\boldsymbol{B},\, \boldsymbol{\Sigma};|\boldsymbol{\Sigma}|=1}{\operatorname{argmin}} \sum_{j=1}^{h} d_{(j)}^2(\boldsymbol{B}, \boldsymbol{\Sigma}), \tag{5}$$

where $d_{(j)}^2$, $j = 1, \ldots, n$, denote the ordered squared distances. Jung (2005) proposed a similar "least-trimmed Mahalanobis squares regression estimator".

Agulló *et al.* (2006) have shown further that any $\boldsymbol{B}$ which minimizes the determinant of the MCD scatter estimate of its regression residuals is a solution of (5). For $k > 1$ the finite sample replacement breakdown point of the MLTS estimator is equal to $\min(N - h + 1, h - g(\underline{\boldsymbol{Z}}_N))/N$, where $g(\underline{\boldsymbol{Z}}_N)$ is the maximal number of observations in the sample $\underline{\boldsymbol{Z}}_N$ lying on the same hyperplane through the origin of $\mathbb{R}^{k+m+1}$ and $h > g(\underline{\boldsymbol{Z}}_N)$ (Agulló *et al.*, 2006).

Both, the MLTS and the MCD regression estimator have a maximal breakdown point of $\lfloor (N - (k + m) + 1)/2 \rfloor / N$. However, these optimal breakdown points can only be attained if the data is in general position.

Additionally, for both of the above MCD-based regression methods reweighting steps are recommended as the raw estimators suffer from low efficiency.

3.2 Regression-based multivariate online filters

The routine application of procedures for online signal extraction affords the existence of a unique solution, to be found within short computation time, and a high robustness with respect to outliers. Additionally, a satisfactory finite sample efficiency under the Gaussian and under some other distributions is important. Although the application of robust univariate regression techniques to the com-

ponents in the multivariate model leads to a loss of efficiency if the error components are correlated, this is an appealing approach due to resulting high finite sample breakdown points and fast computability.
Within the class of affine equivariant multivariate regression methods only the MCD-based methods possess high breakdown points. Note, that for MCD regression the common distribution of the response and regressor variables must be elliptical in order to get Fisher consistent estimates of $(\boldsymbol{\mu}, \boldsymbol{\beta})$. However, in model (2) the only regressor is time, i.e. we have equally spaced design points.

Computational aspects
A disadvantage of MCD-based regression methods is that an exact algorithm for computing the MCD needs $O(N^{k(k+3)/2})$ time (Bernholt and Fischer, 2004). An online application of MCD-based filters is therefore only possible when based on the fast but heuristic Fast-MCD algorithm by Rousseeuw and Van Driessen (1999), which yields approximative solutions.

Sample size versus efficiency and breakdown point
Online procedures are most effective if the time delay of the estimate is as small as possible while also ensuring that the signal is sufficiently smooth and robust. A short delay is achieved by choosing the window width small. This means that within a time window for the multivariate regression the sample size N might only be slightly larger than the dimension k. However, for small sample sizes MCD-based estimators are known to be not very efficient. Also, the optimal finite sample breakdown point of the affine equivariant MCD-based regression functionals is close to 0.5 for large N if the data is in general position, but for small sample sizes it can be much lower.

General position of the data
In model (3) it is typically assumed that the covariance matrix of the error terms has full rank k. The finite sample breakdown point of the MLTS functional depends on the maximal number of observations within a smaller subspace of $\mathbb{R}^{k+m+1}$. If $h \geq [(N+k+m+1)/2]$ is the size of the optimal subsample used for the calculation of the MCD estimate and if h or more observations lie in the same hyperplane, the regularity condition $g(\underline{\boldsymbol{Z}}_N) < h$ is violated and the MLTS estimator is not well defined. The MCD estimate then degenerates to a singular matrix. At least one dimension of the response variable is lost and the link between response and regressor can be described on a space with dimension less than k.
In practice, the assumption that the data is in general position is often not fulfilled if the observations of a time series are measured on a discrete scale. If the error covariance matrix in model (2) is allowed to have rank $r \leq k$, a robust and affine equivariant estimation of the regression parameters could be achieved by first estimating the rank r and then transforming the observations into the corresponding $r-$dimensional subspace, possibly by means of a robust PCA (Li and Chen, 1985; Croux and Ruiz-Gazen, 1996, 2005; Hubert *et al.*, 2005). MCD-based regression can thereafter be performed based on the principal components.

However, this approach needs too much computation time to be applicable in the online situation.
An ad-hoc solution, often used in data mining is to add negligible noise to the observations. In order to avoid inliers, Koivunen (1996) applies a MCD-based location filter to discrete data with added noise. Simulations even show the paradox effect of a lower mean squared error of MCD estimates if additional random noise has been added to the data. This is due to the resulting higher dimension: The relative efficiency of the MCD estimator increases with increasing dimension (Croux and Haesbroeck, 1999) and this property is inherited by MCD-based regression estimators (Croux *et al.*, 2001; Rousseeuw *et al.*, 2004; Agulló *et al.*, 2006) even if the additional variables are completely random. However, as thereby the choice of the optimal subsample for the MCD estimator may change this is not always appropriate.

The above discussion shows, that there is no multivariate regression procedure for signal extraction with all desirable properties: high robustness, high relative efficiency due to affine equivariance, fast computation, unique solution, and the ability to cope with data that are not in general position.

3.3 A new multivariate robust online filter

In this section, we present an alternative procedure for multivariate online signal extraction which is based on the idea of univariate TRM filters (Bernholt *et al.*, 2006).
First, we will weaken the requirement of affine equivariance. Applying the univariate TRM filters to multivariate time series completely neglects the possibility of high correlations among the error components such that separate componentwise trimming procedures may not be able to detect outliers related to the multivariate dependence structure. We suggest the following multivariate generalization:

1. Within each time window $\{t-w,\dots,t,\dots,t+w\}$ use the RM functional $T_{RM}=(\tilde{\mu}(t),\tilde{\beta}(t))$ in order to estimate the local level $\mu_j(t)$ and the local slope $\beta_j(t)$ for each component $y_j(\cdot)$, $j=1,\dots,k$, that is
$$\begin{aligned}\tilde{\beta}_j^{RM}(t) &= med_{s\in\{-w,\dots,w\}}\left(med_{v\neq s,v\in\{-w,\dots,w\}}\frac{y_j(t+s)-y_j(t+v)}{s-v}\right),\\ \tilde{\mu}_j^{RM}(t) &= med_{s\in\{-w,\dots,w\}}\left(y_j(t+s)-\tilde{\beta}^{RM}(t)s\right).\end{aligned}$$
Combine these estimates to initial multivariate estimators $\tilde{\boldsymbol{\beta}}^{RM}(t)=(\tilde{\beta}_1^{RM}(t),\dots,\tilde{\beta}_k^{RM}(t))^\mathsf{T}$ and $\tilde{\boldsymbol{\mu}}^{RM}(t)=(\tilde{\mu}_1^{RM}(t),\dots,\tilde{\mu}_k^{RM}(t))^\mathsf{T}$.
2. Compute multivariate residuals of the regression lines within the current time window: $\boldsymbol{r}(t+s)=\boldsymbol{y}(t+s)-\tilde{\boldsymbol{\mu}}^{RM}(t)-s\tilde{\boldsymbol{\beta}}^{RM}(t)$, $s=-w,\dots,w$.

3. Obtain robust estimates of the local covariance matrix $\tilde{\boldsymbol{\Sigma}}(t)$ of the error terms based on the residuals $\boldsymbol{r}(t+s)$, $s = -w, \ldots, w$.
4. Find the subset of time points within the time window corresponding to residuals, whose squared Mahalanobis distance w.r.t. the local covariance structure is lower than a given value a_N, that is $S_t = \{s = -w, \ldots, w : \boldsymbol{r}(t+s)^{\mathsf{T}} \hat{\boldsymbol{\Sigma}}(t)^{-1} \boldsymbol{r}(t+s) \leq a_N\}$.
5. Obtain estimates of the local level $\tilde{\boldsymbol{\mu}}^{TRM-LS}(t)$ and slope $\tilde{\boldsymbol{\beta}}^{TRM-LS}(t)$ by means of multivariate LS regression from the trimmed observations $\{(t+s, \boldsymbol{y}(t+s)),\ s \in S_t\}$ within the actual time window.

The filter based on this procedure is called *multivariate TRM-LS filter.*

In step 3 a robust estimation of the local covariance matrix $\boldsymbol{\Sigma}(t)$ based on a small sample is needed. From the discussion in subsection 3.2 affine equivariant estimators may not be appropriate here. A robust estimator of the covariance matrix, that is fast to compute although not affine equivariant, is the orthogonalized Gnanadesikan-Kettenring estimator (OGK estimator; Maronna and Zamar, 2002). In order to estimate the covariance between a pair of random variables X and Y (Gnanadesikan and Kettenring, 1972) use the fact that $\mathsf{Cov}(X,Y) = (\sigma(X+Y)^2 - \sigma(X-Y)^2)/4$, where $\sigma(\cdot)$ denotes the standard deviation. Maronna and Zamar (2002) have modified the covariance matrix resulting from the estimated pairwise Gnanadesikan-Kettenring covariances, based on some robust scale estimate $\sigma(\cdot)$, such that a positive definite and approximately affine equivariant matrix is achieved. The computation of the OGK$_\sigma$ estimator for sample variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N \in \mathbb{R}^k$, and a robust univariate scale functional $\sigma(\cdot)$ requires the following steps

i. Scale the sample variables via $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{D}^{-1}$, where $\boldsymbol{D} = diag(\sigma(X_1), \ldots, \sigma(X_k))$.
ii. Apply the Gnanadesikan-Kettenring estimator to the columns of the scaled variables $\boldsymbol{Y}$ and obtain a robust correlation matrix $\boldsymbol{R}$ of $\boldsymbol{X}$ with $R_{jj} = 1$ and $R_{ij} = (\sigma(Y_i + Y_j)^2 - \sigma(Y_i - Y_j)^2)/4$, $i \neq j$.
iii. Perform a spectral decomposition $\boldsymbol{R} = \boldsymbol{E}\boldsymbol{\Lambda}\boldsymbol{E}^{\mathsf{T}}$, where $\boldsymbol{\Lambda} = diag(\lambda_1, \ldots, \lambda_k)$ contains the ordered eigenvalues and $\boldsymbol{E}$ contains the corresponding eigenvectors of $\boldsymbol{R}$.
iv. Define $\boldsymbol{A} = \boldsymbol{D}\boldsymbol{E}$ and $\boldsymbol{Z} = \boldsymbol{X}(\boldsymbol{A}^{\mathsf{T}})^{-1}$. With $\boldsymbol{\Gamma} = diag(\sigma(Z_1)^2, \ldots, \sigma(Z_k)^2)$ the OGK$_\sigma$ estimator is defined as OGK$_\sigma(\boldsymbol{X}) = \boldsymbol{A}\boldsymbol{\Gamma}\boldsymbol{A}^{\mathsf{T}}$.

If, due to inliers, for some variables X_j, $j = 1, \ldots, k$, or Z_j, $j = 1, \ldots, k$, the univariate scale estimate become zero in step i. or iv., the estimated covariance matrix becomes singular. As the estimate of the covariance matrix is only used to trim the residuals, we can replace values of zero by a small lower threshold ϑ in order to ensure invertibility of the matrix. The OGK$_\sigma$ estimator is found based on $\sigma(\cdot) = \max(\tilde{\sigma}(\cdot), \vartheta)$, where $\tilde{\sigma}(\cdot)$ is a univariate scale functional with optimal finite sample breakdown point of 50% and ϑ is an appropriate lower threshold for the variability in each direction, e. g. $\vartheta = 0.02$.

Various highly robust univariate scale functionals are discussed in Gather and

Fried (2003) w.r.t. application in online filtering methods. For the computation of the OGK estimator we will only consider the well known MAD estimator $\sigma_{MAD} = c_N^{MAD}\,\mathrm{med}(|x_1 - \tilde{\mu}|, \ldots, |x_N - \tilde{\mu}|)$ and the Q_N estimator (Rousseeuw and Croux, 1993) $\sigma_{Q_N} = c_N^{Q_N}\{|x_i - x_j| : 1 \leq i < j \leq N\}_{(h)}$, $h = \binom{[N/2]+1}{2}$. The constants c_N^{MAD} and $c_N^{Q_N}$ are correction factors, chosen to achieve unbiasedness under Gaussian noise. An advantage of the MAD estimator is the existence of an update algorithm, that affords only $O(\log N)$ time. Ma and Genton (2001) recommend to use the Q_N estimator for the computation of the OGK estimator and Gather and Fried (2003) also describe a good performance of the Q_N estimator in the presence of inliers and level shifts.

In step 4 of the procedure above an upper trimming threshold a_N is required. Typical choices are $a_N = \chi_k^2(\beta)$, where $\chi_k^2(\beta)$ is the $\beta-$quantile of a χ^2-distribution with k degrees of freedom or $a_N = \chi_k^2(\beta)\,\mathrm{med}(d(-w), \ldots, d(w))/\chi_k^2(0.5)$, where $d(s) = \boldsymbol{r}(t+s)^\mathsf{T}\hat{\boldsymbol{\Sigma}}(t)^{-1}\boldsymbol{r}(t+s)$, $s = -w, \ldots, w$, (Maronna and Zamar, 2002). As the maximum possible finite-sample explosion breakdown point of the OGK estimator is equal to that of the univariate scale estimator $\sigma(\cdot)$, the OGK$_\sigma$ estimator based on the MAD or the Q_N estimator has a maximal breakdown point of 50%, if the data does not show ties within the windows.

3.3.1 Performance of the OGK estimator in small samples

The OGK$_\sigma$ estimator is highly robust, flexible and fast to compute, but not affine equivariant. To get an idea of its performance when applied to small samples we compare the OGK$_\sigma$ estimator based on the MAD and the Q_N estimator to the empirical covariance matrix and the Fast-MCD estimator. We evaluate a measure of the sphericity of $\boldsymbol{\Psi} = \boldsymbol{\Sigma}^{-1/2}\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1/2}$ which is chosen as condition number $\varphi = \mathit{cond}(\boldsymbol{\Psi}) = ||\boldsymbol{\Psi}||_2||\boldsymbol{\Psi}^{-1}||_2$ of $\boldsymbol{\Psi}$, where $\hat{\boldsymbol{\Sigma}}$ is an estimate of $\boldsymbol{\Sigma}$. This quantity measures the mean deviation of the estimation by $\hat{\boldsymbol{\Sigma}}$ and is invariant under affine transformations.
In the simulation study we will consider rather small sample sizes $N = 31, 51$ and 101 which for instance occur as window widths in intensive care settings. The observations are generated from a 10-variate normal distribution with expectation $\mathbf{0}$ and covariance matrices $\boldsymbol{\Sigma}_j = (1 - c_j) \cdot I_{10} + c_j \cdot 1_{10}1_{10}^\mathsf{T}$, where $c_j = \frac{j}{10}$, $j = 0, 5, 9$. Additionally, samples are contaminated as follows: $(1-\varepsilon)N$,, fixed $\varepsilon = 0.2$, observations are drawn from the above normal distributions and the remaining εN observations are generated from a normal distribution with $\mathcal{N}(l\boldsymbol{\Sigma}_j^{1/2}\boldsymbol{v}_j, \tau^2\boldsymbol{\Sigma}_j)$, $j = 0, 5, 9$, where $\boldsymbol{v}_j$ is the scaled eigenvector of the smallest eigenvalue of $\boldsymbol{\Sigma}_j$. We further choose $\tau = 0.1$, such that the contaminated observations are close to each other. The amount l is varied, with $l \in \{5, 7, 10, 15, 20, 40\}$. For each sample the logarithm of the condition number of $\boldsymbol{\Sigma}^{-1/2}\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1/2}$ is obtained, where $\hat{\boldsymbol{\Sigma}}$ is the corresponding estimator. Ideally, the logarithm of the condition number is 0. Note, that the condition numbers of affine equivariant

covariance estimators, such as the empirical covariance and the Fast-MCD, do not depend on the correlation structure.

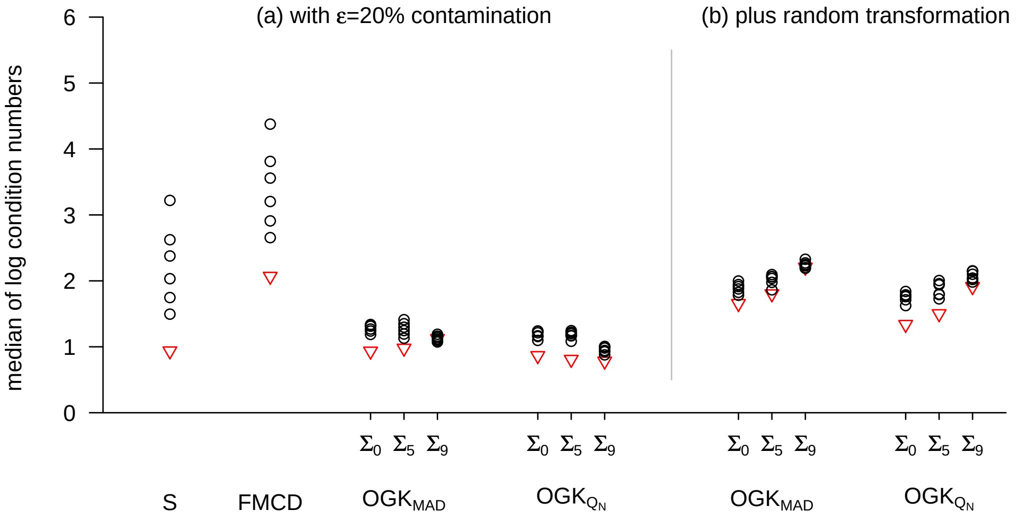

Fig. 1. *Median of log condition numbers for covariance estimators at normal distributions with different covariance structures without (triangles) and with (circles) contamination and under additional random transformations, sample size* $N = 31$

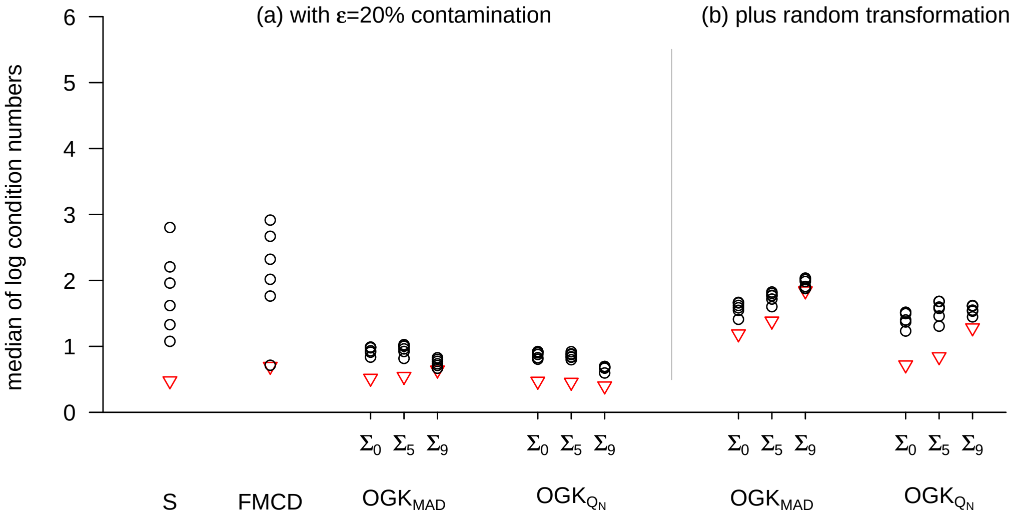

Fig. 2. *Median of log condition numbers for covariance estimators at normal distributions with different covariance structures without (triangles) and with (circles) contamination and under random transformations, sample size* $N = 101$

Figure 1 shows the median of the logarithm of the condition numbers for each of the estimators and the different distributions for a sample size of $N = 31$. Triangles represent results for the non-contaminated observations, circles represent results from samples with $\varepsilon = 20\%$ contamination, where l varies.

Without contamination, the OGK$_\sigma$ estimators perform almost as well as the empirical covariance matrix $\boldsymbol{S}$, though increasing the correlation between the variables reduces their performance. As expected, the Fast-MCD estimator performs worst for small sample sizes. While the empirical covariance matrix is highly affected by the contaminated observations, the influence on the sphericity of the OGK$_\sigma$ estimator is small, and the OGK$_{Q_N}$ performs slightly better than the OGK$_{MAD}$.

We get similar results for sample sizes of $N = 51$ (not shown) and $N = 101$ (cf. Figure 2). The Fast-MCD estimator yields better results with increasing sample size.

In order to investigate how much the OGK$_\sigma$ suffers from the lack of equivariance, the samples were additionally transformed by randomly generated orthogonal matrices (Maronna and Zamar, 2002). The performance of OGK$_\sigma$ under transformation is then measured by the condition numbers. The medians of the logarithm of the condition numbers are also shown in Figures 1, and 2, respectively. As expected, the condition numbers increase for the estimation under transformations due to the lack of affine equivariance. However, for the sample size $N = 31$ the OGK$_\sigma$ under transformation is still better than the Fast-MCD. This advantage gets lost for larger sample sizes.

3.4 Simulation study

We have conducted a simulation study to compare some robust regression estimators with respect to their finite sample efficiency under different distributions and dependence structures of the error terms and their use for online signal extraction.

The following methods are included: the univariate trimmed repeated median T_{TRM} estimator with scale estimation based on the Q$_N$ estimator, the MCD and MLTS regression estimators T_{MCD} and T_{MLTS} with reweighting steps, and the proposed method based on LS estimation after multivariate trimming of the repeated median residuals T_{TRM-LS}. To guarantee a high finite sample breakdown point, we calculate the MCD estimator based on a subsample of size $h = [(N + k + 2)/2]$. The trimming constant δ for the reweighting step is chosen as $\delta = 0.975$. For the multivariate TRM-LS signal extraction procedure the OGK$_{Q_N}$ estimator is used for trimming with a trimming threshold $d_N = \chi^2_{10}(0.95)\,\mathrm{med}(d(-w), \ldots, d(w))/\chi^2_{10}(0.5)$, where $d(s) = \boldsymbol{r}(t + s)^\mathsf{T}\hat{\boldsymbol{\Sigma}}(t)^{-1}\boldsymbol{r}(t + s)$, $s = -w, \ldots, w$. The observations are generated via

$$\boldsymbol{X}(t) = \boldsymbol{\mu} + \boldsymbol{\beta}t + \boldsymbol{\varepsilon}(t), \quad t = -w, \ldots, w, \tag{6}$$

with response $\boldsymbol{X}(t) \in \mathbb{R}^k$ and $k = 10$. As all regression functionals are regression equivariant, we set the regression coefficients to $\boldsymbol{\mu} = \boldsymbol{\beta} = \mathbf{0} \in \mathbb{R}^{10}$. The errors $\boldsymbol{\varepsilon}(t) \in \mathbb{R}^{10}$ are generated from a 10−dimensional normal distribution with expectation $\mathbf{0}$ and from a $t-$distribution with three degrees of freedom (t_3), while the covariance matrices are chosen according to the following schemes: $\mathbf{\Sigma}_j = (1 - c_j) \cdot I_{10} + c_j \cdot 1_{10}1_{10}^{\mathsf{T}}$, where $c_j = \frac{j}{10}$, $j = 0, 2, 4, 6, 8, 9$. Thus, $\mathbf{\Sigma}_0$ describes independently distributed random variables of a $10-$dimensional normal or $t-$distribution while the other covariance matrices are chosen according to uniform correlation models with stepwise increasing correlation. The sample sizes are chosen as $N = 31$ ($N = 51$, 101, 501). For each model and sample size, 5,000 samples are generated. As performance criterion we calculate the relative efficiency of the respective estimators to the LS estimator, which is here defined as 20th square root ($2 \times k = 20$) of the estimated ratio of Wilk's generalized variances of the two estimators (Chakraborty, 1999, 2003; Ollila *et al.*, 2003).

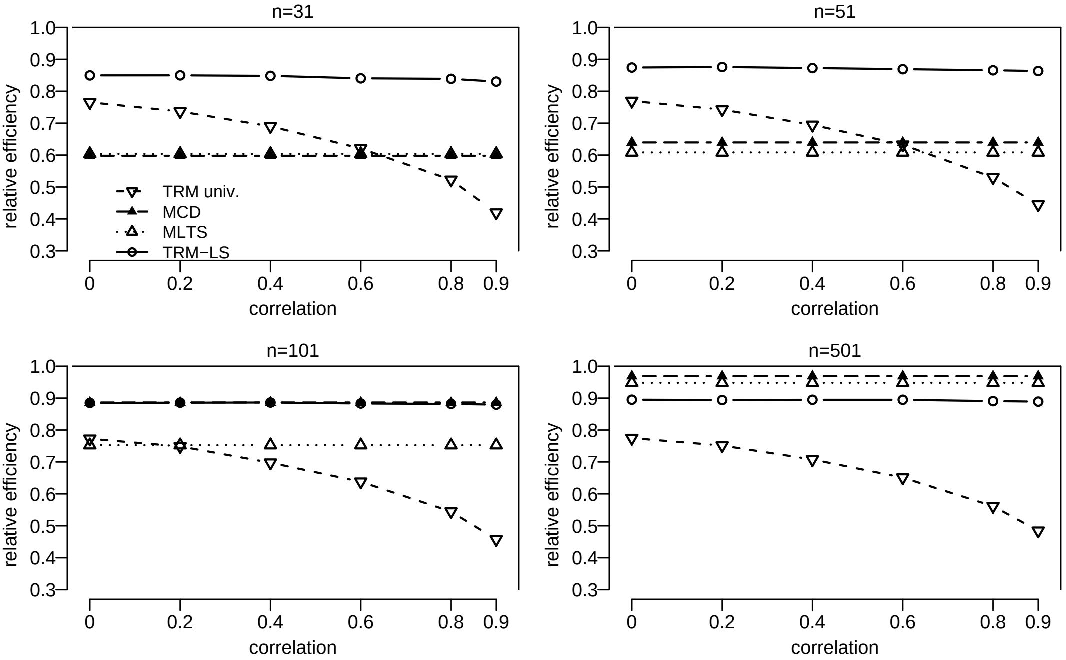

Fig. 3. *Relative efficiencies of the regression estimators compared to the LS estimator for different sample sizes and correlation structures at a normal distribution*

The results in Figures 3 and 4 show that for fixed sample size N the empirical relative finite sample efficiencies of the MCD-based regression functionals do not depend on the dependence structure of the error components due to their affine equivariance. As expected, the relative efficiencies of the MCD-based methods are rather low for small sample sizes. For t_3-distributed error terms with heavy tails the MCD-based estimates are more efficient compared to the very sensitive LS estimator.

Also, the relative efficiency of the univariate TRM estimator decreases rapidly with increasing correlation between the error components. However, for low to moderately high correlation between the error components we find similar finite sample efficiencies for the MCD-based estimators and the univariate TRM estimator.

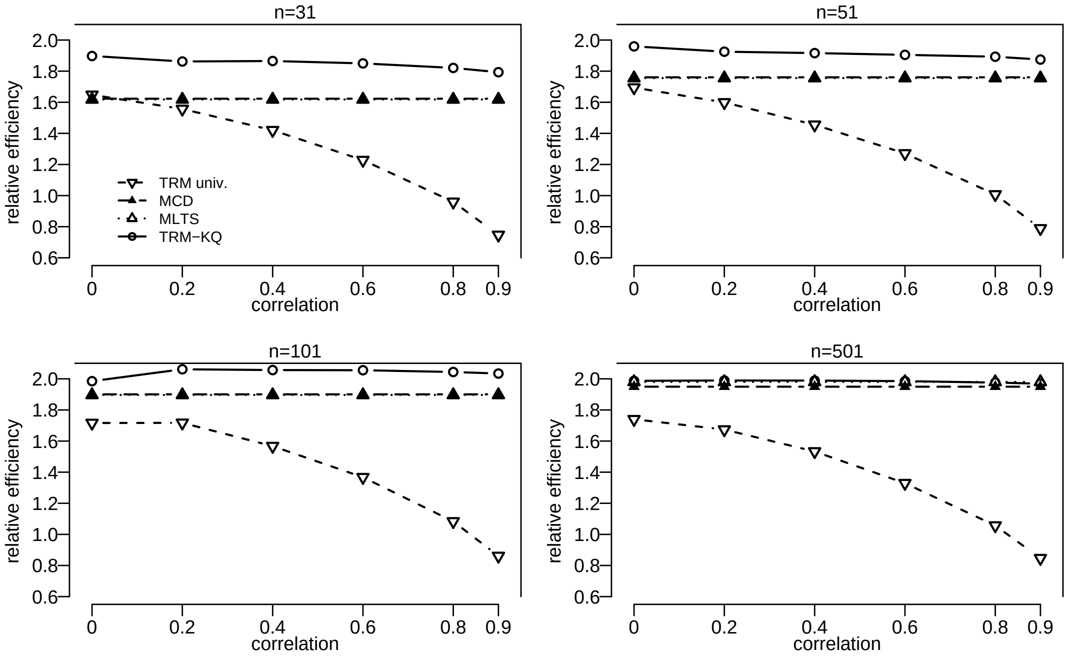

Fig. 4. *Relative efficiencies of the regression estimators to the LS estimator for different sample sizes and correlation structures at a t_3 distribution*

The new multivariate TRM-LS estimator performs best in all situations. Although it is not affine equivariant the relative efficiency remains large even in situations where the correlation between the error components is high, in particular, for small sample sizes . For larger sample sizes (here: N=501) the relative efficiency is comparable to that of the MCD-based regression estimators.

Summarizing it can be concluded that the TRM-LS two-step procedure offers a fast, highly robust, and highly efficient online signal extraction even if window widths, and thus sample sizes are small and the data is not necessarily in general position.

3.5 Application

In this subsection the new multivariate TRM-LS filter is applied to a time series of highly correlated physiological variables, such as blood pressures and heart rate, that has been recorded for a patient on an intensive care unit at the Klinikum Dortmund. The aim is to preserve clinically relevant patterns such as sudden level shifts and trends, while noise and irrelevant artifacts are removed. Figure 5 shows such a time series with variables of the hemodynamic system. We apply the TRM-LS filter to this time series in order to extract the underlying signal. In practice, this signal extraction has to be performed in real time. Figure 6 shows a part of the data together with the extracted signal.

As can be seen, the approximated signal preserves clinically relevant patterns of the hemodynamic data without being influenced by spikes and noise.

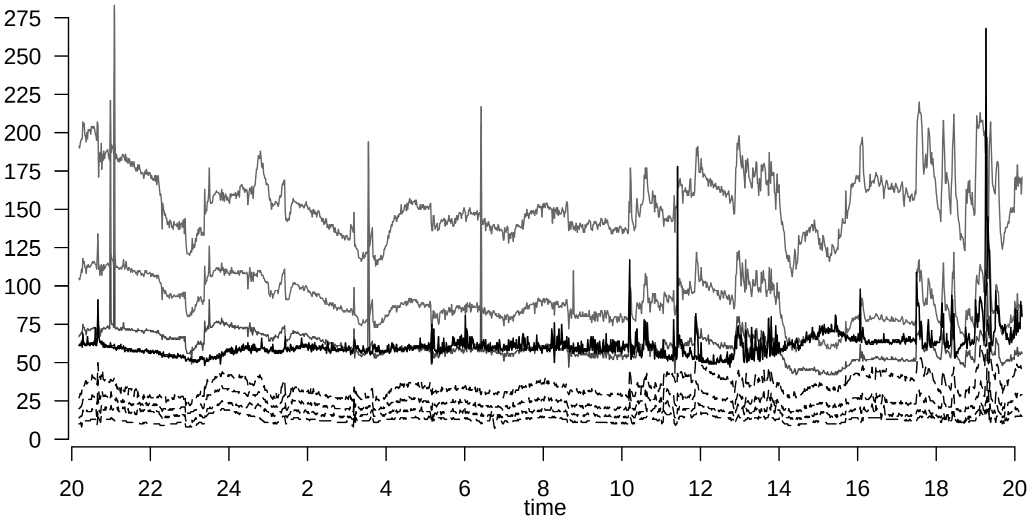

Fig. 5. *Time series with nine hemodynamic variables of a patient in intensive care: arterial blood pressures (grey), heart rate, pulse (black), pulmonary artery blood pressures and central venous blood pressure (dashed)*

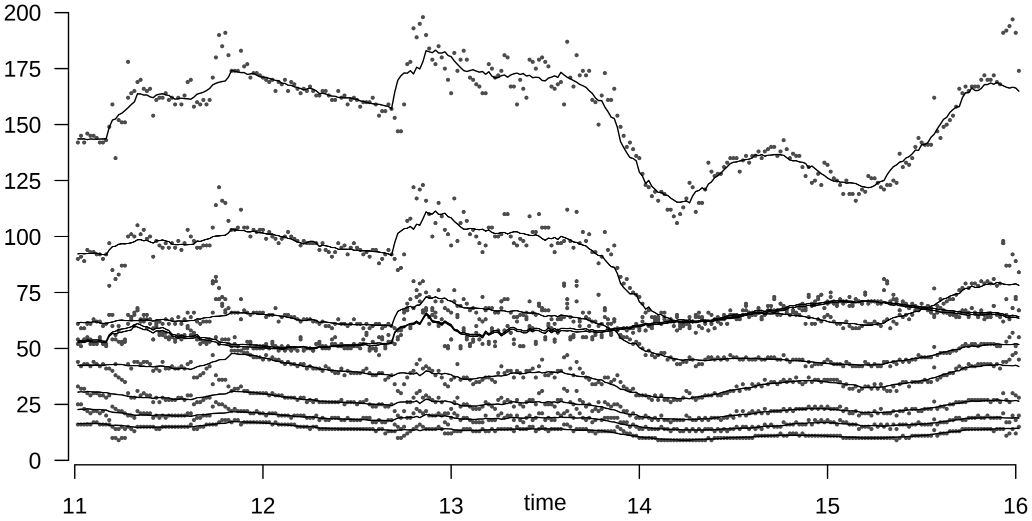

Fig. 6. *Extract of the hemodynamic time series (dotted) from figure 5 with extracted signals: arterial blood pressures, heart rate, pulse, pulmonary artery blood pressures and central venous blood pressure*

4 Discussion and conclusion

We have proposed a new two-step procedure for multivariate signal extraction, which is fast and highly robust. Our method uses a moving outer window and is based on least squares regression estimates, that are obtained from trimmed observations after a componentwise RM regression in an inner time window. The correlation structure between the error components is taken into account by means of orthogonalized Gnanadesikan-Kettenring covariance estimates, that are fast to compute and highly robust. The resulting method for online signal extraction is not affine equivariant but has very good efficiency properties if short time windows are used. Moreover, it can be used for discretely measured data with low variability as well as in situations with many outliers.
Further ongoing research will investigate the estimation of the signal at the end of the respective time windows and the possibility to adjust the window widths according to the structure of the time series. This is especially important for an improved preservation and detection of level shifts.

References

AGULLÓ, J., CROUX, C., VAN AELST, S. (2006) The multivariate least-trimmed squares estimator, *Journal of Multivariate Analysis*, doi:10.1016/j.jmva.2006.06.005.

BAI, Z. D., CHEN, N. R., MIAO, B. Q., RAO, C. R. (1990) Asymptotic theory of least distance estimate in multivariate linear models, *Statistics*, **21**, 503–529.

BERNHOLT, T., FISCHER, P. (2004) The complexity of computing the MCD-estimator, *Theoretical Computer Science*, **326**, 383–398.

BERNHOLT, T., FRIED, R., GATHER, U., WEGENER, I. (2006) Modified repeated median filters, *Statistics and Computing*, **16**, 177–192.

BICKEL, P. J. (1964): On some alternative estimates for shift in the $p-$variate one sample problem, *Annals of Mathematical Statistics*, **35**, 1079–1090.

CHAKRABORTY, B. (1999) On multivariate median regression, *Bernoulli*, **5**, 683–703.

CHAKRABORTY, B. (2003) On multivariate quantile regression, *Journal of Statistical Planning and Inference*, **110**, 109–132.

CHAKRABORTY, B., CHAUDHURI, P. (1997) On multivariate rank regression, In: *L1-Statistical Procedures and Related Topics*, IMS Lecture Notes-Monograph Series, Vol. **31**, Dodge, Y. (ed.), 399–414.

CRACK, T. F., LEDOIT, O. (1996) Robust structure without predictability: the "compass rose" pattern of the stock market, *The Journal of Finance*, **51**, 751–762.

CROUX, C., DEHON, C., ROUSSEEUW, P. J., VAN AELST, S. (2001) Robust estimation of the conditional median function at elliptical models, *Statistics and Probability Letters*, **51**, 361–368.

CROUX, C., HAESBROECK, G. (1999) Influence function and efficiency of the minimum covariance determinant scatter matrix estimator, *Journal of Multivariate Analysis*, **71**, 161–190.

CROUX, C., RUIZ-GAZEN, A. (1996) A fast algorithm for robust principal components based on projection pursuit, In: *Proceedings in Computational Statistics, COMPSTAT*, Physica–Verlag, 211–216.

CROUX, C., RUIZ-GAZEN, A. (2005) High breakdown estimators for principal components: the projection-pursuit approach revisited, *Journal of Multivariate Analysis*, **95**, 206–226.

DAVIES, P. L. (1987) Asymptotic behavior of S-estimates of multivariate location parameters and dispersion matrices, *Annals of Statistics*, **15**, 1269–1292.

DAVIES, P. L., FRIED, R., GATHER, U. (2004) Robust signal extraction from on–line monitoring data, *Journal of Statistical Planning and Inference*, **122**, 65–78.

DONOHO, D. L. (1982) Breakdown properties of multivariate location estimators, PH.D. qualifying paper, Harvard university.

EDGEWORTH, F. J. (1887) A new method of reducing observations relating to several questions, *Phil. Mag.*, **24**, 184–191.

FRIED, R. (2004) Robust filtering of time series with trends, *Nonparametric Statistics*, **16**, 313–328.

FRIED, R., BERNHOLT, T., GATHER, U. (2006) Repeated median and hybrid filters, *Computational Statistics and Data Analysis*, **50**, 2313–2338.

GATHER, U., FRIED, R. (2003) Robust estimation of scale for local linear temporal trends, In: *Proceedings of the 4th International Conference on Mathematical Statistics PROBASTAT 2002*, Stulajter, F. (ed.), Tatra Mountain Mathematical Publications, **26**, 87–101.

GATHER, U., FRIED, R. (2004) Methods and algorithms for robust filtering, In: *Proceedings in Computational Statistics COMPSTAT 2004*, Antoch, J. (ed.), Physica-Verlag, Heidelberg, 159–170.

GATHER, U., FRIED, R., LANIUS, V. (2006) Robust detail-preserving signal extraction, In: Handbook of Time Series Analysis, Schelter, B., Winterhalder, M., Timmer, J. (eds.), Weinheim: Wiley-VCH, Chapter 6, 143–169.

GATHER, U., SCHETTLINGER, K., FRIED, R. (2006) Online signal extraction by robust linear regression, *Computational Statistics*, **21**, 33–51.

GNANADESIKAN, R., KETTENRING, J. R. (1972) Robust estimates, residuals, and outlier detection with multiresponse data, *Biometrics*, **28**, 81–124.

HAMPEL, F. R. (1975) Beyond location parameters: Robust concepts and methods, *Bulletin of the Int. Statist. Inst.*, **46**, 375–382.

HUBERT, M., ROUSSEEUW, P. J., VANDEN BRANDEN, K. (2005) ROBPCA: a new approach to robust principal component analysis, *Technometrics*, **47**, 64–79.

JUNG, K.-M. (2005) Multivariate least-trimmed squares regression estimator, *Computational Statistics and Data Analysis*, **48**, 307–316.

KENT, J. T., TYLER, D. E. (1996): Constrained M-estimation for multivariate location and scatter, *Annals of Statistics*, **24**, 1346–1370.

KOENKER, R., PORTNOY, S. (1990) M-estimation of multivariate regressions, *Journal of the American Statistical Association*, **85**, 1060–1068.

KOIVUNEN, V. (1996) Nonlinear filtering of multivariate images under robust error criterion, *IEEE Transactions on Image Processing*, **5**, 1054–1060.

LANIUS, V. (2005) Statistische Extraktion relevanter Informationen aus multivariaten Online-Monitoring-Daten der Intensivmedizin, Ph.D. thesis, Department of Statistics, University of Dortmund.

LEE, Y., KASSAM, S. (1985) Generalized median filtering and related nonlinear filtering techniques, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **33**, 672–683.

LI, G., CHEN, Z. (1985) Projection-pursuit approach to robust dispersion matrices and principal components: primary theory and Monte Carlo, *Journal of the American Statistical Society*, **80**, 759–766.

LOPUHAÄ, H. P. (1989) On the relation between S-estimators and M-estimators of multivariate location and covariance, *Annals of Statistics*, **17**, 1662–1683.

MA, Y., GENTON, M. G. (2001) Highly robust estimation of dispersion matrices, *Journal of Multivariate Analysis*, **78**, 11–36.

MARONNA, R. A., ZAMAR, R. H. (2002) Robust estimates of location and dispersion for high–dimensional datasets, *Technometrics*, **44**, 307–317.

OLLILA, E., OJA, H., KOIVUNEN, V. (2003) Estimates of regression coefficients based on sign covariance matrix, *Journal of the American Statistical Association*, **98**, 90–98.

RAO, C. R. (1988) Methodology based on the L_1-norm in statistical inference, *Sankayā, Series A*, **50**, 289–313.

ROUSSEEUW, P. J. (1983) Multivariate estimation with high breakdown point, In: Grossmann, W., Pflug, G., Vincze, I., Wertz, W. (eds.) *Proceedings of the 4th Pannonian Symposium on Mathematical Statistics and Probability, Vol. B*,D. Reidel Publishing Company, Dordrecht.

ROUSSEEUW, P. J. (1984) Least median of squares regression, *Journal of the American Statistical Association*, **79**, 871–880.

ROUSSEEUW, P. J. (1985) Multivariate estimation with high breakdown point, In: *Mathematical Statistics and Applications*, **8**, Grossmann, W., Pflug, G., Vincze, I., Wertz, W. (eds.), Reidel, Dordrecht, 283–297.

ROUSSEEUW, P. J., Croux, C. (1993) Alternatives to the median absolute deviation, *Journal of the American Statistical Association*, **88**, 1273–1283.

ROUSSEEUW, P. J., Hubert, M. (1999) Regression Depth, *Journal of the American Statistical Association*, **94**, 388–402.

ROUSSEEUW, P. J., VAN AELST, S., VAN DRIESSEN, K., AGULLÓ, J. (2004) Robust multivariate regression, *Technometrics*, **46**, 293–305.

ROUSSEEUW, P. J., VAN DRIESSEN, K. (1999) A fast algorithm for the minimum covariance determinant estimator, *Technometrics*, **41**, 212–223.

STAHEL, W. A. (1981) Breakdown of covariance estimators, Research Report 31, Fachgruppe für Stochastik, ETH Zürich.

SIEGEL, A. F. (1982) Robust regression using repeated medians, *Biometrika*, **68**, 242–244.

TUKEY, J. W. (1977) *Exploratory data analysis,* Addison-Wesley, Reading, Massachusetts.