Multivariate real time signal extraction by a robust adaptive regression filter

Matthias Borowski, Karen Schettlinger, Ursula Gather

Multivariate real time signal extraction by a robust adaptive regression filter

Journal:	*Communications in Statistics - Simulation and Computation*
Manuscript ID:	LSSP-2008-0123.R2
Manuscript Type:	Original Paper
Date Submitted by the Author:	26-Sep-2008
Complete List of Authors:	Borowski, Matthias; Technische Universität Dortmund, Fakultät Statistik Schettlinger, Karen; Technische Universität Dortmund, Fakultät Statistik Gather, Ursula; Technische Universität Dortmund, Fakultät Statistik
Keywords:	robust regression, multivariate time series, online monitoring, window width adaption, missing values
Abstract:	We propose a new regression-based filter for extracting signals online in moving windows from multivariate high frequency time series. This fast and robust filtering procedure considers the local covariance structure between the single time series components. It tackles the bias variance trade-off problem for the optimal choice of the window width by choosing the size of the window adaptively, depending on the current data situation. Furthermore, the signals are estimated at the recent point in time. Moreover, we present an advanced algorithm of our filter for replacing missing observations in real time. Thus it can be applied in online-monitoring practice.

Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.

MULTIVARIATE REAL TIME SIGNAL EXTRACTION BY A ROBUST ADAPTIVE REGRESSION FILTER

Matthias Borowski, Karen Schettlinger, Ursula Gather
Fakultät Statistik, Technische Universität Dortmund
D-44221 Dortmund
borowski@statistik.uni-dortmund.de
schettlinger@statistik.uni-dortmund.de
gather@statistik.uni-dortmund.de

Key Words: robust regression; multivariate time series; online monitoring; window width adaption; missing values.

ABSTRACT

We propose a new regression-based filter for extracting signals online from multivariate high frequency time series. It separates relevant signals of several variables from noise and (multivariate) outliers.

Unlike parallel univariate filters, the new procedure takes into account the local covariance structure between the single time series components. It is based on high-breakdown estimates, which makes it robust against (patches of) outliers in one or several of the components as well as against outliers with respect to the multivariate covariance structure. Moreover, the trade-off problem between bias and variance for the optimal choice of the window width is approached by choosing the size of the window adaptively, depending on the current data situation.

Furthermore, we present an advanced algorithm of our filtering procedure that includes the replacement of missing observations in real time. Thus the new procedure can be applied in online-monitoring practice. Applications to physiological time series from intensive care show the practical effect of the proposed filtering technique.

1 INTRODUCTION

In intensive care the patient's condition is supervised by an online-monitoring system that provides measurements of several physiological variables once per second. The measured data are noisy non-stationary multivariate time series with patterns such as trends and level changes as well as steady periods. Moreover, the variables are correlated, and the time series may contain technically induced outliers or measurement artifacts and missing values. Common alarm systems are based on lower and upper thresholds, set for each variable. Those systems trigger an alarm if an observation lies outside the specified thresholds. Because of frequent outliers, many false positive alarms occur. This high rate of false alarms induces a desensitization of the clinical staff to relevant alarms.

We develop a new multivariate, robust, and adaptive regression-based filter for separating clinically relevant signals from noise and outliers in real time; the rate of false positive alarms can be lowered by applying the alarm-thresholds to the online filtered signals instead of applying them to the raw measurements.

Consider a multivariate time series of dimension k, i. e., a sequence of observations $\boldsymbol{y}(t)$ of T random variables $\boldsymbol{Y}(t) = (Y_1(t), \ldots, Y_k(t))^\top \in \mathbb{R}^k$, $t = 1, \ldots, T$, for which we assume, that it can be decomposed into a true but unknown underlying signal overlaid by noise and outliers, i. e.,

$$\boldsymbol{Y}(t) \;=\; \boldsymbol{\mu}(t) + \boldsymbol{\varepsilon}(t) + \boldsymbol{\eta}(t), \quad t = 1, \ldots, T. \tag{1}$$

In this simple additive working model $\boldsymbol{\mu}(t) = (\mu_1(t), \ldots, \mu_k(t))^\top$ denotes the k-dimensional underlying signal at time t, whose components are assumed to vary smoothly most of the time but can also show sudden level shifts and trend changes. The observational noise arises from $\boldsymbol{\varepsilon}(t) \in \mathbb{R}^k$, where $\varepsilon_1(t), \ldots, \varepsilon_k(t)$ are errors generated by a symmetric distribution with zero median and time-dependent covariance matrix $\mathrm{Cov}(\boldsymbol{\varepsilon}(t)) = \boldsymbol{\Sigma}(t) \in \mathbb{R}^{k \times k}$. The errors of some components of $\boldsymbol{Y}(t)$ may be correlated, i. e., $\mathrm{Cov}\,[\varepsilon_i(t), \varepsilon_j(t)] = \sigma_{ij}(t)$, may be unequal to zero for $i \neq j$. $\boldsymbol{\eta}(t) \in \mathbb{R}^k$ denotes an outlier term, i. e., impulsive spiky noise that can appear for more than one component at different points in time.

Since time series from intensive care often show trends, Davies, Fried, and Gather (2004) find that regression-based filters, which approximate the signal by a locally linear function within a moving time window, lead to better results than location-based filters like the common moving average or the running median (Tukey, 1977). Due to the frequent occurrence of outliers, it is reasonable to use robust regression filters.

We consider two kinds of moving window filters: ones that estimate the signal with a certain time delay, called *delayed*, and filters that estimate the signal at the present point in time, called *online*.

In a simulation study Davies et al. compare delayed univariate time series filters based on robust regression techniques with respect to robustness, efficiency, and computing time; a filter based on *Repeated Median* (RM) regression (Siegel, 1982) provides best compromise results. Gather, Schettlinger, and Fried (2006) show that the online version of the RM filter (*online* RM, oRM) also outperforms other online filters. For a time window of width n containing the observations $\{y(t-n+1), \ldots, y(t)\}$ of a univariate time series, the oRM regression functional $T_{oRM} = \left(\hat{\beta}^{oRM}, \hat{\mu}^{oRM}\right)$ is defined by

$$\hat{\beta}^{oRM}(t) = \underset{s\in\{1,\ldots,n\}}{\text{med}} \left\{ \underset{v\neq s,\ v\in\{1,\ldots,n\}}{\text{med}} \frac{y(t-n+s)-y(t-n+v)}{s-v} \right\} \tag{2}$$

$$\text{and } \hat{\mu}^{oRM}(t) = \underset{s\in\{1,\ldots,n\}}{\text{med}} \left\{ y(t-n+s) - \hat{\beta}^{oRM}(t)\cdot(s-n) \right\} \tag{3}$$

so that the oRM regression line is given by

$$\hat{y}(t-n+s) = \hat{\mu}^{oRM}(t) + \hat{\beta}^{oRM}(t)\cdot(s-n), \quad s = 1, \ldots, n.$$

The oRM functional has a finite sample replacement breakdown point of $\lfloor n/2 \rfloor / n \approx 50\%$, which is the highest possible value for a regression equivariant functional (Rousseeuw and Leroy, 1987). The oRM filter also provides good efficiency even at non-contaminated samples (Gather, Schettlinger, and Fried, 2006) and needs little computing time (Bernholt and Fried, 2003).

The (o)RM filter is applied in a moving window of fixed width, leading to a bias variance trade-off: large windows lead to signal estimations with low variability, which is desired when

the observed time series is smooth and does not show abrupt level shifts or trend changes; on the other hand, a small window width yields signal estimations with small bias, and abrupt trend changes and level shifts are spotted with a small time delay. For real time application in intensive care, a fixed optimal choice of the window width does not exist because the data structure is not known beforehand.

Schettlinger, Fried, and Gather (2008) propose the univariate *adaptive online Repeated Median* (aoRM) filter, which tackles the trade-off problem by choosing the window width adaptively depending on the current data situation. However, correlations between the variables are not taken into account when this univariate filter is applied separately to each component of a multivariate time series.

Lanius and Gather (2007) propose the multivariate *Trimmed Repeated Median-Least Squares* (TRM-LS) regression as a fast, robust, efficient but non-adaptive two-step filtering procedure based on RM and multivariate Least Squares regression. This filter determines the local covariance structure in order to detect and remove (multivariate) outliers. However, the signal vector is estimated at the center of a moving time window $\{t-w,\ldots,t,\ldots,t+w\}$ of a fixed odd width $n=2w+1$ which leads to the bias variance trade-off described above. Furthermore, application of this filter causes signal estimations with a delay of w time units.

Our new filtering procedure combines the advantages of the univariate aoRM filter and the multivariate TRM-LS regression: it is multivariate, works online, and the window width is chosen adaptively.

After describing the univariate aoRM filter and an online version of the multivariate TRM-LS method, we introduce the newly combined *adaptive online Trimmed Repeated Median-Least Squares* (aoTRM-LS) filter and an algorithm that includes the treatment of missing values in Section 2. Section 3 demonstrates the use of the filtering procedure by means of applications to multivariate online-monitoring time series from intensive care. Furthermore, we propose two options that lead to improved filtering outputs for the examined data. Finally, Section 4 provides a summary and a brief discussion.

2 THE NEW FILTERING PROCEDURE

2.1 *The adaptive online Repeated Median (aoRM) filter*

The aoRM filter, proposed by Schettlinger, Fried, and Gather (2008), is based on an oRM estimate (2, 3) calculated in a moving time window for which a data-driven window width selection takes place at each point in time based on an idea by Gather and Fried (2004). The filtering procedure is able to trace sudden changes such as level shifts or trend changes with small time delay while also yielding a smooth representation when the time series shows only long-term trends and slow changes.

Corresponding to model (1) with $k = 1$, Schettlinger et al. assume that a univariate time series can be represented by a decomposition into an underlying signal, observational noise, and an outlier-generating process. As a general working model they assume that the univariate signal within a time window $\{t-n+1,\ldots,t\}$ of length n, $n \leq t \leq T$, can be approximated by a straight line:

$$Y(t-n+s) \;=\; \mu(t) + \beta(t)\cdot(s-n) + \varepsilon(t,s) + \eta(t,s), \quad s = 1,\ldots,n. \tag{4}$$

Here $\mu(t)$ is the signal level at the rightmost position of the window sample, i. e., at the recent point in time t, and $\beta(t)$ is the associated slope within the time window; $\varepsilon(t,s)$ denotes symmetric observational noise with zero median and time-dependent variance and $\eta(t,s)$ an outlier-generating process.

While for the simple oRM filter the window width n is fixed, for the aoRM procedure it can vary over time and hence is denoted by $n(t)$. The main steps to determine the aoRM signal estimate $\hat{\mu}^{aoRM}(t)$ at time t are the following:

1. Approximate the signal at time t by $\hat{\mu}^{oRM}(t)$, i. e., by an oRM signal estimate (3) calculated from the observations in the time window $\{t-n(t)+1,\ldots,t\}$.

2. Evaluate the signal estimate $\hat{\mu}^{oRM}(t)$:
   If $\hat{\mu}^{oRM}(t)$ is adequate, store the signal estimation, referred to as $\hat{\mu}^{aoRM}(t)$.
   If $\hat{\mu}^{oRM}(t)$ is not adequate, decrease the window width by one, i. e., set $n(t)$ to $n(t)-1$ and go to step 1.

The aoRM procedure requires the specification of a minimum and maximum window width such that $n(t) \in \{n_{min}, \ldots, n_{max}\}$. The lower bound n_{min} ensures robustness against a certain number of outliers in each time window, and the upper bound n_{max} limits the computing time. These values are user-specific and depend on the application. For example, for the application to high frequency online-monitoring time series as in Section 3, we set $n_{min} = 50$ and $n_{max} = 100$. For the first iteration we use the minimal window width, i. e., we get the first signal estimation at time $t = n(t) = n_{min}$. However, any other value in $\{n_{min}, \ldots, n_{max}\}$ is possible.

At step 2 the oRM signal estimate from step 1 is evaluated. Schettlinger et al. use a test procedure based on the fact that an oRM fit results in an equal number of positive and negative residuals. If this equality is not achieved for a small number n_{I_t} of the most recent residuals, the signal estimate could differ distinctly from the observed time series and, therefore, it cannot be considered as adequate. In this case the window width is reduced to $n(t) - 1$ by removing the oldest, i. e., leftmost observation. Then $\mu(t)$ is re-estimated in this smaller window by the oRM level (3), and the test is performed again. This is repeated until either the signs of the most recent residuals are 'balanced' or the window width equals the lower bound n_{min}. An explicit explanation of the test procedure at step 2 is given below.

In order to update the window for the next point in time $t + 1$, the new window width is set to $n(t + 1) = \min\{n(t) + 1, n_{max}\}$. That is, the observation at time $t + 1$ is incorporated into the window sample, and if $n(t) + 1$ exceeds the maximum window width, the oldest observation is removed from the window. A flow chart for the complete *aoRM algorithm* (A1) is shown in Figure 1.

The 'test of appropriateness' at step 2 is based on the fact that the (o)RM regression results in as many positive as negative residuals if the data come from a continuous distribution (Gather and Fried, 2004). Then the median of the (o)RM error distribution in the time window $\{t - n(t) + 1, \ldots, t\}$ is zero. Schettlinger et al. claim that, if the residual signs are also balanced for a small number n_{I_t} of the most recent observations, the signal estimate

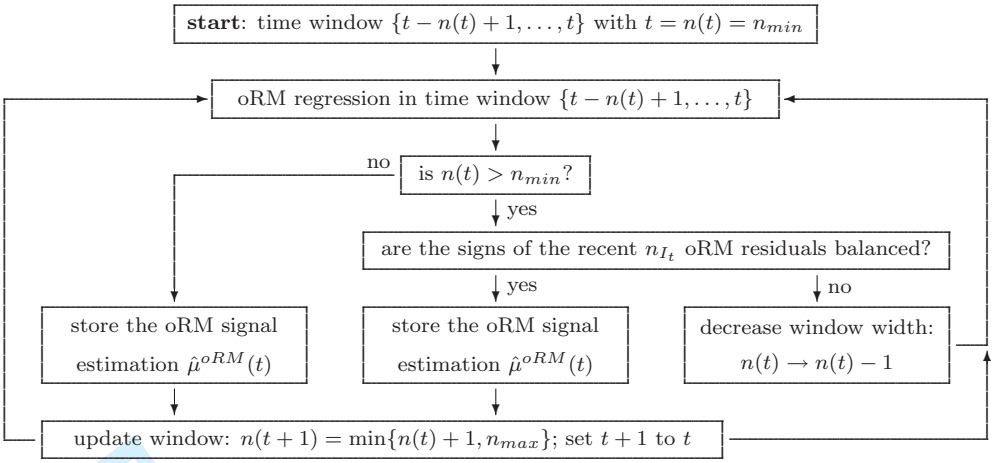

Figure 1: *The aoRM algorithm* (A1).

$\hat{\mu}^{oRM}(t)$ can be considered as adequate. To check this, they test whether the median $\text{med}_{\varepsilon}^{I_t}$ of the n_{I_t} most recent errors $\varepsilon(t,s)$, $s \in I_t$, $I_t = \{t-n_{I_t}+1,\ldots,t\}$, is equal to zero or not:

$$\text{H}_0: \ \text{med}_{\varepsilon}^{I_t} = 0 \quad \text{vs.} \quad \text{H}_1: \ \text{med}_{\varepsilon}^{I_t} \neq 0.$$

As test statistic the sum of the n_{I_t} most recent residual signs is used, i. e.,

$$T \ = \sum_{s \in I_t} \text{sgn}\left[r^{oRM}(t,s)\right],$$

where $r^{oRM}(t,s)$, $s = 1,\ldots,n(t)$, denote the residuals from an oRM fit (2, 3) in the time window $\{t-n(t)+1,\ldots,t\}$, and $\text{sgn}(\cdot)$ is the sign function with $\text{sgn}(0) = 0$. If $|T|$ is too large, either the negative or the positive residuals prevail within the subset I_t, and H_0 is rejected. As critical values for the test decision modified quantiles of the distribution of T, derived by means of simulations, are used.

Based on a simulation study by Schettlinger et al., we suggest to choose a fixed value for n_{I_t}. If $n_{I_t} > n_{min}/2$, we set $n_{I_t} = \min\{n_{I_t}, \lfloor n(t)/2 \rfloor\}$ in order to prevent that the subset I_t includes more than half of the residuals within the time window. However, n_{I_t} determines the number of shifted or trend-changed observations that is required to entail a reduction of the window width. Hence, the choice of n_{I_t} depends on the application. For high-frequency measurements from clinical online-monitoring good results are achieved using $n_{I_t} = 20$ or $n_{I_t} = 30$.

2.2 *The online Trimmed Repeated Median-Least Squares (oTRM-LS) filter*

The second source of our new procedure is based on the TRM-LS regression proposed by Lanius and Gather (2007). This is a multivariate time series filter based on the idea of univariate *Trimmed* RM filters (Bernholt, Fried, Gather, and Wegener, 2006). Lanius and Gather assume that each signal component of the multivariate time series is locally linear. They use the TRM-LS filter to fit k straight lines to the k-variate time series in a moving window $\{t-w,\ldots,t,\ldots,t+w\}$ of a predetermined fixed width $n=2w+1$. Then the k levels of the regression lines at time t form the k-dimensional signal estimation vector $\hat{\boldsymbol{\mu}}^{TRM-LS}(t)\in\mathbb{R}^k$. Thus the signal is estimated with a delay of w time units, and only odd window widths $n=2w+1$ can be chosen.

Based on the TRM-LS filtering procedure we develop the multivariate *online* TRM-LS filter (oTRM-LS), that uses the time window $\{t-n+1,\ldots,t\}$, also allowing for even window widths n. Corresponding to (4) in the univariate case we assume

$$\boldsymbol{Y}(t-n+s) \;=\; \boldsymbol{\mu}(t)+\boldsymbol{\beta}(t)\cdot(s-n)+\boldsymbol{\varepsilon}(t,s)+\boldsymbol{\eta}(t,s),\quad s=1,\ldots,n. \tag{5}$$

Here $\boldsymbol{\mu}(t)\in\mathbb{R}^k$ is the vector of the k signal levels at time t and $\boldsymbol{\beta}(t)\in\mathbb{R}^k$ the vector of the k associated slopes in the time window; $\boldsymbol{\eta}(t,s)\in\mathbb{R}^k$ denotes an outlier term, i. e., impulsive spiky noise and $\boldsymbol{\varepsilon}(t,s)\in\mathbb{R}^k$ symmetric observational noise with zero median and time-dependent covariance matrix. The noise of some components may be correlated, i. e., $\mathrm{Cov}\,[\varepsilon_i(t,s),\varepsilon_j(t,s)]=\sigma_{ij}(t,s)$ may be unequal to zero for $i\neq j$.

In order to determine the oTRM-LS signal vector $\hat{\boldsymbol{\mu}}^{oTRM-LS}(t)$ at time t, $t\geq n$, the oTRM-LS algorithm (A2) applies the following steps:

1. In the time window $\{t-n+1,\ldots,t\}$ determine the k univariate oRM estimates of the level $\hat{\mu}_i^{oRM}(t)$ from (3) and the slope $\hat{\beta}_i^{oRM}(t)$ from (2) for each of the variables $Y_i,\ i=1,\ldots,k$. Then the level and slope vector are given by

$$\begin{aligned}\hat{\boldsymbol{\mu}}^{oRM}(t) &= (\hat{\mu}_1^{oRM}(t),\ldots,\hat{\mu}_k^{oRM}(t))^\top\\ \text{and}\quad \hat{\boldsymbol{\beta}}^{oRM}(t) &= (\hat{\beta}_1^{oRM}(t),\ldots,\hat{\beta}_k^{oRM}(t))^\top.\end{aligned}$$

2. Determine the k-dimensional oRM residual vectors within the time window by

$$\boldsymbol{r}^{oRM}(t-n+s) \;=\; \boldsymbol{y}(t-n+s) - \left[\hat{\boldsymbol{\mu}}^{oRM}(t) + \hat{\boldsymbol{\beta}}^{oRM}(t)\cdot(s-n)\right], \quad s=1,\ldots,n.$$

3. Use a robust method to estimate the local error covariance matrix $\boldsymbol{\Sigma}(t) \in \mathbb{R}^{k\times k}$ based on the sample of residuals $\boldsymbol{r}^{oRM}(t-n+s) \in \mathbb{R}^k,\ s=1,\ldots,n$.

4. Determine the subset $S_t \subset \{1,\ldots,n\}$ of time points within the window, corresponding to those oRM residual vectors, which possess a squared Mahalanobis distance w. r. t. the local structure of covariance, that is not larger than a fixed value d_0, i. e.,

$$S_t \;:=\; \left\{ s=1,\ldots,n \;:\; \boldsymbol{r}^{oRM}(t-n+s)^\top\, \hat{\boldsymbol{\Sigma}}(t)^{-1}\, \boldsymbol{r}^{oRM}(t-n+s) \le d_0 \right\}.$$

5. Based on the trimmed window sample $\{\boldsymbol{y}(t-n+s) \,:\, s\in S_t\}$ perform a multivariate Least Squares regression to obtain estimates of the k signal levels at time t, referred to as $\hat{\boldsymbol{\mu}}^{oTRM-LS}(t) \in \mathbb{R}^k$.

At step 3 the local error covariance matrix $\boldsymbol{\Sigma}(t)$ is estimated based on the residual vectors in the time window. To avoid a masking effect caused by (multivariate) outliers, a robust estimator should be used. Lanius and Gather suggest the fast computable *orthogonalized Gnanadesikan-Kettenring* estimator (OGK) by Maronna and Zamar (2002). The maximum possible explosion breakdown point of the OGK is equal to that of a univariate estimator of scale the OGK depends on. In a comparison study, Lanius and Gather (2007) find the robust Q_n (Rousseeuw and Croux, 1993) to be a suitable univariate scale estimator. It possesses a maximal breakdown point of 50% if the data do not show ties within the time window.

The Q_n estimation of the local univariate scale $\sigma_i(t)$, $i=1,\ldots,k$, is determined based on the oRM residuals $\left\{r_i^{oRM}(t-n+1),\ldots,r_i^{oRM}(t)\right\}$. Due to collinear data, the estimate $\hat{\sigma}_i^{Q_n}(t)$ is possibly very small or equal to zero. Thus the OGK_{Q_n} estimation of the error covariance matrix $\boldsymbol{\Sigma}(t)$ may become singular. In order to obtain a non-singular $\hat{\boldsymbol{\Sigma}}(t)$, Lanius and Gather compute the OGK_{Q_n} estimation based on

$$\hat{\sigma}_i^{Q_n}(\cdot) = \max\{\hat{\sigma}_i^{Q_n}(\cdot), \vartheta\}, \tag{6}$$

where ϑ is an appropriate lower threshold for the variability, for example $\vartheta = 0.02$.

At step 4 of the oTRM-LS algorithm (A2) an upper bound d_0 must be chosen. For symmetric observational noise with zero median, each squared Mahalanobis distance

$$d(s) = \boldsymbol{r}^{oRM}(t-n+s)^\top \hat{\boldsymbol{\Sigma}}(t)^{-1} \boldsymbol{r}^{oRM}(t-n+s), \quad s = 1, \ldots, n, \tag{7}$$

approximately equals the sum of k squared observations from a standard normal distribution. Hence, a typical choice is $d_0 = \chi^2_{k,\alpha}$, where $\chi^2_{k,\alpha}$ is the α-quantile of a χ^2-distribution with k degrees of freedom. We follow a proposal by Maronna and Zamar (2002), where d_0 is adapted via the median of the distances $d(1), \ldots, d(n)$:

$$d_0 \;=\; \frac{\chi^2_{k;\alpha} \cdot \mathrm{med}_{s=1,\ldots,n}\{d(s)\}}{\chi^2_{k;0,5}}.$$

2.3 *The adaptive oTRM-LS (aoTRM-LS) filter*

The new aoTRM-LS filter evolves from a combination of the univariate aoRM and the multivariate oTRM-LS filter and adopts the same working assumption (5) as in the previous section. It requires the prior specification of the same input values as the aoRM filter, i. e., the number n_{I_t} of residuals within the window that are used for the test of adequacy of the signal estimation and the extreme values for the window widths n_{min} and n_{max}. The aoTRM-LS algorithm (A3) works as follows:

0. **Start**: set $t = n(t) = n_{min}$.

1. In the time window $\{t - n(t) + 1, \ldots, t\}$ obtain an adapted individual window width $n_i(t) \in \{n_{min}, \ldots, n(t)\}$, $i = 1, \ldots, k$, for each component by the univariate aoRM procedure.

2. Determine an *overall window width* $n_{ov}(t) := \min\{n_1(t), \ldots, n_k(t)\}$.

3. Apply the oTRM-LS algorithm (A2) to the multivariate sample in the time window $\{t - n_{ov}(t) + 1, \ldots, t\}$ and store the signal estimation vector at time t, referred to as $\hat{\boldsymbol{\mu}}^{aoTRM-LS}(t) \in \mathbb{R}^k$.

4. **Update process**:

   Incorporate the next observation vector $\boldsymbol{y}(t+1)$.

   Set $n(t+1) = \min\{n_{ov}(t)+1\,,\, n_{max}\}$.

   Set $t+1$ to t.

   Go to step 1.

At step 1 the univariate aoRM algorithm (A1) is applied to each of the k univariate time series in the window $\{t-n(t)+1,\ldots,t\}$ to determine the individual window width $n_i(t)$ for each variable $Y_1,\ldots,Y_k$.

Based on the $n_i(t)$, at step 2 an *overall window width* $n_{ov}(t)$ is evaluated, which is the size of the multivariate window sample at time t. In order to ensure that the working assumption (4) of a locally linear signal is true for each individual time series component such that the multivariate assumption (5) can also assumed to be valid, the overall window width $n_{ov}(t)$ is chosen as the minimum of the k individual window widths.

Step 3 consists of the application of the multivariate oTRM-LS algorithm (A2) to the k-variate time series in the time window $\{t-n_{ov}(t)+1,\ldots,t\}$ to obtain the signal estimation vector $\hat{\boldsymbol{\mu}}^{aoTRM-LS}(t)$. Hence, at time t the output of the aoTRM-LS filter is equal to that of the oTRM-LS filter if $n_{ov}(t)$ is equal to the fixed width n used for the oTRM-LS filter.

At step 4 the window sample is updated for the next point in time $t+1$. This is done by incorporating the observation vector $\boldsymbol{y}(t+1)$ into the window sample so that the window width is increased by one to $n(t+1) = n_{ov}(t)+1$. If the width of the updated time window is larger than the maximum bound n_{max}, the oldest, i. e., leftmost observation vector is removed from the window sample. Afterwards, we set $t+1$ to t and start the next iteration.

2.4 *The treatment of missing values*

In intensive care the monitored time series frequently show missing values either at single or at successive points in time and either concerning one, several, or all components. A missing observation of the ith component is denoted by $y_i(\cdot) = (\bullet)$, $i = 1,\ldots,k$.

The oTRM-LS algorithm (A2), which is performed at step 3 of the aoTRM-LS algorithm (A3), includes the estimation of a local error covariance matrix $\mathbf{\Sigma}(t) \in \mathbb{R}^{k\times k}$ based on the residuals of k univariate oRM regressions. This estimation cannot be done if there are missing values in the observation vectors and thus missing values in the residual vectors. Obviously, there is a loss of information if only residual vectors without missing values are used for estimating $\mathbf{\Sigma}(t)$.

Since it is possible to perform the oRM regression if at least two non-missing observations are present within a time window $\{t-n+1,\ldots,t\}$, we suggest to replace missing observations $y_i(t-n+s) = (\bullet)$ by the corresponding level of the oRM regression line

$$\hat{y}_i(t-n+s) = \hat{\mu}_i^{oRM}(t) + \hat{\beta}_i^{oRM}(t)\cdot(s-n) \tag{8}$$

with $i = 1,\ldots,k$ and $s = 1,\ldots,n$. Thus the corresponding residuals are zero, and $\mathbf{\Sigma}(t)$ can be determined by the OGK_{Q_n} estimator. Obviously, the OGK_{Q_n} estimation $\hat{\mathbf{\Sigma}}(t)$ is 'compressed' since zero residuals are inliers in this situation. However, an implosion of $\hat{\mathbf{\Sigma}}(t)$ is excluded by (6), and zero residuals inhibit a masking effect such that there is no loss of robustness associated with the replacement of missing values.

The substitution of missing observations by the respective oRM level guarantees the applicability of the oTRM-LS regression (and hence an aoTRM-LS output) if there are at least two non-missing window observations. However, it is possible that even less than two observations are available in the time window. Furthermore, we are interested in a reliable signal estimation. Therefore, we need a certain minimum number of non-missing values in a time window. Moreover, we have to keep in mind that the signal is estimated at the present point in time t, and the aoRM test is based on the n_{I_t} most recent observations. Hence, we apply the aoTRM-LS signal extraction at time t only for those variables Y_i which offer at least q observations at the recent n_q points in time $\{t-n_q+1,\ldots,t\}$ with $n_q \le n_{min}$. That is, at time t we consider only those $k(t)$ variables Y_i with $i \in \mathfrak{I}(t) \subset \{1,\ldots,k\}$, where

$$\mathfrak{I}(t) \;:=\; \Big\{ i = 1,\ldots,k \;:\; \#\{y_i(t-n_q+s) \neq (\bullet),\; s = 1,\ldots,n_q\} \ge q \Big\}$$

and $k(t) := \#\{\mathfrak{I}(t)\}$. At time t the signal is not estimated for those $k-k(t)$ variables Y_i

with $i \notin \mathfrak{I}(t)$, i. e., the associated entries in the signal estimation vector $\hat{\boldsymbol{\mu}}^{aoTRM-LS}(t)$ are missing values.

If there are more than $q^* = n_q - q$ observations of a variable missing at the recent n_q points in time, the signal is not estimated for that component at the respective time point t. For example: if we choose $n_q = 30$ and $q = 25$, a patch of $q^* + 1 = 6$ subsequent missing observations causes an interruption of the signal estimation for a duration of $q = 25$ points in time. Due to these considerations, for high-frequency online-monitoring data we suggest to choose $n_q \in [20\,,\,30] \cap \mathbb{N}$ and $q \in [0.5n_q\,,\,0.75n_q] \cap \mathbb{N}$.

In the following, we present an advanced aoTRM-LS algorithm (A3a, Figure 2) that contains the treatment of missing observations and therefore yields an output in any data situation. Note that the output is possibly a missing value if there is too little data information.

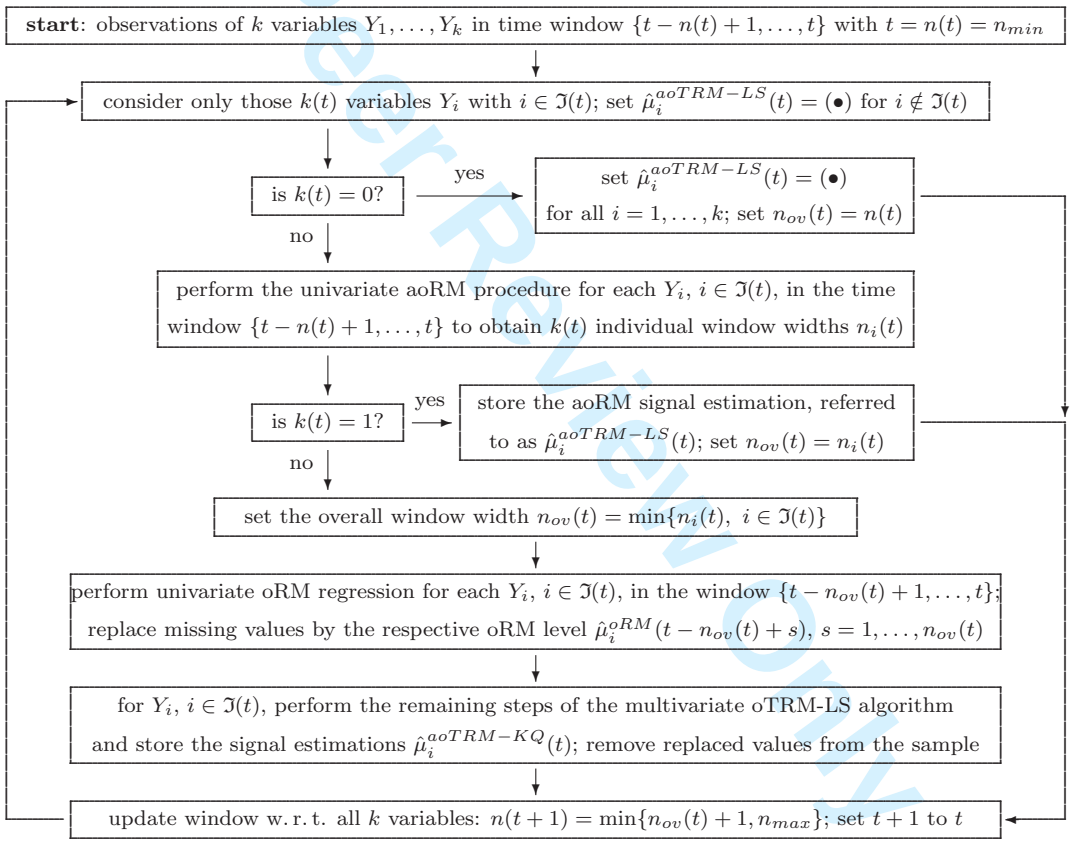

Figure 2: *The aoTRM-LS algorithm including the treatment of missing values (A3a).*

At each time t we consider only those $k(t)$ variables Y_i with $i \in \mathfrak{I}(t)$ and set $\hat{\mu}_i^{aoTRM-LS}(t) = (\bullet)$ for $i \notin \mathfrak{I}(t)$. If $k(t) = 0$, we set $n_{ov}(t) = n(t)$ and $\hat{\mu}_i^{aoTRM-LS}(t) = (\bullet)$ for all $i = 1, \ldots, k$, and the iteration at time t is finished. Otherwise, the univariate aoRM procedure is performed for each Y_i, $i \in \mathfrak{I}(t)$, in the time window $\{t - n(t) + 1, \ldots, t\}$ to obtain the $k(t)$ individual window widths $n_i(t)$. If $k(t) = 1$, the aoRM signal estimation $\hat{\mu}_i^{aoRM}(t)$, $i \in \mathfrak{I}(t)$, is stored, we set the overall window width $n_{ov}(t) = n_i(t)$, and the iteration at time t ends. If $k(t) \geq 2$, we set $n_{ov}(t) = \min\{n_i(t),\ i \in \mathfrak{I}(t)\}$ and perform the univariate oRM regression for each of the $k(t)$ variables in the time window $\{t - n_{ov}(t) + 1, \ldots, t\}$ based on the non-missing window observations. Then missing observations are replaced by the corresponding level of the oRM regression line (8), and the remaining steps of the oTRM-LS algorithm can be executed to obtain the $k(t)$ signal estimations $\hat{\mu}_i^{aoTRM-LS}(t)$, $i \in \mathfrak{I}(t)$. Afterwards the replacements of the missing observations are removed from the window sample. Finally, the update process in done as described in A3 regarding all k variables.

3 APPLICATION

In this section we apply the proposed aoTRM-LS filter to online-monitoring data measured at a frequency of once per second on an intensive care unit and suggest further options for improving the signal filtering.

We extract the signals retrospectively from an observed multivariate time series of a patient including the variables

- systolic, mean, and diastolic arterial blood pressure (ABP.S, ABP.M, ABP.D),
- heart rate (HR) and pulse (PLS),
- and systolic, mean, and diastolic pulmonary artery blood pressure (PBP.S, PBP.M, PBP.D).

The aim is to filter out noise and irrelevant outliers while clinically relevant level shifts and trends are preserved. In practice, the signal filtering has to be performed in real time.

Figure 3 shows 600 seconds of the observed time series (grey). In order to ease the visualization, the particular univariate time series are shifted up- or downwards, respectively, by a fixed amount. The plotted aoTRM-LS signal extraction (black) is based on the filter settings $n_{min} = 50$, $n_{max} = 100$, and $n_{I_t} = 20$. We choose $n_q = 20$ and $q = 15$ for the treatment of missing values.

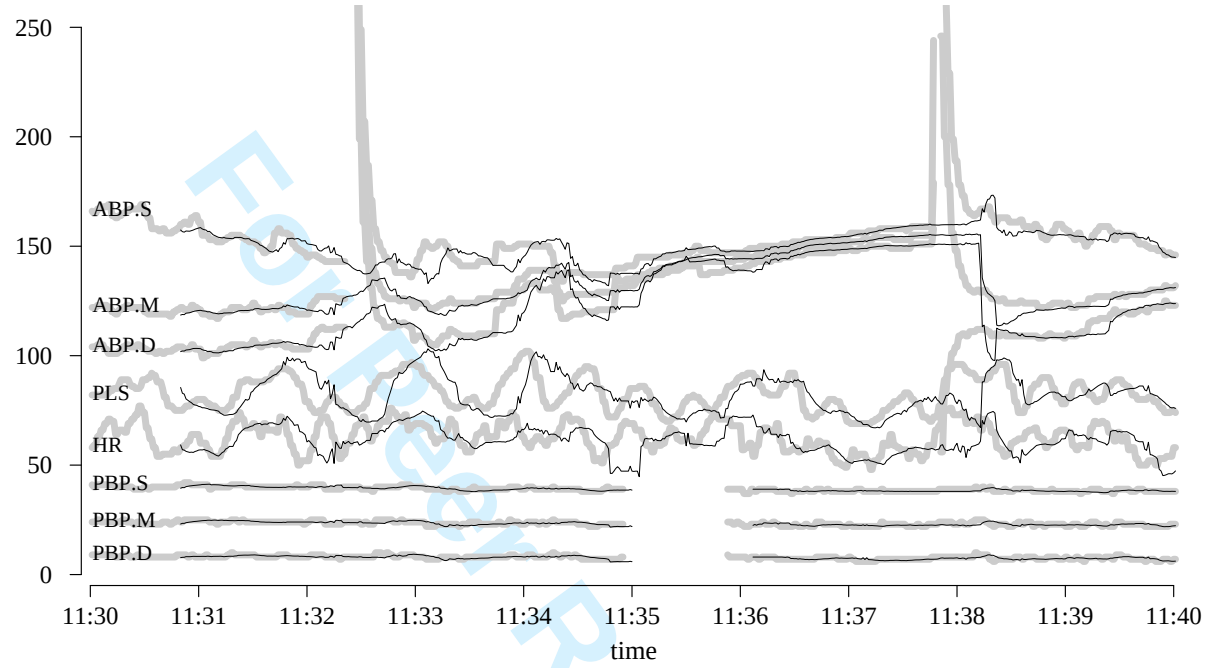

Figure 3: *Online-monitoring measurements (grey) of eight physiological variables and aoTRM-LS signal extractions (black).*

The unfiltered time series of the arterial blood pressures exhibit two conspicuous peaks around time 11:32:30 and 11:38, that both caused a threshold alarm. Since the peaks are assessed by a physician to be clinically irrelevant, we regard the given alarms as false positive. Furthermore, several observations are missing, for example concerning the pulmonary artery blood pressures after 11:35.

The aoTRM-LS filter yields the first signal estimation at time $t = n_{min} = 50$. The observed time series is smoothed, i. e., observational noise is suppressed, and the two peaks are neglected. For the first peak the filter detects 10 and for the second peak 14 subse-

quent outlying observations. These outliers are removed from the window sample before the multivariate Least Squares regression is performed. Thus two false alarms would have been avoided within these ten minutes if the filter had been applied in real time.

As can be seen for the pulmonary artery blood pressures after 11:35, the signal is not estimated when too many observations are missing. Since we choose $n_q = 20$ and $q = 15$, the signal estimation is interrupted once $n_q - q + 1 = 6$ of the 20 most recent window observations are missing, and it is resumed as soon as $q = 15$ new non-missing measurements are present in the time window.

Since the signal is estimated robustly at the right end of the time window, level shifts and trend changes are traced with a certain time delay. This delay depends on the chosen inputs n_{I_t} and n_{min}: the smaller n_{I_t}, the sooner level shifts and trend changes cause a decrease of the window width; the smaller n_{min}, the sooner level shifts and trend changes are traced.

Although the filter neglects noise and outliers as requested, it shows two drawbacks:

1. After sudden level shifts or trend changes the filtering outputs often deviate distinctly from the measurements and even exceed the range of the window observations.

2. When we apply the aoTRM-LS and the oTRM-LS filter to multivariate online-monitoring time series from intensive care, we observe that the signal estimations bear conspicuous similarity to each other if the minimum window width n_{min} of the aoTRM-LS and the fixed window width n of the oTRM-LS filter are chosen equally. That is, there is no obvious effect of the window width adaption for this type of data.

In order to overcome these drawbacks, we suggest two simple options for improving the signal estimation.

3.1 *Improvements*

After sudden level shifts or trend changes in the observed time series, the signal estimates often deviate distinctly from the measurements. Such deviations occur since the signal is estimated robustly: a trend change is not detected until there are enough observations

that follow the new trend; until then the filter carries forward the old trend so that the signal estimation is possibly greater (less) than the maximum (minimum) of the window observations. Sudden level shifts induce such distinct deviations in a similar way.

We suggest to restrict the signal estimations to the observational range of the window sample, i. e., we set

$$\hat{\mu}_i^{aoTRM-LS}(t) \;=\; \max\left\{y_i^{min}(t)\,,\,\min\left\{\hat{\mu}_i(t)\,,\,y_i^{max}(t)\right\}\right\} \tag{9}$$

where $y_i^{min}(t)$ is the minimum and $y_i^{max}(t)$ the maximum of the observations of variable Y_i, $i \in \mathfrak{I}(t)$, in the time window $\{t - n_{ov}(t) + 1, \ldots, t\}$. This *restrict-to-range-rule* may cause constant signal values over some time. However, since the filter output never exceeds the range of the window sample, we obtain less biased signal estimations.

The above-described similarity of the oTRM-LS and aoTRM-LS signal estimates is induced by the fact that the aoTRM-LS output at time t is equal to that of the oTRM-LS filter if $n_{ov}(t) = n$. Intensive care online-monitoring time series often exhibit structural changes so that at least one individual window width is close to or equal to n_{min} most of the time. Since the overall window width $n_{ov}(t)$ is determined by the minimum of the individual window widths $n_i(t)$, it is $n_{ov}(t) \approx n_{min}$ for a large part of time points t.

In order to determine $n_{ov}(t)$, we cannot replace the minimum by the mean or median, for instance, due to the assumption of local linearity in (5). Our proposed solution is based on the fact that the observed variables in clinical online-monitoring show a certain block dependence structure, see Figure 3. Here we find three blocks of highly correlated variables: the block of arterial blood pressures, the block of pulmonary artery blood pressures, and the block of heart rate and pulse. Hence, for 'instable' multivariate time series that exhibit a known block dependence structure, we suggest to apply the aoTRM-LS filter not to the whole multivariate time series but separately to each particular correlation block. Thus a small $n_i(t)$ effects only the variables in the corresponding block but not those that do not refer to that block. For intensive care time series the proposed *blockwise* aoTRM-LS filtering yields signal estimates which are less volatile than the 'simple' aoTRM-LS signal estimates.

In order to improve the filter output for the online-monitoring time series from Figure 3, we apply the aoTRM-LS filter separately to the three blocks composed of the arterial blood pressures, of the pulmonary artery blood pressures, and of heart rate and pulse, and restrict the estimates to the observational range by (9). The input settings are chosen according to the first application, i. e., $n_{min} = 50$, $n_{max} = 100$, $n_{I_t} = 20$, $n_q = 20$, and $q = 15$. Figure 4 shows the online-monitoring data (grey) and the extracted signals (black).

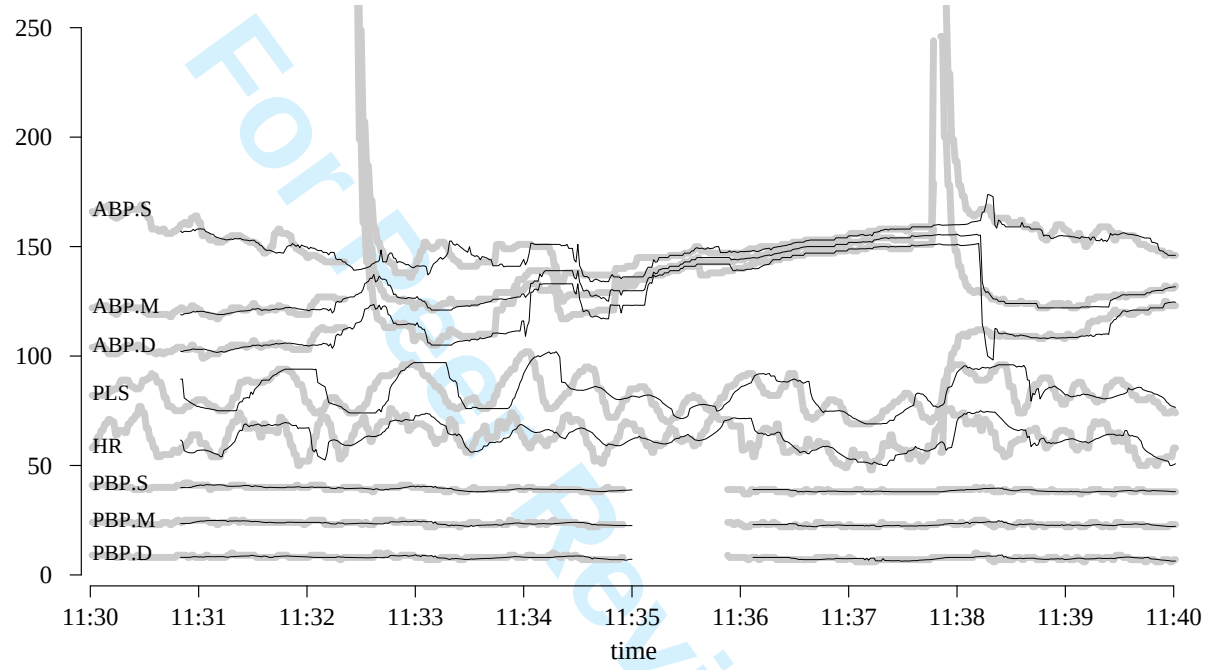

Figure 4: *Online-monitoring measurements (grey) of eight physiological variables and signal extractions by blockwise aoTRM-LS filtering with restrict-to-range-rule (black).*

The blockwise aoTRM-LS filtering results in smoother signal extractions compared to the signals filtered by the 'simple' aoTRM-LS method in Figure 3, as can be seen concerning the time series of heart rate and pulse. The effect of the restrict-to-range-rule (9) is obvious when the signal estimations are constant. At some of these positions the signal estimations from Figure 3 deviate distinctly from the measurements, e. g., the pulse signal extractions around time 11:32 and 11:33.

4 SUMMARY

The adaptive online Trimmed Repeated Median-Least Squares filter (aoTRM-LS) is developed specifically for the online extraction of relevant signals from noisy non-stationary multivariate time series. It combines the advantages of two existent filters: the univariate adaptive online Repeated Median filter (aoRM) and the delayed multivariate Trimmed Repeated Median-Least Squares regression (TRM-LS). The aoTRM-LS filter robustly separates relevant signals of k-variate time series from noise and outliers at the right end point of a moving time window whose width is adjusted according to the present data structure. Furthermore, the local covariance structure of the variables is considered for the online signal estimation and detection of multivariate outliers.

The problem of frequently occurring missing observations, which arise for example in physiological online-monitoring time series from intensive care, is overcome by a simple replacement strategy which works in real time. However, in order to guarantee that the signal estimation is based on a sufficiently large and current set of observations, the signal is extracted only if enough recent non-missing observations are present.

Applications to physiological time series from intensive care show that the aoTRM-LS filter can be improved further: firstly, by a simple bounding rule that restricts the signal estimations to the observational range and thus diminishes the bias; secondly, the variability of the signal extraction can be reduced by applying the filter to separated blocks of highly correlated variables. However, the variables must possess a well-known block dependence structure for this purpose.

Most intensive care online-monitoring units apply threshold alarm systems that trigger an alarm when either the upper or lower threshold is crossed. Due to frequent outliers these alarm systems involve up to 90% false positive alarms (Chambrin et al., 1999). This non-satisfying low specificity can be expected to be considerably increased by the aoTRM-LS filter if the alarm thresholds are applied not to the measurements but to the online separated signals instead. However, since correct alarms must not be suppressed, a sensitivity of

100% is required. That is, the filter must be able to distinguish between clinically relevant structural changes in the data and (patches of) irrelevant outliers. If we define the difference between a structural change and a patch of outliers by the number of deviant observations, we can choose the inputs n_{I_t} and n_{min} accordingly and obtain 100% specificity and sensitivity.

The new aoTRM-LS filter is provided in the R-package `robfilter` (Fried and Schettlinger, 2008) by the function `madore.filter()`. This package also offers several univariate robust filtering methods, for example the oRM and the aoRM filter.

ACKNOWLEDGMENT

We thank the anonymous referee for the comments, and we gratefully acknowledge the support of the Deutsche Forschungsgemeinschaft (SFB 475, 'Reduction of complexity in multivariate data structures').

BIBLIOGRAPHY

Bernholt, T. and Fried, R. (2003). Computing the update of the repeated median regression line in linear time. *Information Processing Letters*, **88**(3), 111–117.

Bernholt, T., Fried, R., Gather, U., and Wegener, I. (2006). Modified repeated median filters. *Statistics and Computing*, **16**, 177–192.

Chambrin, M., Ravaux, P., Calvelo-Aros, D., Jaborska, A., Chopin, C., and Boniface, B. (1999). Multicentric study of monitoring alarms in the adult intensive care unit (ICU): a descriptive analysis. *Intensive Care Medicine*, **25**, 1360–1366.

Davies, P., Fried, R., and Gather, U. (2004). Robust signal extraction for on-line monitoring data. *Journal of Statistical Planning and Inference, Special Issue: Contemporary Data Analysis: Theory and Methods in Honor of John W. Tukey*, **122**, 65–78.

Fried, R. (2004). Robust filtering of time series with trends. *Journal of Nonparametric Statistics*, **16**(3), 313–328.

Fried, R. and Schettlinger, K. (2008). `robfilter`: *Robust Time Series Filters.* R package, version 2.1.

Gather, U. and Fried, R. (2004). Methods and algorithms for robust filtering. In: Antoch, J. ed., *Proceedings in Computational Statistics COMPSTAT 2008*, Heidelberg: Physica, 159–170.

Gather, U., Schettlinger, K., and Fried, R. (2006). Online signal extraction by robust linear regression. *Computational Statistics*, **21**(1), 33–51.

Lanius, V. and Gather, U. (2007). Robust online signal extraction from multivariate time series. *Technical Report 38/07, SFB 475: Reduction of Complexity in Multivariate Data Structures, Technische Universität Dortmund, Germany.*

Maronna, A. and Zamar, R. (2002). Robust estimates of location and dispersion for high-dimensional data sets. *Technometrics*, **44**(4), 307–317.

Rousseeuw, P. and Leroy, A. (1987). *Robust Regression and Outlier Detection.* New York: Wiley.

Rousseeuw, P. and Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, **88**(424), 1273–1283.

Schettlinger, K., Fried, R., and Gather, U. (2008). Real time signal processing by adaptive repeated median filters. *Submitted.*

Siegel, A. (1982). Robust regression using repeated medians. *Biometrika*, **69**, 242–244.

Tukey, J. (1977). *Exploratory Data Analysis.* Reading, Massachusetts: Addison Wesley.