

# Adaptive and Robust Signal Extraction

Roland Fried<sup>1</sup>, Ursula Gather<sup>1</sup>, and Karen Schettlinger<sup>1</sup>

Department of Statistics, TU Dortmund,  
44221 Dortmund, Germany  
(e-mail: {fried, gather, schettlinger}@statistik.uni-dortmund.de)

**Abstract.** We discuss filtering procedures for robust signal extraction from noisy time series. Special attention is given to relevant signal details like level shifts, turning points and volatility changes. Such local patterns provide the most important information in applications like intensive care, finance and ecology. Data-adaptive filters improve the preservation and detection of such changes. Real time application to high frequency data is made feasible by construction of fast algorithms.

**Keywords.** Time series, Robust filtering, Outliers, Level shifts, Trends.

## 1 Introduction

Signal processing is an important task in modern statistical data analysis. Usually, the basic interest is in the underlying level ( $\mu_t$ ) of an observed time series ( $y_t$ ), which is overlaid by noise ( $e_t$ ):

$$y_t = \mu_t + e_t, \quad t \in \mathbb{Z}.$$

The level is assumed to vary smoothly most of the time with only a few abrupt shifts. These changes are of special interest since they often provide important information and point at a change in the data generating mechanism. For instance, level shifts indicate relevant events in intensive care, onsets of trends in ecology, and turning points and volatility changes in finance.

We discuss window-based robust filtering procedures which adapt to local signal details and thus allow improved preservation of relevant data patterns. Simplicity is often an overriding concern since operators need to understand the filtering process. Section 2 reviews Tukey's running median, which works well if the signal is locally almost constant within each time window. Section 3 discusses repeated median filters, which are based on a linear approximation of the signal within each window and thus achieve better performance in trend periods. Both filters can be improved by choosing the window width adaptively, or by weighting the data in the time window according to their distance to the target point at which the signal is estimated. For automatic detection of level shifts we can compare local medians calculated from delayed time windows, and adapt the filter locally if a shift is detected. Section 4 gives some conclusions and points at further work.

## 2 Running medians

Tukey (1977) recommended standard median filters, or running medians, for robust extraction of the underlying level ( $\mu_t$ ) from an observed time series ( $y_t$ ). The standard median filter estimates the level  $\mu_t$  in the center of a moving time window  $y_{t-m}, \dots, y_t, \dots, y_{t+m}$  of fixed width  $n = 2m + 1$ ,

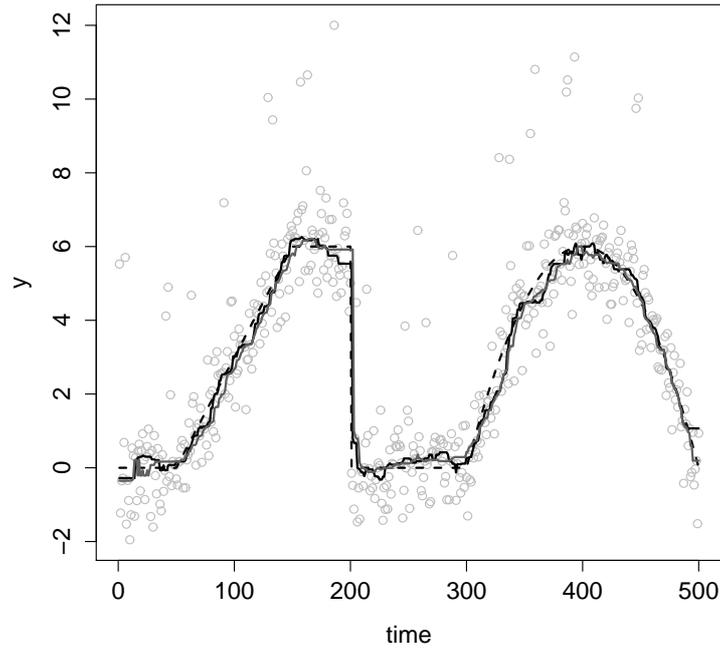
$$\hat{\mu}_t = \text{med}(y_{t-m}, \dots, y_t, \dots, y_{t+m}), \quad t \in \mathbb{Z}.$$

With respect to the choice of the window width  $n$ , there is the usual trade-off between bias and variance in nonparametric smoothing: the larger  $n$ , the smaller the variance but the larger the bias, if the level  $\mu_{t+j}$ ,  $j = -m, \dots, m$ , is not constant or at least linear within each time window. A further concern is the robustness of the filter since the running median can resist up to  $m$  outliers when the signal is constant, while it preserves a shift from one constant level to another one if at least  $m + 1$  subsequent observations deviate. In full online analysis, where we estimate the level  $\mu_{t+m}$  at the end of the window, the filter output follows the true signal with a delay of  $m$  time points if the level is not locally constant within each window.

The advantages of long and of short time windows can be combined by adapting the window width to the local signal characteristics. Simple and robust rules for this can be formulated in terms of the signs of the local residuals with respect to (w.r.t.) the estimated level  $\hat{\mu}_t$ . If the assumption of a locally constant signal is appropriate within the current time window, the corresponding residual signs should be distributed randomly. More precisely, if the noise variables are independent and identically distributed, then the number of positive residuals  $y_{t+j} - \hat{\mu}_t$  in any subwindow of width  $l < n$  follows a hypergeometric distribution with parameters  $l$ ,  $n$  and  $m$  if the level is locally constant. This allows for conducting simple automatic tests: when using a symmetric window of width  $n_t = 2m_t + 1$  for estimation of the level  $\mu_t$ , we check the assumption of a locally constant level analyzing the residual signs at the  $2\lceil m_t/2 \rceil$  outmost positions  $j = -m_t, \dots, -m_t + \lceil m_t/2 \rceil - 1$  and  $j = m_t - \lceil m_t/2 \rceil + 1, \dots, m_t$ , and reject the hypothesis of a constant level if this number is larger than the  $\alpha/2$ -percentile of the mentioned hypergeometric distribution, for a small value of  $\alpha$  like  $\alpha = 5\%$ . One proceeds similarly if the number of negative residuals in the same subwindow is too large. If the hypothesis is rejected, the window width is reduced to  $n'_t = n_t - 2$ , until a lower limit  $n_m$  is reached. Otherwise, the estimate  $\hat{\mu}_t$  is stored and the window is increased by including  $y_{t+m+1}$  and  $y_{t+m+2}$  in it and moving the window center to the next time point  $t + 1$ . To reduce computation time, the window width can be restricted by an upper limit  $n_M$ .

Figure 1 depicts the outputs of two different running medians applied to a time series containing nonlinear and linear trends as well as a level shift and 5% additive outliers of fixed size 5 inserted at positions chosen at random. The running median with fixed width  $n = 21$  gives at least a rough

guess of the underlying signal. The running median with adaptive width varying within  $11 \leq n_t \leq 41$  preserves the level shift at  $t = 200$  somewhat better, and it results in a smoother signal output between times  $t = 200$  and  $t = 300$ , where the level is constant. However, during the trend periods neither of them provides very good results since the underlying assumption of a locally constant level is fulfilled only for very short windows, meaning that the variance can be reduced only marginally.



**Fig. 1.** Time series (dots), underlying level (dashed) and the estimates from a running median with fixed width  $n = 21$  (black) and an adaptive width (grey).

Another possibility is to test at each time  $t$  whether a level shift has occurred at this time or not. In Fried (2007) several rules for outlier-resistant shift detection in locally constant signals are investigated; comparisons of local medians, standardized by Rousseeuw and Croux' (1993)  $Q_n$ -estimator of scale are found to provide good efficiency and high robustness.

### 3 Repeated median filters

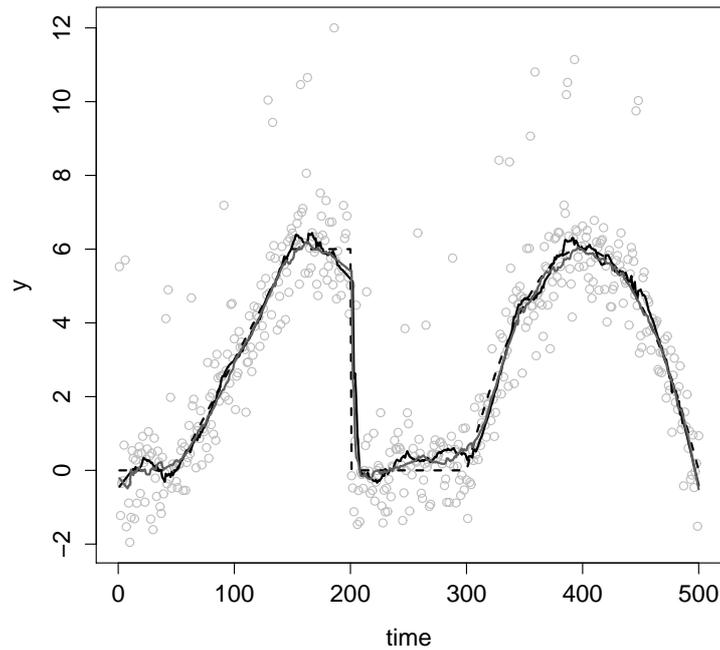
Davies et al. (2004) suggest to improve the shortcomings of running medians during trend periods by fitting a local linear trend to the data in a moving time window using robust regression. In a comparative study they find Siegel's

(1982) repeated median  $(\hat{\mu}_t^{RM}, \hat{\beta}_t^{RM})$ ,

$$\begin{aligned}\hat{\mu}_t^{RM} &= \text{med}(y_{t-m} + m\hat{\beta}_t^{RM}, \dots, y_{t+m} - m\hat{\beta}_t^{RM}) \\ \hat{\beta}_t^{RM} &= \text{med}_{i=-m, \dots, m} \text{med}_{j \neq i} \frac{y_{t+i} - y_{t+j}}{i - j},\end{aligned}$$

to give more stable results than some competitors.

Figure 2 shows the output of a repeated median filter with width  $n = 25$  applied to the above simulated time series. Fitting a straight line to the data in a moving time window obviously improves the results during trend periods. However, it also has problems whenever a straight line does not fit the data in each window well, see the trend change at time  $t = 380$  and the shift at  $t = 200$ . Again we can adapt the window width to the local structure of the data performing a local residual analysis as before. Note that the residuals w.r.t. the repeated median regression line are dependent, implying that the hypergeometric distribution provides only a rough approximation for the distribution of the number of positive or negative residuals in selected parts of the window. Nevertheless, the arising adaption rules typically lead to better level approximations than fixed windows, see Figure 2.



**Fig. 2.** Time series (dots), underlying level (dashed) and the outputs of repeated median filters with fixed width  $n = 25$  (black) and an adaptive width (grey).

The previous considerations yield a delayed analysis using symmetric time windows. Application of locally linear fits allows to overcome the time delays caused by locally constant fits. The good performance of the repeated median in the context of full online analysis is confirmed by Gather, Schettlinger and Fried (2006). Figure 3 depicts the output of an online version of the repeated median filter with a fixed width  $n = 19$ . The filter output is much more wiggly than in the delayed case. Moreover, the trend and the level changes are tracked with some delay, see times  $t = 150$  and  $t = 200$ , because only information on the past is used. Note, however, that the filter approximates the signal well without time delay whenever the assumption of a locally linear trend is valid.

Schettlinger et al. (2008) suggest to analyze subwindows consisting of a fixed number of the most recent residuals for adjusting the window width, because the resulting rules result in better adaption to local signal details than e.g. symmetric subwindows. The number of residuals should be chosen according to the permitted time delay of adaption and the number of outliers which the filter should resist.

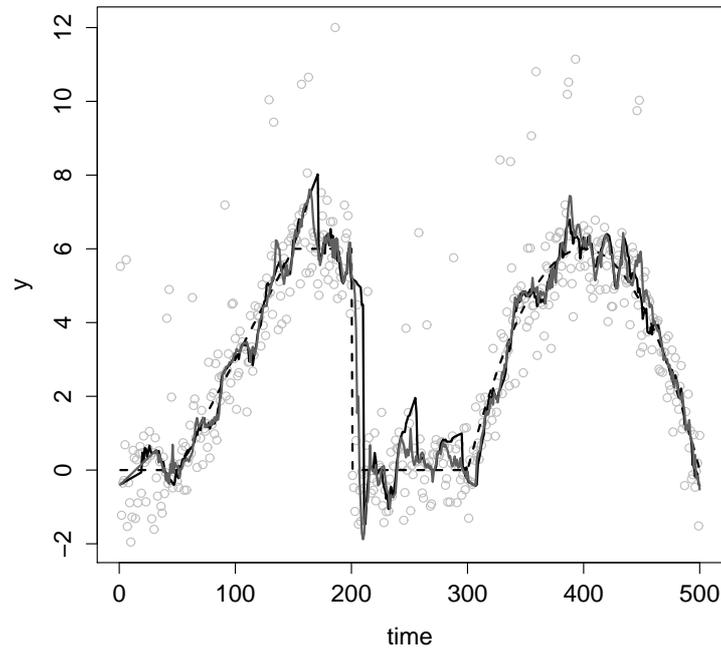
Another improvement in full online analysis is achieved by weighting the observations in the time window according to their temporal distances to the target time at which the signal is estimated. Fried, Einbeck and Gather (2007) apply this idea to repeated median filters, using triangular weights since these allow application of a modified version of the fast update algorithm constructed by Bernholt and Fried (2003). Figure 3 shows that the arising weighted repeated median filters, here with width  $n = 30$ , improve the efficiency and the stability of the filter output without increasing time delays in tracking the level or trend.

## 4 Conclusions

We have presented robust filtering procedures which allow to track the underlying trend of a time series. It is commonly agreed in nonparametric smoothing that locally linear fits are preferable to locally constant fits, among other reasons because they are less prone to boundary bias. Our results verify these findings in the context of robust signal extraction. In particular, running medians can be improved by repeated median filters during trend periods, and the latter even allow a full online analysis. Further improvements are possible by adapting the window width to local signal characteristics. More advances are outlined in Gather, Fried and Lanius (2006). Many of these are implemented in the R-package *robfilter* (R Development Core Team (2007)).

## References

- Bernholt, T. and Fried, R. (2003): Computing the Update of the Repeated Median Regression Line in Linear Time. *Information Processing Letters*, 88, 111–117.



**Fig. 3.** Time series (dots), underlying level (dashed) and the estimates from an online repeated median (black) and a weighted repeated median filter (grey).

- Davies, P.L., Fried, R. and Gather, U. (2004): Robust Signal Extraction for On-line Monitoring Data. *Journal of Statistical Planning and Inference*, 122, 65–78.
- Fried, R. (2007): On Robust Shift Detection in Time Series. *Computational Statistics and Data Analysis*, 52, 1063–1074.
- Fried, R., Einbeck, J. and Gather, U. (2007): Weighted Repeated Median Smoothing and Filtering. *Journal of the American Statistical Association*, 480, 1300–1308.
- Gather, U., Fried, R. and Lanius, V. (2006): Robust and Detail-Preserving Signal Extraction. In: B. Schelter, M. Winterhalder and J. Timmer (Eds.): *Handbook of Time Series Analysis*. Wiley, New York, 131–158.
- Gather, U., Schettlinger, K. and Fried, R. (2006): Online Signal Extraction by Robust Linear Regression. *Computational Statistics*, 21, 33–51.
- R Development Core Team (2007): *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.
- Rousseeuw, P. and Croux, Ch. (1993): Alternatives to the Median Absolute Deviation. *Journal of the American Statistical Association*, 88, 1273–1283.
- Schettlinger, K., Fried, R. and Gather, U. (2008): Real Time Signal Processing by Adaptive Repeated Median Filters. *Preprint, Technische Universität Dortmund*, 44221 Dortmund, Germany.
- Siegel, A.F. (1982): Robust Regression using Repeated Medians. *Biometrika*, 68, 242–244.
- Tukey, J.W. (1977): *Exploratory Data Analysis*. Addison-Wesley. Reading, Mass. (preliminary edition 1971).