Robust Preprocessing of Time Series with Trends

Roland Fried Ursula Gather
Department of Statistics, Universität Dortmund
{fried,gather}@statistik.uni-dortmund.de

Michael Imhoff
Klinikum Dortmund gGmbH
mike@imhoff.de

P.L. Davies
Mathematics Department, Universität Essen
laurie.davies@uni-essen.de

Abstract: Physiological time series measured in intensive care exhibit trends, level changes and periods of relative constancy. This signal is overlaid with a high level of noise and many measurement artifacts, and there are dependencies between the different items measured. We develop a method which allows a reliable denoising of the data and which can separate artifacts from relevant changes in the patients condition. For clinical online application the method has to be automatized and work in real time.
Key words: Medical data analysis, online monitoring, level shifts, outliers.

1 Introduction

Modern technical equipment allows online recording of many variables. In intensive care physiological variables like the heart rate and several blood pressures are measured at least every minute. The therapeutical interventions by the physician are mainly based on these data. In order to provide suitable bedside decision support automatic methods are needed which detect clinical relevant patterns like level shifts and monotonic trends and distinguish them from minor short-term fluctuations and measurement artifacts. The online detection of such patterns from these time series with (patchy) outliers is also a basic step for further data analysis [MIB+00]). Median filtering works well as long as there is no substantial trend in the data. Improvements may be possible by approximating the data by a local linear trend since in view of high sampling frequencies most changes occur gradually.

We develop a robust approach for decomposing a time series with structural changes into a time-varying mean and additive noise (see [DFG02]). This approach is designed to work online. We report results from simulations and provide applications to real physiologic time series as measured in intensive care.

2 Robust Approximation of a Local Linear Trend

Let $y_1, \ldots, y_N$ be real valued data measured at time points $t = 1, \ldots, N$. We assume that there is an underlying signal μ_t, $t = 1, \ldots, N$, that is overlaid by additive noise. In order to separate signal and noise we assume that the signal is simple, i.e. smooth with possibly a few sudden changes, while the noise $E_1, \ldots, E_N$ consists of independent random variables with mean zero. Hence, we consider a decomposition

$$Y_t = \mu_t + E_t$$

of the corresponding random variables Y_t, $t = 1, \ldots, N$, where the noise variance

$$Var(E_t) = \sigma_t^2$$

may slowly vary in time.

We approximate the signal within a time window of small to moderate length $n = 2m + 1$ by a local linear trend $\mu_{t+i} \approx \mu_t + \beta_t i, i = -m, \ldots, m$. W.r.t. the proper choice of the window width we have to search a compromise between a small variance (m large) and a small bias and the time delay possible in the respective application (m small). For our clinical application we move a time window of length $n = 2m + 1 = 31$ minutes through the time series to approximate $(\mu_t, \beta_t, \sigma_t)$. For simplicity we renumber the observations in the current time window by $y_{-m}, \ldots, y_m$ dropping the index t.

In data measured in intensive care there are large patches of measurement artifacts because of e.g. the drawing of blood samples. Therefore we use high breakdown point regression methods for robust approximation of the signal within each time window, namely the least median of squares functional [Ham75, Rou84]

$$T_{LMS} = \text{argmin}\{(a, b) : \text{Median}(y_i - a - bi)^2\},$$

or the repeated median functional T_{RM} [Sie82]

$$\begin{aligned} \tilde{\beta}_{RM} &= med_i \left(med_{j \neq i} \frac{y_i - y_j}{i - j} \right), \\ \tilde{\mu}_{RM} &= med_i (y_i - \tilde{\beta}_{RM} i). \end{aligned}$$

The breakdown point of these methods is the optimal $\lfloor n/2 \rfloor / n$ for a regression equivariant estimator. In case of a small to moderately large number of outliers, T_{RM} has smaller variance and mean square error MSE than T_{LMS} (see [DFG02]). On the other hand, T_{LMS} resists a large number of 30% or more outliers better than T_{RM} showing a much smaller bias and MSE then. T_{LMS} may even be less influenced by large outliers than by smaller ones as it often completely ignores the former. T_{RM} shows the intuitive behavior that larger outliers have a more serious effect. Therefore, replacing detected outliers may well improve the performance of T_{RM}, while this is not necessarily true for T_{LMS}. An advantage of T_{RM} is its smaller computation time. While a straightforward implementation of T_{RM} results in an $O(n^2)$ time algorithm, [Ber02] describes an algorithm for the update which needs $O(n)$ time only and $O(n^2)$ space.

3 Detection of Outliers and Level Shifts

Outliers and shifts can be detected by comparing the residuals $r_i = y_i - \tilde{\mu} - \tilde{\beta} i$ to an estimate $\tilde{\sigma}$ of σ. Such an estimate can be obtained from the past residuals $r_{-m}, \ldots, r_m$, while outliers can be detected online using the scaled distance $r_{m+1}/\tilde{\sigma}$. Gather and Fried [GF02] inspect some explicit robust scale estimators which can be calculated in $O(n \log n)$ time for this purpose. It turns out that the length of the shortest half [Grü88, RL88]

$$LSH = c_1 \cdot min\{|r_{(i+m)} - r_{(i)}|; i = 1, \ldots, n - m\}$$

and Rousseeuw and Croux's [RC93] suggestion

$$QN = c_2 \cdot \{|r_i - r_j| : -m \leq i < j \leq m\}_{(h)}, \; h = \binom{m+1}{2}.$$

are particularly interesting in our context. Here, $r_{(1)}, \ldots, r_{(n)}$ are the ordered residuals and c_1 and c_2 are small sample correction factors. The breakdown point of both methods is $\lfloor n/2 \rfloor / n$. The LSH shows extremely good resistance against a large percentage of outliers. On the other hand, the QN performs better for inliers, e.g. for identical measurements in consequence of small variability relatively to the measurement scale, and it is rather well-behaved in case of a level shift. Moreover, these methods, particularly QN, are less variable than other explicit high breakdown point methods.

Since we can specify lower bounds for clinically relevant changes in physiological variables we choose the LSH in the following. Some preliminary studies show that setting large detected outliers to their prediction $\tilde{\mu}_{RM} + \tilde{\beta}_{RM}(m+1)$ gives better results than other heuristical outlier replacement strategies. Therefore, we use a modified series with $3\tilde{\sigma}$-outliers replaced by these values when using the T_{RM}, and compare the results to those obtained using T_{LMS} without outlier replacement.

The detection of sudden shifts in the underlying signal is an important task in online monitoring. For online monitoring, it is often hard we to distinguish level shifts from large patches of outliers. Imhoff et al. [IBGL98] state that five subsequent observations which are of about the same size and differ substantially from the proceeding observations are often clinically relevant. However, such a rule of thumb is not robust itself as it may fail because of isolated outliers occurring briefly after a shift. Therefore, we base our rule for shift detection on all residuals in the right half of the time window. We consider a positive level shift to occur if more than half of these residuals are larger than a clinical relevant threshold, and we define a negative level shift accordingly. When detecting a level shift we restart the filtering procedure beginning at the earliest outlying residual and use the previous level and slope approximates to bridge the gap.

4 Application

Now we apply the previously described filtering procedures based on the T_{LMS} and on the T_{RM} with outlier replacement to some examples. We first discuss a simulated time series

of length 500 comparing the outcomes of the filtering procedure to the 'true' values. Here, a constant period is followed by a linear trend and another constant period, before a sudden shift occurs. This signal is overlaid by Gaussian white noise with unit variance, and 10% randomly chosen observations have been replaced by additive outliers of size 6σ. These outliers are organized in patches of 3 ($5\times$), 2 ($10\times$) and 1 ($15\times$) subsequent outliers.

Both methods resist the inserted outliers and track the underlying signal well. The 5σ-shift is detected as soon as possible at $t = 408$ and dated correctly at $t = 400$. The method based on the T_{RM} with outlier replacement shows generally less variability for both the level and the slope (not shown here).

The second example is a real physiologic time series representing heart rate. Again, the T_{LMS} is much more volatile than the T_{RM}, and as opposed to the T_{RM} without outlier replacement [DFG02] the positive outliers do not cause the T_{RM} to overestimate the signal now. The T_{LMS} exhibits a large spike at $t = 63$ due to a special pattern in the data. The slope approximates almost constantly signal a monotonic decrease up to $t = 140$. Again the T_{RM} with outlier replacement outperforms the T_{LMS}.

5 Conclusion

The online extraction of a signal which is corrupted by noise and artifacts is an important basic step in the analysis of data measured in intensive care. High breakdown point methods should be applied as there may be long outlier patches. The T_{LMS} is able to resist many outliers without showing a considerable bias but it is very variable and computationally expensive. Moreover, it can be seriously mislead by special patterns in the data. The T_{RM} can also withstand some outliers and it is computationally much less expensive. Its performance can be further improved using rules for outlier and level shift detection.

Acknowledgments:
The financial support of the Deutsche Forschungsgemeinschaft (SFB 475, "Reduction of complexity in multivariate data structures") is gratefully acknowledged.

Literaturverzeichnis

[Ber02] T. Bernholt. Computing the update of the repeated median regression line in linear time. Technical report, Department of Computer Science, University of Dortmund, Germany, 2002.

[DFG02] P.L. Davies, R. Fried, and U Gather. Robust signal extraction for on-line monitoring data. Technical report, SFB 475, University of Dortmund, Germany, 2 2002.

[GF02] U. Gather and R. Fried. Robust scale estimation in the presence of local linear temporal trends. Technical report, Department of Statistics, University of Dortmund, Germany, 2002. Preprint.

[Grü88] R. Grübel. The length of the shorth. *Annals of Statistics*, 16:619–628, 1988.

[Ham75] F.R. Hampel. Beyond location parameters: Robust concepts and methods. *Bulletin of the Int. Statist. Inst.*, 46:375–382, 1975.

[IBGL98] M. Imhoff, M. Bauer, U. Gather, and D. Löhlein. Statistical pattern detection in univariate time series of intensive care on-line monitoring data. *Intensive Care Medicine*, 24:1305–1314, 1998.

[MIB+00] Katharina Morik, Michael Imhoff, Peter Brockhausen, Thorsten Joachims, and Ursula Gather. Knowledge Discovery and Knowledge Validation in Intensive Care. *Artificial Intelligence in Medicine*, 19:225–249, 2000.

[RC93] P.J. Rousseuw and C.W. Croux. Alternatives to the median absolute deviation. *J. Americ. Statist. Assoc.*, 88:1273–1283, 1993.

[RL88] P.J. Rousseeuw and A.M Leroy. A robust scale estimator based on the shortest half. *Statistica Neerlandica*, 42:103–116, 1988.

[Rou84] P.J. Rousseeuw. Least median of squares regression. *J. Amer. Statist. Assoc.*, 79:871–880, 1984.

[Sie82] A.F Siegel. Robust regression using repeated medians. *Biometrika*, 68:242–244, 1982.

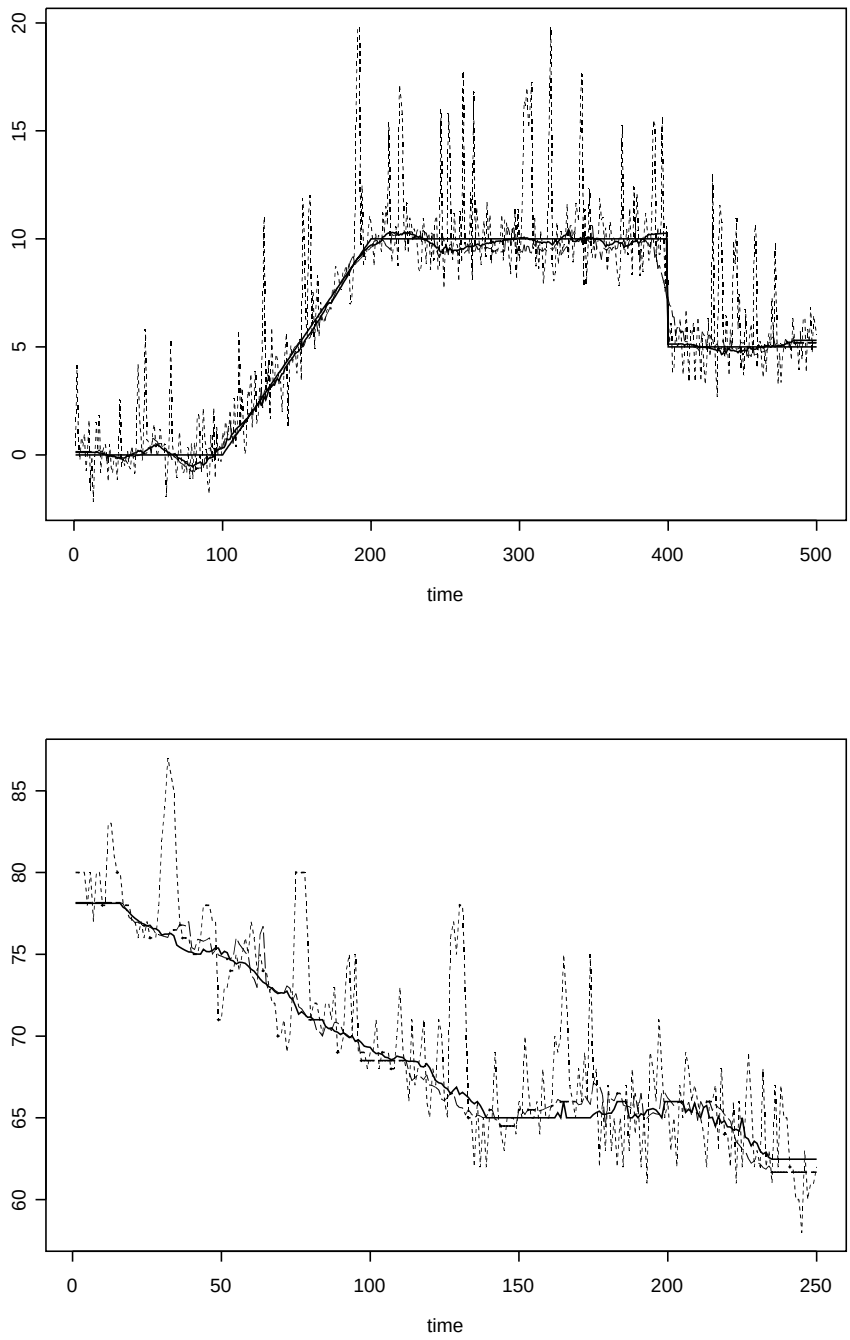

Figure 1: Top: Simulated time series (dotted), underlying level (bold solid) and level approximates: $\tilde{\mu}_{RM}$ (solid), $\tilde{\mu}_{LMS}$ (dashed). Bottom: Real time series representing heart rate (dotted) and level approximates: $\tilde{\mu}_{RM}$ (solid), $\tilde{\mu}_{LMS}$ (dashed).