Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics and Data Analysis 53 (2009) 2354–2362
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Online analysis of time series by the Q_n estimator

Robin Nunkesser[a,*], Roland Fried[b], Karen Schettlinger[b], Ursula Gather[b]

[a] *Fakultät für Informatik, TU Dortmund, 44221 Dortmund, Germany*
[b] *Fakultät Statistik, TU Dortmund, 44221 Dortmund, Germany*

Available online 4 March 2008

Abstract

A fast update algorithm for online calculation of the Q_n scale estimator is presented. This algorithm allows robust analysis of high-frequency time series in real time. It provides reliable estimates of a time-varying volatility even if many large outliers are present and it offers good efficiency in the case of clean Gaussian data.

1. Introduction

The increasing availability of high-frequency data in financial markets and many other fields requires fast, automatic, and reliable methods which can extract the relevant information from the data in real time. Measurement artifacts can influence the output of such analyses severely. High-frequency data are especially susceptible to errors: as stated by Brownlees and Gallo (2006), "the higher the velocity in trading, the higher the probability that some error will be committed in reporting trading information".

Many preprocessing procedures for automatic data cleaning and outlier detection have been suggested. As it is well known that non-robust estimators like empirical means and standard deviations can be strongly mislead by outliers, we should not rely on such methods in automated applications, not even for data cleaning. Robust estimators which are able to resist isolated outliers and patches of outlying values should be preferred. Robust methods even allow us to work with the raw data. Nevertheless, the use of robust methods in time series analysis is not widely established yet, especially in the online context. The computational demands of naive algorithms are typically much higher than those of non-robust methods, causing computation times which are unacceptably large, especially in applications to ultra-high-frequency data.

In this paper we discuss robust scale estimators in time series analysis, which allow us to extract possibly time-varying volatilities in the presence of outliers, see Gather and Fried (2003) and Gelper et al. (2007). These scale estimators can also be applied to estimate the autocorrelations within the process (Ma and Genton, 2000). Moreover, they can be used to standardize test statistics, e.g. for robust level shift detection (Fried, 2007; Nunkesser et al., in press).

* Corresponding author. Tel.: +49 231 755 5132; fax: +49 231 755 2047.
E-mail addresses: robin.nunkesser@tu-dortmund.de (R. Nunkesser), fried@statistik.tu-dortmund.de (R. Fried), schettlinger@statistik.tu-dortmund.de (K. Schettlinger), gather@statistik.tu-dortmund.de (U. Gather).

0167-9473/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2008.02.027

We focus on the robust Q_n estimator of scale (Rousseeuw and Croux, 1993) applied to time series. This estimator is defined as a multiple of an order statistic of all pairwise absolute differences between data points $x_1, \ldots, x_n \in \mathbb{R}$:

$$Q_n(x_1, \ldots, x_n) = c_n \cdot \{|x_i - x_j| : 1 \le i < j \le n\}_{(\ell)}, \qquad \ell = \binom{\lfloor n/2 \rfloor + 1}{2}, \tag{1}$$

where c_n denotes a finite-sample correction factor achieving unbiasedness at Gaussian samples. For data in general position the breakdown point of Q_n is about 50%, i.e. the estimate is bounded and stays away from zero even if almost 50% of the data are contaminated arbitrarily. Another, more widely-known scale estimator with this property is the *median absolute deviation about the median* (MAD):

$$\mathrm{MAD} = d_n \cdot \mathrm{med}\,\{|x_i - \mathrm{med}\{x_1, \ldots, x_n\}| : i = 1, \ldots, n\}, \tag{2}$$

where again d_n yields unbiasedness at Gaussian samples. For independent Gaussian data, Q_n is less variable than other high-breakdown point scale estimators. Its asymptotic efficiency of 82% (relative to the empirical standard deviation) is much larger than the asymptotic efficiency of the MAD, which is only 36%. A drawback of Q_n has been its computational complexity. Calculation of the MAD from n data points needs $O(n)$ computation time, and its value can be updated in $\mathcal{O}(\log n)$ time when applying it to moving time windows of n subsequent observations when analyzing locally stationary data (Bernholt et al., 2006). On the other hand, a straightforward implementation of Q_n would result in a computation time of $\mathcal{O}(n^2)$. Croux and Rousseeuw (1992) provide an offline algorithm for Q_n which needs $\mathcal{O}(n \log n)$ time.

Another class of robust estimators of scale which combine high breakdown point and good efficiency are τ estimators (Maronna and Zamar, 2002). They are defined by

$$\tau(x_1, \ldots, x_n) = \frac{\hat{\sigma}^2}{n} \sum_{i=1}^{n} \rho\left(\frac{x_i - \hat{\mu}}{\hat{\sigma}}\right), \tag{3}$$

where $\hat{\sigma}$ is a highly robust initial estimate of scale, $\hat{\mu}$ is a robust location estimate and ρ is a weight function. The τ estimator implemented in the R package `robustbase` uses the MAD as initial scale estimate, a weighted mean with weights based on Tukey's biweight applied to robustly scaled distances from the sample median, and $\rho_c(u) = \min\{c^2, u^2\}$ with the default value $c = 3$.

Section 2 describes a new update algorithm for the Q_n estimator. This algorithm is easy to implement and allows online application since it is substantially faster in practice than the offline algorithm. It allows us to incorporate incoming new observations and to remove old data quickly when using a moving time window. It can also be used for online computation of the Hodges–Lehmann location estimator and the *medcouple* estimator (Brys et al., 2004). Section 3 compares the performance of the Q_n estimator with the MAD, a trimmed standard deviation, and a τ estimator of scale for online extraction of time-varying volatilities. Computation times are analyzed for different window widths to show the practical relevance of the new update algorithm. Finally, Section 4 gives concluding remarks.

2. An update algorithm for the Q_n estimator

To analyze the scale of a time series $x_1, \ldots, x_N$ online, we apply the Q_n estimator (1) at each time t to a time window of length $n \le N$, which contains the observations $x_{t-n+1}, \ldots, x_t$. Instead of calculating Q_n for each window from scratch, we use an *update* algorithm. This means that for each move of the window from t to $t+1$ all stored information concerning the oldest observation x_{t-n+1} is deleted and new information concerning the incoming observation x_{t+1} is inserted. Note that this algorithm is not restricted to moving time windows; it can also handle arbitrary sequences of deletions and insertions of data points.

For offline computation of Q_n, Croux and Rousseeuw (1992) suggest the algorithm of Johnson and Mizoguchi (1978) with an optimal running time of $\mathcal{O}(n \log n)$ for n observations. Therefore, an optimal update algorithm for the Q_n estimator needs at least $\mathcal{O}(\log n)$ time for insertion or deletion.

In the following, we construct an online version of the algorithm of Johnson and Mizoguchi (1978). It computes an arbitrary, say kth order statistic in a multiset of form $\mathcal{X}+\mathcal{Y}$, where $\mathcal{X}+\mathcal{Y}$ is the multiset $\{x_i + y_i \mid x_i \in \mathcal{X} \text{ and } y_i \in \mathcal{Y}\}$ for $\mathcal{X} = (x_1, \ldots, x_n)$ and $\mathcal{Y} = (y_1, \ldots, y_n)$ n-tuples of real numbers. This algorithm can be used to compute Q_n, the Hodges–Lehmann location estimator (HL), and the MC_n estimator (see Brys et al. (2004) or Nunkesser et al. (in press)).