Bull. Aust. Math. Soc. **92** (2015), 173–175
doi:10.1017/S0004972715000283

QUANTILE BASED ESTIMATION OF SCALE AND DEPENDENCE

GARTH TARR

(Received 5 March 2015; first published online 29 April 2015)

2010 *Mathematics subject classification*: primary 62G35; secondary 62G05.

Keywords and phrases: quantile regression, robust scale estimation, long-range dependence, precision matrix estimation, cellwise contamination.

The sample quantile has a long history in statistics. The aim of this thesis is to explore some further applications of quantiles as simple, convenient and robust alternatives to classical procedures. The first application we consider is estimating confidence intervals for quantile regression coefficients, however, the core of this thesis is the development of a new, quantile based, robust scale estimator and its extension to autocovariance estimation in the time series setting and precision matrix estimation in the multivariate setting.

Chapter 1 addresses the need for reliable confidence intervals for quantile regression coefficients, particularly in small samples. The existing methods for constructing confidence intervals tend to be based on complex asymptotic arguments and little is known about their finite sample performance. We consider taking xy-pair bootstrap samples and calculating the corresponding quantile regression coefficient estimates for each sample. Instead of estimating a covariance matrix based on these bootstrap samples, our approach is to take the appropriate upper and lower quantiles of the bootstrap sample estimates as the bounds of the confidence interval. The resulting confidence interval estimate is not necessarily symmetric, only covers admissible parameter values and is shown to have good coverage properties. This work demonstrates the competitive performance of our quantile based approach in a broad range of model designs with a focus on small and moderate sample sizes. These results were published in [5].

A reliable estimate of the scale of the residuals from a regression model is often of interest, whether it be parametrically estimating confidence intervals, determining a goodness-of-fit measure, performing model selection, or identifying

Thesis submitted to The University of Sydney in January 2014; degree approved on 11 June 2014; supervisors Neville C. Weber and Samuel Müller.

unusual observations. The robustness of quantile regression parameter estimates to y-outliers does not extend to the error distribution. Extreme observations in the y space yield outlying residuals which can interfere with subsequent analyses. This led us to consider the more fundamental issue of robust estimation of scale.

Chapter 2 forms the core of this thesis with its investigation into robust estimation of scale. Common robust estimators of scale such as the interquartile range (IQR) and the median absolute deviation from the median are inefficient when the observations come from a Gaussian distribution. Rousseeuw and Croux [4] propose a more efficient robust scale estimator, Q_n, which is now widely used. We present an even more efficient robust scale estimator, P_n, based on a linear combination of U-quantiles. In its standard form the estimator P_n is proportional to the IQR of the pairwise means and can be thought of as the scale analogue of the Hodges–Lehmann estimator of location, the median of the pairwise means. When the underlying distribution is Gaussian, the Hodges–Lehmann estimator is considerably more efficient than the median, but it is not as robust. Similarly, P_n trades some robustness for significantly higher Gaussian efficiency than the IQR.

In the theoretical treatment, P_n is considered as a special case of a more general class of estimators based on the difference of two quantiles of the pairwise means. For this class of estimators, assuming the observations are independent and identically distributed, we show that the influence function is bounded and establish asymptotic normality. Further extensions to P_n incorporate adaptive trimming to achieve the maximal breakdown value of 50%. The resulting adaptively trimmed scale estimator has enhanced performance at extremely heavy-tailed distributions and is shown to be triefficient across Tukey's three corner distributions amongst the set of estimators considered. The adaptively trimmed P_n also yields good results in the multivariate setting discussed in Chapter 4.

The primary advantage of P_n over competing estimators is its high efficiency at the Gaussian distribution whilst maintaining desirable robustness and efficiency properties at moderately heavy-tailed and contaminated distributions. The desirable efficiency properties of P_n are shown to be even more marked over competing scale estimators in finite samples. The results of Chapter 2 have been published in [6].

Chapter 3 extends our robust scale estimator to the bivariate setting in a natural way as proposed by Gnanadesikan and Kettenring [1]. In doing so we move from estimating scale to estimating dependence. We show that the resulting covariance estimator inherits the robustness and efficiency properties of the underlying scale estimator.

Motivated by the potential to extend the efficiency and robustness properties of P_n to the time series setting, Chapter 3 also considers the problem of estimating scale and autocovariance in dependent processes. We establish the asymptotic normality of P_n under short- and mildly long-range dependent Gaussian processes. In the case of extreme long-range dependence, we prove a non-Gaussian limit result for the IQR, consistent with results found previously for the sample standard deviation and Q_n. In contrast with the results of Lévy-Leduc *et al.* [2] for a single U-quantile, namely Q_n, the proof for the IQR, a difference of two quantiles, relies on the higher-order

https://doi.org/10.1017/S0004972715000283 Published online by Cambridge University Press

terms in the Bahadur representation of Wu [9]. Simulation suggests that an equivalent result holds for P_n; we state the conjectured result which will require the analogous Bahadur representation for U-quantiles under long-range dependence. It is reasonably straightforward to extend the asymptotic results for the robust scale estimator to the corresponding robust autocovariance estimators. Various results from this chapter appear in [8].

Classical robust estimators assume that contamination occurs within a subset of the observations; however, in recent years there has been interest in developing robust estimators that perform well under scattered contamination. Chapter 4 looks at the problem of estimating covariance and precision matrices under cellwise contamination. This form of contamination is prevalent in large, automatically generated data sets, found in data mining and bioinformatics, where there is often little quality control over the inputs. A pairwise approach is shown to perform well under much higher levels of contamination than standard robust techniques would allow. Rather than using the orthogonalised Gnanadesikan and Kettenring procedure from [3], we consider a method that transforms a symmetric matrix of pairwise covariances to the 'nearest' covariance matrix (in a Frobenius norm sense). We combine this method with various regularisation routines purpose built for precision matrix estimation. This approach works well with high levels of scattered contamination and has the advantage of being able to impose sparsity on the resulting precision matrix. The results from this chapter have been published in [7].

References

[1] R. Gnanadesikan and J. R. Kettenring, 'Robust estimates, residuals, and outlier detection with multiresponse data', *Biometrics* **28**(1) (1972), 81–124.

[2] C. Lévy-Leduc, H. Boistard, E. Moulines, M. S. Taqqu and V. A. Reisen, 'Robust estimation of the scale and of the autocovariance function of Gaussian short- and long-range dependent processes', *J. Time Series Anal.* **32**(2) (2011), 135–156.

[3] R. A. Maronna and R. H. Zamar, 'Robust estimates of location and dispersion for high-dimensional datasets', *Technometrics* **44**(4) (2002), 307–317.

[4] P. J. Rousseeuw and C. Croux, 'Alternatives to the median absolute deviation', *J. Amer. Statist. Assoc.* **88**(424) (1993), 1273–1283.

[5] G. Tarr, 'Small sample performance of quantile regression confidence intervals', *J. Stat. Comput. Simul.* **82**(1) (2012), 81–94.

[6] G. Tarr, S. Müller and N. C. Weber, 'A robust scale estimator based on pairwise means', *J. Nonparametr. Stat.* **24**(1) (2012), 187–199.

[7] G. Tarr, S. Müller and N. C. Weber, 'Robust estimation of precision matrices under cellwise contamination', *Comput. Statist. Data Anal.* (2015), to appear; doi:10.1016/j.csda.2015.02.005.

[8] G. Tarr, N. C. Weber and S. Müller, 'The difference of symmetric quantiles under long range dependence', *Statist. Probab. Lett.* **98** (2015), 144–150.

[9] W. B. Wu, 'On the Bahadur representation of sample quantiles for dependent sequences', *Ann. Statist.* **33**(4) (2005), 1934–1963.

GARTH TARR, School of Mathematics and Statistics, The University of Sydney, NSW 2006, Australia
e-mail: garth.tarr@gmail.com

https://doi.org/10.1017/S0004972715000283 Published online by Cambridge University Press