QUANTILE BASED ESTIMATION OF SCALE AND DEPENDENCE

GARTH TARR

A thesis submitted in fulfilment of
the requirements for the degree of
Doctor of Philosophy

School of Mathematics and Statistics

Faculty of Science

University of Sydney

19 May 2014

Far better an approximate answer to the right question,
which is often vague, than the exact answer to the wrong question,
which can always be made precise.

— John Wilder Tukey (1962)

ABSTRACT

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 Tarr (2012).

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 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 (MAD) are inefficient when the observations come from a Gaussian distribution.

Rousseeuw and Croux (1993) 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, which is proportional to the IQR of the pairwise means. The estimator P_n is 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 however it is not as robust – similarly, P_n trades some robustness for significantly higher Gaussian efficiency.

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 this chapter have been published in Tarr, Müller and Weber (2012) and presented at ICORS 2011.

Chapter 3 extends our robust scale estimator to the bivariate setting in a natural way as proposed by Gnanadesikan and Kettenring (1972). 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. (2011c) for a single U-quantile, namely Q_n, the proof for the IQR, a difference of two quantiles, relies on the higher order terms in the Bahadur representation of Wu (2005). 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 have been presented at ASC 2012 and EMS 2013.

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. 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 (Maronna and Zamar, 2002), 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. Some preliminary results from this chapter have been presented at ICORS 2013.

PUBLICATIONS AND PRESENTATIONS

Some of the ideas and figures from the first two chapters of this thesis have appeared in the following publications:

Tarr, G., Müller, S., and Weber, N. C. (2012). A robust scale estimator based on pairwise means. *Journal of Nonparametric Statistics*, **24**(1), 187–199.

Tarr, G. (2012). Small sample performance of quantile regression confidence intervals. *Journal of Statistical Computation and Simulation*, **82**(1), 81–94.

A number of the results have been presented at the following conferences:

Robust scale and autocovariance estimation. *European Meeting of Statisticians*, 2013, Budapest, Hungary.

Robust covariance estimation via quantiles of pairwise means and applications. *International Conference on Robust Statistics*, 2013, St Petersburg, Russia.

Robust scale estimation with extensions. *Young Statisticians Conference*, 2013, Melbourne, Australia.

Robust covariance estimation with P_n. *Australian Statistical Conference*, 2012, Adelaide, Australia.

Efficient and robust scale estimation. *International Conference on Robust Statistics*, 2011, Valladolid, Spain.

Seminar presentations have been given at the following universities:

Robust estimation of scale and covariance with P_n and its application to principal components analysis. University of NSW, School of Mathematics and Statistics, 2013, Sydney, Australia.

Robust estimation of scale and covariance with P_n and its application to precision matrix estimation. University of Sydney, School of Mathematics and Statistics, 2013, Sydney, Australia.

ACKNOWLEDGMENTS

I have been fortunate to have the unfailing support of a number of very talented and generous people over my time at the University of Sydney and the last four years in particular.

Firstly, to Neville Weber and Samuel Müller, I could not have asked for more dedicated, patient and knowledgable supervisors. Their complementary skills have led to an interesting and varied thesis. Neville's breadth of statistical and probabilistic understanding and his expertise in asymptotic arguments gave me the impetus and courage to continue to push outside my comfort zone. Samuel's initial idea for the new scale estimator gave rise to the core of the thesis and his background in robustness provided inspiration for new directions.

Their fanatical/rigorous/enthusiastic/fervent/dogged/rabid attention to detail lead to significant improvements in clarity and structure. They helped make me a better statistician and a better communicator. Though after suffering through many years of grammatical abuse, I fear I may have worn down Neville's resolve against the split infinitive.

I have been extremely lucky to have had many travel opportunities to network and peddle my statistical wares. I am grateful for the financial support I received from the University of Sydney's Postgraduate Research Support Scheme, the School of Mathematics and Statistics' statistics research group, the Statistical Society of Australia through their Golden Jubilee Travel Grant and my supervisors.

I also want to thank my fellow PhD candidates, colleagues and students at the University of Sydney for making it an enjoyable few years. Ellis Patrick and Emi Tanaka who shared the journey with me from the beginning; many coffees with Justin Wishart, Patrick Noble, Luke Cameron-Clarke and Kellie Morrison; discussions at Hermanns with Michael Stewart, John Ormerod, Lisa Cameron and more recently Sanjaya Dissanayake, Tom Porter and Shila Ghazanfar.

My eternal gratitude goes to my partner Georgie Philpott and my parents for their support and for defending me when my grandmother invariably asks why I am still at university.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

ACF	autocorrelation function
CLIME	constrained L_1 minimisation for inverse covariance matrix estimation
CDF	cumulative distribution function
CLT	central limit theorem
GLASSO	graphical lasso
IQR	interquartile range
LRD	long range dependent
MAD	median absolute deviation from the median
MCD	minimum covariance determinant
MSE	mean square error
MLE	maximum likelihood estimate
NPD	nearest positive definite
OGK	Orthogonalised Gnanadesikan Kettenring
OGKw	reweighted Orthogonalised Gnanadesikan Kettenring
PD	positive definite
PRIAL	percentage relative improvement in average loss
PSD	positive semidefinite
QUIC	quadratic inverse covariance
SD	sample standard deviation
SRD	short range dependent

NOMENCLATURE

ϵ	a small number
$\mathrm{IF}(x; T, F)$	influence function
$\lfloor x \rfloor$	the integer part of x
$\boldsymbol{A}^{\intercal}$	matrix transpose of $\boldsymbol{A}$
$\mathbb{I}$	indicator function
$\boldsymbol{S}$	classical sample covariance matrix
$\boldsymbol{\Sigma}$	covariance matrix
$\boldsymbol{\Theta}$	precision matrix
Φ	standard Gaussian cumulative distribution function
ϕ	standard Gaussian probability density function
$\xrightarrow{\mathcal{D}}$	converges in distribution
$\xrightarrow{p}$	converges in probability
ε	proportion of contamination
ε^*	breakdown value
F	cumulative distribution function
f	probability density function
F_n	empirical distribution function
G	cumulative distribution function of the kernels of a U-statistic
G_n	empirical distribution function of the kernels of a U-statistic
H_k	kth Hermite polynomial
u_i	random error term

1

QUANTILE REGRESSION CONFIDENCE INTERVALS

1.1 INTRODUCTION

Quantile regression, first introduced by Koenker and Bassett (1978), provides an alternative to least squares with numerous advantages, not the least being the ability to estimate the full conditional quantile function. However, a major drawback to the use of quantile regression is the lack of agreement on a single unified method for conducting inference on the parameters. The purpose of this chapter is to highlight the xy-pair bootstrap as a widely applicable method that has comparable performance to other more complicated confidence interval construction techniques.

Numerous methods have been proposed, beginning with the direct estimation for standard errors of the parameters in the original Koenker and Bassett (1978) paper which was subsequently extended to the independent but not identically distributed case by Hendricks and Koenker (1992). In 1992, Gutenbrunner and Jurečková related the quantile regression parameters with the linear program used to calculate them, showing that the dual to the linear program allowed the application of rank inversion methodology to calculate confidence intervals directly. This was generalised by Koenker and Machado (1999) to allow for non-identically distributed errors.

Naturally, resampling methods have also been applied to the quantile regression inference problem. One approach would be the residual bootstrap, however Efron and Tibshirani (1993) showed it to be severely lacking when the model does not satisfy the independent and identically distributed assumption. Using the nonparametric xy-pair bootstrap to estimate an asymptotic covariance matrix for the estimated parameters is an option. Another possibility is to use the percentile bootstrap which was shown to have asymptotically correct empirical coverage probabilities by Hahn (1995). However, the percentile bootstrap method has received little attention since then.

Other, more abstract, resampling methods have been proposed. Parzen, Wei and Ying (1994) suggested exploiting the asymptotically pivotal subgradient condition. Another example is the Markov chain marginal bootstrap by He and Hu (2002), applied to quantile regression estimates in Kocherginsky, He and Mu (2005) and extended to the non-identically distributed case in Kocherginsky and He (2007). Further, a version of the generalised or weighted bootstrap with unit exponential weights has been explored by Chamberlain and Imbens (2003) and implemented in the R package `quantreg` (Koenker, 2013).

More recently, Feng, He and Hu (2011) provide a class of weight distributions for the wild bootstrap that is asymptotically valid for quantile regression. For nonparametric quantile estimates of regression functions, Song, Ritov and Härdle (2012) bootstrap the empirical distribution function of the residuals to obtain confidence bands.

This chapter presents the results from a simulation study based on Kocherginsky, He and Mu (2005) but with important extensions. We utilise simplified models so as to highlight the relative strengths and weaknesses of each of the various approaches. All the methods considered currently have routines available in the R package `quantreg` (Koenker, 2013) or, as in the case of the percentile bootstrap, are straightforward to code. Kocherginsky, He and Mu (2005) provide an overview of most of these models, however, they do not consider the percentile bootstrap and the exponentially weighted bootstrap was not available at the time.

In this chapter we include an additional performance index, the sample standard deviation (SD) of the estimated confidence interval lengths, to gain a better understanding of the variability in the estimated lengths. Furthermore we restrict attention to small sample sizes and utilise innovative graphics to provide an overview of how the techniques perform across a range of quantiles. This is particularly important, as one of the main attractions of quantile regression is investigating what the conditional quantile function is at the lower and upper quantiles, rather than just using the median and estimating a L_1 regression model, also known as least absolute deviations regression.

We will show that the percentile bootstrap performs at least as well as, and often better than, the other more complex resampling methods across a wide variety of model designs and quantiles. Other methods that generally perform quite well include the rank inversion techniques and the exponentially weighted bootstrap. Some of the more complex resampling techniques were not designed for use in models with small sample sizes. In this chapter of the thesis we confirm that their performance can be compromised if the sample size is too small. In particular, the primary use of the Markov Chain marginal bootstrap is in high dimensional models which are beyond the scope of this chapter.

This chapter is based around the material presented in Tarr (2012). Section 1.2 presents a brief overview of each of the methods, paying particular attention to the percentile bootstrap method. Section 1.3 presents some results from our simulation study. In Section 1.4 we present the evaluation of the simulation study and conclude with a positive comment on the performance of the percentile bootstrap method.

1.2 OVERVIEW OF CONFIDENCE INTERVAL CONSTRUCTION TECHNIQUES

This section gives an overview of the confidence interval construction techniques currently available, with more detail provided for the percentile bootstrap which has received limited exposure in the quantile regression literature. Kocherginsky, He and Mu (2005) and Koenker (2005) provide further detail about most of the methods considered in this chapter.

Consider the paired observations $(\boldsymbol{x}_i, Y_i)$ where $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})^\top$ is the $k \times 1$ covariate vector and Y_i is the response for $i = 1, \ldots, n$. The relationship between Y_i and $\boldsymbol{x}_i$ is modelled by a linear regression,

$$Y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + u_i.$$

The error terms, u_i, are assumed to be independent from an unknown error distribution, F. We aim to estimate the τth conditional quantile function,

$$F^{-1}_{Y_i|\boldsymbol{x}_i}(\tau) = \boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau.$$

Koenker and Bassett (1978) introduced the check function, shown in Figure 1.1,

$$\rho_\tau(u) = u\big(\tau - \mathbb{I}(u < 0)\big), \quad \tau \in (0, 1),$$

and showed that it can be used to find $\hat{\boldsymbol{\beta}}_\tau$ by solving,

$$\min_{\beta \in \mathbb{R}^k} \sum_{i=1}^{n} \rho_\tau(Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}). \tag{1.1}$$

The following subsections give an overview of the various methods used to construct confidence intervals for $\beta_{j,\tau}$, the jth component of the regression quantile vector $\boldsymbol{\beta}_\tau$.

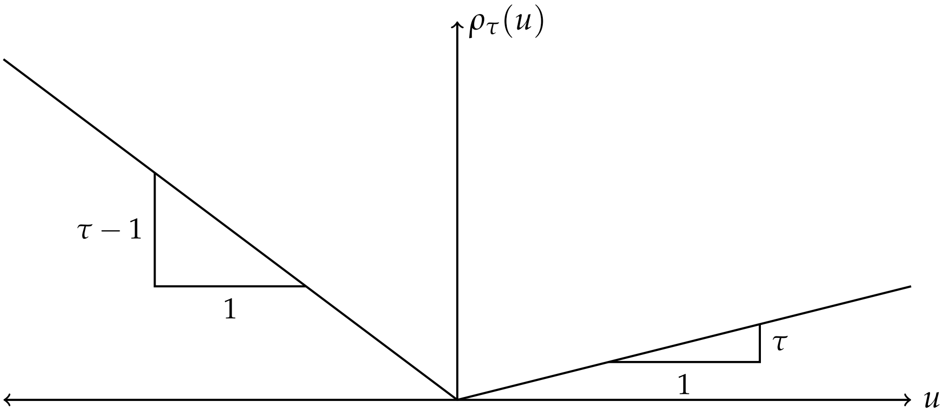


Figure 1.1: The check function, $\rho_\tau(u)$.

1.2.1 *Direct estimation*

Direct estimation of the parameter standard errors under independent and identically distributed errors (henceforth referred to as the iid method) was proposed in the original paper by Koenker and Bassett (1978). Under the iid assumption, the errors, u_i, are taken to be independent and identically distributed with cumulative distribution function (CDF) F, probability density function $f = F'$ and with $f(F^{-1}(x)) > 0$ for x in a neighbourhood of τ.

This method is inherently based on estimating the sparsity function,

$$s(\tau) = \frac{1}{f\big(F^{-1}(\tau)\big)}, \tag{1.2}$$

which gives a measure of the density of the observations near the quantile of interest. The sparsity function is estimated using a difference quotient which in turn utilises a bandwidth parameter to select the range over which the difference quotient is, in a sense, averaged.

Hendricks and Koenker (1992) consider the case where the errors are independent but no longer identically distributed (nid), that is $Y_i = \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}_\tau + u_i$ where $u_i \sim F_i$. This method also relies on an estimate of the sparsity as does the nid rank score method discussed in the following section.

1.2.2 *Rank inversion*

The rank score method avoids direct estimation of the asymptotic covariance matrix of the estimated coefficients and arises naturally from the linear programming techniques used to find estimates for quantile regression coefficients. Koenker (1994) details the iid approach to rank inversion tests, extending the work on rank based inference for linear regression models by Gutenbrunner, Jurečková et al. (1993), which builds on Gutenbrunner and Jurečková (1992). More recently, Portnoy (2012) provides "nearly root-n" results that can be used to strengthen the theoretical results for these rank based procedures.

Koenker and Machado (1999) relax the identically distributed error assumption and consider the location-scale shift model used by Gutenbrunner and Jurečková (1992),

$$Y_i = \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta} + \sigma_i u_i, \tag{1.3}$$

where $\sigma_i = \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\alpha}$ and the $\{u_i\}$ are assumed to be independent and identically distributed with distribution function F.

Under the rank inversion method, confidence interval estimates for a single parameter are found by the process of *inverting* the appropriate test statistic; moving from one simplex pivot to the next to obtain an interval in which the test statistic is such that the null hypothesis, $H_0 : \beta_{j,\tau} = b$, is not rejected. The resulting interval is not necessarily symmetric.

1.2.3 *Resampling methods*

As with rank inversion techniques, the resampling methods also avoid direct estimation of the covariance matrix. The xy-pair bootstrap begins with a bootstrap data set, $(\boldsymbol{x}_1^*, Y_1^*), (\boldsymbol{x}_2^*, Y_2^*), \ldots, (\boldsymbol{x}_n^*, Y_n^*)$, generated by random sampling with replacement from the observed sample and then calculating a bootstrap regression coefficient,

$$\hat{\boldsymbol{\beta}}_\tau^* = \underset{\boldsymbol{\beta}_\tau \in \mathbb{R}^k}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau(Y_i^* - \boldsymbol{x}_i^{*\mathsf{T}}\boldsymbol{\beta}_\tau).$$

Repeating this process B times yields the coefficient vectors $\hat{\boldsymbol{\beta}}_{\tau,1}^*, \ldots, \hat{\boldsymbol{\beta}}_{\tau,B}^*$, which is then used to construct an estimate of the variance of $\hat{\boldsymbol{\beta}}_\tau$,

$$\frac{1}{B}\sum_{b=1}^{B}\left(\hat{\boldsymbol{\beta}}_{\tau,b}^* - \hat{\boldsymbol{\beta}}_\tau\right)\left(\hat{\boldsymbol{\beta}}_{\tau,b}^* - \hat{\boldsymbol{\beta}}_\tau\right)^{\mathsf{T}}.$$

One alternative proposed by Parzen, Wei and Ying (1994) is to bootstrap the estimating equations. Another is to use the Markov chain marginal bootstrap (mcmb) approach of He and Hu (2002) which was extended to the quantile regression setting by Kocherginsky, He and Mu (2005). Additionally, the generalised bootstrap with weights, sampled independently from a standard exponential distribution, applied to the objective function, (1.1), is also considered (see Chamberlain and Imbens (2003); Chen et al. (2008) for details).

Efron and Tibshirani (1993) outline how the percentile interval bootstrap is constructed in the univariate case. In the quantile regression setting, the procedure begins in the same way as for the xy-pair bootstrap to obtain the B bootstrap estimated coefficient vectors, $\hat{\boldsymbol{\beta}}^*_{\tau,1}, \ldots, \hat{\boldsymbol{\beta}}^*_{\tau,B}$. However, instead of estimating a covariance matrix, let $\hat{G}_j$ be the empirical distribution function of $\hat{\beta}^*_{j,\tau}$, the jth element of $\hat{\boldsymbol{\beta}}^*_{\tau}$, $j = 1, \ldots, k$. The $1 - 2\alpha$ percentile interval for β_j is defined by the α and $1 - \alpha$ percentiles of $\hat{G}_j$,

$$\left[\hat{G}_j^{-1}(\alpha), \hat{G}_j^{-1}(1-\alpha)\right] = \left[\hat{\beta}^{*(\alpha)}_{j,\tau}, \hat{\beta}^{*(1-\alpha)}_{j,\tau}\right].$$

The attractions of this approach over many of the others considered above are its simplicity and the fact that the confidence interval covers only feasible parameter values. Furthermore, as we are resampling the $(\boldsymbol{x}_i, Y_i)$ pairs, no assumptions about variance homogeneity need to be made which allows some robustness to heteroskedasticity.

Importantly, the percentile method provides correct asymptotic coverage probabilities under quite general conditions. Hahn (1995) links a general M-estimator convergence result from Arcones and Giné (1992) to the quantile regression case to show that the asymptotic empirical coverage probability of the confidence interval constructed by the bootstrap percentile method is equal to the nominal coverage probability. Hahn (1995) goes on to point out that this result does not require that the error term is independent of the regressor: the bootstrap distribution is a valid approximation even when the conditional density of u_i given $\boldsymbol{x}_i$ depends on $\boldsymbol{x}_i$.

Hahn (1995) notes that this weak convergence result does not imply that the second moment converges to the asymptotic second moment. Interestingly, the bootstrap second moment of the simple sample median may diverge to ∞ even though the bootstrap distribution itself converges (Ghosh et al., 1984). This may explain why the percentile bootstrap outperforms the xy-pair bootstrap in some models.

The simulation study presented in the next section demonstrates that the percentile bootstrap gives quite reasonable empirical coverage probabilities for a broad range of model designs, even when the error distribution has a limited number of finite moments. It is especially interesting to note that these results hold for small sample sizes – indicating that the asymptotic approximations hold quite generally in practice.

1.3 SIMULATION STUDY

This section provides some guidance as to how well each of the previously mentioned methods of quantile regression confidence interval construction perform in practice. Our study differs from Kocherginsky, He and Mu (2005) in four ways: (i) here we consider sample sizes in the range $50 \leq n \leq 200$, whereas the smallest sample size in their study was $n = 200$, which may be more realistic for a broader range of problems; (ii) the models selected here are chosen to emphasise how particular model traits, such as heavy tailed errors or skewed covariates, affect confidence interval construction; (iii) the variability of the estimated confidence interval lengths that each technique produces is used to help identify differences in the techniques; and (iv) we add the approach taken by Parzen, Wei and Ying (1994), the exponentially weighted bootstrap and the percentile bootstrap but exclude the kernel smoothed density approach as Kocherginsky, He and Mu (2005) concluded that it was inadequate which we confirmed in preliminary simulations not reported here.

We firstly consider simple linear regressions with heavy tailed errors that commonly arise in economic and financial data as well as statistical physics, automatic signal detection and telecommunications (Adler, Feldman and Taqqu, 1998). The analysis is extended by considering models that incorporate highly skewed covariates as well as slightly more complex multivariate models where the error term is a function of one of the covariates and we also consider correlated covariates.

For each model a basic Monte Carlo experiment is performed, where a data set with n observations is generated defining the covariates, then the dependent variable is constructed before confidence interval estimates are found using each of the various techniques. We perform $N = 1000$ simulation runs and store the confidence interval estimates having nominal coverage level arbitrarily set at 0.9. For the resampling techniques, the number of resamples is set to $B = 1000$. The mean and SD of the estimated confidence interval lengths under each technique is calculated along with the empirical coverage probability which is defined to be the proportion of confidence intervals that contain the true parameter value.

For each model the process outlined above has been run for conditional quantiles, $\tau = 0.1, 0.2, \ldots, 0.9$, over sample sizes, $n = 50, 100, 150$ and 200. A complete set of results is available for the interested reader, however only a representative subset of these is presented below.

1.3.1 *Heavy tailed errors*

The first set of models exhibit heavy tailed errors. The models take the form,

$$Y_i = \beta_0 + \beta_1 x_i + u_i, \quad i = 1, \ldots, n, \tag{1.4}$$

where $\beta_0 = \beta_1 = 10$, x_i are either fixed or random covariates and $u_i \sim t_\nu$, $\nu \in \{1, 2, 3, 4, 5, \infty\}$, i.e. including the limiting case of Gaussian errors. Figure 1.2 gives a plot of the empirical coverage probabilities for a model with $n = 50$ fixed covariates drawn from $\mathcal{U}(0, 5)$, a uniform distribution with support $(0, 5)$, and the u_i are t_1 distributed, i.e. Cauchy errors.

The percentile bootstrap (pbs) provides remarkably more consistent estimated coverage probabilities than the other methods. Indeed the Markov chain marginal bootstrap (mcmb), Parzen Wei and Ying bootstrap (pwy), direct estimation assuming iid errors (iid), rank inversion assuming iid errors (riid) and rank inversion not assuming identically distributed errors (rnid) methods exhibit 'V' shaped empirical coverage probabilities of varying degrees over the range of τ considered. The direct estimation not assuming identically distributed errors (nid), unit exponential weighted bootstrap (wxy) and paired bootstrap (xy) methods also provide consistent estimated coverages, however they are not as close to the nominal coverage as the simple percentile bootstrap. These observations are replicated in models with random covariates.

As the tail of the error distribution becomes slightly less heavy $\nu \in \{2, 3, 4, 5\}$, a number of models become acceptable. The percentile bootstrap still performs admirably in terms of estimated coverage probabilities, as does the wxy and xy. The rank inversion methods underestimate the coverage probabilities somewhat for the intercept and less markedly for the slope coefficient. The mcmb, pwy and iid methods still exhibit strong 'V' shaped empirical coverage probabilities. Tables A.1 through A.4 in Appendix A give details for $\tau = 0.3$. Overall the trend is for the lengths and their SDs to shrink as the error distribution becomes less heavy tailed. The coverage probabilities do not seem to keep improving, once finite mean and variance becomes a feature of the error distribution, there is little improvement in coverage by adding on additional finite moments.

When the error distribution is extremely heavy tailed, for example Cauchy distributed, a curious phenomenon occurs as the sample size increases. Coverage probability results for this scenario when $n = 50$ are found in Figure 1.2 and those for $n = 100$, 150 and 200 are included in Figures A.1, A.2 and A.3 in Appendix A. When $n = 50$ most methods tend to overestimate

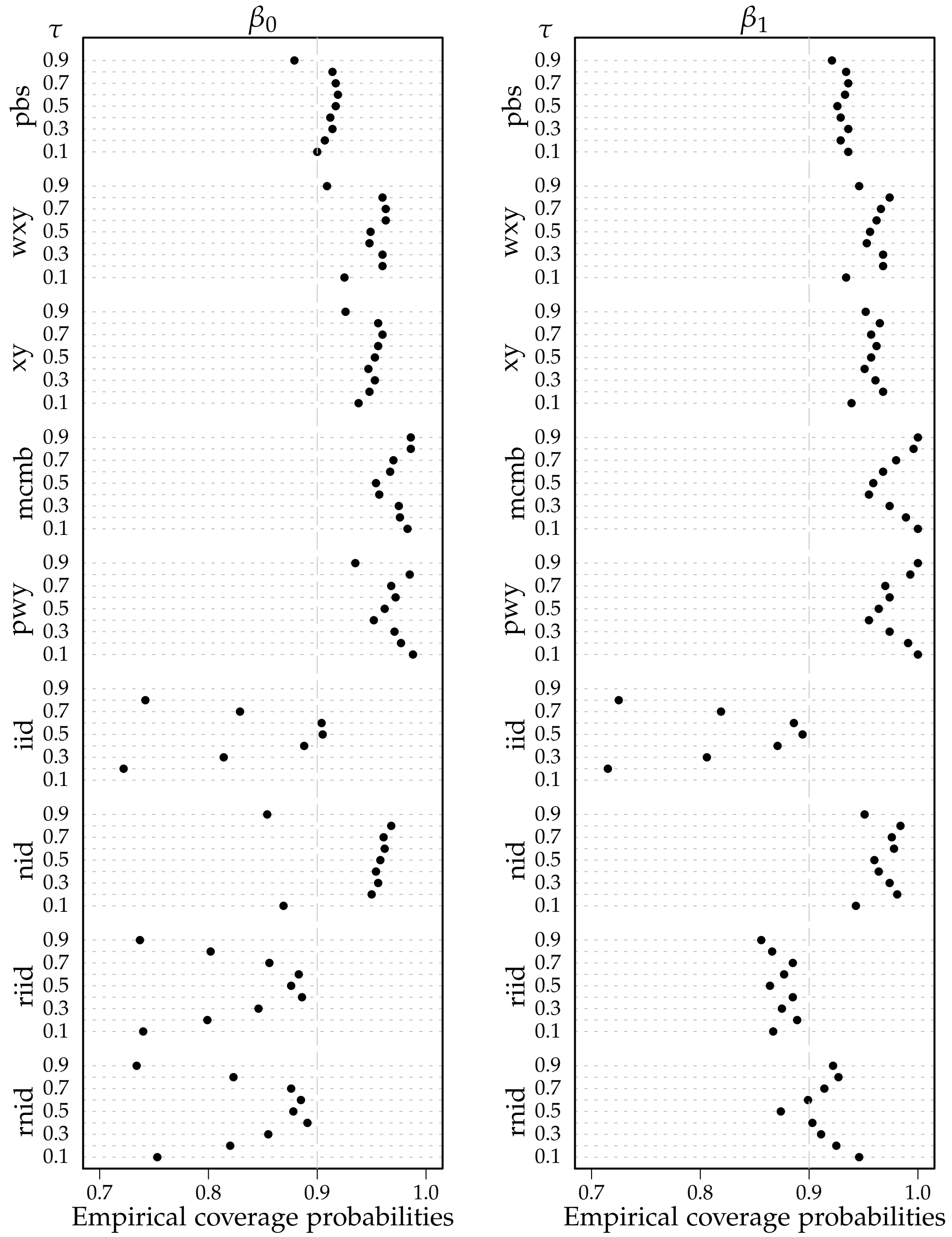


Figure 1.2: Empirical coverage probabilities for the model $Y_i = \beta_0 + \beta_1 x_i + u_i$, where $x_i \sim \mathcal{U}(0,5)$, $u_i \sim t_1$ and $\beta_0 = \beta_1 = 10$. The sample size is $n = 50$, the number of resamples in each of the bootstrap methods is $B = 1000$, the number of Monte Carlo simulations is $N = 1000$. The nominal coverage is 0.9.

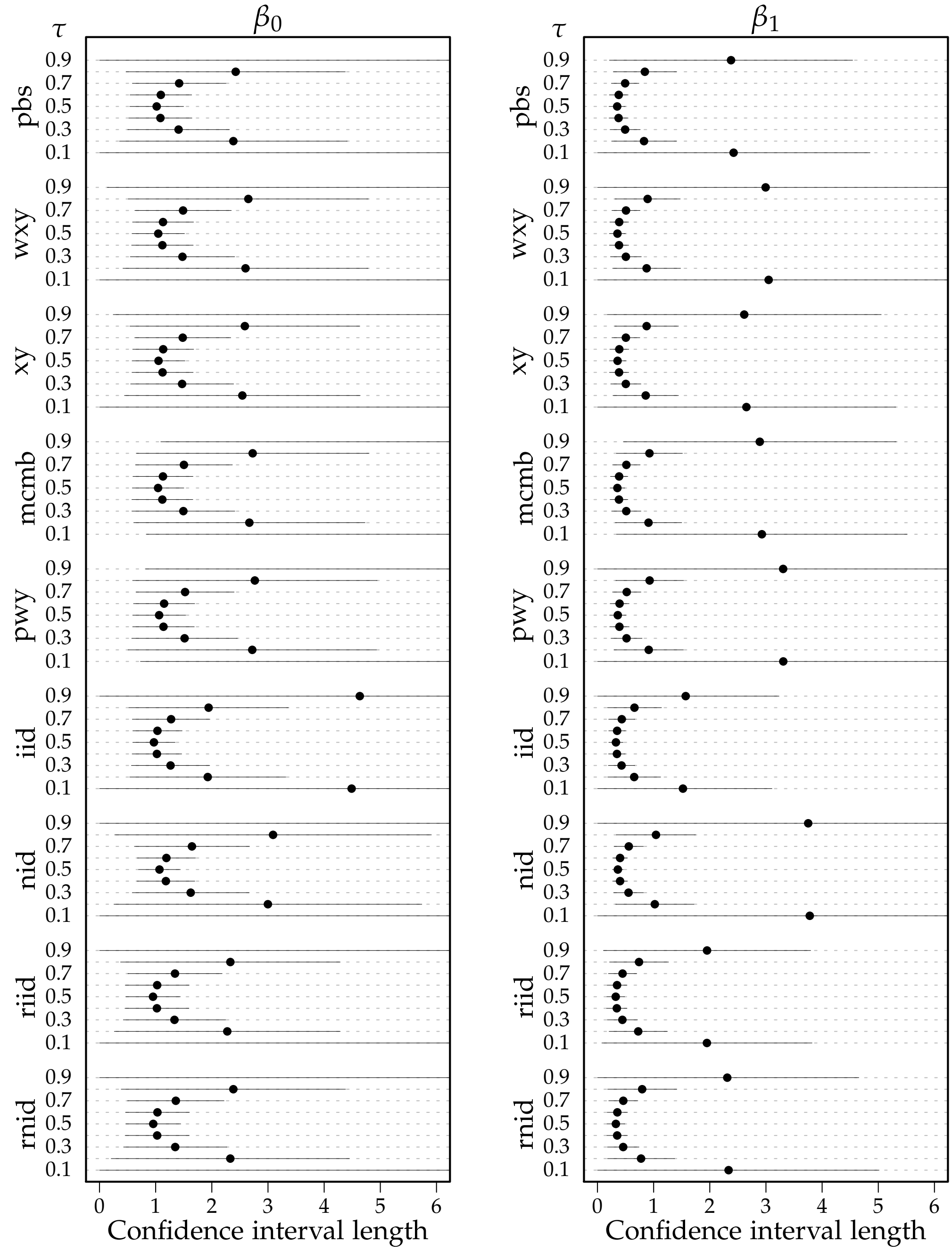


Figure 1.3: Average confidence interval lengths for the model $Y_i = \beta_0 + \beta_1 x_i + u_i$, where $x_i \sim \mathcal{U}(0,5)$, $\beta_0 = \beta_1 = 10$ and $u_i \sim t_1$. The sample size is $n = 150$, the number of resamples in each of the bootstrap methods is $B = 1000$, the number of Monte Carlo simulations is $N = 1000$. The circles represent the mean length, the horizontal lines are a guide to the variability of the mean length estimate and represent a naive 95% confidence interval constructed as the mean length ± two times the SD of the lengths.

the true coverage probabilities for the intercept and slope parameters. As the sample size increases, most methods perform well for central τ values but, for the intercept, their coverage probabilities fall below the nominal level at moderate and extreme τ while, for the slope, they remain relatively unaffected. Across all sample sizes considered, the pbs and rank inversion methods give good coverage probabilities for the slope parameter.

We conjecture that this behaviour is related to the inherent difficulty associated with estimating quantiles of heavy tailed distributions. Indeed the intercept in a quantile regression model actually estimates $\beta_0 + F_u^{-1}(\tau)$. The average lengths are typically twice as long for the intercept than the slope coefficient over all τ for $\nu \in \{1, 2, 3, 4, 5\}$. However, the variability of the lengths decreases as n grows. This would suggest that the confidence intervals are tightening as n increases but the point estimates are not converging to the true parameter values as quickly, leading to poor coverage. While not shown in the figures, as the degrees of freedom increase this phenomenon becomes somewhat more moderate and as $\nu \to \infty$, i.e. for Gaussian errors, the 'V' behaviour is no longer present and all methods perform well, with the exception of the iid method.

More generally with regard to the average lengths, all methods considered experience difficulty constructing concise confidence interval lengths at $\tau \in \{0.1, 0.9\}$. Figure 1.3 demonstrates this point for a set of fixed uniform covariates with t_1 distributed errors and $n = 150$. At more moderate τ, the variability of the average lengths decreases for all models. As would be expected, the average lengths and their associated variability decrease as the sample size increases and as τ becomes more moderate. Interestingly in Figure 1.3 most models are performing very similarly in terms of estimated lengths and their variability. The iid method appears to be doing quite well in terms of length, however the empirical coverage probabilities are well below the nominal level.

1.3.2 *Skewed covariates*

The next class of models considered take the same form as equation (1.4) with $u_i \sim \mathcal{N}(0, 1)$ and the covariates are sampled from a highly skewed distribution. We considered $x_i \sim \chi_1^2$, χ_2^2 and log normal with mean 0 and variance 1 on the log scale. Skewed covariates may cause issues for quantile regression estimates particularly with low sample sizes at extreme τ. The smattering of outlying observations in the tail are likely to wreak havoc on the stability of the extreme quantile estimates. Therefore, a priori, we would assume that confidence intervals for both β_0 and β_1 will be affected quite significantly in the case of small n.

Figure 1.4 is indicative of the coverage patterns exhibited by the various techniques in the presence of skewed covariates for $n = 50$. In this particular example, we have covariates sampled from the χ_1^2 distribution. The resampling methods, with the exception of the mcmb and pwy approach, all perform quite well in terms of estimated coverages for both the slope and the intercept, they continue to work well as n increases. The mcmb and pwy approaches exhibit a strong 'V' shape which is only tempered at $n = 200$ and for the extremely heavy tailed log normal distribution. Even at $n = 200$, the mcmb and pwy methods continue to result in empirical coverage probabilities higher than the nominal level for all $\tau \in \{0.1, 0.2 \ldots, 0.9\}$.

As expected, when it comes to estimating the slope coefficient, the iid and nid methods that rely heavily on standard asymptotic normality theory perform sub-optimally across the whole range of τ considered even at $n = 200$. Indeed, the iid method yields empirical coverage probabilities lower than the nominal level even for the intercept parameter.

The rank inversion methods perform well over all sample sizes and τ, however they tend to yield empirical coverage probabilities below the nominal level. Figure 1.5 plots the lengths of the estimated confidence intervals for $n = 100$. The key feature here is the variability inherent in the mcmb and to a lesser extent, pwy, riid and rnid methods. It is generally preferable for confidence intervals to err on the side of conservatism which is why, in the case of skewed covariates, the pbs, xy and wxy methods appear to be the best performers in terms of estimated coverage and are all equally well behaved in terms of confidence interval length.

1.3.3 *Multiple regression and heteroskedasticity*

Here we consider a more general functional form with two covariates and allow for the possibility of heteroskedasticity,

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + (1 + \alpha x_{i1}) u_i. \tag{1.5}$$

Models where x_1 and x_2 are independent are considered as well as models where x_1 and x_2 are bivariate Gaussian with variance 1 and various values for the correlation between x_1 and x_2.

Firstly, considering models with no heteroskedasticity, i.e. $\alpha = 0$ but letting x_1 and x_2 come from a bivariate Gaussian distribution, we find that all

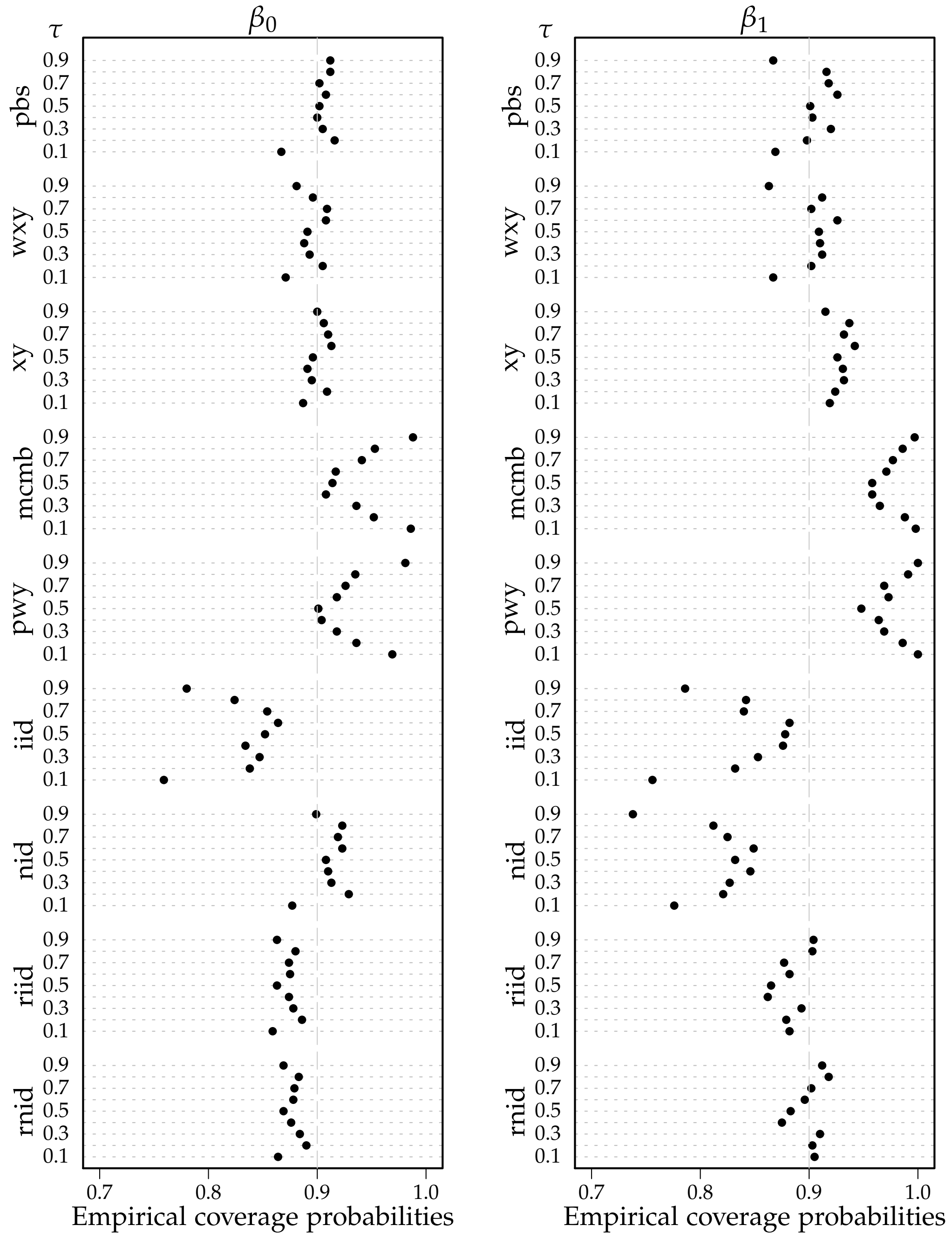


Figure 1.4: Empirical coverage probabilities for the model $Y_i = \beta_0 + \beta_1 x_i + u_i$, where $x_i \sim \chi_1^2$, $u_i \sim \mathcal{N}(0,1)$ and $\beta_0 = \beta_1 = 10$. The sample size is $n = 50$, the number of resamples in each of the bootstrap methods is $B = 1000$, the number of Monte Carlo simulations is $N = 1000$. The nominal coverage is 0.9.

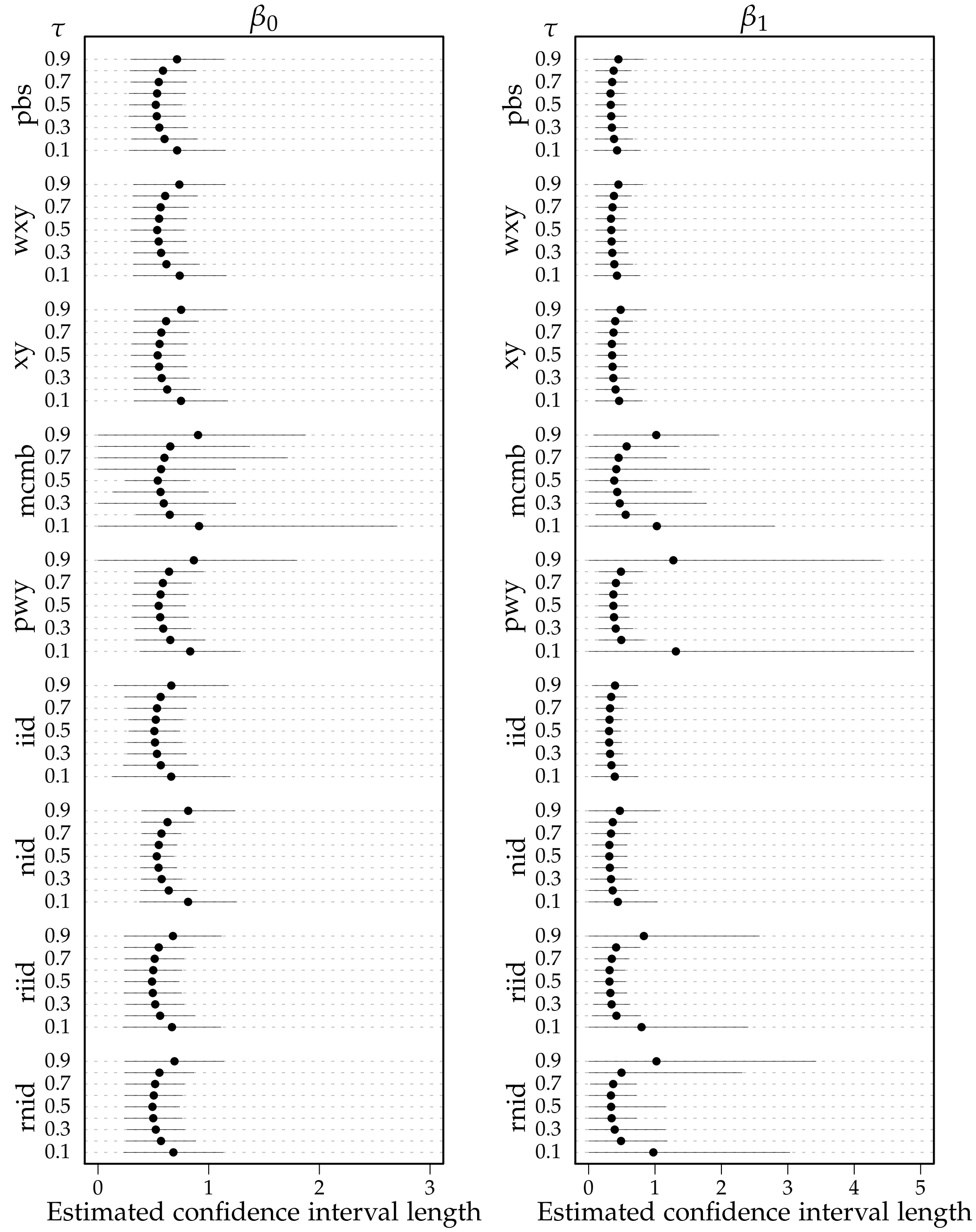


Figure 1.5: Average confidence interval lengths for the model $Y_i = \beta_0 + \beta_1 x_i + u_i$, where $x_i \sim \chi^2_1$, $u_i \sim \mathcal{N}(0,1)$ and $\beta_0 = \beta_1 = 10$. The sample size is $n = 100$, the number of resamples in each of the bootstrap methods is $B = 1000$, the number of Monte Carlo simulations is $N = 1000$. The circles represent the mean length, the horizontal lines are a guide to the variability of the mean length estimate and represent a naive 95% confidence interval constructed as the mean length ± two times the SD of the lengths.

of the techniques, with the exception of the iid method, perform well even when x_1 and x_2 have correlation as high as 0.9.

However, when we introduce heteroskedasticity, even in a simple linear regression type scenario, i.e. $\beta_2 = 0$, the results are quite different. As expected, the methods that rely on the independently distributed error assumption, iid, riid and mcmb, do very poorly in terms of coverage for both the intercept and the slope coefficient. It is interesting that also at $\alpha = 0.5$, the methods designed to be robust to the independent error assumption begin to falter at high and low quantiles. The resampling techniques, pbs, wxy and xy, perform reasonably well for $\tau \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$ over all n. When the error is highly correlated with the covariate, e.g. $\alpha > 0.5$, extreme caution needs to be exercised when conducting inference on the slope parameter away from $\tau = 0.5$.

In the multivariate case, these results still hold. Figure 1.6 demonstrates both of the above points with a sample size of $n = 200$. Here we have $\alpha = 0.5$, and the failure of the methods relying on the iid assumption is evident – further, even the models designed to be robust to this assumption perform poorly for $\tau \in \{0.1, 0.9\}$. The resampling techniques (with the exception of the mcmb method which requires the errors to be independent) deserve special mention. Looking at the slope coefficient of the variable that is not directly related to the error term, the resampling methods (with the exception of the mcmb method which requires the errors to be independent) are quite consistent in their slight over estimation of the true coverage whilst the nid and rank inversion methods all perform quite well. Looking at all three coefficients jointly over the range of τ, it is difficult to ignore the performance of the percentile bootstrap. In terms of lengths, all methods perform quite similarly, however, the rank inversion methods exhibit more variability in their estimates than the resampling techniques.

Introducing higher correlation between the covariates, does not noticeably affect the empirical coverage probabilities, though the lengths of the confidence intervals tend to increase. The major insight is that in the presence of high correlation between the covariates, the coverage will be largely unaffected, though it is likely that the length of the confidence interval will be greater than in the uncorrelated case.

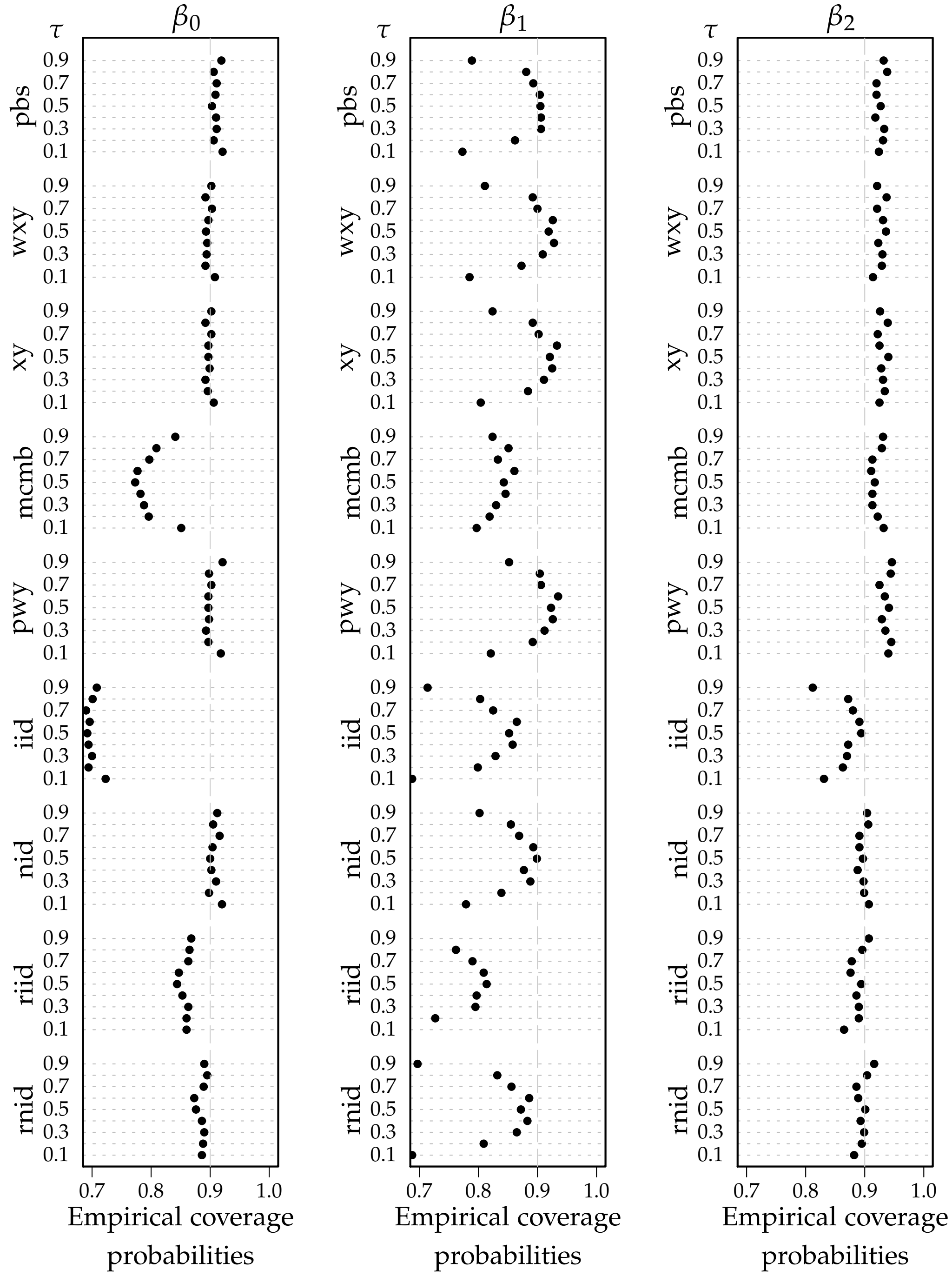


Figure 1.6: Empirical coverage probabilities for the model defined by equation (1.5), where x_1 and x_2 are multivariate Gaussian, both with variance 1 and correlation coefficient 0.5; $u_i \sim \mathcal{N}(0, 1)$; $\alpha = 0.5$ and $\beta_0 = \beta_1 = \beta_2 = 10$. The sample size is $n = 200$, the number of resamples in each of the bootstrap methods is $B = 1000$, the number of Monte Carlo simulations is $N = 1000$. The nominal coverage is 0.9.

1.4 CONCLUSION

The aim of this chapter was to revisit and extend the analysis performed by Kocherginsky, He and Mu (2005), incorporating additional techniques and evaluating their effectiveness for sample sizes $n \leq 200$ and over a broad range of conditional quantiles. We considered simple models with only a few parameters so as to isolate the effect of the various pathologies. This is an important distinction to make as some of the methods considered were designed primarily for use in large samples. For example the mcmb method is most valuable when estimating high dimensional models.

The percentile bootstrap was found to be a sound performer exhibiting significant robustness to heteroskedasticity, heavy tailed error distributions and skewed covariates. In most cases it was found to be better than, or at least on par with, more complex resampling techniques as well as the nid and rank inversion methods in terms of coverage probabilities, average lengths and the variability of those lengths. The performance of the percentile bootstrap was somewhat surprising given its simplicity relative to the other resampling techniques. However, Hahn (1995) foreshadowed this result, showing the asymptotic empirical coverage probability of the bootstrap percentile method to be equal to the nominal coverage probability even when the error term is not independent of the regressor. It is noted in the R documentation for the `quantreg` (Koenker, 2013) package that the percentile method is a "refinement that is still unimplemented", though it is not difficult to code directly.

All methods did not perform well when estimating the intercept for high and low quantiles in the presence of heavy tailed errors. Also, when there is strong heteroskedasticity caused by one variable, the estimated coverage probabilities for the corresponding coefficient can be severely underestimated at high and low quantiles. When the error term is dependent on one or more covariates and extreme τ is of interest, caution should be used when constructing confidence intervals, even for large n.

The nid method generally performed quite well, except in the presence of heavy tailed covariates. This was also noted in Kocherginsky, He and Mu (2005) where they similarly found the nid method tended to underestimate the coverage. The reason being that when a covariate is too heavy tailed, the asymptotic normality of $\hat{\boldsymbol{\beta}}_\tau$ would fail. In this case, the percentile bootstrap or rnid would be suggested, noting that the percentile bootstrap gives consistently tighter lengths than the rnid approach.

The mcmb approach uses the MCMB-A algorithm which is not robust in the presence of heteroskedasticity. At high and low τ even in iid models, the mcmb approach did quite poorly and as such would not be recommended in all but the most well behaved of models when $n \leq 200$.

As expected the iid and riid methods did not perform well when the iid assumption was violated and did not outperform the more robust methods in the standard iid error models. The xy-pair bootstrap and pwy techniques performed well, except when there were correlated covariates.

The rank test inversion methods, riid and rnid, occasionally generated confidence intervals of infinite length, particularly with small sample sizes, heavy tailed covariates and extreme τ. If this is observed in practice, the percentile or paired bootstrap or nid method would be an appropriate alternative.

The confidence interval lengths for the rank inversion methods were often observed to be far more variable than the other methods. This has much to do with the relatively small sample sizes and the way the confidence intervals are generated by inverting a test statistic. With less data points, particularly at extreme τ, the inversion process has to search further afield to find appropriate upper and lower bounds. In the other methods, a standard error is estimated and a standard symmetric confidence interval is calculated. The xy-pair bootstrap and the nid method on average gave the smallest confidence intervals whilst maintaining acceptable coverage performance.

To summarise, the percentile bootstrap approach is a simple, intuitive and viable alternative for constructing confidence interval estimates in the quantile regression setting. Hahn (1995) showed that the percentile bootstrap provides asymptotically correct coverage probabilities and we found its small sample performance to be impressive.

2 SCALE ESTIMATION

2.1 INTRODUCTION

A reliable estimate of the scale of a data set is a fundamental issue in statistics. For example 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 unusual observations. The robustness of quantile regression parameter estimates, described in Chapter 1, 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 problem of finding reliable robust estimates of scale.

While scale parameters are sometimes treated as nuisance parameters, a talk given by Raymond Carroll at the University of Technology, Sydney, in June 2013 shared the same title as his 2003 paper, "Variances are not always nuisance parameters" indicating that finding reliable estimates of scale is just as important a problem in 2013 as it was in 2003.

Robust estimates of scale are important for a range of applications, from true scale problems, to outlier identification, and as auxiliary parameters for more involved analyses. Recent work concerning robust scale estimation includes Boente, Ruiz and Zamar (2010), Wu and Zuo (2008) and Van Aelst, Willems and Zamar (2013).

There are two aims in formulating a robust estimator: the first is to reduce the potential bias caused by outliers and the second is to maintain efficiency when there are no outliers present. These two aims are generally in conflict with one another. In the scale setting, the median absolute deviation from the median (MAD) is commonly used in practice, despite its poor Gaussian efficiency. The estimator Q_n (Rousseeuw and Croux, 1993) is a significant improvement on the MAD in terms of efficiency whilst maintaining a high level of robustness. This chapter presents an alternative robust scale estimator which trades some robustness for desirable efficiency properties.

We propose a new robust scale estimator, the pairwise mean scale estimator P_n, which combines familiar features from a number of commonly used robust estimators and possesses surprising efficiency properties (Tarr, Müller and Weber, 2012). In contrast to Q_n, which utilises pairwise differences, P_n is based on pairwise means. The most basic form of P_n is calculated as the IQR of the pairwise means, yielding a scale estimator that can be viewed as a natural complement to the Hodges-Lehmann location estimator (Hodges and Lehmann, 1963). A generalisation, $P_n(\tau)$, considers the distance between the $(1 \pm \tau)/2$ quantiles of the empirical distribution of the pairwise means. Unless otherwise specified, the notation P_n is equivalent to $P_n(0.5)$. Extensions are investigated that are based around Winsorising and trimming. We also implement a form of adaptive trimming which is shown to achieve the maximal breakdown value of 50%.

Our investigation into the efficiency properties of the pairwise mean scale estimator is based on Randal (2008) who calculates efficiencies of estimators relative to the corresponding maximum likelihood estimators at a particular distribution. This method facilitates easier comparison than the method used in the seminal study by Lax (1985).

The pairwise mean scale estimator fits into the family of generalised L-statistics (GL-statistics) which encompasses broad classes of statistics of interest in nonparametric estimation; in particular, L-statistics, U-statistics and U-quantile statistics (Serfling, 1984). Thus, a wide range of statistics related to scale estimation are embedded into a single unified class. For example, the IQR; variance; trimmed and Winsorised variance; and Q_n all fit within the class of GL-statistics.

M-estimators are an important exception to the class of GL-statistics. The MAD is the most prominent robust scale estimator that sits under the M-estimator umbrella. An advantage of P_n over M-estimators of scale is that it does not require a location estimate.

The next section outlines an important family of statistics and introduces some common decomposition techniques used throughout the thesis to derive limiting distributions. A review of some existing scale estimators and their properties are given in Section 2.3. Section 2.4 formally defines the estimator P_n along with possible generalisations and its breakdown value is found. The influence function and asymptotic normality of P_n are also derived in Section 2.4. In addition to being an intuitive estimate of scale, one of the primary advantages of P_n is its high efficiency over a broad range of distributions. The results of a simulation study are given in Section 2.5 which show how P_n compares favourably with other robust estimates of scale.

2.2 BACKGROUND THEORY

This section outlines some theory that will be used throughout the thesis. We begin by introducing U-statistics and U-quantile statistics. The Hoeffding decomposition, a classical approach to working with U-statistics, and Hermite polynomials, a classical approach to decomposing Gaussian processes, are both briefly discussed. These techniques will be most relevant in Chapter 3. Finally, a broader class of statistics that encompasses both U- and U-quantile statistics is introduced.

2.2.1 *U- and U-quantile statistics*

Let $X_1, \dots, X_n$ be a set of iid observations and g be a symmetric bivariate kernel, $g : \mathbb{R}^2 \mapsto \mathbb{R}$. Hoeffding (1948) defines a U-statistic (of order 2) as,

$$\frac{2}{n(n-1)} \sum_{i<j} g(X_i, X_j).$$

An example of a U-statistic is the sample variance which is a U-statistic with $g(x, y) = (x - y)^2/2$.

A U-quantile statistic is a quantile of the distribution function of the kernels of the U-statistic. Let $G(t) = \mathrm{P}\left(g(X_1, X_2) \leq t\right)$ be the CDF of the kernels with corresponding empirical distribution function,

$$G_n(t) = \frac{2}{n(n-1)} \sum_{i<j} \mathbb{I}\{g(X_i, X_j) \leq t\}, \quad \text{for } t \in I, \tag{2.1}$$

for some interval $I \subseteq \mathbb{R}$. Note that (2.1) is a U-statistic with kernel,

$$h(x, y; t) = \mathbb{I}\{g(x, y) \leq t\}, \quad \forall\, x, y \in \mathbb{R} \quad \text{and} \quad t \in I. \tag{2.2}$$

For $p \in (0, 1)$, the corresponding sample U-quantile is,

$$G_n^{-1}(p) = \inf\{t : G_n(t) \geq p\}.$$

2.2.1.1 *Hoeffding decomposition*

The Hoeffding decomposition is the classical mode of attack for analysing the asymptotics of non-degenerate U-statistics. Let $h_1(x; t) = \mathbb{E}[h(x, X_1; t)] - G(t)$. For all $t \in I$, write the difference,

$$G_n(t) - G(t) = \frac{1}{n(n-1)} \sum_{i \neq j} \left[h(X_i, X_j; t) - G(t)\right], \tag{2.3}$$

as,

$$G_n(t) - G(t) = W_n(t) + R_n(t),$$

where,

$$W_n(t) = \frac{2}{n} \sum_{i=1}^{n} h_1(X_i; t),$$

and,

$$R_n(t) = \frac{1}{n(n-1)} \sum_{i \neq j} \left[h(X_i, X_j; t) - h_1(X_i; t) - h_1(X_j; t) - G(t) \right]. \tag{2.4}$$

The function $h_1(x; t)$ is defined for all $x \in \mathbb{R}$ and $t \in I$ and if $X_1, \dots, X_n$ are independent standard Gaussian,

$$h_1(x; t) = \int h(x, y; t) \phi(y) \, \mathrm{d}y - G(t),$$

and hence, $\mathbb{E}[h_1(X_2; t)] = 0$.

2.2.1.2 *Hermite decomposition*

As Beran (1994) notes, the classical approach based on the Hoeffding decomposition is not always appropriate for establishing the asymptotic behaviour of U-statistics when working with long range dependent (LRD) sequences, as considered in Chapter 3. An alternative approach is to use a decomposition based on Hermite polynomials. The kth Hermite polynomial, $H_k(x)$, is defined as,

$$H_k(x) = (-1)^k e^{x^2/2} \left[\frac{\mathrm{d}^k}{\mathrm{d}x^k} e^{-x^2/2} \right].$$

Using this definition, the first four Hermite polynomials are: $H_0(x) = 1$, $H_1(x) = x$, $H_2(x) = x^2 - 1$ and $H_3(x) = x^3 - 3x$.

Hermite polynomials build an orthogonal basis. To see this let Z be a standard Gaussian random variable then,

$$\mathbb{E}[H_k(Z) H_k(Z)] = k!,$$

and for all $k \neq j$,

$$\mathbb{E}[H_k(Z) H_j(Z)] = 0.$$

Furthermore, let $\mathcal{J}$ be the set of functions J such that $\mathbb{E}[J(Z)] = 0$ and $\mathbb{E}[J^2(Z)] < \infty$ then every function $J \in \mathcal{J}$ can be written as

$$J(Z) = \sum_{k=0}^{\infty} \frac{\alpha_k}{k!} H_k(Z),$$

with Hermite coefficients $\alpha_k = \mathbb{E}[J(Z)H_k(Z)]$. The *Hermite rank* of a function J is defined as $m = \inf\{k \geq 1 : \alpha_k \neq 0\}$. We therefore have,

$$J(Z) = \sum_{k=m}^{\infty} \frac{\alpha_k}{k!} H_k(Z).$$

Let X and Y be independent standard Gaussian random variables. The kernel function defined in (2.2) can be expanded in a bivariate Hermite polynomial basis as follows,

$$h(X,Y;t) = \mathbb{I}\{g(X,Y) \leq t\} = \sum_{p,q\geq 0} \frac{\alpha_{p,q}(t)}{p!q!} H_p(X)H_q(Y), \tag{2.5}$$

for $t \in I$, where, $\alpha_{p,q}(t) = \mathbb{E}\left[h(X,Y;t)H_p(X)H_q(Y)\right]$. The constant term in the bivariate Hermite decomposition (2.5) is given by $\alpha_{0,0}(t)$,

$$\alpha_{0,0}(t) = G(t) = \int\int h(x,y;t)\phi(x)\phi(y)\,\mathrm{d}x\,\mathrm{d}y \quad \text{for all } t \in I.$$

As we will see in Chapter 3, the Hermite rank of the class of functions $\{h(\cdot,\cdot;t) - G(t), t \in I\}$ plays a crucial role in understanding the asymptotic behaviour of empirical quantiles of the U-process G_n. Analogously to the univariate case, the Hermite rank of the bivariate function $h(\cdot,\cdot;t)$ is defined as the smallest positive integer $m(t)$ such that there exist p and q satisfying $p+q = m(t)$ and $\alpha_{p,q}(t) \neq 0$. Thus we can write (2.5) as,

$$h(X,Y;t) - G(t) = \sum_{\substack{p,q\geq 0 \\ p+q\geq m(t)}} \frac{\alpha_{p,q}(t)}{p!q!} H_p(X)H_q(Y).$$

Further details on Hermite polynomials and their properties can be found, among others, in Beran (1994), Giraitis, Koul and Surgailis (2012) and Beran et al. (2013).

2.2.2 *Generalised L-statistics*

Serfling (1984) introduces an extension of U-quantile statistics known as generalised L-statistics (GL-statistics). Many robust scale estimators are nested within the class of GL-statistics, including our new scale estimator P_n.

Again restricting attention to symmetric bivariate kernels, $g(x_1, x_2)$, the GL-functional is defined as,

$$T(G) = \int_0^1 w(p)G^{-1}(p)\,\mathrm{d}p + \sum_{j=1}^{d} a_j G^{-1}(p_j), \tag{2.6}$$

where w is a function for the smooth weighting of $G^{-1}(p)$ and a_j are discrete coefficients for $G^{-1}(p_j)$ (Serfling, 1984). Evaluation of the GL-functional at the corresponding sample distribution of $g(X_1, X_2)$, yields the following representation of GL-statistic,

$$T_n = T(G_n) = \int_0^1 w(p) G_n^{-1}(p) \mathrm{d}p + \sum_{j=1}^{d} a_j G_n^{-1}(p_j). \tag{2.7}$$

The class of generalised L-statistics (GL-statistics) encompasses L-statistics, U-statistics and U-quantile statistics and as such covers a range of scale estimates. Janssen, Serfling and Veraverbeke (1984) and Serfling (1984) study the asymptotic properties of GL-statistics in some detail.

2.3 REVIEW OF ROBUST SCALE ESTIMATION THEORY

We define any estimator, S_n, that is shift invariant and scale equivariant to be a scale estimator. That is, for a set of n observations, $\boldsymbol{x}$, and any constant $c \in \mathbb{R}$, $S_n(\boldsymbol{x} + c) = S_n(\boldsymbol{x})$ and $S_n(c\boldsymbol{x}) = |c| S_n(\boldsymbol{x})$.

A traditional way to measure the spread of a data set $\boldsymbol{x}$ is the (non-robust) sample standard deviation,

$$\text{SD}(\boldsymbol{x}) = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2},$$

where $\bar{x} = n^{-1} \sum_{i=1}^{n} x_i$ is the sample mean. However, it is well known that the SD is not robust to outlying observations. This section outlines some key features of robust estimators and defines some common robust alternatives to the SD, many of which fit within the class of GL-statistics.

2.3.1 *Measures of robustness*

The core of this thesis focusses on robust estimation techniques. An introduction to the philosophy underlying robust procedures is found in Morgenthaler (2007). In general, an estimator is said to be robust if it is relatively unaffected by arbitrary corruption to some small proportion of observations. The proportion of corrupted observations in the data set will be referred to as ε. The *most robust* estimators return bounded estimates as $\varepsilon \uparrow 0.5$, i.e. up to half of the data set may experience arbitrary corruption. In practice, it is not likely that univariate samples will have such a high level of contamination so we do not restrict attention solely to the most robust methods.

Two popular ways of describing the robustness properties of estimators are the breakdown value and the influence function.

2.3.1.1 *Breakdown value*

The breakdown value of an estimator, ε^*, is the smallest value of ε for which the estimator, T_n, when applied to the ε-corrupted sample $\widetilde{\boldsymbol{x}}$, can be forced to the boundary of the parameter space.

Hodges (1967) and Hampel (1971) were the first to propose and define the concept of a breakdown value. Hubert and Debruyne (2009) provide a recent overview of its history. The classical definition of the asymptotic contamination breakdown value of a location estimator T_n at $\boldsymbol{x}$ is

$$\varepsilon^*(\boldsymbol{x}, T_n) = \inf\{\varepsilon | b(\varepsilon; \boldsymbol{x}, T_n) = \infty\},$$

where b is the maximum bias that can be caused by ε-corruption:

$$b(\varepsilon; \boldsymbol{x}, T_n) = \sup |T_n(\widetilde{\boldsymbol{x}}) - T_n(\boldsymbol{x})|,$$

and the supremum is taken over the set of all ε-corrupted samples $\widetilde{\boldsymbol{x}}$.

For a scale estimator, S_n, we need to adjust the classical definition slightly, to include the possibility of implosion, i.e. returning a value of zero. In finite samples this is defined by,

$$\varepsilon^*(\boldsymbol{x}, S_n) = \max\left\{\frac{m}{n} : \sup_{\widetilde{\boldsymbol{x}}} S_n(\widetilde{\boldsymbol{x}}) < \infty \text{ and } \inf_{\widetilde{\boldsymbol{x}}} S_n(\widetilde{\boldsymbol{x}}) > 0\right\},$$

where m is the number of observations in $\boldsymbol{x}$ replaced with arbitrary values (Huber, 1981, p. 110).

2.3.1.2 *Influence function*

Hampel (1974) defines the influence function for a functional T at the distribution F as,

$$\text{IF}(x; T, F) = \lim_{\epsilon \downarrow 0} \frac{T((1-\epsilon)F + \epsilon\delta_x) - T(F)}{\epsilon}, \tag{2.8}$$

where the distribution δ_x has all its mass at x. The influence function is essentially the first order Gâteaux derivative of a functional T at a distribution F in the direction of δ_x. It represents the effect of a point mass contamination at x on the estimate, in a sense, capturing the asymptotic bias caused by the contamination.

An estimator is said to be B-robust (bias-robust) if it has a bounded influence function, i.e. the influence function does not go to infinity as $x \to \pm\infty$. The absolute maximum limit of the influence function is known as the gross-error sensitivity, $\gamma^*(T, F) = \sup_x |\,\mathrm{IF}(x; T, F)|$, which measures the worst approximate influence a fixed amount of contamination can have on the value of the estimator, i.e. it represents an approximate bound for the asymptotic bias of the estimator. Hence, if $\gamma^*(T, F)$ is finite, then T is said to be B-robust.

Another way to characterise the robustness of an estimator is through its maximum bias curve. This is a generalisation of gross error sensitivity over various levels of contamination, ε, within a specified ε-contaminated family of distributions. Early work on maximum bias curves in the context of scale estimation can be found in Martin and Zamar (1989, 1993).

The influence function is also useful in finding the asymptotic variance of an estimator, which equals the expected square of the influence function,

$$\mathrm{var}(T, F) = \int_{\mathbb{R}} \mathrm{IF}^2(x; T, F)\,\mathrm{d}F(x).$$

Huber (1964) notes that the asymptotic relative efficiency is an important criteria that can be used to choose between competing estimators. Once the asymptotic variance has been calculated, the asymptotic relative efficiency between two estimators T and S at a distribution F is defined as,

$$\mathrm{ARE}(T, S, F) = \frac{\mathrm{var}(S, F)}{\mathrm{var}(T, F)}.$$

2.3.2 *Existing scale estimators*

There is an extensive literature on scale estimation. By way of review, Huber and Ronchetti (2009, Chapter 5) outline a number of scale estimates falling into the M-, L- and R-statistic classes. This section continues by outlining some important scale estimators that will be referred to throughout the thesis.

2.3.2.1 *Interquartile range*

The interquartile range (IQR) was an early attempt to robustify scale estimation, see Hojo (1931), and is still widely taught and referred to in practice. The IQR can be defined simply as $\mathrm{IQR}(\boldsymbol{x}) = x_{(n-m+1)} - x_{(m)}$, where $m = \lfloor n/4 \rfloor$. However, there are as many ways to calculate the IQR as there are ways to calculate quantiles, see Hyndman and Fan (1996) for a summary.

Importantly, the IQR does not require an estimate of location, i.e. the data need not be centred. The IQR, as a linear combination of quantiles, can be viewed as a L-estimate, which sits within the class of GL-statistics.

Adapting the general result for the τ-quartile range in Hampel et al. (1986, p. 110), the influence function for the IQR at $F = \Phi$ is,

$$\mathrm{IF}(x; \mathrm{IQR}, \Phi) = \frac{\mathrm{sign}(|x| - \Phi^{-1}(\frac{3}{4}))}{4\Phi^{-1}(\frac{3}{4})\phi(\Phi^{-1}(\frac{3}{4}))}.$$

The gross error sensitivity of the IQR is $\gamma^* = 1.167$, and hence the IQR is B-robust. Furthermore, its asymptotic breakdown value is $\varepsilon^* = \frac{1}{4}$ and its asymptotic efficiency relative to the SD is 0.367 (Hampel et al., 1986).

2.3.2.2 S_n

Croux and Rousseeuw (1992) propose the scale estimator S_n, defined as,

$$S_n(\boldsymbol{x}) = c_n 1.1926 \operatorname*{Med}_i \left\{ \operatorname*{Med}_j |X_i - X_j| \right\}. \tag{2.9}$$

This should be read as follows: for each i we compute the median of $\{|X_i - X_j|; j = 1, \ldots, n\}$. This yields n numbers, of which the median is then found. The factor 1.1926 is for consistency at Gaussian distributions and c_n is a finite sample correction factor. This estimator also achieves a breakdown value of $\varepsilon^* = \lfloor n/2 \rfloor / n$, which is the best possible value for location invariant and scale equivariant estimators.

In order to check whether the factor 1.1926, obtained by means of an asymptotic argument, succeeds in making S_n approximately unbiased for finite samples, Croux and Rousseeuw (1992) perform a simulation study. For each n they generate 10,000 samples of size n Gaussian observations and then compute the average value of (2.9) and the standard error on that value. For n even, there is practically no bias. However, for n odd a small bias appears. Hence, the correction factor c_n is explicitly given for $2 \leq n \leq 9$ and for $n > 9$ is defined as,

$$c_n = \begin{cases} \dfrac{n}{n - 0.9} & \text{for } n \text{ odd} \\ 1 \quad \text{for } n \text{ even.} \end{cases}$$

In order to be able to give c_n with three decimal places, they repeat the simulation for small n with 200,000 replications.

2.3.2.3 Q_n

Croux and Rousseeuw (1992) and Rousseeuw and Croux (1993) also consider another estimator, Q_n, which is more commonly used in practice than S_n. Like the IQR, Q_n does not require centring, however it has a higher breakdown value and is more efficient than the IQR at the Gaussian distribution. The estimator Q_n is based only on the differences between the data values,

$$Q_n = d_n 2.2191\{|x_i - x_j|; i < j\}_{(k)}, \tag{2.10}$$

i.e. the kth largest of the $|x_i - x_j|$ for $i < j$ multiplied by an asymptotic correction factor, 2.2191, and a finite correction factor d_n.

Croux and Rousseeuw (1992) show that Q_n achieves the maximal asymptotic breakdown value, $\varepsilon^* \to 0.5$ as $n \to \infty$, for a range of k however, the most efficient choice is to use $k = \binom{h}{2}$ where $h = \lfloor n/2 \rfloor + 1$ which gives,

$$k = \frac{(\lfloor n/2 \rfloor + 1)\lfloor n/2 \rfloor}{2} \approx \binom{n}{2}\frac{1}{4}.$$

Thus, Q_n is approximately the first quartile of the $|x_i - x_j|$'s (for $i < j$ there are $\binom{n}{2}$ absolute differences) and as such can be asymptotically represented by a GL-statistic with kernel, $g(x,y) = |x - y|$. If the kth order statistic is replaced with the median, this would be equivalent to the median interpoint distance mentioned by Bickel and Lehmann (1979), the breakdown point of which is lower (about 29%).

The influence function for Q_n at a distribution F is given by,

$$\text{IF}\,(x; Q, F) = d\frac{\frac{1}{4} - F(x + d^{-1}) + F(x - d^{-1})}{\int f(x + d^{-1}) f(x)\,\mathrm{d}x}, \tag{2.11}$$

where d is a correction factor specific to each F. At the Gaussian distribution, $d = 2.2191$ and the gross error sensitivity is $\gamma^* = 2.07$.

Croux and Rousseeuw (1992) show that the factor 2.2191 in (2.10) is necessary for asymptotic consistency for the standard deviation when the underlying observations follow a standard Gaussian distribution. A simulation study over 10,000 samples was conducted to find the additional correction factor d_n. It is specified explicitly for $n \leq 9$, and for $n > 9$ is taken to be defined as,

$$d_n = \begin{cases} \dfrac{n}{n + 1.4} & \text{for } n \text{ odd} \\ \dfrac{n}{n + 3.8} & \text{for } n \text{ even.} \end{cases}$$

2.3.2.4 *Trimmed standard deviations*

Trimming classical non-robust estimators is an alternative way to achieve robustness while and maintain a reasonably high level of efficiency, see Welsh and Morrison (1990) for examples of trimmed scale estimators. More recently, Wu and Zuo (2008) consider the properties of scaled-deviation trimmed and Winsorised standard deviations.

For a given point x, the scaled deviation of x to the centre of a distribution F is given by,

$$D(x,F) = \frac{x - T(F)}{S(F)},$$

where $T(F)$ and $S(F)$ are some robust location and scale functionals.

The points are trimmed based on the absolute value of this scaled deviation. Let A be the event $\{|D(x,F)| \leq \beta\}$, where β is some arbitrary parameter. The β scaled-deviations trimmed variance functional is defined as,

$$S^2(F) = c\frac{\int_A w(D(x,F))(x - T(F))^2 \,\mathrm{d}F(x)}{\int_A w(D(x,F)) \,\mathrm{d}F(x)}, \tag{2.12}$$

where c is the consistency coefficient and $T(F)$ is a similarly trimmed mean functional. Further, $0 < \beta \leq \infty$ and $w(D(x,F))$ is an even-bounded weight function on $[-\infty, \infty]$ so that the denominator is positive. Here points that are a robust distance ($\beta S(F)$) away from the robust centre $T(F)$ are trimmed and the remaining observations are reweighted. When w is a non-zero constant, S^2 is the usual variance after trimming. Furthermore, as $\beta \to \infty$, T and S^2 approaches the usual mean and variance and $c \to 1$.

A straightforward extension is to use Winsorisation instead of trimming – the difference being that outlying observations are replaced with cutting-point values instead of simply being eliminated. Wu and Zuo (2008) also find the influence functions for the randomly trimmed and Winsorised variance.

Note that when using scaled-deviation trimming or Winsorisation the proportion of the trimmed points for a fixed β, $\mathrm{P}(|D(X,F)| > \beta)$, is not fixed but F-dependent. In finite samples, the proportion of sample points trimmed is not fixed but random, so $T(F_n)$ and $S(F_n)$ are data adaptive.

We implement a similar form of adaptive trimming to improve the robustness of P_n in Section 2.4 which is later shown to be a good basis for precision matrix estimation in Chapter 4.

2.3.2.5 *M-estimators*

Another important class of statistics are M-estimators, which are a natural extension of maximum likelihood estimate (MLE) type estimators. In general, an M-estimate of scale is any estimate, $\hat{\sigma}$, satisfying an equation of the form,

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{x_i}{\hat{\sigma}}\right) = \delta, \tag{2.13}$$

where ρ is a nondecreasing function of $|x|$ with $\rho(0) = 0$, and δ is a positive constant. Note that in order for (2.13) to have a solution, we must have $0 < \delta < \rho(\infty)$. If ρ is bounded it will be assumed without loss of generality that $\rho(\infty) = 1$ and hence $\delta \in (0,1)$.

When ρ is the step function,

$$\rho(t) = \mathbb{I}\{|t| > c\}, \tag{2.14}$$

where c is a positive constant and $\delta = 0.5$, we have $\hat{\sigma} = \text{Med}\{|\boldsymbol{x}|\}/c$.

2.3.2.6 *Median absolute deviation*

A very robust choice for ρ is (2.14) with $c = \Phi^{-1}(\frac{3}{4}) \approx 0.675$ to make it consistent for the standard deviation at the Gaussian distribution. In the general case with unknown location parameter, we can use the median as a robust estimate of location, which yields the median absolute deviation from the median (MAD),

$$\text{MAD}(\boldsymbol{x}) = \frac{1}{0.675}\text{Med}\{|\boldsymbol{x} - \text{Med}\{\boldsymbol{x}\}|\}. \tag{2.15}$$

The MAD is used extensively in practice, indeed we will use it when adding a random trimming component to P_n in Section 2.4.

Hampel (1974) gives detail about the influence function for the MAD. The influence function for the MAD at the Gaussian distribution is the same as for the IQR,

$$\text{IF}(x;\text{MAD},\Phi) = \frac{\text{sign}\left(|x| - \Phi^{-1}(\frac{3}{4})\right)}{4\Phi^{-1}(\frac{3}{4})\phi\left(\Phi^{-1}(\frac{3}{4})\right)}, \tag{2.16}$$

and is plotted in Figure 2.4 (p. 39). It follows that the IQR and MAD share some of the same asymptotic properties at the Gaussian. In particular, the gross error sensitivity of the MAD at the Gaussian distribution is $\gamma^* \approx 1.167$, therefore the MAD is B-robust. Using the fact that the asymptotic variance of an estimator is the expected square of the influence function, $\text{var}(\text{MAD},\Phi) \approx 1.361$ and hence the asymptotic efficiency of the MAD relative to the SD at the Gaussian distribution is $\text{ARE}(\text{MAD},\text{SD},\Phi) \approx 0.367$.

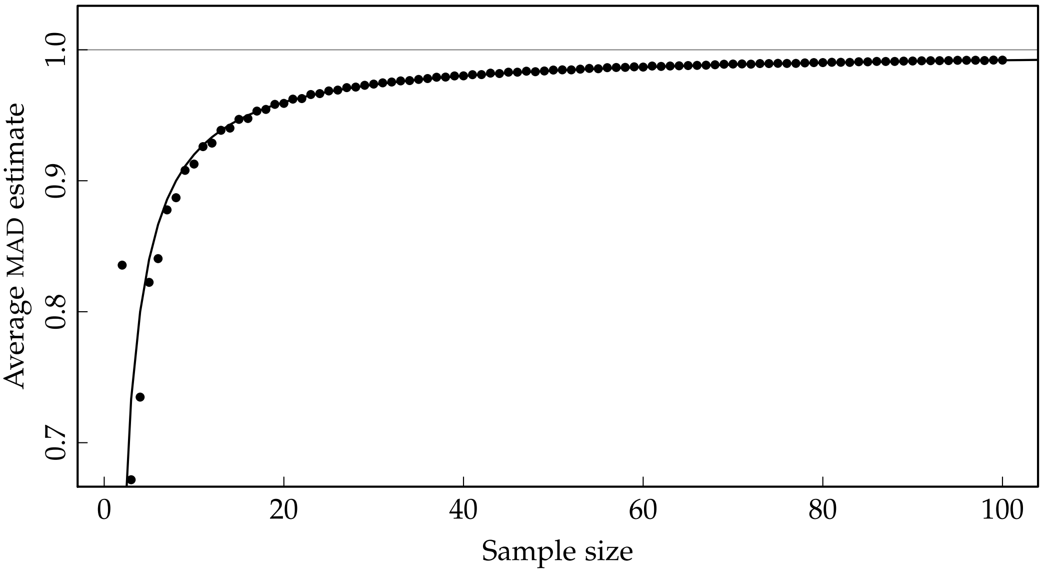


Figure 2.1: Average MAD estimates over one million replications of different sample sizes from a standard Gaussian distribution. The black line represents the finite sample correction factor and the grey line is the true parameter value.

The key difference between the MAD and the IQR is that the MAD possesses maximal asymptotic breakdown value, $\varepsilon^* = 0.5$.

For symmetric distributions, the MAD is asymptotically equivalent to one-half of the interquartile distance. Hall and Welsh (1985), Welsh (1986) and more recently, Mazumder and Serfling (2009) study the asymptotic properties of the MAD and its link with the semi-IQR.

With correction factors to ensure consistency for the standard deviation at the Gaussian distribution, the MAD can be redefined as,

$$\text{MAD}(\boldsymbol{x}) = b_n 1.4826 \times \text{Med}\{|\boldsymbol{x} - \text{Med}\{\boldsymbol{x}\}|\}, \tag{2.17}$$

where the b_n factors are chosen to make MAD_n approximately unbiased in finite samples. Croux and Rousseeuw (1992) use simulation to find finite sample correction factors for the MAD. In particular, they specify them explicitly for $2 \leq n \leq 9$ and if $n > 9$ then use the approximation,

$$b_n = \frac{n}{n - 0.8}.$$

In R, the only correction implemented is an asymptotic one, i.e. $b_n = 1$ for all n. The need for a finite correction factor is demonstrated in Figure 2.1.

2.3.2.7 *τ-scale*

Yohai and Zamar (1988) introduce the class of τ-estimates which have a high breakdown value and controllable efficiency at the Gaussian distribution. If $\hat{\sigma}(\boldsymbol{x})$ is a robust M-estimator of scale, i.e. a solution of equation (2.13), then the τ-scale is defined as,

$$\tau(\boldsymbol{x})^2 = \hat{\sigma}^2(\boldsymbol{x})\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{x_i}{\hat{\sigma}(\boldsymbol{x})}\right).$$

Maronna and Zamar (2002) propose a version of the τ-scale that will be considered in Chapter 4. Namely, the initial M-estimator of scale is the MAD, $\hat{\sigma}(\boldsymbol{x}) = \text{MAD}(\boldsymbol{x})$. The τ-scale estimate is,

$$\tau^2(\boldsymbol{x}) = d\,\text{MAD}(\boldsymbol{x})\frac{1}{n}\sum_{i=1}^{n}\rho_{c_1}\left(\frac{x_i - \hat{\mu}(\boldsymbol{x})}{\text{MAD}(\boldsymbol{x})}\right), \quad \text{where} \quad \hat{\mu}(\boldsymbol{x}) = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}.$$

Note that d is a asymptotic correction factor to ensure consistency at the Gaussian distribution, $\rho_{c_1}(v) = \min(c_1^2, v^2)$ with default $c_1 = 3$ and the weights are calculated as,

$$w_i = w_{c_2}\left(\frac{x_i - \text{Med}(\boldsymbol{x})}{\text{MAD}(\boldsymbol{x})}\right),$$

where $w_{c_2}(u) = \max(0, (1-(u/c_2)^2)^2)$ and the default is $c_2 = 4.5$.

2.4 A SCALE ESTIMATOR BASED ON PAIRWISE MEANS

2.4.1 *The estimator P_n*

Given a set of n observations, $\boldsymbol{x} = (x_1, \dots, x_n)$, the set of $\binom{n}{2}$ pairwise means is $\{g(x_i, x_j), 1 \le i < j \le n\}$, where $g(x_1, x_2) = (x_1 + x_2)/2$. Let G_n be the empirical distribution function of the pairwise means,

$$G_n(t) = \frac{2}{n(n-1)}\sum_{i<j}\mathbb{I}\{g(x_i, x_j) \le t\}, \quad \text{for } t \in \mathbb{R}.$$

The estimator $P_n(\tau)$ is defined as

$$P_n(\boldsymbol{x}, \tau) = P_n(\tau) = c_\tau\left[G_n^{-1}\left(\frac{1+\tau}{2}\right) - G_n^{-1}\left(\frac{1-\tau}{2}\right)\right], \tag{2.18}$$

where c_τ is a correction factor to make $P_n(\tau)$ consistent for the standard deviation when the underlying observations are Gaussian and $0 < \tau \le 1$. By this definition, $P_n(\tau)$ is the range of the middle $\tau \times 100\%$ of G_n.

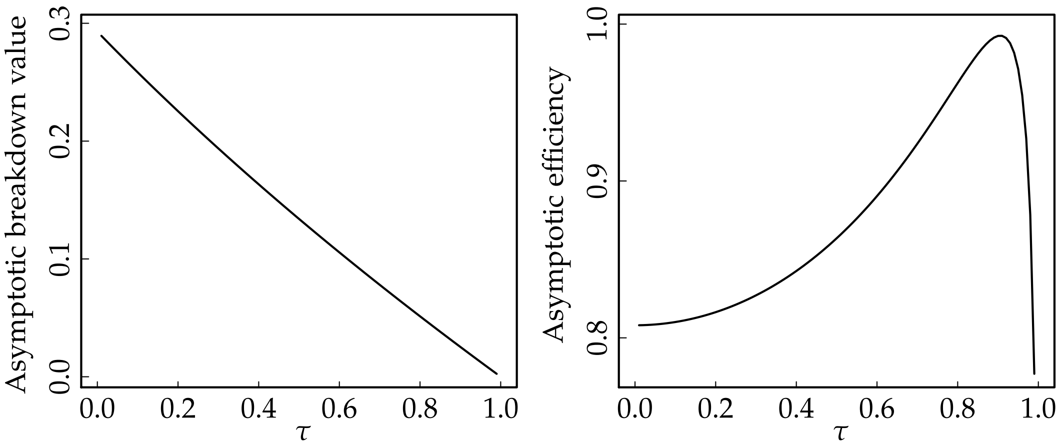


Figure 2.2: Breakdown value and Gaussian efficiency of $P_n(\tau)$.

The notion of working with quantiles of pairwise means is not new. The Hodges-Lehmann estimate, the median of the pairwise means, is a well known robust estimator of location (Hodges and Lehmann, 1963). The pairwise mean scale estimator, P_n, can be thought of as the scale analogue of this location estimate. Following the algorithm set out in Johnson and Mizoguchi (1978), in a similar manner to the Hodges-Lehmann estimate, it is possible to compute P_n in linearithmic time, i.e. the worst case complexity of the algorithm is $O(n \log n)$.

One attraction of P_n is its simplicity, and we will show that it has desirable efficiency and robustness properties. Robustness, in the form of a non-zero breakdown value, is guaranteed by taking the difference of quantiles of the resulting pairwise mean distribution.

The estimator $P_n(\tau)$ will break down if at least $(1-\tau)/2 \times 100\%$ of the pairwise means are contaminated. Arbitrarily changing m of the original observations leaves $n-m$ fixed and $\binom{n-m}{2} = (n-m)(n-m-1)/2$ pairwise means remain uncontaminated. Hence, $P_n(\tau)$ will remain bounded so long as more than $(1+\tau)/2 \times 100\%$ of the pairwise means are unaffected, i.e.,

$$\frac{1}{2}(n-m)(n-m-1) > \frac{1+\tau}{2}\binom{n}{2} = \frac{1}{4}(1+\tau)n(n-1).$$

Setting $m \approx n\varepsilon^*$, for large n we have,

$$(n-n\varepsilon^*)(n-n\varepsilon^*-1) > \frac{1}{2}(1+\tau)(n^2-n).$$

Thus,

$$\varepsilon^* < 1 - \sqrt{\frac{1+\tau}{2}} + O(n^{-1}).$$

The asymptotic breakdown value of $P_n(\tau)$ is $\varepsilon^* \approx 1 - \sqrt{(1+\tau)/2}$ and it is clear that when τ decreases the breakdown value increases. At one extreme, as $\tau \to 1$, $P_n(\tau)$ converges to the range[1] and the breakdown value goes to 0. At the other extreme, as $\tau \to 0$, the breakdown point of P_n is the same as that of the Hodges-Lehmann estimate of location, which has a well known breakdown value of $\varepsilon^* \approx 0.29$.

Figure 2.2 shows the trade off between Gaussian asymptotic efficiency and breakdown value for $P_n(\tau)$ over a range of τ. In general, as τ increases the efficiency also increases but the breakdown point decreases. However, as $\tau \to 1$, i.e. as $P_n(\tau)$ approaches the range, there is a marked decrease in efficiency.

It is also important to consider the performance of $P_n(\tau)$ at heavier tailed distributions. A common heavy tailed distribution used in the robustness literature, dating back to the seminal study of Andrews et al. (1972), is the slash distribution. Let $Z \sim \mathcal{N}(0,1)$ and U be an independently distributed uniform random variable on the interval $[0,1]$ then,

$$X = \mu + \sigma \frac{Z}{U},$$

is known as a slash random variable with location parameter μ and scale parameter σ. Like the Cauchy, the slash distribution is symmetric about its median with tails decaying slowly enough such that it does not have finite mean or variance. For further details see Rogers and Tukey (1972).

Figure 2.3 shows the finite sample relative efficiency of $P_n(\tau)$ at both the standard Gaussian distribution and the standard slash distribution, where $\mu = 0$ and $\sigma = 1$. While increasing τ in finite samples leads to increasing Gaussian efficiency, it also leads to much worse performance at the slash. From this, and the results given in Section 2.5 we conclude that $\tau = 0.5$ performs well over a large range of distributions and is readily interpretable as the IQR of the pairwise means. Hence, we define the pairwise mean scale estimator P_n as,

$$P_n = P_n(0.5) = c\left[G_n^{-1}(\tfrac{3}{4}) - G_n^{-1}(\tfrac{1}{4})\right], \tag{2.19}$$

Under this definition, P_n has an asymptotic breakdown value of $\varepsilon^* \approx 0.134$.

Trimming is a common technique used to increase the robustness of non-robust estimators, see for example Stigler (1977). Wu and Zuo (2008, 2009) show that adaptive trimming of location and scale estimates, as outlined briefly in Section 2.3.2.4, increases efficiency over fixed trimming and can

[1] The range of the pairwise means is the same as the range of the original data.

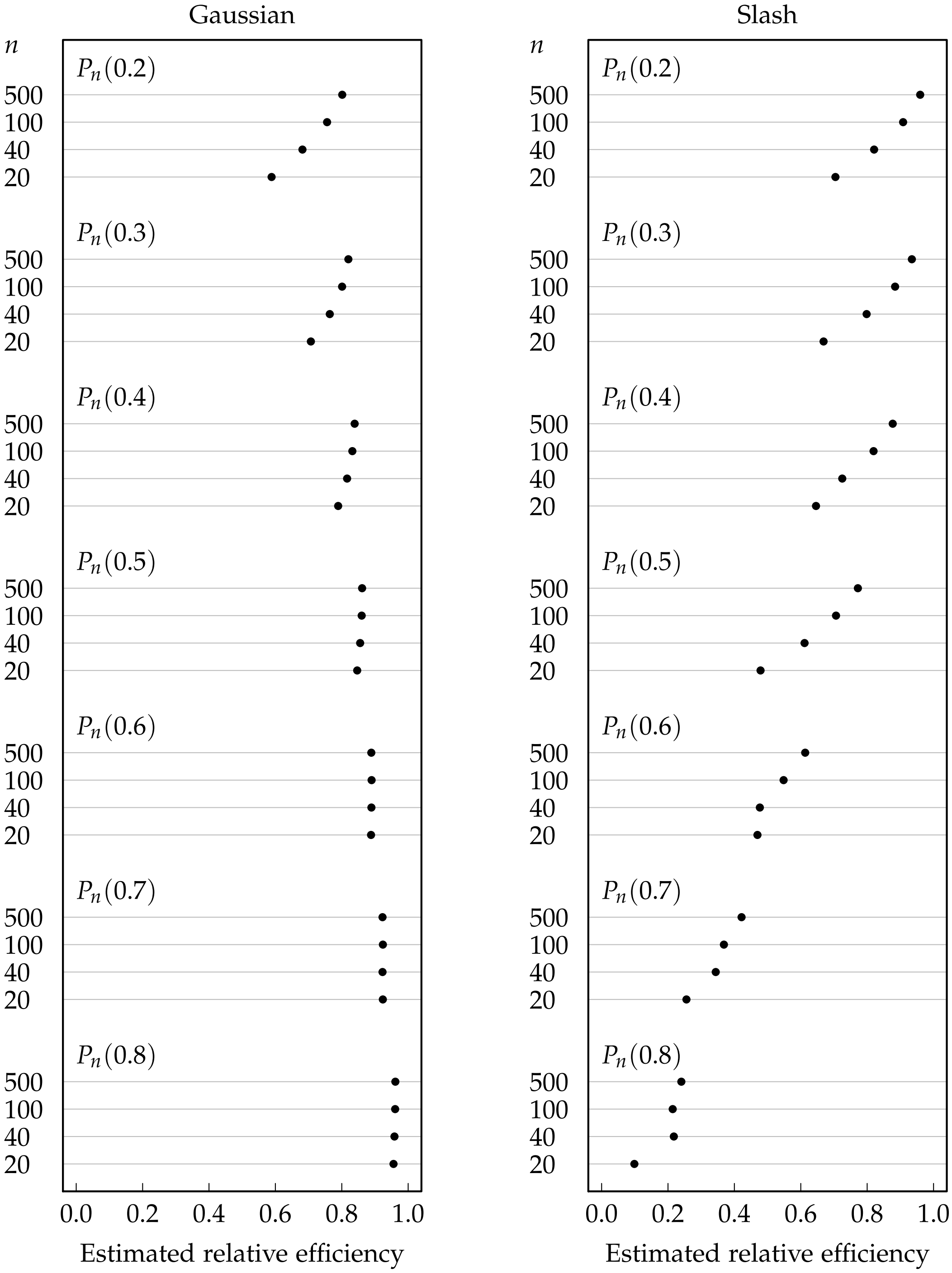


Figure 2.3: Finite sample relative efficiency of $P_n(\tau)$.

improve robustness by achieving the best possible breakdown point for a sensible choice of the tuning parameter.

The $P_n(\tau)$ estimator is inherently robust to a moderate number of outliers. When this may not be seen as robust enough, adaptive trimming may be implemented to achieve a 50% breakdown value. Furthermore, fixed trimming and Winsorising may be used to increase efficiency at extremely heavy tailed distributions such as the slash or Cauchy. Symmetrically trimming a fixed proportion of the data will only increase the breakdown value if the total proportion of the original observations trimmed is greater than two times the original breakdown value.

In the context of P_n we can trim the original data points or the pairwise means. Trimming $\gamma \times 100\%$ of the original data is equivalent to discarding $\lfloor \gamma n \rfloor$ data points. When we calculate the pairwise means from the remaining $n - \lfloor \gamma n \rfloor$ observations we have $(n - \lfloor \gamma n \rfloor)(n - \lfloor \gamma n \rfloor - 1)/2$ pairwise means. If instead trimming occurs after the pairwise means are calculated, i.e. $\alpha \times 100\%$ of the pairwise means are trimmed, $n(n-1)/2 - \lfloor \alpha n(n-1)/2 \rfloor$ pairwise means are left. Therefore, if we wish to make the proportion of pairwise means remaining comparable we need to set,

$$\alpha \approx 1 - \frac{(1-\gamma)((1-\gamma)n - 1)}{n-1},$$

which is approximately $1 - (1-\gamma)^2$ for large n.

Winsorisation may also be used on the original data points, the resulting number of pairwise means is still equal to $\binom{n}{2}$. Winsorisation yields efficiency gains similar to trimming and as such it is not considered in Section 2.5. Of course, Winsorising the pairwise means will give identical results to $P_n(\tau)$ whenever the proportion of pairwise means Winsorised is less than $(1-\tau)/2$.

Whilst trimming may aid in increasing efficiency at heavy tailed distributions, adaptive trimming of P_n, denoted by $\widetilde{P}_n$, is required to simultaneously achieve the maximal breakdown value while preserving high efficiency. Specifically, for preliminary high breakdown location and scale estimates, $m(\boldsymbol{x})$ and $s(\boldsymbol{x})$ respectively, an observation, x_i, is trimmed if,

$$\frac{|x_i - m(\boldsymbol{x})|}{s(\boldsymbol{x})} > d, \tag{2.20}$$

where d is an arbitrary constant. Note that d needs to be sufficiently large such that not all the observations are trimmed. The metric on the left hand side of (2.20) is the absolute value of the generalised scaled deviation, as defined in Wu and Zuo (2009). Simulations suggest a value of $d = 5$ represents a good trade off between achieving high efficiency at heavy tailed distributions whilst maintaining high efficiency at light tailed distributions. This is in agreement with Wu and Zuo (2009) who recommend a value for the tuning parameter of between 4 and 7.

The estimator $\widetilde{P}_n$ will inherit its breakdown value from the minimum breakdown value of the preliminary estimates,

$$\varepsilon^*(\boldsymbol{x}, \widetilde{P}_n) = \min\{\varepsilon^*(\boldsymbol{x}, m), \varepsilon^*(\boldsymbol{x}, s)\}.$$

Choosing estimators with 50% breakdown values, for example setting $m(\boldsymbol{x})$ to be the median or Huber's M-estimate of location and $s(\boldsymbol{x})$ to be the MAD or Q_n, translates to a 50% breakdown value for $\widetilde{P}_n$.

Adaptive trimming of the pairwise means is another alternative to the standard P_n statistic. Using auxiliary estimates of location and scale, each with a breakdown value of 50% to adaptively trim the kernels yields a pairwise mean scale statistic with a breakdown value of 0.29, the same as that for the Hodges-Lehmann estimate of location.

2.4.2 *Properties of P_n*

This section considers some of the properties of P_n. We begin by finding the limiting value of $P_n(\tau)$ defined in (2.18) as $n \to \infty$ and provide correction factors to ensure consistency for the standard deviation in large samples when the distribution of the underlying observations is Gaussian. We also show evidence of finite sample bias and suggest finite sample correction factors for P_n.

By exploiting the generalised L-statistic structure of $P_n(\tau)$, we find the influence function and infer related properties such as the asymptotic efficiency and gross error sensitivity for $P_n(\tau)$. We also establish the asymptotic normality of $P_n(\tau)$.

2.4.2.1 *Correction factors*

As noted in (2.18), a correction factor, c_τ is required to ensure $P_n(\tau)$ is consistent for the standard deviation in the Gaussian case. Without loss of generality, let F, the distribution of the underlying observations, be centred at zero. The CDF of the pairwise means, $(X_1 + X_2)/2$, is given by,

$$G(t) = \int_{-\infty}^{\infty} \int_{-\infty}^{2t-u} f(x) f(u)\, \mathrm{d}x\, \mathrm{d}u = \int_{-\infty}^{\infty} F(2t-u) f(u)\, \mathrm{d}u. \qquad (2.21)$$

When the underlying data follow a Gaussian distribution with CDF, Φ, and density, $\Phi' = \phi$, (2.21) can be written as,

$$G_\Phi(t) = \int_{-\infty}^{\infty} \Phi(2t - x)\phi(x)\,\mathrm{d}x.$$

The correction factor for a given $\tau \in (0,1)$ is,

$$1/c_\tau = G_\Phi^{-1}((1+\tau)/2) - G_\Phi^{-1}((1-\tau)/2). \tag{2.22}$$

The expression in (2.22) can easily be obtained using numerical integration. In particular for P_n, the corresponding asymptotic correction factor is, $c_{0.5} = c \approx 1/0.9539 \approx 1.0483$.

The finite sample correction factors are applied after the large sample correction has been made. Finite sample correction factors for P_n, $c_{n,0.5}$, have been found analytically for $n = 3$ and 4: $c_{3,0.5} = 1.13$ and $c_{4,0.5} = 1.30$ (see Section B.1 in Appendix B for details). The finite sample correction factors exhibit jumps and are not monotonically increasing in n, instead they exhibit a periodic pattern attributable to the order statistic method used to find the quantiles. For $5 \leq n < 40$, finite sample correction factors have been found using simulation, a sample of these is given in Table B.1 in Appendix B. For $n \geq 40$ the small sample correction factor for P_n is well approximated by $c_{n,0.5} = n/(n - 0.7)$.

2.4.2.2 *Influence function*

Noting that $P_n(\tau)$ is a GL-statistic and using equation (2.15) from Serfling (1984) we have the following result.

Result 2.1. *If F has derivative $f > 0$ on $[F^{-1}(\epsilon), F^{-1}(1-\epsilon)]$ for all $\epsilon > 0$, the influence function for $P_n(\tau)$ is,*

$$\mathrm{IF}(x; P_n(\tau), F) = c_\tau \left[\frac{(1+\tau)/2 - F(2G^{-1}((1+\tau)/2) - x)}{\int f(2G^{-1}((1+\tau)/2) - x) f(x)\,\mathrm{d}x} - \frac{(1-\tau)/2 - F(2G^{-1}((1-\tau)/2) - x)}{\int f(2G^{-1}((1-\tau)/2) - x) f(x)\,\mathrm{d}x} \right].$$

Figure 2.4 plots the influence functions for P_n, Q_n, the MAD and the SD when the underlying data are Gaussian. Figure 2.4 shows that P_n has a gross error sensitivity of $\gamma^* = 2.33$, slightly higher than that of Q_n, $\gamma^* = 2.07$ at the Gaussian distribution. As Hampel (1974) notes, the asymptotic variance of an estimator approaches its minimum as the influence function approaches a multiple of the log likelihood derivative. Hence, when the underlying observations are Gaussian, the closer an estimator's influence function

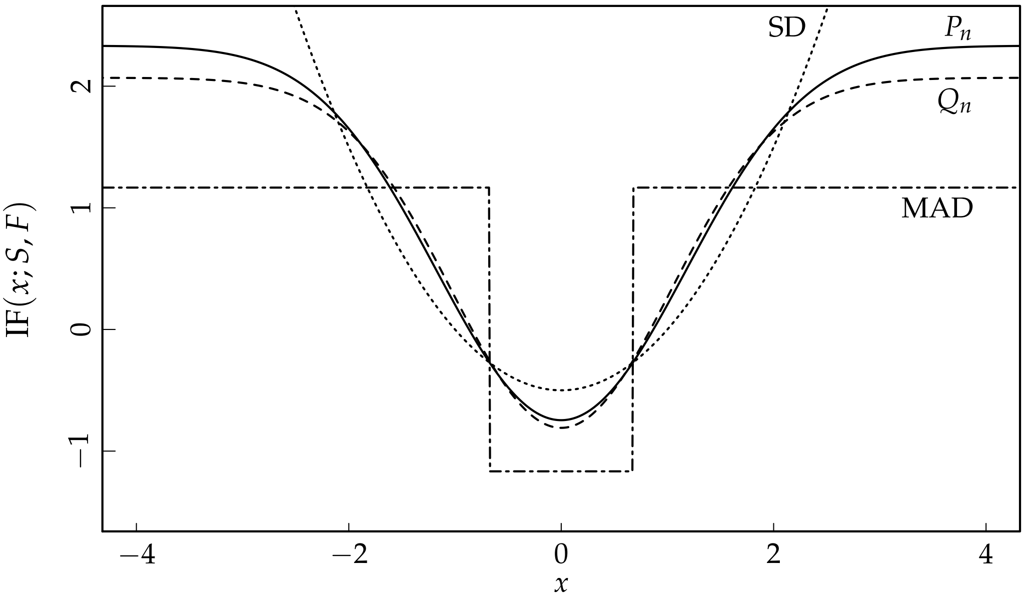


Figure 2.4: Influence functions of the SD, Q_n, P_n and the MAD when the model distribution is Gaussian.

is to that of the SD, the more efficient it will be. In Figure 2.4 the influence function of P_n is almost always closer to that of the SD than Q_n. This is reflected in our calculation of the asymptotic variance, which is found as the expected square of the influence function. Numerical integration yields,

$$V = \int_{\mathbb{R}} \mathrm{IF}^2(x; P_n, \Phi)\, \mathrm{d}\Phi(x) = 0.579. \tag{2.23}$$

This equates to an asymptotic efficiency of 0.86 as compared with 0.82 for Q_n and 0.37 for the MAD at the Gaussian distribution. Thus, despite having a higher gross error sensitivity, P_n is a more efficient estimator than Q_n at the Gaussian distribution.

It is also important to consider efficiencies at distributions other than the Gaussian. If Y follows a Student's t distribution, it can be generalised as a location-scale model, $X = \mu + \sigma Y$, with density,

$$f(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}\sigma}\left(1 + \frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right)^{-(\nu+1)/2},$$

where $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}\,\mathrm{d}x$ is the gamma function. It is important to note that in this setting σ is a scale parameter, not the standard deviation, so it still exists and can be consistently estimated even when $1 \leq \nu < 2$. Setting $\mu = 0$, Bachmaier (2000) adapts the work of Fisher (1922) to show that the Fisher information for the scale parameter of a scaled t distribution is,

$$\mathcal{I}(\sigma) = \frac{2\nu}{(\nu+3)\sigma^2}.$$

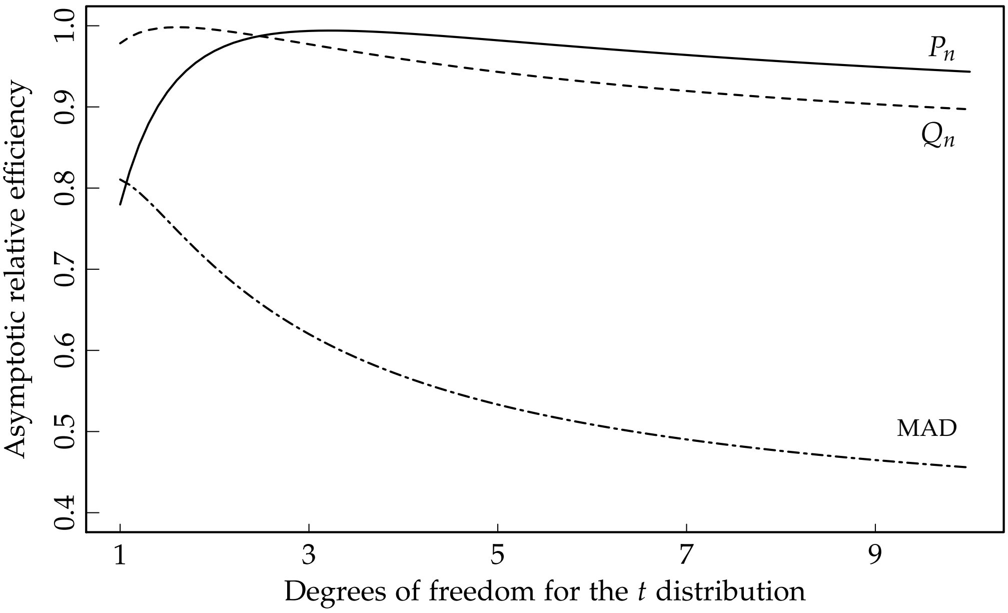


Figure 2.5: Asymptotic relative efficiency of Q_n, P_n and the MAD for t distributions with degrees of freedom ranging between 1 and 10.

Figure 2.5 shows the relative efficiency of P_n, Q_n and the MAD at t distributions with degrees of freedom ranging between 1 (the Cauchy distribution) and 10 with scale parameter $\sigma = 1$, i.e. standardised t distributions. The asymptotic relative efficiencies are calculated as the product of the inverse of the Fisher information and the asymptotic variance, calculated by replacing Φ in (2.23) with the CDF of the corresponding t distribution. It is clear from Figure 2.5 that t distributions with degrees of freedom more than approximately 2.5, P_n is more efficient than Q_n. At extremely heavy tailed t distributions, including the Cauchy distribution, Q_n is asymptotically more efficient than P_n. The MAD also performs better than P_n at the Cauchy, however its efficiency decays substantially as the degrees of freedom increase.

At the exponential distribution the asymptotic relative efficiency of P_n to Q_n is 0.77. Furthermore, the gross error sensitivity for P_n is $\gamma^* = 3.968$ which compares with $\gamma^* = 2.317$ for Q_n at the exponential. However, it is not the case that P_n is necessarily worse than Q_n for skewed distributions. When the underlying distribution of the data is χ^2_1, the asymptotic relative efficiency of P_n to Q_n is 1.43.

The performance of P_n is more attractive than Q_n for discrete distributions. In the limit, Q_n will equal zero, and therefore fail to provide a valid estimate of scale, whenever more than 25% of the pairwise differences equal zero. For a discrete distribution with k distinct possible outcomes, $x_1, x_2, \ldots, x_k$, and probability mass function $\mathrm{P}(X = x_j) = p_j$, for $j = 1, 2, \ldots, k$, the expected proportion of pairwise differences equal to zero is $\sum_j p_j^2$. In particular,

$$\sum_j p_j^2 > \tfrac{1}{4} \iff \lim_{n\to\infty} \mathrm{P}(Q_n = 0) = 1.$$

For example for the Poisson distribution with expected value 1, $\sum_j p_j^2 = 0.31$ and hence $Q_n \xrightarrow{p} 0$. In contrast, for P_n to return a scale estimate of zero, the IQR of the pairwise means must be equal to zero, that is, more than 50% of the pairwise means must be equal. Except for trivial two point distributions, we have shown that $P_n = 0$ implies $Q_n = 0$ (see Section B.2 in Appendix B for details). The pairwise averaging process helps smooth the underlying discrete distribution which results in P_n being a better robust estimator than Q_n in these situations.

2.4.2.3 *Asymptotic normality*

As defined in Section 2.2, let $G(t) = \mathrm{P}(g(X_1, X_2) \leq t)$ and $G_n(t)$ be the CDF and empirical distribution function of a symmetric, bivariate U-statistic. This section proves the following result on the asymptotic normality of $P_n(\tau)$.

Result 2.2. *Let $a = (1-\tau)/2$, $b = (1+\tau)/2$ for $0 < \tau < 1$. If G has derivative $G' > 0$ on $[G^{-1}(b) - \varepsilon, G^{-1}(a) + \epsilon]$ for some $\epsilon > 0$, then as $n \to \infty$,*

$$\sqrt{n}(P_n(\tau) - \theta) \xrightarrow{\mathcal{D}} \mathcal{N}(0, c_\tau^2 V),$$

where $\theta = c_\tau \left(G^{-1}(b) - G^{-1}(a)\right)$, $V = v(a,a) + v(b,b) - 2v(a,b)$, and

$$v(a,b) = 4\frac{\int F(2G^{-1}(a) - x)F(2G^{-1}(b) - x)\,\mathrm{d}F(x) - ab}{G'(G^{-1}(a))G'(G^{-1}(b))}.$$

To establish the limiting distribution, we note that P_n is a linear combination of two U-quantile statistics. Hence, we first consider the empirical U-process defined as,

$$\left(\sqrt{n}\,(G_n(t) - G(t))\right)_{t\in\mathbb{R}}.$$

Silverman (1976, Theorem B) proves that in this context, $\sqrt{n}(G_n(\cdot) - G(\cdot))$ converges weakly in the Skorohod topology to an almost surely continuous, zero-mean Gaussian process, W, with covariance function,

$$\mathbb{E}W(s)W(t) = 4\,\mathrm{P}(g(X_1, X_2) \leq s, g(X_1, X_3) \leq t) - 4G(s)G(t), \qquad (2.24)$$

for all $s, t \in \mathbb{R}$.

Hence, for $0 < p < q < 1$, if G', the derivative of G, is strictly positive on the interval $[G^{-1}(p) - \varepsilon, G^{-1}(q) + \epsilon]$ for some $\epsilon > 0$, then we can use the inverse map to show,

$$\sqrt{n}\left(G_n^{-1}(\cdot) - G^{-1}(\cdot)\right) \xrightarrow{\mathcal{D}} \frac{W(G^{-1}(\cdot))}{G'(G^{-1}(\cdot))},$$

where W is a mean zero Gaussian process with covariance function defined in (2.24). See for example, van der Vaart and Wellner (1996, Section 3.9.4.2). Serfling (2002) summarises stronger results concerning empirical processes of U-statistic structure.

For $P_n(\tau)$ we use the pairwise mean kernel, $g(x,y) = (x+y)/2$. Conditioning on $X_1 = x$ the covariance function (2.24) can be rewritten as,

$$\operatorname{cov}(W(t), W(s)) = 4\int F(2s-x)F(2t-x)\,\mathrm{d}F(x) - 4G(s)G(t),$$

where F is the distribution function of the underlying observations.

The limiting covariance function for the quantiles used in the construction of $P_n(\tau)$, $n\operatorname{cov}(G_n^{-1}(a), G_n^{-1}(b))$, is,

$$v(a,b) = 4\frac{\int F(2G^{-1}(a)-x)F(2G^{-1}(b)-x)\,\mathrm{d}F(x) - ab}{G'(G^{-1}(a))G'(G^{-1}(b))},$$

for $a = (1-\tau)/2$, $b = (1+\tau)/2$ and $0 < \tau < 1$. Hence,

$$\sqrt{n}(P_n(\tau) - \theta) \xrightarrow{\mathcal{D}} \mathcal{N}(0, c_\tau^2 V),$$

where $\theta = c_\tau\left(G^{-1}(b) - G^{-1}(a)\right)$ and $V = v(a,a) + v(b,b) - 2v(a,b)$ both depend on F.

Noting that the derivative of $G(t)$ is,

$$G'(t) = \int 2f(2t-u)f(u)\,\mathrm{d}u, \tag{2.25}$$

it can be shown that the asymptotic variance found here is equivalent to the expected square of the influence function discussed previously.

Furthermore, from the general results in Serfling (2002, Section 12.3.4), the almost sure behaviour of $P_n(\tau)$ can be deduced from the Bahadur representation for the U-quantiles.

2.5 RELATIVE EFFICIENCIES IN FINITE SAMPLES

The asymptotic variance of P_n has been found and corresponding asymptotic efficiencies have been deduced in the previous section. In particular, it has been shown that, asymptotically, P_n is more efficient than Q_n for t distributions with more than approximately 2.5 degrees of freedom. However, it is often of practical interest to determine small sample relative efficiencies. This section considers the finite sample relative efficiency of P_n, and variants thereof, using configural polysampling. It also examines the finite relative efficiencies over the same range of t distributions considered asymptotically in Figure 2.5.

2.5.1 *Design*

Under configural polysampling, estimators are evaluated at particular distributions chosen to exhibit more extreme characteristics than what might be observed in practice. For example, the Gaussian distribution has tails that die off rapidly, and the Cauchy has tails that die off extremely slowly. Such *extreme* distributions will be referred to as *corners*. If it can be shown that an estimator performs well over all the corners considered, it is a fair assumption that the estimator will perform at least as well at intermediate distributions.

A key performance measure for estimators is the polyefficiency, the minimum efficiency that an estimator achieves over a selection of corners. Yatracos (1991) shows that high polyefficiency over a finite selection of corners implies at least as high an efficiency at any convex combination of these corners.

Scale estimates are computed for samples of size n from each of Tukey's three corners: the Gaussian corner where observations are sampled independently from a standard Gaussian distribution; the one wild corner, where $n-1$ observations are independent Gaussian and the remaining observation is scaled by a factor of 10; and the slash corner where observations are constructed as the ratio of an independent Gaussian random variable and an independent standard uniform random variable. The slash distribution has Cauchy-like tails, but is considered to be more generally representative of real data as it is less peaked at the median than the Cauchy.

Improving on the methodology set out in Lax (1985), Randal (2008) proposes using MLEs as the benchmark against which all estimators are compared. Previously, efficiencies were typically measured relative to the most efficient estimator considered for each distribution. MLEs are asymptotically efficient and may be used as a common reference case in future simulation studies.

Kafadar (1982) gives the MLE for the slash distribution and Randal and Thomson (2004) describe how to find the MLE for the one wild distribution by nesting it within the class of Gaussian compound scale models and applying an EM algorithm. Randal (2008) demonstrates how to use a similar EM algorithm approach to find the MLE of scale for t distributions. These algorithms are not part of an R package, however, they are reasonably straightforward to implement and we include our R code in Appendix B.4.

Efficiencies are estimated over m independent samples as,

$$\widehat{\text{eff}}(T) = \frac{\widehat{\text{var}}\,(\ln\hat{\sigma}_1, \ldots, \ln\hat{\sigma}_m)}{\widehat{\text{var}}\,(\ln S(\boldsymbol{x}_1), \ldots, \ln S(\boldsymbol{x}_m))}, \tag{2.26}$$

where for each $j = 1, \ldots, m$, the $\boldsymbol{x}_j$ are independent samples of size n, $\hat{\sigma}_j$ is the scale MLE and $S(\boldsymbol{x}_j)$ is the proposed scale estimate.

Rousseeuw and Croux (1993) propose an alternative efficiency measure, based on the standardised variances. This measure was also considered and gave results largely in agreement with the Lax relative efficiencies and are therefore not reported here. The key difference between these two measures of efficiency is the presence of a log transformation in (2.26) which acts to stabilise the variance estimates. This is not present in the alternative measure which therefore heavily penalises inefficient scale estimates in heavy tailed data. Correction factors do not play a role in determining efficiency in either measure. Therefore, so far as efficiency is concerned, it is not an issue that the estimators are defined with consistency factors for the standard Gaussian only.

2.5.2 *Results*

Figure 2.6 presents the estimated relative efficiencies over t distributions with degrees of freedom ranging from 1 to 10 in increments of 0.5. The curves for P_n and Q_n are similar to the asymptotic efficiencies in Figure 2.5. Of particular interest is the speed with which P_n overtakes the adaptively trimmed P_n. For t distributions with 2 or more degrees of freedom, P_n is more efficient than $\widetilde{P}_n$, the adaptively trimmed form of P_n with tuning parameter $d = 5$. Furthermore, for t distributions with 3 or more degrees of freedom, P_n is more efficient than Q_n.

For skewed distributions, we obtain similar results in finite samples to what was observed asymptotically. In particular, the results are highly dependent on the type of skewed distribution. At the exponential, in samples of size $n = 20$, P_n is 0.87 times as efficient as Q_n. Which is an improvement

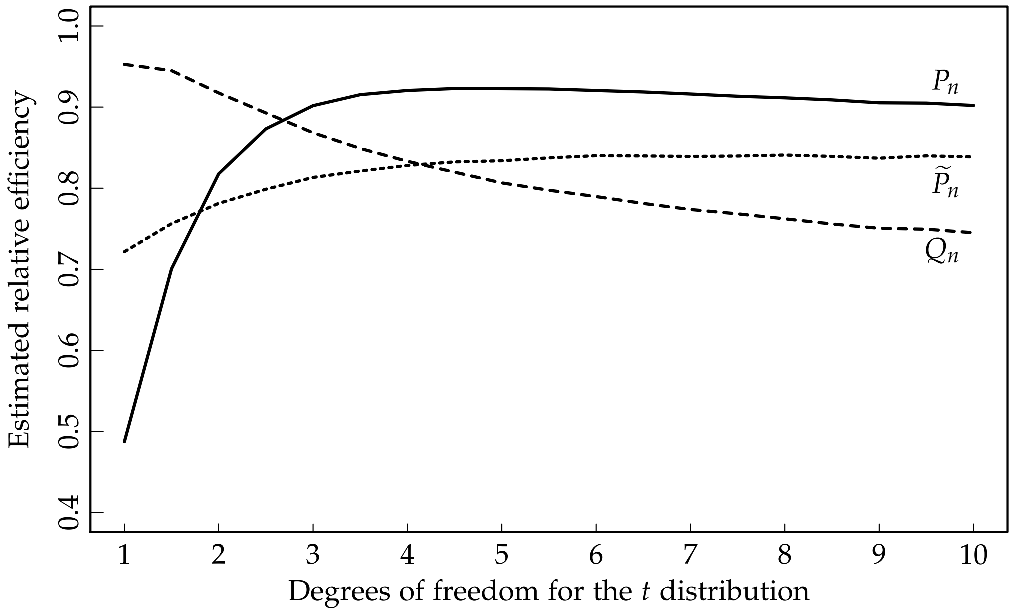


Figure 2.6: Finite sample relative efficiency, estimated from one million t distributed samples of size $n = 20$, for Q_n, P_n and adaptively trimmed P_n with tuning parameter $d = 5$, $\widetilde{P}_n$.

over the asymptotic result. Whereas at the χ^2_1 distribution, in samples of size $n = 20$, P_n is 1.21 times more efficient than Q_n.

Figure 2.7 presents the relative efficiencies of a variety of scale estimators at Tukey's three corner distributions in samples of size $n = 20$. The estimators considered are: P_n; $\widetilde{P}_n$, the adaptively trimmed P_n with tuning parameter $d = 5$; $\breve{P}_n$ is symmetric fixed trimming where 5% of the observations have been trimmed off both tails of the original data, for $n = 20$ this means that the maximum and minimum values have been deleted; Q_n and S_n (Rousseeuw and Croux, 1993); the MAD; and the IQR. The estimator P_n is observed to be the most efficient at the Gaussian and one wild corners, however, it performs poorly at the slash corner. Among the estimators considered, $\widetilde{P}_n$ with tuning parameter $d = 5$ has highest minimum efficiency across all three corners, therefore the estimator is known to be triefficient. Its lowest efficiency occurs at the one wild with 72%. However as noted in Figure 2.6, the efficiency gain from using $\widetilde{P}_n$ over P_n disappears as the tails become slightly less heavy.

The triefficiencies over various sample sizes are reported in Table 2.2. The adaptively trimmed P_n with $d = 5$ is in fact triefficient over the range of sample sizes considered. While the triefficiencies of most estimators increases with sample size, those of the MAD and IQR do not, reflecting their exceptionally poor finite sample and asymptotic Gaussian efficiency.

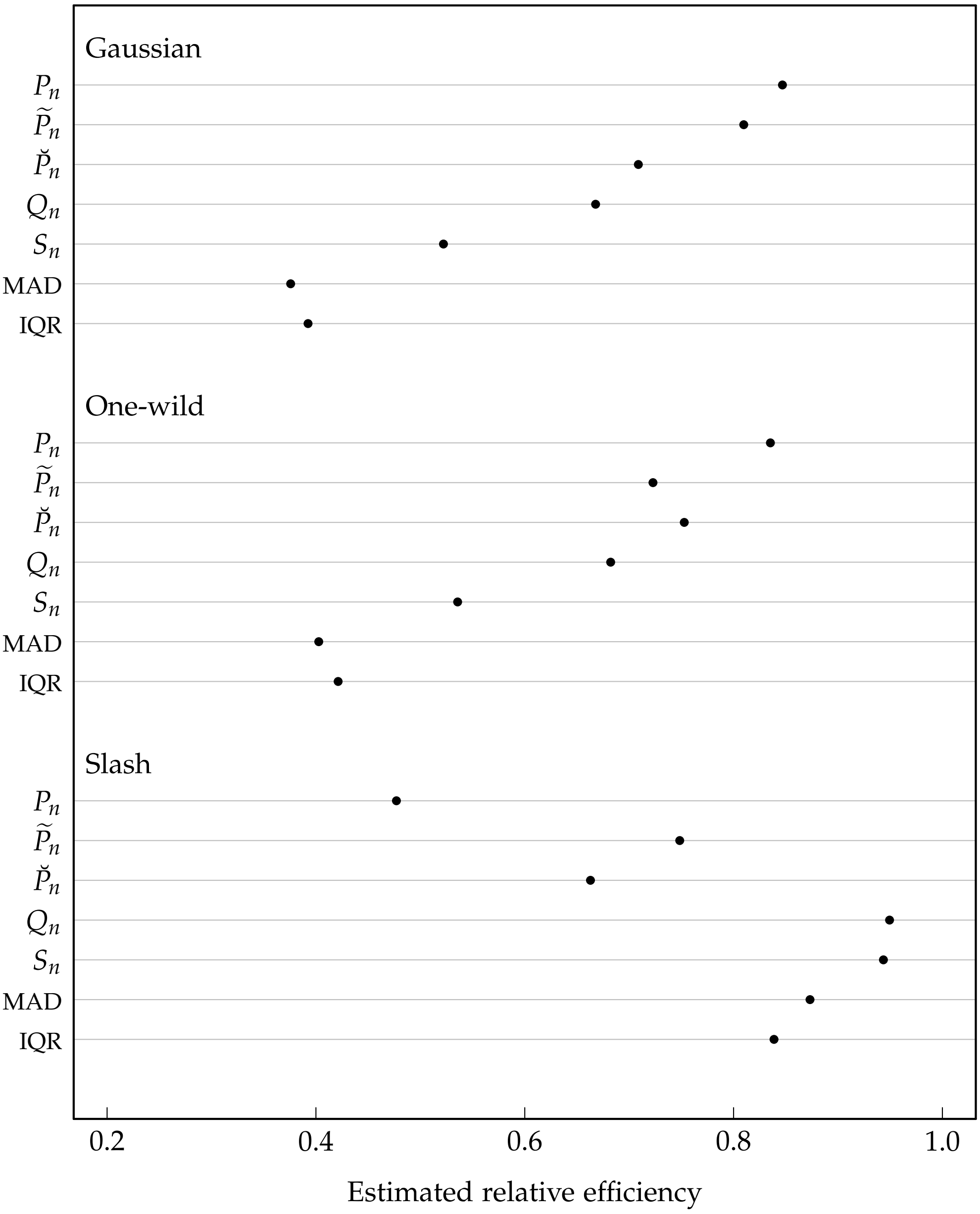


Figure 2.7: Estimated efficiencies relative to maximum likelihood scale estimates for samples of size $n = 20$, calculated over one million independent samples. The various pairwise mean scale estimators are: P_n; $\widetilde{P}_n$, the adaptively trimmed P_n with tuning parameter $d = 5$; and $\check{P}_n$ is symmetric fixed trimming where 5% of the observations have been trimmed off both tails of the original data.

Gaussian	n				
	20	50	100	500	1000
P_n	0.853	0.854	0.862	0.860	0.858
$\widetilde{P}_n$ with $d = 5$	0.819	0.850	0.861	0.860	0.858
$\widetilde{P}_n$ with $d = 3$	0.566	0.678	0.725	0.753	0.757
$\breve{P}_n$	0.718	0.747	0.730	0.726	0.724
Q_n	0.677	0.737	0.778	0.809	0.812
MAD	0.376	0.368	0.374	0.365	0.363
IQR	0.396	0.375	0.376	0.366	0.362

One wild	n				
	20	50	100	500	1000
P_n	0.834	0.864	0.860	0.870	0.867
$\widetilde{P}_n$ with $d = 5$	0.723	0.817	0.839	0.865	0.865
$\widetilde{P}_n$ with $d = 3$	0.574	0.692	0.725	0.763	0.765
$\breve{P}_n$	0.756	0.777	0.738	0.737	0.733
Q_n	0.680	0.751	0.781	0.817	0.820
MAD	0.400	0.380	0.380	0.373	0.372
IQR	0.420	0.386	0.380	0.373	0.372

Slash	n				
	20	50	100	500	1000
P_n	0.480	0.626	0.703	0.773	0.786
$\widetilde{P}_n$ with $d = 5$	0.748	0.785	0.809	0.833	0.829
$\widetilde{P}_n$ with $d = 3$	0.766	0.799	0.809	0.822	0.826
$\breve{P}_n$	0.665	0.738	0.820	0.876	0.886
Q_n	0.945	0.950	0.959	0.961	0.963
MAD	0.876	0.851	0.847	0.837	0.844
IQR	0.839	0.830	0.839	0.839	0.843

Table 2.1: Estimated relative efficiencies relative to maximum likelihood scale estimates for the various scale estimators at each of Tukey's three corners calculated over one million independent samples. Note that $\breve{P}_n$ is symmetric fixed trimming where 5% of the observations have been trimmed off both tails of the original data.

n	20	50	100	500	1000
P_n	0.480	0.626	0.703	0.773	0.786
$\widetilde{P}_n$ with $d = 5$	0.723	0.785	0.809	0.833	0.829
$\widetilde{P}_n$ with $d = 3$	0.566	0.678	0.725	0.753	0.757
$\breve{P}_n$	0.665	0.738	0.730	0.726	0.724
Q_n	0.677	0.737	0.778	0.809	0.812
MAD	0.376	0.368	0.374	0.365	0.363
IQR	0.396	0.375	0.376	0.366	0.362

Table 2.2: Triefficiencies of various scale estimators over a range of sample sizes.

Moderate amounts of fixed trimming, for example 5% at each tail as shown in Figure 2.7 does markedly improve the efficiency of P_n at the slash corner. However, it also compromises the efficiency of P_n at the Gaussian distribution. In other simulations we have conducted, this result is not noticeably improved as the sample size increases. Therefore, adaptive trimming with a sensible choice of the tuning parameter is recommended over fixed trimming.

We have considered the impact that the tuning parameter d has on the efficiency of $\widetilde{P}_n$. Intuitively, as d increases, more observations remain in the sample which, for distributions that are not extremely heavy tailed, tends to increase efficiency. However, if the data are sampled from an extremely heavy tailed distribution, in small samples the opposite tends be true – removing more of the observations in the tails actually increases the efficiency of the scale estimate. Table 2.1 shows that in samples of size $n = 20$ using an adaptive trimming parameter of $d = 5$ results in a Gaussian relative efficiency of 0.82, however using $d = 3$ results in a much lower Gaussian relative efficiency of 0.57. In contrast at the slash corner, $d = 5$ gives a slightly lower relative efficiency of 0.75 compared with the $d = 3$ of 0.77. As the sample size grows, the distinction between $d = 5$ and $d = 3$ is negligible at the slash corner, though at the Gaussian distribution using $d = 5$ results in a much more efficient estimator.

From additional results not summarised in Figure 2.7 or Table 2.1, we find that random trimming of the kernels, i.e. the pairwise means, achieves similar results as randomly trimming the original data at the Gaussian and one wild corners, however, randomly trimming the original data leads to higher efficiency at the slash.

The small sample performance of P_n is also better than the performance of Q_n for discrete distributions. Consider, for example, the case of a binomial distribution, $X \sim \mathcal{B}(6, 0.4)$. In the limit neither Q_n nor P_n will converge to zero. However, if samples of size $n = 20$ are drawn from this distribution, Q_n will return a scale estimate of zero, on average 12% of the time due to the discrete nature of the data. In contrast, P_n returns zero less than 0.1% of the time. Apart from some trivial cases, if $P_n = 0$ in finite samples this implies that $Q_n = 0$ (see Section B.2 in Appendix B for details).

Importantly, P_n possesses comparable small sample efficiency to that obtained asymptotically. It performs particularly well at the Gaussian and one wild as well as heavy tailed distributions such as the t distribution with degrees of freedom greater than approximately 2.5. Even though $\widetilde{P}_n$ with a trimming parameter of $d = 5$ is triefficient amongst the scale estimators considered here, P_n would still be preferred as a scale estimator in small samples.

As the sample size increases the efficiency of $\widetilde{P}_n$ with trimming parameter $d = 5$ approaches that of P_n at the Gaussian distribution and $\widetilde{P}_n$ remains quite efficient at the slash, indicating that a mild amount of adaptive trimming could be a worthwhile thing to do. The utility of an adaptively trimmed version of P_n is revisited in Chapter 4.

2.6 CONCLUSION

This chapter details the scale estimator, based on the difference of two order statistics of the empirical distribution function of the pairwise means. Hence $P_n(\tau)$ fits into the GL-statistic framework, alongside many other well known scale estimators. Choosing $\tau = 0.5$ results in the estimator P_n which possesses reasonable robustness properties, whilst maintaining high efficiencies at Gaussian distributions and has an intuitive interpretation: the IQR of the pairwise means.

We have found that P_n, in its standard form, has a breakdown value of 13%. Its influence function more closely approximates that of the SD at the Gaussian distribution, leading to a relatively high asymptotic efficiency of 86%. We have also found the gross error sensitivity and demonstrated the asymptotic normality of P_n. Furthermore, when the underlying distribution is discrete, P_n is more robust to repeated observations than Q_n.

In finite samples, P_n also performs admirably. In samples of size $n = 20$, at the Gaussian distribution, P_n is 27% more efficient than Q_n and 22% more efficient at the one wild corner. P_n maintains its efficiency advantage over Q_n even when the underlying distribution of the data has quite heavy tails.

To summarise, P_n is a simple and intuitive robust scale estimator, that is not tailored to be most efficient at any particular distribution, rather it maintains high efficiency over a wide range of distributions.

3 COVARIANCE AND AUTOCOVARIANCE ESTIMATION

3.1 INTRODUCTION

Motivated by the potential to extend the efficiency and robustness properties of P_n to the time series setting, this chapter considers the problem of estimating scale and autocovariance in dependent processes. We begin by considering a general device commonly used to extend scale estimators to the covariance setting, before outlining the properties of the resulting covariance estimator when P_n is used as the underlying scale estimator. Standard results are used to show that the asymptotic efficiency and robustness properties carry through and simulation results are provided for the robust correlation coefficient. We also discuss some of the implications for robustness when the covariance estimator is used in a time series setting as an autocovariance estimator.

The properties of the scale estimator P_n and the resulting autocovariance estimator under both short and long range dependent Gaussian processes are discussed. Standard asymptotic results are found to hold in the short range dependent (SRD) setting. In the long range dependent (LRD) setting we establish the asymptotic normality of P_n under short and mildly long range dependent Gaussian processes. Under extreme long range dependence, we motivate a complication in establishing the limiting distribution for P_n by first proving a non-Gaussian limit result for the IQR, consistent with results for other common scale estimators, such as the SD and Q_n. In contrast with the results of Lévy-Leduc et al. (2011c) for a single U-quantile, namely Q_n, the proof for the IQR, a difference of two quantiles, relies on the higher order terms in the Bahadur representation of Wu (2005). Furthermore, we posit that P_n, the IQR of the pairwise means, has a similar behaviour, though the proof relies on an analogous conjectured Bahadur representation for U-quantiles under long range dependence. The asymptotic results for the robust scale estimator extend to the corresponding robust autocovariance estimators. Our theoretical results are illustrated with simulations.

Most of the literature concerning robust covariance estimation is focussed on the multivariate setting, i.e. estimating covariance or dispersion matrices. This topic will be addressed in Chapter 4.

3.2 ROBUST COVARIANCE ESTIMATION

A simple method for turning scale estimators into covariance estimators was introduced by Gnanadesikan and Kettenring (1972) and brought to prominence in the context of robust estimation by Ma and Genton (2000, 2001). The idea is based on the identity,

$$\operatorname{cov}(X,Y) = \frac{1}{4\alpha\beta}\left[\operatorname{var}(\alpha X + \beta Y) - \operatorname{var}(\alpha X - \beta Y)\right], \tag{3.1}$$

where X and Y are random variables. In general, X and Y may have different scales, hence it is standard to let $\alpha = 1/\sqrt{\operatorname{var}(X)}$ and $\beta = 1/\sqrt{\operatorname{var}(Y)}$. A robust covariance estimator is found by replacing the variance in (3.1) with (squared) robust scale estimators. A nice feature of (3.1) is that it is location free, in the sense that no location parameter needs to be estimated when the robust scale estimator is also location free, such as Q_n, IQR or P_n.

The covariance estimator based on P_n is obtained by replacing the variance terms in (3.1) with the square of P_n,

$$\gamma_P(X,Y) = \frac{1}{4\alpha\beta}\left[P_n^2(\alpha X + \beta Y) - P_n^2(\alpha X - \beta Y)\right], \tag{3.2}$$

where $\alpha = 1/P_n(X)$ and $\beta = 1/P_n(Y)$. The correlation could also be estimated as,

$$\frac{P_n^2(\alpha X + \beta Y) - P_n^2(\alpha X - \beta Y)}{4\alpha\beta P_n(X)P_n(Y)},$$

however this estimator does not necessarily satisfy the Cauchy-Schwarz inequality and so the estimated correlation coefficient would not necessarily lie in the range $[-1,1]$. To ensure that the estimated correlation coefficient lies in the appropriate range, Gnanadesikan and Kettenring (1972) propose using an estimator of the form,

$$\rho_P(X,Y) = \frac{P_n^2(\alpha X + \beta Y) - P_n^2(\alpha X - \beta Y)}{P_n^2(\alpha X + \beta Y) + P_n^2(\alpha X - \beta Y)}.$$

While the covariance given by (3.2) depends on the correction factor built into the P_n estimator to ensure consistency at the Gaussian distribution, see equation (2.19), the correlation is independent of the choice of correction factor.

An advantage of covariance estimates based on the Gnanadesikan and Kettenring device, applying equation (3.1), is that the statistical properties of the resulting estimator are relatively straightforward to derive.

3.2.1 *Properties*

Genton and Ma (1999) explore the robustness properties of dispersion estimators constructed using the identity (3.1). In the general case, let $S_n(F_n)$ be a scale estimator with corresponding statistical functional $S(F)$. Let γ_S be a statistical functional of covariance corresponding to a covariance estimate $\hat{\gamma}_S$ based on (3.1),

$$\gamma_S(\boldsymbol{F}) = \frac{1}{4\alpha\beta}\left[S^2(F_+) - S^2(F_-)\right],$$

where $\boldsymbol{F}$ is a bivariate distribution with marginal distributions F_X and F_Y, and $F_\pm$ denotes the distribution of $\alpha X \pm \beta Y$ with scale $\sigma_\pm$ respectively.

3.2.1.1 *Influence function*

Genton and Ma (1999) show that the influence function of γ_S can be defined in terms of the influence function of S as follows,

$$\begin{aligned}\mathrm{IF}(x,y;\gamma_S,\boldsymbol{F}) = \frac{1}{2\alpha\beta}\big[&S(F_+)\,\mathrm{IF}(\alpha x + \beta y; S, F_+) \\ &- S(F_-)\,\mathrm{IF}(\alpha x - \beta y; S, F_-)\big]. \qquad (3.3)\end{aligned}$$

Note that this formulation of a bivariate influence function in terms of two univariate influence functions requires the unidimensional Dirac function δ_x be generalised to a bidimensional Dirac function where the perturbations depend on the choice of the covariance estimator, i.e. along the $\alpha x + \beta y$ and $\alpha x - \beta y$ directions. Since α and β are arbitrary non-zero constants, we can simply set $\alpha = 1/\sigma_X$ and $\beta = 1/\sigma_Y$, where $\sigma_X = \sqrt{\text{var(X)}}$ and $\sigma_Y = \sqrt{\text{var}(Y)}$, in which case (3.3) can be rewritten as,

$$\begin{aligned}\mathrm{IF}(x,y;\gamma_S,\boldsymbol{F}) = \frac{\sigma_X\sigma_Y}{2}\Bigg[&\sigma_+\,\mathrm{IF}\left(\frac{x}{\sigma_X} + \frac{y}{\sigma_Y}; S, F_+\right) \\ &- \sigma_-\,\mathrm{IF}\left(\frac{x}{\sigma_X} - \frac{y}{\sigma_Y}; S, F_-\right)\Bigg]. \qquad (3.4)\end{aligned}$$

Ma and Genton (2001) use the general form, (3.4), to state the influence function for covariance estimators based on Q_n,

$$\mathrm{IF}(x,y;\gamma_Q,\boldsymbol{F}) = \frac{\sigma_X\sigma_Y}{2}\left[\sigma_+\,\mathrm{IF}\left(\frac{x}{\sigma_X} + \frac{y}{\sigma_Y}; Q, F_+\right) - \sigma_-\,\mathrm{IF}\left(\frac{x}{\sigma_X} - \frac{y}{\sigma_Y}; Q, F_-\right)\right],$$

where $F_{\pm}$ is the distribution function of $\frac{X}{\sigma_X} \pm \frac{Y}{\sigma_Y}$ respectively, and the influence function for Q_n is given in (2.11).

We can similarly use the general form, (3.4), to find the influence functions for the covariance estimator based on P_n,

$$\mathrm{IF}(x,y;\gamma_P,\boldsymbol{F}) = \frac{\sigma_X\sigma_Y}{2}\left[\sigma_+\,\mathrm{IF}\left(\frac{x}{\sigma_X}+\frac{y}{\sigma_Y};P,F_+\right) - \sigma_-\,\mathrm{IF}\left(\frac{x}{\sigma_X}-\frac{y}{\sigma_Y};P,F_-\right)\right].$$

The influence function for $P_n(\tau)$ has been found previously in Result 2.1. For $P_n = P_n(0.5)$ this simplifies to,

$$\mathrm{IF}(x;P,F) = c\left[\frac{\frac{3}{4} - F(2G^{-1}(\frac{3}{4})-x)}{\int f(2G^{-1}(\frac{3}{4})-x)f(x)dx} - \frac{\frac{1}{4} - F(2G^{-1}(\frac{1}{4})-x)}{\int f(2G^{-1}(\frac{1}{4})-x)f(x)dx}\right].$$

We will focus specifically on Gaussian processes, in which case,

$$\begin{aligned}\int \phi(2G_{\Phi}^{-1}(\tfrac{3}{4})-x)\phi(x)dx &= \int \phi(2G_{\Phi}^{-1}(\tfrac{1}{4})-x)\phi(x)dx \\ &= \phi(\sqrt{2}G_{\Phi}^{-1}(\tfrac{3}{4}))/\sqrt{2}.\end{aligned}$$

Hence we have,

$$\mathrm{IF}(x;P,\Phi) = c\left[\frac{\frac{1}{2} - \Phi(2G_{\Phi}^{-1}(\frac{3}{4})-x) + \Phi(2G_{\Phi}^{-1}(\frac{1}{4})-x)}{\phi(\sqrt{2}G_{\Phi}^{-1}(\frac{3}{4}))/\sqrt{2}}\right]. \tag{3.5}$$

Importantly, the influence function of the covariance estimator is bounded and hence, the covariance estimator based on P_n is B-robust, i.e. has finite gross error sensitivity.

3.2.1.2 *Asymptotic variance*

As with scale estimators, the asymptotic variance of covariance and correlation estimators can be found as the expected square of the influence function (Hampel et al., 1986). Furthermore, if the scale estimator S is consistent and asymptotically normally distributed, then the covariance estimator, γ_S will also be consistent and asymptotically normal with asymptotic variance given by:

$$n^{-1}V(\gamma_S,\boldsymbol{F}) = \int \mathrm{IF}^2(x,y;\gamma_S,\boldsymbol{F})\,\mathrm{d}\boldsymbol{F}(x,y).$$

Genton and Ma (1999) show that the asymptotic variance of dispersion estimators is directly proportional to that of the underlying scale estimator. This reinforces the importance of using a highly efficient scale estimator,

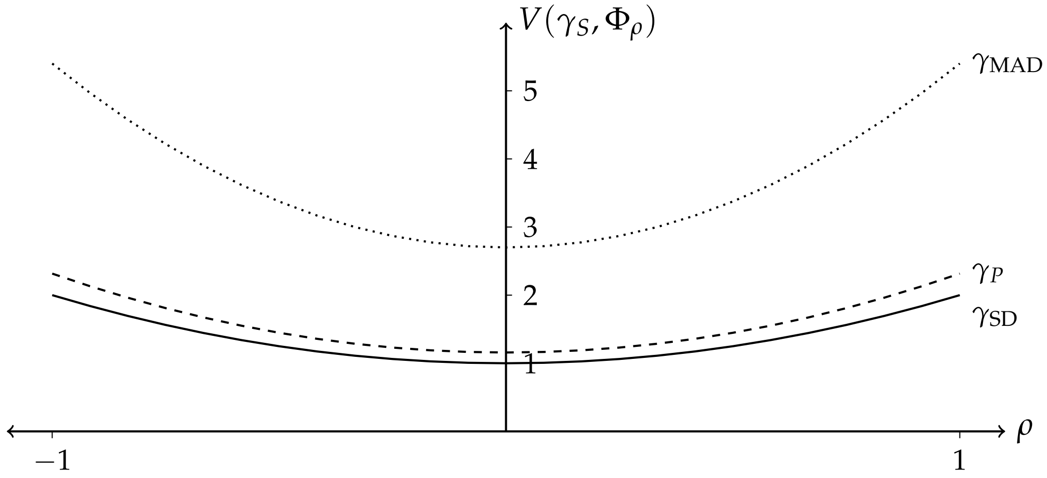


Figure 3.1: Asymptotic variance of covariance estimators, γ_S, based on various scale estimators at the bivariate Gaussian distribution over a range of correlations $-1 \leq \rho \leq 1$.

the efficiency benefits carry through to covariance estimation. In particular, restricting attention to bivariate Gaussian distributions, Φ_ρ,

$$V(\gamma_S, \Phi_\rho) = 2(1+\rho^2)V(S, \Phi), \tag{3.6}$$

where $V(S, \Phi)$ is the asymptotic variance of $\sqrt{n}S$ at Φ.

The asymptotic variance for $\sqrt{n}P_n$ at the Gaussian distribution has previously been shown, in (2.23), to be $V(P, \Phi) = 0.579$ which compares with $V(\text{SD}, \Phi) = 0.5$, $V(Q, \Phi) = 0.61$ and $V(\text{MAD}, \Phi) = 1.35$. This results in a modest increase in asymptotic variance for γ_P over γ_{SD} as shown Figure 3.1. Indeed, given that the asymptotic variance of covariance (and correlation) estimators is directly proportional to that of the underlying scale estimator, the asymptotic efficiency of γ_P and ρ_P compared to the classical covariance estimator at the Gaussian distribution is 86%, the same as that of P_n regardless of the level of correlation in the underlying bivariate Gaussian distribution.

3.2.1.3 *Breakdown value*

The notion of a breakdown value needs to be reinterpreted slightly in the context of (3.1). Ma and Genton (2000) define the breakdown value for a covariance estimator as follows. Let $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ and $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ be two samples of size n. Let $\boldsymbol{Z} = [\boldsymbol{x}, \boldsymbol{y}]$ and let $\widetilde{\boldsymbol{Z}}$ be obtained by replacing any m pairs of $\boldsymbol{Z}$ by arbitrary values. The columns of $\widetilde{\boldsymbol{Z}}$ are then referred to as $\widetilde{\boldsymbol{x}}$ and $\widetilde{\boldsymbol{y}}$. The sample breakdown point of a covariance estimator,

$$\hat{\gamma}_S(\boldsymbol{Z}) = \frac{1}{\alpha\beta}\left[S_n^2(\alpha\boldsymbol{x}+\beta\boldsymbol{y}) - S_n^2(\alpha\boldsymbol{x}-\beta\boldsymbol{y})\right],$$

based on a scale estimator S_n is,

$$\begin{aligned}\varepsilon^*(\hat{\gamma}_S(\boldsymbol{Z})) = \max\Bigg\{\frac{m}{n} : \sup_{\widetilde{\boldsymbol{Z}}}\hat{\gamma}_S(\widetilde{\boldsymbol{Z}}) < \infty \text{ and } \inf_{\widetilde{\boldsymbol{Z}}}\hat{\gamma}_S(\widetilde{\boldsymbol{Z}}) > -\infty \text{ and}\\ \inf_{\widetilde{\boldsymbol{Z}}} S_n(\alpha\widetilde{\boldsymbol{x}}+\beta\widetilde{\boldsymbol{y}}) > 0 \text{ and } \inf_{\widetilde{\boldsymbol{Z}}} S_n(\alpha\widetilde{\boldsymbol{x}}-\beta\widetilde{\boldsymbol{y}}) > 0\Bigg\}.\end{aligned}$$

Hence, the breakdown value of γ_P is 13.4%, inherited directly from P_n (see Section 2.4 for details). To see that this is indeed the case, let x_i (or y_i) be contaminated, then so too is $\alpha x_i \pm \beta y_i$. Hence if we consider the observations as pairs $(x_1, y_1), \ldots, (x_n, y_n)$, then at most 13.4% of the observed pairs can contain contaminated data before the P_n estimator breaks down, and hence γ_P breaks down.

3.2.2 *Simulations*

Having established some asymptotic efficiency results for the covariance estimator at the Gaussian distribution and shown that the robustness and efficiency properties do indeed flow through from the univariate to the bivariate case, we now present some finite sample results for the correlation estimates when the underlying data follows a bivariate t distribution.

There are a number of ways to define a multivariate t distribution, see Kotz (2004) for a recent summary. The form considered in this section is the most common and natural generalisation of the univariate t distribution. A p-dimensional random vector is said to follow a p-variate t distribution with mean vector $\boldsymbol{\mu}$, correlation matrix $\boldsymbol{R}$ and ν degrees of freedom if its joint probability density function is given by,

$$f(\boldsymbol{x}) = \frac{\Gamma((\nu+p)/2)}{(\pi\nu)^{p/2}\Gamma(\nu/2)|\boldsymbol{R}|^{1/2}}\left[1+\frac{1}{\nu}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{R}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]^{-(\nu+p)/2}.$$

As scale and covariance is the primary feature of interest, without loss of generality, we will restrict attention to the central multivariate t distribution where $\boldsymbol{\mu} = \boldsymbol{0}$. Note that in this setting, if $p = 1$ and $\boldsymbol{R} = 1$ then we recover the univariate t distribution. Furthermore, as $\nu \to \infty$ we recover the multivariate Gaussian distribution.

In the simulations that follow, the robust estimates are compared to the (non-robust) MLE calculated assuming that the degrees of freedom parameter is known. This is consistent with how the simulations were performed in Section 2.5, to give a reproducible base line for future comparisons. The MLE of the multivariate t was calculated in R using the `cov.trob` function from the `MASS` package (Kent, Tyler and Vard, 1994; Venables and Ripley, 2002). Comparisons are only made for $\nu > 3$ as the MLE covariance estimates are unstable for $1 \leq \nu < 3$. Indeed for $\nu \leq 2$ we would be estimating a scatter matrix, which is defined more generally than a covariance matrix, as second order moments no longer exist, even though the density of a multivariate t distribution is still well defined for all $\nu > 0$.

Figure 3.2 shows the efficiency of various scale estimates relative to the MLE for bivariate t distributions with correlation equal to 0.5 over a range of degrees of freedom. Over the observed range of degrees of freedom, ρ_P is consistently more efficient than the other robust methods. As the degrees of freedom increase, the efficiency of the non-robust SD based method, ρ_{SD}, increases and will approach parity with the MLE as $\nu \to \infty$. Most of the robust estimators exhibit a slight downward trend in their relative efficiency as the degrees of freedom parameter increases, however ρ_P is reasonably invariant to changes in the degrees of freedom over the range considered here. As expected, ρ_{MAD} performs substantially worse than similar estimates based on P_n and Q_n and the minimum covariance determinant (MCD) methods are even less efficient.

Figure 3.3 shows the relative mean square errors (MSEs). There is no discernible difference between the pattern of efficiencies in Figure 3.2 and the pattern of MSEs in Figure 3.3 indicating that any potential bias introduced by using robust estimators is negligible. This is to be expected, given the correction factors used in the underlying scale estimators cancel when the correlations are calculated.

Figure 3.4 considers what happens to the relative efficiencies of the robust estimators as the level of dependence increases. For all estimators, efficiency appears to decrease as the level of dependence increases while keeping the degrees of freedom parameter fixed at $\nu = 5$. As the dependence level increases, it appears that the MLE does a better job than the robust measures of utilising what information is available information than the robust measures, potentially as a result of the decrease in effective iid sample size. Regardless of the level of dependence ρ_P remains substantially more efficient than ρ_Q. In Figure 3.2, ρ_P is more efficient than ρ_{SD} when $\nu = 5$ with correl-

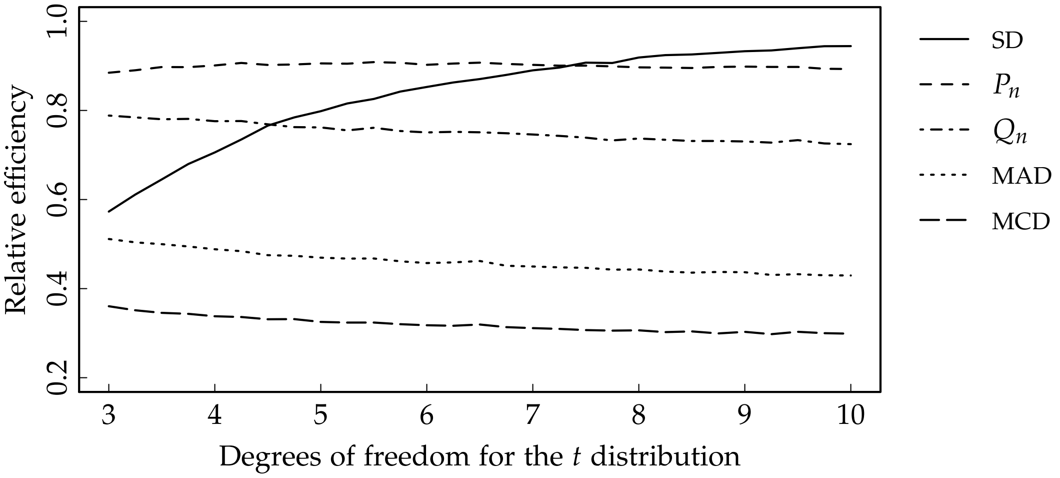


Figure 3.2: Estimated relative efficiencies for the correlation estimates over $N = 100,000$ samples of size $n = 20$ from bivariate t distributions with correlation equal to 0.5. The efficiencies are measured relative to the MLE with known degrees of freedom.

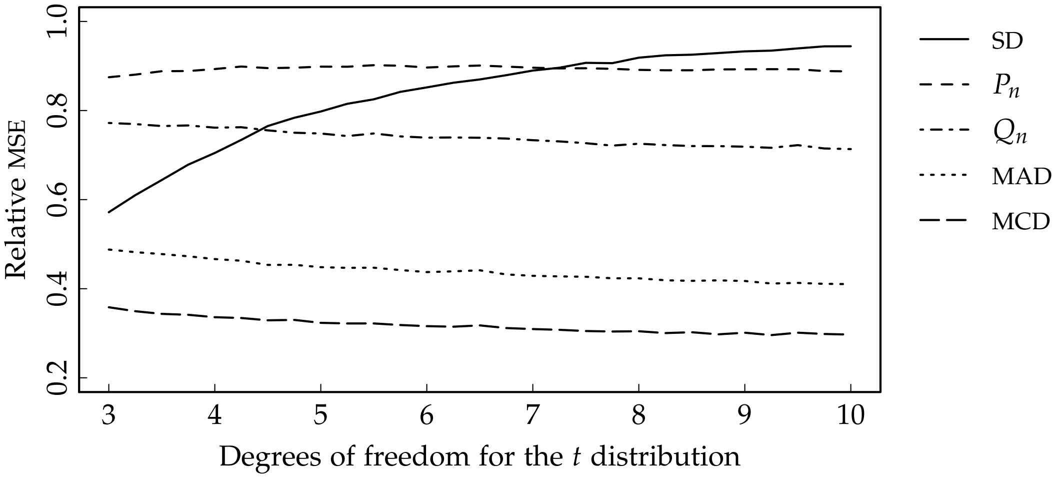


Figure 3.3: Estimated relative MSEs for the correlation estimates over $N = 100,000$ samples of size $n = 20$ from bivariate t distributions with correlation equal to 0.5. The efficiencies are measured relative to the MLE with known degrees of freedom.

ation equal to 0.5. Figure 3.4 shows that the same is true regardless of the level of dependence.

Furthermore, it appears again that there is negligible bias present in the robust covariance estimators, as indicated by the agreement between the relative MSEs in Figure 3.5 with the relative efficiencies in Figure 3.4.

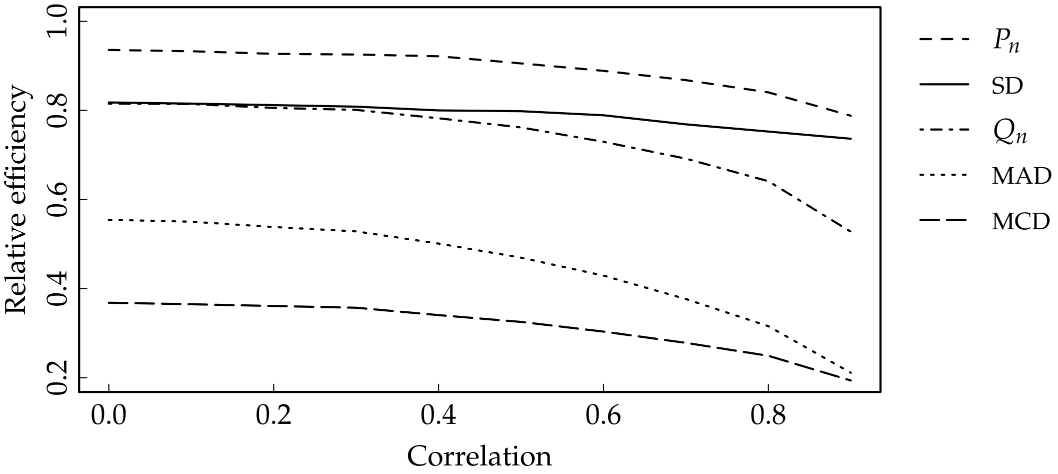


Figure 3.4: Estimated relative efficiencies for correlation estimates over $N = 100,000$ samples of size $n = 20$ from bivariate t distributions with $\nu = 5$ degrees of freedom and correlations ranging from 0 to 0.9. The efficiencies are measured relative to the MLE with known degrees of freedom.

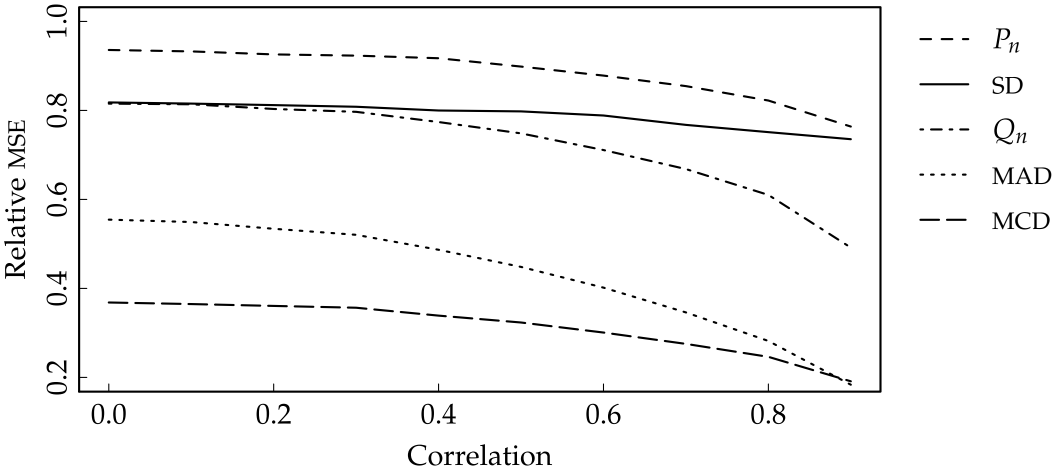


Figure 3.5: Estimated relative MSEs for correlation estimates over $N = 100,000$ samples of size $n = 20$ from bivariate t distributions with $\nu = 5$ degrees of freedom and correlations ranging from 0 to 0.9. The efficiencies are measured relative to the MLE with known degrees of freedom.

3.3 robust autocovariance estimation

In this section we will consider the problem of estimating scale and autocovariance in a time series setting. This is an important problem that has many applications including financial data, hydrology, quality control, and signal processing.

To ensure that the quantities we are aiming to estimate do not change over time, we will restrict attention to time series observations, $\{X_i\}_{i\geq 1}$, that satisfy the hypothesis of second-order stationarity:

A. $\mathbb{E}(X_i) = \mu$ for all $i = 1, 2, \ldots$;

B. $\operatorname{var}(X_i^2) = \sigma^2 < \infty$ for all $i = 1, 2, \ldots$; and

C. $\operatorname{cov}(X_{i+h}, X_i) = \gamma(h)$ for all $i, h \in \mathbb{Z}$.

Applying the identity from Gnanadesikan and Kettenring (1972) in a time series setting we can construct robust autocovariance estimators based on equation (3.1). That is, autocovariance estimators based on the identity,

$$\gamma(h) = \frac{1}{4}\left[\operatorname{var}(X_1 + X_{h+1}) - \operatorname{var}(X_1 - X_{h+1})\right]. \tag{3.7}$$

The autocovariance at a particular lag, h, measures the degree of second order variation between observations that are a fixed distance, h, apart.

Let $\boldsymbol{x}_{a:b} = (x_a, \ldots, x_b)$ denote a sequence of $b - a + 1$ observations. In the simplest case we can substitute a standard (non-robust) estimate of the sample variance,

$$\begin{aligned}\hat{\gamma}_{\hat{\sigma}}(h) &= \frac{1}{4}\left[\hat{\sigma}^2_{n-h}(\boldsymbol{x}_{1:n-h} + \boldsymbol{x}_{h+1:n}) - \hat{\sigma}^2_{n-h}(\boldsymbol{x}_{1:n-h} - \boldsymbol{x}_{h+1:n})\right] \\ &= \frac{1}{n-h}\sum_{i=1}^{n-h}(x_i - \bar{x}_{1:n-h})(x_{i+h} - \bar{x}_{h+1:n}), \quad 0 \leq h \leq n-1, \end{aligned} \tag{3.8}$$

where

$$\hat{\sigma}^2_n(\boldsymbol{x}_n) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_{1:n})^2$$

and $\bar{x}_{a:b} = (b - a + 1)^{-1}\sum_{i=a}^{b} x_i$. The resulting estimator (3.8) is asymptotically equivalent to the classical autocovariance estimator. Ma and Genton (2000) suggest a robust alternative by plugging in Q_n,

$$\hat{\gamma}_Q(h) = \frac{1}{4}\left[Q^2_{n-h}(\boldsymbol{x}_{1:n-h} + \boldsymbol{x}_{h+1:n}) - Q^2_{n-h}(\boldsymbol{x}_{1:n-h} - \boldsymbol{x}_{h+1:n})\right].$$

In an attempt to improve the efficiency whilst maintaining a level of robustness, we propose a similar autocovariance estimator based on P_n,

$$\hat{\gamma}_P(h) = \frac{1}{4}\left[P^2_{n-h}(\boldsymbol{x}_{1:n-h} + \boldsymbol{x}_{h+1:n}) - P^2_{n-h}(\boldsymbol{x}_{1:n-h} - \boldsymbol{x}_{h+1:n})\right]. \tag{3.9}$$

In contrast to scale estimators, autocovariance estimators are not invariant to permutations of the underlying observations – which complicates the characterisation of their breakdown values. Ma and Genton (2000) reinterpret the breakdown point of a covariance estimator, as defined in Section 3.2.1.3, to incorporate the importance of temporal disturbances on autocovariance estimators. Let $\boldsymbol{x} = (x_1, \dots, x_n)^\mathsf{T}$ be a time series of length n. Let $\widetilde{\boldsymbol{x}}$ be obtained by replacing any m observations of $\boldsymbol{x}$ by arbitrary values. Denote I_m a subset of size m of $\{1, \dots, n\}$. The temporal sample breakdown point of an autocovariance estimator $\hat{\gamma}(h)$ is:

$$\begin{aligned}\varepsilon^*(\hat{\gamma}(h)) = \max\Bigg\{\frac{m}{n} : &\sup_{I_m}\sup_{\widetilde{x}} S_{n-h}(\widetilde{\boldsymbol{u}}) < \infty \text{ and } \inf_{I_m}\inf_{\widetilde{x}} S_{n-h}(\widetilde{\boldsymbol{u}}) > 0 \text{ and} \\ &\sup_{I_m}\sup_{\widetilde{x}} S_{n-h}(\widetilde{\boldsymbol{v}}) < \infty \text{ and } \inf_{I_m}\inf_{\widetilde{x}} S_{n-h}(\widetilde{\boldsymbol{v}}) > 0\Bigg\}\end{aligned}$$

where $\widetilde{\boldsymbol{u}} = \widetilde{\boldsymbol{x}}_{1:n-h} + \widetilde{\boldsymbol{x}}_{h+1:n}$, $\widetilde{\boldsymbol{v}} = \widetilde{\boldsymbol{x}}_{1:n-h} - \widetilde{\boldsymbol{x}}_{h+1:n}$.

This definition highlights the importance of the positioning of the corruption taken into account through the supremum and infimum on the set of arrangements, I_m. The breakdown value represents the worst case scenario. A particular example of this is given in Figure 3.6 which considers the autocovariance at lag $h = 2$ from a sample of size $n = 13$. The autocovariance estimator works with the data sets, $\boldsymbol{x}_{1:11} \pm \boldsymbol{x}_{3:13}$. If four observations are contaminated in an appropriate way so as to cause maximum damage, this will

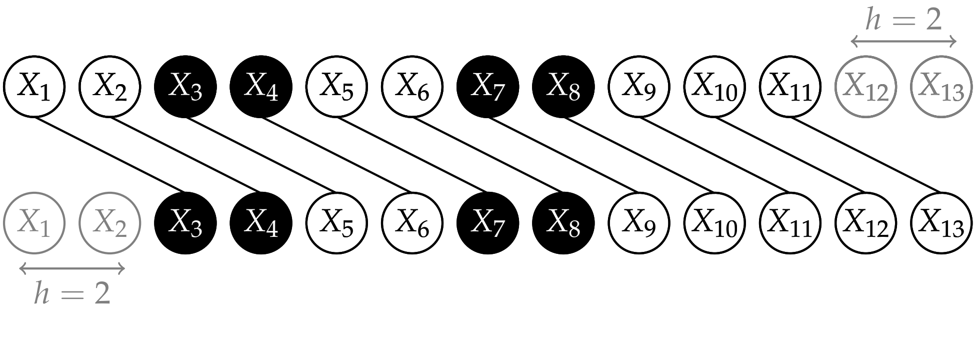


Figure 3.6: Illustration of the effect of contamination on autocovariance estimators. The black observations have been contaminated. If appropriately arranged, the contamination of four observations leads to eight pairs of observations being contaminated.

result in 8 contaminated pairs. Ma and Genton (2000, Proposition 1) show that asymptotically the breakdown value for autocovariance estimators is half that of the corresponding covariance estimator. Hence, the asymptotic breakdown value for $\hat{\gamma}_P$ is 6.7%.

The influence function for the autocovariance estimator has the same general form as in the covariance case, (3.4). Importantly, the influence function of $\hat{\gamma}_P$ remains bounded and as such the estimator has finite gross error sensitivity. In the remainder of this chapter, attention will focus on Gaussian processes hence an important special case of (3.4) is given by,

$$\begin{aligned}\mathrm{IF}(x, y; \gamma_P, \Phi_\rho) = {} & (\gamma(0)+\gamma(h))\,\mathrm{IF}\left(\frac{x+y}{\sqrt{2(\gamma(0)+\gamma(h))}}; P, \Phi\right) \\ & - (\gamma(0)-\gamma(h))\,\mathrm{IF}\left(\frac{x-y}{\sqrt{2(\gamma(0)-\gamma(h))}}; P, \Phi\right), \quad (3.10)\end{aligned}$$

where the influence function of P_n at Φ is defined in (3.5).

The asymptotic distribution and efficiency properties of P_n and $\hat{\gamma}_P$ depend on the level of dependence in the underlying time series. The influence functions are a key ingredient when it comes to proving limit results for $\hat{\gamma}_P$ under short and long range dependent Gaussian processes in Sections 3.4 and 3.5.

The results under SRD follow closely what has been done by Lévy-Leduc et al. (2011c) for $\hat{\gamma}_Q$. However, in the LRD setting, some novel results are found and conjectures drawn that contrast with the limiting distribution of the Hodges-Lehmann estimator established by Lévy-Leduc et al. (2011a). These somewhat surprising results follow as P_n, the difference of two U-quantile statistics, has a distinct asymptotic behaviour to that of the individual U-quantiles.

Furthermore, we observe the same limiting behaviour for a number of robust scale and autocovariance estimators under a certain type of LRD process. We show that the limiting result for the IQR is the same as that for Q_n and the SD and we conjecture that the same is true for P_n and the MAD. We also show how the results for the autocovariance estimators follow from the results for the underlying scale estimators.

3.4 SHORT RANGE DEPENDENT GAUSSIAN PROCESSES

A sequence of observations whose autocovariance function is summable is known as a short range dependent (SRD) time series and is said to have short memory. Formally, let $\{X_i\}_{i\geq 1}$ be a stationary time series and let γ_h be its autocovariance function at lag h. The sequence $\{X_i\}_{i\geq 1}$ is said to be of *short memory* if $\sum_{h\geq 1}|\gamma_h| < \infty$.

Assumption 3.1. $\{X_i\}_{i\geq 1}$ *is a stationary mean-zero Gaussian process with autocovariance sequence* $\gamma(h) = \mathbb{E}(X_1 X_{h+1})$ *satisfying* $\sum_{h\geq 1}|\gamma(h)| < \infty$.

The asymptotic normality for P_n and $\hat{\gamma}_P$ under the conditions outlined in Assumption 3.1 is established in this section. We follow the same general approach that Lévy-Leduc et al. (2011c) use to establish similar results for Q_n. The derivation is a little more involved, owing to the nature of P_n as a difference of two U-quantiles, rather than a single U-quantile as is the case for Q_n.

3.4.1 *Results for P_n*

Before considering the asymptotic distribution of the autocovariance estimator $\hat{\gamma}_P$, we first need results regarding the underlying scale estimator, P_n. This section is dedicated to the proof of the following result regarding the asymptotic distribution of the scale estimator P_n under SRD Gaussian processes.

Result 3.1. *Under Assumption 3.1, P_n satisfies the following central limit theorem (CLT):*

$$\sqrt{n}(P_n - \sigma) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \widetilde{\sigma}^2),$$

where $\sigma = \sqrt{\gamma(0)}$ *and the limiting variance* $\widetilde{\sigma}^2$ *is given by*

$$\widetilde{\sigma}^2 = \sigma^2 \mathbb{E}\left[\mathrm{IF}^2\left(\frac{X_1}{\sigma}; P, \Phi\right)\right] + 2\sigma^2 \sum_{k\geq 1} \mathbb{E}\left[\mathrm{IF}\left(\frac{X_1}{\sigma}; P, \Phi\right)\mathrm{IF}\left(\frac{X_{k+1}}{\sigma}; P, \Phi\right)\right].$$

The proof of Result 3.1 requires the following structure. Let I be a compact interval of $\mathbb{R}$ and let $\mathcal{D}(I)$ be the space of all functions that are right continuous and whose limits from the left exist everywhere on I, i.e. the set of càdlàg functions. Let $\mathcal{M}([-\infty, \infty])$ be the set of CDFs on $[-\infty, \infty]$. Equip both $\mathcal{D}$ and $\mathcal{M}$ with the topology of uniform convergence and denote the uniform norm by $||\cdot||_\infty$.

We need to define functionals in terms of the underlying distribution. First the pairwise mean mapping,

$$\begin{aligned} T_1: \quad & \mathcal{M}([-\infty,\infty]) \to \mathcal{D}([-\infty,\infty]) \\ & F \mapsto \left\{ r \mapsto G(r) = \int\int \mathbb{I}\{x+y \leq 2r\} dF(x) dF(y) \right\}. \end{aligned}$$

Secondly the quantile mappings,

$$\begin{aligned} T_2: \quad & \mathcal{D}([-\infty,\infty]) \to \mathbb{R} \\ & G \mapsto G^{-1}(\tfrac{1}{4}), \end{aligned}$$

and,

$$\begin{aligned} T_3: \quad & \mathcal{D}([-\infty,\infty]) \to \mathbb{R} \\ & G \mapsto G^{-1}(\tfrac{3}{4}). \end{aligned}$$

Finally we need the lower and upper quartiles of the pairwise means,

$$\begin{aligned} T_L = T_2 \circ T_1: \quad & \mathcal{M}([-\infty,\infty]) \to \mathbb{R} \\ & F \mapsto G^{-1}(\tfrac{1}{4}), \end{aligned}$$

and,

$$\begin{aligned} T_U = T_3 \circ T_1: \quad & \mathcal{M}([-\infty,\infty]) \to \mathbb{R} \\ & F \mapsto G^{-1}(\tfrac{3}{4}). \end{aligned}$$

The scale estimator P_n can now be expressed as,

$$P_n = c\,[T_U(F_n) - T_L(F_n)] = c\,(T_U - T_L)\,(F_n),$$

which is a consistent estimator of,

$$P(\Phi_\sigma) = c(T_U - T_L)(\Phi_\sigma) = \sigma,$$

where Φ_σ is a Gaussian distribution with mean zero and variance σ^2. Recall that the correction factor $c = 1.0483$, defined in Section 2.4.2.1, ensures P_n is consistent for the standard deviation at the Gaussian distribution. For the remainder of this chapter, without loss of generality, let $\sigma = 1$ and $P(\Phi) = P(\Phi_1)$.

The proof begins by showing that T_1, T_2 and T_3 and consequently T_U and T_L are Hadamard differentiable[2] defined continuously on $\mathcal{D}(I)$. We then

[2] Refer to Section C.1 in Appendix C (p. 151) for a definition of Hadamard differentiability.

express the estimator P_n as a sum of influence functions with an asymptotically negligible remainder term. A CLT for the sum of influence functions gives the required result.

To show that T_1 is Hadamard differentiable, let $\{g_t\} \in \mathcal{D}(I)$ be a sequence of càdlàg functions with bounded variations such that $||g_t - g||_\infty \to 0$ as $t \to 0$, where g is also a càdlàg function. Now for any $r \in \mathbb{R}$, consider

$$\begin{aligned}\frac{T_1(F+tg_t)[r] - T_1(F)[r]}{t} &= t^{-1}\big[\int\int \mathbb{I}\{x+y \le 2r\}\,\mathrm{d}[F+tg_t](x)\,\mathrm{d}[F+tg_t](y) \\ &\qquad - \int\int \mathbb{I}\{x+y \le 2r\}\,\mathrm{d}F(x)\,\mathrm{d}F(y)\big] \\ &= 2\int\int \mathbb{I}\{x+y \le 2r\}\,\mathrm{d}F(x)\,\mathrm{d}g_t(y) \\ &\qquad + t\int\int \mathbb{I}\{x+y \le 2r\}\,\mathrm{d}g_t(x)\,\mathrm{d}g_t(y).\end{aligned}$$

Furthermore,

$$\begin{aligned}&\big|\int\int \mathbb{I}\{x+y \le 2r\}\,\mathrm{d}F(x)\,\mathrm{d}g_t(y) - \int\int \mathbb{I}\{x+y \le 2r\}\,\mathrm{d}F(x)\,\mathrm{d}g(y)\big| \\ &\qquad = \left|\int_{\mathbb{R}}\int_{-\infty}^{2r-x} \mathrm{d}g_t(y)\,\mathrm{d}F(x) - \int_{\mathbb{R}}\int_{-\infty}^{2r-x} \mathrm{d}g(y)\,\mathrm{d}F(x)\right| \\ &\qquad \le 2||g_t - g||_\infty \to 0 \quad \text{as } t \to 0.\end{aligned}$$

Hence, the Hadamard derivative of T_1 at g is given by,

$$(\mathrm{D}_g\, T_1(F))[r] = \lim_{t\to 0} \frac{T_1(F+tg_t)[r] - T_1(F)[r]}{t} = 2\int_{\mathbb{R}}\int_{-\infty}^{2r-x} \mathrm{d}g(y)\,\mathrm{d}F(x),$$

where D_g is the Hadamard derivative operator.

The inverse mappings T_2 and T_3 are shown to be Hadamard differentiable in van der Vaart (1998, Lemma 21.3) and also van der Vaart and Wellner (1996, Lemma 3.9.20). Applying the chain rule for the inverse map, van der Vaart (1998, Theorem 20.9), we have,

$$\begin{aligned}\mathrm{D}_g\, T_L(F) = \mathrm{D}_g(T_2 \circ T_1)(F) &= -\frac{(\mathrm{D}_g\, T_1(F))[T_L(F)]}{(T_1(F))'[T_L(F)]} \\ &= -\frac{2\int_{\mathbb{R}}\int_{-\infty}^{2T_L(F)-x} \mathrm{d}g(y)\,\mathrm{d}F(x)}{(T_1(F))'[T_L(F)]}. \qquad (3.11)\end{aligned}$$

Similarly,

$$\mathrm{D}_g\, T_U(F) = \mathrm{D}_g(T_3 \circ T_1)(F) = -\frac{2\int_{\mathbb{R}}\int_{-\infty}^{2T_U(F)-x} \mathrm{d}g(y)\,\mathrm{d}F(x)}{(T_1(F))'[T_U(F)]}. \qquad (3.12)$$

Hence, $\mathrm{D}_g(T_U - T_L)(F)$ is a continuous function of g and is defined on $\mathcal{D}(I)$, and as such $(T_U - T_L)(F)$ is Hadamard differentiable.

If for some sequence of numbers $a_n \to \infty$ we have $a_n(F_n - F)$ converging in distribution, then we can apply the functional delta method of van der Vaart (1998, Theorem 20.8), to write,

$$\begin{aligned} a_n(P_n - P) &= a_n c(T_U - T_L)(F_n - F) \\ &= c\,\mathrm{D}_{\{a_n(F_n - F)\}}(T_U - T_L)(F) + o_p(1). \end{aligned} \tag{3.13}$$

When $F = \Phi$, under the conditions of Assumption 3.1, we can apply Csörgó and Mielniczuk (1996) to show that $\sqrt{n}(F_n - \Phi)$ converges in distribution to a Gaussian process in $\mathcal{D}(I)$. Hence, (3.13) is valid with $a_n = \sqrt{n}$.

Using the results in (3.11) and (3.12) we have,

$$\begin{aligned} \mathrm{D}_{\{\sqrt{n}(F_n-\Phi)\}}(T_U - T_L)(\Phi) &= \sqrt{n}\left[\mathrm{D}_{\{F_n-\Phi\}}(T_U)(\Phi) - \mathrm{D}_{\{F_n-\Phi\}}(T_L)(\Phi)\right] \\ &= 2\sqrt{n}\left[-\frac{\int (F_n - \Phi)(2T_U(\Phi) - x)\,\mathrm{d}\Phi(x)}{(T_1(\Phi))'[T_U(\Phi)]}\right. \\ &\qquad \left. + \frac{\int (F_n - \Phi)(2T_L(\Phi) - x)\,\mathrm{d}\Phi(x)}{(T_1(\Phi))'[T_L(\Phi)]}\right] \\ &= 2\sqrt{n}\,[A - A_n]\,, \end{aligned}$$

where,

$$A = \frac{\int_{\mathbb{R}} \Phi(2T_U(\Phi) - x)\,\mathrm{d}\Phi(x)}{(T_1(\Phi))'[T_U(\Phi)]} - \frac{\int_{\mathbb{R}} \Phi(2T_L(\Phi) - x)\,\mathrm{d}\Phi(x)}{(T_1(\Phi))'[T_L(\Phi)]},$$

and,

$$A_n = \frac{\int_{\mathbb{R}} F_n(2T_U(\Phi) - x)\,\mathrm{d}\Phi(x)}{(T_1(\Phi))'[T_U(\Phi)]} - \frac{\int_{\mathbb{R}} F_n(2T_L(\Phi) - x)\,\mathrm{d}\Phi(x)}{(T_1(\Phi))'[T_L(\Phi)]}.$$

Note that the numerators in A can be simplified,

$$A = \frac{3/4}{(T_1(\Phi))'[T_U(\Phi)]} - \frac{1/4}{(T_1(\Phi))'[T_L(\Phi)]}.$$

The numerators in A_n can be rewritten as,

$$\frac{1}{n}\sum_{i=1}^{n} \Phi(2T_U(\Phi) - X_i) \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} \Phi(2T_L(\Phi) - X_i).$$

Noting the symmetry in T_U and T_L we have,

$$\begin{aligned} (T_1(\Phi))'[T_U(\Phi)] = (T_1(\Phi))'[T_L(\Phi)] &= G'_\Phi(T_U(\Phi)) \\ &= 2\int \phi(2T_U(\Phi) - x)\phi(x)dx \\ &= \sqrt{2}\phi(\sqrt{2}T_U(\Phi)). \end{aligned}$$

Hence, from (3.13) we have that,

$$\begin{aligned}\sqrt{n}(P_n - P) &= c\,\mathrm{D}_{\sqrt{n}(F_n-\Phi)}(T_U - T_L)(\Phi) + o_p(1)\\ &= \frac{c}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{\frac{1}{2} - \Phi(2T_U(\Phi) - X_i) + \Phi(2T_L(\Phi) - X_i)}{\phi(\sqrt{2}T_U(\Phi))/\sqrt{2}}\right] + o_p(1)\\ &= \frac{c}{\sqrt{n}}\sum_{i=1}^{n}\mathrm{IF}(X_i; P, \Phi) + o_p(1), \qquad (3.14)\end{aligned}$$

where the influence function for P_n at the Gaussian distribution was given in (3.5). A CLT for $n^{-1/2}\sum_{i=1}^{n}\mathrm{IF}(X_i; P, \Phi)$ follows from Arcones (1994, Theorem 4).[3] Hence, under Assumption 3.1,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathrm{IF}(X_i; P, \Phi) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \tilde{\sigma}^2),$$

where

$$\tilde{\sigma}^2 = \mathbb{E}\left[\mathrm{IF}^2(X_1; P, \Phi)^2\right] + 2\sum_{k\geq 1}\mathbb{E}\left[\mathrm{IF}(X_1; P, \Phi)\,\mathrm{IF}(X_{1+k}; P, \Phi)\right].$$

Result 3.1, where observations have standard deviation, $\sigma = \sqrt{\gamma(0)}$, follows by noting that $\mathrm{IF}(x; S, \Phi_\sigma) = \sigma\,\mathrm{IF}(x/\sigma; S, \Phi)$ from Proposition C.1 in Appendix 3 (p. 151).

3.4.2 *Results for $\hat{\gamma}_P(h)$*

Having established the limiting distribution of the scale estimator, P_n, it is relatively straightforward to find the limiting distribution of the autocovariance function $\hat{\gamma}_P$. This section proves the following result regarding the asymptotic distribution of the autocovariance estimator $\hat{\gamma}_P$ under SRD Gaussian processes.

Result 3.2. *Under Assumption 3.1, $\hat{\gamma}_P(h)$ satisfies the following* CLT*,*

$$\sqrt{n}(\hat{\gamma}_P(h) - \gamma(h)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \breve{\sigma}^2(h)),$$

where

$$\begin{aligned}\breve{\sigma}^2(h) = \mathbb{E}&\left[\mathrm{IF}^2(X_1, X_{1+h}; \gamma_P, \Phi_\rho)\right]\\ &+ 2\sum_{k\geq 1}\mathbb{E}\left[\mathrm{IF}(X_1, X_{1+h}; \gamma_P, \Phi_\rho)\,\mathrm{IF}(X_{k+1}, X_{k+1+h}; \gamma_P, \Phi_\rho)\right]\end{aligned}$$

and $\mathrm{IF}(x, y; \gamma_P, \Phi_\rho)$ *is defined in equation* (3.10).

[3] See Section C.3 in Appendix C (page 158) for details. In particular, we show that the Hermite rank of the influence function equals 2.

Let Φ_+ and Φ_- denote the CDFs of $\{X_i + X_{i+h}\}_{i\geq 1}$ and $\{X_i - X_{i+h}\}_{i\geq 1}$, respectively. Similarly let $F_{+,n-h}$ and $F_{-,n-h}$ denote the empirical distribution functions of $\{X_i + X_{i+h}\}_{1\leq i\leq n-h}$ and $\{X_i - X_{i+h}\}_{1\leq i\leq n-h}$. Also let $P_{\pm,n-h} = P^2_{n-h}(\boldsymbol{x}_{1:n-h} \pm \boldsymbol{x}_{h+1:n})$. Since the underlying Gaussian process, $\{X_i\}_{i\geq 1}$, satisfies the SRD assumptions, the same holds for $\{X_i + X_{i+h}\}_{i\geq 1}$ and $\{X_i - X_{i+h}\}_{i\geq 1}$. Furthermore, with the appropriate correction factor we have that $P(\Phi_\pm) = \sqrt{2(\gamma(0) \pm \gamma(h))}$ and hence,

$$\gamma(h) = \frac{P^2(\Phi_+) - P^2(\Phi_-)}{4}.$$

We apply Csörgó and Mielniczuk (1996) to show that $\sqrt{n-h}(F_{\pm,n-h} - \Phi_\pm)$ converges in distribution to Gaussian processes in $\mathcal{D}(I)$. Thus, using a similar argument to above,

$$\sqrt{n-h}\,(P_{\pm,n-h} - P(\Phi_\pm)) = \frac{1}{\sqrt{n-h}} \sum_{i=1}^{n-h} \mathrm{IF}(X_i \pm X_{i+h}; P, \Phi_\pm) + o_p(1).$$

Using the functional delta method, van der Vaart (1998, Theorem 3.1), with $b(x) = x^2$ and $b'(x) = 2x$, we obtain

$$\sqrt{n-h}\left(P^2_{\pm,n-h} - P^2(\Phi_\pm)\right) = \frac{2P(\Phi_\pm)}{\sqrt{n-h}} \sum_{i=1}^{n-h} \mathrm{IF}(X_i \pm X_{i+h}; P, \Phi_\pm) + o_p(1).$$

Applying the basic properties of influence functions, outlined in Proposition C.1 (p. 151), we have that,

$$P(\Phi_\pm)\,\mathrm{IF}(X_1 \pm X_{1+h}; P, \Phi_\pm) = 2(\gamma(0) \pm \gamma(h))\,\mathrm{IF}\left(\frac{X_1 \pm X_{1+h}}{\sqrt{2(\gamma(0) \pm \gamma(h))}}; P, \Phi\right).$$

Hence, recalling from (3.10) that,

$$\hat{\gamma}_P(h) = \frac{1}{4}\left[P^2_{n-h}(\boldsymbol{x}_{1:n-h} + \boldsymbol{x}_{h+1:n}) - P^2_{n-h}(\boldsymbol{x}_{1:n-h} - \boldsymbol{x}_{h+1:n})\right]$$

has influence function,

$$\begin{aligned}\mathrm{IF}(X_1, X_{1+h}; \gamma_P, \Phi_\rho) = {} & (\gamma(0) + \gamma(h))\,\mathrm{IF}\left(\frac{X_1 + X_{1+h}}{\sqrt{2(\gamma(0) + \gamma(h))}}; P, \Phi\right) \\ & - (\gamma(0) - \gamma(h))\,\mathrm{IF}\left(\frac{X_1 - X_{1+h}}{\sqrt{2(\gamma(0) - \gamma(h))}}; P, \Phi\right),\end{aligned}$$

we have the following asymptotic expansion,

$$\sqrt{n-h}\,(\hat{\gamma}_P(h) - \gamma(h)) = \frac{1}{\sqrt{n-h}} \sum_{i=1}^{n-h} \mathrm{IF}(X_i, X_{i+h}; \gamma_P, \Phi_\rho) + o_p(1).$$

The asymptotic normality follows from a CLT for,

$$\frac{1}{\sqrt{n-h}} \sum_{i=1}^{n-h} \mathrm{IF}(X_i, X_{i+h}; \gamma_P, \Phi_\rho),$$

established by noting that the Hermite rank of $\mathrm{IF}(x, y; \gamma_P, \Phi)$ is $m = 2$ and applying Arcones (1994, Theorem 4), which is reproduced in a simplified form as Theorem C.2 in Appendix C. Hence,

$$\sqrt{n}(\hat{\gamma}_P(h) - \gamma(h)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \breve{\sigma}^2(h)),$$

where

$$\begin{aligned}\breve{\sigma}^2(h) = \mathbb{E}\left[\mathrm{IF}^2(X_1, X_{1+h}; \gamma_P, \Phi_\rho)\right] \\ + 2\sum_{k\geq 1} \mathbb{E}\left[\mathrm{IF}(X_1, X_{1+h}; \gamma_P, \Phi_\rho)\, \mathrm{IF}(X_{k+1}, X_{k+1+h}; \gamma_P, \Phi_\rho)\right].\end{aligned}$$

3.5 LONG RANGE DEPENDENT GAUSSIAN PROCESSES

LRD time series form an important class of dependent observations. The distinction between short and long range dependence is due to the behaviour of the autocovariance decay rate. In particular, let $\{X_i\}_{i\geq 1}$ be a stationary time series and let γ_k be its autocovariance function at lag k. The sequence $\{X_i\}_{i\geq 1}$ is said to be of *long memory* if $\sum_{h\geq 1} |\gamma_h| = \infty$, i.e. its autocovariances are not summable. LRD processes are characterised by spurious trends, self-similarity and an autocorrelation function that exhibits a slow hyperbolic decay, $\rho(h) \sim h^{-D}$ for $0 < D < 1$.[4]

Applications involving LRD processes are found in a diverse range of settings, such as the analysis of network traffic, heartbeat fluctuations, wind turbine output, climate and financial data, to name just a few. For some recent examples we refer to Park et al. (2011), Ercan, Kavvas and Abbasov (2013) and Beran et al. (2013).

The presence of extreme events in LRD processes is also of philosophical interest. Bunde et al. (2005) discuss how long term correlations represent a natural mechanism for the clustering of extreme events. This may make it difficult to ascertain whether unusual observations or clusters are outliers or genuine observations that are part of the LRD data generating process.

[4] The notation $a_n \sim b_n$ as $n \to \infty$ for two real-valued sequences a_n and b_n means that $a_n/b_n \to 1$. Similarly for functions, $g(x) \sim h(x)$ as $x \to x_0$ means that $g(x)/h(x) \to 1$ as $x \to x_0$.

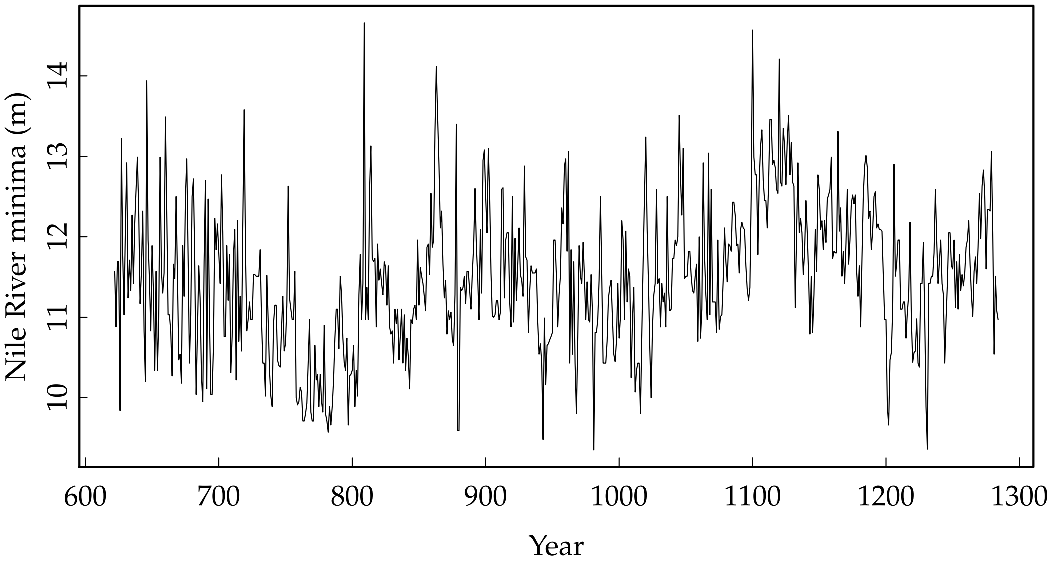


Figure 3.7: Annual Nile river minima (in meters) for the period 622–1284.

Hence, robust procedures that sidestep the issue of identifying outliers are inherently valuable. More recently, Franzke et al. (2012) discuss how self-similar processes can arise as a result of both long range dependence and non-Gaussianity.

A classic example of a LRD process is the data set consisting of yearly minimum water levels of the Nile river at the Roda Gauge. Most of the observations were taken in a structure known as the Nilometer on the southern tip of Roda island in central Cairo. Toussoun (1925) reports an uninterrupted series of observations that were recorded over the years 622–1284; these are plotted in Figure 3.7. Post 1284, the observations become sparse, for example there were only 17 measurements during the sixteenth century. Furthermore, the construction of dams in the 20th century ended the dominance of nature in determining the level of the Nile. The level of the Nile river was, and remains, of vital interest to the inhabitants of the region. More globally, Eltahir and Wang (1999) found a strong relationship between the Nile water levels and the El Niño phenomenon.

Statistically, this data set is historically very important as it is the focus of Hurst (1951), a seminal paper in the field of LRD processes. It is also important from a robustness perspective. Whitcher et al. (2002) suggest that there exists heterogeneity in the variance of the series, with a change point at 720 AD. This date coincides with historical records that indicate construction of the Nilometer on Roda island and a change in the way the series was measured. More recently, Chareka, Matarise and Turner (2006) investigate

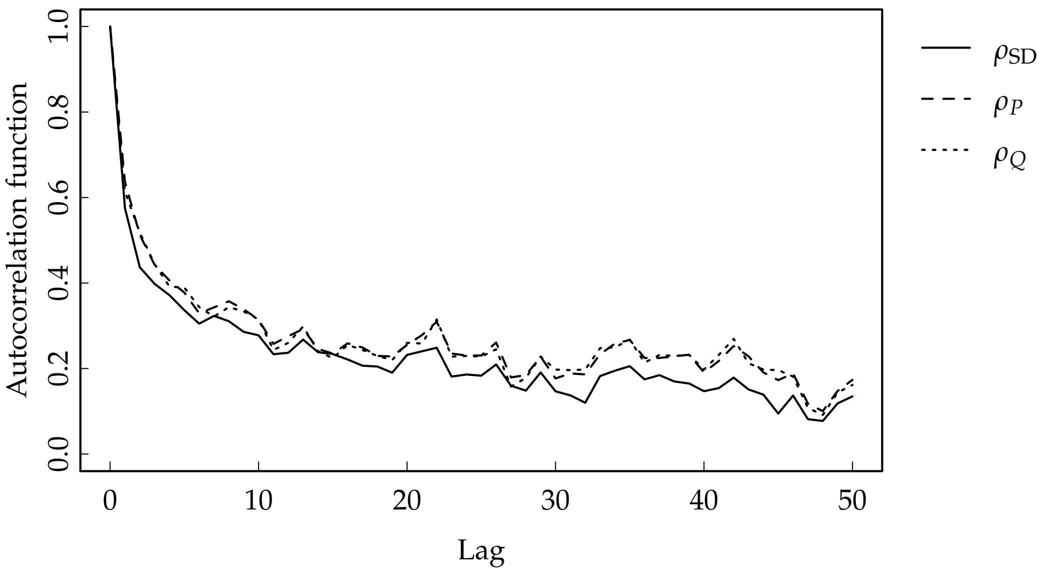


Figure 3.8: ACF of the annual Nile river minima for the period 622–1284.

whether or not there are outliers present in this data set. They find that the observations in years 646 and 809 are most likely outliers, and potentially 878 AD.

The autocorrelation functions (ACFs) plotted in Figure 3.8 show no appreciable difference between $\hat{\rho}_P$ and $\hat{\rho}_Q$, though both robust methods find marginally higher levels of dependence than the classical autocorrelation method. Regardless, the slow decay rate of the autocorrelation function is apparent.

The distribution of estimators under LRD processes is important for practical reasons and is interesting theoretically as it is common for non-normal limit distributions to arise. Lévy-Leduc et al. (2011c) investigate the limiting distribution of Q_n and associated robust autocovariance estimators under dependent processes. They find the SD and Q_n have the same limiting distribution, with no loss of asymptotic efficiency under extreme long range dependence. Similar results about the relative performance of robust estimators to classical estimators also hold for location estimators. See Lévy-Leduc et al. (2011a) for details about the Hodges-Lehmann estimator and Beran (1991) for M-estimates of location.

This section continues with a discussion of the various ways a LRD process can be parameterised. A new limit result for the IQR is presented, comparable with those for Q_n and the SD. We also partially show similar results hold for the scale estimator P_n and the autocovariance estimator $\hat{\gamma}_P$. The validity of these results is demonstrated through simulation.

3.5.1 *Parameterisation*

There are a number of ways to parameterise LRD processes. A brief overview is provided below, for further details see Giraitis, Koul and Surgailis (2012) or Beran et al. (2013).

As in Section 3.4, we restrict attention to Gaussian processes. In particular, let $\{X_i\}_{i\geq 0}$ be the stationary mean-zero linear Gaussian process,

$$X_i = \sum_{j=0}^{\infty} a_j \varepsilon_{i-j}, \qquad \text{for } i \in \mathbb{N},$$

where $a_0 = 1$, $\sum_{j=0}^{\infty} a_j^2 < \infty$ and the innovations, $\{\varepsilon_i\}_{i\in\mathbb{Z}}$, are iid mean-zero Gaussian random variables with variance $\text{var}(\varepsilon_1) = \sigma_\varepsilon^2 < \infty$.

When $\sum_{j=0}^{\infty} |a_j| < \infty$ and $\sum_{j=0}^{\infty} a_j \neq 0$ then we are in the SRD setting of Section 3.4. If instead,

$$a_j \sim j^{d-1} L_a(j) \quad \text{and} \quad 0 < d < \frac{1}{2}, \tag{3.15}$$

where L_a is a slowly varying function at infinity[5] then we have a *long memory process*. In particular, an implication of (3.15) is that $\sum_{j=0}^{\infty} a_j = \infty$ and by Karamata's theorem (see for example Resnick (1987, Theorem 0.6)),

$$\gamma(h) = \mathbb{E}(X_1 X_{1+h}) \sim h^{-D} L_\gamma(h),$$

where $D = 1 - 2d$ is the autocovariance decay rate, $0 < D < 1$, and

$$L_\gamma(h) = \sigma_\varepsilon^2 L_a^2(h) k(D),$$

where, $k(D) = \text{Beta}(D, d) = \Gamma(D)\Gamma(d)/\Gamma(D + d)$.

An important family of LRD processes are the fractionally integrated autoregressive moving average (ARFIMA) processes (Granger and Joyeux, 1980; Hosking, 1981). The ARFIMA$(0, d, 0)$ model is defined as,

$$X_i = \sum_{j=0}^{\infty} a_j \varepsilon_{i-j} = (1 - \mathrm{B})^{-d} \varepsilon_i,$$

where B is the backshift operator, $\mathrm{B}\,\varepsilon_i = \varepsilon_{i-1}$. Consider the series expansion,

$$(1 - z)^{-d} = \sum_{j=0}^{\infty} (-1)^j \binom{-d}{j} z^j = \sum_{j=0}^{\infty} a_j z^j.$$

[5] A function L is called slowly varying (at infinity in Karamata's sense) if it is positive and measurable for large enough x, and for any $u > 0$, $L(ux) \sim L(x)$ as $x \to \infty$.

Noting the relation $\Gamma(z-n) = (-1)^n\Gamma(z)\Gamma(1-z)/\Gamma(n-z+1)$, for $n \in \mathbb{Z}$, we have,

$$a_j = (-1)^j \binom{-d}{j} = \frac{(-1)^j\Gamma(1-d)}{\Gamma(j+1)\Gamma(1-d-j)} = \frac{\Gamma(j+d)}{\Gamma(j+1)\Gamma(d)}.$$

Applying Stirling's approximation, $\Gamma(s+1) \approx (2\pi s)^{1/2}e^{-s}s^s$ for large s gives,

$$a_j \sim \frac{1}{\Gamma(d)} j^{d-1}, \quad \text{as } j \to \infty.$$

Hence, $L_a \sim 1/\Gamma(d)$, and $\gamma(h) \sim c_\gamma h^{-D}$, with

$$c_\gamma = \frac{\sigma_\varepsilon^2}{\Gamma^2(d)} k(D) = \sigma_\varepsilon^2 \frac{\Gamma(D)}{\Gamma(d)\Gamma(1-d)}.$$

Beran (1994, p. 63) provides exact results for the autocovariance function of an ARFIMA$(0,d,0)$,

$$\gamma(h) = \sigma_\varepsilon^2 \frac{(-1)^h\Gamma(D)}{\Gamma(1+h-d)\Gamma(1-h-d)} = \sigma_\varepsilon^2 \frac{\Gamma(h+d)\Gamma(D)}{\Gamma(d)\Gamma(1-d)\Gamma(1+h-d)},$$

with an important special case being,

$$\gamma(0) = \sigma_\varepsilon^2 \frac{\Gamma(D)}{\Gamma^2(1-d)}.$$

Another popular parameterisation of LRD sequences is the Hurst parameter or self-similarity parameter $H = d + \frac{1}{2}$, however, for the remainder of this section, we will employ the autocovariance decay rate parameterisation and base results on the following assumption.

Assumption 3.2. *$\{X_i\}_{i\geq 1}$ is a stationary mean-zero Gaussian process with autocovariance sequence $\gamma(h) = \mathbb{E}(X_1X_{h+1})$ satisfying $\gamma(h) = h^{-D}L_\gamma(h)$ for $0 < D < 1$, where L_γ is slowly varying at infinity and is positive for large h.*

3.5.2 *Hoeffding versus Hermite decomposition*

Throughout this section, we rely heavily on the Hoeffding and Hermite decompositions of the empirical distribution function of the pairwise means outlined in Section 2.2. Recall, $h(x,y;t) = \mathbb{I}\{g(x,y) \leq t\}$ and,

$$G_n(t) - G(t) = \frac{1}{n(n-1)} \sum_{i \neq j} \left[h(X_i, X_j; t) - G(t)\right]. \tag{3.16}$$

In the applications that follow it is sufficient to consider the cases where the Hermite rank is $m = 1$ or 2 as this covers the specific estimators of interest. For LRD processes, when $D < \frac{1}{m}$, and $m = 1$ or 2, the remainder term in the Hoeffding decomposition of (3.16), stated explicitly in equation (2.4), is no longer of lower order with respect to the leading term, W_n, which means that the limiting distribution is not solely determined by W_n. Hence to study the case where $D < \frac{1}{m}$ we decompose G_n in terms of the Hermite polynomials,

$$[G_n(t) - G(t)] = \frac{2}{n(n-1)}\widetilde{W}_n(t) + \widetilde{R}_n(t), \tag{3.17}$$

where

$$\widetilde{W}_n(t) = \sum_{i<j} \sum_{\substack{p,q\geq 0 \\ p+q\leq m}} \frac{\alpha_{p,q}(t)}{p!q!} H_p(X_i)H_q(X_j), \tag{3.18}$$

and

$$\widetilde{R}_n(t) = \sum_{i<j} \sum_{\substack{p,q\geq 0 \\ p+q> m}} \frac{\alpha_{p,q}(t)}{p!q!} H_p(X_i)H_q(X_j). \tag{3.19}$$

In general, the limiting behaviour of the quantile $G_n^{-1}(p)$ will depend on the Hermite rank m and the range of the index D associated with the underlying dependence structure (Beran, 1994).

Some existing results from Lévy-Leduc et al. (2011a,b,c) employing these decompositions are briefly stated in Appendix C. These include some technical limit results for U-processes under long range dependence with applications to the scale estimator Q_n. The remainder of this section is dedicated to establishing similar results for P_n. Note that the key difference is that Q_n is a single U-quantile statistic whereas P_n is a linear combination of U-quantiles which appears to complicate the derivation of limit results.

In obtaining the limiting distribution for P_n we need to link the empirical distribution function of the adjusted pairwise means to its inverse using the functional delta method (van der Vaart, 1998, Theorem 20.8), which Hössjer and Mielniczuk (1995) show is applicable under long range dependence. However, when $0 < D < \frac{1}{2}$ it turns out that when we take linear combinations of these quantities, a first order approximation is no longer sufficient as the first order terms cancel, necessitating a higher order approximation. Thus the higher order terms now play a role in the limiting distribution. The need for higher order terms is highlighted in the following section where we find the limiting distribution of the IQR under LRD with $0 < D < \frac{1}{2}$. To do this we apply the Bahadur representation of Wu (2005) for quantiles of LRD processes.

3.5.3 *Interquartile range*

Given a set of data, $\boldsymbol{x} = \{X_i\}_{1 \leq i \leq n}$, the IQR is defined as,

$$T_n(\boldsymbol{x}) = F_n^{-1}(\tfrac{3}{4}) - F_n^{-1}(\tfrac{1}{4}), \tag{3.20}$$

where $F_n(t) = n^{-1} \sum_{i=1}^{n} \mathbb{I}\{X_i \leq t\}$. Furthermore, define the scale parameter $\theta = F^{-1}(\frac{3}{4}) - F^{-1}(\frac{1}{4})$.

The aim of this section is to prove the following result.

Result 3.3. *Under Assumption 3.2 with* $0 < D < \frac{1}{2}$*, as* $n \to \infty$*,*

$$\frac{k(D)n^D}{L_\gamma(n)}(T_n - \theta) \xrightarrow{\mathcal{D}} \frac{\theta}{2}\left(Z_{2,D}(1) - Z_{1,D}^2(1)\right).$$

The limit processes in Result 3.3 are the standard fractional Brownian motion $\{Z_{1,D}(t)\}_{0 \leq t \leq 1}$ and the Rosenblatt process $\{Z_{2,D}(t)\}_{0 \leq t \leq 1}$.

The Hermite expansion for $h(X;t) = \mathbb{I}\{X \leq t\}$, for $t \in \mathbb{R}$ is,

$$\mathbb{I}\{X \leq t\} = \sum_{k=0}^{\infty} \frac{\alpha_k(t)}{k!} H_k(X),$$

where the first three Hermite coefficients, $\alpha_0(t) = \Phi(t)$, $\alpha_1(t) = -\phi(t)$ and $\alpha_2(t) = -t\phi(t)$ are found explicitly in Section C.2.2 in Appendix C.

Now the empirical distribution function can be written in terms of this Hermite expansion,

$$\begin{aligned} F_n(t) &= \frac{1}{n} \sum_{i=1}^{n} \sum_{k=0}^{\infty} \frac{\alpha_k(t)}{k!} H_k(X_i) \\ &= \alpha_0(t) + \frac{1}{n} \sum_{i=1}^{n} \alpha_1(t) X_i + \frac{1}{n} \sum_{i=1}^{n} \frac{\alpha_2(t)}{2}(X_i^2 - 1) + R_n(t), \end{aligned}$$

where

$$R_n(t) = \frac{1}{n} \sum_{i=1}^{n} \sum_{k \geq 3} \frac{\alpha_k(t)}{k!} H_k(X_i).$$

Hence, we have,

$$\Phi(t) - F_n(t) = \phi(t)\bar{X}_n + \frac{t\phi(t)}{2} \frac{1}{n} \sum_{i=1}^{n} (X_i^2 - 1) - R_n(t). \tag{3.21}$$

The convergence of partial sums under long range dependence is given in Taqqu (1975) using the reduction principle, where for $0 < D < \frac{1}{m}$ only the first non-zero term in the Hermite expansion determines the asymptotic distribution. That is, for any fixed t, the limiting distribution of,

$$\frac{n^{mD/2}}{n\sqrt{L_m(n)}}\sum_{i=1}^{n}[h(X_i;t)-\Phi(t)],$$

is the same as the limiting distribution of,

$$\frac{n^{mD/2}}{n\sqrt{L_m(n)}}\sum_{i=1}^{n}\frac{\alpha_m(t)}{m!}H_m(X_i),$$

where

$$L_m(n)=\frac{2m!L_\gamma^m(n)}{(Dm+1)(Dm+2)}.$$

The Hermite rank of the class of functions $\{h(\cdot;t)-\Phi(t);\ t\in\mathbb{R}\}$ is equal to one as $\alpha_1(t)\neq 0$ for all $t\in\mathbb{R}$. So it is only the first term in the expansion that determines the limiting distribution of the quantiles.

From (3.21), we can apply the reduction principle result with $m=2$ to get,

$$\frac{n^D}{L_\gamma(n)}\left([F_n(t)-\Phi(t)]+\phi(t)\bar{X}_n+\frac{t\phi(t)}{2}\frac{1}{n}\sum_{i=1}^{n}(X_i^2-1)\right)\to 0,$$

in probability, or equivalently, $\frac{n^D}{L_\gamma(n)}R_n(t)=o_p(1)$.

When investigating the difference of two quantiles, we will see that the first two Hermite polynomials in the expansion (3.21) are required. To establish this, we use the Bahadur representation of sample quantiles of LRD processes established by Wu (2005).

Before stating the Bahadur result we need to introduce some additional notation. Under Assumption 3.2, let $\Upsilon_n=\sqrt{n}\sum_{k=1}^{n}k^{-D-1/2}L_a^2(k)$ and

$$\sigma_{n,1}^2=\mathbb{E}(|n\bar{X}_n|^2)=\text{var}\left(\sum_{i=1}^{n}X_i\right)\sim\frac{L_\gamma(n)n^{2-D}}{(1-D)(2-D)}. \tag{3.22}$$

For the derivation of the above result, see, for example Beran et al. (2013, Lemma 4.9). By Karamata's theorem,

$$\Upsilon_n\sim\begin{cases}\dfrac{n^{1-D}L_a^2(n)}{1/2-D} & \text{if } 0<D<\frac{1}{2},\\ \sqrt{n}L^*(n) & \text{if } D=\frac{1}{2},\text{ or}\\ \sqrt{n}\sum_{k=1}^{\infty}k^{-D-1/2}L_a^2(k) & \text{if } \frac{1}{2}<D<1,\end{cases}$$

where $L^*(n) = \sum_{k=1}^{n} L_a^2(k)/k$ is also a slowly varying function. Furthermore, let,

$$A_n(D) = \begin{cases} \Psi_n^2(\log n)(\log\log n)^2 & \text{if} \quad 0 < D < \frac{1}{2}, \text{ or} \\ \Psi_n^2(\log n)^3(\log\log n)^2 & \text{if} \quad \frac{1}{2} \leq D < 1. \end{cases}$$

We can now write down Theorem 3 from Wu (2005) adapted under the assumption of a Gaussian LRD process.

Theorem 3.1. *Assume* $\inf_{p_0 \leq p \leq p_1} f(F^{-1}(p)) > 0$ *for some* $0 < p_0 < p_1 < 1$, *and let* $b_n = \sigma_{n,1}(\log n)^{1/2}(\log\log n)/n$, *then,*

$$\sup_{p_0 \leq p \leq p_1} \left| F_n^{-1}(p) - F^{-1}(p) - \frac{p - F_n(F^{-1}(p))}{f(F^{-1}(p))} - \frac{\bar{X}_n^2}{2} \frac{f'(F^{-1}(p))}{f(F^{-1}(p))} \right|$$
$$= O_{a.s.} \left[b_n^3 + \frac{\sqrt{b_n \log n}}{\sqrt{n}} + \frac{b_n \sqrt{A_n(D)}}{n} \right].$$

As Wu (2005) notes, the three terms in the $O_{a.s.}$ bound have different orders of magnitude and correspondingly the term that dominates the bound changes depending on the range of D. The error bound of Theorem 3.1 is summarised as,

$$O[n^{\max(-D/4-1/2, -3D/2)} L_1(n)] = O[n^{h(D)} L_1(n)],$$

for some slowly varying function $L_1(n)$.

We can now prove Result 3.3. By Theorem 3.1 we can say,

$$F_n^{-1}(p) - r = \frac{p - F_n(r)}{\phi(r)} + \frac{\bar{X}_n^2}{2} \frac{\phi'(r)}{\phi(r)} + O(n^{h(D)} L_1(n)).$$

Noting that $\phi'(r)/\phi(r) = -r$, we have

$$F_n^{-1}(p) - r = \frac{p - F_n(r)}{\phi(r)} - r\frac{\bar{X}_n^2}{2} + O(n^{h(D)} L_1(n)).$$

Furthermore, using equation (3.21), we have

$$p - F_n(r) = \phi(r)\bar{X}_n + \frac{r\phi(r)}{2} \frac{1}{n} \sum_{i=1}^{n} (X_i^2 - 1) - R_n(r).$$

Turning attention to the IQR, let $s = F^{-1}(\frac{1}{4})$ and $t = F^{-1}(\frac{3}{4})$,

$$\begin{aligned}
&[F_n^{-1}(\tfrac{3}{4}) - F_n^{-1}(\tfrac{1}{4})] - [t-s] \qquad (3.23)\\
&= \left(\frac{\frac{3}{4} - F_n(t)}{\phi(t)} - t\frac{\bar{X}_n^2}{2}\right) - \left(\frac{\frac{1}{4} - F_n(s)}{\phi(s)} - s\frac{\bar{X}_n^2}{2}\right) + O(n^{h(D)}L_1(n))\\
&= \left(\frac{\phi(t)\bar{X}_n + \frac{t\phi(t)}{2}\frac{1}{n}\sum_{i=1}^n (X_i^2 - 1) - R_n(t)}{\phi(t)} - t\frac{\bar{X}_n^2}{2}\right)\\
&\quad - \left(\frac{\phi(s)\bar{X}_n + \frac{s\phi(s)}{2}\frac{1}{n}\sum_{i=1}^n (X_i^2 - 1) - R_n(s)}{\phi(s)} - s\frac{\bar{X}_n^2}{2}\right) + O(n^{h(D)}L_1(n))\\
&= \left(\frac{t}{2n}\sum_{i=1}^n (X_i^2 - 1) - \frac{t}{2}\bar{X}_n^2\right) - \frac{R_n(t)}{\phi(t)}\\
&\quad - \left(\frac{s}{2n}\sum_{i=1}^n (X_i^2 - 1) - \frac{s}{2}\bar{X}_n^2\right) + \frac{R_n(s)}{\phi(s)} + O(n^{h(D)}L_1(n))\\
&= \frac{t-s}{2n^2}\left(n\sum_{i=1}^n (X_i^2 - 1) - \sum_{1\le i,j\le n} X_i X_j\right) - \frac{R_n(t)}{\phi(t)} + \frac{R_n(s)}{\phi(s)} + O(n^{h(D)}L_1(n)). \qquad (3.24)
\end{aligned}$$

Now, multiply both sides of (3.24) by $n^D / L_\gamma(n)$, so that the left hand side becomes,

$$\frac{n^D}{L_\gamma(n)}\left([F_n^{-1}(\tfrac{3}{4}) - F_n^{-1}(\tfrac{1}{4})] - [t-s]\right).$$

Then, on the RHS we know that $\frac{n^D}{L_\gamma(n)}R_n(t) = o_p(1)$ and $\frac{n^D}{L_\gamma(n)}R_n(s) = o_p(1)$. Furthermore,

$$\frac{n^D}{L_\gamma(n)}n^{h(D)}L_1(n) = O\left(n^{\max\{-D/2, 3D/4-1/2\}}\frac{L_1(n)}{L_\gamma(n)}\right) \to 0, \quad \text{as } n \to \infty,$$

for $0 < D < \frac{1}{2}$. Hence,

$$\frac{k(D)n^D}{L_\gamma(n)}([F_n^{-1}(\tfrac{3}{4}) - F_n^{-1}(\tfrac{1}{4})] - [t-s])$$

can be written as

$$\frac{k(D)n^D}{n^2 L_\gamma(n)}\frac{t-s}{2}\left[n\sum_{i=1}^n (X_i^2 - 1) - \sum_{1\le i,j\le n} X_i X_j\right] + o_p(1),$$

which, by Lemma C.4 in Appendix C, gives the required proof.

3.5.4 *Results for P_n*

This section establishes the following result, which is not a trivial application of Lévy-Leduc et al. (2011c), owing to the complex nature of P_n as the difference of two U-quantiles.

Result 3.4. *Under Assumption 3.2, if $D > \frac{1}{2}$, as $n \to \infty$,*

$$\sqrt{n}(P_n - \sigma) \xrightarrow{\mathcal{D}} \mathcal{N}(0, c^2 V),$$

where $\sigma = \sqrt{\gamma(0)}$ and

$$c^2 V = \mathbb{E}\left(\mathrm{IF}^2(X_1; P, \Phi)\right) + 2 \sum_{\ell \geq 1} \mathbb{E}\left(\mathrm{IF}(X_1; P, \Phi)\, \mathrm{IF}(X_{\ell+1}; P, \Phi)\right).$$

Note that this is exactly the same as for the SRD setting, however the proof is different.

We also posit the following conjecture, the proof of which requires a U-quantile analogue of Wu (2005).

Conjecture 3.1. *Under Assumption 3.2, if $D < \frac{1}{2}$, as $n \to \infty$,*

$$\frac{k(D)n^D}{L_\gamma(n)}(P_n - \sigma) \xrightarrow{\mathcal{D}} \frac{\sigma}{2}\left(Z_{2,D}(1) - Z_{1,D}^2(1)\right).$$

The conjectured limiting distribution is a linear combination of the standard fractional Brownian motion $\{Z_{1,D}(t)\}_{0 \leq t \leq 1}$ and the Rosenblatt process $\{Z_{2,D}(t)\}_{0 \leq t \leq 1}$.

When written as a GL-statistic, the scale estimator P_n shares the same kernel function as the Hodges-Lehmann estimator of location. The asymptotic properties of the Hodges-Lehmann estimator when applied to a LRD process are discussed in Section C.4.2.1 in Appendix C. In particular, we show that estimated central quantiles[6] of a LRD process follow a Gaussian distribution. Hence, one may naïvely believe that the autocovariance estimator based on P_n for LRD data will similarly follow a Gaussian distribution as it is proportional to the difference of two quantiles of the pairwise mean distribution, each of which are Gaussian for all $D \in (0, 1)$. However, it turns out that this is not the case – instead, as with the SD and Q_n we need to consider two cases, $D < \frac{1}{2}$ and $D > \frac{1}{2}$, and we need to employ some subtle arguments

[6] It is important to note that the quantiles referred to do not include extremes, such as the minima and maxima, as these will have their own limiting behaviour. Specifically the technical results in Section C.4.1 in Appendix C rely on the quantiles lying in a compact interval of $\mathbb{R}$.

in the derivation of the limiting distribution in each case. The first trick will be to adjust the pairwise mean kernel function so that we can more easily apply existing results.

3.5.4.1 *Augmenting the kernel*

Section C.4.2.1 in Appendix C discusses the Hodges-Lehmann estimator with pairwise mean kernel, $h(x, y; r) = \mathbb{I}\{x + y \leq 2r\}$, and notes that the Hermite rank for the class of functions $\{h(\cdot, \cdot; r) - G(r);\, r \in \mathbb{R}\}$ equals one. The aim of this section is to manipulate the kernel so that we can apply existing results and rely on the symmetry in the difference of the quantiles to erase the spurious artefacts created by the augmented kernel.

Consider the augmented kernel,

$$h^*(x, y; r) = \mathbb{I}\{x + y \leq 2r\} + (x + y)\frac{\phi(r\sqrt{2})}{\sqrt{2}}. \tag{3.25}$$

Let X and Y be independent standard Gaussian random variables. Hence, $\alpha^*_{0,0}(r) = G^*(r) = \mathbb{E}h^*(X, Y; r) = G(r) = \alpha_{0,0}(r)$. However, the Hermite rank for the class of functions $\{h^*(\cdot, \cdot; r) - G^*(r); r \in I\}$ is $m = 2$ as

$$\begin{aligned}
\alpha^*_{1,0}(r) &= \mathbb{E}[Xh^*(X, Y; r)] \\
&= \mathbb{E}[Xh(X, Y; r)] + \mathbb{E}[X^2 + XY]\frac{\phi(r\sqrt{2})}{\sqrt{2}} \\
&= \alpha_{1,0}(r) + \frac{\phi(r\sqrt{2})}{\sqrt{2}} = 0,
\end{aligned}$$

using Result C.2 in Appendix C.2.3. Similarly, $\alpha^*_{0,1}(r) = 0$, however,

$$\begin{aligned}
\alpha^*_{1,1}(r) &= \mathbb{E}[XYh^*(X, Y; r)] \\
&= \mathbb{E}[XYh(X, Y; r)] + \mathbb{E}[X^2Y + XY^2]\frac{\phi(r\sqrt{2})}{\sqrt{2}} \\
&= \alpha_{1,1}(r) \neq 0,
\end{aligned}$$

using Result C.3 in Appendix C.2.3. Furthermore,

$$\begin{aligned}
\alpha^*_{2,0}(r) &= \mathbb{E}[(X^2 - 1)h^*(X, Y; r)] \\
&= \mathbb{E}[(X^2 - 1)h(X, Y; r)] + \mathbb{E}[(X^2 - 1)(X + Y)]\frac{\phi(r\sqrt{2})}{\sqrt{2}} \\
&= \alpha_{2,0}(r) = \alpha_{0,2}(r) \neq 0.
\end{aligned}$$

Consider the Hoeffding decomposition of the augmented kernel, (3.25), as follows,

$$\begin{aligned} G_n^*(r) - G(r) &= \frac{1}{n(n-1)} \sum_{i \neq j} \left[h^*(X_i, X_j; r) - G(r) \right] \\ &= W_n^*(r) + R_n^*(r), \end{aligned}$$

where

$$W_n^*(r) = \frac{2}{n} \sum_{i=1}^{n} h_1^*(X_i; r), \tag{3.26}$$

and

$$\begin{aligned} h_1^*(x;r) &= \int h(x,y;r)^* \phi(y)\,\mathrm{d}y - G(r) \\ &= \int \left[\mathbb{I}\{x + y \leq 2r\} + \frac{x}{\sqrt{2}} \phi(r\sqrt{2}) + \frac{y}{\sqrt{2}} \phi(r\sqrt{2}) \right] \phi(y)\,\mathrm{d}y - G(r) \\ &= h_1(x;r) + x \frac{\phi(r\sqrt{2})}{\sqrt{2}}, \end{aligned}$$

with $h_1(x;r)$ as defined in Section 2.2. Furthermore,

$$\begin{aligned} R_n^*(r) &= \frac{1}{n(n-1)} \sum_{i \neq j} \left[h^*(X_i, X_j; r) - h_1^*(X_i; r) - h_1^*(X_j; r) - G(r) \right] \\ &= R_n(r), \end{aligned}$$

as defined in Section 2.2. Now consider,

$$\begin{aligned} \frac{G_n^*(r) - G(r)}{G'(r)} &= \frac{W_n^*(r) + R_n^*(r)}{G'(r)} \\ &= \frac{\frac{2}{n} \sum_{i=1}^{n} h_1^*(X_i; r) + R_n(r)}{G'(r)} \\ &= \frac{\frac{2}{n} \sum_{i=1}^{n} \left[h_1(X_i; r) + \phi(r\sqrt{2}) X_i / \sqrt{2} \right] + R_n(r)}{G'(r)} \\ &= \frac{\frac{2}{n} \sum_{i=1}^{n} h_1(X_i; r) + R_n(r)}{G'(r)} + \frac{2\phi(r\sqrt{2})}{nG'(r)\sqrt{2}} \sum_{i=1}^{n} X_i \\ &= \frac{G_n(r) - G(r)}{G'(r)} + \frac{1}{n} \sum_{i=1}^{n} X_i. \end{aligned}$$

The equality holds as $R_n^*(r) = R_n(r)$ and using Result C.1 from Appendix C, $G'(r) = \sqrt{2}\phi(t\sqrt{2})$.

We can thus deduce that $[G_n^{*-1}(p) - G_n^{*-1}(q)]$ and $[G_n^{-1}(p) - G_n^{-1}(q)]$ have the same asymptotic distribution. Letting $t = G^{-1}(p)$, $s = G^{-1}(q)$,

$$\begin{aligned} \frac{G_n^*(t) - G(t)}{G'(t)} - \frac{G_n^*(s) - G(s)}{G'(s)} &= \frac{G_n(t) - G(t)}{G'(t)} + \bar{X}_n - \frac{G_n(s) - G(s)}{G'(s)} - \bar{X}_n \\ &= \frac{G_n(t) - G(t)}{G'(t)} - \frac{G_n(s) - G(s)}{G'(s)}. \end{aligned} \tag{3.27}$$

Now, consider the functional delta method (van der Vaart, 1998, Theorem 20.8), for an appropriate sequence, a_n, we have,

$$a_n\left(G_n^{*-1}(p) - G^{-1}(p)\right) = -a_n\left[\frac{G_n^*(G^{-1}(p)) - G(G^{-1}(p))}{G'(G^{-1}(p))}\right] + o_p(1). \tag{3.28}$$

Hence, using (3.27) with the delta method applied to,

$$a_n\left([G_n^{*-1}(p) - G^{-1}(p)] - [G_n^{*-1}(q) - G^{-1}(q)]\right),$$

we can conclude that $[G_n^{*-1}(p) - G_n^{*-1}(q)]$ and $[G_n^{-1}(p) - G_n^{-1}(q)]$ have the same asymptotic distribution.

3.5.4.2 P_n when $D > \frac{1}{2}$

As the Hermite rank of the augmented kernel is $m = 2$, we can employ Lemma 9 from Lévy-Leduc et al. (2011a) which states that, $(\sqrt{n}W_n^*(r))_{r\in I}$ converges weakly in $\mathcal{D}(I)$ to a zero mean Gaussian process, W, with covariance structure $\mathbb{E}[W(s)W(t)]$ given by,

$$\begin{aligned} &4\,\mathrm{cov}(h_1^*(X_1;s), h_1^*(X_1;t)) \\ &\quad + 4\sum_{\ell\geq 1}\left[\mathrm{cov}(h_1^*(X_1;s), h_1^*(X_{\ell+1};t)) + \mathrm{cov}(h_1^*(X_1;t), h_1^*(X_{\ell+1};s))\right]. \end{aligned}$$

Furthermore, Lévy-Leduc et al. (2011a, Theorem 1) have that the U-process $\left(\sqrt{n}(G_n^*(r) - G(r))\right)_{r\in I}$ also converges in $\mathcal{D}(I)$ to W. Hence,

$$\sqrt{n}\left(G_n^{*-1}(\cdot) - G^{-1}(\cdot)\right) \xrightarrow{\mathcal{D}} \frac{W(G^{-1}(\cdot))}{G'(G^{-1}(\cdot))}.$$

Therefore,

$$\sqrt{n}\left([G_n^{*-1}(p) - G_n^{*-1}(q)] - [G^{-1}(p) - G^{-1}(q)]\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, V),$$

where $V = v(p,p) + v(q,q) - 2v(p,q)$ and $v(p,q) = n\,\mathrm{cov}(G_n^{-1}(p), G_n^{-1}(q))$ is defined as

$$\begin{aligned} \frac{4}{G'(s)G'(t)}\Bigg[&\mathrm{cov}(h_1^*(X_1;s), h_1^*(X_1;t)) \\ &+ \sum_{\ell\geq 1}\left[\mathrm{cov}(h_1^*(X_1;s), h_1^*(X_{\ell+1};t)) + \mathrm{cov}(h_1^*(X_1;t), h_1^*(X_{\ell+1};s))\right]\Bigg], \end{aligned}$$

with $t = G^{-1}(p)$ and $s = G^{-1}(q)$. Hence, as $[G_n^{*-1}(p) - G_n^{*-1}(q)]$ and $[G_n^{-1}(p) - G_n^{-1}(q)]$ have the same asymptotic distribution, we can also conclude that,

$$\sqrt{n}([G_n^{-1}(p) - G_n^{-1}(q)] - [G^{-1}(p) - G^{-1}(q)]) \xrightarrow{\mathcal{D}} \mathcal{N}(0, V).$$

To complete the proof, note that

$$\begin{aligned}
&\operatorname{cov}(h_1^*(X_1; s), h_1^*(X_1; t)) \\
&\quad = \mathbb{E}\Bigg[\left(\Phi(2s - X_1) + X_1\frac{\phi(s\sqrt{2})}{\sqrt{2}} - \mathbb{E}\Phi(2s - X_1)\right) \times \\
&\qquad\qquad \left(\Phi(2t - X_1) + X_1\frac{\phi(t\sqrt{2})}{\sqrt{2}} - \mathbb{E}\Phi(2t - X_1)\right)\Bigg] \\
&\quad = \mathbb{E}[\Phi(2s - X_1)\Phi(2t - X_1)] + \frac{\phi(s\sqrt{2})}{\sqrt{2}}\mathbb{E}[X_1\Phi(2t - X_1)] \\
&\qquad - \mathbb{E}\Phi(2t - X_1)\mathbb{E}\Phi(2s - X_1) + \frac{\phi(t\sqrt{2})}{\sqrt{2}}\mathbb{E}[X_1\Phi(2s - X_1)] \\
&\qquad + \frac{\phi(s\sqrt{2})\phi(t\sqrt{2})}{2}.
\end{aligned}$$

From Lemma C.1, $\mathbb{E}(\Phi(2G^{-1}(a) - X)) = a$ and restricting attention to $p = \frac{3}{4}$ and $q = \frac{1}{4}$ with $t = G^{-1}(\frac{3}{4})$ and $s = G^{-1}(\frac{1}{4})$, we have from Lemma C.2 that $\mathbb{E}(X\Phi(2t - X)) = \mathbb{E}(X\Phi(2s - X))$. Furthermore, from the symmetry of the Gaussian density we have $s = -t$ and $\phi(s) = \phi(t)$, hence we can simplify further,

$$\begin{aligned}
\operatorname{cov}(h_1^*(X_1; s), h_1^*(X_1; t)) &= \mathbb{E}[\Phi(2s - X_1)\Phi(2t - X_1)] + \tfrac{3}{16} \\
&\qquad + \sqrt{2}\phi(s\sqrt{2})\mathbb{E}[X_1\Phi(2s - X_1)] + \frac{\phi^2(s\sqrt{2})}{2}.
\end{aligned}$$

Thus,

$$\begin{aligned}
&\operatorname{var}(h_1^*(X_1; s)) + \operatorname{var}(h_1^*(X_1; t)) - 2\operatorname{cov}(h_1^*(X_1; s), h_1^*(X_1; t)) \\
&\quad = \mathbb{E}\left[\Phi(2s - X)\right]^2 + \mathbb{E}\left[\Phi(2t - X)\right]^2 - 2\mathbb{E}\left[\Phi(2s - X)\Phi(2t - X)\right] - \tfrac{1}{4} \\
&\quad = \mathbb{E}\left[\tfrac{1}{2} - \Phi(2t - X) + \Phi(2s - X)\right]^2.
\end{aligned}$$

For a particular lag, ℓ, using similar results to those implemented above, elementary, though tedious, algebra shows that

$$\begin{aligned}
&2\operatorname{cov}(h_1^*(X_1; t), h_1^*(X_{\ell+1}; t)) + 2\operatorname{cov}(h_1^*(X_1; s), h_1^*(X_{\ell+1}; s)) \\
&- 2\left[\operatorname{cov}(h_1^*(X_1; s), h_1^*(X_{\ell+1}; t)) + \operatorname{cov}(h_1^*(X_1; t), h_1^*(X_{\ell+1}; s))\right] \\
&= \mathbb{E}\left(\frac{1}{2} - \Phi(2t - X_1) + \Phi(2s - X_1)\right)\left(\frac{1}{2} - \Phi(2t - X_{\ell+1}) + \Phi(2s - X_{\ell+1})\right).
\end{aligned}$$

Hence,

$$c^2 V = \mathbb{E}\left(\mathrm{IF}^2(X_1; P, \Phi)\right) + 2\sum_{\ell \geq 1} \mathbb{E}\left(\mathrm{IF}(X_1; P, \Phi)\,\mathrm{IF}(X_{\ell+1}; P, \Phi)\right),$$

where $\mathrm{IF}(x; P, \Phi)$ was defined in (3.5). The $D > \frac{1}{2}$ case in Result 3.4 follows.

3.5.4.3 P_n *when* $D < \frac{1}{2}$

This section considers the limiting distribution of P_n under LRD when $D < \frac{1}{2}$ assuming that an unproven conjecture holds true. The basic idea is the same as for the IQR, which required a Bahadur representation for quantiles under long range dependence. To our knowledge, an equivalent result for U-quantiles is still to be established.

Consider the augmented kernel defined in (3.25) and write the Hermite expansion of $G_n^*(r) - G(r)$ as,

$$G_n^*(r) - G(r) = \frac{1}{n(n-1)}\left[\widetilde{W}_n^*(r) + \widetilde{R}_n^*(r)\right], \tag{3.29}$$

where

$$\widetilde{W}_n^*(r) = \alpha_{1,1}^*(r)\sum_{i \neq j} X_i X_j + \alpha_{2,0}^*(r)\sum_{i \neq j}(X_i^2 - 1),$$

and

$$\widetilde{R}_n^*(r) = \sum_{i \neq j}\sum_{\substack{p,q \geq 0 \\ p+q > 2}} \frac{\alpha_{p,q}^*(r)}{p!q!} H_p(X_i) H_q(X_j).$$

There has been some progress towards a Bahadur representation for U-quantiles under strongly dependent sequences. In particular, Wendler (2011, 2012) has a Bahadur representation for U-quantiles and GL-statistics of strongly mixing random variables and functionals of absolutely regular sequences. Whilst these are not long range dependent sequences, the Hodges-Lehmann estimator is considered and the pairwise mean kernel function is shown to satisfy the required assumptions. Given the regularity of the pairwise mean kernel, it is not unreasonable to assume that it would also satisfy the assumptions for a Bahadur representation of the corresponding U-quantile under LRD. The following conjecture, has a similar flavour to the Bahadur representation for quantiles under LRD given by Wu (2005) in Theorem 3.1.

Conjecture 3.2. *Under Assumption 3.2, with* $h(x, y; r) = \mathbb{I}\{x + y \leq 2r\}$,

$$G_n^{-1}(p) - r = \frac{p - G_n(r)}{G'(r)} + \frac{\bar{X}_n^2}{2}\frac{G''(r)}{G'(r)} + O(n^{h(D)} L_1(n))$$

where $h(D) + D < 0$ *for* $0 < D < \frac{1}{2}$ *and* $L_1(n)$ *is some slowly varying function.*

Noting that $G''(r)/G'(r) = -2r$, letting $t = G^{-1}(\frac{3}{4})$ and $s = G^{-1}(\frac{1}{4})$, assuming Conjecture 3.2 to be true and using the Hermite expansion from (3.29) we have,

$$\begin{aligned}
&\left(G_n^{*-1}(\tfrac{3}{4}) - G^{-1}(\tfrac{3}{4})\right) - \left(G^{-1}(\tfrac{1}{4}) - G_n^{*-1}(\tfrac{1}{4})\right) \\
&\quad= \left(\frac{G(t) - G_n^*(t)}{G'(t)} - t\bar{X}_n^2\right) - \left(\frac{G_n^*(s) - G(s)}{G'(s)} - s\bar{X}_n^2\right) + O(n^{h(D)}L_\gamma(n)) \\
&\quad= \frac{\widetilde{W}_n^*(s) + \widetilde{R}_n^*(s)}{n(n-1)G'(s)} - \frac{\widetilde{W}_n^*(t) + \widetilde{R}_n^*(t)}{n(n-1)G'(t)} + (s-t)\bar{X}_n^2 + O(n^{h(D)}L_\gamma(n)) \\
&\quad= \frac{\alpha_{1,1}^*(s)\sum_{i\neq j} X_iX_j + \alpha_{2,0}^*(s)\sum_{i\neq j}(X_i^2 - 1)}{n(n-1)G'(s)} \\
&\qquad - \frac{\alpha_{1,1}^*(t)\sum_{i\neq j} X_iX_j + \alpha_{2,0}^*(t)\sum_{i\neq j}(X_i^2 - 1)}{n(n-1)G'(t)} \\
&\qquad + \frac{\widetilde{R}_n^*(s) - \widetilde{R}_n^*(t)}{n(n-1)G'(t)} - (t-s)\bar{X}_n^2 + O(n^{h(D)}L_\gamma(n)).
\end{aligned} \tag{3.30}$$

Recalling that $\alpha_{1,1}^*(r) = \alpha_{1,1}(r)$ and $\alpha_{2,0}^*(r) = \alpha_{2,0}(r)$, noting from Result C.4, $\alpha_{2,0}(r) = \alpha_{1,1}(r)$ and from Result C.3, $\frac{\alpha_{1,1}(r)}{G'(r)} = -\frac{r}{2}$, we have,

$$\begin{aligned}
&\frac{\alpha_{1,1}^*(s)\sum_{i\neq j} X_iX_j + \alpha_{2,0}^*(s)\sum_{i\neq j}(X_i^2 - 1)}{G'(s)} \\
&\qquad\qquad - \frac{\alpha_{1,1}^*(t)\sum_{i\neq j} X_iX_j + \alpha_{2,0}^*(t)\sum_{i\neq j}(X_i^2 - 1)}{G'(t)} \\
&\quad= \frac{\alpha_{1,1}(s)\left(\sum_{i\neq j} X_iX_j + \sum_{i\neq j}(X_i^2 - 1)\right)}{G'(s)} \\
&\qquad\qquad - \frac{\alpha_{1,1}(t)\left(\sum_{i\neq j} X_iX_j + \sum_{i\neq j}(X_i^2 - 1)\right)}{G'(t)} \\
&\quad= \frac{t-s}{2}\left(\sum_{i\neq j} X_iX_j + \sum_{i\neq j}(X_i^2 - 1)\right).
\end{aligned}$$

We know from Lévy-Leduc et al. (2011a, Lemma 18) that,

$$\frac{n^D}{n(n-1)L_\gamma(n)}\widetilde{R}_n^*(r) = o_p(1).$$

Furthermore, for $0 < D < \frac{1}{2}$, Conjecture 3.2 assumes that, $n^{h(D)+D} \to 0$ as $n \to \infty$. Hence, multiplying both sides of (3.30) by $n^D / L_\gamma(n)$, gives,

$$\begin{aligned}
&\frac{n^D}{L_\gamma(n)} \left(G_n^{*-1}(\tfrac{3}{4}) - G_n^{*-1}(\tfrac{1}{4}) - [G^{-1}(\tfrac{3}{4})) - G^{-1}(\tfrac{1}{4})] \right) \\
&= \frac{n^D(t-s)}{2n(n-1)L_\gamma(n)} \left[\sum_{i \neq j} X_i X_j + (n-1) \sum_{i=1}^{n} (X_i^2 - 1) - 2n(n-1)\bar{X}_n^2 \right] + o_p(1) \\
&= \frac{n^D(t-s)}{2n(n-1)L_\gamma(n)} \left[\sum_{i \neq j} X_i X_j + (n-1) \sum_{i=1}^{n} (X_i^2 - 1) - \frac{2(n-1)}{n} \sum_{i,j} X_i X_j \right] + o_p(1) \\
&= \frac{n^D(t-s)}{2n(n-1)L_\gamma(n)} \left[n \sum_{i=1}^{n} (X_i^2 - 1) - \sum_{i,j} X_i X_j - \sum_{i=1}^{n} (2X_i^2 - 1) \right] + o_p(1) \\
&= \frac{n^D(t-s)}{2n(n-1)L_\gamma(n)} \left[n \sum_{i=1}^{n} (X_i^2 - 1) - \sum_{i,j} X_i X_j \right] + o_p(1)
\end{aligned}$$

Thus, noting (3.27),

$$\frac{k(D)n^D}{L_\gamma(n)} ([G_n^{-1}(p) - G^{-1}(p)] - [G_n^{-1}(q) - G^{-1}(q)])$$

can be written as,

$$\frac{k(D)n^D}{L_\gamma(n)} \frac{t-s}{2n(n-1)} \left[n \sum_{i=1}^{n} (X_i^2 - 1) - \sum_{i,j} X_i X_j \right] + o_p(1),$$

which, similarly to the IQR case, by Lemma C.4, converges in distribution to

$$\frac{t-s}{2} \left[Z_{2,D}(1) - Z_{1,D}^2(1) \right].$$

Hence, with an appropriate correction factor, Conjecture 3.1 follows.

3.5.5 *Results for $\hat{\gamma}_P(h)$*

This section proves the following result for the autocovariance estimator, $\hat{\gamma}_P(h)$, defined by (3.9).

Result 3.5. *Under Assumption 3.2, as $n \to \infty$,*

A. *If $D > \frac{1}{2}$,*

$$\sqrt{n}(\hat{\gamma}_P(h) - \gamma(h)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \breve{\sigma}^2(h)),$$

where

$$\begin{aligned}\breve{\sigma}^2(h) = \mathbb{E}&\left[\mathrm{IF}^2(X_1, X_{1+h}; \gamma_P, \Phi)\right] \\ &+ 2\sum_{k\geq 1} \mathbb{E}\left[\mathrm{IF}(X_1, X_{1+h}; \gamma_P, \Phi)\, \mathrm{IF}(X_{k+1}, X_{k+1+h}; \gamma_P, \Phi)\right].\end{aligned}$$

B. *If $D < \frac{1}{2}$ and Conjecture 3.1 holds then,*

$$\frac{k(D)n^D}{\widetilde{L}(n)}(\hat{\gamma}_P(h) - \gamma(h)) \xrightarrow{\mathcal{D}} \frac{\gamma(0) + \gamma(h)}{2}(Z_{2,D}(1) - Z^2_{1,D}(1)),$$

where, $\widetilde{L}(n) = 2L_\gamma(n) + (1 + \frac{h}{n})^{-D} L_\gamma(n + h) + (1 - \frac{h}{n})^{-D} L_\gamma(n - h)$.

We begin by analysing the properties of the sequences $\{X_i + X_{i+h}\}_{i\geq 1}$ and $\{X_i - X_{i+h}\}_{i\geq 1}$, before applying results found earlier in this section to the resulting sequences.

The sequence $\{X_i + X_{i+h}\}_{i\geq 1}$ has a slowly decaying autocovariance function similar to $\{X_i\}_{i\geq 1}$ with $L_\gamma = L$ replaced by some slowly varying function $\widetilde{L}$. To see this, let $\gamma_+(k)$ be the autocovariance function for the sequence $\{X_i + X_{i+h}\}_{i\geq 1}$, then

$$\begin{aligned}\gamma_+(k) &= \mathbb{E}\left[(X_1 + X_{1+h})(X_{k+1} + X_{1+h+k})\right] \\ &= 2\gamma(k) + \gamma(k + h) + \gamma(k - h) \\ &= 2k^{-D} L(k) + (k + h)^{-D} L(k + h) + (k - h)^{-D} L(k - h) \\ &= k^{-D}\left[2L(k) + (1 + h/k)^{-D} L(k + h) + (1 - h/k)^{-D} L(k - h)\right] \\ &= k^{-D}\widetilde{L}(k).\end{aligned}$$

Hence results for P_n applied to $\{X_i + X_{i+h}\}_{i\geq 1}$ have the same form as those for P_n applied to $\{X_i\}_{i\geq 1}$.

On the other hand, the sequence $\{X_i - X_{i+h}\}_{i\geq 1}$ has a fast autocovariance decay rate. Consider,

$$\begin{aligned}\gamma_-(k) &= \mathbb{E}\left[(X_1 - X_{1+h})(X_{k+1} - X_{1+h+k})\right] \\ &= 2\gamma(k) - \gamma(k+h) - \gamma(k-h) \\ &= 2k^{-D}L(k) - \Big[(k+h)^{-D}L(k+h) + (k-h)^{-D}L(k-h)\Big] \\ &= 2k^{-D}L(k) - k^{-D}\Big[(1+h/k)^{-D}L(k+h) + (1-h/k)^{-D}L(k-h)\Big].\end{aligned}$$

Hence, using a binomial series (Abramowitz and Stegun, 1973, p. 822), for $|x| < 1$,

$$(1-x)^{-m-1} = \sum_{n=m}^{\infty}\binom{n}{m}x^{n-m},$$

we have,

$$\begin{aligned}(1+h/k)^{-D} &= \sum_{j=0}^{\infty}\binom{D+j-1}{j}\left(\frac{h}{k}\right)^{j}(-1)^{j} \\ &= 1 - \frac{h}{k}\binom{D}{1} + \frac{h^2}{k^2}\binom{D+1}{2} - \frac{h^3}{k^3}\binom{D+2}{3} + \ldots,\end{aligned}$$

and similarly,

$$\begin{aligned}(1-h/k)^{-D} &= \sum_{j=0}^{\infty}\binom{D+j-1}{j}\left(\frac{h}{k}\right)^{j} \\ &= 1 + \frac{h}{k}\binom{D}{1} + \frac{h^2}{k^2}\binom{D+1}{2} + \frac{h^3}{k^3}\binom{D+2}{3} + \ldots.\end{aligned}$$

Following Lévy-Leduc et al. (2011c) let us impose some quite reasonable restrictions on the slowly varying function L. Assume $L_i(x) = x^i L^{(i)}(x)$ satisfies $L_i(x)/x^{\epsilon} = O(1)$, for some $\epsilon \in (0, D)$, as $x \to \infty$ for all $i = 0, 1, 2, 3$ where $L^{(i)}$ is the ith derivative of L.

Using a series expansion, we have,

$$L(k+h) = L(k) + \frac{L^{(1)}(k)}{1!}h + \frac{L^{(2)}(k)}{2!}h^2 + \frac{L^{(3)}(k)}{3!}h^3 + \ldots,$$

and similarly,

$$L(k-h) = L(k) - \frac{L^{(1)}(k)}{1!}h + \frac{L^{(2)}(k)}{2!}h^2 - \frac{L^{(3)}(k)}{3!}h^3 + \ldots.$$

Thus,

$$\begin{aligned}
&(1+h/k)^{-D}L(k+h)+(1-h/k)^{-D}L(k-h)\\
&=2\Bigg[L(k)+L(k)\frac{h^2}{k^2}\binom{D+1}{2}-L^{(1)}(k)h\frac{h}{k}\binom{D}{1}-L^{(1)}(k)h\frac{h^3}{k^3}\binom{D+2}{3}\\
&\qquad+\frac{L^{(2)}(k)h^2}{2}+\frac{L^{(2)}h^2}{2}\frac{h^2}{k^2}\binom{D+1}{2}\\
&\qquad-\frac{L^{(3)}(k)h^3}{3!}\frac{h}{k}\binom{D}{1}-\frac{L^{(3)}(k)h^3}{3!}\frac{h^3}{k^3}\binom{D+2}{3}+\ldots\Bigg].
\end{aligned}$$

Hence,

$$\begin{aligned}
\gamma_-(k)&=2k^{-D}L(k)-k^{-D}\Big[(1+h/k)^{-D}L(k+h)+(1-h/k)^{-D}L(k-h)\Big]\\
&=-2k^{-2-D}L(k)h^2\binom{D+1}{2}+2k^{-1-D}L^{(1)}(k)h^2\binom{D}{1}\\
&\qquad+2k^{-3-D}L^{(1)}(k)h^4\binom{D+2}{3}-2k^{-D}\frac{L^{(2)}(k)h^2}{2}\\
&\qquad-2k^{-2-D}\frac{L^{(2)}(k)h^4}{2}\binom{D+1}{2}+2k^{-1-D}\frac{L^{(3)}(k)h^4}{3!}\binom{D}{1}+\ldots\\
&\sim O(k^{-2-D+\epsilon}).
\end{aligned}$$

Having established the autocovariance decay rate for the sequences $\{X_i \pm X_{i+h}\}_{i\geq 1}$, we can now look to the asymptotic behaviour of P_n when applied to these sequences.

3.5.5.1 $\hat{\gamma}_P(h)$ *when* $D > \frac{1}{2}$

Let $p = \frac{3}{4}$ and $q = \frac{1}{4}$ with $t = G^{-1}(p)$ and $s = G^{-1}(q)$. As the series $\{X_i + X_{i+h}\}_{i\geq 1}$ behaves similarly to $\{X_i\}_{i\geq 1}$ for all $0 < D < 1$ we can apply results from Section 3.5.4.2,

$$\begin{aligned}
\sqrt{n}(P_n-P)&=c\sqrt{n}\Big[(G_n^{-1}(p)-G_n^{-1}(q))-(G^{-1}(p)-G^{-1}(q))\Big]\\
&=c\sqrt{n}\left[\frac{G(t)-G_n(t)}{G'(t)}-\frac{G(s)-G_n(s)}{G'(s)}\right]+o_p(1)\\
&=c\sqrt{n}\left[\frac{2\sum_i h_1(X_i;s)}{nG'(s)}-\frac{2\sum_i h_1(X_i;t)}{nG'(t)}-\frac{R_n(t)}{G'(t)}+\frac{R_n(s)}{G'(s)}\right]+o_p(1)\\
&=\frac{2c}{\sqrt{n}}\left[\sum_{i=1}^{n}\frac{\Phi(2s-X_i)-\frac{1}{4}-\Phi(2t-X_i)+0.75}{G'(t)}\right]+o_p(1)\\
&=\frac{c}{\sqrt{n}}\sum_{i=1}^{n}\mathrm{IF}(X_i;P,\Phi)+o_p(1),
\end{aligned}$$

noting that $G'(s) = G'(t) = \sqrt{2}\phi(t\sqrt{2})$ and $\sqrt{n}R_n(t) = o_p(1)$. Hence, for $D > \frac{1}{2}$,

$$\sqrt{n-h}(P_{+,n-h} - P(\Phi_+)) = \frac{c}{\sqrt{n-h}} \sum_{i=1}^{n-h} \mathrm{IF}(X_i + X_{i+h}; P, \Phi_+) + o_p(1).$$

As $\{X_i - X_{i+h}\}_{i\geq 1}$ behaves like a SRD process, that is, it satisfies Assumption 3.1, we can use results directly from Section 3.4. In particular, we have that $\sqrt{n}(F_{-,n-h} - \Phi_-)$ converges in distribution to a Gaussian process in $\mathcal{D}(I)$ (Csörgó and Mielniczuk, 1996). Hence, we can use the expansion (3.14),

$$\sqrt{n-h}(P_{-,n-h} - P(\Phi_+)) = \frac{c}{\sqrt{n-h}} \sum_{i=1}^{n-h} \mathrm{IF}(X_i - X_{i+h}; P, \Phi_-) + o_p(1).$$

Using the delta method, $\hat{\gamma}_P(h)$ satisfies the following asymptotic expansion, just as we had in the SRD case in Section 3.4.2,

$$\sqrt{n-h}\,(\hat{\gamma}_P(h) - \gamma(h)) = \frac{1}{\sqrt{n-h}} \sum_{i=1}^{n-h} \mathrm{IF}(X_i, X_{i+h}; \gamma_P, \Phi) + o_p(1)$$

where $\mathrm{IF}(X_i, X_{i+h}; \gamma_P, \Phi)$ is defined in (3.10). As in the SRD case, the asymptotic normality follows from a CLT for $\frac{1}{\sqrt{n-h}} \sum_{i=1}^{n-h} \mathrm{IF}(X_i, X_{i+h}; \gamma_P, \Phi)$, established by noting that the Hermite rank is $m = 2$ and applying Theorem C.2 (page 159).

3.5.5.2 $\hat{\gamma}_P(h)$ *when* $D < \frac{1}{2}$

We can write, $\hat{\gamma}_P(h) - \gamma(h) = A_n^+ - A_n^-$, where,

$$A_n^{\pm} = \frac{1}{4}\left[P^2_{\pm,n-h} - P^2(\Phi_\pm)\right].$$

As $\{X_i - X_{i-h}\}_{i\geq 1}$ satisfies Assumption 3.1, $\sqrt{n}(F_{-,n-h} - \Phi_-)$ converges in distribution to a Gaussian process in $\mathcal{D}(I)$ (Csörgó and Mielniczuk, 1996) with a standard convergence rate of $\sqrt{n}$. Hence,

$$\sqrt{n-h}(P_{-,n-h} - P(\Phi_-)) = O_p(1),$$

and similarly, via the delta method,

$$\sqrt{n-h}(P^2_{-,n-h} - P^2(\Phi_-)) = O_p(1).$$

Therefore, for $D < \frac{1}{2}$,

$$\frac{k(D)(n-h)^{D-1/2}}{\widetilde{L}(n-h)}\sqrt{n-h}A_n^- = o_p(1).$$

Now if we consider A_n^+, we know that $\{X_i + X_{i+h}\}_{i\geq 1}$ exhibits LRD behaviour in a similar way to $\{X_i\}_{i\geq 1}$, and hence using the approach from Section 3.5.4.3,

$$\frac{k(D)(n-h)^D}{\widetilde{L}(n-h)}\left(P_{+,n-h} - P(\Phi_+)\right) \xrightarrow{D} \frac{P(\Phi_+)}{2}\left[Z_{2,D}(1) - Z_{1,D}^2(1)\right].$$

Using the delta method we have,

$$\frac{k(D)(n-h)^D}{\widetilde{L}(n-h)}\left(P_{+,n-h}^2 - P^2(\Phi_+)\right) \xrightarrow{\mathcal{D}} P^2(\Phi_+)\left[Z_{2,D}(1) - Z_{1,D}^2(1)\right].$$

Hence,

$$\frac{k(D)(n-h)^D}{\widetilde{L}(n-h)}A_n^+ \xrightarrow{\mathcal{D}} \frac{P^2(\Phi_+)}{4}(Z_{2,D}(1) - Z_{1,D}^2(1)).$$

Combining these results, the A_n^+ term determines the asymptotic distribution of $\hat{\gamma}_P$ as A_n^- is dominated by the convergence rate of A_n^+. Hence, noting that $P^2(\Phi_+) = \text{var}(X_1 + X_h) = 2(\gamma(0) + \gamma(h))$,

$$\frac{k(D)(n-h)^D}{\widetilde{L}(n-h)}\left(\hat{\gamma}_P(h) - \gamma(h)\right) \xrightarrow{\mathcal{D}} \frac{\gamma(0) + \gamma(h)}{2}(Z_{2,D}(1) - Z_{1,D}^2(1)).$$

3.6 SIMULATIONS

This section presents a selection of simulation results to highlight the relative efficiencies for both short and long range dependent processes. We also illustrate the difference in limiting distributions for LRD processes and explore the impact contamination has on the classical estimators relative to robust techniques.

Table 3.1 gives the relative MSEs for robust autocovariance estimators based on P_n and Q_n when the underlying data follow a SRD AR(1) process, $X_i = \alpha X_{i-1} + \varepsilon_i$, where $\{\varepsilon_i\}_{i\geq 1}$ are iid Gaussian random variables and $\alpha = 0.5$. In this setting, there is negligible bias, so the relative efficiencies and relative MSEs are the same. In small samples, the method based on P_n has an advantage over Q_n, similar to that found in Section 2.5. The efficiencies were reasonably constant for the variance and first and second order autocovariances. For both the methods based on P_n and Q_n, the relative MSEs increased as the sample size increases, as is the case in the iid setting.

In the LRD setting, Figure 3.9 shows an example of the types of time series considered. The top left plot shows an example of an ARFIMA$(0, 0.1, 0)$ process which corresponds to $D = 0.8$, hence scale and autocovariance estimators have a normal limiting distribution. The top right plot shows an

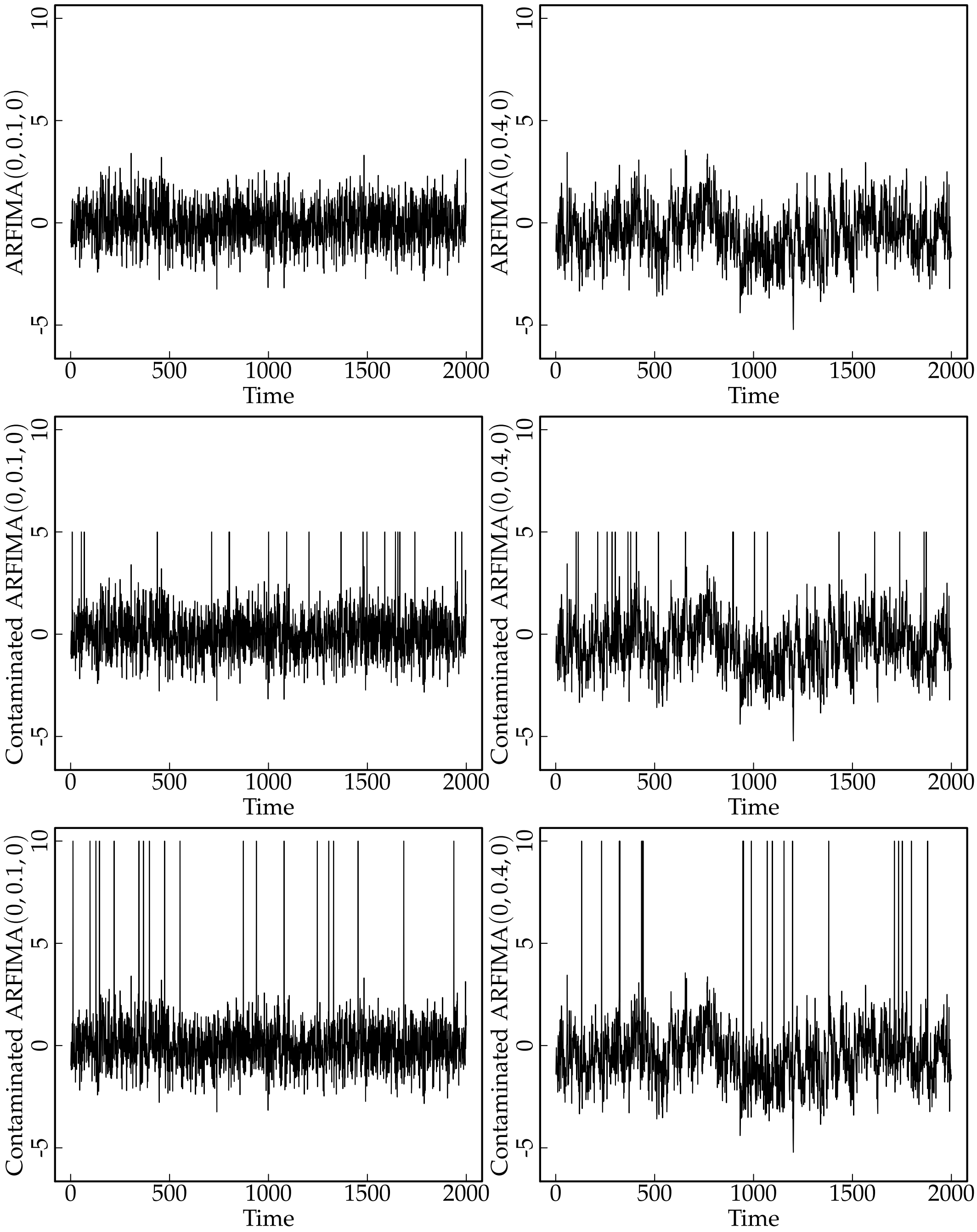


Figure 3.9: Example data series for an ARFIMA$(0, 0.1, 0)$ model (left) and ARFIMA$(0, 0.4, 0)$ model (right). The same series with 1% of observations have contamination fixed at 5 (middle) and 10 (bottom).

ARFIMA$(0, 0.4, 0)$, hence $D = 0.2$, it appears less stationary than the top left plot and the distributions of the scale and autocovariance estimators are expected to have a non-normal limit. In the middle panel, the same data set has been subject to 1% random contamination where all contaminated observations have been set to the value 5. If the outliers were jittered, it would be far more difficult to identify them in the $D < \frac{1}{2}$ plot, where the variability is much higher, than in the $D > \frac{1}{2}$ plot. Indeed, on the right, where $D < \frac{1}{2}$, one *true* observation is less than -5. In the bottom panel, 1% of observations have been fixed with the value 10. In this case, the outliers are extreme and clearly different to the rest of the series. The length of the series is $n = 2000$.

The empirical densities of normalised scale estimates corresponding to the scenarios found in Figure 3.9 are given in Figure 3.10. The top left panel, with no contamination, demonstrates the normal limiting distribution for the SD, P_n, Q_n and IQR. In terms of efficiency, P_n is 87% efficient relative to the SD, Q_n is 82% efficient and the IQR has only 38% efficiency. The heavier tails are a clear indicator of the relative inefficiency of the IQR.

All densities are centred at zero indicating that the scale estimators are consistently estimating the true value. This can also be seen in Table 3.2 where the variance and first and second order autocovariance estimates are unbiased for the true parameter values for all estimators over all sample sizes, in fact the three methods are virtually indistinguishable to 2 decimal places. As such, the relative MSEs are very similar to the relative efficiencies, see Table C.1 in Appendix C (page 167) for more detailed results.

When 1% contamination is added to the model, there is a limited amount of bias introduced in the robust methods, however, the SD experiences a

	P_n			Q_n		
n	$\gamma(0)$	$\gamma(1)$	$\gamma(2)$	$\gamma(0)$	$\gamma(1)$	$\gamma(2)$
20	0.80	0.74	0.67	0.72	0.71	0.61
50	0.86	0.82	0.78	0.80	0.80	0.76
100	0.88	0.86	0.84	0.83	0.83	0.81
500	0.90	0.89	0.89	0.87	0.86	0.86
1000	0.90	0.89	0.89	0.87	0.86	0.86

Table 3.1: MSEs of robust estimators relative to the classical methods over $100,000$ replications from an AR(1) process with $\alpha = 0.5$.

significant amount of bias. For the robust methods, there is negligible difference between the middle and bottom plots where the outliers have been moved from moderate (at 5) to extreme (at 10), indicating that they have reached their maximum bias for the 1% one-sided contamination and are resilient to the magnitude of the outliers. For a fixed proportion of contamination, the SD will continue to increase as the size of the outliers increases.

On the right hand side of Figure 3.10 where $D < \frac{1}{2}$, we clearly have a non-normal limit distribution as the empirical distributions of the normalised estimates are right skewed. This is not surprising given that it is a combination of a Rosenblatt process and a χ^2 distribution. In the top panel, we see again that the IQR appears to be slightly less efficient than the SD, but the difference is less marked than in the $D > \frac{1}{2}$ case. In fact, P_n and Q_n both have a similar relative efficiency of 93% and the IQR is 76% efficient relative to the SD.

It is important to note that there is significant bias in all the estimators when $D < \frac{1}{2}$. This is to be expected given that the limiting distribution not having mean zero. In particular Lévy-Leduc et al. (2011a) show that,

$$\mathbb{E}(Z_{2,D}(1) - Z_{1,D}^2(1)) = -\frac{2k(D)}{(1-D)(2-D)}.$$

This is highlighted even more dramatically in Table 3.3, where the estimated variances are, on average, well below the true values and similarly for the estimated autocovariances. There is also evidence of finite sample bias with slow convergence towards the limiting value. As a result, the MSEs are primarily driven by the bias, see Table C.2 in Appendix C for further detail.

When 1% contamination fixed at 5 is introduced, the SD begins to shift to the right, though the difference is not as marked as in the $D > \frac{1}{2}$ case.

	P_n			Q_n			SD		
n	$\gamma(0)$	$\gamma(1)$	$\gamma(2)$	$\gamma(0)$	$\gamma(1)$	$\gamma(2)$	$\gamma(0)$	$\gamma(1)$	$\gamma(2)$
100	1.01	0.09	0.04	1.01	0.09	0.04	0.99	0.09	0.04
500	1.02	0.11	0.06	1.02	0.11	0.06	1.01	0.11	0.06
1000	1.02	0.11	0.06	1.02	0.11	0.06	1.02	0.11	0.06

Table 3.2: Average estimates over 100,000 replications from an ARFIMA$(0, 0.1, 0)$ process with no contamination. The true values are $\gamma(0) = 1.02$, $\gamma(1) = 0.11$ and $\gamma(2) = 0.07$.

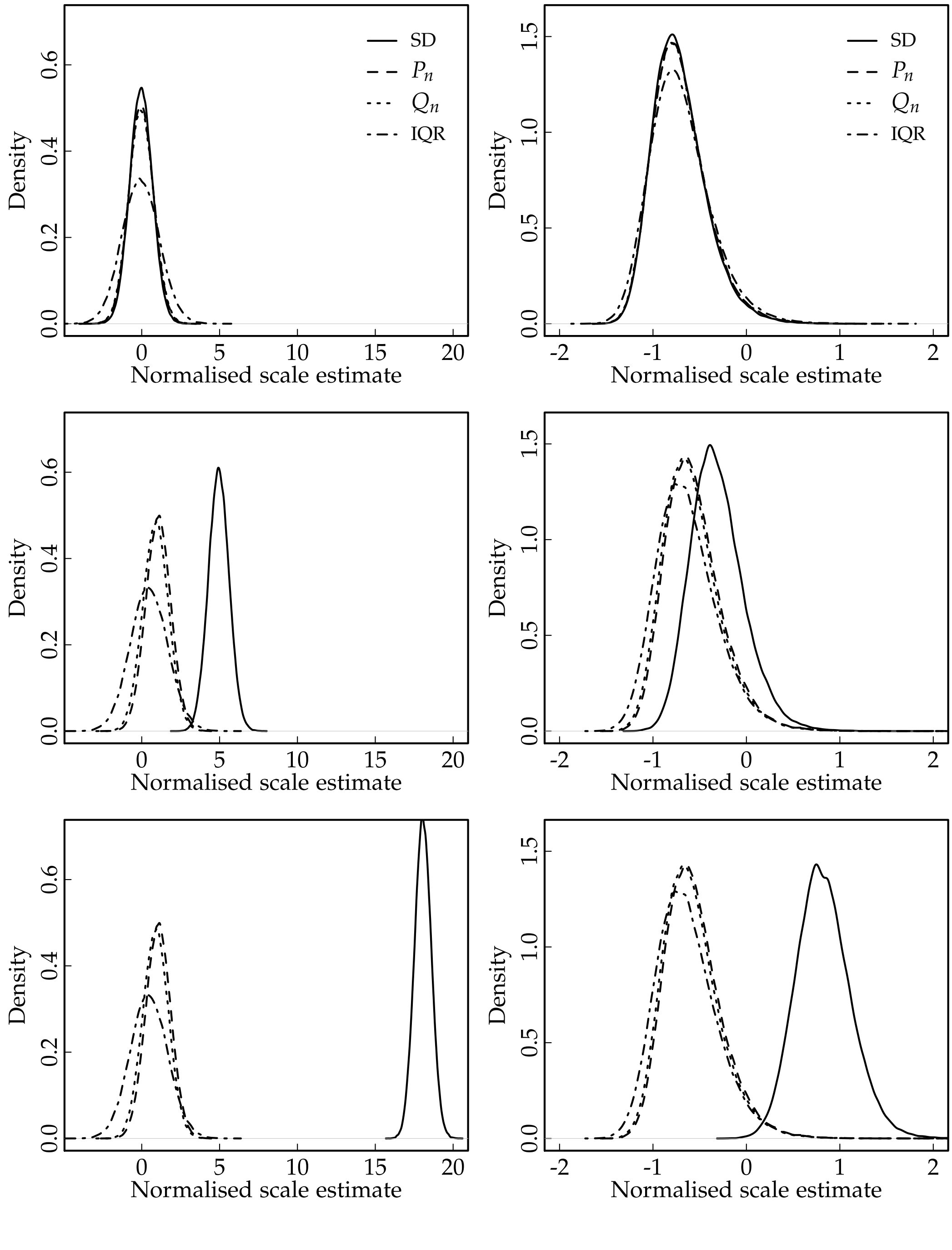


Figure 3.10: Empirical densities of normalised scale estimates for ARFIMA$(0, 0.1, 0)$ models, $\sqrt{n}(S - \sigma)$, (left) and ARFIMA$(0, 0.4, 0)$ models, $n^D(S - \sigma)$, (right) over 100,000 replications. The top row is based on estimates from samples with no contamination, the samples for the middle row have 1% contamination fixed at 5 and the samples for the bottom row have 1% contamination fixed at 10.

	P_n			Q_n			SD		
n	$\gamma(0)$	$\gamma(1)$	$\gamma(2)$	$\gamma(0)$	$\gamma(1)$	$\gamma(2)$	$\gamma(0)$	$\gamma(1)$	$\gamma(2)$
100	1.34	0.64	0.47	1.33	0.62	0.45	1.30	0.60	0.43
500	1.53	0.84	0.67	1.52	0.83	0.66	1.51	0.82	0.65
1000	1.59	0.90	0.73	1.59	0.90	0.73	1.58	0.89	0.72

Table 3.3: Average estimates over 100,000 replications from an ARFIMA$(0, 0.4, 0)$ process with no contamination. The true values are $\gamma(0) = 2.07$, $\gamma(1) = 1.38$ and $\gamma(2) = 1.21$.

The shift is more evident when the contamination is fixed at 10. The robust methods remain relatively unaffected by the contamination.

Figure 3.11 presents similar results for the first order autocovariances. In the $D > \frac{1}{2}$ case, the contamination leads to reduced efficiency in the SD with a slight negative bias. This behaviour is due to the corruption interfering with the ability of the classical method to discern the dependence structure between observations one lag apart, hence it is tending towards zero when the true value is $\gamma(1) = \frac{2}{3}$. Furthermore, the classical method is increasingly likely to return extreme estimates as the magnitude of the corruption increases. In the $D < \frac{1}{2}$, the behaviour is less marked, with all estimators returning broadly similar results with the various levels of contamination. An issue in practice, would be that the calculation of autocorrelations relies on an estimate of the scale (specifically the variance), which is failing for the standard deviation in contaminated samples.

3.7 CONCLUSION

In this chapter we confirmed that the good robustness and efficiency properties of P_n carried through to the covariance and autocovariance setting. Straightforward application of existing theory enabled us to write down the breakdown value, influence function and asymptotic efficiency for $\hat{\gamma}_P$. We also established that correlation estimates based on P_n had good finite sample efficiency relative to the SD and other robust estimates for bivariate t distributions. The relative efficiencies in the bivariate setting closely mirrored what was previously found in Chapter 2.

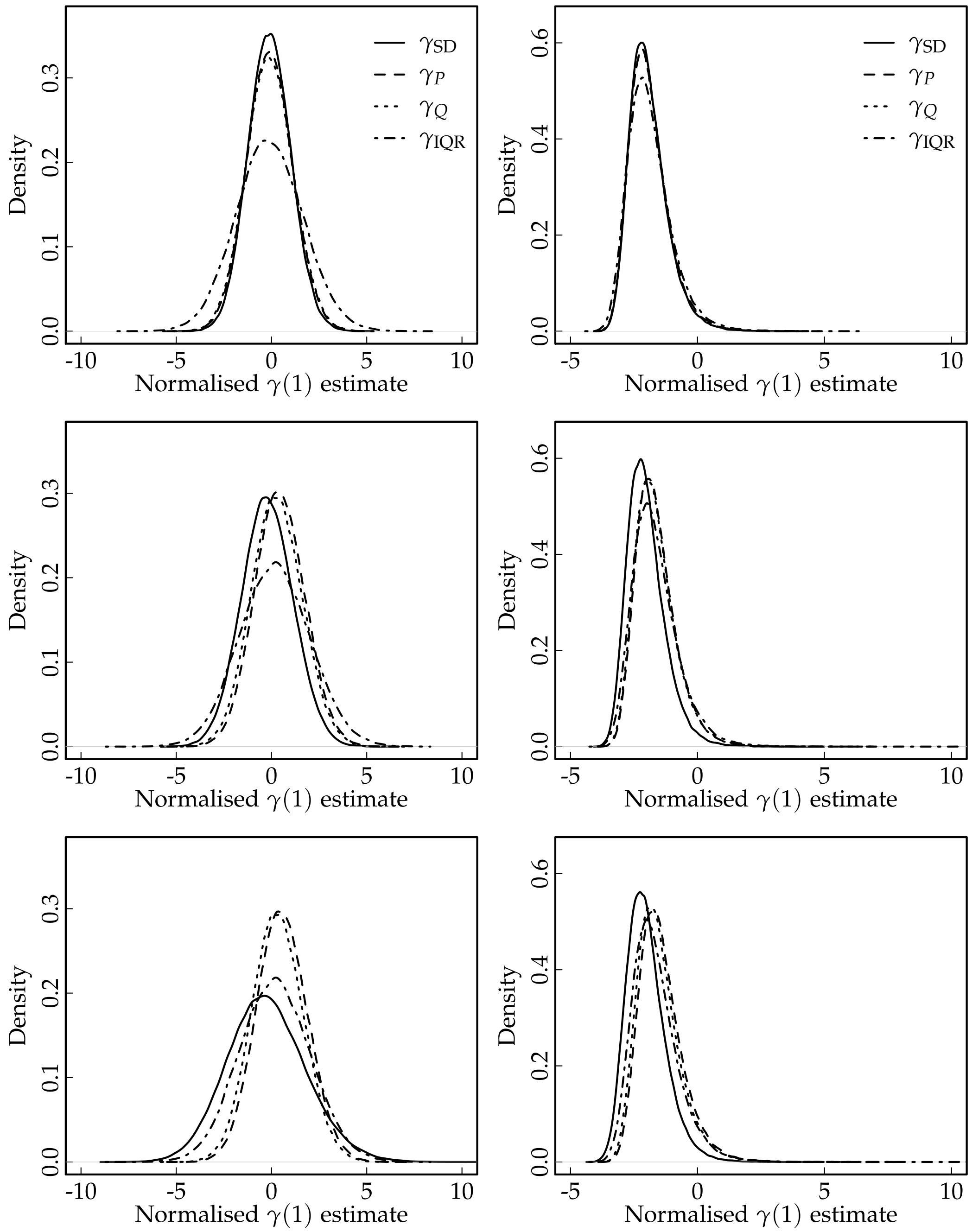


Figure 3.11: Empirical densities of normalised autocovariance estimates for ARFIMA$(0, 0.1, 0)$ models (left) and ARFIMA$(0, 0.4, 0)$ models (right) over 100,000 replications. The top row has no contamination, the middle has 1% contamination fixed at 5 and the bottom row has 1% contamination a fixed at 10.

We found the asymptotic distribution for P_n under short and long range dependence when the level of dependence was not too high ($D > \frac{1}{2}$). While the resulting limit was the same in both situations, the derivation is quite different. The SRD approach mirrored closely that of Lévy-Leduc et al. (2011c) for a single U-quantile, which was applied to obtain results for Q_n. However, the derivation of results for P_n was somewhat more involved as we needed to deal with the difference of two U-quantiles. Our LRD approach required a subtle augmentation of the kernel function in order to apply existing results.

In the long range dependence setting with extreme levels of dependence, $D < \frac{1}{2}$, we conjectured that the limiting distribution of P_n is the same as that previously established for other common scale estimators such as the SD and Q_n. This conjecture is supported by simulations.

Furthermore, we established that the IQR similarly follows the same non-normal distribution (a combination of the Rosenblatt and χ^2 distributions), which, to the best of our knowledge has not previously been established. In proving the non-normal limit for the IQR we relied upon a Bahadur representation for the sample quantiles.

In the simulations we demonstrated that the robust methods perform comparably to the classical approach when there is no contamination and the robust estimates remain bounded when contamination is introduced into the model. There is significant bias for all estimators when $D < \frac{1}{2}$ owing to the asymmetric limiting distribution which does not have mean zero.

4 COVARIANCE AND PRECISION MATRIX ESTIMATION

4.1 INTRODUCTION

Robust estimation of covariance matrices is one of the most challenging and fundamental issues in modern applied statistics. Through a natural extension to the covariance estimators considered in Chapter 3, this chapter investigates the contribution robust scale estimators can have in the estimation of the dependence structure in multivariate data sets.

Robust estimation of covariance matrices has received much attention in the past, notably the minimum volume ellipsoid and MCD estimators, projection type estimators and M-estimators, see Hubert, Rousseeuw and Van Aelst (2008) for a survey. Furthermore, research into covariance matrix estimation and its applications is ongoing, see for example Filzmoser, Ruiz-Gazen and Thomas-Agnan (2014) who use the MCD estimator to construct robust Mahalanobis distances to identify local multivariate outliers; Hubert, Rousseeuw and Vakili (2014) who study the shape bias of a range of existing robust covariance matrix estimators; or Cator and Lopuhaä (2010, 2012) who consider asymptotic expansions and establish asymptotic normality for general MCD estimators.

An alternative is to estimate the covariance matrix in a component-wise manner based on a robust estimator of scale as outlined by Ma and Genton (2001). It is well known that the resulting symmetric matrix is not guaranteed to be positive definite (PD). Methods to ensure the resulting estimator is PD have previously been explored by Rousseeuw and Molenberghs (1993) with notable updates in the robustness literature by Maronna and Zamar (2002) and quite separately in the finance literature by Higham (2002).

Often it is the precision matrix, the inverse of the covariance matrix, that is the statistic of interest, for example in linear discriminant analysis or Gaussian graphical model selection. Rather than focussing on the covariance matrix, this chapter is primarily concerned with robustly estimating the precision matrix. Whilst there is an obvious link between covariance matrices and precision matrices, it is not obvious that a good (robust) estimator for one results in a good estimator for the other. We will employ robust pairwise covariance matrices as a starting point for various regularisation techniques to facilitate the estimation of robust, potentially sparse, precision matrices. We also discuss why it is often inappropriate to directly invert existing pairwise covariance matrix estimators to find robust non-sparse precision matrices.

A detailed simulation study has been undertaken to assess the performance of a variety of estimators over a number of scenarios and levels of p while keeping the sample size fixed. Our results are distilled from a comprehensive range of performance indices which will be introduced and their applicability to the various scenarios discussed.

We analyse the impact of cellwise contamination on the estimators. While cellwise contamination is common in the missingness literature, it is an emerging area of research in robust statistics, defined formally in the robustness context by Alqallaf et al. (2009). This alternative form of contamination represents a philosophical divergence from the traditional approach to robustness, which is primarily concerned with contaminated observation vectors.

It will be shown that the pairwise nature of the covariance estimates enables the resulting precision matrix to have a higher level of robustness than when using standard robust covariance matrix estimation procedures in the presence of scattered contamination. This is a novel result and a significant first step towards dealing with cellwise contamination in this context.

The remainder of this chapter is structured as follows. Section 4.2 outlines the scattered contamination model and highlights why standard robust techniques fail in this setting. Section 4.3 considers the suitability of various performance indices that are later used to assess the performance of various covariance and precision matrix estimators. Sections 4.4 and 4.5 provide basic theory for existing pairwise covariance matrix estimation techniques and regularisation routines and outlines our new procedure which combines robust pairwise covariance matrix estimation with regularisation. Section 4.6 presents the results of an extensive simulation study and Section 4.7 summarises the important findings.

4.2 SCATTERED CONTAMINATION

Classically, even the *most robust* procedures are designed such that they only work when at most 50% of the rows in the $n \times p$ data matrix have contamination present. In the discussion section of a survey paper on robust statistics, Maronna and Yohai specifically highlight elementwise contamination as an important focus for future developments in robustness (Morgenthaler, 2007).

Alqallaf et al. (2009) formally outline the cellwise contamination model as an extension of the standard Tukey-Huber contamination model which was first introduced in the univariate location-scale setup. In the multivariate setup, we observe the random vector,

$$\boldsymbol{x} = (\boldsymbol{I} - \boldsymbol{B})\boldsymbol{y} + \boldsymbol{B}\boldsymbol{z},$$

where $\boldsymbol{y} \sim \boldsymbol{F}$, the distribution of *well-behaved* data, $\boldsymbol{z} \sim \boldsymbol{G}$, some outlier generating distribution and $\boldsymbol{B} = \operatorname{diag}(B_1, \ldots, B_p)$ is a diagonal matrix, where $B_1, \ldots, B_p$ are Bernoulli random variables, $P(B_i = 1) = \varepsilon_i$. For simplicity, we will assume that $\boldsymbol{y}$, $\boldsymbol{B}$ and $\boldsymbol{z}$ are independent. This is similar to the missing completely at random model, where the missingness does not depend on the values of $\boldsymbol{y}$, see, for example, Little and Rubin (2002).

Clearly, it is the structure of $\boldsymbol{B}$ that determines the contamination model. If $B_1, \ldots, B_p$ are fully dependent, then $\boldsymbol{B} = B\boldsymbol{I}_p$, where $B \sim \mathcal{B}(1, \varepsilon)$, and we recover the *fully dependent* contamination model, the standard model on which classical robust procedures are based. In this setting, the probability that an observation is uncontaminated, $1 - \varepsilon$, is independent of the dimensionality. Furthermore, the proportion of contaminated observations is preserved under affine equivariant transformations.

In contrast, if $B_1, \ldots, B_p$ are mutually independent we have the *fully independent* contamination model, where each element of $\boldsymbol{x}$ comes from $\boldsymbol{F}$ or $\boldsymbol{G}$ independently of the other $p - 1$ elements in $\boldsymbol{x}$. In this setting, it may be be unreasonable to assume that less than half the rows have contamination. Furthermore, if p is large and there is only one outlier in an observation vector, then down-weighting the entire observation may be wasteful.

If the data matrix is randomly contaminated in this elementwise manner, as the number of variables increases, the chance that more than half the rows are contaminated increases exponentially. Formally, let ε be the probability that any particular element in a data matrix is contaminated. Assuming the contamination is randomly scattered throughout the data matrix, the probability that any particular row has no contamination is $(1 - \varepsilon)^p$,

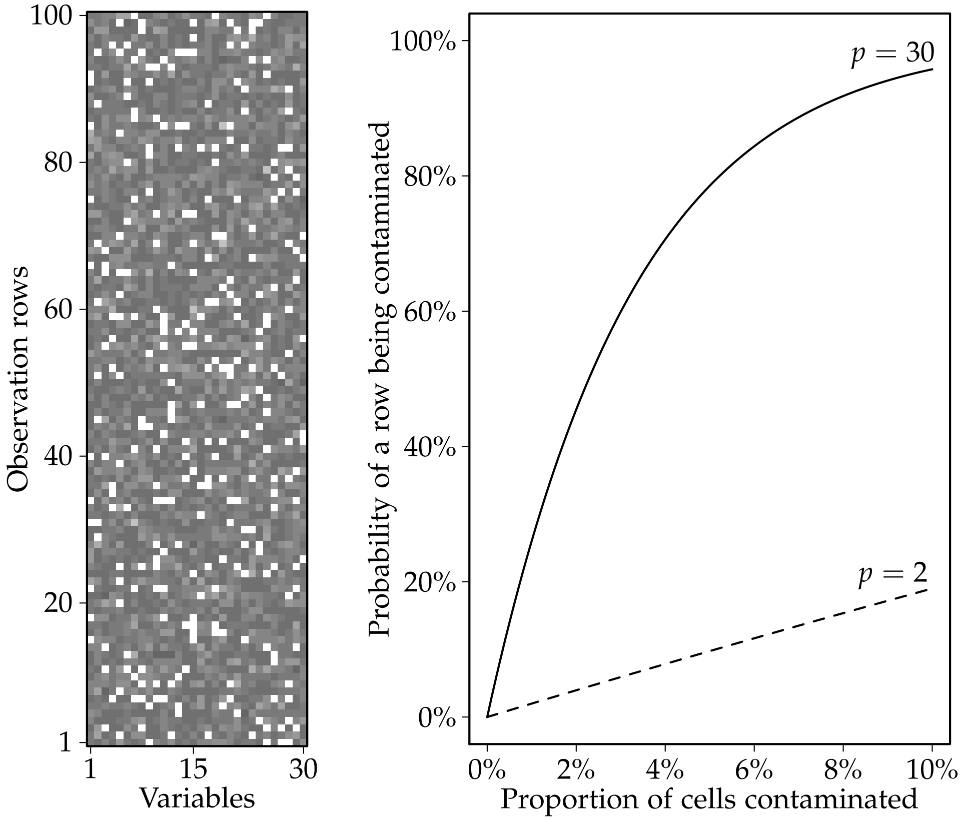


Figure 4.1: On the left, heat map of a data matrix with 30 variables and 100 observations. 10% of the observations have been contaminated (represented in white) scattered throughout the data set. On the right, the probability that any particular row (observations) in the data matrix that will be contaminated, $1 - (1 - \varepsilon)^p$, over a range of ε, the proportion of cells effected by scattered contamination.

which quickly decays towards zero even for small values of ε. For example, if $p = 30$ and $\varepsilon = 0.1$, then the probability that any particular row remains uncontaminated is only 4%. This is demonstrated graphically in Figure 4.1. The plot on the left shows a 100×30 data matrix where 10% of the cells have been contaminated, the white cells. While virtually all the rows of the data matrix have at least one contaminated element, the majority of cells remain uncontaminated in the sense that they are still *real* measurements from the underlying data generating process. Even if $\varepsilon = 0.03$, the probability that any particular row is uncontaminated is 40%, however with a sample size of 100, this translates to a 98% chance that at least half the rows are contaminated, in which case standard robust methods fail.

It is important to note that the fully independent contamination model lacks affine equivariance, in the sense that linear combinations of columns of a contaminated data set result in "outlier propagation" (Alqallaf et al., 2009). As such, affine equivariance is not an achievable outcome for any estimator in this setting.

Existing research into the problem of scattered contamination has focussed on coordinatewise procedures, that only operate on one column at a time. Croux, Filzmoser et al. (2003) consider an approach based on "alternating regressions" using weighted L_1 regression, Maronna and Yohai (2008) use a coordinatewise procedure for principal component analysis. Liu et al. (2003) have an application involving the singular value decomposition of microarray data and De la Torre and Black (2001) consider cellwise contamination in the context of computer vision.

We show that a pairwise approach is able to cope with much higher levels of scattered contamination than existing classical robust estimators. In our simulations we do not use the fully independent contamination model, rather, we impose restrictions on the amount of contamination in each variable. As such the contamination is no longer strictly independent, however, the advantage is that we are able to assess the impact over various *known* levels of contamination in each variable.

4.3 PERFORMANCE INDICES

The following sections require a way to assess the performance of various covariance and precision matrix estimators under cellwise contamination. There are a range of possible ways to measure how *close* an estimated matrix is to the true value. In order to assess the performance of our proposed estimators, we first need to identify which performance indices are appropriate. One class of performance indices considered are matrix norms which measure the *size* of a matrix. The second class looks at how closely the estimated precision (or covariance) matrix reflects the nature of the theoretical precision (covariance) matrix, through either the determinant, condition number or an overall entropy loss index. Let $\boldsymbol{\Sigma}$ denote the true covariance matrix and $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$ denote the true precision matrix. In this section we define the performance measures and consider their appropriateness in the context of scattered contamination.

4.3.1 *Matrix norms*

We need only consider square matrices. Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be $p \times p$ matrices. Matrix norms are mappings which satisfy the following criteria (Gentle, 2007):

i. $||\boldsymbol{A}|| \geq 0$ and $||\boldsymbol{A}|| = 0 \iff \boldsymbol{A} = \boldsymbol{0}$;

ii. $||c\boldsymbol{A}|| = |c|\, ||\boldsymbol{A}||$ for $c \in \mathbb{R}$;

iii. $||\boldsymbol{A} + \boldsymbol{B}|| \leq ||\boldsymbol{A}|| + ||\boldsymbol{B}||$; and

iv. $||\boldsymbol{AB}|| \leq ||\boldsymbol{A}||\, ||\boldsymbol{B}||$.

The Frobenius norm is perhaps the most common matrix norm, it is an element-wise norm, the Euclidean norm of $\boldsymbol{A}$ treated as if it were a vector of length p^2,

$$||\boldsymbol{A}||_F = \sqrt{\sum_{i,j} |a_{ij}|^2} = \sqrt{\operatorname{tr}(\boldsymbol{A}^\top \boldsymbol{A})}.$$

Note that $||\boldsymbol{A}^\top||_F = ||\boldsymbol{A}||_F$.

An alternative way of constructing a matrix norm is to take a vector norm and use it to generate a matrix norm of the form:

$$||\boldsymbol{A}|| = \sup_{||\boldsymbol{x}||=1} ||\boldsymbol{A}\boldsymbol{x}||,$$

where $||\cdot||$ on the left is the induced (or operator) norm and $||\cdot||$ on the right is a vector norm. Examples of induced norms are the L_p norms:

- The one norm or the column sum norm, is the maximum column-wise sum after taking the absolute values of all the elements in the matrix,

$$||\boldsymbol{A}||_1 = \max_j \sum_{i=1}^{n} |a_{ij}|.$$

- The infinity norm or the row sum norm, is the maximum row-wise sum after taking the absolute values of all the elements in the matrix,

$$||\boldsymbol{A}||_\infty = \max_i \sum_{j=1}^{n} |a_{ij}|.$$

  Note that $||\boldsymbol{A}^\mathsf{T}||_\infty = ||\boldsymbol{A}||_1$ so if $\boldsymbol{A}$ is symmetric then the one norm and the infinity norm coincide.

- The spectral norm, $||\boldsymbol{A}||_2 = \sigma_{\max}(\boldsymbol{A})$, where $\sigma_{\max}(\boldsymbol{A})$ is the largest singular value of $\boldsymbol{A}$. When $\boldsymbol{A}$ is nonsingular $||\boldsymbol{A}^{-1}||_2 = 1/\sigma_{\min}(\boldsymbol{A})$ where $\sigma_{\min}(\boldsymbol{A})$ is the smallest singular value of $\boldsymbol{A}$.

  Note that $\sigma_{\max}(\boldsymbol{A}) = \sqrt{\lambda_{\max}(\boldsymbol{A}^\mathsf{T}\boldsymbol{A})}$ where $\lambda_{\max}(\boldsymbol{A}^\mathsf{T}\boldsymbol{A})$ is the largest eigenvalue of $\boldsymbol{A}^\mathsf{T}\boldsymbol{A}$. Hence, as the positive eigenvalues of $\boldsymbol{A}^\mathsf{T}\boldsymbol{A}$ are the same as those of $\boldsymbol{A}\boldsymbol{A}^\mathsf{T}$, singular values are invariant to matrix transposition and we have $||\boldsymbol{A}^\mathsf{T}||_2 = ||\boldsymbol{A}||_2$.

In our experiments, we apply the matrix norms to $\boldsymbol{A} = \boldsymbol{\Sigma}_0 - \boldsymbol{I}$ where $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}$ or $\boldsymbol{A} = \boldsymbol{\Theta}_0 - \boldsymbol{I}$ where $\boldsymbol{\Theta}_0 = \boldsymbol{\Theta}^{-1}\hat{\boldsymbol{\Theta}}$. While $\boldsymbol{\Sigma}$ and $\boldsymbol{\Theta}$ and their corresponding estimates are symmetric, it is not the case that the product of two symmetric matrices yields a symmetric matrix, hence in general $\boldsymbol{\Sigma}_0 \neq \boldsymbol{\Sigma}_0^\mathsf{T}$ and similarly $\boldsymbol{\Theta}_0 \neq \boldsymbol{\Theta}_0^\mathsf{T}$, so in practice the one norm and the infinity norm may yield different results. Furthermore, while $\boldsymbol{\Sigma}$, $\boldsymbol{\Theta}$, $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{\Theta}}$ are typically[7] PSD, it is also not the case that the product of two of these will also be PSD. Meenakshi and Rajian (1999) make the point that if $\boldsymbol{A}$ and $\boldsymbol{B}$ are PSD then $\boldsymbol{AB}$ will also be PSD if and only if $\boldsymbol{AB}$ is normal, i.e. $(\boldsymbol{AB})^\mathsf{T}\boldsymbol{AB} = \boldsymbol{AB}(\boldsymbol{AB})^\mathsf{T}$ which is typically not the case here.

For each norm, one may naïvely assume that the closer to zero the better (as they are all positive), however, Section 4.3.3 demonstrates that this is not always the case, particularly when estimating precision matrices in the presence of outliers.

[7] The estimated covariance matrices will only be non-positive semidefinite (PSD) in the case where $\hat{\boldsymbol{\Sigma}}$ has been obtained in a component-wise manner and has not had the OGK or NPD method applied to it.

4.3.2 *Other indices*

Aside from matrix norms, there are a few other commonly employed performance indices. Let $\boldsymbol{\Gamma}$ be a $p \times p$ matrix representing either the precision or the covariance matrix with corresponding estimate, $\hat{\boldsymbol{\Gamma}}$. We apply the indices in this section to standardised covariance or precision matrices, $\boldsymbol{\Gamma}_0 = \boldsymbol{\Gamma}^{-1}\hat{\boldsymbol{\Gamma}}$. If $\hat{\boldsymbol{\Gamma}}$ is *close* to $\boldsymbol{\Gamma}$ then $\boldsymbol{\Gamma}_0$ should be close to an identity matrix.

4.3.2.1 *Entropy loss*

The entropy loss, as suggested by Stein (1956) and featured in James and Stein (1961) and also Dey and Srinivasan (1985), is defined as,

$$\begin{aligned} L_1(\boldsymbol{\Gamma}, \hat{\boldsymbol{\Gamma}}) &= \operatorname{tr}(\boldsymbol{\Gamma}^{-1}\hat{\boldsymbol{\Gamma}}) - \log\det(\boldsymbol{\Gamma}^{-1}\hat{\boldsymbol{\Gamma}}) - p \\ &= \sum_{i=1}^{p} (\lambda_i - \log \lambda_i) - p, \end{aligned}$$

where λ_i, $i = 1, \ldots, p$, are the eigenvalues of $\boldsymbol{\Gamma}_0$. Stein (1956) notes that this function is “somewhat arbitrary” but it is convex in $\hat{\boldsymbol{\Gamma}}$ and assuming that $\boldsymbol{\Gamma}$ and $\hat{\boldsymbol{\Gamma}}$ are PSD, $L_1(\boldsymbol{\Gamma}, \hat{\boldsymbol{\Gamma}}) \geq 0$ with equality if and only if $\boldsymbol{\Gamma} = \hat{\boldsymbol{\Gamma}}$. If $\hat{\boldsymbol{\Gamma}} = \boldsymbol{S}$, the sample covariance matrix and $p > n$, then $\boldsymbol{S}$ is singular and therefore $L_1(\boldsymbol{\Sigma}, \boldsymbol{S}) = \infty$.

The entropy loss is not as “arbitrary” as it may seem at first. Note that the Kullback-Leibler divergence from $\mathcal{N}_1(\boldsymbol{\mu}, \boldsymbol{\Sigma}_1)$ to $\mathcal{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}_2)$ is,

$$D_{\text{KL}}(\mathcal{N}_1, \mathcal{N}_2) = \frac{1}{2} L_1(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2).$$

The close link between the entropy loss and the Kullback-Leibler or Bregman divergence loss is shown in a Bayesian context by Gupta and Srivastava (2010).

There is also a clear link between the entropy loss and the likelihood ratio test for $H_0 : \boldsymbol{\Sigma} = \boldsymbol{\Sigma}^\star$ with unknown mean assuming the data come from a Gaussian distribution,

$$-2(l_1 - l_0) = nL_1(\boldsymbol{\Sigma}^\star, \boldsymbol{S}),$$

where l_0 and l_1 are the log likelihoods under the null and alternative hypotheses, see, for example Mardia, Kent and Bibby (1979, p. 126).

The entropy loss is used extensively as a basis for developing and assessing improved precision and covariance matrix estimators, for example in Lin and Perlman (1985), Yang and Berger (1994) and more recently, Won et al. (2013).

4.3.2.2 *Log determinant*

In the multivariate Gaussian setting, Wilks (1932) names the determinant of the covariance matrix, $L_D(\boldsymbol{\Sigma})$, the generalised variance. The generalised precision is similarly defined as, $L_D(\boldsymbol{\Theta})$. This idea can be used as the basis for a performance index. Consider the log of the determinant of the standardised covariance or precision matrix,

$$L_D(\boldsymbol{\Gamma}_0) = \log\det(\boldsymbol{\Gamma}_0) = \sum_{i=1}^{p} \log \lambda_i,$$

where λ_i, $i = 1, \ldots, p$, are the eigenvalues of $\boldsymbol{\Gamma}_0$. The determinant of an identity matrix is 1, so the optimal value of $L_D(\boldsymbol{\Gamma}_0)$ is 0. Positive (negative) log determinant results indicate that the generalised variance or precision is being over (under) estimated. Noting that, $\boldsymbol{\Sigma}_0^{-1} = \hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma} = \boldsymbol{\Theta}_0^{\mathsf{T}}$, we have, $\det(\boldsymbol{\Theta}_0^{\mathsf{T}}) = \det(\boldsymbol{\Theta}_0)$ and $\det(\boldsymbol{\Sigma}_0^{-1}) = [\det(\boldsymbol{\Sigma}_0)]^{-1}$ and hence,

$$L_D(\boldsymbol{\Theta}_0) = -L_D(\boldsymbol{\Sigma}_0).$$

Thus, methods that underestimate the generalised variance will overestimate the generalised precision.

The log determinant is a very crude performance index which can be dominated by one eigenvalue that is very close to zero. Furthermore, it is incorporated as part of the entropy loss, $L_1(\boldsymbol{\Gamma}, \hat{\boldsymbol{\Gamma}})$, so there is little need to focus on it in the results of the simulation studies presented later.

4.3.2.3 *Log condition number*

Formally, the condition number of a square matrix is the product of the norm of the matrix and the norm of its inverse,

$$\kappa(\boldsymbol{\Gamma}_0) = ||\boldsymbol{\Gamma}_0^{-1}|| \cdot ||\boldsymbol{\Gamma}_0||,$$

and hence depends on the choice of matrix-norm. It is common to use the spectral norm, in which case the condition number is the ratio of the largest to the smallest non-zero singular value of the matrix.

The condition number associated with the systems of equations, $\boldsymbol{Ax} = \boldsymbol{b}$, gives a bound on how inaccurate the solution may be. A system is said to be *ill-conditioned* if small changes in the inputs, $\boldsymbol{A}$ and $\boldsymbol{b}$, result in large changes in the solution, $\boldsymbol{x}$. Consider the performance index defined as the log of the condition number of $\boldsymbol{\Gamma}_0$,

$$L_\kappa(\boldsymbol{\Gamma}_0) = \log \kappa(\boldsymbol{\Gamma}_0) = \log(\sigma_{\max}(\boldsymbol{\Gamma}_0)) - \log(\sigma_{\min}(\boldsymbol{\Gamma}_0)).$$

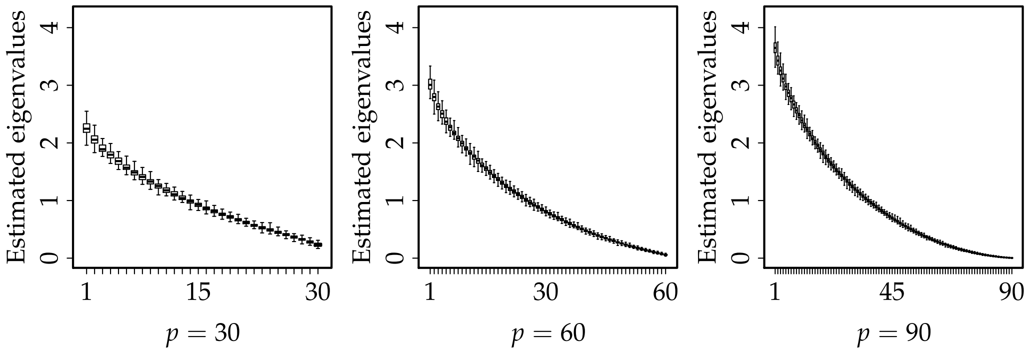


Figure 4.2: A series of boxplots showing the eigenvalues of the sample covariance matrix, $\boldsymbol{S}$ over $N = 100$ samples from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ with $n = 100$ and $p = 30$, 60 and 90.

The condition number for an identity matrix is 1 and the condition number for a singular matrix is infinity, so the log condition number index ranges from zero (best) to infinity (worst).

Gentle (2007) notes that while the condition number of a matrix provides a useful indication of its ability to solve linear equations accurately, it can be misleading at times when the rows (or columns) of the matrix have very different scales. That is, the condition number can be changed by simply scaling the rows or columns which does not actually make a linear system of equations any better or worse conditioned. This is known as artificial ill-conditioning.

In the context of the sample covariance matrix, $\boldsymbol{S}$, Ledoit and Wolf (2004) note that "when the ratio p/n is less than one but not negligible, the sample covariance matrix is invertible but numerically ill-conditioned, which means that inverting it amplifies estimation error dramatically." Won et al. (2013) go further stating that "the eigenstructure [of $\boldsymbol{S}$] tends to be systematically distorted unless p/n is extremely small, resulting in numerically ill-conditioned estimators for $\boldsymbol{\Sigma}$." Figure 4.2 demonstrates the systematic deterioration in the eigenstructure as $p/n \to 1$. The eigenvalues of the true covariance matrix are all identically 1, however this is not reflected in the eigenvalues of the estimated sample covariance matrices.

As with the log determinant, it is not expected that the log condition number will be a particularly discerning performance index. In assessing whether a robust estimator provides reasonable estimates, it will be enough to note that the log condition number remains bounded.

4.3.2.4 *Quadratic loss*

Another index that is frequently used in the literature to assess the performance of covariance matrix estimators is the quadratic loss. The exact specification varies from paper to paper, for example Ledoit and Wolf (2004) define it as,

$$L_2(\hat{\boldsymbol{\Gamma}}) = ||\hat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}||_F^2,$$

However, their use of this performance index is confounded with their proposed estimator, in the sense that they were constructing an estimator that minimised the expected quadratic loss, and then used the expected quadratic loss to evaluate the performance of their estimator.

A specification of the quadratic loss, more in line with the entropy loss, is used in Won et al. (2013),

$$L_Q(\hat{\boldsymbol{\Gamma}}, \boldsymbol{\Gamma}) = ||\hat{\boldsymbol{\Gamma}}\boldsymbol{\Gamma}^{-1} - \boldsymbol{I}||_F^2.$$

It is obvious that the quadratic loss is intrinsically linked to the Frobenius norm, so there is no need to include it as a separate performance index in our simulations.

4.3.3 *Behaviour of performance indices*

The performance indices outlined in this section are typically used to compare competing estimators in uncontaminated data sets. Contaminated data will have potentially severe implications for structure and size of the estimated precision and covariance matrices, and it is not clear how these indices will behave in such settings. As such, we begin our investigation by exploring how these indices react to the presence of gross outliers in a data set.

The model used to assess the behaviour of the various performance indices is typical of that which will be used in the simulation study, $n = 100$ observations drawn from the standard multivariate Gaussian distribution, $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$, with $p = 30$.

4.3.3.1 *Inflated variances*

This section explores how the various performance indices react if we artificially inflate the variance of the first variable, i.e. increase the value of s_{11} in the sample covariance matrix, $\boldsymbol{S} = (s_{ij})$, based on a single sample of uncontaminated data. In this simple case, where $\boldsymbol{\Sigma} = \boldsymbol{\Theta} = \boldsymbol{I}$, the matrix norms are applied to $\boldsymbol{S} - \boldsymbol{I}$ or $\boldsymbol{S}^{-1} - \boldsymbol{I}$ and the entropy loss, log determinant and log

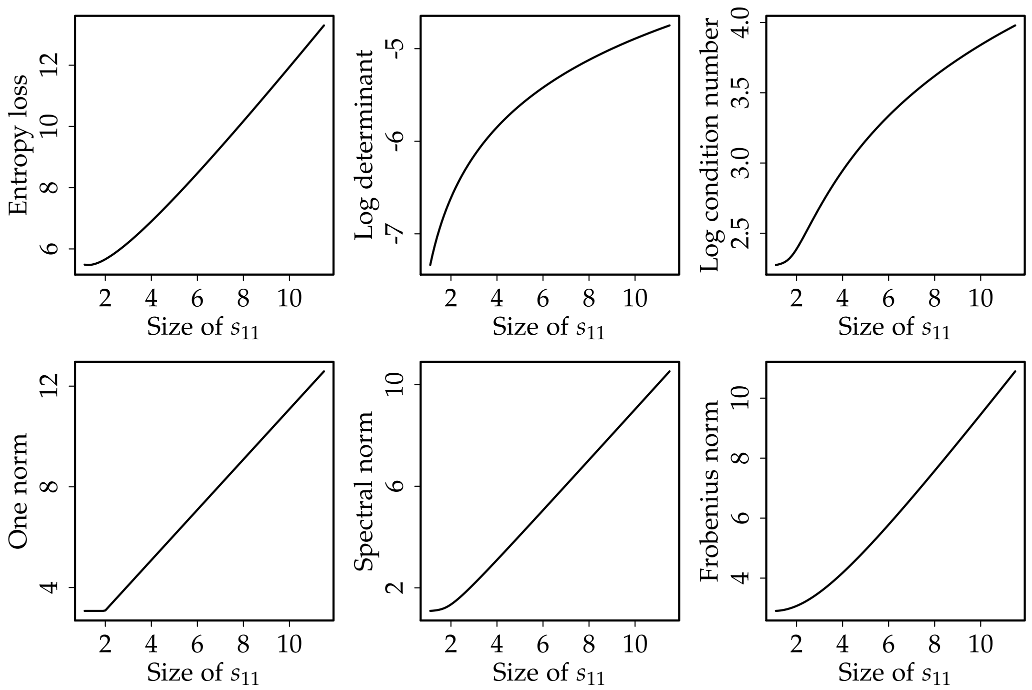


Figure 4.3: The impact of artificially inflating the size of the top left element of the sample covariance matrix, s_{11}, on each of the performance indices when applied to the resulting *covariance* matrix.

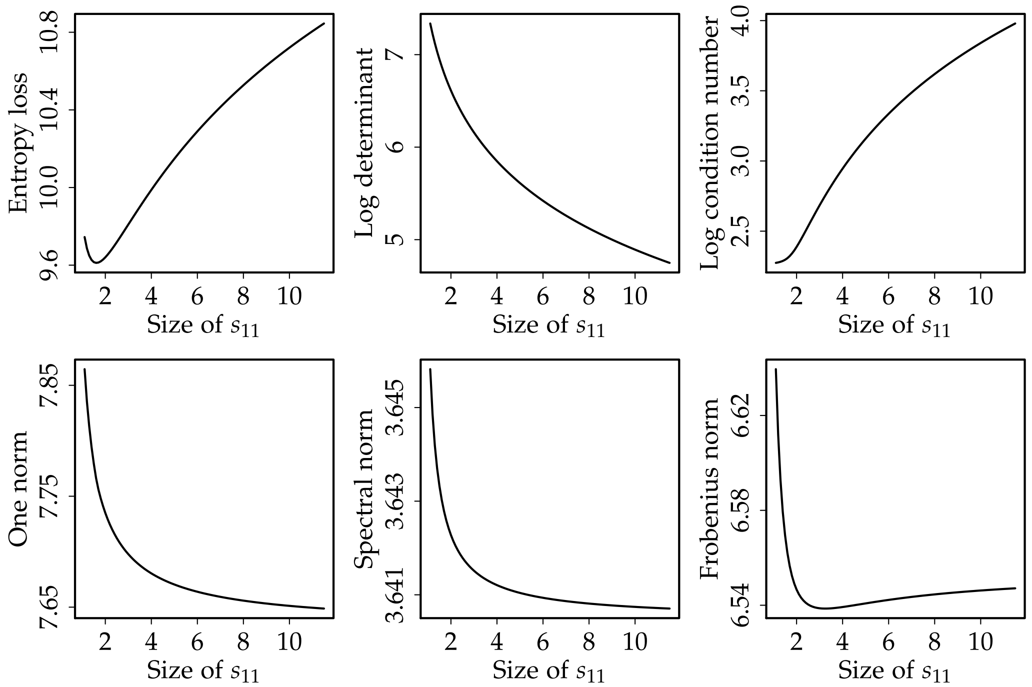


Figure 4.4: The impact of artificially inflating the size of the top left element of the sample covariance matrix, s_{11}, on each of the performance indices when applied to the resulting *precision* matrix.

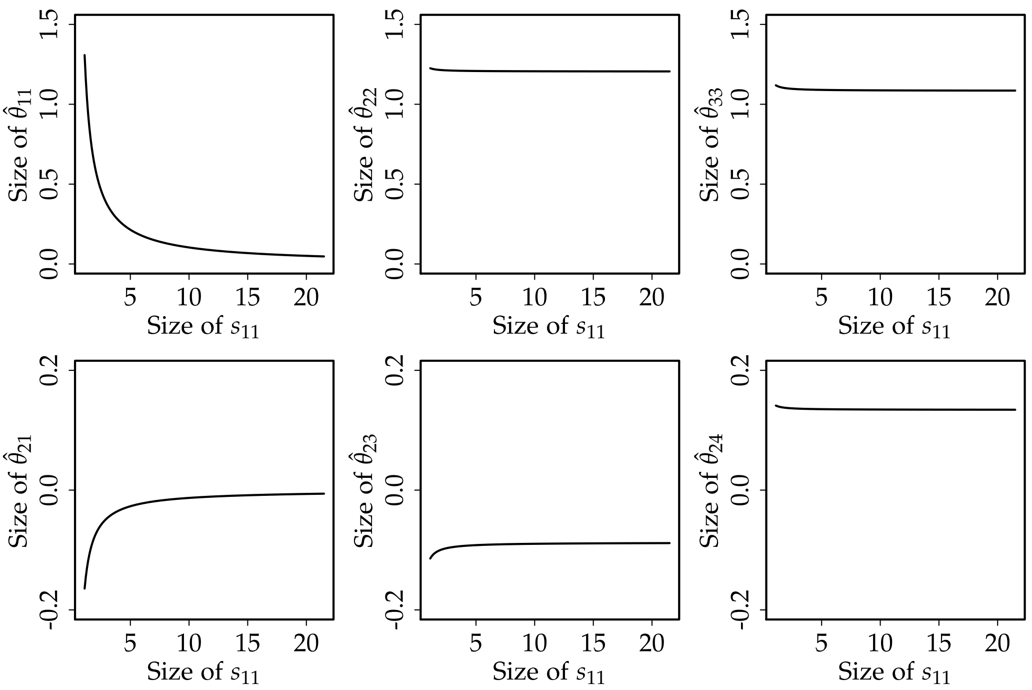


Figure 4.5: The change in few elements of $\hat{\Theta} = S^{-1}$ as s_{11} is artificially inflated. Main diagonal elements are in the top row and the off diagonal elements are in the bottom row. Note that $S = (s_{ij})$ and $\hat{\Theta} = (\hat{\theta}_{ij})$.

condition number are simply applied to S or S^{-1}. Note that in this setting the one norm and the infinity norm will give identical results and so only the results for the one norm are shown.

Figure 4.3 shows the behaviour of the various indices when applied to these adjusted covariance matrices and Figure 4.4 presents the same for the resulting precision matrices, $\hat{\Theta} = S^{-1}$. The horizontal axis shows the size of s_{11}, the artificially inflated variance of the first variable in the sample covariance matrix.

In Figure 4.3, the majority of the performance indices behave similarly in the covariance case – there is an overall positive trend as the variance of the first variable increases. The spectral norm and Frobenius norm both increase uniformly with s_{11}. The one norm, and correspondingly the infinity norm (not shown), remains flat as the sum of the absolute value of the elements in another column (or row) remains larger than the first column (row) up until the point where $s_{11} \approx 2$, at which point the one norm (and infinity norm) increase linearly with s_{11}. For large s_{11}, it is clear that all considered performance indices register that the adjusted S matrix is no longer *close* to the true value, $\boldsymbol{I}$.

In stark contrast, Figure 4.4 shows the indices when applied to the resulting precision matrix (after the first entry in the covariance matrix has been artificially inflated). As expected, the condition number of the resulting inverse is identical to that of the original covariance matrix and $L_D(\boldsymbol{S}^{-1}) = -L_D(\boldsymbol{S})$, however the other indices exhibit somewhat different behaviour. In particular the matrix norms tend to decrease, rather than increase. This is explained by noting that as the first element in the covariance matrix is artificially inflated, the first row and column of the precision matrix decay towards zero while the other elements are more or less constant. This is demonstrated in Figure 4.5 where $\hat{\theta}_{11}$ and $\hat{\theta}_{21}$ both tend towards zero whereas the other main diagonal and off diagonal elements remain quite stable as the level of contamination increases. However, the Frobenius norm, an elementwise norm, exhibits a minimum turning point before levelling off. This is due to the first row and column converging rapidly to zero, often from relatively large starting points, whereas the convergence of the other elements is not as drastic and not necessarily shrinking towards zero, hence the upward trend.

The entropy loss also broadly exhibits similar behaviour in both Figure 4.3 and 4.4. The minimum turning point in Figure 4.4 is somewhat similar to that of the Frobenius norm and is explained by noting that this is due to the competition between the sum of eigenvalues and the sum of the logs of the eigenvalues – the sum of the eigenvalues decays quite quickly before levelling off as s_{11} increases, whereas the sum of the logs of the eigenvalues decays much more slowly, as shown in Figure 4.6. Regardless, it is clear that the entropy loss tends to reflects the impact of the inflated variance in both the covariance matrix and the resulting precision matrix.

4.3.3.2 *Contamination in the data*

Instead of directly manipulating the estimated covariance matrix, consider introducing contamination into the original data set and observing what effect that has on the performance metrics applied to the covariance and resulting precision matrix. For each level of contamination we take $N = 1000$ samples from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. For each sample we estimate the classical sample covariance matrix, $\boldsymbol{S}$, and take the inverse to obtain $\hat{\boldsymbol{\Theta}} = \boldsymbol{S}^{-1}$.

Figures 4.7 and 4.8 show the behaviour of the various loss indices over $N = 1000$ replications. The horizontal axis represents the number of contaminated observations within each of the $p = 30$ variables. Starting with an uncontaminated multivariate Gaussian distribution, we then progress-

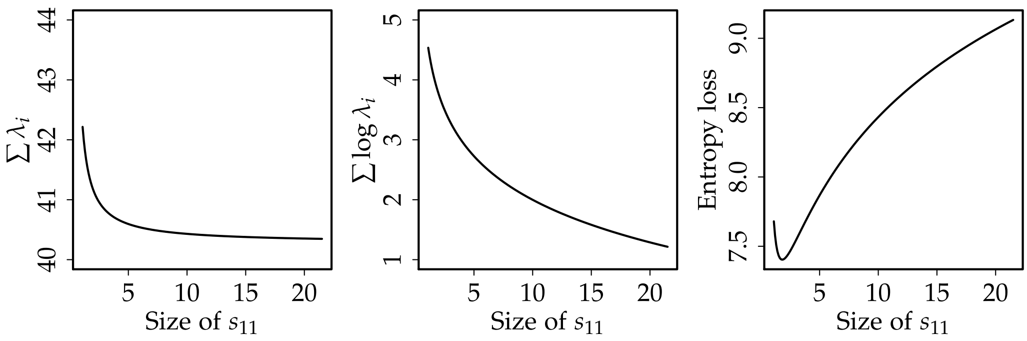


Figure 4.6: The relative speed of decay for the sum of the eigenvalues and the sum of the log of the eigenvalues, which results in the non-monotonic behaviour of the entropy loss, $\sum(\lambda_i - \log \lambda_i) - p$, where λ_i are the eigenvalues of $\hat{\boldsymbol{\Theta}} = \boldsymbol{S}^{-1}$.

ively add one contaminated observation to each variable until there are 24 contaminated observations within each variable. The contamination is performed by assigning each randomly selected cell the value of 10.

In general, scattered outlying contamination will destroy any existing dependence structure and inflate the main diagonal of the covariance matrix, resulting in increases for the entropy loss and matrix norms applied to the covariance as seen in Figure 4.7. The log determinant of the estimated covariance matrix trends upwards, demonstrating the over estimated generalised variance with increasing levels of contamination. Apart from a relatively minor spike when there is only one contaminated observation in each variable, the condition number is not adversely affected by increasing levels of contamination, reflecting the stabilised eigenvalues of the resulting covariance matrix. This demonstrates that in this setting, the condition number is not an appropriate index against which to compare the performance of competing robust estimators.

The interpretation of the performance indices when they are applied to the resulting precision matrix is more complicated. We see in Figure 4.8 that there is still structure present in the precision matrix, in the sense that there is a main diagonal behaving distinctly from the off diagonal elements. However, all the elements tend to shrink towards zero. Hence, for large amounts of contamination, $\hat{\boldsymbol{\Theta}} - \boldsymbol{I} \approx -\boldsymbol{I}$ and so the matrix norms tend to converge to $|| - \boldsymbol{I}||$.

As in the previous scenario, the Frobenius norm exhibits a minimal turning point before plateauing. This is explained by noting that while the introduction of contamination has an immediate shrinkage effect on the main

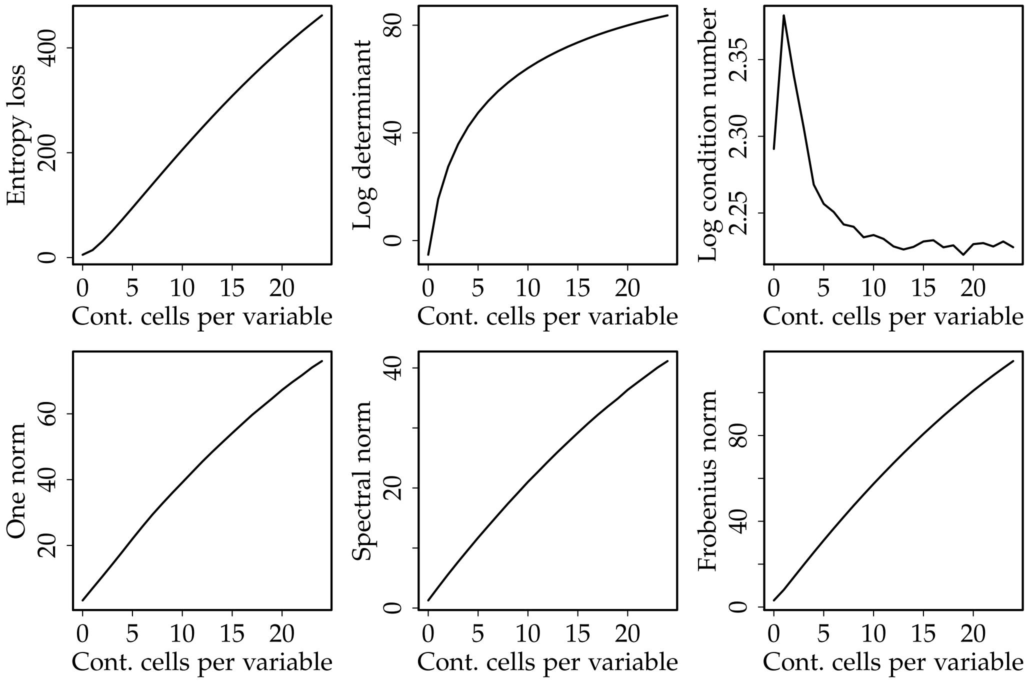


Figure 4.7: The impact of randomly contaminating a certain number of cells in each variable of a 100 × 30 data matrix on the various performance indices when applied to the resulting *covariance* matrix.

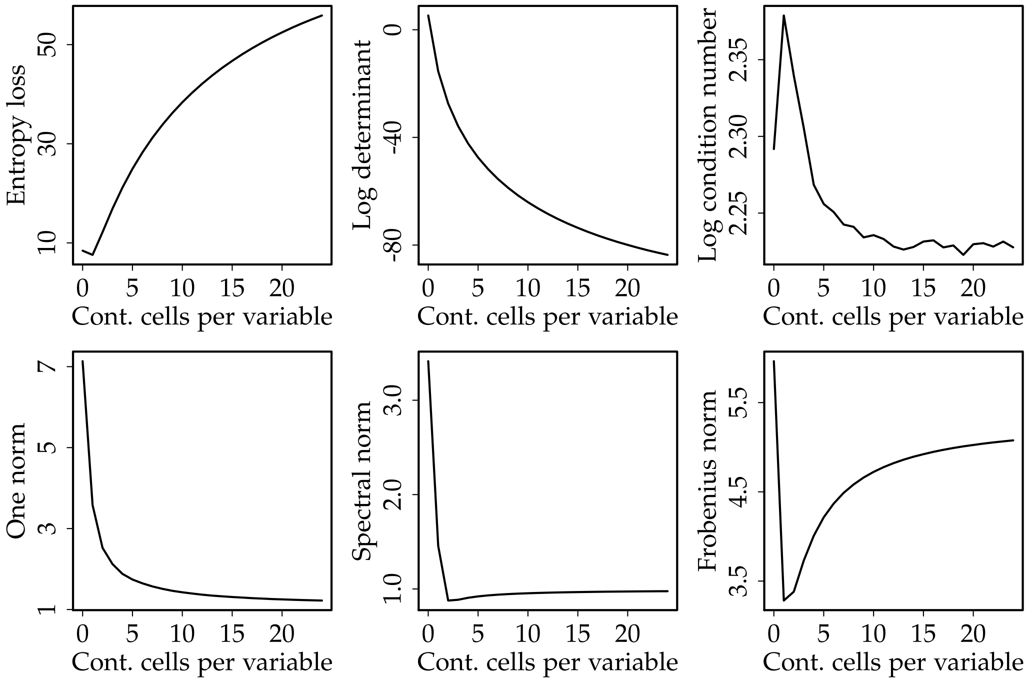


Figure 4.8: The impact of randomly contaminating a certain number of cells in each variable of a 100 × 30 data matrix on the various performance indices when applied to the resulting *precision* matrix.

diagonal of the precision matrix, including one or two influential observations in each variable induces an artificially high level of correlation between some variables. Hence, it can take some time for the off diagonal elements to stabilise. When more than a few outlying cells are present in each variable, the artificial correlation structure wanes and hence the Frobenius norm applied to the resulting precision matrix trends towards $||\boldsymbol{I}||_F = \sqrt{p}$. Hence, in this contamination setting the matrix norms appear to be useful only when applied to the covariance matrix, not the precision matrix.

As in the previous scenario, the entropy loss behaves consistently for both the covariance and precision matrix. Similarly to Figure 4.4, it exhibits a slight drop when only one cell in each variable is contaminated after which it increases as the proportion of contaminated cells grows. As such, the entropy loss is the preferred performance index when looking across both covariance and precision matrix estimators.

4.4 PAIRWISE COVARIANCE MATRIX ESTIMATION

The premise behind pairwise covariance matrix estimation is to take the $\binom{p}{2}$ pairs of variables and robustly estimate the covariance between each pair, as discussed in Chapter 3. As we will show, the primary advantage is robustness to scattered contamination in the data set. The main disadvantage of this approach is that the resulting symmetric matrix is not guaranteed to be PSD or affine equivariant.

Given a stochastic $n \times p$ matrix $\boldsymbol{X}$ and a fixed nonsingular $p \times p$ matrix $\boldsymbol{A}$, an estimator $\hat{\boldsymbol{\Sigma}}(\cdot)$ is equivariant if $\hat{\boldsymbol{\Sigma}}(\boldsymbol{X}\boldsymbol{A}) = \boldsymbol{A}\hat{\boldsymbol{\Sigma}}(\boldsymbol{X})\boldsymbol{A}^\top$. As noted earlier, in the scattered contamination model, affine equivariance is unachievable as there is the potential for all rows to have a contaminated cell, hence linear combinations of the rows propagate the contamination.

4.4.1 *OGK procedure*

To overcome the lack of positive semidefiniteness, Maronna and Zamar (2002) propose a modification based on the observation that the eigenvalues of the covariance matrix are the variances along the directions given by the respective eigenvectors. The procedure is known as the Orthogonalised Gnanadesikan Kettenring (OGK) estimator and is as follows.

A. Let $\boldsymbol{D} = \text{diag}(s(\boldsymbol{x}_1), \ldots, s(\boldsymbol{x}_p))$, where s is a robust scale estimator and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ are the columns of $\boldsymbol{X}$. Standardise $\boldsymbol{X}$ such that $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{D}^{-1}$.

B. Compute the "correlation matrix" applying $s(\cdot)$ to the columns of $\boldsymbol{Y}$ to obtain $\boldsymbol{U} = \big(u_{jk}\big)$, where $u_{jj} = 1$ and

$$u_{jk} = \frac{1}{4}(s(\boldsymbol{y}_j + \boldsymbol{y}_k)^2 - s(\boldsymbol{y}_j - \boldsymbol{y}_k)^2), \quad \text{for } j \neq k.$$

C. Find the eigenvalues λ_j and orthonormal eigenvectors $\boldsymbol{e}_j$ of $\boldsymbol{U}$. Let $\boldsymbol{E} = \big[\boldsymbol{e}_1, \ldots, \boldsymbol{e}_p\big]$, so $\boldsymbol{U} \equiv \boldsymbol{E}\boldsymbol{\Lambda}\boldsymbol{E}^\mathsf{T}$ with $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_p)$.

D. Let $\boldsymbol{Z} = \boldsymbol{Y}\boldsymbol{E}$ be the matrix of "principal components" vectors with corresponding robust variance estimates, $\hat{\boldsymbol{\Lambda}} = \operatorname{diag}(s(\boldsymbol{z}_1)^2, \ldots, s(\boldsymbol{z}_p)^2)$.

E. The estimated covariance matrix is then, $\hat{\boldsymbol{\Sigma}} = \boldsymbol{D}^2\boldsymbol{E}\hat{\boldsymbol{\Lambda}}\boldsymbol{E}^\mathsf{T}$.

Note that even if the original covariance matrix was already PD (all eigenvalues are positive) applying the OGK procedure will not necessarily return the same matrix due to the use of $\hat{\boldsymbol{\Lambda}}$ as an approximation to $\boldsymbol{\Lambda}$. In essence, the eigenvalues of $\boldsymbol{\Lambda}$ are replaced by robust variance estimates which are guaranteed to be positive ensuring that the resulting matrix is PD.

Maronna and Zamar (2002) and Maronna, Martin and Yohai (2006, p. 207) suggest that the OGK estimator can be improved by iterating the procedure and then using this estimate to find robust Mahalanobis distances for each observation vector. These are then used to screen for outliers before applying the classical covariance estimator to the cleaned data. This is done in an effort to increase efficiency and to make the result "more equivariant". The robust Mahalanobis distances are,

$$d_i = d(\boldsymbol{x}_i) = (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})^\mathsf{T}\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}),$$

for some robust location estimate $\hat{\boldsymbol{\mu}}$. Let w be a weight function, and define $\hat{\boldsymbol{\mu}}_w$ and $\hat{\boldsymbol{\Sigma}}_w$ as the weighted mean and covariance matrix, where each $\boldsymbol{x}_i$ has weight $w(d_i)$, that is,

$$\hat{\boldsymbol{\mu}}_w = \frac{\sum_i w(d_i)\boldsymbol{x}_i}{\sum_i w(d_i)} \quad \text{and} \quad \hat{\boldsymbol{\Sigma}}_w = \frac{\sum_i w(d_i)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_w)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_w)^\mathsf{T}}{\sum_i w(d_i)}.$$

The standard weight function is "hard rejection," with $w(d) = \mathbb{I}\{d \leq d_0\}$ where

$$d_0 = \frac{\chi_p^2(\beta)\operatorname{Med}\{d_1, \ldots, d_n\}}{\chi_p^2(0.5)},$$

and $\chi_p^2(\beta)$ is the β quantile of the χ^2 distribution with p degrees of freedom. In their simulations, Maronna and Zamar (2002) find that $\beta = 0.9$ generally yields good results. This is used as the default in the `robustbase` R package although Martin Maechler "strongly believes that the hard threshold currently in use is too arbitrary, and further that soft thresholding should be used instead" (Rousseeuw, Croux et al., 2013).

In terms of the impact of not being affine equivariant, Maronna and Zamar (2002) note that "although the worst case may differ from the original data, for most transformations the results are very similar" and "the lack of equivariance is not a serious concern in our estimates".

Regardless, neither the OGK method nor the reweighted OGK method is able to cope with scattered contamination. The issue of outlier propagation means that the number of contaminated "principal components", i.e. columns of $\boldsymbol{Z} = \boldsymbol{Y}\boldsymbol{E}$, could easily be greater than 50% even for small levels of cellwise contamination. Hence, the robust variance estimates that are used in place of the eigenvalues will no longer be valid estimates – they will be in breakdown. Furthermore, the reweighting step will often needlessly exclude many observation vectors where there is only one contaminated cell.

4.4.2 *NPD procedure*

Higham (2002) considers the problem of computing the nearest positive definite (NPD) matrix to a given symmetric matrix. The motivation stems from finance, where sample covariance matrices are constructed from vectors of stock returns, however, the problem arises when not all stocks are observed every day. In this setting, classical covariances may be computed on a pairwise basis using data drawn only from days where both stocks have data available. The resulting covariance matrix is not guaranteed to be PD because it has been built from inconsistent data sets. Motivated by the same problem, Løland et al. (2013) propose both a pseudo-likelihood and a Bayesian approach to find PD estimates of pairwise correlation matrices. However, their approach relies on expert knowledge to formulate priors for the pairwise covariances.

In contrast to the OGK procedure, if the initial symmetric matrix is PD, then the NPD method simply returns the original pairwise covariance matrix. A potential advantage of the NPD method over the OGK procedure is its ability to cope with scattered contamination.

Formally, for an arbitrary symmetric $p \times p$ matrix $\boldsymbol{A}$, the aim is to find the distance

$$\gamma(\boldsymbol{A}) = \min\{||\boldsymbol{A} - \boldsymbol{W}||_F : \boldsymbol{W} \text{ is a symmetric PD matrix}\}, \tag{4.1}$$

and the resulting matrix that achieves this minimum distance. Higham (2002) uses the Frobenius norm as it is "the easiest norm to work with for this problem and also being the natural choice from the statistical point of view".

Framing the problem in terms of projecting from the set of symmetric matrices onto the set of symmetric PD matrices, with respect to the Frobenius norm, facilitates the use of standard results in approximation theory (Luenberger, 1969, p. 69). In particular, it follows that the minimum in (4.1) is achieved and that the minimiser is achieved at a unique matrix, $\hat{\boldsymbol{W}}$.

While Higham (2002) considers a variety of weighting mechanisms, in the simplest case without specifying any weights, the procedure is quite straightforward. Let, $\hat{\boldsymbol{W}} = \boldsymbol{E}\hat{\boldsymbol{\Lambda}}\boldsymbol{E}^{\mathsf{T}}$, where $\boldsymbol{E}\boldsymbol{\Lambda}\boldsymbol{E}^{\mathsf{T}}$ is the spectral decomposition of $\boldsymbol{A}$, with $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_p)$ and $\hat{\boldsymbol{\Lambda}} = \text{diag}(\max\{\lambda_i, \delta\})$, where δ is a small positive constant. The NPD procedure is similar to the OGK procedure in that it performs a spectral decomposition and then updates the eigenvalues to ensure that the result is PD. However, it does not rely on linear transformations of the original dataset and hence is not affected by the "outlier propagation" issue associated with cellwise contamination.

4.4.3 *Comparing OGK and NPD*

To compare the OGK and NPD methods in the presence of scattered contamination we performed a simulation study. The underlying data generating process was multivariate Gaussian, $\mathcal{N}(\mathbf{0}, \boldsymbol{\Theta}^{-1})$, with $n = 100$ and $p = 60$ where the true precision matrix $\boldsymbol{\Theta}$ had a randomly generated sparsity pattern. Outliers, generated from a t_{10} distribution scaled by a factor of $\sqrt{10}$ for moderate outliers or 10 for extreme outliers, were introduced incrementally to each variable. This is a similar contaminating design to that used in the full simulation study in Section 4.6. For each level of contamination, $N = 100$ replications of the experiment were performed, with new data generated at each replication and outliers randomly allocated.

Figure 4.9 shows the percentage relative improvement in the average Frobenius norm relative to $\hat{\boldsymbol{\Sigma}}_0$, the classical covariance estimator with no contamination. Of the OGK methods, only that which uses P_n as the underlying scale estimator is shown, however the OGK methods based on Q_n, τ-scale also behaved similarly poorly. Special mention should be made of the OGK method paired with the MAD, which performed substantially worse than

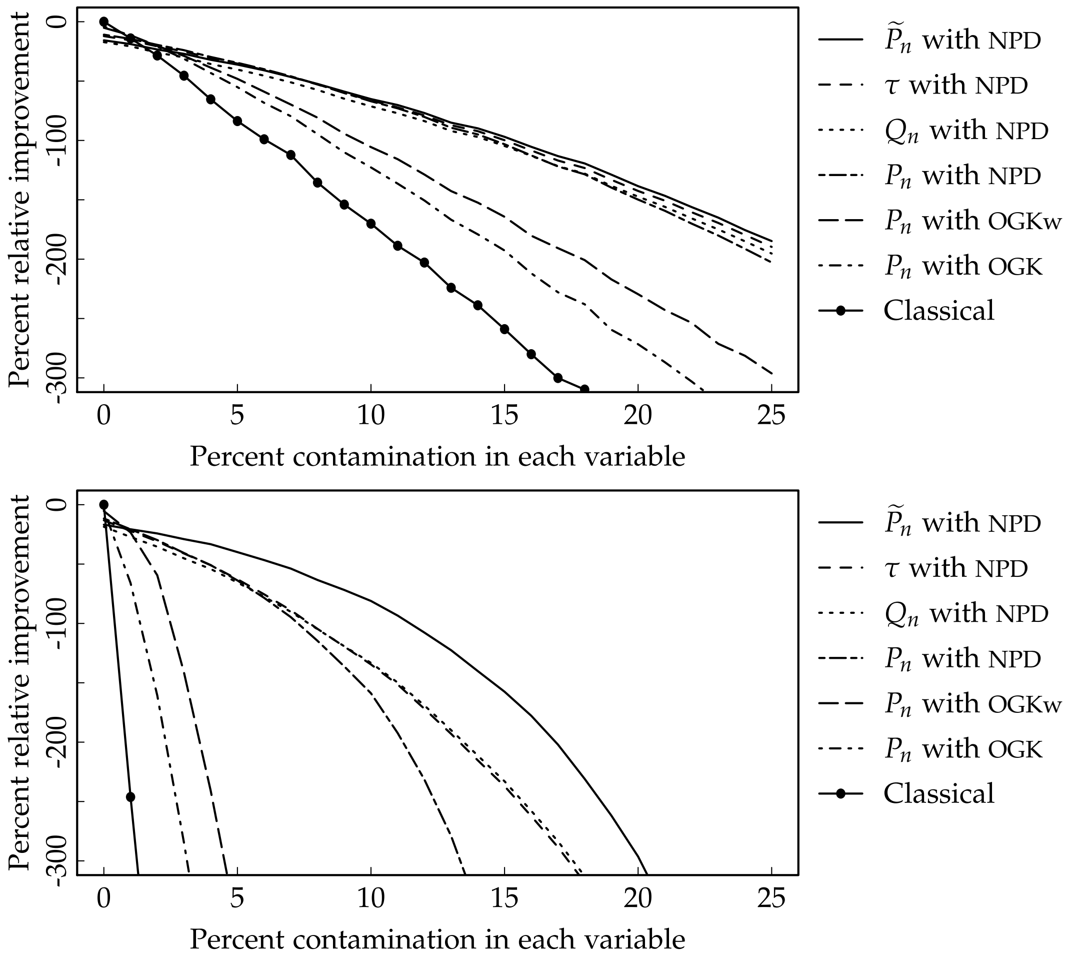


Figure 4.9: Percent relative improvement in average Frobenius norm over $N = 100$ replications for various covariance estimation techniques relative to the classical method with no contamination for $p = 60$. The outliers are independently generated from a t_{10} distribution scaled by $\sqrt{10}$ (top) or 10 (bottom).

the other robust scale estimators considered. The reweighted Orthogonalised Gnanadesikan Kettenring (OGKw) methods performed only marginally better, again only the OGKw method with P_n is shown, but it is extremely close to the performance of the other OGKw methods, including that based on the MAD, as the reweighting step identifies and deletes the same outlying observation vectors. Comparing the top and bottom panels of Figure 4.9 it is clear that the performance of both the standard and reweighted OGK methods deteriorate significantly when the extremity of the outliers increases, reflecting the inability of these methods to deal with scattered contamination.

For moderate outliers the NPD methods based on P_n, Q_n and the τ-scale estimator perform similarly, however for more extreme outliers the more robust estimators Q_n and τ-scale perform better than P_n after a certain level of contamination is reached. In contrast, the NPD method based on the adaptively trimmed P_n with trimming parameter $d = 3$, henceforth denoted $\widetilde{P}_n$, performs remarkably well, largely due to the ease with which it identifies and trims any pairs of observations with an outlier. This performance is representative of the behaviour under the other matrix norms.

Despite the good performance with respect to the norms, the NPD method often results in estimated matrices with a number of extremely small eigenvalues which give poorly conditioned estimates, i.e. the condition number of these estimators is very high as is the entropy loss, which involves the log of the eigenvalues. In general, it is not recommended to use either the OGK nor the NPD in isolation when there is scattered contamination present. Even in the presence of standard row-wise contamination, the NPD method is not recommended due to its propensity to return poorly conditioned estimates.

4.5 PRECISION MATRIX ESTIMATION

Many statistical procedures are primarily concerned with the precision matrix, the inverse of a covariance matrix, rather than the covariance matrix itself. For example, finding Mahalanobis distances and performing linear discriminant analysis both require an estimate of $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$. Finding good precision matrix estimates has been a focus of many investigators over a long period of time, the first major contribution being Dempster (1972).

There is extensive interest in estimating *sparse* precision matrices in moderate and high dimensions. We will restrict attention to cases where the uncontaminated data come from a Gaussian distribution, that is we have a Gaussian graphical model (Lauritzen, 1996). Under this model pairwise conditional independence between variables X_j and X_k holds if and only if $\theta_{jk} = 0$, hence inferring linkages between variables corresponds to identifying the nonzero elements of $\boldsymbol{\Theta} = (\theta_{jk})$.

The following routines take as an input an estimated covariance matrix and output a regularised precision matrix. In Section 4.6 we demonstrate the advantages of using a robust pairwise covariance matrix estimate as the input to these regularisation routines.

4.5.1 *GLASSO*

A natural way to estimate $\boldsymbol{\Theta}$ is by maximising the log-likelihood of the data. With Gaussian observations, the log-likelihood takes the form,

$$\log|\boldsymbol{\Theta}| - \operatorname{tr}(\boldsymbol{S}\boldsymbol{\Theta}), \tag{4.2}$$

where $\boldsymbol{S}$ is an estimate of the covariance matrix of the data. Maximising (4.2) with respect to $\boldsymbol{\Theta}$ leads to the MLE, $\boldsymbol{S}^{-1}$. In general, $\boldsymbol{S}^{-1}$ will not be sparse, in the sense that it will contain no elements exactly equal to zero. Furthermore in $p > n$ situations $\boldsymbol{S}$ will be singular so the MLE cannot be computed. Yuan and Lin (2007) consider minimising the penalised negative log-likelihood,

$$\operatorname{tr}(\boldsymbol{S}\boldsymbol{\Theta}) - \log|\boldsymbol{\Theta}| + \lambda \textstyle\sum_{i,j} |\theta_{ij}|, \tag{4.3}$$

over the set of PD matrices where λ is a tuning parameter to control the amount of shrinkage. This is the graphical lasso (GLASSO) (Friedman, Hastie and Tibshirani, 2008), which has two major advantages over (4.2): the solution is PD for all $\lambda > 0$ even if $\boldsymbol{S}$ is singular, and for large values of λ the resulting estimate, $\hat{\boldsymbol{\Theta}}$, will be sparse.

4.5.2 *QUIC*

The quadratic inverse covariance (QUIC) method solves the same minimisation problem as the GLASSO. The improvement in speed comes from noticing that the Gaussian log-likelihood component of (4.3) is twice differentiable and strictly convex which lends itself to a quadratic approximation and hence faster convergence (Hsieh et al., 2011). On the other hand, the penalty term is convex but not differentiable and so is treated separately. Algorithm 1 briefly outlines the procedure.

Algorithm 1 QUIC

```
Input: Symmetric matrix $\hat{\boldsymbol{\Sigma}}_0$, scalar $\lambda$, stopping tolerance $\epsilon$
1: for $i = 0, 1, \ldots$ do
2:     Compute $\boldsymbol{\Sigma}_t = \boldsymbol{\Theta}_t^{-1}$
3:     Find second order approximation
4:     Partition variables into free and fixed sets
5:     Find Newton direction $\boldsymbol{D}_t$ using coordinate descent over the set of
       free variables (lasso problem)
6:     Determine the step-size $\alpha$ such that $\boldsymbol{\Theta}_{t+1} = \boldsymbol{\Theta}_t + \alpha \boldsymbol{D}_t$ is PD and the
       objective function sufficiently decreases
7: end for
Output: Sequence of $\boldsymbol{\Theta}_t$
```

The QUIC routine explicitly includes a step that ensures positive definiteness of the precision matrix for each iteration. Furthermore, as implemented in the R package `QUIC`, it accepts a symmetric matrix as its input, which was the main reason we initially considered this method (other than its speed advantages) as we entertained the idea that we could input robust pairwise covariance matrices directly into the QUIC routine.

Preliminary results were promising, particularly for low levels of scattered contamination.[8] However, further investigation found that this approach does not perform as well as might be expected for moderate and high levels of scattered contamination. This is understandable given that step 6 in the QUIC routine, the step that ensures the result is PD, is quite crude and was never designed to make such a significant correction as required when converting a symmetric matrix full of pairwise covariances to be PD. The QUIC routine essentially moves the estimated inverse as far as necessary in a predetermined direction to ensure that the result is PD. This correction works well when the symmetric matrix is not *too far* from being PD, i.e. when the negative eigenvalues are very small. However when there is a high level of contamination in the data, the symmetric matrix of pairwise covariances may be *far* from PD which leads to extremely poor performance.

We subsequently found that applying the NPD method to the pairwise covariance matrix before using the QUIC routine resulted in substantially better results. Hence, we show in Section 4.6, that it is preferable to use a method designed specifically for the purpose of converting a symmetric matrix to a PD matrix, such as the NPD method, before applying the QUIC

[8] These preliminary results were presented at ICORS 2013.

routine. Furthermore, as expected given that they are solving the same minimisation problem, the QUIC estimates are virtually indistinguishable from the standard GLASSO approach in all scenarios considered here.

4.5.3 *CLIME*

We established that the best way forward is to preprocess the symmetric matrix of pairwise covariances. This opens up the possibility of using regularisation routines other than QUIC. One such alternative is constrained L_1 minimisation for inverse covariance matrix estimation (CLIME), implemented in the R package `clime` (Cai, Liu and Luo, 2011, 2012).

The CLIME routine uses linear programming to solve the following (convex) optimisation problem,

$$\boldsymbol{\Theta}^\star = \min |\boldsymbol{\Theta}|_1 \quad \text{subject to: } |\boldsymbol{S}\boldsymbol{\Theta} - \boldsymbol{I}|_\infty \leq \lambda,$$

where $\boldsymbol{S}$ is the sample covariance matrix and $|\boldsymbol{A}|_1 = \sum_{i,j} |a_{ij}|$ is the *element-wise* L_1 norm of a matrix, $\boldsymbol{A}$, and $|\boldsymbol{A}|_\infty = \max_{1 \leq i \leq p, 1 \leq j \leq q} |a_{ij}|$ is the *element-wise* infinity norm. No symmetry requirements are placed on $\boldsymbol{\Theta}^\star$ so a symmetrising step is applied to obtain the final solution, $\hat{\boldsymbol{\Theta}}$,

$$\hat{\theta}_{ij} = \hat{\theta}_{ji} = \theta^\star_{ij} \mathbb{I}\{|\theta^\star_{ij}| \leq |\theta^\star_{ji}|\} + \theta^\star_{ji} \mathbb{I}\{|\theta^\star_{ij}| > |\theta^\star_{ji}|\}.$$

Theorem 1 of Cai, Liu and Luo (2011) shows that the resulting $\hat{\boldsymbol{\Theta}}$ is PD with high probability.

The simulation study in the next section shows that there is little difference between using CLIME and QUIC – the key point is that both appear to perform well in the presence of scattered contamination when the input matrix is based on pairwise robust covariance estimates that has been made PD using the NPD routine.

4.6 SIMULATION STUDY

This section presents the results of an extensive simulation study to assess how well various robust covariance estimation techniques perform when used as an input to the regularisation routines outlined previously.

The proposed estimator begins by finding the covariances between all $\binom{p}{2}$ pairs of variables. For the scale estimator underlying the robust covariance estimator, we have considered Q_n, the τ-scale, the MAD and the IQR. We also consider the P_n estimator, defined in Chapter 2, and two adaptively trimmed variants $\widetilde{P}_n$, with trimming parameters $d = 5$ and $d = 3$, see Section 2.4 for further details about these estimators. The pairwise covariances are then arranged in a symmetric, though not necessarily PD, matrix. The symmetric matrix is transformed to a PD matrix using either the OGK method or the NPD method before being input into the GLASSO, QUIC or CLIME regularisation routines. For comparison purposes we also included the classical covariance estimator and the MCD as initial covariance matrix estimates.

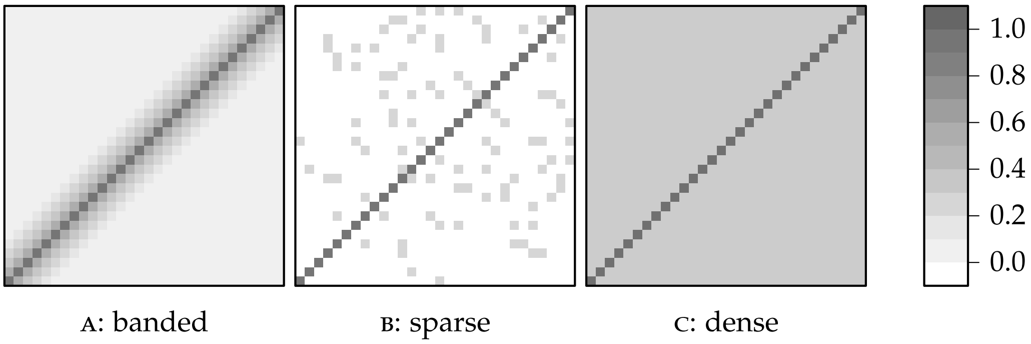
4.6.1 *Design*

Throughout the investigation, a number of different simulations under numerous designs have been performed. Three types of precision matrices have been selected as the basis for the data generating process. They represent a broad range of scenarios that occur in practice and are similar to those used in Cai, Liu and Luo (2011) and Hsieh et al. (2011). The simulated data comes from a multivariate Gaussian distribution with $n = 100$ observations, $\mathcal{N}(\mathbf{0}, \boldsymbol{\Theta}^{-1})$. The scenarios outlined below are shown in Figure 4.10 and the code for generating each is given in Appendix D.

A. Banded precision matrices, with elements $\theta_{ij} = 0.6^{|i-j|}$, such that the values of the entries decay the further they are from the main diagonal.

B. Sparse precision matrices with randomly allocated non-zero entries, where $\boldsymbol{\Theta} = \boldsymbol{B} + \delta \boldsymbol{I}$ with each off diagonal entry in $\boldsymbol{B}$ generated independently, where $P(b_{ij} = 0.5) = 0.1$ and $P(b_{ij} = 0) = 0.9$ and δ is chosen such that the condition number of the matrix equals p. The matrix is then standardised to have diagonal components equal to one. This scenario will be referred to as scattered sparsity.

C. Dense precision matrices, where $\boldsymbol{\Theta}$ has all off diagonal elements equal to 0.5 and diagonal elements equal to 1.



A: banded B: sparse C: dense

Figure 4.10: Heat maps of the three kinds of precision matrices used to generate the data.

The outliers were generated independently for each variable. In our simulations we allow the number of contaminated observations within each variable to increase up to a maximum of 25 observations (out of $n = 100$). In this way we have complete control over the total number of contaminated cells. The distribution of the outliers is a t_{10} distribution scaled by either a factor of 10 for extreme outliers or $\sqrt{10}$ for moderate outliers. The moderate outliers are perhaps closer to what one might expect in a real data set. However, the focus here is primarily on the extreme outliers where the overwhelming majority of the *unusual observations* lie well outside the cloud of standard observations. The extreme nature of the outliers serves to clearly demark estimators that have effectively broken down from those that are still capable of giving *ballpark correct* results. In both cases, the outliers are symmetrically distributed. An example of what this contaminating model looks like in the $p = 2$ case is shown in Figure 4.11 where there is 10% scattered contamination in both variables. Usually only one component in each observation vector is contaminated, but of course, on average 1% of observation vectors will have both components contaminated.

Each of the regularisation routines require a tuning parameter. At each replication, the tuning parameter was obtained by training on a separate (uncontaminated) randomly generated data set drawn from the true data generating process. For the training data, a sequence of precision matrices were obtained and the value of the tuning parameter corresponding to the smallest entropy loss was then used for that replication. In practice, there was a small amount of variability in the choice of tuning parameter within each scenario and dimensionality. Furthermore, the QUIC and GLASSO routines almost always picked the same tuning parameter and the CLIME routine chose tuning parameters in a similar region.

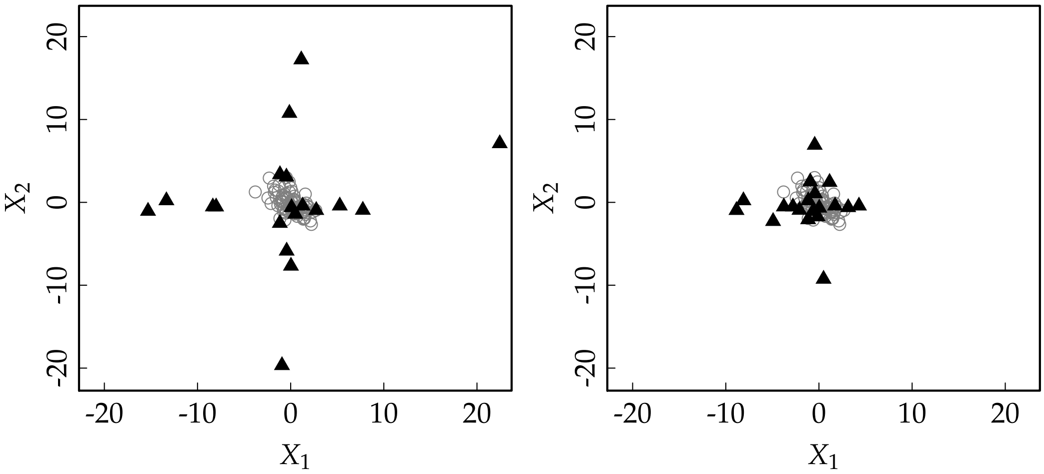


Figure 4.11: An example data set with 10% scattered contamination in each variable. Observations with at least one contaminated component are represented by ▲. The contaminating distribution is t_{10} scaled by a factor of 10 (left) or $\sqrt{10}$ (right).

As in Lin and Perlman (1985), for the entropy loss we report the results in terms of the percentage relative improvement in average loss (PRIAL),

$$\text{PRIAL}(\hat{\boldsymbol{\Theta}}) = \frac{L_1(\boldsymbol{\Theta}, \hat{\boldsymbol{\Theta}}_0) - L_1(\boldsymbol{\Theta}, \hat{\boldsymbol{\Theta}})}{L_1(\boldsymbol{\Theta}, \hat{\boldsymbol{\Theta}}_0)} \times 100,$$

where $\hat{\boldsymbol{\Theta}}_0$ is the estimated precision matrix after a regularisation technique has been applied to the classical sample covariance matrix for uncontaminated data. It is important to note that this is an extremely harsh benchmark to set.

4.6.2 *Results*

4.6.2.1 *No contamination*

Any good robust method should give comparable results to the classical non-robust method it is replacing when presented with a *clean* dataset. Table 4.1 presents the PRIAL results. As the PRIAL results are relative to the base case for each routine, Table 4.1 cannot be used to compare the performance of the CLIME routine to the QUIC routine.

In the uncontaminated case, the OGK method substantially outperforms the NPD method. Overall, the methods appear to improve as the dimensionality increases, however, this is more a reflection of the deteriorating absolute performance of the baseline classical covariance matrix estimate.

For $p = 30$, the pairwise methods outperform the MCD, however the MCD method uses $\lfloor n+p+1 \rfloor /2$ observations so when $p = 90$ the resulting estimator is the classical covariance estimate applied to 95 out of a total $n = 100$ observations. Hence, it is not surprising that the PRIAL for the MCD method is so close to zero. In fact, the MCD is not recommended for use when $n < 2p$ (Rousseeuw, Croux et al., 2013).

The reweighted OGK (OGKw) methods essentially perform outlier detection and deletion before returning a classical covariance estimate of the *cleaned* data set. The performance of these methods is broadly similar over all the various initial scale estimates. Though not shown in Table 4.1, the MAD performs particularly poorly under both the OGK and the NPD corrections and would not be recommended for use.

As would be expected, given the solid Gaussian performance of P_n (see Section 2.4), the methods based on P_n outperform those based on the τ-scale and Q_n. The relative deterioration in performance for the robust methods compared to the classical method is comparable to that in the simple univariate scale case. Recall, the univariate scale estimator P_n has a Gaussian relative efficiency of 86%.

		$p = 30$		$p = 60$		$p = 90$	
		CLIME	QUIC	CLIME	QUIC	CLIME	QUIC
τ-scale	OGK	-17.1	-13.0	-15.1	-9.5	-12.7	-8.1
	OGKw	-34.8	-26.9	-29.3	-15.9	-20.3	-15.6
	NPD	-31.5	-25.3	-31.3	-16.7	-27.6	-15.4
Q_n	OGK	-16.5	-13.6	-13.3	-10.5	-11.7	-9.3
	OGKw	-34.7	-27.0	-28.5	-15.6	-12.3	-14.6
	NPD	-40.6	-32.5	-36.9	-21.9	-31.4	-20.3
P_n	OGK	-13.6	-11.8	-12.7	-9.6	-10.9	-8.8
	OGKw	-33.7	-26.2	-27.1	-14.9	-17.5	-14.5
	NPD	-19.9	-15.7	-18.4	-11.4	-18.2	-11.2
MCD		-53.1	-56.2	-19.5	-19.5	-3.7	-3.7

Table 4.1: PRIAL results for the various estimators when there is no contamination present.

4.6.2.2 *Cellwise contamination*

There are a number of ways to compare and contrast the various estimators. We consider data with $n = 100$ observations from three different data generating processes, contaminated with either moderate or extreme outliers across four dimensions, $p = 15$, 30, 60 or 90. We implement an array of initial covariance estimation techniques and process these through the GLASSO, QUIC and CLIME regularisation routines. Finally, as outlined in Section 4.3.3, there are a number of performance indices that are considered. This section extracts and synthesises the key results.

We first consider the effect of dimensionality on the performance of the various estimators. A typical example is shown in Figure 4.12 where we plot the PRIAL results for the precision matrix resulting from the CLIME procedure for various input covariance matrices across different amounts of extreme contamination in each variable. The original data was generated assuming a banded precision matrix, however the trend holds true for scattered sparsity and dense precision matrices as well as for the QUIC and GLASSO procedures.

For relatively low dimensions, such as in the top panel of Figure 4.12 where $p = 15$, there is clearly an advantage to using the NPD method over the OGK method once there is more than a few percent of observations in each variable being contaminated. To avoid clutter, only the OGK method with P_n has been included in the plots, however, it is representative of the performance of the other scale estimators when used in conjunction with the OGK method.

As the dimensionality increases, the OGK and the MCD methods deteriorate faster. When $p = 90$, as outlined in the previous section, the MCD method behaves like the classical method. The OGK method performs similarly poorly as outlier propagation can lead to more than half of the elements in each principle component vector being contaminated. Hence, the eigenvalues in the spectral decomposition are replaced with robust estimates of scale that may no longer be valid.

Remarkably, the NPD methods perform consistently well. Their performance, relative to the classical method with no contamination improves as the number of variables increases. The raw entropy loss plots are found in Figure D.1 in Appendix D. The P_n based method performs well for low levels of contamination, however once the proportion of contaminated cells is greater than 10% it does not perform as well as the other pairwise methods due to its lower breakdown value.

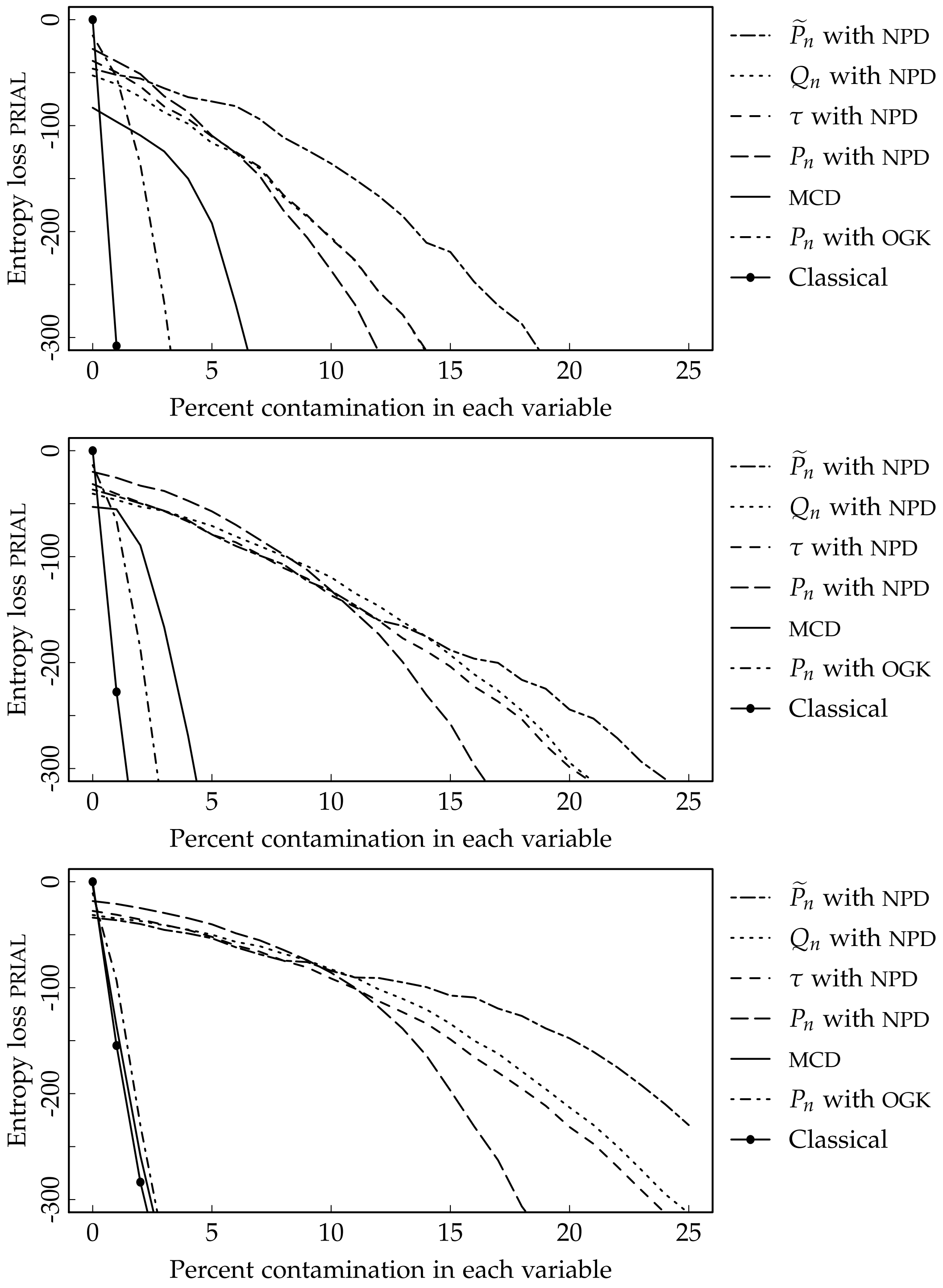


Figure 4.12: PRIAL results for a selection of estimators applied to data generated with a banded precision matrix with extreme outliers for $p = 15$ (top), $p = 30$ (middle) and $p = 90$ (bottom) using the CLIME regularisation procedure.

It is interesting to note that the adaptively trimmed P_n with adaptive trimming parameter $d = 3$, $\widetilde{P}_n$, follows a trajectory that is somewhat different to the rest of the NPD type estimators. It maintains a relatively high level of performance even for quite high levels of contamination. This is due to the extreme nature of the contamination making the adaptive trimming extremely effective in identifying and excluding the errant observations. The advantage of $\widetilde{P}_n$ is lost when the contaminating distribution has only moderately sized outliers, in which case all the NPD pairwise methods perform comparably because $\widetilde{P}_n$ does almost no trimming.

To summarise, for $p = 30$, $p = 60$ (shown in the top panel of Figure 4.13) and $p = 90$, using a pairwise method in conjunction with the NPD procedure as an input into the CLIME regularisation routine, the increase in entropy loss can be contained to less than double that of the classical method without contamination if the proportion of cellwise contamination is less than 10%.

The same pattern holds true when using the QUIC or the GLASSO regularisation routines. To demonstrate this consider Figure 4.13 where the PRIAL results are shown for CLIME, QUIC and the GLASSO under the banded precision matrix scenario with extreme outliers and $p = 60$. As would be expected the QUIC and GLASSO results are essentially identical, and largely consistent with the CLIME results in the top panel. Looking at the raw entropy loss numbers shown in Figure D.2 in Appendix D, the CLIME method gives slightly lower average entropy loss measurements, particularly for very high levels of contamination. In practice it does not matter what regularisation routine is used, the benefits of taking a pairwise approach to covariance estimation in the presence of scattered contamination will still hold.

The NPD pairwise approach is a major improvement over standard robust estimators. An example of this is given in Figure 4.14 where we present the average PRIAL results for the QUIC estimator with $p = 30$ across all three scenarios. The raw entropy loss numbers are found in Figure D.3 in Appendix D. Across all scenarios the same general pattern holds, the classical method and the OGK and MCD methods fail quite rapidly whereas the NPD approach offers much greater resilience to the scattered contamination.

For the banded precision matrix scenario, top panel of Figure 4.14, the NPD based methods under the various robust scale estimators give similar results with P_n having a slight advantage over the others for low levels of contamination whereas Q_n has an advantage for higher contamination proportions.

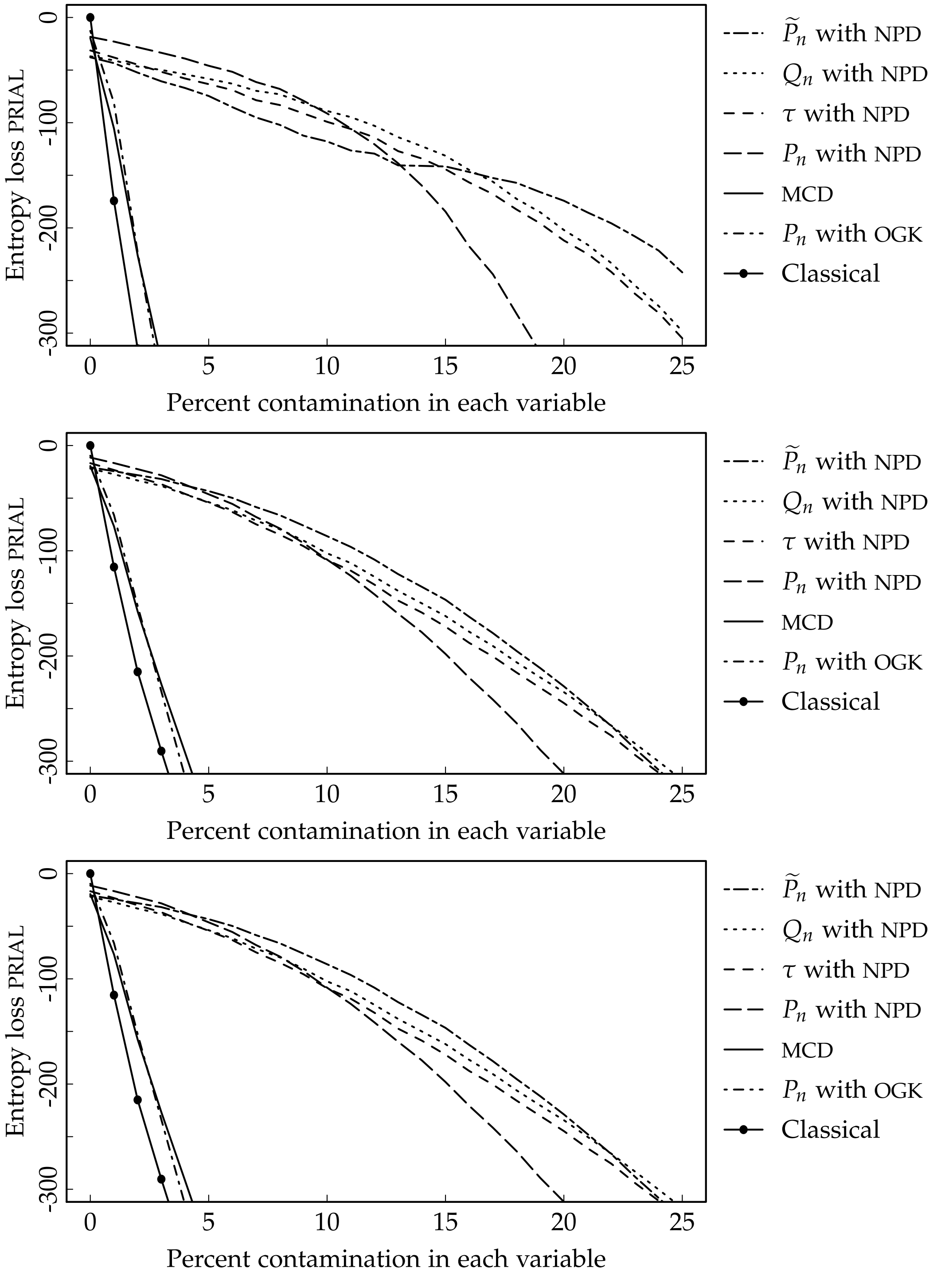


Figure 4.13: PRIAL results for a selection of estimators applied to data generated with a banded precision matrix with extreme outliers for $p = 60$ using CLIME (top), QUIC (middle) and GLASSO (bottom).

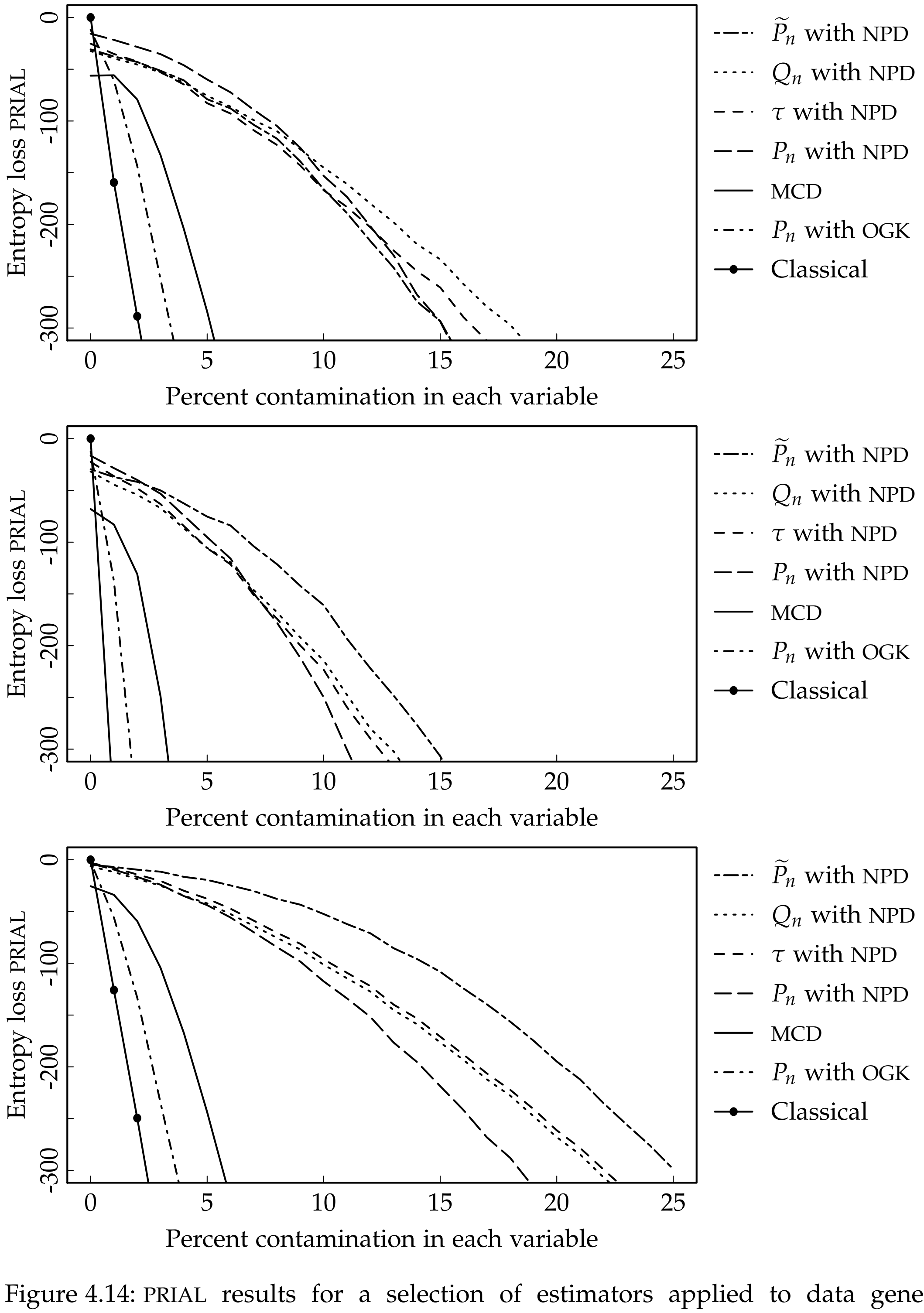


Figure 4.14: PRIAL results for a selection of estimators applied to data generated with a banded precision matrix (top), scattered precision matrix (middle) and dense precision matrix (bottom) with extreme outliers for $p = 30$ using the QUIC routine.

For the scattered precision matrix and the dense precision matrix scenarios, $\widetilde{P}_n$ gives the best results, however this is primarily due to the extreme nature of the outliers. The advantage of the adaptive trimming procedure is lost when the outliers are not so extreme, as demonstrated in Figure D.4 in Appendix D. However, it should be noted that when the outliers are not so extreme, the adaptive trimming approach still performs no worse than the other NPD methods.

We previously established that matrix norms are not a good performance measure for precision matrices. In terms of the other performance indicators, for all scenarios considered the log condition number remained bounded, suggesting that all three regularisation routines return well conditioned precision matrix estimates regardless of the level of contamination or the data generating process.

The NPD also performed well in terms of the log determinant index. Figure 4.15 shows the average log determinant of the precision matrices resulting from the GLASSO being applied to the various covariance matrix estimates under the three different scenarios. As with the entropy loss, there appears to be an advantage to using $\widetilde{P}_n$ over the other scale estimators in each of the scenarios. Unlike with the entropy loss, the advantage of the adaptive trimming procedure is still evident even when the contamination is less extreme, as seen in Figure D.5.

If we take the inverse of the resulting estimated precision matrix, we obtain a regularised estimate of the covariance matrix and can compare that to the *true* covariance matrix. Figure 4.16 presents the average Frobenius norm results for the resulting estimated covariance matrices after regularisation using the CLIME procedure. We see similar trends to those outlined earlier. While using P_n alone does not perform well when the amount of contamination in each variable is large, the adaptive trimming procedure gives excellent results. The other pairwise methods also perform quite well. However, as we would expect, the classical method and standard robust techniques, MCD and OGK fail quite rapidly. In general, the results for the matrix norms applied to the estimated covariance matrices are very similar to those of the entropy loss for the estimated precision matrix.

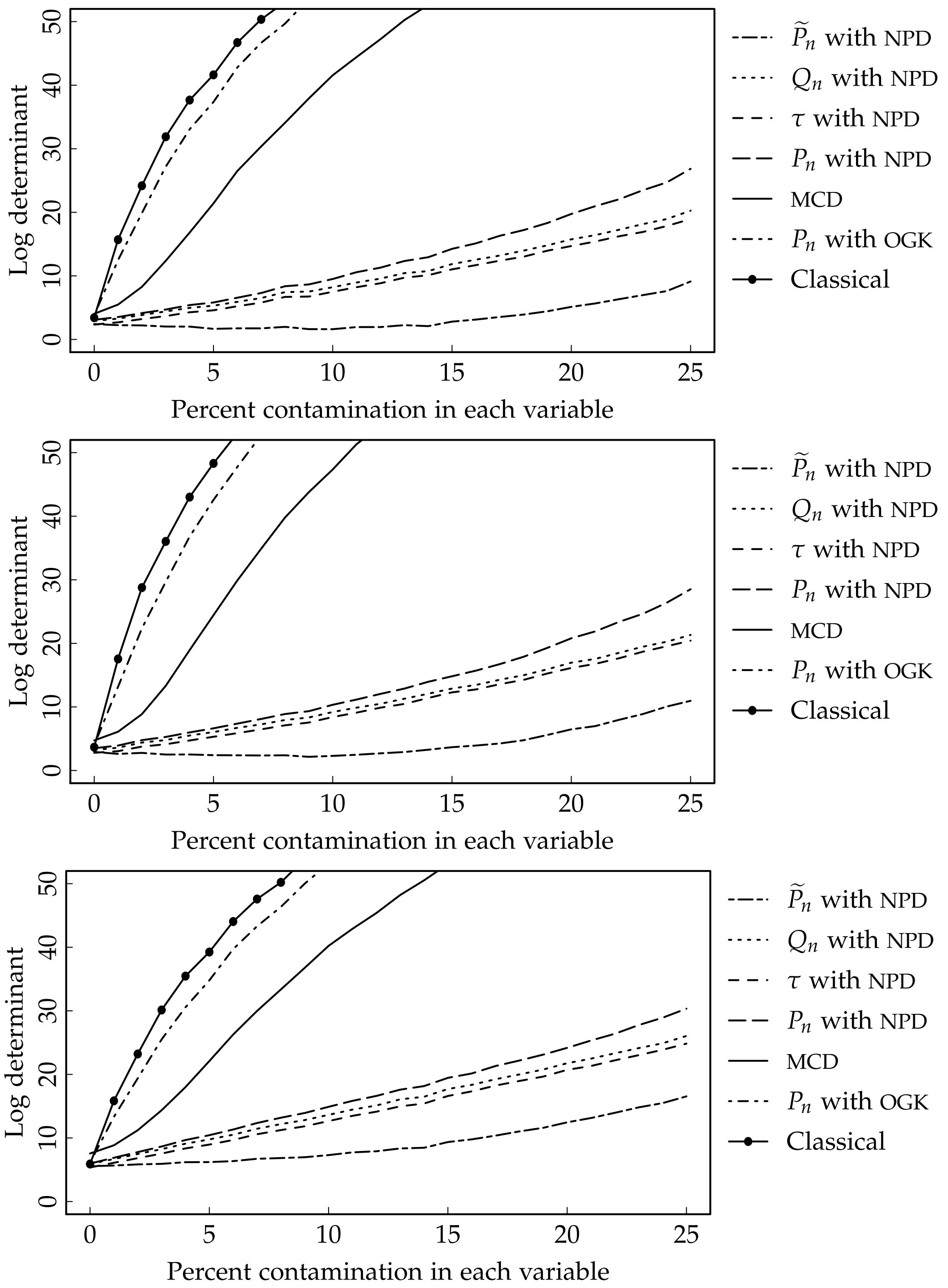


Figure 4.15: Log determinant results for the GLASSO with extreme outliers, $p = 30$ for the banded (top), scattered sparsity (middle) and dense (bottom) precision matrix scenarios.

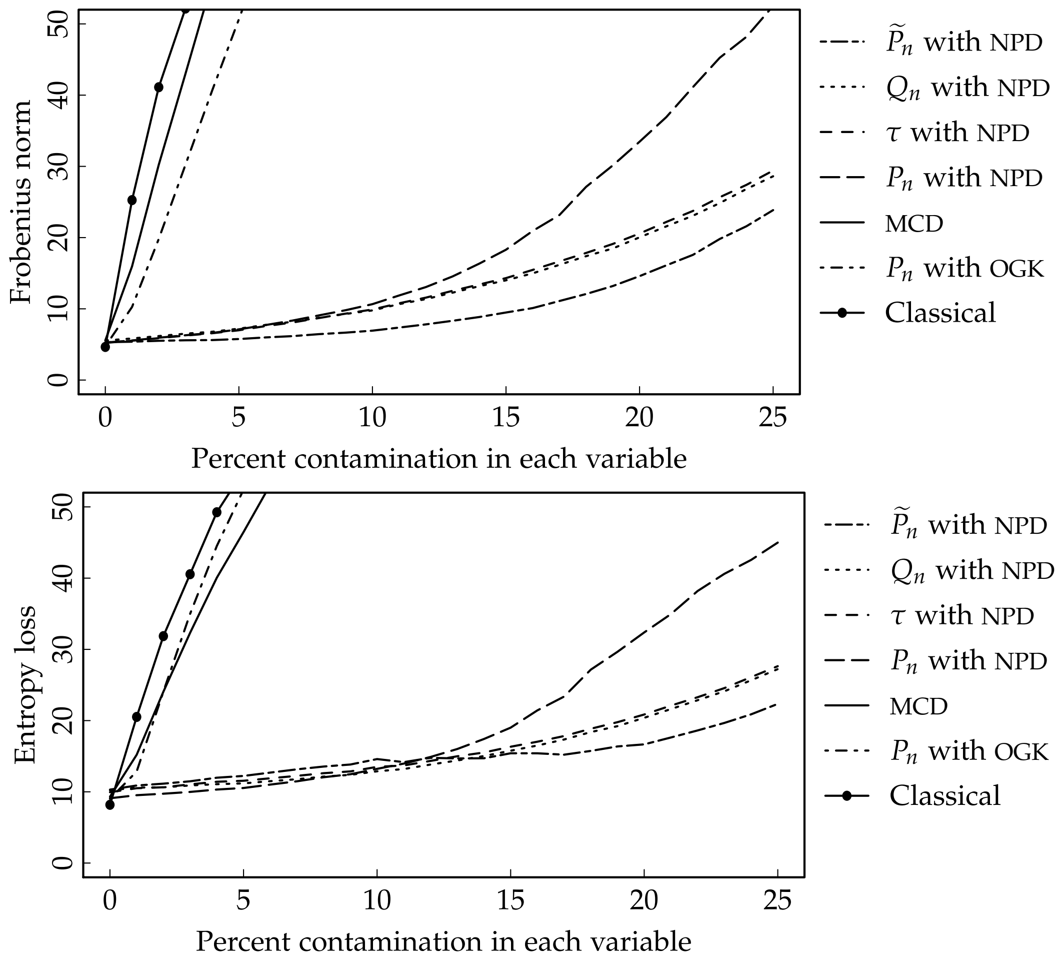


Figure 4.16: Average entropy loss results for the precision matrices resulting from the CLIME procedure (top) and average Frobenius norm results for the resulting covariance matrix estimates (bottom) for $p = 60$ with scattered sparsity and extreme outliers.

4.7 CONCLUSION

A pairwise approach to covariance estimation has a natural resilience to the type of scattered contamination seen in high dimensional scenarios where classical robust procedures, such as the MCD, tend to fail.

We have shown that combining robust pairwise covariance estimation with the NPD method and regularisation techniques such as the CLIME, QUIC or GLASSO yield precision matrices robust to cellwise contamination. The additional advantages of the regularisation techniques, such as the promotion of sparsity also carry through. Furthermore, it did not appear to matter which of the three considered regularisation routines was applied, as all gave broadly similar results in the various scenarios considered. This is comforting given the current pace of research in the field of regularised precision matrix estimation, with new procedures being suggested frequently.

We considered a broad range of scenarios: from dense precision matrices, as is typically found in standard analyses with $n \gg p$; to banded precision matrices that often occur in time series settings and may also be representative of scenarios with block diagonal precision matrices; as well as scattered sparsity, where the linkages between variables are not known beforehand and can show up anywhere within the precision matrix.

After careful consideration of the various performance indices available in the multivariate setting, our primary choice was the entropy loss. When appropriate, we showed that the entropy loss returned similar conclusions to other performance indices, such as the Frobenius norm and log determinant. An interesting further investigation would be to see how well the robust methods perform in terms of Gaussian graphical discovery rates.

The scenarios considered allowed for quite high levels of arbitrary contamination in multivariate data sets. As such, the pairwise techniques based on the standard P_n estimator unsurprisingly did not perform as well as Q_n and τ-scale estimators, however, the adaptively trimmed P_n, $\widetilde{P}_n$ with trimming parameter $d = 3$ typically performed extremely well, due to its ability to detect and trim extreme outliers in bivariate space.

APPENDICES

A

QUANTILE REGRESSION CONFIDENCE INTERVALS

This appendix includes summary figures and tables to provide additional evidence for some of the claims made in Chapter 1.

Figures A.1, A.2 and A.3 show the deteriorating coverage performance for the intercept parameter estimates of a bivariate regression model with Cauchy distributed errors as the sample size increases. The pbs and rank inversion methods seem to be most affected, however when $n = 200$, not all methods are affected. Interestingly, the slope parameter estimates are less affected by changes in the sample size, but for $n = 200$ the pbs and rank inversion methods give the best empirical coverage probabilities for the slope parameter – the trade off being that the empirical coverage probabilities for the intercept are considerably lower than the nominal level.

Tables A.1 to A.4 correspond to the set of models defined by the regression function $y_i = \beta_0 + \beta_1 x_i + u_i$, where $x_i \sim \text{Uniform}(0, 5)$ and $\beta_0 = \beta_1 = 10$. The sample size is $n = 100$, the number of resamples in each of the bootstrap methods is $B = 1000$, the number of Monte Carlo simulations is $N = 1000$. The conditional quantile function is estimated at $\tau = 0.3$. The only difference between the four tables is the specification of the error distribution. Define l_i to be the confidence interval length corresponding to β_i for $i = 0, 1$.

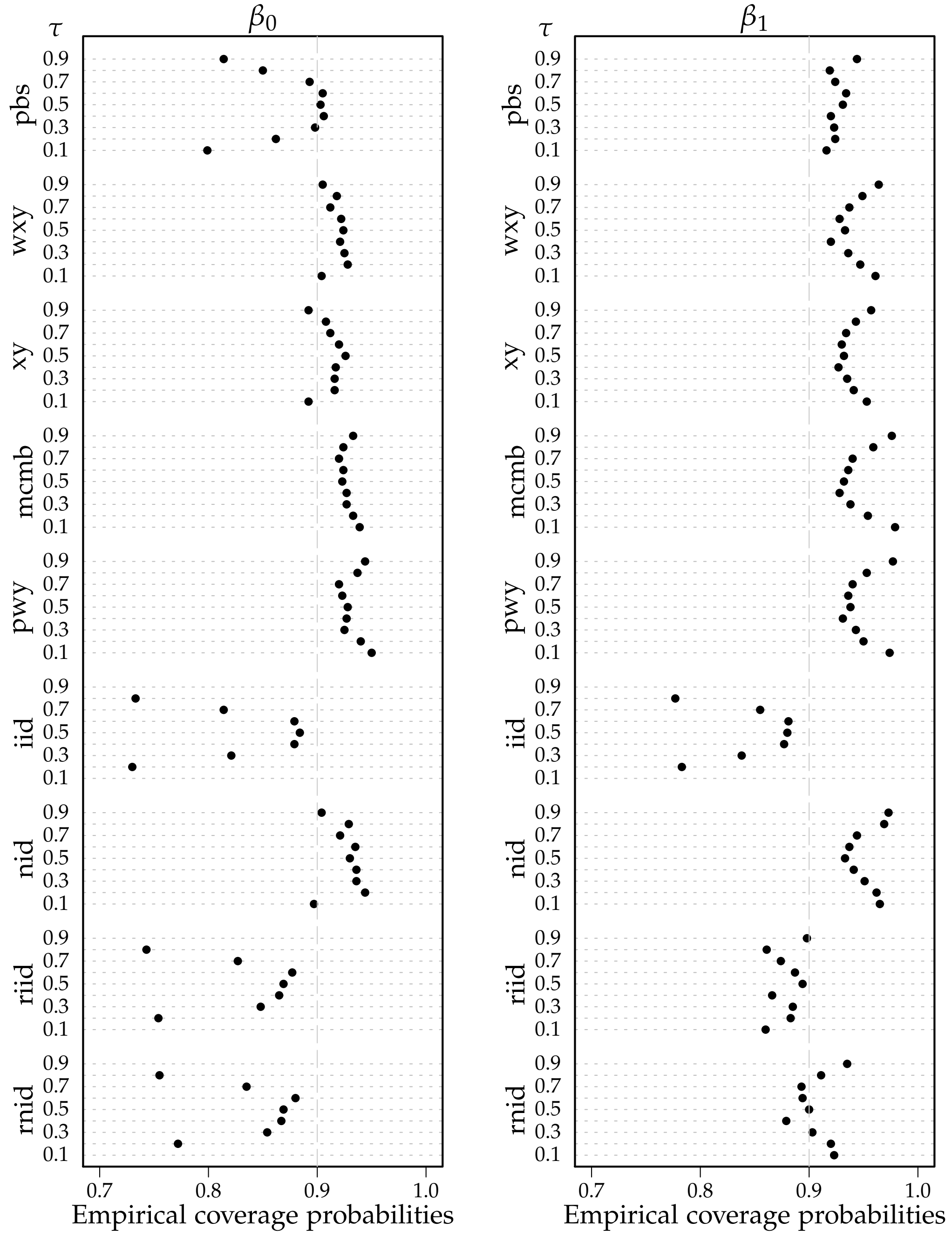


Figure A.1: Empirical coverage probabilities for the model $Y_i = \beta_0 + \beta_1 x_i + u_i$, where $x_i \sim \mathcal{U}(0,5)$, $u_i \sim t_1$ and $\beta_0 = \beta_1 = 10$. The sample size is $n = 100$, the number of resamples in each of the bootstrap methods is $B = 1000$, the number of Monte Carlo simulations is $N = 1000$. The nominal coverage is 0.9.

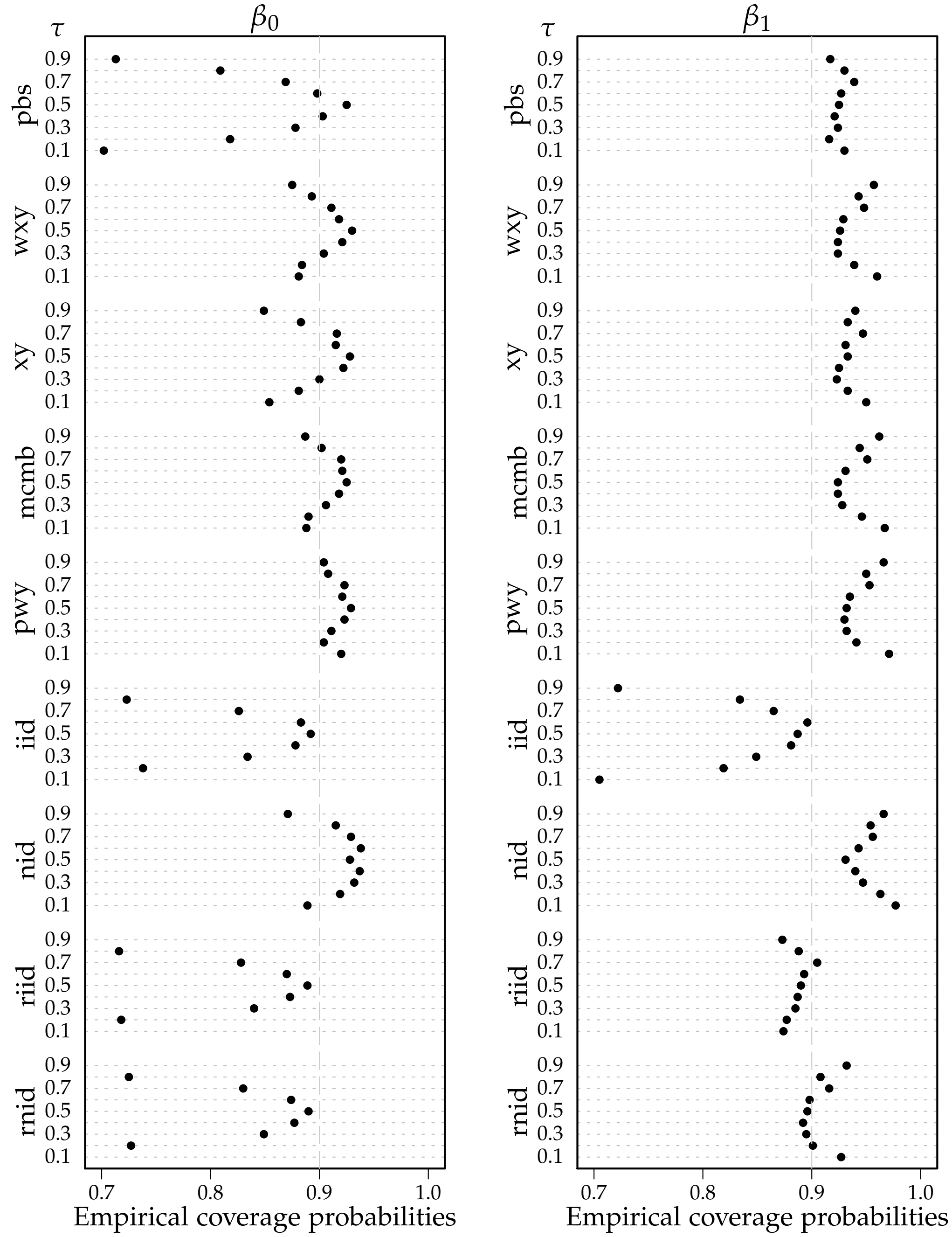


Figure A.2: Empirical coverage probabilities for the model $Y_i = \beta_0 + \beta_1 x_i + u_i$, where $x_i \sim \mathcal{U}(0, 5)$, $u_i \sim t_1$ and $\beta_0 = \beta_1 = 10$. The sample size is $n = 150$, the number of resamples in each of the bootstrap methods is $B = 1000$, the number of Monte Carlo simulations is $N = 1000$. The nominal coverage is 0.9.

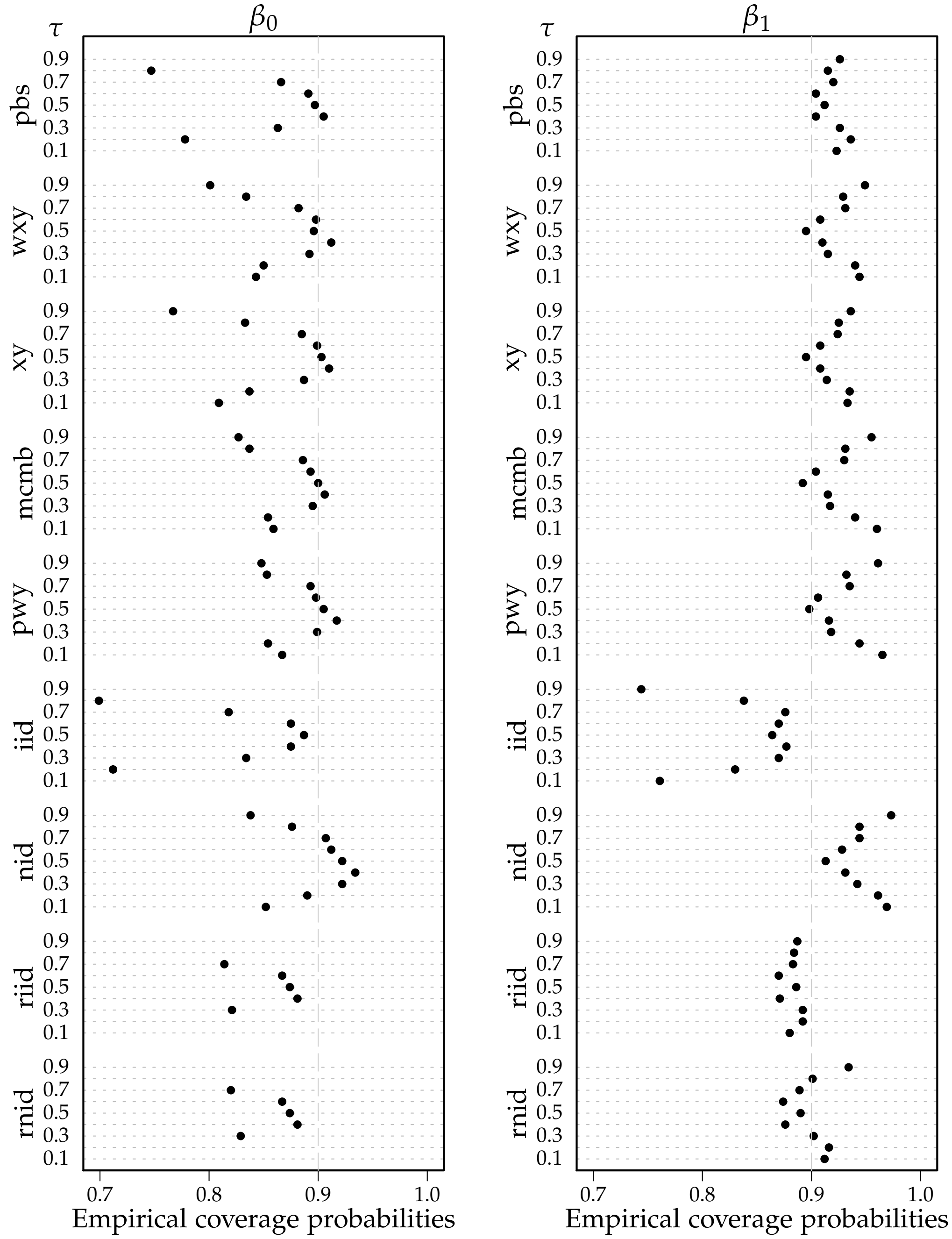


Figure A.3: Empirical coverage probabilities for the model $Y_i = \beta_0 + \beta_1 x_i + u_i$, where $x_i \sim \mathcal{U}(0,5)$, $u_i \sim t_1$ and $\beta_0 = \beta_1 = 10$. The sample size is $n = 200$, the number of resamples in each of the bootstrap methods is $B = 1000$, the number of Monte Carlo simulations is $N = 1000$. The nominal coverage is 0.9.

	Average		SD		Coverage	
	l_0	l_1	l_0	l_1	β_0	β_1
iid	0.713	0.719	0.221	0.230	0.744	0.868
nid	1.086	0.970	0.368	0.339	0.930	0.939
riid	0.779	0.768	0.259	0.270	0.694	0.892
rnid	0.802	0.834	0.266	0.366	0.710	0.905
xy	0.899	0.907	0.277	0.289	0.848	0.944
pwy	0.945	0.938	0.293	0.290	0.870	0.949
mcmb	0.871	0.873	0.259	0.268	0.831	0.934
wxy	0.919	0.913	0.287	0.296	0.865	0.943
pbs	0.858	0.870	0.265	0.282	0.752	0.933

Table A.1: Summary output for Cauchy distributed errors with $n = 100$ and $\tau = 0.3$.

	Average		SD		Coverage	
	l_0	l_1	l_0	l_1	β_0	β_1
iid	0.498	0.504	0.133	0.139	0.869	0.859
nid	0.592	0.563	0.114	0.124	0.926	0.908
riid	0.497	0.516	0.126	0.169	0.866	0.884
rnid	0.504	0.532	0.126	0.177	0.871	0.893
xy	0.554	0.581	0.121	0.155	0.903	0.919
pwy	0.569	0.598	0.122	0.160	0.901	0.934
mcmb	0.546	0.574	0.117	0.152	0.899	0.918
wxy	0.557	0.575	0.120	0.154	0.904	0.916
pbs	0.537	0.564	0.118	0.153	0.915	0.929

Table A.2: Summary output for t_3 distributed errors when $n = 100$ and $\tau = 0.3$.

	Average		SD		Coverage	
	l_0	l_1	l_0	l_1	β_0	β_1
iid	0.476	0.477	0.124	0.128	0.881	0.869
nid	0.538	0.511	0.091	0.101	0.923	0.902
riid	0.462	0.473	0.116	0.150	0.877	0.892
rnid	0.468	0.484	0.116	0.153	0.885	0.901
xy	0.515	0.530	0.106	0.136	0.918	0.916
pwy	0.528	0.546	0.109	0.141	0.923	0.924
mcmb	0.510	0.526	0.106	0.135	0.914	0.907
wxy	0.517	0.524	0.107	0.136	0.920	0.913
pbs	0.501	0.513	0.107	0.134	0.924	0.938

Table A.3: Summary output for t_5 distributed errors when $n = 100$ and $\tau = 0.3$.

	Average		SD		Coverage	
	l_0	l_1	l_0	l_1	β_0	β_1
iid	0.427	0.432	0.107	0.112	0.851	0.873
nid	0.470	0.451	0.071	0.084	0.913	0.886
riid	0.414	0.425	0.100	0.124	0.867	0.875
rnid	0.419	0.437	0.100	0.153	0.869	0.883
xy	0.460	0.474	0.093	0.110	0.901	0.900
pwy	0.469	0.488	0.092	0.114	0.911	0.908
mcmb	0.453	0.469	0.090	0.109	0.896	0.899
wxy	0.460	0.467	0.092	0.110	0.904	0.896
pbs	0.447	0.459	0.091	0.109	0.909	0.910

Table A.4: Summary output for $\mathcal{N}(0,1)$ errors when $n = 100$ and $\tau = 0.3$.

B SCALE ESTIMATION

B.1 FINITE SAMPLE CORRECTION FACTORS

To ensure that P_n is consistent for the standard deviation at the Gaussian distribution, correction factors are introduced. As such the pairwise mean scale estimator will be redefined as:

$$P_n(\tau, \boldsymbol{x}) = c_{n,\tau} c_\tau \left[G_n^{-1}((1+\tau)/2) - G_n^{-1}((1-\tau)/2) \right], \qquad \text{(B.1)}$$

where c_τ is the large sample correction factor and $c_{n,\tau}$ is a small sample correction factor. We shall restrict attention to $\tau = 0.5$, the standard P_n estimator and assume that the observations come from a Gaussian distribution. For $n \geq 5$ the small sample correction factors have been found through simulation with one million replications and are provided in Table B.1. We have also explicitly found the correction factors for P_n at the Gaussian distribution for samples of size $n = 3$ and 4.

n	$c_{n,0.5}$	n	$c_{n,0.5}$	n	$c_{n,0.5}$	n	$c_{n,0.5}$
3	1.128	13	1.057	23	1.036	33	1.021
4	1.303	14	1.040	24	1.030	34	1.023
5	1.109	15	1.061	25	1.029	35	1.018
6	1.064	16	1.047	26	1.032	36	1.020
7	1.166	17	1.043	27	1.023	37	1.019
8	1.103	18	1.048	28	1.025	38	1.017
9	1.087	19	1.031	29	1.024	39	1.020
10	1.105	20	1.037	30	1.021	40	1.018
11	1.047	21	1.035	31	1.026	41	1.017
12	1.063	22	1.028	32	1.022	42	1.018

Table B.1: Finite sample correction factors applied to P_n at the Gaussian distribution to ensure approximate unbiasedness.

Assume that observations, $X_1, \ldots, X_n$, are drawn from $F = \Phi$, the standard Gaussian population. For $n = 3$ there are $\binom{3}{2} = 3$ pairwise means in the empirical distribution function, therefore

$$\begin{aligned}\mathbb{E}\left[G_3^{-1}(\tfrac{3}{4}) - G_3^{-1}(\tfrac{1}{4})\right] &= \mathbb{E}\left[\left(\frac{X_i + X_j}{2}\right)_{(3)} - \left(\frac{X_i + X_j}{2}\right)_{(1)}\right] \\ &= \mathbb{E}\left[\frac{X_{(2)} + X_{(3)} - X_{(1)} - X_{(2)}}{2}\right] \\ &= \frac{1}{2}\mathbb{E}(X_{(3)} - X_{(1)}) \\ &= \mathbb{E}(X_{(3)}) \quad \text{(by symmetry)} \\ &= 0.846\end{aligned}$$

In samples of size $n = 3$ from a standard Gaussian distribution the expected value of the maximum is $\mathbb{E}(X_{(3)}) \approx 0.846$. After applying the large sample correction factor,

$$\mathbb{E}\left(c_{0.5}\left[G_3^{-1}(\tfrac{3}{4}) - G_3^{-1}(\tfrac{1}{4})\right]\right) \approx 0.887.$$

Therefore $c_{3,0.5} \approx 1/0.887 \approx 1.127$.

When $n = 4$ we have $\binom{4}{2} = 6$ pairwise means. The IQR of the pairwise means is therefore, $\big((X_i + X_j)/2\big)_{(5)} - \big((X_i + X_j)/2\big)_{(2)}$. Assuming no ties, $\big((X_i + X_j)/2\big)_{(5)} = (X_{(4)} + X_{(2)})/2$. Similarly we have $\big((X_i + X_j)/2\big)_{(2)} = (X_{(1)} + X_{(3)})/2$. So the expected value of the IQR of the pairwise means is

$$\begin{aligned}\mathbb{E}\left[\left(\frac{X_i + X_j}{2}\right)_{(5)} - \left(\frac{X_i + X_j}{2}\right)_{(2)}\right] &= \mathbb{E}\left[\frac{X_{(4)} + X_{(2)} - X_{(1)} - X_{(3)}}{2}\right] \\ &= \mathbb{E}(X_{(4)}) + \mathbb{E}(X_{(2)}) \quad \text{(by symmetry)} \\ &= 0.732.\end{aligned}$$

After applying the large sample correction factor,

$$\mathbb{E}\left(c_{0.5}\left[G_4^{-1}(\tfrac{3}{4}) - G_4^{-1}(\tfrac{1}{4})\right]\right) \approx 0.767,$$

so $c_{4,0.5} \approx 1/0.767 \approx 1.303$.

B.2 IMPLOSION BREAKDOWN

This section shows that $P_n = 0$ implies $Q_n = 0$. The design is constructed as follows. Let $x_1, x_2, \ldots,$ be the distinct ordered outcomes from a discrete random variable, each occurring with probabilities $p_1, p_2, \ldots,$ respectively. Table B.2 outlines the probabilities associated with pairs of observations obtained from this distribution. The set of outcomes is replicated once across the columns, A, and once down the rows, B. To construct each pair, one observation is obtained from A and the other from B.

In large samples, for Q_n to return a value of zero, $\sum_i p_i^2 \geq \frac{1}{4}$, where the sum is over all possible outcomes. For P_n to fail, we require at least half of the pairs to have a common sum, S, say. Let $D = \{(i_A, i_B); i_A \leq i_B : x_{i_A} + x_{i_B} = S\}$. For P_n to fail we require $2\sum_{(i,j)\in D} p_i p_j \left(1 - \frac{1}{2}\mathbb{I}\{i = j\}\right) \geq 0.5$. Note that because we have distinct observations, for each x_{i_A} selected from the column set, there is at most one corresponding x_{i_B} selected from the row set. That is, for each row or column, there is at most one pair entry that can sum to S.

Now, consider the probabilities associated with any x_i and x_j, $p_i \in (0, 1)$ and $p_j \in (0, 1 - p_i)$ respectively. We know $p_i p_j \leq \max\{p_i^2, p_j^2\}$, hence, if we take the sum over all the pairs of outcomes that sum to a constant, we have

$$\sum_{(i,j)\in D} p_i p_j \leq \sum_{(i,j)\in D} \max\{p_i^2, p_j^2\} \leq \sum_i p_i^2. \tag{B.2}$$

The last inequality follows by noting that $\sum_{(i,j)\in D} \max\{p_i^2, p_j^2\}$ is a sum of distinct elements from $p_1^2, p_2^2 \ldots$. No subscript is repeated as for any given B sample value there is at most one A sample value giving the required sum for $x_{i_A} + x_{i_B}$. Hence if P_n fails, $\sum_i p_i^2 \geq \frac{1}{4}$ and so Q_n will also fail.

	A	x_1	x_2	x_3	x_4	$\cdots$
B		p_1	p_2	p_3	p_4	$\cdots$
x_1	p_1	p_1^2				
x_2	p_2	p_2p_1	p_2^2			
x_3	p_3	p_3p_1	p_3p_2	p_3^2		
x_4	p_4	p_4p_1	p_4p_2	p_4p_3	p_4^2	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\ddots$

Table B.2: Lower triangular elements for the sampling distribution of pairs. The (i, j)th element in the table is $P(X_A = x_i, X_B = x_j) = p_i p_j$.

B.3 PAIRWISE MEAN SCALE ESTIMATOR CODE FOR R

The R code below outlines a basic implentation for the standard $P_n = P_n(0.5)$ estimator. Speed improvements can be made by using the `pair.sum()` function from the `ICSNP` package which performs the required computation in C. Alternatively, it is technically feasible to compute P_n in $O(n \log n)$ time following the algorithm set out in Johnson and Mizoguchi (1978).

```
Pn = function(y){
 n = length(y)
 if(n <= 2){
      warning("Your sample size is too small")
      return()
 }
 y.pairs = outer(y,y,"+")
 y.pairs = y.pairs[lower.tri(y.pairs)]/2
 # require(ICSNP) # loads the pair.sum C function
 # y.pairs = pair.sum(matrix(y))/2
 const = 1/0.9539 # asymptotic correction factor
 scale.est=const*as.numeric(diff(quantile(y.pairs,c(1/4,3/4),type=1)))
 # Correction factors obtained through simulation over 1 million
     replications
 correction.factors =
 c(1.128,1.303,1.109,1.064,1.166,1.103,1.087,1.105,1.047,1.063,1.057,
   1.040,1.061,1.047,1.043,1.048,1.031,1.037,1.035,1.028,1.036,1.030,
   1.029,1.032,1.023,1.025,1.024,1.021,1.026,1.022,1.021,1.023,1.018,
   1.020,1.019,1.017,1.020,1.018,1.017,1.018,1.015,1.016,1.016,1.014,
   1.016,1.015,1.014,1.015)
 if(n <= 40){
  scale.est = scale.est*correction.factors[n-2]
  # n-2 as the first element of the correction.factors vector is for n=3
  } else if(n > 40) scale.est = scale.est*n/(n-0.7)
      return(scale.est)
}
#### Example ####
x = rnorm(100)
Pn(x)
```

B.4 EM ALGORITHM IMPLEMENTATION

The following R code implements the EM algorithm for some Gaussian compound scale models as proposed by Randal and Thomson (2004) and Randal (2008).

```r
## EM Algorithm to find scale estimates ##
Estep = function(x,mu,sigma2,type,p,k,df){
      E = NULL
      if(type=="slash"){
            diff.x = abs(x-mu)
            U2 = (x-mu)^2/sigma2
            E = 2*(U2^(-1)) - (exp(U2/2)-1)^(-1)
            check = diff.x<0.0001
            E[check] = 0.5
            return(E)
      } else if(type=="wild"){
            A = (exp(99/200*(x-mu)^2/sigma2))
            E = 1 - 99/100*A/(sum(A))
            if(!is.finite(sum(E))){
            n = length(x)
            E = c(rep(1,n-1),1-99/100)
            }
            return(E)
      } else if(type=="cn"){
            A = exp(1/2*(1-k^-2)*(x-mu)^2/sigma2)
            E = 1 - (1-k^-2)*A/(k*(1/p - 1)+A)
            E[E=="NaN"] = 1
            return(E)
      } else if(type=="t"){
            E = (df+1)/df/(1+(x-mu)^2/(df*sigma2))
            return(E)
      }
}
EM.alg = function(x,mu=median(x),sigma2=mad(x),type,p,k,df){
      iter = 1
      if(type == "normal"){
            mu = mean(x)
            sigma2 = var(x)
            sigma = sd(x)
            mu.store = mu
            sigma2.store = sigma2
            } else if(type == "exponential"){
                  mu = mean(x)
                  sigma2 = mu^2
                  sigma = mu
                  mu.store = mu
                  sigma2.store = sigma2
            } else if(type == "chisq"){
                  mu = mean(x)
                  s = log(mu) - sum(log(x))/length(x)
                  k = (3-s+sqrt((s-3)^2+24*s))/(12*s)
                  sigma2 = k*2^2
                  sigma = sqrt(sigma2)
                  mu.store = mu
                  sigma2.store = sigma2
            } else {
      mu.store = mu
      sigma2.store = sigma2
      diff = 1
      n = length(x)
      i=1:n
      while(diff>0.0001){
                  E = Estep(x,mu,sigma2,type,p,k,df)
                  new.mu = sum(E*x)/sum(E)
                  new.sigma2 = sum(E*(x-mu)^2)/n
                  diff = max(abs(mu-new.mu),abs(sigma2-new.sigma2))
                  mu = new.mu
                  sigma2 = new.sigma2
                  mu.store = c(mu.store,mu)
                  sigma2.store = c(sigma2.store,sigma2)
                  iter = iter+1
            if(iter==500){
                  break
                  }
            }
      sigma = sqrt(new.sigma2)
      }
      return(list(mu=mu, sigma=sigma, sigma2=sigma2, mu.store=mu.store,
                              sigma2.store=sigma2.store, iter = iter))
}
#### Examples ####
## Slash ##
n=100
EM.alg(rnorm(n)/runif(n),type="slash")
## One Wild ##
y = c(rnorm(n-1),rnorm(1,0,10))
EM.alg(y,type="wild")
## Gaussian ##
EM.alg(rnorm(n),type="normal")
## Contaminated Normal (0.1,10) ##
indicator = rbinom(n,1,0.1)
cndata = indicator*rnorm(n,0,10) + (1-indicator)*rnorm(n,0,1)
EM.alg(cndata,type = "cn",p=0.1,k=10)
## t distribution ##
EM.alg(rt(n,df=5),type="t",df = df)
```

C

COVARIANCE AND AUTOCOVARIANCE ESTIMATION

C.1 TECHNICAL RESULTS

C.1.1 *Hadamard differentiability*

In Section 3.4.1 we required the Hadamard differentiability of a number of maps. Hadamard differentiability is defined in van der Vaart (1998, p. 296) as follows.

Let $\mathbb{D}$ and $\mathbb{E}$ be normed spaces. A map $T : \mathbb{D}_T \subset \mathbb{D} \mapsto \mathbb{E}$ is Hadamard differentiable at $\theta \in \mathbb{D}_T$ if there exists a continuous linear map $T'_\theta : \mathbb{D} \mapsto \mathbb{E}$:

$$\left\| \frac{T(\theta + tg_t) - T(\theta)}{t} - T'_\theta(g) \right\|_{\mathbb{E}} \to 0,$$

as $t \downarrow 0$ for all converging sequences $g_t \to g$ such that $\theta + tg_t \in \mathbb{D}_T$ for all small $t > 0$.

C.1.2 *Influence functions*

Ma and Genton (2000) prove the following proposition, used in Section 3.4.2. The proof follows from the definition of the influence function as a directional derivative.

Proposition C.1. *Suppose $\Phi_\sigma = \mathcal{N}(0, \sigma^2)$ and $\Phi = \mathcal{N}(0,1)$. Let S be a statistical functional of scale, and hence S^2 is a statistical functional of variance. More generally, consider a functional of S denoted by $h(S)$. Then,*

$$\begin{aligned} \mathrm{IF}(x; S, \Phi_\sigma) &= \sigma\, \mathrm{IF}\left(x/\sigma; S, \Phi\right) \\ \mathrm{IF}(x; S^2, \Phi_\sigma) &= \sigma^2\, \mathrm{IF}\left(x/\sigma; S^2, \Phi\right) \\ \mathrm{IF}(x; h(S), \Phi_\sigma) &= h'(S(F))\, \mathrm{IF}\left(x/\sigma; S, \Phi\right). \end{aligned}$$

C.2 HERMITE RANKS

Most of the limit theorems considered in Chapter 3 require knowledge of the Hermite rank of a function. Recall from Section 2.2.1.2 that the Hermite rank of a function J is defined as $m = \inf\{k \geq 1 : \alpha_k \neq 0\}$ where $\alpha_k = \mathbb{E}[J(Z)H_k(Z)]$, $Z \sim \mathcal{N}(0,1)$, $\mathbb{E}[J(Z)] = 0$ and $H_k(\cdot)$ is the kth Hermite polynomial:

$$J(Z) = \sum_{k \geq m} \frac{\alpha_k}{k!} H_k(Z).$$

C.2.1 *Influence function of P_n*

This section sets out to prove that the Hermite rank of the influence function of P_n when the observations come from a standard Gaussian distribution is $m = 2$. Recall from equation (3.5),

$$\mathrm{IF}(x; P, \Phi) = c \left[\frac{\frac{1}{2} - \Phi(2G_\Phi^{-1}(\frac{3}{4}) - x) + \Phi(2G_\Phi^{-1}(\frac{1}{4}) - x)}{\phi(\sqrt{2}G_\Phi^{-1}(\frac{3}{4}))/\sqrt{2}} \right].$$

Let $Z \sim \mathcal{N}(0,1)$ and $Z_1, Z_2, \ldots, Z_n$ be an independent sample of size n from a standard Gaussian distribution. To establish that the Hermite rank is $m = 2$ we show that,

A. $\mathbb{E}(\mathrm{IF}(Z; P, \Phi)) = 0$;

B. $\mathbb{E}(Z\ \mathrm{IF}(Z; P, \Phi)) = 0$; and

C. $\mathbb{E}(Z^2\ \mathrm{IF}(Z; P, \Phi)) \neq 0$.

Lemma C.1. $\mathbb{E}(\mathrm{IF}(Z; P, \Phi)) = 0$.

Proof. It is sufficient to show that

$$\mathbb{E}\left(\Phi(2G_\Phi^{-1}(\tfrac{3}{4}) - Z) - \Phi(2G_\Phi^{-1}(\tfrac{1}{4}) - Z)\right) = \tfrac{1}{2}.$$

The CDF of pairwise means is,

$$G_\Phi(t) = P((Z_1 + Z_2)/2 \leq t) = \int_{\mathbb{R}} \Phi(2t - u)\phi(u)du. \tag{C.1}$$

Hence,

$$\begin{aligned}\mathbb{E}[\Phi(2G_\Phi^{-1}(\tfrac{3}{4}) - Z)] &= \int_{\mathbb{R}} \Phi(2G_\Phi^{-1}(\tfrac{3}{4}) - x)\phi(x)dx \\ &= G_\Phi(G_\Phi^{-1}(\tfrac{3}{4})) \\ &= \tfrac{3}{4}.\end{aligned}$$

Similarly, $\mathbb{E}[\Phi(2G_\Phi^{-1}(\frac{1}{4}) - Z)] = \frac{1}{4}$ and hence, $\mathbb{E}(\mathrm{IF}(Z; P, \Phi)) = 0$. □

Lemma C.2. $\mathbb{E}(Z \text{ IF}(Z; P, \Phi)) = 0$.

Proof. Noting that $\mathbb{E}(Z) = 0$, it is sufficient to show that

$$\mathbb{E}\left[Z\Phi(2G_\Phi^{-1}(\tfrac{3}{4}) - Z) - Z\Phi(2G_\Phi^{-1}(\tfrac{1}{4}) - Z)\right] = 0.$$

Consider,

$$\mathbb{E}[Z\Phi(2t - Z)] = \int x\phi(x)\Phi(2t - x)\,\mathrm{d}x$$

using the relation $x\phi(x) = -\phi'(x)$ and integrating by parts,

$$\begin{aligned}
&= [-\phi(x)\Phi(2t - x)]_{-\infty}^{\infty} - \int \phi(2t - x)\phi(x)\,\mathrm{d}x \\
&= -\frac{1}{2\pi}\int \exp\left\{-\frac{(2t - x)^2 + x^2}{2}\right\}\mathrm{d}x \\
&= -\frac{\phi(t\sqrt{2})}{\sqrt{2}}.
\end{aligned}$$

In particular, we have

$$\begin{aligned}
\mathbb{E}&\left[Z\Phi(2G_\Phi^{-1}(\tfrac{3}{4}) - Z) - Z\Phi(2G_\Phi^{-1}(\tfrac{1}{4}) - Z)\right] \\
&= -\frac{\phi(G_\Phi^{-1}(\tfrac{3}{4})\sqrt{2})}{\sqrt{2}} + \frac{\phi(G_\Phi^{-1}(\tfrac{1}{4})\sqrt{2})}{\sqrt{2}} = 0,
\end{aligned}$$

by symmetry, $G_\Phi^{-1}(\tfrac{3}{4}) = -G_\Phi^{-1}(\tfrac{1}{4})$ and $\phi(a) = \phi(-a)$. □

Lemma C.3. $\mathbb{E}(Z^2 \text{ IF}(Z; P, \Phi)) \neq 0$.

Proof. First note that,

$$\mathbb{E}[(Z^2 - 1)\Phi(2t - Z)] = \int (x^2 - 1)\phi(x)\Phi(2t - x)\,\mathrm{d}x$$

consider $\frac{\mathrm{d}}{\mathrm{d}x}x\phi(x) = -(x^2 - 1)\phi(x)$ and using integration by parts,

$$\begin{aligned}
&= [-x\phi(x)\Phi(2t - x)]_{-\infty}^{\infty} - \int x\phi(x)\phi(2t - x)\,\mathrm{d}x \\
&= -\frac{1}{2\pi}\int x \exp\left\{-\frac{1}{2}\left[2(t - x)^2 + 2t^2\right]\right\}\mathrm{d}x \\
&= -\frac{t\phi(t\sqrt{2})}{\sqrt{2}}.
\end{aligned}$$

Furthermore, using equation (C.1) we obtain,

$$\begin{aligned}
\mathbb{E}[Z^2\Phi(2t - Z)] &= \mathbb{E}[(Z^2 - 1)\Phi(2t - Z)] + \mathbb{E}[\Phi(2t - Z)] \\
&= -\frac{t\phi(t\sqrt{2})}{\sqrt{2}} + G_\Phi(t).
\end{aligned}$$

Hence,

$$\mathbb{E}[Z^2\Phi(2G_\Phi^{-1}(\tfrac{3}{4}) - Z)] = -\frac{G_\Phi^{-1}(\frac{3}{4})\phi(G_\Phi^{-1}(\frac{3}{4})/\sqrt{2})}{\sqrt{2}} + \tfrac{3}{4},$$

and

$$\mathbb{E}[Z^2\Phi(2G_\Phi^{-1}(\tfrac{1}{4}) - Z)] = -\frac{G_\Phi^{-1}(\frac{1}{4})\phi(G_\Phi^{-1}(\frac{1}{4})/\sqrt{2})}{\sqrt{2}} + \tfrac{1}{4}.$$

Also, by symmetry, $G_\Phi^{-1}(\frac{3}{4}) = -G_\Phi^{-1}(\frac{1}{4})$ and $\phi(a) = \phi(-a)$. Hence, as $\mathbb{E}(Z^2) = 1$,

$$\begin{aligned}\mathbb{E}[Z^2 \text{ IF}(Z; P, \Phi)] &= \mathbb{E}\left[cZ^2\left(\frac{\frac{1}{2} - \Phi(2G_\Phi^{-1}(\frac{3}{4}) - Z) + \Phi(2G_\Phi^{-1}(\frac{1}{4}) - Z)}{\phi(\sqrt{2}G_\Phi^{-1}(\frac{3}{4}))/\sqrt{2}}\right)\right] \\ &= \frac{c\sqrt{2}G_\Phi^{-1}(\frac{3}{4})\phi(G_\Phi^{-1}(\frac{3}{4})/\sqrt{2})}{\phi(\sqrt{2}G_\Phi^{-1}(\frac{3}{4}))/\sqrt{2}} > 0.\end{aligned}$$

□

C.2.2 *Empirical distribution function*

Let $Z \sim \mathcal{N}(0,1)$. The Hermite expansion for $h(Z;t) = \mathbb{I}\{Z \leq t\}$ for $t \in \mathbb{R}$ is,

$$\mathbb{I}\{Z \leq t\} = \sum_{k=0}^{\infty} \frac{\alpha_k(t)}{k!} H_k(Z).$$

The first three Hermite coefficients are given by,

$$\alpha_0(t) = \mathbb{E}[h(Z;t)H_0(Z)] = \mathbb{E}[\mathbb{I}\{Z \leq t\}] = \Phi(t),$$

$$\begin{aligned}\alpha_1(t) = \mathbb{E}[h(Z;t)H_1(Z)] &= \int \mathbb{I}\{z \leq t\} z\phi(z) dz \\ &= \int_{-\infty}^{t} z \frac{e^{-z^2/2}}{\sqrt{2\pi}} dz \\ &= \frac{1}{\sqrt{2\pi}} \int_{\infty}^{t^2/2} e^{-u} du \\ &= -\phi(t),\end{aligned}$$

and

$$\begin{aligned}\alpha_2(t) = \mathbb{E}[h(Z;t)H_2(Z)] &= \mathbb{E}[\mathbb{I}\{Z \leq t\}(Z^2 - 1)] \\ &= \int_{-\infty}^{t} (z^2 - 1)\phi(z) dz\end{aligned}$$

note that $\frac{\mathrm{d}}{\mathrm{d}z} z\phi(x) = -(z^2 - 1)\phi(z)$ and integrate both sides,

$$= -t\phi(t).$$

C.2.3 *Pairwise mean distribution function*

Let $g(x,y) = (x+y)/2$ be the pairwise mean kernel and let $h(x,y;t) = \mathbb{I}\{g(x,y) \leq t\}$ for $t \in \mathbb{R}$. As shown in Section 2.2, if X and Y are independent standard Gaussian random variables, $h(\cdot,\cdot;t)$ can be expanded in a bivariate Hermite polynomial basis as follows,

$$h(X,Y;t) = \mathbb{I}\{g(X,Y) \leq t\} = \sum_{p,q \geq 0} \frac{\alpha_{p,q}(t)}{p!q!} H_p(X) H_q(Y),$$

with Hermite coefficients,

$$\alpha_{p,q}(t) = \mathbb{E}\left[h(X,Y;t) H_p(X) H_q(Y)\right].$$

First note that $\alpha_{0,0}(t) = G_\Phi(t)$.

Result C.1. *Consider, X, Y independent standard Gaussian random variables,*

$$G_\Phi(t) = P\left(X + Y \leq 2t\right) = \Phi(t\sqrt{2}),$$

with derivative,

$$G'_\Phi(t) = \sqrt{2}\phi(t\sqrt{2}).$$

Result C.2. *When $g(x,y) = (x+y)/2$ and X, Y are independent standard Gaussian variables, the $\alpha_{1,0}(t)$ and $\alpha_{0,1}(t)$ Hermite coefficients are given by,*

$$\alpha_{1,0}(t) = \alpha_{0,1}(t) = \mathbb{E}\left[X\,\mathbb{I}\{X+Y \leq 2t\}\right] = -\frac{\phi(t\sqrt{2})}{\sqrt{2}}.$$

Furthermore, noting Result C.1, for all $t \in \mathbb{R}$,

$$\frac{\alpha_{1,0}(t)}{G'_\Phi(t)} = -\frac{1}{2}$$

Proof. Consider,

$$\begin{aligned}\alpha_{1,0}(t) = \alpha_{0,1}(t) &= \mathbb{E}\left[X\,\mathbb{I}\{X+Y \leq 2t\}\right] \\ &= \int\int x\mathbb{I}\{x+y \leq 2t\}\phi(x)\phi(y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int x\phi(x)\Phi(2t-x)\,\mathrm{d}x\end{aligned}$$

using the relation $x\phi(x) = -\phi'(x)$ and integrating by parts we have

$$\begin{aligned}
&= [-\phi(x)\Phi(2t-x)]_{-\infty}^{\infty} - \int \phi(2t-x)\phi(x)\,\mathrm{d}x \\
&= -\frac{1}{2\pi}\int \exp\left\{-\frac{(2t-x)^2+x^2}{2}\right\}\mathrm{d}x \\
&= -\frac{\phi(t\sqrt{2})}{\sqrt{2}}. \qquad \square
\end{aligned}$$

Result C.3. *When $g(x,y) = (x+y)/2$ and X, Y are independent standard Gaussian variables, the $\alpha_{1,1}(t)$ Hermite coefficient is given by,*

$$\alpha_{1,1}(t) = \mathbb{E}(XY\mathbb{I}\{X+Y \le 2t\}) = -\frac{t\phi(t\sqrt{2})}{\sqrt{2}}.$$

Furthermore, noting Result C.1, for all $t \in \mathbb{R}$,

$$\frac{\alpha_{1,1}(t)}{G'_{\Phi}(t)} = -\frac{t}{2}.$$

Proof. Consider,

$$\begin{aligned}
\alpha_{1,1}(t) &= \mathbb{E}(XY\mathbb{I}\{X+Y \le 2t\}) \\
&= \int\int xy\mathbb{I}\{x+y \le 2t\}\phi(x)\phi(y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int x\phi(x)\left(\int_{-\infty}^{2t-x} y\frac{1}{\sqrt{2\pi}}e^{-y^2/2}\,\mathrm{d}y\right)\mathrm{d}x \\
&= -\frac{1}{\sqrt{2\pi}}\int x\phi(x)e^{-(2t-x)^2/2}\,\mathrm{d}x \\
&= -\frac{\sigma}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\frac{t^2}{\sigma^2}\right\}\int x\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2}\frac{(x-t)^2}{\sigma^2}\right\}\mathrm{d}x
\end{aligned}$$

where $\sigma^2 = \frac{1}{2}$. Now let $W \sim \mathcal{N}(t, \sigma^2)$,

$$\begin{aligned}
&= -\frac{\sigma}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\frac{t^2}{\sigma^2}\right\}\mathbb{E}(W) \\
&= -\frac{t\phi(t\sqrt{2})}{\sqrt{2}}. \qquad \square
\end{aligned}$$

Result C.4. *When $g(x,y) = (x+y)/2$ and X, Y are independent standard Gaussian variables, the $\alpha_{2,0}(t)$ and $\alpha_{0,2}(t)$ Hermite coefficients satisfy,*

$$\alpha_{2,0}(t) = \alpha_{0,2}(t) = \alpha_{1,1}(t).$$

Proof. Consider,

$$\begin{aligned}\alpha_{2,0}(t) &= \mathbb{E}\left[(X^2-1)\mathbb{I}\{X+Y\leq 2t\}\right]\\ &= \int (x^2-1)\phi(x)\Phi(2t-x)\,\mathrm{d}x\end{aligned}$$

noting that $\frac{\mathrm{d}}{\mathrm{d}x}x\phi(x) = -(x^2-1)\phi(x)$ and using integration by parts we have,

$$\begin{aligned} &= -\frac{1}{2\pi}\int x\exp\left\{-\frac{1}{2}\left[2(t-x)^2+2t^2\right]\right\}\mathrm{d}x\\ &= -\frac{t\phi(t\sqrt{2})}{\sqrt{2}}\\ &= \alpha_{1,1}(t).\end{aligned}$$

The final equality draws on Result C.3. □

C.3 RESULTS FOR SRD PROCESSES

C.3.1 *Convergence of empirical distribution functions*

In the context of short range dependent Gaussian sequences, Csörgó and Mielniczuk (1996) show that the empirical process $\sqrt{n}(F_n - \Phi_\sigma)$ converges in distribution to a Gaussian process in the space $\mathcal{D}([-\infty, \infty])$, the space of càdlàg functions. Note that Φ_σ is the CDF of a Gaussian random variable with mean zero and standard deviation σ.

Let $\{X_i\}_{i\geq 1}$ be a stationary zero-mean Gaussian sequence with covariance function $\gamma(h) = \mathbb{E}(X_1 X_{h+1})$. Define the empirical distribution function, $F_n(r) = n^{-1}\sum_{i=1}^{n} \mathbb{I}\{X_i \leq r\}$. Also let,

$$\begin{aligned} F_n(r) - F(r) &= \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{I}\{X_i \leq r\} - F(r)\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{k=m}^{\infty}\frac{\alpha_k(r)}{k!}H_k(X_i) \end{aligned}$$

where $\alpha_k(r) = \mathbb{E}[(\mathbb{I}\{X \leq r\} - F(r))H_k(X)]$ and m is the Hermite rank of the class of functions $\{\mathbb{I}\{X_i \leq r\} - F(r)]; r \in \mathbb{R}\}$.

Theorem C.1 (Csörgó and Mielniczuk (1996))**.** *If $\sum_{h=1}^{\infty}|\gamma(h)|^m < \infty$, then $\sqrt{n}(F_n - \Phi_\sigma)$ converges in distribution in $\mathcal{D}([-\infty, \infty])$ to a mean-zero Gaussian process, W, with covariance function,*

$$\mathbb{E}[W(s)W(t)] = \sum_{q=m}^{\infty}\frac{\alpha_q(s)\alpha_q(t)}{q!}\left[\gamma(0) + 2\sum_{h=1}^{\infty}\gamma^q(h)\right],$$

where $\alpha_q(r) = \int\left[\mathbb{I}\{\sigma x \leq r\} - \Phi_\sigma(r)\right]H_q(x)\mathrm{d}\Phi(x)$ for all $r \in \mathbb{R}$.

C.3.2 *Central limit theorem*

Arcones (1994, Theorem 4) provides a CLT result for more general functions operating on vectors of short range dependent processes. It is adapted to our needs in the following theorem.

Theorem C.2 (Arcones (1994, adapted from Theorem 4))**.** *Let $\{X_i\}_{i\geq 1}$ be a stationary mean-zero Gaussian sequence in $\mathbb{R}$. Let f be a function on $\mathbb{R}$ with Hermite rank m, $1 \leq m < \infty$. Suppose that $\sum_{h=-\infty}^{\infty} |\gamma(h)|^m < \infty$, then,*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (f(X_i) - \mathbb{E}f(X_i)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2(f)),$$

where

$$\begin{aligned}\sigma^2(f) =& \mathbb{E}\left[(f(X_1) - \mathbb{E}f(X_1))^2\right] \\ &+ 2\sum_{k=1}^{\infty} \mathbb{E}\left[(f(X_1) - \mathbb{E}f(X_1))(f(X_{1+k}) - \mathbb{E}f(X_{1+k}))\right].\end{aligned}$$

C.3.3 *Existing results for estimators*

The asymptotics of classical scale and autocovariance estimators under Gaussian SRD processes are well established.

Result C.5 (Lévy-Leduc et al., 2011c, Proposition 1)**.** *Under Assumption 3.1,*

$$\sqrt{n}(\hat{\sigma}_n - \sigma) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \tilde{\sigma}_{cl}^2),$$

where

$$\tilde{\sigma}_{cl}^2 = \frac{1}{2\gamma(0)}\left(\gamma^2(0) + 2\sum_{h\geq 1}\gamma^2(h)\right).$$

Lévy-Leduc et al. (2011c) also establish an analogous limit result for Q_n.

Result C.6 (Lévy-Leduc et al., 2011c, Theorem 1)**.** *Under Assumption 3.1, Q_n satisfies the following* CLT*,*

$$\sqrt{n}(Q_n - \sigma) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \tilde{\sigma}_Q^2),$$

where $\sigma = \sqrt{\gamma(0)}$ and the limiting variance $\tilde{\sigma}^2$ is given by

$$\tilde{\sigma}_Q^2 = \sigma^2\mathbb{E}\left[\mathrm{IF}^2(X_1/\sigma; Q, \Phi)\right] + 2\sigma^2\sum_{k\geq 1}\mathbb{E}\left[\mathrm{IF}(X_1/\sigma; Q, \Phi)\,\mathrm{IF}(X_{k+1}/\sigma; Q, \Phi)\right],$$

where $\mathrm{IF}(x; Q, \Phi)$ *is defined in equation* (2.11).

Standard techniques can be applied to transfer the limiting results of scale estimators to the corresponding autocovariance estimators.

Result C.7 (Lévy-Leduc et al., 2011c, Proposition 2). *Under Assumption 3.1, for a given non-negative integer h, as $n \to \infty$,*

$$\sqrt{n}(\hat{\gamma}(h) - \gamma(h)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \breve{\sigma}^2_{cl}(h)),$$

where

$$\breve{\sigma}^2_{cl}(h) = \gamma^2(0) + \gamma^2(h) + 2\sum_{k\geq 1}\gamma^2(k) + 2\sum_{k\geq 1}\gamma(k+h)\gamma(k-h).$$

The application of Arcones (1994, Theorem 4) to this setting follows by noting, as in Bartlett (1981, p. 302),

$$\begin{aligned}\mathbb{E}(X_u X_{u+s} X_{u+v} X_{u+v+t}) = {} & \mathbb{E}(X_u X_{u+s})\mathbb{E}(X_{u+v} X_{u+v+t}) \\ & + \mathbb{E}(X_u X_{u+v+t})\mathbb{E}(X_{u+s} X_{u+v}) \\ & + \mathbb{E}(X_u X_{u+v})\mathbb{E}(X_{u+s} X_{u+v+t}) + \kappa_{v,s,t}\end{aligned}$$

where $\kappa_{v,s,t}$ is the fourth order cumulant between X_u, X_{u+s}, X_{u+v} and X_{u+v+t}, which is necessarily zero for Gaussian processes.

Result C.8 (Lévy-Leduc et al., 2011c, Theorem 4). *Under Assumption 3.1, $\hat{\gamma}_Q(h)$ satisfies the following CLT,*

$$\sqrt{n}(\hat{\gamma}_Q(h) - \gamma(h)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \breve{\sigma}^2_Q(h)),$$

where

$$\begin{aligned}\breve{\sigma}^2_Q(h) = {} & \mathbb{E}\left[\mathrm{IF}^2(X_1, X_{1+h}; \gamma_Q, \Phi)\right] \\ & + 2\sum_{k\geq 1}\mathbb{E}\left[\mathrm{IF}(X_1, X_{1+h}; \gamma_Q, \Phi)\,\mathrm{IF}(X_{k+1}, X_{k+1+h}; \gamma_Q, \Phi)\right],\end{aligned}$$

and

$$\begin{aligned}\mathrm{IF}(x, y; \gamma_Q, \Phi) = {} & (\gamma(0) + \gamma(h))\,\mathrm{IF}\left(\frac{x+y}{\sqrt{2(\gamma(0)+\gamma(h))}}; Q, \Phi\right) \\ & - (\gamma(0) - \gamma(h))\,\mathrm{IF}\left(\frac{x-y}{\sqrt{2(\gamma(0)-\gamma(h))}}; Q, \Phi\right).\end{aligned}$$

C.4 RESULTS FOR LRD PROCESSES

C.4.1 *Limit results for U-processes*

Lévy-Leduc et al. (2011a,b,c) explore the limiting behaviour of U-processes of LRD sequences. Of particular interest in scale and autocovariance estimation are processes that have Hermite rank $m = 1$ or $m = 2$. The following theorem considers the case where $m = 2$ and $D > \frac{1}{2}$. Recall that D is the autocovariance decay rate in the parameterisation of the LRD process. The notation $h(\cdot, \cdot; t)$ and $h_1(x;t)$ follows from Section 2.2.

Theorem C.3 (Lévy-Leduc et al., 2011a, Theorem 1)**.** *Let I be a compact interval of $\mathbb{R}$. Suppose that the Hermite rank of the class of function $\{h(\cdot, \cdot; t) - G(t); t \in I\}$ is $m = 2$ and that $\{X_i\}_{i \geq 1}$ is a Gaussian LRD process with $D > \frac{1}{2}$. Assume that h and h_1 satisfy the three following conditions:*

A. *There exists a positive constant C such that for all s, t in I, u, v in $\mathbb{R}$,*

$$\mathbb{E}[|h(X+u, Y+v; s) - h(X+u, Y+v; t)|] \leq C|t-s|,$$

where (X, Y) is a standard Gaussian vector.

B. *There exists a positive constant C such that for all $k \geq 1$,*

$$\mathbb{E}[|h(X_1+u, X_{1+k}+v; t) - h(X_1, X_{1+k}; t)|] \leq C(|u| + |v|),$$

and $\mathbb{E}[|h(X_1, X_{1+k}; s) - h(X_1, X_{1+k}; t)|] \leq C|t-s|$.

C. *There exists a positive constant C such that for all t, s in I, and x, u, v in $\mathbb{R}$,*

$$|h_1(x+u; t) - h_1(x+v; t)| \leq C(|u| + |v|),$$

and $|h_1(x;s) - h_1(x;t)| \leq C|t-s|$.

Then the U-process,

$$\left(\sqrt{n}(G_n(t) - G(t))\right)_{t \in I}$$

converges weakly in $\mathcal{D}(I)$ equipped with the topology of uniform convergence to the zero mean Gaussian process $(W(t))_{t \in I}$ with covariance, $\mathbb{E}(W(s)W(t))$, given by,

$$\begin{aligned} &4\operatorname{cov}(h_1(X_1;s), h_1(X_1;t)) \\ &\quad + 4\sum_{\ell \geq 1} \left[\operatorname{cov}(h_1(X_1;s), h_1(X_{\ell+1};t)) + \operatorname{cov}(h_1(X_1;t), h_1(X_{\ell+1};s))\right]. \end{aligned} \tag{C.2}$$

The conditions in Theorem C.3 are satisfied for any linear function g where, $h(\cdot,\cdot;t) = \mathbb{I}\{g(\cdot,\cdot) \leq t\}$, and in particular for $g(x,y) = (x+y)/2$.

Corollary C.1 (Lévy-Leduc et al., 2011a, Corollary 3). *Let p be a fixed real number in $(0,1)$. Assume that the conditions of Theorem* (C.3) *are satisfied. Suppose also that there exists some t in I such that $G(t) = p$, that G is differentiable at t and that $G'(t)$ is non null. Then in the limit as $n \to \infty$,*

$$\sqrt{n}\left(G_n^{-1}(p) - G^{-1}(p)\right) \xrightarrow{\mathcal{D}} -\frac{W(G^{-1}(p))}{G'(G^{-1}(p))},$$

where W is a Gaussian process having covariance structure given by (C.2).

The following theorem outlines the limit result for U-processes when $m = 1$ or 2 and $D < \frac{1}{m}$.

Theorem C.4 (Lévy-Leduc et al., 2011a, Theorem 2). *Let I be a compact interval of $\mathbb{R}$. Suppose that $\{X_i\}_{i\geq 1}$ is a LRD process with $D < \frac{1}{m}$, where $m = 1$ or 2 is the Hermite rank of the class of functions $\{h(\cdot,\cdot;t) - G(t), t \in I\}$. Assume the following:*

A. *There exists a positive constant C such that, for all $k \geq 1$ and for all s, t in I,*

$$\mathbb{E}[|h(X_1, X_{1+k}; s) - h(X_1, X_{1+k}; t)|] \leq C|t-s|. \tag{C.3}$$

B. *G is a Lipschitz function.*

C. *The function $\widetilde{\Lambda}$ is also a Lipschitz function, where for all $s \in I$,*

$$\widetilde{\Lambda}(s) = E[h(X,Y;s)(|X| + |XY| + |X^2 - 1|)], \tag{C.4}$$

where X and Y are independent standard Gaussian random variables.

Then

$$\left(n^{mD/2} L(n)^{-m/2} (G_n(t) - G(t))\right)_{t\in I}$$

converges weakly in $\mathcal{D}(I)$, equipped with the topology of uniform convergence, to

$$\left(2\alpha_{1,0}(t) k(D)^{-1/2} Z_{1,D}(1)\right)_{t\in I} \quad \text{if } m = 1,$$

and to

$$\left(k(D)^{-1}\left[\alpha_{1,1}(t) Z_{1,D}(1)^2 + \alpha_{2,0}(t) Z_{2,D}(1)\right]\right)_{t\in I} \quad \text{if } m = 2,$$

where $k(D) = \mathrm{Beta}((1-D)/2, D)$.

The limit processes appearing in Theorem C.4 are the standard fractional Brownian motion $\{Z_{1,D}(t)\}_{0\leq t\leq 1}$ and the Rosenblatt process $\{Z_{2,D}(t)\}_{0\leq t\leq 1}$. They are defined through multiple Wiener-Itô integrals and given by

$$Z_{1,D}(t) = \int_{\mathbb{R}} \left[\int_0^t (u-x)_+^{-(D+1)/2} \, \mathrm{d}u \right] \mathrm{d}B(x), \quad 0 < D < 1, \tag{C.5}$$

and

$$Z_{2,D}(t) = \int_{\mathbb{R}}' \left[\int_0^t (u-x)_+^{-(D+1)/2} (u-y)_+^{-(D+1)/2} \, \mathrm{d}u \right] \mathrm{d}B(x)\, \mathrm{d}B(y), \tag{C.6}$$

for $0 < D < \frac{1}{2}$, where $x_+ = \max\{x, 0\}$ and B is the standard Brownian motion. The symbol $\int'$ means that the domain of integration excludes the diagonal. Note that $Z_{1,D}$ and $Z_{2,D}$ are dependent but uncorrelated. A recent exposition on these types of processes in this context can be found in Pipiras and Taqqu (2010), Taqqu (2011) and Taqqu and Veillette (2013).

The inverse map (see for example van der Vaart and Wellner, 1996) can be used to transfer these convergence results to U-quantiles.

Corollary C.2 (Lévy-Leduc et al., 2011a, Corollary 4). *Let p be a fixed real number in $(0,1)$. Assume that the conditions of Theorem* (C.4) *are satisfied. Suppose also that there exists some t in I such that $G(t) = p$, that G is differentiable at t and that $G'(t)$ is non null. Then, as $n \to \infty$,*

$$\frac{n^{mD/2}}{L^{-m/2}(n)} (G_n^{-1}(p) - G^{-1}(p))$$

converges in distribution to

$$-2k(D)^{-1/2} \frac{\alpha_{1,0}(G^{-1}(p))}{G'(G^{-1}(p))} Z_{1,D}(1), \quad \text{if } m = 1,$$

and to

$$-k(D)^{-1} \frac{\alpha_{1,1}(G^{-1}(p)) Z_{1,D}(1)^2 + \alpha_{2,0}(G^{-1}(p)) Z_{2,D}(1)}{G'(G^{-1}(p))}, \quad \text{if } m = 2.$$

The following lemma is also useful for finding the limit distribution of statistics that have been decomposed using a Hermite expansion with $m = 2$ as in Sections 3.5.3 and 3.5.4.3.

Lemma C.4 (Lemma 15 from Lévy-Leduc et al. (2011a)). *Under Assumption 3.2, with $D < \frac{1}{2}$, let a and b be two real constants, then as n tends to infinity,*

$$k(D) \frac{n^{D-2}}{L_\gamma(n)} \left[an \sum_{i=1}^{n} (X_i^2 - 1) + b \sum_{1\leq i,j\leq n} X_i X_j \right] \xrightarrow{\mathcal{D}} \left[a Z_{2,D}(1) + b Z_{1,D}^2(1) \right].$$

C.4.2 *Existing results for estimators*

C.4.2.1 *Hodges-Lehmann estimator*

Lévy-Leduc et al. (2011a) consider the problem of estimating the location parameter of a LRD Gaussian process. Assume that the process $\{Y_i\}_{i\geq 1}$ satisfies $Y_i = \theta + X_i$ where $\{X_i\}_{i\geq 1}$ satisfies Assumption 3.2. Using the notation of Section 2.2, let $g(x,y) = (x+y)/2$ and thus $h(x,y;t) = \mathbb{I}\{x+y \leq 2t\}$. The Hodges-Lehmann estimator can then be expressed as $\hat{\theta}_{HL} = \theta + G_n^{-1}(\frac{1}{2})$, where

$$G_n(r) = \frac{2}{n(n-1)} \sum_{i<j} \mathbb{I}\{X_i + X_j \leq 2r\}.$$

Result C.9. *Under Assumption 3.2, Lévy-Leduc et al. (2011a) demonstrate that, for all* $0 < D < 1$,

$$\sqrt{\frac{n^D}{L(n)}}(\hat{\theta}_{HL} - \theta) \xrightarrow{\mathcal{D}} \frac{Z_{1,D}(1)}{\sqrt{k(D)}},$$

a zero-mean Gaussian random variable with variance $2(1-D)^{-1}(2-D)^{-1}$.

To show this result, Lévy-Leduc et al. (2011a) show that Hermite rank of the class of functions $\{h(\cdot,\cdot;t) - G(t);\, t \in \mathbb{R}\}$ is $m = 1$. This can be seen as a special case of Result C.2 where for all $t \in \mathbb{R}$, $\alpha_{1,0}(t)$ and $\alpha_{0,1}(t)$ are not equal to zero. After the appropriate assumptions have been checked, they apply Theorem C.4, and because $m = 1$ and $D < \frac{1}{m}$ obtain a single Gaussian limit result for all $D \in (0,1)$,

$$\left(n^{D/2} L(n)^{-1/2} (G_n(t) - G(t))\right)_{t\in\mathbb{R}}$$

converges weakly in $\mathcal{D}([-\infty, +\infty])$, equipped with the uniform norm, to

$$\left(-\sqrt{2}k(D)^{-1/2}\phi(t\sqrt{2})Z_{1,D}(1)\right)_{t\in\mathbb{R}}.$$

Corollary C.2, noting Result C.1, therefore implies that

$$n^{D/2} L(n)^{-1/2} (\hat{\theta}_{HL} - \theta) \xrightarrow{\mathcal{D}} k(D)^{-1/2} Z_{1,D}(1),$$

As Lévy-Leduc et al. (2011a) note, this is the same limiting distribution as the sample mean and hence in the long-memory framework with $0 < D < 1$, the Hodges-Lehmann estimator converges to θ at the same rate as the sample mean, $\bar{X}_n$ and there is no loss of efficiency. A similar result was also proved in Beran (1991) for location M-estimators.

More generally, we need not consider simply the median, rather a similar limit results holds for central quantiles of the pairwise mean distribution.

C.4.2.2 *Scale and autocovariance estimators*

Using the results in Section C.4.1, Lévy-Leduc et al. (2011c) derive the limiting distribution for Q_n and autocovariance estimator, $\hat{\gamma}_Q$.

Result C.10 (Lévy-Leduc et al., 2011c, Theorem 3). *Under Assumption 3.2, Q_n satisfies the following limit theorems as $n \to \infty$,*

A. *If $D > \frac{1}{2}$,*

$$\sqrt{n}(Q_n - \sigma) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \widetilde{\sigma}^2),$$

where $\sigma = \sqrt{\gamma(0)}$ and the limiting variance $\widetilde{\sigma}^2$ is given by

$$\sigma^2 \mathbb{E}\left[\mathrm{IF}^2(X_1/\sigma; Q, \Phi)\right] + 2\sigma^2 \sum_{k \geq 1} \mathbb{E}\left[\mathrm{IF}(X_1/\sigma; Q, \Phi)\,\mathrm{IF}(X_{k+1}/\sigma; Q, \Phi)\right].$$

B. *If $D < \frac{1}{2}$,*

$$\frac{k(D)n^D}{L(n)}(Q_n - \sigma) \xrightarrow{\mathcal{D}} \frac{\sigma}{2}\left(Z_{2,D}(1) - Z_{1,D}^2(1)\right).$$

Note that the limiting distribution for Q_n when $D > \frac{1}{2}$ the same as that found in the SRD setting, Result C.6.

It is instructive to compare the robust estimators with the classical estimators. Lévy-Leduc et al. (2011c) show the following convergence result for the SD.

Result C.11 (Lévy-Leduc et al., 2011c, Proposition 3). *Under Assumption 3.2, as $n \to \infty$,*

A. *If $D > \frac{1}{2}$, $\sqrt{n}(\hat{\sigma}_n - \sigma) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \widetilde{\sigma}_{cl}^2)$, where $\widetilde{\sigma}_{cl}^2$ is the same as in the SRD case, Result C.5.*

B. *If $D < \frac{1}{2}$,*

$$\frac{k(D)n^D}{L(n)}(\hat{\sigma}_n - \sigma) \xrightarrow{\mathcal{D}} \frac{\sigma}{2}\left(Z_{2,D}(1) - Z_{1,D}^2(1)\right).$$

Note that the limiting distribution of $\hat{\sigma}_n$ and Q_n are identical and there is no loss of (asymptotic) efficiency when $D < \frac{1}{2}$. Furthermore, the limiting distribution for $D < \frac{1}{2}$ is not centred and is asymmetric.

Result C.12 (Lévy-Leduc et al., 2011c, Theorem 4). *Under Assumption 3.2 with the additional requirement that L has three continuous derivatives and $L_i(x) = x^i L^{(i)}(x)$ satisfies $L_i(x)/x^\epsilon = O(1)$, for some $\epsilon \in (0, D)$, as $x \to \infty$, for all $i = 0, 1, 2, 3$, where $L^{(i)}$ denotes the ith derivative of L. Let h be a non-negative integer. Then $\hat{\gamma}_Q(h)$ satisfies the following limit theorems as $n \to \infty$.*

A. *If $D > \frac{1}{2}$, $\sqrt{n}(\hat{\gamma}_Q(h) - \gamma(h)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \breve{\sigma}^2(h))$, where*

$$\begin{aligned}\breve{\sigma}^2(h) = {} & \mathbb{E}\left[\mathrm{IF}^2(X_1, X_{1+h}; \gamma_Q, \Phi)\right] \\ & + 2\sum_{k \geq 1} \mathbb{E}\left[\mathrm{IF}(X_1, X_{1+h}; \gamma_Q, \Phi)\, \mathrm{IF}(X_{k+1}, X_{k+1+h}; \gamma_Q, \Phi)\right].\end{aligned}$$

B. *If $D < \frac{1}{2}$,*

$$\frac{k(D) n^D}{\widetilde{L}(n)}(\hat{\gamma}_Q(h) - \gamma(h)) \xrightarrow{\mathcal{D}} \frac{\gamma(0) + \gamma(h)}{2}(Z_{2,D}(1) - Z_{1,D}^2(1))$$

where $k(D) = \mathrm{Beta}((1-D)/2, D)$ and

$$\widetilde{L}(n) = 2L(n) + L(n+h)(1 + h/n)^{-D} + L(n-h)(1 - h/n)^{-D}.$$

As for Q_n, when $D > \frac{1}{2}$ the limit result for $\hat{\gamma}_Q$ is the same as in the SRD case. The assumptions made on L_i are satisfied if L is the logarithmic function or a power thereof. Following Taqqu (2003), if we restrict attention to an ARFIMA$(0, d, 0)$ process, then

$$L(n) = \frac{\Gamma(D)}{\Gamma(d)\Gamma(1-d)} = a,$$

which, for fixed d is a constant and therefore slowly varying. Furthermore, as $n \to \infty$, $\widetilde{L}(n) \to 4a$.

Result C.13 (Lévy-Leduc et al., 2011c, Proposition 4). *Under Assumption 3.2, as $n \to \infty$,*

A. *If $D > \frac{1}{2}$, $\sqrt{n}(\hat{\gamma}(h) - \gamma(h)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \breve{\sigma}_{cl}^2(h))$, where $\breve{\sigma}_{cl}^2(h)$ is the same as in the SRD case, Result C.7.*

B. *If $D < \frac{1}{2}$,*

$$\frac{k(D) n^D}{\widetilde{L}(n)}(\hat{\gamma}(h) - \gamma(h)) \xrightarrow{\mathcal{D}} \frac{\gamma(0) + \gamma(h)}{2}(Z_{2,D}(1) - Z_{1,D}^2(1))$$

where,

$$\widetilde{L}(n) = 2L(n) + L(n+h)(1 + h/n)^{-D} + L(n-h)(1 - h/n)^{-D}.$$

C.5 RELATIVE EFFICIENCIES

	P_n			Q_n		
n	$\gamma(0)$	$\gamma(1)$	$\gamma(2)$	$\gamma(0)$	$\gamma(1)$	$\gamma(2)$
	Relative efficiencies					
100	0.84	0.81	0.80	0.76	0.75	0.73
500	0.86	0.86	0.86	0.81	0.82	0.81
1000	0.86	0.87	0.87	0.82	0.83	0.82
	Mean square errors					
100	0.86	0.83	0.82	0.78	0.76	0.74
500	0.86	0.87	0.86	0.81	0.82	0.82
1000	0.87	0.87	0.87	0.82	0.83	0.83

Table C.1: Efficiencies and MSEs of the robust estimators relative to the classical methods over 100,000 replications from an ARFIMA$(0, 0.1, 0)$ process.

	P_n			Q_n		
n	$\gamma(0)$	$\gamma(1)$	$\gamma(2)$	$\gamma(0)$	$\gamma(1)$	$\gamma(2)$
	Relative efficiencies					
100	0.84	0.79	0.77	0.83	0.85	0.83
500	0.90	0.87	0.86	0.90	0.90	0.90
1000	0.91	0.89	0.89	0.91	0.91	0.91
	Mean square errors					
100	1.05	1.06	1.05	1.04	1.02	1.02
500	1.03	1.03	1.03	1.02	1.01	1.01
1000	1.02	1.02	1.03	1.01	1.01	1.01

Table C.2: Efficiencies and MSEs of the robust estimators relative to the classical methods over 100,000 replications from an ARFIMA$(0, 0.4, 0)$ process.

D

COVARIANCE AND PRECISION MATRIX ESTIMATION

D.1 GENERATING PRECISION MATRICES

The R code below generates the covariance matrices used in the simulation study in Section 4.6.

```r
# Model 1 # Banded precision matrix structure
Theta1 = function(p,a=0.6){
  Theta = matrix(NA,p,p)
  for(i in 1:p){
    for(j in 1:p){
      Theta[i,j] = a^abs(i-j)
    }
  }
  return(Theta)
}
# Model 2 # Randomly scattered sparse precision matrix
Theta2 = function(p,seed){
  if(!missing(seed)) set.seed(seed)
  B = matrix(NA,p,p)
  for(i in 1:p){
    for(j in 1:i){
      B[i,j] = B[j,i] = sample(x=c(0.5,0),size=1,prob=c(0.1,0.9))
    }
  }
  diag(B) = 0
  Ident = diag(p)
  cond.Theta = function(d) {(kappa(B + d*Ident)-p)^2}
  delta = optim(par=1,fn=cond.Theta,lower=0,upper=1000,method="Brent")$
      par
  Theta = B+delta*Ident
  #kappa(Theta) check that the condition number equals p
  return(cov2cor(Theta))
}
# Model 3 # dense precision matrix
Theta3 = function(p){
  Theta = matrix(0.5,p,p)
  diag(Theta)=1
  return(Theta)
}
```

D.2 ENTROPY LOSS RESULTS

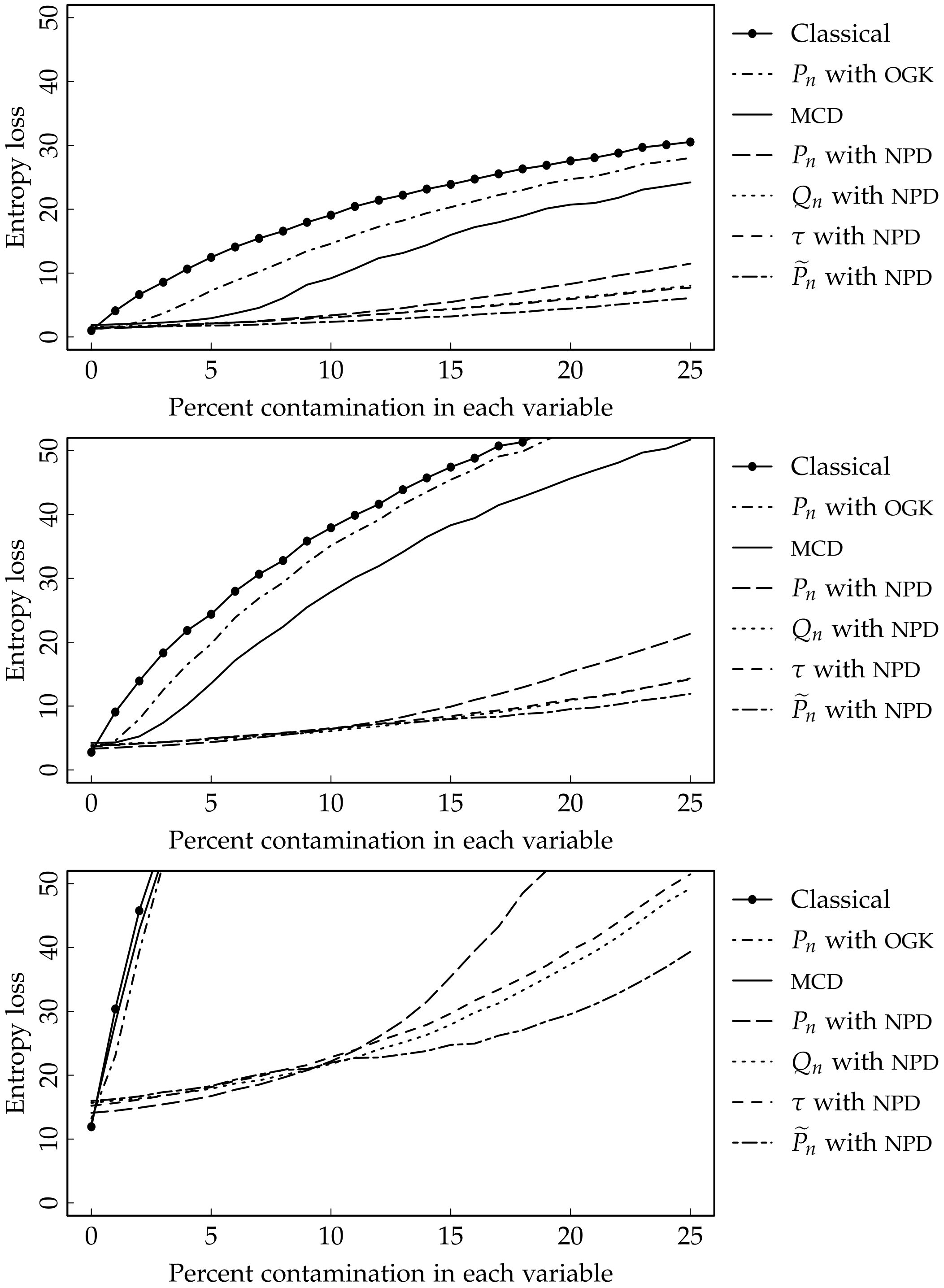


Figure D.1: Raw entropy loss results for the CLIME routine applied to data generated with a banded precision matrix for $p = 15$ (top), $p = 30$ (middle) and $p = 90$ (bottom) with extreme outliers.

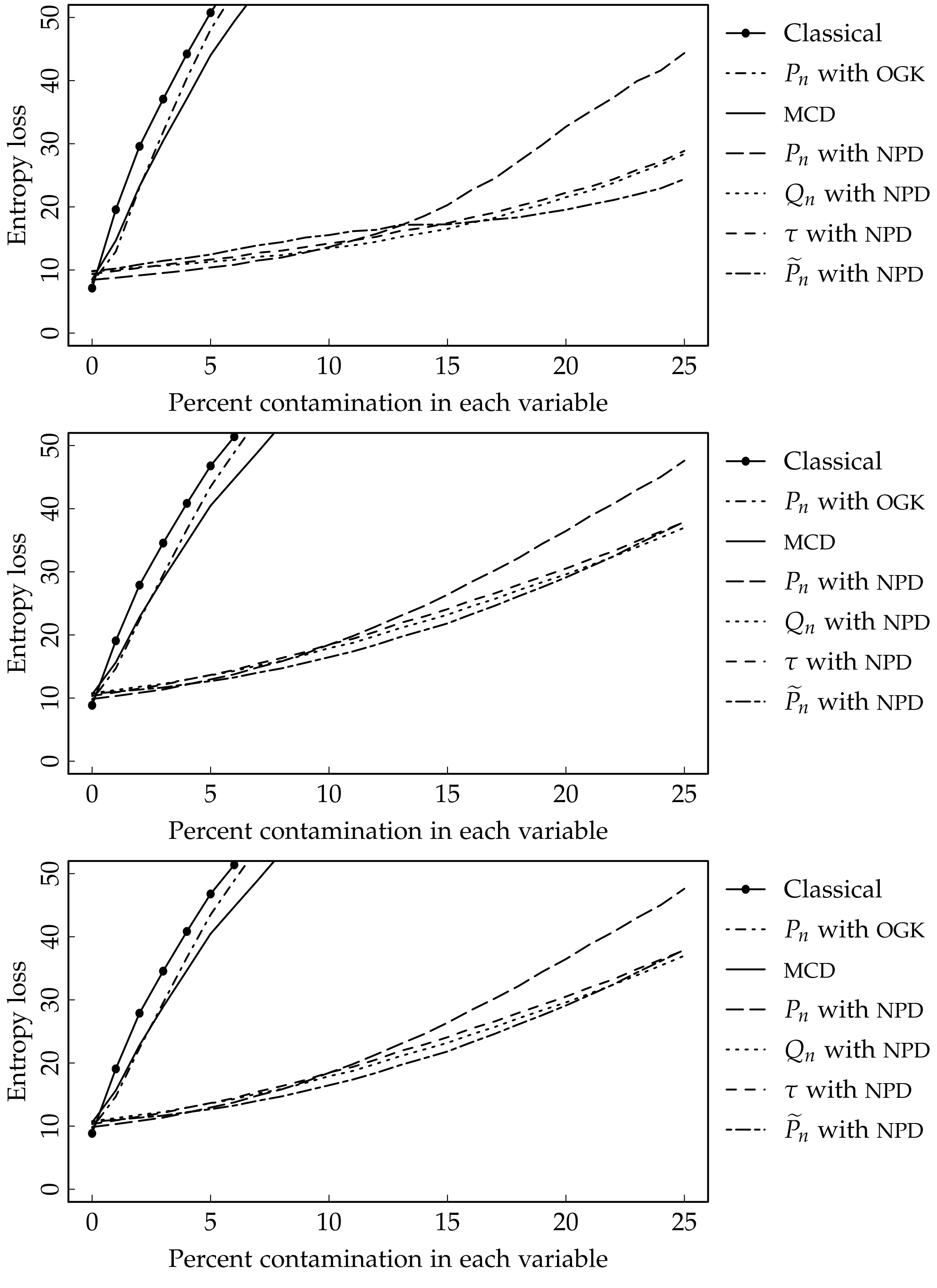


Figure D.2: Raw entropy loss results for data generated with a banded precision matrix with extreme outliers for $p = 60$ using CLIME (top), QUIC (middle) and GLASSO (bottom).

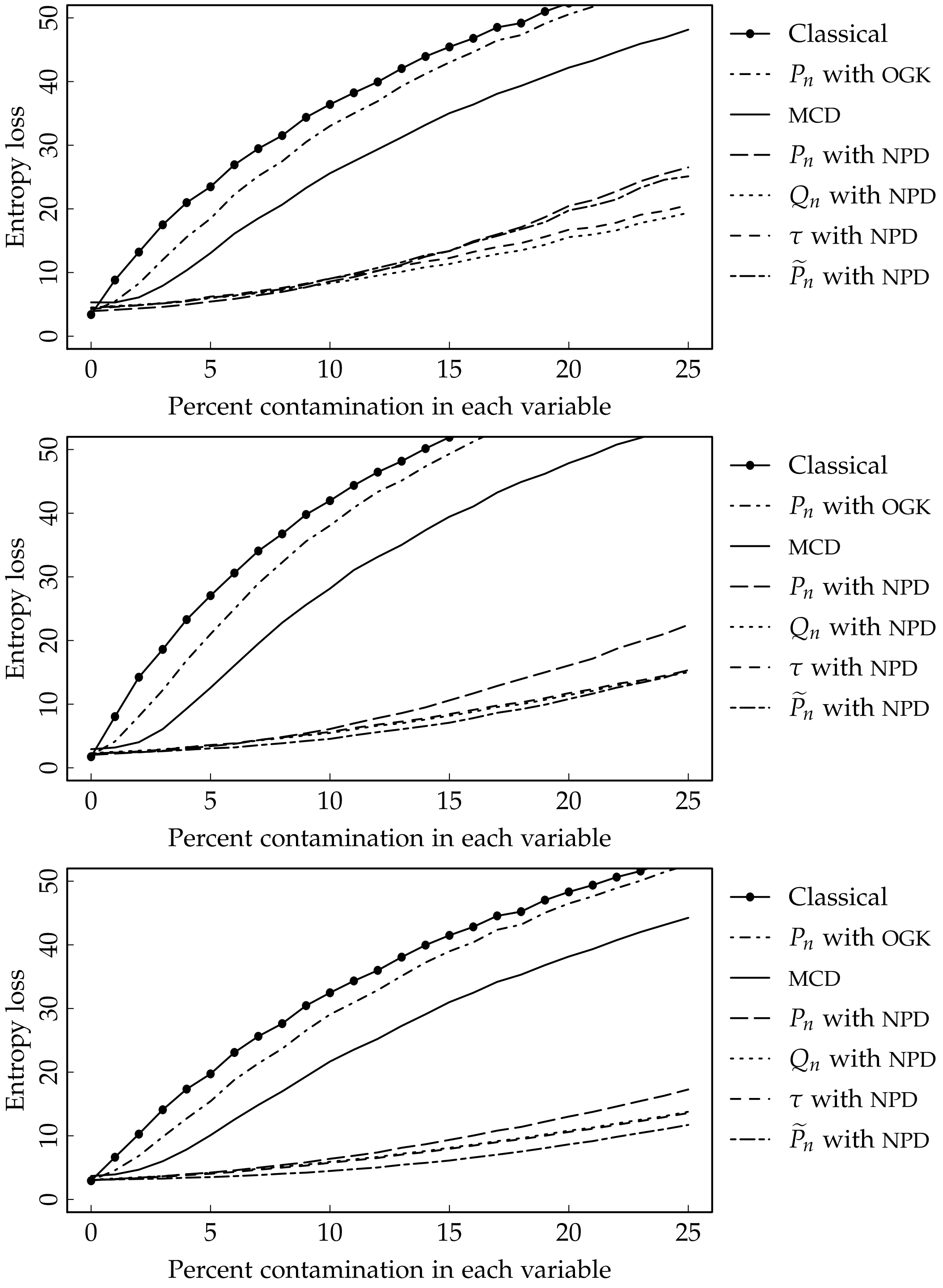


Figure D.3: Raw entropy loss results for data generated with a banded precision matrix (top), scattered precision matrix (middle) and dense precision matrix (bottom) with extreme outliers for $p = 30$ using the QUIC routine.

D.3 PRIAL RESULTS

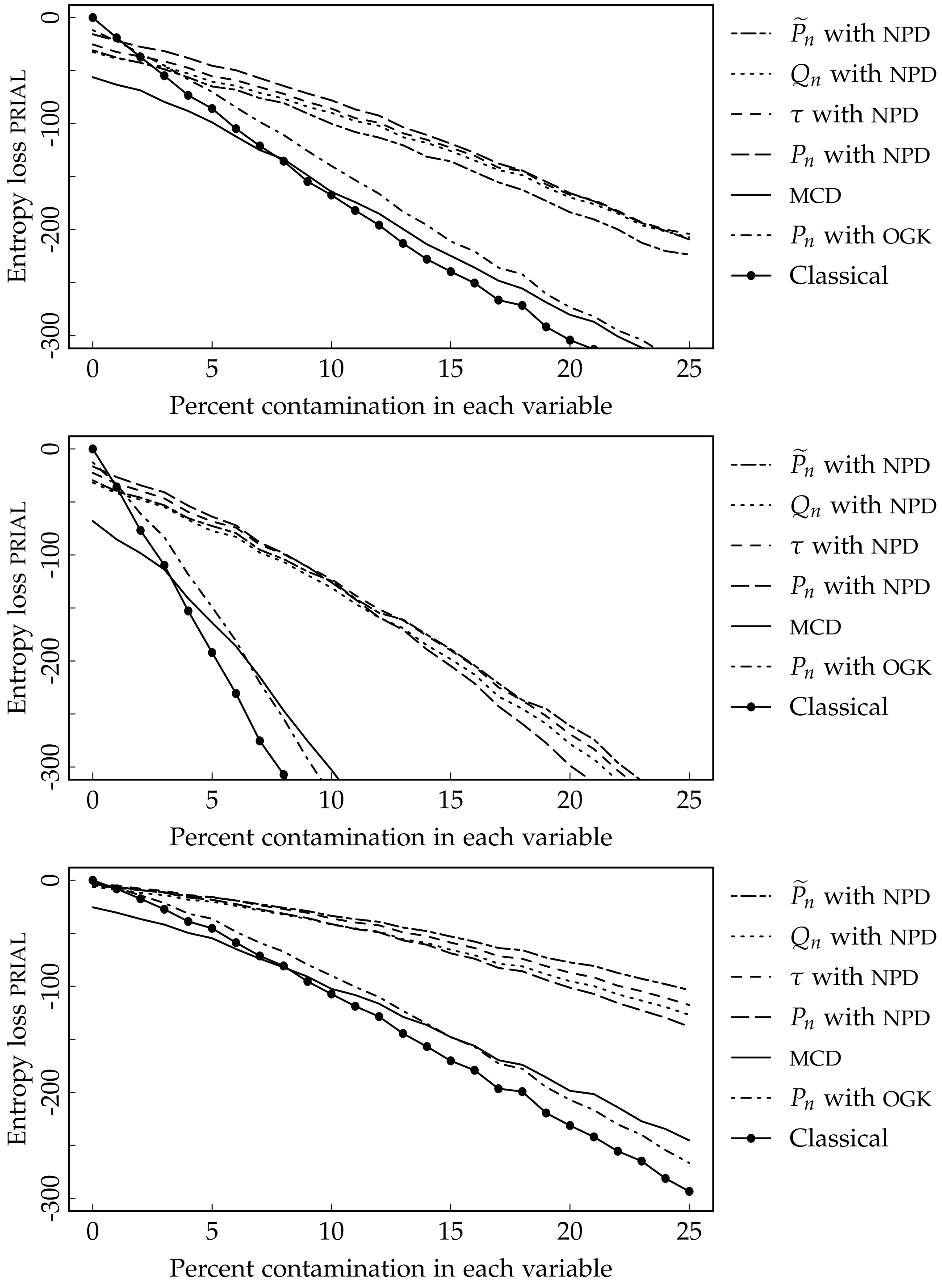


Figure D.4: PRIAL results for data generated with a banded precision matrix (top), scattered precision matrix (middle) and dense precision matrix (bottom) for $p = 30$ using the QUIC routine with moderate outliers.

D.4 LOG DETERMINANT RESULTS

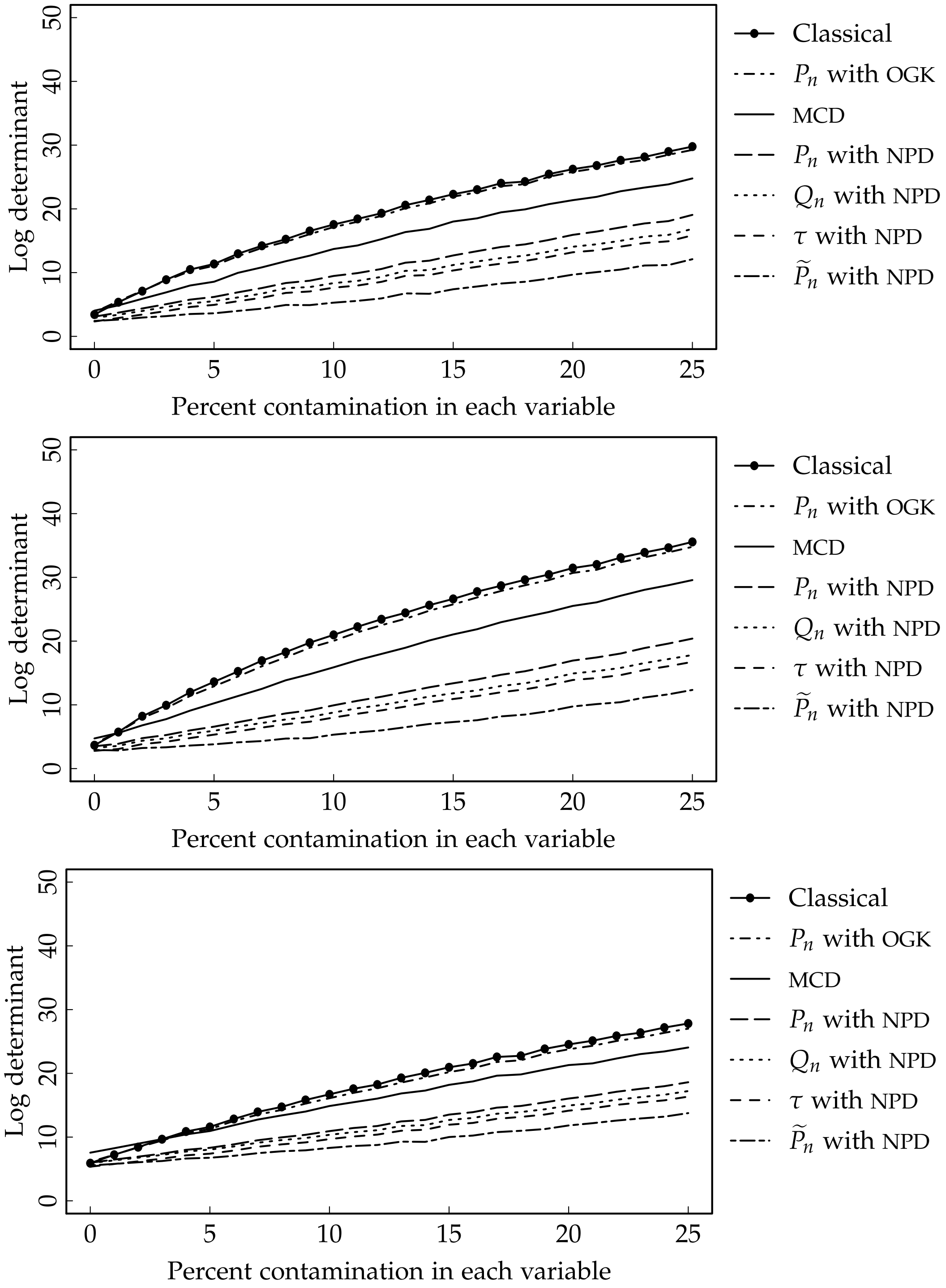


Figure D.5: Log determinant results for the GLASSO with moderate outliers, $p = 30$ for the banded (top), scattered sparsity (middle) and dense (bottom) precision matrix scenarios.

BIBLIOGRAPHY

Abramowitz, M. and Stegun, I. A., eds. (1973). *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*. New York: Dover Publications.

Adler, R., Feldman, R. and Taqqu, M. S. (1998). *A Practical Guide to Heavy Tails: Statistical Techniques and Applications*. Boston: Birkhäuser.

Alqallaf, F., Van Aelst, S., Yohai, V. J. and Zamar, R. H. (2009). Propagation of outliers in multivariate data. *The Annals of Statistics*, **37** (1), 311–331.

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H. and Tukey, J. W. (1972). *Robust Estimates of Location: Survey and Advances*. Princeton: Princeton University Press.

Arcones, M. A. (1994). Limit theorems for nonlinear functionals of a stationary Gaussian sequence of vectors. *The Annals of Probability*, **22** (4), 2242–2274.

Arcones, M. A. and Giné, E. (1992). On the bootstrap of M-estimators and other statistical functionals. In: *Exploring the Limits of Bootstrap*. LePage, R. and Billard, L. (eds.). New York: Wiley, pp. 13–47.

Bachmaier, M. (2000). Efficiency comparison of M-estimates for scale at t-distributions. *Statistical Papers*, **41** (1), 53–64.

Bartlett, M. (1981). *An Introduction to Stochastic Processes with Special Reference to Methods and Applications*. New York: Cambridge University Press.

Beran, J. (1991). M-estimators of location for Gaussian and related processes with slowly decaying serial correlations. *Journal of the American Statistical Association*, **86** (415), 704–708.

– (1994). *Statistics for Long-Memory Processes*. New York: Chapman & Hall.

Beran, J., Feng, Y., Ghosh, S. and Kulik, R. (2013). *Long-Memory Processes: Probabilistic Properties and Statistical Methods*. Berlin: Springer.

Bickel, P. J. and Lehmann, E. L. (1979). Descriptive statistics for nonparametric models. IV: Spread. In: *Contributions to Statistics, Hájek Memorial Volume*. Jurecková, J. (ed.), pp. 33–40.

Boente, G., Ruiz, M. and Zamar, R. H. (2010). On a robust local estimator for the scale function in heteroscedastic nonparametric regression. *Statistics & Probability Letters*, **80** (15-16), 1185–1195.

Bunde, A., Eichner, J. F., Kantelhardt, J. W. and Havlin, S. (2005). Long-term memory: A natural mechanism for the clustering of extreme events and anomalous residual times in climate records. *Physical Review Letters*, **94**, 048701.

Cai, T., Liu, W. and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, **106** (494), 594–607.

– (2012). *clime: Constrained ℓ_1-minimization for Inverse (covariance) Matrix Estimation*. R package version 0.4.1.

Carroll, R. J. (2003). Variances are not always nuisance parameters. *Biometrics*, **59** (2), 211–220.

Cator, E. A. and Lopuhaä, H. P. (2010). Asymptotic expansion of the minimum covariance determinant estimators. *Journal of Multivariate Analysis*, **101** (10), 2372–2388.

– (2012). Central limit theorem and influence function for the MCD estimators at general multivariate distributions. *Bernoulli*, **18** (2), 520–551.

Chamberlain, G. and Imbens, G. W. (2003). Nonparametric applications of Bayesian inference. *Journal of Business & Economic Statistics*, **21** (1), 12–18.

Chareka, P., Matarise, F. and Turner, R. (2006). A test for additive outliers applicable to long-memory time series. *Journal of Economic Dynamics and Control*, **30** (4), 595–621.

Chen, K., Ying, Z., Zhang, H. and Zhao, L. (2008). Analysis of least absolute deviation. *Biometrika*, **95** (1), 107–122.

Croux, C., Filzmoser, P., Pison, G. and Rousseeuw, P. J. (2003). Fitting multiplicative models by robust alternating regressions. *Statistics and Computing*, **13** (1), 23–36.

Croux, C. and Rousseeuw, P. J. (1992). Time-efficient algorithms for two highly robust estimators of scale. In: *Computational Statistics*. Dodge, Y. and Whittaker, J. (eds.). Vol. 1. Heidelberg: Physica-Verlag, pp. 411–428.

Csörgó, S. and Mielniczuk, J. (1996). The empirical process of a short-range dependent stationary sequence under Gaussian subordination. *Probability Theory and Related Fields*, **104** (1), 15–25.

De la Torre, F. and Black, M. J. (2001). Robust principal component analysis for computer vision. In: *International Conference on Computer Vision*. Vol. 1. Vancouver: IEEE, pp. 362–369.

Dempster, A. P. (1972). Covariance selection. *Biometrics*, **28** (1), 157–175.

Dey, D. K. and Srinivasan, C. (1985). Estimation of a covariance matrix under Stein's loss. *The Annals of Statistics*, **13** (4), 1581–1591.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Eltahir, E. A. B. and Wang, G. (1999). Nilometers, El Niño, and climate variability. *Geophysical Research Letters*, **26** (4), 489–492.

Ercan, A., Kavvas, M. L. and Abbasov, R. K. (2013). *Long-Range Dependence and Sea Level Forecasting*. Cham: Springer International Publishing, pp. 7–10.

Feng, X., He, X. and Hu, J. (2011). Wild bootstrap for quantile regression. *Biometrika*, **98** (4), 995–999.

Filzmoser, P., Ruiz-Gazen, A. and Thomas-Agnan, C. (2014). Identification of local multivariate outliers. *Statistical Papers*, **55** (1), 29–47.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London: Series A*, **222** (594-604), 309–368.

Franzke, C. L. E., Graves, T., Watkins, N. W., Gramacy, R. B. and Hughes, C. (2012). Robustness of estimators of long-range dependence and self-similarity under non-Gaussianity. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **370** (1962), 1250–1267.

Friedman, J. H., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9** (3), 432–441.

Gentle, J. (2007). *Matrix Algebra Theory, Computations and Applications in Statistics*. New York: Springer.

Genton, M. G. and Ma, Y. (1999). Robustness properties of dispersion estimators. *Statistics & Probability Letters*, **44** (4), 343–350.

Ghosh, M., Parr, W. C., Singh, K. and Babu, G. J. (1984). A note on bootstrapping the sample median. *The Annals of Statistics*, **12** (3), 1130–1135.

Giraitis, L., Koul, H. L. and Surgailis, D. (2012). *Large Sample Inference for Long Memory Processes*. London: Imperial College Press.

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

Granger, C. W. J. and Joyeux, R. (1980). An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis*, **1** (1), 15–29.

Gupta, M. and Srivastava, S. (2010). Parametric Bayesian estimation of differential entropy and relative entropy. *Entropy*, **12** (4), 818–843.

Gutenbrunner, C. and Jurečková, J. (1992). Regression rank scores and regression quantiles. *The Annals of Statistics*, **20** (1), 305–330.

Gutenbrunner, C., Jurečková, J., Koenker, R. and Portnoy, S. (1993). Tests of linear hypotheses based on regression rank scores. *Journal of Nonparametric Statistics*, **2** (4), 307–331.

Hahn, J. (1995). Bootstrapping quantile regression estimators. *Econometric Theory*, **11** (1), 105–121.

Hall, P. and Welsh, A. H. (1985). Limit theorems for the median deviation. *Annals of the Institute of Statistical Mathematics*, **37** (1), 27–36.

Hampel, F. R. (1971). A general qualitative definition of robustness. *The Annals of Mathematical Statistics*, **42** (6), 1887–1896.

– (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, **69** (346), 383–393.

Hampel, F. R., Ronchetti, E., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. New York: Wiley.

He, X. and Hu, F. (2002). Markov chain marginal bootstrap. *Journal of the American Statistical Association*, **97** (459), 783–795.

Hendricks, W. and Koenker, R. (1992). Hierarchical spline models for conditional quantiles and the demand for electricity. *Journal of the American Statistical Association*, **87** (417), 58–68.

Higham, N. J. (2002). Computing the nearest correlation matrix – a problem from finance. *IMA Journal of Numerical Analysis*, **22** (3), 329–343.

Hodges, J. L. (1967). Efficiency in normal samples and tolerance of extreme values for some estimates of location. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statististics and Probability*. Le Cam, L. M. and Neyman, J. (eds.). Vol. 1. Berkeley: University of California Press, pp. 163–186.

Hodges, J. L. and Lehmann, E. L. (1963). Estimates of location based on rank tests. *The Annals of Mathematical Statistics*, **34** (2), 598–611.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, **19** (3), 293–325.

Hojo, T. (1931). Distribution of the median, quartiles and interquartile distance in samples from a normal population. *Biometrika*, **23** (3-4), 315–363.

Hosking, J. R. M. (1981). Fractional differencing. *Biometrika*, **68** (1), 165–176.

Hössjer, O. and Mielniczuk, J. (1995). Delta method for long-range dependent observations. *Journal of Nonparametric Statistics*, **5** (1), 75–82.

Hsieh, C.-J., Sustik, M. A., Dhillon, I. S. and Ravikumar, P. K. (2011). Sparse inverse covariance matrix estimation using quadratic approximation. In: *Advances in Neural Information Processing Systems*. Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F. and Weinberger, K. (eds.). Vol. 24, pp. 2330–2338.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, **35** (1), 73–101.

Huber, P. J. (1981). *Robust Statistics*. Wiley series in probability and mathematical statistics. New York: Wiley.

Huber, P. J. and Ronchetti, E. (2009). *Robust Statistics*. 2nd ed. New York: Wiley.

Hubert, M. and Debruyne, M. (2009). Breakdown value. *Wiley Interdisciplinary Reviews: Computational Statistics*, **1** (3), 296–302.

Hubert, M., Rousseeuw, P. J. and Vakili, K. (2014). Shape bias of robust covariance estimators: an empirical study. *Statistical Papers*, **55** (1), 15–28.

Hubert, M., Rousseeuw, P. J. and Van Aelst, S. (2008). High-breakdown robust multivariate methods. *Statistical Science*, **23** (1), 92–119.

Hurst, H. E. (1951). Long term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers*, **116**, 770–799.

Hyndman, R. J. and Fan, Y. (1996). Sample quantiles in statistical packages. *American Statistician*, **50** (4), 361–365.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In: *Proceedings of the Fourth Berkeley Symposium on Mathematical Statististics and Probability*. Neyman, J. (ed.). Vol. 1. Berkeley: University of California Press, pp. 361–379.

Janssen, P., Serfling, R. and Veraverbeke, N. (1984). Asymptotic normality for a general class of statistical functions and applications to measures of spread. *The Annals of Statistics*, **12** (4), 1369–1379.

Johnson, D. B. and Mizoguchi, T. (1978). Selecting the Kth element in $X + Y$ and $X_1 + X_2 + \ldots + X_m$. *SIAM Journal on Computing*, **7** (2), 147–153.

Kafadar, K. (1982). A biweight approach to the one-sample problem. *Journal of the American Statistical Association*, **77** (378), 416–424.

Kent, J. T., Tyler, D. E. and Vard, Y. (1994). A curious likelihood identity for the multivariate t-distribution. *Communications in Statistics – Simulation and Computation*, **23** (2), 441–453.

Kocherginsky, M. and He, X. (2007). Extensions of the Markov chain marginal bootstrap. *Statistics & Probability Letters*, **77** (12), 1258–1268.

Kocherginsky, M., He, X. and Mu, Y. (2005). Practical confidence intervals for regression quantiles. *Journal of Computational and Graphical Statistics*, **14** (1), 41–55.

Koenker, R. (1994). Confidence intervals for regression quantiles. In: *Proceedings of the Fifth Prague Symposium on Asymptotic Statistics*. Mandl, P. and Huskova, M. (eds.). Heidelberg: Physica-Verlag, pp. 349–359.

Koenker, R. (2005). *Quantile Regression*. Cambridge: Cambridge University Press.

– (2013). *quantreg: Quantile Regression*. R package version 5.05.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, **46** (1), 33–50.

Koenker, R. and Machado, J. A. F. (1999). Goodness of fit and related inference processes for quantile regression regression. *Journal of the American Statistical Association*, **94** (448), 1296–1310.

Kotz, S. (2004). *Multivariate t Distributions and their Applications*. Cambridge: Cambridge University Press.

Lauritzen, S. (1996). *Graphical Models*. New York: Oxford University Press.

Lax, D. (1985). Robust estimators of scale: Finite-sample performance in long-tailed symmetric distributions. *Journal of the American Statistical Association*, **80** (391), 736–741.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, **88** (2), 365–411.

Lévy-Leduc, C., Boistard, H., Moulines, E., Taqqu, M. S. and Reisen, V. A. (2011a). Asymptotic properties of *U*-processes under long-range dependence. *The Annals of Statistics*, **39** (3), 1399–1426.

– (2011b). Large sample behaviour of some well-known robust estimators under long-range dependence. *Statistics*, **45** (1), 59–71.

– (2011c). Robust estimation of the scale and of the autocovariance function of Gaussian short- and long-range dependent processes. *Journal of Time Series Analysis*, **32** (2), 135–156.

Lin, S. P. and Perlman, M. D. (1985). A Monte Carlo comparison of four estimators of a covariance matrix. In: *Proceedings of the Sixth International Symposium on Multivariate Analysis*. Krishnaiah, P. R. (ed.). Vol. 6. Amsterdam: North-Holland, pp. 411–429.

Little, R. and Rubin, D. (2002). *Statistical Analysis with Missing Data*. 2nd ed. Hoboken: Wiley.

Liu, L., Hawkins, D. M., Ghosh, S. and Young, S. S. (2003). Robust singular value decomposition analysis of microarray data. *Proceedings of the National Academy of Sciences of the United States of America*, **100** (23), 13167–13172.

Løland, A., Huseby, R. B., Hjort, N. L. and Frigessi, A. (2013). Statistical corrections of invalid correlation matrices. *Scandinavian Journal of Statistics*, **40** (4), 807–824.

Luenberger, D. (1969). *Optimization by Vector Space Methods*. New York: Wiley.

Ma, Y. and Genton, M. G. (2000). Highly robust estimation of the autocovariance function. *Journal of Time Series Analysis*, **21** (6), 663–684.

– (2001). Highly robust estimation of dispersion matrices. *Journal of Multivariate Analysis*, **78** (1), 11–36.

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979). *Multivariate Analysis*. Probability and Mathematical Statistics. London: Academic Press.

Maronna, R. A., Martin, R. D. and Yohai, V. J. (2006). *Robust Statistics*. London: Wiley.

Maronna, R. A. and Yohai, V. J. (2008). Robust low-rank approximation of data matrices with elementwise contamination. *Technometrics*, **50** (3), 295–304.

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

Martin, R. D. and Zamar, R. H. (1989). Asymptotically min-max bias robust M-estimates of scale for positive random variables. *Journal of the American Statistical Association*, **84** (406), 494–501.

– (1993). Bias robust estimation of scale. *The Annals of Statistics*, **21** (2), 991–1017.

Mazumder, S. and Serfling, R. (2009). Bahadur representations for the median absolute deviation and its modifications. *Statistics & Probability Letters*, **79** (16), 1774–1783.

Meenakshi, A. and Rajian, C. (1999). On a product of positive semidefinite matrices. *Linear Algebra and its Applications*, **295** (1–3), 3–6.

Morgenthaler, S. (2007). A survey of robust statistics. *Statistical Methods & Applications*, **15** (3), 271–293.

Park, C. et al. (2011). Long-range dependence analysis of internet traffic. *Journal of Applied Statistics*, **38** (7), 1407–1433.

Parzen, M. I., Wei, L. J. and Ying, Z. (1994). A resampling method based on pivotal estimating functions. *Biometrika*, **81** (2), 341–350.

Pipiras, V. and Taqqu, M. S. (2010). Regularization and integral representations of Hermite processes. *Statistics & Probability Letters*, **80** (23–24), 2014–2023.

Portnoy, S. (2012). Nearly root-n approximation for regression quantile processes. *The Annals of Statistics*, **40** (3), 1714–1736.

Randal, J. A. (2008). A reinvestigation of robust scale estimation in finite samples. *Computational Statistics & Data Analysis*, **52** (11), 5014–5021.

Randal, J. A. and Thomson, P. J. (2004). Maximum likelihood estimation for Tukey's three corners. *Computational Statistics & Data Analysis*, **46** (4), 677–687.

Resnick, S. I. (1987). *Extreme Values, Regular Variation, and Point Processes*. New York: Springer.

Rogers, W. H. and Tukey, J. W. (1972). Understanding some long-tailed symmetrical distributions. *Statistica Neerlandica*, **26** (3), 211–226.

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

Rousseeuw, P. J., Croux, C. et al. (2013). *robustbase: Basic Robust Statistics*. R package version 0.9-10.

Rousseeuw, P. J. and Molenberghs, G. (1993). Transformation of non positive semidefinite correlation matrices. *Communications in Statistics – Theory and Methods*, **22** (4), 965–984.

Serfling, R. (1984). Generalized *L*-, *M*-, and *R*-statistics. *The Annals of Statistics*, **12** (1), 76–86.

– (2002). Robust estimation via generalized *L*-statistics: theory, applications, and perspectives. In: *Advances on Methodological and Applied Aspects of Probability and Statistics*. Balakrishnan, N. (ed.). New York: Taylor & Francis, pp. 197–217.

Silverman, B. W. (1976). Limit theorems for dissociated random variables. *Advances in Applied Probability*, **8** (4), 806–819.

Song, S., Ritov, Y. and Härdle, W. K. (2012). Bootstrap confidence bands and partial linear quantile regression. *Journal of Multivariate Analysis*, **107**, 244–262.

Stein, C. (1956). *Some Problems in Multivariate Analysis, Part I*. Tech. rep. 6. Stanford: Stanford University.

Stigler, S. (1977). Do robust estimators work with real data? With discussion and a reply by the author. *The Annals of Statistics*, **5** (6), 1055–1098.

Taqqu, M. S. (1975). Weak convergence to fractional Brownian motion and to the Rosenblatt process. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **31** (4), 287–302.

– (2003). Fractional brownian motion and long-range dependence. In: *Theory and Applications of Long-Range Dependence*. Doukhan, P., Oppenheim, G. and Taqqu, M. S. (eds.). Boston: Birkhäuser, pp. 5–38.

– (2011). The Rosenblatt process. In: *Selected works of Murray Rosenblatt*. Davis, R. A., Lii, K.-S. and Politis, D. N. (eds.). Selected Works in Probability and Statistics. New York: Springer, pp. 29–45.

Taqqu, M. S. and Veillette, M. S. (2013). Properties and numerical evaluation of the Rosenblatt distribution. *Bernoulli*, **19** (3), 982–1005.

Tarr, G. (2012). Small sample performance of quantile regression confidence intervals. *Journal of Statistical Computation and Simulation*, **82** (1), 81–94.

Tarr, G., Müller, S. and Weber, N. C. (2012). A robust scale estimator based on pairwise means. *Journal of Nonparametric Statistics*, **24** (1), 187–199.

Toussoun, O. (1925). Mémoire sur l'histoire du Nil. In: *Mémoire de l'Institut d'Egypte*. Vol. 18, pp. 366–404.

Van Aelst, S., Willems, G. and Zamar, R. H. (2013). Robust and efficient estimation of the residual scale in linear regression. *Journal of Multivariate Analysis*, **116**, 278–296.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer Series in Statistics. New York: Springer.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. 4th ed. New York: Springer.

Welsh, A. H. (1986). Bahadur representations for robust scale estimators based on regression residuals. *The Annals of Statistics*, **14** (3), 1246–1251.

Welsh, A. H. and Morrison, H. L. (1990). Robust L-estimation of scale with an application in astronomy. *Journal of the American Statistical Association*, **85** (411), 729–743.

Wendler, M. (2011). Bahadur representation for U-quantiles of dependent data. *Journal of Multivariate Analysis*, **102** (6), 1064–1079.

– (2012). U-processes, U-quantile processes and generalized linear statistics of dependent data. *Stochastic Processes and their Applications*, **122** (3), 787–807.

Whitcher, B., Byers, S. D., Guttorp, P. and Percival, D. B. (2002). Testing for homogeneity of variance in time series: Long memory, wavelets, and the Nile River. *Water Resources Research*, **38** (5), 1054–1070.

Wilks, S. S. (1932). Certain generalizations in the analysis of variance. *Biometrika*, **24** (3-4), 471–494.

Won, J.-H., Lim, J., Kim, S.-J. and Rajaratnam, B. (2013). Condition-number-regularized covariance estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **75** (3), 427–450.

Wu, M. and Zuo, Y. (2008). Trimmed and Winsorized standard deviations based on a scaled deviation. *Journal of Nonparametric Statistics*, **20** (4), 319–335.

Wu, M. and Zuo, Y. (2009). Trimmed and Winsorized means based on a scaled deviation. *Journal of Statistical Planning and Inference*, **139** (2), 350–365.

Wu, W. B. (2005). On the Bahadur representation of sample quantiles for dependent sequences. *The Annals of Statistics*, **33** (4), 1934–1963.

Yang, R. and Berger, J. O. (1994). Estimation of a covariance matrix using the reference prior. *The Annals of Statistics*, **22** (3), 1195–1211.

Yatracos, Y. G. (1991). A note on Tukey's polyefficiency. *Biometrika*, **78** (3), 702–703.

Yohai, V. J. and Zamar, R. H. (1988). High breakdown-point estimates of regression by means of the minimization of an efficient scale. *Journal of the American Statistical Association*, **83** (402), 406–413.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, **94** (1), 19–35.