An Extension of a Method of Hardin and Rocke, with an Application to Multivariate Outlier Detection via the IRMCD Method of Cerioli

Christopher G. Green[1] and Doug Martin[2]

Working Paper

Revision date: **July 23, 2017**

Abstract

Hardin and Rocke investigated the distribution of the robust Mahalanobis squared distance (RSD) computed using the minimum covariance determinant (MCD) estimator. They showed that the distribution of RSDs for outlying observations not part of the MCD subset is well-approximated by an F distribution. They developed a methodology to adjust an asymptotic formula for the degrees of freedom parameters of this F distribution to provide correct parameter values in small-to-moderate samples. This methodology was developed for the maximum breakdown point version of the MCD, which is based on approximately half of the observations. Whether the approximation remains accurate for the MCD using larger subsets of the data is an open question. We show that their approximation works quite well for the more general MCD, but can be noticeably inaccurate for sample sizes less than 250 and when the MCD estimate uses nearly all of the observations. Motivated by the desire to apply RSD-based outlier detection tests to financial asset return and factor exposure data sets whose typical sample sizes are smaller than 250, we develop a more general correction procedure that is accurate across a wider range of sample sizes and MCD subset sizes than the Hardin and Rocke approach. We use our approach to extend Cerioli's IRMCD procedure for accurate RSD-based outlier tests to arbitrary MCD subset sizes.

Keywords: outlier detection, Mahalanobis distances, minimum covariance determinant, robust estimation

[1]E-mail address for correspondence: `christopher.g.green@gmail.com`
[2]Department of Applied Mathematics, Box 353925, Univerity of Washington, Seattle, WA 98195-3925, USA.

1 Introduction

Detection and mitigation of outliers in multivariate data remains a challenging problem. A common method of detecting outliers in multivariate data is through the use of Mahalanobis distances. Mahalanobis distances, introduced in Mahalanobis (1936), measure the distance of an observation from the mean of a distribution, weighted by the correlation information contained in the covariance matrix (Seber, 1984). If $\mathbf{x}$ is an observation from a multivariate distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, the Mahalanobis squared distance (MSD) of $\mathbf{x}$ from $\boldsymbol{\mu}$ is defined as

$$D^2 \equiv (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}). \tag{1}$$

When $\mathbf{x}$ is ν-dimensional multivariate normal with known mean and covariance, the population MSD is distributed as a chi-squared χ^2_ν random variable with ν degrees of freedom (Mardia et al., 1979; Seber, 1984). This suggests a test of deviation from the multivariate normal assumption: compare an observation's MSD to an appropriate quantile of the chi-squared distribution. An observation may be an outlier if its associated value of D^2 is larger than some critical threshold derived from the distribution of D^2.

In common practice the unknown mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ are replaced by their classical estimates $\hat{\boldsymbol{\mu}} = \overline{\mathbf{x}}$, the coordinate-wise sample mean, and

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T, \tag{2}$$

the sample covariance matrix. When the $\mathbf{x}_i$ are multivariate normal, the resulting sample MSDs are approximately chi-squared for "moderate" values of n, but in higher dimensions larger sample sizes are needed for the approximation to be reasonably accurate. Small (1978) shows that in dimension $\nu = 4$, the chi-squared approximation is noticeably inaccurate in sample sizes as small as $n = 100$. Gnanadesikan and Kettenring (1972) showed (using an earlier result of Wilks (1962)) that the exact distribution of the sample MSDs in this situation is a scaled Beta distribution. In practice, however, the chi-squared approximation is used, either for simplicity or due to a lack of awareness that the accuracy of the approximation depends on the dimension of the data.

Since the classical covariance estimator (2) is not robust to outliers (see, for instance, Maronna et al. (2006)), using it in the Mahalanobis distance metric could lead to some good observations being flagged as outliers (known as "swamping" in the literature) (Rousseeuw and van Zomeren, 1990, 1991; Becker and Gather, 1999; Peña and Prieto, 2001). Moreover, when there are multiple outliers, the classical Mahalanobis distance metric may lead to "masking" of moderate outliers by one extreme outlier (Pearson and Chandra Sekar, 1936; Rocke and Woodruff, 1996). This suggests replacing the sample mean and covariance estimate in Equation (1) with estimates of location and dispersion that are robust to outliers. We will refer to the resulting distance metric as the robust Mahalanobis squared distance (RSD). The robust estimates downweight or ignore the outliers, and thus provide a better representation of the location and dispersion of the majority of the data. Non-outlying points should hence be closer to the location estimate than outlying points, and outlying points should have larger distances than expected under the multivariate normal model.

It remains to calculate an approximate sampling distribution for RSDs in order to identify these outliers. Unfortunately, determining appropriate critical values for the Mahalanobis distance test is more challenging in the robust case than in the classical case. The exact finite-sample distribution is not known for any of the common robust dispersion estimates. The distributional assumption used to test the distances in the

classical case, namely that the distances are independent and identically distributed (IID) chi-squared χ^2_ν random variables, only holds asymptotically in the robust case when the dispersion estimate is consistent for $\mathbf{\Sigma}$ (Mardia et al., 1979; Serfling, 1980; Seber, 1984). As we discuss below, the sample sizes needed to justify using the asymptotic approximation increase as the dimension of the data increases.

The problem of calculating good approximations to the sampling distribution of RSDs has been studied most extensively for the minimum covariance determinant (MCD) estimator introduced by Rousseeuw (1985). Briefly, for $0 < \gamma < 1/2$, the MCD(γ) dispersion estimate is the sample covariance of the subset of $h \approx (1-\gamma)n$ observations whose covariance matrix has the smallest determinant, over all possible h element subsets of the n observations. For the MCD estimate, it is known that using χ^2_ν quantiles for critical values can lead to many more false positives than expected in small to moderate samples, especially when the data set actually does not contain any outliers (Rousseeuw and van Zomeren, 1991; Becker and Gather, 2001). In fact, Cerioli et al. (2009) found that the use of the χ^2_ν approximation leads to a serious problem for MCD-based distance tests for outlyingness: the realized false positive rates of the tests can be substantially larger than the nominal false positive rates even in moderate sample sizes.

Cerioli et al. (2009) looked at how well MCD-based Mahalanobis distances performed both in an individual testing framework ("is this observation an outlier?") and under a simultaneous testing framework ("are there any outliers in the data?"). First they conducted a simulation experiment in which each observation was tested for outlyingness at some nominal test size (say, $\alpha = 0.01$). We expect to see about $\lfloor \alpha n \rfloor$ incorrectly flagged observations on average. Their simulations show this is not the case for the MCD with χ^2_ν critical values. Testing MCD-based distances against χ^2_ν critical values requires large sample sizes to be reliably accurate, with the needed sample size increasing with dimension ν. For small to moderate sample sizes the χ^2_ν critical values can give significantly more false positives than expected based on the nominal test size: in dimension $\nu = 10$ the average false positive rate is about 5 times too large for $n = 200$, and about 13 times too large for $n = 100$. (Further details are available in Appendix A of Green (2017).)

Cerioli et al. (2009) then looked at the accuracy of tests of the intersection null hypothesis

$$H_0 : \{\mathbf{x}_1 \sim N(\boldsymbol{\mu}, \mathbf{\Sigma})\} \cap \cdots \cap \{\mathbf{x}_n \sim N(\boldsymbol{\mu}, \mathbf{\Sigma})\} \tag{3}$$

that examines whether there are any outliers in the data (as opposed to whether a given observation is outlying). The obvious way to perform this test is via comparison of the largest RSD in the set of observations to an appropriate quantile at a Bonferroni-corrected size α/n. The quantile could come from the χ^2_ν distribution, as done in Becker and Gather (1999, 2001), or the scaled F distribution derived by Hardin and Rocke (2005). Again via a simulation study, Cerioli et al. showed that the χ^2_ν quantile works poorly for testing the intersection hypothesis with the maximum breakdown point case of the MCD, with false positive rates 50–100 times too large for small samples in dimension $\nu = 10$. Subsequently, Cerioli (2010) developed a methodology, the Iterated Reweighted MCD (IRMCD), that yields RSD-based tests for outliers with the correct false positive rates for both the individual and intersection tests. Cerioli's approach (described in Section 4.3) works for the MCD estimator and relies upon the distributional approximation developed by Hardin and Rocke (2005).

For financial applications, however, we would not want to use the maximum-breakdown point case of MCD, as it discards nearly half of the data to compute the estimate. We would recommend that a practitioner use the MCD with 90% or more (i.e., $\gamma \leq 0.10$) of the observations, depending on the sample size. This choice of trimming would only exclude extreme outliers from the estimate. Although Cerioli (2010) presents tests of the IRMCD methodology for MCD(0.25), the methodology depends on the distributional approximation

developed by Hardin and Rocke (2005). That distributional approximation uses a correction developed only for the maximum-breakdown point case of MCD. We were not aware of any studies examining how well the Hardin-Rocke correction works for the more general version of the MCD, so we conducted simulations to test the accuracy of the approximation outside of its original design parameters. We found that the Hardin-Rocke approximation works well in moderate-to-large ($n > 500$) samples for the general version of the MCD, but that it is unreliable in smaller samples and/or when 90% or more of the data is used to compute the estimate. Thus, in order to use IRMCD safely for the MCD in general, we developed an improved approximation for the distribution of MCD-based RSDs for outlying points. We show our correction methodology is more accurate than the Hardin-Rocke approach for MCD(γ) for γ as small as 0.005. We validate our approach using simulated data and via tests of the IRMCD approach.

The remainder of the paper is organized as follows. Section 2 reviews technical details on the MCD estimate, the Hardin-Rocke distributional approximation, and Cerioli's IRMCD procedure. Section 3 describes the Hardin-Rocke method for estimating the Wishart degrees of freedom parameter needed to use their distribution approximation, and describes our improved method that is more accurate than the Hardin-Rocke method for a wide range of sample sizes, dimensions, and trimming fractions. Section 4 presents several tests of our model. Section 5 concludes with a discussion of potential future improvements.

2 Technical Background

2.1 The MCD Estimate

Rousseeuw (1985) introduced the minimum covariance determinant (MCD) robust dispersion estimate. Given n observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ of dimension ν and a subset of size $h \leq n$, the (non-reweighted or raw) *MCD subset* of the observations is defined by a set of indices $\{j_1, \ldots, j_h\}$ such that the determinant of the sample covariance of the observations $\mathbf{x}_{j_1}, \ldots, \mathbf{x}_{j_h}$ is minimal over all subsets of observations of size h:

$$\det \hat{\mathbf{\Sigma}} \left(\mathbf{x}_{j_1}, \ldots \mathbf{x}_{j_h} \right) \leq \det \hat{\mathbf{\Sigma}} \left(\mathbf{x}_{k_1}, \ldots \mathbf{x}_{k_h} \right),$$

for any subset $\{k_1, \ldots, k_h\}$ of $\{1, \ldots, n\}$ with cardinality h and satisfying $1 \leq k_1 < \cdots < k_h \leq n$. The *MCD estimate* of the dispersion matrix of the data is then the sample covariance matrix S_{MCD} of the MCD subset, and the MCD estimate of the location vector is the sample mean $\overline{X}_{MCD}$ of the MCD subset.

Croux and Haesbroeck (1999) demonstrate that the efficiency of the raw MCD is rather low for the maximum breakdown point case, especially in small dimensions. Cerioli therefore uses a reweighted MCD in his IRMCD procedure. Reweighting the observations using the raw MCD estimate can increase the efficiency of the estimate while preserving its breakdown point (Lopuhaä, 1999; Croux and Haesbroeck, 1999). A "reweighted" MCD is calculated by computing the "raw" MCD based on the given observations and then excluding observations based on their RSD (using χ^2_ν critical values). The reweighted MCD estimate is then the classical mean and covariance of the remaining observations.

The MCD is computationally difficult because it involves a combinatorial optimization problem. In practice most MCD implementations actually compute an approximate solution by optimizing over a random subset of all possible size-h subsets of the n observations. Rousseeuw and van Driessen (1999) developed the *fastMCD* algorithm based upon this idea. The *fastMCD* algorithm is used in the `covMcd` function in the `R` package `robustbase` and is used in all calculations below.

Although we have defined the MCD in terms of the number of observations h used to compute the

estimate, it is often convenient to think of the MCD in terms of the asymptotic fraction γ, $0 < \gamma < 1/2$, of the data trimmed from the MCD estimate, as this fraction controls its properties such as its breakdown point and efficiency. In the `R` function `covMcd` implementing the MCD, one specifies $1 - \gamma$, the asymptotic fraction of observations used in the MCD, as an input parameter. The value h is then computed from $1 - \gamma$ as

$$\begin{aligned} h &= \lfloor 2n_2 - n + 2(n - n_2)(1 - \gamma) \rfloor \\ &= \lfloor (2n_2 - n)\gamma + n(1 - \gamma) \rfloor , \end{aligned} \tag{4}$$

where

$$n_2 = \left\lfloor \frac{n + \nu + 1}{2} \right\rfloor .$$

If n is even, then

$$n_2 = \frac{n}{2} + \left\lfloor \frac{\nu + 1}{2} \right\rfloor ,$$

and, after a bit of algebra, we have

$$\begin{aligned} h &= \left\lfloor n - \left(n - 2 \left\lfloor \frac{\nu + 1}{2} \right\rfloor \right) \gamma \right\rfloor \\ &= n - \left\lceil \left(n - 2 \left\lfloor \frac{\nu + 1}{2} \right\rfloor \right) \gamma \right\rceil . \end{aligned} \tag{5}$$

Similarly if n is odd we can show that

$$h = n - \left\lceil \left(n - 2 \left\lfloor \frac{\nu + 1}{2} \right\rfloor - 1 \right) \gamma \right\rceil .$$

When $n \gg \nu$, the quantity $1 - h/n$ will be approximately equal to γ, so that $h \approx (1 - \gamma)n$ and the MCD estimate trims approximately $n\gamma$ observations. This motivates our use of γ as an approximate or asymptotic "trimming fraction" (Maechler, 2016).

The definition (4) ensures that in smaller samples the value of h computed using (4) will be strictly smaller than n, even if γ is very small. In the n even case, rearranging (5) yields

$$n - h = \left\lceil \left(n - 2 \left\lfloor \frac{\nu + 1}{2} \right\rfloor \right) \gamma \right\rceil . \tag{6}$$

The right hand side will not vanish unless $\gamma = 0$ or $n = 2 \left\lfloor \frac{\nu+1}{2} \right\rfloor$. The MCD is not recommended in situations where $n < 2\nu$, so the latter situation never occurs provided one follows this recommendation. Thus we have $h < n$ for any non-degenerate case of MCD(γ).

The number of observations $n - h$ not used in the MCD subset can still be quite different from $n\gamma$, however, when γ is small and/or n is small. For example, suppose again that n is even and that $\gamma = 1/N$

for an integer N. Plugging $\gamma = 1/N$ into (6) yields

$$n - h = \left\lceil \frac{\left(n - 2\left\lfloor \frac{\nu+1}{2} \right\rfloor\right)}{N} \right\rceil.$$

For $1 + 2\left\lfloor \frac{\nu+1}{2} \right\rfloor \leq n \leq N + 2\left\lfloor \frac{\nu+1}{2} \right\rfloor$, the right-hand side of this equation will be equal to 1, i.e., MCD($1/N$) will exclude exactly 1 point. Again, in practice we would not use the MCD with $n < 2\nu$, so a more practical range is

$$\max\left\{1 + 2\left\lfloor \frac{\nu+1}{2} \right\rfloor, 2\nu\right\} \leq n \leq N + 2\left\lfloor \frac{\nu+1}{2} \right\rfloor.$$

This range depends on the dimension ν and the value of $\gamma = 1/N$, and is much larger for smaller values of γ (i.e., larger values of N). For example, for $\gamma = 0.25 = 1/4$ and $\nu = 2$, the range is $4 \leq n \leq 6$, while for $\nu = 20$ the range will be empty since there are no even $n > 40$ that satisfy the condition above when $N = 4$. For $\gamma = 0.01 = 1/100$ and $\nu = 2$, we will have $n - h = 1$ when $4 \leq n \leq 102$. When $\nu = 20$ the corresponding range is $40 \leq n \leq 120$.

We thus emphasize that γ is an asymptotic trimming fraction. In the remainder of this paper, we will denote the MCD estimate based on the asymptotic fraction $1 - \gamma$ of the observations by MCD(γ), with the above caveats in mind.

In the most commonly used version of the MCD(γ) estimate, the subsample size is set to $h_{MBP} = \lfloor (n + \nu + 1)/2 \rfloor$, so that $1 - h_{MBP}/n \approx 1/2$ when $n \gg \nu$. With this subsample size the MCD achieves the maximum possible breakdown point of 1/2 for large samples. We will use the notation MCD(γ^*) to refer to the maximum breakdown point case of the MCD.

2.2 The Hardin-Rocke Distributional Approximation

Hardin and Rocke (2005) studied the distribution of (non-reweighted) MCD-based RSDs for the MCD(γ^*) estimator. Their work was motivated by previous studies such as Rousseeuw and van Zomeren (1991) that showed that the χ^2_ν critical values can be too small in sample sizes $n \leq 50$ in dimensions $\nu \leq 4$, resulting in many observations being incorrectly flagged as outliers. Hardin and Rocke established that, when the observations $\mathbf{x}_i$ arise from a ν-dimensional multivariate normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the RSDs for observations not in the MCD subset are approximately independent of the RSDs for the MCD subset, and that the non-MCD subset distances are approximately F distributed rather than χ^2_ν distributed. Their argument rests upon the assumption that the distribution of the scaled MCD(γ^*) estimate dispersion matrix S_{MCD} is well-approximated by a ν-dimensional Wishart distribution:

$$\frac{m}{c} S_{MCD} \sim \text{Wishart}_\nu(m, \boldsymbol{\Sigma}), \tag{7}$$

where ν is the known dimension of the observations, m is an unknown Wishart degrees of freedom parameter and c is an unknown consistency constant. Recall that the sample covariance matrix (2) of n observations from a ν-dimensional multivariate normal distribution follows a scaled ν-dimensional Wishart distribution with $n - 1$ degrees of freedom. The MCD(γ^*) estimate S_{MCD} is the sample covariance of the MCD subset of observations, which is well-modeled by a multivariate normal distribution (assuming the subset does not possess strong non-linear structure). It is thus reasonable to assume S_{MCD} follows a Wishart distribution, but with an unknown degrees of freedom parameter.

Hardin and Rocke then show that the sample RSDs for outlying points are approximately F-distributed

after suitable scaling:

$$\frac{c(m-\nu+1)}{m\nu} D^2_{S_{MCD}}\left(X_i, \bar{X}_{MCD}\right) \sim F_{\nu,m-\nu+1}. \tag{8}$$

This F distribution provides more accurate critical values for testing RSDs than the χ^2_ν distribution.

3 Estimating the Wishart Degrees of Freedom Parameter in the Hardin-Rocke F Distribution

In order to use the distribution (8) for MCD(γ^*) or more generally, MCD(γ), we must determine the parameters c and m. Simulation is the most accurate means of estimating the parameters c and m but obviously not convenient for everyday use of the Hardin-Rocke F distribution. In this section we will review the approach developed by Hardin and Rocke to estimate m for use with the MCD(γ^*). We will then show that their method is inaccurate for small samples $n \leq 250$ and for the more general MCD(γ) with small γ (e.g., $\gamma = 0.05$). Finally, we will develop a better model that works reliably across a wide range of sample sizes, dimensions, and trimming fractions.

3.1 The Hardin-Rocke Adjustment to the Asymptotic Degrees of Freedom

Hardin and Rocke note that if S_{MCD} has the scaled Wishart distribution (7), then its diagonal elements s_{jj} will be distributed as

$$mc^{-1}s_{jj} \sim \sigma_{jj}\chi^2_m,$$

where σ_{jj} are the diagonal elements of $\mathbf{\Sigma}$. The MCD estimate is affine equivariant, so one can assume $\boldsymbol{\mu} = \mathbf{0}$, a vector of zeros, and $\mathbf{\Sigma} = \mathbf{I}$, the identity matrix with $\sigma_{jj} = 1$. Since a χ^2_m random variable has mean m and variance $2m$, we can use the method of moments to estimate m.

$$\begin{aligned} E\left[mc^{-1}s_{jj}\right] &= m \\ \text{Var}\left(mc^{-1}s_{jj}\right) &= 2m \\ CV = \frac{\sqrt{\text{Var}(s_{jj})}}{E(s_{jj})} &= \frac{c\sqrt{2/m}}{c} = \sqrt{\frac{2}{m}} \end{aligned} \tag{9}$$

where CV is the coefficient of variation. Therefore

$$m = \frac{2}{CV^2}. \tag{10}$$

Croux and Haesbroeck (1999) derive the influence function for S_{MCD} in the general MCD(γ) case and use it to calculate the asymptotic variance of S_{MCD}. This calculation provides asymptotic formulas for the variance of s_{jj} that can be used to estimate CV, and hence, m in large samples. The Appendix to Hardin and Rocke (2005) summarizes the asymptotic formulas c_{asy} and $m_{\text{asy}}(n, \nu, \gamma)$ for c and m, respectively. We reproduce their formulas again here for the reader's convenience.[3] Here $\gamma \approx 1 - h/n$ is the approximate fraction of observations trimmed by the MCD as in Section 2.1.

[3] Our notation here is slightly different from that of Hardin and Rocke (2005). We use ν to represent the dimension rather than p, and we refer to the fraction of observations trimmed from the MCD as γ rather than α.

The constant $c(\nu, \gamma)$ is defined as

$$c(\nu, \gamma) = \frac{1-\gamma}{P\left(\chi^2_{\nu+2} \le q(\nu, 1-\gamma)\right)},$$

where $q(\nu, 1-\gamma)$ is the $1-\gamma$ quantile of a χ^2_ν distribution and satisfies $1-\gamma = P(\chi^2_\nu \le q(\nu, 1-\gamma))$. The asymptotic consistency constant c_{asy} is defined as the reciprocal of $c(\nu, \gamma)$:[4]

$$c_{\text{asy}} = 1/c(\nu, \gamma). \tag{11}$$

The asymptotic coefficient of variation is given by

$$CV^2_{\text{asy}} = c(\nu, \gamma)^2 v(\nu, \gamma),$$

where $v(\nu, \gamma)$ is the asymptotic variance of the s_{jj}. (The formula for $v(\nu, \gamma)$ is provided in Appendix A.) Thus from (10) we have

$$m_{\text{asy}}(n, \nu, \gamma) = \frac{2}{c(\nu, \gamma)^2 v(\nu, \gamma)}. \tag{12}$$

Our notation reflects that $m_{\text{asy}}(n, \nu, \gamma)$ is actually function of n, ν, and γ, even though Hardin and Rocke only considered the $\gamma = \gamma^*$ case.

Croux and Haesbroeck's formula for c_{asy} is reliable for small samples, but this is not the case for $m_{\text{asy}}(n, \nu, \gamma)$. Thus we need a way to estimate m accurately for small to moderate sample sizes (e.g., $30 \le n \le 250$). Hardin and Rocke estimated the values of m for the MCD(γ^*) estimator via simulation for sample sizes $n = 50, 100, 250, 500, 750, 1000$ and dimensions $\nu = 3, 5, 7, 10, 15, 20$. Their procedure is as follows.

1. Simulate $N = 1000$ random samples of size n from a ν-dimensional multivariate normal $\text{N}(\mathbf{0}, \mathbf{I})$.
2. For each random sample, calculate the MCD(γ^*) estimate S_{MCD}. Retain the ν diagonal elements s_{jj} from each S_{MCD}. There will be a total of $N\nu$ such values from all the simulations.
3. Calculate the estimate $\tilde{c}_{\text{sim}}(n, \nu, \gamma^*)$ of c as the sample mean of the $N\nu$ s_{jj} values.
4. Calculate the sample variance $\tilde{v}_{\text{sim}}(n, \nu, \gamma^*)$ of the $N\nu$ s_{jj} and use it to calculate an estimate $\widetilde{CV}_{\text{sim}}(n, \nu, \gamma^*)^2$ of the coefficient of variation.
5. Calculate an estimate $\tilde{m}_{\text{sim}}(n, \nu, \gamma^*)$ of m using (10) as

$$\tilde{m}_{\text{sim}}(n, \nu, \gamma^*) = \frac{2}{\widetilde{CV}_{\text{sim}}(n, \nu, \gamma^*)} = \frac{2\tilde{c}_{\text{sim}}(n, \nu, \gamma^*)^2}{\tilde{v}_{\text{sim}}(n, \nu, \gamma^*)}.$$

Obviously $\tilde{m}_{\text{sim}}(n, \nu, \gamma^*)$ is a function of n and ν, but it is also a function of γ in general since the MCD(γ) estimator in Step 2 could be used with with any value of γ.

[4] Different authors define the consistency constant differently, hence the need for an extra constant here.

Hardin and Rocke then fit the following model to the simulated $\tilde{m}_{\text{sim}}(n, \nu, \gamma^*)$ using bivariate least squares regression to estimate the true m from the Croux-Haesbroeck asymptotic $m_{\text{asy}}(n, \nu, \gamma^*)$ for the $\gamma = \gamma^*$ case:

$$\log\left(\frac{\tilde{m}_{\text{sim}}(n, \nu, \gamma^*)}{m_{\text{asy}}(n, \nu, \gamma^*)}\right) = \beta_0 + \beta_1\nu + \beta_2 \log n + \epsilon_{n,\nu}, \qquad \epsilon_{n,\nu} \overset{iid}{\sim} N(0, 1)$$

where ϵ is an error term. They used the 36 values of $\tilde{m}_{\text{sim}}(n, \nu, \gamma^*)$ to compute values of $\log\left(\tilde{m}_{\text{sim}}(n, \nu, \gamma^*)/m_{\text{asy}}(n, \nu, \gamma^*)\right)$, which were then regressed on the corresponding 36 pairs of predictors $(\nu, \log(n))$ for the 6 values of ν and 6 values of n stated above. The final fitted model is

$$\log\left(\frac{m}{m_{\text{asy}}(n, \nu, \gamma^*)}\right) = 0.725 - 0.00663\nu - 0.0780 \log(n). \tag{13}$$

We will refer to the above formula to estimate m from $m_{\text{asy}}(n, \nu, \gamma)$ as the "Hardin-Rocke adjustment".

Hardin-Rocke established via simulation that their method gives more accurate results, in terms of detecting an appropriate number of outliers, for the MCD-based RSD tests than the standard χ^2_ν-based tests. The simulation study of Cerioli et al. (2009) further affirmed that, for sample sizes $n > 100$ and even dimensions up to $\nu = 12$, the Hardin-Rocke quantiles were more accurate for testing individual observations for outlyingness than the χ^2_ν quantiles for the MCD(γ^*) case. Unfortunately, their study also showed that Hardin-Rocke approach can still result in too many false positives for sample sizes $n \leq 100$. There is also the question of how well the Hardin-Rocke adjustment works for small values of values of γ other than γ^*. While the formulas for c_{asy} and $m_{\text{asy}}(n, \nu, \gamma)$ are valid for arbitrary values of γ, Hardin and Rocke's simulated values $\tilde{m}_{\text{sim}}(n, \nu, \gamma)$ were estimated using the MCD(γ^*). It is not clear from the Hardin and Rocke paper how well their approximation (13) works for other fractions γ, nor have we seen any research into this matter.

In the next section we show that the Hardin-Rocke adjustment (13) does not work well for sample sizes less than 250 when $\gamma \in \{0.25, 0.05, 0.01\}$. The ensuing sections will then detail our development of a new model that works more reliably across a larger range of sample sizes, dimensions, and trimming fractions.

3.2 Testing the Hardin-Rocke Adjustment for Other Values of γ

First, we consider how the 0.01 critical value, i.e., the 0.99 quantile, from the Hardin-Rocke scaled F distribution varies with the input parameters m and ν. For dimensions $\nu = 5, 10, 20$ and integer values of m satisifying $\nu \leq m \leq 20\nu$, we calculated the logarithm of the 0.99 quantile of the Hardin-Rocke F distribution given in (8). Figure 1 shows how the logarithm of the 0.99 quantile depends on the Wishart degrees of freedom parameter m for $\nu = 5, 10, 20$. For fixed values of dimension ν, larger values of m lead to smaller quantiles. Thus if we overpredict m, the quantiles of the F distribution will be too small, and we will reject more observations than we should.

Next we examine how well the Hardin-Rocke adjustment (13) estimates the true value of m for γ other than γ^*. We estimated $\tilde{m}_{\text{sim}}(n, \nu, \gamma)$ using a simulation similar to that performed by Hardin and Rocke (described in the previous subsection) but extended to include the MCD(γ) for several values of γ other than γ^* and more coverage of small sample sizes.[5] We simulated $N = 5000$ draws of size n from a multivariate normal distribution $N(\mathbf{0}, \mathbf{I}_\nu)$ with dimensions $\nu = 3, 5, 7, 10, 15, 20$ and sample sizes $n = 50, 100, 250, 500, 750, 1000$. We calculated the MCD($\gamma$) subset of each simulated data set for $0.05 \leq \gamma \leq 0.45$ in increments of 0.05, as well as maximum breakdown point case γ^* and the extreme cases of $\gamma = 0.01$ and $\gamma = 0.005$. In order to understand well how the Hardin-Rocke adjustment worked in small samples, we also included sample sizes

[5] Additional details on the simulation computations are available in Appendix B.

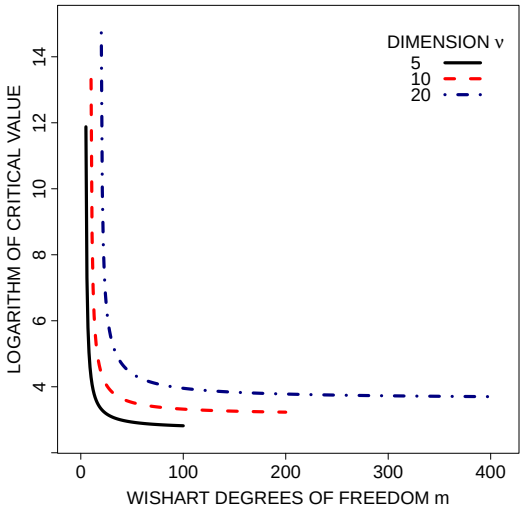


Figure 1: Logarithms of 0.99 quantiles produced from Hardin-Rocke scaled F distribution (vertical axis) as a function of the Wishart degrees of freedom parameter m (horizontal axis). The quantiles are shown for several values of dimension ν (plot symbols and colors).

$n = 3\nu, 5\nu, 7\nu, 9\nu, 11\nu$ for the above dimensions and values of γ.[6] We remind the reader that, as discussed in Section 2.1, γ is an asymptotic trimming fraction. When n is small or γ is small, the number of observations excluded from the MCD(γ) subset can be different from the asymptotic value of $n\gamma$. For example, when $\nu = 3$ and $n = 3\nu = 9$, one observation is excluded from the MCD(0.01) subset, even though the value $\lfloor 0.01 \times 9 \rfloor = 0$ might suggest that no observations will be excluded.

For each simulated data set and each value of γ we calculate the estimate $\tilde{m}_{\text{sim}}(n, \nu, \gamma)$ of the Wishart degrees of freedom m using Hardin and Rocke's simulation procedure (described in the previous section). The consistency constant c is estimated by the asymptotic version c_{asy} (Equation (11)).

We first considered how well the Hardin-Rocke adjustment estimated m for $\gamma < \gamma^*$. Figures 2–4 show, for MCD(γ) with $\gamma = 0.25$, $\gamma = 0.05$, and $\gamma = 0.01$, respectively, the ratio of the Wishart degrees of freedom m estimates obtained from simulation to those obtained from the Hardin-Rocke adjustment to the asymptotic degrees of freedom. The range of sample sizes in each figure is constrained to $n \leq 250$ to highlight the behavior of the Hardin-Rocke adjustment in the smaller sample sizes typically encountered in financial applications, e.g., $n = 60$ (five years of monthly returns) or $n = 252$ (one year of daily returns). We will briefly describe the behavior for $n > 250$ as well, even though this range is not reflected in the figures.

In the $\gamma = 0.25$ case, the Hardin-Rocke adjustment leads to values of m that can be as much as 1.3 times too large for sample sizes smaller than $n = 250$. As the sample size increases beyond $n = 250$, the Hardin-Rocke estimated values of m are closer to the simulation values, with the convergence to equality requiring larger sample sizes in lower dimensions. For the smaller trimming fractions $\gamma = 0.05$ and $\gamma = 0.01$, on the other hand, the Hardin-Rocke adjustment over-estimates m by a factor as large as 2.5. The performance of the adjustment steadily improves with sample size, however. Convergence to equality between the two methods also takes a bit longer with the smaller trimming fractions.

Next we looked at whether the above inaccuracy in estimating m translated into meaningful differences in

[6] We use dimension-dependent sample sizes for small-sample coverage to avoid a subtle problem with fixed sample sizes like $n = 25$: the MCD may be infeasible when $n < 2\nu$. The `R` function `covMcd` will helpfully warn the user about such small sample sizes.

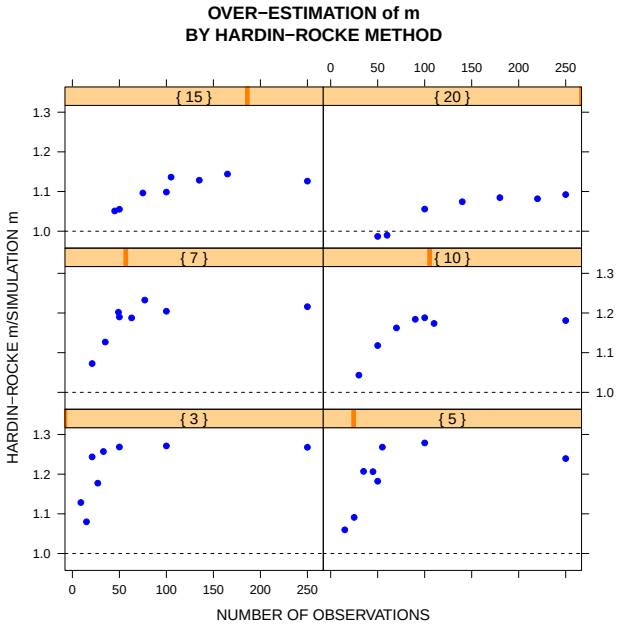


Figure 2: Comparison of Wishart degrees of freedom parameter m estimated via simulation and Hardin-Rocke approach with $\gamma = 0.25$. The ratio of the degrees of freedom parameters coming from the Hardin-Rocke approach to those resulting from the simulation is shown (stratified by dimension ν). Sample size is plotted on the horizontal axis. Sample sizes shown in the plot are the dimension-dependent values $n = 3\nu, 5\nu, 7\nu, 9\nu$, and 11ν (which hence vary between panels), as well as the fixed values $n = 50, 100, 250$. Not shown are ratios for the sample sizes $n = 500, 750, 1000$. The dimension ν for each subgroup is shown in the yellow bars at the top of each subplot.

the critical values for testing RSDs. Figures 5–7 show how the resulting 0.01 critical values computed using Hardin and Rocke's F distribution using the simulated and Hardin-Rocke estimated values of m compare for $\gamma = 0.25$, $\gamma = 0.05$, and $\gamma = 0.01$ respectively. The overprediction of m seen in Figures 2–4 translates into critical values that are smaller than they should be, as we would expect from Figure 1. In small samples $n < 250$ and small dimensions $\nu \leq 5$ the critical values are typically about 80% as large as they should be based on the value of m estimated from the simulation. For the smaller values of γ it takes slightly larger sample sizes for the two methods to produce approximately equal critical values.

Overall we observe that the Hardin-Rocke adjustment (13) is quite accurate for producing 0.01 critical values for sample sizes of at least 250 and $\gamma \in \{0.25, 0.05, 0.01\}$, but can result in critical values that are much too small for sample sizes less than 100 and a bit too small for $100 < n \leq 250$. The inaccuracy is worse for the smaller trimming fractions $\gamma = 0.05$ and $\gamma = 0.01$ compared to the $\gamma = 0.25$ case.[7]

Thus using the Hardin-Rocke adjustment for small values of γ, e.g., $\gamma = 0.05$ or $\gamma = 0.01$, and/or with $n \leq 250$ will result in flagging too many observations as outliers. This is concerning for our intended use of RSD-based outlier tests in financial applications: it is quite common in financial applications to encounter sample sizes $n \leq 100$ (e.g., 2 years of weekly data or 5 years of monthly data), and financial practitioners are often keen to use small values of γ. For financial applications of RSDs it is crucial to have an accurate reference distribution for detecting potential outliers via RSDs in small samples and with small values of γ. Therefore in the next section we develop a more general formula to estimate the true degrees of freedom parameter m from the asymptotic value $m_{\text{asy}}(n, \nu, \gamma)$ that remains accurate across a wider range of sample

[7] We observed similar results for the 0.025 and 0.05 critical values.

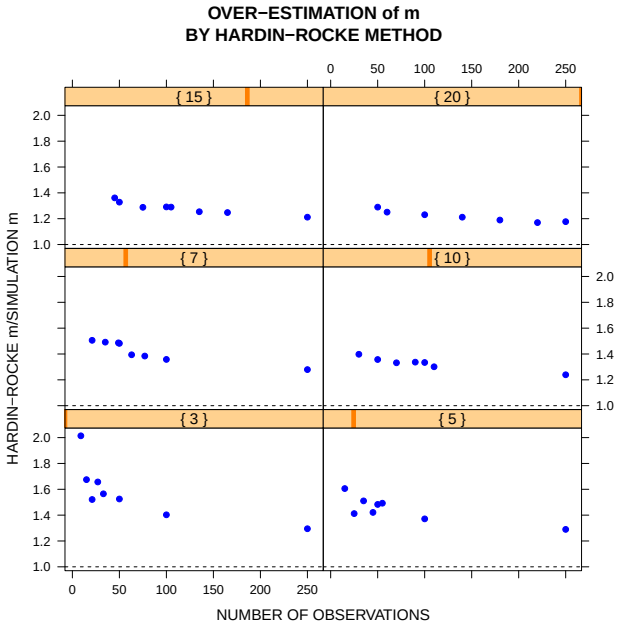


Figure 3: Comparison of Wishart degrees of freedom parameter m estimated via simulation and Hardin-Rocke approach with $\gamma = 0.05$. The plot setup is identical to that of Figure 2.

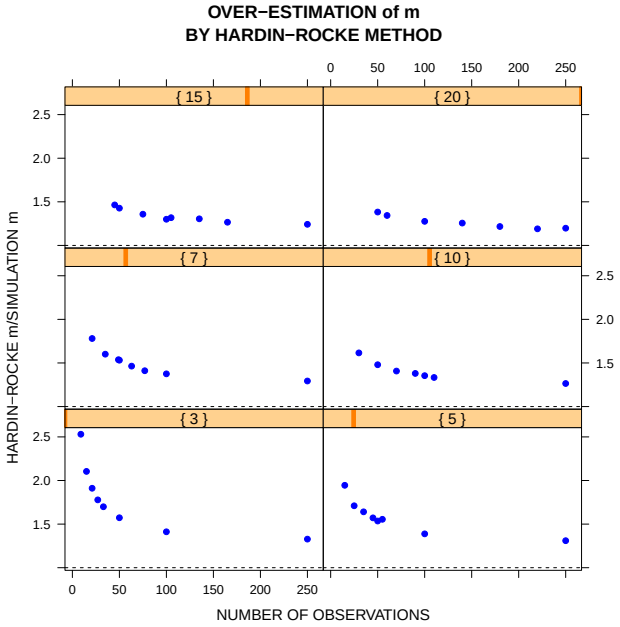


Figure 4: Comparison of Wishart degrees of freedom parameter m estimated via simulation and Hardin-Rocke approach with $\gamma = 0.01$. The plot setup is identical to that of Figure 2.

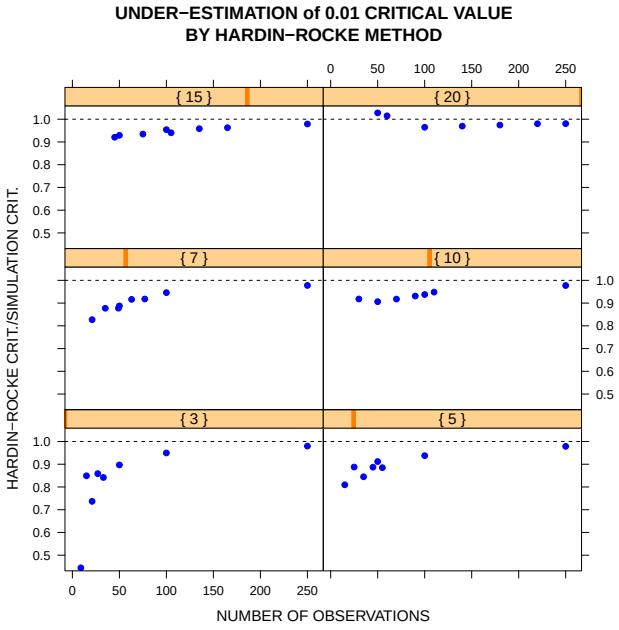


Figure 5: Comparison of 0.01 critical values produced using Wishart degrees of freedom parameter m estimated via simulation and Hardin-Rocke approach with $\gamma = 0.25$. Critical values are calculated using the scaled F distributional approximation of Hardin and Rocke with each degrees of freedom parameter estimate. The ratio of the Hardin-Rocke critical values to those resulting from the simulation is shown (stratified by dimension ν). The dotted line at a ratio of 1 indicates when the two critical values are approximately equal. Sample size is plotted on the horizontal axis. The pattern of sample sizes used here is identical to that used in Figure 2. The dimension ν is shown in the yellow bars at the top of each subplot.

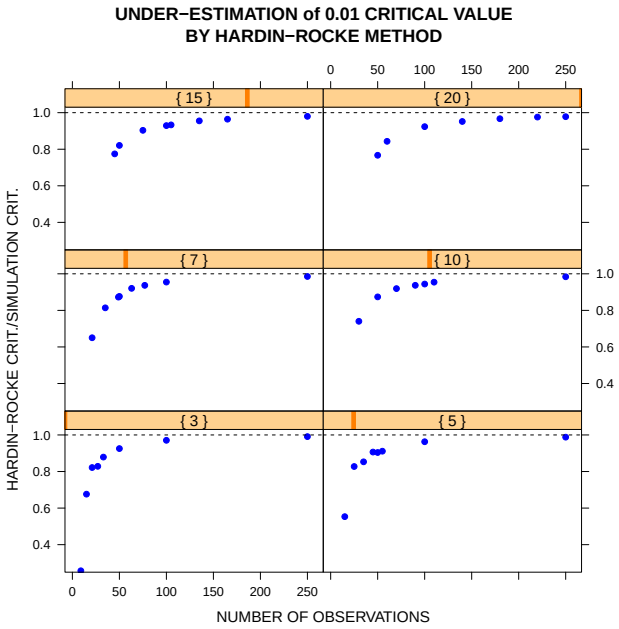


Figure 6: Comparison of 0.01 critical values produced from Wishart degrees of freedom parameter estimated via simulation and Hardin-Rocke approach with $\gamma = 0.05$. The plot setup is identical to that of Figure 5.

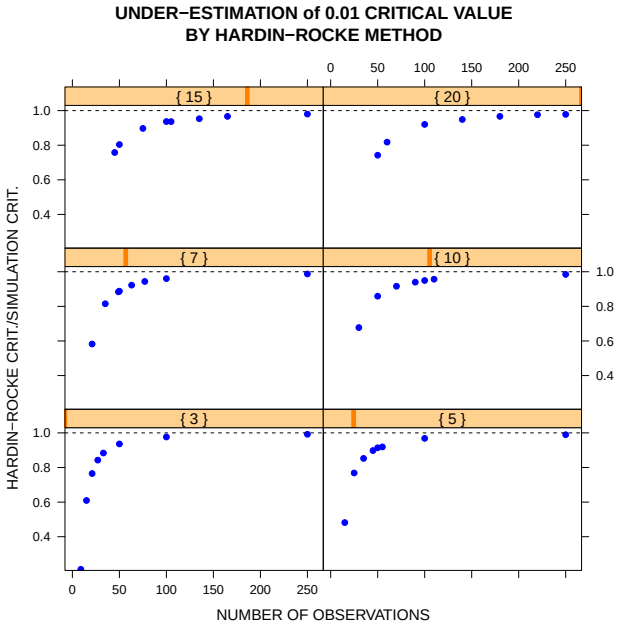


Figure 7: Comparison of 0.01 critical values produced from Wishart degrees of freedom parameter estimated via simulation and Hardin-Rocke approach with $\gamma = 0.01$. The plot setup is identical to that of Figure 5.

sizes n, dimensions ν, and trimming fractions γ.

3.3 An Improved Adjustment to the Asymptotic Degrees of Freedom

We start our search for a better adjustment formula with some exploratory data analysis. Figure 8 shows how the estimated values of $\tilde{m}_{\text{sim}}(n, \nu, \gamma)$ from our simulation compare to the asymptotic values $m_{\text{asy}}(n, \nu, \gamma)$ for varying levels of γ and dimension ν. The plots suggests the log ratio of the true m to $m_{\text{asy}}(n, \nu, \gamma)$ decays inversely with a power of sample size n that depends on $1-\gamma$. This is in sharp contrast to the model used in the Hardin-Rocke adjustment, which posited that the log ratio varied with $\log(n)$ and did not allow for any dependence of m on γ. Furthermore, with respect to the correct dependence of m on n, we know that since the asymptotic formula should approach the true value of m as $n \to \infty$, the quantity $\log(m/m_{\text{asy}}(n, \nu, \gamma))$ should go to zero as $n \to \infty$. In the Hardin-Rocke adjustment, however, $\log(m/m_{\text{asy}}(n, \nu, \gamma))$ goes to $\pm\infty$ as $n \to \infty$, depending on the sign of β_2, the coefficient of $\log(n)$ in (13).

In their analysis, Hardin and Rocke found that the dependence of $\log(m/m_{\text{asy}}(n, \nu, \gamma))$ on the dimension ν was weak. We see that in our data as well, as is evidenced by the stacking of the points in each plot of Figure 8. Finally the sign of the dependence relation changes for $n \leq 100$ when $\gamma \leq 0.1$. Here the MCD(γ) estimator discards very few observations and becomes more like the sample covariance estimator.[8]

Based on the above observations, we propose the following power model for estimating m from $m_{\text{asy}}(n, \nu, \gamma)$ in the general γ case:

$$\log\left(\frac{\tilde{m}_{\text{sim}}(n, \nu, \gamma)}{m_{\text{asy}}(n, \nu, \gamma)}\right) = \frac{\beta_0 + \beta_1(1-\gamma) + \beta_2\nu}{n^{\beta_3+\beta_4(1-\gamma)}} + \epsilon_{n,\nu,\gamma}, \qquad \epsilon_{n,\nu,\gamma} \overset{iid}{\sim} N(0,1). \tag{14}$$

[8]The change in the shape of the log ratio curves for $\gamma \leq 0.05$ does not appear to be an artifact of the simulation: we ran the experiment for small samples and $\gamma \leq 0.05$ multiple times, and observed very consistent behavior across the experimental runs.

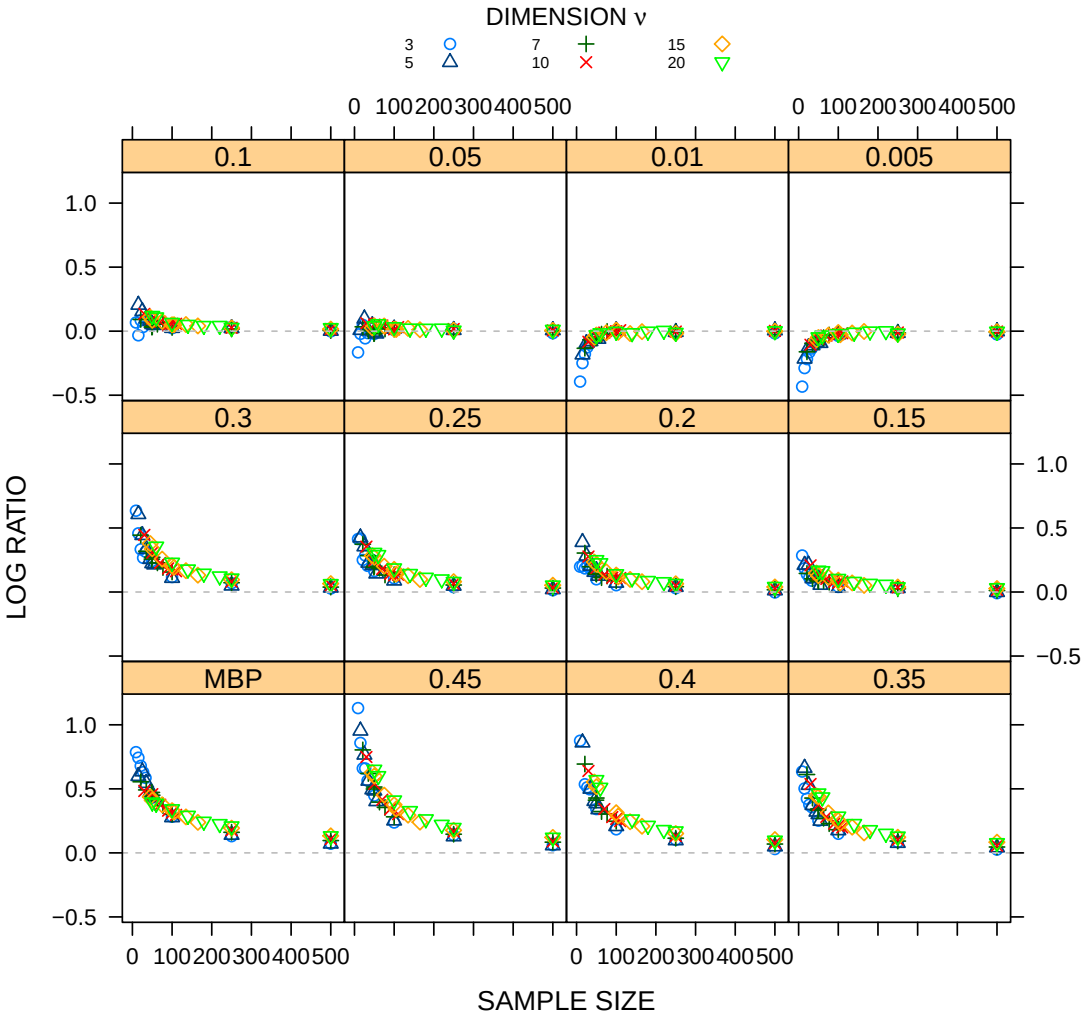


Figure 8: Logarithm of the ratio of the Wishart degrees of freedom estimated via simulation to the degrees of freedom calculated from the asymptotic formula, plotted against sample size and stratified by γ (printed in the yellow headers) and dimension (given by the plot symbols in each plot).

We fit this model in `R` using nonlinear least squares (available via the `nls` function) using the $\tilde{m}_{\text{sim}}(n, \nu, \gamma)$ values from our expanded simulation as well as the corresponding values of n, ν, and γ. The final model fit is

$$\log\left(\frac{m}{m_{\text{asy}}(n,\nu,\gamma)}\right) = \frac{12.746 - 14.546(1-\gamma) + 0.127\nu}{n^{0.559+0.149(1-\gamma)}}, \tag{15}$$

and hence our improved adjustment model for estimating m from $m_{\text{asy}}(n, \nu, \gamma)$ is

$$\tilde{m} = m_{\text{asy}}(n,\nu,\gamma)\exp\left(\frac{12.746 - 14.546(1-\gamma) + 0.127\nu}{n^{0.559+0.149(1-\gamma)}}\right). \tag{16}$$

Table 1 provides the regression coefficients along with their standard errors. All the regression coefficients are highly significant.

Table 1: Estimated coefficients, and their standard errors, for the model described by Equation (14).

Coefficient	Estimate	Std. Error	t-Statistic
β_0	12.746	0.305	41.8
β_1	-14.546	0.368	-39.5
β_2	0.127	0.007	17.5
β_3	0.559	0.011	49.2
β_4	0.149	0.018	8.2

4 Validation of the Improved Adjustment Model

4.1 Out-of-Sample Validation of the Hardin-Rocke Extension

To validate the fitted model (15), we used the same simulation procedure used in Section 3.3 with a different parameter set: we simulated 5000 draws of size n from a multivariate normal distribution $N(\mathbf{0}, \mathbf{I}_\nu)$ with dimensions $\nu = 2, 3, 5, 8, 11, 16, 22$ and sample sizes $n = 50, 150, 300, 500, 750, 1000$, as well as the dimension-dependent sample sizes $n = 4\nu, 6\nu, 8\nu, 10\nu, 12\nu$. For each sample we computed the MCD(γ) subset for $0.05 \leq \gamma \leq 0.45$ in increments of 0.05, as well as the extreme cases of $\gamma \in \{0.01, 0.005\}$. We estimate $\tilde{m}_{\text{sim}}(n, \nu, \gamma)$ as before for each combination of parameters. We then use our new model to estimate m from $m_{\text{asy}}(n, \nu, \gamma)$ for the corresponding values of n, ν, γ. With the output of this experiment we can examine how well the new model predicts the Wishart degrees of freedom parameter m for general γ and compare the new model's performance to that of the Hardin-Rocke model for $\gamma = \gamma^*$.

Figures 9, 10, and 11 show how well our proposed method estimates the Wishart degrees of freedom parameter m relative to the Hardin-Rocke method on the out-of-sample data set for $\gamma = 0.25, 0.05$, and 0.01 respectively. Each plot shows the ratios of the value of m estimated using each method to the simulated value $\tilde{m}_{\text{sim}}(n, \nu, \gamma)$ for a given combination of the n and ν values used in our out-of-sample testing. Our proposed method is generally more accurate for estimating m than the Hardin-Rocke method, as evidenced by the red triangles plotting near a ratio of 1.

Figures 12, 13, and 14 show how the better estimates of m from our proposed method translate into 0.01 critical values from the Hardin-Rocke F distribution for $\gamma = 0.25, 0.05$, and 0.01 respectively. Using our out-of-sample data set, we calculated 0.01 critical values using the simulated m, the value of m estimated from the Hardin-Rocke method, and the value of m estimated using our proposed method. The plot shows

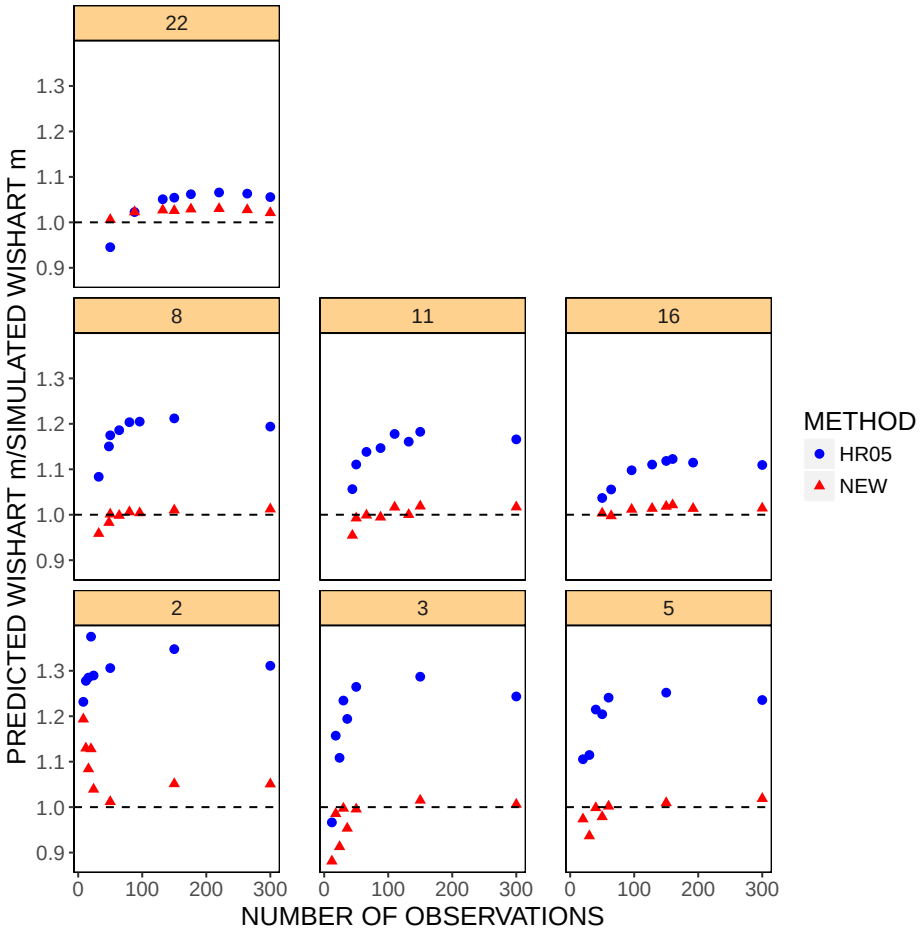


Figure 9: Out of sample comparison of estimated Wishart degrees of freedom parameter m to simulated value $\tilde{m}_{\text{sim}}(n, \nu, \gamma)$ using the Hardin-Rocke method and the proposed method with $\gamma = 0.25$. The plot shows the ratio of the degrees of freedom parameter m estimated using a given method to the simulated value $\tilde{m}_{\text{sim}}(n, \nu, \gamma)$, stratified by dimension ν. Blue dots represent the estimate with the Hardin-Rocke method, while red triangles represent the estimate with our proposed method. Sample size is plotted on the horizontal axis. Sample sizes shown in the plot are the dimension-dependent values $n = 2\nu, 4\nu, 6\nu, 8\nu, 10\nu, 12\nu$ (which hence vary between panels), as well as the fixed values $n = 50, 150, 300$. The dimension ν for each subgroup is shown in the yellow bars at the top of each subplot. The dashed line indicates the ideal ratio of 1.

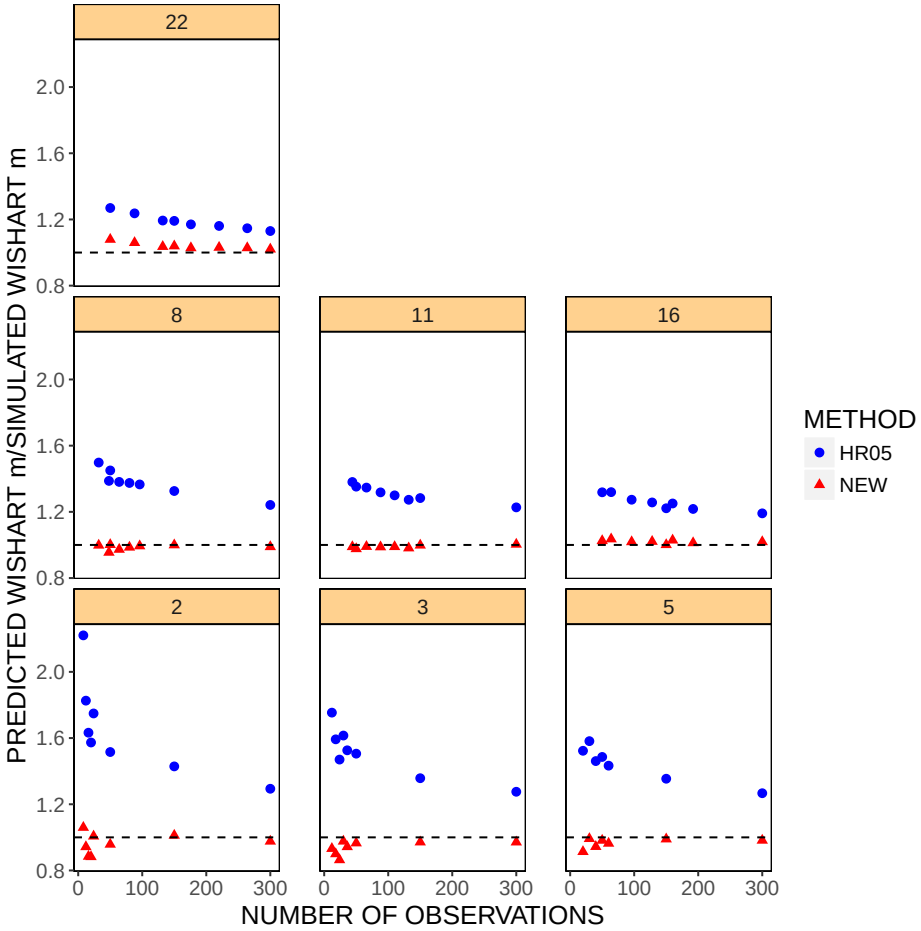


Figure 10: Out of sample comparison of estimated Wishart degrees of freedom parameter m to simulated value $\tilde{m}_{\text{sim}}(n, \nu, \gamma)$ using the Hardin-Rocke method and the proposed method with $\gamma = 0.05$. The plot setup is identical to Figure 9.

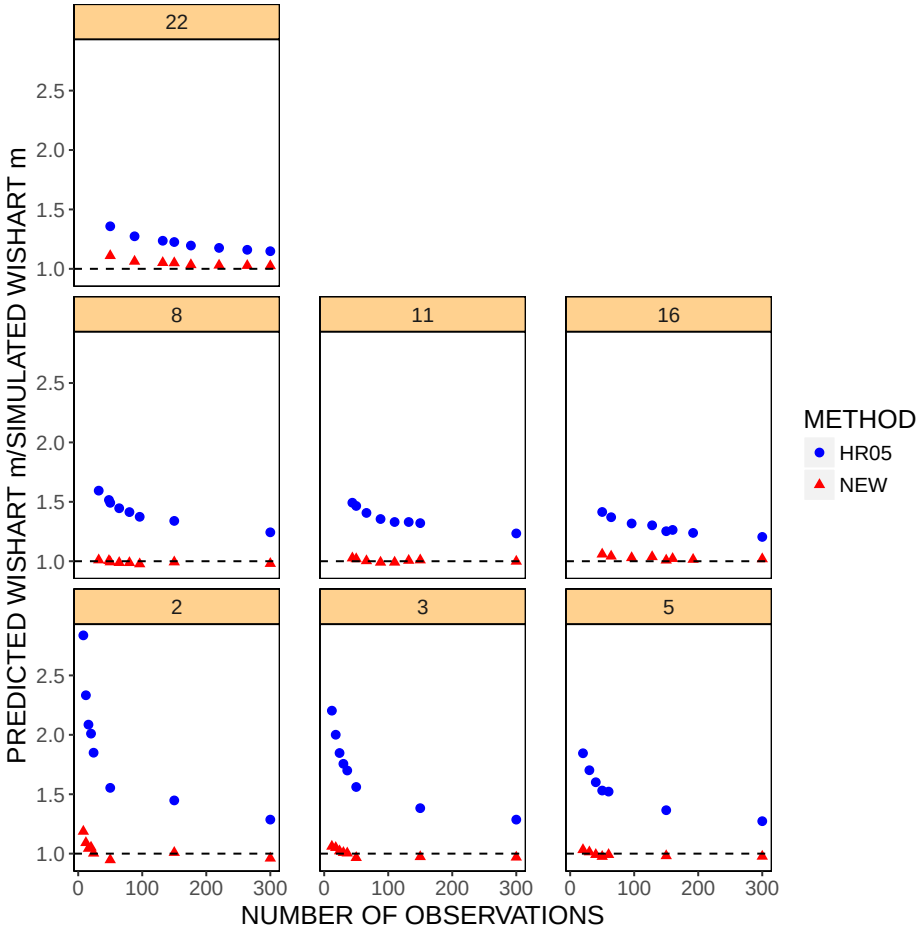


Figure 11: Out of sample comparison of estimated Wishart degrees of freedom parameter m to simulated value $\tilde{m}_{\text{sim}}(n, \nu, \gamma)$ using the Hardin-Rocke method and the proposed method with $\gamma = 0.01$. The plot setup is identical to Figure 9.

the ratios of the critical value computed from the estimated m to that computed using the simulated m for the Hardin-Rocke method (blue dots) and our proposed method (red triangles) using each combination of n and ν in the out-of-sample data set. Our proposed method generally results in much more accurate critical values, particularly for $\gamma = 0.05$ and $\gamma = 0.01$. Our results for 0.001 critical values were very similar and are not shown to conserve space.

Figure 15 shows how the proposed methodology performs relative to the Hardin-Rocke methodology for the maximum breakdown point case $\gamma = \gamma^*$. As it turns out, the performance of our method depends strongly on the ratio n/ν of the sample size to dimension, so our figure is structured accordingly. The proposed correction is much more accurate (as evidenced by medians closer to 0) and much less variable (as evidenced by smaller boxplot heights).[9] A Mann-Whitney test of the hypothesis that the median difference in the log-ratio of the predicted m to the simulated m between the Hardin-Rocke method and the proposed method is 0 has a p-value of 0.028. If we conduct the same test within each n/ν group, the p-values are as follows: $(0, 5] : 0.002$; $(5, 10] : 1.2 \times 10^{-7}$; $(10, 20] : 0.021$; and $(20, \infty) : 5 \times 10^{-5}$. Thus the new method is generally a modest improvement over Hardin and Rocke (2005) in the maximum breakdown point case $\gamma = \gamma^*$, and a strong improvement for moderate values of n/ν and very large values of n/ν.

Finally, Figure 16 shows the out-of-sample performance, as measured by the logarithm of the ratio of the predicted m to the simulated m, of our proposed improvement to the Hardin-Rocke methodology for the values of γ tested.[10] Again, the performance of our method depends on the ratio n/ν, so our figure reflects this grouping. Generally the proposed method is very good when the sample size is between 5 and 20 times the dimension: there is not much bias (the median ratios are close to 0) and not much dispersion in the correction factors (as evidenced by the tight boxplot widths). For small samples ($n < 5\nu$) the new method is generally good for $0.05 \leq \gamma \leq 0.35$, but shows some slight bias downward (meaning the corrected m is smaller than the simulation suggests it should be) for $\gamma > 0.35$ and bias upward for $\gamma < 0.05$. In very large samples $n > 20\nu$ and for $0.3 \leq \gamma \leq \gamma^*$ our method overestimates m slightly. The median ratio over all cases is approximately 1.01, so our model tends to overpredict m by 1% in general.

Overall, when the number of observations n is small compared to the dimension ν, the new method still underpredicts the degrees of freedom parameter m slightly. For large samples the new method still overpredicts m, but is more accurate on average than the Hardin-Rocke approach.

4.2 Testing that Our Model Gives the Correct False Positive Rates

As further validation of the fitted model, we ran a simulation experiment similar to that used by Hardin and Rocke (2005) to create Tables 1 and 2 in their paper. We generated 5000 draws of size n from an uncontaminated multivariate normal distribution $N(\mathbf{0}, \mathbf{I}_\nu)$ with dimension ν for sample sizes $n = 50, 100, 250, 500, 1000$ and $\nu = 5, 10, 20$. For each observation in a sample, we computed the MCD(γ)-based RSDs for $\gamma = \gamma^*, 0.35, 0.25, 0.10, 0.05, 0.01$. We tested observations for outlyingness at the α level by comparing these RSDs to the $1 - \alpha$ quantile of the Hardin-Rocke F distribution with degrees of freedom m calculated using the Hardin-Rocke adjustment (13) and using the new method (15) developed in this paper. Since the data contains no outliers by construction, any outliers detected are false positives. We thus evaluate the performance of the two methods for estimating m by comparing the empirically observed false positive rate from the simulated data to the true value α. While we know the limitations of this exercise from the work of

[9] The large outlier for our new method in the $0 < n/\nu \leq 5$ group corresponds to the case $n = 8$ and $\nu = 2$. The large outliers for our new method in the $5 < n/\nu \leq 10$ group correspond to dimension $\nu = 2$ with sample sizes $n = 12, 16, 20$.

[10] Full results are available in Table 5 in Appendix D.

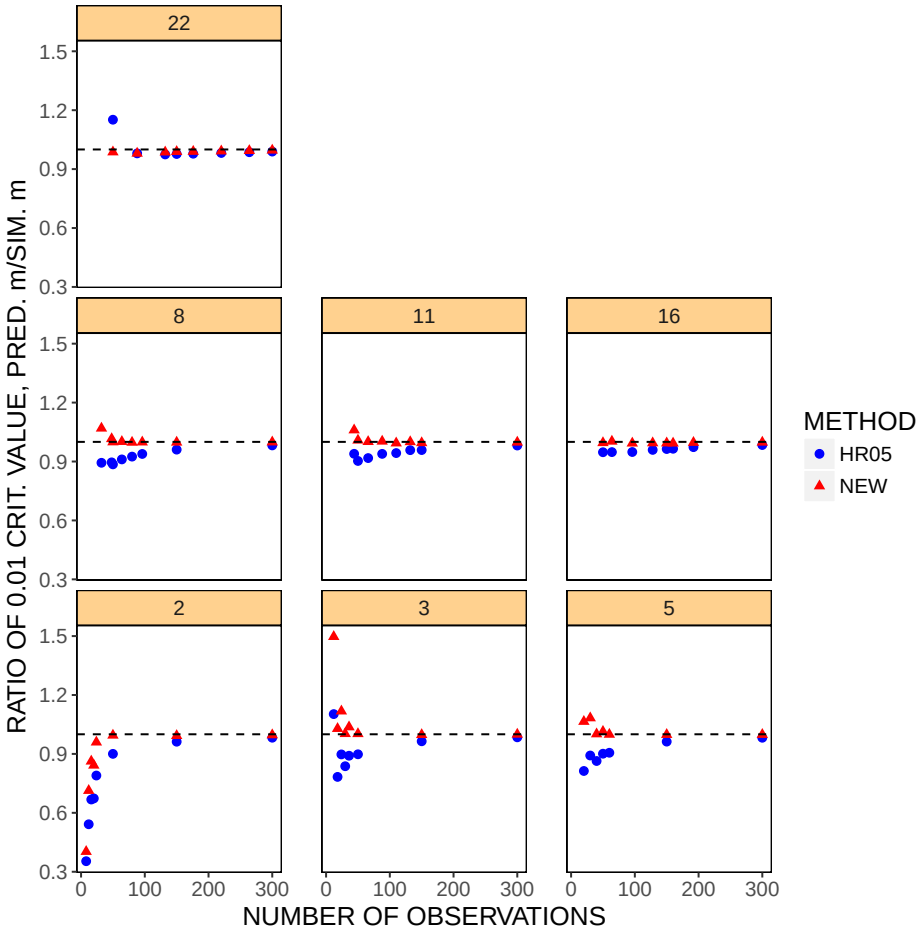


Figure 12: Out of sample comparison of 0.01 critical values from the Hardin-Rocke F distribution computed using the estimated Wishart degrees of freedom parameter m from the Hardin-Rocke method and the proposed method with $\gamma = 0.25$. The plot shows the ratio of the 0.01 critical value computed using the estimated value of m to the 0.01 critical value computed using the simulated value of m for each method, stratified by dimension ν. Blue dots represent the estimate with the Hardin-Rocke method, while red triangles represent the estimate with our proposed method. Sample size is plotted on the horizontal axis. The pattern of sample sizes used here is identical to that used in Figure 9. The dimension ν for each subgroup is shown in the yellow bars at the top of each subplot. The dashed line indicates the ideal ratio of 1.

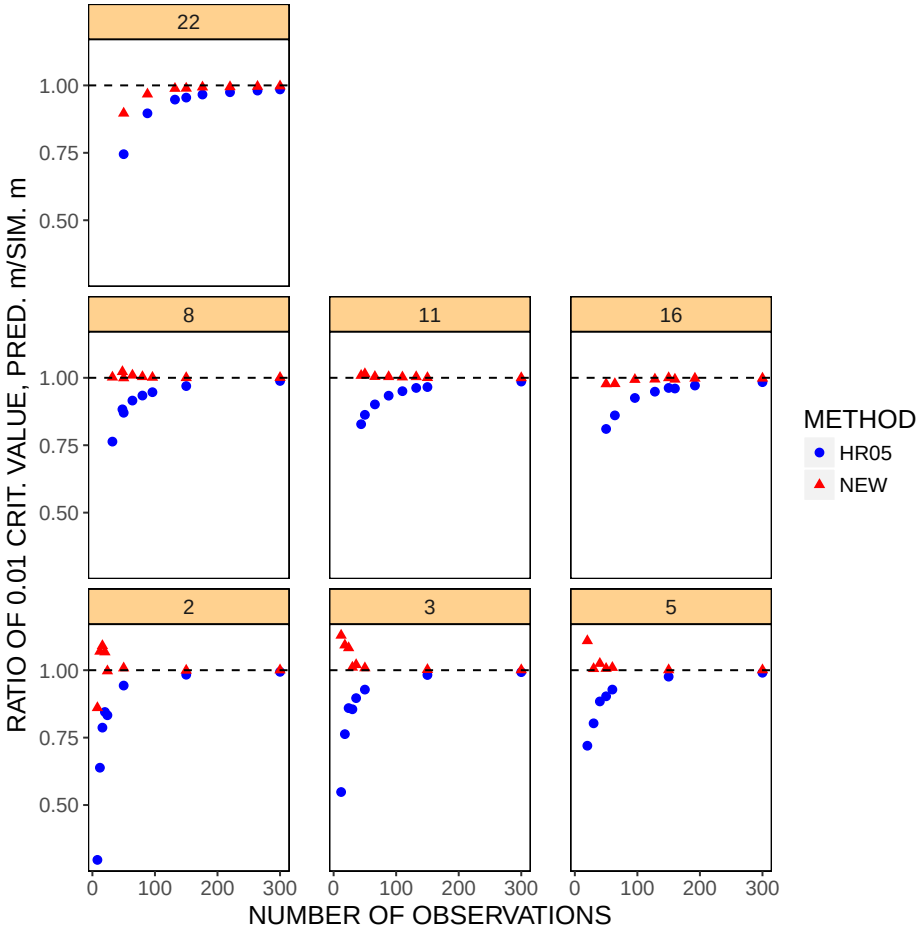


Figure 13: Out of sample comparison of 0.01 critical values from the Hardin-Rocke F distribution computed using the estimated Wishart degrees of freedom parameter m from the Hardin-Rocke method and the proposed method with $\gamma = 0.05$. The plot setup is identical to that of Figure 12.

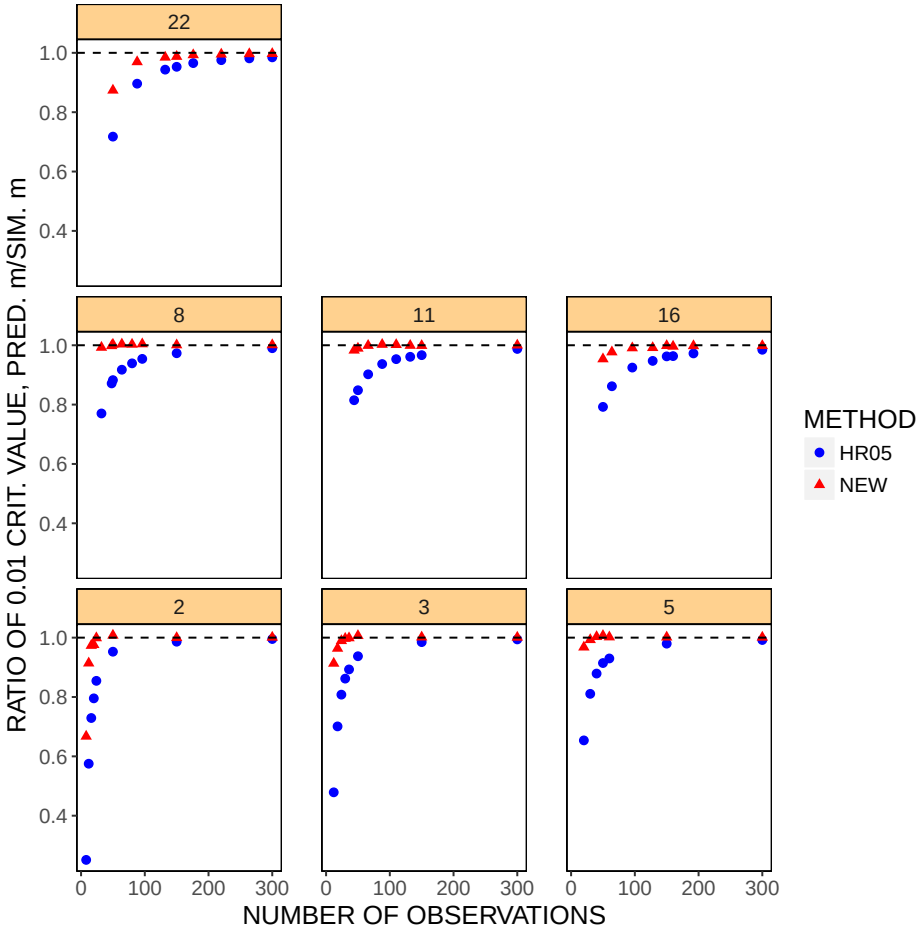
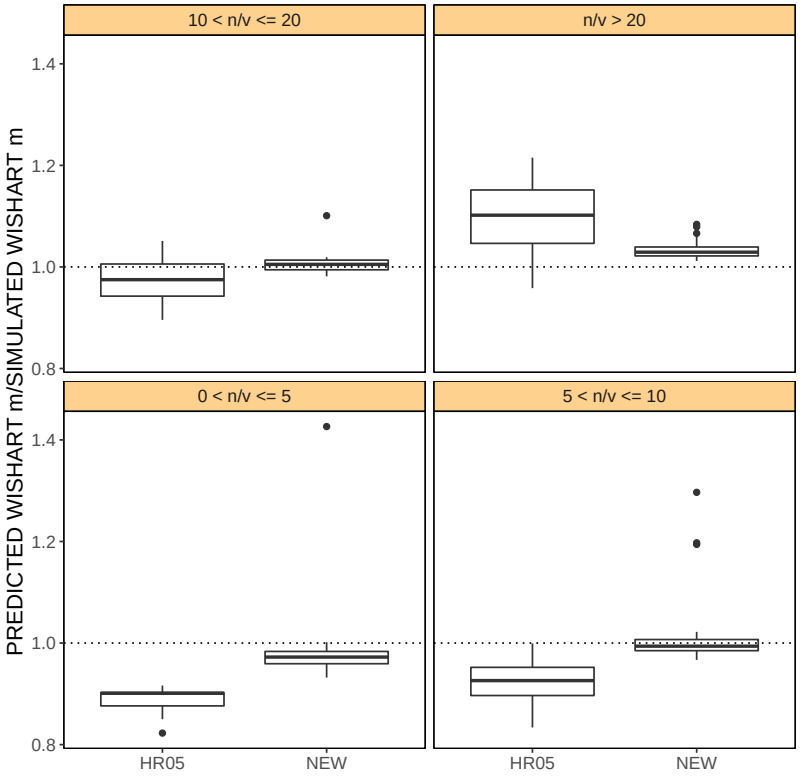


Figure 14: Out of sample comparison of 0.01 critical values from the Hardin-Rocke F distribution computed using the estimated Wishart degrees of freedom parameter m from the Hardin-Rocke method and the proposed method with $\gamma = 0.01$. The plot setup is identical to that of Figure 12.



n/ν	Pairs (ν, n)
$(0, 5]$	$(2, 8)$, $(3, 12)$, $(5, 20)$, $(8, 32)$, $(11, 44)$, $(11, 50)$, $(16, 50)$, $(16, 64)$, $(22, 50)$, $(22, 88)$
$(5, 10]$	$(2, 12)$, $(2, 16)$, $(2, 20)$, $(3, 18)$, $(3, 24)$, $(3, 30)$, $(5, 30)$, $(5, 40)$, $(5, 50)$, $(8, 48)$, $(8, 50)$, $(8, 64)$, $(8, 80)$, $(11, 66)$, $(11, 88)$, $(11, 110)$, $(16, 96)$, $(16, 128)$, $(16, 150)$, $(16, 160)$, $(22, 132)$, $(22, 150)$, $(22, 176)$, $(22, 220)$
$(10, 20]$	$(2, 24)$, $(3, 36)$, $(3, 50)$, $(5, 60)$, $(8, 96)$, $(8, 150)$, $(11, 132)$, $(11, 150)$, $(16, 192)$, $(16, 300)$, $(22, 264)$, $(22, 300)$
> 20	$(2, 50)$, $(2, 150)$, $(2, 300)$, $(2, 500)$, $(2, 750)$, $(2, 1000)$, $(3, 150)$, $(3, 300)$, $(3, 500)$, $(3, 750)$, $(3, 1000)$, $(5, 150)$, $(5, 300)$, $(5, 500)$, $(5, 750)$, $(5, 1000)$, $(8, 300)$, $(8, 500)$, $(8, 750)$, $(8, 1000)$, $(11, 300)$, $(11, 500)$, $(11, 750)$, $(11, 1000)$, $(16, 500)$, $(16, 750)$, $(16, 1000)$, $(22, 500)$, $(22, 750)$, $(22, 1000)$

Figure 15: Boxplot showing performance of the proposed correction methodology (NEW) against that of the Hardin-Rocke methodology (HR05) for the maximum breakdown point case $\gamma = \gamma^*$, stratified by the ratio n/ν of observations to variables. Performance is measured by the ratio of the predicted Wishart degrees of freedom value to the value computed via the simulation methodology used in Hardin and Rocke (2005). For the reader's convenience, the pairs (ν, n) of dimensions and sample sizes that fall into each n/ν bin are listed in the table below the plot.

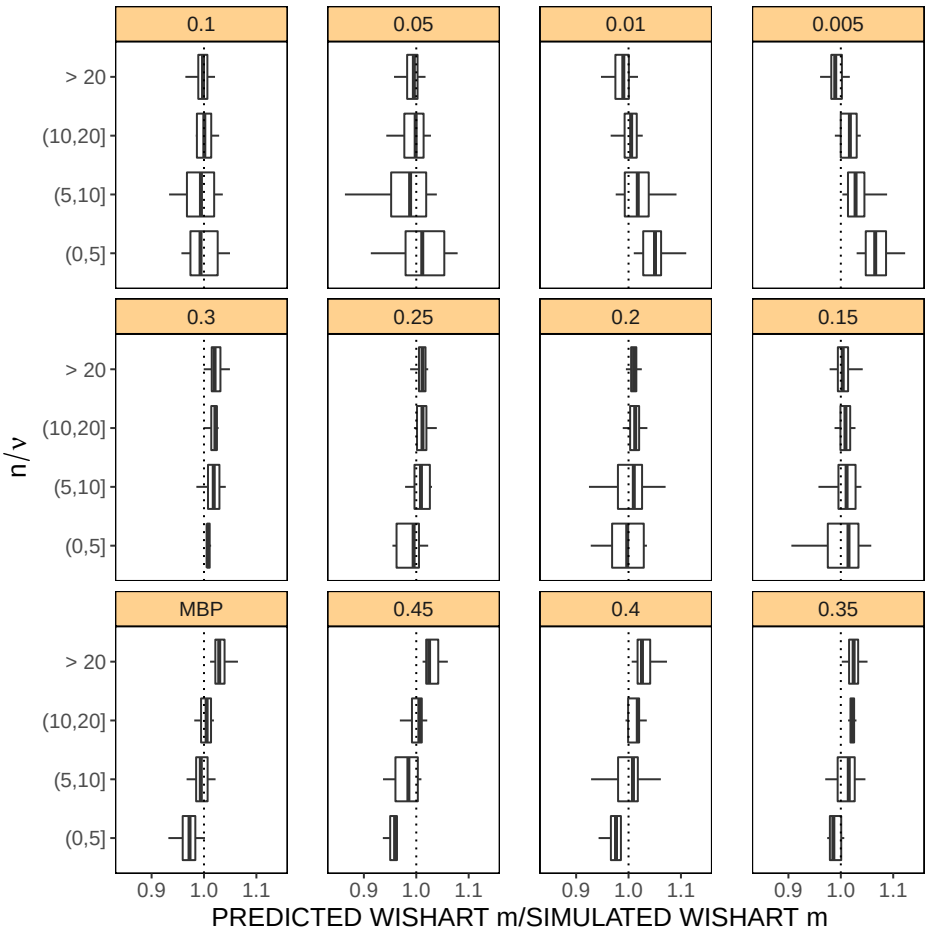


Figure 16: Boxplots showing the range of out of sample performance of the proposed correction methodology, stratified by γ (yellow box) and the ratio n/ν of observations to variables (vertical axis). Performance is measured by the ratio of the predicted Wishart degrees of freedom value to the value computed via the simulation methodology used in Hardin and Rocke (2005). The dashed vertical lines at 1 correspond to perfect agreement between prediction and simulation. Outliers are omitted from the plot to highlight the overall performance of the method. The pairs (ν, n) of dimensions and sample sizes that fall into each bin are identical to those used in Figure 15.

Cerioli et al. (2009), this test does provide another comparison of our method to that of Hardin and Rocke.

Tables 2 and 3 show the results of testing how well each method of predicting m translates to outlier detection using the above test. (The results for $n = 1000$ are similar to those for $n = 500$ and are omitted to save space.) For $n = 250$ or $n = 500$, the Hardin-Rocke method leads to false positive rates that are smaller than expected as γ gets closer to 0 or as dimension ν increases. For those sample sizes our proposed method gives false positive rates that are closer to the ideal values of α for most γ values. Only in the $\gamma = 0.01$ case does our method become noticeably inaccurate, and even then it is still more accurate than the original Hardin-Rocke approach.

For small samples ($n = 100$), our method gives false positive rates that are close to ideal for $\gamma = 0.05, 0.10, 0.25$, while the Hardin-Rocke method yields false positive rates that are too small. For $\gamma = 0.35$ our method has a higher false positive rate than expected, while the Hardin-Rocke method has a lower-than-expected rate. At the maximum breakdown point case $\gamma = \gamma^*$ both methods exhibit higher false positive rates than expected, and there is no clear winner between the two. Neither method is accurate for $\gamma = 0.01$, but our method is far closer to the true α.

Table 2: Mean percentage of simulated data, for selected sample sizes n and dimensions ν, with MCD(γ)-based RSDs exceeding the 0.05 quantile produced using the Hardin-Rocke method (HR05) and the proposed correction method (NEW). Ideally, each percentage should be close to 5%. Standard errors are given in parentheses and are also expressed in percentages. Compare to Table 1 of Hardin and Rocke (2005), which showed the results results of using their method in the maximum breakdown point case (MBP) of MCD. Results for $n = 1000$ are similar to those for $n = 500$ and are omitted to save space.

Dimension (ν)	**n = 50**		**n = 100**		**n = 250**		**n = 500**	
	HR05	NEW	HR05	NEW	HR05	NEW	HR05	NEW
$\gamma =$ **MBP**								
5	6.28 (0.07)	6.76 (0.07)	6.42 (0.05)	6.26 (0.05)	5.67 (0.03)	5.36 (0.03)	5.27 (0.02)	5.02 (0.02)
10	6.79 (0.07)	8.46 (0.07)	6.67 (0.05)	6.94 (0.05)	5.55 (0.03)	5.36 (0.03)	5.15 (0.02)	5.02 (0.02)
20	4.99 (0.05)	8.20 (0.05)	4.39 (0.04)	6.57 (0.04)	4.55 (0.02)	5.14 (0.02)	4.65 (0.01)	4.93 (0.02)
$\gamma = 0.35$								
5	2.89 (0.04)	6.14 (0.06)	3.19 (0.03)	5.35 (0.04)	3.89 (0.02)	5.10 (0.02)	4.30 (0.01)	5.00 (0.02)
10	4.25 (0.05)	7.64 (0.06)	3.44 (0.03)	6.13 (0.04)	3.61 (0.02)	5.06 (0.02)	4.07 (0.01)	4.98 (0.01)
20	8.33 (0.06)	9.66 (0.06)	2.87 (0.03)	6.40 (0.04)	2.69 (0.02)	5.13 (0.02)	3.34 (0.01)	4.94 (0.01)
$\gamma = 0.25$								
5	1.14 (0.02)	4.87 (0.05)	2.04 (0.02)	5.11 (0.03)	3.26 (0.02)	5.03 (0.02)	3.93 (0.01)	4.97 (0.01)
10	1.50 (0.03)	6.08 (0.05)	1.85 (0.02)	5.49 (0.03)	2.81 (0.02)	5.01 (0.02)	3.61 (0.01)	4.97 (0.01)
20	3.63 (0.04)	8.25 (0.05)	1.15 (0.02)	5.87 (0.04)	1.81 (0.01)	5.06 (0.02)	2.74 (0.01)	4.94 (0.01)
$\gamma = 0.10$								
5	0.38 (0.01)	4.57 (0.04)	1.18 (0.02)	4.87 (0.03)	2.71 (0.01)	4.91 (0.02)	3.64 (0.01)	4.97 (0.01)
10	0.26 (0.01)	4.35 (0.03)	0.76 (0.01)	4.82 (0.03)	2.12 (0.01)	4.92 (0.02)	3.18 (0.01)	4.98 (0.01)
20	0.32 (0.01)	4.67 (0.02)	0.27 (0.01)	5.08 (0.02)	1.09 (0.01)	5.00 (0.02)	2.18 (0.01)	4.97 (0.01)
$\gamma = 0.05$								
5	0.23 (0.01)	3.99 (0.03)	0.99 (0.01)	4.15 (0.02)	2.58 (0.01)	4.64 (0.01)	3.51 (0.01)	4.75 (0.01)
10	0.12 (0.01)	3.14 (0.02)	0.61 (0.01)	4.17 (0.02)	1.96 (0.01)	4.39 (0.01)	3.06 (0.01)	4.65 (0.01)
20	0.15 (0.01)	3.64 (0.01)	0.18 (0.01)	3.67 (0.01)	0.94 (0.01)	4.40 (0.01)	2.03 (0.01)	4.51 (0.01)
$\gamma = 0.01$								
5	0.15 (0.01)	2.04 (0.01)	0.60 (0.01)	2.68 (0.02)	1.86 (0.01)	3.96 (0.01)	3.05 (0.01)	4.45 (0.01)
10	0.07 (0.01)	1.86 (0.01)	0.35 (0.01)	1.37 (0.01)	1.23 (0.01)	3.02 (0.01)	2.23 (0.01)	3.95 (0.01)
20	0.06 (0.00)	1.96 (0.00)	0.08 (0.00)	1.00 (0.00)	0.75 (0.01)	1.58 (0.01)	1.11 (0.00)	2.87 (0.01)

Table 3: Mean percentage of simulated data, for selected sample sizes n and dimensions ν, with MCD(γ)-based RSDs exceeding the 0.01 quantile produced using the Hardin-Rocke method (HR05) and the proposed correction method (NEW). Ideally, each percentage should be close to 1%. Standard errors are given in parentheses and are also expressed in percentages. Compare to Table 2 of Hardin and Rocke (2005), which showed the results results of using their method in the maximum breakdown point case (MBP) of MCD. Results for $n = 1000$ are similar to those for $n = 500$ and are omitted to save space.

Dimension (ν)	**n = 50**		**n = 100**		**n = 250**		**n = 500**	
	HR05	NEW	HR05	NEW	HR05	NEW	HR05	NEW
$\gamma =$ **MBP**								
5	1.39 (0.03)	1.63 (0.04)	1.65 (0.03)	1.51 (0.03)	1.29 (0.01)	1.15 (0.01)	1.11 (0.01)	1.05 (0.01)
10	1.93 (0.04)	2.71 (0.05)	1.89 (0.03)	1.90 (0.03)	1.25 (0.01)	1.18 (0.01)	1.07 (0.01)	1.03 (0.01)
20	1.45 (0.03)	3.25 (0.04)	1.02 (0.02)	1.87 (0.03)	0.93 (0.01)	1.12 (0.01)	0.93 (0.01)	1.01 (0.01)
$\gamma = 0.35$								
5	0.39 (0.01)	1.44 (0.03)	0.48 (0.01)	1.16 (0.02)	0.65 (0.01)	1.04 (0.01)	0.77 (0.01)	1.01 (0.01)
10	0.93 (0.02)	2.27 (0.04)	0.60 (0.01)	1.52 (0.02)	0.60 (0.01)	1.07 (0.01)	0.71 (0.01)	0.99 (0.01)
20	3.33 (0.05)	4.21 (0.05)	0.52 (0.01)	1.70 (0.02)	0.42 (0.01)	1.08 (0.01)	0.55 (0.00)	0.99 (0.01)
$\gamma = 0.25$								
5	0.07 (0.01)	0.98 (0.02)	0.21 (0.01)	1.06 (0.02)	0.46 (0.01)	1.01 (0.01)	0.65 (0.01)	1.01 (0.01)
10	0.17 (0.01)	1.47 (0.03)	0.21 (0.01)	1.22 (0.02)	0.40 (0.01)	1.03 (0.01)	0.60 (0.01)	1.01 (0.01)
20	0.84 (0.02)	3.36 (0.04)	0.13 (0.01)	1.41 (0.02)	0.22 (0.00)	1.02 (0.01)	0.40 (0.00)	0.99 (0.01)
$\gamma = 0.10$								
5	0.01 (0.00)	0.87 (0.02)	0.07 (0.00)	0.97 (0.01)	0.33 (0.01)	0.98 (0.01)	0.56 (0.00)	1.00 (0.01)
10	0.01 (0.00)	0.91 (0.02)	0.04 (0.00)	0.99 (0.01)	0.24 (0.00)	0.99 (0.01)	0.46 (0.00)	0.98 (0.01)
20	0.02 (0.00)	1.41 (0.02)	0.01 (0.00)	1.15 (0.02)	0.10 (0.00)	1.00 (0.01)	0.28 (0.00)	1.00 (0.01)
$\gamma = 0.05$								
5	0.01 (0.00)	0.86 (0.02)	0.04 (0.00)	0.92 (0.01)	0.29 (0.00)	0.97 (0.01)	0.53 (0.00)	1.00 (0.01)
10	0.00 (0.00)	0.79 (0.02)	0.03 (0.00)	0.97 (0.01)	0.21 (0.00)	0.98 (0.01)	0.44 (0.00)	0.99 (0.01)
20	0.00 (0.00)	1.42 (0.02)	0.01 (0.00)	1.07 (0.01)	0.08 (0.00)	1.03 (0.01)	0.25 (0.00)	0.99 (0.01)
$\gamma = 0.01$								
5	0.00 (0.00)	0.65 (0.01)	0.03 (0.00)	0.63 (0.01)	0.27 (0.00)	0.86 (0.01)	0.50 (0.00)	0.90 (0.00)
10	0.00 (0.00)	0.71 (0.01)	0.02 (0.00)	0.61 (0.01)	0.18 (0.00)	0.84 (0.01)	0.41 (0.00)	0.86 (0.00)
20	0.00 (0.00)	1.08 (0.01)	0.00 (0.00)	0.66 (0.01)	0.07 (0.00)	0.87 (0.01)	0.24 (0.00)	0.84 (0.00)

In very small samples ($n = 50$) neither method is particularly accurate: the Hardin-Rocke method tends to yield false positive rates that are too low, while our method yields rates that are too high for $\gamma \geq 0.25$. For $\gamma = 0.10$ or $\gamma = 0.05$ the false positve rate from our method is a bit smaller than the nominal size α, but it is much closer to the truth than the rate resulting from the Hardin-Rocke method. The extreme case of $n = 50$ and $\gamma = 0.01$ is particularly challenging for both methods.

One takeaway from the tables for finance practitioners is that for samples of size $n = 50$, one should not use MCD(γ) with $\gamma < 0.01$, especially if the dimension ν is larger than 10. Likewise, for $n = 100$, $\gamma = 0.05$ is about as small as one can go and maintain fairly accurate false positive rates.

4.3 Extension of FSRMCD and IRMCD to Arbitrary γ

4.3.1 Cerioli's FSRMCD and IRMCD Methodologies

Cerioli (2010) developed two methods for conducting accurate outlier tests using MCD-based RSDs, namely, the Finite Sample Reweighted MCD and Iterated Reweighted MCD procedures. The Finite Sample Reweighted MCD (FSRMCD) methodology is designed to control the family-wise error rate (FWER) for the set of individual outlier tests

$$H_{0i} : \mathbf{x}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \qquad i = 1, \ldots, n. \tag{17}$$

The FWER is the probability that at least one of these hypotheses is rejected incorrectly. A well-known approach to controlling the FWER of a set of tests is Bonferroni correction. Suppose we wish to achieve a FWER of α_1. If we test each individual hypothesis H_{0i} at the $\alpha = \alpha_1/n$ level rather than the α_1 level, the FWER is guaranteed to be no more than α_1 (by Bonferroni's inequality). The Bonferroni correction is conservative and does not require us to assume the tests are independent. It is hence widely applicable. When the tests of the H_{0i} are independent, the Šidák (1967) correction gives an exact FWER of α_1 by testing each individual hypothesis H_{0i} at the $\alpha = 1-(1-\alpha_1)^{1/n}$ level. The FSRMCD uses the Šidák correction and the Hardin-Rocke distributional approximation to provide good control over the FWER of the individual RSD tests and the correct size for the intersection hypothesis (Equation (3)).

As above, let α be the nominal size at which each individual hypothesis H_{0i} is tested, and let α_1 be the nominal size for testing the intersection hypothesis. The FSRMCD method proceeds as follows.

1. For a given h or γ, compute the raw MCD(γ) on the data.

2. Compute RSDs based on the raw MCD. Test each observation at the 0.025 level for outlyingness using the Hardin-Rocke distribution.[11] Rejected observations are assigned weight 0, while all other observations receive weight 1.

3. Compute the reweighted MCD estimate using the weights from Step 2.

4. Test RSDs based on the reweighted MCD using a distribution conditional on the weight of the corresponding observation from Step 2: for observations receiving weight 1, we test RSDs against a scaled Beta distribution. For observations with weight 0, we test RSDs against a scaled F distribution. These tests are performed using a nominal size of α, e.g., $\alpha = 0.01$.

As Cerioli (2010) points out, the FSRMCD procedure unfortunately has low power. The Iterated Reweighted MCD (IRMCD) test improves the power of FSRMCD by adding an additional step to the process.

[11] The value of 0.025 is based on a recommendation in Rousseeuw and van Driessen (1999) for the reweighted MCD.

Let α_1 be the desired nominal size of the intersection test. Then $\alpha = 1 - (1 - \alpha_1)^{1/n}$ is the Šidák-corrected size for the individual hypothesis tests.

4. In Step 4 of FSRMCD, test all RSDs using the conditional distribution at the α level.

5. If no observations are rejected by this test, we conclude that there is no evidence of outliers in the data. If at least one observation is rejected, we then test each observation at the α_1 level using the distribution from Step 4. Any observation that fails its test is flagged as an outlier.

The first test ensures IRMCD will have the same false positive rate as FSRMCD for the intersection test, while the second test improves our ability to correctly identify outliers when they are present in the data set.

4.3.2 Modifying FSRMCD and IRMCD for Arbitrary γ

The FSRMCD and IRMCD procedures depend on the Hardin-Rocke methodology, which was only defined for the maximum breakdown point case $\gamma = \gamma^*$. As we showed in Tables 2 and 3, the Hardin-Rocke estimator for m can lead to false-positive rates that are much too small for $\gamma \in \{0.01, 0.05, 0.25\}$ and sample sizes less than 250. Our improved adjustment method performs much better than the Hardin-Rocke adjustment across a wide range of sample sizes, dimensions, and trimming fractions. We thus implemented and tested modified versions of FSRMCD and IRMCD using our improved adjustment. We will then be able to use the modified versions in financial studies.

Simulations similar to those in Cerioli (2010) were run to verify the accuracy of modified implmentation. We drew $N = 5000$ independent samples from an $N(\mathbf{0}, \mathbf{I}_\nu)$ distribution, and estimated the size of the intersection test (3) as the fraction of samples for which the null hypothesis is incorrectly rejected at the 0.01 level. We focused on the cases $\gamma \in \{\gamma^*, 0.25, 0.05, 0.01\}$: the former two for comparison with Cerioli's results, and $\gamma \in \{0.05, 0.01\}$ for use in later chapters.[12]

Table 4 shows the results of testing our implementation of the finite-sample and iteratively reweighted MCD estimators (FSRMCD and IRMCD, respectively) defined in Cerioli (2010). Overall our implementation gives the right sizes empirically, and it produces results consistent with those presented in Table 1 and 2 of that paper. (Table 6 in Appendix E provides standard deviations for the entries in the table.)

Power calculations for our modified implementation of IRMCD are discussed in Green (2017)

5 Discussion

Our modified version of the Hardin-Rocke adjustment to the asymptotic degrees of freedom parameter estimate performs very well in general: in the out-of-sample tests portrayed in Figure 16 our predicted m was larger than the simulated m by only 1%, on average, across all combinations of sample size, dimension, and γ tested. The new method is more accurate, on average, than the Hardin and Rocke (2005) method, and performs more consistently across a variety of sample sizes and dimensions.

For small samples $n < 5\nu$ there is still some bias, i.e., the predicted m tends to be too small for γ near γ^*, and too large for γ near 0. Likewise for large samples $n > 20\nu$ the predicted m tends to be too large for γ near γ^* and a little too small for γ near 0. The deviations are not terribly large, though. For instance, for small samples and $\gamma = 0.005$ the predicted value is 1.06 times the simulated value on average, which means

[12] The simulations and the analysis were performed on a laptop running Windows 7 Ultimate SP 1 with an Intel® Core™ i7-3740QM processor running at 2.7GHz and 32GB of RAM.

Table 4: Results of simulation tests of FSRMCD and IRMCD implementations. The table shows the estimated size for testing the hypothesis of no outliers in the data at the nominal size of 0.01. Ideally each entry should be close to 0.01. The size is estimated using 5000 simulations for each combination of sample size n and dimension ν. Compare to Table 1 of Cerioli (2010). (Table 6 in Appendix E provides standard deviations for the entries in the table.)

Dimension	Method	$n = 40$	$n = 60$	$n = 90$	$n = 125$	$n = 200$	$n = 400$
$\gamma = \gamma^*$							
$\nu = 5$	FSRMCD	0.013	0.013	0.014	0.012	0.013	0.010
	IRMCD	0.015	0.011	0.013	0.011	0.011	0.009
$\nu = 10$	FSRMCD	0.023	0.012	0.009	0.010	0.008	0.008
	IRMCD	0.020	0.014	0.010	0.010	0.008	0.008
$\nu = 15$	FSRMCD	0.020	0.012	0.009	0.011	0.009	0.009
	IRMCD	0.023	0.011	0.009	0.012	0.009	0.009
$\gamma = 0.25$							
$\nu = 5$	FSRMCD	0.013	0.012	0.011	0.010	0.012	0.009
	IRMCD	0.013	0.014	0.012	0.012	0.010	0.011
$\nu = 10$	FSRMCD	0.013	0.013	0.012	0.014	0.010	0.010
	IRMCD	0.015	0.011	0.007	0.010	0.012	0.008
$\nu = 15$	FSRMCD	0.012	0.012	0.011	0.007	0.009	0.008
	IRMCD	0.012	0.012	0.012	0.009	0.010	0.010
$\gamma = 0.05$							
$\nu = 5$	FSRMCD	0.010	0.011	0.012	0.011	0.011	0.012
	IRMCD	0.011	0.012	0.010	0.011	0.011	0.010
$\nu = 10$	FSRMCD	0.011	0.011	0.013	0.009	0.012	0.010
	IRMCD	0.013	0.013	0.011	0.014	0.013	0.010
$\nu = 15$	FSRMCD	0.019	0.013	0.015	0.012	0.011	0.013
	IRMCD	0.017	0.011	0.015	0.009	0.012	0.009
$\gamma = 0.01$							
$\nu = 5$	FSRMCD	0.006	0.008	0.012	0.010	0.006	0.011
	IRMCD	0.006	0.009	0.008	0.008	0.010	0.011
$\nu = 10$	FSRMCD	0.007	0.009	0.005	0.010	0.009	0.010
	IRMCD	0.007	0.007	0.009	0.006	0.007	0.009
$\nu = 15$	FSRMCD	0.009	0.008	0.005	0.007	0.008	0.010
	IRMCD	0.008	0.007	0.009	0.011	0.009	0.010

a true m of 50 is predicted to be 53; this translates into critical values that are 1-2% too small in dimensions less than 10. In higher dimensions, e.g., larger than 20, the difference in the critical values will be larger and might have a more noticable impact on outlier detection.

Due to the computational requirements of the simulations done here, we were only able to run the full experiment once. Thus, we do not know how variable the simulated m can be in general.[13] However, in the process of investigating the behavior of the simulated m for γ near 0, we did run the $\gamma \leq 0.1$ cases several times. As the sample size n gets larger, we observed more variation in the simulated value of m; however this does not seem to translate into much variation in the resulting 0.01 critical values. For small sample sizes ($n < 100$) or when n is a small multiple of ν, there can be a wider range of critical values resulting from the simulated m values. The MCD estimate with $\gamma \leq 0.1$ is discarding relatively few observations, so a potential improvement to our methodology might consider an alternative approach to calculating the distribution of the MCD estimate in such cases.

6 Conclusions and Further Research

We have extended the Hardin and Rocke (2005) methodology for estimating parameters of their F distribution to the general MCD(γ) estimator, thereby ensuring that the FSRMCD and IRMCD outlier detection methodologies introduced by Cerioli (2010) give the right test sizes for arbitrary γ (as long as the sample size is not very small compared to the dimension).

For some applications the MCD may not be the best robust dispersion estimate to use. Maronna et al. (2006) recommend the use of so-called S-estimators over the MCD based on a simulation study detailed in their Chapter 6.8. They demonstrate that certain types of S-estimators offer a better balance of bias and variability than the MCD. Briefly, an S-estimate $\left(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}\right)$ of multivariate location and dispersion tries to minimize a univariate robust scale estimate $\hat{\sigma}$ of the RSDs (based on $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\Sigma}}$) subject to constraints on the determinant of the dispersion estimate $\tilde{\boldsymbol{\Sigma}}$. The Maronna et al. (2006) study considered S-estimators based on two different robust scale estimates $\hat{\sigma}$: one defined using the Tukey bisquare ρ function and another based on the Rocke (1996) biflat ρ function. The bisquare-based S-estimator can be configured to have the maximum asymptotic breakdown point of 1/2, but as the dimension ν increases it becomes more efficient, and hence, more biased and less robust to outliers. The Rocke-type S-estimator was designed to approximately maintain a desired level of efficiency and robustness as the dimension of the data increases. (These estimators are discussed in greater detail in Appendix A of Green (2017).) Not surprisingly, the simulations of Maronna et al. show that the bisquare S-estimator is preferred to the MCD for dimension $\nu < 10$, while the Rocke-type S-estimator is preferred for dimension $\nu \geq 10$.

Furthermore, Alqallaf et al. (2009) points out that the MCD is based on the so-called Tukey-Huber Contamination Model.[14] The Tukey-Huber Contamination Model assumes that whether a given observation $\mathbf{x}_i$ is contaminated (i.e., comes from a distribution different from the other observations) is independent of whether any other observation $\mathbf{x}_j$ is contaminated, but if an observation $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,\nu})$ is contaminated then all of its coordinates $x_{i,k}$ are assumed to be contaminated. Typically in the Tukey-Huber Contamination Model the (uncontaminated) bulk of the data is assumed to follow a multivariate normal distribution. Some of the implications of the above assumption are hence that (a) most observations fit the multivariate normal assumption well; (b) outlying observations can be detected and trimmed in a multivariate manner; and (c)

[13] Recall that the commonly used *fastMCD* procedure of Rousseeuw and van Driessen (1999) involves random sampling as well, which is an additional source of variability in the m estimates.

[14] Agostinelli and Yohai (2017) provide a review of the the Tukey-Huber and Independent Contamination Models.

affine equivariance can be invoked to justify studying robustness and outlier detection only for a multivariate normal distribution with mean vector $\mathbf{0}$ and the identity matrix $\mathbf{I}_\nu$ for covariance.

In many applications, observations may only be outlying in a few coordinates, however, and a significant fraction of observations may exhibit some degree of contamination. Alqallaf et al. introduce a more flexible contamination model, the Independent Contamination Model (ICM), that allows not only the observations $\mathbf{x}_i$ to be contaminated independently of one another, but also the coordinates x_{i,k_1} to be contaminated independently of any other coordinates x_{i,k_2} within a given observation. Alqallaf et al. demonstrate that the MCD performs poorly under this contamination model: while MCD(γ^*) has asymptotic breakdown point 1/2 under the Tukey-Huber Model, it can exhibit a breakdown point near 0 under the ICM. Hence RSDs based on the MCD under the ICM might not be much more robust to outliers than Mahalanobis distances based on the sample mean and covariance. Robust estimators that build up an estimate of the dispersion matrix from consideration of pairs of observations are better suited to analyzing data whose outlier structure is more accurately captured by the ICM. For example, the Orthogonalized Gnanadesikan-Kettenring (OGK) robust dispersion estimator, developed by Gnanadesikan and Kettenring (1972), Devlin et al. (1981), and Maronna and Zamar (2002) is well-known estimator based on pairwise robust covariance analysis. (Appendix A of Green (2017) provides additional detail on the OGK estimator.) The quadrant correlation is another common robust dispersion estimate based on pairwise analyses (Huber, 1981).

In a previous paper (Martin et al., 2010) we used OGK-based RSDs to investigate the existence and prevalence of multivariate outliers in the type of financial data used to build fundamental factor models. Given the results of Cerioli et al. (2009) for the maximum-breakdown point version of the MCD, however, it was of interest to understand whether OGK and other robust dispersion-based estimates suffered from the same problem. In a companion study (documented in Appendix A of Green (2017)) we showed that several other robust dispersion estimates exhibit, to varying degrees, the problems with the RSD test for outliers that Cerioli et al. (2009) found for the MCD estimate. The results of the simulation show that the S-estimators and the OGK also suffer from inflated average false positive rates like the MCD, for both the individual and intersection tests. The OGK performs better than the MCD, in that average false positive rates for OGK-based RSDs are inflated much less than the rates for MCD-based RSDs, and the inflation factor is roughly independent of the dimension ν.

Thus, correction methodologies are also needed for other robust dispersion estimators such as S-estimators and the OGK estimate. A correction methodology for the OGK estimator would be valuable due to the comparative computational simplicity of the OGK in higher dimensions and its appeal in dealing with componentwise contamination scenarios. We are not aware of a correction procedure for the OGK, however, and the IRMCD method does not obviously apply as the OGK and MCD estimates have very different structure. Thus it seems for the time, OGK-based RSDs cannot be safely used for financial applications unless the sample sizes are large ($n \geq 500$). For the moment, MCD-based distances with the IRMCD procuedure are our only viable option for reliable RSD-based tests of outlyingness.

We have only considered outlier detection in a multivariate normal framework in this paper. Real data, especially financial data, often exhibit skewness and heavy tails that give rise to outliers. In such cases it becomes more difficult to define what an outlier is and to identify them in the data. An important research direction for the future is outlier detection in more general univariate and multivariate distributions such as elliptical and skewed elliptical distributions. We refer the reader to the recent book of Azzalini and Capitanio (2014) and the references therein for further discussion of the latter.

Extreme value theory has also proven to be quite useful for modeling skewed and heavy-tailed financial

data. Some initial work on the compatibility of robust methods and extreme value theory has been done by several authors. Vandewalle et al. (2004) showed how to construct a robust estimator of the tail-index of a Pareto-type distribution using robust regression techniques. Dell'Aquila and Embrechts (2006) showed how to use robust methods to construct estimators for extreme value distributions that are not highly influenced by observations that do not conform to same distribution as the bulk of the data. Goegebeur et al. (2014) proposed a robust estimator for extreme quantiles of heavy-tailed distributions. Additional research on applications of outlier detection in the context of extreme value models would be very beneficial to financial practitioners focused on risk management.

A Croux-Haesbroeck Formulas for the Asymptotic Variance of the MCD Estimate

Croux and Haesbroeck (1999) derive the influence function for the MCD estimate S_{MCD} under the assumption of observations with a multivariate elliptical distribution. This influence function can be used to calculate the variance of the MCD estimate, and hence, the variance of the diagonal elements s_{jj} that was needed to derive the method of moments estimate m in Section 3.1. Hardin and Rocke (2005) calculated the variance of the s_{jj} for the specific case of a multivariate normal distribution using the Croux-Haesbroeck result, and provided their formulas in an appendix to their paper. We reproduce these formulas here for the reader's convenience.

Here $\gamma = 1 - h/n$ is the (asymptotic) fraction of observations trimmed by the MCD as in the main text, and $q(\nu, 1-\gamma)$ is the $1-\gamma$ quantile of a χ^2_ν distribution and satisfies $1 - \gamma = P(\chi^2_\nu \leq q(\nu, 1-\gamma))$.

$$
\begin{aligned}
c(\nu,\gamma) &= \frac{1-\gamma}{P\left(\chi^2_{\nu+2} \leq q(\nu,1-\gamma)\right)} \\
c_2(\nu,\gamma) &= \frac{-P\left(\chi^2_{\nu+2} \leq q(\nu,1-\gamma)\right)}{2} \\
c_3(\nu,\gamma) &= \frac{-P\left(\chi^2_{\nu+4} \leq q(\nu,1-\gamma)\right)}{2} \\
c_4(\nu,\gamma) &= 3c_3(\nu,\gamma) \\
b_1(\nu,\gamma) &= \frac{c(\nu,\gamma)(c_3(\nu,\gamma) - c_4(\nu,\gamma)}{1-\gamma}
\end{aligned}
$$

$$
\begin{aligned}
b_2(\nu,\gamma) &= \frac{1}{2} + \frac{c(\nu,\gamma)}{1-\gamma}\left(c_3(\nu,\gamma) - \frac{q(\nu,1-\gamma)}{\nu}\left(c_2(\nu,\gamma) + \frac{1-\gamma}{2}\right)\right) \\
v_1(\nu,\gamma) &= (1-\gamma)b_1(\nu,\gamma)^2\left(\gamma\left(\frac{c(\nu,\gamma)q(\nu,1-\gamma)}{\nu} - 1\right)^2 - 1\right) - \\
&\qquad 2c_3(\nu,\gamma)c(\nu,\gamma)^2\Big(3\left(b_1(\nu,\gamma) - \nu b_2(\nu,\gamma)\right)^2 + \\
&\qquad\qquad (\nu+2)b_2(\nu,\gamma)(2b_1(\nu,\gamma) - \nu b_2(\nu,\gamma))\Big) \\
v_2(\nu,\gamma) &= \left(b_1(\nu,\gamma)\left(b_1(\nu,\gamma) - \nu b_2(\nu,\gamma)\right)(1-\gamma)\right)^2 c(\nu,\gamma)^2 \\
v(\nu,\gamma) &= \frac{v_1(\nu,\gamma)}{n v_2(\nu,\gamma)}.
\end{aligned}
$$

B Replicating the Hardin-Rocke Extension Simulations

The simulations used to build and to validate our Hardin-Rocke extension were performed on a 16-node computing cluster managed by the University of Washington Department of Statistics. Each node has an 8-core, Intel Xeon® E5410 2.33GhZ processor and 16GB of RAM, and runs Debian Linux 7.1. We used R 3.0.2 (64-bit) to conduct the simulations. We implemented the simulation and verification steps in two packages, `CerioliOutlierDetection` and `HardinRockeExtensionSimulations`, described below.

Data analysis, modeling, and plotting were performed on a laptop running Windows 7 Ultimate SP 1 with an Intel® Core™ i7-3740QM processor running at 2.7GHz and 32GB of RAM. A full listing of packages used (and their versions) is provided below to aid reproducibility of our results.

B.1 R Session Details

```
> sessionInfo()
R version 3.0.2 (2013-09-25)
Platform: x86_64-w64-mingw32/x64 (64-bit)

locale:
[1] LC_COLLATE=English_United States.1252
[2] LC_CTYPE=English_United States.1252
[3] LC_MONETARY=English_United States.1252
[4] LC_NUMERIC=C
[5] LC_TIME=English_United States.1252

attached base packages:
[1] parallel  stats     graphics  grDevices utils     datasets
[7] methods   base

other attached packages:
[1] HardinRockeExtensionSimulations_1.0      rrcov_1.3-4
[3] pcaPP_1.9-49                             mvtnorm_0.9-9997
[5] abind_1.4-0                              CerioliOutlierDetection_1.0.0
[7] robustbase_0.90-2

loaded via a namespace (and not attached):
[1] DEoptimR_1.0-1 stats4_3.0.2
```

B.2 The `CerioliOutlierDetection` R Package

This `R` package implements the outlier detection methodology of Cerioli (2010) based on Mahalanobis distances and the minimum covariance determinant (MCD) estimate of dispersion. It also implements the extension to Hardin and Rocke (2005) developed in this paper. The package is available on CRAN (Green and Martin, 2014).

B.3 The `HardinRockeExtensionSimulations` R Package

This package contains scripts to perform the simulations described in this paper. It can be downloaded via `git` or a web browser from Christopher Green's GitHub repository:

`http://christopherggreen.github.io/HardinRockeExtensionSimulations/`

The easiest way to install this package in `R` is via the `devtools` package:

```
> require(devtools)
> install_github("christopherggreen/HardinRockeExtensionSimulations")
```

C Simulated Degrees of Freedom and Consistency Factor

A table containing the Wishart degrees of freedom parameter m and consistency factor c calculated via simulation is available in the `HardinRockeExtensionSimulations` package described above. These values were used to fit the model shown in Equation (15).

D Full Results of Out of Sample Tests of Proposed Modification to Hardin and Rocke (2005) Methodology

Table 5 provides the out of sample results from testing the model shown in Equation (15). The table shows the ratio of the predicted degrees of freedom to the simulated degrees of freedom.

Table 5: Out of sample performance of the proposed improvement to the Hardin-Rocke methodology, as measured by the ratio of the predicted degrees of freedom to the simulated degrees of freedom. Blank cells correspond to combinations of n, ν, and γ that were not part of the out of sample tests. The data in this table is depicted in Figure 16.

		$\gamma =$												
ν	n	MBP	0.45	0.4	0.35	0.3	0.25	0.2	0.15	0.1	0.05	0.01	0.005	0
Cases where $n \leq 5\nu$														
2	8	1.426	1.242	1.367	1.485	1.116	1.194	1.272	0.906	0.971	1.059	1.187	1.217	0.683
3	12	0.962	0.997	1.092	0.938	1.012	0.881	0.948	1.024	0.839	0.932	1.061	1.090	0.769
5	20	0.932	0.963	0.955		1.007	0.974	0.928	1.006	0.957	0.913	1.032	1.057	0.866
8	32	0.958	0.936	0.969	1.001	1.014	0.959	0.962	0.971	0.983	0.998	1.010	1.030	0.922
11	44	0.953	0.928	0.943	0.979	0.965	0.955	0.988	0.971	1.003	0.989	1.027	1.044	0.974
11	50	0.967	0.962	0.966	0.986	0.988	0.993	0.992	0.987	0.984	0.976	1.020	1.038	0.978
16	50	0.979	0.956	0.970	0.974	1.006	1.003	1.003	1.035	1.029	1.025	1.059	1.074	1.018
16	64	0.978	0.959	0.986	0.988	1.006	0.997	1.016	1.036	1.018	1.036	1.042	1.057	1.014
22	50	1.002	0.948	0.982	0.981	1.008	1.006	1.033	1.058	1.050	1.079	1.110	1.123	1.068
22	88	0.985	0.960	0.985	1.007	1.007	1.023	1.035	1.029	1.046	1.060	1.063	1.076	1.049

Table 5: *(continued)*

		$\gamma =$												
ν	n	MBP	0.45	0.4	0.35	0.3	0.25	0.2	0.15	0.1	0.05	0.01	0.005	0
Cases where $5\nu < n \leq 10\nu$														
2	12	1.297	1.178	1.062	1.206	1.011	1.130	0.924	1.029	0.829	0.943	1.092	1.124	0.755
2	16	1.197	1.147	1.099	1.026	1.183	1.084	0.983	1.121	0.998	0.885	1.043	1.077	0.801
2	20	1.194		1.214	1.205	1.181	1.128	1.071	0.998	0.933	0.882	1.053	1.088	0.842
3	18	0.986	0.983	0.960	0.932	1.042	0.985	0.930	1.038	0.969	0.900	1.050	1.082	0.832
3	24	0.983	0.890	0.928	0.927	0.924	0.913	0.897	0.883	0.870	0.864	1.022	1.055	0.857
3	30	0.999	0.945	0.991	1.047	0.968	0.997	1.022	0.925	0.947	0.976	1.008	1.042	0.886
5	30	0.974	0.937		0.996	1.012	0.937	0.948	0.957	0.964	0.992	1.013	1.040	0.909
5	40	0.967	0.963	0.968	0.970	1.027	0.999	0.959	1.000	0.974	0.944	0.992	1.019	0.925
5	50	0.988	0.962	0.987	1.012	1.028	0.979	0.981	0.981	0.979	0.981	0.975	1.003	0.926
8	48	0.972	0.953	0.966	0.984	0.985	0.983	0.977	0.966	0.958	0.955	1.005	1.026	0.953
8	50	0.983	0.954	0.973	0.989	1.001	1.001	1.001	0.998	0.990	1.001	0.993	1.015	0.953
8	64	1.007	1.005	1.024	1.020	1.010	0.998	1.021	1.008	0.991	0.973	0.988	1.010	0.958
8	80	1.016		1.009	1.036	1.026	1.006	1.017	0.989	1.007	0.986	0.987	1.010	0.972
11	66	0.987	0.965	0.972	0.976	1.009	0.999	0.988	1.014	1.001	0.990	1.003	1.022	0.975
11	88	0.982	0.991	0.998	1.004	1.003	0.995	1.008	0.999	0.994	0.988	0.991	1.010	0.974
11	110	0.994	1.003	1.008	1.018	1.022	1.017	1.014	1.001	0.995	0.989	0.991	1.010	0.990
16	96	0.996	0.971	0.988	0.995	1.004	1.012	1.011	1.014	1.017	1.018	1.029	1.045	1.019
16	128	0.992	1.000	1.012	1.022	1.018	1.014	1.027	1.015	1.028	1.021	1.037	1.027	1.012
16	150	0.994	1.003	1.012	1.008	1.019	1.018	1.009	1.015	1.008	1.001	1.008	1.005	0.996
16	160	1.006		1.017	1.029	1.033	1.021	1.026	1.024	1.027	1.029	1.022	1.018	1.011
22	132	0.992	0.987	1.001	1.012	1.014	1.027	1.026	1.033	1.026	1.036	1.052	1.044	1.030
22	150	1.007	0.995	1.018	1.018	1.029	1.026	1.031	1.039	1.029	1.039	1.050	1.045	1.035
22	176	1.005	1.010	1.011	1.018	1.030	1.029	1.031	1.039	1.036	1.028	1.033	1.032	1.024
22	220	1.022		1.031	1.034	1.033	1.030	1.030	1.028	1.027	1.031	1.030	1.029	1.026

Table 5: *(continued)*

		$\gamma =$												
ν	n	MBP	0.45	0.4	0.35	0.3	0.25	0.2	0.15	0.1	0.05	0.01	0.005	0

Table 5: *(continued)*

		$\gamma =$												
ν	n	MBP	0.45	0.4	0.35	0.3	0.25	0.2	0.15	0.1	0.05	0.01	0.005	0
Cases where $10\nu < n \leq 20\nu$														
2	24	1.101	0.994	1.033	1.042	1.055	1.039	1.036	1.028	1.013	1.007	1.003	1.038	0.827
3	36	0.986	0.969	0.946	1.014	0.992	0.953	0.998	0.938	0.988	0.943	1.004	1.038	0.913
3	50	0.981	0.952	0.994	1.030	0.981	0.996	1.012	0.940	0.944	0.965	0.966	1.000	0.917
5	60	0.995		1.001	0.994	1.025	1.002	0.976	1.003	0.987	0.963	0.991	1.020	0.957
8	96	1.012	1.004	1.035	1.023	1.028	1.004	1.005	1.009	0.985	0.994	0.977	1.000	0.969
8	150	1.019	1.006	1.018	1.020	1.019	1.010	1.005	1.007	0.999	1.000	0.993	0.989	0.979
11	132	0.992	0.990	0.995	1.000	1.001	1.000	0.989	0.988	0.984	0.981	1.006	0.998	0.984
11	150	1.003	1.007	1.015	1.022	1.022	1.019	1.014	1.009	1.003	0.998	1.010	1.003	0.989
16	192	1.001	1.010	1.019	1.023	1.018	1.014	1.019	1.016	1.015	1.012	1.015	1.015	1.009
16	300	1.018	1.021	1.024	1.027	1.022	1.014	1.014	1.017	1.013	1.019	1.019	1.025	1.016
22	264	1.007	1.011	1.019	1.024	1.024	1.027	1.028	1.027	1.029	1.028	1.027	1.030	1.021
22	300	1.010	1.017	1.018	1.025	1.023	1.021	1.024	1.021	1.020	1.020	1.025	1.031	1.021

Table 5: *(continued)*

		$\gamma =$												
ν	n	MBP	0.45	0.4	0.35	0.3	0.25	0.2	0.15	0.1	0.05	0.01	0.005	0
Cases where $n > 20\nu$														
2	50	1.030	1.021	1.095	1.047	1.070	1.012	1.021	1.023	0.953	0.958	0.947	0.983	0.886
2	150	1.079	1.079	1.095	1.077	1.096	1.051	1.060	1.042	1.010	1.012	1.009	0.998	0.986
2	300	1.084	1.085	1.087	1.061	1.050	1.051	1.025	1.005	0.983	0.976	0.961	0.974	0.968
2	500	1.064	1.059	1.060	1.030	1.012	1.002	1.003	0.979	0.964	0.960	0.959	0.966	0.955
2	750	1.066	1.060	1.073	1.062	1.034	1.002	0.995	1.001	0.992	0.991	0.987	0.985	0.992
2	1000	1.040	1.023	1.006	1.002	0.988	0.988	0.996	0.992	0.979	0.974	0.973	0.969	0.972
3	150	1.038	1.029	1.043	1.051	1.022	1.015	1.013	0.987	0.975	0.971	0.973	0.960	0.946
3	300	1.037	1.043	1.036	1.024	1.013	1.006	1.015	0.990	0.971	0.971	0.969	0.984	0.977
3	500	1.037	1.033	1.027	1.025	1.027	1.023	1.035	1.014	1.004	0.997	0.980	0.981	0.979
3	750	1.023	1.028	1.024	1.017	1.018	0.999	1.007	0.993	0.993	0.985	0.974	0.975	0.971
3	1000	1.065	1.045	1.046	1.039	1.029	1.012	1.013	1.016	1.007	0.999	1.001	1.007	1.004
5	150	1.032	1.023	1.032	1.030	1.039	1.009	1.007	1.000	0.992	0.989	0.981	0.971	0.966
5	300	1.055	1.056	1.042	1.034	1.035	1.019	1.008	1.008	0.991	0.982	0.977	0.987	0.973
5	500	1.037	1.050	1.038	1.027	1.033	1.041	1.023	1.028	1.019	1.007	1.002	1.006	0.999
5	750	1.026	1.022	1.012	1.014	1.019	1.006	1.002	1.001	1.000	0.997	0.995	0.994	0.992
5	1000	1.032	1.035	1.031	1.028	1.016	1.012	0.997	0.991	0.985	0.974	0.970	0.973	0.972
8	300	1.042	1.040	1.039	1.042	1.032	1.012	1.010	0.993	0.990	0.989	0.980	0.988	0.978
8	500	1.022	1.022	1.018	1.020	1.013	1.004	1.000	0.987	0.989	0.987	0.981	0.984	0.976
8	750	1.022	1.020	1.018	1.017	1.015	1.011	1.011	1.004	1.003	0.996	0.989	0.988	0.985
8	1000	1.024	1.017	1.010	1.011	1.007	1.005	1.005	0.999	1.004	0.995	0.996	0.997	0.996

Table 5: *(continued)*

		$\gamma =$												
ν	n	MBP	0.45	0.4	0.35	0.3	0.25	0.2	0.15	0.1	0.05	0.01	0.005	0
11	300	1.012	1.016	1.022	1.018	1.020	1.017	1.012	1.010	1.004	1.004	0.999	1.005	0.994
11	500	1.012	1.014	1.017	1.015	1.011	1.005	1.007	1.003	0.998	0.999	0.993	0.997	0.992
11	750	1.022	1.016	1.015	1.013	1.017	1.010	1.005	1.000	0.998	0.993	0.997	0.996	0.993
11	1000	1.018	1.019	1.017	1.008	1.000	0.995	0.998	0.998	0.998	0.995	0.991	0.990	0.991
16	500	1.021	1.017	1.016	1.022	1.022	1.014	1.011	1.010	1.011	1.011	1.011	1.015	1.011
16	750	1.018	1.012	1.010	1.010	1.018	1.016	1.015	1.016	1.009	1.001	1.003	1.003	1.001
16	1000	1.022	1.018	1.018	1.015	1.015	1.008	1.009	1.006	1.006	1.003	1.000	1.000	1.000
22	500	1.015	1.020	1.024	1.024	1.024	1.022	1.018	1.017	1.014	1.013	1.013	1.016	1.011
22	750	1.021	1.026	1.027	1.026	1.024	1.023	1.022	1.025	1.021	1.018	1.018	1.017	1.015
22	1000	1.028	1.028	1.025	1.025	1.020	1.018	1.015	1.013	1.009	1.006	1.006	1.008	1.007

Table 6: Monte Carlo standard deviations of simulation tests of FSRMCD and IRMCD implementations. Standard errors for the quantities in Table 4 can be obtained by dividing the corresponding entries in this table by $\sqrt{5000}$.

Dimension	Method	$n = 40$	$n = 60$	$n = 90$	$n = 125$	$n = 200$	$n = 400$
$\gamma = \gamma^*$							
$\nu = 5$	FSRMCD	0.113	0.115	0.119	0.107	0.112	0.100
	IRMCD	0.123	0.105	0.113	0.105	0.102	0.092
$\nu = 10$	FSRMCD	0.150	0.111	0.095	0.100	0.091	0.089
	IRMCD	0.141	0.118	0.100	0.098	0.090	0.091
$\nu = 15$	FSRMCD	0.141	0.111	0.097	0.104	0.092	0.093
	IRMCD	0.149	0.103	0.097	0.108	0.093	0.092
$\gamma = 0.25$							
$\nu = 5$	FSRMCD	0.113	0.108	0.102	0.099	0.111	0.097
	IRMCD	0.115	0.118	0.108	0.108	0.101	0.102
$\nu = 10$	FSRMCD	0.113	0.112	0.110	0.118	0.101	0.100
	IRMCD	0.120	0.103	0.082	0.100	0.110	0.091
$\nu = 15$	FSRMCD	0.108	0.108	0.102	0.086	0.092	0.089
	IRMCD	0.111	0.107	0.109	0.095	0.100	0.099
$\gamma = 0.05$							
$\nu = 5$	FSRMCD	0.100	0.105	0.108	0.105	0.105	0.107
	IRMCD	0.105	0.109	0.101	0.102	0.102	0.100
$\nu = 10$	FSRMCD	0.106	0.106	0.115	0.097	0.109	0.101
	IRMCD	0.114	0.112	0.102	0.118	0.115	0.100
$\nu = 15$	FSRMCD	0.136	0.113	0.120	0.111	0.104	0.113
	IRMCD	0.131	0.105	0.122	0.097	0.109	0.093
$\gamma = 0.01$							
$\nu = 5$	FSRMCD	0.077	0.090	0.109	0.100	0.076	0.105
	IRMCD	0.076	0.095	0.089	0.087	0.100	0.105
$\nu = 10$	FSRMCD	0.085	0.097	0.071	0.101	0.092	0.098
	IRMCD	0.085	0.081	0.094	0.080	0.086	0.095
$\nu = 15$	FSRMCD	0.093	0.088	0.073	0.083	0.090	0.098
	IRMCD	0.090	0.085	0.092	0.103	0.092	0.099

E Standard Deviations for FSRMCD and IRMCD Simulation Tests

Table 6 provides standard deviations for the simulation results presented in Table 4. Standard errors for entries in the latter table can be calculated by dividing the corresponding entry of this table by $\sqrt{5000}$.

References

Claudio Agostinelli and Victor J. Yohai. Composite robust estimators for linear mixed models. *Journal of the American Statistical Association*, 111(516):1764–1774, 2017.

Fatemah Alqallaf, Stefan van Aelst, Victor J. Yohai, and Ruben H. Zamar. Propagation of outliers in multivariate data. *Annals of Statistics*, 37(1):311–331, 2009.

Adelchi Azzalini and Antonella Capitanio. *The Skew-Normal and Related Familites*. Cambridge University Press, Cambridge, UK, 2014.

C. Becker and U. Gather. The masking breakdown point of multivariate outlier identification rules. *Journal of the American Statistical Association*, 94(447):947–955, 1999.

C. Becker and U. Gather. The largest nonidentifier outlier: A comparison of multivariate simultaneous outlier identification rules. *Computational Statistical & Data Analysis*, 36:119–127, 2001.

Andrea Cerioli. Multivariate outlier detection with high-breakdown estimators. *Journal of the American Statistical Association*, 105(489):147–156, 2010.

Andrea Cerioli, Marco Riani, and Anthony C. Atkinson. Controlling the size of multivariate outlier tests with the MCD estimator of scatter. *Statistical Computing*, 19:341–353, 2009.

Christophe Croux and Gentiane Haesbroeck. Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. *Journal of Multivariate Analysis*, 71:161–190, 1999.

Rosario Dell'Aquila and Paul Embrechts. Extremes and robustness: A contradiction? *Financial Markets and Portfolio Management*, 20:103–108, 2006.

S. J. Devlin, R. Gnanadesikan, and J. R. Kettenring. Robust estimation of dispersion matrices and principal components. *Journal of the American Statistical Association*, 76(374):354–362, 1981.

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

Yuri Goegebeur, Armelle Guillou, and Andréhette Verster. Robust and asymptotically unbiased estimation of extreme quantiles for heavy tailed distributions. *Statistics and Probability Letters*, 87:108–114, 2014.

Christopher G. Green. *Applications of Robust Statistical Methods in Quantitative Finance*. PhD thesis, University of Washington, Department of Statistics, 2017.

Christopher G. Green and R. Douglas Martin. `CerioliOutlierDetection`: Outlier detection using the iterated RMCD method of Cerioli (2010), 2014.

J. Hardin and D. M. Rocke. The distribution of robust distances. *Journal of Computational and Graphical Statistics*, 14:928–946, 2005.

Peter J. Huber. *Robust Statistics*. John Wiley & Sons, New York, 1981.

Hendrik Lopuhaä. Asymptotics of reweighted estimators of multivariate location and scatter. *Annals of Statistics*, 27(5):1638–1665, 1999.

Martin Maechler. *covMcd()—Considerations about Generalizing the FastMCD*, 2016. Vignette included in the `robustbase` package. Accessed January 3, 2017.

P. C. Mahalanobis. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2:49–55, 1936.

K. Mardia, J. Kent, and J. Bibby. *Multivariate Analysis*. Academic Press, New York, 1979.

Ricardo Maronna, R. Douglas Martin, and Victor Yohai. *Robust Statistics: Theory and Practice*. John Wiley & Sons, New York, 2006.

Ricardo A. Maronna and Ruben H. Zamar. Robust estimates of location and disperion for high-dimensional datasets. *Technometrics*, 44(4):307–317, November 2002.

R. Douglas Martin, Andrew Clark, and Christopher G. Green. Robust portfolio construction. In John Guerard, editor, *Handbook of Portfolio Construction: Contemporary Applications of Markowitz Techniques*, pages 337–382. Springer-Verlag, London, 2010.

E. Pearson and C. Chandra Sekar. The efficiency of statistical tools and a criterion for the rejection of outlying observations. *Biometrika*, 28:308–320, 1936.

Daniel Peña and Francisco J. Prieto. Multivariate outlier detection and robust covariance matrix estimation. *Technometrics*, 43(3):286–300, 2001.

D. Rocke and D. Woodruff. Identification of outliers in multivariate data. *Journal of the American Statistical Association*, 91:1047–1061, 1996.

D. M. Rocke. Robustness properties of s-estimators of multivariate location and shape in high dimension. *The Annals of Statistics*, 24:1327–1345, 1996.

Peter J. Rousseeuw. Multivariate estimation with high breakdown point. In W. Grossmann, G. Pflug, I. Vincze, and W. Wertz, editors, *Proc. of the 4th Pannonian Symp. on Math. Stat., Bad Tatzmannsdorf, Austria, 4-10 September, 1983*, volume B, pages 283–297. Kluwer Academic Publishers, 1985.

Peter J. Rousseeuw and Katrien van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, 1999.

Peter J. Rousseeuw and Bert C. van Zomeren. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85(411):633–639, 1990.

Peter J. Rousseeuw and Bert C. van Zomeren. Robust distances: Simulation and cutoff values. In W. Stahel and S. Weisberg, editors, *Directions in Robust Statistics and Diagnostics*, volume 2. Springer-Verlag, New York, 1991.

G. A. F. Seber. *Multivariate Observations*. John Wiley and Sons, New York, 1984.

R. Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley, New York, 1980.

Zbynek Šidák. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633, 1967.

N. J. H. Small. Plotting squared radii. *Biometrika*, 65:657–658, 1978.

B. Vandewalle, J. Beirlant, and M. Hubert. A robust estimator of the tail index based on an exponential regression model. In M. Hubert, G. Pison, A. Struyf, and S. van Aelst, editors, *Theory and Application of Recent Robust Methods*, pages 367–376. Birkhauser, Basel, 2004.

S. Wilks. *Mathematical Statistics*. Wiley, New York, 1962.