Applications of Robust Statistical Methods in Quantitative Finance

Christopher George Green

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

University of Washington

2017

Reading Committee:

R. Douglas Martin, Chair

Andrew F. Siegel

Fang Han

Program Authorized to Offer Degree:
Statistics

University of Washington

Abstract

Applications of Robust Statistical Methods in Quantitative Finance

Christopher George Green

Chair of the Supervisory Committee:
Professor R. Douglas Martin
Applied Mathematics

Financial asset returns and fundamental factor exposure data often contain outliers, observations that are inconsistent with the majority of the data. Both academic finance researchers and quantitative finance professionals are well aware of the occurrence of outliers in financial data, and seek to limit the influence of such observations in data analyses. Commonly used outlier mitigation techniques assume that it is sufficient to deal with outliers in each variable separately. Such approaches can easily miss multivariate outliers, observations that are outlying in higher dimensions without being outlying in any individual variable. Robust statistical methods are a better approach to building reliable financial models in the presence of multivariate outliers, but they are unfortunately underused by academic researchers and practitioners.

This dissertation motivates greater use of robust statistical methods in quantitative finance research via two applications to outlier detection and asset pricing research. We first demonstrate the use of robust Mahalanobis distances (RSDs) based on the minimum covariance determinant (MCD) robust mean and covariance estimates to detect multivariate outliers in asset returns time series data and fundamental factor exposure data. We improve upon a result of Hardin and Rocke for approximating the distribution of such distances, and use our result to improve the accuracy of the Iterated Reweighted MCD (IRMCD) technique of Cerioli for testing MCD-based RSDs with sample sizes as small as $n = 60$ and with high-

efficiency versions of the MCD. We show that, with our improvements, outlier detection via RSDs combined with IRMCD is more accurate than both common univariate approaches and multivariate Mahalanobis distances based on the classical sample mean and covariance estimates.

Second, we illustrate the benefits of robust MM-regression for empirically testing factor-based asset pricing models by revisiting the classic 1992 asset pricing study of Fama and French with data updated through December 2015. Our analysis using cross-sectional robust MM-regression reveals the surprising extent to which influential outliers, mainly small firms with isolated large returns, drove some of the main conclusions of the Fama and French study. Specifically, we demonstrate that the relationship between average returns and firm size is positive for nearly all stocks. The negative relationship found by Fama and French and most other asset pricing studies arises from a small percentage, usually less than 2%, of small stocks each month with unusually large returns. Similarly, we find a significant and complex relationship between average returns and firm betas, in contrast to Fama and French's assertion of the lack of such a relationship. We furthermore find that there is a non-trivial interaction between beta and size that must be included in an asset pricing model to fully explain the relationship between average returns and beta. Finally, while we confirm the positive relationship between average returns and firm book-to-market ratios found by Fama and French, we also confirm results due to Loughan demonstrating that this relationship is only significant in smaller stocks. Overall our robust regression analysis demonstrates the danger of relying solely upon classical statistical methods, such as least squares regression, in empirical asset pricing studies and encourages the use of modern robust methods in asset pricing research.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Number | Page

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I would like to thank my family for their patience and unconditional support during nearly two decades of graduate school.

I would like express my gratitude to my advisor, Doug Martin, for his continued support over more than a decade and his seemingly endless supply of new ideas.

I would like to thank the members of my dissertation committee, namely, Professors Andrew F. Siegel, Fang Han, Eric Zivot, and Les Atlas, for their willingness to serve and for their helpful comments on my work. I would also like to thank Professors Marina Meilă, Werner Stuetzle, and Tilmann Gneiting for serving on past iterations of the committee and for providing comments on earlier versions of this dissertation.

I would like to thank the Department of Statistics for its support, both financial and moral, over the years, as well as for providing me with opportunities to learn new and interesting things.

I thank my employer, the Washington State Investment Board, for their support and flexibility. I am especially grateful for their willingness to allow me to take an extended sabbatical to complete my dissertation, and for providing me with a "real-world" prospective on my research.

I thank Dr. John B. Guerard, Jr., of McKinley Capital Management for providing data used in Chapter 3 as well as helpful comments on the dissertation.

I thank Lew Thorson of the UW Foster School of Business for his help with WRDS.

Parts of this research were supported by funding from the National Security Agency and the National Science Foundation (specifically, ATM-0724721 during Summer 2008), as well as from the UW Applied Mathematics Department.

Chapter 1

INTRODUCTION

1.1 Overview of Robust Statistics

Robust statistical methods are statistical procedures which are designed to perform well at an assumed model and at "small" deviations from this model. Such deviations are commonly due to outliers in the data or to misspecification of the underlying mechanism for the data. The field of robust statistics was developed largely in response to problems with classical methods (such as sample means and covariances) in the presence of outliers and asymmetric, heavy-tailed distributions. Robust inference in such situations can be more reliable than classical inference. Robust methods also serve a valuable diagnostic role: if one fits a model to data using non-robust and robust methods and the answers differ significantly, one immediately knows that there are some unusual observations that need to be investigated before any conclusions are drawn from the research. John Tukey (1979) expressed this sentiment well: "Just which robust and resistant methods you use is NOT important—what IS important is that you use SOME. It is perfectly proper to use both classical and robust/resistant methods routinely, and only worry when they differ enough to matter. BUT when they differ, you should think HARD."

Robust statistical methods have been studied in the statistical community since the 1960s, starting with the foundational papers of Tukey (1960) introducing the concept of efficiency robustness; Huber (1964, 1973) introducing the notion of an M-estimator; and Hampel (1968, 1974) introducing idea of the influence function. Early texts such as Huber (1981) and Hampel et al. (1986) cover the theoretical and historical underpinnings of robust statistics, while Rousseeuw and Leroy (1987) focused entirely on robust regression. Over the ensuing decades there have been substantial advances in the theory underlying robust methods, as well as significant computational improvements. Several new texts offer the novice an easy path to learning not only the basics of robust statistics but also the latest advances. Maronna et al. (2006) provides broad coverage of modern robust statistical theory and methods, including discussion of bias robust methods and advances in computational techniques. The forthcoming update (Maronna et al., 2017) will detail important new results that have appeared in the literature since 2006. The classic book of Huber (1981) has also been updated (Huber and Ronchetti, 2009) to reflect new results since the early 1980s.

At the heart of robust methods is the tradeoff between bias and variance. If we evaluate the performance of an estimate $\tilde{\theta}$ of a univariate parameter θ by its mean squared error (MSE), we can decompose the MSE into two terms: the variance of the estimate and the square of the bias of the estimate. Many robust methods seek to control one or both of these measures (possibly only asymptotically for large samples) in order to provide reliable estimates of θ in the presence of outliers or heavier-than-expected tails. For instance, we might design an estimator that is *minimax bias* optimal, in the sense that it minimizes the largest possible bias of the estimator over some set of possible data distributions. This also leads to the concept of the *breakdown point* of an estimator, which is roughly the fraction of observations that can be completely corrupted while still keeping the maximum bias of the estimate finite. Section 3.5 of Maronna et al. (2006) provides further background on bias optimality, as does Section 4.2 of this dissertation.

If the estimate is unbiased, then variance is the only quantity we need to control. Typically, there is a minimum possible asymptotic variance for an estimator in a given class of estimates (e.g., provided by the Cramér-Rao lower bound), so it is more convenient to control the variance of our robust estimate relative to this lower bound. The *efficiency* of a candidate estimate, relative to another, "optimal" estimate, is defined as the ratio of the optimal estimate's variance to that of our candidate estimate's variance. In common robust estimation situations the optimal estimate is the maximum likelihood estimate when the data are uncontaminated (e.g., arise from a multivariate normal distribution).

Generally, though, our set of possible estimators will be biased for θ, so we must balance bias and variance/efficiency in selecting an estimator. We can find estimators that have the desired balance via optimization. For example, we can minimize the maximum bias of the estimate subject to a constraint on efficiency when the data arise from a multivariate normal distribution with no contamination. Sections 3.4–3.6 of Maronna et al. (2006) provide further details on bias, efficiency, and the bias-efficiency tradeoff, as well as corresponding definitions for multivariate estimators. We will also give a brief overview of these topics in Section 4.2 of this dissertation.

The tradeoff between bias and efficiency has been key to the design of robust methods for regression and covariance estimation, the two main tools we will use throughout this dissertation. The need for a robust regression methodology comes from the known poor performance of least squares (LS) regression in the presence of outliers in either the response or explanatory variables. A single outlier can lead to arbitrarily large bias in the estimated regression coefficients $\boldsymbol{\beta}$ and hence the LS estimator has breakdown point 0 (Maronna et al., 2006). Even if such an outlier does not lead to the catastrophic failure of the estimated coefficients, it often yields very misleading regression results. The sensitivity of the LS estimator to outliers is a consequence of its loss function $\rho(u) = u^2$, which magnifies the importance of large residuals in the LS objective (the sum of the squared residuals). This allows the maximum asymptotic bias of the LS estimator to become arbitrarily large. Many attempts at improving this situation have focused on minimizing the sum of a different function of the residuals, one that puts less weight on very large residuals. Huber (1973) introduced the concept of a *regression M-estimator*, an estimator for $\boldsymbol{\beta}$ that minimizes the sum of a function $\rho(u)$ of the residuals that grows more slowly than the quadratic u^2 or is bounded for large values of u.

Huber's original proposed $\rho(u)$ function (detailed later in Section 4.2) and similar functions that are not bounded are robust to outliers in the response variables, but not in the explanatory variables. Such estimates still have breakdown point 0 like the LS estimate. In fact, Martin et al. (1989) showed that loss functions must be bounded to limit the bias that could be caused by outliers. This shortcoming of early M-estimators led to the development of other robust regression approaches, such as least trimmed squares (Rousseeuw, 1984) and S-estimates (Rousseeuw and Yohai, 1984), which can have breakdown points as high as 1/2 and hence can be very robust to outliers in any of the variables. These estimators were not very efficient, however, meaning that the standard errors of the resulting regression coefficient estimates would be much larger than those of the LS estimates when the residuals were normally distributed.

The regression *MM-estimator*, developed by Yohai (1987) and refined by Yohai et al. (1991), offered a better solution to these problems. The MM-estimator combines a high-breakdown point initial S-estimate with a high efficiency final M-estimate. The choice of loss function $\rho(u)$ in each step is key to obtaining an estimator with both properties. Yohai and Zamar (1997) and Svarc et al. (2002) derived a minimax bias optimal loss function that minimizes the maximum asymptotic bias (under certain types of departures from normality) while ensuring a minimum efficiency when the data are normally distributed. With proper choice of tuning constants the resulting MM-estimator can have breakdown point 1/2 and high efficiency. The MM-estimator is available in common statistical software packages such as `R`, `SAS`, and `Stata`, and we will use it in Chapter 4 extensively. (Section 2 of that chapter provides a more detailed explanation of the MM-estimator and its properties.)

Like the LS estimate of the regression coefficents, the sample covariance matrix is known to be susceptible to outliers in the observations. One can trace a path through the development of robust estimates of the dispersion matrix similar to that of robust regression. Maronna (1976) developed multivariate M-estimators of dispersion by generalizing the structure of the maximum likelihood estimator of the dispersion matrix of an elliptical distribution. This estimator unfortunately has low breakdown point (at most $1/(\nu+1)$ for ν-dimensional data) as shown by Huber (1977, 1981) and Stahel (1981). The search for higher breakdown point estimators lead to several estimators whose goal is to minimize some measure of the dispersion of the Mahalanobis squared distances (MSDs) of the observations. The minimum volume ellipsoid estimator of Rousseeuw (1983, 1984) minimizes the median of the MSDs. Davies (1987) introduced S-estimators of dispersion, which minimize a smooth bounded function $\rho(u)$ of the MSDs. Rousseeuw (1985) introduced the minimum covariance determinant (MCD) estimator, which minimizes a trimmed mean of the MSDs. The MVE, S-estimators, and the MCD can all be tuned to yield an estimator with breakdown point 1/2. The MVE is very inefficient, however, and is no longer commonly used. The MCD is more efficient than the MVE for the same choice of tuning parameter (the subset size), but one must still sacrifice some efficiency for higher breakdown point. The efficiency of an S-estimate depends on the choice of loss function $\rho(u)$. (Chapter 6 of Maronna et al. (2006) provides more details on all of these estimators.) These estimators are also available in common statistical packages.

High-breakdown point robust dispersion estimators are commonly used in outlier detection settings, particularly for detecting outliers via MSDs. Of the MVE, S-estimators, and the MCD, the MCD has historically been the most commonly used esitmator for this purpose, due to a wealth of literature on MCD-based MSDs and the existence of a fast approximate algorithm to compute the MCD (Rousseeuw and van Driessen, 1999). It is not always the best dispersion estimator of the lot, however: Maronna et al. (2006) show, via a simulation experiment in their Section 6.8, that for certain choices of $\rho(u)$ the corresponding S-estimator offers a better balance of bias and variability than the MCD for estimating location and dispersion under a point-mass contaminated multivariate normal model. Outlier detection using MSDs based on S-estimators might therefore be more accurate than detection with MCD-based MSDs. In practice, however, the sampling distribution of MSDs based on a robust estimate of dispersion has only been well-studied in the MCD case. Work by Hardin and Rocke (2005), Cerioli et al. (2009), Cerioli (2010), and others has led to a calibrated MCD-based detection methodology that has the correct false positive rates for testing observations for outlyingness. This methodology does not obviously apply to distances based on S-estimators. We therefore employ these calibrated MCD-based robust squared distances in our outlier detection work (Chapters 2 and 3).

1.2 *Motivation for this Dissertation*

The volatile nature of financial markets ensures that financial data will almost always contain some observations that are seemingly inconsistent with the rest of the data, be they data errors or legitimate but one-time events such as market crashes, unanticipated mergers, and natural disasters. This is supported by a wealth of empirical evidence of outliers in asset returns data and in factor exposures data, e.g., see Chapter 6 of Scherer and Martin (2005) or Martin et al. (2010), as well as numerous research papers proposing non-normal distributional models for asset returns.[1] Quantitative finance professionals involved in portfolio construction and management are well-aware of the presence of outliers in financial data and the damage they can potentially cause.

Given the maturity of robust statistical methods and the availability of high-quality software implementations, one might expect that robust methods would be part of every practitioner's modeling toolbox. The quantitative finance community, however, has historically been largely unaware of robust statistical methods and/or unsure of how to use them. Outlier mitigation methodologies in quantitative finance have been limited to univariate approaches such as trimming or Winsorizing each variable separately. Such one-dimensional outlier mitigation methods are not adequate for dealing with multivariate outliers, that is, observations that are outlying in higher-dimensional views of the data without being outlying in any specific marginal variable.

Robust statistical methods have not been used very much in academic financial research either, except in very simple single factor models.[2] While there are a substantial number

---

[1] Some of the many models in the literature include mixtures of normal distributions (e.g., McNeil et al. (2005)), skewed-t distributions (e.g., Azzalini and Capitanio (2014)), and α-stable distributions (e.g., Rachev and Mittnik (2000)).

[2] It is difficult to estimate how many quantitative finance and econometrics papers use robust statistics due to the fact that the word "robust" has several different meanings in finance and econometrics. For example, standard errors can be robust to heteroskedasticity and autocorrelation without being robust to outliers. A methodology might be described as "robust" to extreme values without being robust in the statistical sense we use here.

of academic papers applying robust regression to the estimation of CAPM betas,[3] robust regression has seen very little published use in the empirical evaluation of multiple factor-based asset pricing models. The three papers Knez and Ready (1997), Chou et al. (2004), and Garza-Gómez et al. (2001), which all use least trimmed squares regression, as well as the recent paper Winker et al. (2011) using least median squares regression, are the only examples of which we are aware that apply robust regression in an empirical asset pricing context.[4]

In light of the above observations, a high-level goal of this dissertation is to encourage more widespread use of robust statistical methods in quantitative finance. One step towards this goal is to provide a reliable robust method for detecting multivariate outliers in the type of asset returns and factor exposure data used for portfolio management and construction and in empirical asset pricing studies. We do this via the introduction of an improved method for detecting multivariate outliers and the demonstration of its superiority to existing techniques. We illustrate the utility of our method for detecting unusual times in multivariate returns data and unusual assets in factor exposure data.

We then turn our attention to applications of robust regression to empirical asset pricing studies. We use MM-regression to show how outliers in cross-sectional returns and factor exposures can distort risk premia estimated via least squares. In some cases, our conclusions

---

[3]Early work in this area by Sharpe (1971) and Cornell and Dietrich (1978) employed the least absolute deviation regression. Connolly (1989) and Bowie and Bradfield (1998) used regression M-estimators in their studies of CAPM betas. More recent papers by Martin and Simin (2003) and Bailer et al. (2011) use regression MM-estimators. Also worth mentioning are the Theil-Sen robust regression methodology (Theil, 1950a,b,c; Sen, 1968) used by Philips (2012) and a Bayesian approach due to Genton and Ronchetti (2008) that offers a compromise between a robust estimate of beta and the least squares estimate. The Theil-Sen regression only has a breakdown point of 29.3% in a single factor model context, however. Siegel (1982) discusses a robust regression methodology based on repeated medians that has an asymptotic breakdown point of 50%.

[4]Robust statistical methods have been used more widely in economics and econometrics, though they are still not commonplace. Zaman et al. (2001) and Cížek and Härdle (2008) provide (somewhat dated) surveys of robust regression in econometrics. Sapra (2003) presents three applications of the S-estimator of regression. Bramati and Croux (2007) looks at applications of robust regression to panel data. Atkinson (2009) revisits Zaman et al. (2001) using the forward search method for outlier detection. Colombier (2009) applies MM-regression to an investigation of fiscal policies.

from MM-regression about whether certain risk factors are priced are very different from those obtained using least squares. Overall, the results of our analyses speak to the value of robust methods in quantitative financial research, and will hopefully lead to greater adoption of such methods by the academic and industry communities.

1.3 Multiple Factor Models

Multiple factor models are linear regression models with one of the following three forms.

- Time series factor models are based on observable time series of changes in macroeconomic measures or returns on market indices, hedge funds, or other portfolios. In the case of macroeconomic measures, these time series factor models are usually called macroeconomic factor models.

- Cross-sectional factor models are based on unobservable returns for firm characteristics such as market capitalization, industry sector, country of domicile, or accounting measures such as sales and the ratio of earnings to price. Such factors are used in asset pricing applications, portfolio optimization, and risk management.

- Latent or statistical factor models are based on factors that may not correspond to any recognizable financial measure, but provide a meaningful statistical explanation of asset returns.

The general form of a factor model is

$$r_{i,t} = \alpha_i + \beta_{1,i} f_{1,t} + \beta_{2,i} f_{2,t} + \cdots + \beta_{K,i} f_{K,t} + \epsilon_{i,t}, \qquad \epsilon_{i,t} \sim N(0, \sigma_i^2). \tag{1.1}$$

Here $r_{i,t}$ is the known return on asset i, $i = 1, \ldots, N$, at time t, $t = 1, \ldots, T$. The intercept, α_i, captures the expected asset return when the factor returns are zero or when the factor returns are not linearly related to the asset returns. $f_{k,t}$ is the return on the kth factor at time t. $\beta_{k,i}$ is the factor beta (also known as factor loading or factor exposure) for asset i on the kth factor. Finally, the residual term $\epsilon_{i,t}$ represents an asset-specific return that is not accounted for by the model. Further details on the model and its assumptions can be found in any standard investment management or financial economics textbook, e.g., Zivot and Wang (2003).

In the case of a macroeconomic multiple factor model, the $f_{k,t}$ series are known, and multiple linear regression across time for each asset is used to estimate the unknown coefficients $\beta_{k,i}$ of the model. For fundamental factor models, the coefficients $\beta_{k,i}$ are known, and multiple linear regression across assets for each time is used to estimate the unknown factor returns $f_{k,t}$. In a statistical factor model, neither the factor betas nor the factor returns are known a priori. Principal components analysis and related techniques are often used to compute both pieces of the model simultaneously. Combinations of these types of models are also possible, though they are more challenging to estimate. Stroyny (2005) provides some guidance on how to build such a model.

Factor models can be "single factor models" if asset returns are assumed to have only one common driver of risk/return, or "multiple factor models" if there is more than one factor. The Capital Asset Pricing Model (CAPM), developed by Sharpe (1964), Treynor (1961), Lintner (1965), and Mossin (1966), is the prototypical single factor model, with the return on the market portfolio serving as the sole factor.

Multiple factor models are popular for forecasting asset returns and asset risks (e.g., as inputs to a portfolio construction process),[5] as well as explaining the drivers of past portfolio risk and performance. Factor models reduce the number of parameters to estimate when forecasting returns or estimating asset covariances. It is often easier for the practitioner to formulate opinions about the future direction/level of the factors (e.g., oil prices or a firm's revenue) than about the assets themselves. Furthermore, the inherent dimensionality reduction of a factor model improves the numerical stability of covariance estimation in high dimensions. For example, the sample covariance matrix of 1000 assets can be difficult to compute due to its size and too numerically unstable to be useful. The covariance matrix

[5] See, for instance, the books Zivot and Wang (2003); Grinold and Kahn (2000); Litterman (2003); Tsay (2005); Elton et al. (2006); Chincarini and Kim (2006); Connor et al. (2010); Campbell et al. (1997).

could be estimated in a much more stable way using an estimated covariance matrix for a small number of factors and the structure of the factor model.

Least squares regression is commonly used to estimate factor models models due to its ease of use and optimality—the least squares estimate of the vector of regression coefficients $\boldsymbol{\beta}$ is the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$ (as a consequence of the Gauss-Markov theorem), and has the smallest variance of any unbiased estimate of $\boldsymbol{\beta}$ when the residuals are normally distributed.[6] Least squares regression, however, is known to be sensitive to the presence of outliers in the independent and dependent variables (e.g., see Ruppert and Carroll (1980); Koenker (1982); Rousseeuw and Leroy (1987)). This sensitivity arises from the fact that the objective of least squares is to minimize the sum of the squared residuals. A large residual due to an outlier will make the sum of all the squared residuals much larger, so the optimization algorithm will adjust the regression coefficients to reduce the size of this residual. In the process, other residuals may become larger, meaning the resulting regression model fits those observations worse than it would if the outlier were not there.

Outliers can arise quite easily in the factor model estimation process. For instance, in the CAPM both the historical asset return series r_t and the historical market return series $r_{m,t}$ might contain outliers due to asset-specific events and market-wide events. These outliers might distort the estimated beta to the extent that the fitted model is not representative of the behavior of the asset at any point in time. As another example, fundamental factor models often have many factors that are derived from the same quantities, e.g., several ratios of accounting measures to stock price. A price outlier can thus lead to an outlier in several variables. Such multivariate outliers can distort the estimated factor model, leading to poor return forecasts and a misallocation of risk between the common factors and the asset-specific residuals.

We can consider replacing least squares regression in the factor model estimation process with some form of robust regression to limit the impact of outliers. For asset returns forecas-

[6] Briefly, the best linear unbiased estimator of $\boldsymbol{\beta}$ has the smallest variance of all linear unbiased estimators of $\boldsymbol{\beta}$.

ting, several papers by John Guerard and colleagues (Bloch et al., 1993; Guerard et al., 2015; Guerard, 2016; Guerard et al., 2016) have explored the use of robust regression based on the Tukey biweight function. (We will use one of these factor models in Chapter 3.) On the other hand, there is little published work exploring how to construct multiple factor models for risk forecasting using robust regression. The commercial risk model vendor Axioma uses a Huber M-estimator to construct their Axioma Robust Risk Model (Guerard, 2017). The Huber M-estimator is not robust to outliers in the independent variables, however.

Another approach to dealing with potential outliers in single and multiple factor models is to detect outliers in the independent variables (the factor returns or factor betas) using an outlier detection method, and then remove these outliers or shrink them to more reasonable values. For example, MSCI uses Winsorization in the calculation of their style-based equity indices (MSCI, 2016, page 10). Stephan et al. (2001) constructed a multiple fundamental factor model for European stocks using the skipped Huber method applied to each factor beta. Approaches like these are commonly used by practitioners but only address extreme observations one variable at a time. They do not address the problem of multivariate outliers that, as we shall see in Chapter 3, are often present in the data.

Empirical Asset Pricing Models

Asset pricing models are used to estimate a "fair" value for an asset that compensates an investor for the risk of the asset. There are many pricing models based on single and multiple linear factor models. CAPM is the most well-known single factor asset pricing model. The Arbitrage Pricing Theory (APT) developed by Ross (1976) and Roll and Ross (1980, 1984) and the 3-factor model of Fama and French (1993) are common examples of multiple factor asset pricing models.

The main focus of empirical asset pricing is developing and testing such models on observed market data. For example, if CAPM is true, stock returns and stock betas will be linearly related. We might therefore validate CAPM by regressing stock returns on their betas and testing whether the intercept and slope in the regression are signficantly different from zero.

A significant non-zero intercept would suggest there is some factor other than beta that is needed to explain the variability in stock returns. This idea is the heart of the Fama and MacBeth (1973) or cross-sectional regression approach to testing factor-based asset pricing models: at each point in time, stock returns are regressed on their factor betas, resulting in a time series of regression coefficients. We compute averages of the regression coefficients and test the average coefficients for signficance. If the average regression coefficient on a given factor beta is not significant, the corresponding factor was not relevant for pricing returns over the time period considered.

The original Fama and MacBeth (1973) study was designed to test CAPM. Throughout the late 1970s and 1980s, many researchers found that beta could not fully explain the cross-section of U. S. stock returns, contrary to the assertion of CAPM. Other firm characteristics such as firm size (Banz, 1981; Keim, 1983), the book-to-market ratio (Stattman, 1980; Rosenberg et al., 1985; Chan et al., 1992),[7] firm leverage (Bhandari, 1988), and the ratio of firm earnings to stock price (Basu, 1983) were shown to have clear influence on average returns. These other characteristics were known as "anomalies" in that they did not fit the CAPM. In a famous paper, Fama and French (1992) used cross-sectional least squares regression to investigate the above CAPM anomalies more extensively. They concluded that firm size and the book-to-market ratio were the most important missing factors from the CAPM (out of the four mentioned above), and used this insight to develop their 3-factor asset pricing model in Fama and French (1993).

Since least squares regression is not robust to outliers, however, it is possible that the empirical tests of an asset pricing model using the Fama-MacBeth technique will be distorted by outliers. Replacing the least squares regression with some form of robust regression would render an asset pricing study less susceptible to erroneous conclusions driven by outliers. Indeed, Knez and Ready (1997) used an early robust regression method, least trimmed squares regression, to illustrate the effects of outliers in the Fama and French (1992) study.

[7] The book-to-market ratio is the ratio of a firm's book equity (i.e., its value from an accounting standpoint) to its market value.

While Fama and French (1992) found a negative relationship between average returns and firm size in the U. S., Knez and Ready (1997) determined that the relationship is positive for most stocks. Furthermore, they found that a small number of firms and a small number of months with unusual returns were responsible for the negative relationship found by Fama and French. Knez and Ready's findings were confirmed over a longer time period by Chou et al. (2004), also using least trimmed squares in U.S. equity markets through 2001. Garza-Gómez et al. (2001) applied a similar cross-sectional least trimmed squares regression to the Japanese equity markets through 1995, and also found that the relationship between average returns and firm size was strongly driven by a handful of unusual firms and time periods.

Least trimmed squares regression is very robust to outliers, but is not very efficient compared to least squares. Other robust regression approaches like MM-regression offer a better compromise of robustness to outliers and efficiency when the data are normally distributed. Bailer and Martin (2007) explored the use of cross-sectional robust MM-regression for testing factor-based asset pricing models, but there does not seem to be any other documented use of MM-regression in the asset pricing literature. We will use MM-regression in Chapter 4 to revisit the Fama and French (1992) study and extend the analysis through 2015.

1.4 Multivariate Outlier Detection

Multivariate outlier detection has recently found its way into financial applications, albeit in a limited capacity. The few papers that have been published in this area have used Mahalanobis squared distances for outlier detection. The Mahalanobis squared distance (MSD) measures how far away an observation is from the center of the data, taking into account the relative dispersion of each variable (by weighting distances from the center using the inverse square root of the covariance matrix). Formally, the MSD is defined as

$$D^2 \equiv (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}), \tag{1.2}$$

where $\mathbf{x}$ is an observation, $\boldsymbol{\mu}$ is the mean of the observations, and $\boldsymbol{\Sigma}$ is the covariance matrix of the observations. In practice we must replace the unknown true mean and covariance with estimates, such as the sample mean and covariance. We can then identify outliers by looking for observations whose squared distance is larger than a threshold determined by the distribution of D^2.

Chow et al. (1999) is one of the earliest financial studies to advocate the use of MSDs based on the sample mean and covariance matrix for identifying multivariate outliers. Kritzman and Li (2010) used such distances to identify outlying or "turbulent" time periods in multivariate financial time series.

Since sample means and covariances can be unduly influenced by outliers, MSDs based on these estimates may be misleading. The resulting distances could incorrectly identify valid observations as outliers or fail to identify true outliers. Robust mean and covariance estimates can potentially improve upon the situation: the robust estimates would be less affected by outliers and hence would better represent the center and dispersion of the non-outlying bulk of the data. So-called robust squared distances (RSDs) would, in principle, do a better job of detecting multivariate outliers. For instance, Boudt et al. (2008) used RSDs based on a robust mean and covariance estimate to detect and shrink outliers in data prior to estimating value-at-risk and expected shortfall. There have not been many other applications of such "robust distances" in finance, however. We will present some motivating examples in Chapter 3.

1.5 Other Potential Applications of Robust Methods in Quantitative Finance

Generally, anywhere mean and covariance estimates are needed, one can consider using robust mean and robust covariance estimates instead of the usual sample mean and covariance. Scherer and Martin (2005) devotes an entire chapter to such applications in portfolio optimization. The review of Martin et al. (2010) presents more extensive applications to equity portfolio management.

We note in passing that our use of the term "robust" is not related to the use of the term "robust" in so-called "robust portfolio optimization" as developed by Goldfarb and Iyengar (1993); Erdogan et al. (2004); Ceria and Stubbs (2006); Garlappi et al. (2007) and critiqued in Scherer (2007). We use the term "robust" to refer to estimators that are resistant to outliers and model misspecification, while robust portfolio optimization refers to a method of constructing an optimal portfolio in the presence of parameter uncertainty.

1.6 Main Contributions of this Dissertation

This dissertation is divided into two parts. The first half (Chapters 2 and 3) focuses on using robust squared distances (RSDs) to reliably detect multivariate outliers with a specified false positive rate in the type of asset returns and factor exposure data used to conduct empirical asset pricing studies and to construct and manage equity portfolios.

Chapter 2 develops a new approach to calibrating RSDs based on the minimum covariance determinant (MCD) that improves upon the prior approach of Hardin and Rocke (2005) in sample sizes less than 250 and when the MCD uses a fraction of the observations greater than the one achieving the highest breakdown point of approximately 1/2. This improvement, when combined with the Iterated Reweighted MCD technique developed by Cerioli (2010), makes MCD-based RSDs more accurate for the data set sizes typically encountered in portfolio management and asset pricing research applications.

In Chapter 3 we use the improved detection method developed in Chapter 2 to illustrate how the standard MSDs based on the sample mean and covariance can fail to detect many multivariate outliers in financial data. We show that our RSD approach identifies many more multivariate outliers in asset returns and factor exposure data, outliers missed by the standard approach. We demonstrate our approach on multivariate hedge funds and commodities portfolios, where outlying times might indicate a breakdown or change point of the usual relationships between assets; and on factor exposure data, where outlying assets correspond to firms with unusual market or accounting data. We also show how one-dimensional outlier approaches like trimming and Winsorization can miss multivariate outliers, and argue that our multivariate approach is strongly preferable to these one-variable-at-a-time approaches.

In the second part of this dissertation, we show that a theoretically well-justified robust MM-regression should be used as a complement to least squares in an empirical asset pricing context. The robust regression will reveal which conclusions about risk premia are driven by highly influential outliers. Chapter 4 revisits the classic paper of Fama and French (1992) (FF92) using MM-regression and extends the analysis through the end of 2015. Contrary to FF92, we show that beta is still a significant predictor of average stock returns. We verify that, when all stocks are considered, average returns decrease with firm size as documented by FF92, but show that this result is driven by smaller firms with large isolated returns. When the influence of such outliers is controlled via cross-sectional MM-regression, average returns increase with firm size for most stocks. We show that the value effect, that average returns increase with a firm's book-to-market ratio, holds even in the robust regression case. We also demonstrate that this effect is largely confined to small stocks in modern financial markets: the effect vanished from moderately-sized stocks after 1980, and was never present in large stocks. Somewhat surprisingly, we show that the relationship between average stock returns, beta, and firm size is non-linear: there is a non-trivial interaction between beta and size that must be captured in an asset pricing model to explain the cross-section of average returns. The results in this chapter are strong evidence of the utility of optimal robust MM-regression for asset pricing work, and should encourage the use of such robust regression in other popular asset pricing models.

Chapter 5 summarizes the dissertation and offers suggestions for future research.

Appendix A summarizes a preliminary experiment done prior to Chapters 2 and 3. This simulation study extends work done by Cerioli et al. (2009) for MCD-based RSDs to distances based on several other robust dispersion estimates. Cerioli et al. (2009) pointed out how MCD-based distances could have incorrect false positive rates in small sample sizes and higher dimensional data, and motivated the calibration method developed in Cerioli (2010). We show that RSDs based on three other dispersion estimates also suffer from the same problem, to varying degrees, and are in need of calibration. The results of this study motivated our decision to use calibrated MCD-based RSDs in Chapter 3 rather than RSDs based on other dispersion estimates, as well as our work in Chapter 2 to improve the Hardin-Rocke methodology for our purposes.

Chapter 2

CALIBRATED MINIMUM COVARIANCE DETERMINANT ROBUST DISTANCES

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. In this chapter, 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.

2.1 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

$$D^2 \equiv (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}). \tag{2.1}$$

Compare this definition to the usual Euclidean distance of $\mathbf{x}$ from $\boldsymbol{\mu}$:

$$(\mathbf{x} - \boldsymbol{\mu})^T (\mathbf{x} - \boldsymbol{\mu}).$$

The extra $\boldsymbol{\Sigma}^{-1}$ factor captures the fact that the distribution may not look the same in each direction. For example, it may be more dispersed in one direction, so an observation that is far from the mean in a Euclidean sense may not be "unusually" far away once covariance information is taken into account.

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

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T, \tag{2.2}$$

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 (2.2) 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 (2.1) 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 $\boldsymbol{\Sigma}$ (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.)

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

$$H_0 : \{\mathbf{x}_1 \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})\} \cap \cdots \cap \{\mathbf{x}_n \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})\} \tag{2.3}$$

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 2.4.3) 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 2.2 reviews technical details on the MCD estimate, the Hardin-Rocke distributional approximation, and Cerioli's IRMCD procedure. Section 2.3 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 2.4 presents several tests of our model. Section 2.5 concludes with a discussion of potential future improvements.

2.2 Technical Background

2.2.1 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.[1] 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

$$\begin{aligned} h &= \lfloor 2n_2 - n + 2(n-n_2)(1-\gamma) \rfloor \\ &= \lfloor (2n_2 - n)\gamma + n(1-\gamma) \rfloor , \end{aligned} \tag{2.4}$$

where

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

[1]For instance, when the observations come from a 5-dimensional multivariate normal $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and contain no outliers, the maximum breakdown point version of the MCD is only about 13% efficient for estimating diagonal elements of the true covariance matrix $\boldsymbol{\Sigma}$ compared to the usual sample covariance estimate, and 11% efficient for the off-diagonal elements.

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

$$\begin{aligned} h &= \left\lfloor n - \left(n - 2\left\lfloor \frac{\nu+1}{2} \right\rfloor \right) \gamma \right\rfloor \\ &= n - \left\lceil \left(n - 2\left\lfloor \frac{\nu+1}{2} \right\rfloor \right) \gamma \right\rceil. \end{aligned} \tag{2.5}$$

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 (2.4) ensures that in smaller samples the value of h computed using (2.4) will be strictly smaller than n, even if γ is very small. In the n even case, rearranging (2.5) yields

$$n - h = \left\lceil \left(n - 2\left\lfloor \frac{\nu+1}{2} \right\rfloor \right) \gamma \right\rceil. \tag{2.6}$$

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 (2.6) 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.

2.2.2 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:

$$\frac{m}{c} S_{MCD} \sim \text{Wishart}_\nu(m, \boldsymbol{\Sigma}), \tag{2.7}$$

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 (2.2) 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:

$$\frac{c(m-\nu+1)}{m\nu} D^2_{S_{MCD}}\left(X_i, \bar{X}_{MCD}\right) \sim F_{\nu, m-\nu+1}. \tag{2.8}$$

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

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

In order to use the distribution (2.8) 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.

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

Hardin and Rocke note that if S_{MCD} has the scaled Wishart distribution (2.7), 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.

$$\begin{aligned} E\left[mc^{-1}s_{jj}\right] &= m \\ \text{Var}\left(mc^{-1}s_{jj}\right) &= 2m \\ CV = \frac{\sqrt{\text{Var}(s_{jj})}}{E(s_{jj})} &= \frac{c\sqrt{2/m}}{c} = \sqrt{\frac{2}{m}} \end{aligned} \tag{2.9}$$

where CV is the coefficient of variation. Therefore

$$m = \frac{2}{CV^2}. \tag{2.10}$$

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.[2] Here $\gamma \approx 1 - h/n$ is the approximate fraction of observations trimmed by the MCD as in Section 2.2.1.

[2]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

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

where $q(\nu, 1-\gamma)$ is the $1-\gamma$ quantile of a χ^2_ν distribution and satisfies $1-\gamma = P(\chi^2_\nu \leq q(\nu, 1-\gamma))$. The asymptotic consistency constant c_{asy} is defined as the reciprocal of $c(\nu, \gamma)$:[3]

$$c_{\text{asy}} = 1/c(\nu, \gamma). \tag{2.11}$$

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 2.A.) Thus from (2.10) we have

$$m_{\text{asy}}(n, \nu, \gamma) = \frac{2}{c(\nu, \gamma)^2 v(\nu, \gamma)}. \tag{2.12}$$

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., $30 \leq n \leq 250$). 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 $\mathrm{N}(\mathbf{0}, \mathbf{I})$.

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

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 (2.10) 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.

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

$$\log\left(\frac{m}{m_{\text{asy}}(n, \nu, \gamma^*)}\right) = 0.725 - 0.00663\nu - 0.0780\log(n). \qquad (2.13)$$

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 (2.13) 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 (2.13) 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.

2.3.2 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 (2.8). Figure 2.1 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 (2.13) 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.[4] 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

[4] Additional details on the simulation computations are available in Appendix 2.B.

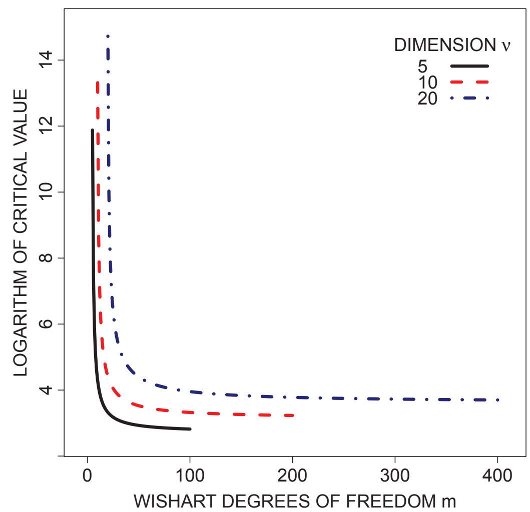


Figure 2.1: Logarithms of 0.99 quantiles produced from Hardin-Rocke scaled F distribution (vertical axis) as a function of the Wishart degrees of freedom parameter m (horizontal axis). The quantiles are shown for several values of dimension ν (plot symbols and colors).

calculated the MCD(γ) 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 $n = 3\nu, 5\nu, 7\nu, 9\nu, 11\nu$ for the above dimensions and values of γ.[5] We remind the reader that, as discussed in Section 2.2.1, 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

[5] 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.

c_{asy} (Equation (2.11)).

We first considered how well the Hardin-Rocke adjustment estimated m for $\gamma < \gamma^*$. Figures 2.2–2.4 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 the critical values for testing RSDs. Figures 2.5–2.7 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 2.2–2.4 translates into critical values that are smaller than they should be, as we would expect from Figure 2.1. 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.

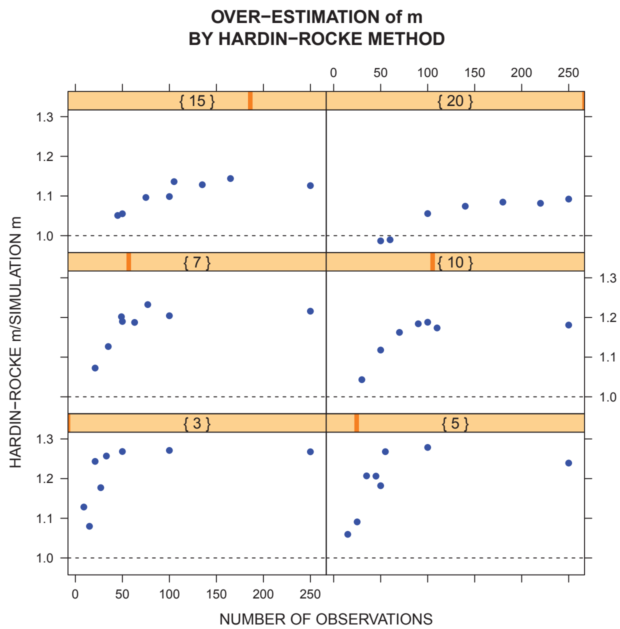


Figure 2.2: Comparison of Wishart degrees of freedom parameter m estimated via simulation and Hardin-Rocke approach with $\gamma = 0.25$. The ratio of the degrees of freedom parameters coming from the Hardin-Rocke approach to those resulting from the simulation is shown (stratified by dimension ν). Sample 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.

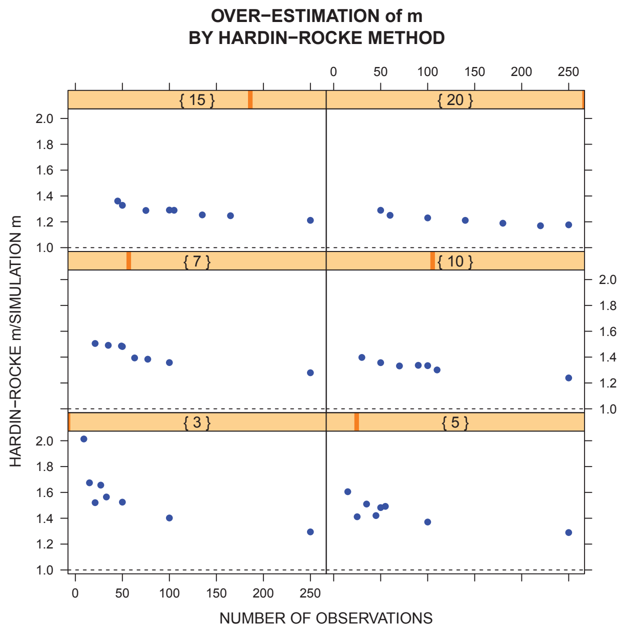


Figure 2.3: Comparison of Wishart degrees of freedom parameter m estimated via simulation and Hardin-Rocke approach with $\gamma = 0.05$. The plot setup is identical to that of Figure 2.2.

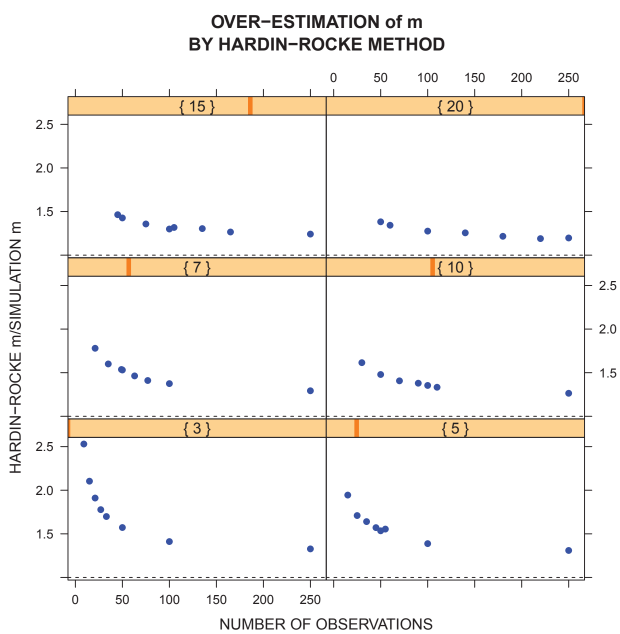


Figure 2.4: Comparison of Wishart degrees of freedom parameter m estimated via simulation and Hardin-Rocke approach with $\gamma = 0.01$. The plot setup is identical to that of Figure 2.2.

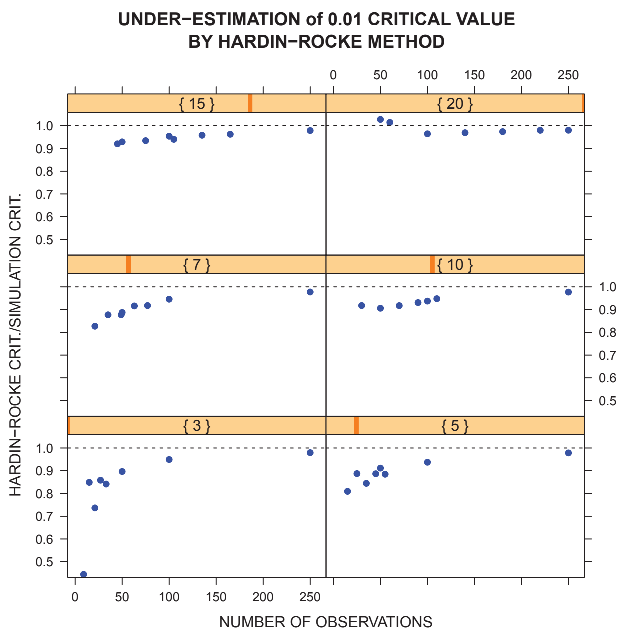


Figure 2.5: Comparison of 0.01 critical values produced using Wishart degrees of freedom parameter m estimated via simulation and Hardin-Rocke approach with $\gamma = 0.25$. Critical values are calculated using the scaled F distributional approximation of Hardin and Rocke with each degrees of freedom 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 2.2. The dimension ν is shown in the yellow bars at the top of each subplot.

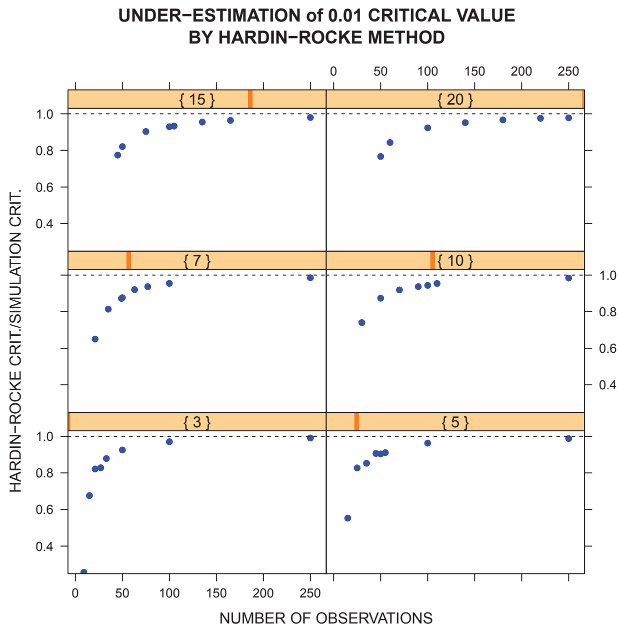


Figure 2.6: Comparison of 0.01 critical values produced from Wishart degrees of freedom parameter estimated via simulation and Hardin-Rocke approach with $\gamma = 0.05$. The plot setup is identical to that of Figure 2.5.

Overall we observe that the Hardin-Rocke adjustment (2.13) 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.[6]

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

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

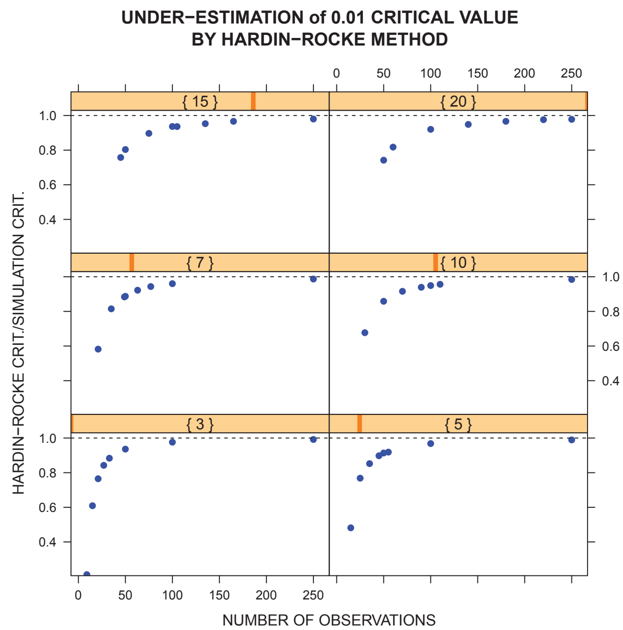


Figure 2.7: Comparison of 0.01 critical values produced from Wishart degrees of freedom parameter estimated via simulation and Hardin-Rocke approach with $\gamma = 0.01$. The plot setup is identical to that of Figure 2.5.

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 sizes n, dimensions ν, and trimming fractions γ.

2.3.3 An Improved Adjustment to the Asymptotic Degrees of Freedom

We start our search for a better adjustment formula with some exploratory data analysis. Figure 2.8 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 (2.13).

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 2.8. 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.[7]

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:

$$\log\left(\frac{\tilde{m}_{\text{sim}}(n, \nu, \gamma)}{m_{\text{asy}}(n, \nu, \gamma)}\right) = \frac{\beta_0 + \beta_1(1-\gamma) + \beta_2\nu}{n^{\beta_3+\beta_4(1-\gamma)}} + \epsilon_{n,\nu,\gamma}, \qquad \epsilon_{n,\nu,\gamma} \overset{iid}{\sim} N(0,1). \tag{2.14}$$

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

$$\log\left(\frac{m}{m_{\text{asy}}(n, \nu, \gamma)}\right) = \frac{12.746 - 14.546(1-\gamma) + 0.127\nu}{n^{0.559+0.149(1-\gamma)}}, \tag{2.15}$$

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

$$\tilde{m} = m_{\text{asy}}(n, \nu, \gamma) \exp\left(\frac{12.746 - 14.546(1-\gamma) + 0.127\nu}{n^{0.559+0.149(1-\gamma)}}\right). \tag{2.16}$$

Table 2.1 provides the regression coefficients along with their standard errors. All the regression coefficients are highly significant.

[7]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.

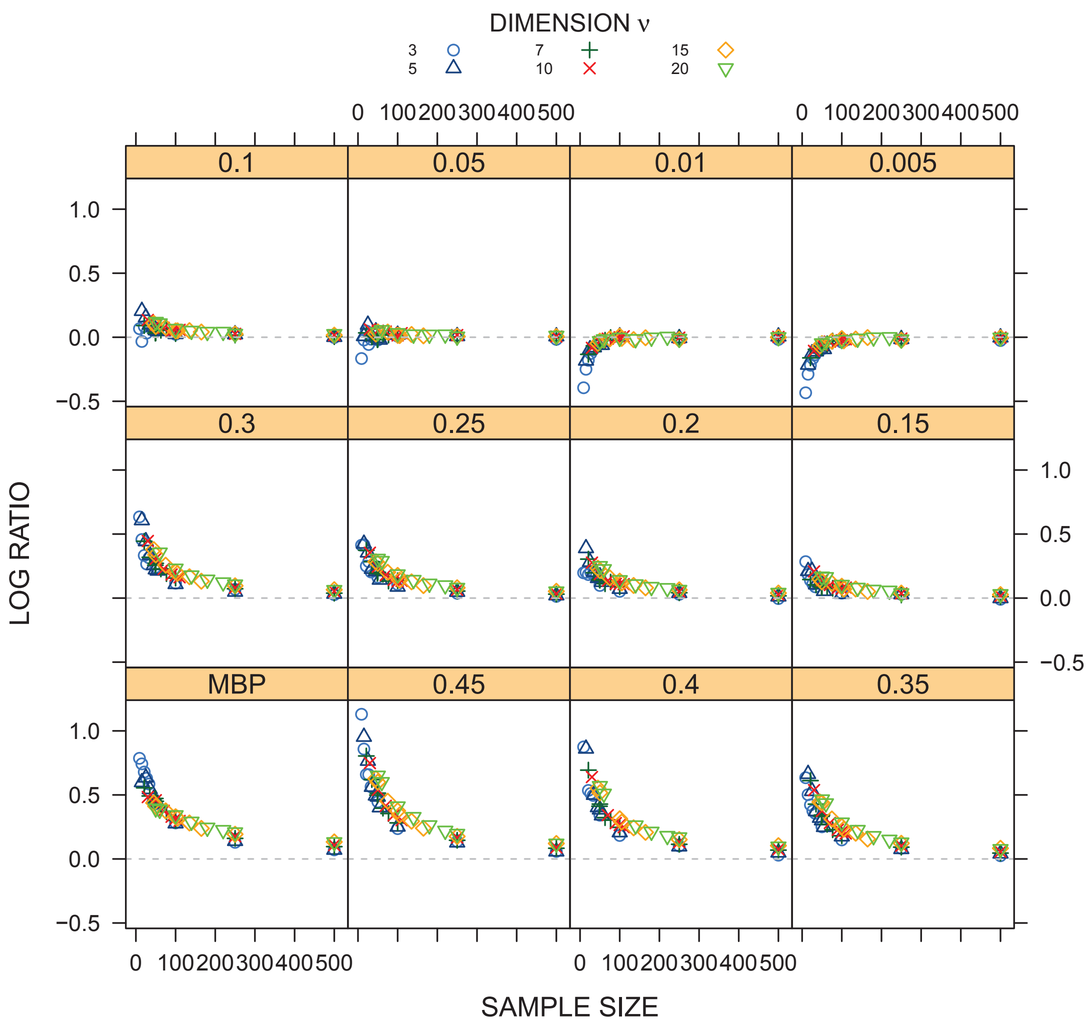


Figure 2.8: Logarithm of the ratio of the Wishart degrees of freedom estimated via simulation to the degrees of freedom calculated from the asymptotic formula, plotted against sample size and stratified by γ (printed in the yellow headers) and dimension (given by the plot symbols in each plot).

Table 2.1: Estimated coefficients, and their standard errors, for the model described by Equation (2.14).

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

2.4 *Validation of the Improved Adjustment Model*

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

To validate the fitted model (2.15), we used the same simulation procedure used in Section 2.3.3 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 $0.05 \le \gamma \le 0.45$ 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 2.9, 2.10, and 2.11 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 2.12, 2.13, and 2.14 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 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 2.15 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).[8] 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 2.16 shows the out-of-sample performance, as measured by the logarithm

[8] 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 $n = 12, 16, 20$.

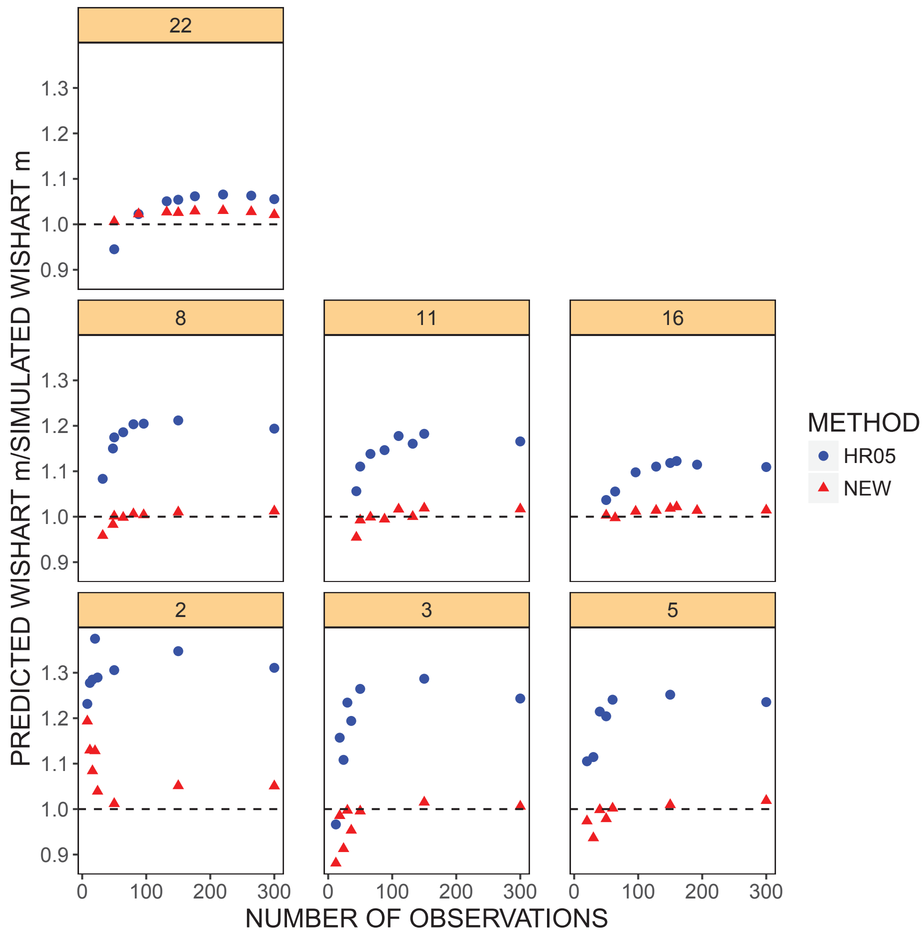


Figure 2.9: Out of sample comparison of estimated Wishart degrees of freedom parameter m to simulated value $\tilde{m}_{\text{sim}}(n, \nu, \gamma)$ using the Hardin-Rocke method and the proposed method with $\gamma = 0.25$. The plot shows the ratio of the degrees of freedom parameter m estimated using 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.

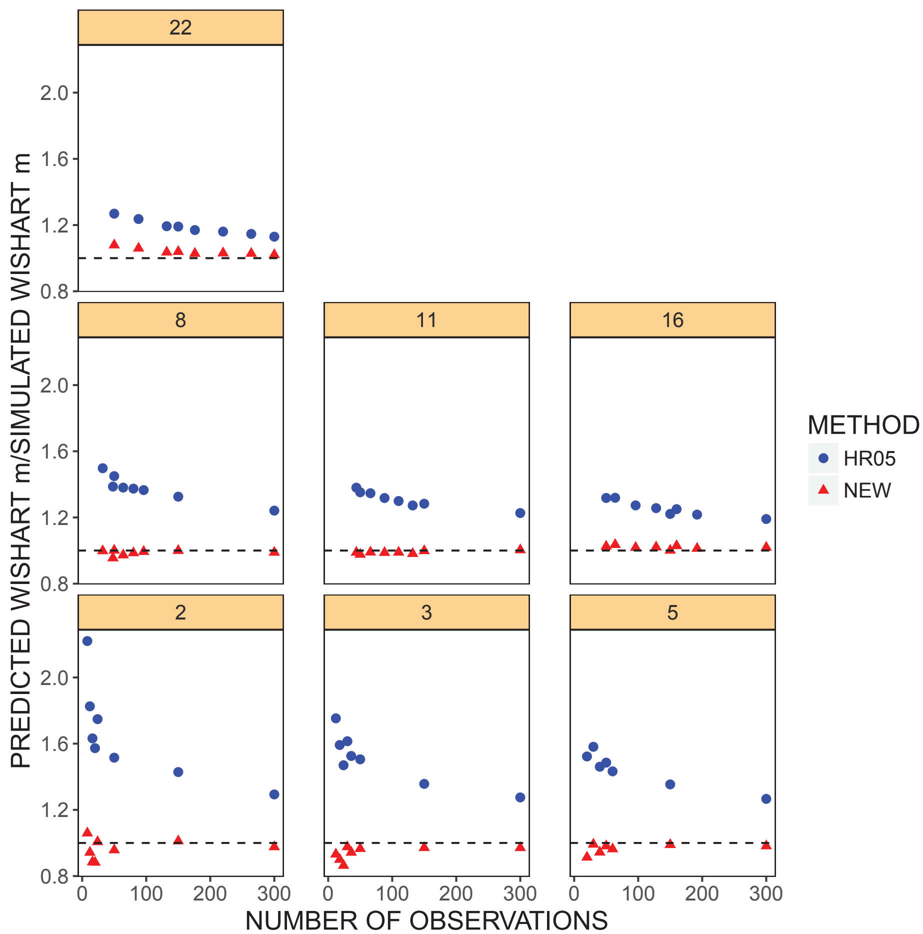


Figure 2.10: Out of sample comparison of estimated Wishart degrees of freedom parameter m to simulated value $\tilde{m}_{\text{sim}}(n, \nu, \gamma)$ using the Hardin-Rocke method and the proposed method with $\gamma = 0.05$. The plot setup is identical to Figure 2.9.

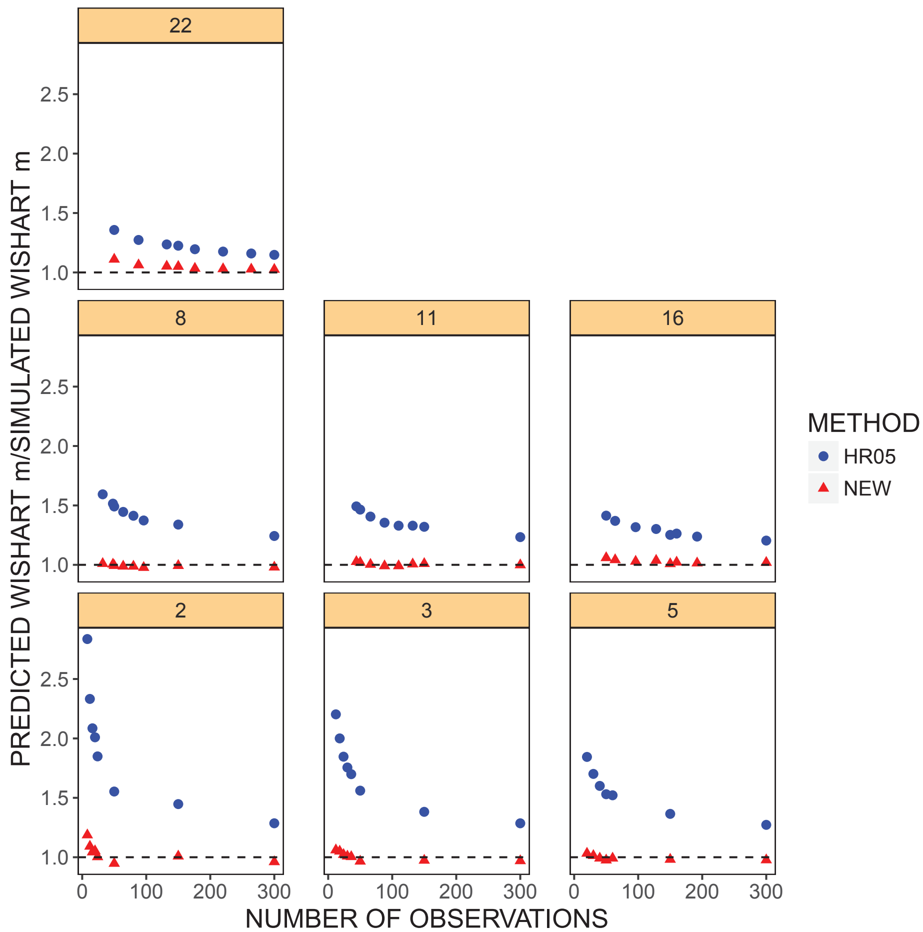


Figure 2.11: Out of sample comparison of estimated Wishart degrees of freedom parameter m to simulated value $\tilde{m}_{\text{sim}}(n, \nu, \gamma)$ using the Hardin-Rocke method and the proposed method with $\gamma = 0.01$. The plot setup is identical to Figure 2.9.

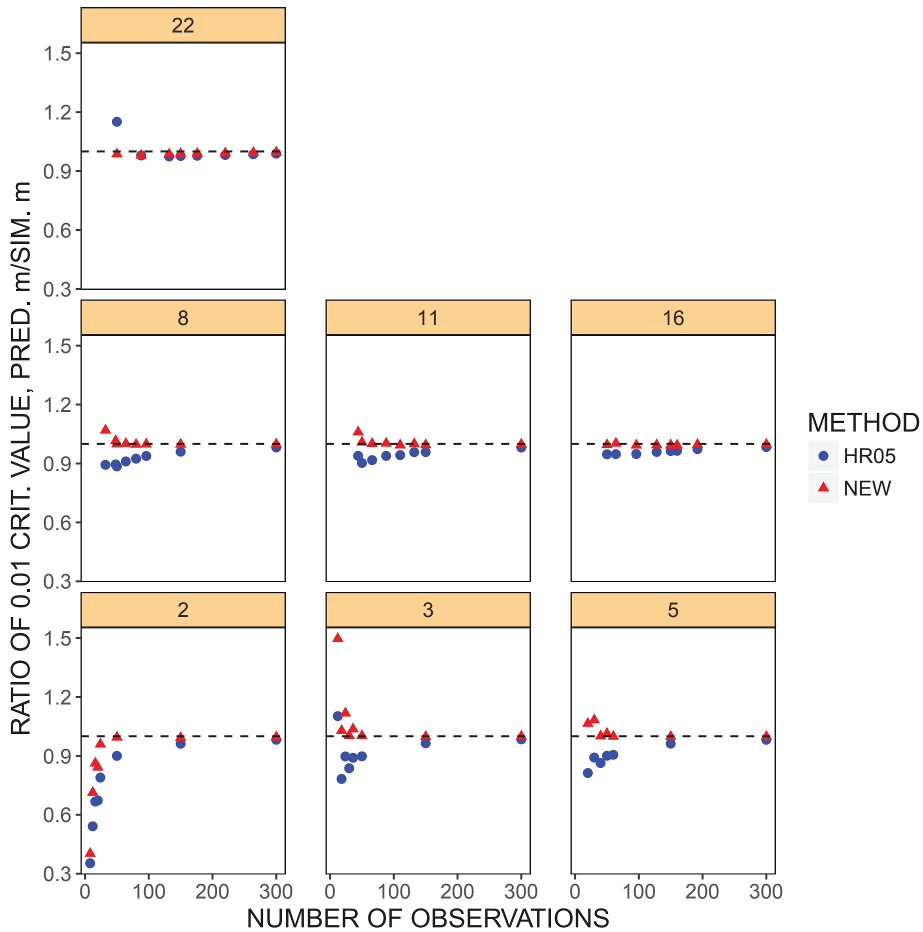


Figure 2.12: Out of sample comparison of 0.01 critical values from the Hardin-Rocke F distribution computed using the estimated Wishart degrees of freedom parameter m from the Hardin-Rocke method and the proposed method with $\gamma = 0.25$. The plot shows the ratio of the 0.01 critical value computed 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 2.9. 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.

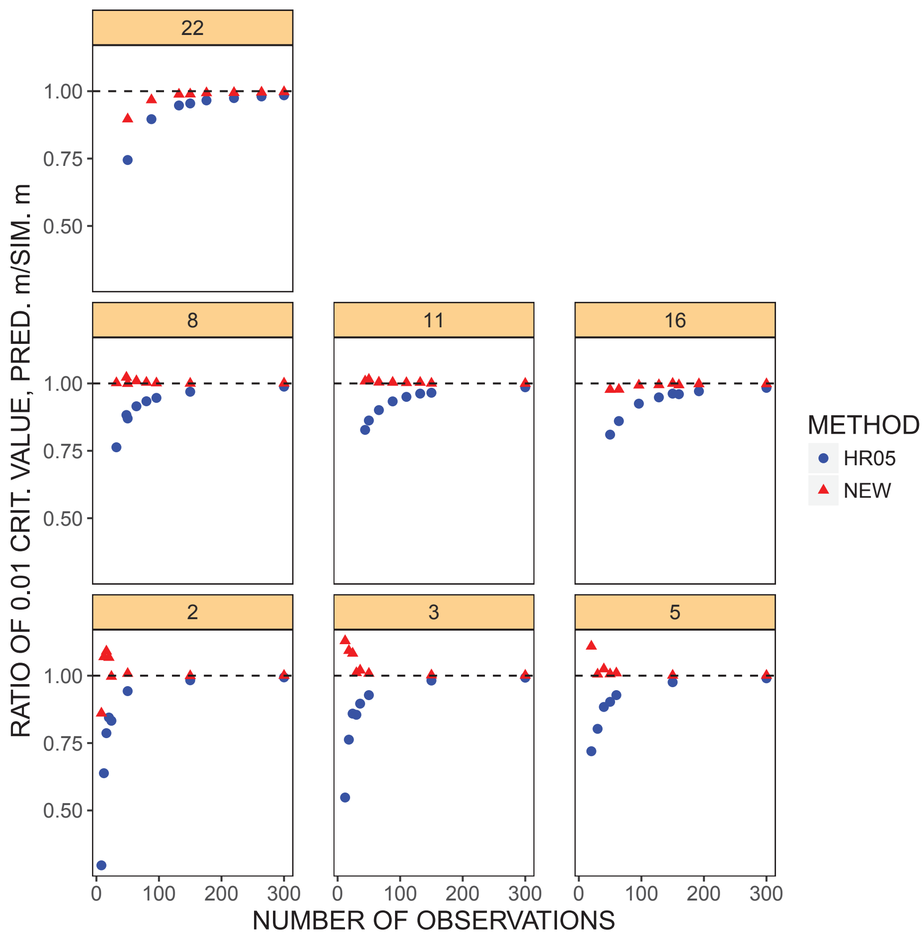


Figure 2.13: Out of sample comparison of 0.01 critical values from the Hardin-Rocke F distribution computed using the estimated Wishart degrees of freedom parameter m from the Hardin-Rocke method and the proposed method with $\gamma = 0.05$. The plot setup is identical to that of Figure 2.12.

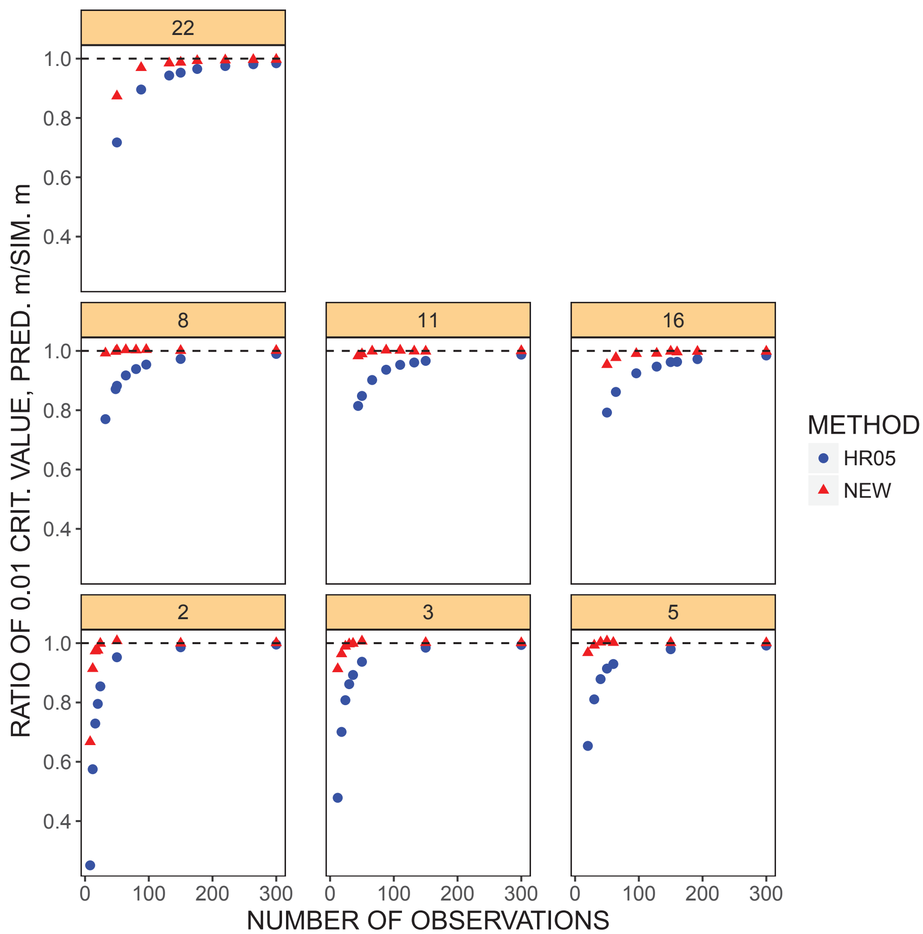
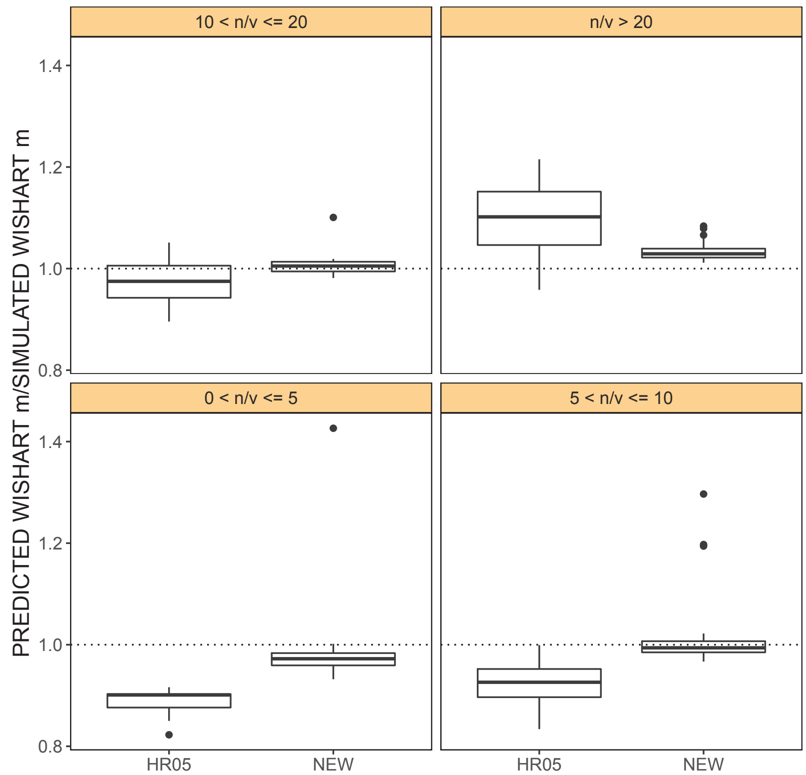


Figure 2.14: Out of sample comparison of 0.01 critical values from the Hardin-Rocke F distribution computed using the estimated Wishart degrees of freedom parameter m from the Hardin-Rocke method and the proposed method with $\gamma = 0.01$. The plot setup is identical to that of Figure 2.12.



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 2.15: 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.

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.[9] 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.

2.4.2 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 (2.13) and using the new method (2.15) developed in this paper. Since the data contains no outliers by construction, any outliers detected are false positives.

[9] Full results are available in Table 2.5 in Appendix 2.D.

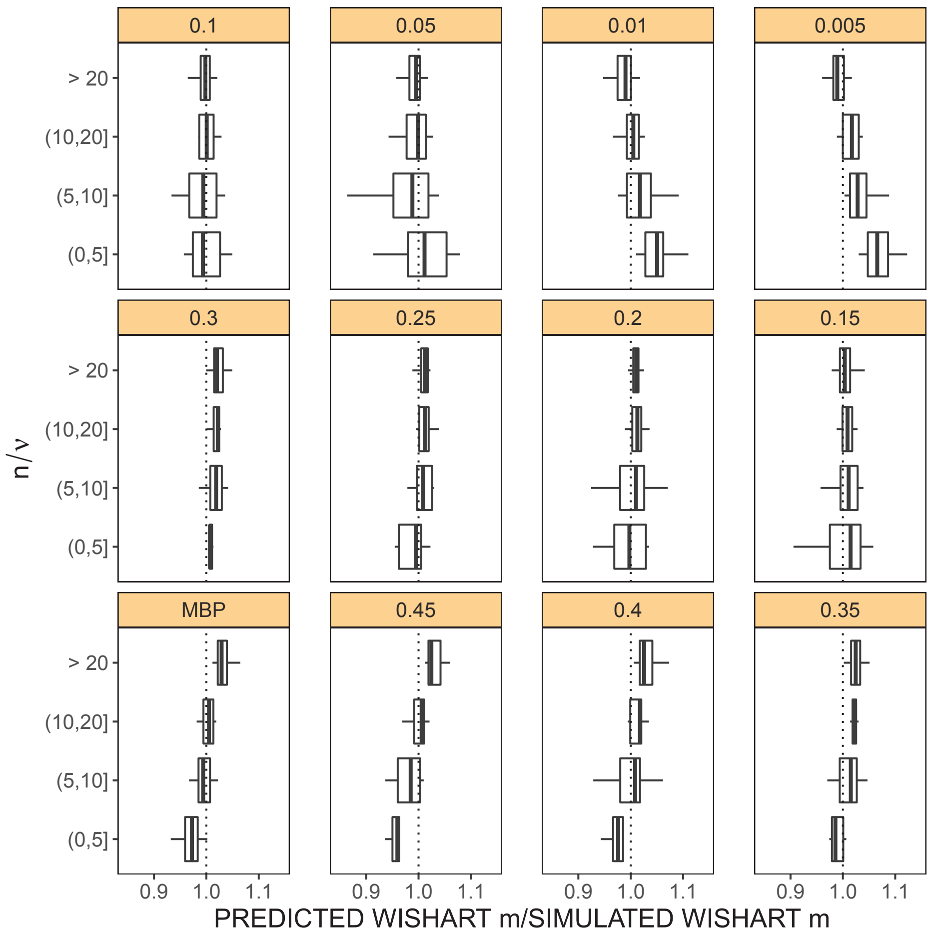


Figure 2.16: Boxplots showing the range of out of sample performance of the proposed correction methodology, stratified by γ (yellow box) and the ratio n/ν of observations to variables (vertical axis). Performance is measured by the ratio of the predicted Wishart degrees of freedom value to the value 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 2.15.

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

Tables 2.2 and 2.3 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 2.2: 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 2.3: 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.

2.4.3 Extension of FSRMCD and IRMCD to Arbitrary γ

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

$$H_{0i} : \mathbf{x}_i \sim N\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right), \qquad i = 1, \ldots, n. \tag{2.17}$$

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 (2.3)).

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.[10] 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. 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

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

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.

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 2.2 and 2.3, 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 such as the one to be presented in Chapter 3.

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 (2.3) 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.[11]

Table 2.4 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 2.6 in Appendix 2.E provides standard deviations for the entries in the table.)

[11]The simulations and the analysis were performed on a laptop running Windows 7 Ultimate SP 1 with

Table 2.4: 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 2.6 in Appendix 2.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

Power calculations for our modified implementation of IRMCD are discussed in Appendix 2.F.

2.5 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 2.16 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 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.[12] 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

an Intel® Core™ i7-3740QM processor running at 2.7GHz and 32GB of RAM.

[12]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.

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.

2.6 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.) 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.[13] The Tukey-Huber Contamination Model assumes that

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

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) 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 provides additional detail on the OGK estimator.) The quadrant cor-

---

Models.

relation 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) 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.

APPENDIX

2.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 2.3.1. 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 $1 - \gamma = P(\chi^2_\nu \leq q(\nu, 1-\gamma))$.

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

$$b_2(\nu,\gamma) = \frac{1}{2} + \frac{c(\nu,\gamma)}{1-\gamma}\left(c_3(\nu,\gamma) - \frac{q(\nu,1-\gamma)}{\nu}\left(c_2(\nu,\gamma) + \frac{1-\gamma}{2}\right)\right)$$

$$v_1(\nu,\gamma) = (1-\gamma)b_1(\nu,\gamma)^2\left(\gamma\left(\frac{c(\nu,\gamma)q(\nu,1-\gamma)}{\nu} - 1\right)^2 - 1\right) -$$
$$2c_3(\nu,\gamma)c(\nu,\gamma)^2\left(3\left(b_1(\nu,\gamma) - \nu b_2(\nu,\gamma)\right)^2 +\right.$$
$$\left.(\nu+2)b_2(\nu,\gamma)(2b_1(\nu,\gamma) - \nu b_2(\nu,\gamma))\right)$$

$$v_2(\nu,\gamma) = \left(b_1(\nu,\gamma)\left(b_1(\nu,\gamma) - \nu b_2(\nu,\gamma)\right)(1-\gamma)\right)^2 c(\nu,\gamma)^2$$

$$v(\nu,\gamma) = \frac{v_1(\nu,\gamma)}{n v_2(\nu,\gamma)}.$$

2.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.

2.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
```

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).

2.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")
```

2.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 (2.15).

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

Table 2.5 provides the out of sample results from testing the model shown in Equation (2.15). The table shows the ratio of the predicted degrees of freedom to the simulated degrees of freedom.

Table 2.5: 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 2.16.

		$\gamma =$												
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 2.5: *(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 2.5: *(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 2.5: *(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 2.5: *(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

2.E Standard Deviations for FSRMCD and IRMCD Simulation Tests

Table 2.6 provides standard deviations for the simulation results presented in Table 2.4. Standard errors for entries in the latter table can be calculated by dividing the corresponding entry of this table by $\sqrt{5000}$.

Table 2.6: Monte Carlo standard deviations of simulation tests of FSRMCD and IRMCD implementations. Standard errors for the quantities in Table 2.4 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

2.F Power Calculations for IRMCD with the Modified Estimator of m

2.F.1 Overview

We examined the power of IRMCD with our modified estimator of the Wishart parameter m to detect outliers when they are present in the data. We are particularly interested in its performance relative to IRMCD with the Hardin-Rocke estimator of m. We investigate the power of these outlier detection rules using simulation studies identical to those used in Cerioli (2010), with minor modifications to allow for $\gamma < \gamma^*$. We review the details of the simulation below for the convenience of the reader.

2.F.2 Methodology

Let τ be the fraction of observations that are "contaminated". Given a sample size n and dimension ν, we generate 5000 ν-dimensional samples of size n from a mixture distribution: $n(1-\tau)$ of the observations are drawn from a multivariate normal distribution $N(\mathbf{0}, \mathbf{I})$, while the remaining $n\tau$ observations come from a contaminating distribution. We will consider three types of contaminating distributions:

- A location-shift contamination model $N(\lambda\mathbf{1}, \mathbf{I})$ where the expected values of all variables are shifted by $\lambda > 0$;
- A radial contamination model $N(\mathbf{0}, \psi\mathbf{I})$ where marginal variances of the variables are inflated by $\psi > 0$; and
- A t distribution contamination model where observations are generated from a multivariate t distribution with $\zeta \geq 1$ degrees of freedom.

We consider these scenarios with a small amount of contamintation $\tau = 0.05$ and a moderate amount of contamination $\tau = 0.20$.

Let α_1 be the nominal size of the outlier detection tests. In each scenario, we compute MCD(γ)-based RSDs and test observations for outlyingness at the α_1 level using each of the following four approaches.

RMCD Distances are calculated using the reweighted MCD(γ) and tested against the $1-\alpha$ quantile of a χ^2_ν distribution, where $1 - \alpha = (1 - \alpha_1)^{(1/n)}$.

RMCD_ind Distances are calculated using the reweighted MCD(γ) and tested against the $1 - \alpha_1$ quantile of a χ^2_ν distribution.

IRMCD_HR IRMCD(γ) using the Hardin-Rocke estimator of the Wishart parameter m.

IRMCD_GM IRMCD(γ) using the estimator of the Wishart parameter m developed in this chapter.

We consider three values of the MCD trimming parameter γ: γ^*, 0.25, and 0.05.

We calculate the power of each approach to detect outliers as the ratio of the number of contaminated observations detected to the total number of contaminated observations. Note that, in contrast to Cerioli (2010), we do not define the power for the cases of no contamination (i.e., $\lambda = 0$, $\psi = 1$, and $\zeta = \infty$) to be the empirical false positive rate for testing the hypothesis of no outliers in the data. We also use more values of the contamination parameter (λ, ψ, or ζ) as appropriate. Our results for small values of contamination will thus look different from those presented in Cerioli (2010).

2.F.3 Results

Figures 2.17 and 2.20 show the power of the four approaches for detecting a shift in location for MCD(γ^*)-based RSDs. Figures 2.18 and 2.21 show the corresponding results for the MCD(0.25)-based RSDs, and Figures 2.19 and 2.22 show the results for the MCD(0.05)-based RSDs. For the $\gamma = \gamma^*$ and $\gamma = 0.25$ cases, all methods perform well, in all sample size and dimension combinations considered, for detecting a shift of at least 2 when the amount of contamination is small ($\tau = 0.05$). For the larger fraction of contamination $\tau = 0.20$, both IRMCD approaches are less powerful than the simple chi-squared based rules (RMCD and RMCD_ind) in small samples ($n = 60$ and $n = 120$). When the dimension is $\nu = 20$ all of the rules show markedly decreased power to detect a location shift, with the IRMCD approaches again being less powerful than the chi-squared approaches.

The choice of $\gamma = 0.25$ generally results in lower power, other things being equal, compared to $\gamma = \gamma^*$. This is not surprising given that the breakdown point of the MCD(γ) covariance estimate is γ. The MCD(0.25)-based distances are hence more influenced by the outliers themselves, which inhibits our ability to detect more moderate outliers. Moving to an even smaller trimming fraction, $\gamma = 0.05$, makes it even harder for the four approaches to detect outliers in the data sets. When the contamination fraction is $\tau = 0.20$ but the trimming fraction is only $\gamma = 0.05$, all of the methods are rather useless for detecting shifts in location (Figure 2.22).

Figures 2.23 and 2.26 show the results for detecting observations arising from a marginal distribution with a larger variance using MCD(γ^*)-based RSDs. Figures 2.24 and 2.27 show the results for $\gamma = 0.25$, and Figures 2.25 and 2.28 show the results for $\gamma = 0.05$. In small samples $n = 60$ and small dimensions, neither IRMCD approach is as powerful as the RMCD approaches, but as the sample size increases the four methods agree more closely for both contamination fractions. As the dimension increases, however, the two IRMCD methods are actually more powerful than the chi-squared approaches for moderate amounts of contamination, particularly in smaller samples. This true for the three values of γ we tested, but the difference between the approaches is larger for $\gamma = 0.25$ and $\gamma = 0.05$.

Figures 2.29 and 2.32 show the results for detecting observations arising from a t distribution using MCD(γ^*)-based RSDs. Figures 2.30 and 2.33 show the results for $\gamma = 0.25$, and Figures 2.31 and 2.34 show the results for $\gamma = 0.05$. In contrast to the location-shift and variance-inflation cases, here power decreases for all methods as the t degrees of freedom parameter ζ increases: with larger values of ζ the t distribution looks more like the normal distribution from which the non-contaminated sample is drawn, so it becomes harder to distinguish the “outliers” from the non-outliers.

Overall, we see that the IRMCD is less powerful, compared to the chi-squared approaches, for detecting contaminated values from a t distribution, for all values of γ considered. For this type of contamination, however, IRMCD and the chi-squared approaches are closer in power for smaller values of γ.

2.F.4 Summary

Under all the contamination models considered here, the power of the IRMCD with the Hardin-Rocke estimator of m and with the estimator of m developed in this chapter is approximately the same. Hence our improved estimator of m leads to more accurate false positive rates in small samples and with small values of γ without comprising the power of the original IRMCD.

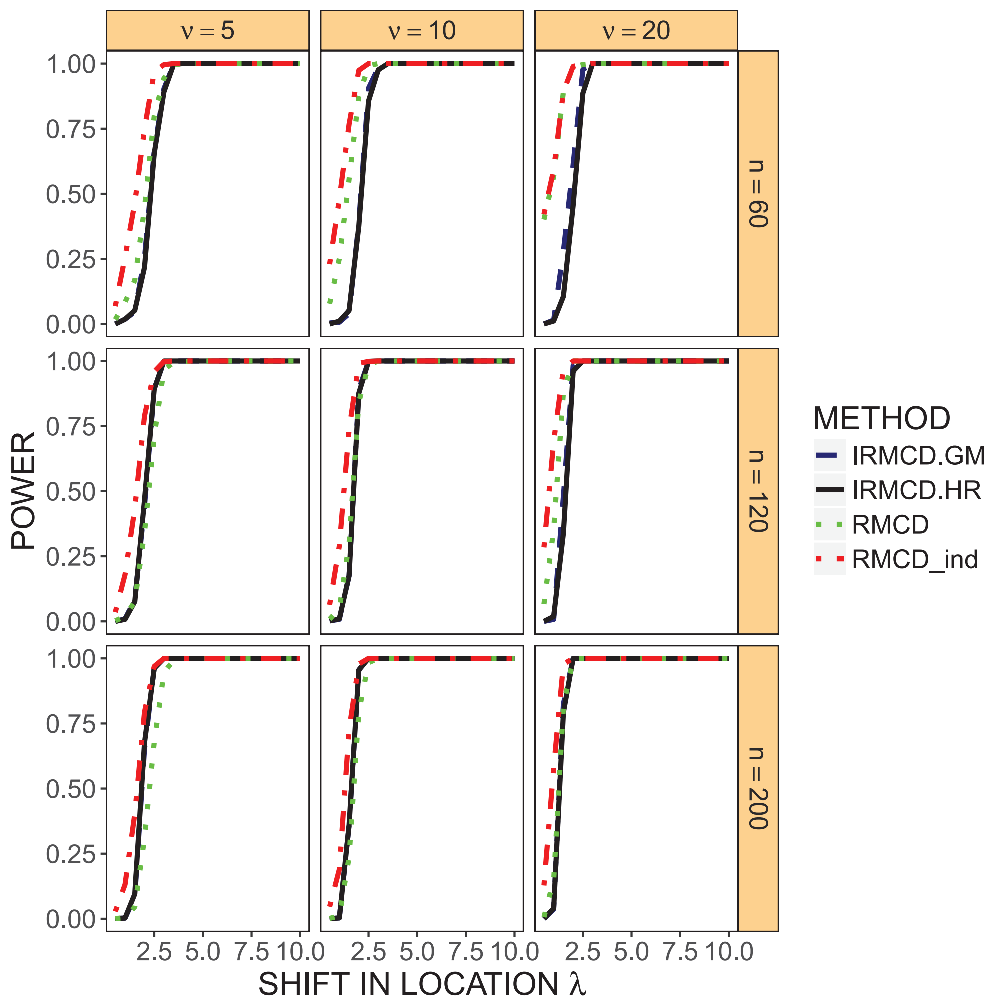


Figure 2.17: Power of MCD-based outlier detection rules under a multivariate location-shift contamination model for $\gamma = \gamma^*$ and contamination rate $\tau = 0.05$. Results are shown for dimensions $\nu = 5, 10, 20$, depicted in columns, and sample sizes $n = 60, 120, 200$, depicted in rows. The four outlier detection rules shown are the IRMCD with our modified estimator of the Wishart parameter m (IRMCD.GM, blue dashed line); IRMCD with the original Hardin-Rocke estimator of m (IRMCD.HR, black solid line); RSDs based on the reweighted MCD and tested against a chi-squared distribution with a multiplicity-correction to the test size (RMCD, green dashed line); and RSDs based on the reweighted MCD and tested against a chi-squared distribution (RMCD_ind, red dashed line).

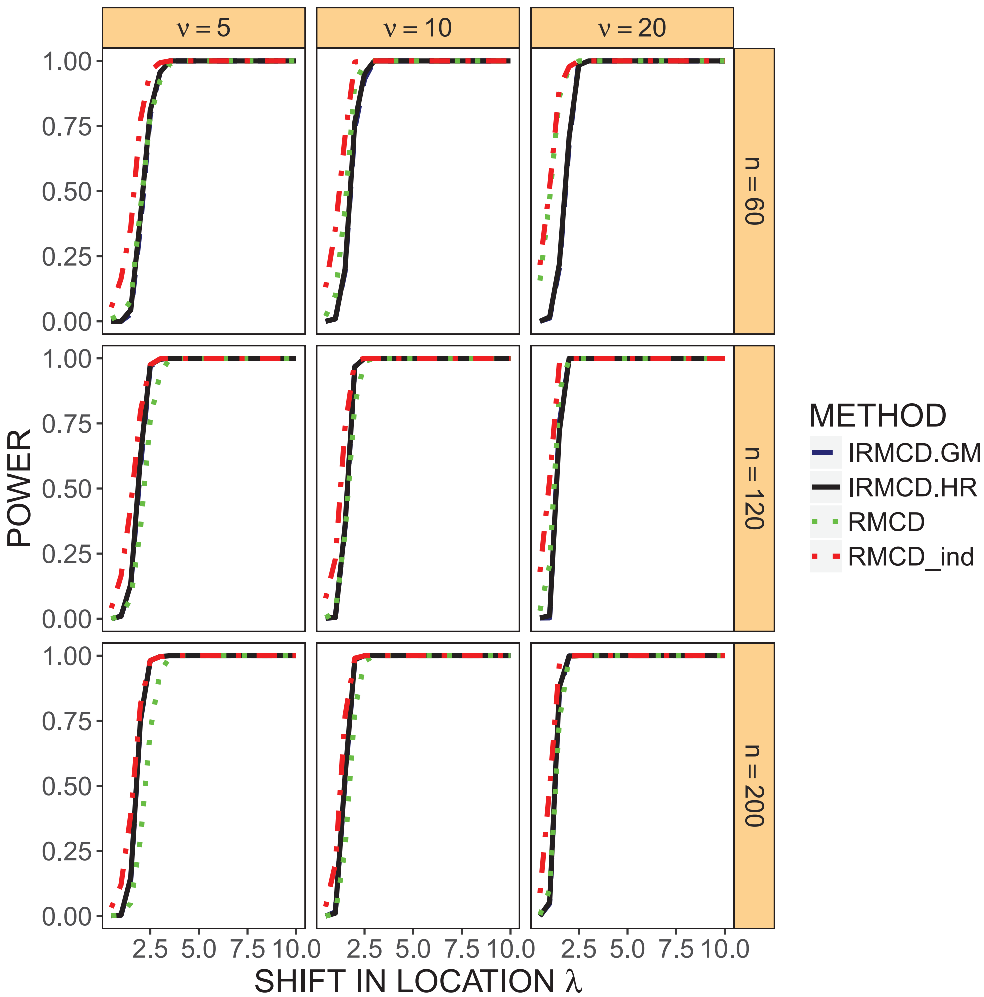


Figure 2.18: Power of MCD-based outlier detection rules under a multivariate location-shift contamination model for $\gamma = 0.25$ and contamination rate $\tau = 0.05$. The plot setup is identical to that of Figure 2.17.

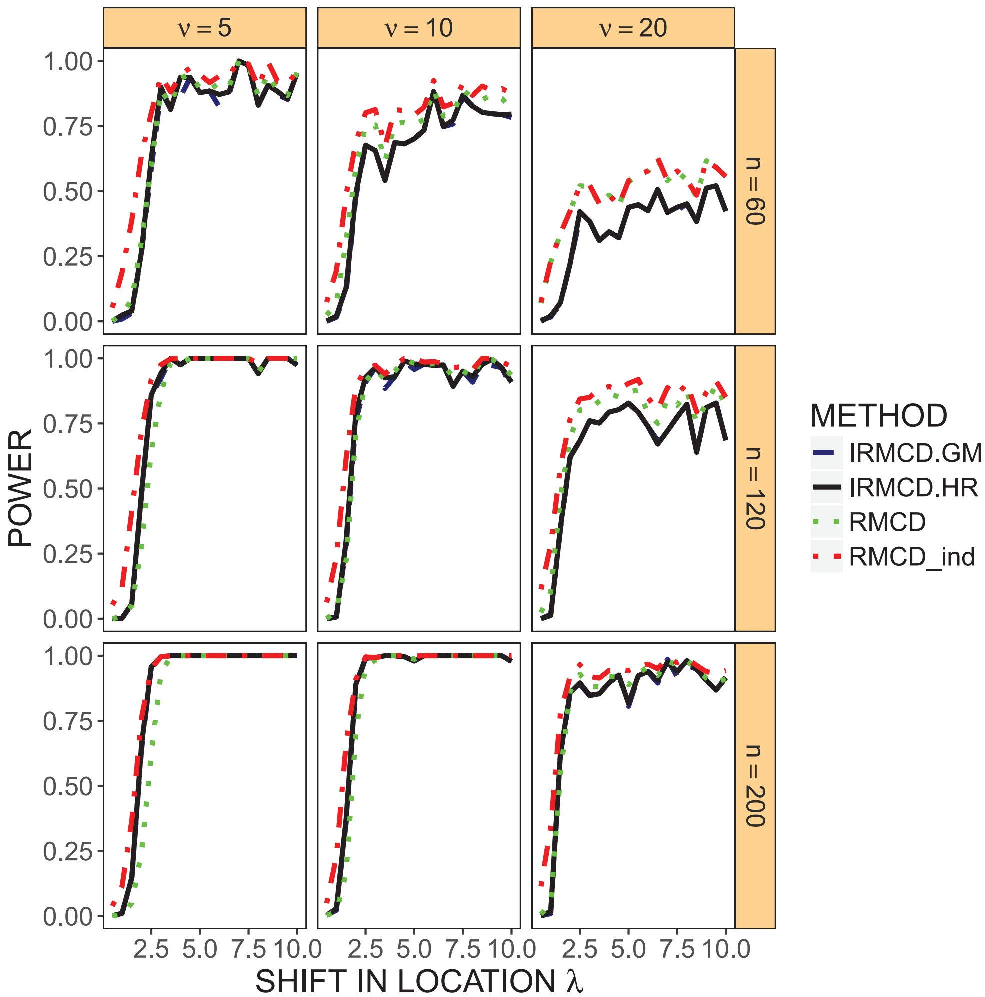


Figure 2.19: Power of MCD-based outlier detection rules under a multivariate location-shift contamination model for $\gamma = 0.05$ and contamination rate $\tau = 0.05$. The plot setup is identical to that of Figure 2.17.

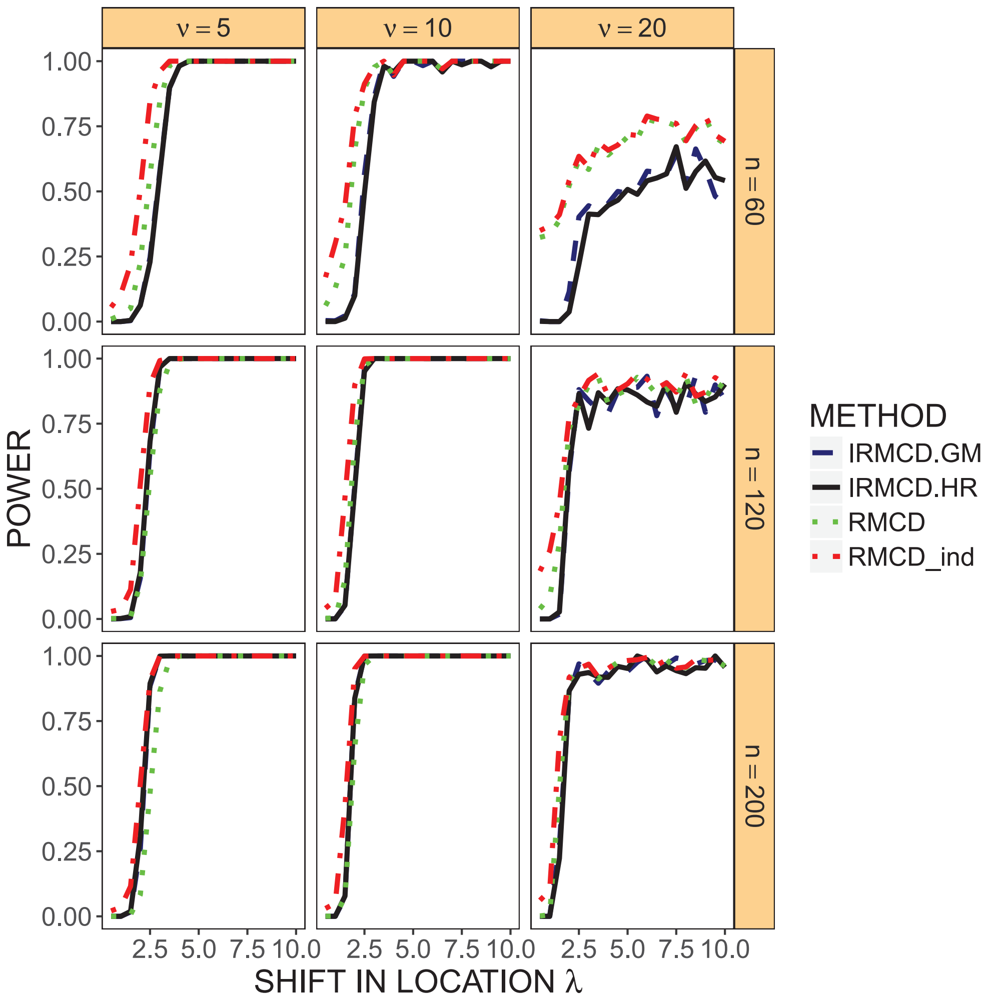


Figure 2.20: Power of MCD-based outlier detection rules under a multivariate location-shift contamination model for $\gamma = \gamma^*$ and contamination rate $\tau = 0.20$. The plot setup is identical to that of Figure 2.17.

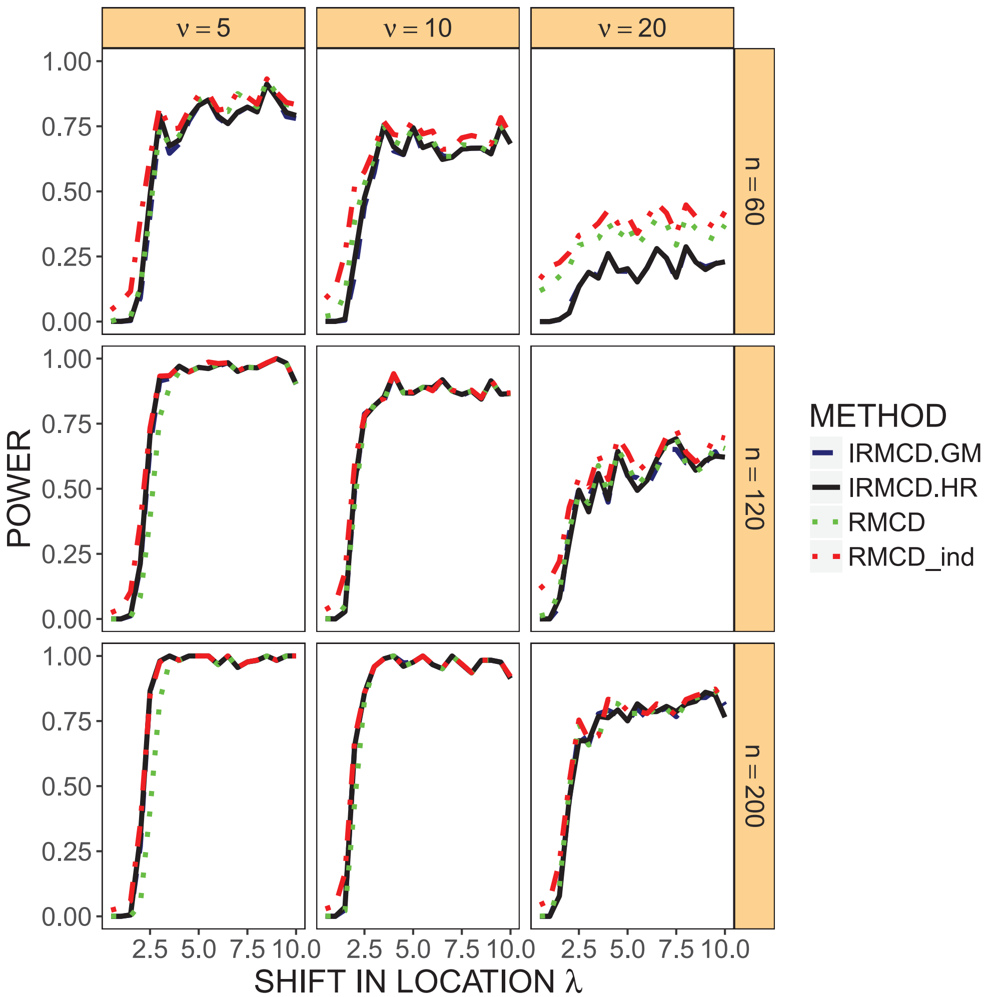


Figure 2.21: Power of MCD-based outlier detection rules under a multivariate location-shift contamination model for $\gamma = 0.25$ and contamination rate $\tau = 0.20$. The plot setup is identical to that of Figure 2.17.

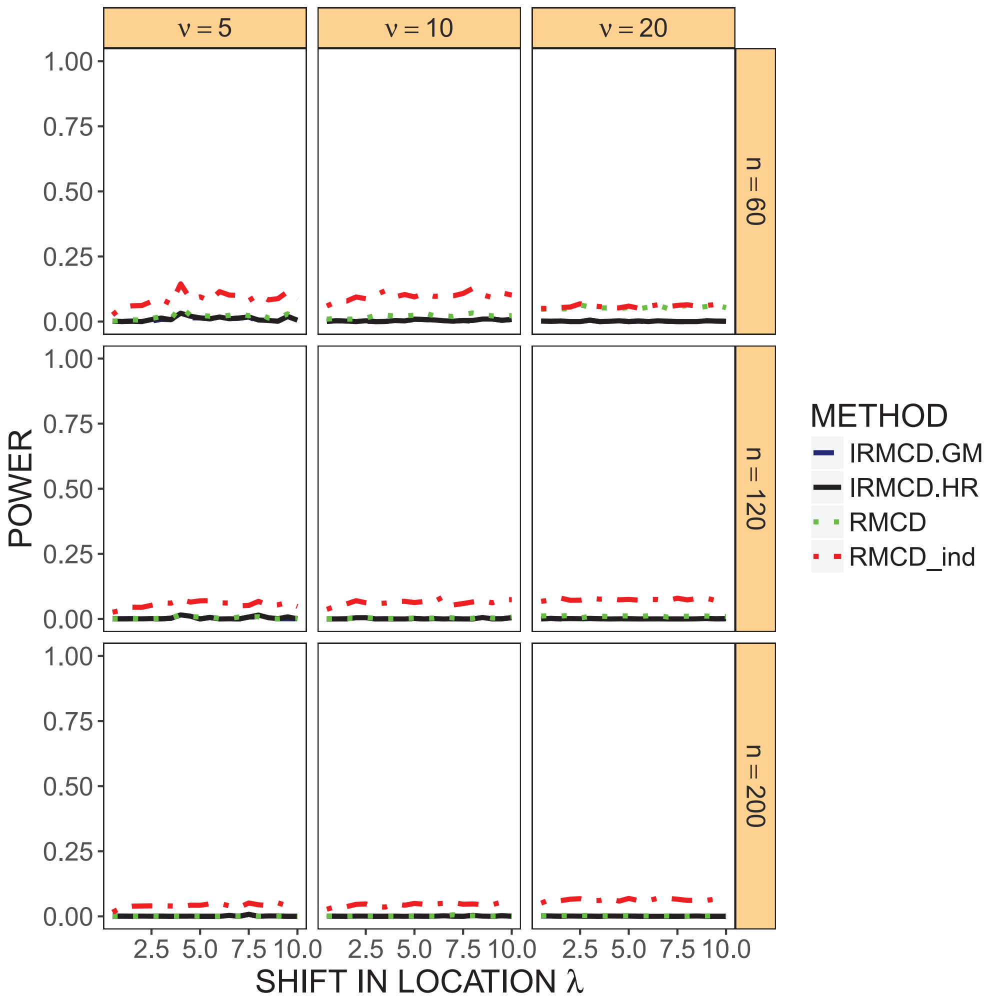


Figure 2.22: Power of MCD-based outlier detection rules under a multivariate location-shift contamination model for $\gamma = 0.05$ and contamination rate $\tau = 0.20$. The plot setup is identical to that of Figure 2.17.

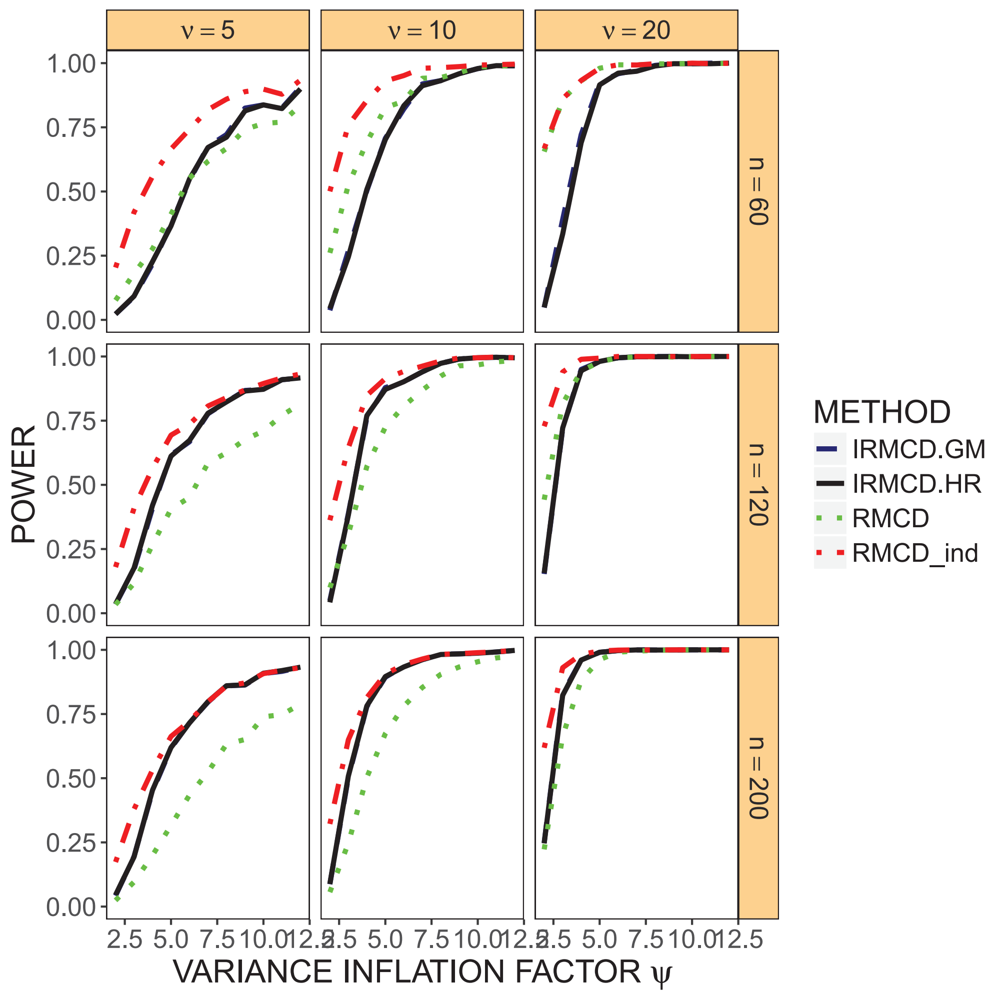


Figure 2.23: Power of MCD-based outlier detection rules under a multivariate radial contamination model for $\gamma = \gamma^*$ and contamination rate $\tau = 0.05$. Results are shown for dimensions $\nu = 5, 10, 20$, depicted in columns, and sample sizes $n = 60, 120, 200$, depicted in rows. The four outlier detection rules shown are the IRMCD with our modified estimator of the Wishart parameter m (IRMCD.GM, blue dashed line); IRMCD with the original Hardin-Rocke estimator of m (IRMCD.HR, black solid line); RSDs based on the reweighted MCD and tested against a chi-squared distribution with a multiplicity-correction to the test size (RMCD, green dashed line); and RSDs based on the reweighted MCD and tested against a chi-squared distribution (RMCD_ind, red dashed line).

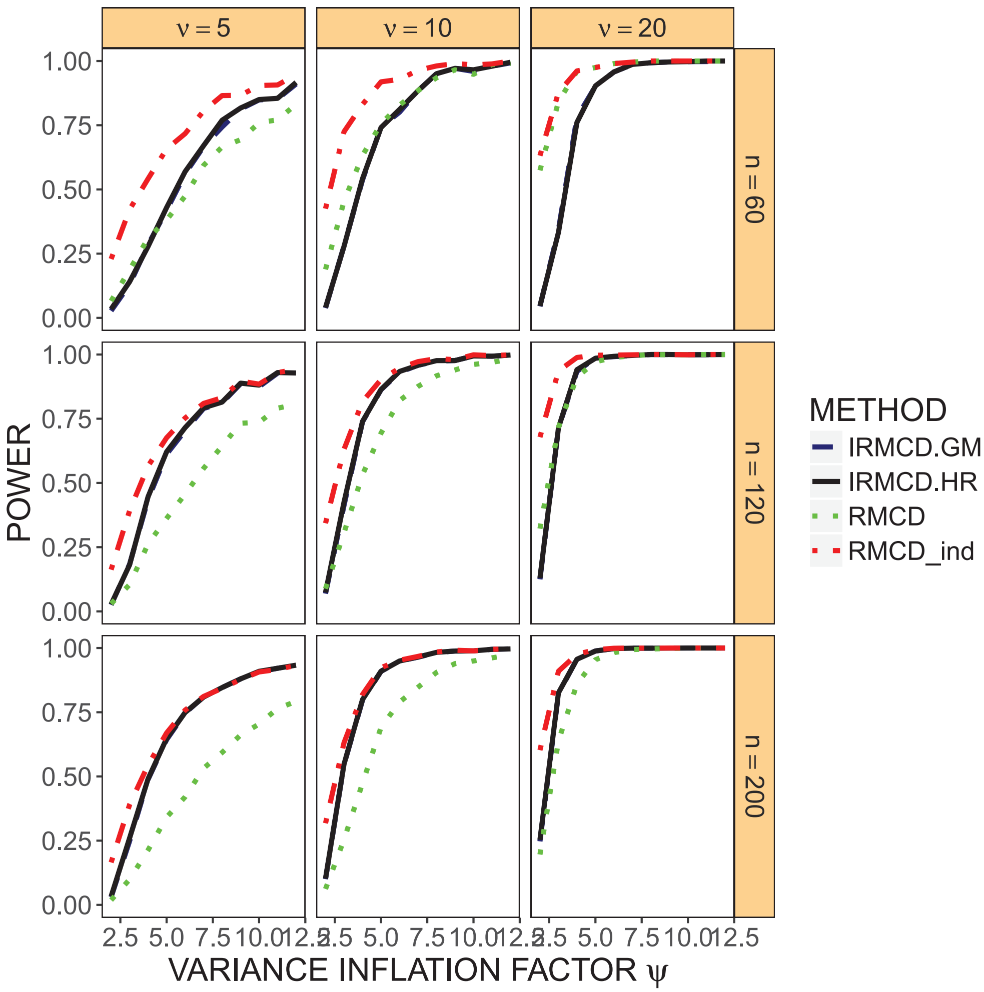


Figure 2.24: Power of MCD-based outlier detection rules under a multivariate radial contamination model for $\gamma = 0.25$ and contamination rate $\tau = 0.05$. The plot setup is identical to that of Figure 2.23.

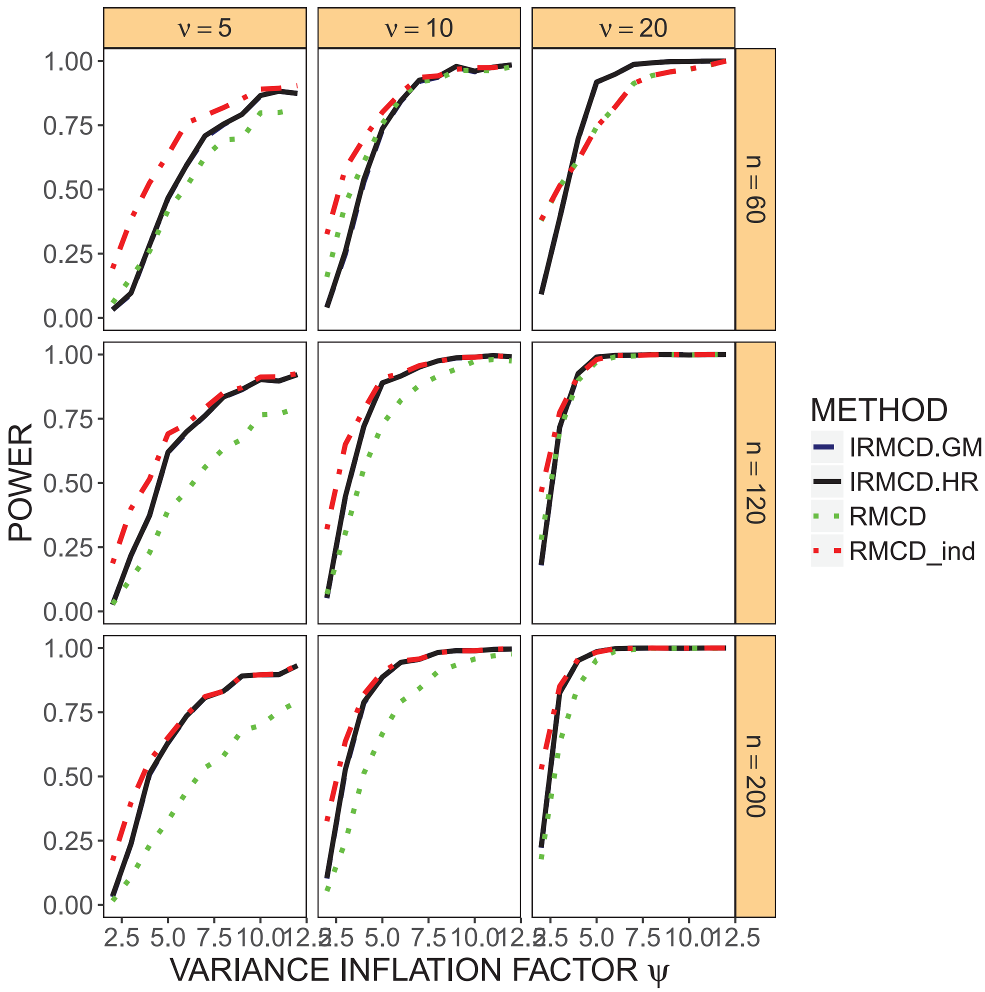


Figure 2.25: Power of MCD-based outlier detection rules under a multivariate radial contamination model for $\gamma = 0.05$ and contamination rate $\tau = 0.05$. The plot setup is identical to that of Figure 2.23.

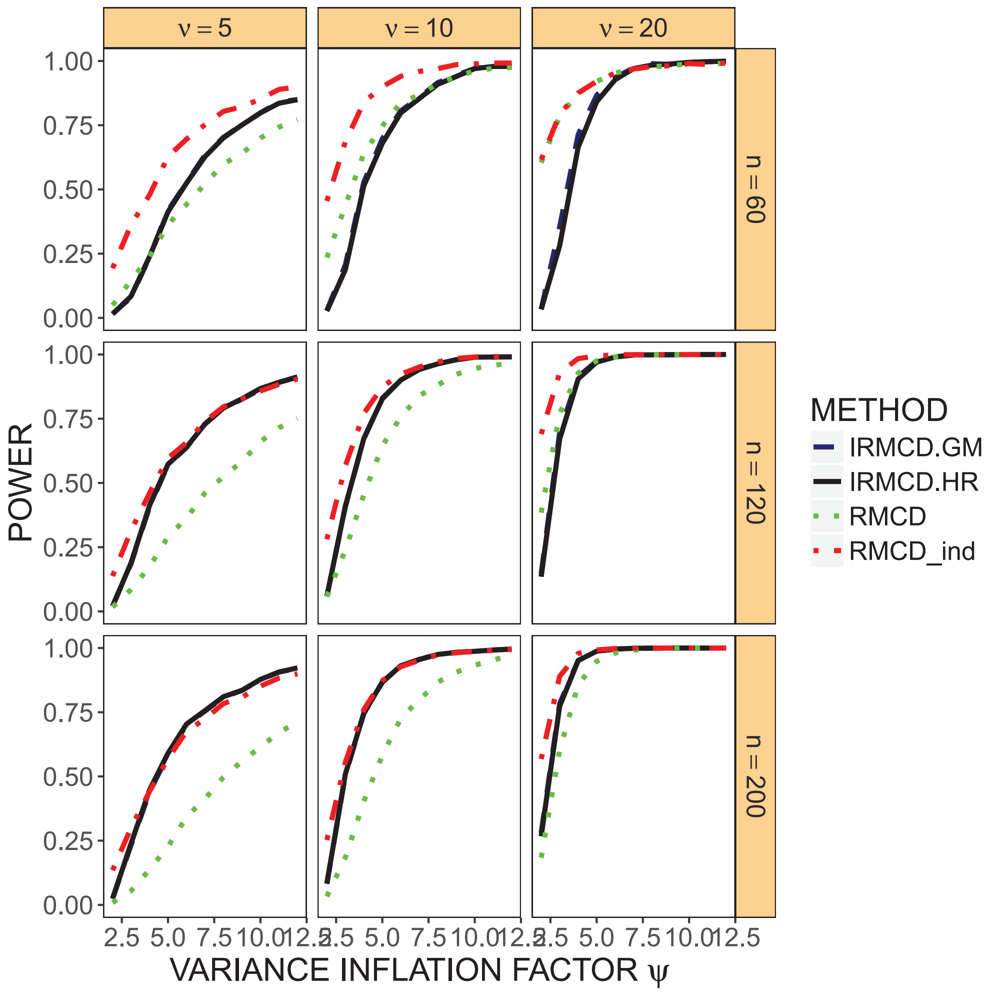


Figure 2.26: Power of MCD-based outlier detection rules under a multivariate radial contamination model for $\gamma = \gamma^*$ and contamination rate $\tau = 0.20$. The plot setup is identical to that of Figure 2.23.

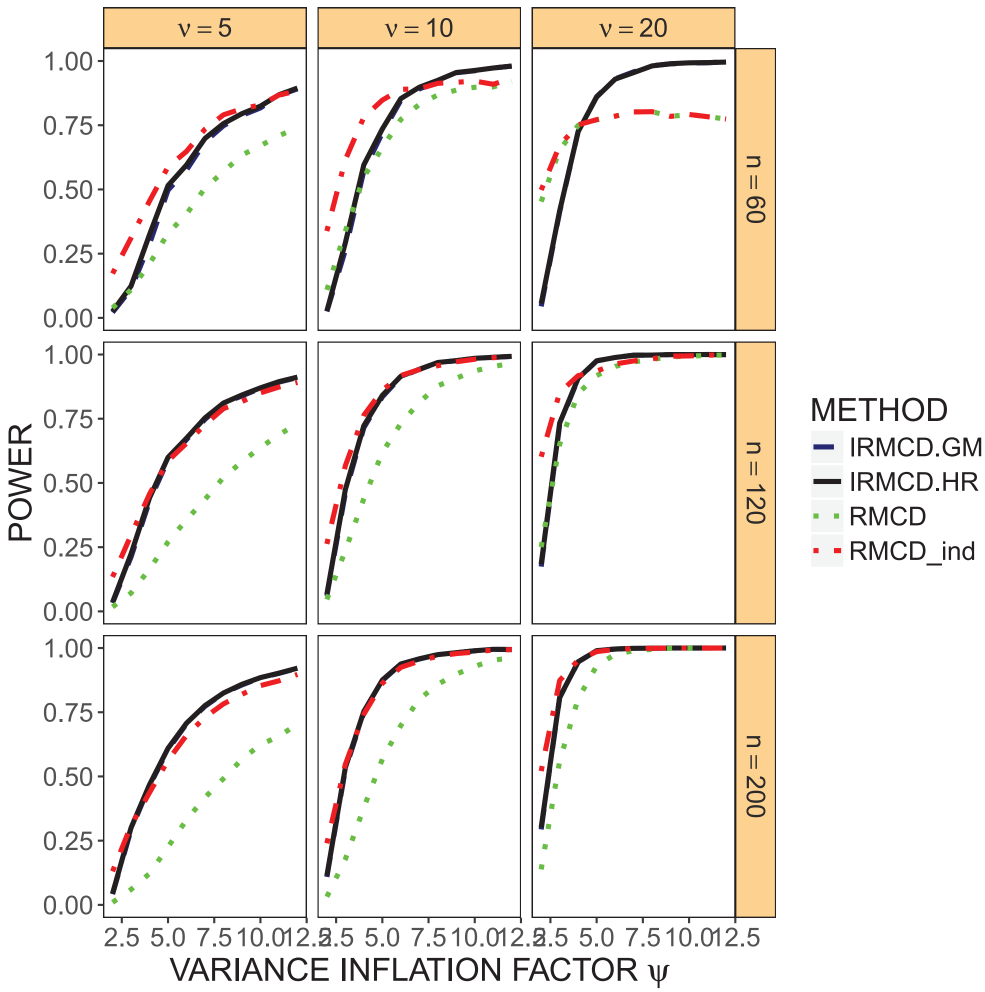


Figure 2.27: Power of MCD-based outlier detection rules under a multivariate radial contamination model for $\gamma = 0.25$ and contamination rate $\tau = 0.20$. The plot setup is identical to that of Figure 2.23.

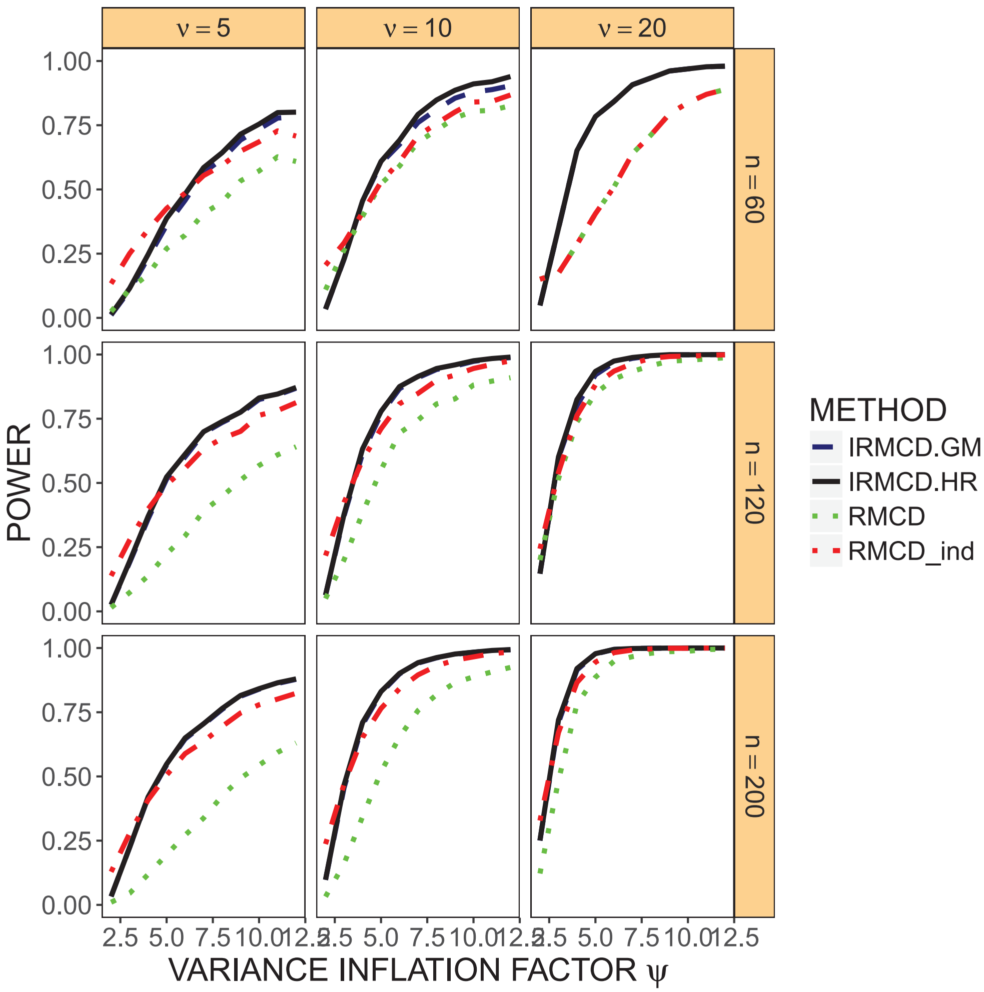


Figure 2.28: Power of MCD-based outlier detection rules under a multivariate radial contamination model for $\gamma = 0.05$ and contamination rate $\tau = 0.20$. The plot setup is identical to that of Figure 2.23.

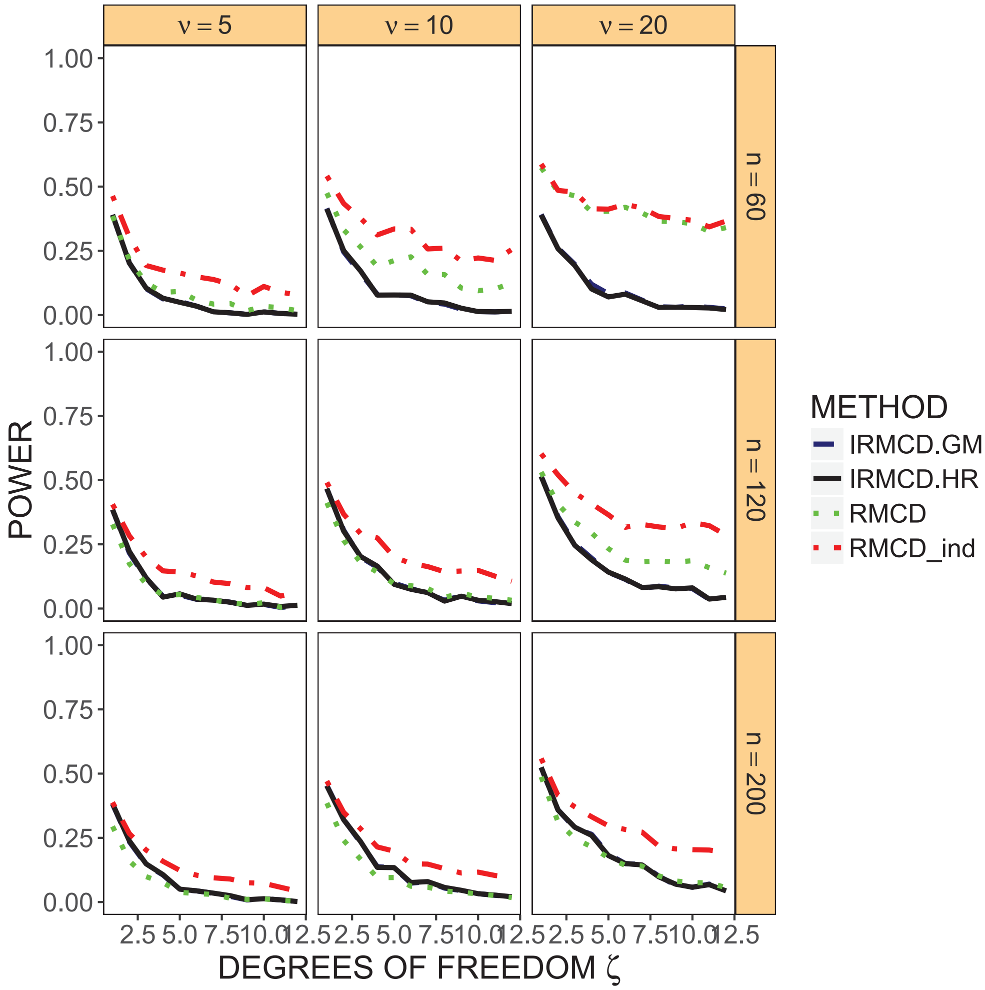


Figure 2.29: Power of MCD-based outlier detection rules under a multivariate t-distribution contamination model for $\gamma = \gamma^*$ and contamination rate $\tau = 0.05$. Results are shown for dimensions $\nu = 5, 10, 20$, depicted in columns, and sample sizes $n = 60, 120, 200$, depicted in rows. The four outlier detection rules shown are the IRMCD with our modified estimator of the Wishart parameter m (IRMCD.GM, blue dashed line); IRMCD with the original Hardin-Rocke estimator of m (IRMCD.HR, black solid line); RSDs based on the reweighted MCD and tested against a chi-squared distribution with a multiplicity-correction to the test size (RMCD, green dashed line); and RSDs based on the reweighted MCD and tested against a chi-squared distribution (RMCD_ind, red dashed line).

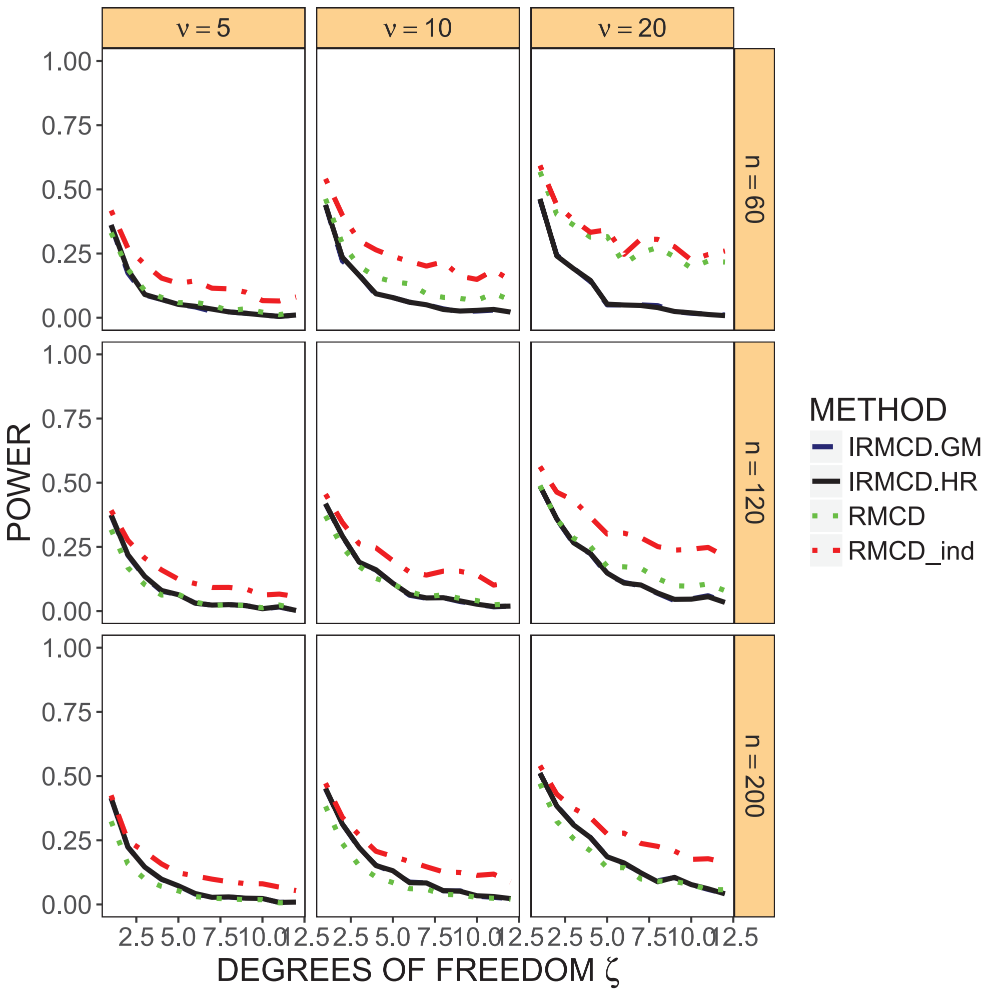


Figure 2.30: Power of MCD-based outlier detection rules under a multivariate t-distribution contamination model for $\gamma = 0.25$ and contamination rate $\tau = 0.05$. The plot setup is identical to that of Figure 2.29.

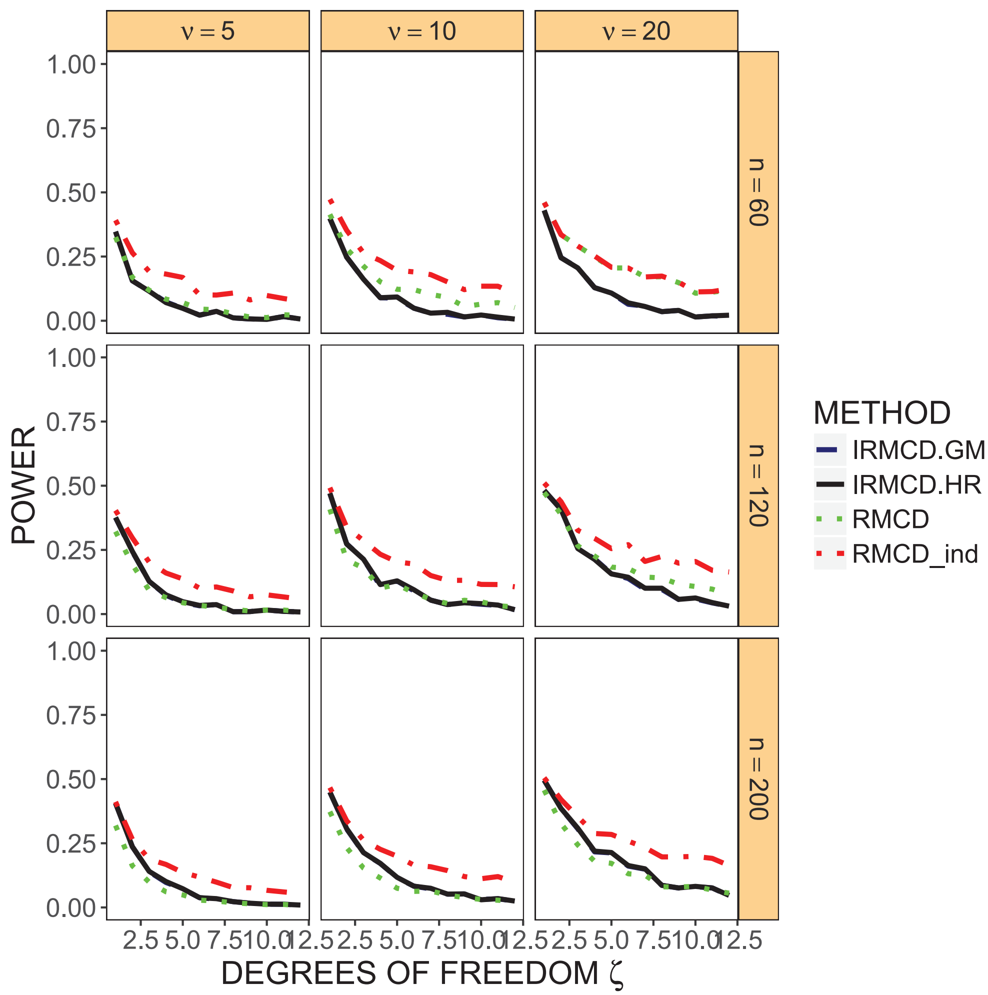


Figure 2.31: Power of MCD-based outlier detection rules under a multivariate t-distribution contamination model for $\gamma = 0.05$ and contamination rate $\tau = 0.05$. The plot setup is identical to that of Figure 2.29.

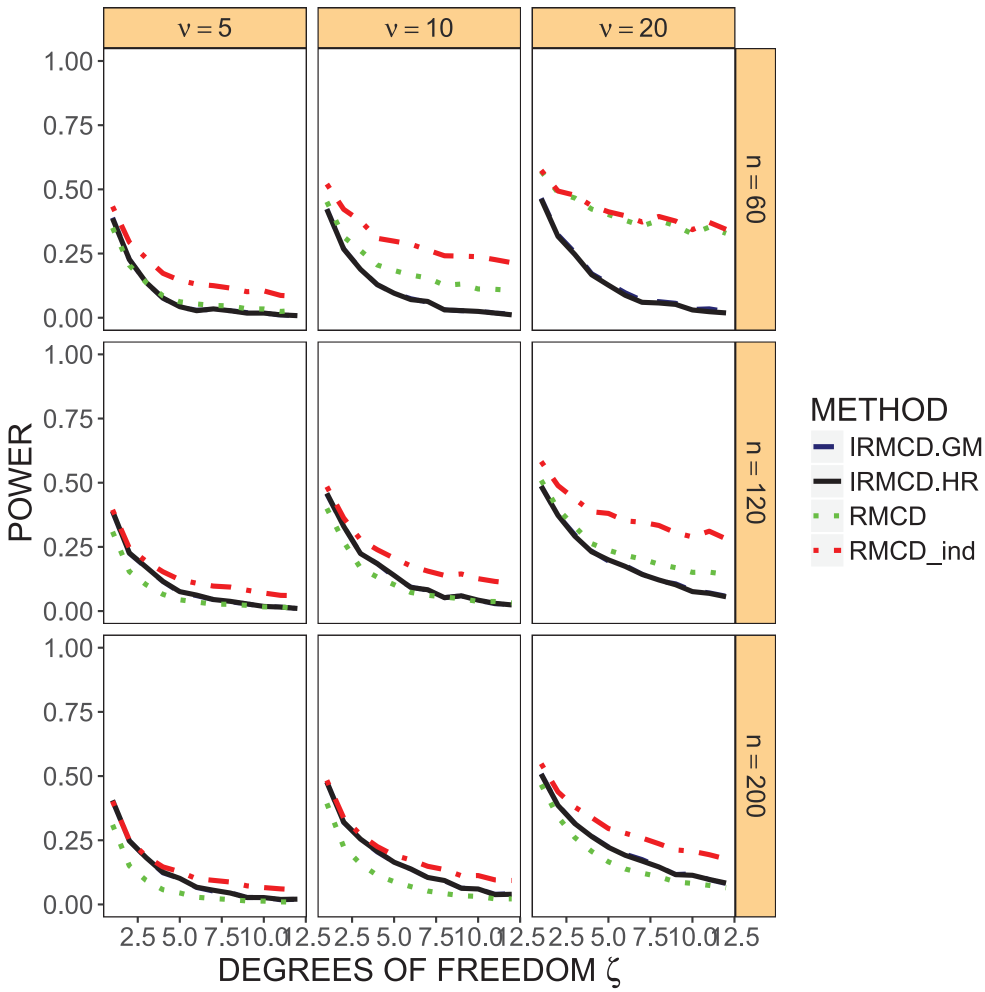


Figure 2.32: Power of MCD-based outlier detection rules under a multivariate t-distribution contamination model for $\gamma = \gamma^*$ and contamination rate $\tau = 0.20$. The plot setup is identical to that of Figure 2.29.

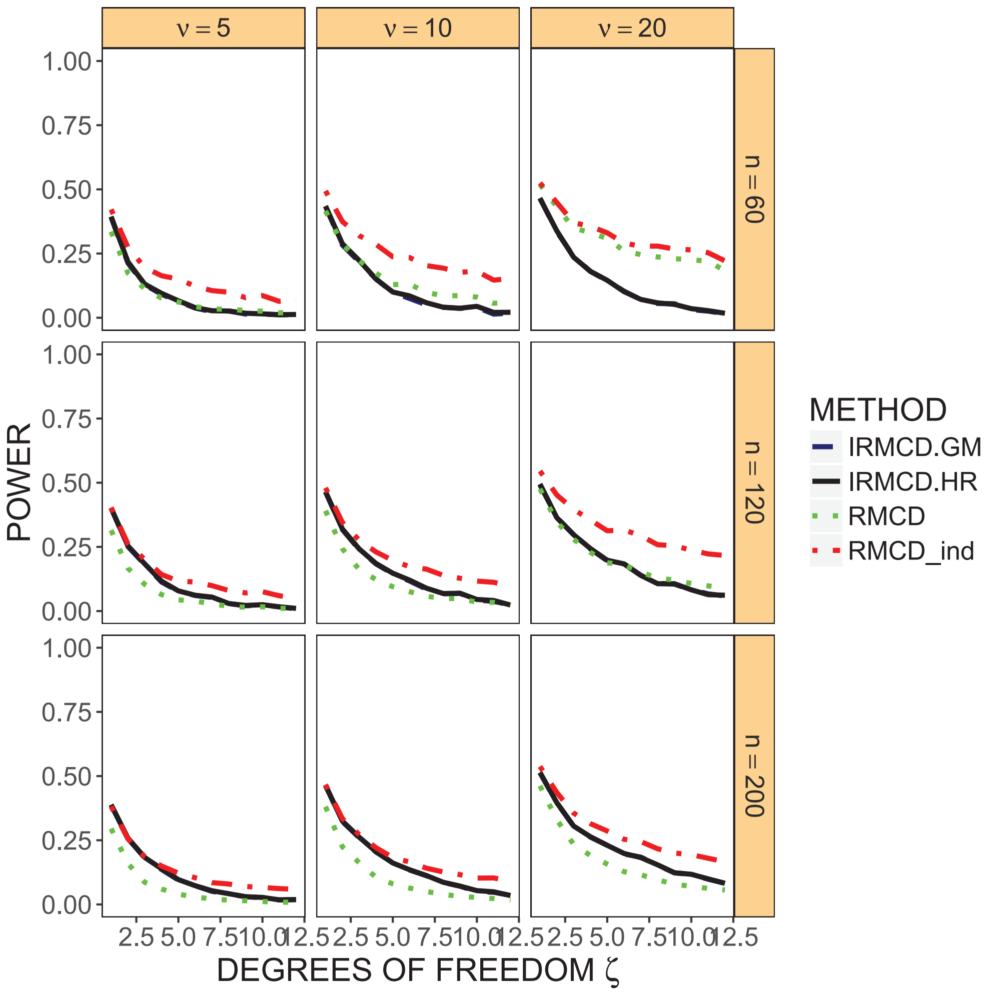


Figure 2.33: Power of MCD-based outlier detection rules under a multivariate t-distribution contamination model for $\gamma = 0.25$ and contamination rate $\tau = 0.20$. The plot setup is identical to that of Figure 2.29.

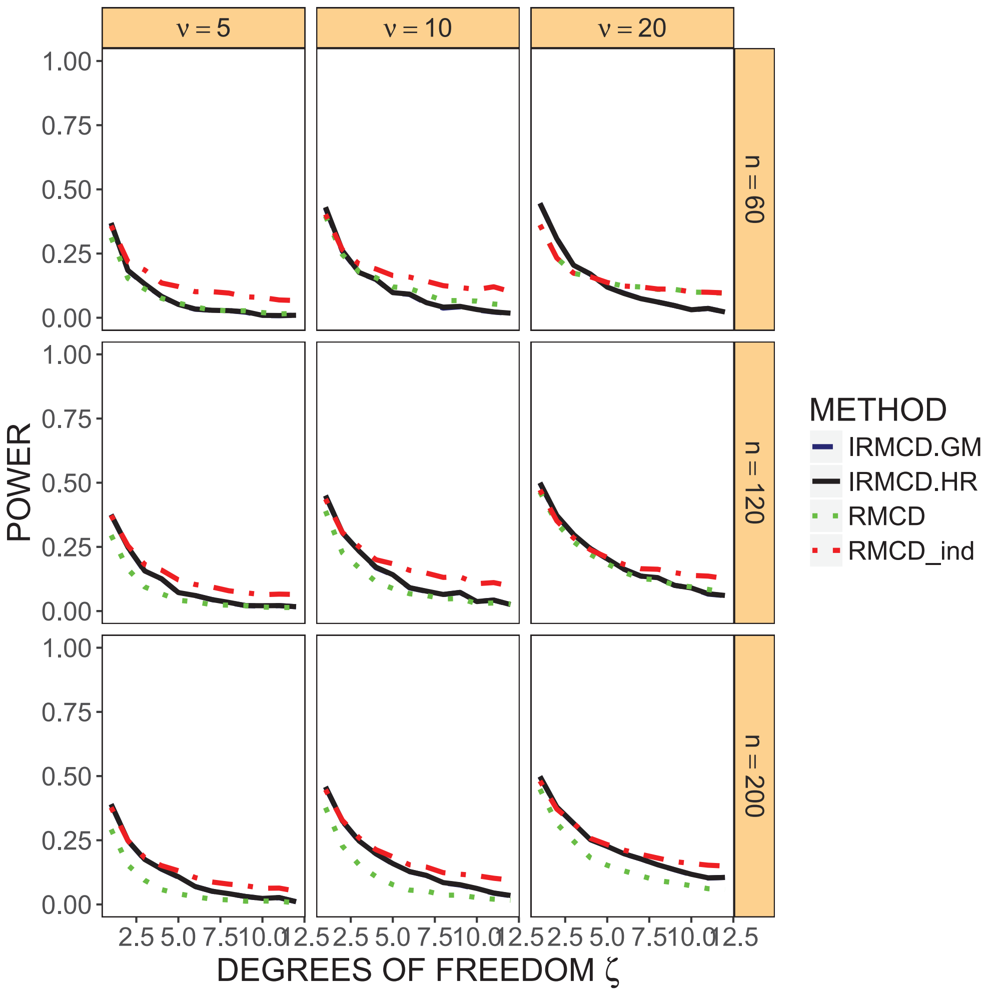


Figure 2.34: Power of MCD-based outlier detection rules under a multivariate t-distribution contamination model for $\gamma = 0.05$ and contamination rate $\tau = 0.20$. The plot setup is identical to that of Figure 2.29.

Chapter 3

ROBUST DETECTION OF MULTIVARIATE OUTLIERS IN ASSET RETURNS AND RISK FACTORS DATA

Abstract

It is well-known that outliers exist in the type of multivariate data used by financial practitioners for portfolio construction and management. Typically, outliers are addressed prior to model fitting by applying some combination of trimming and/or Winsorization to each individual variable. This approach may fail to detect and/or mitigate multivariate outliers. Existing literature documents the use of the robust Mahalanobis squared distance (RSD) based on the minimum covariance determinant (MCD) estimator to detect and to shrink multivariate outliers in financial data. We use MCD-based RSDs, along with the Iterated Reweighted MCD methodology of Cerioli, to illustrate the presence of outliers in the type of asset return and firm fundamental data that equity portfolio managers would use to build and monitor portfolios. We demonstrate how RSDs based on the MCD estimate are superior to Mahalanobis distances based on the classical mean and covariance estimates for detecting multivariate outliers. In the process, we show that univariate trimming and Winsorization are insufficient to deal with multivariate outliers in financial data.

3.1 Introduction

It is well-established in the statistics literature that outliers can adversely affect the outcome of classical statistical methods such as parameter estimation and hypothesis testing. By now, most equity portfolio managers and commercial equity portfolio management software providers recognize that outliers in the asset returns and factor model exposures can cause problems for portfolio construction and monitoring. It is common practice to reduce the influence of outliers in multidimensional asset returns and/or factor model exposures data by trimming or Winsorizing each one dimensional component of the data. Trimming refers to the removal of the extreme values in the data (e.g., by sorting the data and deleting the smallest and largest 1% of the observations). Winsorization, on the other hand, reduces extreme values rather than removing them outright: after sorting the data, the smallest and largest values are replaced by a predetermined empirical quantile.[1] Grinold and Kahn (2000) recommends examining observations more than three standard deviations from the mean of the sample for validity, then trimming those that seem dubious while Winsorizing legitimate values (pages 382–383). MSCI uses Winsorization in the calculation of their style-based equity indices (MSCI, 2016, page 10).

Trimming or Winsorizing one dimension at a time in this manner only addresses univariate outliers, however. So-called multivariate outliers are observations that are far from the bulk of the observations in dimension $\nu > 1$. For example, Kritzman and Li (2010) provides examples of so-called “turbulent” periods in multivariate returns data where the returns on assets that are typically positively correlated move in the opposite direction. The returns to the individual assets during these times may not be very outlying in the context of their individual return histories, but their joint movement during these times is unusual. Such situations can lead to incorrect portfolio allocations, as documented by Kritzman and Li (2010) and Martin et al. (2010), and inaccurate risk forecasts, as documented by Boudt et al. (2008).

Multivariate outliers can also arise quite easily in the construction of empirical asset pricing models and fundamental factor models. Suppose, for instance, that a firm has an outlier in its price history. (This outlier could be a data error—failure to correct for a stock split, for instance—or a legitimate value caused by unexpected news—a merger, a patent approval, etc.) If the price outlier is a price level shift it will lead to an isolated returns

[1] For example, to perform 1% Winsorization of a data set with 1000 observations, the observations are first sorted from smallest to largest. Then, the 10 smallest observations are replaced by the 11th smallest, and the 10 largest observations are replaced by the 11th largest.

outlier, and if it is an isolated price outlier it will result in a pair of returns outliers of opposite sign.[2] An empirical asset pricing model or a fundamental factor model might have several factors derived from price and/or returns, such as ratios of accounting variables to price or momentum factors. Hence, the price outlier can lead to a complex multivariate outlier in the factor exposures data. Detecting and appropriately adjusting for such multivariate outliers during the model construction process is crucial to creating reliable factor models.

Depending on the geometric configuration of the observations, multivariate outliers may not be outlying in each individual coordinate, and may hence be unaffected by one-dimensional trimming and Winsorization. Figure 3.1 shows a simple example of how this can happen. We simulated 100 returns from a bivariate normal distribution with both marginal means equal to 2, both marginal variances equal to 1, and correlation 0.75. We then added four returns to our simulated data set (the blue squares) that are bivariate outliers but not univariate outliers in this data configuration. The dashed lines indicate the 1% and 99% percentiles of each marginal variable (excluding the bivariate outliers). The red triangles are the observations that would be deleted (if we were trimming by 1%) or replaced with the percentiles indicated by the dashed lines (if we were Winsorizing by 1%). The black dots are observations simulated from the bivariate normal that are left unchanged by trimming or Winsorizing. The blue squares are not outlying in either variable, and are hence also not touched by trimming or Winsorization despite being clearly far from the bulk of the data.

Table 3.1 shows how the sample correlation coefficient of the data with and without the two-dimensional outliers changes after 1% trimming and 1% Winsorization of each variable. In this example, the estimated correlation between the two variables without the two-dimensional outliers is slightly lower after one-dimensional trimming, and about the same after one-dimensional Winsorization. The sample correlation coefficient of the two

[2] For instance, in October 2008, Porsche revealed it controlled nearly 75% of the outstanding stock in Volkswagen, creating a short-squeeze and temporarily driving the stock price of Volkswagen up five-fold (Norris, 2008). Another example of a transitory outlier occurred in September 2008 when an investment analyst mistakenly republished a 2002 news article about the bankruptcy of United Airlines in a newsletter, leading to a 75% drop in the stock price of United Airlines during the day (Zetter, 2008).

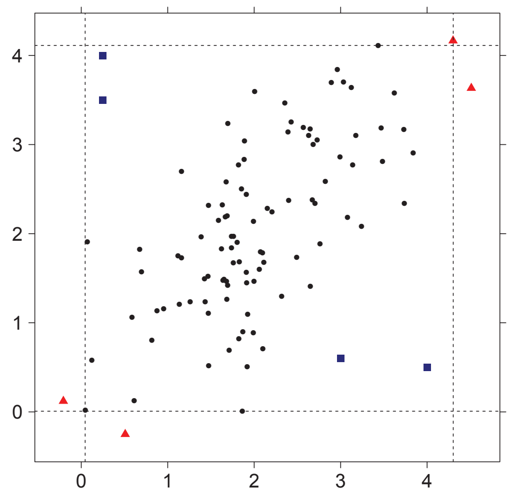


Figure 3.1: Example of two-dimensional outliers (blue squares) that are missed by one-dimensional trimming or Winsorization. The black circles are points that are not outlying and are unchanged by trimming and Winsorization. The red triangles are points that would be removed via trimming or replaced via Winsorization. The dashed lines correspond to the 1% and 99% percentiles used in 1% trimming or Winsorization.

variables including the two-dimensional outliers is substantially lower (0.526) than the correlation for the data set without the two-dimensional outliers (reported in the first row of Table 3.1). One-dimensional Winsorization of the augmented data set does not appreciably change the estimated correlation, and one-dimensional trimming makes the estimate even smaller. Clearly the two-dimensional outliers distort the estimated correlation coefficient, and the one-dimensional approaches do nothing to remedy the problem.

One-dimensional outlier detection and mitigation approaches also become less effective as the dimensionality of the data increases. Visualization of the data is difficult in dimensions higher than $\nu = 2$, so it is harder to spot multivariate outliers in exploratory data analysis without dynamic tools such as GGobi (Cook and Swayne, 2007). Furthermore, trimming or Winsorization of each variable separately defines an ν-dimensional "box" (e.g., see the dashed lines in Figure 3.1). In higher dimensions there can be more "empty space" inside this box, depending on the configuration of the observations. The empty space provides more

Case	Sample Corr. Coef.
Data without 2-D outliers, no trimming or Winsorization	0.705
Data without 2-D outliers after 1-D 1% trimming	0.647
Data without 2-D outliers after 1-D 1% Winsorization	0.704
Data with 2-D outliers, no trimming or Winsorization	0.526
Data with 2-D outliers after 1-D 1% trimming	0.438
Data with 2-D outliers after 1-D 1% Winsorization	0.522

Table 3.1: Sample correlation coefficients estimated using the data in Figure 3.1 with and without trimming, Winsorization, and two-dimensional outliers.

locations where observations can be outlying in higher dimensions without being outlying in any individual coordinate, and hence more places where such observations can escape mitigation by one-dimensional trimming or Winsorization.

A better approach to detecting outliers in multivariate financial data is the Mahalanobis squared distance (MSD), which was introduced in Chapter 2. Kritzman and Li (2010), building on earlier work by Chow et al. (1999), used sample MSDs to identify "turbulent" time periods in multivariate financial time series. The sample MSDs for each time period are given by

$$d_t^2 \equiv (\mathbf{x}_t - \overline{\mathbf{x}})^T S^{-1} (\mathbf{x}_t - \overline{\mathbf{x}}), \tag{3.1}$$

where $\overline{\mathbf{x}}$ and S are the sample mean and covariance, respectively. A time period with index t is declared "turbulent" if d_t^2 is larger than the 75th percentile of a chi-squared distribution χ_T^2 with T degrees of freedom. Kritzman and Li also show how to use sample MSDs for more reliable portfolio construction: they shrink returns flagged as outliers by the MSD criterion to the return on the minimum variance portfolio.

One drawback of using this version of the MSD is that the sample mean and covariance estimators are not robust to outliers (see, for instance, Rousseeuw and Leroy (1987) or Maronna et al. (2006)). As we discussed in Chapter 2, the so-called robust Mahalanobis squared distance (RSD) is a more reliable distance metric in the presence of outliers. We compute RSDs by replacing the sample mean and covariance estimate in Equation (3.1) with robust estimates of location and covariance that are not much influenced by outliers. The minimum covariance determinant MCD(γ) of Rousseeuw (1985) is frequently used to calculate the robust location estimate $\overline{\mathbf{x}}_{MCD}$ and covariance estimate S_{MCD} needed for RSDs. As in Chapter 2 we write this particular RSD as

$$d_t^2 \equiv \left(\mathbf{x}_t - \overline{\mathbf{x}}_{MCD}\right)^T S_{MCD}^{-1} \left(\mathbf{x}_t - \overline{\mathbf{x}}_{MCD}\right).$$

With this modification to the MSD, non-outlying points should be closer to the location estimate than outlying points, and outlying points should have larger RSDs than expected under the multivariate normal model. Scherer and Martin (2005) and Martin et al. (2010) provide several examples illustrating the use of MCD-based RSDs and demonstrating how they can find outliers missed by MSDs based on the sample mean and covariance.

Boudt et al. (2008) used MCD(γ^*)-based RSDs d_t^2 across a specified time interval to detect and shrink outliers in historical asset returns data as a pre-processing step prior to estimating modified value-at-risk (VaR) and expected shortfall (ES). They compare the distances d_t^2 to the larger of the 0.999 quantile from a chi-squared distribution and a quantile (chosen based on the loss level of the VaR/ES calculation) from the empirical distribution of the distances. This approach flags only extreme returns outliers for mitigation. The returns outliers identified during this process are then shrunk so as to have an RSD no greater than the detection threshold. This leads to VaR and ES forecasts that are not much influenced by market crashes and other multivariate returns outliers.

Getting accurate outlier detections with MCD-based RSDs rests on having a good approximation to the sampling distribution of the RSDs. The sampling distribution of MCD-based RSDs is approximately χ^2_ν for very large samples $n \geq 1000$, but it is known that for $n \leq 250$ this approximation can be quite inaccurate. In Chapter 2 we developed IRMCD2, a modified version of the the Iterated Reweighted MCD (IRMCD) of Cerioli (2010) that yields an RSD-based outlier test with accurate false positive rates. Just as for the IRMCD method, we consider two sets of outlier tests for the IRMCD2: individual hypothesis tests of each observation's outlier status

$$H_{0i} : \mathbf{x}_i \sim N\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right), \qquad i = 1, \ldots, n \tag{3.2}$$

and an "intersection" test of whether there are any outliers in the data set

$$H_0 : \{\mathbf{x}_1 \sim N\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)\} \cap \cdots \cap \{\mathbf{x}_n \sim N\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)\}. \tag{3.3}$$

Let α_1 be the desired false positive rate of the intersection test H_0. Then $\alpha = 1-(1-\alpha_1)^{1/n}$ is the corrected false positive rate for the individual hypothesis tests H_{0i}. The IRMCD2 method proceeds as follows.

1. 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 RSDs and the Hardin and Rocke (2005) F distribution:
$$\frac{c(m-\nu+1)}{m\nu} D^2_{S_{MCD}}\left(\mathbf{x}_i, \overline{\mathbf{x}}_{MCD}\right) \sim F_{\nu,m-\nu+1}.$$
Here ν is the known dimension, $\overline{\mathbf{x}}_{MCD}$ and S_{MCD} are the MCD mean and covariance estimates, respectively, and m is the degrees of freeom parameter for a Wishart distribution. The latter is estimated using the improved methodology developed in Chapter 2. Rejected observations are assigned weight 0, while all other observations receive weight 1.

3. Compute the weighted sample mean and sample covariance of the data using the weights from Step 2. (This is known as the "reweighted MCD".)

4. Test RSDs based on the weighted covariance estimate 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[3]
$$d_i^2 \sim \frac{(w-1)^2}{w} \text{Beta}\left(\frac{\nu}{2}, \frac{w-v-1}{2}\right), \tag{3.4}$$

[3]Gnanadesikan and Kettenring (1972) pointed out that the exact distribution of an MSD computed using the sample mean and covariance is a scaled Beta distribution, a result proved earlier by Wilks (1962). Atkinson et al. (2004) provides an easier proof of the result.

with w equal to the number of observations receiving weight 1.

For observations with weight 0, we test RSDs against a scaled F distribution:

$$d_i^2 \sim \frac{w+1}{w}\frac{(w-1)\nu}{w-\nu}F_{\nu,w-\nu}. \tag{3.5}$$

These tests are performed using a false positive rate of α.

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.

We have implemented the IRMCD2 method, with the extensions discussed in Chapter 2 to improve accuracy when using $\gamma \leq 0.05$ and $n \leq 250$, in the `R` package `CerioliOutlierDetection` (Green and Martin, 2014). In what follows we will refer to the combination of MCD-based RSDs with the IRMCD2 detection approach as RSD-IRMCD2, and use the notation RSD-IRMCD2(γ) to refer to distances computed using MCD(γ) and tested using IRMCD2.

In this chapter we will motivate the use of RSD-IRMCD2 for accurate detection of multivariate outliers in portfolio returns and factor data.[4] We will show, via several examples, that RSD-IRMCD2 is superior to Mahalanobis distances based on classical means and covariance and tested against chi-squared quantiles. In the process, we will demonstrate that multivariate outliers in asset returns and in cross-sectional factor models are a frequently occurring phenomena about which asset pricing researchers and portfolio managers need to be very concerned.

We will illustrate the RSD-IRMCD2 outlier detection methodology using (a) example historical returns based data sets similar to what one would use for portfolio construction and monitoring; and (b) example multiple fundamental factor model data that one would use

[4] Another important potential application of RSD-IRMCD2 that we do not cover here is the sequential detection of outliers, also known as an "unusual movement test" (Scherer and Martin, 2005). This can done using a moving window and RSD-IRMCD2.

for returns forecasting and portfolio risk and performance assessment. We detect potential outliers using MSDs based on classical and robust estimates of location and dispersion. For the multivariate returns examples in Section 3.2, we ran our MCD-based outlier detection tests using $\gamma = 0.10$ due to the relatively small sample sizes of the returns data sets. For the factor exposure examples, we considered $\gamma = \gamma^*$, $\gamma = 0.25$, and $\gamma = 0.05$. With smaller values of γ, the MCD(γ) estimator trims only the most extreme outliers and is hence very conservative, but does not offer much protection against a larger number of outliers. By considering three trimming fractions we can offer the practitioner some guidance on good choices of γ for everyday use.[5]

An early version of this work was presented in abbreviated fashion in Section 11.6.2 of Martin et al. (2010). In that analysis, multivariate outliers were identified via RSDs and critical values determined via the asymptotic chi-squared distribution for such distances. Since the publication of that paper, the shortcomings of the chi-squared distribution for testing MSDs have come to light through the work of Cerioli et al. (2009). Thus, we have redone and extended some of the analyses of that paper using our IRMCD2 method to test whether those earlier results still hold. This chapter also provides much more detail on the nature of the outliers than before and includes new analyses of multivariate outliers in the factor exposures of cross-sectional factor models for asset pricing and portfolio construction.

The remainder of this chapter is organized as follows. Section 3.2 illustrates how to use RSD-IRMCD2 outlier detection to find unusual times in multivariate returns time series. Section 3.3 shows the advantages of the technique for detecting multivariate outlying assets in a four factor asset pricing model. Section 3.4 illustrates the methodology on a ten factor fundamental factor model used for asset return forecasts. Section 3.5 summarizes our findings and suggests topics for further research. Appendix 3.A provides some examples of how one-dimensional trimming and Winsorization fail to address multivariate outliers in the four

[5] All calculations were performed using a laptop running Windows 7 Ultimate SP 1 and `R` 3.1.3 with an Intel® Core™ i7-3740QM processor running at 2.7GHz and 32GB of RAM. The MCD(γ) is available in `R` via the function `covMcd` in the `robustbase` package (Rousseeuw et al., 2016).

factor asset pricing model. Appendix 3.B describes the construction of the data set used in Sections 3.4 and 3.A.

3.2 Outlier Detection in Asset Returns

The goal of outlier detection for the returns data sets is to find unusual times. We compute the mean vector and covariance estimate of the historical returns over time, and compute one set of MSDs, indexed by time, for each detection methodology. We consider (a) MSDs based on the classical mean and covariance and tested against chi-squared quantiles; (b) MCD-based RSDs tested against chi-squared quantiles; and (c) the RSD-IRMCD2 method. Although we do not recommend testing MCD-based RSDs against chi-squared quantiles, we include this approach here to illustrate how it compares to RSD-IRMCD2. In the tests against chi-squared quantiles, we test distances for significance using a nominal false detection rate of $\alpha = 1\%$. In the RSD-IRMCD2 tests we use a nominal false detection rate of $\alpha_1 = 1\%$ for the intersection hypothesis. The corresponding rate for the first set of IRMCD2 tests (Step 4 of the IRMCD2 procedure) is computed as $\alpha = 1 - (1 - \alpha_1)^{1/T}$, where T is the number of time periods. If at least one time period is rejected in this step, the distances are retested at the nominal rate $\alpha_1 = 1\%$.

We illustrate the use of RSD-IRMCD2 on the commodity and hedge fund data sets used in Martin et al. (2010).

3.2.1 Commodity Data

Description of the Data Set

The commodity data set consists of monthly returns on seventeen commodity contracts selected from the Reuters CRB Index over the period February 1999–December 2008. The seventeen commodities are the following.

Petroleum Products crude oil, heating oil, natural gas

Metals gold, silver, copper, platinum

Soft Commodities corn, soybeans, wheat, cattle, hogs, cocoa, coffee, sugar, cotton, orange juice

Outlier Detection Results

Figure 3.2 shows the Mahalanobis distances at each month for the classical mean and covariance (left panel), robust MCD(0.10) (middle panel), and RSD-IRMCD2(0.10) (right panel), respectively. The first two use the quantile $\chi^2_{17,0.99}$ as the threshold for identifying outliers, while the third uses the IRMCD2 approach with an equivalent nominal false positive rate of 1% for the intersection hypothesis test. Recall that the IRMCD2 approach uses a different detection threshold depending on whether an observation was included in the reweighted MCD calculation: for observations receiving weight 1 are thresholded against a scaled Beta distrubtion (Equation (3.4)), while observations with weight 0 are tested against a scaled F distribution (Equation (3.5)).

The distances based on the classical estimates, tested using the chi-squared threshold, only identify three possible outliers: January 2000, September 2008, and October 2008. The latter two outliers correspond to the height of the 2008 financial crisis. By replacing the classical estimates with the MCD(0.10) estimates, we detect those three outliers as well as eight other possible outliers: June 1999, November 1999, January 2001, December 2003, May 2004, July 2004, March 2008, and November 2008. Using the RSD-IRMCD2(0.10) methodology instead of the chi-squared threshold tells us that the June 1999, November 1999, March 2008, November 2008 are likely false alarms, however.

The 17-dimensional nature of the data makes visualization impractical. Pairwise scatterplots alone will not be sufficient to spot multidimensional outliers. They are, at best, helpful to understand the configuration of some, but not all, of the outliers. Figure 3.3 shows pairwise plots for a subset of the commodities: heating oil, crude oil, copper, gold, and platinum. The three outliers found by both the classical Mahalanobis distances and

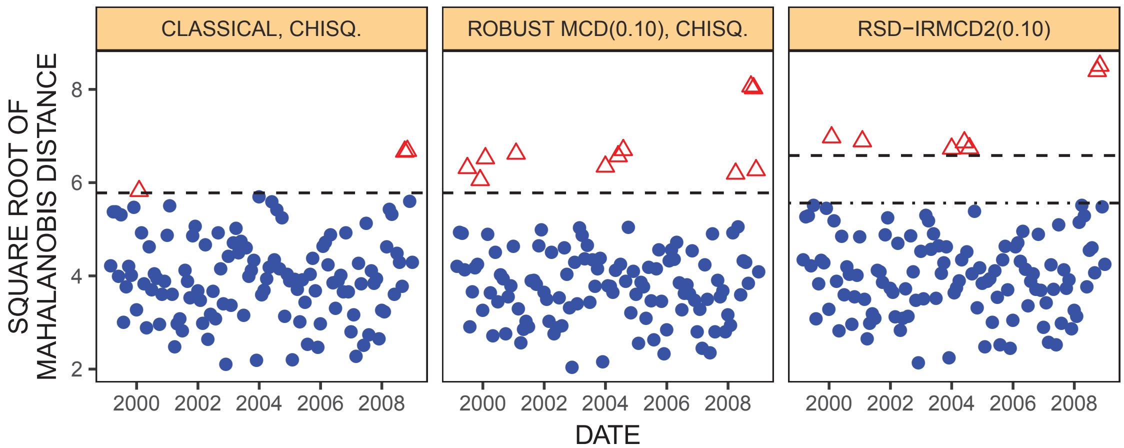


Figure 3.2: Mahalanobis distance plots for commodity data. In each panel we plot the distance (computed using the method listed at the top of the panel) for each time. Blue dots correspond to non-outlying times, while red triangles indicate possible outliers. The methods used are, from left to right, the sample mean and covariance with a chi-squared threshold; the MCD(0.10) location and covariance estimate with a chi-squared threshold; and the RSD-IRMCD2(0.10) approach. The dashed line(s) in each panel indicates the detection threshold(s). Recall from Section 3.1 that the RSD-IRMCD2(0.10) method has two such thresholds given by the distributions in Equations (3.4) and (3.5). The former is shown as a "dot-dashed" line in the third panel, while the latter is shown as a dashed line.

RSD-IRMCD2(0.10) are marked with green triangles, while the additional outliers found only by RSD-IRMCD2(0.10) are marked with red stars. October 2008 (one of the green triangles in the lower left corner of each plot) is a very obvious outlier in this view of the data, and is outlying in each of the five variables shown. September 2008 is also an obvious outlier for the three metals, but is not outlying in the joint plot of heating oil and crude oil. The configuration of the January 2000 outlier is less obvious: that date is a positive outlier for heating oil, but is not especially outlying in any other variable, nor is it the only positive outlier for heating oil.[6] Rather, it is the relationship between heating oil and the

[6] According to historical data from the U.S. Energy Information Administration, the spot residential price for No. 2 Heating Oil jumped from \$1.193/gallon on January 17, 2000, to \$1.615/gallon by January 24, 2000, on increased demand due to a particularly severe winter storm (U.S. Energy Information Administration, 2017; Mariner-Volpe, 2000). This jump is the cause of the large positive outlier for January 2000 in the heating oil futures returns series.

other commodities during January 2000 that is abnormal. This behavior is to be expected when external circumstances, e.g., unusually cold weather, have a disproportionate effect on a single commodity.

Figure 3.4 shows pairwise plots for subset of the soft commodities: The December 2003 outlier is a very obvious outlier in the cattle futures time series, and possibly only in cattle futures.[7] Yet it was not found via the classical distances with the chi-squared quantiles: the two 2008 outliers are much more extreme in magnitude, and they mask more moderate outliers. The RSD-IRMCD2 approach is much less susceptible to this phenomenon. Likewise, the May 2004 outlier is mainly an outlier in soybeans that is masked by more extreme outliers.[8] Again, these times are outlying due to commodity-specific events that lead to a temporary breakdown in the usual relationships between the commodities.

3.2.2 Hedge fund data

Description of the Data Set

The hedge fund data set consists of monthly returns on fourteen hedge funds over the period October 1999–September 2004. This data set was included in the FinAnalytica, Inc. (`http://www.finanalytica.com/`) Cognity portfolio optimization and risk management software. The hedge funds have been anonymized and will be referred to as F1, F2, . . . , F14.

Outlier Detection Results

Figure 3.5 shows the Mahalanobis distances for the classical mean and covariance (left panel), robust MCD(0.10) (middle panel), and RSD-IRMCD2(0.10) (right panel). Again the first two use the quantile $\chi^2_{14,0.99}$ to flag outliers while the RSD-IRMCD2 case is calibrated to have

---

[7] In December 2003 the U.S. Agriculture Department confirmed a case of mad cow disease in the U.S. cattle herd. This lead to a 21% drop in cattle futures by the end of 2003 (U.S. Department of Agriculture, 2003; The Associated Press, 2003). Hence, this outlier may be specific to the cattle futures time series.

[8] Soybean prices experienced a sharp decline in May 2004 due to a combination of overproduction and diminshed demand (Ash and Dohlman, 2005).

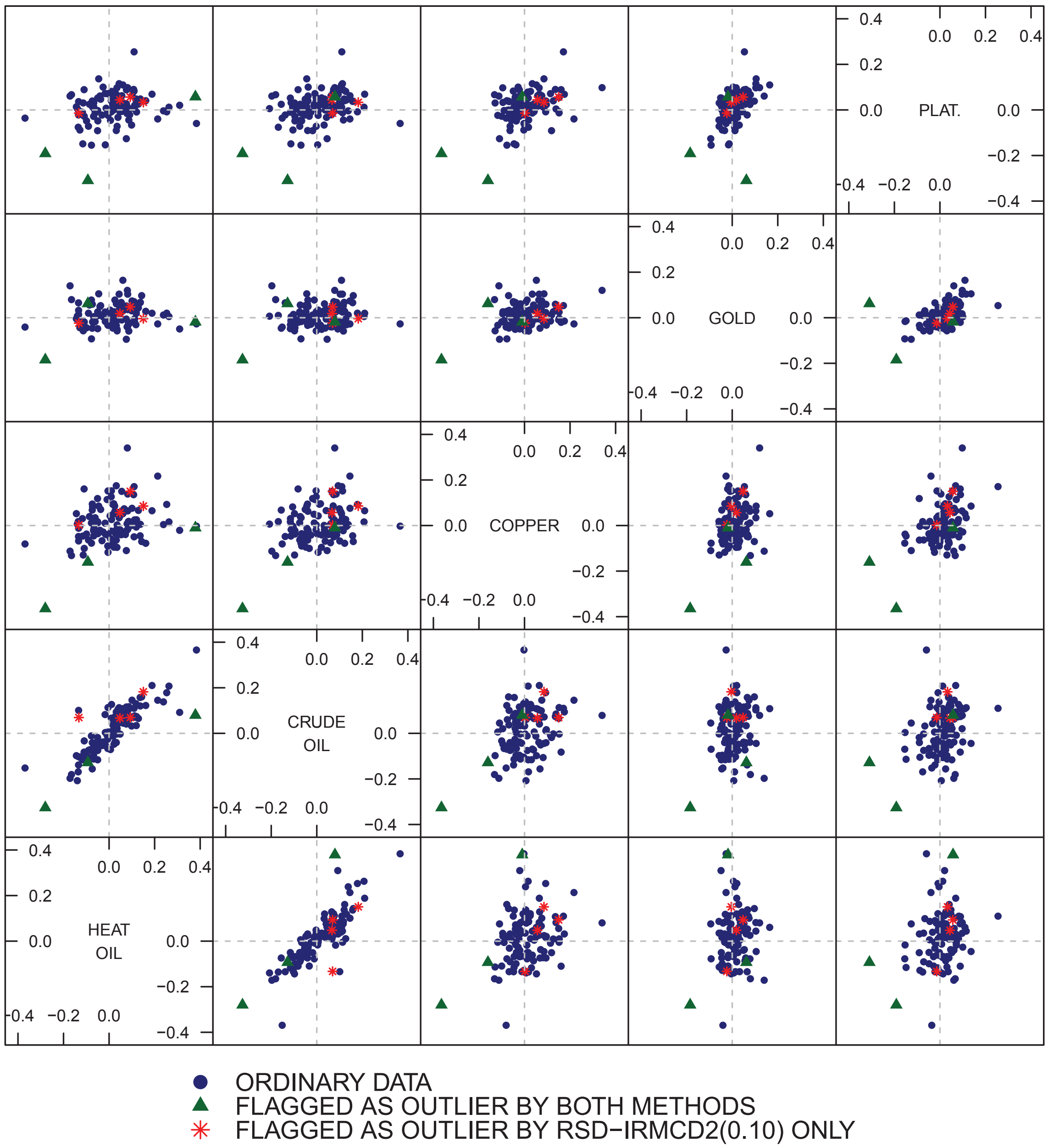


Figure 3.3: Pairwise scatterplots of selected petroleum products and metals. From bottom left to top right: heating oil, crude oil, copper, gold, and platinum. Observations flagged as outliers by Mahalanobis distances using classical estimates and the RSD-IRMCD2(0.10) method are plotted as green triangles. Outliers detected only by RSD-IRMCD2(0.10) are plotted as red stars. Non-outlying points are plotted as blue dots.

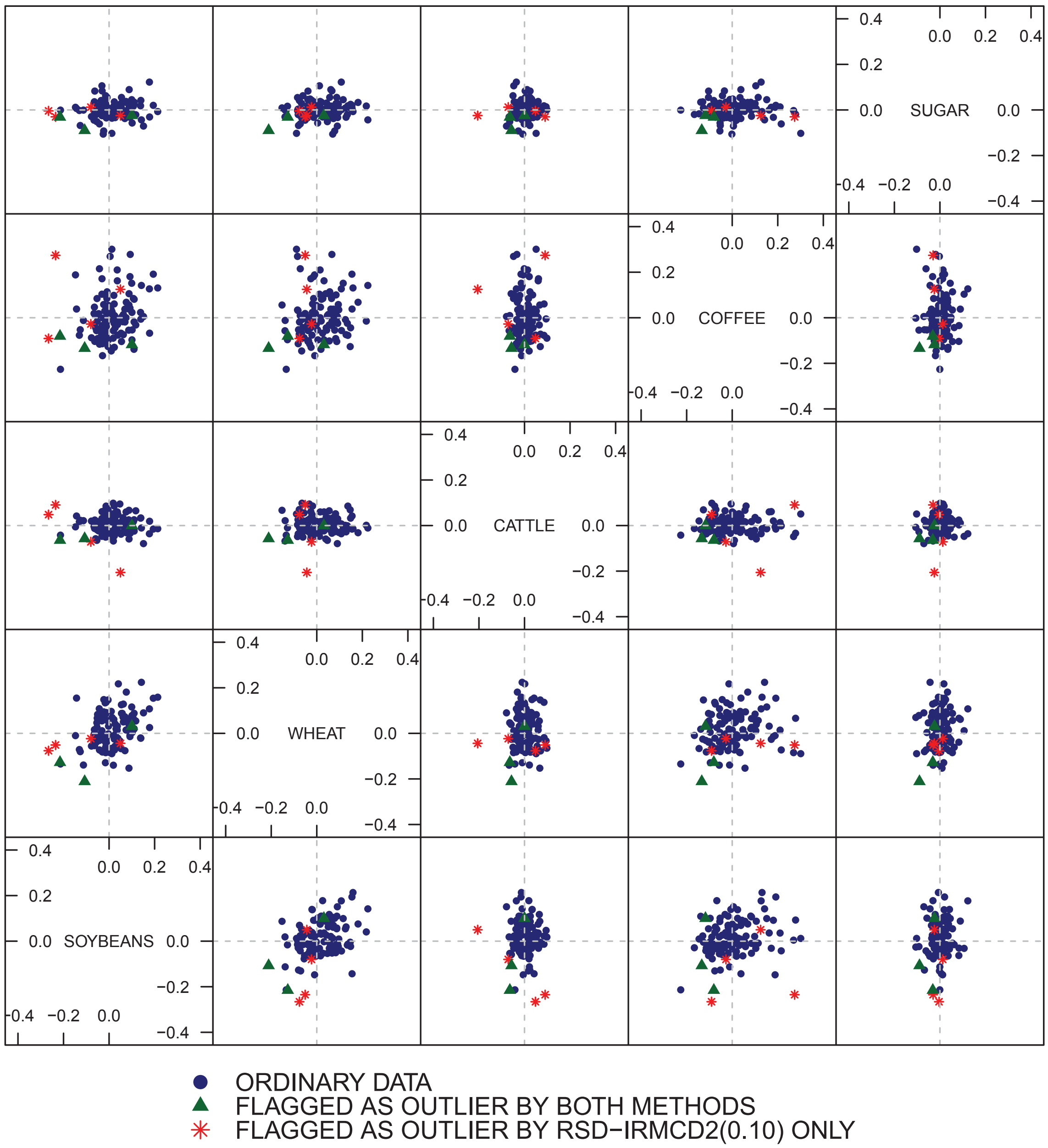


Figure 3.4: Pairwise scatterplots of selected soft commodities. From bottom left to top right: soybeans, wheat, cattle, coffee, and sugar. The plot setup is identical to that used in Figure 3.3.

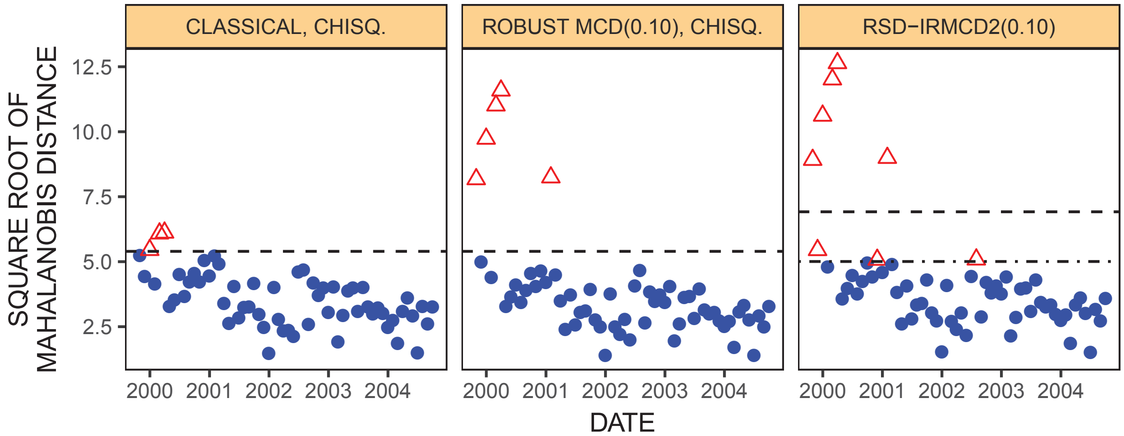


Figure 3.5: Mahalanobis distance plots for hedge fund data. The plot setup is identical to that used in Figure 3.2.

a false positive rate of 1% for the intersection hypothesis that the data set is outlier-free.

With the classical estimates we only find three outliers, all associated with the dot-com period: December 1999, February 2000, and March 2000. The MCD(0.10)-based distances tested against the chi-squared quantile finds two additional outliers at October 1999 and January 2001. The RSD-IRMCD2 approach finds the same set of five outliers and hence offers some reassurance that they are not false positives. It also finds three additional moderate outliers at November 1988, November 2000, and July 2002. These are observations that received weight 1 in the reweighted MCD portion of the IRMCD2 procedure, and the detection threshold is determined by a scaled Beta distribution.

Figures 3.6 and 3.7 show pairwise scatterplots of subsets of the hedge funds. (We have omitted funds F1, F7, F12, and F13 to keep the plots readable.) In these figures the three outliers found by both the classical distances and RSD-IRMCD2(0.10) are marked with green triangles, while the additional two outliers found by the robust method are marked with red stars. The three outliers found by the classical approach are very clearly outlying in most pairwise plots. The two additional outliers found using either of the robust approaches are extreme outliers in some pairwise plots and moderate in others. The classical approach fails to detect these because of the influence of the very extreme outliers (December 1999, February 2000, and March 2000) on the classical sample covariance matrix estimate and the resulting distortions to the Mahalanobis distances. Furthermore, we again see that the pairwise plots are not an effective means of finding moderate outliers.

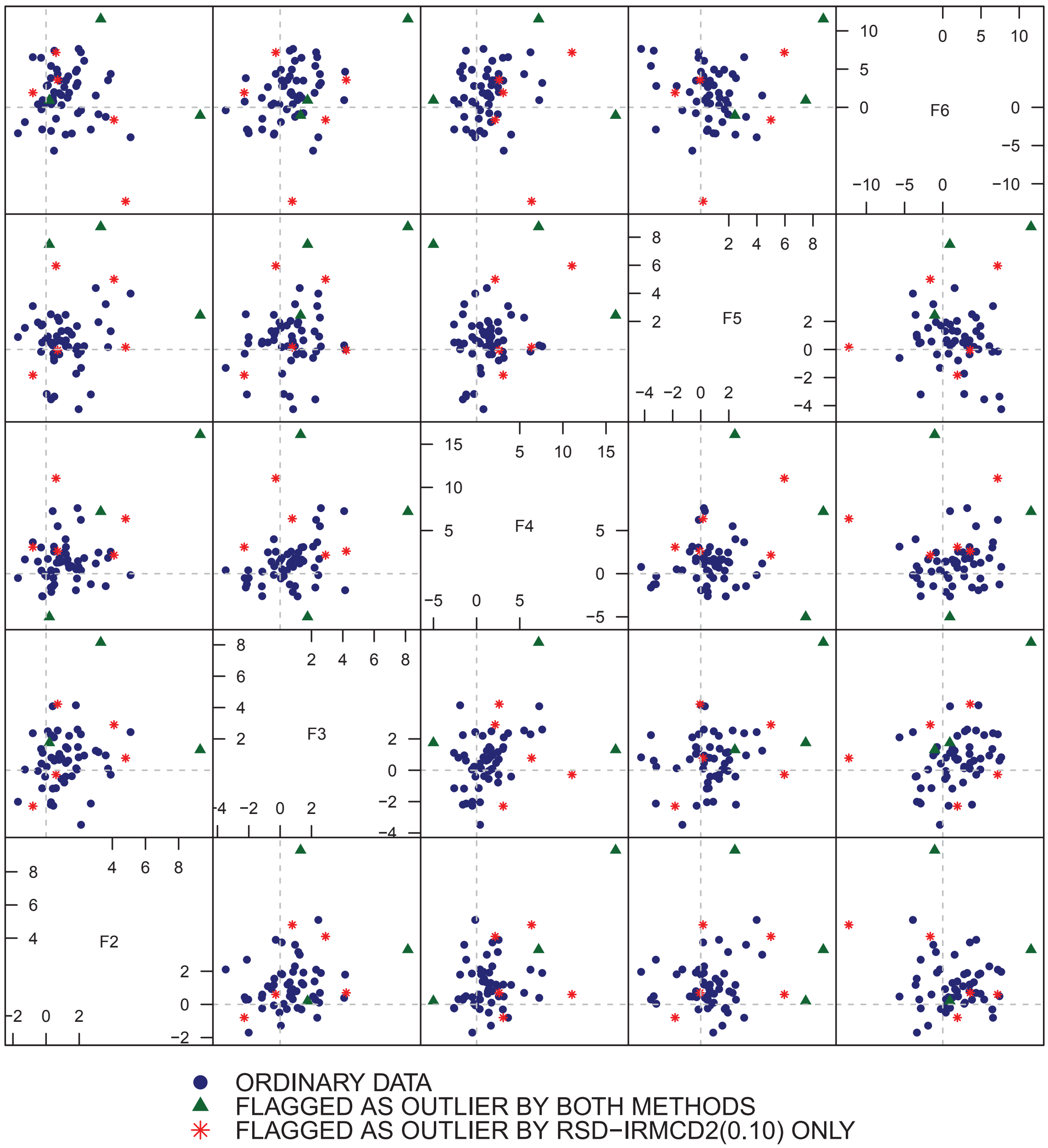


Figure 3.6: Pairwise plots for hedge funds F2–F6. The plot setup is identical to that used in Figure 3.3.

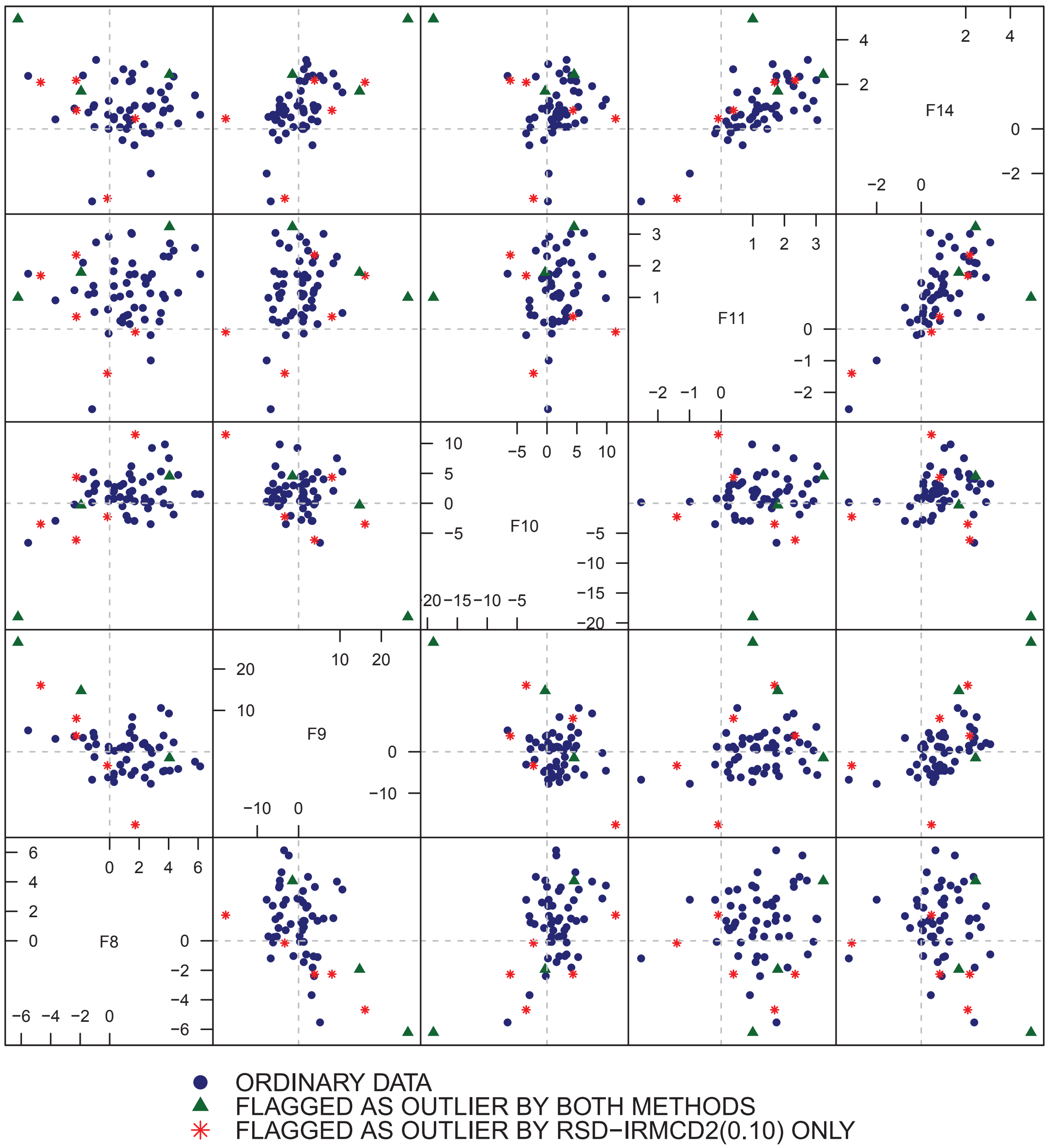


Figure 3.7: Pairwise plots for hedge funds F8-F11 and F14. The plot setup is identical to that used in Figure 3.3.

3.3 Outlier Detection in a Four-Factor Asset Pricing Model

We will illustrate the RSD-IRMCD2 method on a four-dimensional multivariate data set of factor exposures (also called "loadings") for an empirical asset pricing model. An asset pricing model based on these factors would be estimated using cross-sectional regression. At each point in time t, we would regress the N asset returns on a set of K factor exposures:

$$r_{i,t} = \alpha_{i,t} + \beta_{i,1,t} f_{1,t} + \cdots + \beta_{i,k,t} f_{k,t} + \epsilon_{i,t}. \tag{3.6}$$

Here $r_{i,t}$ is the return on the ith stock at time t; $\beta_{i,j,t}$ is the exposure of the ith stock to the jth factor at time t; $f_{j,t}$ is the return to the jth factor at time t; $\alpha_{i,t}$ is a stock-specific return in excess of the factor returns, and $\epsilon_{i,t}$ is the (transitory) residual return on the ith stock at time t. (See, for instance, Zivot and Wang (2003) for more details.)

In a fundamental factor model, the factor exposures $\beta_{i,j,t}$ are known, and the factor returns $f_{j,t}$ must be estimated. Thus at each point in time, the explanatory data for the regression consists of an $N \times K$ matrix of exposures on each factor for each asset. (In this section, $K = 4$.) Our goal in this application of RSD-IRMCD2 is to detect outlying assets at each time, i.e., four-dimensional vectors of factor exposures that do not fit in well with the rest of the data. We will discuss both point-in-time results and trends in the percent of outliers detected over time.

3.3.1 The Data

We use factor exposure data for four common factors used by the well-known Fama and French (1993) and Carhart (1997) asset pricing models for U.S. stocks over the period December 1, 1985, to December 31, 2012.[9]

- A "size" factor (denoted "LOGME" herein), the natural logarithm of the firm's market capitalization (in millions), defined as the product of the number of shares outstanding and the share price.

[9] These four factors are also commonly used in fundamental factor models for portfolio construction and risk management.

- The book-to-market ratio ("BOOK2MARKET"), the ratio of a firm's book value to its market capitalization. This compares how much a company is worth "on paper" to the market's perception of the company's total value, and can be used by investors to identify over- or under-valued companies, as well as to define "value" and "growth" investment strategies.

- The earnings-to-price ratio ("EARN2PRICE"), the ratio of a company's trailing 12-month earnings per share to its price. The earnings-to-price ratio, also known as the earnings yield, allows investors to compare a company's earnings to bond yields (e.g., to see if they are being compensated for taking equity risk). The reciprocal of this measure, the price-to-earnings ratio, captures how much the market is willing to pay for a firm's future earnings, and has a long history of use as a valuation signal.

- A "momentum" factor ("MA12"), a 12-month moving average of a firm's past (raw) returns. This is intended to capture the observed phenomenon of trends in stock returns, as documented by Jegadeesh and Titman (1993, 2001) and others. Several researchers (e.g., Jegadeesh (1990)) have documented a reversal effect in the previous month's returns, so we omit the most recent month in calculating our momentum factor.

Appendix 3.B explains how the data set was constructed.

We divide our data set into groups based on a firm's market capitalization. It is generally accepted that the shape of the distribution of a company's stock returns tends to vary with the size of a company, e.g., large, established companies (like Microsoft or Wal-Mart) tend to have less variability in their price returns (and more normally distributed returns) while small companies that are still growing have more volatile returns (and more non-normally distributed returns).[10] This higher volality and non-normality in smaller capitalization stocks can be caused by outliers, though they are certainly not the only cause of the phenomenon.

[10]Some empirical support for this observation can be found in Amaya et al. (2013) and Blau et al. (2013).

Thus, we might see outlier dynamics in smaller stocks that are significantly different from those observed in larger stocks. We therefore split the factor exposure data set into quartile data sets according to the market capitalization of a firm as measured on June 30 of each year.[11] (Further details are available in Appendix 3.B.)

The first market capitalization quartile contains lots of outliers, as very small stocks may be thinly traded and can exhibit very wild swings in their returns. We will find many outliers in such data using our method, which might bias the results of our overall study in favor of RSD-IRMCD2. Moreover, many portfolio managers would not include these stocks in their portfolio construction universe due to the high transaction costs and capacity constraints of such small stocks. We therefore exclude all stocks in the first market capitalization quartile from our study.

3.3.2 Outlier Detection Results

For the factor model data sets, we estimate the mean and covariance matrix of the factor exposures over assets in each month using four methods.

- The "classical" method, which uses the sample mean and covariance to compute Mahalanobis distances, and tests the squared distances against a chi-squared quantile.

- The RSD-IRMCD2(0.50) method, which uses the location and covariance estimates from the MCD(γ^*) estimator to compute RSDs, and tests the RSDs using the IRMCD2 method. (Recall from Chapter 2 that $\gamma^* \approx 0.50$ in large samples.)

- The RSD-IRMCD2(0.25) method, which uses the location and covariance estimates from the MCD(0.25) estimator to compute RSDs, and tests the RSDs using the IRMCD2 method.

[11] We also conducted the experiment using (a) quintiles as breakpoints; and (b) typical breakpoints for "smallcap", "midcap", and "largecap" stocks. We obtained qualitatively similar results in each case, so it does not appear that our findings are strongly influenced by how the stocks are partitioned into groups.

- The RSD-IRMCD2(0.05) method, which uses the location and covariance estimates from the MCD(0.05) estimator to compute RSDs, and tests the RSDs using the IRMCD2 method.

Among MCD(γ) estimators, the maximum breakdown point MCD(γ^*) estimator is the most robust to contamination (in the sense of the breakdown point: see Lopuhaä and Rousseeuw (1991)) but is not a very efficient estimator when the underlying data are multivariate normal. The MCD(0.25) has been proposed as a compromise between efficiency and breakdown point (see, for example, Croux and Haesbroeck (1999)), and is another commonly used version of the MCD estimator. We included the MCD(0.05) estimator as well, as some practitioners may be hesistant to discard a significant fraction of the data. It is very efficient when the data follow a normal distribution, but has a 5% breakdown point. All of the MCD variants are implemented in the `rrcov` `R` library (Todorov and Filzmoser, 2009).

In each month, we flag assets whose four-dimensional vector of factor exposures is outlying. As described in Section 3.1, the IRMCD2 methodology ensures the false positive rate of the outlier detection test based on RSDs in each month is accurate, e.g., if we use $\alpha_1 = 2.5\%$ in the intersection test, we expect to detect outliers purely by chance 2.5% of the time.[12] We are running the outlier detection tests once a month for 325 months (the number of months in the data set), however, so we expect to declare incorrectly that the cross-sections for $0.025 \times 325 \approx 8$ months have at least one outlier. This is also true for the Mahalanobis distances based on the classical mean and covariance and tested against chi-squared quantiles. In order to reduce the occurrence of such false alarms, we run each month's test using a Bonferroni-corrected significance level of $2.5\%/325 \approx 0.008\%$. For the classical distances, we compare squared distances against the $1 - 0.025/325 \approx 0.99992$ quantile of the chi-squared χ^2_4 distribution. For the RSD-IRMCD2 cases, we use $\alpha_1 = 0.008\%$ in the intersection test. At this level of significance we would expect to see 8 observations in a

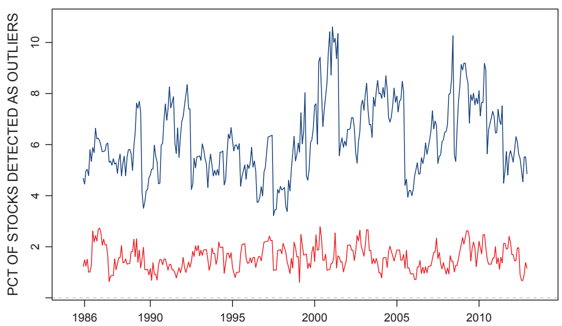
[12]Cerioli et al. (2009) showed that testing MCD-based RSDs against quantiles from a chi-squared distribution had a higher-than-expected false positive rate, especially in sample sizes less than 500. The IRMCD2 method was developed to correct this problem.

sample of size $100,000$ (i.e., the number of stocks multipled by the number of time points) flagged as outliers purely by chance.

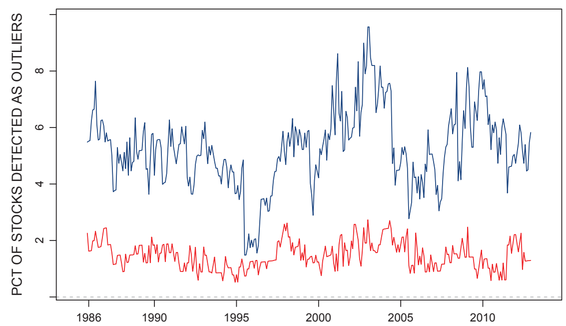
Figures 3.8a–3.8c show the percentage of observations in the four-factor fundamental factor model data sets that were flagged as outliers using (a) the classical method (blue lines) with the Bonferroni-adjusted quantile $\chi^2_{4,1-(0.025/325)}$; and (b) the RSD-IRMCD2(0.05) method (red lines) at a Bonferroni-adjusted nominal false positive rate of 2.5%/325. (The plots for the RSD-IRMCD2(0.50) and RSD-IRMCD2(0.25) methods are similar to those for RSD-IRMCD2(0.05) and are omitted to make the charts easier to read.) The robust method generally finds more outliers than the classical method, and often finds significantly more outliers. Table 3.2 shows the maximum number of outliers detected by each method for each data set. The classical approach never finds more than 13 outliers in any data set at any time.

Surprisingly, the percentage of stocks identified as outliers by the classical method did not change dramatically over time even though the stock market experienced some major volatility episodes during the time period covered by the sample. The robust methods, on the other hand, detected more violations of multivariate normality at many points in time. The robust estimates from Figures 3.8a-3.8c suggest a market that is more volatile and non-normal from the late 1990s onwards. Even if we restrict our attention to the largest, most liquid stocks of the Quartile 4 group (Figure 3.8c), there are notable peaks in the robust outlier series from 2000–2004, and again from 2008–2011. These results correspond more closely to the actual volatility of the market during that time than the classical method results.

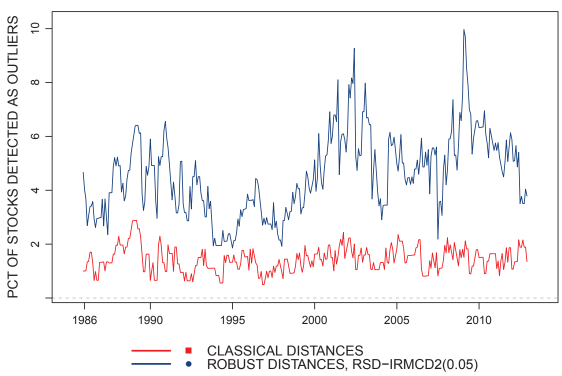
Table 3.3 provides the actual numbers depicted in Figure 3.8c for during 2000–2004. The number of outliers detected by all methods is elevated for the 2000–2004 period. Starting in 2000, however, the robust methods consistently flag more observations as outlying than the classical method. Even the RSD-IRMCD2(0.05) method, which discards less observations in the MCD estimates compared to the RSD-IRMCD2(0.25) and RSD-IRMCD2(0.50) methods, flags two to three times as many points as the classical method for most months during the



(a) Quartile 2



(b) Quartile 3



(c) Quartile 4

Figure 3.8: Percentage of stocks detected as outliers by classical and robust Mahalanobis distances over time for the three largest quartile groups.

2000–2004.

The detection rates in the smaller stocks (Quartiles 2 and 3) are consistently higher for the robust methods. In Figure 3.8b the RSD-IRMCD2(0.05) approach flags many more observations in the Quartile 3 group as outlying from the late 1990s through about 2005, and again during the 2008–2011 financial crisis period. In the Quartile 2 group (Figure 3.8a), there is a large spike in the RSD-IRMCD2(0.05) detection rate immediately following the dot-com era. Table 3.4 shows the actual numbers depicted in Figure 3.8a during late 1999 through late 2001. Again we see that the robust methods are flagging far more points as potential outliers. Even RSD-IRMCD2(0.05) tends to flag at least twice as many outliers as the classical approach. The robust approaches are much more in tune with the volatility and non-normality experienced during and after the dot-com crash. If one believes the classical method, the dynamics of the market were largely unchanged over the last 25 years.

Figure 3.9 shows pairwise scatterplots of the Quartile 4 data for June 2002, in which every pair of variables is plotted together with outliers marked as red stars. (For example, the upper left cell of the classical method result is a plot of the earnings-to-price exposures against the moving average exposures, with points flagged as outliers by the classical method marked with a red asterisk.) We can see some obvious one-dimensional outliers (e.g., in the earnings-to-price and moving average exposures) that are detected by all four methods. Only the robust method, though, is successful at picking out the multivariate outliers that do not fit with the majority of the data despite being only mildly extreme in any one dimension. The RSD-IRMCD2 method is also less susceptible to masking effects.

For another example, we consider the Quartile 2 group data at December 2000, a point in time at which the RSD-IRMCD2(0.05) detected a large number of outliers (see Figure 3.8a) but the classical method did not. Figure 3.10 shows two pairwise scatterplots at this time for the Quartile 2 group data. The classical method only detects a handful of extreme, one-dimensional outliers. The RSD-IRMCD2(0.05) method detects far more outliers, but the nature of the outliers is not evident in the two-dimensional scatterplots. Figure 3.11 presents three-dimensional scatterplots that shed some light on the outliers: here the outliers detected by both the classical and robust methods are shown as green triangles, those detected only by RSD-IRMCD2(0.05) are shown as red asterisks, and non-outlying points are shown as blue dots. Most of the outliers detected are characterized by a negative earnings-to-price ratio and a relatively smaller size. The classical method is only flagging the most extreme of the observations and is missing the moderate outliers (which are masked by the extreme ones).

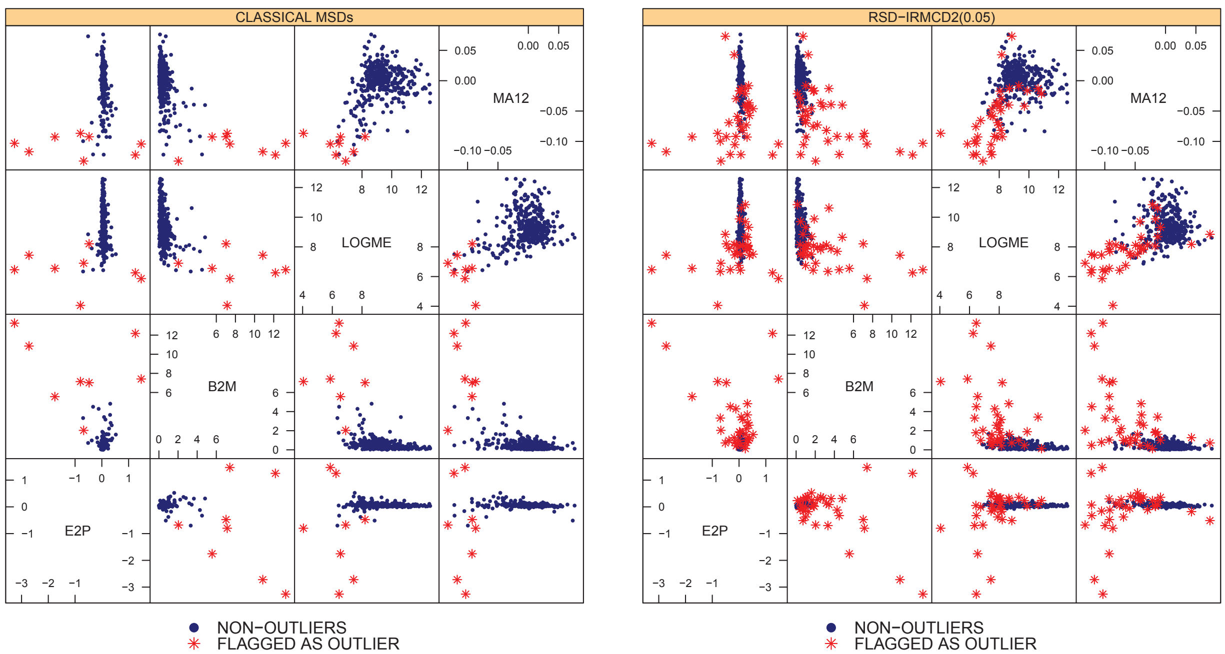


Figure 3.9: Pairwise scatterplots of Quartile 4 group data during June 2002. Outliers are shown as red asterisks, while regular data values are shown as blue dots. Left panel: classical estimate. Right panel: RSD-IRMCD2(0.05) estimate. Factors shown are earnings-to-price (E2P), book-to-market (B2M), size (LOGME), and momentum (MA12).

Table 3.2: Maximum number of outliers detected (MAX) and corresponding dates for each capitalization group and detection method.

Classical		RSD-IRMCD2(0.50)		RSD-IRMCD2(0.25)		RSD-IRMCD2(0.05)	
			Quartile 2				
MAX	Date(s)	MAX	Date(s)	MAX	Date(s)	MAX	Date(s)
13	Apr 1997 May 2000 Oct 2002 Mar 2003 Apr 2003	77	Oct 1999	65	Jun 2002	47	Dec 2000 Feb 2001
			Quartile 3				
MAX	Date(s)	MAX	Date(s)	MAX	Date(s)	MAX	Date(s)
10	Dec 1997 Feb 1998 Jan 2003 May 2004	58	Mar 2003	49	Jan 2003 Feb 2003	35	Jan 2003 Feb 2003
			Quartile 4				
MAX	Date(s)	MAX	Date(s)	MAX	Date(s)	MAX	Date(s)
10	Oct 2001	61	Feb 1999	46	Jun 2002 Feb 2009	37	Jun 2002

Table 3.3: Number of outliers detected in untransformed Quartile 4 data set during 2000–2004 using classical method (C), RSD-IRMCD2(0.50) method, RSD-IRMCD2(0.25) method, and RSD-IRMCD2(0.05) method.

		RSD-IRMCD2					RSD-IRMCD2		
Date	C	0.50	0.25	0.05	Date	C	0.50	0.25	0.05
Jul 2000	5	39	31	17	Jul 2002	4	25	20	20
Aug 2000	7	47	42	20	Aug 2002	4	27	21	18
Sep 2000	6	45	42	22	Sep 2002	6	29	22	21
Oct 2000	6	47	40	22	Oct 2002	6	33	29	20
Nov 2000	8	51	39	24	Nov 2002	7	31	26	20
Dec 2000	8	46	42	28	Dec 2002	5	36	28	26
Jan 2001	6	44	39	23	Jan 2003	6	38	32	26
Feb 2001	6	50	40	24	Feb 2003	6	39	36	30
Mar 2001	4	48	41	27	Mar 2003	7	37	30	25
Apr 2001	7	49	43	27	Apr 2003	6	35	32	25
May 2001	5	43	35	26	May 2003	6	34	30	24
Jun 2001	6	49	41	32	Jun 2003	4	33	30	24
Jul 2001	8	38	28	19	Jul 2003	4	24	21	14
Aug 2001	9	43	31	24	Aug 2003	5	30	23	18
Sep 2001	8	53	37	25	Sep 2003	4	26	28	21
Oct 2001	10	47	40	25	Oct 2003	4	27	24	18
Nov 2001	6	50	37	24	Nov 2003	4	23	21	15
Dec 2001	7	45	30	22	Dec 2003	4	20	19	13
Jan 2002	8	44	28	25	Jan 2004	5	20	20	14
Feb 2002	9	52	40	32	Feb 2004	5	15	15	11
Mar 2002	9	49	37	30	Mar 2004	5	19	18	13
Apr 2002	7	47	40	33	Apr 2004	4	22	17	13
May 2002	6	55	44	32	May 2004	7	26	22	13
Jun 2002	8	53	46	37	Jun 2004	5	24	18	13

Table 3.4: Number of outliers detected in untransformed Quartile 2 data set during 1999–2001 using classical method (C), RSD-IRMCD2(0.50) method, RSD-IRMCD2(0.25) method, and RSD-IRMCD2(0.05) method.

		RSD-IRMCD2					RSD-IRMCD2		
Date	C	0.50	0.25	0.05	Date	C	0.50	0.25	0.05
Jan 1999	8	54	42	30	Jul 2000	7	46	39	32
Feb 1999	3	54	43	28	Aug 2000	7	45	40	35
Mar 1999	12	54	47	35	Sep 2000	8	49	44	37
Apr 1999	10	51	41	29	Oct 2000	5	54	52	39
May 1999	8	54	46	31	Nov 2000	5	65	62	44
Jun 1999	8	53	47	38	Dec 2000	5	63	60	47
Jul 1999	9	67	62	25	Jan 2001	6	59	56	39
Aug 1999	6	68	60	24	Feb 2001	6	60	57	47
Sep 1999	7	71	58	26	Mar 2001	7	62	58	44
Oct 1999	6	77	58	31	Apr 2001	11	62	54	44
Nov 1999	9	70	55	31	May 2001	5	57	49	40
Dec 1999	10	66	58	33	Jun 2001	7	58	55	44
Jan 2000	7	58	55	37	Jul 2001	8	51	41	28
Feb 2000	12	62	53	37	Aug 2001	7	50	37	30
Mar 2000	9	64	53	29	Sep 2001	7	50	46	31
Apr 2000	9	70	62	44	Oct 2001	5	55	46	29
May 2000	13	67	57	44	Nov 2001	6	51	40	30
Jun 2000	11	66	57	38	Dec 2001	7	53	41	29

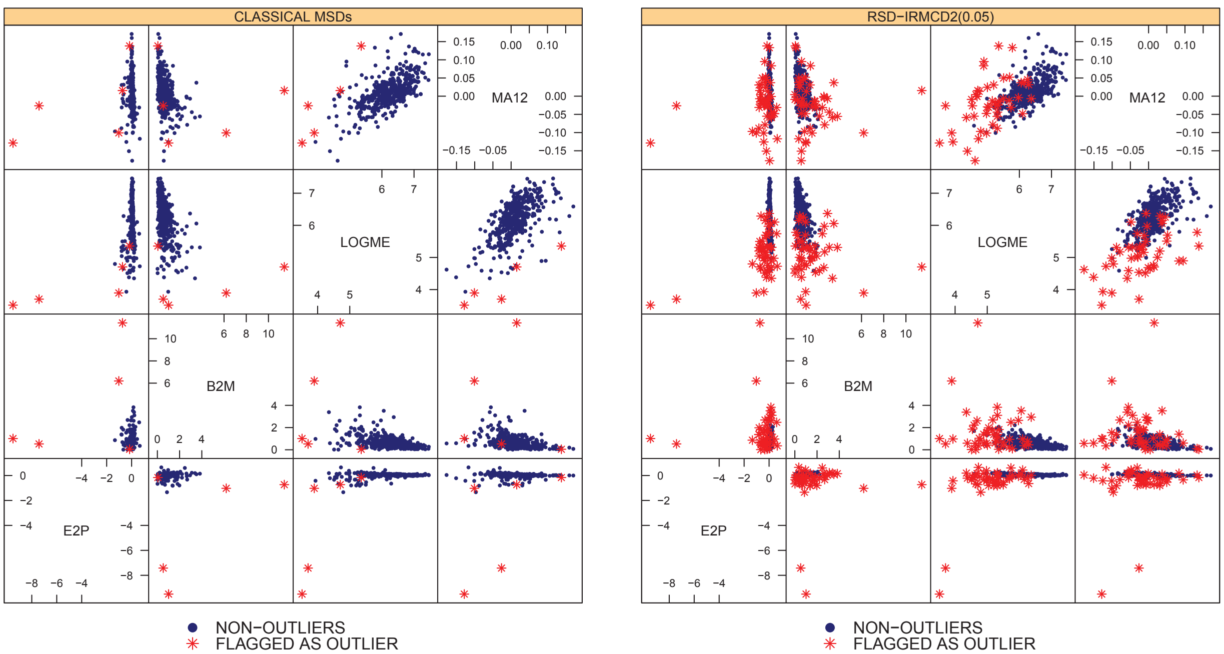


Figure 3.10: Pairwise scatterplots of Quartile 2 group data during December 2000. Outliers are shown as red asterisks, while regular data values are shown as blue dots. The plot setup is identical to that used in Figure 3.9. Left panel: classical estimate. Right panel: RSD-IRMCD2(0.05) estimate.

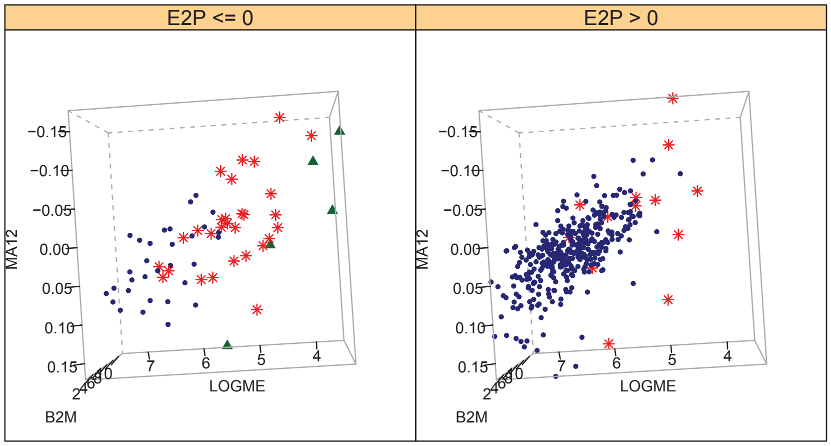


Figure 3.11: Three-dimensional scatterplots of Quartile 2 data during December 2000, stratified by the sign of the earnings-to-price (E2P) variable, with outliers detected by the classical and RSD-IRMCD2(0.05). The variables shown in the plot are BOOK2MARKET (B2M), LOGME, and MA12. The green triangles are points flagged as outliers by both methods, while the red asterisks are only flagged by the RSD-IRMCD2(0.05) method. (Blue dots are non-outlying data values as before.) The plot has been rotated to highlight the location of the outlying observations relative to the non-outlying observations.

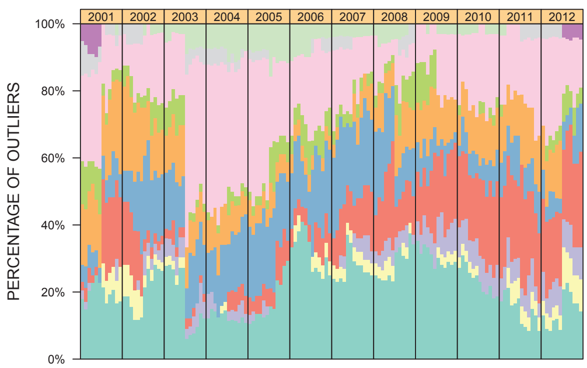
Next we can investigate the characteristics of the firms flagged as outliers. Figures 3.12a-3.12c show the distribution of the industry sectors (as defined by the GICS scheme) of firms flagged as outliers by the RSD-IRMCD2(0.05) method in the unaltered data.[13] In early 2001 technology firms were a large percentage of the stocks detected as outliers. Around the middle of 2001 we see the emergence of more financial and health care firms in the outliers. This trend persists until mid-2003, when the majority of the outliers reverts to technology firms. We also see a fairly steady percentage of outlying firms coming from the consumer discretionary sector (turquoise). This suggests that the market dynamics after the dotcom-crash partially explains the sharp increase in the number of outliers in the Quartile 2 data in 2001–2003.

Technology firms are also a large percentage of the outliers after the dot-com crisis for the

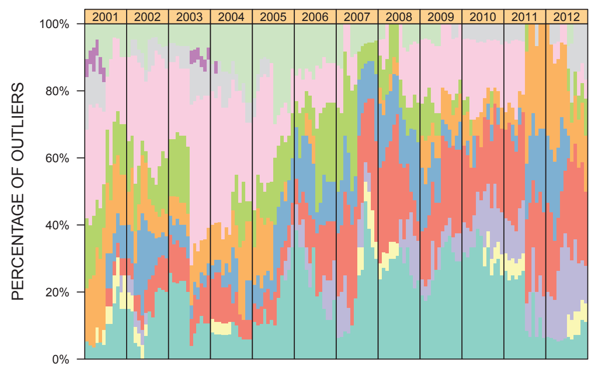
[13]Industry classifications were obtained from the Compustat Xpressfeed database, Bloomberg, and manual research. GICS classifications were not available consistently prior to 2001 from these sources.

larger firms in the Quartile 3 and Quartile 4 groups. We also see that consumer discretionary firms comprise a significant portion of the outliers in those groups during the 2005–2007 period prior to the financial crisis. Once the financial crisis starts to unfold in 2008, we see a shift towards financial firms as the main source of outliers, with financial firms being a big driver of the number of outliers in the largest stock group (Quartile 4).

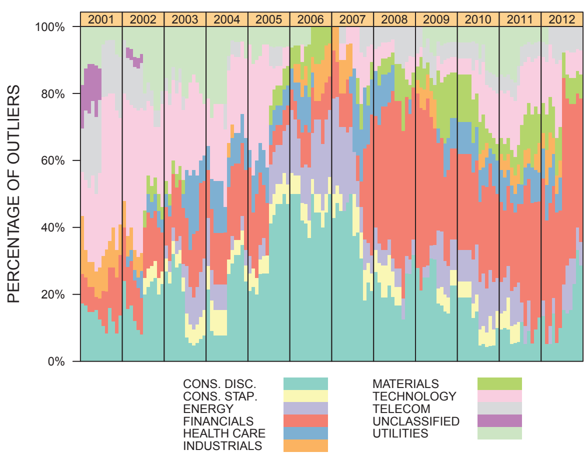
Finally, although we have focused on examples of factor exposure cross-sections containing a large number of multivariate outliers, it is certainly not the case that every cross-section will have a large number of outliers, or any outliers at all. The RSD-IRMCD2 approach is still more accurate than the classical distance approach even when there are relatively few multivariate outliers in the data set. Figure 3.13 shows an example of this phenomenon at July 2007: neither the classical method nor the RSD-IRMCD2 method flag many points as outliers due to the configuration of the observations in four-dimensional space. The robust approach, however, picks up a few moderate outliers that are missed by the classical approach due to masking by extreme, isolated observations. Hence, the RSD-IRMCD2 method is still preferred to the classical distance method even if one believes the data to be relatively outlier-free.



(a) Quartile 2



(b) Quartile 3



(c) Quartile 4

Figure 3.12: Industry distribution of outliers detected in the unaltered data by the RSD-IRMCD2(0.05) method. Top to bottom: (a) Quartile 2 data; (b) Quartile 3 data; (c) Quartile 4 data. Each "column" in the bar chart shows the percentages of the outliers, in each month, that come from each industry sector. The data for each year is grouped for ease of interpretation. The color scheme was produced using the `RColorBrewer` package (Neuwirth, 2014).

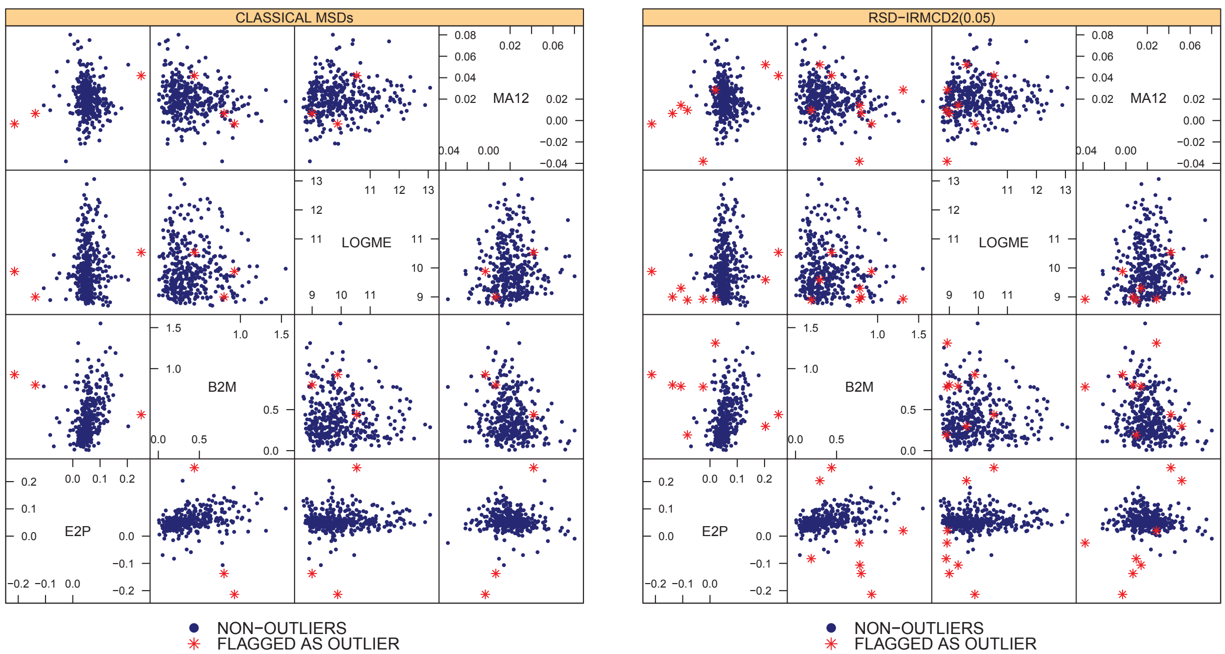


Figure 3.13: Pairwise scatterplots of Quartile 4 group data during July 2007 with outliers marked as in Figure 3.9. Results from using the classical Mahalanobis distances are shown in the left panel, while results from the RSD-IRMCD2(0.05)-based distances are shown on the right. The results for the RSD-IRMCD2(0.50) and RSD-IRMCD2(0.25) estimates are similar to the RSD-IRMCD2(0.05) result, and are omitted to conserve space.

3.4 *Outlier Detection in a Ten-Factor Fundamental Factor Model*

3.4.1 Description of the Data Sets

We use the U.S. Expected Returns (USER) data set discussed in Guerard et al. (2015). The data is provided courtesy of John Guerard. This data set consists of ten fundamental factors used to forecast expected returns for U.S. stocks.

- EP, the earnings-to-price ratio;
- BP, the book-to-price ratio;
- CP, the cash flow-to-price ratio;
- SP, the net sales-to-price ratio;
- REP, the current earnings-to-price ratio divided by the average earnings-to-price ratio over the past five years;
- RBP, the current book-to-price ratio divided by the average book-to-price ratio over the past five years;
- RCP, the current cash flow-to-price ratio divided by the average cash flow-to-price ratio over the past five years;
- RSP, the current net sales-to-price ratio divided by the average net sales-to-price ratio over the past five years;
- CTEF, the composite earnings forecasting variable developed in Guerard et al. (1997); and
- PM, a price momentum factor.

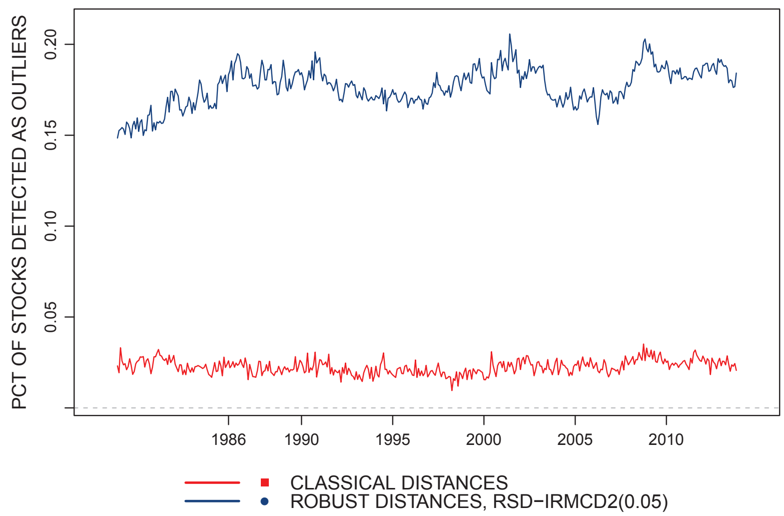


Figure 3.14: Percentage of stocks detected as outliers by classical and robust Mahalanobis distances over time.

Further details on these factors and the construction of the USER data set can be found in Guerard et al. (2015).

3.4.2 Outlier Detection Results

We use the same four techinques to detect multivariate outliers in the USER data as we did with the four-factor factor model data in the previous section. We again flag outlying assets in each month, which necessitates a Bonferroni adjustment to maintain our desired overall false positive rate of 2.5%. The USER data set contains 405 months of data, so we effectively use the upper $2.5\%/405 \approx 0.006\%$ percentile of the appropriate distribution in our outlier tests.

Figure 3.14 shows the percentage of observations in the ten-factor USER factor model data set that were flagged as outliers using the classical method (blue lines) and the RSD-IRMCD2(0.05) method (red lines). Once again we see that the robust method finds many more outliers in the data set than the classical method. Table 3.5 shows the maximum number of outliers detected by each method for each data set.

Table 3.5: Maximum number of outliers detected (MAX) and corresponding dates for each capitalization group and detection method.

Classical		RSD-IRMCD2(0.50)		RSD-IRMCD2(0.25)		RSD-IRMCD2(0.05)	
MAX	Date(s)	MAX	Date(s)	MAX	Date(s)	MAX	Date(s)
120	Oct 2008	1910	Aug 1998	1325	Dec 1998	776	Dec 1998 Oct 1999

Due to the number of stocks in each monthly cross-section of the USER data, it is difficult to visualize the geometry of the outlying and non-outlying points via pairwise scatterplots as we have done previously. Figure 3.15 shows the classical and RSD-IRMCD2(0.05) distances calculated for the USER data as of December 2008. RSD-IRMCD2(0.05) detects a very extreme outlier far away from the bulk of the data, and many extreme outliers that are quite far away from the center of the data. We have trimmed the range of the y-axis in the bottom pair of plots to show more of the outlying and non-outlying points. The RSD-IRMCD2(0.05) method is much better than the classical distances at identifying multivariate outliers even in large data sets.

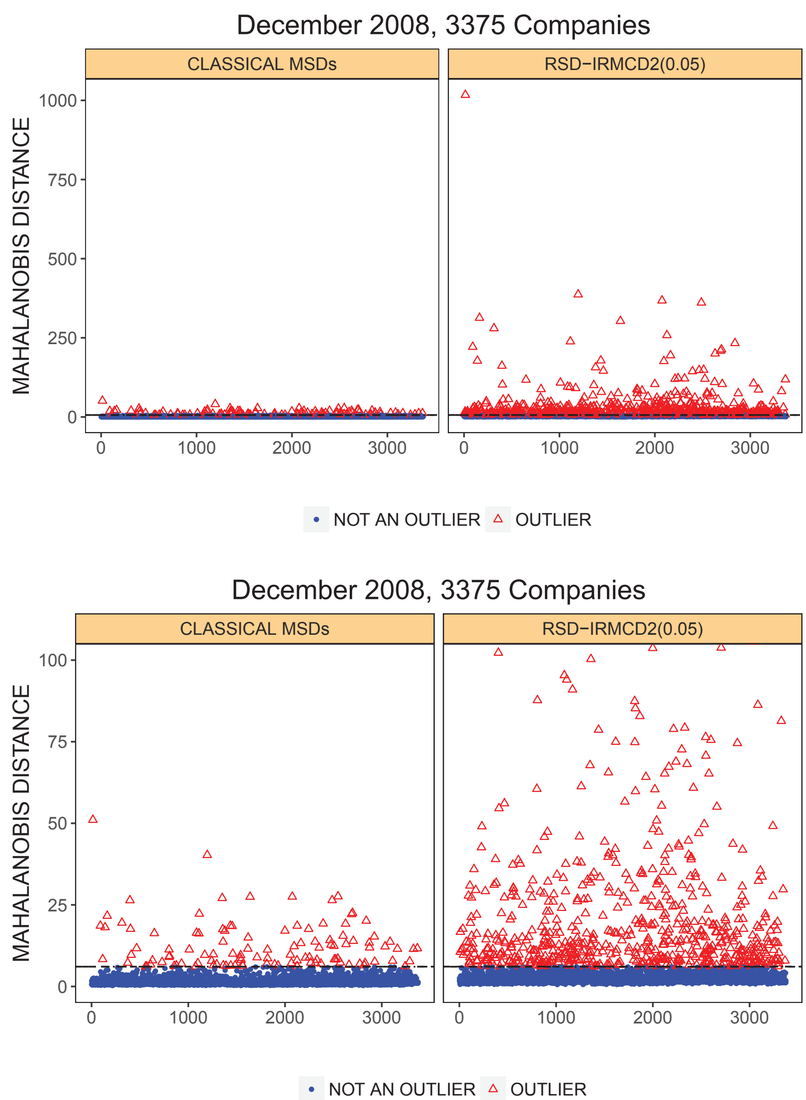


Figure 3.15: Mahalanobis distances calculated for the USER data as of December 2008. Outliers are marked with red triangles, while non-outlying points are marked with blue dots. The top pair of plots show the full range of the distances. In the bottom pair of plots only distances less than 100 are shown.

3.5 Conclusions and Further Research

Our analyses demonstrate that using MSDs based on the classical sample mean and sample covariance to detect outliers in asset returns and fundamental factor model data will often fail to detect moderate outliers. The classical estimates are not robust to outliers, and hence the resulting MSDs are also not robust to outliers. Extreme outliers in the data lead to masking of moderate outliers when the classical MSDs are used to detect outliers. Distances based on the MCD robust location and covariance estimate result in RSDs that are not much influenced by outliers. The asymptotic chi-squared distribution of RSDs in large samples does not provide accurate quantiles for testing RSDs based on the smaller sample sizes (e.g., $n \leq 250$) commonly encountered in finance, however. Our IRMCD2 methodology, which improves upon the IRMCD methodology introduced by Cerioli (2010), provides an accurate approximation to the finite-sample distribution of MCD-based RSDs for sample sizes $60 \leq n \leq 250$. Robust squared distances based on the MCD and tested using the IRMCD2 methodology provide a highly reliable means of identifying extreme and moderate outliers in portfolio returns and in fundamental factor model exposure data.

In practice, a portfolio manager can use our improved outlier detection methodology to identify multivariate outliers in asset returns and factor exposure data prior to portfolio construction or model fitting. She can then decide the best way to handle the outliers—perhaps deleting or shrinking a small number of outliers but resorting to further robust statistical methods, e.g., regression, when there is evidence of a large number of outliers. The portfolio manager can also construct RSDs in a rolling fashion to identify new observations that are inconsistent with historical data. The portfolio manager first estimates a robust mean and covariance for the data up to the present, then uses this mean and covariance to compute an RSD for the new observation as well as conduct the IRMCD2 tests of that distance.

As we stated earlier, this chapter re-examines some of the work done in Martin et al. (2010) with the more accurate IRMCD2 methodology. While the results of this chapter qualitatively agree with those of the earlier paper, our results here reinforce that one cannot rely upon classical MSDs, as advocated by Chow et al. (1999), Kritzman and Li (2010), and others, to detect all the multivariate outliers in portfolio returns or factor model data. Furthermore, our study validates earlier work by Cerioli et al. (2009) that chi-squared quantiles are not reliable for testing MCD-based RSDs on the smaller sample sizes common in finanical data. The IRMCD2 methodology used herein provides accurate multivariate outlier detection for our chosen applications to portfolio returns and factor model data.

Implicit in the outlier tests used here is the assumption that the bulk of the data follow an approximate normal distribution. For monthly data this may be a reasonable assumption, but at higher frequencies this is likely not true. Higher frequency financial data tends to have heavier tails even in the absence of outliers. The tests may therefore flag a large number of observations as outlying even though they would not be atypical when the heavy-tailed nature of the data is taken into account. The development of an accurate approximation to the finite-sample distribution of RSDs for data from a multivariate elliptical distribution would be helpful for working with higher frequency data.

Finally, we note that there are numerous other methods of detecting anomalous observations, such as the approach of Willems et al. (2009) (which was also based on the MCD) and the so-called “grand-tour” approach (Buja and Asimov (1986); Cook et al. (1995)), which involves looking at all lower-dimensional (say two-dimensional or three-dimensional) slices of a multidimensional data set. These methods can also be very helpful in understanding the structure of the outliers in multidimensional data set. A rigorous comparison of a more varied set of approaches to outlier detection in the context of portfolio and factor model construction would be an ambitious undertaking for the interested researcher.

APPENDIX

3.A The Inadequacy of One-Dimensional Trimming and Winsorization

In this section, we illustrate how one-dimensional trimming and Winsorization can alter the nature of the data set and thereby affect the results of outlier detection. Trimming or Winsorizing one variable at a time is standard practice for dealing with multivariate financial data, but as we saw in Figure 3.1 some multivariate outliers will not be affected by one-dimensional trimming or Winsorization since they are only outlying in higher-dimensional views of the data. Moreover, one-dimensional Winsorization can introduce artificial structure to the data: an observation that is outlying in k variables gets mapped to a hyperplane of dimension $\nu - k$, e.g., an edge ($k = 1$) or a corner ($k = 2$) of the box ($\nu = 2$) created by the dashed lines in Figure 3.1. The resulting Winsorized data set possesses structure that was not there prior to Winsorization, structure that can potentially have an adverse effect on estimation methods. For example, the presence of a large subset of observations that lie in the same hyperplane can make an estimated covariance matrix singular,[14] invalidating the asymptotic distribution theory one typically uses for computing confidence intervals and test statistics.

We examine the potential impacts of one-dimensional trimming and Winsorization using our four-factor fundamental factor model data set from Secton 3.3. We will use the same classical and RSD-IRMCD2 detection methods used in that section to flag outliers in the data after trimming each variable and after Winsorizing each variable. We will then analyze the detected outliers to determine if any of them were outlying before trimming and/or Winsorizing was applied.

[14]Section 6.2.2 of Maronna et al. (2006) discusses the relationship between the size of the subset lying in a hyperplane and the invertibility of the covariance estimate.

3.A.1 Data Setup

To investigate the effects of univariate trimming and Winsorization, we need versions of the four-factor fundamental factor model data set in which the factor exposures at each time point were trimmed or Winsorized by $\delta\%$. Suppose there are n assets at a given point in time. For each factor exposure, we sort the observations from smallest to largest. Then we examine the $n\delta$ largest and $n\delta$ smallest observations.

- To trim by $\delta\%$, we simply remove the corresponding assets from this time period.
- To Winsorize by $\delta\%$, we replace the $n\delta$ largest values of the factor exposure by the $n\delta$ upper quantile of the data, and the $n\delta$ smallest values by the $n\delta$ lower quantile of the data.

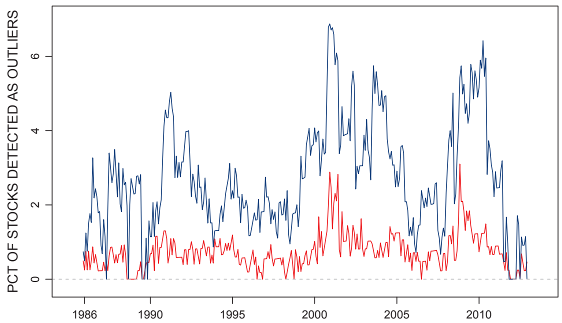
Trimming reduces the influence of the largest and smallest values on an estimate to that of the $n\delta$ upper and lower quantiles. Winsorization was developed with the same goal in mind, but it introduces discontinuities into the influence function (e.g., for the Winsorized mean) at the cutoff quantiles (Hampel, 1974).[15] For the present study we trimmed and Winsorized by $\delta = 2.5\%$. This represents a compromise between removing very large outliers in the data and preserving the heavy tails and skewness that often characterize financial data.

In our discussions below we will refer to the data set used in Section 3.3 without any trimming or Winsorizing as the "unaltered" or "original" data set.

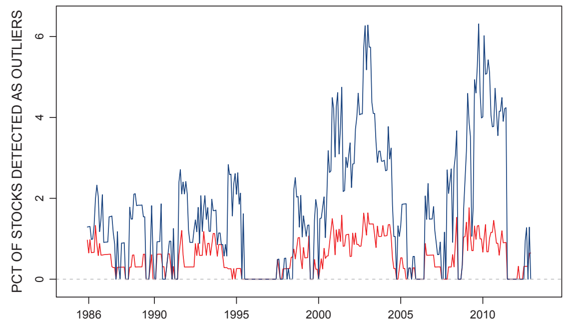
3.A.2 Outlier Detection Results—Trimmed Data

Figures 3.16a–3.16c show the results of the outlier detection procedures on the data after trimming each factor separately. After trimming, the classical and robust approaches yield similar results more often, but can still give very different results during periods of market stress. Table 3.6 shows the maximum number of outliers detected by the classical and robust

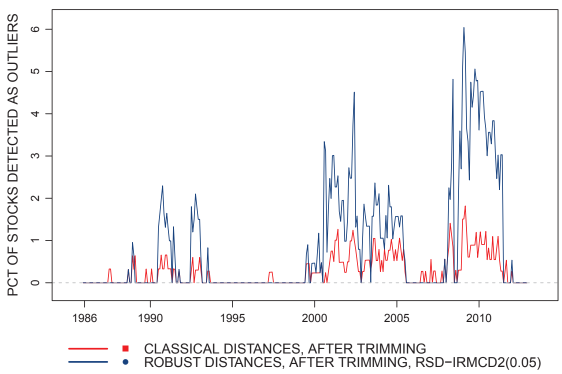
[15] Martin et al. (2010) provides a plot of the influence function that shows the discontinuities.



(a) Quartile 2



(b) Quartile 3



(c) Quartile 4

Figure 3.16: Percentage of stocks detected as outliers by classical and robust Mahalanobis distances over time after one-dimensional trimming.

distance methods. Once again, the period after the dot-com era and the period of the recent financial crisis show elevated numbers of outliers. Overall, though, the number of outliers detected by each method is less than in the corresponding unaltered data case. Note, though, that since trimming removes observations, the data set has changed. The observations flagged as outliers now may not be the same ones flagged before. We will investigate this question in Section 3.A.4.

December 2000 is again an interesting time for Quartile 2 group data. Figure 3.17 shows pairwise scatterplots for that group at this time. A cursory comparison of these scatterplots with those from the analysis on the data without any trimming (Figure 3.10), in particular the plots of LOGME against MA12, suggests that some multivariate outliers may not have been touched by the trimming procedure. This is, in fact, the case. Figure 3.18 (top panel) shows the pairwise scatterplot for the unaltered data with RSD-IRMCD2(0.05) outliers marked. The green triangles correspond to outlying observations that would have been removed by one-dimensional trimming, while the red asterisks are outlying observations (from the unaltered data) that would remain in the trimmed sample. There are clearly many observations that are flagged as outliers using the robust methods that remain outliers after one-dimensional trimming. Some of these are one-dimensional outliers that were masked by the more extreme one-dimensional outliers, and some are truly multivariate in nature. The bottom panels split the data on the sign of the EARN2PRICE factor (as we did in Figure 3.11) to demonstrate that many of the outliers belong to the same cluster of observations with a negative earnings-to-price ratio and a relatively smaller size.

Table 3.6: Maximum number of outliers detected (MAX) and corresponding dates for each capitalization group and detection method after one-dimensional trimming.

Classical		RSD-IRMCD2(0.50)		RSD-IRMCD2(0.25)		RSD-IRMCD2(0.05)	
			Quartile 2				
MAX	Date(s)	MAX	Date(s)	MAX	Date(s)	MAX	Date(s)
13	Dec 2000	56	Jun 2002	51	Jun 2002	31	Nov 2000 Dec 2000
			Quartile 3				
MAX	Date(s)	MAX	Date(s)	MAX	Date(s)	MAX	Date(s)
6	Oct 2002 Jan 2003	48	Aug 2002	35	Aug 2002 Nov 2002	23	Nov 2002 Jan 2003
			Quartile 4				
MAX	Date(s)	MAX	Date(s)	MAX	Date(s)	MAX	Date(s)
6	Mar 2009	40	May 2002	31	Jun 2008	20	Feb 2009

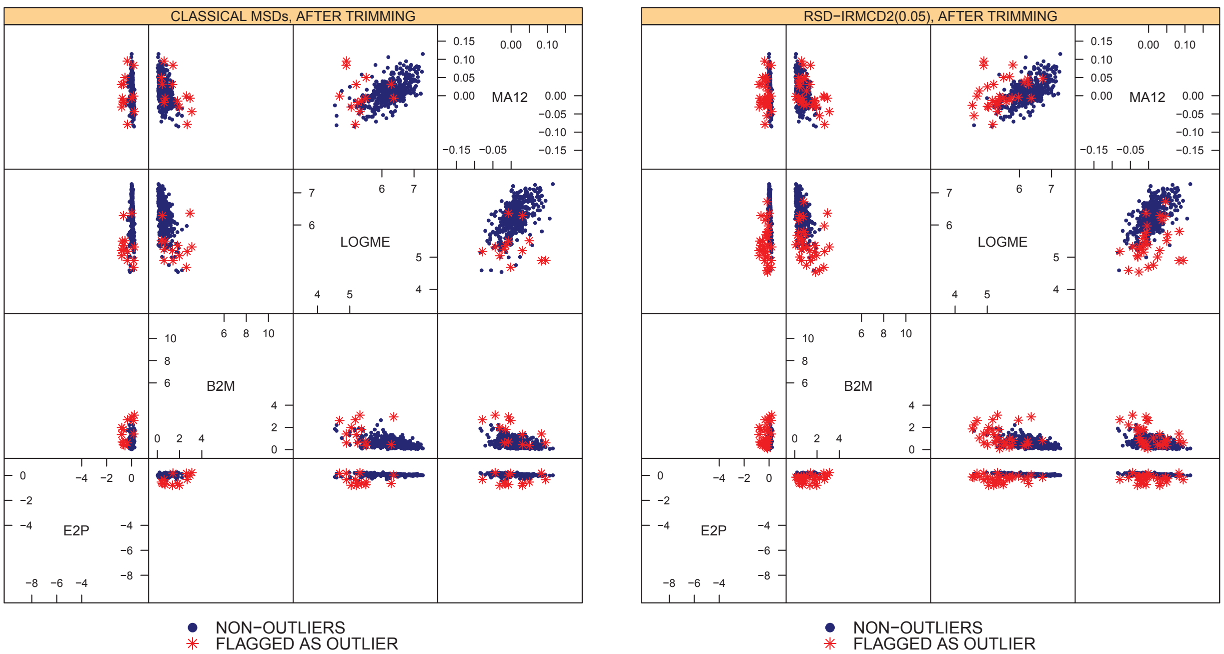


Figure 3.17: Pairwise scatterplots of Quartile 2 group data during December 2000 after univariate trimming. Outliers are shown as red asterisks, while regular data values are shown as blue dots. The classical estimate is shown on the left, while the RSD-IRMCD2(0.05) estimate is shown on the right. The results for the RSD-IRMCD2(0.50) estimate and RSD-IRMCD2(0.25) estimate are similar to that of the RSD-IRMCD2(0.05) estimate, and are omitted to save space.

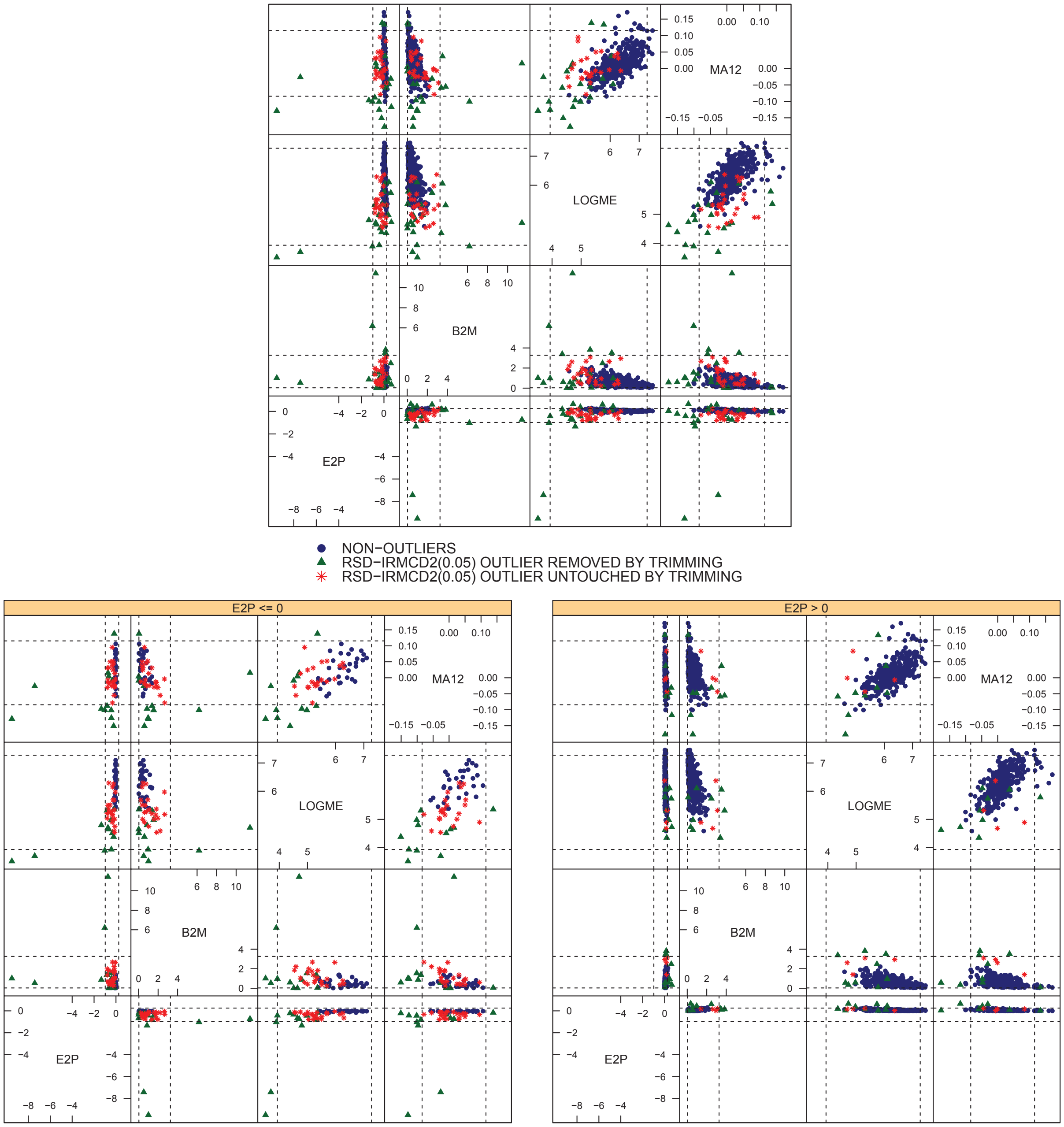
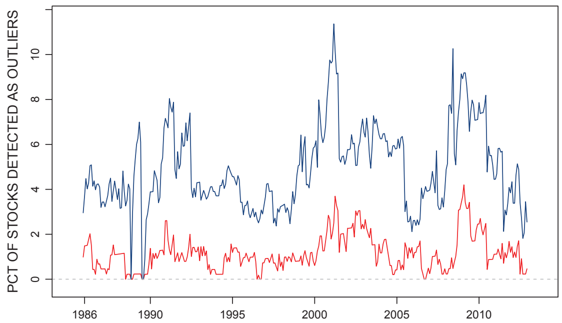
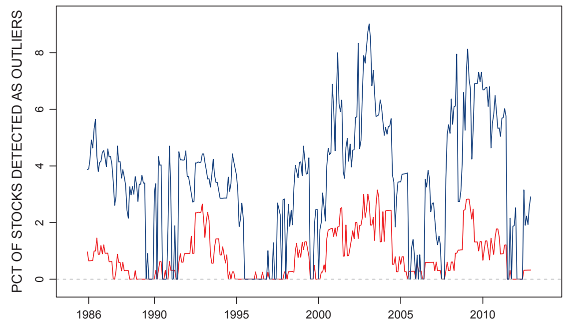


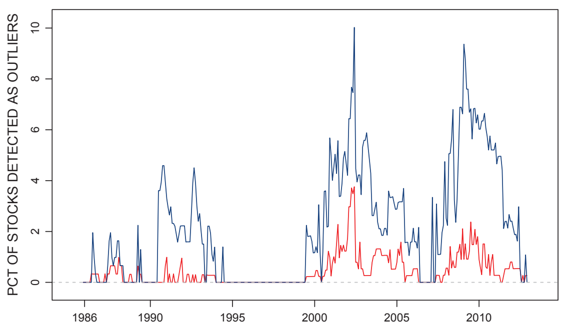
Figure 3.18: (Top) Pairwise scatterplot of the unaltered Quartile 2 group data during December 2000 with outliers detected by the RSD-IRMCD2(0.05) method. Points marked with green triangles are outliers that would have been deleted by the trimming procedure. Red asterisks are points that are still flagged as outliers by the RSD-IRMCD2(0.05) after trimming. (Blue dots are non-outlying points.) The black lines represent 2.5% trimming bounds for each variable. (Bottom) The same scatterplot, with observations split based on the sign of the EARN2PRICE factor. Observations with negative earnings yield are shown in the left panel, and observations with positive earnings yield are shown in the right panel.



(a) Quartile 2



(b) Quartile 3



(c) Quartile 4

Figure 3.19: Percentage of stocks detected as outliers by classical and robust Mahalanobis distances over time after one-dimensional Winsorization.

3.A.3 Outlier Detection Results—Winsorized Data

Figures 3.19a–3.19c show the results of the outlier detection procedures on the Winsorized data. Table 3.7 shows the maximum number of outliers detected by each method. Winsorization did not produce as significant a reduction in the number of outliers detected, relative to the unaltered data, as trimming did. At times, the number of outliers in the Winsorized data is nearly as high as in the unaltered data.

As we indicated earlier, one-dimensional Winsorization changes the structure of the data by creating clusters of observations on the Winsorization boundaries, i.e., the edges of the boxes defined by the 2.5% and 97.5% percentiles of each variable. Figure 3.20 shows pairwise scatterplots of the factor exposure cross-section data for April 2001 with outliers detected

Table 3.7: Maximum number of outliers detected (MAX) and corresponding dates for each capitalization group and detection method after one-dimensional Winsorization.

Classical		RSD-IRMCD2(0.50)		RSD-IRMCD2(0.25)		RSD-IRMCD2(0.05)	
			Quartile 2				
MAX	Date(s)	MAX	Date(s)	MAX	Date(s)	MAX	Date(s)
16	Apr 2001 Feb 2009	73	Jun 2002	66	Jun 2002	50	Mar 2001
			Quartile 3				
MAX	Date(s)	MAX	Date(s)	MAX	Date(s)	MAX	Date(s)
12	Aug 2003	58	Feb 2003	48	Oct 2002 Jan 2003 Feb 2003 Mar 2003	33	Feb 2003
			Quartile 4				
MAX	Date(s)	MAX	Date(s)	MAX	Date(s)	MAX	Date(s)
15	Apr 2002 Jun 2002	61	May 2002	47	Jun 2002	40	Jun 2002

using classical distances in each panel. The clustering of previously outlying observations along the Winsorization boundaries is quite clear. Many of the outliers detected in the original data set are also outlying after Winsorization because they still violate the underlying assumption of multivariate normality. Prior to Winsorization, these assets were outliers due to their position: they were farther away from the center of the cross-section data than one would expect for multivariate normal data. After Winsorization, they are outliers for a different reason: under multivariate normality observations it is unlikely that so many observations would fall into the same hyperplane.

The situation worsens if we switch to the RSD-IRMCD2 detection method, as even more

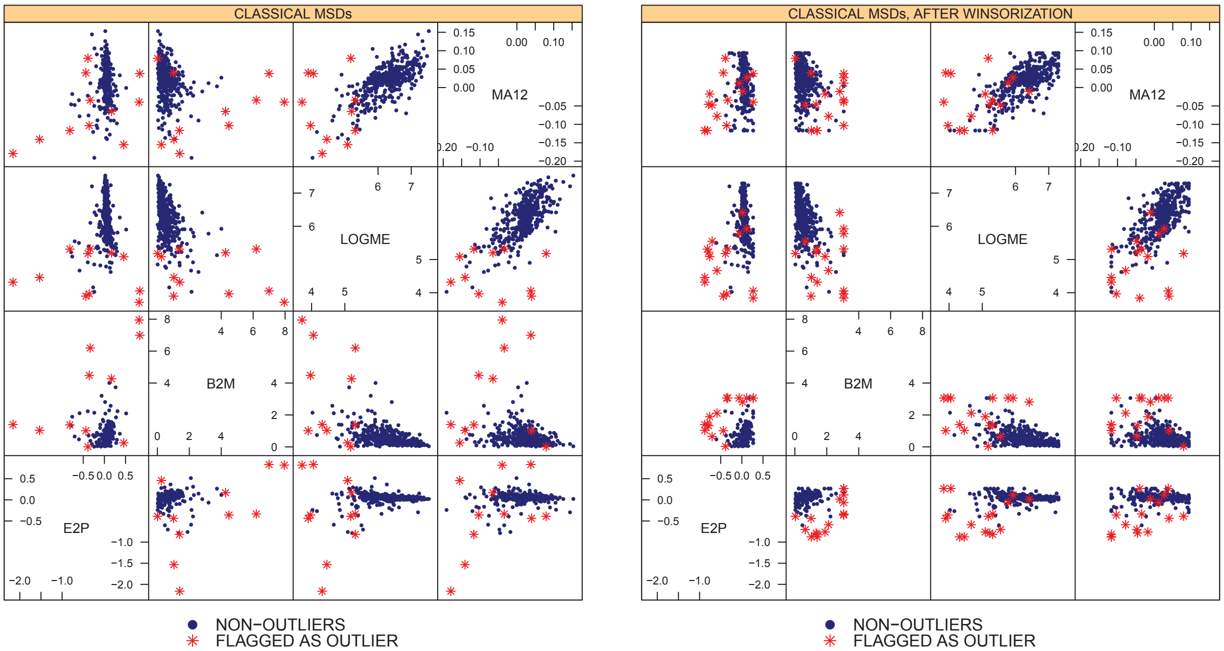


Figure 3.20: Pairwise scatterplots of Quartile 2 group data during April 2001 before and after univariate Winsorization. Outliers are shown as red asterisks, while regular data values are shown as blue dots. Left: classical method on data prior to Winsorization. Right: classical method on Winsorized data.

assets are flagged as outlying. Figure 3.21 shows pairwise scatterplots of the Quartile 4 group data for June 2002 after Winsorization for the classical and RSD-IRMCD2(0.05) outlier tests. Figures 3.22 shows similar scatterplots for the Quartile 2 group data for December 2000. We again see that Winsorizaton creates clusters of observations along the Winsorization boundaries, which would be unlikely to occur if the observations were truly multivariate normal distributed. Hence many of the points on the boundary are flagged as outliers.

3.A.4 One-Dimensional Trimming and Winsorization are Not Adequate for Dealing with All Multivariate Outliers

The results in the trimmed and Winsorized data cases show that multivariate outliers will often still exist after applying these one-dimensional methods. As we noted above, however,

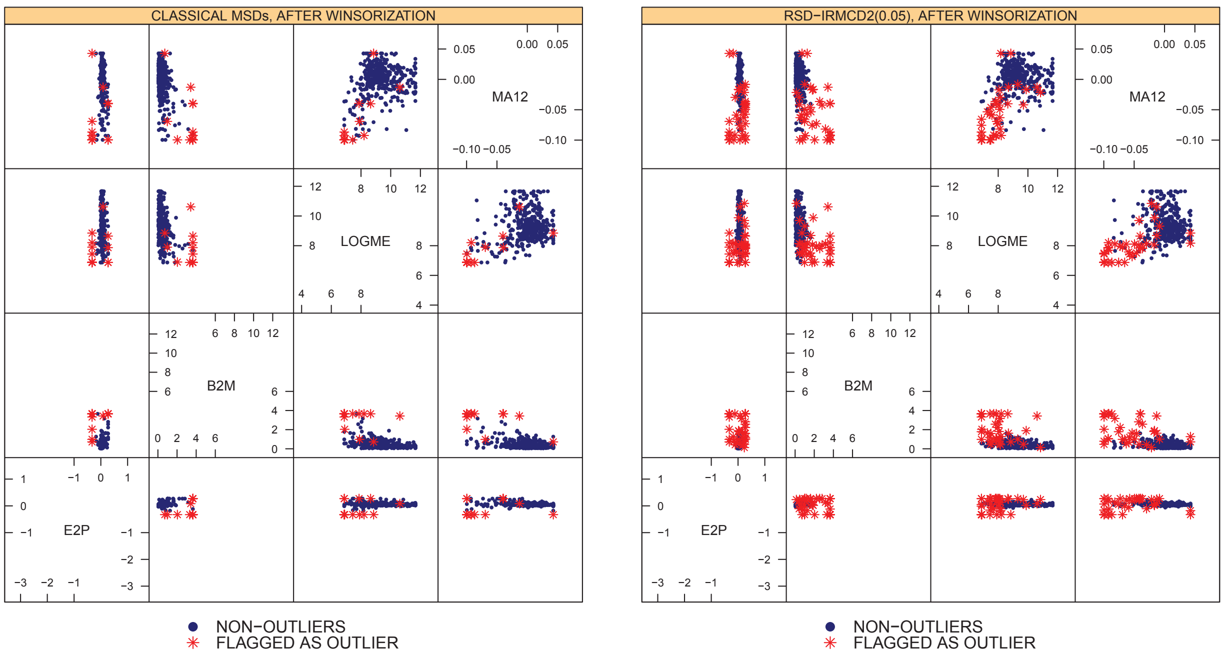


Figure 3.21: Pairwise scatterplots of Quartile 4 group data during June 2002 after univariate Winsorization. Outliers are shown as red asterisks, while regular data values are shown as blue dots. Left panel: classical estimate. Right panel: RSD-IRMCD2(0.05) estimate.

the use of either of these methods changes the structure of the data, so the results above do not tell us much about the persistence of outliers after mitigation. The outliers found in the trimmed data set, for instance, may not have been outlying in the untrimmed data set. To investigate this question, we examined each observation's status as an outlier in the original unaltered data set, the trimmed data set, and the Winsorized data set with each of the detection methods. For a given detection method, we can group the observations according to the data sets in which they were found to be outlying, and then count the number of observations in each of the seven combinations of "unaltered", "trimmed", and "winsorized".[16]

Figure 3.23 depicts the results of this calculation for the Quartile 2 data over 2000–2001, one of the periods found earlier with a high number of outliers. The "ALL" row of Figure

[16]We exclude the eighth case, observations that were not deemed outlying in any of the data sets, as it is not relevant here.

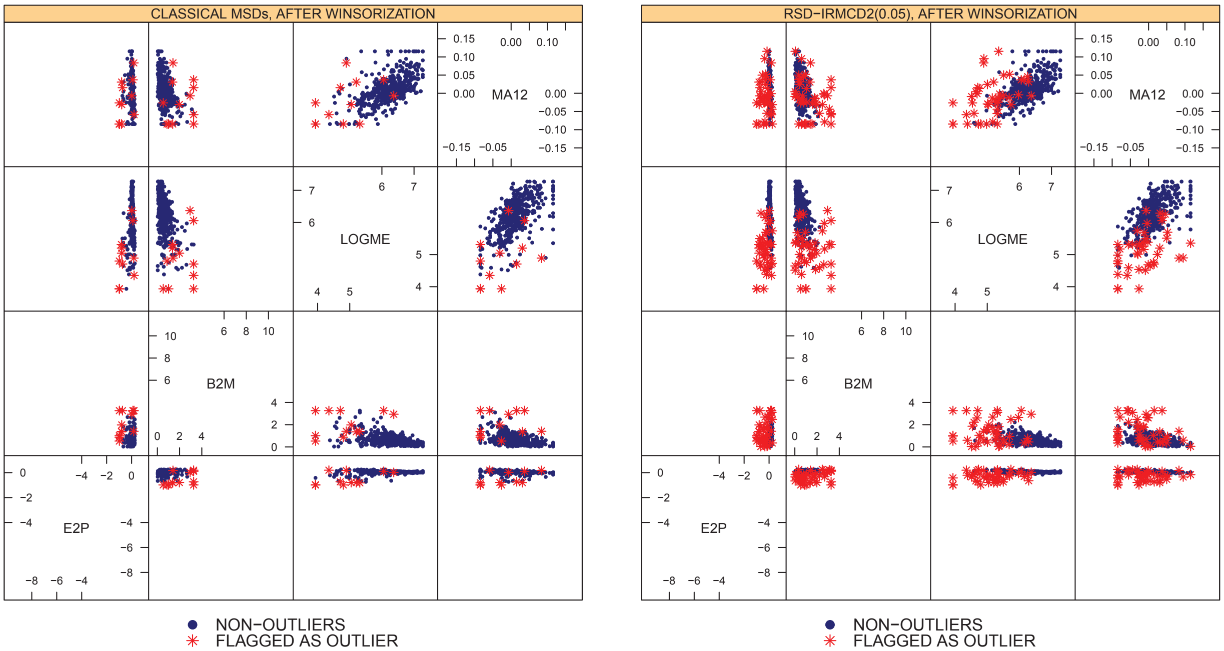


Figure 3.22: Pairwise scatterplots of Quartile 2 group data during December 2000 after univariate Winsorization. The plot setup is identical to that used in Figure 3.21 Left panel: classical estimate. Right panel: RSD-IRMCD2(0.05) estimate.

3.23 represents observations that are outlying in the original data set and are still outlying after trimming and Winsorization. This means that neither trimming nor Winsorizing each variable by 2.5% eliminated those observations, as even larger observations dictated the trimming/Winsorization boundaries. (The example shown in Figure 3.18 illustrates how this happens.) For instance, in December 2000 the "ALL" row of the RSD-IRMCD2(0.05) panel indicates that there were 26 outliers in the unaltered data set that were still flagged as outliers by the RSD-IRMCD2(0.05) method both after one-dimensional trimming and after one-dimensional Winsorization. In the case of Winsorization, moderate outliers were either unaffected or were projected to the "box" created by Winsorization, after which they were still judged as a poor fit for the multivariate normal model by the detection methodology. Many outliers detected by the RSD-IRMCD2(0.05) method in the unaltered data persist after Winsorization, as evidenced by the large numbers of outliers reported in the "ALL" and “UNALT. AND WINS.” rows. The classical method, on the other hand, puts very few observations in the “ALL” category, and fewer in the “UNALT. AND WINS.” category than the RSD-IRMCD2(0.05) method.

Figure 3.24 illustrates how this can happen via pairwise scatterplots of the Quartile 2 data group at December 2000 with outliers flagged using the RSD-IRMCD2(0.05) method and the observations coded by the same categories used in Figure 3.23. The pink/blue inverted triangles in the figure are observations in the “ALL” category. With the classical method, moderate outliers are masked by extreme outliers (such as those indicated by the blue Xs). After the extreme outliers are trimmed, the more moderate outliers may be detected by the classical method, depending on their location. On the other hand, the robust approach flags these outliers in the unaltered data, and continues to flag them after trimming and Winsorization. As we discussed earlier (see Figure 3.11), there is a cluster of multivariate outliers (mostly with negative earnings-to-price ratio) in the data at this time. The classical approach only catches the more extreme outliers in this group due to the masking effect, while the robust approach flags the entire cluster.

The non-blank cells in the “TRIM ONLY” rows of each panel indicate that there are multivariate outliers (as determined by either of the classical and robust squared distances) in the data after trimming that were not found prior to trimming. These are shown in Figure 3.24 as yellow stars. One-dimensional trimming “peels away” the extreme outliers in the data, revealing more moderate outliers. It also changes the structure of the data, however, so observations that would not have been considered outlying prior trimming might now be outlying in the modified data set. These are the observations that fall into the “TRIM ONLY” category.

The “UNALT. AND TRIM” row is blank for both distance methods for this particular data set and time range. An extreme outlier beyond the trimming boundaries in the original data will be removed by trimming, and hence would not show up in the trimmed data set. A moderate outlier inside the trimming boundaries, however, will be untouched by trimming. In the trimmed data set it can be outlying or non-outlying. In the latter case, it will not

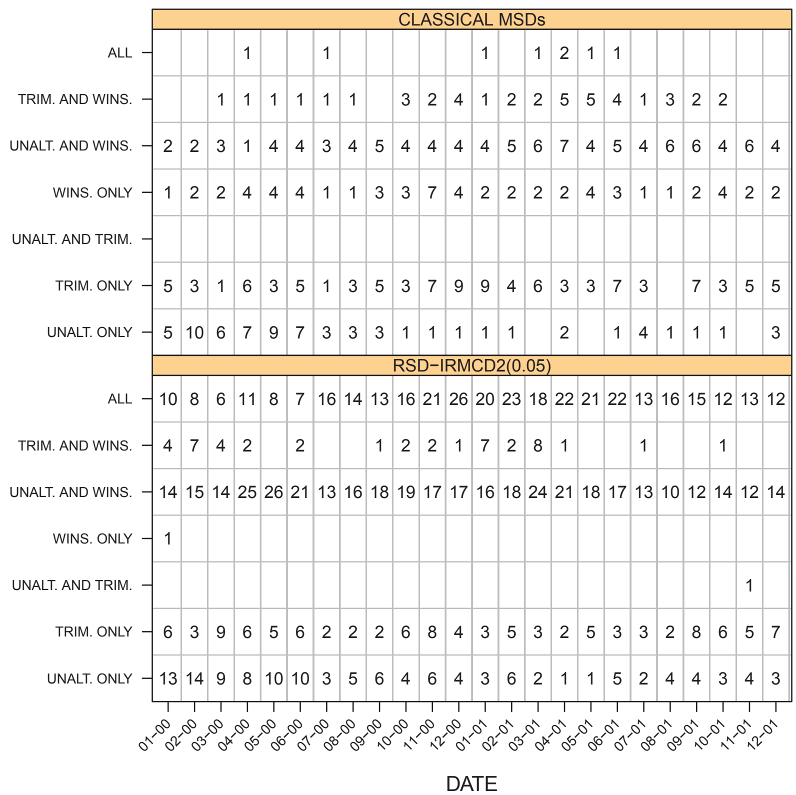


Figure 3.23: Status of outliers detected in the Quartile 2 group data, 2000–2001, using the classical and robust RSD-IRMCD2(0.05) methods. From bottom to top in each panel, the cells show the number of observations in each month's data that were identified as outliers in the unaltered data only (UNALT. ONLY); the trimmed data only (TRIM. ONLY); the unaltered and the trimmed data sets (UNALT. AND TRIM); the Winsorized data only (WINS. ONLY); the unaltered and Winsorized (UNALT. AND WINS) data sets; the trimmed data and the Winsorized data (TRIM. AND WINS.); and in all of the data sets (ALL).

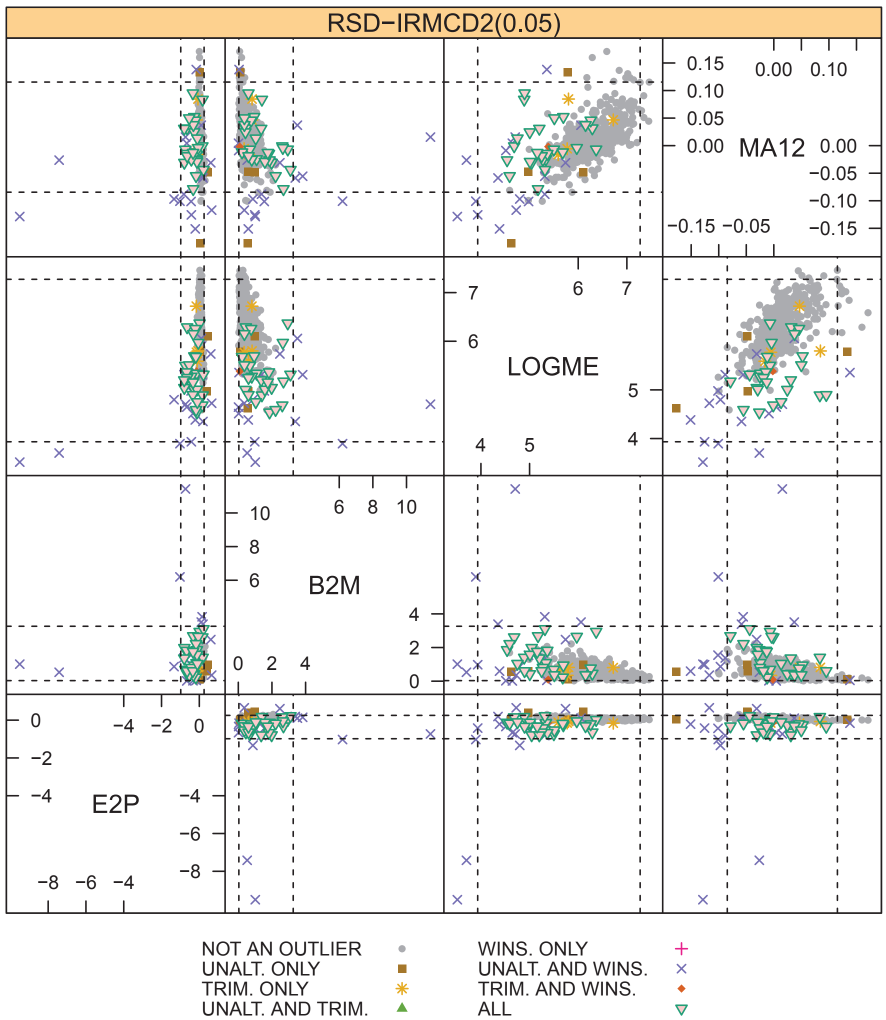


Figure 3.24: Status of outliers detected in the Quartile 2 group data at December 2000 using the RSD-IRMCD2(0.05) methods. The observations are coded by types used in Figure 3.23. The dashed lines depict the 2.5% trimming boundaries for each variable. The color scheme was produced using the `RColorBrewer` package (Neuwirth, 2014).

fall into the "UNALT. AND TRIM" category. In the former case, the observation will often also be outlying in the Winsorized version of the data, and will hence show up in the "ALL" category.

3.A.5 Summary

The results of this section provide strong evidence that one-dimensional trimming and Winsorization are insufficient to deal with many of the outliers found in the type of asset returns and fundamental factor model data used to build and monitor portfolios. While the one-dimensional approaches may be an adequate means of dealing with outliers in individual variables, they fail to detect multivariate outliers. Such multivariate outliers can be difficult to spot in high-dimensional data, where standard methods for visualizing the data provide little insight. Algorithmic outlier detection methods, such as the RSD-IRMCD2 method presented here, are a more effective means of finding outlying observations in multivariate financial data.

3.B Construction of the Four-Factor Asset Pricing Model Data Set

We aquired historical monthly market capitalization (ME) data from CRSP (CRSP, 2015b) on U.S. firms listed on one of the three major exchanges (NYSE, AMEX, and NASDAQ) at any time between December 1, 1985 and December 31, 2012.[17] We removed any assets that were not common stocks (such as ADRs and closed-end funds).[18]

The historical ME data for each selected firm was then augmented by its historical accounting data taken from the Compustat annual file (Compustat, 2015). We joined CRSP data to Compustat data using the linkages provided in the CRSP/Compustat Merged database (CRSP, 2015a). Since firms do not complete their audited financials instanteously (and

[17]For firms (identified by the variable PERMCO in the CRSP database) with more than one security (identified by a PERMNO) trading at a given date, we aggregated market capitalization data over all such securities for that date and assigned this aggregate market capitalization to the PERMNO with the largest market capitalization. See Palacios and Vora (2011) for more details.

[18]In the CRSP database we excluded securities with a share code (SHRCD) other than 10 or 11.

historically may not have been very timely in completing them), we must lag the accounting data when merging it with the market data. We follow the approach of Asness and Frazzini (2013) for combining the accounting data and the market data: the accounting data for a firm having a fiscal year-end in calendar year $t-1$ is "fixed" for that year on December 31, and assumed known to the market six months thereafter on June 30 of year t. The firm's accounting data is held constant over the next twelve months (July 1 of year t through June 30 of year $t+1$). Ratios involving accounting and market data use these "fixed" accounting values, but *current* market data.[19] Asness and Frazzini showed that using the current market data to compute a firm's book value-to-price ratio (equivalently, its book-to-market ratio), rather than lagging the market data as well per Fama and French (1992), yielded a better forecast of a firm's book value-to-price ratio at its next fiscal year-end (which is usually not observable in June). We adopt their approach here rather than the Fama and French approach as the former is more in line with current practices in empirical asset pricing and portfolio construction.

We computed the four factor exposures described in Section 3.3.1: size, book-to-market ratio,[20] earnings-to-price ratio, and 12-month momentum. (For some empirical justification of these particular factors, see Fama and French (1992) and the references therein.)

Next, we divide our data set into groups based on a firm's market capitalization as noted in Section 3.3.1. We split our merged CRSP-Compustat data set into four data sets according to the market capitalization of a firm: on June 30 of year t, we order all firms in our data set that are listed on the NYSE by market capitalization and calculate quartiles of market capitalization. We then assign each firm in our data set (including AMEX and NASDAQ firms) on June 30 of year t to one of the four groups based on its market capitalization on June 30 of year t. To avoid stocks moving between classes too frequently, stocks remained

---

[19]Thus, for example, if a firm's fiscal year 2005 ends June 30, 2005, we assume its 2005 accounting data can be used by the market starting July 1, 2006, and remains constant until June 30, 2007. A ratio of book value to price calculated on July 31, 2006, would use the June 30, 2005, book value and the July 31, 2006, market price.

[20]We calculate book value using the methodology given in Fama and French (1993) and Davis et al. (2000).

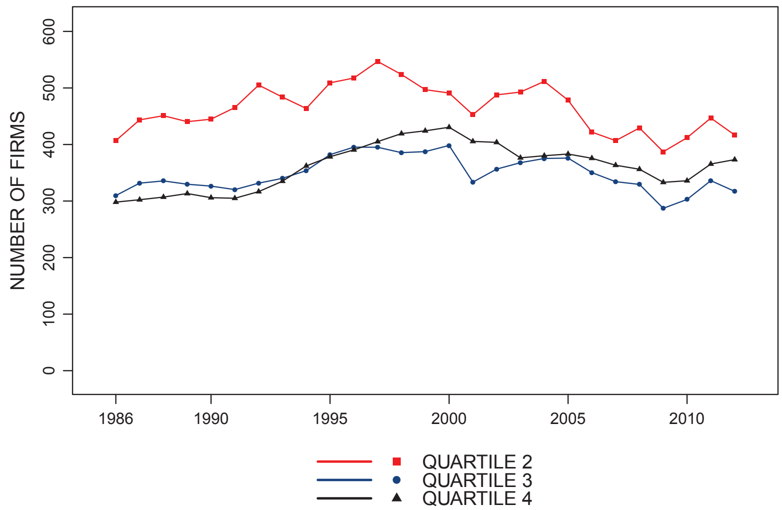


Figure 3.25: Number of firms in each capitalization quartile over time, prior to any trimming or Winsorization.

in their groups from July 1 of year t to June 30 of year $t+1$. We drop all stocks in the first market capitalization quartile: these very small stocks may be thinly traded and can exhibit very wild swings in their returns.

At any month end we drop stocks (a) with less than two years of accounting data; (b) with negative book value; (c) any missing factor data for that month end; and (d) less than 24 prior months of factor data. These measures do introduce some survivor bias, and omit some thinly-traded small capitalization stocks, but without taking such steps the amount of missing data would make direct calculations difficult in some places. Furthermore, the data associated with such short-lived and irregularly traded companies can be fairly abnormal. Such companies could potentially make the outlier situation of the present study worse than it would be for a "typical" manager, and could give an unfair advantage to the robust methods in our tests. We felt that the study would be more valuable if we constrained ourselves to a universe that a typical manager would use for portfolio construction.

Figure 3.25 shows the number of stocks in each capitalization group over time, after the screening criteria have been applied.

Chapter 4

FAMA-FRENCH 1992, REDUX

Abstract

Robust statistical methods provide estimates that are not much influenced by a small percentage of outliers but perform in a nearly optimal manner for normally distributed data. Unfortunately, such methods are rarely used in quantitative finance research despite their potential utility, particularly in empirical asset pricing studies. As a means of stimulating the use of robust statistical methods in such studies and in quantitative finance in general, we demonstrate the efficacy of using a theoretically well-justified robust regression method in the cross-sectional regressions often used in empirical asset pricing studies. We compare the results of using both least squares and robust regression methods for the models presented in Fama and French (1992) (FF92), as well as some extensions to these models, over the time period 1963–2015 and subsets thereof. Our analysis clearly demonstrates that a very small fraction of outliers, in the returns and/or the factors, often distorts least squares cross-sectional regression estimates sufficiently enough to result in misleading conclusions as to whether a risk factor is priced. We reconfirm previously reported robust regression results demonstrating a positive relationship between average equity returns and firm size during the period 1963–1990 of the FF92 study, and show that this relationship continues to hold through 2015. Furthermore, we show that, once the impact of extreme outliers is eliminated by use of robust regression, the size effect is significant in most months, not just in January as was previously shown by other researchers. We confirm and extend the FF92 results and other previous work demonstrating a positive relationship between average returns and the book-to-market ratio, and clarify that the relationship is driven largely by small stocks. Contrary to FF92 and other previous empirical studies, we find a significant and negative

relationship between average returns and beta for most U.S. equities between 1963 and 2015 when the influence of a small fraction of outliers is eliminated by robust regression. Furthermore, we show that there exists a highly significant interaction between size and beta that is needed to fully explain the dependence of returns on size and beta. Unlike FF92, who focused on positive earnings-to-price, we find that total earnings-to-price factor is highly significant in explaining the cross-section of returns for small to moderately-sized stocks both alone and when included along with the size and book-to-market factors. Finally, we also show that the time series of cross-sectional regression slopes contain large outliers, and that a robust location estimate provides a better representation of the "typical" slope than the sample mean.

4.1 Introduction

The landmark paper of Fama and French (1992) (FF92) used cross-sectional least squares regression analysis to argue that beta was insufficient to explain average returns as predicted by the Capital Asset Pricing Model (CAPM) of Sharpe (1964) and Lintner (1965). FF92 argued that adding (the logarithm of) a firm's market capitalization ("size") and book-to-market ratio explained more of the variation in the cross-section of stock returns in the U.S. during the years 1963–1990. Furthermore, their analysis indicated (a) the existence of a negative relationship between size and average returns, the so-called "size effect", previously documented by Banz (1981), Reinganum (1981), and Keim (1983); and (b) a positive relationship between a firm's book-to-market ratio and its average returns, the so-called "value effect", previously described in Stattman (1980), Rosenberg et al. (1985), and Chan et al. (1992). Their findings led them to develop (in another seminal paper, Fama and French (1993)) a three-factor model using beta, size, and book-to-market that is still used extensively today for portfolio analysis and benchmarking.

In the years after its publication several researchers questioned the findings of the original FF92 study. Several studies (Horowitz et al. (2000a), Horowitz et al. (2000b), Gu (2003), Easterday et al. (2009), and many others) have documented that the size effect is very strong in the month of January but weaker or non-existent for non-January months. Horowitz et al. (2000b) also found that the size effect is driven by the smallest stocks. Many researchers, among them Eleswarapu and Reinganum (1993), Dichev (1998), Chan et al. (2000), Amihud (2002), and Schwert (2003), found evidence that the strength of the size effect diminished over time, particularly after numerous publications in the 1990s studying the effect. (Crain (2011) and van Dijk (2011) provide reviews of the literature on the size effect.) As for the value effect, Loughran (1997) demonstrated that the overall value effect was driven largely by smaller stocks, and that the book-to-market factor had little explanatory power for large stocks. Loughran (1997) also found that January was a significant driver of the book-to-market effect. Later work by Fama and French (such as Davis et al. (2000), Fama and French (1996), and Fama and French (2008)) attempted to address some of these concerns. All of these issues with the conclusions of FF92 (and of FF93) were identified mainly through careful but tedious analysis of the data and the cross-sectional regression results.

It is well-known in the robust statistical literature that least squares (LS) regression is not robust in this sense, as it is sensitive to outliers in both the explanatory and response variables. Even a single outlier can severely distort the regression coefficients, leading to a model that is not representative of the vast majority of the data. A highly effective way to determine whether the LS cross-sectional regression results are influenced by a small fraction of outliers is to compute cross-sectional robust regressions in addition to the LS regression.[1] Comparing the robust and LS results will then reveal observations that may be distorting the LS regression analysis.

Outliers caused by one-time firm and market events are known to be common in asset returns and risk factor exposures. Indeed, robust methods have already shown promise for estimating simple time series factor models and correlations. A number of authors (Martin

[1] The term "robust" as used in this paper is not related to the use of the term "robust" in so-called "robust portfolio optimization" described in papers such as Goldfarb and Iyengar (1993), Erdogan et al. (2004), Ceria and Stubbs (2006), and others. We use the term "robust" to refer to estimators that are not much influenced by outliers, while robust portfolio optimization refers to a method of constructing an optimal portfolio in the presence of uncertainty about asset return and covariance forecasts.

and Simin (2003); Bailer et al. (2011), among many others) have demonstrated how robust regression can yield more reliable estimates of beta in the CAPM model. Martin et al. (2010) presents several applications of robust statistics to equity portfolio management. Scherer and Martin (2005) devote an entire chapter to applications of robust methods to portfolio analysis and optimization.

In the context of asset pricing research, robust regression is potentially a powerful tool for investigating whether risk premia are representative of all stocks or are misleading for the vast majority of stocks due to the influence of a small fraction of stocks. These influential observations are typically not data errors. Instead outliers represent unusual events, e.g., the stock price of a small pharmaceutical firm skyrocketing after getting FDA approval for a drug, or accounting charges arising from a one-time corporate restructuring. Outlying returns may arise simply from the often fat-tailed nature of asset returns. By limiting the impact of such infrequent outliers, we believe that asset pricing research results will do a better job of explaining the behavior of the vast majority of stocks, most of the time, while at the same time making it easy to identify situations where different models may be more appropriate.

We are not the first researchers to adopt this viewpoint in the context of empirical asset pricing studies. Several researchers used an early robust regression method called least trimmed squares (LTS) to reanalyze the FF92 study. Knez and Ready (1997) (KR97) determined that (a) the relationship between firm size and average returns is actually positive, not negative, for most stocks; and (b) Fama and French's results for size were driven by firm-level outliers associated with 1% of firms each month and by outlying cross-sectional regression estimates in 5% of the months used to compute the average risk premium on size. Chou et al. (2004) (CCW04) confirmed these findings in U.S. equity markets through 2001. Garza-Gómez et al. (2001) applied Knez and Ready's analysis in Japanese equity markets through 1995, finding that the least squares estimated size effect there was also strongly driven by a handful of influential firms/time periods. However such studies as those above are a rarity in the empirical asset pricing literature.

The present study builds upon the above prior studies that use LTS robust regression by using robust MM-regression, a modern robust regression method that is preferred to LTS by virtue of its theoretical properties. Thus we repeat the FF92 study with robust MM-regression for the 1963–1990 time period covered in that study, and compare the robust regression results to their LS results. We report striking differences in the conclusions drawn for the LS results versus the robust MM-regression results. Furthermore, during the 25 years since the Fama and French paper, and 20 years since the Knez and Ready paper, equity markets have experienced several boom and bust cycles and numerous structural changes. It is also of interest to understand whether the results of FF92 and our robust regressions still hold over the extended period 1990–2015. We note that the CRSP and Compustat data sources used in the original study have been cleaned up over the years,[2] and we now have the ability to analyze much larger data sets. With the availability of modern robust statistical methods, combined with more than 50 years of "good" data available and more computing power, it is easier to uncover long-term persistent trends, as well as anomalous local regime shifts, in risk premia.

This chapter's primary contribution to the asset pricing field is its demonstration of the utility of robust MM-regression in empirical asset pricing studies. We illustrate how to apply this technique by extending the FF92 analysis through December 2015. In the process we confirm the findings of Knez and Ready, namely that Fama and French's conclusions on the negative relationship between expected returns and firm size were driven by a small number of outliers. We show that the premium on firm size is positive for most stocks and most months, and we show that the result found by Fama and French was driven by a very small fraction of stocks each month. We show that the value effect is not driven by extreme outliers, but confirm the findings of Loughran (1997) and others that it is mainly concentrated in smaller stocks. We find that the value effect does not persist for larger stocks after 1980s, but it remains significant for the smallest stocks through the end of 2015. We demonstrate

[2]This has helped reduce some of the issues arising from bias in the data, such as the survivorship biases documented by Davis (1996), Shumway (1997), and Shumway and Warther (1999).

a complex and significant relationship between size and beta that offers new insight into the question of whether beta is still important for asset pricing models. Finally, we demonstrate that the negative earnings-to-price indicator and the positive earnings-to-price results found in FF92 were influenced by outliers, and show that an unmodified earnings-to-price is relevant for pricing smaller stocks.

This chapter proceeds as follows. Section 4.2 provides a self-contained introduction to robust regression. Section 4.3 details how we constructed our data set and performed our regression analyses. Section 4.4 presents the results of our single-factor regression analyses, while Section 4.5 presents the results for several multi-factor models. Section 4.6 discusses some extensions to models originally discussed in FF92. Section 4.7 presents an analysis of the sensitivity of our results to the choice of robust regression parameters and methodology. Section 4.8 summarizes the results of the chapter and offers some potential follow up research ideas.

4.2 Robust Regression

Robust regression is a means of producing reasonable estimates of regression coefficients in the presence of outliers in the data. For the reader's convenience we provide a brief overview of robust regression here. More details can be found in the texts Huber (1981), Hampel et al. (1986), Maronna et al. (2006), and the forthcoming Maronna et al. (2017).

4.2.1 Robustness Concepts

Efficiency Robustness

Tukey (1960) introduced the concept of efficiency as a tool for comparing estimators of a parameter. The *relative efficiency* under normality of one univariate estimate $\tilde{\theta}_n$ of θ relative to the MLE $\hat{\theta}_n$ is defined by the ratio

$$\mathrm{EFF}_n = \frac{\mathrm{Var}\left(\hat{\theta}_n\right)}{\mathrm{Var}\left(\tilde{\theta}_n\right)},$$

where the subscript n emphasizes that the estimates are based on a sample of size n, and the variances are computed assuming independent, normally distributed observations. The measure is made relative to the MLE since the MLE has the smallest possible asymptotic variance (given by the Cramér-Rao lower bound). The *asymptotic relative efficiency* EFF_∞ of the estimate is simply the limit of EFF_n as $n \to \infty$.

As a simple example of calculating efficiency, consider the efficiency of the sample median $\ddot{x}$ as an estimate of the center μ of a normal distribution $N(\mu, \sigma^2)$. The MLE $\hat{\mu}$ of μ is easily shown to be the sample mean $\bar{x}$, which has variance σ^2/n. The variance of the sample median can be shown to be $\pi\sigma^2/(2n)$. The efficiency of the median is thus $2/\pi \approx 0.64$. This means the sample mean can estimate μ in this setup just as accurately as the median but with 36% fewer observations. Alternatively, the median needs $\pi/2 \approx 1.57$ times as many observations to achieve the same level of accuracy as the sample mean under the assumption of normally distributed observations.

For a multivariate estimate $\tilde{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta}$, the efficiency relative to the MLE $\hat{\boldsymbol{\theta}}_n$ is defined by Maronna et al. (2006) as

$$\text{EFF}_n = \min_{\mathbf{x} \neq 0} \frac{\mathbf{x}^T \operatorname{Cov}\left(\hat{\boldsymbol{\theta}}_n\right) \mathbf{x}}{\mathbf{x}^T \operatorname{Cov}\left(\tilde{\boldsymbol{\theta}}_n\right) \mathbf{x}}.$$

For a given vector $\mathbf{x}$ the ratio measures the efficiency of the univariate estimate $\mathbf{x}^T\tilde{\boldsymbol{\theta}}_n\mathbf{x}$ of the quantity $\mathbf{x}^T\boldsymbol{\theta}\mathbf{x}$. The multivariate efficiency above is then the "worst case" univariate efficiency over all nonzero vectors $\mathbf{x}$.

Bias Robustness

Another way to evaluate an estimator is via its *bias*. For a univariate estimate $\tilde{\theta}$ of a parameter θ, the bias of $\tilde{\theta}$ is the difference between the expected value of $\tilde{\theta}$ and the target value:

$$\text{BIAS}\left(\tilde{\theta}\right) = E_\theta \tilde{\theta} - \theta.$$

It is a standard result in statistics that the mean squared error (MSE) of $\tilde{\theta}$ for estimating θ can be decomposed into the variance of the estimate and the square of its bias:

$$\text{MSE}\left(\tilde{\theta}\right) = E_\theta\left(\tilde{\theta} - \theta\right)^2 = E_\theta\left(\tilde{\theta} - E_\theta\tilde{\theta}\right)^2 + \left(E_\theta\tilde{\theta} - \theta\right)^2 = \text{Var}\left(\tilde{\theta}\right) + \text{BIAS}\left(\tilde{\theta}\right)^2.$$

This decomposition makes plain the trade-off between variance (and by extension, efficiency) and bias. For a fixed level of MSE, decreasing the variance of an estimate (equivalently, increasing its efficiency relative to the MLE) forces the bias to increase. It is often the case in statistical inference that we must balance efficiency and bias when selecting an estimator.

We can evaluate how resistant an estimate is to deviations from the assumed distribution by studying how its bias changes under such deviations. Huber (1964) first studied estimators of location that controlled the maximum bias over a family of distributions "near" an assumed model. For $0 < \epsilon < 1$, define the distribution F_ϵ as the mixture

$$F_\epsilon = (1 - \epsilon)F_\theta + \epsilon H, \tag{4.1}$$

where F_θ is the assumed distribution of the observations, parameterized by θ. The distribution F_ϵ is a "contaminated" version of F, where a small fraction of observations is replaced by observations from a different distribution, H.[3] The maximum asymptotic bias of $\tilde{\theta}$, over all such distributions F_ϵ, is then a measure of how far $\tilde{\theta}$ can be from θ under the contamination model (4.1). An estimate that minimizes the maximum bias over such a contamination model is called *bias robust*.

Huber (1964) showed that the sample median minimized the maximum bias that could be encountered over all perturbed distributions F_ϵ defined above and all location equivariant estimators when the assumed distribution F is symmetric and unimodal.[4] Martin and Zamar (1989, 1993) showed that the so-called "Shorth" estimate minimized the maximum bias over all M-estimates (defined below) of scale with arbitrary location.

---

[3] Huber (1991) points out that the contaminating observations are not necessarily bad data to be discarded—they may indicate that the assumed distribution is misspecified, for instance.

[4] An estimate $\tilde{\theta}(x_1, \ldots, x_n)$ is *location equivariant* if $\tilde{\theta}(x_1 + c, \ldots, x_n + c) = \tilde{\theta}(x_1, \ldots, x_n) + c$ for any constant c.

Martin et al. (1989) introduced a notion of bias for multivariate estimators:

$$\text{BIAS}\left(\tilde{\theta}\right) = \sqrt{\left(\tilde{\theta} - \theta\right)^T \mathbf{\Sigma}^{-1} \left(\tilde{\theta} - \theta\right)},$$

where $\mathbf{\Sigma}$ is the covariance matrix of the true distribution. Under this definition, they derived a robust regression estimator that minimizes the maximum bias over all M-estimates with arbitrary scale.

Breakdown Point

We also can measure how robust an estimate is to outliers via its breakdown point. Formally, the *breakdown point* of an estimate is the largest value of ϵ in (4.1) for which the maximum bias of the estimate is finite for all possible distributions H (Hampel, 1968). Intuitively, the breakdown point of an estimate is the largest fraction of the data which can be modified without substantially changing the estimate (Huber, 1981). For values of ϵ larger than the breakdown point, the bias of $\tilde{\theta}$ can become infinite, at which point the estimate is said to "break down". The sample mean, for instance, has breakdown point 0: replacing a single observation with ∞ would make the mean infinite, and $1/n \to 0$ as $n \to \infty$. Contrast this with the sample median: we could replace all observations larger than the median with ∞ without changing the median, so the median will have the highest possible breakdown point of 1/2 in large samples.

Martin et al. (1989) points out that a high breakdown point estimate can still exhibit large bias for fractions of contamination ϵ less than the breakdown point. Thus one often tries to design robust estimates that have high breakdown point, good control over bias for smaller fractions of contamination, and when possible, high efficiency for normally distributed data. As we shall see below, this goal is achievable using robust MM-regression.

4.2.2 Lack of Robustness of Least Squares

Let $(\mathbf{X}_i, y_i)$ be the observations, with $\mathbf{X}_i$ a $p \times 1$ dimensional vector (possibly including an intercept term). Assume the observations are independent and identically distributed with distribution function F. Suppose we wish to fit a linear regression model of the form

$$y_i = \mathbf{X}_i^T \boldsymbol{\beta} + r_i, \quad i = 1, \ldots, n,$$

with residuals $r_i = y_i - \mathbf{X}_i^T \boldsymbol{\beta}$. Assume that the residuals have expected value 0 and are independent of the regressors $\mathbf{X}_i$.

The ordinary least squares (LS) estimate $\boldsymbol{\beta}_{\text{LS}}$ of the coefficient vector $\boldsymbol{\beta} \in \mathbb{R}^{p \times 1}$ is obtained by minimizing the sum of the squared residuals:

$$\tilde{\boldsymbol{\beta}}_{\text{LS}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} r_i^2 \left(\boldsymbol{\beta} \right).$$

The LS estimate has several optimal properties—it is the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$, in the sense that it has the smallest variance of all linear unbiased estimators of $\boldsymbol{\beta}$. When the residuals are normally distributed, the LS estimate is the maximum likelihood estimate of $\boldsymbol{\beta}$, and is hence 100% efficient, as it has the smallest possible variance among all linear and nonlinear estimates. The LS estimate is well-known and well-understood, and is available in practically any data analysis software one might encounter.

The LS estimate of $\boldsymbol{\beta}$ is not robust to outliers in the response variable y_i or explanatory variables $\mathbf{X}_i$ however. In fact the LS estimate has breakdown point 0, as even a single outlier can lead to arbitrarily large bias (Maronna et al., 2006). Even if such an outlier does not lead to the catastrophic failure of $\boldsymbol{\beta}$, it often yields very misleading regression results. Suppose one observation $(\mathbf{X}_j, y_j)$ is outlying in some way, and let $\boldsymbol{\beta}^*$ be the regression coefficients estimated from the data if $(\mathbf{X}_j, y_j)$ is omitted. If the response variable y_j and/or the explanatory variables $\mathbf{X}_j$ are inconsistent with this model, the residual r_j for this observation will be large. Squaring it will make it even larger, and this residual will dominate the overall sum of squared residuals. The LS algorithm will "tilt" the regression model towards $(\mathbf{X}_j, y_j)$ so as to reduce the overall sum of squares, leading to a model that fits the outlier better, but possibly fits the non-outlying data points worse (depending on the configuration of the data). LS may thus yield a model that is a poor fit for most, if not all, of the observations.

4.2.3 Huber M-Estimator and Least Trimmed Squares

The sensitivity of the LS regression estimate to outliers is a consequence of its unbounded loss function $\rho(x) = x^2$, which magnifies the importance of large residuals in the objective. This allows the maximum asymptotic bias of the LS estimate to become arbitrarily large. We can improve this situation by minimizing the sum of a different function of the residuals, one that puts less weight on very large residuals. Huber (1973) introduced the concept of a *regression M-estimate*, an estimate for $\boldsymbol{\beta}$ that minimizes the sum

$$\sum_{i=1}^{n} \rho\left(\frac{r_i\left(\boldsymbol{\beta}\right)}{\tilde{\sigma}}\right). \tag{4.2}$$

Here $\tilde{\sigma}$ is a scale estimate, and $\rho(x)$ is a loss function that is selected to optimize resistance to outliers in some way. The LS estimate is an M-estimate with loss function $\rho(x) = x^2$, and setting $\rho(x) = |x|$ yields the least absolute deviation (LAD), or L^1 estimate, of $\boldsymbol{\beta}$. Huber (1964) found that the family of loss functions defined by function

$$\rho(x;k) = \begin{cases} x^2, & |x| \leq k \\ 2k|x| - k^2, & |x| > k \end{cases} \tag{4.3}$$

minimized the maximum asymptotic variance of the M-estimate. Figure 4.1 shows the shape of this loss function and its derivative $\psi(x;k)$. The loss function is quadratic near the origin like the LS loss function, but linear beyond $x = \pm k$ like the LAD estimate. The linear portion of the loss function reduces the impact of large residuals in the calculation of the M-estimate. The breakdown point of this estimate is still 0, however, as the loss function is unbounded. The magnitude of large residuals is not limited, so the bias of the slope estimates can be arbitrarily large. In general, M-estimates as defined above have breakdown point 0, and are not robust to outliers in the explanatory variables $\mathbf{X}_i$.

This shortcoming of M-estimates led researchers to develop other approaches. Rousseeuw (1984) introduced least trimmed squares (LTS) regression. In LTS regression, we order the residuals from smallest to largest, remove the extreme values, and minimize the sum of the

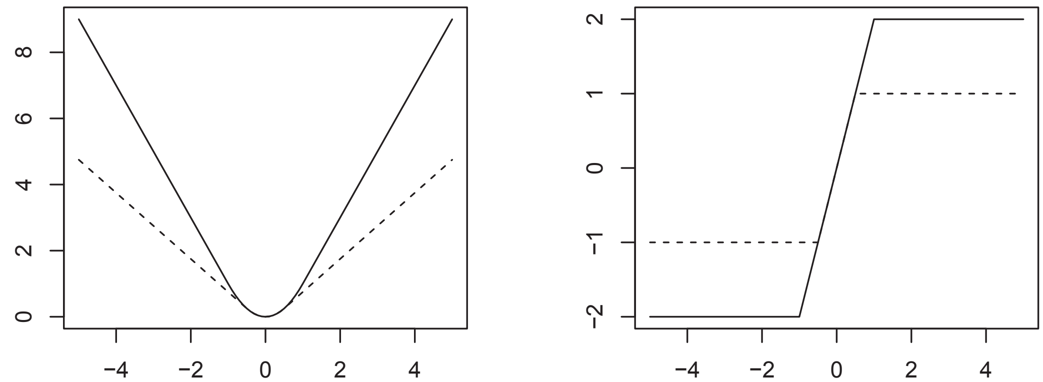


Figure 4.1: Huber loss functions for $k = 1$ (solid lines) and $k = 1/2$ (dashed lines). Left panel: Huber loss function $\rho(x)$, as defined in Equation (4.3). Right panel: The derivative $\psi(x)$ of the Huber loss function.

remaining squared residuals. If we denote the ith smallest squared residual as $r^2_{(i)}$, we can express the LTS estimate as

$$\tilde{\boldsymbol{\beta}}_{\text{LTS}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{h} r^2_{(i)}\left(\boldsymbol{\beta}\right),$$

where $1 \leq h \leq n$ controls how much trimming is done. (For $h = n$ we recover the usual LS regression estimate.) Rousseeuw (1984) recommended using $h = \lfloor n/2 \rfloor + \lfloor (p+1)/2 \rfloor$ to achieve an estimate $\tilde{\boldsymbol{\beta}}_{\text{LTS}}$ with breakdown point 1/2.

LTS, however, does not yield a very *efficient* estimate of $\boldsymbol{\beta}$, in the sense that when the residuals are normally distributed, the LTS estimate of $\boldsymbol{\beta}$ has a much higher sampling variance than the LS estimate. Rousseeuw and Leroy (1987) derived the asymptotic variance of the LTS estimate and showed that LTS with the maximum breakdown point choice of h above has an efficiency of only 7%, which means that the standard error of the LTS estimate of $\boldsymbol{\beta}$ is about $\sqrt{1/0.07} \approx 3.8$ times larger than the standard error of the LS estimate. Table 4.1 shows the efficiency of the LTS estimate for various trimming fractions. Here we again see the tradeoff between robustness to outliers and efficiency: higher trimming fractions lead to LTS estimates with higher breakdown points but lower efficiency. In other words, trimming very little data leads to an estimate that is nearly as efficient as LS for normally distributed

	Trimming Percentage					
	50%	25%	10%	5%	1%	0.1%
Efficiency	7%	28%	56%	72%	92%	99%

Table 4.1: Efficiency of the LTS estimate for several trimming fractions.

data, but very sensitive to departures from normality.

4.2.4 MM-Estimates

Yohai (1987) developed a modification to the M-estimate, called the *regression MM-estimate*, that is robust to outliers in the explanatory variables as well as the response variables. Unlike M-estimates and LTS, MM-estimates can have both high breakdown points and high efficiency when the residuals are normally distributed. The MM-regression procedure, developed by Yohai (1987) and refined by Yohai et al. (1991), proceeds as follows.

1. Compute an initial M-estimate $\tilde{\boldsymbol{\beta}}_0$ of $\boldsymbol{\beta}$ with high breakdown point.

2. Estimate the residual scale σ using the residuals from the initial estimate. We compute $\tilde{\sigma}$ of $\{r_i(\tilde{\boldsymbol{\beta}}_0)\}$ using a robust estimate with high breakdown point.

3. The final MM-estimate $\tilde{\boldsymbol{\beta}}_1$ of $\boldsymbol{\beta}$ is obtained by numerically solving the equations

$$\sum_{i=1}^{n} \mathbf{X}_i \psi_{1,c} \left(\frac{r_i(\boldsymbol{\beta})}{\tilde{\sigma}} \right) = 0$$

   for $\boldsymbol{\beta}$. Here $\psi_{1,c}$ is the derivative of a loss function $\rho_{1,c}$. We choose the loss function parameter c to achieve our desired efficiency when the data are normally distributed.

   The solution must satisfy the additional condition that the corresponding objective is smaller at the solution than at the initial estimate:

$$\sum_{i=1}^{n} \rho_{1,c} \left(\frac{r_i(\tilde{\boldsymbol{\beta}}_1)}{\tilde{\sigma}} \right) \leq \sum_{i=1}^{n} \rho_{1,c} \left(\frac{r_i(\tilde{\boldsymbol{\beta}}_0)}{\tilde{\sigma}} \right).$$

The final equation may have multiple roots, but Yohai (1987) shows that we do not need to find the global minimum to obtain a high breakdown point and high efficiency estimate: any local minimum close to the initial estimate will be an MM-estimate. Furthermore, such a local minimum will inherit the high breakdown point, consistency, and equivariance properties of the initial estimate, and will be just as efficient as the global minimum.

The initial estimates $\tilde{\boldsymbol{\beta}}_0$ and $\tilde{\sigma}$ are found using the approach of Rousseeuw and Yohai (1984): $\tilde{\boldsymbol{\beta}}_0$ is the solution to the equation

$$\frac{1}{n}\sum_{i=1}^{n}\rho_{0,c}\left(\frac{r_i(\tilde{\boldsymbol{\beta}})}{\tilde{\sigma}}\right) = 0.5$$

where the loss function $\rho_{0,c}$ and tuning parameter c are chosen to yield an estimate $\tilde{\sigma}$ with breakdown point 1/2. Solving this problem can be quite hard, as the "good" choices for loss functions are bounded, non-convex functions. Yohai et al. (1991) developed a resampling-based algorithm for solving this problem in a reasonable amount of time.

Martin et al. (1989) showed that good loss functions ρ should be bounded to limit the bias that could be caused by outliers, and symmetric to treat positive and negative outliers in the same fashion. The corresponding $\psi = \rho'$ functions should give full weight to "good" data and zero weight to outliers, with a smooth transition between these cases to downweight moderate outliers. Yohai and Zamar (1997) derived an optimal loss function that minimizes the maximum asymptotic bias under certain types of departures from normality while guaranteeing a minimum efficiency when the data are normally distributed. Svarc et al. (2002) provided a piecewise polynomial approximation to the Yohai-Zamar function that is computationally more tractable and is used in actual implementations of MM-regression. The loss function and its derivative are shown in Equations (4.4) and (4.5), respectively.

$$\rho(r) = \begin{cases} \frac{r^2}{2}, & |r/c| \leq 2 \\ c^2\left[1.792 - \frac{1.944}{2}(r/c)^2 + \frac{1.728}{4}(r/c)^4 - \right. & \\ \qquad \left. \frac{0.312}{6}(r/c)^6 + \frac{0.016}{8}(r/c)^8\right], & 2 < |r/c| \leq 3 \\ 3.25c^2, & |r/c| > 3 \end{cases} \tag{4.4}$$

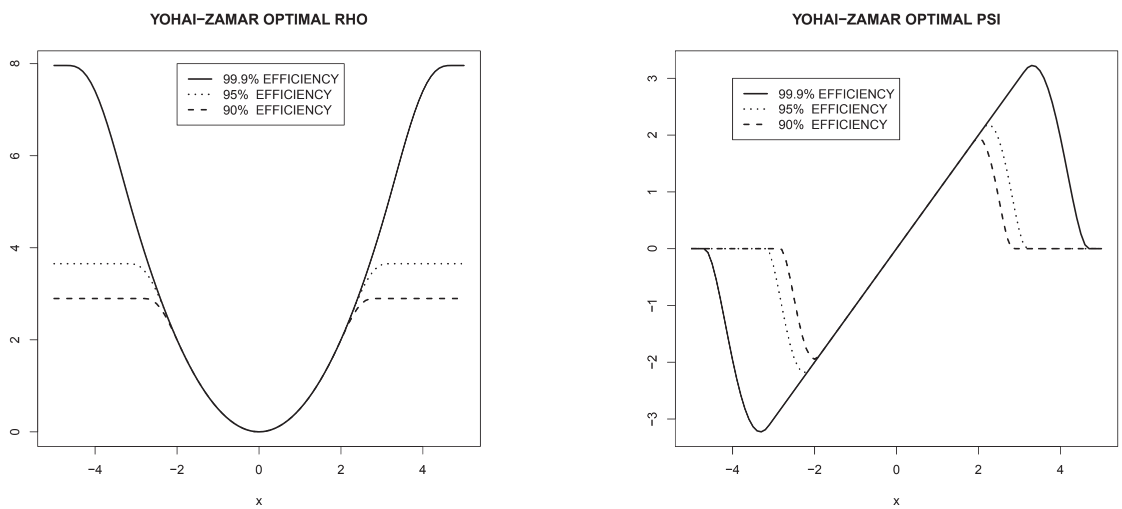


Figure 4.2: Piecewise approximation to the bias optimal ρ and ψ functions of Yohai and Zamar (1997) for several common efficiencies.

$$\psi(r) = \begin{cases} r, & |r/c| \leq 2 \\ c\left[-1.944\,(r/c) + 1.728\,(r/c)^3 - \right. & \\ \qquad \left. 0.312\,(r/c)^5 + 0.016\,(r/c)^7\right], & 2 < |r/c| \leq 3 \\ 0, & |r/c| > 3 \end{cases} \tag{4.5}$$

By using this function during Steps 1–3 (with different tuning constants c for the initial and final estimates) we obtain an MM-estimate $\tilde{\boldsymbol{\beta}}_1$ with maximal breakdown point 1/2 and high efficiency.

Figure 4.2 plots this function for various values of the tuning parameter c, corresponding to commonly used efficiencies. The c parameter controls where the ψ function begins to downweight observations that are moderately outlying and where it begins to reject (i.e., assign zero weight to) observations that are extreme outliers. It leaves residuals less than $2c$ unchanged, and rejects residuals larger than $3c$. The rapidly redescending $\psi(x)$ function downweights moderate residuals that fall between $2c$ and $3c$.

The tuning constant c also controls the breakdown point and efficiency of the M-estimates

Table 4.2: Tuning constants c, rejection threshold for scaled residuals, and the probability $P(|N(0,1)| > 3c)$ of rejecting extreme values under normality for the Yohai-Zamar loss function at various efficiencies.

	Efficiency			
	90%	95%	99%	99.9%
c	0.944	1.06	1.29	1.565
Hard Rejection Threshold ($3c$)	2.832	3.181	3.869	4.695
$P(\lvert N(0,1)\rvert > 3c)$	0.462%	0.147%	0.011%	0.0003%

obtained in each step of the MM-regression procedure. Using $c = 0.4047$ in Steps 1 and 2 yields a scale estimate with breakdown point 1/2 but low efficiency. In Step 3 we chose a different value of c to achieve our desired efficiency when the data are normally distributed. Table 4.2 shows the values of c that yield commonly used efficiencies.[5] The table also provides the threshold beyond which observations will be rejected, and the fraction of observations $P(|N(0,1)| > 3c)$ expected to be rejected in the presence of normally distributed observations.

Another popular choice of loss function for MM-regression is the Tukey bisquare function, whose $\rho(r)$ and $\psi(r)$ functions are shown in Equations (4.6) and (4.7).

$$\rho(r) = \begin{cases} (r/c)^6 - 3\,(r/c)^4 + 3\,(r/c)^2\,, & |r/c| \le 1 \\ 1, & |r/c| > 1 \end{cases} \tag{4.6}$$

$$\psi(r) = \begin{cases} \frac{6}{c}\,(r/c)^5 - \frac{12}{c}\,(r/c)^3 + \frac{6}{c}\,(r/c)\,, & |r/c| \le 1 \\ 0, & |r/c| > 1. \end{cases} \tag{4.7}$$

For the bisquare, we set $c = 1.548$ in Steps 1 and 2 to achieve a breakdown point of 1/2, and then chose a different c in Step 3 to hit an efficiency target. Table 4.3 provides the relevant tuning and rejection constants for the bisquare loss function.

[5] The c values can be obtained using the `lmRob.effvy` function from the `robust` `R` package.

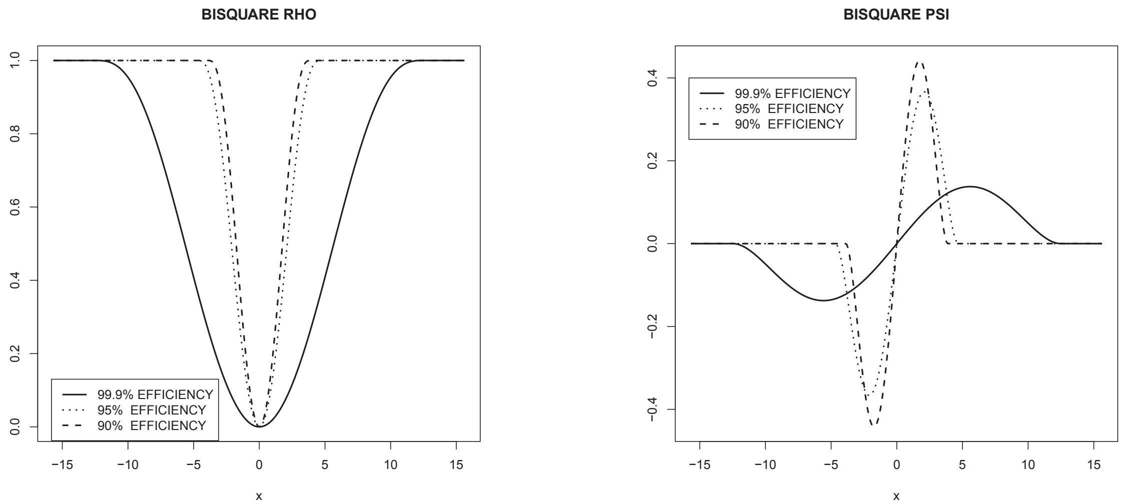


Figure 4.3: Tukey bisquare ρ and ψ functions for several common efficiencies.

With normally distributed residuals, the MM-estimate of $\boldsymbol{\beta}$ is consistent, but not as efficient as the LS estimate. The loss of efficiency relative to least squares is insurance against the bias that can be created when the distribution of the residuals is non-normal. The MM-estimate remains consistent and efficient under non-normality, though it should be noted that asymmetry in the distribution of residuals can introduce bias in the MM-estimate of the intercept of a regression model (Maronna et al., 2006).

Table 4.3: Tuning constants c, rejection threshold for scaled residuals, and the probability $P(|N(0,1)| > c)$ of rejecting extreme values under normality for the Tukey bisquare loss function at various efficiencies.

	Efficiency			
	90%	95%	99%	99.9%
c	3.883	4.685	7.041	12.482
Hard Rejection Threshold (c)	3.883	4.685	7.041	12.482
$P(\|N(0,1)\| > c)$	1×10^{-2} %	3×10^{-4}%	2×10^{-10}%	9×10^{-34}%

4.2.5 Computing MM-Estimators

MM-regression is available in R via the function lmRob in the robust package (Wang et al., 2014) and via the function lmrob in the robustbase package (Rousseeuw et al., 2016). We used the lmRob version throughout the present study. MM-regression is available in SAS via the ROBUSTREG procedure, and in Stata via the contributed package mmregress created by Verardi and Croux (2009).

4.3 Data and Estimation Methodology

Our study analyzes the extent to which the cross-section of stock returns are explained by one or more of several firm characteristics using the same Fama and MacBeth (1973) cross-sectional regression method used by FF92, as well as in many other empirical asset pricing studies.[6] We use the following firm characteristics treated by FF92: a firm's CAPM beta to the market, its size (the logarithm of its market equity in millions on June 30 of each year), a value measure (the logarithm of the ratio of its book value to its market value), two leverage measures (the logarithm of the ratio of total assets to market value; the logarithm of the ratio of total assets to book value), and two earnings measures (an indicator for stocks with negative earnings; the ratio of positive earnings to price). In addition we also study the relationship of the cross-section of returns to the ratio of earnings to price (regardless of whether earnings are negative or positive). All of these except beta will be known quantities in our data set; beta, on the other hand, must be estimated via regression. This creates an errors-in-variables problem that is well-known in econometrics and in the empirical finance literature. We use FF92's approach to this problem, namely, the creation of size-beta portfolios, in our study. This approach is described below in Section 4.3.3.

The monthly data set for our replication of the FF92 study starts on July 31, 1963, and ends on December 31, 1990. Our extended analysis uses a data set ending on December 31, 2015. Our analysis also considers the period from January 31, 1980, through December

[6] See also Fama (1976); Shanken (1992); Cochrane (2001) for more details on Fama-MacBeth regression.

31, 2015, as it excludes an anomalous highly non-stationary and volatile time period in the vicinity of 1970. (Examples of such behavior will be discussed in Sections 4.4 and 4.5.)

4.3.1 Construction of the Data Set

To effect a proper comparison of our robust approach to the LS-based cross-sectional approach used not only in FF92 but also most empirical asset pricing studies, our data set is constructed generally in the same manner as in FF92. Our process for building the data set is based on the code for the Fama and French (1993) study that was written by Palacios and Vora (2011). Data sets were built using SAS 9.4 on the WRDS cloud servers (Wharton School, 1993). We summarize key points here for the convenience of the reader.

We acquired via CRSP (2015b) stock prices, returns, and market equity for common stocks listed on NYSE, AMEX, or NASDAQ starting in 1959. At the time we built the data set, stock data through the end of 2015 was available via CRSP. We exclude financial firms (identified by SIC codes 6000–6999) as is typical in asset pricing studies. We include delisting returns in our study to reduce survivorship bias in our data set. We set aside December prices and market equities for use in ratios, and June market equities for assigning stocks to size deciles.

Next we calculate the pre-ranking betas of FF92. On each June 30, we perform a time-series regression of a stock's returns $r_{i,t}$ for the prior five years on the CRSP value-weighted index for the same period m_t and lagged by one month m_{t-1}:

$$r_{i,t} = \alpha + \beta_0 m_t + \beta_1 m_{t-1}.$$

Stocks with less than 24 months of returns data are not assigned a beta for this June, while stocks with 24-60 months of data use as much data as possible in the regression. A firm's pre-ranking beta is estimated using the sum beta approach of Dimson (1979) as $\beta_i = \beta_0 + \beta_1$.

Fama and French used least squares regression (LS) to compute the betas. Previous research by Martin and Simin (2003) and Bailer et al. (2011) has shown that stock betas computed using LS can be adversely influenced by outliers. Thus we also perform the regression above using the robust MM-regression described earlier (with 99.9% efficiency). Hence at each June 30, each stock with sufficient history will have a size (logarithm of market equity in millions), an LS pre-ranking beta, and a robust pre-ranking beta.

Company fundamentals and accounting variables were obtained from the Compustat (2015) Xpressfeed annual database starting in 1959. We use the variable constructions as in FF92, with minor adjustments to deal with missing values.[7] Ratios of accounting variables, namely, book-to-market, earnings-to-price, assets-to-market, and assets-to-book, are calculated using December market equity or prices.[8]

The Compustat and CRSP databases are joined using the PERMNO-gvkey linking information contained in the CRSP-Compustat Merged database maintained by WRDS (CRSP, 2015a). We use the same 6–18 month lagging approach as FF92, even through such a long lag is no longer necessary and is inconsistent with current industry practice.[9] After joining the CRSP and Compustat data sets, we apply the filter described in FF92: as of July 1 of year t, a stock must have (a) non-missing CRSP stock price for December of year $t-1$ and June of year t; at least 24 non-missing monthly returns for the previous 60 months; and non-missing total assets, book equity, and earnings for (calendar) year $t-1$.

We assign each stock to a size and a beta decile on June 30 of each year, again as described in FF92. We assign stocks to both LS beta deciles and robust-regression beta deciles as a means of testing whether outliers strongly influence the characteristics of the beta deciles.

---

[7]Book equity is stockholder's equity (SEQ) plus balance-sheet deferred taxes (TXDB); if stockholder's equity is missing, we estimate it by common equity (CEQ), plus the par value of preferred stock (PSTK). If either of those are missing, we estimate book equity as total assets (AT) minus total liabilities (LT). Earnings (in millions) is calculated as earnings per share excluding extraordinary items (EPSPX) times common shares (CSHPRI), plus income-statement deferred taxes (TXDI) minus preferred dividends (DVP). (Xpressfeed variable names are given in parentheses.)

[8]Ratios of accounting variables to market variables combine annually reported data with monthly data. Furthermore, since firm fiscal year ends vary, the accounting variables are measured at different times. Fama and French (1992) chose to use December market equity to compute accounting ratios (such as book-to-market); their tests showed that using a firm's market equity at its fiscal year end did not significantly change the results of their analysis. We followed their setup for our data set. It should be noted that this is no longer common practice; see Asness and Frazzini (2013) and Breitschwerdt (2015) for commentary on this matter.

[9]See Asness and Frazzini (2013) for an empirical discussion of this matter.

4.3.2 Size-Beta Portfolio Formation and Calculation of Post-Ranking Betas

On the principle that the estimation error of a portfolio will be lower than the estimation error of a single stock, we investigate the relationship between returns and beta using portfolio-level betas rather than firm-level betas. Chan and Chen (1988) found empirical evidence that stock betas varied significantly with firm size, and thus advocated analyzing returns and beta within size-based portfolios. FF92 pointed out that there can still be meaningful variations in beta within size groups that is obscured by the strong correlation between size and beta in the Chan and Chen (1988) approach. We therefore follow the FF92 approach to beta: each June 30, we form 100 equally-weighted portfolios based on combinations of 10 size deciles and 10 pre-ranking beta deciles. We compute the return (including dividends and the delisting return) on this portfolio for each month, giving us a time series of returns on each size-beta portfolio.

Next, we estimate, as FF92 did, "post-ranking" betas for each portfolio over the entire time frame by regressing these time series on the CRSP value-weighted index. As before, we regress on both the market proxy and a one-month lagged version, and then estimate the portfolio beta by summing the slopes from the regression. This "post-ranking" beta is then assigned to all stocks within a size-beta group. Note that (a) the assignment is constant between the yearly rebalances on June 30; and (b) there are only 100 post-ranking betas.

The goal of the size-beta sort was to reduce size-related variation in the (post-ranking) betas to be used in the cross-sectional regressions, thereby obviating the need to use a more complicated errors-in-variables regression method. The assumption is that the "true" betas for stocks within a size-beta group will not be far from the "average" beta for the group. We test this by computing (at each June 30) the standard deviation, over the assets within a size-beta group, of the pre-ranking betas. Table 4.4 shows the time-series averages of these standard deviations (with the extreme beta deciles split in half). We find that there is not much variation in the pre-ranking betas within the size groups except in the 5th and 95th percentiles. This suggests that our post-ranking beta variable will work as designed in general, except possibly for stocks with very low or very high betas.

4.3.3 Estimation Methodologies

Once post-ranking betas are computed, we compute cross-sectional regressions of stock returns on various combinations of firm characteristics, for each of the three time periods we study (1963–1990, 1963–2015, and 1980–2015).[10] The size-beta portfolios are formed every June 30 after the market close and are implemented the following July 1 at the opening of the markets. Firm characteristics such as accounting ratios are held constant for the next 12 months, and we analyze the relationships between the returns on stocks in each portfolio over the next 12 months and the predictive variables.

After performing this regression over all months of one of the three time periods, we obtain a time series of regression slopes. FF92 reported their average slopes and t-statistics (average slope divided by its standard error) for their analyses in their Table III, and considered a factor to be "priced" by the market if its t-statistic exceeded 1.96 in absolute value (i.e., if the average slope was significant at the 0.05 level).

The original Fama and MacBeth study used LS to compute the cross-sectional regressions. As we discussed in Section 4.2, however, LS can yield poor estimates of regression parameters in the presence of outliers in either the response variable or the explanatory variables. Fama and French trimmed the accounting ratios in their study by 0.5% at each end to deal with this problem. Univariate trimming of the regressors, however, would neither address outliers in the returns nor multivariate outliers in the regressors.[11] Furthermore, the trimming percentage is subjective: one might obtain markedly different regression results simply by trimming more observations away.

---

[10] The cross-sectional regressions and ensuing analyses were performed using a laptop running Windows 7 Ultimate SP 1 and R 3.2.4 with an Intel® Core™ i7-3740QM processor running at 2.7GHz and 32GB of RAM.

[11] Additionally, FF92 did not apply trimming to the beta and size regressors, so any outliers in those variables could have affected the LS analyses.

Table 4.4: Time-series average standard deviation, across assets, of June pre-ranking betas with top and bottom beta deciles split, 1963–1990.

	β-01a	β-01b	β-02	β-03	β-04	β-05	β-06	β-07	β-08	β-09	β-10a	β-10b
ME01	0.36	0.05	0.06	0.04	0.04	0.04	0.04	0.04	0.06	0.08	0.08	0.55
ME02	0.21	0.06	0.06	0.04	0.04	0.04	0.04	0.05	0.06	0.08	0.08	0.45
ME03	0.23	0.05	0.06	0.05	0.04	0.03	0.05	0.05	0.06	0.08	0.08	0.39
ME04	0.20	0.06	0.07	0.05	0.04	0.04	0.05	0.04	0.06	0.08	0.08	0.33
ME05	0.17	0.05	0.08	0.06	0.04	0.04	0.04	0.05	0.06	0.08	0.08	0.30
ME06	0.09	0.04	0.08	0.05	0.04	0.04	0.04	0.04	0.05	0.08	0.07	0.31
ME07	0.10	0.03	0.08	0.05	0.04	0.04	0.04	0.04	0.06	0.09	0.09	0.26
ME08	0.08	0.03	0.06	0.05	0.04	0.04	0.04	0.04	0.05	0.07	0.07	0.31
ME09	0.09	0.03	0.06	0.05	0.03	0.04	0.03	0.04	0.04	0.07	0.06	0.23
ME10	0.07	0.04	0.04	0.04	0.03	0.03	0.04	0.03	0.04	0.06	0.06	0.17

Throughout our analysis we use the robust MM-regression described earlier, with 99.9% efficiency, in each cross-sectional regression. The choice of 99.9% efficiency is motivated by a desire to eliminate the influence of a small fraction of extreme outliers in the data while still offering performance essentially identical to LS regression when the data contain no outliers or are close to normally distributed. (For comparison, we have also computed 95%, 99% efficiency robust MM-regressions for each cross-section. We show in Section 4.7 that one draws the same conclusions from the robust MM-regression with any of these efficiencies.)

We note that KR97 and CCW04 used LTS for each cross-sectional regression, with varying degrees of trimming, to mitigate the effects of outliers. As we discussed in Section 4.2, however, the LTS estimator can be very inefficient, and lacks many of the optimality properties possessed by robust MM-regression estimators.

Robust cross-sectional regression protects against the adverse influence of firm-level outliers in the monthly data sets, thereby providing a good fit to the bulk of the data for each month. Whether estimated via LS, LTS, or robust regression, however, some monthly slopes may be outliers in the time series of slopes. Furthermore, the time series of slopes may exhibit non-normality and serial correlation. While the significance level of the classical t-statistic is robust to outliers and heavy-tailed distributions, the t-statistic may be less powerful in these circumstances than other tests. Serial correlation is a more serious problem, as neither the level nor the power of the t-statistic is robust to violations of the assumption of independent observations.[12] This issue can be remedied by replacing the sample mean of the slopes with a robust location estimate, and using a corresponding t-statistic that is robust with respect to both significance level and power. We examine this alternative approach in Section 4.6.

4.3.4 Size-Beta Portfolio Statistics

Figures 4.4 and 4.5 show the time-series average returns (over July 1963–December 1990) for each size-beta group. The former plot shows returns versus size for each beta decile, while

[12] Lehmann (2004) discusses the behavior of the t-test under non-normality and dependent observations.

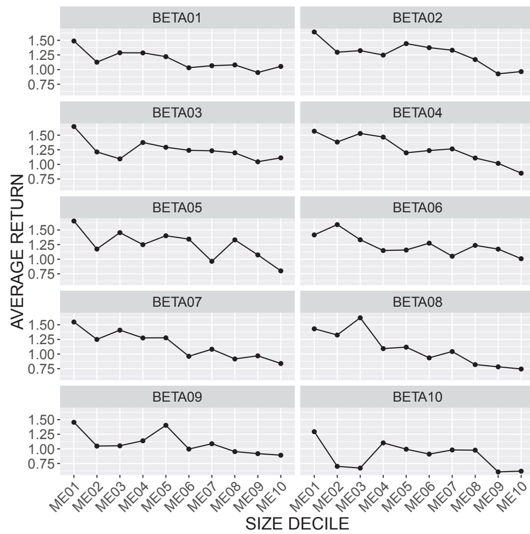


Figure 4.4: Time-series averages of monthly returns versus size within beta deciles.

the latter plot shows returns versus beta for each size decile. The data for this analysis is also shown in Table 4.38 of Appendix 4.C and should be compared to the data presented in Table Ia of FF92. Overall our average portfolio returns are typically close to those of FF92, but can differ by as much as 60 basis points in absolute value for some combinations of size and beta. Both CRSP and Compustat have updated and backfilled their database since 1992, so we are working with slightly different data than FF92. As we will see below, our regression results agree rather well with those of FF92 despite these differences.[13]

4.4 Robust Regression Analysis of Single-Factor Fama-French Models

We first discuss the results of our LS and robust univariate regression analyses of the individual Fama-French factors, namely book-to-market, size, beta, earnings-to-price, and leverage, for the three time periods: the FF92 1963–1990 interval, and the intervals 1963–2015 and 1980–2015. Overall, the results for the longer time periods are fairly similar to those for the 1963–1990 period, which suggests that the relationships between returns and the factors

[13] We were also able to replicate many of the tables in FF92 quite closely. Appendix 4.C details our replication results.

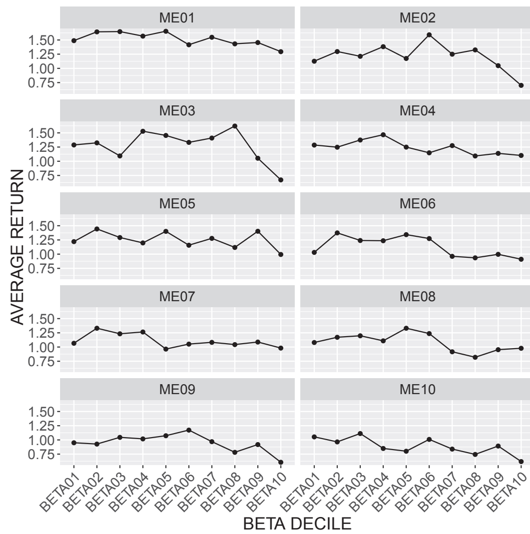


Figure 4.5: Time-series averages of monthly returns versus beta within size deciles.

found by both the LS and robust regressions have been reasonably stable over a period of time nearly twice as long as that of the original FF92 study.

We note that our LS results for the 1963–1990 period generally agree quite well with those of FF92, which indicates that we replicated their data set sufficiently accurately. The few cases where our LS results do not match those of FF92 very well are partially explained by our use of delisting returns in our data set. For instance, without delisting returns our average slope for the post-ranking beta factor increases from 0.05 to 0.09, which is closer to the FF92 average of 0.15. We believe that the remaining modest differences are likely due to revisions to the CRSP and Compustat data sets since 1992, but this would be rather difficult to verify.

On the other hand, our robust model fitting results often differ substantially from those of FF92. Our results are quite consistent, however, with prior uses of robust regression for the FF92 models by KR97 over 1963–1990 and CCW04 over 1963–2001 for the factors considered in each of these papers.

4.4.1 Book-to-Market

Table 4.5 and Figure 4.6 show the average slopes and t-statistics over all monthly cross-sectional regressions of stock returns on the book-to-market factor. The robust regressions do not suggest that firm-level outliers are substantially influencing the relationship between returns and the book-to-market factor. This agrees with the findings of KR97 and CCW04 using LTS regression. The book-to-market factor is significant in all the regressions considered, according to the t-statistic. This remains true in the extended time periods, suggesting that the relationship between returns and book-to-market has been fairly stable over time. As in FF92 we trimmed the data set by removing observations corresponding to the smallest 0.5% and largest 0.5% of the book-to-market factor prior to the LS regressions to mitigate the potential presence of extreme outliers. Our LS and robust results confirm that, for the book-to-market regression, this preliminary trimming was sufficient to deal with book-to-market outliers, if any. The distributions of the LS and robust slopes, shown in Figure 4.7, are quite similar, and the distribution of the paired differences in slopes (LS − robust) does not exhibit the large differences in slopes we would expect to see if there were numerous influential outliers in the data set.

Loughran (1997), however, found that the results of FF92 for book-to-market were driven by a January effect and a small stock effect. Figure 4.8 shows a heatmap of the regression slopes by month and by year. There is some evidence of a January effect, particularly in the large positive slopes observed in the mid-1970s. Figure 4.9 shows a hexbin scatterplot (developed by Carr et al. (1987)) of the cross-section data for January 1975, along with the fitted LS and robust regression lines. While there are a small number of returns outliers biasing the LS slope upwards, the large positive LS and robust slopes at this time do not appear to be driven solely by outliers.

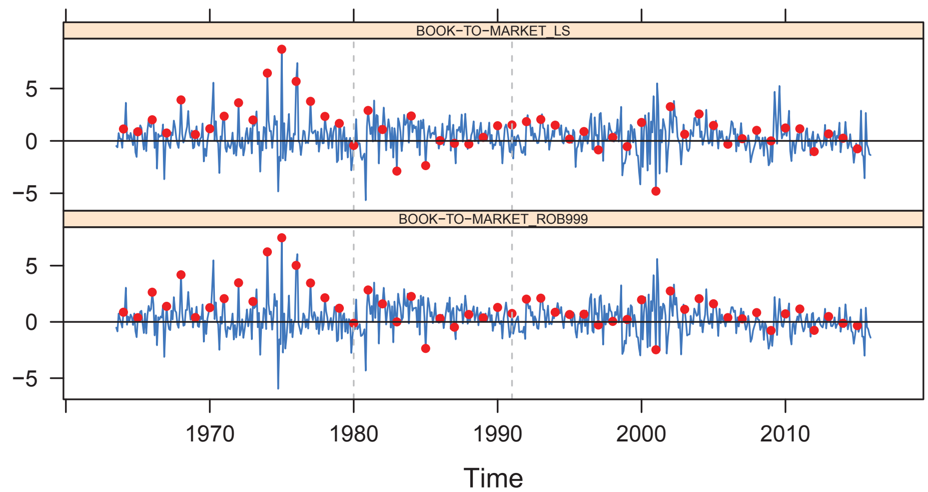


Figure 4.6: Time series of slopes from regression of returns on book-to-market. The top half of the chart shows the time series of LS slopes, and the bottom half shows the robust regression slopes. The left gray vertical line in these plots indicates the beginning of the 1980–2015 period, while the right gray vertical line indicates the end of the 1963–1990 period used by Fama and French. Red dots indicate slopes for January months.

Summary Statistics for
Book-to-Market Slopes

	LS	Robust
Minimum	-5.66	-5.94
1st Quartile	-0.43	-0.48
Median	0.37	0.28
Mean	0.39	0.30
3rd Quartile	1.21	1.04
Maximum	8.76	7.48

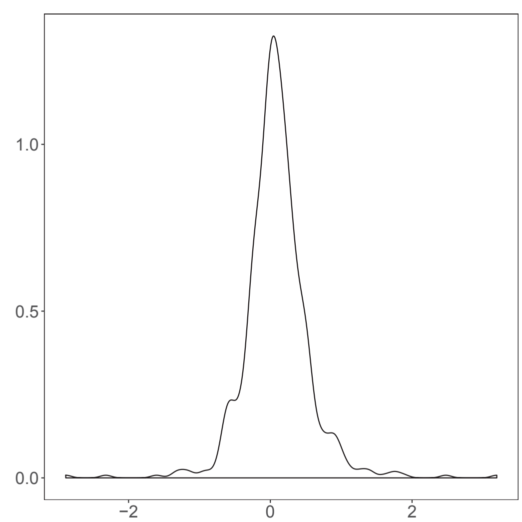


Figure 4.7: Statistics on monthly LS and robust slopes on book-to-market, 1963–2015. Left panel: summary statistics for each series. Right panel: kernel density estimate of the distribution of paired differences (LS − robust) in monthly slopes.

Table 4.5: Average intercepts and slopes (t-statistics) from cross-sectional regressions of stock returns on book-to-market over three time periods: the 1963–1990 period used by FF92; the full period 1963–2015; and the period 1980–2015. Results from the LS regression of FF92 are shown in the rows labeled "LS (FF92)"; results from our LS regression are shown in the rows labeled "LS (GM)"; results from the LTS with 1% trimming regression of KR97 are shown in the rows labeled "LTS 1% (KR97)"; results from the LTS with 5% trimming regression of CCW04 are labeled "LTS 5% (CCW04)"; and results from our 99.9% efficient robust MM-regression are shown in the rows labeled "Robust (GM)". The analysis performed by CCW04 actually spans July 1963–December 2001. Average intercepts were not provided by FF92, KR97, or CCW04. Note that the t-statistics shown here are uncorrected for serial correlation.

		Book-to-Market		
Factor	Method	1963–1990	1963–2015	1980–2015
Intercept	LS (FF92)			
	LS (GM)	1.33 (4.22)	1.41 (6.15)	1.42 (5.25)
	LTS 5% (CCW04)			
	LTS 1% (KR97)			
	Robust (GM)	0.70 (2.31)	0.54 (2.53)	0.39 (1.58)
ln(BE/ME)	LS (FF92)	0.50 (5.71)		
	LS (GM)	0.47 (5.51)	0.39 (6.59)	0.37 (5.48)
	LTS 5% (CCW04)	0.39 (7.05)		
	LTS 1% (KR97)	0.48 (5.73)		
	Robust (GM)	0.44 (5.59)	0.30 (5.63)	0.30 (4.99)

Figure 4.10 shows t-statistics computed using only the regressions for each month during each of the periods 1963–2015 and 1980–2015. It is notable that the figures for the LS and robust regressions of returns on book-to-market are extremely similar, again indicating that the LS estimates are not overly influenced by outliers. We see that over the years 1963–2015 the book-to-market factor was strongly significant, at a Bonferroni-corrected 0.05 level,[14] from January to March for both the LS and robust methods, and also during July for the LS method only. This supports Loughran's conclusions about the importance of January for the book-to-market effect, but indicates that book-to-market was substantially important during March, and to a lesser extent, February.[15] However, when we omit the years 1963–1979, only the March effect persists.

The heatmap (Figure 4.8) and the plot of the time series of slopes (Figure 4.6) suggest that the large positive January slopes (indicated by the red dots) in the vicinity of 1975 are the cause of the January effect when using the entire 1963–2015 data set. After 1980 there are few such large January slopes, which presumably leads to the disappearance of the January effects in the month-by-month analysis. The implication of the latter observation is that the importance of whatever caused the strong January and February effects diminished after 1980. It is not clear why a March effect persists in the 1980–2015 period.

We will address the small stock question below in our discussion of the size and book-to-market model.

[14]When multiple hypotheses are tested at the same time for a given significance level α, the probability of rejecting one or more hypothesis purely by chance can be greater than α. The Bonferroni correction is a conservative means of addressing this issue: each of the n individual hypotheses are tested at the adjusted significance level α/n. This guarantees an overall false positive rate no greater than α for the entire set of tests. Further details can be found in standard textbooks on statistics or econometrics, for instance, Stock and Watson (2007).

[15]Since the size-beta portfolios are rebalanced every July 1, we suspect the "July" effect in the LS slopes may be induced by the data construction.

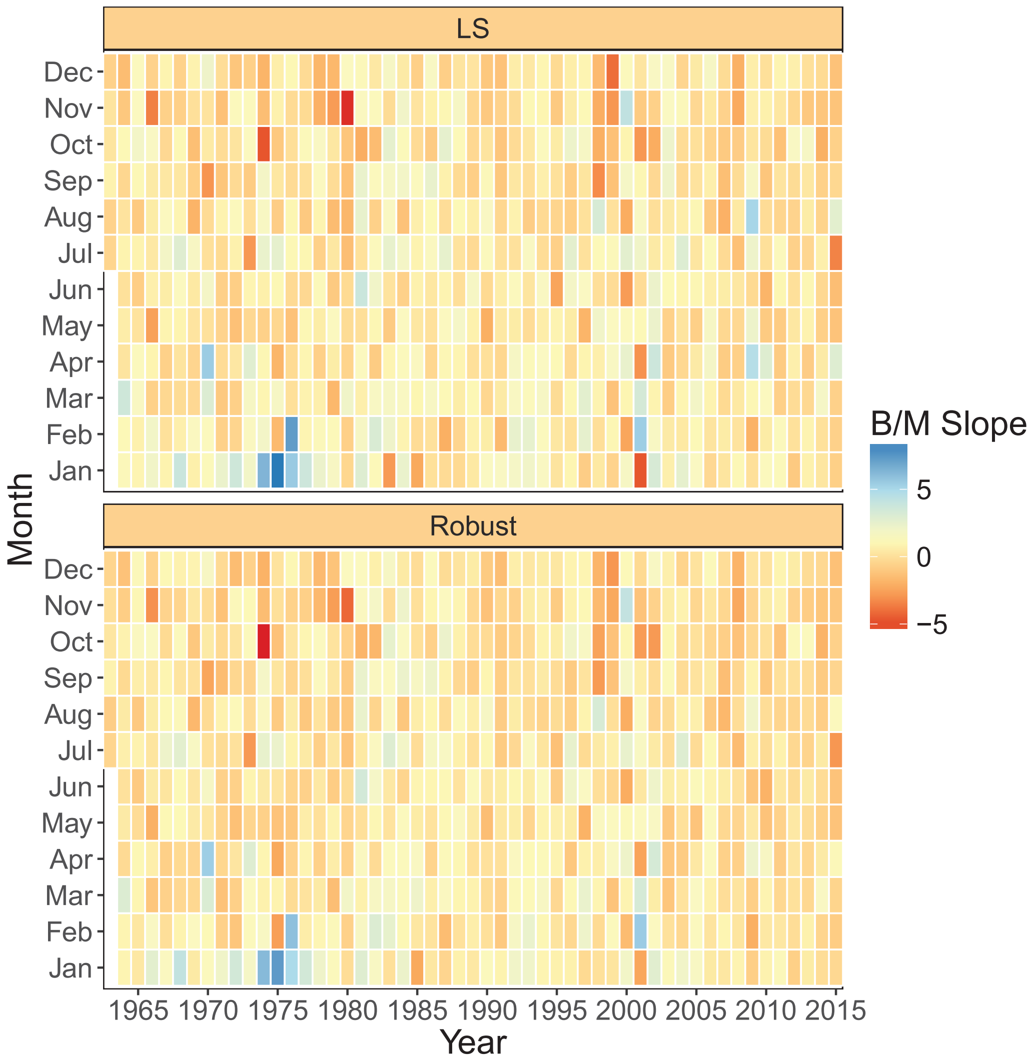


Figure 4.8: Heatmap of monthly slopes on book-to-market for LS and robust regressions. Each cell is colored according to the size of the slope at that particular month (row index) and year (column index).

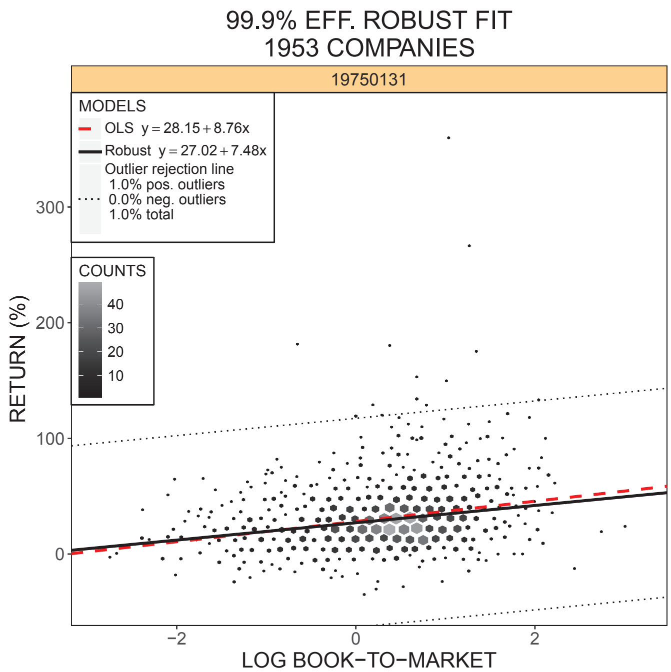


Figure 4.9: Hexbin scatterplot showing stock returns versus the logarithm of book-to-market values, January 1975. Nearby points are collapsed into hexagonal bins to avoid overplotting. Bin colors and sizes reflect the number of points falling into each bin. The LS (red dashed) and robust regression (black solid) lines for the regression of returns on the logarithm of book-to-market are also shown.

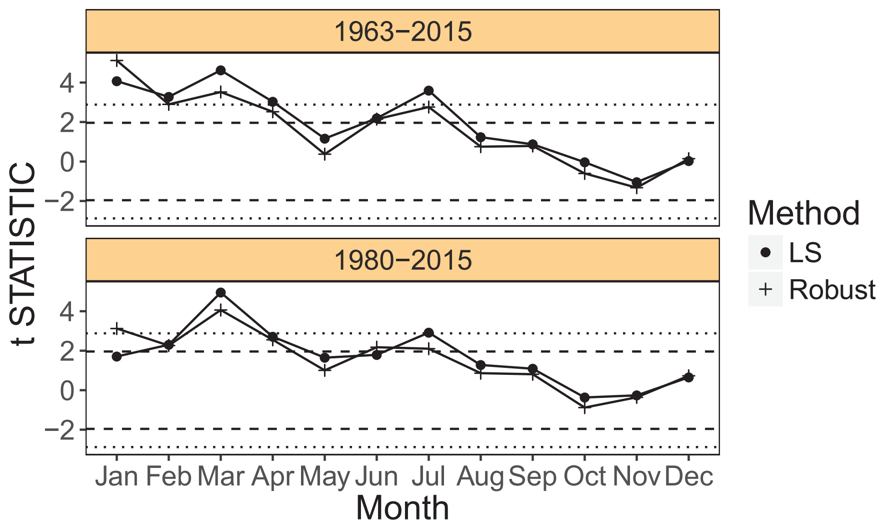


Figure 4.10: Monthly t-statistics for the book-to-market factor. t-statistics are calculated using only slopes from regressions with a given month. The dashed lines indicate the standard cutoff values for the 0.05 significance level (± 1.96), while the dotted lines indicate cutoff values for the Bonferroni-corrected nominal 0.05 significance level. Top panel: 1963–2015. Bottom panel: 1980–2015.

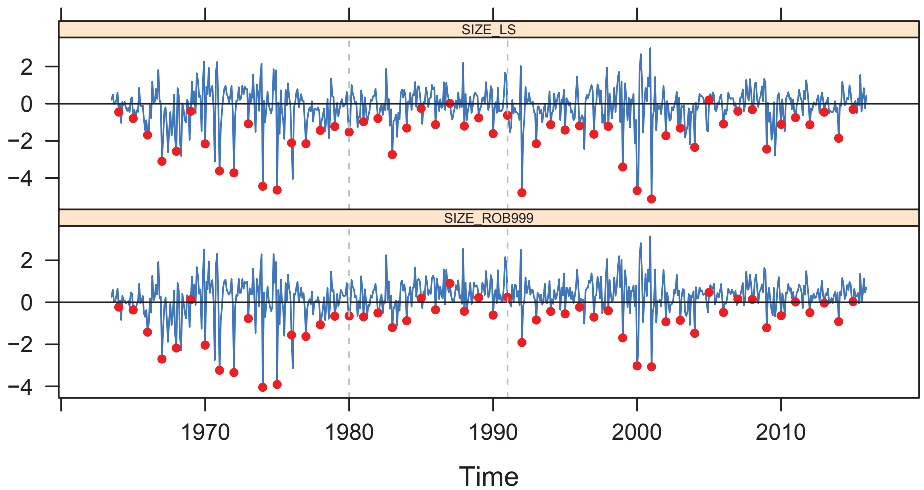


Figure 4.11: Time series of slopes from regression of returns on (June) size. The plot setup is identical to that of Figure 4.6.

4.4.2 Size

Figure 4.11 presents the time series of regression slopes for the size factor for the three time periods, and Table 4.6 presents the average slopes and t-statistics for each regression method. The LS regressions indicate that, on average, stock returns decrease with increasing firm size. As a consequence of this well-known result, it has been commonly accepted by practitioners that there is a negative relationship between average returns and firm size. Our robust regression results, on the other hand, indicate quite the opposite: returns are increasing with firm size for the majority of stocks and time periods. We are not the first to find this result: as shown in Table 4.6 both KR97 with 1% LTS regression and CCW04 with 5% LTS regression found a positive relationship between average returns and size over the periods 1963–1990 and 1963–2001, respectively. Together these results illustrate the consistency of statistically robust regression methods across choices of method and with respect to choice of parameters for a given method. (We explore this point further in Section 4.7.)

Figure 4.12 provides summary statistics (left panel) on the LS and robust slopes during the period 1963–2015, and a kernel density estimate of the distribution of the paired differen-

Table 4.6: Average intercepts and slopes and t-statistics from regressions of returns on size for the three periods. The CCW04 analysis covers the period July 1963–December 2001.

		Size		
Factor	Method	1963–1990	1963–2015	1980–2015
Intercept	LS (FF92)			
	LS (GM)	1.70 (3.57)	1.85 (5.10)	1.71 (4.05)
	LTS 5% (CCW04)			
	LTS 1% (KR97)			
	Robust (GM)	−0.32 (−0.72)	−1.04 (−3.42)	−1.77 (−5.45)
Size	LS (FF92)	−0.15 (−2.58)		
	LS (GM)	−0.13 (−2.33)	−0.14 (−3.45)	−0.10 (−2.25)
	LTS 5% (CCW04)	0.22 (5.79)		
	LTS 1% (KR97)	0.14 (2.63)		
	Robust (GM)	0.21 (4.01)	0.28 (8.44)	0.39 (11.48)

ces (right panel). We see that the LS slopes tend to be more negative than the robust slopes. The numerous large and negative LS slopes drive the average slope downward. A paired t-test of the LS and robust slopes would reject the hypothesis that the average difference between the slopes is 0 (with a t statistic of 17.8).[16]

Figure 4.13 shows hexbin scatterplots of the returns and size data at November 1998, as well as the LS (dashed) and robust (solid) regression lines at this time. The vertical lines indicate the size decile breakpoints in effect at this time. As a result of using only NYSE stocks to define the size deciles, nearly half of the stocks at this time are in the lowest size decile. The LS and robust regression approaches disagree about the sign of the slope on size for this month. As noted in the figure legend, 3.3% of all stocks have been rejected (as

[16]The non-parametric Wilcoxon signed rank-test, which is robust to outliers, also rejects the hypothesis that the difference between the LS and robust slopes has mean 0 ($W = 569180$, $p \approx 0$).

Summary Statistics for Size Slopes

	LS	Robust
Minimum	-5.12	-4.04
1st Quartile	-0.51	-0.04
Median	-0.03	0.33
Mean	-0.14	0.28
3rd Quartile	0.43	0.71
Maximum	2.99	3.14

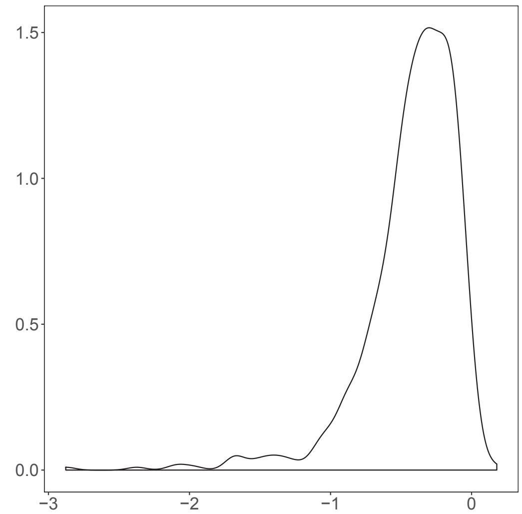


Figure 4.12: Statistics on monthly LS and robust slopes on size, 1963–2015. Left panel: summary statistics for each series. Right panel: kernel density estimate of the distribution of paired differences (LS − robust) in monthly slopes. A paired t-test of the LS and robust slopes rejects the hypothesis that the average difference is 0 with a t statistic of 17.8.

indicated by the parallel dotted lines in the figure) by our robust regression. Almost all of these rejected "outliers" lie in the lowest size decile and have monthly returns in excess of 50%. In the LS regression these observations bias the slope on size downward, leading us to believe that small stocks should earn higher returns on average than large stocks. The robust regression, on the other hand, ignores these unusual firms and consequently finds that average returns increase with size for most firms.

The unusually high returns exhibited by the rejected firms are generally a "small firm" phenomenon, it being much easier to go from \$2 per share to \$3 than from \$20 to \$30. This is an effect that investors would seldom rely upon for similar future gains, since it is unlikely that such stocks would continue to experience such large percentage-wise returns. Correspondingly, returns forecasts based on the LS fit would be overly optimistic for small firms. In summary, the robust regression of returns on size is not much influenced by the occasional huge returns of these small size firms and hence accurately describes how the cross

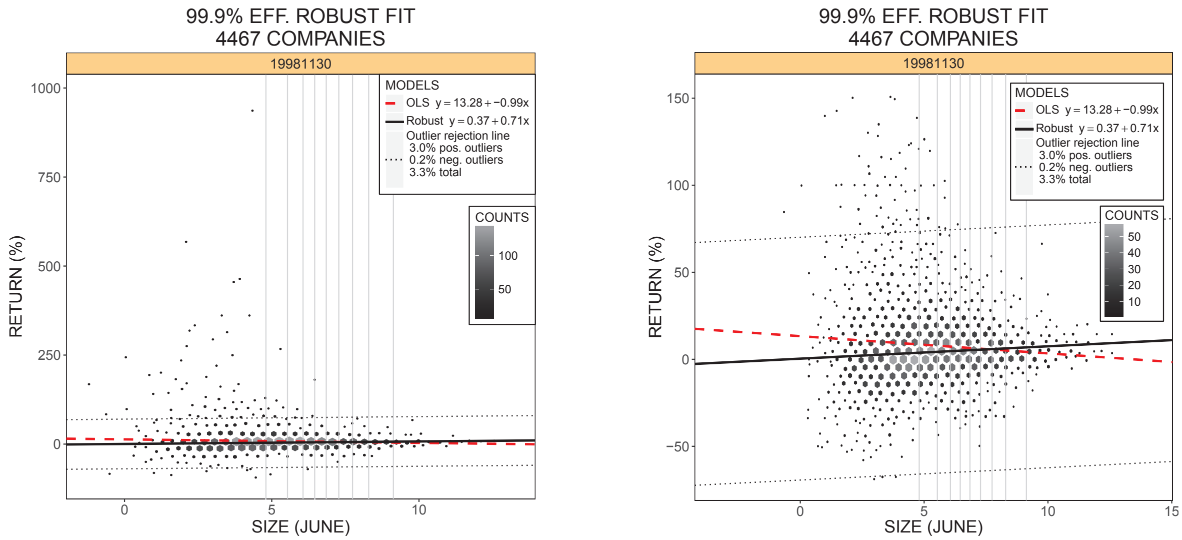


Figure 4.13: Hexbin scatterplots showing stock returns versus size values, November 1998. The plot setup follows that of Figure 4.9. The left plot shows the full range of returns values. In the right plot, the vertical range has been trimmed by 0.1% on each side to show more detail in the main point cloud. The additional grey vertical lines in these plots indicate the size decile breakpoints effective at this time. Roughly 48% of the stocks at this time fall into the lower decile.

section of returns varies with size for the vast majority of the firms.

Next, we investigated the "January effect" for size, i.e., that the explanatory power for the size factor is largely due to returns in January. Figure 4.14 shows a heatmap of the LS and robust slopes by month and by year. The chart makes the "January effect" quite obvious: January slopes for both the LS and robust regressions tend to be negative, with some particularly negative slopes around the time of the dot-com bubble. What is striking about the visualization, however, is how overwhelmingly positive the robust slopes are outside of January. This suggests that the average robust slopes will be positive outside of January, and may be significant.

We formally investigated these calendar effects by computing t-statistics for slopes in each month over the periods 1963–2015 and 1980–2015. Figure 4.15 shows the monthly t-statistics from the LS and robust regressions. Both LS and robust regressions (solid dots and plus

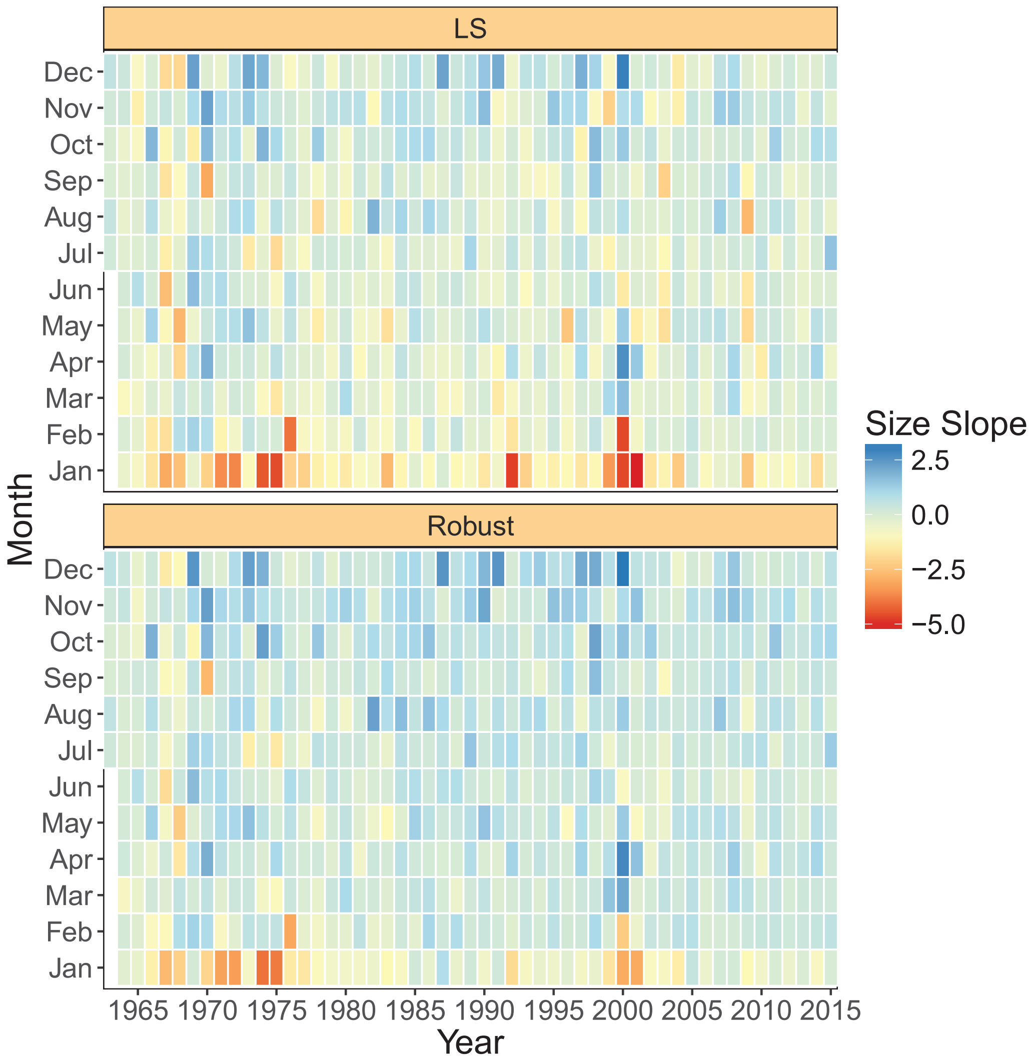


Figure 4.14: Heatmap of monthly slopes on size for LS and robust regressions.

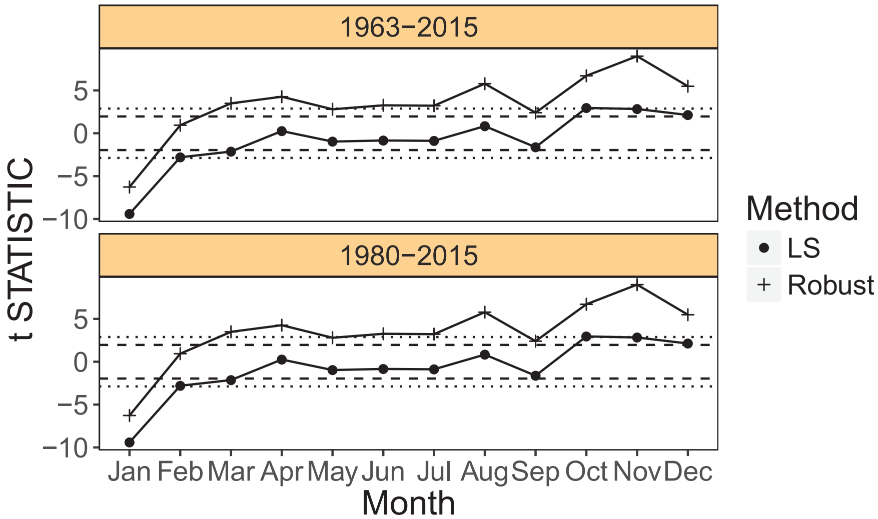


Figure 4.15: Monthly t-statistics for the size factor. The t-statistics are calculated using only slopes from regressions with a given month. The dashed lines indicate the standard cutoff values for the 0.05 significance level (± 1.96), while the dotted lines indicate cutoff values for the Bonferroni-corrected nominal 0.05 significance level. Top panel: 1963–2015. Bottom panel: 1980–2015.

signs, respectively, in the figure) find a strong, negative, January effect (and a borderline February effect) using a Bonferroni-corrected 0.05 significance level. This agrees with what we observed in Figure 4.14, with the results presented in Panel B of Tables 1 and 2 of CCW04 for LS regressions, and with the accepted notion of the January effect for size.

However, the LS and robust results for the rest of the year are quite different. According to the LS regressions, size is significant only in the month of January. In stark contrast to the LS result, our robust regression (plus signs in Figure 4.15) finds that size is significant in nearly all months. According to the robust regressions, the relationship between average returns and size is negative in January, but turns positive by the start of the second quarter of the year, with a strong positive relationship at the end of the year. CCW04 also found that there was a significant and positive relationship between returns and size in non-January months using 5% LTS (presented in Table 2 of that paper). Our robust regression approach confirms their results through 2015 and provides additional insight into the month-to-month dynamics of the size effect.

Our result that average returns increase with increasing size for nearly all stocks is at odds with much of the literature, which documents a negative relationship between returns and size via cross-sectional LS regressions and via returns on equally-weighted portfolios of stocks grouped by size decile. Indeed, as we show below, equally-weighted portfolios of stocks in our data set grouped by size deciles, whether these deciles are defined by NYSE stocks only or by all stocks, exhibit the same decreasing trend in returns with increasing size that is found in nearly every other study of the size effect. The key to reconciling our robust results with the accepted view of the size effect lies in the nature of the robust regression: since a small number of outlying stocks are rejected in each monthly cross-sectional regression, we need to consider a similar approach to forming size decile portfolios.

The return on an equally-weighted portfolio is simply the sample mean of the returns of its constituents. We can define a "robust" mean using our robust regression estimate simply by fitting a model containing only an intercept to the stock returns in a given combination of month and size decile. The intercept term is then a robust estimate of the "center" of the distribution of the returns within this combination of month and size decile. Furthermore, the regression produces observation weights (determined by the derivative $\psi(x)$ of the underlying loss function): non-outlying stocks have weight 1, extremely outlying stocks have weight 0, and moderately outlying stocks have a weight somewhere between 0 and 1. After normalizing the weights to sum to 1, we obtain a "robust mean" portfolio that omits stocks with returns that are outlying (in this month and size decile). The efficiency of the regression will again control how many stocks are flagged as outliers and hence omitted from the portfolio.

Table 4.7 presents the time series average monthly returns on the equally-weighted portfolio and the "robust mean" portfolio using the 99.9% efficient robust MM-estimate we have been using throughout our analysis. We further enhance this discussion by including a "robust mean" portfolio formed with 99.99% efficiency, and a "trimmed mean" portfolio formed using the 1% trimmed mean return for each monthly size group. The former portfolio rejects fewer outliers than the 99.9% efficient one, while the latter portfolio places equal weight on stocks remaining after excluding stocks with the top and bottom 1% of returns for that month and size group. Figure 4.16 presents a graphical summary of the average returns in Table 4.7 for the period 1963–2015. In addition to the NYSE-based size deciles we have been using so far, we also consider size deciles based on the entire stock market each June (subject to the inclusion conditions described in Section 4.3). The latter deciles correspond more closely to nature of the size factor as used in the cross-sectional regressions.

As we would expect, the average returns on the equally-weighted portfolios decline with increasing size, regardless of time period or how the deciles are defined. The average returns on the 99.9% efficient robust mean portfolio, on the other hand, tend to increase with increasing size decile, in line with our regression results. With the NYSE-defined deciles, there is a sharp increase in average returns from the smallest stocks to moderately-sized stocks, but afterwards returns on the robust mean portfolio are rather flat. The all-market deciles exhibit a much stronger positive relationship for the 99.9% efficient robust mean portfolio. We continue to see this pattern in the 99.99% efficient robust mean portfolio, which excludes even fewer stocks each month. This tells us, once again, that the negative relationship exhibited via equally-weighted portfolio returns and via the cross-sectional LS regression is driven by a small number of stocks each month.

The differences between the trends in the portfolio returns are clearly driven by the bottom half of the stocks by size. The greatest differences in average returns appear in the smallest decile. If we consider only stocks in deciles 2-10 (regardless of decile definition), returns on the equally-weighted portfolio are much less negative in trend, and returns on the 1% trimmed mean portfolio are fairly flat across size deciles. As firm size increases, the average returns of the four portfolios agree rather closely, which suggests the returns for larger deciles are less influenced by extreme monthly returns.[17]

[17] Another interesting observation from this analysis is that average returns in the largest two size decile portfolios are slightly lower than the returns from the middle deciles.

Table 4.7: Time series average returns, in percent, on portfolios formed on size deciles over the periods 1963–1990, 1963–2015, and 1980–2015. Size deciles are calculated using either only NYSE stocks or using all stocks (subject to the inclusion criteria stated in Section 4.3). Returns for four portfolio weighting schemes are shown: equal weight on each stock in a given size decile in a given month; equal weight on stocks remaining after removing the smallest and largest 1% of returns within a month-size decile combination ("Trim 1%"); weights calculated using 99.9% efficient robust regression; and weights calculated using 99.99% efficient robust regression.

		Size Deciles									
Deciles	Weighting	ME01	ME02	ME03	ME04	ME05	ME06	ME07	ME08	ME09	ME10
						1963–1990					
NYSE	Equal Wgt.	1.44	1.15	1.24	1.22	1.23	1.11	1.10	1.08	0.94	0.89
	Trim 1%	1.09	1.04	1.15	1.15	1.17	1.07	1.07	1.04	0.90	0.88
	Robust 99.9%	0.45	0.78	0.94	0.99	1.00	0.95	0.99	0.99	0.85	0.85
	Robust 99.99%	0.65	0.89	1.02	1.04	1.05	1.00	1.02	1.01	0.87	0.86
ALL	Equal Wgt.	2.12	1.42	1.24	1.13	1.10	1.11	1.10	1.12	1.06	0.94
	Trim 1%	1.66	1.13	1.01	0.94	0.93	1.02	1.00	1.04	1.01	0.90
	Robust 99.9%	0.14	0.33	0.45	0.55	0.62	0.79	0.82	0.92	0.94	0.87
	Robust 99.99%	0.49	0.57	0.64	0.69	0.74	0.88	0.89	0.98	0.97	0.89
						1963–2015					
NYSE	Equal Wgt.	1.50	1.14	1.20	1.18	1.20	1.11	1.12	1.06	1.00	0.89
	Trim 1%	1.01	1.00	1.10	1.10	1.14	1.06	1.08	1.03	0.96	0.88
	Robust 99.9%	0.13	0.67	0.87	0.94	0.97	0.92	0.98	0.96	0.89	0.85
	Robust 99.99%	0.35	0.79	0.94	0.99	1.02	0.96	1.01	0.98	0.91	0.86
ALL	Equal Wgt.	2.30	1.41	1.35	1.25	1.19	1.14	1.12	1.11	1.09	0.96
	Trim 1%	1.54	1.01	1.03	1.00	0.98	1.01	1.01	1.04	1.03	0.92
	Robust 99.9%	-0.39	-0.01	0.26	0.45	0.57	0.73	0.79	0.90	0.94	0.88
	Robust 99.99%	-0.02	0.25	0.48	0.61	0.71	0.82	0.87	0.94	0.97	0.89

Table 4.7: *(continued)*

Deciles	Weighting	Size Deciles ME01	ME02	ME03	ME04	ME05	ME06	ME07	ME08	ME09	ME10
						1980–2015					
NYSE	Equal Wgt.	1.39	1.08	1.17	1.14	1.19	1.18	1.21	1.13	1.14	1.03
	Trim 1%	0.80	0.91	1.05	1.07	1.13	1.13	1.16	1.10	1.09	1.01
	Robust 99.9%	-0.25	0.56	0.83	0.91	0.97	1.00	1.04	1.02	1.02	0.99
	Robust 99.99%	-0.03	0.67	0.89	0.95	1.00	1.02	1.07	1.04	1.03	0.99
ALL	Equal Wgt.	2.28	1.28	1.25	1.16	1.11	1.06	1.08	1.12	1.17	1.09
	Trim 1%	1.29	0.80	0.87	0.87	0.88	0.91	0.97	1.04	1.11	1.05
	Robust 99.9%	-1.16	-0.44	-0.04	0.23	0.39	0.61	0.74	0.89	0.99	0.99
	Robust 99.99%	-0.75	-0.16	0.19	0.40	0.54	0.71	0.81	0.93	1.02	1.01

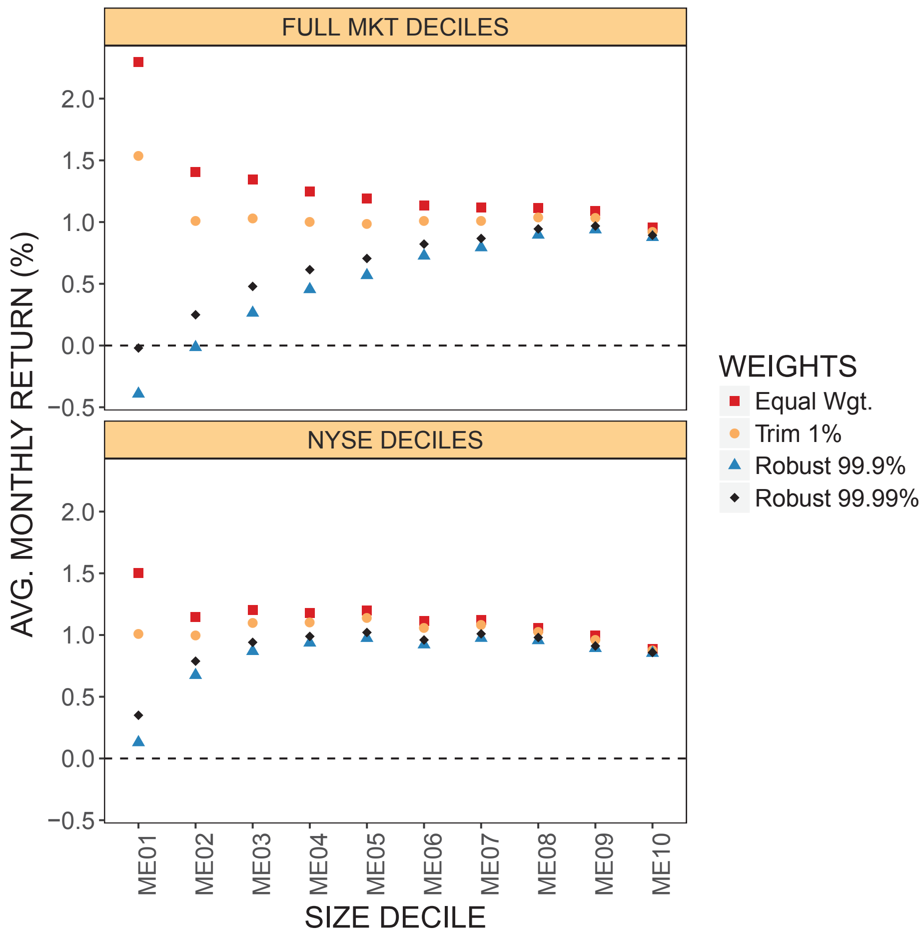


Figure 4.16: Time series average monthly returns, in percent, on size decile portfolios for the four portfolio weights and two decile definitions over the period 1963–2015. The data presented here also appears in the middle section of Table 4.7.

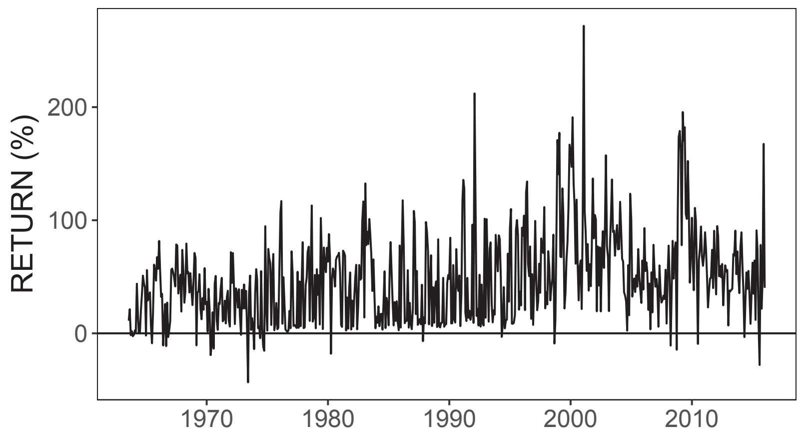


Figure 4.17: Time series of the monthly returns, in percent, on the outlier portfolio constructed using the 99.9% robust mean portfolio within the smallest NYSE decile ("ME01").

As for the stocks excluded each month by the 99.9% efficient robust mean, Table 4.8 shows the time series average return on an equally-weighted portfolio of the rejected stocks in each combination of month and size decile. Figure 4.17 shows the time series of monthly returns for the rejected stock portfolio formed using the smallest NYSE size decile. The smaller deciles exhibit very high average returns and numerous months with double-digit returns, which further confirms that the large returns on the first decile equally-weighted portfolios in Table 4.7 are driven by outliers. Note that this return would be difficult to achieve in practice, as predicting which stocks will earn such large returns is difficult, and transaction costs for trading such small stocks may be prohibitively high.

Figure 4.18 presents histograms of the percentage of stocks rejected each month over the period 1963–2015 within each size group for the 99.9% robust mean portfolio with the NYSE deciles. For size deciles 2–10, about 1-1.5% of stocks are rejected each month, on average. In the smallest deciles, 3.8% of stocks are rejected on average, but in some months the percentage can be quite higher. Figure 4.19 plots the percentage of stocks rejected over time for the smallest NYSE size decile. There is a sharp uptick in the number of outliers detected in the late 1970s, and a longer period from the mid-1980s until the mid-1990s with

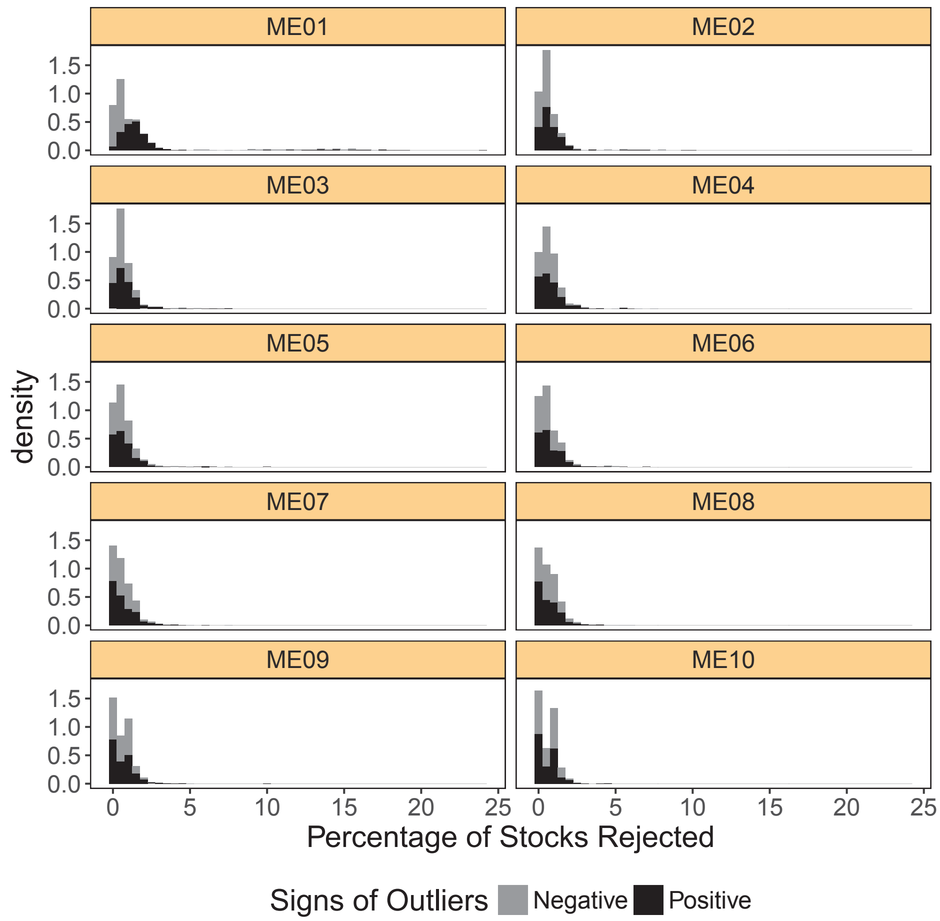


Figure 4.18: Histograms of the percentage of stocks rejected each month by the 99.9% robust mean portfolio with NYSE deciles over the period 1963–2015. Relative frequencies for positive and negative outliers are shown in black and gray, respectively.

a large number of outliers and high month-to-month volatility in the percentage of outliers. This behavior suggests the existence of clusters of small stocks during these periods with very different returns from the majority of stocks. It is further interesting that frequency of outliers drops back down to its pre-1975 average level after 2000.

It is important to point out that our robust mean portfolio is not investible, as we do not know the stock returns ahead of time and thus do not know which stocks will have outlying returns. Similarly, we cannot invest in the portfolio of monthly outliers which is seemingly

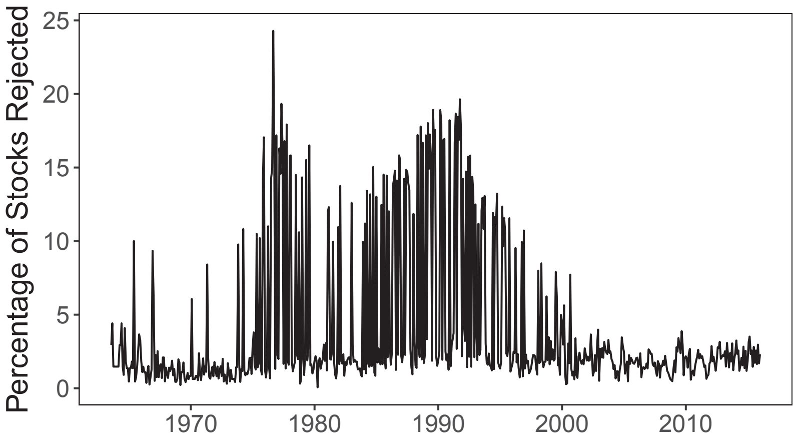


Figure 4.19: Time series of the percentage of stocks rejected each month by the 99.9% robust mean portfolio within the smallest NYSE decile ("ME01").

responsible for the "smallcap premium" implied by LS regressions and equally-weighted portfolios. Harvesting the smallcap premium seemingly requires either extraordinary skill in forecasting which small firms will experience explosive growth in their stock price, or the ability to hold the entire small stock market. These stocks tend to be illiquid, thinly traded, and capacity-constrained, however, which would make holding the entire market difficult.

Nonetheless, the robust mean portfolios reinforce our conclusion from the cross-sectional regressions: the apparent negative relationship between average returns and size that we observe when all stocks are used in the analysis is an artifact of a small number of stocks with unusually large (positive or negative) returns. For most stocks, average returns increase with increasing firm size.

Finally, the portfolio analysis above suggests we examine the impact of the smallest stocks on our regression results. Tables 4.9 and 4.10 show how our LS and robust regression results change if we omit, for each monthly cross-section, all stocks falling into the bottom 5th and 10th size percentiles as defined by the NYSE breakpoints. After removing the bottom 5% of stocks each month by size, the average LS slope on size is no longer significant in any time period. The average robust slope is no longer significant for the 1963–1990 period if small stocks are removed, but it remains positive and significant for the longer period 1963–2015 and the 1980–2015 period. The LS and robust results suggest that for the 1963–1990 period the size effect, regardless of its direction, was driven entirely by small firms. After 1980, though, a positive relationship between average returns and firm size exists for nearly all firms and is not confined to the smallest firms.

Table 4.8: Time series average returns, in percent, on equally-weighted portfolios of outliers rejected by the robust means within size deciles over the periods 1963–1990, 1963–2015, and 1980–2015. For each month and size decile combination, we calculate the return on an equally-weighted portfolio of stocks identified as outliers via the robust mean. Note that in some month/size decile combinations there may be a single outlier or no outliers at all. If there are no rejected stocks for a given month/size decile, we define the return for that month as 0.

		Size Deciles									
Deciles	Weighting	ME01	ME02	ME03	ME04	ME05	ME06	ME07	ME08	ME09	ME10
					1963–1990						
NYSE	Robust 99.9%	34.46	16.44	13.02	10.12	9.82	5.70	4.52	3.95	4.24	2.32
	Robust 99.99%	34.88	13.54	11.74	8.84	7.79	4.09	3.70	3.63	3.62	1.83
ALL	Robust 99.9%	18.35	26.31	22.90	19.49	18.71	13.66	15.00	11.60	8.21	6.15
	Robust 99.99%	16.72	23.76	19.91	16.59	17.21	11.19	13.15	10.53	7.29	5.53
					1963–2015						
NYSE	Robust 99.9%	48.22	22.45	16.17	11.06	9.60	8.06	5.79	4.49	4.86	2.08
	Robust 99.99%	53.65	21.09	15.79	10.19	8.05	6.78	5.37	4.29	4.33	1.90
ALL	Robust 99.9%	44.54	37.46	31.81	27.46	24.68	18.59	17.24	11.90	10.00	6.17
	Robust 99.99%	47.20	37.84	31.81	26.46	24.30	17.28	16.32	12.03	9.70	5.88
					1980–2015						
NYSE	Robust 99.9%	55.82	26.57	18.11	11.82	10.65	9.07	7.16	5.22	6.16	2.63
	Robust 99.99%	64.82	26.13	18.58	11.07	9.23	7.80	6.89	5.10	5.66	2.52
ALL	Robust 99.9%	57.44	46.22	38.77	32.69	28.53	21.58	19.16	13.72	12.00	7.84
	Robust 99.99%	62.65	48.56	40.69	32.51	28.87	20.93	19.01	14.73	11.93	7.86

Table 4.9: Average slopes and t-statistics from regressions of returns on size excluding lower 5th percentile for the three periods.

		Size (Excluding Lower 5th Percentile)		
Factor	Method	1963–1990	1963–2015	1980–2015
Intercept	LS	1.26 (2.48)	1.33 (3.27)	1.06 (2.14)
	Robust	0.33 (0.67)	−0.02 (−0.06)	−0.45 (−1.03)
Size	LS	−0.05 (−0.95)	−0.05 (−1.23)	0.01 (0.20)
	Robust	0.08 (1.64)	0.12 (3.44)	0.19 (4.84)

Table 4.10: Average slopes and t-statistics from regressions of returns on size excluding lower decile for the three periods.

		Size (Excluding Lower Decile)		
Factor	Method	1963–1990	1963–2015	1980–2015
Intercept	LS	1.36 (2.65)	1.37 (3.31)	1.13 (2.22)
	Robust	0.54 (1.07)	0.26 (0.68)	−0.08 (−0.17)
Size	LS	−0.07 (−1.25)	−0.05 (−1.39)	0.00 (0.00)
	Robust	0.05 (0.94)	0.08 (2.25)	0.14 (3.41)

4.4.3 Beta Factors

Figure 4.20 presents the time series of regression slopes for the beta factor for the three time periods. It is interesting to note that there is considerable volatility in the slopes in the early to mid-1970s, with some particularly large slopes in the months of January. The longer time series (bottom panel) exhibits even larger January slopes around the end of the dot-com bubble, with large fluctuations in slopes before and after that period.[18]

Table 4.11 presents the average slopes and t-statistics for each regression method and shows once again that our LS results for 1963–1990 are consistent with those of FF92 in indicating that beta does not explain the cross-section of returns. The robust regression, however, tells us something quite different and perhaps a bit shocking, namely that returns decrease (increase) with increasing (decreasing) beta for the 1963–1990 period as well as the other two periods, each of which has a larger t-statistic value than the 1963–1990 period.[19] This negative relationship between returns and beta portfolios is consistent with the low beta anomaly reported by Black (1972), Black et al. (1972), Ang et al. (2006), and many others.

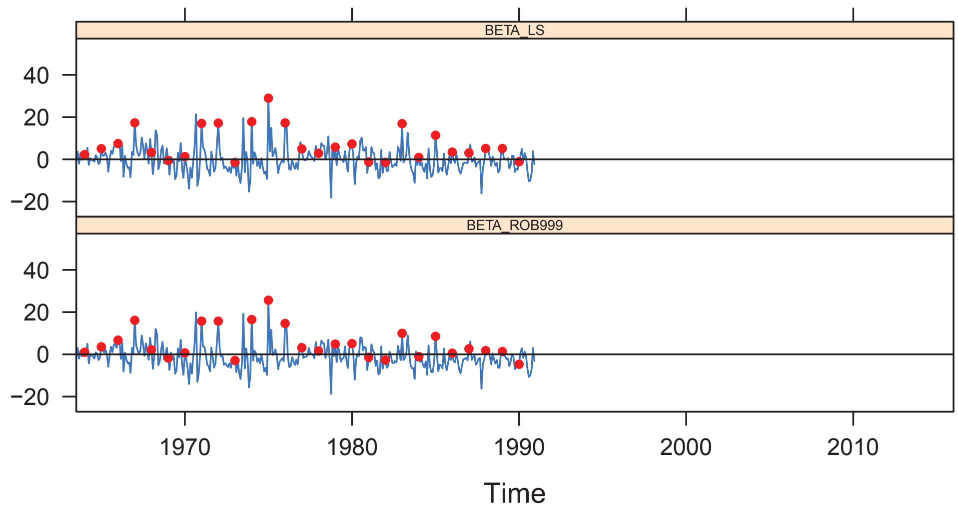
The distribution of robust slopes on beta during the period 1963–2015 tends to be shifted in a more negative direction relative to the distribution of LS slopes (Figure 4.21). A paired t-test of the LS and robust slopes would reject the hypothesis (at the 0.05 level) that the average difference between the slopes is 0 (with a t statistic of 3.23).[20] Moreover, when both

---

[18]Since the post-ranking betas are computed by regressing size-beta portfolio returns on market returns for the entire time period for each of the time periods 1963–1990 and 1963–2015, these two time periods can have a different set of post-ranking betas. Thus unlike the other factors, the results for regressing returns on beta portfolios for 1963–1990 may be different than those for the 1963–1990 subperiod of 1963–2015. In order to determine how different the results can be we provided two separate time series charts. A comparison the upper pair of time series plots with the lower ones for the period 1963–1990 suggests the effect of having possibly different post-ranking betas is inconsequential to our conclusions about the role of beta.

[19]We also analyzed this relationship using the robust MM-regression with 95% and 99% efficiency, and found similar results. Details are provided in Section 4.7. Furthermore, using 99.9% efficient robust pre- and post-ranking betas in the analysis does not change our results in a meaningful way.

[20]The Wilcoxon signed rank-test also rejects the hypothesis that the difference between the LS and robust slopes has mean 0 ($W = 450450$, $p = 0.00004$).



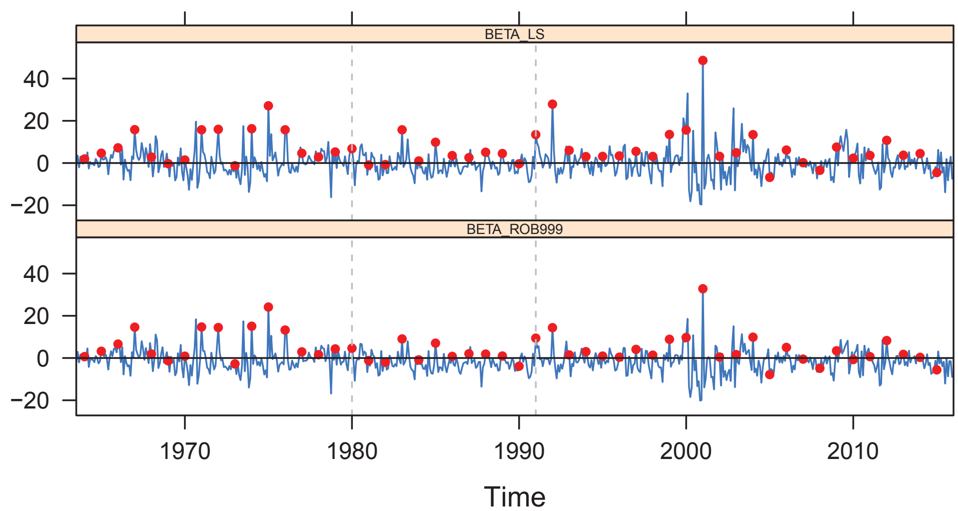


Figure 4.20: Results of regression of returns on post-ranking betas. Top panel: Time series of slopes computed using post-ranking betas from the 1963–1990 data set. The LS series are shown in the top half of the chart, and the robust series are shown in the bottom half. The plot setup is identical to that of Figure 4.6. Bottom panel: Time series of slopes computed using post-ranking betas from the 1963–2015 data set.

Table 4.11: Average intercepts and slopes and t-statistics from regressions of returns on post-ranking betas for the three periods. The CCW04 analysis covers the period July 1963–December 2001.

Factor	Method	Beta 1963–1990	1963–2015	1980–2015
Intercept	LS (FF92)			
	LS (GM)	1.19 (4.96)	1.07 (5.07)	1.36 (4.94)
	LTS 5% (CCW04)			
	Robust (GM)	2.01 (8.78)	2.11 (11.71)	2.57 (11.33)
Beta	LS (FF92)	0.15 (0.46)		
	LS (GM)	0.05 (0.14)	0.16 (0.64)	−0.08 (−0.25)
	LTS 5% (CCW04)	−0.40 (−1.73)		
	Robust (GM)	−1.08 (−3.47)	−1.25 (−5.72)	−1.70 (−6.66)

slopes are positive, the LS slope can be larger than the robust slope by as much as 15.8 (for January 2001), and there are several months (during the dot-com bubble, the 2009 market rebound after the 2008 financial crisis, and January 1992 again) where the difference is larger than 5.

Figure 4.22 shows a heatmap of the monthly LS and robust slopes. There is evidence of a positive January effect, possibly driven by a handful of large positive outliers in 1975 and the end of the dot-com bubble. Otherwise, the slopes appear to be generally negative or close to zero. Figure 4.23 shows monthly t-statistics for slopes on post-ranking beta for the periods 1963–2015 and 1980–2015. The t-statistics for both LS and our robust regressions confirm the existence of strong, positive effects for beta during the month of January during both time periods. Outside of January, the LS slopes are not significant, and would lead us to conclude that beta only matters in January. The t-statistics for the robust slopes, however, confirm a significant negative relationship between average returns and beta in

Summary Statistics for Beta Slopes

	LS	Robust
Minimum	-19.68	-20.18
1st Quartile	-3.53	-4.24
Median	-0.25	-1.36
Mean	0.16	-1.25
3rd Quartile	3.07	1.44
Maximum	48.63	32.86

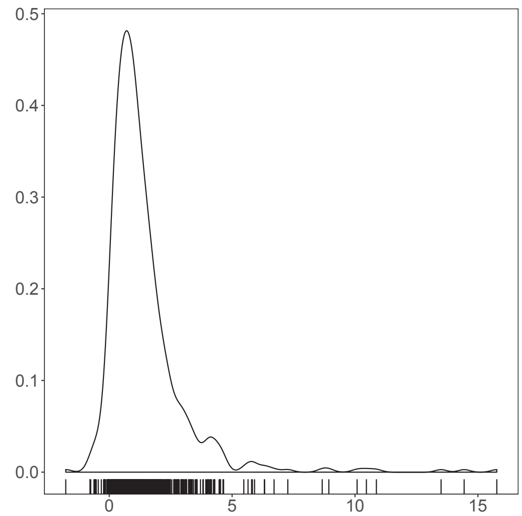


Figure 4.21: Statistics on monthly LS and robust slopes on beta, 1963–2015. Left panel: summary statistics for each series. Right panel: kernel density estimate of the distribution of paired differences (LS − robust) in monthly slopes. A paired t-test of the LS and robust slopes rejects the hypothesis that the average difference is 0 with a t statistic of 3.2.

several months over both time periods, particularly at the ends of calendar quarters. The latter finding agrees with the results obtained by CCW04 using LTS with 5% trimming for non-January months during the period 1963–2001. Hence, the relationship between average returns and firm betas is negative most of the year for most firms, and positive during the month of January for all firms.

Figure 4.24 shows scatterplots of the returns and post-ranking beta cross-section data for November 2000–February 2001, as well as the LS and robust regression lines.[21] The November, December, and February cross-sections show very few outliers, so the LS and robust regressions yield similar models. On the other hand, in the January 2001 cross-section, a small number of stocks in the largest beta portfolios have very large returns that result in the LS regression line slope being much larger than that of the robust line. Such

[21]The "columnar" nature of the scatterplot is due to the construction of the post-ranking beta values: there are only 100 unique post-ranking beta values for the 1963–1990 and 1963–2015 periods. The binning of nearby values into hexagons further conceals the individual beta columns.

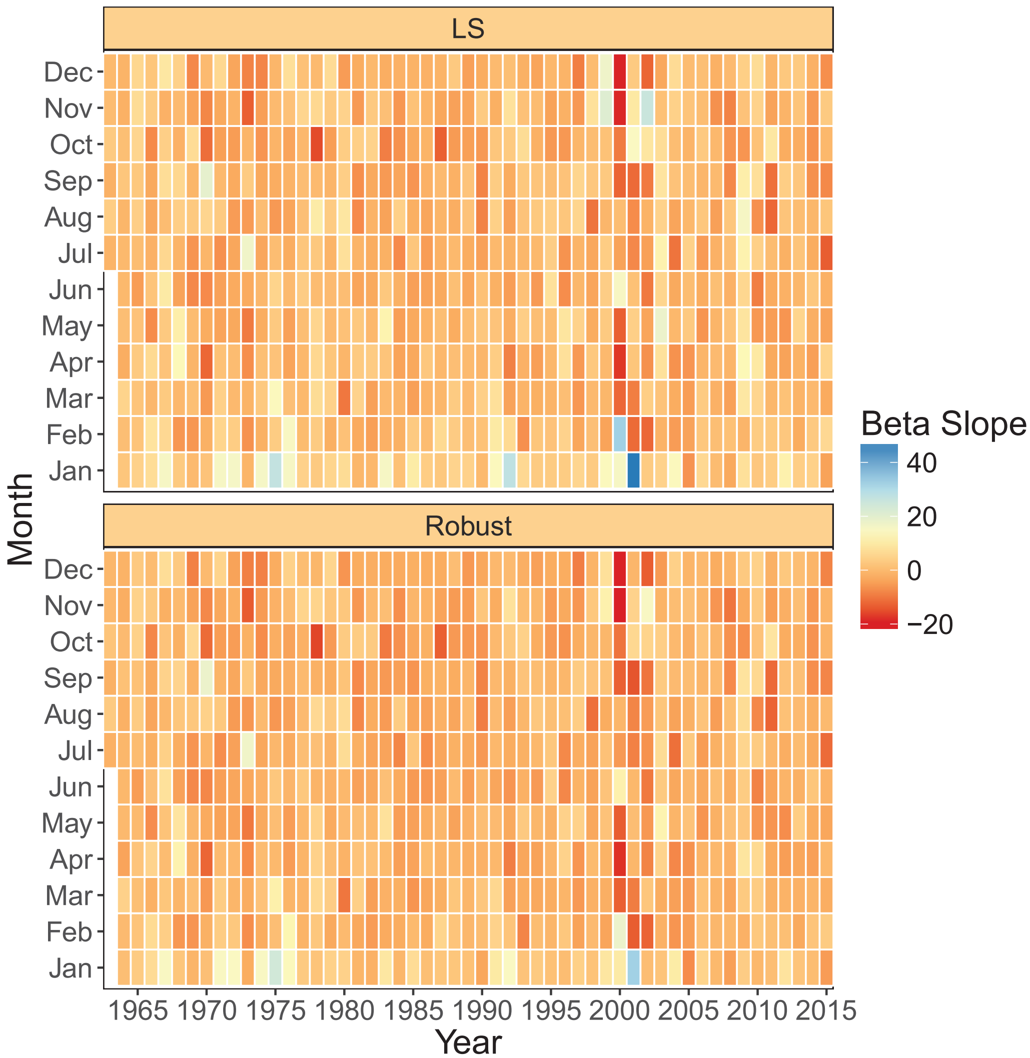


Figure 4.22: Heatmap of monthly slopes on beta for LS and robust regressions.

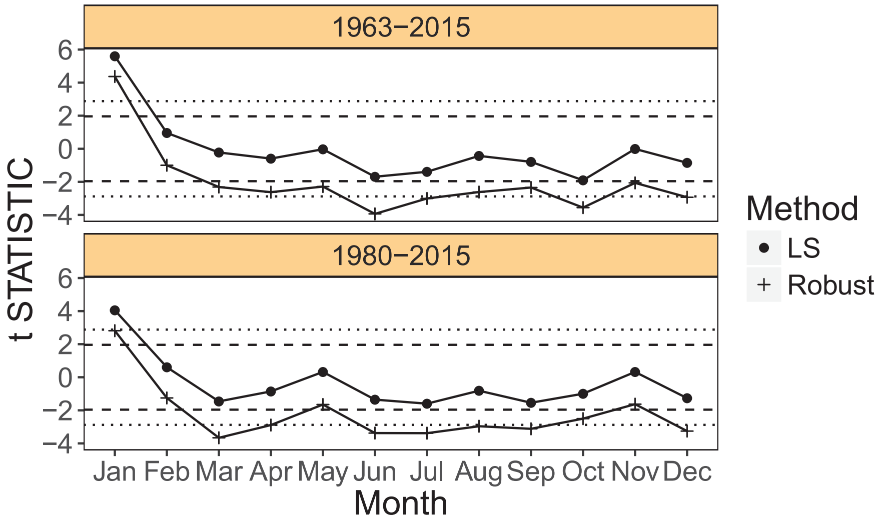


Figure 4.23: Monthly t-statistics for the post-ranking beta factor. t-statistics are calculated using only slopes from regressions with a given month. The dashed lines indicate the standard cutoff values for the 0.05 significance level (± 1.96), while the dotted lines indicate cutoff values for the Bonferroni-corrected nominal 0.05 significance level. Top panel: 1963–2015. Bottom panel: 1980–2015.

combinations of large values of beta portfolios and large returns act as leverage points with respect to their influence on the LS estimate. That being said, the sequence of scatterplots clearly shows how different the return-beta relationship is in January from the rest of the year. The switch to a positive relationship, and ensuing return to a negative relationship, do not appear to be explained by outliers or other data oddities. At least for the dot-com period, unusually large returns on high beta stocks for the month of January seems to be the culprit.

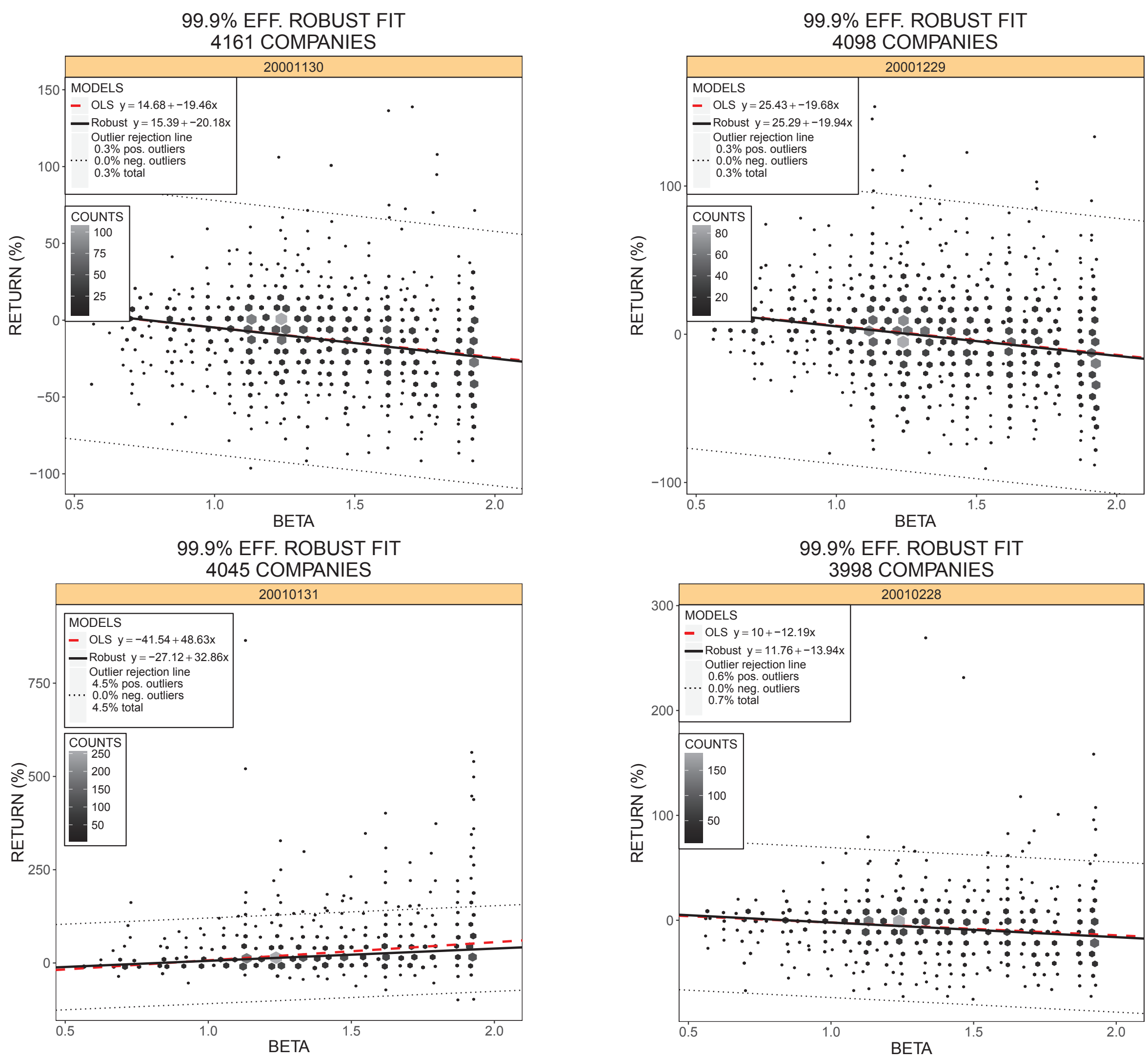


Figure 4.24: Hexbin scatterplots showing stock returns versus post-ranking beta values, November 2000–February 2001. Clockwise from upper left: November 2000, December 2000, February 2001, and January 2001. The plot setup is the same as in Figure 4.9. No trimming is applied to the vertical axis.

Table 4.12: Average intercepts and slopes and t-statistics from regressions of returns on earnings for the three periods.

		Earnings-to-Price		
Factor	Method	1963–1990	1963–2015	1980–2015
Intercept	LS (FF92)			
	LS (GM)	0.84 (2.44)	0.99 (4.33)	1.07 (4.08)
	Robust (GM)	0.42 (1.27)	0.63 (2.90)	0.74 (2.99)
E/P Dummy	LS (FF92)	0.57 (2.28)		
	LS (GM)	0.52 (2.14)	0.37 (2.06)	0.03 (0.15)
	Robust (GM)	−1.07 (−5.00)	−1.43 (−9.80)	−1.94 (−13.02)
E+/P	LS (FF92)	4.72 (4.57)		
	LS (GM)	5.78 (4.96)	3.73 (5.08)	2.40 (3.39)
	Robust (GM)	5.58 (5.38)	2.58 (4.26)	0.65 (1.41)

4.4.4 Earnings Factors

Figure 4.25 and Table 4.12 present the regression results for the two earnings factors, the negative earnings indicator (E/P Dummy) and positive-earnings-to-price (E+/P). The LS and robust regressions both agree that negative earnings had a significant relationship to returns during the period 1963–1990, but they disagree on the direction: LS suggests that negative earnings increase average returns while robust regression suggests negative earnings decrease expected returns. The robust regression suggests that the relationship between average returns and earnings-to-price is not U-shaped, as FF92 claimed (page 445). Both regressions agree during this time period that average returns increase with increasing earnings-to-price ratio if earnings are positive.

Each of the LS and robust regression results over the longer period 1963–2015 is largely consistent with the corresponding results over the original period 1963–1990. When we

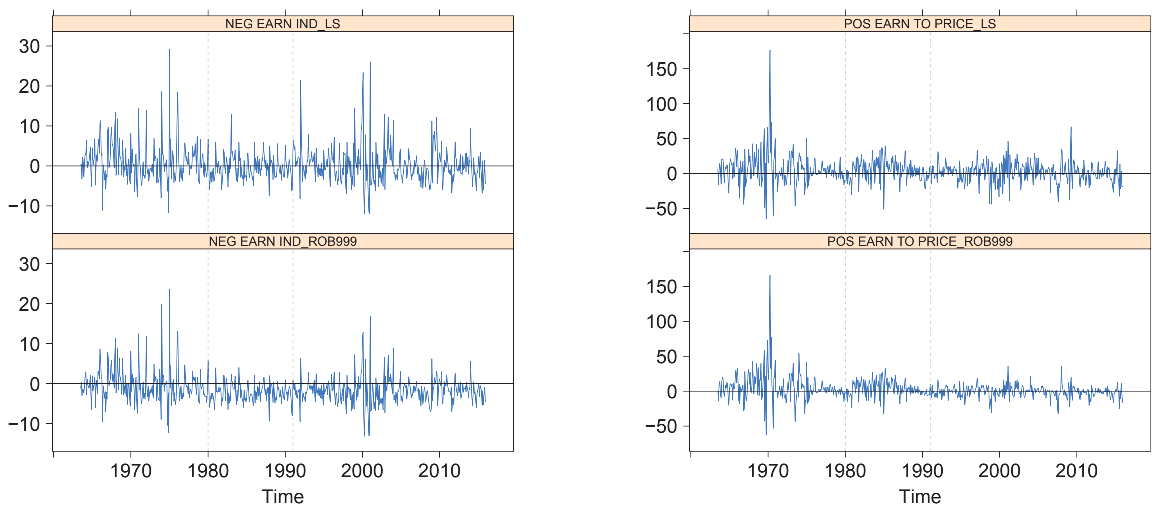


Figure 4.25: Time series of slopes from regression of returns on the two earnings-to-price measures. Left panel: the negative earnings indicator, E/P dummy. Right panel: the positive-earnings-to-price ratio, E+/P.

omit the 1963–1979 period, however, the LS risk premium for stocks with negative earnings (E/P Dummy) vanishes. The time series of LS and robust slopes, as shown in Figure 4.25, both exhibit many large isolated spikes and time-varying volatility. These outliers and the heteroskedasticity may distort the sample means and t-statistics of the monthly slopes. For example, the time series of LS slopes for the negative earnings factor exhibits a large spike in January 1975 (a slope of 29.1). If we omit this month, the average LS slope for 1963–1990 drops to 0.43, with a t-statistic of 1.90. For the period 1963–2015, the average LS slope drops to 0.33, with a t-statistic 1.87. The robust slope at January 1975 is also large (23.56), but the average robust slopes for the time periods 1963–1990 and 1963–2015 after removing the January 1975 observation are -1.47 and -1.14, respectively, with t statistics of -10.46 and -5.70. Thus, it is dangerous to draw firm conclusions here on the relevance of the earnings-to-price factors based on the average slope and t-statistic. In Section 4.6.1 we will revisit this analysis with statistical tools better suited for the task.

A deeper examination of the structure of the negative earnings indicator, however, sug-

Table 4.13: Average number of stocks having positive or negative earnings within each size decile over the period 1963–2015.

	Size Deciles									
	1	2	3	4	5	6	7	8	9	10
Positive Earnings	770	258	194	166	146	133	125	123	122	119
Negative Earnings	550	91	45	30	20	15	12	10	7	5

gests that both the LS and robust regression results are misleading for a different reason: the regression results are driven almost entirely by small stocks. Table 4.13 displays the average number of stocks with positive and negative earnings within each size decile over the period 1963–2015. On average there are relatively few stocks with negative earnings in size deciles 3 and larger. Furthermore, the average over 52 years conceals months in the larger size deciles for which there are zero stocks with negative earnings. For these months and size deciles it is not possible to carry out a regression of returns on the two earnings factors (or any regression involving the E/P Dummy variable for that matter), as there are not enough observations to estimate the coefficient for the negative earnings indicator (within those month/size decile combinations). It is only within size decile 1 that we can reliably fit the earnings model. The slopes on negative earnings in the overall model depicted in Table 4.12 are thus overwhelmingly representative of how the market prices negative earnings in small stocks, rather than all stocks.

LS finds a significant and positive average premium on positive earnings-to-price ratio stocks across all three time periods. The robust regression confirms this finding in the periods 1963–1990 and 1963–2015, but the average slope on positive earnings-to-price ratio from the robust regressions over the period 1980–2015 is not significant. This suggests the premium found in the longer time periods may have been driven by something happening during 1963–

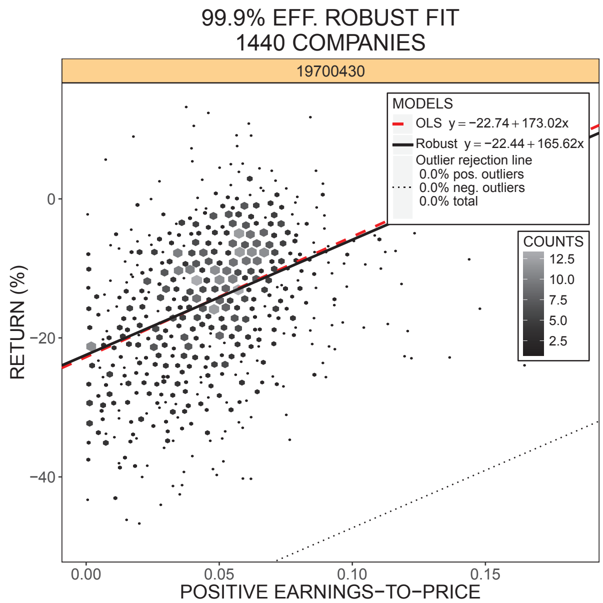


Figure 4.26: Hexbin scatterplot showing stock returns versus E+/P values, April 1970. The plot setup is the same as in Figure 4.9.

1979, and/or that market participants eventually priced the anomaly out of existence after 1979. An obvious suspect for the former explanation is the large positive outlier in the slope on positive earnings-to-price for April 1970 with either regression method. The scatterplot of returns versus E+/P values (Figure 4.26) does not reveal any significant firm-level outliers at this time. Few observations, if any, lie outside the rejection boundaries for the 99.9% efficient robust regression. Removing the April 1970 slope for each of the LS and robust regressions changes the average slopes over 1963–2015 to 3.45 and 2.31, respectively, with t-statistics 5.07 and 4.24. Thus the spike at April 1970 alone does not explain the non-significance of the robust regression results over 1980–2015.

The slopes in the late 1960s and early 1970s are, in fact, all very large. Figure 4.27 shows summary statistics for the regression slopes before and after January 1980, as well as kernel density estimates of the paired differences of slopes for each subperiod. During the period 1963–1979 there are many outlying times like April 1970 that impact both LS and robust fairly equally as evidenced by the summary statistics and density estimates. These

Summary Statistics for E+/P Slopes

	LS		Robust	
	63–79	80–15	63–79	80–15
Minimum	-65.0	-51.2	-62.8	-32.8
1st Quartile	-5.1	-6.4	-2.9	-4.4
Median	4.5	3.1	3.2	0.01
Mean	6.6	2.4	6.8	0.7
3rd Quartile	16.5	10.2	14.2	5.8
Maximum	177.0	67.0	166.6	35.8

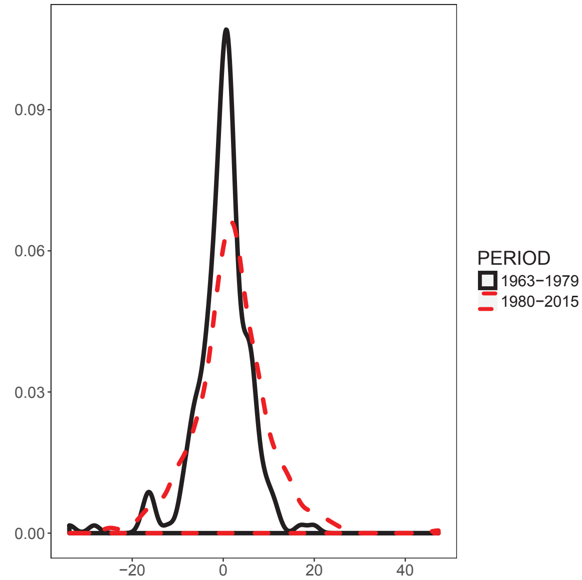


Figure 4.27: Statistics on monthly LS and robust slopes on E+/P, 1963–2015. Left panel: Summary statistics for slopes within each of the two periods July 1963–December 1979 and January 1980–December 2015. Right panel: Kernel density estimate of paired differences (LS - robust) in monthly slopes on E+/P cross-sectional regressions for each of the periods July 1963–December 1979 (solid black line) and January 1980–December 2015 (dashed red line).

abnormally large slopes (LS and robust) in the late 1960s and early 1970s are not driven by firm-specific outliers. Rather, the large slopes arise from low earnings-to-price ratios at these times for all stocks. In contrast, the distributions of the LS and robust slopes during the 1980–2015 period are more dissimilar: the robust slopes have a smaller range than the LS slopes and are centered at approximately 0. These differences are due more to firm-specific outliers in the data for this period. The average slope of the LS regressions over 1980–2015 is thus misleading, as it suggests a market-wide risk premium for positive earnings-to-price that in reality may have only existed for a small number of stocks. It seems likely that the market gradually arbitraged away the risk premium for positive earnings-to-price after the publication of research studies pointing out its existence.

Table 4.14: Average intercept and slopes (t-statistics) from cross-sectional regressions of stock returns on leverage.

Factor	Method	Leverage 1963–1990	1963–2015	1980–2015
Intercept	LS (FF92)			
	LS (GM)	1.44 (5.12)	1.48 (6.76)	1.52 (5.54)
	Robust (GM)	1.03 (3.71)	0.75 (3.65)	0.57 (2.27)
Market Leverage	LS (FF92)	0.50 (5.69)		
	LS (GM)	0.47 (5.53)	0.39 (6.31)	0.36 (5.01)
	Robust (GM)	0.42 (5.37)	0.28 (5.08)	0.27 (4.22)
Book Leverage	LS (FF92)	−0.57 (−5.34)		
	LS (GM)	−0.62 (−5.30)	−0.50 (−7.00)	−0.51 (−7.32)
	Robust (GM)	−0.91 (−9.14)	−0.61 (−10.01)	−0.53 (−9.74)

4.4.5 Leverage Factors

Table 4.14 shows the average slopes and t-statistics for the market leverage and book leverage factors for the three time periods. FF92 trimmed the leverage factors by 0.5% on each side to deal with potential outliers, and the robust regression confirms a lack of extreme outliers in the market leverage factor in any of the time periods. On the other hand, the book leverage factor shows some evidence of mild influence by outliers during 1963–1990. Furthermore, in contrast to the LS regression, the robust regression assigns a much higher risk premium on the book leverage factor than market leverage factor during the period of the FF92 study. Whereas in Fama and French's analysis the combined leverage factor is approximately equal to the book-to-market factor, the robust regression suggests that for the vast majority of stocks and most time periods, book leverage may capture some aspect of average stock returns beyond what is contained in the book-to-market factor.

4.4.6 Summary

Our LS results on the above single factor models agree with those in FF92 for the period 1963–1990, which is good evidence that we have replicated their data and methodology correctly. The robust regression results yield valuable insight into how influenced each factor's results are by firm level outliers.

All regressions support the existence of a positive relationship between returns and the book-to-market ratio, across all time periods and regardless of which regression methodology we use. However, closer inspection of the time series of slopes over 1963–2015 confirms the existence of a January effect, as found by Loughran and others. The robust regression suggests the effect actually exists through the entire first quarter. The January and February effects are driven by large positive slopes in the mid-1970s, and do not persist after 1980. There is a March effect, however, that persists into the 1980–2015. We have not seen this effect mentioned in the literature previously, and currently do not have an explanation for this effect.

The LS and robust methods disagree about the nature of the size effect. The LS approach finds a negative relationship between average returns and firm size, in line with the findings of FF92 and many others. Our robust regression suggests that the relationship is positive for the vast majority of stocks. We confirmed our regression findings in average returns in size-sorted portfolios constructed to limit the impact of unusually large positive and negative returns. The commonly accepted negative relationship and premium for small stocks can be attributed to a small number of particularly influential stocks each month that exhibit unusually high returns at isolated points in time. Capturing this premium requires skill in forecasting which stocks will have large returns or the ability to implement a diverse portfolio of relatively illiquid and capacity-constrained stocks. Neither of these options seems realistic for most investors. The positive relationship suggested by the robust approach is a better representation of how average returns and firm size relate most of the time. That being said, the robust regression also confirms the classic negative January effect for the size factor.

Starting around March, however, the size relationship becomes positive for most stocks, and remains so until the end of the year. Again, this intriguing pattern is masked in the LS regressions by a very small number of outliers.

Beta is also a source of disagreement for the LS and robust methods. Our LS results agree with those of FF92, that there is no real relationship between average returns and firm betas. The robust regression, on the other hand, tells us that average returns generally decrease with increasing beta. In the Sections 4.5 and 4.6 we will see that this relationship is more complex than it appears here, with an interesting co-dependence on firm size.

Table 4.15: Average intercepts and slopes (t-statistics) from cross-sectional regressions of stock returns on beta and size.

			Beta and Size	
Factor	Method	1963–1990	1963–2015	1980–2015
Intercept	LS (FF92)			
	LS (GM)	2.47 (7.25)	2.37 (8.44)	2.47 (6.86)
	Robust (GM)	0.72 (2.22)	0.03 (0.10)	−0.37 (−1.10)
Beta	LS (FF92)	−0.37 (−1.21)		
	LS (GM)	−0.45 (−1.49)	−0.28 (−1.15)	−0.43 (−1.38)
	Robust (GM)	−0.63 (−2.12)	−0.61 (−2.76)	−0.81 (−2.90)
Size	LS (FF92)	−0.17 (−3.41)		
	LS (GM)	−0.16 (−3.19)	−0.16 (−4.44)	−0.13 (−3.10)
	Robust (GM)	0.16 (3.42)	0.24 (7.29)	0.33 (9.11)

4.5 Robust Regression Analysis of Multi-Factor Fama-French Models

4.5.1 Beta and Size

When we consider beta and size together (see Table 4.15), our LS results agree well with FF92 for the time period 1963–1990: the average slopes for both beta and size are negative, with a significant t-statistic for the size factor. The LS results for the longer time periods are similar. Our robust regression results, however, differ from those of FF92: the average slope on beta is still significant and negative, but the average slope on size is now positive and significant. The robust results do suggest that the beta effect is weaker once size is taken into account for all three time periods (compare with Table 4.11). One way to understand this effect is to modify the model to include a size-beta interaction term. We analyze such a model in Section 4.6.3 below.

Table 4.16: Average intercepts and slopes (t-statistics) from cross-sectional regressions of stock returns on size and book-to-market.

		Size and Book-to-Market		
Factor	Method	1963–1990	1963–2015	1980–2015
Intercept	LS (FF92)			
	LS (GM)	1.68 (3.59)	1.84 (5.10)	1.74 (4.12)
	Robust (GM)	−0.34 (−0.80)	−1.04 (−3.46)	−1.71 (−5.25)
Size	LS (FF92)	−0.11 (−1.99)		
	LS (GM)	−0.10 (−1.74)	−0.11 (−2.53)	−0.08 (−1.49)
	Robust (GM)	0.25 (4.96)	0.33 (9.53)	0.44 (11.84)
ln(BE/ME)	LS (FF92)	0.35 (4.44)		
	LS (GM)	0.35 (4.40)	0.28 (4.32)	0.30 (3.70)
	Robust (GM)	0.53 (7.44)	0.49 (8.94)	0.54 (8.23)

4.5.2 Size and Book-to-Market

Table 4.16 shows that the bivariate regression on size and book-to-market is consistent with the univariate regressions on each factor: book-to-market is consistently priced across all time periods and by both LS and robust regressions. Once again our LS results are consistent with those of FF92 for both size and book-to-market. The LS and robust approaches again disagree about the sign and significance of the size premium. It is interesting to note that our LS regressions indicate that, when book-to-market is also in the model, size is priced in this model over the full period 1963–2015, but not in either of the smaller periods. The robust regression indicates that size is consistently significant for all three time periods. The robust regressions indicate that both size and book-to-market are needed to explain the cross-section of average returns.

In their Table V Fama and French (1992) presented time series average monthly returns on equally-weighted portfolios of stocks within deciles of size and book-to-market. Their results support their LS cross-sectional regression findings of a positive relationship between average returns and book-to-market for fixed values of size, and a negative relationship between average returns and size for fixed values of book-to-market. Our robust cross-sectional regression results on the relationship between returns, size, and book-to-market are at odds with this result. As we demonstrated earlier in our discussion of the size effect, however, the negative relationship between returns and size in the time series average returns of equally-weighted portfolios is driven by a small number of small stocks with unusually large returns. The same phenomenon explains the discrepancy between our robust regression results here and FF92's Table V.

As we mentioned earlier in Section 4.4.2, Loughran (1997) found that the LS book-to-market effect was mainly significant (a) in January and (b) for smaller growth firms during the period 1963–1995. The latter observation suggests that we consider cross-sectional regressions of returns on book-to-market within each size decile. The results, shown in Table 4.17 and Figure 4.28, validate Loughran's conclusion that the book-to-market factor is primarily relevant for smaller stocks. Within the smallest three size deciles, book-to-market is consistently priced in each time period and by both LS and robust regression. For stocks in size deciles 4–5, book-to-market is still significant at the 0.05 level for both LS and robust methods over the periods 1963–1990 and 1963–2015, but not in the post-1980 era. This is also the case for size decile 6, except that the average LS slope is not significant for the full period 1963–2015. Given that book-to-market is not significant after 1980 for size deciles 4–6, it is likely that the significance for the period 1963–2015 is driven by influential months prior to 1980. Finally, book-to-market is not significant with either regression methodology for size deciles 7–10 in nearly all time periods.

Table 4.17: Average slopes (t-statistics) from cross-sectional regressions of stock returns on book-to-market by size decile.

Size Decile	Method	1963–1990	Book-to-Market 1963–2015	1980–2015
ME01	LS	0.45 (5.06)	0.42 (6.13)	0.44 (5.47)
	Robust	0.61 (7.72)	0.57 (10.51)	0.63 (10.66)
ME02	LS	0.47 (4.39)	0.35 (4.19)	0.34 (3.27)
	Robust	0.59 (6.08)	0.47 (6.24)	0.45 (5.03)
ME03	LS	0.42 (3.52)	0.28 (3.17)	0.27 (2.53)
	Robust	0.48 (4.40)	0.36 (4.40)	0.33 (3.54)
ME04	LS	0.41 (3.57)	0.23 (2.59)	0.18 (1.65)
	Robust	0.36 (3.23)	0.20 (2.46)	0.13 (1.33)
ME05	LS	0.30 (2.55)	0.20 (2.18)	0.14 (1.24)
	Robust	0.35 (3.11)	0.25 (2.95)	0.19 (1.84)
ME06	LS	0.26 (2.16)	0.15 (1.61)	0.07 (0.64)
	Robust	0.30 (2.62)	0.23 (2.75)	0.14 (1.40)
ME07	LS	0.17 (1.49)	0.08 (0.89)	0.02 (0.21)
	Robust	0.10 (0.96)	0.10 (1.32)	0.08 (0.83)
ME08	LS	0.25 (2.06)	0.10 (1.13)	−0.00 (−0.01)
	Robust	0.22 (1.82)	0.10 (1.14)	0.00 (0.02)
ME09	LS	0.03 (0.27)	0.05 (0.55)	0.07 (0.60)
	Robust	0.01 (0.11)	0.03 (0.40)	0.05 (0.45)
ME10	LS	0.04 (0.33)	−0.02 (−0.28)	−0.04 (−0.47)
	Robust	0.03 (0.22)	−0.02 (−0.23)	−0.03 (−0.35)

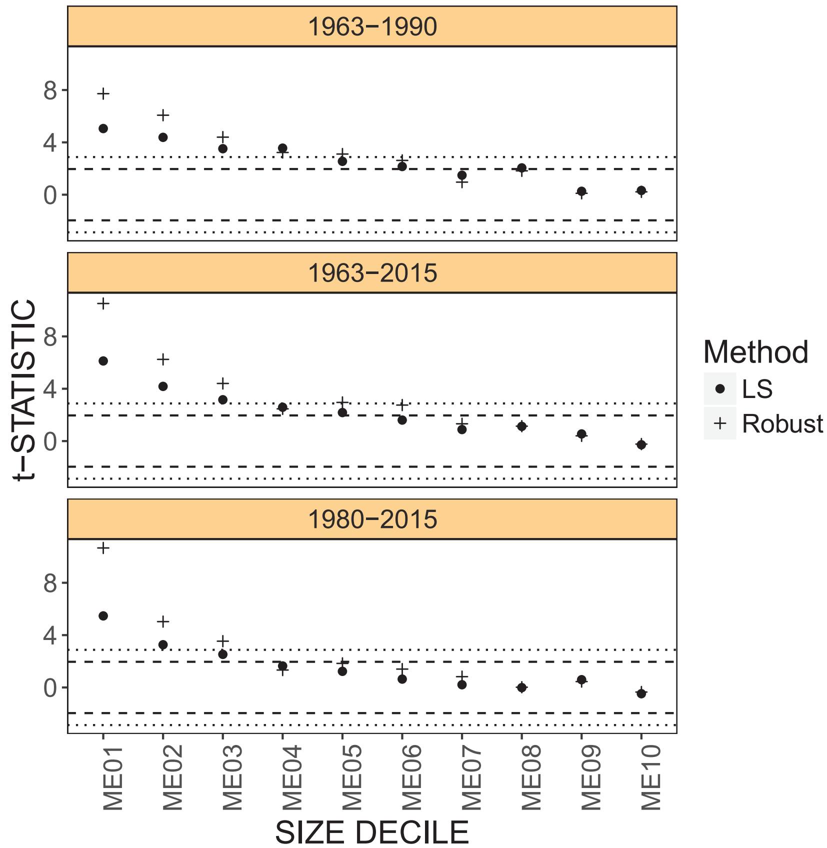


Figure 4.28: Graphical depiction of t-statistics from Table 4.17. The figure shows the t-statistic on book-to-market for each size decile and regression method. The dashed lines indicate the standard cutoff values for the 0.05 significance level (± 1.96), while the dotted lines indicate cutoff values for the Bonferroni-corrected nominal 0.05 significance level.

Exchange Dependency of Size and Book-to-Market Relationship

Figure 4.29 shows the percentage of stocks from each exchange within each size decile over 1963–2015. NASDAQ stocks dominate the smallest size decile and at times represent nearly half of the stocks in the data set for size deciles 2–4. Furthermore, as shown in Figure 4.30, once NASDAQ stocks enter our data set in July 1975,[22] their numbers grow to dominate the

[22]NASDAQ started in 1973, and recall that each month we filter out stocks with less than 24 months of history. Thus July 1975 is the first month for which the data set contains an appreciable number of NASDAQ stocks.

Table 4.18: Average intercepts and slopes (t-statistics) from cross-sectional regressions of stock returns on size and book-to-market by exchange during the period 1963–2015.

		Size and Book-to-Market		
Factor	Method	NYSE	AMEX	NASDAQ
Intercept	LS	1.45 (3.65)	2.22 (5.42)	2.16 (5.52)
	Robust	0.19 (0.53)	−0.94 (−2.79)	−1.71 (−5.68)
Size	LS	−0.06 (−1.56)	−0.27 (−3.90)	−0.16 (−2.79)
	Robust	0.09 (2.69)	0.36 (6.78)	0.57 (12.77)
ln(BE/ME)	LS	0.11 (1.91)	0.30 (3.95)	0.33 (4.18)
	Robust	0.06 (1.04)	0.58 (9.31)	0.69 (11.33)

data set, even though NYSE stocks still compose the bulk of the market capitalization of the data set. This leads us to investigate whether the size and book-to-market factors are priced for each market. Table 4.18 shows the average slopes for size and book-to-market from separate cross-sectional regression analyses for the stocks on each exchange (NYSE, AMEX, NASDAQ) over the period 1963–2015. For AMEX and NASDAQ both size and book-to-market are significant, with average slopes and t-statistics that are in line with the results presented in Table 4.16 for the entire market. For NYSE stocks, however, book-to-market is not significant with either regression methodology, and size is only significant with our robust regression methodology. Hence, the overall regression results reported in Table 4.16 are thus more reflective of the relationship between size and book-to-market for the smaller stocks from NASDAQ than of the relationship across the entire stock market.

In summary, our analysis suggests that the book-to-market "anomaly" (a) never existed for larger stocks; (b) was present in moderately-sized stocks during July 1963–December 1979, but gradually vanished for such stocks thereafter; and (c) has consistently existed for the smallest stocks since 1963. For moderately-sized stocks it is likely that, over time, increased

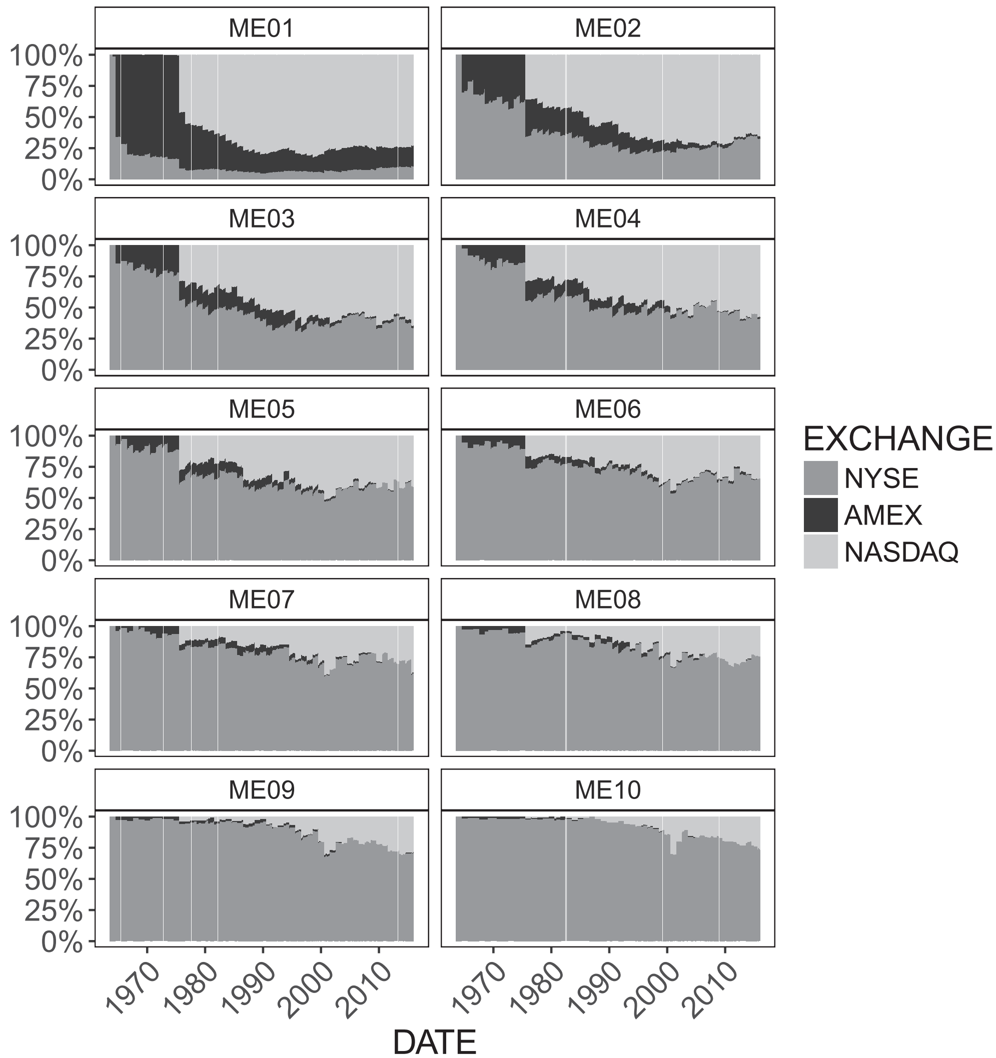


Figure 4.29: Percentage of stocks within each size decile coming from the three exchanges over time.

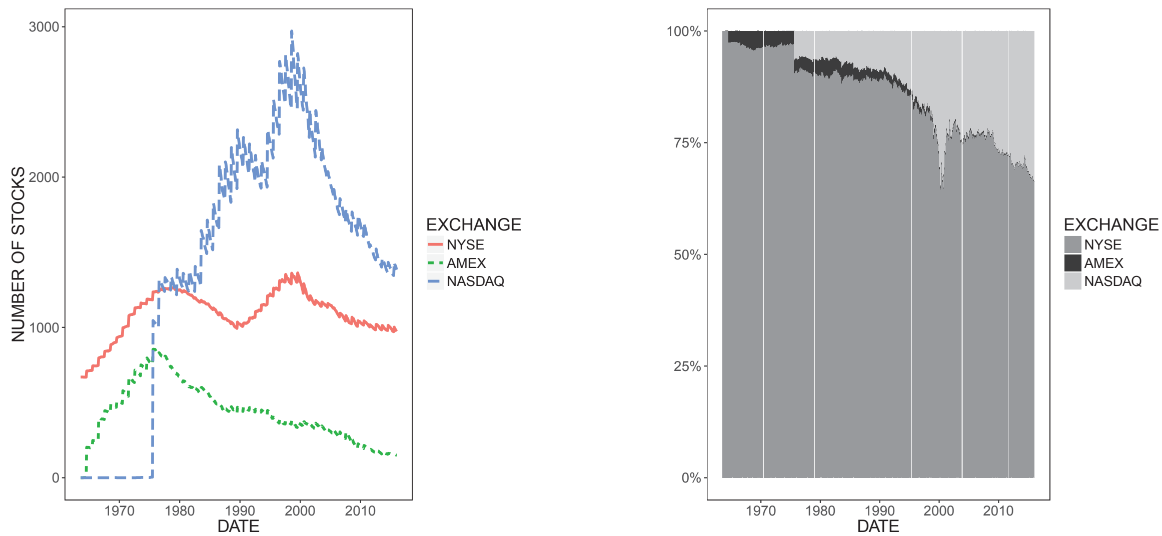


Figure 4.30: Left panel: number of stocks by exchange over time, 1963–2015. Right panel: percentage of total market capitalization of our data set attributable to stocks on each exchange over time, 1963–2015.

access to and efficiency in that segment of the stock market led to the disappearance of the book-to-market effect. On the other hand, the continued presence of the book-to-market premium for the smallest stocks suggests that few market participants have been able to take advantage of their risk premium.

4.5.3 Size, Book-to-Market, and Earnings-to-Price

Table 4.19 shows the results from the cross-sectional regressions on size, book-to-market, and earnings-to-price. The LS and robust results for size and book-to-market are consistent with those from the single factor analyses (Tables 4.5 and 4.6). However for the LS regression the negative earnings indicator (E/P Dummy) is significant only for the time period 1980–2015, and the positive earnings-to-price ratio is only significant for the time period 1963–1990. Based on the LS results, we would conclude, as did FF92, that earnings-to-price measures explain very little of the cross-section of returns once size and book-to-market effects are taken into account, and hence can be dropped from a model for average returns in the interest of parsimony.

The robust regressions, however, tell a slightly different story. The average return premium for negative earnings, while smaller than in the standalone earnings model (Table 4.12), remains consistently negative and significant when size and book-to-market are added to the model. Positive earnings-to-price is still positively related to returns after the addition of size and book-to-market to the model in the 1963–1990 and 1963–2015 periods. During 1980–2015, however, the average slope is nearly zero. We saw in the standalone model that in the late 1960s and early 1970s earnings-to-price ratios were very small, resulting in some very large slopes for both LS and robust regressions. An analysis of the pre- and post-1980 behavior similar to that performed in Section 4.4.4 shows that the significance of the positive earnings-to-price factor in the robust regressions during 1963–1990 and 1963–2015 is driven by market behavior prior to 1980. We conclude that from 1980 onward the positive earnings-to-price factor was not priced, once size and book-to-market factors were taken into account.

As for whether the negative earnings indicator should be included in the model, we have previously seen that the coefficients for this factor are due almost entirely to stocks in the first size decile. In this model, however, the robust regression version of the factor is still priced despite the inclusion of size in the model. Hence it appears that for small stocks, whether a stock's earnings were positive or negative is a relevant factor for explaining the cross-section of stock returns.

4.5.4 Summary

The multifactor analyses revealed more of the size-related dependencies of the beta, book-to-market, and earnings-to-price factors. We found that the relationship between average returns and beta was still significant and negative for the vast majority of stocks even after accounting for size effects. The beta effect is weaker, however, once size is added to the model, and there is evidence that the relationship between returns and beta may be different for small and large stocks. We investigate this matter further in Section 4.6.3.

We affirmed the findings of Loughran (1997) that book-to-market mainly matters for small stocks. Our analysis shows that book-to-market was not relevant for large stocks during 1963–2015 and was of decreasing relevance for moderately size-stocks after 1980. The book-to-market effect remains significant for small stocks, however, even through December 2015.

Finally, we found that the positive earnings-to-price factor was not relevant for average returns after 1980. Average returns for mid- and large capitalization stocks were not impacted by whether a firm's earnings were negative, but small stocks still exhibit a return penalty for negative earnings (regardless of their magnitude).

Table 4.19: Average intercepts and slopes (t-statistics) from cross-sectional regressions of stock returns on size, book-to-market, and earnings-to-price.

		Size, Book-to-Market, and Earnings-to-Price		
Factor	Method	1963–1990	1963–2015	1980–2015
Intercept	LS (FF92)			
	LS (GM)	1.71 (3.63)	1.94 (5.98)	1.93 (5.54)
	Robust (GM)	−0.05 (−0.12)	−0.39 (−1.40)	−0.72 (−2.56)
Size	LS (FF92)	−0.13 (−2.47)		
	LS (GM)	−0.12 (−2.32)	−0.13 (−3.39)	−0.10 (−2.42)
	Robust (GM)	0.18 (3.65)	0.24 (7.36)	0.31 (9.45)
ln(BE/ME)	LS (FF92)	0.33 (4.46)		
	LS (GM)	0.31 (4.17)	0.25 (4.37)	0.28 (3.85)
	Robust (GM)	0.45 (6.72)	0.41 (8.07)	0.46 (7.58)
E/P Dummy	LS (FF92)	−0.14 (−0.90)		
	LS (GM)	−0.14 (−0.95)	−0.14 (−1.18)	−0.30 (−2.01)
	Robust (GM)	−1.03 (−8.59)	−1.15 (−11.91)	−1.43 (−12.40)
E+/P	LS (FF92)	0.87 (1.23)		
	LS (GM)	1.81 (2.14)	0.51 (0.93)	−0.17 (−0.30)
	Robust (GM)	2.83 (4.08)	1.24 (3.00)	−0.03 (−0.10)

4.6 Extensions of Fama-French Models

4.6.1 Outlier and Autocorrelation Robust t-stats

The time series of regression slopes shown throughout this paper show evidence of fat-tailed non-normality, outliers, time-varying volatility, and serial correlation. Figure 4.31 shows normal quantile-quantile (QQ) plots of the LS and robust slopes for the beta, size, book-to-market, and earnings-to-price single factor models. The distributions of the slopes from beta, book-to-market, and earnings-to-price models exhibit longer right tails than a normal distribution would, while the size factor has a heavy left tail.

The Fama-Macbeth method of basing factor pricing conclusions on the time series sample means and classical t-statistics of the cross-sectional regression slopes will not always provide reliable inference under these circumstances. We therefore examine the use of two improvements to the classical approach: replacing the sample mean with a robust location estimate, and using a heteroskedasticity and autocorrelation consistent (HAC) method for estimating standard errors. The former modification protects our estimate of the "typical" factor premium from distortion by outliers and non-normality in the time series of slopes. The latter modification protects our inferences against serial correlation and time-varying volatility in the time series of slopes.

We consider three combinations of these improvements. In the first, we continue to use the sample mean but adjust our standard errors (and hence, t-statistics) using the well-known Newey and West (1987) HAC approach to correcting standard errors of regression coefficients. Software for this technique is readily available, e.g., in the `sandwich` `R` package (Zeileis, 2004). Our second approach replaces the sample mean with a 99.9% efficient robust estimate of location, and uses the Newey-West correction for t-statistics. The robust location estimator, or "robust mean", is a special case of the robust regression estimator we have been using, in which the regression part of the model contains only a constant (the intercept). The third approach combines the robust estimate of location above with the method of Croux et al. (2003a) to obtain outlier-robust and HAC standard errors. This method is

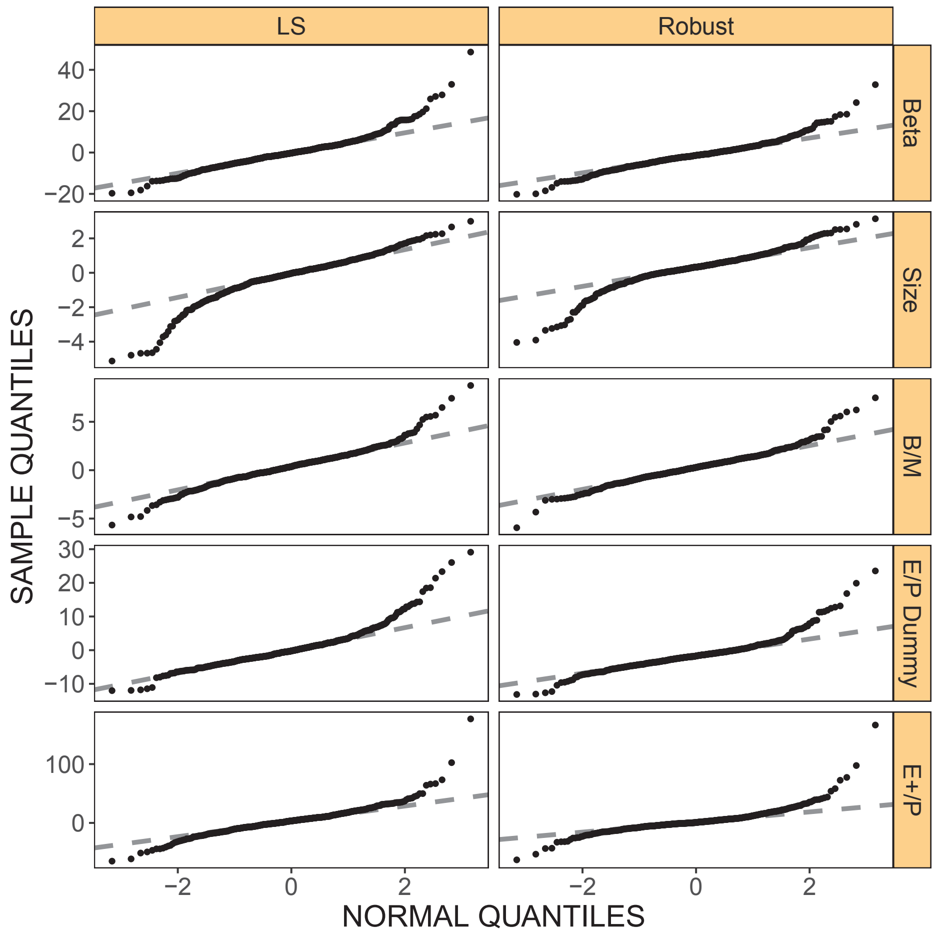


Figure 4.31: Normal quantile-quantile (QQ) plots of monthly slopes from LS and robust regressions for the period 1963–2015. Each row corresponds to the slopes for a single factor (noted on the right). LS slopes are shown in the left-hand column, while robust slopes appear on the right.

specifically designed for the robust location estimate and provides additional robustness against outliers that Newey-West does not. (Appendix 4.A provides a brief overview of the Croux et al. (2003a) method.) Contrasting these three approaches can reveal the extent to which fat-tailed non-normality and/or serial correlation influences the time series of slopes and adversely affects our inference about whether factors are priced.

We present both the sample mean and robust mean of slopes, with HAC-corrected t-statistics, in Tables 4.20-4.26. For brevity we have only included tables for selected univariate models (beta, size, book-to-market, and earnings-to-price) and the three most important multifactor models (beta and size; size and book-to-market; and size, book-to-market, and earnings). Newey-West and Croux et al. HAC-corrected t-statistics are calculated using the raw residuals from either LS or robust regression of the monthly slopes on a constant to obtain either the sample mean (in the LS case) or our "robust mean" (in the robust case).

Table 4.20: Sample means and robust location estimates of slopes from cross-sectional regression of stock returns on beta, with heteroskedasticity and autocorrelation consistent (HAC) t-statistics. For convenience we include the uncorrected t-statistics for the LS and robust cross-sectional regressions shown previously in Table 4.11. The "Regression" column indicates the method used for the monthly cross-sectional regressions. The "Location Estimate" column indicates whether the sample mean or robust location estimate of slopes is reported in that row. The robust location estimate used is based on the same 99.9% efficient regression MM-estimates used throughout in the paper. The "t-Statistic" column indicates which HAC adjustment was used to calculate the t-statistics in that row—none ("uncorrected"), Newey and West (1987), or Croux et al. (2003a).

Regression	Location Estimate	t-Statistic	Post-Ranking Beta 1963–1990	Post-Ranking Beta 1963–2015	Post-Ranking Beta 1980–2015
LS	mean	uncorrected	0.05 (0.14)	0.16 (0.64)	−0.08 (−0.25)
	mean	Newey-West	0.05 (0.13)	0.16 (0.55)	−0.08 (−0.22)
	robust	Newey-West	−0.12 (−0.21)	−0.13 (−0.29)	−0.43 (−0.84)
	robust	Croux et al.	−0.12 (−0.35)	−0.13 (−0.53)	−0.43 (−1.52)
Robust	mean	uncorrected	−1.08 (−3.47)	−1.25 (−5.72)	−1.70 (−6.66)
	mean	Newey-West	−1.08 (−3.18)	−1.25 (−4.93)	−1.70 (−6.04)
	robust	Newey-West	−1.32 (−2.52)	−1.41 (−3.85)	−1.77 (−4.59)
	robust	Croux et al.	−1.32 (−4.30)	−1.41 (−6.94)	−1.77 (−7.49)

Table 4.21: Location estimates and t-statistics of slopes from cross-sectional regressions of stock returns on size. The notation is identical to that used in Table 4.20.

Regression	Location Estimate	t-Statistic	Size 1963–1990		Size 1963–2015		Size 1980–2015	
LS	mean	uncorrected	−0.13	(−2.33)	−0.14	(−3.45)	−0.10	(−2.25)
	mean	Newey-West	−0.13	(−2.17)	−0.14	(−2.97)	−0.10	(−1.99)
	robust	Newey-West	−0.06	(−0.62)	−0.07	(−0.95)	−0.05	(−0.72)
	robust	Croux et al.	−0.06	(−1.18)	−0.07	(−1.87)	−0.05	(−1.33)
Robust	mean	uncorrected	0.21	(4.01)	0.28	(8.44)	0.39	(11.48)
	mean	Newey-West	0.21	(3.83)	0.28	(7.03)	0.39	(11.66)
	robust	Newey-West	0.28	(2.86)	0.34	(5.26)	0.40	(8.29)
	robust	Croux et al.	0.28	(6.29)	0.34	(11.62)	0.40	(12.95)

Table 4.22: Location estimates and t-statistics of slopes from cross-sectional regressions of stock returns on beta and size. The notation is identical to that used in Table 4.20.

Factor	Regression	Location Estimate	t-Statistic	Beta and Size 1963–1990	Beta and Size 1963–2015	Beta and Size 1980–2015
Beta	LS	mean	uncorrected	−0.45 (−1.49)	−0.28 (−1.15)	−0.43 (−1.38)
		mean	Newey-West	−0.45 (−1.39)	−0.28 (−1.11)	−0.43 (−1.34)
		robust	Newey-West	−0.50 (−1.28)	−0.48 (−1.40)	−0.70 (−1.54)
		robust	Croux et al.	−0.50 (−1.63)	−0.48 (−2.11)	−0.70 (−2.42)
	Robust	mean	uncorrected	−0.63 (−2.12)	−0.61 (−2.76)	−0.81 (−2.90)
		mean	Newey-West	−0.63 (−1.99)	−0.61 (−2.50)	−0.81 (−2.62)
		robust	Newey-West	−0.69 (−1.80)	−0.69 (−2.13)	−0.90 (−2.19)
		robust	Croux et al.	−0.69 (−2.30)	−0.69 (−3.21)	−0.90 (−3.34)
Size	LS	mean	uncorrected	−0.16 (−3.19)	−0.16 (−4.44)	−0.13 (−3.10)
		mean	Newey-West	−0.16 (−3.13)	−0.16 (−4.34)	−0.13 (−3.17)
		robust	Newey-West	−0.13 (−1.85)	−0.13 (−2.60)	−0.11 (−2.04)
		robust	Croux et al.	−0.13 (−2.67)	−0.13 (−3.68)	−0.11 (−2.66)
	Robust	mean	uncorrected	0.16 (3.42)	0.24 (7.29)	0.33 (9.11)
		mean	Newey-West	0.16 (3.44)	0.24 (6.42)	0.33 (9.20)
		robust	Newey-West	0.20 (2.54)	0.27 (5.33)	0.34 (7.47)
		robust	Croux et al.	0.20 (4.47)	0.27 (8.62)	0.34 (9.28)

Table 4.23: Location estimates and t-statistics of slopes from cross-sectional regressions of stock returns on book-to-market. The notation is identical to that used in Table 4.20.

Regression	Location Estimate	t-Statistic	Book-to-Market 1963–1990	Book-to-Market 1963–2015	Book-to-Market 1980–2015
LS	mean	uncorrected	0.47 (5.51)	0.39 (6.59)	0.37 (5.48)
	mean	Newey-West	0.47 (5.25)	0.39 (5.72)	0.37 (4.32)
	robust	Newey-West	0.42 (4.03)	0.36 (4.18)	0.38 (3.65)
	robust	Croux et al.	0.42 (5.64)	0.36 (6.51)	0.38 (5.77)
Robust	mean	uncorrected	0.44 (5.59)	0.30 (5.63)	0.30 (4.99)
	mean	Newey-West	0.44 (5.35)	0.30 (4.43)	0.30 (3.37)
	robust	Newey-West	0.40 (3.68)	0.26 (3.26)	0.29 (2.90)
	robust	Croux et al.	0.40 (5.68)	0.26 (5.39)	0.29 (4.98)

Table 4.24: Location estimates and t-statistics of slopes from cross-sectional regressions of stock returns on earnings-to-price. The notation is identical to that used in Table 4.20.

Factor	Regression	Location Estimate	t-Statistic	Earnings-to-Price 1963–1990	Earnings-to-Price 1963–2015	Earnings-to-Price 1980–2015
E/P Dummy	LS	mean	uncorrected	0.52 (2.14)	0.37 (2.06)	0.03 (0.15)
		mean	Newey-West	0.52 (2.00)	0.37 (1.72)	0.03 (0.12)
		robust	Newey-West	0.21 (0.53)	0.05 (0.15)	−0.27 (−0.78)
		robust	Croux et al.	0.21 (0.95)	0.05 (0.28)	−0.27 (−1.48)
	Robust	mean	uncorrected	−1.07 (−5.00)	−1.43 (−9.80)	−1.94 (−13.02)
		mean	Newey-West	−1.07 (−4.76)	−1.43 (−7.88)	−1.94 (−12.20)
		robust	Newey-West	−1.41 (−4.00)	−1.67 (−6.45)	−2.03 (−9.60)
		robust	Croux et al.	−1.41 (−8.04)	−1.67 (−13.84)	−2.03 (−15.66)
E+/P	LS	mean	uncorrected	5.78 (4.96)	3.73 (5.08)	2.40 (3.39)
		mean	Newey-West	5.78 (4.50)	3.73 (3.85)	2.40 (2.67)
		robust	Newey-West	5.05 (2.44)	3.20 (2.42)	2.39 (1.99)
		robust	Croux et al.	5.05 (5.16)	3.20 (5.09)	2.39 (3.37)
	Robust	mean	uncorrected	5.58 (5.38)	2.58 (4.26)	0.65 (1.41)
		mean	Newey-West	5.58 (4.80)	2.58 (2.72)	0.65 (0.96)
		robust	Newey-West	4.67 (2.16)	1.72 (1.06)	0.61 (0.59)
		robust	Croux et al.	4.67 (6.06)	1.72 (3.69)	0.61 (1.36)

Table 4.25: Location estimates and t-statistics of slopes from cross-sectional regressions of stock returns on size and book-to-market. The notation is identical to that used in Table 4.20.

Factor	Regression	Location Estimate	t-Statistic	Size and Book-to-Market 1963–1990	Size and Book-to-Market 1963–2015	Size and Book-to-Market 1980–2015
Size	LS	mean	uncorrected	−0.10 (−1.74)	−0.11 (−2.53)	−0.08 (−1.49)
		mean	Newey-West	−0.10 (−1.61)	−0.11 (−2.23)	−0.08 (−1.43)
		robust	Newey-West	−0.03 (−0.26)	−0.03 (−0.44)	−0.02 (−0.23)
		robust	Croux et al.	−0.03 (−0.49)	−0.03 (−0.87)	−0.02 (−0.43)
	Robust	mean	uncorrected	0.25 (4.96)	0.33 (9.53)	0.44 (11.84)
		mean	Newey-West	0.25 (4.71)	0.33 (8.05)	0.44 (12.75)
		robust	Newey-West	0.33 (3.25)	0.39 (5.86)	0.45 (8.86)
		robust	Croux et al.	0.33 (7.29)	0.39 (13.30)	0.45 (14.36)
ln(BE/ME)	LS	mean	uncorrected	0.35 (4.40)	0.28 (4.32)	0.30 (3.70)
		mean	Newey-West	0.35 (4.04)	0.28 (3.65)	0.30 (3.19)
		robust	Newey-West	0.38 (2.93)	0.33 (3.26)	0.37 (3.06)
		robust	Croux et al.	0.38 (5.04)	0.33 (5.81)	0.37 (5.32)
	Robust	mean	uncorrected	0.53 (7.44)	0.49 (8.94)	0.54 (8.23)
		mean	Newey-West	0.53 (6.87)	0.49 (6.91)	0.54 (6.20)
		robust	Newey-West	0.57 (4.45)	0.50 (5.18)	0.54 (4.90)
		robust	Croux et al.	0.57 (8.46)	0.50 (10.13)	0.54 (9.18)

Table 4.26: Location estimates and t-statistics of slopes from cross-sectional regressions of stock returns on size, book-to-market, and earnings-to-price. The notation is identical to that used in Table 4.20.

				Size, Book-to-Market, and Earnings-to-Price		
Factor	Regression	Location Estimate	t-Statistic	1963–1990	1963–2015	1980–2015
Size	LS	mean	uncorrected	−0.12 (−2.32)	−0.13 (−3.39)	−0.10 (−2.42)
		mean	Newey-West	−0.12 (−2.16)	−0.13 (−3.05)	−0.10 (−2.50)
		robust	Newey-West	−0.04 (−0.40)	−0.06 (−0.96)	−0.06 (−1.12)
		robust	Croux et al.	−0.04 (−0.85)	−0.06 (−1.84)	−0.06 (−1.79)
	Robust	mean	uncorrected	0.18 (3.65)	0.24 (7.36)	0.31 (9.45)
		mean	Newey-West	0.18 (3.51)	0.24 (6.58)	0.31 (10.51)
		robust	Newey-West	0.26 (2.82)	0.29 (5.44)	0.32 (7.98)
		robust	Croux et al.	0.26 (6.14)	0.29 (10.85)	0.32 (11.21)
ln(BE/ME)	LS	mean	uncorrected	0.31 (4.17)	0.25 (4.37)	0.28 (3.85)
		mean	Newey-West	0.31 (3.86)	0.25 (3.68)	0.28 (3.30)
		robust	Newey-West	0.34 (2.80)	0.31 (3.32)	0.35 (3.23)
		robust	Croux et al.	0.34 (4.72)	0.31 (5.67)	0.35 (5.38)
	Robust	mean	uncorrected	0.45 (6.72)	0.41 (8.07)	0.46 (7.58)
		mean	Newey-West	0.45 (6.22)	0.41 (6.12)	0.46 (5.55)
		robust	Newey-West	0.50 (3.86)	0.43 (4.64)	0.47 (4.51)
		robust	Croux et al.	0.50 (7.74)	0.43 (9.04)	0.47 (8.31)

Table 4.26: *(continued)*

Factor	Regression	Location Estimate	t-Statistic	1963–1990		1963–2015		1980–2015	
E/P Dummy	LS	mean	uncorrected	−0.14	(−0.95)	−0.14	(−1.18)	−0.30	(−2.01)
		mean	Newey-West	−0.14	(−0.92)	−0.14	(−0.97)	−0.30	(−1.55)
		robust	Newey-West	−0.24	(−1.07)	−0.30	(−1.38)	−0.50	(−1.76)
		robust	Croux et al.	−0.24	(−1.71)	−0.30	(−2.64)	−0.50	(−3.68)
	Robust	mean	uncorrected	−1.03	(−8.59)	−1.15	(−11.91)	−1.43	(−12.40)
		mean	Newey-West	−1.03	(−7.91)	−1.15	(−8.82)	−1.43	(−9.64)
		robust	Newey-West	−1.09	(−4.66)	−1.22	(−6.85)	−1.52	(−7.82)
		robust	Croux et al.	−1.09	(−9.06)	−1.22	(−13.59)	−1.52	(−14.57)
E+/P	LS	mean	uncorrected	1.81	(2.14)	0.51	(0.93)	−0.17	(−0.30)
		mean	Newey-West	1.81	(2.01)	0.51	(0.70)	−0.17	(−0.24)
		robust	Newey-West	1.27	(0.65)	0.08	(0.07)	−0.31	(−0.32)
		robust	Croux et al.	1.27	(1.69)	0.08	(0.16)	−0.31	(−0.56)
	Robust	mean	uncorrected	2.83	(4.08)	1.24	(3.00)	−0.03	(−0.10)
		mean	Newey-West	2.83	(3.72)	1.24	(1.99)	−0.03	(−0.08)
		robust	Newey-West	2.00	(0.98)	0.69	(0.64)	−0.11	(−0.17)
		robust	Croux et al.	2.00	(3.85)	0.69	(2.21)	−0.11	(−0.34)

Sample means with Newey-West HAC correction

The Newey-West corrected t-statistics for the sample means of both the LS and robust slopes are almost always smaller in absolute value than the uncorrected t-statistics, the only exceptions being a few instances of the sample mean estimators where the uncorrected t-statistic is a little larger in absolute value than the Newey-West corrected t-statistic (with both t-statistics highly significant). This is typically the case for Newey-West t-statistics: the classical t-statistic underestimates the variance of the residuals by failing to account for the presence of positive serial correlation, while the Newey-West t-statistic is based on a residual variance estimate that is usually larger.

Among the six models considered and both LS and robust regressions, we find only two instances where the sample mean of the monthly slopes was no longer significant at the 5% level after the Newey-West correction. Specifically, with LS regression, E/P Dummy is no longer significant in the earnings-to-price model during 1963–2015; and E/P Dummy is no longer significant in the size, book-to-market, and earnings-to-price model during 1980–2015. The cases for the sample mean slopes over the 1980–2015 period where the Newey-West corrected t-statistic is larger than the classical t-statistic are likely caused by outlier-induced negative serial correlation.

Robust location estimates with Newey-West HAC correction

The robust location estimates for the LS slopes, on the other hand, do lead to some different conclusions, particularly about the size effect. Size is not priced during any time period for any of the following three models involving size, regardless of whether the classical or Newey-West corrected t-statistic is used: the single-factor LS model (Table 4.21); the size and book-to-market LS model (Table 4.25); and the size, book-to-market, and earnings-to-price LS model (Table 4.26). This tells us that the average LS slopes and t-statistics for size are strongly influenced by a small number of anomalous months with outlying slope values, and do not represent typical market behavior. This is consistent with our earlier finding (in Section 4.4.2) that the average LS slope was driven largely by outlying slopes from the month of January.

Although the robust location estimate for the LS slopes on beta are still not significant, note that the sign is now negative in all periods. This matches the results obtained with robust regression, and tells us that the positive LS slope on beta was driven by a small number of months with outlying slopes. Given that the sign on the sample mean of the LS slopes for beta was negative during 1980–2015, it is likely that the anomalous months occurred during the period July 1963–December 1979.

It is important to note that the Newey-West t-statistics are typically smaller for the robust means of monthly slopes than for the corresponding sample means. The robust location estimate, by limiting the influence of individual months on the regression coefficient, results in larger residuals for outlying months (compared to their LS counterparts). The standard error estimates from the robust means are usually larger than those obtained using the sample mean, and result in smaller t-statistics.

Robust location estimates with Croux et al. HAC correction

Overall we would reach the same conclusions about factor pricing, based on the robust means, using the Croux et al. t-statistics as we would with the Newey-West t-statistics. Occasionally the Newey-West approach would declare a factor as insignificant over a time period while the Croux et al. approach would find it to be significant (using the same significance level). The Newey-West adjustment is based on classical estimates that are not robust to outliers, so outliers in the time series of slopes will inflate the calculated standard errors. In contrast, the Croux et al. adjusted standard errors are based on a residual scale estimate that is robust to outliers and is a consistent estimator of the standard deviation for normally distributed data. Since these standard errors are the denominators of the t-statistics, the Croux t-statistics tend to be larger in magnitude than the Newey-West t-statistics.

In our analyses, the Croux statistics are typically about twice as large as the Newey-West statistics; for some of the earnings models they are nearly four times as large. Thus while at

5% significance our conclusions may not change, they might at stricter levels of significance. For instance, in the earnings-only model (Table 4.24), we would fail to reject the hypothesis that the robust means from the LS and robust cross-sectional regressions for the period 1963–1990 are statistically indistinguishable from zero based on the Newey-West t-statistics and 1% significance level (with cutoff value 2.58), but with the Croux et al. t-statistics we would still reject the hypothesis even at the 0.1% level. Thus the Croux et al. approach has greater statistical power with the robust location estimate than the Newey-West approach, and it is the preferred approach for HAC standard errors with robust regression.

A careful examination of the results presented Tables 4.20-4.26 for the robust cross-sectional regressions reveals few cases where the robust mean slopes with the Newey-West t-statistics would lead to a different conclusion than the average slopes with uncorrected t-statistics, and no such cases for the robust location estimates with the Croux et al. t-statistics. The combination of robust cross-sectional regressions, robust location estimates for the slopes, and Croux et al. HAC-adjusted t-statistics mitigates extreme outliers both at the firm level and at the time period level, leading us to conclusions about risk premia that hold for the majority of firms in the majority of months. These results, and the contrast with the LS results, help practitioners distinguish which factor risk premia are consistently priced by the market as a whole and which are driven by a small fraction of outlying stocks, unusual months, and/or serial correlation.

Based on the results of this section, we will only report (a) sample means and uncorrected t-statistics; and (b) robust means with Croux et al. t-statistics in the analyses appearing the remainder of this paper.

4.6.2 Returns and Unaltered Earnings-to-Price Ratios

FF92 chose to split the earnings-to-price ratio into a negative earnings indicator (E/P Dummy) and a ratio of positive earnings-to-price on the grounds that negative earnings "are not a proxy for the earnings forecasts embedded in the stock price" (Fama and French, 1992, 444) and that the distribution of returns versus earnings was "U-shaped" (Fama and French, 1992, 445). This is no longer common practice, so we investigated models using the unaltered earnings-to-price ratio, i.e., we do not separate negative and positive earnings. Table 4.27 shows that, if we use average slopes rather than a robust location estimator of slopes, earnings-to-price alone is not a significant predictor of average returns under the LS approach. The robust means of the monthly LS slopes with Croux et al. t-statistics, however, tell us that earnings-to-price is significant during 1963–2015 and 1980–2015. Moreover, our robust cross-sectional regression results uniformly indicate that earnings-to-price was priced during all three time periods.

Table 4.27: Average slopes (t-statistics) from cross-sectional regressions of stock returns on unaltered earnings-to-price.

				Unaltered Earnings-to-Price		
Factor	Regression	Location Estimate	t-Statistic	1963–1990	1963–2015	1980–2015
Intercept	LS	mean	uncorrected	1.20 (3.50)	1.25 (5.20)	1.23 (4.47)
		robust	Croux et al.	1.30 (4.08)	1.30 (5.77)	1.30 (5.12)
	Robust	mean	uncorrected	0.43 (1.32)	0.38 (1.68)	0.29 (1.13)
		robust	Croux et al.	0.54 (1.82)	0.48 (2.31)	0.43 (1.80)
E/P	LS	mean	uncorrected	0.81 (1.14)	0.62 (1.52)	0.41 (1.57)
		robust	Croux et al.	0.32 (0.90)	0.56 (2.41)	0.70 (3.08)
	Robust	mean	uncorrected	2.92 (5.01)	2.47 (7.74)	1.86 (13.18)
		robust	Croux et al.	1.88 (9.73)	1.82 (13.74)	1.71 (15.46)

Table 4.28 shows the regression results for a size, book-to-market, and earnings-to-price model. LS regression would lead us to conclude that only the book-to-market factor matters. The average slope on size is highly significant for the time interval 1963–2015, but at best marginally significant for the other time intervals. In contrast the robust mean LS slope is not consistent for any time interval. With LS regression earnings-to-price is significant only during 1980–2015 according to the average slope with the uncorrected t-statistic, but the robust location estimate with Croux et al. t-statistic indicates it was priced during 1963–2015 and 1980–2015 for most months. The robust cross-sectional regression, on the other hand, supports the existence of significant size, book-to-market, and earnings-to-price factors, consistently during nearly all time periods for both sample mean and robust mean slopes.[23]

[23] Moreover, the average slopes estimated using our robust regression are consistent with those estimated in the earlier single factor models.

Table 4.28: Average slopes (t-statistics) from cross-sectional regressions of stock returns on size, book-to-market, and unaltered earnings-to-price.

Factor	Regression	Location Estimate	t-Statistic	Size, Book-to-Market, and Unaltered Earnings-to-Price 1963–1990	1963–2015	1980–2015
Intercept	LS	mean	uncorrected	1.70 (3.59)	1.88 (5.48)	1.80 (4.67)
		robust	Croux et al.	1.54 (3.42)	1.62 (5.03)	1.61 (4.66)
	Robust	mean	uncorrected	−0.34 (−0.79)	−0.86 (−2.92)	−1.39 (−4.48)
		robust	Croux et al.	−0.58 (−1.48)	−0.99 (−3.68)	−1.29 (−4.53)
Size	LS	mean	uncorrected	−0.11 (−1.97)	−0.12 (−2.99)	−0.09 (−1.92)
		robust	Croux et al.	−0.03 (−0.61)	−0.05 (−1.37)	−0.04 (−1.04)
	Robust	mean	uncorrected	0.22 (4.51)	0.29 (8.70)	0.38 (11.01)
		robust	Croux et al.	0.30 (6.89)	0.35 (12.43)	0.39 (13.21)
ln(BE/ME)	LS	mean	uncorrected	0.35 (4.40)	0.27 (4.23)	0.29 (3.60)
		robust	Croux et al.	0.39 (5.16)	0.33 (5.71)	0.36 (5.12)
	Robust	mean	uncorrected	0.51 (7.16)	0.48 (8.94)	0.55 (8.52)
		robust	Croux et al.	0.56 (8.18)	0.50 (10.12)	0.56 (9.43)
E/P	LS	mean	uncorrected	0.49 (1.01)	0.50 (1.70)	0.48 (2.16)
		robust	Croux et al.	0.35 (1.26)	0.65 (3.44)	0.79 (4.22)
	Robust	mean	uncorrected	2.22 (6.01)	1.96 (9.39)	1.59 (13.22)
		robust	Croux et al.	1.57 (9.13)	1.56 (14.25)	1.51 (14.30)

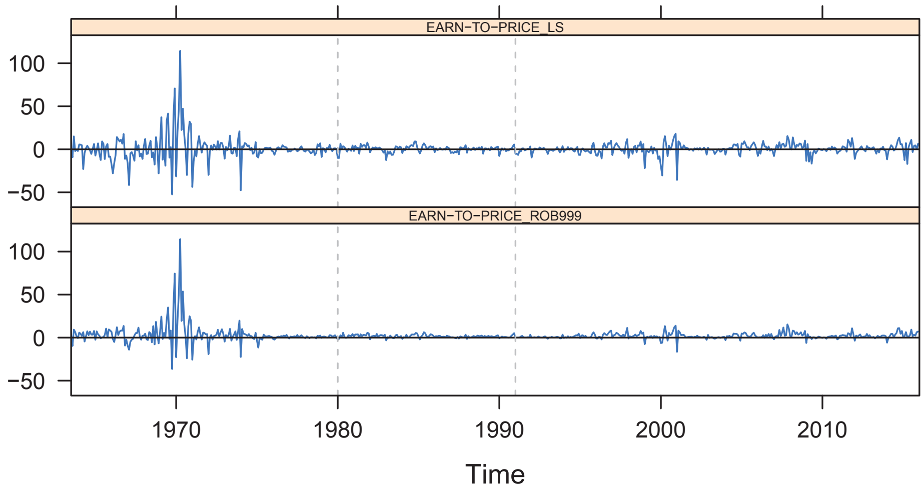


Figure 4.32: Time series of slopes from regression of returns on the unaltered earnings-to-price ratio. The plot setup is identical to that of Figure 4.6.

The difference between the LS and robust regression results for the size and unaltered earnings-to-price factors is again being driven by firm-level outliers in individual months. Figure 4.32 shows the time series of slopes on earnings-to-price in the single factor model depicted in Table 4.27. Again we see the burst of huge positive and negative slopes in 1970 that was discussed earlier in Section 4.4.4. It is firm-level outliers throughout the time periods, however, that lead to differences in the monthly LS and robust slope estimates, differences that translate to different average slopes over time.

We showed previously (Section 4.4.4) that stocks with negative earnings were concentrated in the smallest size decile. A natural question to ask in the present analysis is whether the unaltered earnings-to-price effect exists in all size deciles. Table 4.29 presents the results of the regression of returns on the unaltered earnings-to-price factor within size deciles, and Figure 4.33 summarizes these results graphically.

The average LS slopes for the earnings-to-price factor are significant at an uncorrected 0.05 level for all time periods in size deciles 4, 6, and 8; in the 1980–2015 period only for deciles 2 and 3; and in the 1963–1990 and 1963–2015 periods only for decile 5. As shown in

Figure 4.33 however, most of these slopes are not significant after a Bonferroni correction for the number of tests. The average robust slopes are significant at an uncorrected 0.05 level across all time periods for size deciles 1–6, and in some time periods for deciles 7–9. With the Bonferroni correction the robust slopes are significant for the smallest 40% of stocks during 1980–2015 and the smallest 60% of stocks over 1963–2015. The effect is particularly strong in the smallest decile after 1980. Hence, the unaltered earnings-to-price factor is priced primarily for small-to-moderately sized stocks. This is similar to the behavior we found earlier for the book-to-market factor, and likely also arises from inefficiencies in the stock market for smaller stocks that are hard for practitioners to exploit due to market constraints.

Table 4.29: Average slopes (t-statistics) from cross-sectional regressions of stock returns on unaltered earnings-to-price by size decile.

		Earnings-to-Price		
Size Decile	Method	1963–1990	1963–2015	1980–2015
ME01	LS	−0.11 (−0.18)	0.13 (0.40)	0.34 (1.55)
	Robust	1.43 (3.24)	1.39 (5.79)	1.27 (12.12)
ME02	LS	1.15 (1.01)	0.86 (1.37)	0.98 (2.75)
	Robust	2.56 (2.46)	1.99 (3.48)	1.68 (5.31)
ME03	LS	2.07 (1.42)	1.52 (1.86)	1.41 (2.83)
	Robust	4.71 (3.42)	3.57 (4.61)	2.64 (5.57)
ME04	LS	3.63 (2.12)	2.95 (3.04)	2.22 (3.68)
	Robust	5.11 (3.32)	4.09 (4.65)	3.16 (5.38)
ME05	LS	6.11 (3.33)	3.12 (2.94)	0.30 (0.41)
	Robust	8.10 (4.33)	4.75 (4.37)	1.69 (2.28)
ME06	LS	5.08 (2.58)	3.39 (2.91)	1.87 (2.08)
	Robust	5.76 (2.96)	4.30 (3.78)	2.55 (3.01)
ME07	LS	3.33 (1.88)	2.02 (1.85)	0.45 (0.48)
	Robust	3.09 (1.76)	2.30 (2.17)	0.91 (1.04)
ME08	LS	7.62 (3.41)	4.82 (3.63)	2.10 (2.03)
	Robust	7.84 (3.42)	4.53 (3.38)	1.65 (1.69)
ME09	LS	2.15 (1.13)	1.26 (0.98)	0.87 (0.64)
	Robust	2.97 (1.50)	2.89 (2.26)	2.62 (2.02)
ME10	LS	0.41 (0.20)	1.64 (1.07)	1.83 (1.10)
	Robust	0.26 (0.13)	1.33 (0.84)	1.48 (0.85)

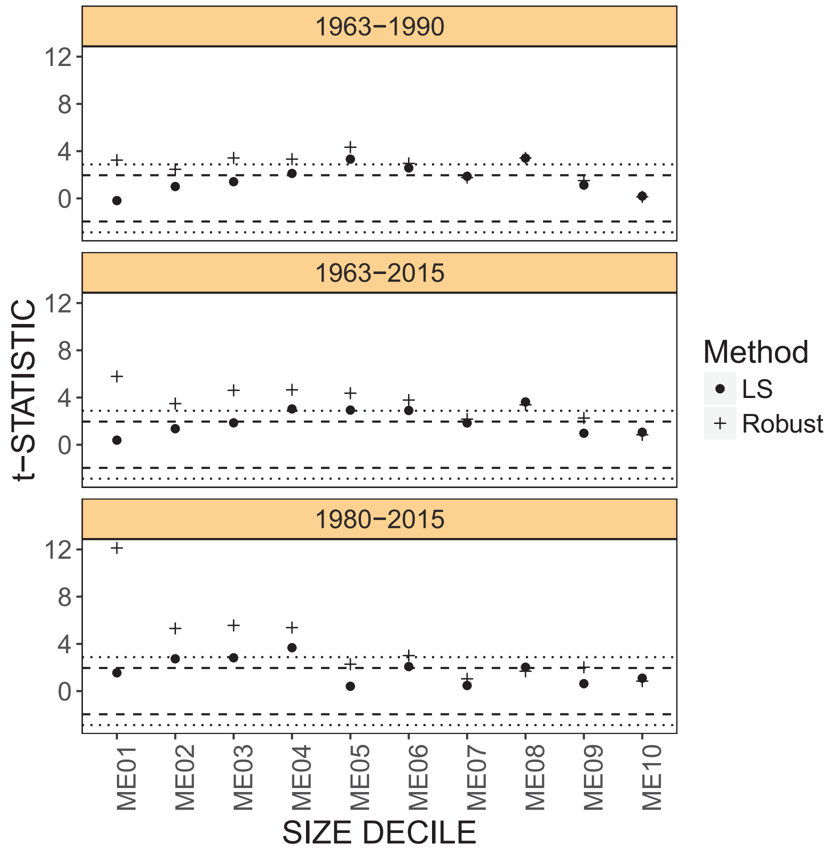


Figure 4.33: Graphical depiction of t-statistics from Table 4.29. The figure shows the t-statistic on earnings-to-price for each size decile and regression method. The dashed lines indicate the standard cutoff values for the 0.05 significance level (± 1.96), while the dotted lines indicate cutoff values for the Bonferroni-corrected nominal 0.05 significance level.

4.6.3 Size and Beta with Interaction

FF92 considered a model for returns containing both size and beta factors and found (using LS) that, on average, beta was not related to returns once the size effect was taken into account. However, our earlier analysis of beta on its own using robust regression showed that there was a significant negative relationship between returns and beta. Moreover, our robust regression analysis of beta and size supported this negative beta effect but suggested the beta effect was weakened by the presence of size in the model. We now explore whether adding an interaction term might help explain how size and beta are related to returns. Table 4.30 shows the results of LS and robust regression models including an interaction term between size and beta.

Table 4.30: Average slopes (t-statistics) from cross-sectional regressions of stock returns on beta, size, beta-size interaction.

Factor	Regression	Location Estimate	t-Statistic	Beta, Size, and Beta-Size Interaction 1963–1990	1963–2015	1980–2015
Intercept	LS	mean	uncorrected	2.01 (5.67)	1.55 (4.18)	1.47 (2.89)
		robust	Croux et al.	2.13 (6.12)	2.02 (6.53)	1.99 (4.55)
	Robust	mean	uncorrected	3.43 (10.89)	3.60 (12.35)	4.18 (10.71)
		robust	Croux et al.	3.58 (11.74)	3.63 (13.87)	4.26 (12.25)
Beta	LS	mean	uncorrected	−0.10 (−0.22)	0.34 (0.90)	0.32 (0.63)
		robust	Croux et al.	−0.48 (−1.13)	−0.44 (−1.41)	−0.71 (−1.77)
	Robust	mean	uncorrected	−2.72 (−7.27)	−3.33 (−12.05)	−4.26 (−12.70)
		robust	Croux et al.	−3.12 (−8.68)	−3.54 (−14.09)	−4.43 (−14.73)
Size	LS	mean	uncorrected	−0.06 (−1.05)	−0.00 (−0.03)	0.06 (0.95)
		robust	Croux et al.	−0.08 (−1.34)	−0.03 (−0.63)	0.03 (0.46)
	Robust	mean	uncorrected	−0.41 (−7.47)	−0.39 (−10.18)	−0.43 (−8.78)
		robust	Croux et al.	−0.41 (−7.47)	−0.40 (−10.39)	−0.44 (−9.04)
Inter.	LS	mean	uncorrected	−0.08 (−1.15)	−0.13 (−2.48)	−0.15 (−2.28)
		robust	Croux et al.	−0.04 (−0.60)	−0.08 (−1.73)	−0.09 (−1.53)
	Robust	mean	uncorrected	0.47 (7.56)	0.50 (12.06)	0.61 (11.80)
		robust	Croux et al.	0.47 (7.47)	0.51 (12.35)	0.62 (12.25)

The LS analysis finds that there was no relationship during 1963–1990 between returns, size, and beta on average once interactions are added to the model. During the periods 1963–2015 and 1980–2015, only the interaction term is significant, but the robust location estimate tells us this significance is driven by a small number of months. The robust regression, on the other hand, finds that all terms in the model are significant: the beta effect is once again negative, but now the size effect is negative, and the interaction term is positive. This implies that the relationship between returns and size varies with beta; likewise, the relationship between returns and beta varies with size.

We can visualize the relationship between returns, size, and beta by considering the "average" robust regression model over 1963–2015: from Table 4.30 this model is given by the equation

$$\begin{aligned}\text{returns} &= 3.63 - 3.54 \times \text{beta} - 0.40 \times \text{size} + 0.51 \times \text{size} \times \text{beta}\\ &= (3.63 - 3.54 \times \text{beta}) - (0.40 - 0.51 \times \text{beta}) \times \text{size}\\ &= (3.63 - 0.40 \times \text{size}) - (3.54 - 0.51 \times \text{size}) \times \text{beta}.\end{aligned}$$

Figure 4.34 plots this equation for fixed beta with varying size and fixed size with varying beta. For fixed beta (left panel) smaller than approximately $0.40/0.51 = 0.8$, returns decrease with increasing size. As beta approaches 0.8, the average regression line flattens out, meaning there is little dependence of returns on size. As beta increases past 0.8, the slope of the fitted line becomes positive, and average returns increase with size. On the other hand, for fixed size (right panel) less than $3.54/0.51 = 6.9$ (corresponding to a market capitalization of about 1 billion), returns decrease with increasing beta. Near a size of 6.9, returns and beta have no relationship. For larger stocks, returns increase with increasing beta. When beta is approximately 0.8, all the lines in the right panel intersect since the coefficient on size will be 0. Likewise, when size is approximately 6.9, all the lines in the left panel intersect since there is no dependence on beta.

Our results showing that the relationship between average returns and beta varies with firm size, and vice versa, are similar in nature to the returns-beta relationship found by

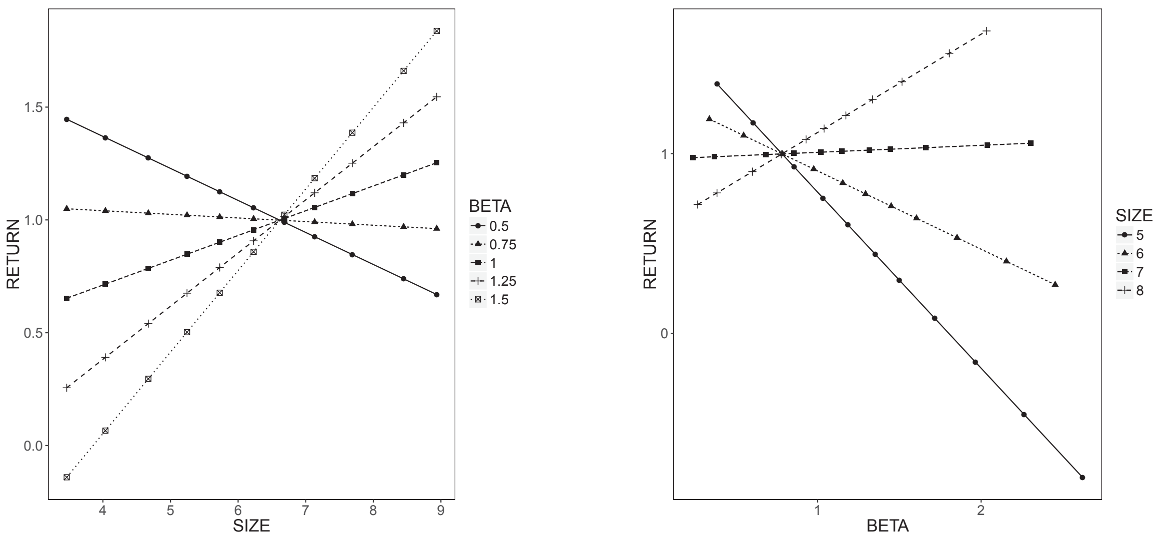


Figure 4.34: Average 99.9% efficient robust regression lines for size-beta interaction model. Left panel: average regression model of returns versus size for several fixed values of beta. Right panel: average model of returns versus beta for several fixed values of size.

Barnes and Hughes (2002) using quantile regression.[24] Barnes and Hughes found that for stocks whose return was near the median return for all stocks, the risk premium on beta was not significant. On the other hand, for stocks that overperformed or underperformed the conditional mean return, beta had a significant relationship with average returns: underperforming firms exhibited a risk penalty on beta while overperforming firms exhibited a risk premium on beta. Our results, together with those of Barnes and Hughes, indicate that the relationship between returns and beta is not the same for all firms. The nonlinear structure of the returns-beta relationship can be obscured, however, by outliers in the cross-sectional data and too narrow of a focus on the relationship at the conditional mean return.

4.6.4 Alternate Measures of Size

Berk (1995b,a, 1997, 2000) argued that the FF92 size factor, defined as the logarithm of

[24]We note in passing that quantile regression, while robust to returns outliers, is not as robust to outliers in the fundamental variables as our MM-regression, especially for larger quantiles.

market capitalization, created an errors-in-variables problem for estimating the relationship between average returns and firm size. Market capitalization is calculated using stock prices, which are subject to errors in the pricing model assumed by market participants (CAPM in this case). The negative relationship between average returns and size found by FF92, Berk argued, could have arisen solely from assuming, incorrectly, that FF92's size measure is known without error. He suggested that other measures of firm size, such as a firm's book equity and its book value of assets, were less noisy proxies for firm size.

Tables 4.31 and 4.32 show the results from our LS and robust cross-sectional regressions of average returns on the logarithm of book equity and the logarithm of the book value of assets, respectively, for each firm. Berk found no statistical evidence of a size premium using non-priced based measures of size. Our LS regressions generally support this finding.[25] Our robust regression, however, finds a significant and positive relationship between average returns and either size measure, consistently across the three time periods considered. This bolsters our earlier findings of a positive relationship between average returns and size as measured by the logarithm of firm market value. It further suggests that for most stocks average returns increase with firm size, regardless of how firm size is defined. A small number of stocks each month do not conform to this pattern, however, and are influential enough to obscure the relationship that holds for most stocks, most of the time.

[25]Berk (2000) reported an average slope of -0.099 with a t-statistic of -1.83 for the LS regression of returns on the logarithm of the book value of assets over the period July 1967–June 1987. Our average LS slope for the period July 1963–December 1990 agrees very well with Berk's result.

Table 4.31: Average slopes (t-statistics) from cross-sectional regressions of stock returns on the logarithm of book equity.

Factor	Regression	Location Estimate	t-Statistic	Book Equity 1963–1990	Book Equity 1963–2015	Book Equity 1980–2015
Intercept	LS	mean	uncorrected	1.49 (3.11)	1.60 (4.44)	1.43 (3.41)
		robust	Croux et al.	1.40 (3.06)	1.39 (4.09)	1.24 (3.25)
	Robust	mean	uncorrected	−0.41 (−0.93)	−1.01 (−3.33)	−1.64 (−5.00)
		robust	Croux et al.	−0.57 (−1.41)	−1.12 (−3.97)	−1.51 (−4.95)
ln(BE)	LS	mean	uncorrected	−0.08 (−1.34)	−0.08 (−1.97)	−0.04 (−0.77)
		robust	Croux et al.	−0.00 (−0.09)	−0.01 (−0.18)	0.02 (0.55)
	Robust	mean	uncorrected	0.26 (5.19)	0.33 (9.58)	0.43 (12.08)
		robust	Croux et al.	0.33 (7.63)	0.38 (13.19)	0.44 (14.91)

Table 4.32: Average slopes (t-statistics) from cross-sectional regressions of stock returns on the logarithm of the book value of assets.

				Book Value of Assets		
Factor	Regression	Location Estimate	t-Statistic	1963–1990	1963–2015	1980–2015
Intercept	LS	mean	uncorrected	1.60 (3.23)	1.67 (4.42)	1.45 (3.29)
		robust	Croux et al.	1.51 (3.16)	1.41 (3.99)	1.19 (3.02)
	Robust	mean	uncorrected	−0.47 (−1.04)	−1.12 (−3.55)	−1.82 (−5.30)
		robust	Croux et al.	−0.66 (−1.57)	−1.29 (−4.43)	−1.75 (−5.61)
ln(BVA)	LS	mean	uncorrected	−0.09 (−1.72)	−0.09 (−2.18)	−0.04 (−0.86)
		robust	Croux et al.	−0.03 (−0.52)	−0.01 (−0.41)	0.02 (0.60)
	Robust	mean	uncorrected	0.23 (4.88)	0.30 (9.12)	0.40 (11.50)
		robust	Croux et al.	0.30 (6.98)	0.35 (12.68)	0.41 (14.60)

4.6.5 Summary

The time series of slopes from the single factor cross-sectional regressions exhibit evidence of heavy-tailed distributions, serial correlation, and heteroskedasticity. Sample averages and ordinary t-statistics may be distorted by anomalous months or regime shifts in time series of slopes. A "robust mean" can reveal when outlying months have undue influence on the sample average slope. The corresponding t-statistic should be calculated using the approach of Croux et al. (2003b) to account for potential heteroskedasticity and autocorrelation in the time series of slopes.

The robust mean of the LS slopes from the single factor model for size indicates that the negative relationship between average returns and size suggested by the sample mean of the LS slopes results from a small number of months with very negative slopes on the size factor. This agrees with our earlier finding that the LS size premium was driven by a strong negative January effect. On the other hand, our robust regression slopes still point to a positive relationship between average returns and firm size for nearly all stocks and nearly all months, even after accounting for potential time series effects and anomalous months.

The LS relationship between average returns and beta is negative, once outlying months are excluded by the robust mean. This brings the LS results closer to agreement with the robust results. As we saw in Section 4.4, there are a small number of months with large positive slopes on beta that obscure the generally negative relationship between returns and beta.

Overall the robust regression slopes are more reliable than the LS slopes. Our conclusions from the robust regression slopes are generally the same whether we consider the sample mean or the robust mean of slopes, and/or whether t-statistics are HAC-corrected. The combination of robust cross-sectional regression, robust means of slopes, and Croux et al. HAC-corrected t-statistics provides a very reliable way to make inferences about risk premia in the presence of outliers, fat-tailed distributions, and time series effects.

The robust regression results show that the risk premium on the unaltered earnings-to-price factor was significant for nearly all stocks over all the time periods. It is not subsumed by the size and beta factors as FF92 claimed based on the LS regression results. Earnings-to-price is mainly meaningful for smaller stocks, however.

The relationship between average returns, size, and beta is more complicated than previous models have indicated, as there is a non-trivial interaction between size and beta. This implies that risk premium on beta varies with firm size, and that risk premium on size varies with beta. For small stocks, average returns decrease with increasing beta, while for large stocks returns increase with increasing beta. For moderately sized stocks there is no evidence of a risk premium on beta. The small stock relationship between returns and beta is identical to the well-known "low-beta anomaly". Our findings suggest that that this anomaly may be more representative of a small-stock phenomenon than a broad market phenomenon.

Finally, we find that the positive relationship between average returns and firm size found using cross-sectional robust regression still holds if we use non-price measures of firm size as suggested by Berk.

4.7 Robustness to Choice of Efficiency and Choice of Robust Loss Function

It is important to know how the robust regression slope statistics change as the efficiency of the regression estimator is changed: lower efficiency yields a smaller maximum bias due to outliers by rejecting larger fractions of the observations in the regression, at the price of a less efficient estimator when the returns are normally distributed. Comparing the slopes from regressions with varying efficiencies gives us a rough understanding of the number of outliers driving the LS results. If the robust regressions disagree with the LS regressions even at high efficiencies, it suggests there are extreme outliers in the data with strong influence on the LS results. On the other hand, when the regressions at various efficiencies agree with the LS results in sign, magnitude, and significance, it gives us confidence that the risk premia under investigation are not substantially driven by outliers.

We also consider a version of the MM-estimator that uses the bisquare loss function instead of the Yohai-Zamar optimal loss function (as discussed in Section 4.2). The bisquare loss function is a common choice for M- and MM-estimation due to its early availability in the literature and software. We study the performance of this MM-estimator relative to that of the optimal loss function-based MM-estimator for each of the three normal distribution efficiencies. Our results will also illustrate how the performance of the bisquare-based MM-estimator changes with the choice of efficiency.

For brevity, we analyze only the univariate beta and size models here. We report robust means of monthly cross-sectional slopes, the latter using the Croux et al. approach for HAC-corrected t-statistics that was discussed in Section 4.6. We omit the sample means and uncorrected t-statistics here for space considerations. Our conclusions from sample means and uncorrected t-statistics are largely the same as those from the robust means and HAC-corrected t-statistics.

4.7.1 Beta

Table 4.33 presents the sample means of monthly slopes, with uncorrected t-statistics, and robust means of slopes with Croux et al.-corrected t-statistics, for the regression of returns on post-ranking betas. (The LS and robust optimal 99.9% results were previously discussed in Sections 4.4.3 and 4.6.1.) The robust regressions all yield negative robust means of slopes for all time periods. Generally as efficiency decreases, the mean slope becomes more negative, and the t-statistic becomes more significant. We also note that the strength of the beta relationship increases over the longer time horizons, uniformly across all robust regression results. Thus our robust regression results from Section 4.4.3 and 4.6.1 are robust to the choice of regression efficiency.

An empirical asset pricing study will usually arrive at the same conclusions using any of the three choices of efficiency presented here. We recommend that practitioners estimate a given factor model using robust regression with all three efficiencies from largest to smallest. If the robust mean slopes from the robust regressions do not exhibit large differences from the robust mean slopes from the LS regression, then the factor model results are unlikely to be driven by firm-level influential outliers. Furthermore, if the sample mean slopes from the robust and LS regressions are also in agreement, then the factor model results are not driven by influential months, and the LS results can be used safely.

4.7.2 Size

Table 4.34 shows the regression results for the size premium. As we have previously discussed in Section 4.6, the significance of the sample mean of the LS slopes for size is driven by a small number of (mostly January) months. The robust mean of the LS slopes, which excludes a small number of outlying months, supports our prior finding. The robust regressions, on the other hand, are consistent across efficiencies and choice of loss functions in demonstrating that average returns increase with increasing size.

Table 4.33: Robust means of slopes (with Croux et al. t-statistics) from cross-sectional regressions of stock returns on beta using varying efficiencies for robust regression and varying types of robust regressions. The regressions used are as follows: least squares ("LS"), our robust MM-regression with Yohai-Zamar optimal loss function and 95%, 99%, and 99.9% efficiencies ("Rob. Opt. 95%", "Rob. Opt. 99%", and "Rob. Opt. 99.99%", respectively), and the robust MM-regression with the bisquare loss function and 95%, 99%, and 99.9% efficiencies ("Rob. Bsq. 95%", "Rob. Bsq. 99%", and "Rob. Bsq. 99.9%").

				Beta	
Regression	Location Estimate	t-Statistic	1963–1990	1963–2015	1980–2015
LS	robust	Croux et al.	−0.12 (−0.35)	−0.13 (−0.53)	−0.43 (−1.52)
Rob. Opt. 95%	robust	Croux et al.	−1.83 (−6.29)	−1.86 (−9.74)	−2.16 (−9.86)
Rob. Opt. 99%	robust	Croux et al.	−1.60 (−5.35)	−1.65 (−8.38)	−1.98 (−8.70)
Rob. Opt. 99.9%	robust	Croux et al.	−1.32 (−4.30)	−1.41 (−6.94)	−1.77 (−7.49)
Rob. Bsq. 95%	robust	Croux et al.	−1.67 (−5.65)	−1.73 (−8.88)	−2.05 (−9.17)
Rob. Bsq. 99%	robust	Croux et al.	−1.26 (−4.06)	−1.35 (−6.57)	−1.71 (−7.17)
Rob. Bsq. 99.9%	robust	Croux et al.	−0.74 (−2.30)	−0.85 (−3.86)	−1.23 (−4.81)

Table 4.34: Average slopes (t-statistics) from cross-sectional regressions of stock returns on size. The table setup is identical to that of Table 4.33.

Regression	Location Estimate	t-Statistic	Size 1963–1990	Size 1963–2015	Size 1980–2015
LS	robust	Croux et al.	−0.06 (−1.18)	−0.07 (−1.87)	−0.05 (−1.33)
Rob. Opt. 95%	robust	Croux et al.	0.39 (9.42)	0.43 (16.09)	0.47 (16.55)
Rob. Opt. 99%	robust	Croux et al.	0.34 (7.93)	0.39 (13.92)	0.44 (14.95)
Rob. Opt. 99.9%	robust	Croux et al.	0.28 (6.29)	0.34 (11.62)	0.40 (12.95)
Rob. Bsq. 95%	robust	Croux et al.	0.36 (8.52)	0.40 (14.75)	0.44 (15.62)
Rob. Bsq. 99%	robust	Croux et al.	0.27 (6.02)	0.32 (11.10)	0.38 (12.38)
Rob. Bsq. 99.9%	robust	Croux et al.	0.14 (2.87)	0.19 (6.15)	0.25 (7.42)

4.8 Concluding Discussion

Overall, our robust cross-sectional regressions lead to different conclusions than the traditional LS cross-sectional regressions for most of the factor models considered in FF92 over their time period 1963–1990 and over the longer time period 1963–2015. Our conclusions from Sections 4.4–4.6 remain largely unchanged if we use the bisquare loss function in our MM-estimator instead of the optimal loss function in the cross-sectional regression. Furthermore, efficiencies of 95%, 99%, or 99.9% all yield similar results for either loss function. This is good evidence that our conclusions from the robust regressions are not due to a particular choice of MM-estimator, subjective removal of observations, or "data mining".

Our robust regression analysis reaffirms that average stock returns increase with increasing book-to-market ratio. This conclusion is supported by cross-sectional LS and robust regressions over the period 1963–2015. However, we also confirm and enhance the findings of Loughran (1997), namely, that the book-to-market effect is mainly a small-firm effect. We find evidence that the book-to-market effect is still positive and strong for small firms through 2015. For moderately-sized firms, the effect appears to have dissipated after 1980, possibly due to publication of numerous papers about the anomaly and subsequent actions by investors to exploit the effect. We also confirm a January effect for book-to-market that in fact exists in the entire first quarter. The January effect does not persist after 1980, however.

In contrast to many previous studies on the size effect, we find, via robust regression, that average returns increase with increasing firm size for nearly all stocks. The relationship remains positive using robust regression if we use non-priced based measures of size as suggested by Berk. We further confirm the existence of a strong negative relationship between returns and size in the month of January that still holds even with our robust regression analysis. We show, however, that the size relationship is positive from March through December for most stocks. The overall negative relationship between average returns and size observed by other researchers is attributable to small stocks with unusually large returns. We can observe a positive relationship between average returns and increasing size in size-sorted portfolios if we remove a small percentage of small firms each month with extreme returns.

Our findings on the size anomaly suggest that the "small stock premium" found by other researchers can only be captured by predicting which small stocks will experience abnormally large returns, or by holding all small stocks, including those that may be relatively illiquid and capacity-constrained. Given the difficulty of implementing either of these approaches, it seems unrealistic to rely on the "small stock premium" to generate returns.

Our cross-sectional regression analysis finds a complex interaction between the size and beta factors. The relationship between average returns and beta varies within size segments, and vice versa. For beta values near one or moderately-sized stocks, the corresponding returns relationships can be quite flat, but these represent inflection points in the returns-beta-size relationship rather than typical behavior. This suggests that the standard approach of looking only at the linear normal distribution conditional mean relationship between returns and beta across all stocks obscures interesting structure in the returns-beta relationship.

Another limitation of the FF92 size-beta analysis is the limited number of "post-ranking beta" values available in the cross-sectional regressions. The size-beta sorting approach used by Fama and French produces only 100 unique values of beta (10 within each size decile), and the post-ranking beta for a stock is held constant from July 1 of year t to June 30 of year $t + 1$. The size-beta sorts were a simple means of dealing with a complex errors-in-variables problem. It would be of interest to understand, using alternative techniques, how our conclusions about the relationship between returns, size, and beta would change if we used firm-level betas in our analysis instead of the post-ranking betas. Other techniques, such as orthogonal or errors-in-variables regression, would allow us to use the firm betas (estimated on a rolling basis) directly in a regression. A robust version of orthogonal regression is also available to mitigate the impact of outliers. Another approach, pioneered by Cederburg and O'Doherty (2015), uses a hierarchical model and Bayesian techniques to deal with the errors-in-variables issue.

Our results show it is unrealistic to assume that an asset pricing model will or should hold for all stocks. We have shown several instances of relationships estimated using all of the data that turned out to be highly influenced by a small fraction of it. We believe that it is more reasonable to look for models that work for "nearly all stocks", and accept that a small number of stocks will not behave according to this model. After all, a model is merely an approximation to reality, and a model that explains most of the data well is certainly preferable to one that is distorted by a small but "vocal" minority of outliers.

In the spirit of this approach, it would be interesting to reconsider the classic 3-factor model of Fama and French (1993) using the robust statistical approach. Fama and French used the findings of their 1992 study to develop the well-known model presented in their 1993 paper. Many similar studies have been done over the years, including Fama and French (2015) which extended Fama and French (1993) with two new factors, profitability and quality. Given our findings on the size effect, in particular on the returns on size-sorted portfolios, it would be interesting to revisit that study using robust methods. By excluding a small number of small stocks each month with unusually large returns, we suspect a "robust" BMS ("Big Minus Small") factor would behave rather differently from the traditional SMB factor.

Overall, we find that the use of classical procedures like LS regressions, sample means, and t-statistics uncorrected for serial correlation and heteroskedasticity can lead to very misleading conclusions about asset pricing relationships. In most of the models we considered in this study, we found that small stocks had a strong influence on the results of asset pricing tests. In several cases limiting the influence of these small stocks led to strikingly different conclusions about pricing anomalies. The advantage of our robust approach is the ease with which one can uncover observations with undue influence on the statistical inference and the associated conclusions about financial markets. Robust statistical methods provide a more reliable inference that is not subject to the influence of a small fraction of outliers.

To summarize, a theoretically justified robust regression method provides a very valuable complement to LS cross-sectional regression for empirical asset pricing studies. A long time ago, when very little research on robust regression methods existed, John Tukey (1979) expressed the value of robust methods quite well: "Just which robust and resistant methods you use is NOT important—what IS important is that you use SOME. It is perfectly proper to use both classical and robust/resistant methods routinely, and only worry when they differ enough to matter. BUT when they differ, you should think HARD.”

APPENDIX

4.A Croux et al. (2003) Robust Standard Errors

This appendix provides a brief explanation of the Croux et al. (2003a) HAC-corrected standard errors used in Section 4.6 for the reader's convenience. Full details can be found in the referenced paper.

Suppose we estimate a linear regression model via MM-regression with loss function $\rho(r)$ and corresponding derivative $\psi(r)$. Let y_t and $\mathbf{X}_t$ be the observations and regressors at time t, and define the standardized residuals

$$r_t = \frac{y_t - \mathbf{X}_t^T \tilde{\beta}_{MM}}{\tilde{\sigma}_S} \qquad \text{and} \qquad r_{0t} = \frac{y_t - \mathbf{X}_t^T \tilde{\beta}_S}{\tilde{\sigma}_S},$$

where $\tilde{\beta}_{MM}$ is the final estimate of the regression coefficients, $\tilde{\beta}_S$ is an initial estimate of the coefficients, and $\tilde{\sigma}_S$ is a residual scale estimate. Furthermore, define

$$\psi_t = \psi(r_t), \quad \rho_t = \rho(r_t), \quad \text{and} \quad \rho_{0t} = \rho(r_{0t})$$

for notational convenience.

Croux et al. (2003a) derives the HAC correction for the MM-estimate $\tilde{\beta}_{MM}$ by recasting the MM-regression procedure as a generalized method of moments (GMM) problem. Let $\theta = (\beta_{MM}^T, \beta_S^T, \sigma_S)$ be the vector of true model parameters, with $\tilde{\theta} = (\tilde{\beta}_{MM}^T, \tilde{\beta}_S^T, \tilde{\sigma}_S)^T$ the corresponding vector of estimates. If we define

$$m_t(\theta) = \begin{pmatrix} \psi_t \mathbf{X}_t \\ \rho_{0t}^T \mathbf{X}_t \\ \rho_{0t} - b \end{pmatrix},$$

then the MM-estimate (and the initial S-estimates) are the solution to the GMM problem

$$\frac{1}{T} \sum_{t=1}^{T} m_t(\tilde{\theta}) = 0.$$

Hansen (1982) established asymptotic normality for GMM estimates: we have

$$\sqrt{T}\left(\tilde{\theta}-\theta\right) \rightarrow_d N(0,V),$$

where the variance matrix V has the form $(G^T\Omega G)^{-1}$, with

$$G = E\left[\frac{\partial m_t(\theta)}{\partial \theta^T}\right] \qquad \text{and} \qquad \Omega = \sum_{j=-\infty}^{\infty} E\left[m_t(\theta)m_{t-j}(\theta)\right].$$

The HAC-corrected asymptotic variance of the MM-estimate $\tilde{\beta}_{MM}$ will be the upper $p\times p$ submatrix of V by our construction of $\tilde{\theta}$. Equations 3.8 and 3.9 of Croux et al. (2003a) provide explicit expressions for the asymptotic variance $Avar(\tilde{\beta}_{MM})$ of $\tilde{\beta}_{MM}$:

$$Avar(\tilde{\beta}_{MM}) = A\sum_{j=-\infty}^{\infty} E\left(\psi_t\psi_{t-j}\mathbf{X}_t\mathbf{X}_{t-j}^T\right)A - a\sum_{j=-\infty}^{\infty} E\left(\rho_{0t}\psi_{t-j}\mathbf{X}_{t-j}^T\right)A - \\ A\sum_{j=-\infty}^{\infty} E\left(\psi_t\rho_{0,t-j}\mathbf{X}_t\right)a^T + \sum_{j=-\infty}^{\infty} E\left(\rho_{0t}\rho_{0,t-j}-b^2\right)aa^T, \quad (4.8)$$

with

$$A = \sigma\left[E(\psi_t^T\mathbf{X}_t\mathbf{X}_t^T)\right]^{-1} \qquad \text{and} \qquad a = A\frac{E(\psi_t^T\mathbf{X}_t r_t)}{E(\rho_{0t}^T r_{it})}.$$

The sample version of (4.8) is obtained by replacing expectations in (4.8) with sample means, and replacing infinite sums with truncated weighted sums. The weights used are the same Bartlett weights used in the Newey and West (1987) HAC variance estimate:

$$w_j = 1 - \frac{|j|}{q+1},$$

where $q = q(T)$ is the index at which the sums are truncated, and $q(T)\rightarrow\infty$ at a slow rate in T.

Thus the sample variance is given by

$$\widetilde{Avar}(\tilde{\beta}_{MM}) = \tilde{A}\left[\frac{1}{T}\sum_{j=-q}^{q}\sum_{t=1}^{T}\left(w_j\psi_t\psi_{t-j}\mathbf{X}_t\mathbf{X}_{t-j}^T\right)\right]\tilde{A}-$$
$$\tilde{a}\left[\frac{1}{T}\sum_{j=-q}^{q}\sum_{t=1}^{T}\left(\rho_{0t}\psi_{t-j}\mathbf{X}_{t-j}^T\right)\right]\tilde{A}-$$
$$\tilde{A}\left[\frac{1}{T}\sum_{j=-q}^{q}\sum_{t=1}^{T}\left(\psi_t\rho_{0,t-j}\mathbf{X}_t\right)\right]\tilde{a}^T+$$
$$\left[\frac{1}{T}\sum_{j=-q}^{q}\sum_{t=1}^{T}\left(\rho_{0t}\rho_{0,t-j}-b^2\right)\right]\tilde{a}\tilde{a}^T$$

with

$$\tilde{A} = \tilde{\sigma}\left(\frac{1}{T}\sum_{t=1}^{T}\psi_t\mathbf{X}_t\mathbf{X}_t^T\right)^{-1}$$
$$\tilde{a} = \tilde{A}\frac{\sum_{t=1}^{T}\psi_t^T\mathbf{X}_t r_t}{\sum_{t=1}^{T}\rho_{0t}^T r_{it}}.$$

The standard error of each coefficient $\tilde{\beta}_{MM,j}$ is then computed using the diagonal elements of the asymptotic variance matrix:

$$\widetilde{SE}(\tilde{\beta}_{MM,j}) = \sqrt{\frac{\widetilde{Avar}(\tilde{\beta}_{MM})_{jj}}{T}}.$$

4.B Statistics on Rejected Outliers

Tables 4.35, 4.36 and 4.37 present summary statistics over time on the number of outliers rejected each month in the regressions on book-to-market, size, and beta, respectively, over the time period 1963–2015, as well as whether they are positive or negative outliers. Figures 4.35, 4.36, and 4.37 show the corresponding kernel density estimates of the distributions of the number of rejected observations over time for each model. Positive outliers are much more common than negative outliers in all three models. Across the entire period and consistently across the three models considered here, our robust regression rejects only 1.4–1.5% of stocks

Table 4.35: Statistics on percentages of observations rejected by our robust regression of returns on book-to-market, 1963–2015. The “Positive Outliers” column reports statistics over time on the monthly percentages of observations whose scaled residual was greater than 3, while the “Negative Outliers” column reports statistics over time for the percentage of observations with a scaled residual less than −3. The “All Outliers” column gives statistics over time for all outliers, positive and negative.

	Positive Outliers	Negative Outliers	All Outliers
Minimum	0.00%	0.00%	0.00%
1st Quartile	0.66%	0.06%	0.90%
Median	1.10%	0.21%	1.42%
Mean	1.20%	0.26%	1.46%
3rd Quartile	1.59%	0.39%	1.92%
Maximum	6.00%	1.73%	6.00%

Table 4.36: Statistics on percentages of observations rejected by our robust regression of returns on size, 1963–2015. The table setup is the same as in Table 4.35.

	Positive Outliers	Negative Outliers	All Outliers
Minimum	0.00%	0.00%	0.00%
1st Quartile	0.73%	0.06%	0.95%
Median	1.20%	0.21%	1.54%
Mean	1.30%	0.27%	1.54%
3rd Quartile	1.73%	0.40%	2.11%
Maximum	6.12%	1.77 %	6.12%

each month, on average. Thus our robust regression is a very conservative way of making reliable inferences in the presence of outliers.

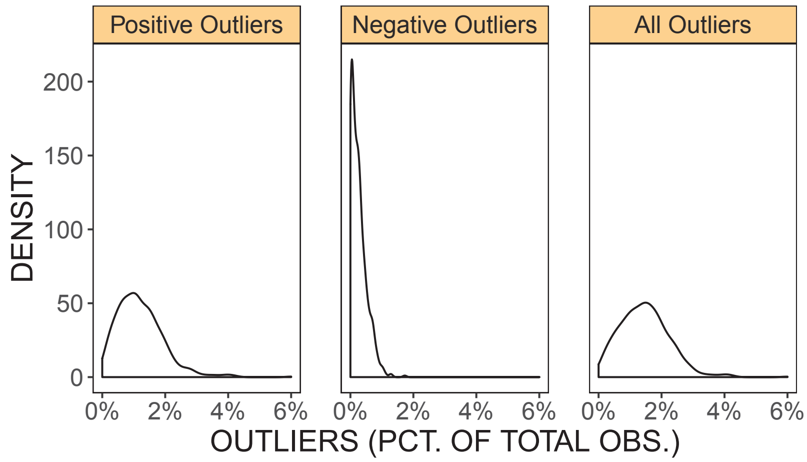


Figure 4.35: Kernel density estimates of the distribution of the number of positive, negative, and total outliers rejected by our robust regression of returns on book-to-market, 1963–2015.

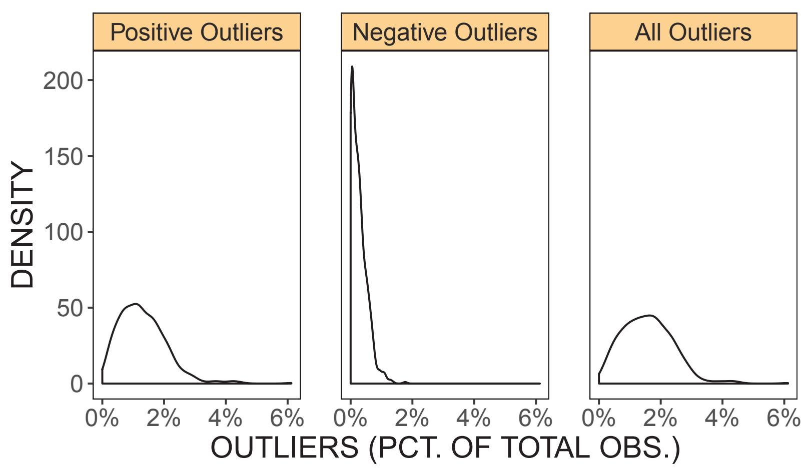


Figure 4.36: Kernel density estimates of the distribution of the number of positive, negative, and total outliers rejected by our robust regression of returns on size, 1963–2015.

Table 4.37: Statistics on percentages of observations rejected by our robust regression of returns on beta, 1963–2015. The table setup is the same as in Table 4.35.

	Positive Outliers	Negative Outliers	All Outliers
Minimum	0.00%	0.00%	0.00%
1st Quartile	0.76%	0.06%	0.98%
Median	1.17%	0.22%	1.49%
Mean	1.28%	0.27%	1.55%
3rd Quartile	1.68%	0.42%	2.04%
Maximum	5.98%	1.57 %	5.98%

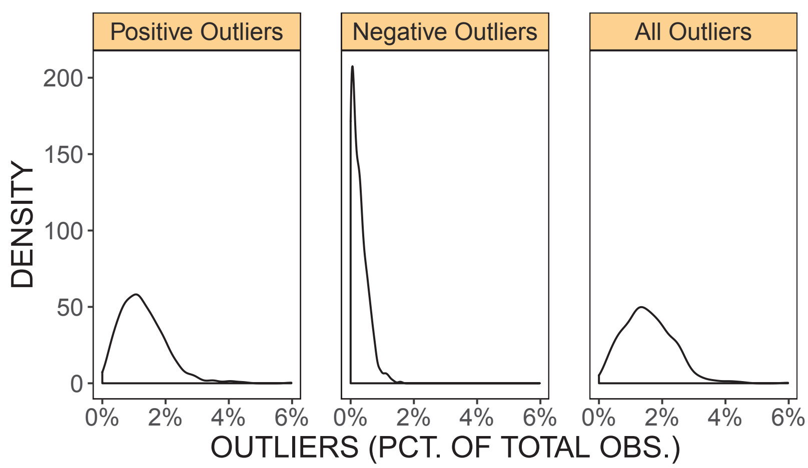


Figure 4.37: Kernel density estimates of the distribution of the number of positive, negative, and total outliers rejected by our robust regression of returns on beta, 1963–2015.

Table 4.38: Average return (including dividends) each month by size-beta decile group, 1963–1990.

	ALL	β-01	β-02	β-03	β-04	β-05	β-06	β-07	β-08	β-09	β-10
ALL	1.23	1.30	1.33	1.31	1.32	1.33	1.31	1.25	1.18	1.18	1.06
ME01	1.47	1.49	1.64	1.65	1.57	1.65	1.42	1.55	1.43	1.46	1.29
ME02	1.15	1.13	1.30	1.21	1.38	1.18	1.59	1.25	1.33	1.05	0.70
ME03	1.23	1.29	1.32	1.09	1.53	1.46	1.33	1.41	1.62	1.05	0.67
ME04	1.22	1.29	1.25	1.37	1.47	1.25	1.15	1.28	1.09	1.14	1.10
ME05	1.22	1.22	1.44	1.29	1.20	1.40	1.16	1.28	1.12	1.40	0.99
ME06	1.11	1.03	1.37	1.24	1.24	1.34	1.27	0.96	0.94	1.00	0.91
ME07	1.10	1.07	1.33	1.23	1.27	0.96	1.05	1.08	1.04	1.09	0.98
ME08	1.07	1.08	1.17	1.20	1.11	1.33	1.24	0.92	0.82	0.95	0.98
ME09	0.94	0.95	0.93	1.05	1.02	1.07	1.17	0.97	0.78	0.92	0.61
ME10	0.89	1.05	0.97	1.11	0.85	0.80	1.01	0.84	0.75	0.89	0.62

4.C Replication of the Fama-French 1992 Study

This appendix provides additional details on the replication of the Fama-French data set for the period 1963–1990. Tables 4.38, 4.39, 4.40, and 4.41 show the time-series average returns, post-ranking betas, size, and number of firms within size and pre-ranking beta deciles. These tables should be compared to Table I (and its caption) of Fama and French (1992). Figures 4.38-4.40 show how well we were able to replicate the three components of Table 1 of Fama and French (1992).

Our average postbeta and size results are very close to those of FF92. Our average portfolio returns are typically close but can differ by as much as 60 basis points in absolute value

Table 4.39: Average postbeta each month by size-beta decile group, 1963–1990.

	ALL	β-01	β-02	β-03	β-04	β-05	β-06	β-07	β-08	β-09	β-10
ALL	1.31	0.90	1.03	1.13	1.21	1.28	1.32	1.38	1.47	1.54	1.71
ME01	1.44	1.10	1.21	1.28	1.36	1.42	1.43	1.51	1.59	1.65	1.76
ME02	1.40	0.95	1.16	1.21	1.31	1.32	1.45	1.47	1.59	1.63	1.72
ME03	1.35	0.85	1.09	1.16	1.22	1.34	1.31	1.40	1.53	1.61	1.81
ME04	1.32	0.80	1.04	1.15	1.17	1.31	1.37	1.39	1.51	1.60	1.75
ME05	1.26	0.62	0.91	1.15	1.15	1.19	1.24	1.42	1.49	1.59	1.66
ME06	1.18	0.56	0.76	1.04	1.07	1.23	1.28	1.31	1.34	1.50	1.64
ME07	1.17	0.60	0.89	1.05	1.08	1.17	1.24	1.30	1.33	1.32	1.67
ME08	1.07	0.50	0.67	0.92	1.01	1.09	1.18	1.17	1.19	1.30	1.60
ME09	1.01	0.54	0.77	0.84	0.93	1.04	1.09	1.11	1.19	1.23	1.40
ME10	0.94	0.55	0.69	0.79	0.94	0.93	0.94	0.99	1.07	1.14	1.39

Table 4.40: Average size each month by size-beta decile group, 1963–1990.

	ALL	β-01	β-02	β-03	β-04	β-05	β-06	β-07	β-08	β-09	β-10
ALL	4.01	3.67	4.14	4.30	4.19	4.21	4.20	4.23	4.12	4.06	3.71
ME01	2.25	2.14	2.28	2.29	2.31	2.30	2.33	2.32	2.30	2.30	2.20
ME02	3.73	3.72	3.73	3.74	3.73	3.71	3.74	3.73	3.73	3.72	3.73
ME03	4.21	4.21	4.21	4.20	4.21	4.20	4.20	4.21	4.21	4.20	4.20
ME04	4.63	4.64	4.63	4.62	4.63	4.63	4.63	4.61	4.63	4.63	4.62
ME05	5.04	5.04	5.04	5.05	5.05	5.04	5.04	5.03	5.04	5.03	5.02
ME06	5.44	5.44	5.45	5.45	5.46	5.45	5.44	5.44	5.44	5.44	5.44
ME07	5.88	5.88	5.89	5.88	5.89	5.88	5.87	5.87	5.86	5.87	5.87
ME08	6.37	6.39	6.37	6.38	6.36	6.37	6.37	6.36	6.35	6.35	6.36
ME09	6.91	6.91	6.94	6.92	6.93	6.90	6.91	6.91	6.91	6.91	6.88
ME10	8.02	8.01	8.17	8.13	8.11	8.12	8.12	8.03	7.90	7.86	7.74

Table 4.41: Average number of firms each month by size-beta decile group, 1963–1990.

	β-01	β-02	β-03	β-04	β-05	β-06	β-07	β-08	β-09	β-10
ME01	202	108	82	90	90	95	85	100	107	193
ME02	28	25	23	21	22	23	26	24	27	39
ME03	18	17	16	16	17	18	17	20	19	26
ME04	16	15	16	15	15	17	15	16	16	20
ME05	13	14	14	14	15	14	15	13	16	17
ME06	13	12	13	12	13	13	13	13	13	14
ME07	12	12	12	12	12	12	12	12	12	13
ME08	12	11	12	12	12	12	11	12	12	12
ME09	12	11	11	11	11	11	11	11	11	12
ME10	12	11	11	11	11	11	11	11	11	11

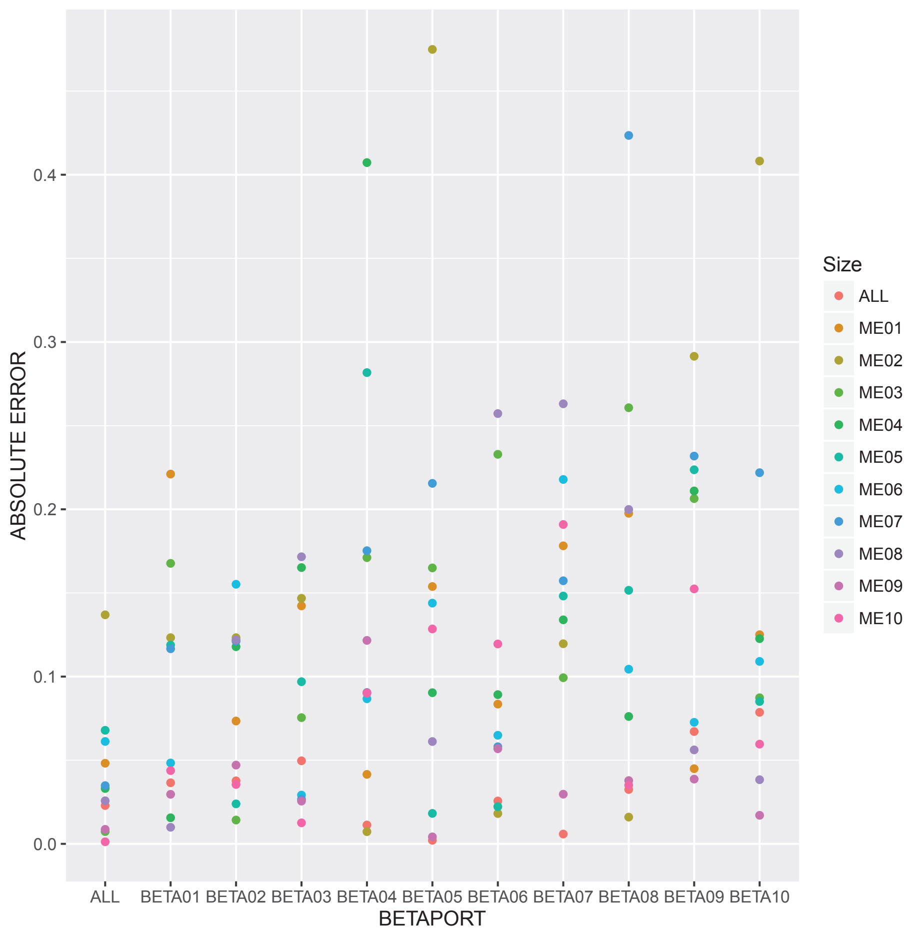


Figure 4.38: Absolute error in replicating the return statistics presented in Table 1a of Fama and French (1992). The chart plots the absolute differences between our time-series average returns on each size-beta portfolio and those of FF92 for the corresponding portfolio.

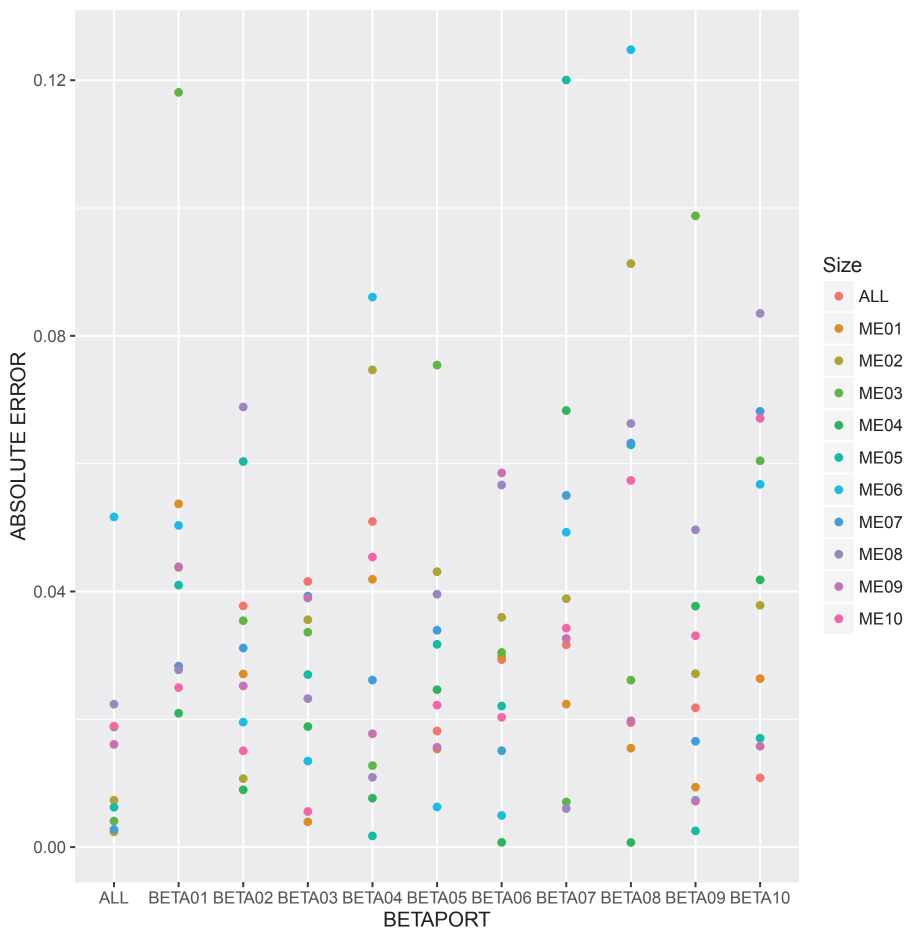


Figure 4.39: Absolute error in replicating the post-ranking beta statistics presented in Table 1b of Fama and French (1992).

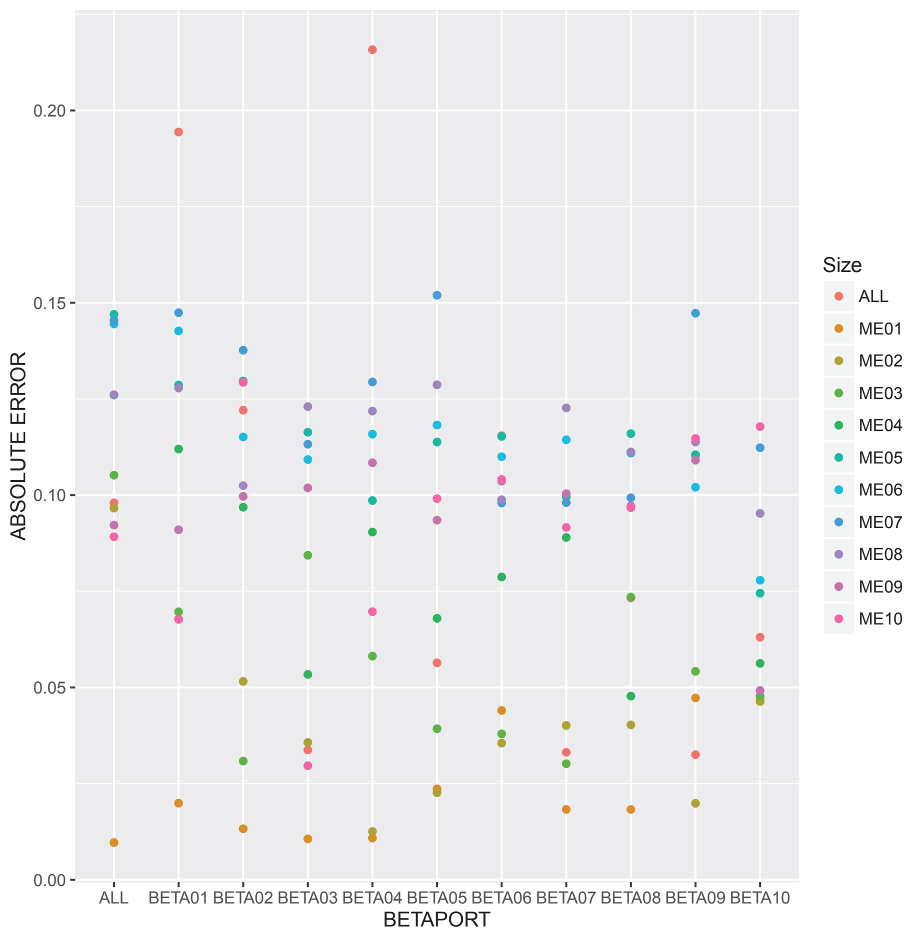


Figure 4.40: Absolute error in replicating the average size statistics presented in Table 1c of Fama and French (1992).

for some combinations of size and beta. Two possible explanations for the difference are (a) CRSP and Compustat have updated and backfilled their databases, so we are working with slightly different data; and (b) we have included delisting returns in our returns calculations, which Fama and French might not have done. The former cause is difficult to investigate without access to the CRSP and Compustat databases as they existed at the time of FF92. As for the latter cause, Table 4.42 shows the average returns on the size-beta portfolios without delisting returns. Figure 4.41 shows how well our average returns without delisting match up to the returns shown in Table 1a of FF92. It does not seem that including or excluding delisting returns yields meaningful differences in the average returns on the size-beta portfolios. We have elected to use delisting returns in our calculations in the main body of the paper, as this helps alleviate survivorship bias in our analysis.

Table 4.43 shows the standard deviations of the monthly returns within each of the size and pre-ranking beta deciles. Unsurprisingly, the volatility of the size-beta portfolio returns generally decreases with size and increases with beta.

We also calculated the pre- and post-ranking betas using our robust regression. Table 4.44 shows the time-series average returns within each size-robust beta group, while Table 4.45 shows the time-series average post-ranking betas. Table 4.46 shows the difference in the average beta estimates (LS - robust). The average returns on the portfolios formed using robust pre-ranking betas are not very different from those computed using the LS betas. The average post-ranking robust betas are generally not too different from their LS counterparts. Overall, this suggests the beta decile breakpoints are not being driven by extreme outliers, though we note there are larger differences between the LS and robust beta estimates for the smallest and largest beta deciles.

Tables 4.47 and 4.48 show average returns, post-ranking betas, sizes, and accounting variables within size deciles and beta deciles. Our results compare favorably to those presented in Table II of FF92.

Table 4.42: Average return (including dividends but not delisting returns) each month by size-beta decile group, 1963–1990.

	ALL	β-01	β-02	β-03	β-04	β-05	β-06	β-07	β-08	β-09	β-10
ALL	1.23	1.30	1.33	1.31	1.32	1.33	1.31	1.23	1.18	1.18	1.06
ME01	1.47	1.49	1.64	1.65	1.57	1.65	1.42	1.50	1.43	1.45	1.30
ME02	1.15	1.13	1.30	1.21	1.38	1.17	1.59	1.25	1.33	1.05	0.70
ME03	1.23	1.29	1.32	1.09	1.53	1.45	1.33	1.41	1.62	1.05	0.67
ME04	1.21	1.29	1.23	1.37	1.46	1.25	1.15	1.28	1.09	1.14	1.10
ME05	1.22	1.22	1.44	1.29	1.20	1.40	1.16	1.28	1.12	1.40	1.00
ME06	1.11	1.03	1.37	1.24	1.24	1.34	1.27	0.96	0.94	1.00	0.91
ME07	1.10	1.07	1.33	1.23	1.27	0.96	1.05	1.08	1.04	1.09	0.98
ME08	1.07	1.08	1.17	1.20	1.11	1.33	1.24	0.92	0.82	0.95	0.98
ME09	0.94	0.95	0.93	1.04	1.02	1.07	1.17	0.97	0.78	0.92	0.61
ME10	0.89	1.05	0.97	1.11	0.85	0.80	1.01	0.84	0.75	0.89	0.62

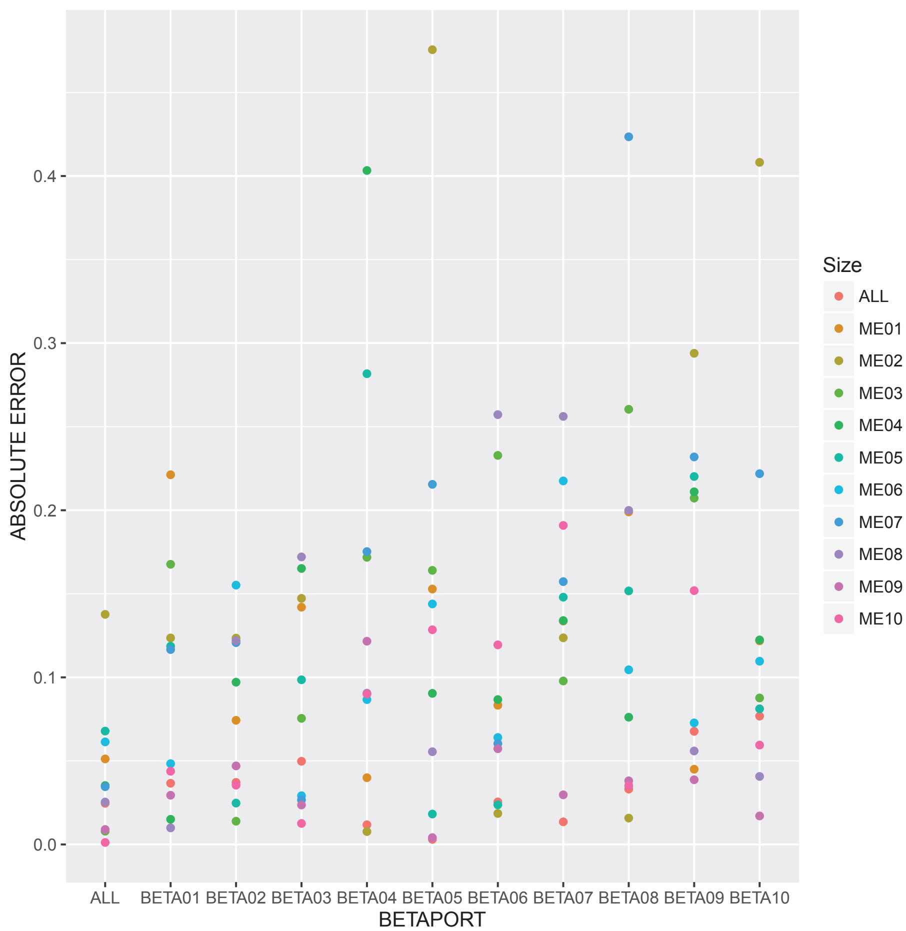


Figure 4.41: Absolute error in replicating the return statistics presented in Table 1a of Fama and French (1992) with delisting returns omitted.

Table 4.43: Standard deviation (over time) of returns each month by size-beta decile group, 1963–1990. The “ALL” row presents the standard deviation of returns on equally-weighted portfolios of all stocks within a beta decile; the “ALL” column presents the standard deviation for equally-weighted portfolios of all stocks within a size decile. The “ALL”-“ALL” cell in the upper-left corner shows the standard deviation of returns on an equally-weighted portfolio of all stocks.

	ALL	β-01	β-02	β-03	β-04	β-05	β-06	β-07	β-08	β-09	β-10
ALL	5.95	4.29	4.67	5.11	5.54	5.82	6.01	6.37	6.68	7.22	8.16
ME01	7.11	5.37	6.11	6.44	6.86	7.18	7.17	7.83	8.01	8.36	9.08
ME02	6.62	5.03	5.92	6.12	6.67	6.55	7.24	7.50	7.56	8.12	8.89
ME03	6.39	4.86	5.41	5.99	6.23	6.76	6.57	7.08	7.57	7.88	9.14
ME04	6.11	4.49	5.25	5.59	6.21	6.21	6.77	6.88	7.41	8.05	8.73
ME05	5.86	4.25	4.79	5.50	5.86	5.99	6.48	6.88	7.28	8.08	8.40
ME06	5.50	3.87	4.49	5.17	5.46	5.89	6.27	6.33	6.88	7.52	8.56
ME07	5.42	4.36	4.82	5.12	5.44	5.89	6.21	6.40	6.49	6.62	8.45
ME08	5.08	4.32	4.09	4.83	5.37	5.44	5.76	6.04	6.17	6.51	8.07
ME09	4.81	4.08	4.38	4.54	4.92	5.22	5.34	5.72	5.81	6.21	7.30
ME10	4.57	4.01	4.19	4.50	4.86	4.85	5.06	5.09	5.44	5.72	6.76

Table 4.44: Average return each month by size-robust beta decile group, 1963–1990.

	ALL	β-01	β-02	β-03	β-04	β-05	β-06	β-07	β-08	β-09	β-10
ALL	1.23	1.31	1.31	1.30	1.36	1.25	1.30	1.24	1.24	1.17	1.06
ME01	1.47	1.56	1.53	1.55	1.69	1.46	1.51	1.44	1.62	1.43	1.27
ME02	1.15	1.11	1.30	1.28	1.25	1.26	1.59	1.26	1.08	1.19	0.74
ME03	1.23	1.16	1.48	1.09	1.59	1.51	1.17	1.60	1.45	1.06	0.68
ME04	1.22	1.29	1.20	1.41	1.49	1.11	1.23	1.17	1.09	1.20	1.06
ME05	1.22	1.20	1.39	1.25	1.34	1.27	1.18	1.21	1.26	1.32	1.02
ME06	1.11	0.97	1.45	1.27	1.16	1.38	1.17	1.00	1.02	1.02	0.84
ME07	1.10	1.04	1.35	1.29	1.23	0.89	1.09	1.06	1.06	1.14	0.97
ME08	1.07	1.06	1.12	1.15	1.29	1.27	1.17	1.03	0.79	0.91	1.00
ME09	0.94	0.94	0.96	0.98	1.08	0.98	1.22	0.99	0.73	0.94	0.64
ME10	0.89	1.04	0.95	1.15	0.83	0.80	0.98	0.82	0.83	0.90	0.59

Table 4.45: Average robust postbeta each month by size-robust beta decile group, 1963–1990.

	ALL	β-01	β-02	β-03	β-04	β-05	β-06	β-07	β-08	β-09	β-10
ALL	1.29	0.93	1.03	1.14	1.21	1.26	1.31	1.38	1.45	1.50	1.67
ME01	1.42	1.13	1.23	1.30	1.34	1.37	1.41	1.50	1.57	1.60	1.70
ME02	1.38	0.98	1.09	1.24	1.29	1.31	1.43	1.49	1.52	1.59	1.70
ME03	1.34	0.88	1.13	1.17	1.24	1.31	1.33	1.41	1.48	1.58	1.74
ME04	1.31	0.78	1.01	1.15	1.19	1.29	1.36	1.41	1.51	1.57	1.72
ME05	1.25	0.63	0.89	1.12	1.19	1.21	1.25	1.38	1.46	1.56	1.67
ME06	1.18	0.56	0.78	1.04	1.09	1.22	1.24	1.29	1.40	1.44	1.63
ME07	1.16	0.59	0.87	1.02	1.11	1.19	1.24	1.28	1.35	1.34	1.60
ME08	1.07	0.50	0.69	0.89	1.02	1.08	1.16	1.19	1.24	1.27	1.63
ME09	1.02	0.60	0.70	0.85	0.98	1.06	1.12	1.11	1.14	1.23	1.39
ME10	0.94	0.55	0.71	0.79	0.92	0.93	0.95	1.01	1.09	1.13	1.37

Table 4.46: Differences in average OLS postbeta and average robust postbeta each month by size-beta decile group, 1963–1990.

	ALL	β-01	β-02	β-03	β-04	β-05	β-06	β-07	β-08	β-09	β-10
ALL	0.01	−0.03	−0.01	−0.01	0.00	0.02	0.01	0.00	0.02	0.04	0.04
ME01	0.03	−0.03	−0.03	−0.02	0.02	0.04	0.01	0.02	0.03	0.05	0.06
ME02	0.02	−0.03	0.07	−0.04	0.03	0.01	0.02	−0.02	0.07	0.04	0.03
ME03	0.01	−0.03	−0.03	−0.01	−0.02	0.02	−0.02	−0.02	0.05	0.02	0.07
ME04	0.01	0.02	0.03	0.01	−0.02	0.02	0.01	−0.02	0.00	0.04	0.03
ME05	0.01	−0.01	0.02	0.03	−0.04	−0.02	−0.01	0.04	0.03	0.03	−0.01
ME06	0.00	0.00	−0.02	0.00	−0.02	0.01	0.03	0.02	−0.06	0.06	0.01
ME07	0.01	0.01	0.01	0.03	−0.03	−0.02	0.00	0.01	−0.02	−0.02	0.06
ME08	0.00	0.00	−0.02	0.03	−0.01	0.01	0.02	−0.01	−0.05	0.03	−0.02
ME09	0.00	−0.06	0.07	−0.01	−0.04	−0.02	−0.03	0.00	0.05	−0.01	0.01
ME10	0.00	0.00	−0.01	0.00	0.02	0.00	−0.01	−0.02	−0.02	0.01	0.01

Table 4.47: Average values each month by size decile group, 1963–1990

	ME01a	ME01b	ME02	ME03	ME04	ME05	ME06	ME07	ME08	ME09	ME10a	ME10b
Return	1.55	1.15	1.15	1.23	1.22	1.22	1.11	1.10	1.07	0.94	0.93	0.85
post β	1.44	1.44	1.40	1.35	1.32	1.26	1.18	1.17	1.07	1.01	0.96	0.92
Size	2.00	3.26	3.73	4.21	4.63	5.04	5.44	5.88	6.37	6.91	7.49	8.56
Log(B2M)	−0.02	−0.22	−0.23	−0.26	−0.31	−0.34	−0.35	−0.39	−0.37	−0.39	−0.49	−0.69
Log(MktLvg)	0.73	0.47	0.46	0.43	0.37	0.34	0.31	0.24	0.27	0.26	0.15	−0.10
Log(BkLvg)	0.74	0.70	0.69	0.69	0.68	0.68	0.66	0.63	0.64	0.65	0.64	0.60
FracNegE2P	0.27	0.14	0.11	0.07	0.06	0.04	0.04	0.03	0.03	0.01	0.02	0.01
E+2P	0.08	0.09	0.09	0.09	0.09	0.09	0.08	0.08	0.08	0.08	0.08	0.08
Firms	938.76	212.83	257.67	183.31	160.04	142.95	128.36	122.05	117.48	113.21	55.99	55.20

Table 4.48: Average values each month by beta decile group, 1963–1990

	β-01a	β-01b	β-02	β-03	β-04	β-05	β-06	β-07	β-08	β-09	β-10a	β-10b
Return	1.30	1.32	1.33	1.31	1.32	1.33	1.31	1.25	1.18	1.18	1.05	1.05
post β	0.91	0.87	1.03	1.13	1.21	1.28	1.32	1.38	1.47	1.54	1.70	1.71
Size	3.50	3.97	4.14	4.30	4.19	4.21	4.20	4.23	4.12	4.06	3.93	3.61
Log(B2M)	−0.09	−0.02	−0.08	−0.16	−0.17	−0.18	−0.21	−0.26	−0.28	−0.35	−0.41	−0.61
Log(MktLvg)	0.65	0.68	0.60	0.49	0.49	0.48	0.45	0.42	0.41	0.37	0.34	0.20
Log(BkLvg)	0.74	0.70	0.68	0.65	0.66	0.66	0.66	0.68	0.70	0.73	0.76	0.81
FracNegE2P	0.14	0.11	0.11	0.10	0.11	0.11	0.12	0.13	0.14	0.14	0.18	0.23
E+2P	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.08	0.08	0.08	0.07	0.07
Firms	208.38	128.89	236.28	209.37	213.61	217.17	225.55	216.60	231.96	243.77	150.36	205.91

Chapter 5

CONCLUSIONS AND FURTHER THOUGHTS

5.1 Summary of the Dissertation

The first half of the dissertation addressed detection of multivariate outliers in the type of asset returns and factor exposure data used by quantitative finance practitioners to construct and manage equity portfolios. In Chapter 2 we reviewed the use of robust squared Mahalanobis distances (RSDs) based on the minimum covariance determinant (MCD) for identifying multivariate outliers. We extended the Hardin and Rocke (2005) method for estimating the parameters of the F distribution used to test MCD-based RSDs. Our improved methodology is more accurate than the Hardin-Rocke method for the small sample sizes $n < 250$ commonly encountered in financial applications, and more accurate when the MCD uses nearly all the observations. We used this enhancement to improve the accuracy of the Iterated Reweighted MCD (IRMCD) outlier detection methodology of Cerioli (2010). Our improved method, which we call IRMCD2, can be reliably used to detect multivariate outliers in financial applications in samples as small as $n = 60$ and/or with low breakdown point versions of the MCD. All of these improvements are available in an `R` package, `CerioliOutlierDetection`, which is available on CRAN.

In Chapter 3 we use IRMCD2 to demonstrate that multivariate outliers are present in asset returns data and factor exposure data. These outliers can be missed by detection methods based on sample means and covariances since those estimators are themselves biased by the outliers. We provide detailed examples of detecting multivariate outliers using IRMCD2 in hedge fund and commodity portfolios, a four-factor asset pricing model, and a ten-factor returns forecasting model. We show, in each case, that our method is more accurate than Mahalanobis distances based on sample means and covariances. Furthermore, we show that the one-dimensional approaches to outlier mitigation that are commonly used in quantitative finance can fail to detect multivariate outliers, especially in high-dimensional data. Our results show that MCD-based RSDs, combined with the IRMCD2, are a very effective means of detecting outlying times in multivariate return time series and outlying assets in factor exposure data.

In the second half of the dissertation we explored the use of robust MM-regression for testing factor-based asset pricing models. In Chapter 4 we revisited the classic asset pricing study of Fama and French (1992) using a high-efficiency robust MM-regression for the cross-sectional regressions, as well as robust methods to estimate the average risk premia over time and their significance. Fama and French evaluated, using cross-sectional least squares regression, whether significant relationships exist between asset returns and several factors—among them beta, size, and book-to-market—in historical stock market data. Fama and French found a negative relationship between average returns and firm size, a positive relationship between average returns and firm book-to-market ratios, and no relationship between average returns and firm betas. Our analysis using robust methods and data through December 2015 showed that the relationship between average returns and firm size is positive for the vast majority of stocks. The negative relationship identified by Fama and French is driven by a small percentage, typically less than 2%, of small stocks each month that have unusually large returns.

On the other hand, we confirmed Fama and French's finding of a positive relationship between average returns and firm book-to-market ratios. We also confirmed, however, a result of Loughran (1997): this positive relationship mainly exists in small stocks. For small stocks this effect still holds through 2015, while for moderately-sized stocks the effect largely vanished after 1980, and for large stocks the effect was never present from 1963 onwards. Thus while a "value effect" does exist, it has been confined to smaller stocks for the past two decades.

Finally, in sharp contrast to Fama and French's results, we showed that the relationship between average returns and firm betas is significant and negative. More importantly, the relationship between average returns, firm betas, and firm size is not linear, and is more complex than assumed by conventional asset pricing models. Our analysis, using robust MM-regression, of a pricing model including beta, size, and a beta-size interaction term shows that the average premia on beta and size are significant and negative for nearly all stocks, while the premium on the interaction term is significant and positive. The relationship between average returns and firm betas hence varies with firm size: for small stocks, average returns tend to decrease with increasing beta, but for large stocks returns tend to increase with increasing beta. For moderately-sized stocks there can be little dependence of average returns on firm betas. Likewise, the relationship between average returns and firm size is significant and positive for low beta stocks, but significant and negative for high beta stocks.

Our analyses in Chapter 4 demonstrate the importance of using highly efficient and robust cross-sectional MM-regression in empirical asset pricing research. The results for the size and beta risk premia in particular show the danger of using only cross-sectional least squares regressions for testing pricing models: a very small percentage of the observations can have very high influence on the test results, leading to erroneous conclusions that do not accurately reflect how the majority of stocks are priced.

5.2 Future Research Topics

There have been several advancements in the field of robust statistics since the publication of Maronna et al. (2006) that could be applied to the analyses presented in this dissertation or other common estimation tasks in quantitative finance. The remainder of this chapter briefly discusses potential improvements to the research conducted in this dissertation, as well as other areas of quantitative finance where robust statistical methods may offer significant improvements over the standard techniques in those areas.

5.2.1 Improvements to Outlier Detection via Robust Distances

In the standard Tukey-Huber model of contamination, observations are considered to be contaminated "case-wise", i.e., an observation is either outlying in all coordinates or none.[1] A simulation experiment presented in Chapter 6.8 of Maronna et al. (2006) studied covariance estimation under this model of contamination, and demonstrated that certain S-estimators, namely the bisquare and Rocke S-estimators, were to be preferred to the MCD. Maronna and Yohai (2017) presents an updated comparison of several robust covariance estimators, with the Rocke S-estimator now being preferred in dimensions $\nu \geq 15$, and an MM-estimator for covariance preferred in dimensions $\nu < 15$. An important and, as of this time, open, question is whether these estimators yield more accurate RSD-based tests of outlyingness than tests using MCD-based RSDs.

As we discussed in Section 2.6, the Tukey-Huber contamination model is unrealistic for some applications: in some data sets contamination may only occur in a few variables within each observation, and a significant fraction of the observations may suffer from such minor contamination. Moreover, a data set might exhibit both partially- and entirely-contaminated observations. Consider, for instance, estimating the covariance of a group of stocks from a multivariate time series of their returns. At times one or more stocks might have an unusual return, independently of all other stocks in the data set. A market-wide event, on the other hand, could produce a vector of returns that is outlying in all coordinates (stocks) simultaneously. The so-called Independent Contamination Model of Alqallaf et al. (2009) provides a framework for robust estimation in the presence of this type of data contamination. Agostinelli et al. (2015) discusses methods for estimation under the Independent Contamination Model. Pairwise robust covariance matrix estimators, such as the OGK estimator of Maronna and Zamar (2002) or the quantile-based scatter estimator of Qiu et al. (2015), may provide better estimates in this scenario than estimators like the MCD that consider all variables simultaneously. Optimal covariance estimation under the Independent Contamination

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

Model is currently an open problem.

Regardless of the contamination model, the inaccurate false positive rates of MCD-based distance tests using chi-squared thresholds raise the question of how accurate distance tests using covariance estimates other than the MCD and chi-squared threshold are. We show in Appendix A that the outlier tests based on the OGK and Rocke S-estimators will have inaccurate false positive rates in smaller samples (e.g., $60 \leq n \leq 250$), just like the MCD-based distance tests. The IRMCD2 approach we used in Chapter 3 is specific to the structure of the MCD, and not applicable to other robust covariance estimates. The development of calibration methods for the Rocke S-estimator, OGK estimator, and other robust covariance estimators remains a barrier to their adoption for outlier detection tests.

5.2.2 Applications of Robust Mean and Covariance Estimators to Portfolio Construction

Based on the results presented in this dissertation, we might consider using a robust mean and covariance matrix as the inputs to a mean-variance portfolio construction process. Chopra and Ziemba (1993) demonstrated that mean-variance optimization is sensitive to errors in the inputs, so it is reasonable to hypothesize that robust estimates might lead to more reliable portfolios than the sample mean and covariance. Several authors (Cavadini et al., 2002; Lauprete, 2001; Lauprete et al., 2002; Vaz-de Melo and Camara, 2003; Gao, 2004; Perret-Gentil and Victoria-Feser, 2003; Welsch and Zhou, 2007; DeMiguel and Nogales, 2009) have examined the use of robust estimation in portfolio construction, either via robust inputs to the usual mean-variance optimization process or by trying to construct a "robust" version of mean-variance optimization.[2] Overall, these studies have found isolated cases where an approach to mean-variance optimization based on robust estimates of mean and variance outperforms the standard approach based the sample mean and variance estimates, but have not established that the robust approach is better or worse in general.

[2]These attempts to "robustify" mean-variance optimization are not related to so-called "robust portfolio optimization" as developed by Goldfarb and Iyengar (1993); Erdogan et al. (2004); Ceria and Stubbs (2006); Garlappi et al. (2007) and critiqued in Scherer (2007).

5.2.3 Applications of Improved Robust Regression Methods to Factor Models

In Chapter 4 we demonstrated the value of robust MM-regression in tests of factor-based empirical asset pricing models. Guerard and colleagues (Bloch et al., 1993; Guerard et al., 2015; Guerard, 2016; Guerard et al., 2016) have explored the use of robust regression in factor models used for asset return forecasts. Not much is published about the use of robust MM-regression in the construction of factor models for risk analysis. Axioma is known to use the Huber regression M-estimator to construct some of their factor models for risk analysis (e.g., see (Guerard, 2017)), but that estimator is not robust to outliers in the independent variables. The proper application of MM-regression to the construction of factor models for risk analysis remains an open problem.

Koller and Stahel (2011) and Maronna and Yohai (2015) present improvements to MM-regression that should be also tested in the cross-sectional regression context. These improvements may be beneficial in empirical asset pricing research, portfolio construction, and risk management. A formal comparison of these new regression methods against the optimal MM-regression method used in Chapter 4 would be a valuable contribution to quantitative finance.

BIBLIOGRAPHY

Claudio Agostinelli and Victor J. Yohai. Composite Robust Estimators for Linear Mixed Models. *Journal of the American Statistical Association*, 111(516):1764–1774, 2017.

Claudio Agostinelli, Andy Leung, Victor J. Yohai, and Ruben H. Zamar. Robust Estimation of Multivariate Location and Scatter in the Presence of Cellwise and Casewise Contamination. *TEST*, 24(3):441–461, 2015.

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.

Diego Amaya, Peter Christoffersen, Kris Jacobs, and Aurelio Vasquez. Does Realized Skewness Predict the Cross-Section of Equity Returns? Available at SSRN: `http://ssrn.com/abstract=1898735`. Version February 21, 2013. Accessed August 20, 2014., 2013.

Y. Amihud. Illiquidity and Stock Returns: Cross-section and Time-series Effects. *Journal of Financial Markets*, 5:31–56, 2002.

D. F. Andrews, R. Gnanadesikan, and J. L. Warner. Methods for Assessing Multivariate Normality. In P. R. Krishnaiah, editor, *Multivariate Analysis*, volume III, pages 95–116. Academic Press, New York, 1973.

Andrew Ang, Robert J. Hodrick, Yuhang Xing, and Xiaoyan Zhang. The Cross-Section of Volatility and Expected Returns. *Journal of Finance*, 61(1):259–299, 2006.

Mark Ash and Erik Dohlman. Oil Crops Situation and Outlook Yearbook, March 2005. Available from `http://usda.mannlib.cornell.edu/usda/ers/OCS-yearbook//2000s/2005/OCS-yearbook-03-21-2005.pdf`. Accessed January 6, 2017.

Clifford S. Asness and Andrea Frazzini. The Devil in HML's Details. *Journal of Portfolio Management*, 39(4):49–68, 2013.

A. C. Atkinson, M. Riani, and A. Cerioli. *Exploring Multivariate Data with the Forward Search*. Springer Verlag, New York, 2004.

Anthony Atkinson. Econometric Applications of the Forward Search in Regression: Robustness, Diagnostics, and Graphics. *Econometric Reviews*, 28(1-1):21–39, 2009.

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

Heiko Bailer and R. Douglas Martin. Fama-MacBeth 1973: Reproduction, Extension, Robustification. *Journal of Economic and Social Measurement*, 32:41–63, 2007.

Heiko Bailer, Tatiana A. Maravina, and R. Douglas Martin. Robust Betas In Asset Management. In Bernd Scherer and Kenneth Winston, editors, *Oxford Handbook of Quantitative Asset Management*, pages 203–242. Oxford University Press, 2011.

R. W. Banz. Anomalies in Relationships between Securities' Yields and Yield Surrogates. *Journal of Financial Economics*, 6:103–126, 1981.

M. L. Barnes and A. W. Hughes. A Quantile Regression Analysis of the Cross Section of Stock Market Returns. Working paper, available from `http://bostonfed.org/economic/wp/wp2002/wp022.pdf`. Accessed August 11, 2014., 2002.

S. Basu. The Relationship between Earnings Yield, Market Value, and Return for NYSE Common Stocks: Further Evidence. *Journal of Financial Economics*, 12:129–156, 1983.

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.

Johnathan B. Berk. An Empirical Reexamination of the Relation between Firm Size and Return. Working paper, University of Washington, 1995a.

Johnathan B. Berk. A View of the Current Status of the Size Anomaly. In Donald B. Keim and William T. Ziemba, editors, *Security Market Imperfections in World Wide Equity Markets*, pages 90–115. Cambridge University Press, Cambridge, 2000.

Jonathan B. Berk. A Critique of Size-Related Anomalies. *The Review of Financial Studies*, 8(2):275–286, 1995b.

Jonathan B. Berk. Does Size Really Matter? *Financial Analysts Journal*, 53(5):12–18, 1997.

Laxmi Chand Bhandari. Debt/Equity Ratio and Expected Common Stock Returns: Empirical Evidence. *Journal of Finance*, 43:507–528, 1988.

Fischer Black. Capital Market Equilibrium with Restricted Borrowing. *Journal of Business*, 45:444–454, 1972.

Fischer Black, M. C. Jensen, and M. Scholes. The Capital Asset Pricing Model: Some Empirical Tests. In M. C. Jensen, editor, *Studies in the Theory of Capital Markets*. Praeger, New York, 1972.

Benjamin M. Blau, Abdullah Masud, and Ryan J. Whitby. Does it Pay to Have Fat Tails? Examining Kurtosis and the Cross-section of Stock Returns. Available from http://huntsman.usu.edu/economicsandfinance/files/uploads/Whitby/Fat_Tails.pdf., 2013.

M. Bloch, J. Guerard, H. Markowitz, P. Todd, and G. Xu. A Comparison of Some Aspects of the U.S. and Japanese Equity Markets. *Japan and the World Economy*, 5(1):3–26, 1993.

Kris Boudt, Brian Peterson, and Christophe Croux. Estimation and Decomposition of Downside Risk for Portfolios with Non-Normal Returns. *The Journal of Risk*, 11(2):79–103, 2008.

David C. Bowie and David J. Bradfield. Robust Estimation of Beta Coefficients: Evidence from a Small Stock Market. *Journal of Business Finance & Accounting*, 25(3-4):439–454, 1998.

Maria Caterina Bramati and Christophe Croux. Robust Estimators for the Fixed Effects Panel Data Model. *Econometrics Journal*, 10(3):521–540, 2007.

Ernest Breitschwerdt. Point-In-Time vs. Lagged Fundamentals: This Time I[t']s Different? *S&P Capital IQ Quantamental Research*, August 2015.

Andreas Buja and Daniel Asimov. Grand Tour Methods: An Outline. In D. M. Allen, editor, *Computer Science and Statistics: Proceedings of the Seventeenth Symposium on The Interface*, pages 63–67, North-Holland, 1986. Elsevier Science Publishers B. V.

J. Y. Campbell, A. W. Lo, and A. C. MacKinlay. *The Econometrics of Financial Markets*. Princeton University Press, Princeton, N. J., 1997.

Mark M. Carhart. On Persistence in Mutual Fund Performance. *Journal of Finance*, 52: 57–82, 1997.

D. B. Carr, R. J. Littlefield, W. L. Nicholson, and J. S. Littlefield. Scatterplot Matrix Techniques for Large N. *Journal of the American Statistical Association*, 82(398):424–436, 1987.

F. Cavadini, A. Sbuelz, and F. Trojani. A Simplified Way of Incorporating Model Risk, Estimation Risk and Robustness in Mean Variance Portfolio Management. Working Paper, University of Southern Switzerland, 2002.

Scott Cederburg and Michael S. O'Doherty. Asset-Pricing Anomalies at the Firm Level. *Journal of Econometrics*, 186(1):113–128, 2015.

Sebastian Ceria and Robert A. Stubbs. Incorporating Estimation Errors in Portfolio Selection: Robust Portfolio Construction. Axioma Research Paper No. 003. Available from http://www.axiomainc.com/downloads/incorporating_estimation.pdf. Accessed September 5, 2009., 2006.

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.

K. C. Chan and Nai-Fu Chen. An Unconditional Asset-Pricing Test and the Role of Firm Size as an Instrumental Variable for Risk. *The Journal of Finance*, 43(2):309–325, 1988.

Louis K. Chan, Yasushi Hamao, and Josef Lakonishok. Fundamentals and Stock Returns in Japan. *Journal of Finance*, 46:1739–1789, 1992.

Louis K. C. Chan, Jason Karceski, and Josef Lakonishok. New Paradigm or Same Old Hype in Equity Investing? *Financial Analysts Journal*, 56(4):23–36, 2000.

Ludwig B. Chincarini and Daehwan Kim. *Quantitative Equity Portfolio Management: An Active Approach to Portfolio Construction and Managment*. McGraw-Hill, 2006.

Vijay K. Chopra and William T. Ziemba. The Effect of Errors in Means, Variances, and Covariances on Optimal Portfolio Choice. *The Journal of Portfolio Management*, pages 6–11, Winter 1993.

P. H. Chou, R. K. Chou, and J. S. Wang. On the Cross-Section of Expected Stock Returns: Fama-French Ten Years Later. *Financial Letters*, 2(1):18–22, 2004.

George Chow, Eric Jacquier, Mark Kritzman, and Kenneth Lowry. Optimal Portfolios in Good Times and Bad. *Financial Analysts Journal*, 55(3):65–73, 1999.

P. Čížek and W. Härdle. Robust Estimators in Econometrics. In Steven N. Durlauf and Lawrence E. Blume, editors, *The New Palgrave Dictionary of Economics*. Palgrave Macmillan, second edition, 2008.

J. H. Cochrane. *Asset Pricing*. Princeton Univ. Press, Princeton, NJ, 2001.

C. Colombier. Growth Effects of Fiscal Policies: An Application of Robust Modified M-estimator. *Applied Economics*, 41(7):899–912, 2009.

Compustat. Standard & Poor's Compustat® Data, 2015. Used with permission. All rights reserved. Accessed March 2015 via Wharton Research Data Services.

R. A. Connolly. An Examination of the Robustness of the Weekend Effect. *Journal of Financial and Quantitative Analysis*, 24(2):133–169, 1989.

Gregory Connor, Lisa R. Goldberg, and Robert A. Korajczyk. *Portfolio Risk Analysis*. Princeton University Press, New Jersey, 2010.

D. Cook, A. Buja, J. Cabrera, and C. Hurley. Grand Tour and Projection Pursuit. *Journal of Computational and Graphical Statistics*, 4:155–172, 1995.

Dianne Cook and Deborah F. Swayne. *Interactive and Dynamic Graphics for Data Analysis: With Examples Using R and GGobi*. Springer-Verlag, New York, 2007.

Bradford Cornell and J. Kimball Dietrich. Mean-Absolute-Deviation Versus Least-Squares Regression Estimation of Beta Coefficients. *The Journal of Financial and Quantitative Analysis*, 13(1):123–131, 1978.

Michael A. Crain. A Literature Review of the Size Effect. Published October 29, 2011. Available at SSRN: `https://ssrn.com/abstract=1710076`, 2011.

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.

Christophe Croux, Geert Dhaene, and Dirk Hoorelbeke. Robust Standard Errors for Robust Estimators. Technical report, K.U. Leuven, CES., 2003a. Discussion Papers Series 03.16.

Christophe Croux, Peter Filzmoser, G. Pison, and Peter J. Rousseeuw. Fitting Multiplicative Models by Robust Alternating Regressions. *Statistical Comput.*, 13:23–36, 2003b.

CRSP. CRSP/Compustat Merged Database, 2015a. Used with permission. All rights reserved. `www.crsp.uchicago.edu`. Accessed March 2015 via Wharton Research Data Services.

CRSP. Center for Research in Security Prices , 2015b. Used with permission. All rights reserved. `www.crsp.uchicago.edu`. Accessed March 2015 via Wharton Research Data Services.

P. L. Davies. Asymptotic Behaviour of S-Estimates of Multivariate Location Parameters and Dispersion Matrices. *The Annals of Statistics*, 15(3):1269–1292, 1987.

James Davis, Eugene F. Fama, and Kenneth R. French. Characteristics, Covariances, and Average Returns: 1929 to 1997. *Journal of Finance*, 55(1):389–406, Feb. 2000.

James L. Davis. The Cross-Section of Stock Returns and Survivorship Bias: Evidence from Delisted Stocks. *The Quarterly Review of Economics and Finance*, 36(3):365–375, 1996.

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

Victor DeMiguel and Francisco J. Nogales. Portfolio Selection with Robust Estimation. *Operations Research*, 57(3):560–577, May–June 2009.

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.

Ilia D. Dichev. Is the Risk of Bankruptcy a Systematic Risk? *The Journal of Finance*, 53 (3):1131–1147, 1998. URL `http://www.jstor.org/stable/117389`.

Elroy Dimson. Risk Measurement when Shares are Subject to Infrequent Trading. *Journal of Financial Economics*, 7(2):197–226, 1979.

Kathryn E. Easterday, Pradyot K. Sen, and Jens A. Stephan. The Persistence of the Small Firm/January Effect: Is it Consistent with Investors' Learning and Arbitrage Efforts? *The Quarterly Review of Economics and Finance*, 49(3):1172–1193, 2009.

Venkat R. Eleswarapu and Marc R. Reinganum. The Seasonal Behavior of the Liquidity Premium in Asset Pricing. *Journal of Financial Economics*, 34(3):373–386, 1993.

E. Elton, M. J. Gruber, Stephen J. Brown, and William N. Goetzmann. *Modern Portfolio Theory and Investment Analysis*. John Wiley & Sons, New York, 7th edition, 2006.

E. Erdogan, D. Goldfarb, and G. Iyengar. Robust Portfolio Management. Technical report, Operations Research Center, Columbia University, 2004. CORC Technical Report TR-2004-11.

Eugene F. Fama. *Foundations of Finance*. Basic Books, New York, 1976.

Eugene F. Fama and Kenneth French. A Five-Factor Asset Pricing Model. *Journal of Financial Economics*, 116:1–22, 2015.

Eugene F. Fama and Kenneth R. French. The Cross-Section of Expected Stock Returns. *The Journal of Finance*, 47(2):427–465, June 1992.

Eugene F. Fama and Kenneth R. French. Common Risk Factors in the Returns on Stocks and Bonds. *Journal of Financial Economics*, 33:3–56, 1993.

Eugene F. Fama and Kenneth R. French. Multifactor Explanations of Asset Pricing Anomalies. *The Journal of Finance*, 51(1):pp. 55–84, 1996. URL `http://www.jstor.org/stable/2329302`.

Eugene F. Fama and Kenneth R. French. Dissecting Anomalies. *The Journal of Finance*, 63 (4):pp. 1653–1678, 2008. URL `http://www.jstor.org/stable/25094486`.

Eugene F. Fama and J. D. MacBeth. Risk, Return, and Equilibrium: Empirical Tests. *Journal of Political Economy*, 81:607–636, 1973.

Weiguo Gao. Portfolio Optimization Based on Robust Estimation Procedures. Master's thesis, Worcester Polytechnic Institute, 2004.

L. Garlappi, R. Uppal, and T. Wang. Portfolio Selection with Parameter and Model Uncertainty: A Multi-Prior Approach. *Review of Financial Studies*, 20(1):41–81, 2007.

Xavier Garza-Gómez, Jiro Hodoshima, and Michio Kunimura. Reexamining the Robustness of the Market Value of Equity. *Asia-Pacific Financial Markets*, 8(2):61–85, 2001.

Marc G. Genton and E. Ronchetti. Robust Prediction of Beta. In E. J. Kontoghiorghes, B. Rustem, and P. Winker, editors, *Computational Methods in Financial Engineering: Essays in Honour of Manfred Gilli*, pages 147–161. Springer-Verlag, 2008.

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.

D. Goldfarb and G. Iyengar. Robust Portfolio Selction Problems. *Mathematics of Operations Research*, pages 1–38, February 1993.

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

R. C. Grinold and R. N. Kahn. *Active Portfolio Management: A Quantitative Approach for Producing Superior Returns and Controlling Risk*. McGraw-Hill, New York, 2nd edition, 2000.

Anthony Yanxiang Gu. The Declining January Effect: Evidences from the U.S. Equity Markets. *The Quarterly Review of Economics and Finance*, 43(2):395–404, 2003.

John Guerard, editor. *Handbook of Portfolio Construction: Contemporary Applications of Markowitz Techniques*. Springer-Verlag, London, 2010.

John Guerard. Investing in Global Markets: Big Data and Applications of Robust Regression. *Frontiers in Applied Mathematics and Statistics*, 1(14), 2016.

John B. Guerard. The Theory of Risk Return and Performance Measurement. In *Portfolio Construction, Measurement, and Efficiency*, pages 1–38. Springer-Verlag, 2017.

John B. Guerard, M. Gultekin, and Bernell K. Stone. The Role of Fundamental Data and Analyst's Earnings Breadth, Forecasts, and Revisions in the Creation of Efficient Portfolios. In A. Chen, editor, *Research in Finance*, volume 15. 1997.

John B. Guerard, Harry Markowitz, and GanLin Xu. Earnings Forecasting in a Global Stock Selection Model and Efficient Portfolio Construction and Management. *International Journal of Forecasting*, 31(2):550–560, 2015.

John B. Guerard, Shijie Deng, Robert A. Gillam, Harry Markowitz, Ziwei Wang, and Ganlin Xu. Investing in Global Equity Markets with Particular Emphasis on Chinese Stocks. Available from `http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2744304`, March 2016.

Frank R. Hampel. *Contributions to the Theory of Robust Estimation*. PhD thesis, University of California, Berkeley, 1968.

Frank R. Hampel. The Influence Curve and its Role in Robust Estimation. *Journal of the American Statistical Association*, 69:383–393, 1974.

Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. *Robust Statistics: the Approach based on Influence Functions*. John Wiley & Sons, New York, 1986.

L. Hansen. Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica*, 50:1029–1054, 1982.

J. Hardin and D. M. Rocke. The Distribution of Robust Distances. *Journal of Computational and Graphical Statistics*, 14:928–946, 2005.

Joel L Horowitz, Tim Loughran, and N.E Savin. Three Analyses of the Firm Size Premium. *Journal of Empirical Finance*, 7(2):143–153, 2000a.

Joel L. Horowitz, Tim Loughran, and N.E. Savin. The Disappearing Size Effect. *Research in Economics*, 54(1):83–100, 2000b.

P. J. Huber. *Robust Statistical Procedures*, volume 27 of *CBMS Regional Conference Series in Applied Mathematics*. SIAM, Philadelphia, PA, 1977.

Peter J. Huber. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35:73–101, 1964.

Peter J. Huber. Robust Regression: Asymptotics, Conjectures and Monte Carlo. *The Annals of Statistics*, 1:799–821, 1973.

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

Peter J. Huber. Between Robustness and Diagnostics. In Werner Stahel and Sanford Weisberg, editors, *Directions in Robust Statistics and Diagnostics. Part I*, pages 121–130. Springer-Verlag Inc, 1991.

Peter J. Huber and Elvezio M. Ronchetti. *Robust Statistics*. John Wiley & Sons, New York, 2nd edition, 2009.

Narasimhan Jegadeesh. Evidence of Predictable Behavior of Security Returns. *Journal of Finance*, 45(3):881–898, 1990.

Narasimhan Jegadeesh and Sheridan Titman. Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency. *Journal of Finance*, 48:65–91, 1993.

Narasimhan Jegadeesh and Sheridan Titman. Profitability of Momentum Strategies: an Evaluation of Alternative Explanations. *Journal of Finance*, 56:699–718, 2001.

D. B. Keim. Size-related Anomalies and Stock Return Seasonality: Further Empirical Evidence. *Journal of Financial Economics*, 12:13–32, June 1983.

Peter Knez and Ready. On the Robustness of Size and Book-to-Market in Cross-Sectional Regressions. *Journal of Finance*, 52(4):1355–1382, September 1997.

Roger Koenker. Robust Methods in Econometrics. *Econometric Review*, 1:213–225, 1982.

Manuel Koller and Werner A. Stahel. Sharpening Wald-type Inference in Robust Regression for Small Samples. *Computational Statistics & Data Analysis*, 55(8):2504–2515, 2011.

Mark Kritzman and Yuanzhen Li. Skulls, Financial Turbulence, and Risk Management. *Financial Analysts Journal*, 66(5):30–41, 2010.

G. J. Lauprete. *Portfolio Risk Minimization under Departures from Normality*. PhD thesis, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA, 2001.

G. J. Lauprete, A. M. Samarov, and R. E. Welsch. Robust Portfolio Optimization. *Metrika*, 55:139–149, 2002.

Erich L. Lehmann. *Elements of Large Sample Theory*. Springer-Verlag, New York, 2004.

John Lintner. The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets. *The Review of Economics and Statistics*, 47(1):13–37, February 1965.

R. Litterman. *Modern Investment Management*. John Wiley & Sons, New Jersey, 2003.

H. P. Lopuhaä and Peter J. Rousseeuw. Breakdown Points of Affine Equivariant Estimators of Multivariate Location and Covariance Matrices. *Annals of Statistics*, 19:229–248, 1991.

Hendrik Lopuhaä. Asymptotics of Reweighted Estimators of Multivariate Location and Scatter. *Annals of Statistics*, 27(5):1638–1665, 1999.

Tim Loughran. Book-To-Market across Firm Size, Exchange, and Seasonality: Is There an Effect? *The Journal of Financial and Quantitative Analysis*, 32(3):249–268, 1997.

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.

Barbara Mariner-Volpe. Natural Gas Markets and EIA's Information Program, March 2000. Presentation to NARUC Winter Conference Gas Subcommittee Meeting. Available from http://www.eia.gov/pub/oil_gas/natural_gas/presentations/2000/markets_information_program/sld001.htm. Accessed January 6, 2017.

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

Ricardo Maronna, R. Douglas Martin, Victor Yohai, and Matías Salibián-Barrera. *Robust Statistics: Theory and Practice*. John Wiley & Sons, New York, 2nd edition, 2017.

Ricardo A. Maronna and Victor J. Yohai. High Finite-Sample Efficiency and Robustness Based on Distance-constrained Maximum Likelihood. *Computational Statistics & Data Analysis*, 83:262–274, 2015.

Ricardo A. Maronna and Victor J. Yohai. Robust and Efficient Estimation of Multivariate Scatter and Location. 2017.

Ricardo A. Maronna and Ruben H. Zamar. Robust Estimates of Location and Disperion for High-Dimensional Datasets. *Technometrics*, 44(4):307–317, November 2002.

Ricardo Antonio Maronna. Robust M-Estimators of Multivariate Location and Scatter. *The Annals of Statistics*, 4(1):51–67, 1976.

R. Douglas Martin and Tim Simin. Outlier Resistant Estimates of Beta. *Financial Analysts Journal*, 59(5):56–69, 2003.

R. Douglas Martin and Ruben H. Zamar. Asymptotically Min-max Bias Robust M-estimates of Scale for Positive Random Variables. *Journal of the American Statistical Association*, 84:494–501, 1989.

R. Douglas Martin and Ruben H. Zamar. Bias Robust Estimation of Scale. *The Annals of Statistics*, 21(2):991–1017, 06 1993.

R. Douglas Martin, Victor J. Yohai, and Ruben H. Zamar. Min-max bias robust regression. *The Annals of Statistics*, 17:1608–1630, 1989.

R. Douglas Martin, Andrew Clark, and Christopher G. Green. Robust Portfolio Construction. In Guerard (2010), pages 337–382.

A. McNeil, R. Frey, and P. Embrechts. *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton University Press, Princeton, NJ, 2005.

J. Mossin. Equilibrium in Capital Asset Markets. *Econometrica*, 34(4):768–783, 1966.

MSCI. MSCI Global Investable Market Value and Growth Index Methodology, 2016. Available from `https://www.msci.com/eqb/methodology/meth_docs/MSCI_Aug16_GIMIVGMethod.pdf`. Accessed December 28, 2016.

Erich Neuwirth. *RColorBrewer: ColorBrewer Palettes*, 2014. URL `http://CRAN.R-project.org/package=RColorBrewer`. R package version 1.1-2.

W. K. Newey and K. D. West. A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica*, 55:703–8, 1987.

Floyd Norris. A Clever Move by Porsche on VW's Stock. *The New York Times*, October 30, 2008.

Luis Palacios and Premal Vora. SAS Program `fama_french_factors_replication.sas`, 2011. A SAS program to replicate the Fama and French (1993) SMB and HML factors. WRDS Research Applications. Original version April 2009, Updated August 2011. Available from `http://wrds-web.wharton.upenn.edu/wrds/research/applications/risk/fama-french/index.cfm`. Accessed April 2014.

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.

Cédric Perret-Gentil and Maria-Pia Victoria-Feser. Robust Mean-Variance Portfolio Selection. FAME Research Paper No. 140, Université de Genève; NCCR - FINRISK working paper no. 173; SSRN `http://ssrn.com/abstract=721509`, 2003.

Thomas K. Philips. Robust Estimates of Betas and Correlations. In Frank J. Fabozzi et al., editor, *Encyclopedia of Financial Models*, pages 437–447. John Wiley & Sons, Inc., 2012.

G. Pison, S. van Aelst, and G. Willems. Small Sample Corrections for LTS and MCD. *Metrika*, 55:111–123, 2002.

Huitong Qiu, Fang Han, Han Liu, and Brian Caffo. Robust Portfolio Optimization. In *Neural Information Processing Systems (NIPS)*, volume 28, 2015.

Svetlozar Rachev and Stefan Mittnik. *Stable Paretian Models in Finance*. John Wiley and Sons, New York, 2000.

M. R. Reinganum. Misspecification of Capital Asset Pricing: Empirical Anomalies Based on Earnings Yields and Market Values. *Journal of Financial Economics*, March 1981.

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.

Richard Roll and Stephen A. Ross. An Empirical Investigation of the Arbitrage Pricing Theory. *Journal of Finance*, 35:1073–1103, 1980.

Richard Roll and Stephen A. Ross. The Arbitrage Pricing Theory Approach to Strategic Portfolio Planning. *Financial Analysts Journal*, pages 14–26, May/June 1984.

Barr Rosenberg, Kenneth Reid, and Ronald Lanstein. Persuasive Evidence of Market Inefficiency. *Journal of Portfolio Management*, 11:9–11, 1985.

Stephen A. Ross. The Arbitrage Theory of Capital Asset Pricing. *Journal of Economic Theory*, 13:341–360, 1976.

Peter J. Rousseeuw. Multivariate Estimation with High Breakdown Point. Technical report, Centre for Statistics and Operations Research, VUB Brussels, 1983. Research Report No. 192.

Peter J. Rousseeuw. Least Median of Squares Regression. *Journal of the American Statistical Association*, 79(388):871–880, 1984.

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 Annick M. Leroy. *Robust Regression and Outlier Detection*. John Wiley & Sons, New York, New York, 1987.

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.

Peter J. Rousseeuw and Victor Yohai. Robust Regression by Means of *S*-estimators. In J. Franke, W. Härdie, and D. Martin, editors, *Robust and Nonlinear Time Series Analysis*, volume 26 of *Lecture Notes in Statistics*, pages 256–272. Springer-Verlag Inc, New York, 1984.

Peter J. Rousseeuw, Christophe Croux, Valentin Todorov, Andreas Ruckstuhl, Matias Salibian-Barrera, Tobias Verbeke, Manuel Koller, and Martin Maechler. `robustbase`*: Basic Robust Statistics*, 2016. R package version 0.92-6.

David Ruppert and R. J. Carroll. Trimmed Least Squares Estimation in the Linear Model. *Journal of the American Statistical Association*, 75:828–838, 1980.

Sunil K. Sapra. High-Breakdown Point Estimation of Some Regression Models. *Applied Economics Letters*, 10(14):875–878, 2003.

Bernd Scherer. Can Robust Portfolio Optimisation Help to Build Better Portfolios? *Journal of Asset Management*, 7(6):374–387, 2007.

Bernd Scherer and R. Douglas Martin. *Introduction to Modern Portfolio Optimization with* NuOPT*,* S-Plus™ , and S+Bayes. Springer-Verlag, New York, 2005.

G. William Schwert. Anomalies and Market Efficiency. In *Financial Markets and Asset Pricing*, volume 1, Part B of *Handbook of the Economics of Finance*, pages 939–974. Elsevier, 2003. Chapter 15.

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

Pranab Kumar Sen. Estimates of the Regression Coefficient Based on Kendall's Tau. *Journal of the American Statistical Association*, 63(324):1379–1389, 1968.

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

J. Shanken. On the Estimation of Beta-Pricing Models. *Review of Financial Studies*, 5:1–33, 1992.

William Sharpe. Mean-Absolute-Deviation Characteristic Lines for Securities and Portfolios. *Management Science*, 18(2):B1–B13, 1971.

William F. Sharpe. Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk. *Journal of Finance*, 19(3):425–442, 1964.

T. Shumway. The Delisting Bias in CRSP Data. *Journal of Finance*, 52:327–340, 1997.

Tyler Shumway and Vincent A. Warther. The Delisting Bias in CRSP's Nasdaq Data and Its Implications for the Size Effect. *The Journal of Finance*, 54(6):2361–2379, 1999.

Zbynek Šidák. Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. *Journal of the American Statistical Association*, 62(318):626–633, 1967.

Andrew F. Siegel. Robust Regression using Repeated Medians. *Biometrika*, 69:242–244, 1982.

N. J. H. Small. Plotting Squared Radii. *Biometrika*, 65:657–658, 1978.

W. A. Stahel. *Robust Estimation: Infinitesimal Optimality and Covariance Matrix Estimators*. PhD thesis, ETH, Zurich, 1981. (in German).

D. Stattman. Book Values and Stock Returns. *The Chicago MBA: A Journal of Selected Papers*, pages 25–45, 1980.

T. G. Stephan, R. Maurer, and M. Dürr. A Multiple Factor Model for European Stocks. In *The Canadian Institute of Actuaries and the Financial section of IAA (2001): Proceedings of the 11th AFIR International Colloquium*, volume 2, pages 633–651, Toronto, 2001.

James H. Stock and Mark W. Watson. *Introduction to Econometrics*. Pearson/Addison Wesley, Boston, 2nd ed. edition, 2007.

Alvin L. Stroyny. Estimating a Combined Linear Factor Model. In John Knight and Stephen Satchell, editors, *Linear Factor Models in Finance*, pages 210–225. Elsevier/Butterworth-Heinemann, Oxford, 2005.

M. Svarc, Victor Yohai, and Ruben Zamar. Optimal Bias-Robust M-estimates of Regression. In Y. Dodge, editor, *Statistical Data Analysis Based on the L1 Norm and Related Methods*, pages 191–200. Birkhäuser, 2002.

The Associated Press. U.S. Traces Mad Cow Case to Canada, December 2003. Originally published December 27, 2003. Available from `http://www.nytimes.com/2003/12/27/national/us-traces-mad-cow-case-to-canada.html`. Accessed January 6, 2017.

Henri Theil. A Rank-Invariant Method of Linear and Polynomial Regression Analysis, Part I. In Royal Netherlands Academy of Arts and Sciences, editors, *Proceedings of the Royal Netherlands Academy of Sciences*, volume 53, pages 386–392, 1950a.

Henri Theil. A Rank-Invariant Method of Linear and Polynomial Regression Analysis, Part II. In Royal Netherlands Academy of Arts and Sciences, editors, *Proceedings of the Royal Netherlands Academy of Sciences*, volume 53, pages 521–525, 1950b.

Henri Theil. A Rank-Invariant Method of Linear and Polynomial Regression Analysis, Part III. In Royal Netherlands Academy of Arts and Sciences, editors, *Proceedings of the Royal Netherlands Academy of Sciences*, volume 53, pages 1397–1412, 1950c.

Valentin Todorov. A Note on the MCD Consistency and Small Sample Correction Factors. Unpublished manuscript, in preparation, 2008.

Valentin Todorov and Peter Filzmoser. An Object Oriented Framework for Robust Multivariate Analysis. *Journal of Statistical Software*, 32(3):1–47, 2009. URL `http://www.jstatsoft.org/v32/i03/`.

Jack Treynor. Towards a Theory of Market Value of Risky Assets. Unpublished Manuscript, 1961.

Ruey S. Tsay. *Analysis of Financial Time Series*. John Wiley & Sons, New Jersey, 2005.

John W. Tukey. A Survey of Sampling from Contaminated Distributions. In I. Olkin, editor, *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, Stanford, CA, June 1960.

John W. Tukey. Robust Techniques for the User. In Robert L. Launer and Graham N. Wilkinson, editors, *Robustness in Statistics*, pages 103–106. Academic Press, New York, 1979.

U.S. Department of Agriculture. USDA Makes Preliminary Diagnosis of BSE, 2003. Original url `http://www.usda.gov/Newsroom/0432.03.html`. Currently available from the Internet Archive at `https://web-beta.archive.org/web/20041117151000/http://www.usda.gov/Newsroom/0432.03.html`. Accessed January 6, 2017.

U.S. Energy Information Administration. Weekly U.S. Weekly No. 2 Heating Oil Residential Price (Data Series), 2017. Available from `https://www.eia.gov/dnav/pet/hist/LeafHandler.ashx?n=PET&s=W_EPD2F_PRS_NUS_DPG&f=W`. Accessed January 6, 2017.

Mathijs A. van Dijk. Is Size Dead? A Review of the Size Effect in Equity Returns. *Journal of Banking & Finance*, 35(12):3263–3274, 2011.

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.

B. Vaz-de Melo and Rene P. Camara. Robust Modeling of Multivariate Financial Data. Coppead Working Paper Series No. 355, 2003.

Vincenzo Verardi and Christophe Croux. *mmregress*, 2009. Stata package.

Jiahui Wang, Ruben Zamar, Alfio Marazzi, Victor Yohai, Matias Salibian-Barrera, Ricardo Maronna, Eric Zivot, David Rocke, Doug Martin, Martin Maechler, and Kjell Konis. *robust: Robust Library*, 2014. R package version 0.4-16.

R. E. Welsch and X. Zhou. Application of Robust Statistics to Asset Allocation Models. *REVSTAT—Statistical Journal*, 5(1):97–114, 2007.

Wharton School. Wharton Research Data Services, 1993. Available at https://wrds-web.wharton.upenn.edu/wrds/.

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

Gert Willems, Harry Joe, and Ruben Zamar. Diagnosing Multivariate Outliers Detected by Robust Estimators. *Journal of Computational and Graphical Statistics*, 18(1):73–91, 2009.

Peter Winker, Marianna Lyra, and Chris Sharpe. Least Median of Squares Estimation by Optimization Heuristics with an Application to the CAPM and a Multi-Factor Model. *Computational Management Science*, 8(1):103–123, 2011.

Victor Yohai. High Breakdown-Point and High Efficiency Robust Estimates for Regression. *The Annals of Statistics*, 15(2):642–656, 1987.

Victor J. Yohai and Ruben H. Zamar. High Breakdown-Point Estimates of Regression by Means of the Minimization of an Efficient Scale. *Journal of the American Statistical Association*, 83(402):406–413, 1988.

Victor J. Yohai and Ruben H. Zamar. Optimally Local Robust M-estimates of Regression. *Journal of Statistical Planning and Inference*, 64(2):309–323, 1997.

Victor J. Yohai, Werner A. Stahel, and Ruben H. Zamar. A Procedure for Robust Estimation and Inference in Linear Regression. In Werner Stahel and Sanford Weisberg, editors, *Directions in Robust Statistics and Diagnostics. Part II*, pages 365–374. Springer-Verlag Inc, 1991. The IMA Volumes in Mathematics and its Applications.

A. Zaman, Peter J. Rousseeuw, and M. Orhan. Econometric Applications of High-Breakdown Robust Regression Techniques. *Economics Letters*, 71:1–8, 2001.

Achim Zeileis. Econometric Computing with HC and HAC Covariance Matrix Estimators. *Journal of Statistical Software*, 11(10):1–17, 2004.

Kim Zetter. Six-Year-Old News Story Causes United Airlines Stock to Plummet. *Wired.com*, September 8, 2008.

Eric Zivot and Jiahui Wang. *Modern Financial Time Series with* S-Plus™. Springer-Verlag, New York, 2003.

Appendix A

EMPIRICAL FALSE POSITIVE RATES OF OUTLIER DETECTION TESTS BASED ON ROBUST SQUARED DISTANCES

A.1 Introduction

As we discussed in Chapter 2, Cerioli et al. (2009) showed that detecting outliers by comparing MCD-based robust Mahalanobis squared distances (RSDs) to chi-squared χ^2_ν quantiles resulted in higher than expected false positive rates in smaller samples (e.g., $n \leq 250$). Their study focused on RSDs based on the maximum-breakdown case of the MCD, which uses approximately half of the data to calculate the location and dispersion estimates. In some applications a practitioner might want to use an MCD estimator based on a larger subset of the data. Moreover, other robust dispersion estimates such as the orthogonalized Gnanadesikan-Kettenring (OGK) pairwise estimator (Gnanadesikan and Kettenring, 1972; Devlin et al., 1981; Maronna and Zamar, 2002), or the class of S-estimators might be more appropriate for some data sets. For example, the OGK estimator is computationally fast and often more practical with higher-dimensional data sets than methods like the MCD that require random resampling. Maronna et al. (2006) show (in their Chapter 6.8) that the bisquare and Rocke-type S-estimators are a better choice than the MCD for estimating location and dispersion under a point-mass contaminated multivariate normal model. Given the results of Cerioli et al. (2009), we were curious whether the MCD with asymptotic trimming fractions smaller than the maximum breakdown point case, as well as the OGK and S-estimators, suffered from the same issues as the maximum breakdown point version of the MCD.

This appendix extends the Cerioli et al. (2009) experiment by conducting similar studies for the MCD with asymptotic trimming fractions $\gamma = 0.25$ and $\gamma = 0.05$, as well as an OGK estimator and two S-estimators. We show that these estimators suffer from the same issues highlighted in Cerioli et al. (2009), to varying degrees: RSDs constructed using any of these estimators and tested against chi-squared quantiles can exhibit false positive rates for outlier detection tests that are much higher than expected. Hence, outlier detection tests based on RSDs using any of these methods need to employ some sort of correction methodology in samples smaller than 250. The results of this study motivated the work done in Chapter 2, and changes to the analyses presented in Chapter 3 from earlier versions presented in Martin et al. (2010).

A.2 Technical Background

Let $\mathbf{x}_i$, $i = 1, \ldots, n$ be the observations, and let $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\Sigma}}$ be estimates of the location vector and dispersion matrix of the data, respectively.

A.2.1 The MCD Estimator

We reviewed the minimum covariance determinant estimator MCD(γ) based on the asymptotic trimming fraction γ in Section 2.1 of Chapter 2. We refer the reader to that section for details on the MCD(γ) estimate.

Recall that γ^* is our notation for the asymptotic trimming fraction corresponding to the maximum breakdown point case of the MCD.

A.2.2 S-estimators

Recall (again from Chapter 2) that the (sample) Mahalanobis squared distance (MSD) of an observation $\mathbf{x}_i$ is defined as

$$d_i^2 = \left(\mathbf{x}_i - \tilde{\boldsymbol{\mu}}\right)^T \tilde{\boldsymbol{\Sigma}}^{-1} \left(\mathbf{x}_i - \tilde{\boldsymbol{\mu}}\right), \quad \text{(A.1)}$$

where $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\Sigma}}$ are estimates of the location vector and dispersion matrix, respectively. An S-estimate $(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$ of the location vector and dispersion matrix of the data $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is the solution to the following minimization problem:

$$\begin{aligned} \text{minimize} \quad & \det \tilde{\boldsymbol{\Sigma}} \\ \text{subject to} \quad & \frac{1}{n}\sum_{i=1}^{n} \rho\left(d_i\right) = b_0. \end{aligned}$$

Here ρ is a loss function that is non-decreasing on $[0, \infty)$, and b_0 is a tuning constant that controls the breakdown point BP of the estimate via the relation $BP = b_0 / \sup_u \rho(u)$ (Lopuhaä and Rousseeuw, 1991). If ρ is not bounded the breakdown point will be zero.

A common choice for ρ is the Tukey bisquare function introduced in Section 4.2 of Chapter 4.[1]

$$\rho(u; c_\nu) = \begin{cases} \frac{u^6}{6c_\nu^4} - \frac{u^4}{2c_\nu^2} + \frac{u^2}{2}, & |u| \le c_\nu \\ \frac{c_\nu^2}{6}, & |u| > c_\nu. \end{cases} \tag{A.2}$$

We find the tuning constant c_ν by (numerically) solving the equation $E\rho(u; c_\nu) = b_0$, where the expectation is with respect to the ν-dimensional normal distribution in our usual data setup. We use the notation c_ν to emphasize that the tuning constant depends on the dimension ν, a fact that will become important below. (When $\nu = 1$, we saw in Section 4.2 that $c_1 = 1.548$ yields a S-estimator with breakdown point 1/2.)

The S-estimator based on the Tukey bisquare has the undesireable property of increasing efficiency *and* bias with increasing dimension. Under the assumption of multivariate normality for the $\mathbf{x}_i$, the distances d_i^2 are χ_ν^2 distributed if the mean and covariance of the $\mathbf{x}_i$ are known, and asymptotically so when the sample mean and covariance are used in Equation (A.1). Hence the mean and standard deviation of the distances will be (asymptotically) $\mu_d = \nu$ and $\sigma_d = \sqrt{2\nu}$, respectively. When configured to yield an estimator with breakdown point 1/2, the Tukey bisquare puts zero weight on observations with distances larger than $c_\nu \sigma_d = c_\nu \sqrt{2\nu}$, which grows with the dimension ν. In higher dimensions there can be outliers that are inside this threshold and thus receive a positive weight in the S-estimate. Thus,

[1]Our parameterization of the bisquare is slightly different in this chapter to be consistent with how the bisquare appears in the S-estimator literature, e.g., Rocke (1996), and in the `R` package `rrcov`.

even though the resulting S-estimate cannot become arbitrarily large (since the S-estimator was configured to have breakdown point 1/2), its bias can be rather large.[2]

Rocke (1996) pointed out that one could avoid the problem above by controlling what he termed the "asymptotic rejection probability" (ARP) of the S-estimator. Motivated by this concept, Rocke defined the rejection point for an S-estimator as the smallest MSD such that any observation with a larger MSD receives a zero weight in the estimate:

$$d_{RP} = \inf\{d_0 | w(d) = 0, \forall d > d_0\},$$

where $w(u) = \psi(u)/u$ is the usual weight function for M- and S-estimation, and $\psi(u) = \rho'(u)$ is the derivative of the loss function $\rho(u)$. The ARP α_ν is then the probability, in large samples, that an observation is larger than the rejection point by chance alone, i.e., that the observation is incorrectly flagged as an outlier.

Theorem 1 of Rocke (1996) shows that, if an S-estimator is based on a (continuous) loss function ρ that can only adjust to the dimension ν of the data via scaling, then the ARP α_ν of the resulting estimator will tend to 0 as $\nu \to \infty$. Specifically, for large ν, Rocke derives the following asymptotic expression for $\log(\alpha_\nu)$:

$$\log(\alpha_\nu) \approx 0.5\nu \left(-2\log(M) - 1/M^2 + 1\right),$$

with M defined by $\rho(M) = BP \times \rho(d_{RP})$. Thus an S-estimator based on the Tukey bisquare, which merely rescales with increasing dimension, will have ARP approximately 0 in large dimension. Moreover, this result tells us that in order to construct an S-estimator whose ARP is independent of the dimension, the shape of the loss function ρ must vary more significantly with dimension.

To this end, Rocke introduced the biflat family of loss functions that can be tuned to achieve a desired breakdown point and ARP in any dimension. Maronna et al. (2006) provide a simplified version of Rocke's original biflat function, and this is used in the `rrcov` package

[2]This also explains why the efficiency of the bisquare S-estimator increases with dimension ν. The S-estimator becomes more like the sample covariance in higher dimensions since the threshold to reject observations is moved farther and farther out, and more observations receive full weight.

and our simulations below. For a desired ARP α, let z be the $1-\alpha$ quantile of a χ^2_ν distribution, and define $\zeta = \min(z/\nu - 1, 1)$. The biflat loss function $\rho(u;\zeta)$ is defined as follows:

$$\rho(u;\zeta) = \begin{cases} 0, & 0 \leq u \leq 1-\zeta \\ \left(\frac{u-1}{4\zeta}\right)\left[3 - \left(\frac{u-1}{\zeta}\right)^2\right] + \frac{1}{2}, & 1-\zeta < u < 1+\zeta \\ 1, & u \geq 1+\zeta. \end{cases} \tag{A.3}$$

Note that the loss function is not merely rescaled with the dimension: the extent of the non-constant "middle" of this loss function depends on the dimension ν through the definition of ζ. This avoids the problem of increasing bias with increasing dimension encountered with the bisquare.

A.2.3 The OGK Estimator

Gnanadesikan and Kettenring (1972) and Devlin et al. (1981) introduced a robust dispersion estimator, the pairwise Gnanadesikan-Kettenring (GK) estimator, based on computing robust correlations between pairs of variables. This would be advantageous in higher dimensions, where other estimators such as the MCD or S-estimators would be computationally difficult. The GK estimator is based on the identity

$$\mathrm{Cov}(x,y) = \frac{1}{4}\left(\mathrm{Var}(x+y) - \mathrm{Var}(x-y)\right).$$

Given a robust univariate estimator of scale $\tilde{\sigma}$, Gnanadesikan and Kettenring standardize each variable by a robust estimate of its scale, then use the above identity with the variance replaced by the "robust" variance $V(u) = \tilde{\sigma}(u)^2$:

$$\mathrm{RCorr}(x,y) = \frac{1}{4}\left(V\left(\frac{x}{\tilde{\sigma}(x)} + \frac{y}{\tilde{\sigma}(y)}\right) - V\left(\frac{x}{\tilde{\sigma}(x)} - \frac{y}{\tilde{\sigma}(y)}\right)\right). \tag{A.4}$$

They then define a robust pairwise covariance estimate by plugging the robust scale and correlation estimates into to the usual relationship between covariance and correlation:

$$\mathrm{RCov}(x,y) = \tilde{\sigma}(x)\tilde{\sigma}(y)\,\mathrm{RCorr}(x,y).$$

While this estimator is fast to compute, even in higher dimensions, it has two large drawbacks: is not affine equivariant, and the resulting robust covariance matrix may not be positive (semi-)definite.

The orthogonalized GK estimator (OGK), devised by Maronna and Zamar (2002), fixes these problems by forcing the identity

$$\mathrm{Var}(a^T x) = a^T \,\mathrm{Var}(x) a,$$

to hold for the eigenvectors a of the pairwise GK estimate of the dispersion matrix. The resulting estimator will be positive definite and approximately equivariant. Given a data matrix $\mathbf{X}$ with n observations (rows) of dimension ν (columns), a robust univariate location estimator $\tilde{\mu}$, and a robust univariate scale estimator $\tilde{\sigma}$, the OGK procedure is as follows.

1. Create a matrix $\mathbf{Y}$ by normalizing each column of $\mathbf{X}$ by its robust scale $\tilde{\sigma}(\mathbf{X}[,j])$. Equivalently, each row $\mathbf{Y}[i,]$ of $\mathbf{Y}$ is given by $\mathbf{Y}[i,] = \mathbf{D}^{-1}\mathbf{X}[i,]$, where $\mathbf{D} = \mathrm{diag}\left(\tilde{\sigma}(\mathbf{X}[,1]), \ldots, \tilde{\sigma}(\mathbf{X}[,\nu])\right)$.

2. Apply Equation (A.4), using $\tilde{\sigma}$ as the robust scale estimate, to pairs of columns of $\mathbf{Y}$ to compute a robust correlation matrix $\mathbf{U}$.

3. Compute an eigendecomposition $\mathbf{E\Lambda E}^T$ of $\mathbf{U}$.

4. Compute the matrix $\mathbf{Z}$ with rows $\mathbf{Z}[i,] = \mathbf{E}^T\mathbf{Y}[i,]$.

5. Calculate the robust location $\tilde{\mu}(\mathbf{Z}[,i])$ and robust scale $\tilde{\sigma}(\mathbf{Z}[,i])$ of each column of $\mathbf{Z}$. Define

$$\mathbf{\Gamma} = \mathrm{diag}\left(\tilde{\sigma}(\mathbf{Z}[,1])^2, \ldots, \tilde{\sigma}(\mathbf{Z}[,\nu])^2\right), \quad \boldsymbol{\delta} = \left(\tilde{\mu}(\mathbf{Z}[,1]), \ldots, \tilde{\mu}(\mathbf{Z}[,\nu])\right).$$

6. Finally, define the OGK estimates for the original data matrix $\mathbf{X}$ as

$$\widehat{\mathbf{\Sigma}}(\mathbf{X}) = \mathbf{DE\Gamma E}^T\mathbf{D}^T, \quad \widehat{\boldsymbol{\mu}}(\mathbf{X}) = \mathbf{DE}\boldsymbol{\delta}.$$

The resulting estimator is scale-equivariant (by Step 1) and positive definite (by the remaining sequence of steps). Maronna and Zamar suggest using the truncated standard deviation of Yohai and Zamar (1988) as the scale estimator $\tilde{\sigma}$, and a weighted mean with weights determined using the bisquare weight function as the location estimator $\tilde{\mu}$.

The breakdown point of the OGK estimator is at least as good as the breakdown points of the underlying location and scale estimates $\tilde{\mu}$ and $\tilde{\sigma}$. As usual, the efficiency of the OGK can be increased using a reweighting step: we compute RSDs using the raw OGK estimate, then assign weight 1 to observations with distances less than the threshold

$$\frac{\chi^2_\nu(\beta)\,\mathrm{med}(d_1,\ldots,d_n)}{\chi^2_\nu(0.5)}, \tag{A.5}$$

and weight 0 to observations with distances larger than the threshold. The reweighted location and dispersion estimates are then the sample mean and covariance of the observations with weight 1. Maronna and Zamar recommend using $\beta = 0.90$ in the above expression for the threshold. This is the default threshold for the reweighted OGK (ROGK) in `R`, as we use this threshold for the ROGK in our experiment below.

A.3 Experimental Setup

Our experimental setup is similar to that of Cerioli et al. (2009): for given values of sample size n and dimension ν we simulate $50,000$ independent samples from a multivariate normal distribution $N(\mathbf{0}, \mathbf{I}_\nu)$. For each sample, we estimate a robust location vector and dispersion matrix using each of the following estimators.

- MCD(γ^*), the maximum breakdown point case of the MCD;
- RMCD(γ^*), MCD(γ^*) followed by one-step reweighting using a hard rejection threshold of $\chi^2_{\nu,0.975}$;
- MCD(0.25);

- RMCD(0.25), MCD(0.25) followed by one-step reweighting using a hard rejection threshold of $\chi^2_{\nu,0.975}$;

- MCD(0.05);

- RMCD(0.05), MCD(0.05) followed by one-step reweighting using a hard rejection threshold of $\chi^2_{\nu,0.975}$;

- OGK, the OGK estimator using the τ-estimators described in Yohai and Zamar (1988);

- ROGK, the OGK estimator followed by one-step reweighting using the hard rejection threshold (Equation (A.5)) based on $\beta = 0.9$;

- the S-estimator with the Tukey bisquare ρ-function calibrated to have asymptotic breakdown point 1/2; and

- the Rocke S-estimator (Rocke, 1996) with asymptotic rejection probability 5%.[3]

The one-step reweighting cases compute RSDs based on the initial "raw" estimate, then assign weight 1 to observations with an RSD less than the stated threshold, and weight 0 to all other observations. The reweighted location and dispersion estimates are then the sample mean and covariance of the observations with weight 1.

For the tests of the intersection hypothesis we also consider versions of the above reweighted estimators with a Bonferroni-adjustment to the hard rejection threshold. (See Section 3.2 of Cerioli et al. (2009) for more details.)

- RMCD(γ^*).CH, the RMCD(γ^*) estimator but with hard rejection threshold $\chi^2_{\nu,1-(0.01/n)}$;

[3]The Rocke estimator, as implemented in the `rrcov` package, had computational difficulties for $n = 50$ and $\nu \geq 12$: the estimation methodology occassionaly encountered singular values and produced an error. The cause of the error seemed to be a bad random subsample with the estimation algorithm; advancing the random number generator's state usually resulted in better subsample. We have logged a bug report about the Rocke estimator implementation with the authors of the `rrcov` package.

- RMCD(0.25).CH, the RMCD(0.25) estimator but with hard rejection threshold $\chi^2_{\nu,1-(0.01/n)}$;
- RMCD(0.05).CH, the RMCD(0.05) estimator but with hard rejection threshold $\chi^2_{\nu,1-(0.01/n)}$; and
- ROGK.CH, the ROGK estimator but with hard rejection threshold based on $\beta = 1 - (0.10/n)$.

All of these estimators, with the exception of the Bonferroni-adjusted versions of the MCD estimators, are implemented in the `rrcov` R library (Todorov and Filzmoser, 2009). For the Bonferroni-adjusted versions of the MCD estimators we used the `covMcd` function from the `robustbase` R package with a slight modification to accommodate the different reweighting scheme.

Given a sample data set, we compute each robust location and dispersion estimate. We then calculate the corresponding RSDs for every observation in the sample. For the tests of the individual hypothesis (Equation 2.17 from Chapter 2), we measure the percentage of observations whose RSD exceeds a chosen critical value. We use a χ^2_ν quantile as a critical value with each estimator. For the MCD estimators (with and without reweighting) we also test RSDs against the Hardin-Rocke F distributional approximation with both the Hardin-Rocke methodology for estimating the Wishart parameter m, and the improved methodology developed in Chapter 2.

For the tests of the intersection hypothesis (Equation 2.3 from Chapter 2), we compare the largest RSD in each sample to a chosen critical value, and record whether it exceeds the threshold for each sample. We use Bonferroni-corrected χ^2_ν quantiles for each set of RSDs: if α is the nominal false positive rate for the intersection test, our test will use the $1 - (\alpha/n)$ quantile of the chi-squared distribution. For the MCD-based RSDs we again also use the Hardin-Rocke F distributional approximation with the Wishart parameter m estimated using the Hardin-Rocke estimator and the improved estimator from Chapter 2. In each case the quantile used to test the largest RSD is the $1 - (\alpha/n)$ quantile.

We used a superset of the values of n and ν used in Cerioli et al. (2009). We used sample sizes $n \in \{50, 75, 100, 150, 200, 300, 500, 1000\}$, and dimensions $\nu \in \{2, 4, 6, 8, 10, 12, 14, 16, 18, 20\}$.[4] We conducted our experiment using nominal test sizes of $\alpha = 0.01, 0.025$, and 0.05.

All calculations 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(R) E5410 2.33GhZ processor and 16GB of RAM, and runs Debian Linux 7.1.

A.4 Results

Results are shown for the nominal test level $\alpha = 0.01$ only, but similar results were obtained for $\alpha = 0.025$ and $\alpha = 0.05$.[5] The complete results of testing the individual and intersection hypotheses for each estimator are available in the accompanying `R` package `HardinRockeExtensionsSimulations`, detailed in the Appendix A.A.

In the figures below the (R)MCD(γ^*) estimator will be denoted "(R)MCDMBP".

A.4.1 Individual Outlier Hypothesis Test Results

MCD

First, our results for the MCD(γ^*) case with the χ^2_ν cutoff values are similar to those of Cerioli et al. (2009): we see in Figure A.1 that the false positive rates for non-reweighted MCD-based RSDs (green squares) are much larger than the nominal size, as much as 25 times too

[4]Cerioli et al. (2009) only considered even dimensions in their study, so we did the same for our main study. We did, however, run a test study using odd dimensions $5 \leq \nu \leq 19$ and only 5,000 simulations runs; the test study showed results very similar to those obtained with even dimensions. We therefore did not expand the experiment to cover the odd dimensions as well.

[5]For $\alpha = 0.025$ and $\alpha = 0.5$ the qualitative shape of the results matches what is shown in the Figures herein, e.g., Figure A.1, but the degree to which an estimator overpredicts the test size declines as the nominal size α increases. For instance, for the MCD(γ^*) case the individiual tests are about 25 times too large for small samples in the $\alpha = 0.01$ case, 10-12 times too large in the $\alpha = 0.025$ case, and 5-6 times too large in the $\alpha = 0.05$ case.

large for small samples ($n \leq 250$) in dimensions as small as $\nu = 8$. The performance of the χ^2_ν quantiles for MCD(γ^*)-based RSDs and fixed n deteriorates with increasing dimension ν. Convergence to the true rate $\alpha = 0.01$ only occurs in very large samples. Distances based on the reweighted MCD (Figure A.2) are better than the non-reweighted version for moderate sample sizes ($250 \leq n \leq 500$) but still very bad for small samples.

Testing MCD(γ^*)-based RSDs against the Hardin-Rocke F distributional approximation and the Hardin-Rocke estimator of m (green squares in Figure A.3) leads to tests that are larger than the nominal size (except in dimensions $\nu \leq 4$), though the performance in small samples is much better than with the χ^2_ν quantile. This agrees with the results presented in Table 1 of Cerioli et al. (2009). As the sample size increases the test size converges to the true rate of $\alpha = 0.01$, but there is noticeable undershooting of the realized false positive rate in moderate sample sizes as the dimension ν increases. This was not seen in the Cerioli et al. study since it only considered dimensions up to $\nu = 12$.

Using the Hardin-Rocke F distribution with the estimator of m developed in Chapter 2 (Figure A.4) also gives realized false positive rates that are much closer to the nominal rate of $\alpha = 0.01$ than the χ^2_ν quantiles. For small dimensions the modified methodology leads to test sizes that are too small. The original Hardin-Rocke approach gave slightly better results for $n < 200$. (Interestingly, both approaches yield realized false positive rates that are too small for $\nu = 2$.) Overall, both approaches are better than the χ^2_ν for testing individual observations for outlyingness with MCD(γ^*)-based RSDs.

Since the MCD(0.25) and MCD(0.05) estimators discard less of the data than the MCD(γ^*), we would expect the resulting dispersion estimate to be closer to the classical sample covariance, and the resulting RSDs to be closer to chi-squared distributed. That is indeed the case for the non-reweighted and reweighted versions of the estimators: Figure A.1 shows that the simulated sizes for the MCD(0.25)- and MCD(0.05)-based distances with χ^2_ν quantiles get closer to the nominal size of $\alpha = 0.01$, especially in small samples. Figure A.2 shows similar behavior for the reweighted versions. These estimators still lead to incorrect test behavior for small samples, however, and the performance degrades in small samples as

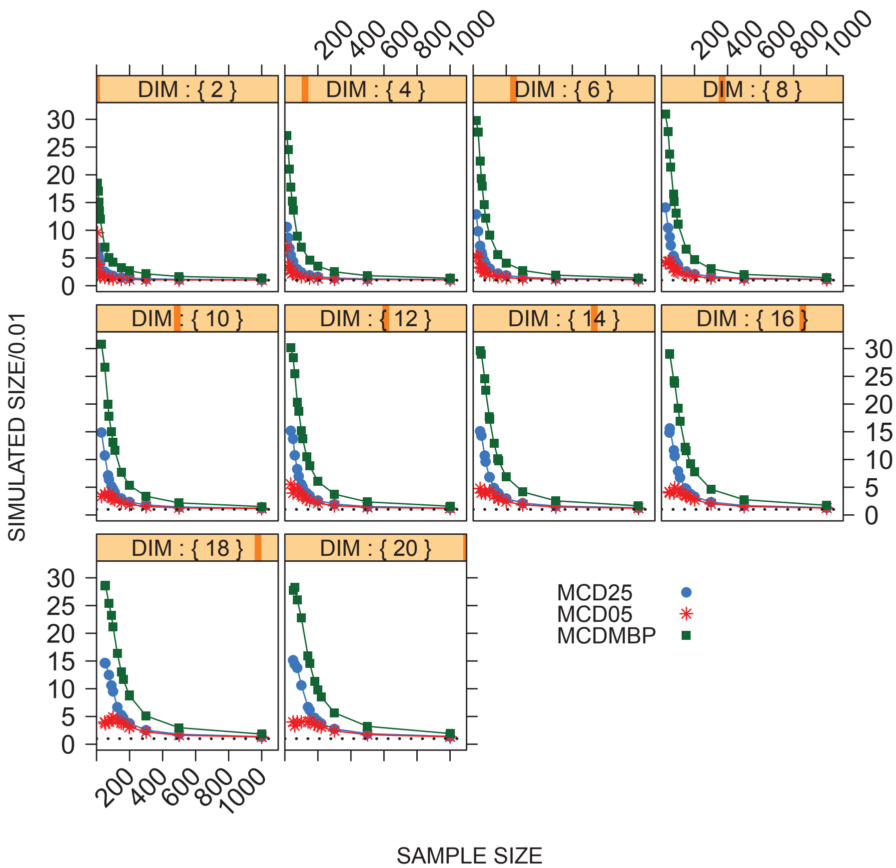


Figure A.1: Simulated sizes of the individual hypothesis tests for the MCD-based RSDs tested against χ^2_ν quantiles. The MCD(γ^*) estimator (denoted "MCDMBP") is represented by the green squares, the MCD(0.25) estimator is represented by the blue dots, and the MCD(0.05) estimator is represented by the red asterisks. Each box presents the data for a given dimension ν (shown in the orange header). The horizontal axis represents sample size n, and the vertical axis represents the simulated size of the test, in units of the nominal size 0.01.

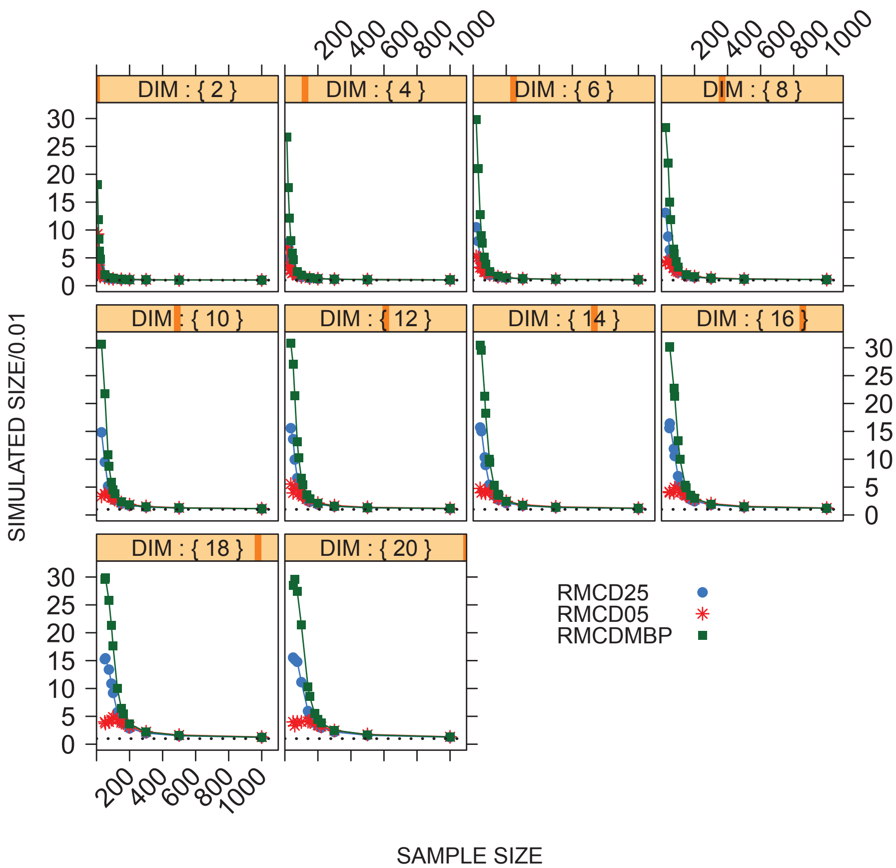


Figure A.2: Simulated sizes of the individual hypothesis tests for RSDs using the one-step reweighted MCD estimator (RMCD) and tested against χ^2_ν quantiles. The RMCD(γ^*) estimator (denoted "RMCDMBP") is represented by the green squares, the RMCD(0.25) estimator is represented by the blue dots, and the RMCD(0.05) is represented by the red asterisks. The setup of the plot is identical to that of Figure A.1.

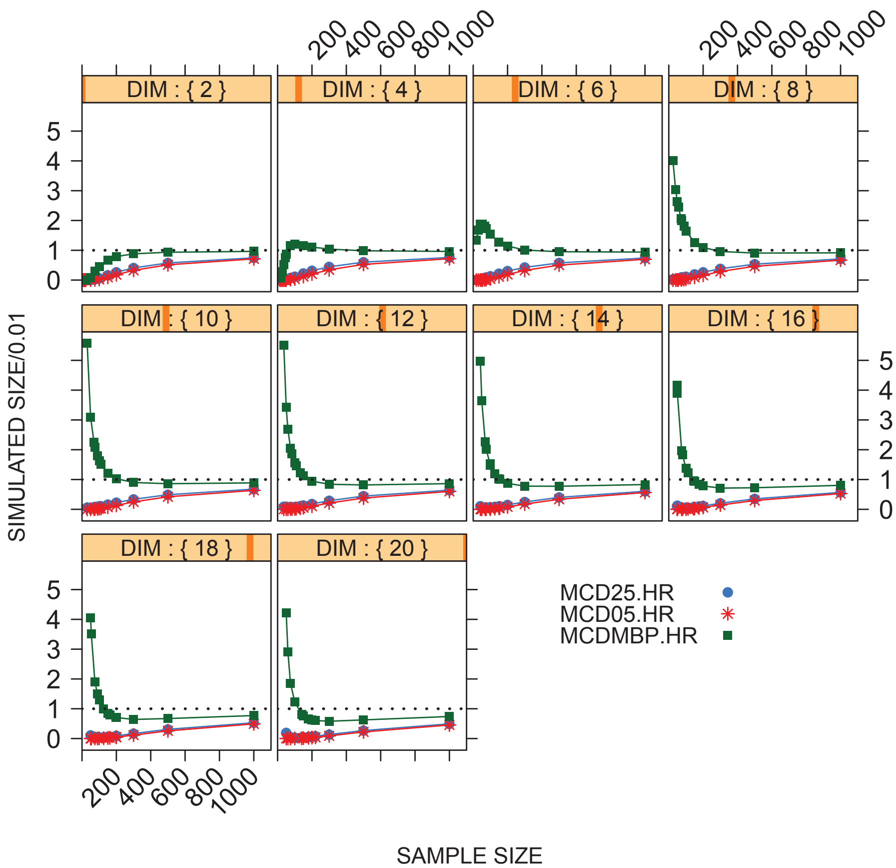


Figure A.3: Simulated sizes of the individual hypothesis tests for MCD-based RSDs compared against the Hardin-Rocke F quantiles with the Hardin-Rocke estimator for m. The setup of the plot is identical to that of Figure A.1.

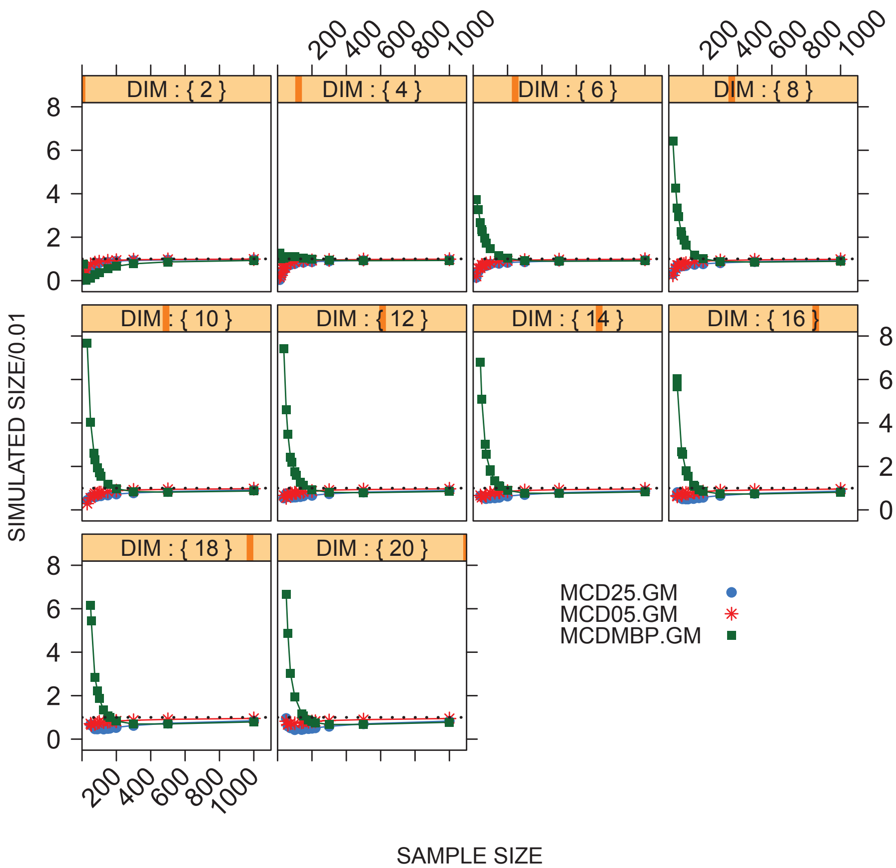


Figure A.4: Simulated sizes of the individual hypothesis tests for MCD-based RSDs compared against the Hardin-Rocke F quantiles with the estimator for m developed in Chapter 2. The setup of the plot is identical to that of Figure A.1.

the dimension increases (which agrees with observations by Andrews et al. (1973) and Small (1978)).

The original Hardin-Rocke estimator of m was only designed for the MCD(γ^*) case. For $\gamma = 0.25$ and $\gamma = 0.05$ it does not give very accurate false positive rates, as shown in Figure A.3. False positive rates are far too small in both cases (blue dots and red asterisks, respectively). The false positive rates converge to the correct nominal rate of α as the sample size n increases, but very slowly. Even for $n = 1000$ the test sizes are too small. With our modified estimator of m (Figure A.4), the observed test sizes are much closer to the correct size, even in small samples.

Figure A.5 shows how the four MCD methods compare for testing the individual hypotheses with data of dimension $\nu = 10$. We see a consistent pattern across values of γ: RSDs based on the raw MCD estimator and tested against χ^2_ν quantiles perform the worst, while distances based on the one-step reweighted version (also tested against χ^2_ν quantiles) are slightly better. The Hardin-Rocke F distributional approximation gives uniformly better results than the χ^2_ν distribution with either estimator of m: the empirical sizes of the tests are about right for moderate sample sizes (though they can be too small for smaller sample sizes). The middle and right panels of Figure A.5 show that for $\gamma = 0.25$ or $\gamma = 0.05$ our improved estimator of m yields tests with sizes much closer to the nominal size of $\alpha = 0.01$ than the Hardin-Rocke estimator of m.

OGK

Distances based on the non-reweighted OGK estimate and tested against chi-squared quantiles (blue dots in Figure A.6) give fairly good results for $n > 150$, even at high dimensions. In smaller samples the empirical false positive rate is not more than 5 times the expected rate. On the other hand, RSDs computed with the reweighted version of the OGK and tested against chi-squared quantiles (red asterisks in Figure A.6) perform very poorly. In small samples their behavior is much like that of the MCD, but even for $n = 1000$ the realized false positive rates are about 3 times too large. In the context of the results for the MCD

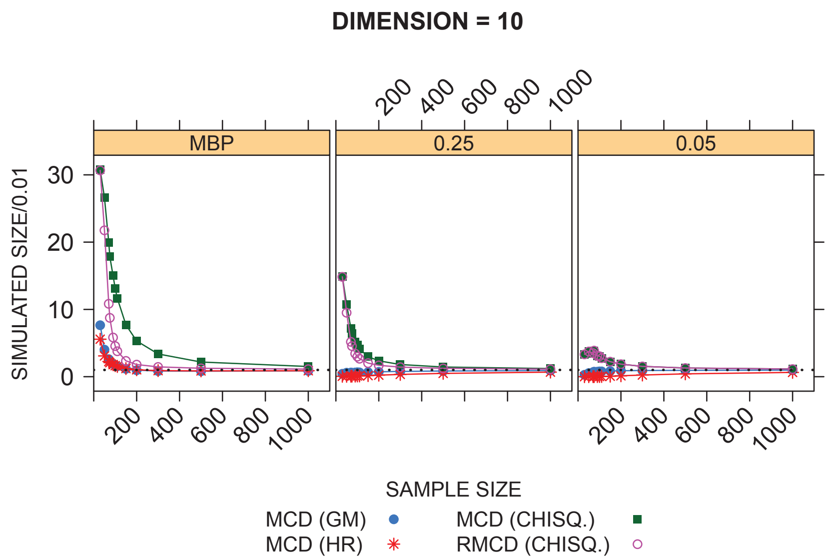


Figure A.5: Comparison of simulated sizes of the individual hypothesis tests for the MCD estimators on data of dimension $\nu = 10$. The fraction of the data used in the MCD is shown in the orange box at the top of each panel.

above, this is not terrible, but it is inconsistent with the behavior of the other estimators that asymptotically lead to tests of the correct size.

S-Estimators

Distances computed using the bisquare S-estimator and tested against chi-squared quantiles, shown as the blue dots in Figure A.7, exhibit test sizes that are larger than they should be, but not terribly so: even in small samples and high dimensions tests are only 3–4 times too large. This is much better than the corresponding tests with the MCD-based RSDs in similar situations. The behavior of the bisquare-based RSDs is fairly consistent across dimensions ν, and is asymptotically correct.

Distances based on the Rocke S-estimator and tested against chi-squared quantiles, shown as red asterisks in Figure A.7, result in test sizes similar to the MCD: they are way too big for small samples (with size increasing with dimension), and approximately correct for large

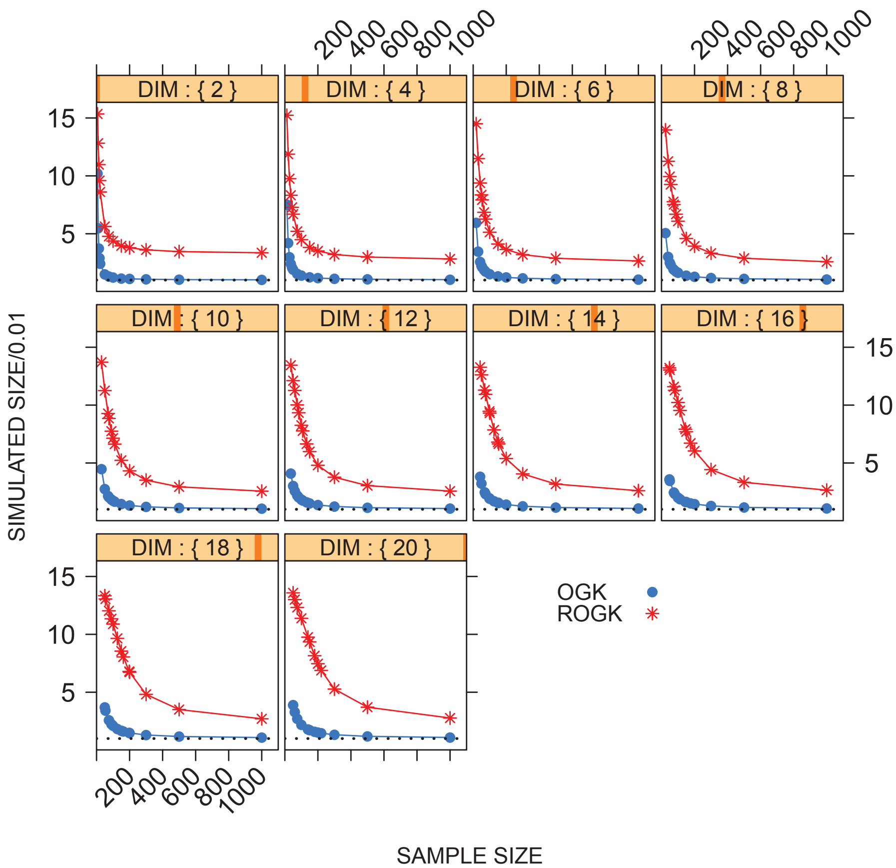


Figure A.6: Simulated sizes of the individual hypothesis tests for RSDs based on the OGK and one-step reweighted OGK estimator (ROGK), both using χ^2_ν quantiles for outlier detection. The OGK estimator is represented by the blue dots, and the ROGK estimator is represented by the red asterisks. The setup of the plot is identical to that of Figure A.1.

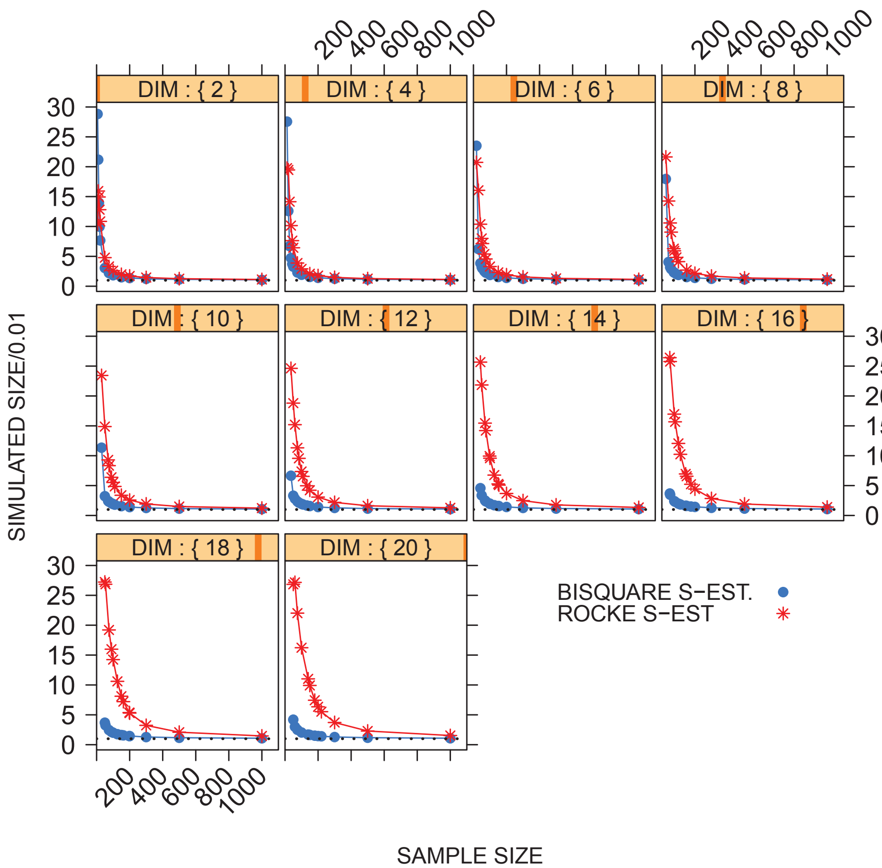


Figure A.7: Simulated sizes of the individual hypothesis tests for RSDs based on the bisquare and Rocke S-estimators using the χ^2_ν for outlier detection. The bisquare estimator is represented by the blue dots, and the Rocke estimator is represented by the red asterisks. The setup of the plot is identical to that of Figure A.1.

n.

Empirical False Positive Rates and Sample Sizes

In reviewing the above results for each robust dispersion estimate we were struck by the repeated occurrence of very high empirical false positive rates in small samples and larger dimensions. Figure A.8 shows how the empirical false positive rates for RSDs based on the robust dispersion estimates above vary with the ratio n/ν of sample size to dimension. When the individual hypothesis test results are viewed this way it becomes clear that the robust distance tests are generally less reliable when the sample size is less than 10 times the dimension, and particularly bad when $n/\nu \leq 5$.[6] The maximum-breakdown point case of the MCD and the two S-estimators are the worst in this respect. The MCD-based RSDs, tested against the Hardin-Rocke F distribution (with either estimator of m), along with the non-reweighted OGK, lead to robust distance tests that are more accurate than the other covariance estimators we consider here, but even these estimators can still have false positive rates 10 times larger than expected in smaller samples.

[6] In our simulation setup, the smallest value of n/ν is 2.5, arising from the case $n = 50$, $\nu = 20$.

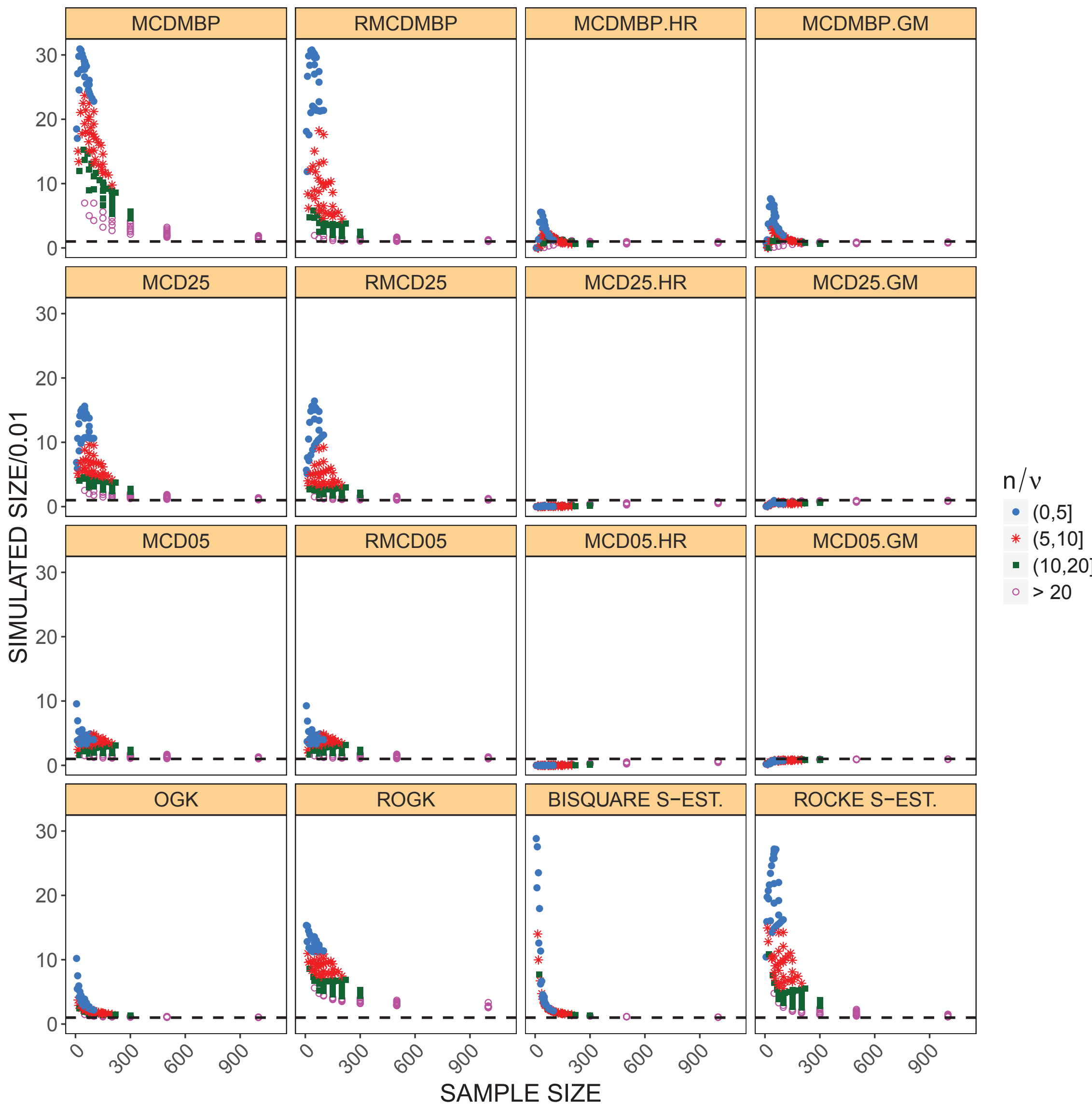


Figure A.8: Simulated sizes of the individual hypothesis tests for all estimators, stratified by the ratio n/ν of sample size to dimension. Each panel shows results for a different robust dispersion estimate (specified in the orange box above the panel). The ratios n/ν are binned into four ranges: $(0, 5]$ (blue filled circles), $(5, 10]$ (red asterisks), $(10, 20]$ (green squares), and > 20 (purple open circles). The false positive rate is specified on the y-axis as a multiple of the nominal rate 0.01.

A.4.2 Intersection Outlier Hypothesis Test Results

MCD

Figure A.9 shows the results of testing the intersection hypothesis with MCD-based distances and the χ^2_ν quantiles. Like Cerioli et al. (2009), we find that the MCD(γ^*)-based RSDs, tested against the chi-squared distribution, result in empirical test sizes that are nearly 100 times too large for $n \leq 250$ in dimensions greater than 8. This is also true for the MCD(0.25) and MCD(0.05) estimators. It is only in very large samples that we start to see the empirical test sizes get close to $\alpha = 0.01$: for $n = 1000$ the empirical test sizes are merely 4-6 times too large.

Figure A.10 shows the corresponding plots for RSDs using the one-step reweighted MCD estimator. Again our results agree with those of Cerioli et al.: the reweighted MCD estimators perform just as badly as the non-reweighted version for small samples. They are a bit better than the non-reweighted case for larger samples, in that the empirical test sizes seem to converge to the correct value faster than in the unweighted case. Making a Bonferroni correction to the hard rejection threshold in the reweighting step does not improve matters significantly (Figure A.11).

Figure A.12 shows the results for tests using the MCD-based RSDs and the Hardin-Rocke F distributional approximation with the Hardin-Rocke estimator of the Wishart parameter m. For dimensions $\nu > 4$, the empirical test sizes are smaller than the corresponding ones from the χ^2_ν, but still too large for small sample sizes. With two-dimensional data the Hardin-Rocke method yields test sizes that are too small. Cerioli et al. observed this odd behavior as well. With our improved estimator of m we see similar behavior (shown in Figure A.13), but worse performance in small samples compared to the original Hardin-Rocke method.

In the MCD(0.25) and MCD(0.05) cases, the tests using distances based on the MCD (with or without a reweighting step) and χ^2_ν quantiles all show empirical test sizes that are nearly 100 times too large in high dimensions and small samples. Using the Hardin-Rocke approach for these values of γ leads to test sizes that are far too small, even in large samples.

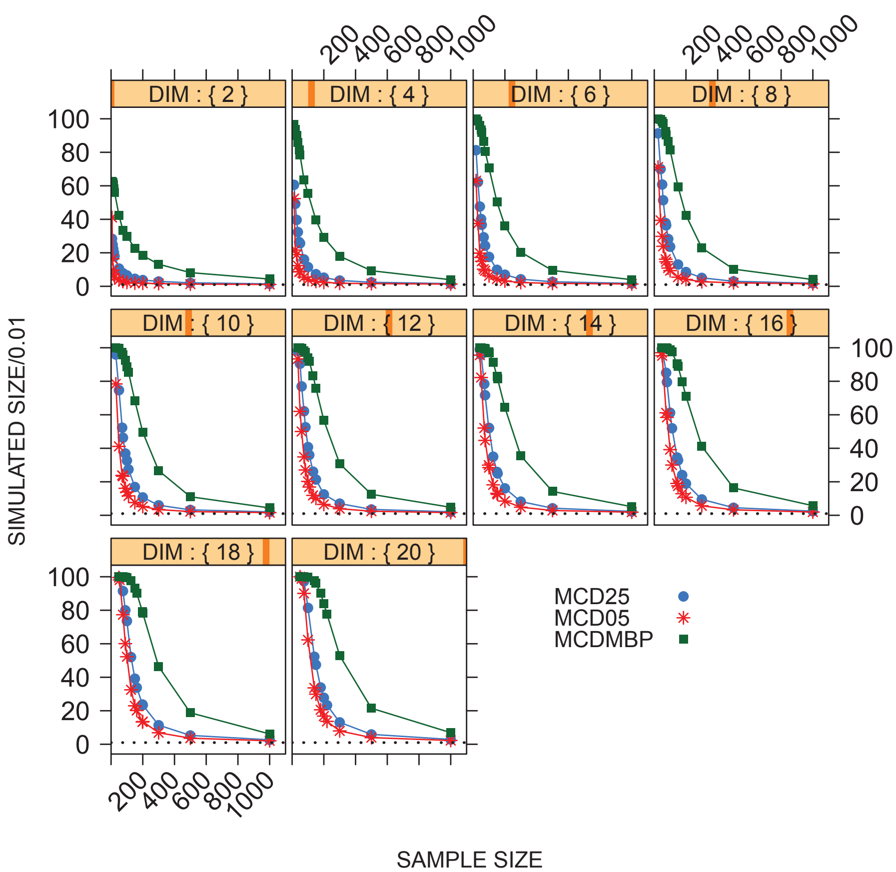


Figure A.9: Simulated sizes of the intersection hypothesis tests for the MCD-based RSDs and using χ^2_ν quantiles for outlier detection. The MCD(γ^*) estimator is represented by the green squares, the MCD(0.25) estimator is represented by the blue dots, and the MCD(0.05) is represented by the red asterisks. The plot setup is identical to that of Figure A.1.

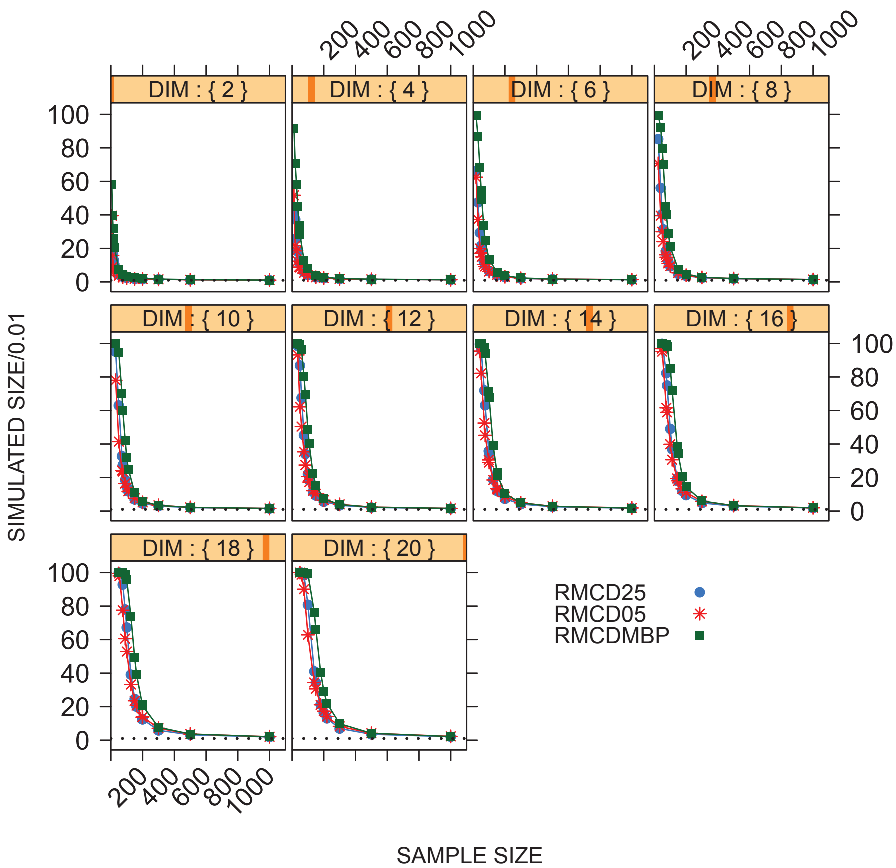


Figure A.10: Simulated sizes of the intersection hypothesis tests for distances based on the one-step reweighted MCD estimator and tested against χ^2_ν quantiles for outlier detection. The RMCD(γ^*) estimator is represented by the green squares, the RMCD(0.25) estimator is represented by the blue dots, and the RMCD(0.05) is represented by the red asterisks. The setup of the plot is identical to that of Figure A.1.

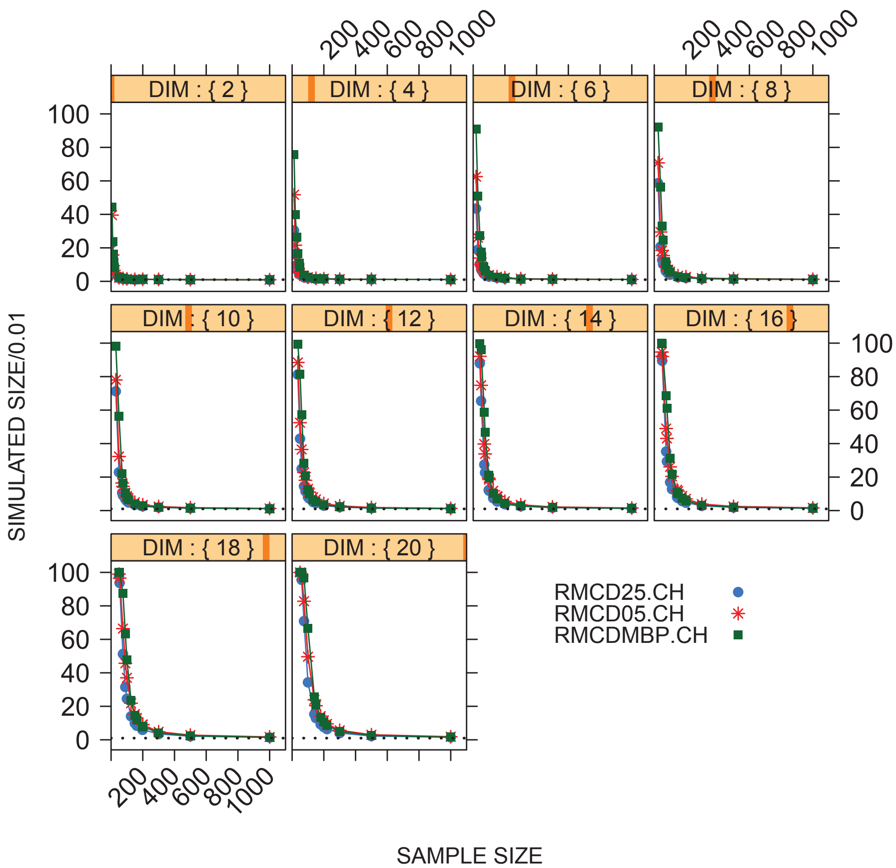


Figure A.11: Simulated sizes of the intersection hypothesis tests for distances based on the one-step reweighted MCD estimator, with a Bonferroni-corrected rejection threshold in the reweighting step (RMCD.CH), and tested against χ^2_ν quantiles for outlier detection. The RMCD.CH(γ^*) estimator is represented by the green squares, the RMCD.CH(0.25) estimator is represented by the blue dots, and the RMCD.CH(0.05) is represented by the red asterisks. The setup of the plot is identical to that of Figure A.1.

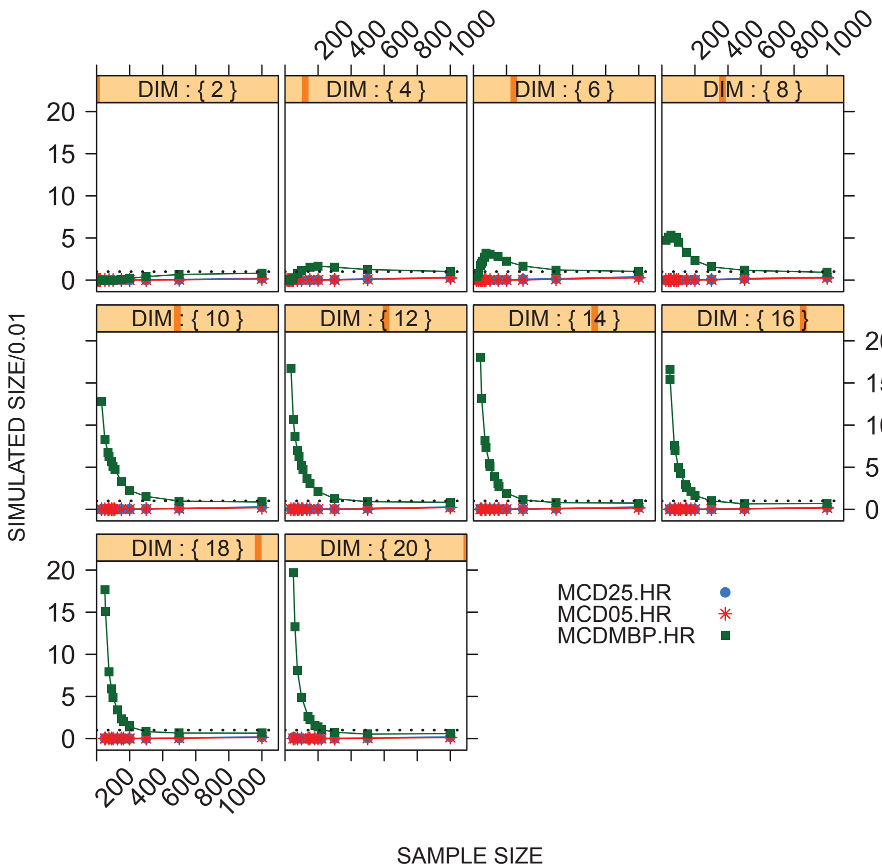


Figure A.12: Simulated sizes of the intersection hypothesis tests for MCD-based RSDs tested against quantiles from the Hardin-Rocke F distributional approximation (using the Hardin-Rocke estimator of m). The setup of the plot is identical to that of Figure A.1.

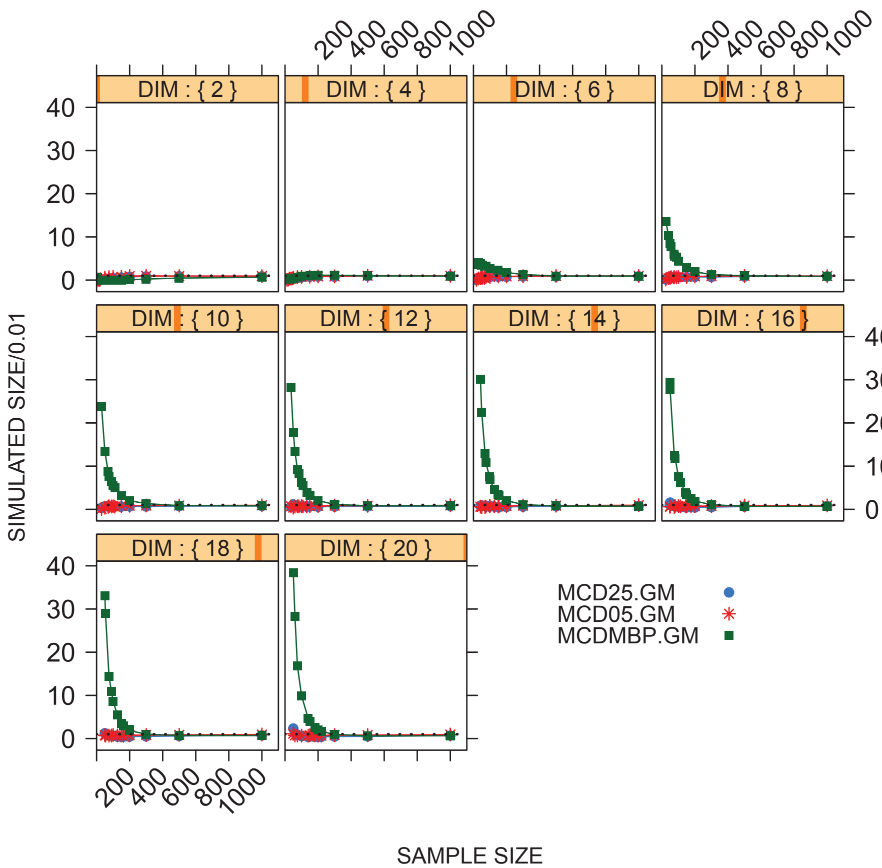


Figure A.13: Simulated sizes of the intersection hypothesis tests for MCD-based RSDs tested against quantiles from the Hardin-Rocke F distributional approximation (using the improved estimator of m from Chapter 2). The setup of the plot is identical to that of Figure A.1.

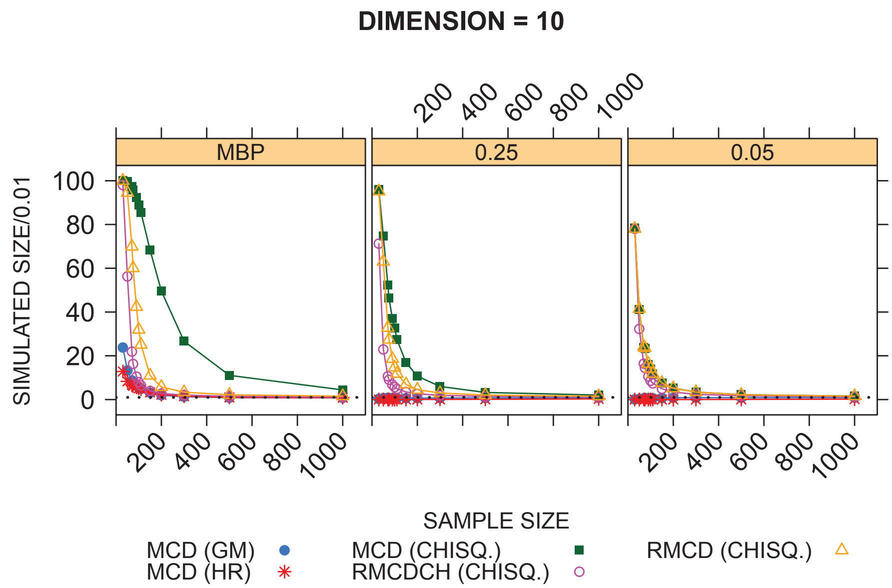


Figure A.14: Comparison of simulated sizes of the intersection hypothesis tests for RSDs based on the MCD estimators on data of dimension $\nu = 10$. The fraction of the data used in the MCD is shown in the orange box at the top of each panel.

The modified approach we developed in Chapter 2 gives much better results for moderate sample sizes.

Figure A.14 shows how the five MCD methods compare for testing the intersection hypothesis with data of dimension $\nu = 10$. We see a consistent pattern across values of γ: RSDs based on the raw MCD estimator and tested against χ^2_ν quantiles perform the worst. One-step reweighting in the MCD estimate improves matters a bit, but not enough to make the resulting RSDs reliable for testing the intersection hypothesis for $n \leq 500$. Testing RSDs against the Hardin-Rocke F distributional approximation with the Hardin-Rocke estimator of m for $\gamma = \gamma^*$, or our estimator of m for $\gamma = 0.25$ and $\gamma = 0.05$, gives much better results, though one should still expect to see incorrect test sizes in small samples.

OGK

Figure A.15 shows the results of the intersection hypothesis tests for RSDs based on the OGK estimators. All of the OGK estimators lead to outlier tests with empirical test sizes that are too big. Once again, the OGK is not as bad as the MCD estimator, in that for small samples the test sizes are only about 20–30 times too large. The reweighting step unfortunately makes the empirical test sizes of the resulting RSDs larger, though the Bonferroni correction does improve matters.

S-Estimators

Figure A.16 shows the results of testing the intersection hypothesis for RSDs based on the bisquare and Rocke S-estimator. Distances based on the bisquare S-estimator yields better test sizes than distances based on the MCD, but the resulting false positive rates are still 50–60 times too big in small samples ($n \leq 250$). Distances based on the Rocke S-estimator perform similarly to those using the MCD estimators.

Empirical False Positive Rates and Sample Sizes

Figure A.17 shows how the empirical false positive rates for tests of the intersection hypothesis using RSDs vary with the ratio n/ν of sample size to dimension. As with the individual hypothesis test results (Figure A.8), using the chi-squared distribution to determine whether outliers are present in the data is very unreliable when the sample size is less than 10 times the dimension. Empirical false positive rates can be 100 times too large when n/ν is smaller than 5, with the maximum-breakdown point case of the MCD and the two S-estimators again showing the worst performance. MCD-based RSDs tested against the Hardin-Rocke F distribution using either the Hardin-Rocke estimator of m or the modified version from Chapter 2, along with the non-reweighted OGK, lead to robust distance tests that are more accurate than the other dispersion estimators we consider here, but even these estimators can still have false positive rates quite larger than expected in smaller samples.

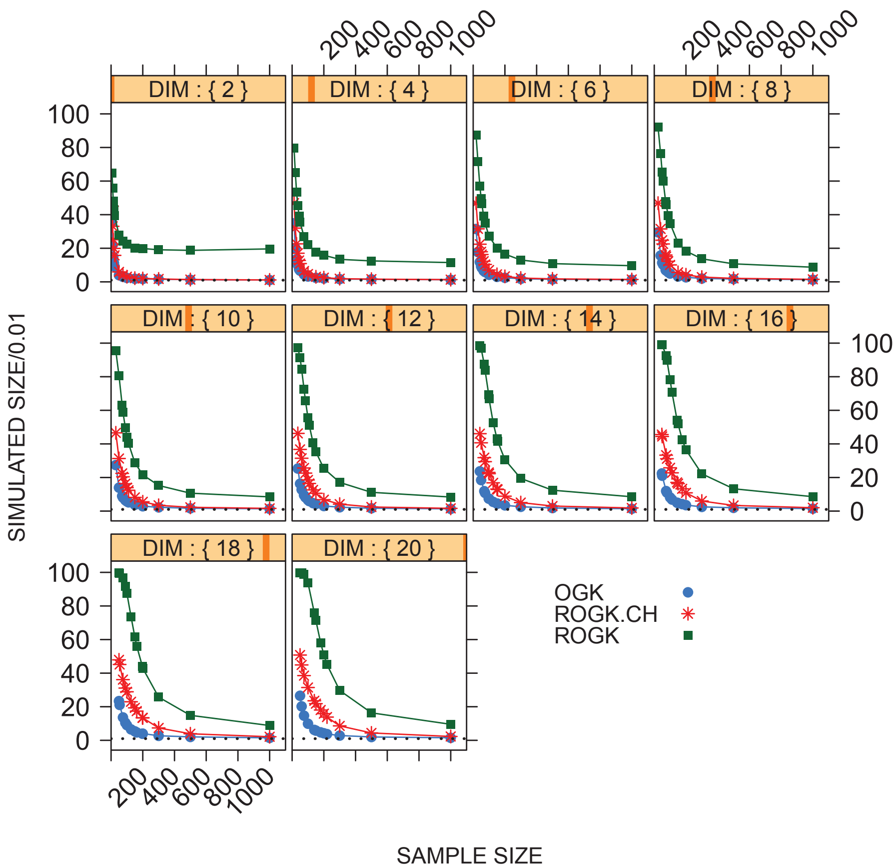


Figure A.15: Simulated sizes of the intersection hypothesis tests for distances based on the OGK and one-step reweighted OGK estimator (with and without a Bonferroni correction in the reweighting step). The estimators shown are the OGK (OGK, blue dots), the one-step reweighted OGK (ROGK, green squares), and the one-step reweighted OGK with a Bonferroni correction to quantile used for reweighting (ROGK.CH, red asterisks). All cases use χ^2_ν quantiles for outlier detection. The setup of the plot is identical to that of Figure A.9.

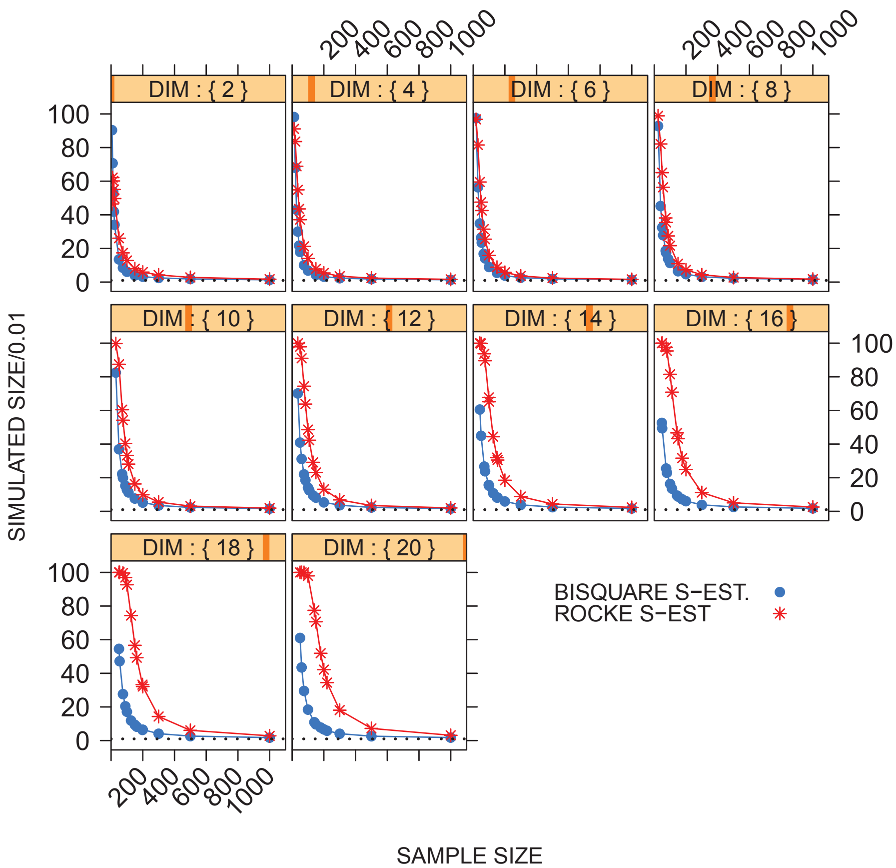


Figure A.16: Simulated sizes of the intersection hypothesis tests for distances based on the bisquare and Rocke S-estimators and tested against χ^2_ν quantiles. The bisquare estimator is represented by the blue dots, and the Rocke estimator is represented by the red asterisks. The setup of the plot is identical to that of Figure A.9.

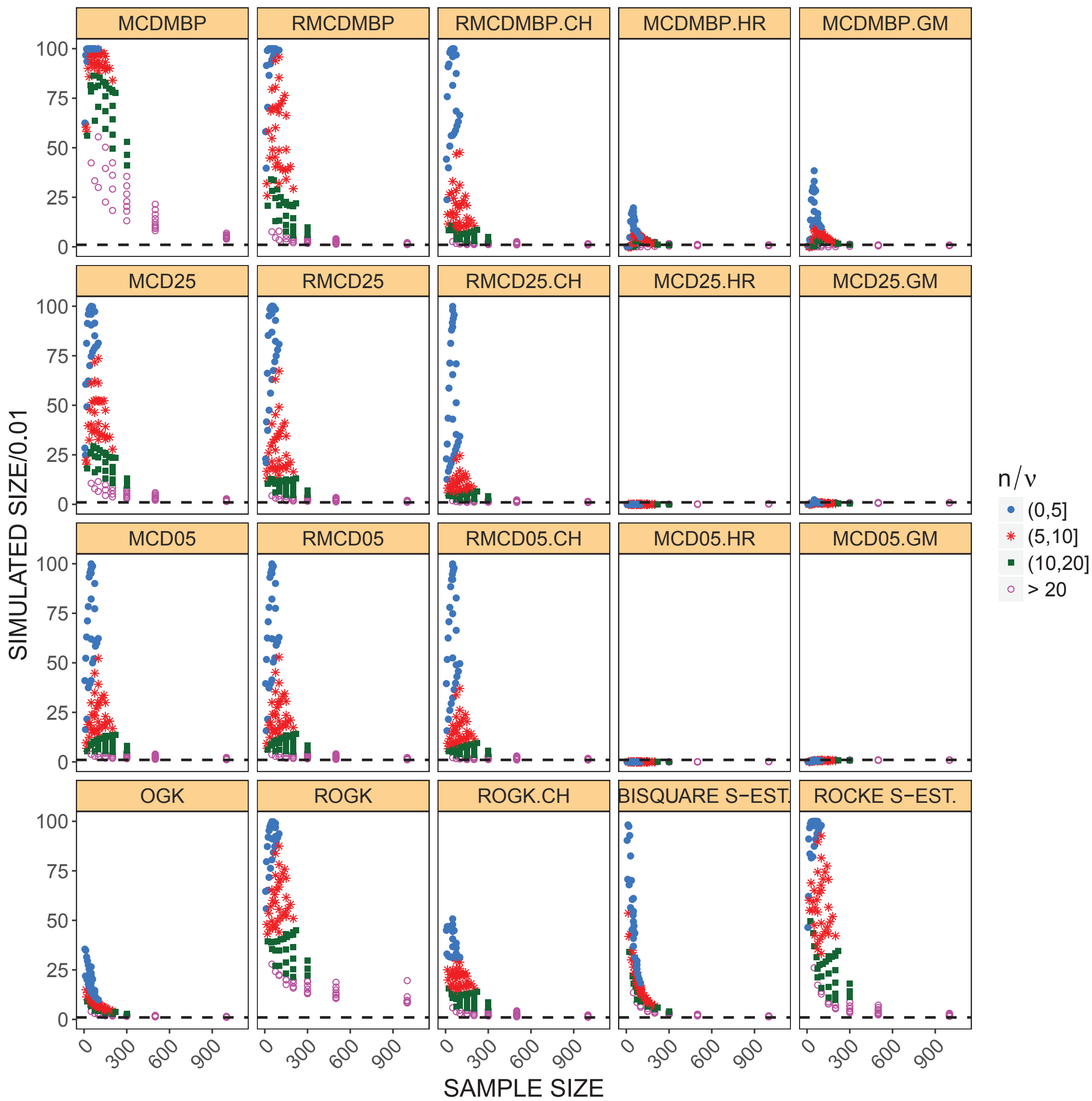


Figure A.17: Simulated sizes of the intersection hypothesis tests for all estimators, stratified by the ratio n/ν of sample size to dimension. Each panel shows results for a different robust dispersion estimate (specified in the orange box about the panel). The ratios n/ν are binned into four ranges: $(0, 5]$ (blue filled circles), $(5, 10]$ (red stars), $(10, 20]$ (green squares), and > 20 (purple open circles). The false positive rate is specified as a multiple of the nominal rate 0.01.

A.5 Discussion

The tests of the individual outlier hypothesis with our choices of robust dispersion estimators mirror the findings of Cerioli et al. (2009). Tests using the chi-squared distribution and RSDs derived from the MCD and reweighted MCD estimators result in far more false rejections than expected in small to moderate sample sizes. The Hardin-Rocke F distributional approximation, with the modified Wishart degrees of freedom estimate developed in Chapter 2, results in better test performance than the original Hardin-Rocke methodology in smaller samples, but can still fail to reject outlying points sufficiently often in moderate samples. Distances based on the OGK, reweighted OGK, and the bisquare estimators all give decent results compared to MCD-based distances, while RSDs based on the Rocke S-estimator (or possibly its `R` implementation) have some difficulties in small samples.

For the intersection hypothesis all estimators are woefully inaccurate for outlier detection in sample sizes smaller than 500. This echoes the result of Cerioli et al. (2009) for the MCD(γ^*) and variants. As discussed in that paper, (i) the chi-squared distribution is not appropriate in small samples for RSDs based on the MCD(γ^*) estimators; and (ii) the small sample corrections for the MCD developed by Pison et al. (2002) may not be sufficiently general. Our study hints that (i) might be true for non-MCD estimators as well. To the best of our knowledge, small sample corrections for the distribution of the non-MCD estimators have not yet been developed. Improving the behavior of all these robust estimators for testing the intersection hypothesis with RSDs seems to hinge on better modeling of the distribution of the RSDs for small to moderate sample sizes.

The reweighted OGK approach did not seem to give the right asymptotic behavior in both the individual and intersection hypothesis tests. The reason for this behavior is not immediately clear from the experiments conducted—perhaps the convergence happens more slowly than for the other estimators, or perhaps after reweighting the chi-squared distribution is no longer the right asymptotic distribution for the resulting Mahalanobis distances. More simulations with larger samples sizes would be an easy, if computationally expensive, way to investigate this further.

A.6 Conclusions and Further Research

We have confirmed the findings of Cerioli et al. (2009), namely that the chi-squared and Hardin-Rocke F distributional approximations for testing for the presence of outliers using MCD(γ^*)-based RSDs are insufficient for sample sizes less than 250. We have also extended their work to cover the MCD(0.25) and MCD(0.05) estimators as well as three other robust dispersion estimators: the OGK, bisquare S-estimator, and the Rocke S-estimator.

Generally speaking, the chi-squared approximation leads to higher false positive rates for sample sizes under 250, regardless of the estimator. For testing individual observations for outlyingness, the MCD tested against the Hardin-Rocke F distributional approximation using our estimator for the Wishart parameter m; the non-reweighted OGK with the chi-squared distribution; and the bisquare S-estimator with the chi-squared distribution are to be preferred. While inaccurate, these methodologies give false positive rates that are closer to the nominal test size than any of the other choices. For testing the intersection hypothesis of no outliers in the data, the MCD with our modification to the Hardin-Rocke methodology was the best approach, even though it can still flag too many outliers in small samples, and not enough in moderately-sized samples.

Small samples were consistently a problem for the outlier tests. For the MCD and its variants the correction factors of Pison et al. (2002) were used, but they are only first order corrections. Todorov (2008) has introduced new first-order corrections for small samples that we did not consider here. These may improve the performance of the MCD variants in the outlier tests. It is possible that higher order corrections are needed as well. Finally, correction factors to reduce the small sample bias of the S-estimators and the OGK estimators are lacking and need to be developed.

APPENDIX

A.A Replicating the Cerioli et al. (2009) Experiment

The full results of our replication and extension of Cerioli et al. (2009) are available in the `R` package `HardinRockeExtensionSimulations`. We provide both the estimated test sizes (the average percentage of outliers detected over 50,000 runs) and the standard deviation of the percentages from the simulation. These results should be compared to their Tables 1 and 3.

Code to replicate this work is available in 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/`

R Session Details

```
> sessionInfo()
R version 3.2.4 Revised (2016-03-16 r70336)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 7 x64 (build 7601) Service Pack 1

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

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

other attached packages:
[1] Hmisc_3.17-4    ggplot2_2.1.0   Formula_1.2-1   survival_2.39-4
[5] lattice_0.20-33
```

```
loaded via a namespace (and not attached):
 [1] Rcpp_0.12.5         chron_2.3-47        grid_3.2.4          plyr_1.8.4
 [5] gtable_0.2.0        acepack_1.3-3.3     scales_0.4.0        data.table_1.9.6
 [9] latticeExtra_0.6-28 rpart_4.1-10        Matrix_1.2-4        splines_3.2.4
[13] RColorBrewer_1.1-2  tools_3.2.4         foreign_0.8-66      munsell_0.4.3
[17] colorspace_1.2-6    cluster_2.0.4       nnet_7.3-12         gridExtra_2.2.1
```

VITA

Christopher George Green obtained his Bachelor's Degree (summa cum laude) in Mathematics from Washington University in St. Louis in May 1999. He obtained a Master's Degree in Mathematics (August 2001) and a Master's Degree in Statistics (December 2011), as well as a Certificate in Computational Finance (August 2007), from the University of Washington in Seattle. In June 2017 he graduated with a Doctor of Philosophy in Statistics from the University of Washington.