Diagnosing the Presence of Multivariate Outliers in Fundamental Factor Data using Calibrated Robust Mahalanobis Distances

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

University of Washington, Seattle, WA, USA

Working Paper
Department of Statistics
University of Washington
Seattle, WA

Revision date: **May 31, 2015**

Abstract

It is well-known that outliers exist in the type of multivariate data used by financial practitioners to construct fundamental factor models of stock returns. 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 to mitigate multivariate outliers, which may not be considered outliers in any individual variable despite being outlying in higher-dimensional views of the data. Existing literature documents the use of the robust Mahalanobis distance (Distance) based on robust covariance matrix estimators to detect and to shrink multivariate outliers in financial data. The minimum covariance determinant (MCD) estimate is commonly used for this purpose; see, for example, Scherer and Martin (2005); Boudt et al. (2008); Martin et al. (2010). For pure detection purposes with reliable false alarm rates, an accurate approximation to the distribution of the squared Distance is needed; Cerioli (2010) has developed a correction methodology which yields outlier detection tests with the correct Type I error behavior. We use the Cerioli approach to illustrate the presence of outliers in the type of firm fundamental/characteristic data that equity portfolio managers would use to build fundamental factor models. We demonstrate how Mahalanobis distances based on the MCD estimate are superior to the distances based on the classical mean and covariance estimates for detecting multivariate outliers. Finally, we provide evidence that univariate trimming and Winsorization are insufficient to deal with multivariate outliers in fundamental factor data.

[1]Address for correspondence: Department of Statistics, Box 354322, University of Washington, Seattle, WA 98195-4322, USA. E-mail: `cggreen@uw.edu`

[2]Department of Applied Mathematics, Box 353925, Univerity of Washington, Seattle, WA 98195-3925, USA.

1 Introduction

It is well-established in the robust statistics literature that outliers can adversely affect the outcome of classical statistical procedures such as parameter estimation and hypothesis testing. By now, most equity portfolio managers and commercial equity portfolio management software developers recognize that outliers in the factor exposures can cause problems in the construction of fundamental/characteristic (multi-) factor models. It is common practice to reduce the influence of outliers in the factor exposures by trimming or Winsorizing each factor exposure; see, for instance, (Grinold and Kahn, 2000, 382–383). 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.[3]

Trimming or Winsorizing the input factor exposures in this manner only addresses univariate outliers, however; multivariate outliers may be unaffected by this procedure. Figure 1 shows a simple example of how this can happen. We simulated 1000 observations from a bivariate normal distribution with both marginal means equal to 2, both marginal variances equal to 1, and correlation 0.75. The dashed lines indicate the 1% and 99% percentiles of each marginal variable. The red triangles are the observations that would have been deleted (if we were trimming by 1%) or replaced (if we were Winsorizing), while the black dots are the observations left unchanged by trimming or Winsorizing.

Due to the configuration of the observations in this case, there is quite a bit of area outside the bulk of the data but within the trimming bounds. The two blue squares indicate two data values we could add to the dataset that would not be trimmed or Winsorized under the 1% criterion. The effect of these two data values on estimation would be very dependent on the structure of the estimator used. More generally, the two blue squares could be replaced with small clusters of observations, corresponding to a data-generating process that is a mixture of several distributions. Neither trimming nor Winsorizing would affect these points much, but they would still be considered multivariate outliers relative to the bulk of the data under any reasonable definition of "outlier".

Of course, one could choose to trim or Winsorize by a larger amount, say 5% or 10%; that may work for this example, but in general this results in more drastic changes to the data set, and hence, to the estimates. In practice, the number of dimensions is much higher (for example, consider a universe of 1000 stocks), so finding multivariate outliers is much harder, and the matter of how much to trim or Winsorize each exposure is entirely subjective.

In addition, Winsorization can introduce artificial structure to the data. Suppose an observation is "outlying" in more than one variable. Winsorization of each variable separately defines an n-dimensional "box" (e.g., see the dashed lines in Figure 1). An observation that is outlying in k variables gets mapped to a hyperplane of dimension $n - k$. (So in Figure 1, some red triangles will be mapped to an edge of the box, while some will get mapped to a corner of the box.) This introduces structure to the data set that was not there prior to Winsorization. This structure can potentially have an adverse effect on estimation methods: for example, a robust covariance estimator such as the FastMCD estimator (Rousseeuw and van Driessen, 1999) will detect the existence of a subset of the data that lies in a lower-dimensional hyperplane. Moreover, the additional structure can make an estimated covariance matrix singular[4], invalidating the asymptotic

[3]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.

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

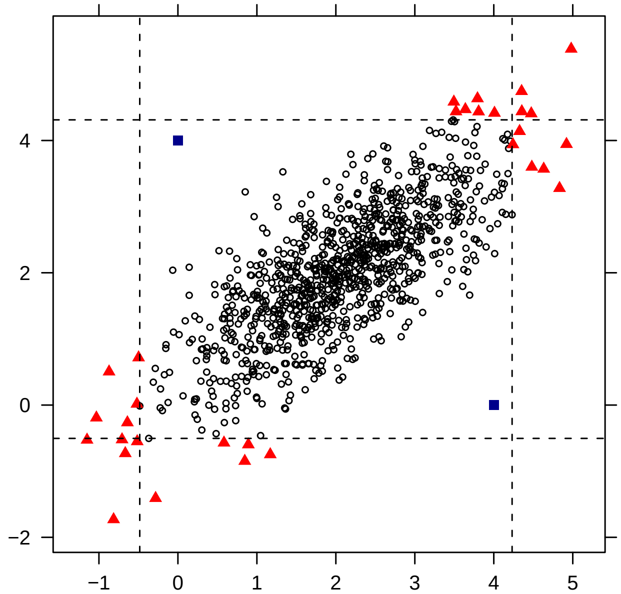


Figure 1: Example of two-dimensional outliers (blue squares) that are missed by one-dimensional trimming or Winsorization.

distribution theory one typically uses for computing confidence intervals and test statistics.

Multivariate outliers can arise quite easily in fundamental/characteristic factor model construction. 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.) The price outlier will lead to a returns outlier (possibly two, if the price outlier is due to some transitory effect)[5]. Price also factors into the computation of market capitalization and earnings-to-price. The returns outlier leads to outliers in a moving average of returns, and the market capitalization outlier could lead to outliers in a book-to-market ratio. This can be particularly problematic for large factor models that have many fundamental factors that are derived from the same quantities.

In a multiple factor model context, it is not well-understood how multivariate outliers affect the calculation of the factor exposures and/or factor returns (depending on the type of factor model). Linear factor models generally take the form

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

where $r_{i,t}$ is the return on the ith stock at time t; $\beta_{i,j}$ is the exposure of the ith stock to the jth factor; $f_{j,t}$ is the return to the jth factor at time t; α_i is a persistent, stock-specific return, 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 the case of a fundamental factor model, it is the factor returns and asset residuals that need to be estimated from the observed exposures and returns. This estimation is commonly performed via (Fama-MacBeth) cross-sectional regression: at each time point, asset returns are regressed on the exposures to the fundamental

the invertibility of the covariance estimate.

[5] 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 such transitory outliers 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).

factors (as of the prior time point) to determine the factor returns for that time point. Generalized least squares (GLS) must be used to account for the differing residual variances among the stocks; since the GLS weights are not known beforehand they too must be estimated. An initial ordinary least squares (OLS) fit can be used to obtain estimates of the residual variances, which can then be used as (inverse) weights in GLS. However, it is known that ordinary least squares (OLS) can be greatly influenced by outliers, in the asset returns and/or in the factor exposures, leading to biased estimates of the coefficients of the regression and of the residuals (see, for example, Rousseeuw and Leroy (1987); Maronna et al. (2006)). Hence, in the presence of outliers, the residuals from the initial OLS run can be too large or too small; since the weights for GLS are estimated from the residuals of the initial OLS run, the resulting GLS factor returns can be poorly estimated as well.

Outliers may also pose a problem prior to the regression: it is common practice to standardize the factor exposures before regression to facilitate the interpretation of the fitted factor model. This standardization is commonly accomplished at each time point for each factor exposure by subtracting the sample mean (over assets) and dividing by the sample standard deviation (again over assets). Since both the sample mean and sample standard deviation can be drastically influenced by outliers, however, the resulting standardized factor exposures may look quite strange when there are outliers in the unstandardized factor exposures. For example, consider 100 points drawn from a normal distribution with mean 0.05. The expected value of the sample mean is 0.05. If 10 of those observations were replaced by 0.50 (either due to a data error or due to a different data generating process), the expected value of the sample mean is easily shown to be 0.095. Subtracting off the larger value as part of the standardization process will still yield an expected sample mean of 0 for the entire data set, but the 90 uncontaminated points will have an expected sample mean of -0.045 post-standardization, not 0. Similarly, the 10 larger outliers will inflate the sample standard deviation, leading to most of the standardized observations being smaller than they would be, absent the data contamination.

The statistical community has developed many estimation and testing procedures that perform “well” in the presence of multivariate outliers, where “well” is defined in some problem-dependent sense. The field of robust statistics was developed largely in response to problems with classical methods in the presence of outliers; see for instance, Huber (1981); Hampel et al. (1986); Maronna et al. (2006). Robust statistical methods are, in general terms, statistical procedures which are designed to perform well at an assumed model (e.g., a multivariate normal distribution) and at small deviations from this model (such as those caused by outliers). While robust methods can certainly be used as replacements for classical methods, they are perhaps more valuable as diagnostic tools: 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 data oddities that need to be investigated before any conclusions are drawn from the research.

Robust statistical methods are also useful for detecting outliers in multivariate data. One traditional method of detecting outliers in multivariate data is through the use of Mahalanobis distances. Mahalanobis distances, introduced by 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). Formally, let $\mathbf{x}$ be an observation from a multivariate distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$; then the squared Mahalanobis distance of $\mathbf{x}$ from $\boldsymbol{\mu}$ is defined as

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

When $\mathbf{x}$ is ν-dimensional multivariate normal with known mean and covariance, the population squared

Mahalanobis distance is distributed as a chi-squared χ^2_ν random variable with ν degrees of freedom (Mardia et al., 1979). This suggests a test of deviation from the multivariate normal assumption: compare an observation's squared Mahalanobis distance to an appropriate quantile of the chi-squared distribution. 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

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T, \tag{3}$$

the sample covariance matrix. When the $\mathbf{x}_i$ are multivariate normal, the resulting sample Mahalanobis distances are approximately chi-squared for "moderate" values of n, but "moderate" increases with the dimension ν (Andrews et al., 1973; Small, 1978). In fact, Cerioli et al. (2009) showed that using the chi-squared distribution can lead to more false positive detections than we would expect (on average) given our chosen Type 1 error rate α. The follow-up paper by Cerioli (2010) describes a modified approach to detect outliers using robust Mahalanobis distances that yields the expected false positive rates; we will use this method in this paper rather than the traditional chi-squared approach.

While robust statistical methods have been studied in the statistical community since the 1960s, they have only recently started to find use in equity portfolio management. Several authors have examined the use of robust estimation in portfolio construction; see Cavadini et al. (2002); Lauprete (2001); Lauprete et al. (2002); Vaz-de Melo and Camara (2003); Perret-Gentil and Victoria-Feser (2003); Welsch and Zhou (2007); DeMiguel and Nogales (2009). Boudt et al. (2008) used Mahalanobis distances to detect and remove outliers in data prior to estimating value-at-risk and expected shortfall. In the context of factor model construction, Knez and Ready (1997) used least-trimmed squares robust regression to illustrate the effects of outliers in Fama-French style factor models. Stephan et al. (2001) constructed a multiple factor model for European stocks using the Barra methodology; they dealt with outliers (in a univariate fashion) via a skipped Huber method. Robust estimation of beta via a single factor model has been studied in some form since 1971; see Bailer et al. (2011) for a concise history of the research in that area as well as an examination of how robust regression can yield more reliable estimates of beta. The dissertation of Bailer (2005) and subsequent paper Bailer and Martin (2007) pioneered the application of robust methods to fitting fundamental factor models using the Fama-Macbeth cross-sectional regression methodology. The text of Scherer and Martin (2005) devotes an entire chapter to applications of robust methods to finance, while the review paper of Martin et al. (2010) presents some recent applications to equity portfolio management.

Robust methods are becoming more widely used in both academic research and the financial industry as more people learn of their value and as more ready-to-use software appears (for example, the `robustbase` and `robust` software packages for the `R` programming language). However, while those in the finance community who build factor models are certainly aware of the need to address outliers in analyses, they seem generally either unaware of multivariate robust statistical methods, or unconvinced of their need in financial modeling.[6] Univariate outlier mitigation approaches remain commonplace in academic and professional work. Hence, the primary contribution of this paper is to provide evidence of multivariate outliers in the market and firm fundamental data that a manager might use to build a factor model. We demonstrate how to use robust Mahalanobis distances to detect such outliers in the data. We also examine how trimming and

[6] There have been many recent papers that take other approaches to outliers in factor models. For example, there are several papers that address skewness and kurtosis in factor modeling, both of which can be caused by outliers (but are not the only cause of such non-normality). A discussion of all these approaches is beyond the scope of this paper.

Winsorization change the structure of the data and number of outliers detected. The analysis of how the multivariate outliers affect the factor model is much more complicated, however; we will address this problem in a future paper.

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 Mahalanobis distances 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 Mahalanobis distances have come to light through the work of Cerioli et al. (2009). Thus, we have redone and extended the analysis of that paper with the corrected methodology presented in Cerioli (2010) and Green and Martin (2014b) to test whether our results still hold. This paper also provides much more detail on the nature of the outliers than we were able to provide before (due to reasons of space).

2 Experimental Setup and Methodology

2.1 Construction of the 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.[7] We removed any assets that were not common stocks (such as ADRs and closed-end funds).[8]

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 instantaneously (and 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.[9] Asness and Frazzini showed that using the current market data to compute a firm's book value-to-price 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). Hence, we adopt their approach here rather than the Fama and French approach.

Next, 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

[7] 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.

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

[9] 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.

volatile returns (and more non-normally distributed returns). (Some empirical support for this observation can be found in Amaya et al. (2013) and Blau et al. (2013).) 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. Thus, we might see outlier dynamics in smaller stocks that are significantly different from those observed in larger stocks. We therefore split the resulting 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.[10] 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 in their groups from July 1 of year t to June 30 of year $t + 1$.

Next, we computed the following factor exposures, all of which are typical factors used in fundamental factor models. (For some empirical justification of these particular factors, see Fama and French (1992) and the references therein.)

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

- The book-to-market ratio ("BOOK2MARKET"), the ratio of a firm's book value to its market capitalization.[11] 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.

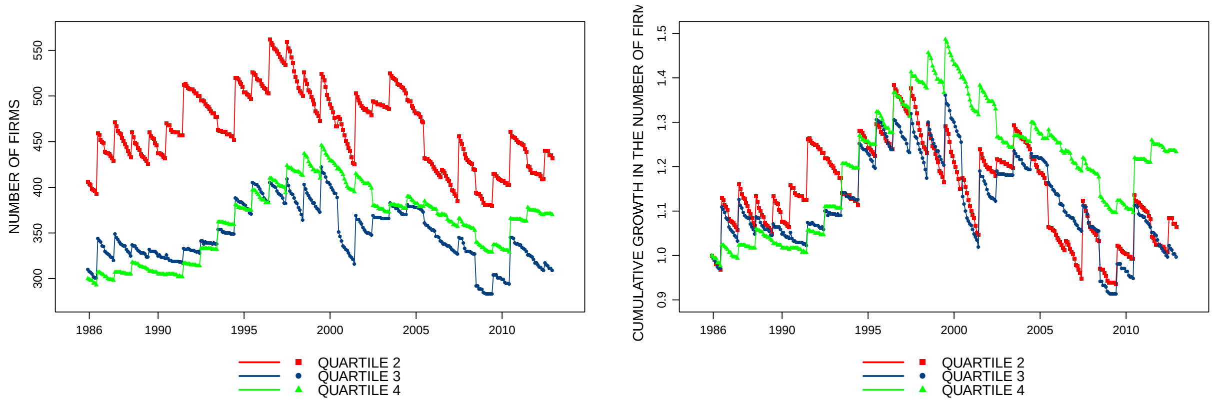
- The earnings-to-price ratio ("EARN2PRICE"), the ratio of a company's (trailing) earnings per share to its price. The earnings-to-price ratio, also known as the earnings yield, allows investors to compare a company's earnings to bond yields (e.g., to see if they are being compensated for taking equity 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 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.

Finally, 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

[10] 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.

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



(a) Sample sizes

(b) Growth in sample sizes over time

Figure 2: Number of firms in each capitalization quartile over time, prior to any trimming or Winsorization. The left chart shows the actual sample sizes for each capitalization class, while the right chart shows the year-over-year percentage growth in the sample sizes starting from December 31, 1985.

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.

Figures 2a and 2b show the number of stocks in each capitalization group over time, after the screening criteria have been applied. (The spikes correspond to the reallocation that occurs each June.)

2.1.1 Trimming and Winsorization

To investigate the effects of univariate trimming and Winsorization, we also needed versions of the data sets in which the factor exposures at each time point were trimmed or Winsorized by $\alpha\%$. 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 $n\alpha$ largest and $n\alpha$ smallest observations.

- To trim by $\alpha\%$, we simply remove these observations from this time slice. (In `R` this is accomplished by setting the factor exposure for those observations to be the missing value `NA`.)
- To Winsorize by $\alpha\%$, we replace the $n\alpha$ largest values by the $n\alpha$ upper quantile of the data, and the $n\alpha$ smallest values by the $n\alpha$ lower quantile of the data.

Trimming reduces the influence of the largest and smallest values on an estimate to that of the $n\alpha$ 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).

For the present study we trimmed and Winsorized by 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.

2.2 Outlier Detection via Mahalanobis Distances

We detect potential outliers in the factor exposures using classical and robust Mahalanobis distances as described in the introduction. In the classical estimation setting, the dispersion matrix is estimated via the classical covariance estimator given by Equation (3). For a robust estimate of the dispersion matrix, we used the following estimators:

- the minimum covariance determinant (MCD) estimator (Rousseeuw, 1985; Rousseeuw and van Driessen, 1999) with the maximum breakdown point fraction (approximately 50%) of observations;
- the MCD(0.75) estimator using 75% of the observations; and
- the MCD(0.95) estimator using 95% of the observations.

In an earlier version of this experiment (Martin et al., 2010) we also examined the pairwise OGK estimator (Gnanadesikan and Kettenring, 1972; Devlin et al., 1981; Maronna and Zamar, 2002), with one step reweighting ($\beta = 0.9$); and an S-estimator with a bisquare $\rho-$function (Maronna et al., 2006). However, for reasons explained in Green and Martin (2015), we have omitted those estimators from this revised experiment.[12]

Among MCD estimators, the maximum breakdown point MCD 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.[13] Using 75% of the observations 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(0.95) 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).

As discussed in Cerioli et al. (2009), Cerioli (2010), and Green and Martin (2014b), the chi-squared distribution provides inaccurate outlier detection rates for small to moderate sample sizes. Furthermore, outlier detection using Mahalanobis distances (even if based on the MCD) and the chi-square distribution is not immune to the problem of masking, wherein an extreme outlier prevents the detection of a moderate outlier (Barnett and Lewis, 1994). We thus employ the iterated reweighted MCD (IRMCD) method of Cerioli (2010) for detection of outliers, with the extensions discussed in Green and Martin (2014b) to maintain accuracy when using more than the maximum breakdown point fraction of the observations. This method is implemented in the `R` package `CerioliOutlierDetection` (Green and Martin, 2014a).

At each time point, we estimate the dispersion matrix of the factor exposures using the classical and robust methods. We perform this estimation on the raw data, the trimmed data, and the Winsorized data. We flag outliers by using the Cerioli method with a nominal false detection rate of 2.5%. In order to deal with multiple comparison issues, we used a Bonferroni adjustment to nominal rate: the Bonferroni-corrected significance level is the $2.5\%/n$ quantile, where $n = 325$ is the number of time points.[14] Thus we effectively use the upper 0.008% percentile in the Cerioli tests for outlyingness; at this level of significance we would expect to see 8 observations in a sample of size $100,000$ (i.e., the number of stocks multipled by the number of time points) flagged as outliers purely by chance.

[12]Briefly, the authors show that testing for outlyingness by comparing Mahalanobis distances calculated with these robust dispersion estimates to chi-squared quantiles suffers from the same problems found in Cerioli et al. (2009) for MCD-based Mahalanobis distances. See Green and Martin (2015) for more details.

[13]The (finite sample) breakdown point of an estimator is, roughly speaking, the smallest fraction of contaminated observations that can cause the estimator to assume values that are arbitrarily far from the parameter estimate that would have been obtained with the uncontaminated data. For instance, the sample mean of n observations can be made arbitrarily large by changing only one observation; it therefore has a breakdown point of $1/n$, which asymptotically tends to 0. The sample median, however, will still give the correct answer even if half the observations are replaced with infinity; it therefore has breakdown point of $1/2$.

The relative efficiency of an estimator $\tilde{\theta}$ of a parameter θ is a way of quantifying how accurately $\tilde{\theta}$ estimates θ relative to some "ideal" estimator $\hat{\theta}$ (e.g., the maximum likelihood estimator). The relative efficiency can be interpreted as the sample size $\tilde{n}$ needed to achieve the same level of accuracy (as measured by the sampling variance of $\tilde{\theta}$) as the $\hat{\theta}$ estimator with sample size $\hat{n}$.

[14]Without the correction the number of observations flagged as outliers would be higher, but some of the flagged observations would be false positives, i.e., flagged purely by chance and not because they were actual outliers.

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.

3 Results

Figures 3a–3c, 10a–10c, and 14a–14c show the number of observations that were flagged as outliers using the calibrated Mahalanobis distance criterion, for each covariance estimate (classical, MCD, MCD(0.75), MCD(0.95)) and cleaning method (none, 2.5% trim, or 2.5% Winsorization).

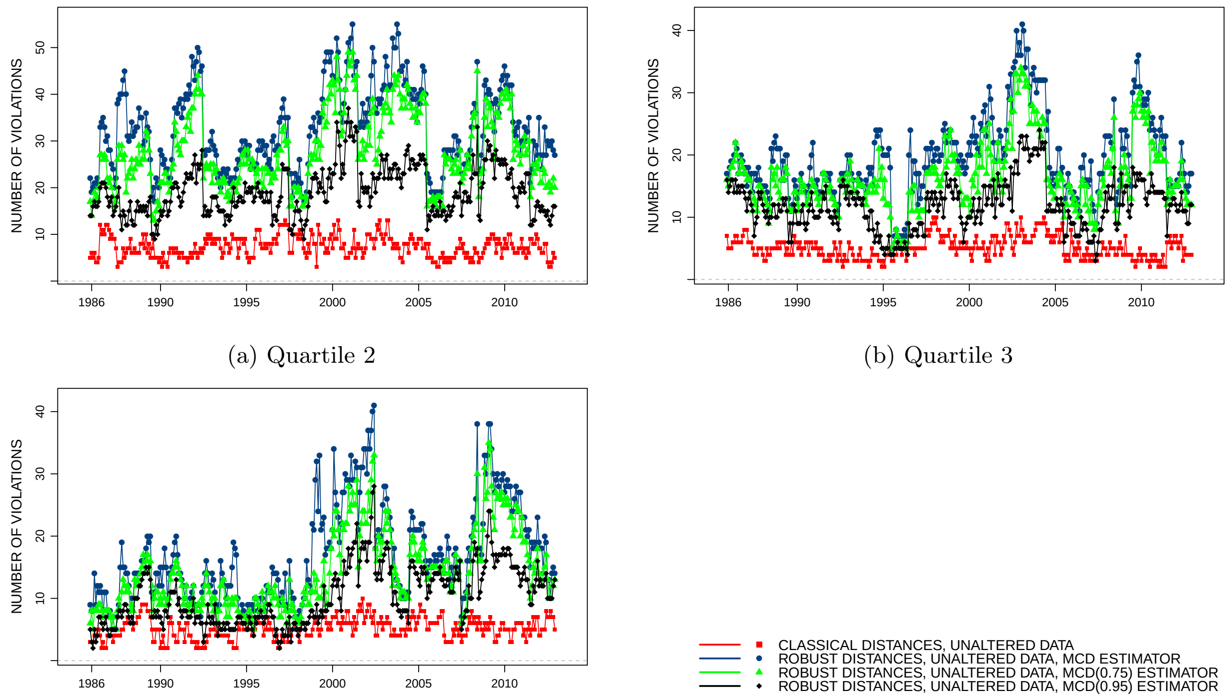
3.1 Unaltered Data

The results for the unaltered data sets (Figures 3a–3c) show that the robust distance methods generally find more outliers than the classical distance method (red line). In each of the data sets, the classical method often flagged far fewer observations as outliers than the robust methods. Table 1 shows the maximum number of outliers detected by each method for each data set.[15]

Furthermore, the number of outliers found 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: for instance, the robust estimates from Figures 3a-3c 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 3c), 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 2 provides the actual numbers depicted in Figure 3c for during 2000–2004. The number of outliers detected by all methods is elevated for this period; starting in 2000, however, the robust methods consistently flag more observations as outlying than the classical method. Even the MCD(0.95) method, which is very conservative compared to the higher-breakdown point MCD(0.75) and MCD methods, flags about twice as many points as the classical method for most months during the 2000–2004.

Whereas in the larger stocks we do not see large differences in the number of outliers detected prior to 1999 (with a few exceptions around the end of the 1980s), the detection rates in the smaller stocks (Quartiles 2 and 3) are consistently higher for the robust methods. In Figure 3b the MCD(0.95) 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 3a), there is a large spike in the MCD(0.95) detection rate immediately following the dot-com era, then again near the end of 2005. Table 3 shows the actual numbers depicted in Figure 3a during late 1999 through late 2001. Again we see that the robust methods are flagging far more points as potential outliers; even the conversative MCD(0.95) tends to flag at least twice as many outliers as the classical approach. Furthermore, this growth cannot be explained solely by an increase in the number of stocks in the sample over time—the peak in Figure 3a during 1999–2001, for instance, persists if we look at the number of violations as a percentage of the number

[15] Somewhat counterintuitively, we note that for the MCD-based methods, the number of outliers detected is not a strictly decreasing function of α, the percentage of observations used; that is, even though MCD(0.75) uses a larger fraction of the data than MCD, it may not be the case that MCD(0.75) flags a fewer number of outliers. The "best" subset of the data in the MCD(0.75) may not contain the MCD subset. Since the theoretical MCD(α) procedure involves combinatoral optimization, and the `fastMCD` implementation uses subsampling to approximate the solution, it is not clear whether this phenomenon is an artifact of the implementation or the underlying geometry of the data. We have not seen mention of the phenomenon in the literature to date.



(a) Quartile 2

(b) Quartile 3

(c) Quartile 4

Figure 3: Number of observations flagged as outliers over time, stratified by dispersion estimate.

of observations on a monthly basis. 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 4 shows pairwise scatterplots of the Quartile 4 data for June 2002, in which every pair of variables is plotted together with outliers marked using our four detection methods. (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.) This type of plot is very useful for understanding the geometry of the outliers in the data. 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 methods, though, are 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 robust 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 all three robust methods detected a large number of outliers (see Figure 3a) but the classical method did not. Figure 5 shows the four pairwise scatterplots at this time for the Quartile 2 group data. The classical method only detects a handful of extreme, one-dimensional outliers. The robust methods detect far more outliers; the nature of the outliers, however, is not evident in the two-dimensional scatterplots. Figure 6 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 squares, those detected only by MCD(0.95) 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

Table 1: Maximum number of outliers detected (MAX) and corresponding dates for each capitalization group and detection method. Unaltered data.

Quartile	Classical		MCD		MCD(0.75)		MCD(0.95)	
	MAX	Date(s)	MAX	Date(s)	MAX	Date(s)	MAX	Date(s)
Quartile 2	13	Apr 1997 May 2000 Oct 2002 Mar 2003 Apr 2003	55	Mar 2001 Oct 2003	49	Dec 2000 Mar 2001	37	Dec 2000
Quartile 3	10	Dec 1997 Feb 1998 Jan 2003 May 2004	41	Feb 2003	34	Jan 2003	24	Feb 2004
Quartile 4	10	Oct 2001	41	Jun 2002	35	Feb 2009	28	Jun 2002

flagging the most “extreme” of the observations and is missing the moderate outliers (which are masked by the “extreme” ones).

Would any of these outliers have been mitigated by trimming each variable individually? Figure 7 shows the June 2002 Quartile 4 group data again, this time with 2.5% trimming bounds shown (black dashed lines). Figure 8 shows a similar plot for the Quartile 2 group data at December 2000. Points outside of the box formed by the trimming bounds would have been discarded prior to factor model construction. All of the points flagged by the classical method (green squares) would have been mitigated by trimming; this would not have been the case for the MCD(0.95) outliers (red asterisks). Certainly trimming by a larger fraction would have caught some of these outliers, but that would have also thrown away more of the “good” data.

Finally, it is certainly not the case that the robust methods will always detect far more outliers in the data than the classical method. If the outliers are largely extreme, isolated, observations that do not mask many moderate outliers, then we can see fairly similar results from the classical and robust approaches. Figure 9 shows an example of this phenomenon at July 2007: neither the classical method nor the MCD methods flag many points as outliers due to the configuration of the observations in 4-dimensional space. The robust approach picks up a few moderate outliers that are missed by the classical approach due to masking, but largely the two approaches agree that there are few outlying observations in the data.

Table 2: Number of outliers detected in untransformed Quartile 4 data set during 2000–2004 using classical method (C), MCD method, MCD(0.75) method, and MCD(0.95) method.

Date	C	MCD	MCD(0.75)	MCD(0.95)	Date	C	MCD	MCD(0.75)	MCD(0.95)
Jul 2000	5	22	14	8	Jul 2002	4	18	18	14
Aug 2000	7	27	22	12	Aug 2002	4	20	16	13
Sep 2000	6	27	21	16	Sep 2002	6	21	20	14
Oct 2000	6	30	22	14	Oct 2002	6	20	19	16
Nov 2000	8	29	22	14	Nov 2002	7	20	17	13
Dec 2000	8	29	28	18	Dec 2002	5	25	19	16
Jan 2001	6	28	24	15	Jan 2003	6	28	21	17
Feb 2001	6	33	25	17	Feb 2003	6	28	24	18
Mar 2001	4	29	27	18	Mar 2003	7	26	22	19
Apr 2001	7	29	25	19	Apr 2003	6	24	24	18
May 2001	5	32	25	19	May 2003	6	23	21	14
Jun 2001	6	31	29	22	Jun 2003	4	21	19	15
Jul 2001	8	21	15	12	Jul 2003	4	15	12	9
Aug 2001	9	27	20	16	Aug 2003	5	18	15	11
Sep 2001	8	31	22	18	Sep 2003	4	13	14	11
Oct 2001	10	31	24	19	Oct 2003	4	13	13	9
Nov 2001	6	34	24	18	Nov 2003	4	12	12	8
Dec 2001	7	34	22	19	Dec 2003	4	12	10	8
Jan 2002	8	30	22	16	Jan 2004	5	10	11	7
Feb 2002	9	37	27	19	Feb 2004	5	11	11	8
Mar 2002	9	34	24	19	Mar 2004	5	10	11	8
Apr 2002	7	37	29	23	Apr 2004	4	10	10	8
May 2002	6	40	32	27	May 2004	7	11	10	8
Jun 2002	8	41	33	28	Jun 2004	5	11	10	6

Table 3: Number of outliers detected in untransformed Quartile 2 data set during 1999–2001 using classical method (C), MCD method, MCD(0.75) method, and MCD(0.95) method.

Date	C	MCD	MCD(0.75)	MCD(0.95)	Date	C	MCD	MCD(0.75)	MCD(0.95)
Jan 1999	8	32	30	23	Jul 2000	7	36	31	22
Feb 1999	3	30	26	22	Aug 2000	7	34	32	25
Mar 1999	12	36	30	19	Sep 2000	8	38	36	23
Apr 1999	10	34	26	21	Oct 2000	5	41	41	30
May 1999	8	35	29	20	Nov 2000	5	47	44	34
Jun 1999	8	39	36	24	Dec 2000	5	51	49	37
Jul 1999	9	45	38	19	Jan 2001	6	49	46	31
Aug 1999	6	47	33	17	Feb 2001	6	52	47	34
Sep 1999	7	49	34	17	Mar 2001	7	55	49	31
Oct 1999	6	49	39	22	Apr 2001	11	44	43	30
Nov 1999	9	47	42	24	May 2001	5	46	41	33
Dec 1999	10	49	40	20	Jun 2001	7	47	44	32
Jan 2000	7	44	43	23	Jul 2001	8	33	26	17
Feb 2000	12	43	42	23	Aug 2001	7	33	27	16
Mar 2000	9	43	36	25	Sep 2001	7	34	27	23
Apr 2000	9	52	48	34	Oct 2001	5	38	30	20
May 2000	13	49	43	32	Nov 2001	6	33	26	19
Jun 2000	11	43	41	29	Dec 2001	7	36	28	20

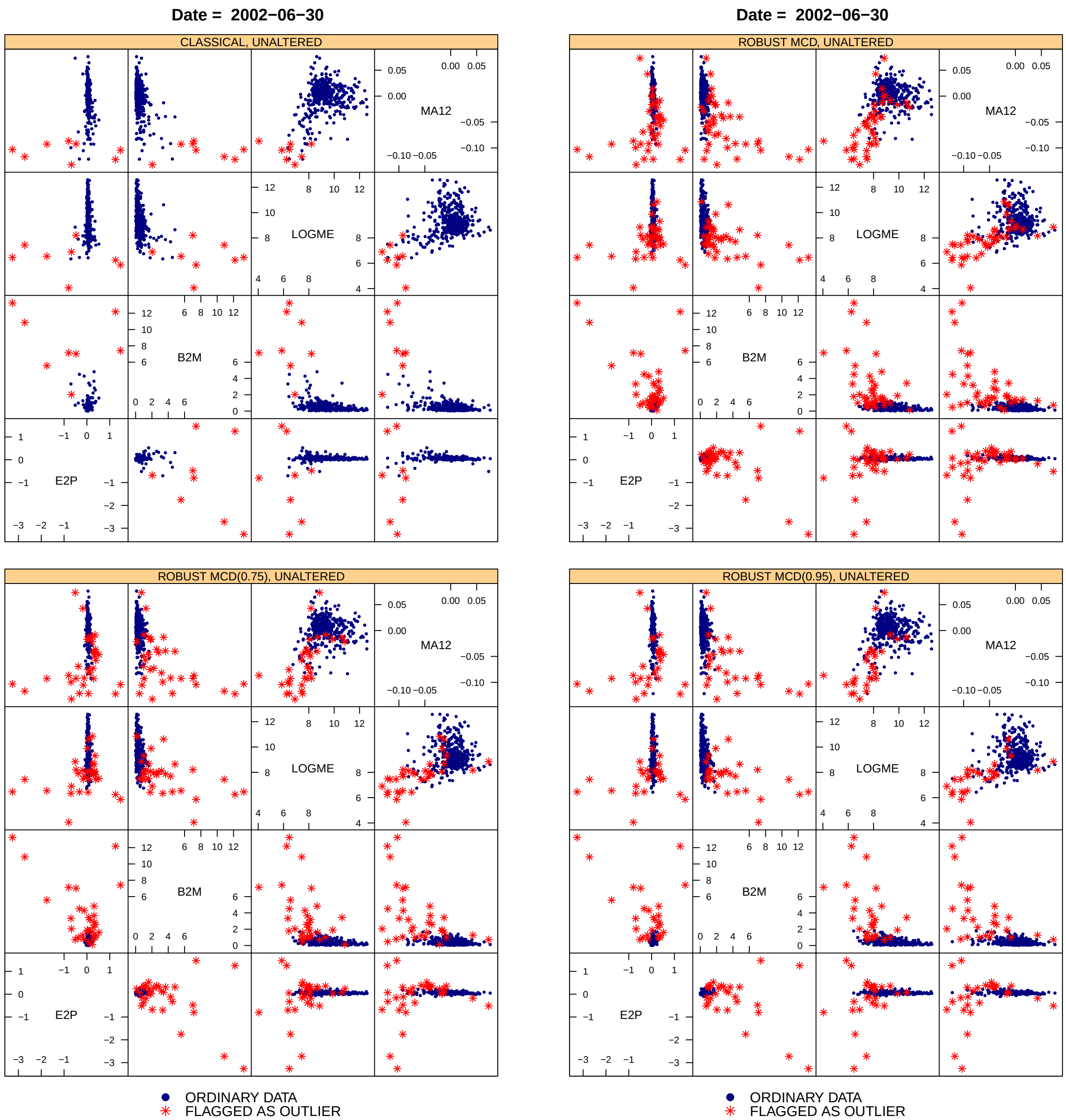


Figure 4: Pairwise scatterplots of Quartile 4 group data during June 2002. Outliers are shown as red asterisks, while regular data values are shown as blue dots. Clockwise from upper left: classical estimate, MCD estimate, MCD(0.95) estimate, MCD(0.75) estimate. Factors shown are earnings-to-price (E2P), book-to-market (B2M), size (LOGME), and momentum (MA12).

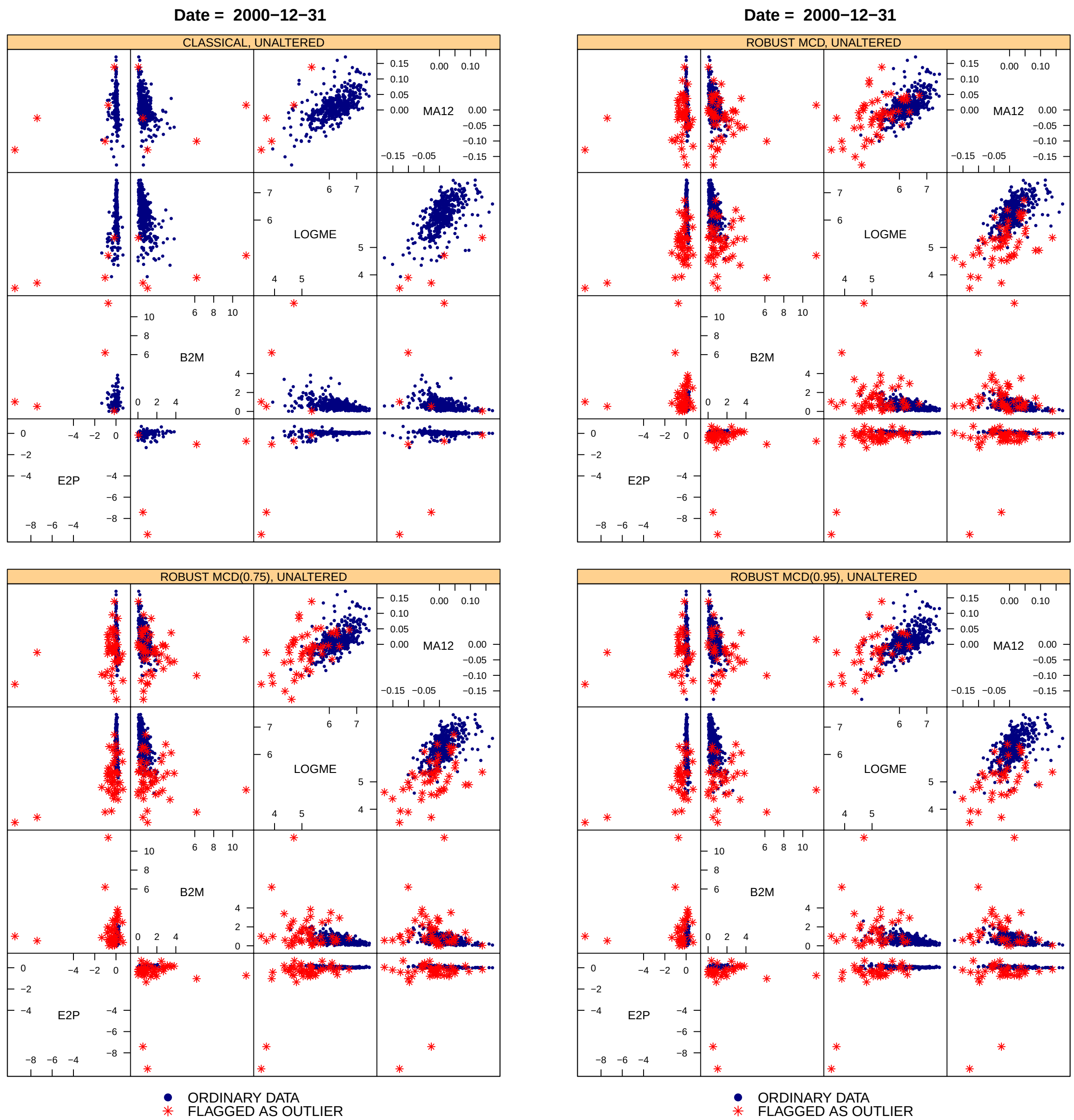


Figure 5: Pairwise scatterplots of Quartile 2 group data during December 2000. Outliers are shown as red asterisks, while regular data values are shown as blue dots. Clockwise from upper left: classical estimate, MCD estimate, MCD(0.95) estimate, MCD(0.75) estimate.

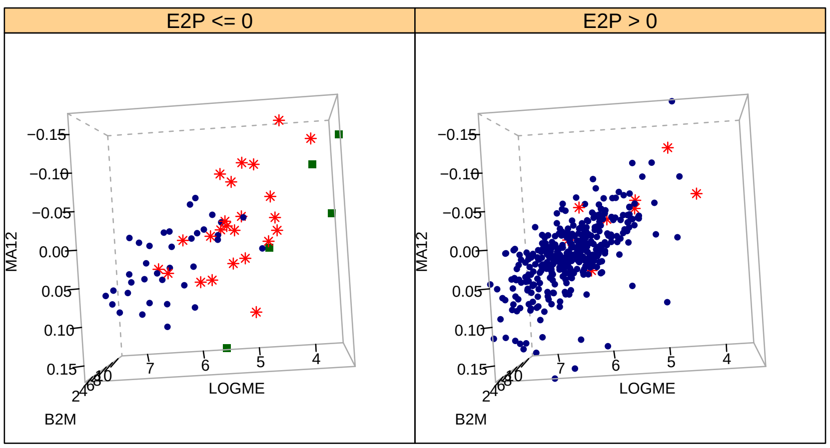


Figure 6: Three-dimensional scatterplots of Quartile 2 data during December 2000, stratified by the sign of the earnings-to-price (E2P) variable, with outliers detected by the classical and MCD(0.95). The variables shown in the plot are BOOK2MARKET (B2M), LOG.MARKET.EQUITY (LOGME), and MA12. The green squares are points flagged as outliers by both methods, while the red asterisks are only flagged by the MCD(0.95) 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.

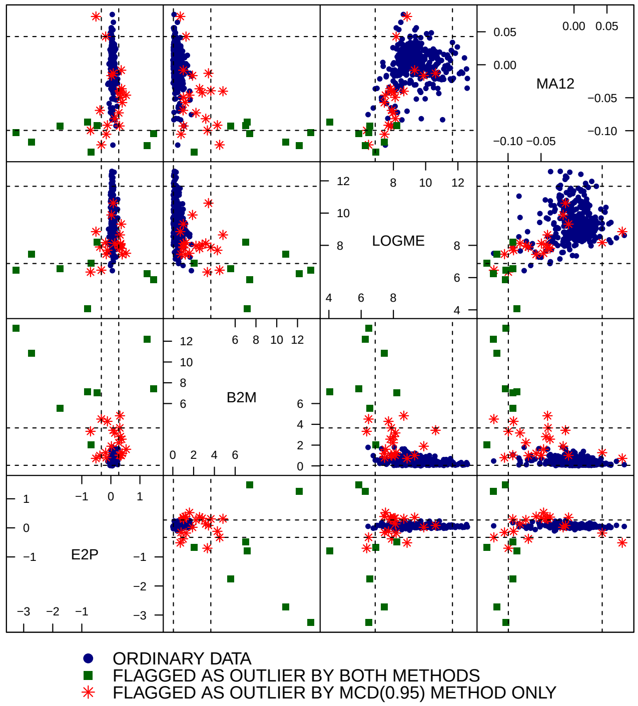


Figure 7: Pairwise scatterplot of Quartile 4 group data during June 2002 with outliers detected by MCD(0.95) method and classical method marked as in Figure 6. The black dashed lines correspond to the cutoff values that would have been used had each variable been trimmed by 2.5%.

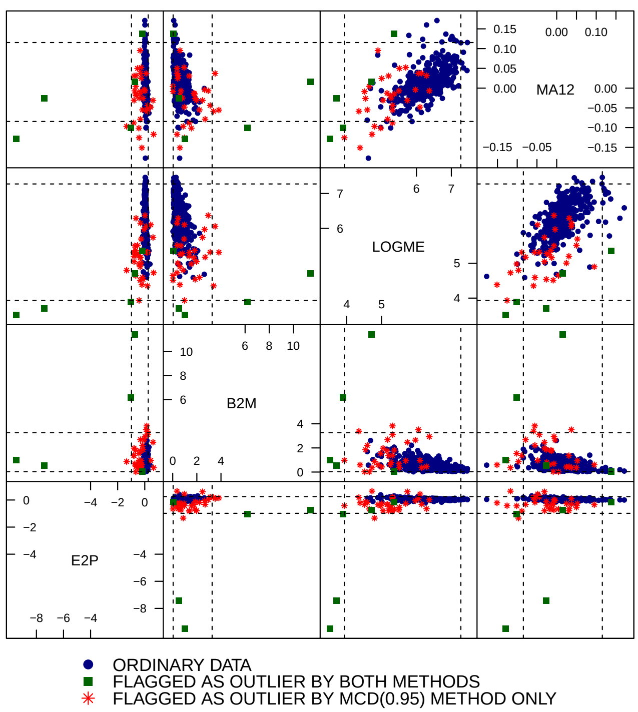


Figure 8: Pairwise scatterplot of Quartile 2 group data during December 2000 with outliers detected by MCD(0.95) method and classical method marked as in Figure 6.

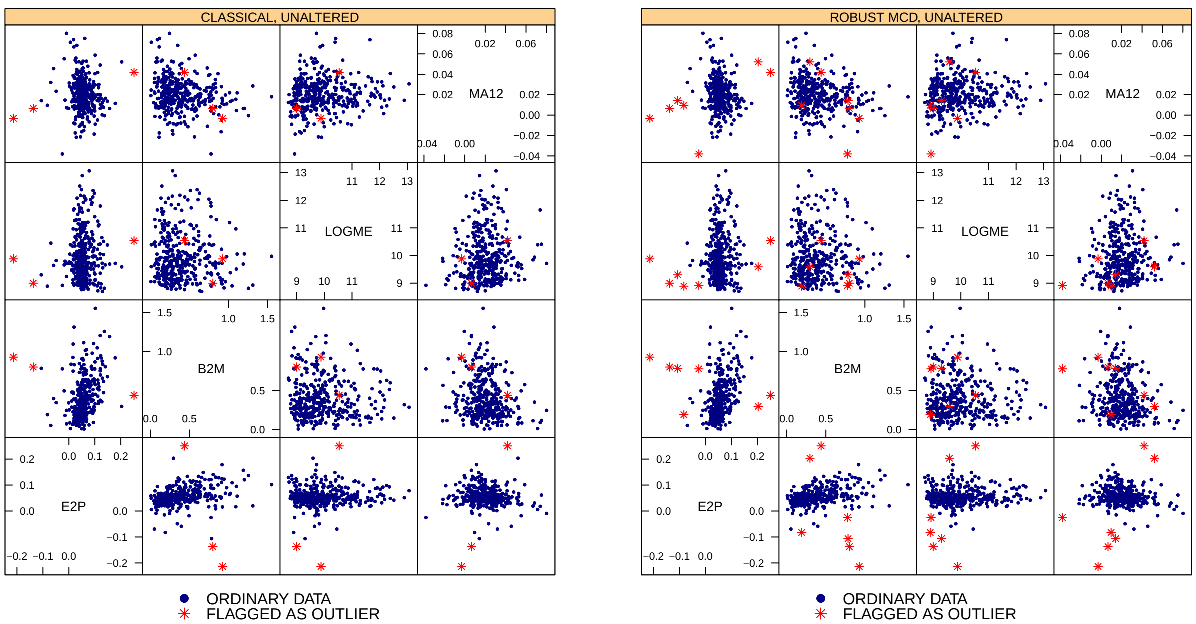
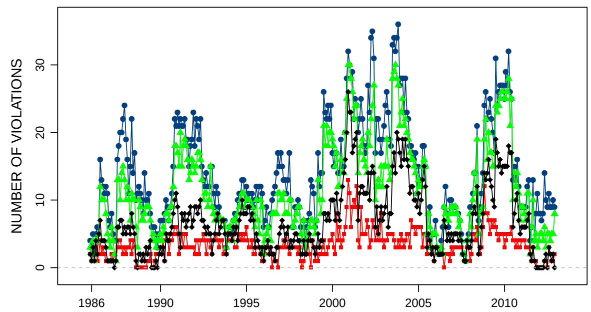
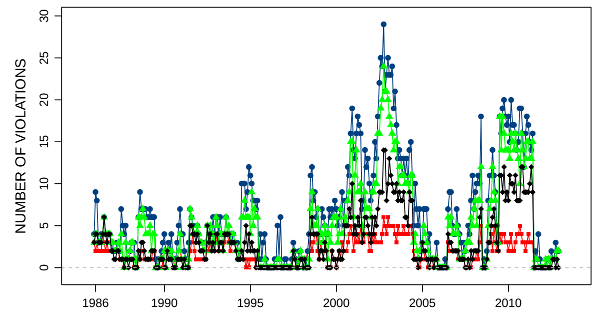


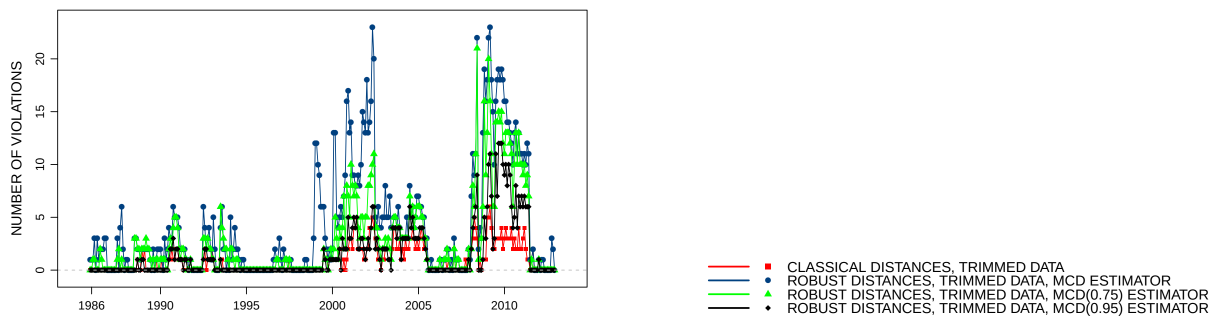
Figure 9: Pairwise scatterplots of Quartile 4 group data during July 2007 with outliers marked as in Figure 4. Results from using the classical Mahalanobis distances are shown in the left panel, while results from the MCD-based distances are shown on the right. The results for the MCD(0.75) and MCD(0.95) estimates are similar to the MCD result, and are omitted for reasons of space.



(a) Quartile 2



(b) Quartile 3



(c) Quartile 4

Figure 10: Number of observations flagged as outliers over time, stratified by dispersion estimate. Trimmed data.

3.2 Trimmed Data

Figures 10a–10c show the results of the outlier detection procedures on the trimmed data. 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 4 shows the maximum number of outliers detected by each method. 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. (More on this later.)

October 2002 is an interesting time for the Quartile 3 group after one-dimensional trimming; all of the detection approaches identify an unusual number of outliers at that time. Figure 11 shows pairwise scatterplots for the Quartile 3 group data at October 2002 with outliers marked; the box-like structure of the data after one-dimensional trimming is apparent.

December 2000 is again an interesting time for Quartile 2 group data. Figure 12 shows pairwise scatterplots for that group at this time. A cursory comparison of these scatterplots with those from the analysis on the unaltered data (Figure 5), 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 13 (top panel) shows the pairwise scatterplot for the untrimmed data with MCD(0.95) outliers marked. The green squares correspond to outlying observations that would have been removed by one-dimensional trimming, while the red asterisks are outlying observations (from the untrimmed data) that would remain

Table 4: Maximum number of outliers detected (MAX) and corresponding dates for each capitalization group and detection method. Trimmed data.

Cap. Class	Classical		MCD		MCD(0.75)		MCD(0.95)	
	MAX	Date(s)	MAX	Date(s)	MAX	Date(s)	MAX	Date(s)
Quartile 2	13	Dec 2000	36	Nov 2003	30	Dec 2000 Jan 2001 Sep 2003	26	Dec 2000
Quartile 3	6	Oct 2002 Jan 2003	29	Oct 2002	24	Oct 2002	14	Oct 2002 Nov 2002
Quartile 4	6	Mar 2009	23	May 2002 Mar 2009	21	Jun 2008	12	Sep 2009 Oct 2009 Nov 2009

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 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 6) 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.

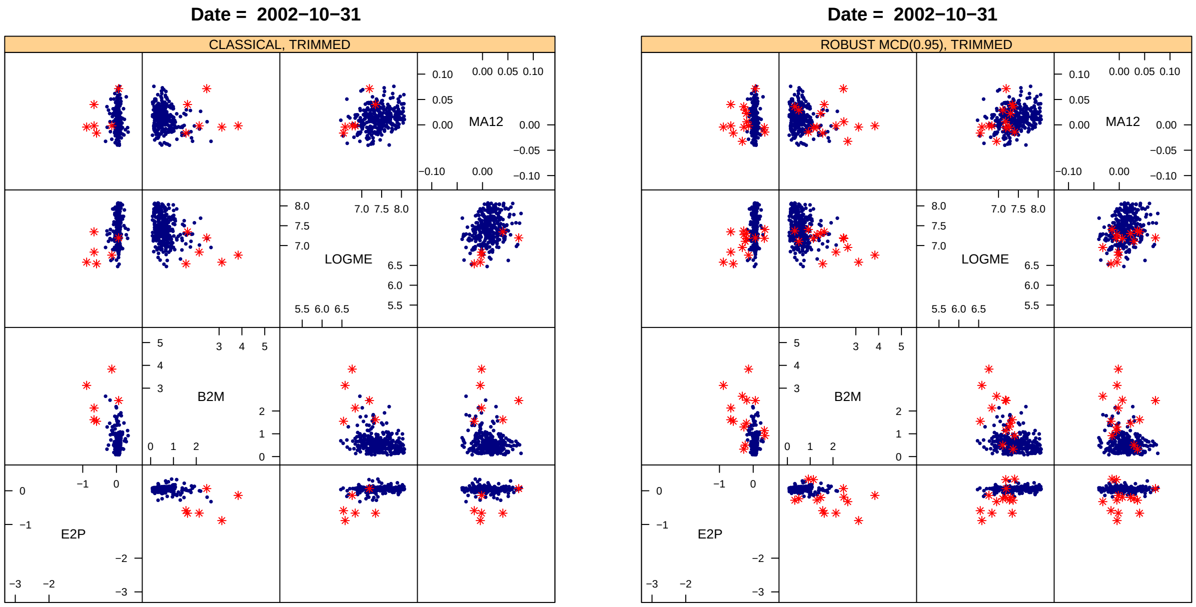


Figure 11: Pairwise scatterplots of Quartile 3 group data during October 2002 after univariate trimming. Outliers are shown as red asterisks, while regular data values are shown as blue dots. The classical estimate is shown on the left, while the MCD(0.95) estimate is shown on the right. The results for the MCD estimate and MCD(0.75) estimate are similar to that of the MCD(0.95) estimate, and are omitted to save space.

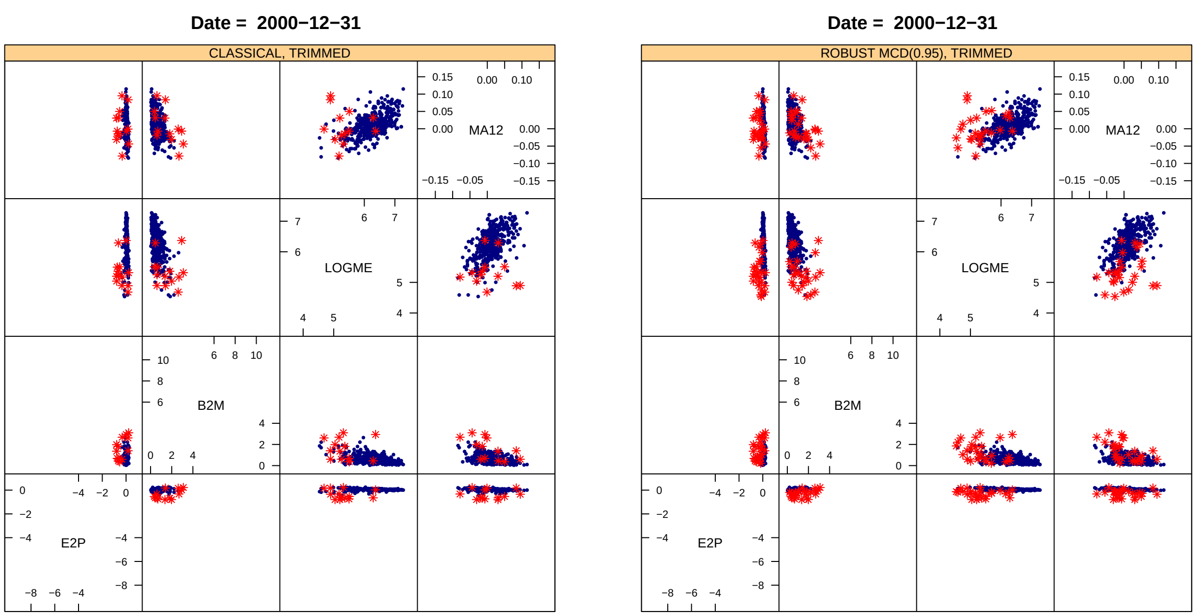


Figure 12: Pairwise scatterplots of Quartile 2 group data during December 2000 after univariate trimming. The plot layout is identical to that of Figure 11. Once again, the results for the MCD estimate and MCD(0.75) estimate are similar to that of the MCD(0.95) estimate, and are omitted to save space.

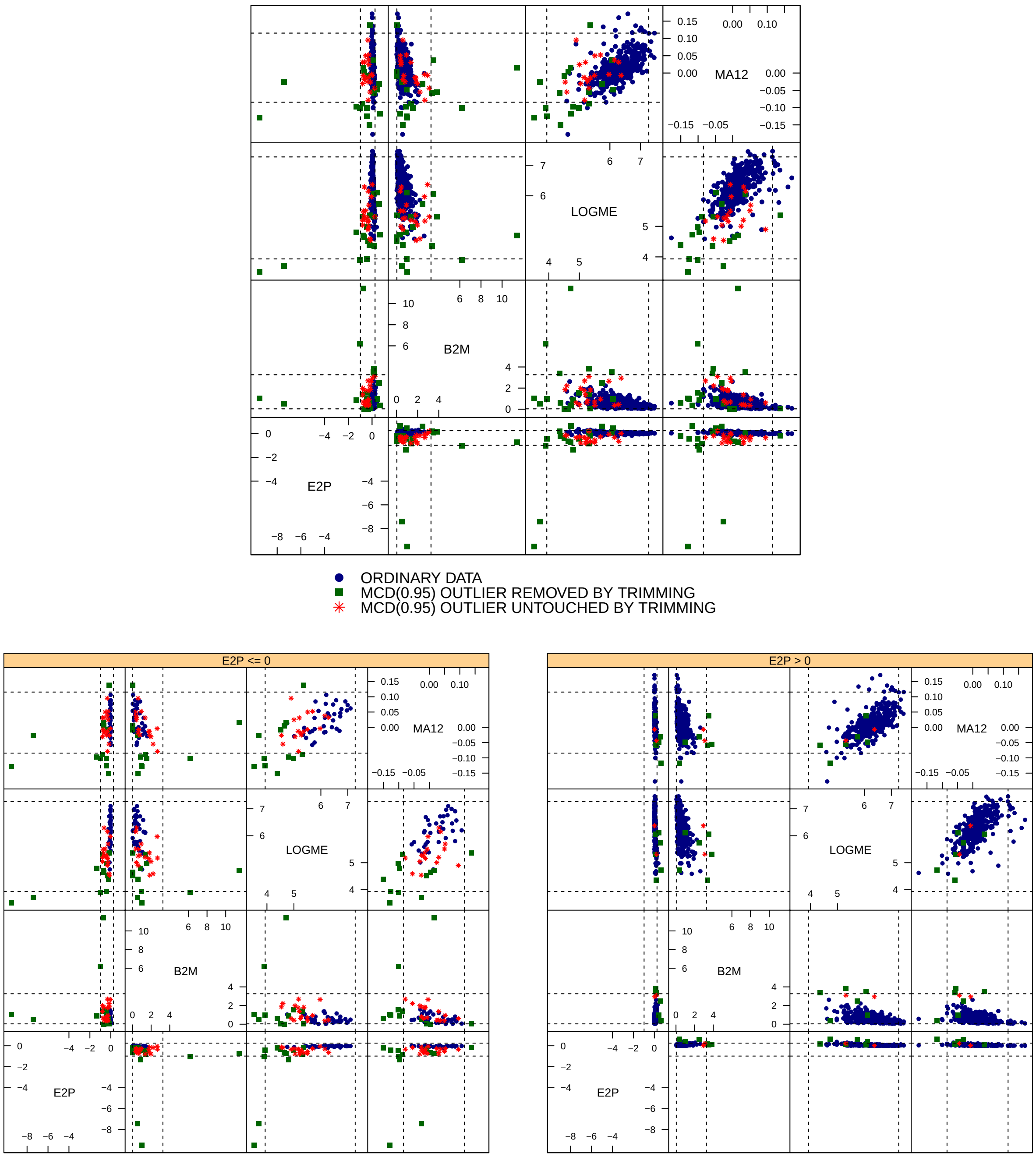
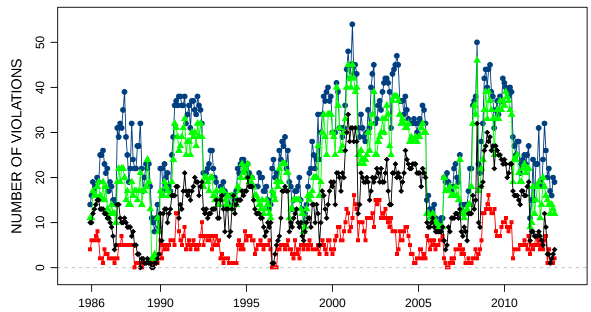
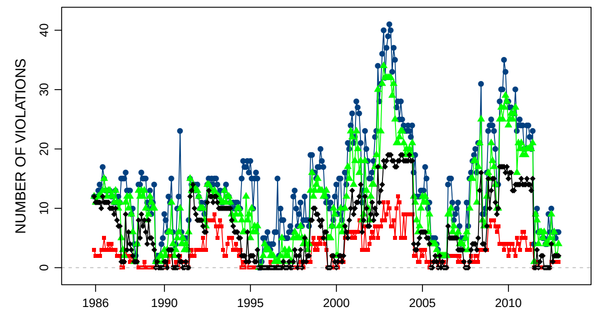


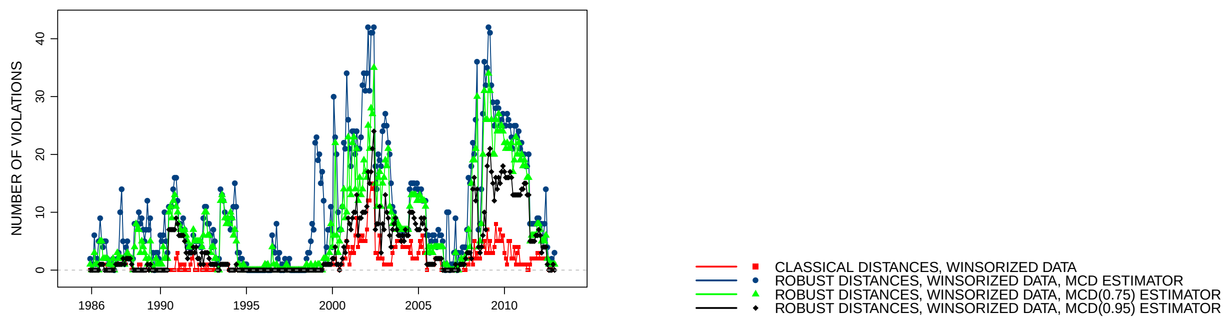
Figure 13: (Top) Pairwise scatterplot of the unaltered Quartile 2 group data during December 2000 with outliers detected by MCD(0.95) method. Points marked with green squares are outliers that would have been mitigated by the trimming procedure. Red asterisks are points that are still flagged as outliers by MCD(0.95) 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 negative earnings yield are shown in the right panel.



(a) Quartile 2



(b) Quartile 3



(c) Quartile 4

Figure 14: Number of observations flagged as outliers over time, stratified by dispersion estimate. Winsorized data.

3.3 Winsorized Data

Figures 14a–14c show the results of the outlier detection procedures on the Winsorized data. Table 5 shows the maximum number of outliers detected by each method. Winsorization did not produce as a significant of 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.

Figure 15 shows pairwise scatterplots of the Quartile 4 group data for June 2002 after Winsorization, while Figure 16 shows the pairwise scatterplots for the Quartile 3 data at October 2002. Just as in the trimmed data case, Winsorization changes the structure of the data: there is now a clustering of points along the boundaries of the 2.5% Winsorization.

Figure 17 (upper panel) shows the outliers detected by the classical method before and after Winsorization in the Quartile 2 group data for April 2001. In the three-dimensional scatterplot (lower panel) we plot the data for three of the variables (earnings-to-price, momentum, book-to-market) before and after Winsorization. This illustration makes it clear that the Winsorization creates clusters of observations along the Winsorization boundaries. It is possible that after Winsorization many of the points on the boundary are flagged as outliers because they violate the underlying multivariate normality assumption.

Table 5: Maximum number of outliers detected (MAX) and corresponding dates for each capitalization group and detection method. Winsorized data.

Cap. Class	Classical		MCD		MCD(0.75)		MCD(0.95)	
	MAX	Date(s)	MAX	Date(s)	MAX	Date(s)	MAX	Date(s)
Quartile 2	16	Apr 2001 Feb 2009	54	Mar 2001	46	Jun 2008	34	Dec 2000
Quartile 3	12	Aug 2003	41	Feb 2003	34	Oct 2002	19	Jan 2003 Feb 2003 Mar 2003 Oct 2003 Nov 2003 Apr 2004
Quartile 4	15	Apr 2002 Jun 2002	42	Feb 2002 Jun 2002 Feb 2009	35	Jun 2002	24	Jun 2002

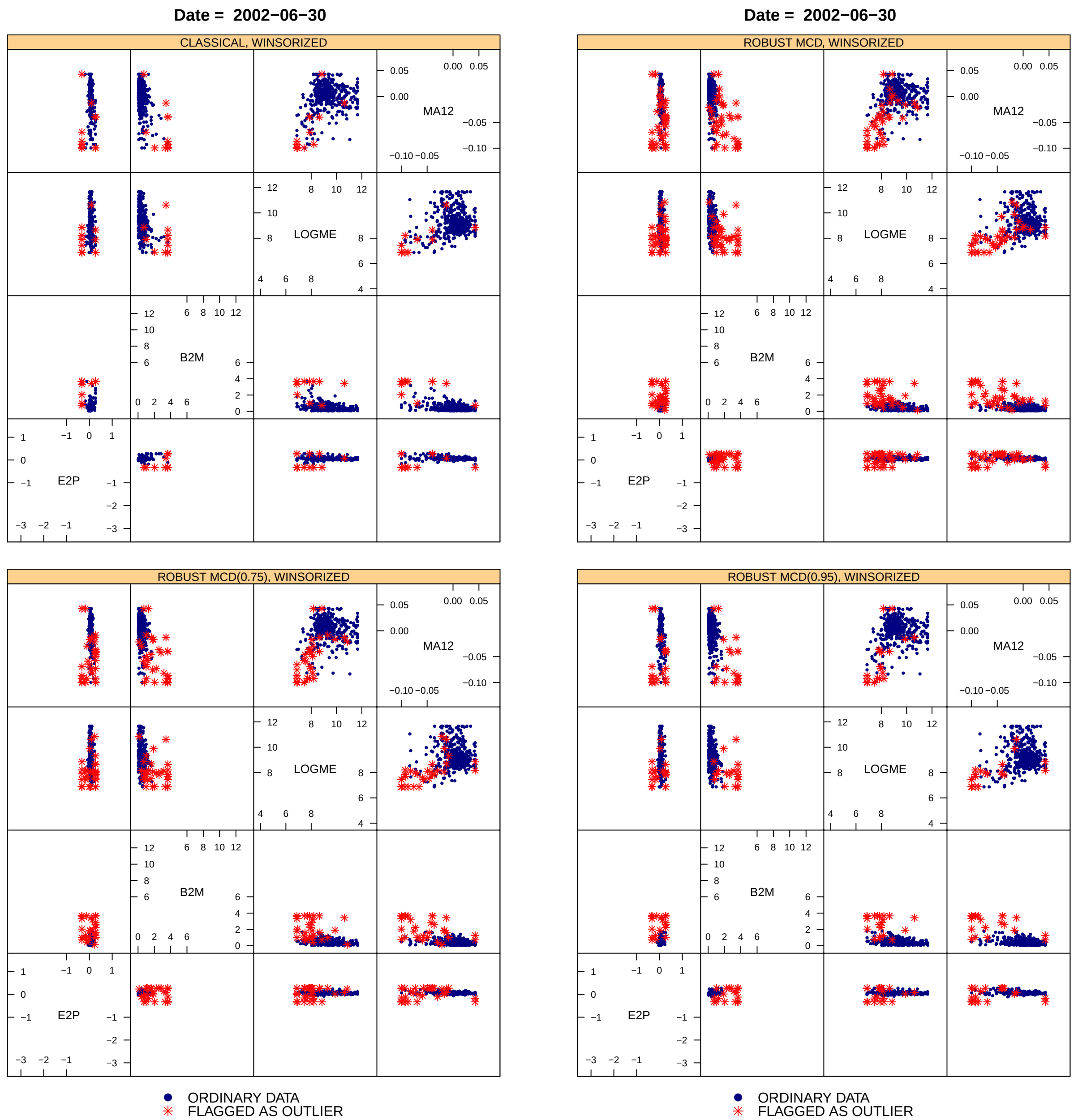


Figure 15: Pairwise scatterplots of Quartile 4 group data during June 2002 after univariate Winsorization. Outliers are shown as red asterisks, while regular data values are shown as blue dots. Clockwise from upper left: classical estimate, MCD estimate, MCD(0.95) estimate, MCD(0.75) estimate.

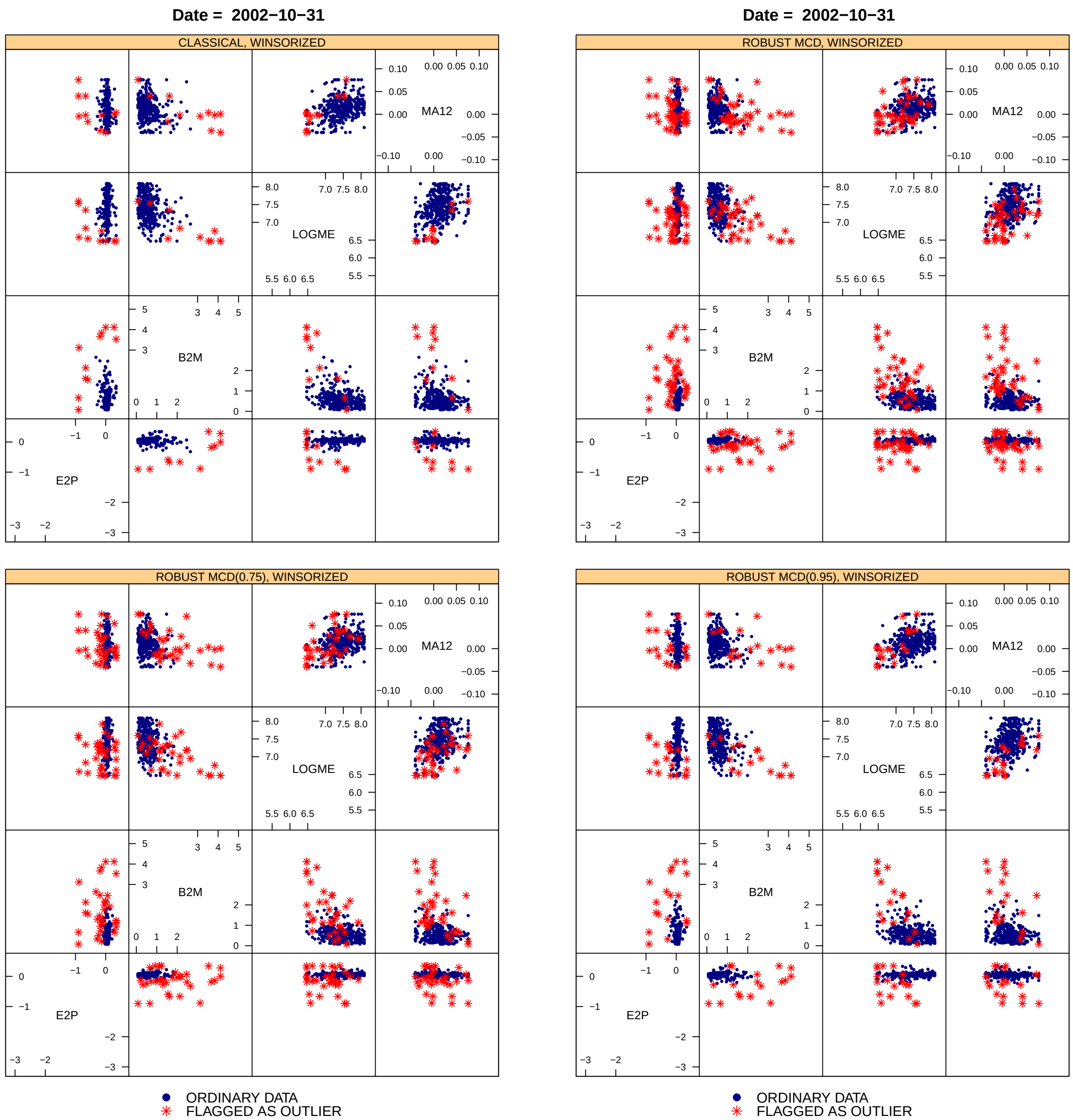


Figure 16: Pairwise scatterplots of Quartile 3 group data during October 2002 after univariate Winsorization. Outliers are shown as red asterisks, while regular data values are shown as blue dots. Clockwise from upper left: classical estimate, MCD estimate, MCD(0.95) estimate, MCD(0.75) estimate.

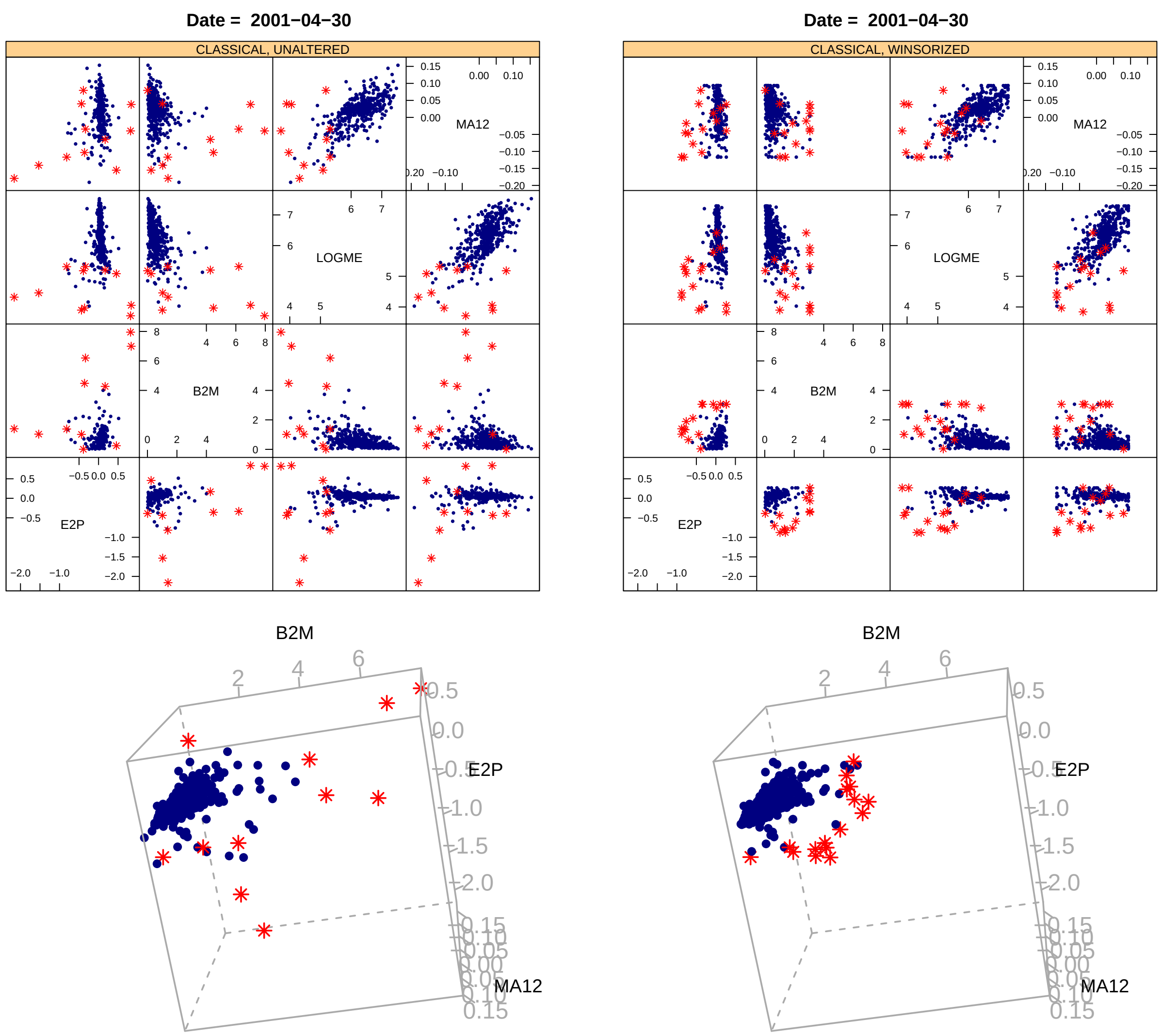


Figure 17: Pairwise and three-dimensional scatterplots of Quartile 2 group data during April 2001 before and after univariate Winsorization. Outliers are shown as red asterisks, while regular data values are shown as blue dots. Top left: classical method on unaltered data. Top right: classical method on Winsorized data. Bottom left: classical method on unaltered data; the variables BOOK2MARKET (B2M), EARN2PRICE (E2P), and MA12 are shown. Bottom right: classical method on Winsorized data.

4 Discussion

It is clear from the analyses presented above that classical covariance-based Mahalanobis distances, combined with chi-squared critical values, are a poor means of detecting multivariate outliers in fundamental factor model data. While glaringly obvious outliers may be found in this way, those outliers will hinder the detection of moderate outliers. The robust covariance approaches, combined with corrected detection thresholds provided by Cerioli (2010), do a far better job of detecting multivariate outliers in the type of factor model data presented here.

The results in the trimmed and Winsorized data cases show that multivariate outliers can still be found after one-dimensional outlier mitigation approaches. As we noted above, however, the mitigation processes change 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 be unrelated to those found in the unaltered data set. To investigate this question, we examined each observation's status as an outlier under each of the data sets (unaltered, trimmed, or winsorized) and 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] Figures 18 and 19 depict the results of this calculation for two of the interesting periods found earlier.

Figure 18 shows how many of the outliers detected in the Quartile 3 group data (during 2002–2003) persist after trimming and/or Winsorization; for instance, in October 2002, there were 10 observations that were deemed outlying in the unaltered data using the MCD(0.95) detection approach that were still considered outlying after trimming and after Winsorization. This means that trimming each variable by 2.5% did not eliminate the observations, as even larger observations dictated the trimming boundaries. (The example shown in Figure 13 illustrates how this happens.) Similarly, the Winsorization boundaries were determined by the most extreme outliers; moderate outliers were either unaffected or were projected to the "box" created by Winsorization, where they still judged as a poor fit for the multivariate normal model by the detection methodologies. Many outliers in the raw data persist after Winsorization.

Likewise, Figure 19 looks at how many of the outliers detected in the Quartile 2 group data (during 2000–2001) persist after trimming and/or Winsorizing. Winsorization again did not have the desired effect: there are significant numbers of outliers in the data (as determined by the robust approach) that remain after Winsorization.

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 approaches) in the data after trimming that were not there prior to trimming. Likewise, the non-blank cells in the "WINS. ONLY" rows correspond to outliers found only after Winsorization. This again suggests that the one-dimensional outlier mitigation approaches are changing the structure of the data. In Figure 20 we show pairwise scatterplots of the Quartile 2 data group at February 2001 (with outliers flagged using the MCD(0.95) approach), a time at which there are outliers that occur only in the trimmed and Winsorized data sets. To illustrate how such outliers arise, we code the points by the same categories used in Figures 18 and 19. The brown squares are extreme outliers in the unaltered data that would be removed by either trimming or Winsorizing, as a practitioner might expect. One-dimensional trimming changes the structure of the data, so that some observations (yellow stars) around the boundaries of the main point cloud are now considered outlying in the altered data set.

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

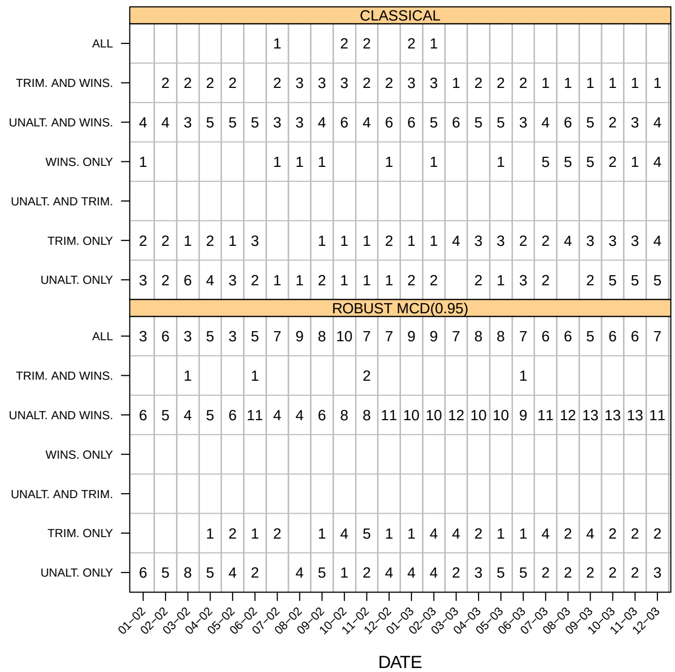


Figure 18: Status of outliers detected in the Quartile 3 group data, 2002–2003, using the classical and robust MCD(0.95) methods. From bottom to top in each panel, the cells show the number of observations in each month's data that were identified as outliers in the unaltered data only (UNALT. ONLY); 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).

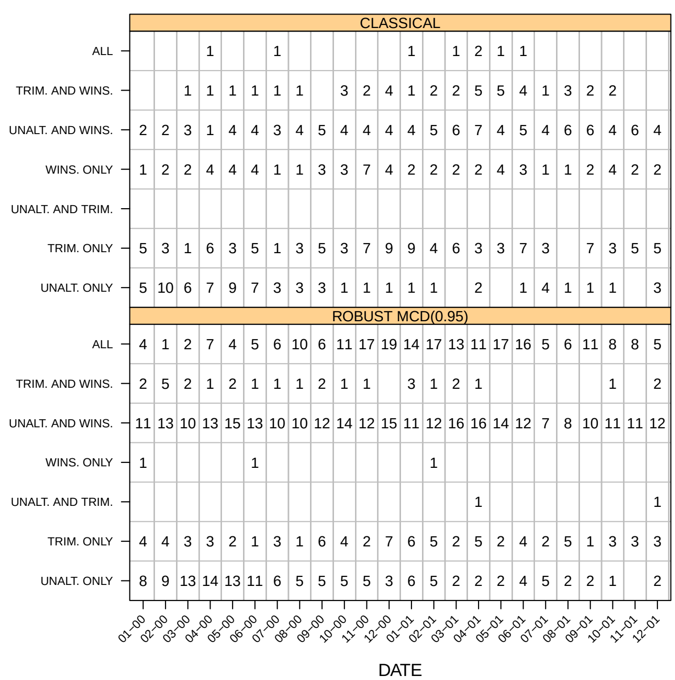


Figure 19: Status of outliers detected in the Quartile 2 group data, 2000–2001, using the classical and robust MCD(0.95) methods. The labeling is as in Figure 18.

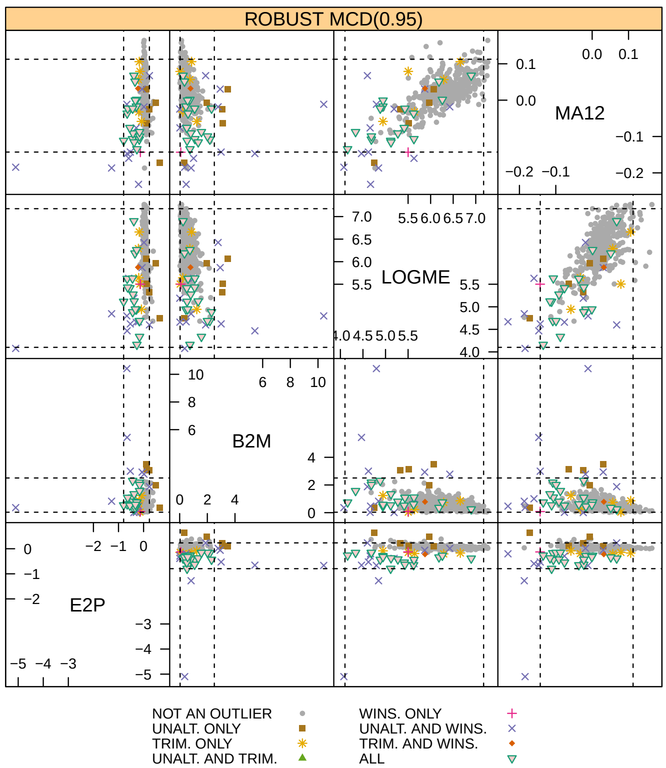


Figure 20: Status of outliers detected in the Quartile 2 group data at February 2001 using the robust MCD(0.95) method. The observations are coded by types used in Figure 18. The dashed lines depict the 2.5% trimming boundaries for each variable. The color scheme was produced using the `RColorBrewer` package (Neuwirth, 2014).

One-dimensional Winsorization also produces a few such points (pink crosses). Winsorization also fails to mitigate some outliers: the blue Xs are outliers in the raw data that are projected to the Winsorization boundaries, where they become outlying relative to the altered data set.

Another interesting difference between the classical and robust approaches is the "ALL" row of Figures 18 and 19, which represents observations that are outlying the original data set and are still outlying after trimming and Winsorization. The classical method finds very few such points, while the robust method consistently finds many of these persistent outliers. Figure 21 shows the Quartile 2 group data at December 2000 with the same coding used in Figure 20. The pink/blue inverted triangles are observations in the "ALL" category. With the classical Mahalanobis distances, moderate outliers are masked by extreme outliers, and are only detected after the extreme outliers are trimmed away (yellow stars). 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 6), 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.

Stepping beyond the issue of detecting multivariate outliers, we should ask why did the number of outliers rise so significantly during the dot-com crisis and the ensuing recession, especially in the smaller stocks. Recall that we are testing outlyingness against a multivariate normal model, so we are assuming at each month-end the bulk of the data follow that law at least approximately. Is it possible that assumption broke down during that time? If so, why?

First, the astute reader will note that our book-to-market factor takes only positive values (since we have excluded firms with negative book value). Hence, this factor's marginal distribution likely does not follow a normal law, which implies the joint distribution of the four factors is likely not multivariate normal. Suppose instead we consider the natural logarithm of the book-to-market ratio in our factor data set; this transformation will tend to make the marginal distribution of that factor more Gaussian, at the expense of the financial interpretation of the factor. We can again run the outlier detection procedure using this alternate factor data set, and compare the results to those obtained on the original factor data set.

Figure 22 shows the number of outliers detected over time in the Quartile 2 group data set using the MCD(0.95) approach on each factor data set (without any trimming or Winsorization). Although slightly more outliers were found at times using the book-to-market factor, the number of outliers detected in each factor data set is largely the same over time. Thus it seems that the lack of normality of our book-to-market ratio is neither a major driver of the peak in the number outliers in the Quartile 2 group data during 1999–2001, nor of the number of outliers detected during the entire time period.

Next we can investigate the characteristics of the firms flagged as outliers. Figures 23a-23c show the distribution of the industry sectors (as defined by the GICS standard) of firms flagged as outliers by the robust MCD(0.95) method in the unaltered data.[17] We can see that from 2001–2003 technology firms (pink) are a much larger percentage of the outliers in the Quartile 2 group data, compared to 1999–2000 and 2004–2005. 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 plays a role in 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 larger firms in the Quartile 3 and Quartile 4 groups. We also see a shift to consumer discretionary firms comprising a

[17] 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.

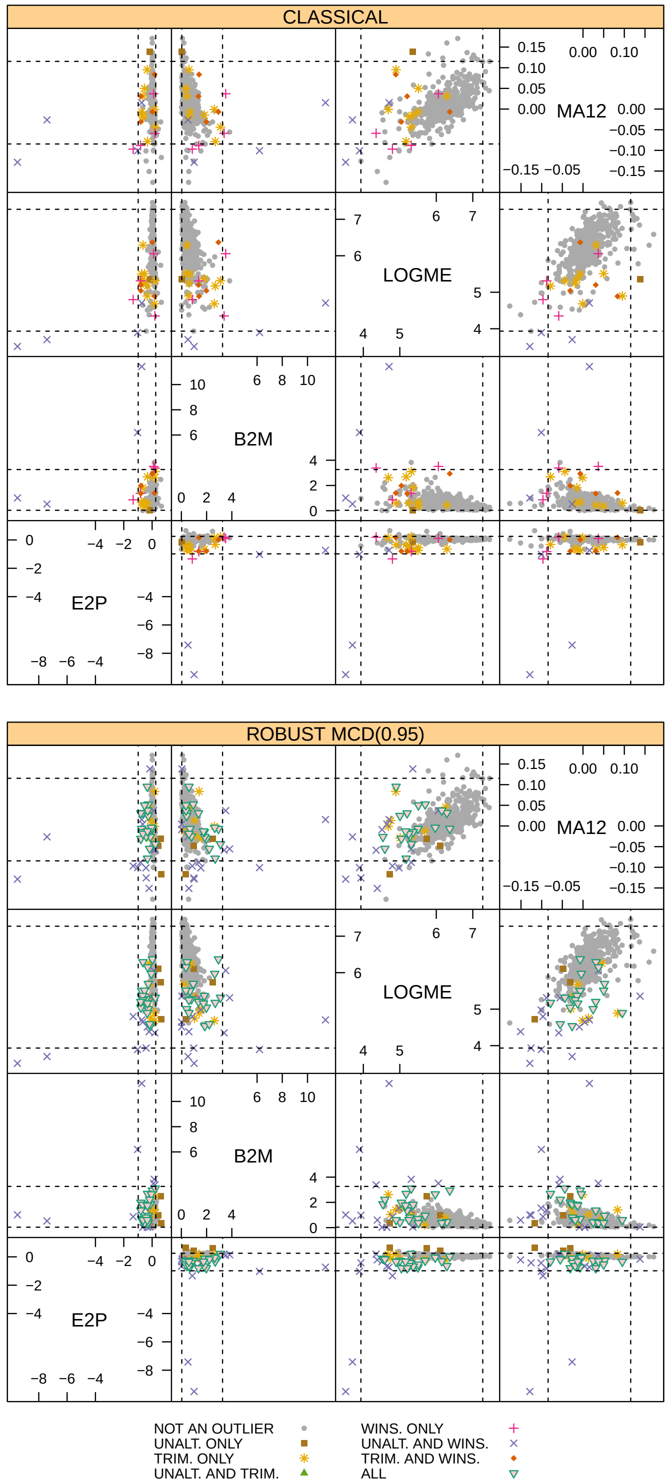


Figure 21: Status of outliers detected in the Quartile 2 group data at December 2000 using the classical and robust MCD(0.95) methods. The observations are coded by types used in Figure 18. The dashed lines depict the 2.5% trimming boundaries for each variable.

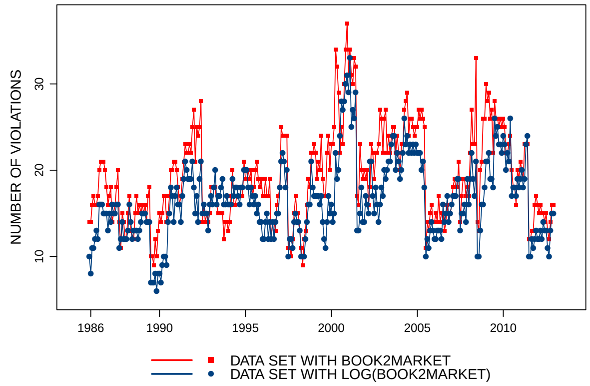


Figure 22: Comparison of the number of observations flagged as outliers over time in unaltered data using the book-to-market ratio as a factor (red squares) versus using the logarithm of the book-to-market ratio as a factor (blue circles).

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, which financial firms being a big driver of the number of outliers in the largest stock group (Quartile 4).

5 Conclusions and Further Research

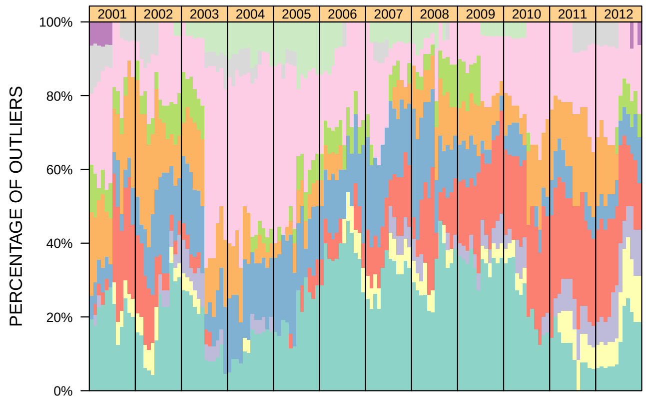
The results of this experiment provide strong evidence that one-dimensional trimming and Winsorization are insufficient to deal with many of the outliers found in the type of data used to build fundamental factor models. While the one-dimensional approaches are an appropriate means of dealing with outliers in individual variables, they fail to mitigate 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 robust Mahalanobis distance methods presented here, are a more effective means of finding outlying observations.

The experiment also demonstrated that one-dimensional Winsorization can actually make the outlier problem worse. One-dimensional Winsorization essentially projects outlying points to the boundary of a ν-dimensional box, where ν is the number of variables. The resulting Winsorized data will violate the assumption of a multivariate normal distribution, or even an multivariate elliptical distribution.

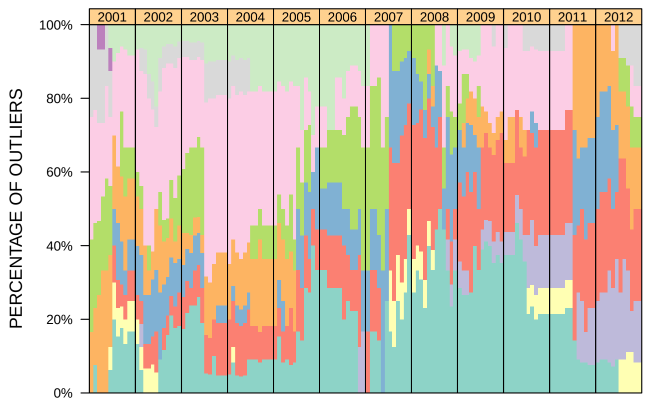
We also illustrated that using Mahalanobis distances to detect multivariate outliers can miss moderate outliers if the distances are computed using the classical mean and covariance estimates. Neither of these estimates is robust to outliers, so extreme outliers in the data will lead to masking of moderate outliers. Using a robust location and dispersion estimate, like the MCD estimate, is a superior approach to identifying both extreme and moderate outliers in fundamental factor data.

As we stated earlier, this paper re-examines some of the work done in Martin et al. (2010) with the corrected Mahalanobis distance distribution derived by Cerioli (2010). The results of this paper qualitatively agree with those of the earlier paper, even though we are now imposing a more stringent detection criterion. This further supports our view that multivariate outliers exist in the firm financial data and cannot be mitigated by the standard one-dimensional approaches to outliers.

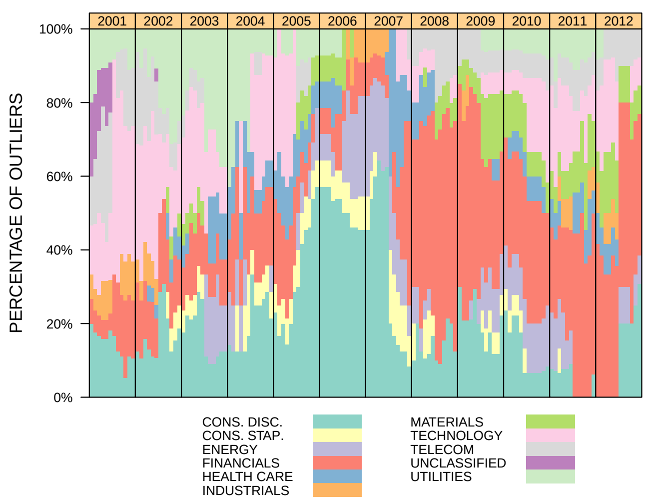
Even though we have shown that outlier mitigation via classical Mahalanobis distances; one-dimensional



(a) Quartile 2



(b) Quartile 3



(c) Quartile 4

Figure 23: Industry distribution of outliers detected in the unaltered data by the robust MCD(0.95) 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).

trimming; and one-dimensional Winsorization are not sufficient to detect or mitigate all the potential outliers in the data (under a multivariate normal model), it is not clear from this experiment what effects these methodologies, or the improvements proposed here, will have on the outputs of a factor model construction process. We will investigate this matter in a future paper.

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 factor model construction would be an ambitious undertaking for the interested researcher.

References

Diego Amaya, Peter Christoffersen, Kris Jacobs, and Aurelio Vasquez. Does realized skewness predict the cross-section of equity returns? Version February 21, 2013. Accessed August 20, 2014. Available at SSRN: `http://ssrn.com/abstract=1898735`, February 2013.

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.

Clifford S. Asness and Andrea Frazzini. The devil in HML's details. *Journal of Portfolio Management*, 39: 49–68, 2013.

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.

Heiko M. Bailer. *Robust Estimation of Factor Models in Finance.* PhD thesis, University of Washington, Department of Statistics, Seattle, WA, 2005.

V. Barnett and T. Lewis. *Outliers in Statistical Data.* John Wiley & Sons, New York, 3rd edition, 1994.

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.

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.

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.

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.

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. doi: 10.1007/s11222-008-9096-5.

Compustat. Standard & Poor's Compustat® Data, 2015. Used with permission. All rights reserved. Accessed March 2015 via Wharton Research Data Services.

D. Cook, A. Buja, J. Cabrera, and C. Hurley. Grand tour and projection pursuit. *Journal of Computational and Graphical Statistics*, 4:155–172, 1995.

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

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.

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.

Victor DeMiguel and Francisco J. Nogales. Portfolio selection with robust estimation. *Operations Research*, 57(3):560–577, May–June 2009. doi: 10.1287/opre.1080.0566.

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.

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.

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

Christopher G. Green and R. Douglas Martin. `CerioliOutlierDetection`: Outlier detection using the iterated RMCD method of Cerioli (2010), 2014a. URL `http://cran.r-project.org/web/packages/CerioliOutlierDetection/index.html`.

Christopher G. Green and R. Douglas Martin. An Extension of a Method of Hardin and Rocke, with an Application to Multivariate Outlier Detection via the IRMCD Method of Cerioli. Working paper, available from `http://students.washington.edu/cggreen/uwstat/papers/cerioli_extension.pdf`, 2014b.

Christopher G. Green and R. Douglas Martin. Untitled paper on Cerioli et al. (2009). Working paper, 2015.

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.

John Guerard, editor. *Handbook of Portfolio Construction: Contemporary Applications of Markowitz Techniques*. Springer-Verlag, London, 2010.

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.

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

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.

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.

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.

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.

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

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

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

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

R. Douglas Martin, Andrew Clark, and Christopher G. Green. Robust portfolio construction. In Guerard (2010), pages 337–382.

Erich Neuwirth. *RColorBrewer: ColorBrewer Palettes*, 2014. URL `http://CRAN.R-project.org/package=RColorBrewer`. R package version 1.1-2.

Floyd Norris. A clever move by Porsche on VW's stock. *The New York Times*, October 30, 2008. Available from `http://www.nytimes.com/2008/10/31/business/worldbusiness/31norris.html`. Accessed 9/18/2009.

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.

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.

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.

Bernd Scherer and R. Douglas Martin. *Introduction to Modern Portfolio Optimization with* NuOPT*,* S-Plus™ , and S+Bayes. Springer-Verlag, New York, 2005. ISBN 0387-210164.

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

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

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.

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/`.

B. Vaz-de Melo and Rene P. Camara. Robust modeling of multivariate financial data. Coppead Working Paper Series No. 355, 2003.

R. E. Welsch and X. Zhou. Application of robust statistics to asset allocation models. *REVSTAT—Statistical Journal*, 5(1):97–114, 2007.

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.

Kim Zetter. Six-year-old news story causes united airlines stock to plummet. *Wired.com*, September 8, 2008. Available from `http://www.wired.com/threatlevel/2008/09/six-year-old-st/`. Accessed 9/18/2009.

Eric Zivot and Jiahui Wang. *Modern Financial Time Series with* S-Plus™ . Springer-Verlag, New York, 2003.