working paper

2008

A New Tool for Robust Estimation and Identification of Unusual Data Points

Christian Garciga and Randal J. Verbrugge

FEDERAL RESERVE BANK OF CLEVELAND

ISSN: 2573-7953

Working Paper 20-08 March 2020

A New Tool for Robust Estimation and Identification of Unusual Data Points

Christian Garciga and Randal J. Verbrugge

Most consistent estimators are what Müller (2007) terms “highly fragile”: prone to total breakdown in the presence of a handful of unusual data points. This compromises inference. Robust estimation is a (seldom-used) solution, but commonly used methods have drawbacks. In this paper, building on methods that are relatively unknown in economics, we provide a new tool for robust estimates of mean and covariance, useful both for robust estimation and for detection of unusual data points. It is relatively fast and useful for large data sets. Our performance testing indicates that our baseline method performs on par with, or better than, two of the currently best available methods, and that it works well on benchmark data sets. We also demonstrate that the issues we discuss are not merely hypothetical, by re-examining a prominent economic study and demonstrating its central results are driven by a set of unusual points.

JEL codes: C3 C4 C5.

Keywords: big data, machine learning, outlier identification, fragility, robust estimation, detMCD, RMVN.

Suggested citation: Garciga, Christian, and Randal Verbrugge. 2020. “A New Tool for Robust Estimation and Identification of Unusual Data Points.” Federal Reserve Bank of Cleveland, Working Paper No. 20-08. https://doi.org/10.26509/frbc-wp-202008.

Christian Garciga is at the Federal Reserve Bank of Cleveland (christian.garciga@clev.frb.org). Randal Verbrugge is at the Federal Reserve Bank of Cleveland and NBER/CRIW (randal.verbrugge@clev.frb.org). The authors thank Mark Bognanni, Todd Clark, Rob Engle, Bruce Fallick, Peter Hinrichs, Maksim Isakin, Ed Knotek, Kurt Lunsford, Phuong Ngo, Ellis Tallman, Murat Tasci, and seminar participants at Cleveland State University and the Federal Reserve Bank of Cleveland.

INTRODUCTION

Many commonly used "consistent" estimators – that is, estimators that are consistent for a particular model – are what Müller (2007) terms "highly fragile": Small contaminations can make the estimator converge in probability to something arbitrarily divergent. For instance, the classical estimator of the covariance matrix is *zero-breakdown*: Sufficient corruption of a *single* data point – a negligible percentage of a large sample – can cause the estimator to break down, that is, to be driven arbitrarily far from the true population covariance matrix. The breakdown value is the minimum fraction of gross outliers that can completely spoil the estimate. A nonzero breakdown value is related to "qualitative robustness"; we define this term more formally below, but intuitively, an estimator is qualitatively robust if a small perturbation of the distribution of *x* – for example, the introduction of a small fraction of data drawn from a different distribution – leads to a small change in the asymptotic distribution of the estimator. Many consistent estimators are not qualitatively robust: Consider contaminating the sample with a small amount of data drawn from a standard Cauchy distribution. The mean is no longer defined! Standard techniques such as multivariate regression or principal components are based on multivariate means, covariance matrices, and least-squares fitting, all of which lack qualitative robustness, all of which can be profoundly influenced by even a few unusual data points (UDPs). Rousseeuw and Leroy (1987) provide many examples where the OLS estimator breaks down completely.[1] Such phenomena do *not* disappear asymptotically.

Often, economists ignore this issue. Sometimes researchers attempt to rectify such estimator sensitivity by removing "outliers," or UDPs, where detection of such points may be done on the basis of being extreme in one dimension of the data (for example, income in excess of $1 million), or perhaps on the basis of traditional outlier-detection techniques such as Cook's *d*. However, such methods can easily fail in multivariate data, as we discuss below; multivariate outliers can be extremely challenging to identify, particularly in large data sets.

Less commonly, analysts make use of robust regression in order to assess whether their findings are reliable. But in practice, most commonly used robust regression techniques are less robust than they appear. For instance, perhaps the most commonly used robust regression method in economics, median regression, is actually zero breakdown (!): A single outlying (*x,y*) pair can force *all* quantile regression hyperplanes to pass through it.[2]

[1] See also Zaman, Rousseeuw, and Orhan (2001) for further discussion and some small-sample examples using a least trimmed squares estimator.

[2] This is a known defect of the entire class of *M*-estimators, and this defect has motivated the search for robust quantile regression methods, such as that of Rousseeuw and Hubert (1999).

Other methods, such as the least median of squares (LMS) estimator (Hampel (1975)), the least trimmed squares (LTS) estimator (Rousseeuw (1984, 1985)), the least trimmed absolute deviations (LTA) estimator (Bassett (1991) and Hossjer (1991)), and the minimum covariance determinant estimator (MCD) (Rousseeuw (1984)), are impractical to compute except in small data sets, since they involve the combinatorial problem of determining the *c* best-fitting data points to include. As a result, so-called "practical" implementations – more accurately, *approximations*, such as "fast LTS" or "fast MCD"[3] – are used. But these approximations do *not* necessarily share the good properties of the actual estimators. Why? Most commonly, such approximations involve "basic resampling," where numerous random samples or "starts" of minimal size are drawn, on which basis distances or residuals can be estimated (and appropriate minimization can take place), generally after refinement steps.[4] Effectively, these approaches attempt to stumble upon a subset of the data that is clean, which can then be used to identify problematic data points. However, Hawkins and Olive (2002) show that resampling algorithms that use a fixed number *K* of starts of bounded size produce *inconsistent* estimators.[5] Furthermore, as those authors show (and as the simulations below reinforce), huge computations are needed for good approximations in large, high-dimensioned samples.[6] When data are systematically contaminated, large sample sizes make things worse for these methods, not better, since they increase computational time without improving the performance of the individual starts.

Fortunately, better methods that address the aforementioned deficiencies now exist. Our paper falls under an "MCD" general approach, one that shares some features with fast MCD, but rather than using numerous *random* starts, our approach instead uses a handful of "initial" robust estimates of the mean and covariance of the sample as the starts. These initial estimates are refined, after which an MCD criterion is applied to select the final robust estimate of the mean and covariance. Equipped with these robust estimates, it is straightforward to conduct subsequent

[3] See Rousseeuw, et al. (2004) for fast LTS and Rousseeuw, and Van Driessen (1999) for fast MCD.

[4] As Hawkins and Olive (2002) observe, refinement steps do not improve theoretical convergence rates, but they can often give dramatic practical improvement.

[5] These authors provide a dramatic example of failure using a real-world data set with only 276 observations and seven variables that had nine outliers: five infants and four toddlers. Even with 30,000+ starts, the final LMS and LTS fits accommodated the five infants. Note that if a random start contains one or more contaminated cases, that start will fail. Olive and Hawkins (2010) further establish that "If the algorithm needs to use many attractors to achieve outlier resistance, then the individual attractors have little outlier resistance. Such estimators include elemental concentration algorithms, heuristic and genetic algorithms, and projection algorithms." Hawkins and Olive (2002) further demonstrate via simulation that finding six-standard-deviation outliers in a 50-dimensional regression is a nontrivial undertaking. They also note that even in theory, the *exact* LTS, LMS, and LTA estimators can pass through outliers. Yohai's two-stage estimators, such as MM (Yohai (1987)), need an initial consistent high breakdown estimator – but are always implemented with inconsistent estimators such as the ones noted above.

[6] Similarly, the forward search – see Atkinson, Riani, and Cerioli (2004) – requires a valid start, such as an LTS-based start; furthermore, it requires long computation times for large data sets.

inference, such as multivariate regression (see below). Further, these estimates allow the straightforward detection of unusual data points, based on (now robustly-estimated) Mahalanobis distances. The analysis of UDPs can lead to a deeper understanding about the data-generating process (see also the discussion in Knez and Ready (1997)), and identification of such data is sometimes the point of an analysis (for example, detecting counterfeits or fraud, identifying superstars). Indeed, as Janson and Verbrugge (2020) discuss in more detail, tools like these can play a critical role in discovering heretofore unknown unobserved variables (or multiple data-generating mechanisms) in the data.

Our chief contribution is to extend previous work in this branch of the literature by integrating insights from a separate branch of the machine learning literature, clustering. This is a notable advance. UDPs that are distributed completely at random represent the least troublesome type of contamination, but contamination that is systematic, that is, contamination consisting of *groupings* of aberrant points, can be both far more problematic and far more challenging to detect. This is due to *masking.* As discussed in more detail in the Appendix, masking occurs when a cluster of anomalous points effectively disguise each other, compromising inference without there being any indication that something has gone wrong. Traditional approaches to outlier detection, such as Cook's *d*, are rarely able to detect such outlier configurations. But these same configurations can also pose a challenge to most methods, as will be evident in our performance comparison. On the face of it, cluster analysis would seem well-suited to addressing this challenge, since its entire aim is to find partitions that best approximate the data. But conventional clustering methods tend to be non-robust and can fail in the presence of aberrant data.

We make several contributions. We develop a novel *robust* clustering technique. We further employ this as a novel robust start in an MCD-based method, along with a second novel robust start that we develop. Are these significant contributions? Following the standard approach in this literature, we assess the relative performance of our new methods on the basis of classic real-world data sets (where the "answer" is "known") and using artificial data that are "contaminated" in various ways. We demonstrate that our methods operate on par with, and in some cases dominate, the heretofore best methods. And we demonstrate that these tools "matter," in that we demonstrate that in the economic study we investigate, UDPs drive the main results (and for most of the data, the opposite conclusions hold). This does not mean that the results are "wrong" — but it does mean that the story is (at least) more nuanced.

Thus, we provide economists with powerful new tools for understanding the data they are confronted with. We believe that methods like ours should be routinely used as part of a scientific (hands-off) robustness test of results – and may also be useful in machine learning applications, where contamination is a severe threat to inference. If evidence for UDPs is uncovered, a researcher then faces the important – though perhaps challenging – decision about what to do next. With knowledge of the set of UDPs in hand, she must determine the best course of action: further study to determine if there is an omitted categorical variable; developing or altering existing theory to take into account the now-richer understanding of the data; winsorizing, data cleaning, or simply focusing on one part of the sample, such as the majority; or perhaps something else.

BACKGROUND

Arguably the two best current practical estimators of multivariate location and dispersion are a reweighted version of the FCH ("fast, consistent, and highly outlier resistant") estimator of Olive and Hawkins (2010), and the detMCD ("deterministic minimum covariance determinant") estimator of Hubert, Rousseeuw, and Verdonck (2012). These are practical algorithms purported to approximate the MCD estimator of Rousseeuw (1984), although this claim is admittedly tenuous (see Olive and Hawkins (2010) and Olive (2017)). These "distance-based" methods are simple, fast, and non-parametric in that they do not require knowledge of the underlying data distribution. Zhang, Olive, and Ye (2012) demonstrate the solid performance of a reweighted FCH estimator entitled RMVN (so termed because it is reweighted and because it is useful for estimating the parameters of outlier-contaminated multivariate normal data.) Both the RMVN and detMCD estimators are practical in that they may be applied to large data sets and estimated relatively quickly; and both have been shown to work well for both simulated and real-world data, although they have never previously (to our knowledge) been compared in a single study. As we explain below, FCH-based estimators are backed by large-sample theory (described below), a rarity in this literature – though small sample performance with contaminated data may be of greater interest. Furthermore, in a sense to be made more precise below, the RMVN estimator is subsumed within the estimator introduced here.

This study makes five contributions. First, it introduces a new multivariate outlier identification technique, simple cluster-based UDP identification (sCBUI), designed to partition data into components that were generated by different multivariate data-generating processes. In the data-mining field, clustering methods are often used for outlier detection (see, for example, Yu, Sheikholeslami, and Zhang (2002), Ghoting, Parthasarathy, and Otey (2006), Bhaduri and Matthews (2011), Pamula, Deka, and Nandi 2011), or Bansal and Chugh (2013)). In this literature, emphasis focuses on efficient anomaly detection, I/O costs, and computational burden. But despite the existence of some general procedures directly aimed at robust covariance matrix estimation (for example, Wang and Raftery (2002)), cluster-based methods have not gained traction in the wider statistics field, nor have they previously been integrated into broader algorithms. This may be partly due to their complexity and/or instability (see Croux and Van Aelst (2002)). sCBUI uses a very simple clustering method, based on distance or density; yet despite this simplicity, it is powerful. For some outlier configurations, taken as a standalone method, sCBUI outperforms all the other methods considered in this study. Thus, it may be useful when there is reason to believe that the data are composed of distinct clusters. But it is also useful when used as a robust start to an MCD-based method, as we describe next.

Second, this study introduces a new "hybrid" estimator of multivariate location and scatter. This multiple-start estimator shares many of the building blocks with RMVN and detMCD, although it has several distinctive features, including two distinctive starts. Since one of these starts is a cluster method, sCBUI, we call it a hybrid method. This method is relatively fast and straightforward to compute. Third, we introduce a second robust start that is completely new.

Fourth, we present simulation evidence to compare the sCBUI method, the hybrid method, RMVN, detMCD, and three other notable methods. To our knowledge, this is the first study to directly compare the relative performance of RMVN and detMCD. The evidence presented here indicates that our hybrid method almost invariably performs on a par with, and in many cases outperforms, RMVN and detMCD, putting this method in good company indeed.

Finally, this study demonstrates the applicability of the new estimators on some real-world data sets. Like the simulation evidence, these applications indicate that the methods perform well. In one case, involving forged currency, sCBUI has a 100 percent detection success rate, compared to rates of around 15 percent by the other methods examined. As a second example, a prominent economic study is re-examined, and we demonstrate that the main results are driven by UDPs.[7] Taken together, these applications indicate that our concerns about robustness are not hypothetical, and that a failure to adequately take into account the possibility of outliers can lead to distorted inference and to misleading conclusions.

We focus here on tasks that are central to many robust statistical procedures: identifying outliers and estimating location and scatter. We do not investigate whether various procedures benefit from "plugging in" these robust estimates. The study of robust methods for more general inference problems is interesting and left to future research.

[7] Previous versions of this paper examined three such studies.

FRAMEWORK

Following Zhang, Olive, and Ye (2012), we describe the general framework as follows. A multivariate location and dispersion (*MLD*) model is a joint distribution for a k x 1 random vector $\boldsymbol{x}$ that is completely specified by a k x 1 population *location* vector μ and a k x k symmetric positive definite population *dispersion* (or *scatter*) matrix Σ. An important *MLD* model is the elliptically contoured distribution $EC_k(\mu, \Sigma, g)$ with probability density function

$$f(z) = g_k |S|^{-1/2} g\left[(z - m)^T S^{-1} (z - m)\right]$$

where z is a $k' \; 1$ dummy vector, $g_K > 0$ is some constant, and g is a known function. The multivariate normal $N_k(\mu, \Sigma)$ is a special case, with $g(z) = (2\pi)^{-k/2} e^{-z/2}$, as is the elliptical k-variate Student distribution with ν degrees of freedom (0 < ν < ∞). Further, x is "spherical about μ" if x has an $EC_k(\mu, c\mathbf{I}, g)$ distribution where $c > 0$ is some constant, and $\mathbf{I}$ is the conformable $(k' \; k)$ identity matrix. In this literature it is often assumed, explicitly or implicitly, that the distribution F of the uncontaminated (or clean, or regular) data is elliptical. For elliptical distributions: a) the contours of constant density are ellipsoids; b) if the mean and variances of X exist, then $E(X) = \mu$ and $\mathrm{Cov}(X) = c\Sigma$ for some constant $c > 0$ (for Normal distributions, $c = 1$); and c) X can be written as $X = AZ + \mu$, with A satisfying $\Sigma = AA^T$ and Z a random variable with a spherical distribution about 0. As discussed in Olive (2008), many classical procedures originally meant for multivariate normal distributions are semi-parametric in that the procedures also perform well on a much larger class of *EC* distributions. (If the data-generating process is not elliptical – for instance, if it is log-normal – a suitable data transformation may be necessary. As will be evident below, selection of a data transformation inevitably requires judgment on the part of the practitioner: A cluster of outliers may suggest an asymmetric but unique data-generating process when, in fact, data are generated from two or more distinct processes. Note that the Box-Cox transformation is not a reliable guide in that it is sensitive to outliers. For a more thorough treatment of this topic and operational suggestions, see Janson and Verbrugge (2020)).

For *EC* distributions, let constants $d > 0$ and $c_X > 0$. Then a dispersion estimator estimates $d\Sigma$, and one such dispersion estimator, a covariance matrix estimator, estimates the covariance matrix $\mathrm{Cov}(x) = c_X\Sigma$. (To fix ideas, we remark that if the bulk of the data is $N_k(\mu, \Sigma)$, then the RMVN estimator will, for certain types of outlier configurations, give what Zhang, Olive and Ye (2012) term a "useful" estimate of (μ, Σ), while FCH itself would estimate $(\mu, d\Sigma)$ for $d > 1$. These estimates are merely "useful" because *consistency cannot generally be claimed when outliers are* *present.*) For multivariate analysis, the classical estimator (x, S) of ($E(x)$, Cov(x)) is the sample mean and sample covariance matrix:

$$\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad S := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T$$

Let $\bar{x}_H$ and S_H denote the classical estimators taken over the subset of the data whose indices are in H. Let the k×1 column vector L_X be a multivariate location estimator, and let the $k \times k$ symmetric positive definite matrix C_X be a dispersion estimator. The notation (L, C) will often be used, suppressing X. The ith *squared sample Mahalanobis distance* (from L) is the scalar

$$D_i^2 \equiv D_i^2(L,C) := (x_i - L)^T C^{-1}(x_i - L) \tag{1}$$

for each observation x_i. (This is sometimes termed the "statistical distance.") The Euclidean distance of x_i from L is $D_i(L, I_k)$. The classical Mahalanobis distance uses the classical estimates $(L,C) = (\bar{x}, S)$. Wilks (1962) showed that under a multivariate normal distribution, D_i^2 follows a scaled Beta distribution:

$$D_i^2 \sim \frac{(n-1)^2}{n} Beta\left(\frac{k}{2}, \frac{n-k-1}{2}\right) \tag{2}$$

(In practice, a chi-squared approximation is often used for simplicity.) For a nominal test of size α, a data point j may be identified as an outlier if D_j^2 exceeds the (1-α) quantile of the scaled Beta distribution (2). If there is only one outlier, this is an accurate and powerful test. But if based upon the *classical* Mahalanobis distance, this test can easily fail in the presence of more than one outlier, owing to masking. One effective way to avoid masking is to use high-breakdown estimators of L and C in (1), and this is the basis of many multivariate outlier tests, including those examined in this study.

Wilks also conjectured that a Bonferroni bound could be used to test outlyingness without much loss of power, and this idea has been formalized as a means of controlling the false discovery rate in Cerioli (2010). However, our results (see Appendix A) indicate that the accompanying power loss is rather severe.

The MCD subset (Rousseeuw (1984)) is defined to be the subsample of h observations, with $n/2 \le h < n$, whose covariance matrix has the smallest determinant. Let $y_{MCD} = \{i_1, \ldots, i_h\}$ denote the indices of the observations in this subset. The MCD estimate of *location* is the average of the MCD subset. The MCD estimate of *scatter* is proportional to the dispersion matrix of this subset, $\hat{\Sigma}_{MCD} = \lambda_{MCD}(h,n,k) S_{MCD}$, where $\lambda_{MCD}(h,n,k)$ is a proportionality constant that makes the estimator consistent and unbiased for Σ when the data are iid $N(m,\mathrm{S})$ (see Pison, Van Aelst, and Willems (2002)). Since finding the MCD is impractical on large data sets, "practical" approaches (or approximations) have been developed.

Several of the most popular robust estimators generate K trial starts $\{(L_{0j},C_{0j}),\ j=1,\ldots,K,\}$ then use the following *concentration* technique. Using start *j*, compute all Mahalanobis distances $D_i^2(L_{0j},C_{0j})$ for all *n* data points. Let $y_{1j} = \{i_{1j},\ldots,i_{kj}\}$ denote the indices of the $\kappa \approx n/2$ cases corresponding to the smallest distances (that is, distances less than or equal to the median distance). Then compute the classical estimators over those κ points, so that at the next iteration, $(L_{1j},C_{1j})=(\bar{x}_{y_{1j}},S_{y_{1j}})$. This iteration is continued for s steps. This results in a sequence of estimators (L_{0j},C_{0j}), …, (L_{sj},C_{sj}). Each concentration step can only decrease the determinant of *C*. Hence, this procedure can either be iterated until "convergence," or alternatively stopped after a predetermined number *s* of steps. For the estimators in this paper, following Zhang, Olive, and Ye (2012), $s=5$ concentration steps are used. In a small abuse of terminology, and to maintain contact with the literature, let (L_{sj},C_{sj}) denote the *j*th attractor. Using the MCD criterion, select the attractor with the smallest determinant $\det(C_{sj})$ as the basis for the final estimator. At this point, various reweighting steps are undertaken that improve the estimates.

As noted above, many (most?) consistency results in the literature are "highly fragile" (Müller 2007): Small contaminations can make the estimator converge in probability to something arbitrarily divergent. The breakdown value of an estimator is closely related to *qualitative robustness*. Let *F* denote the distribution of $\boldsymbol{x}$ and let *G* denote the asymptotic distribution of an estimator θ. The estimator θ is said to be qualitatively robust at *F* if for every $\varepsilon > 0$ there exists a $\delta > 0$, such that if $d(F,F') < \delta \Rightarrow d(G,G') < \varepsilon$, for a suitable metric *d*(.,.) (Hampel et al. (1986) suggest the Levy-Prokhorov metric), and for distribution function F'. The intuition is that if the distribution of $\boldsymbol{x}$ is perturbed only slightly, then the corresponding change in the asymptotic distribution of the estimator *G* should also be small. By letting $F'=(1-\varsigma)F+\varsigma F''$ for $0<\varsigma<<1$, and considering various distributions for F'' such as Cauchy, we can see that many estimators are not qualitatively robust, and their consistency will break down given modest contamination. (See Zähle (2015) for a more general definition of qualitative robustness.)

It is known that multivariate regression is very sensitive to outliers in the data. But given identified outliers and/or a robust estimate of location and scatter, it is straightforward to undertake robust multivariate regression (Rousseeuw et al., (2004)). This is a big deal, since the computational complexity of most "brand-name" robust estimators, such as least trimmed squares, makes them infeasible on large data sets. The least complicated robust regression method is to drop outlier observations at the outset, that is, treat these observations as missing, and proceed as usual. Indeed, finding a set of observations to drop is the approach of most algorithms (see Torti et al., (2012)), and in some situations, this approach is the best one can do. However, one may wonder if this might lead to underestimating estimator standard errors.[8] Accuracy may be enhanced if one makes use of the robust $(\hat{L},\hat{C})$ estimates. The multivariate regression model is given by $\mathbf{y}=\mathbf{B}'\mathbf{x}+\boldsymbol{\alpha}+\boldsymbol{\varepsilon}$ where $\mathbf{y}$ is a q-dimensional variable, $\mathbf{x}$ is a p-dimensional set of predictors, $\mathbf{B}$ is a $(p\times q)$ slope matrix, $\boldsymbol{\alpha}$ is a q-dimensional intercept, and the errors $\boldsymbol{\varepsilon}$ are q-dimensional with mean $\mathbf{0}$ and $\operatorname{cov}(\boldsymbol{\varepsilon})=\boldsymbol{\Sigma}$. Let $\boldsymbol{\mu}=\begin{pmatrix}\boldsymbol{\mu}_{\mathbf{x}} & \boldsymbol{\mu}_{\mathbf{x}}\end{pmatrix}^T$ denote the mean of the joint $(\mathbf{x},\mathbf{y})$ variables, and partition $\boldsymbol{\Sigma}$ accordingly:

$$\boldsymbol{\Sigma}=\begin{pmatrix}\boldsymbol{\Sigma}_{\mathbf{xx}} & \boldsymbol{\Sigma}_{\mathbf{xy}}\\ \boldsymbol{\Sigma}_{\mathbf{yx}} & \boldsymbol{\Sigma}_{\mathbf{yy}}\end{pmatrix}$$

The least squares estimators of $\mathbf{B}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ can be written as functions of the estimators $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ (see Rencher and Christensen (2012), p.362)

$$\hat{\mathbf{B}}=\hat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-1}\hat{\boldsymbol{\Sigma}}_{xy}$$
$$\hat{\boldsymbol{\alpha}}=\hat{\boldsymbol{\mu}}_{\mathbf{y}}-\hat{\mathbf{B}}^T\hat{\boldsymbol{\mu}}_{\mathbf{y}}$$

and it is straightforward to show that

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon}}=\hat{\boldsymbol{\Sigma}}_{\mathbf{yy}}-\hat{\mathbf{B}}^T\hat{\boldsymbol{\Sigma}}_{\mathbf{xx}}\hat{\mathbf{B}}.$$

Hence, robust estimates $(\hat{L},\hat{C})$ of $(\mathbf{x},\mathbf{y})$ may be used in the above expressions to undertake robust multivariate regression.

Note that further improvements are possible, by focusing on the specifics of the regression context; see Rousseeuw et al. (2004) and Olive and Hawkins (2011). However, Blankmeyer (2016) provides evidence that four high-breakdown regression estimators —robust MM (Maronna, Martin, and Yohai (2006)), least trimmed squares (Rousseeuw and Leroy (1987) and Rousseeuw and Van Driessen (2006)), "hbr" (the high breakdown regression method of Olive and Hawkins (2011)), and "fwd" (the forward search regression method of Torti et al., (2012)) — may not

[8] Olive (2003) points out that if outliers form a recurring fraction of the data, analysis based on the cleaned data (with outliers deleted) will typically give rise to misleading inferences. In prediction, outliers will unavoidably be missed, but unadjusted hypothesis tests and interval coverage will only be sized at level α for a hypothetical: No more outliers appear. Thus, one must choose a ¢< a in order to achieve a size of a .

perform well when the true R^2 is mediocre. As noted above, some of these methods are computationally intensive. Blankmeyer's results support the use of the "raw" (non-reweighted, and less efficient) estimators from the MCD family. In his study, the efficient versions of these estimators, which are estimated over a larger fraction of the data, tend to retain unacceptably large biases in this circumstance.

sCBUI (simple Cluster-Based UDP Identification) Algorithm

General description

The sCBUI algorithm, as its name indicates, uses clustering of data to identify UDPs. In the initial stage, Euclidean or classical distance is used to locate candidate clusters, starting with a case in the highest density part of the data. Then its "neighbors," that is, the nearby cases, are used to provide initial covariance estimates for the cluster. Next, data are allocated to this cluster on the basis of Mahalanobis distance. After removing those cases initially assigned to the first cluster, the process is then repeated multiple times, to find up to four clusters. Next, we test whether the clusters are statistically distinguishable. Finally, the largest cluster becomes the basis of the final estimates, with other clusters and excluded points identified as UDPs. Note that in contrast to all the other methods considered here, this method does not use iterated concentration steps. Related to this, the selection of final estimates does not depend in any way on minimizing a covariance determinant.

Procedure

1. Let X be an $n \times k$ data matrix. If $n \le 2000$, let $X' = X$ and $n' = n$. Otherwise, select a simple random sample of size 2000 from the data, and denote this subset X', with $n' = 2000.$[9]
2. Let med$_X$:= MED(X), the coordinate-wise (column) median, and let mad$_X$:= MAD(X), the coordinate-wise median absolute deviation. Let meddist$_X$:= MED(D_i(med$_X$,I_k)), the median Euclidean distance between each observation $\boldsymbol{x_i}$ and med$_X$. Note that med$_X$ as a center and meddist$_X$ as a radius define the *median ball*, which plays a role in the sequel. For each observation $\boldsymbol{x}_i$ Î X¢, define the $n' \times 1$ vector $\boldsymbol{e_i}$ with the *j*th element $e_{ij} := D_j\left(x_i, I_k\right)$, "$\boldsymbol{x}_j$ Î X¢ the vector $\boldsymbol{e_i}$ is the Euclidean distance between the point $\boldsymbol{x_i}$ and all the other

[9] We restrict the *initial* size of the sample, since the initial steps are time-consuming. Subsequent steps use the entire data.

points $\boldsymbol{x_j}$ in $X\text{¢}$. Let $r_i := \frac{1}{2}\left(q_{i,10} + q_{i,25}\right)$ where $q_{i,l}$ denotes the lth order statistic of $\boldsymbol{e_i}$. Define $MED_{10} := \underset{i}{MED}\left(q_{i,10}\right)$. As the median of the 10th percentile of all pairwise distances between points in X', MED_{10} is a global measure of closeness, while ρ_i is a relative measure, the nearest neighbors in X' of a given observation i.

3. Define the $n' \times n'$ *neighbors matrix* M_1 with elements
$$m_{ji} = \begin{cases} 1.5 & \textit{if } e_{ji} \le MED_{10} \\ 1 & \textit{if } e_{ji} > MED_{10} \text{ but } e_{ji} \le \min\left(meddist_X, \rho_i\right) \\ 0 & \textit{otherwise} \end{cases} .$$
In other words: each column i of M_1 contains a 1.5 in position j if $\boldsymbol{x_j}$ lies within distance MED_{10} of $\boldsymbol{x_i}$; (that is, it is close in a global sense to $\boldsymbol{x_i}$); otherwise column i contains a 1 in position j row if $\boldsymbol{x_j}$ lies within distance ρ_i of $\boldsymbol{x_i}$ (that is,, $\boldsymbol{x_j}$ is one of the nearest neighbors of $\boldsymbol{x_i}$); otherwise entry j of column i equals zero (i.e., $\boldsymbol{x_j}$ is not a near neighbor of $\boldsymbol{x_i}$).

4. Let $\boldsymbol{m}_j$ denote row j of M_1, with elements m_{ji}. Define $j^* \text{Î} \underset{j}{\sup}\left(\text{å}_i\, m_{ji}\right)$. In words, $\boldsymbol{x}_{j^*}$ is an observation that lies well within a dense cluster of points, since "many" other observations have $\boldsymbol{x}_{j^*}$ as a nearest neighbor. Let $\boldsymbol{m}_{j^*}$ denote the row of M_1 corresponding to j^*, with elements m_{j^*i}. Define a set of weights denoting all observations for which $\boldsymbol{x}_{j^*}$ is a nearest neighbor, as follows:
$$w_{j^*i} := \begin{cases} 1 & \textit{if } m_{j^*i}\ {}^1\ 0 \\ 0 & \textit{else} \end{cases}$$
These define a subset of data points that are provisionary or founding members of the first cluster. Let $h_1 := \underset{i}{\text{å}}\ w_{j^*i}$, the cardinality of this subset. If $h_1 > k$, then the initial cluster 1 estimate of location, $\hat{m}_1$, is the average of this subset; that is,
$$\hat{m}_1 = \tfrac{1}{h_1} \overset{n\text{¢}}{\underset{i=1}{\text{å}}}\ w_{j^*i}\boldsymbol{x_i} ,$$
while the initial cluster 1 estimate of scatter is the dispersion matrix of this subset:
$$\hat{C}_1 = \tfrac{1}{h_1 - 1} \overset{n\text{¢}}{\underset{i=1}{\text{å}}}\ w_{1i}\left(\boldsymbol{x_i} - \hat{m}_1\right)\left(\boldsymbol{x_i} - \hat{m}_1\right)\text{¢}$$
Given these estimates, we define members of cluster 1 as all points $\boldsymbol{x_i} \in X$ whose squared Mahalanobis distance from $\hat{m}_1$ lies within the 97.5th percentile of a chi-square distribution

with k degrees of freedom: that is, all cases i satisfying $D_i^2\left(\hat{m}_1,\hat{C}_1\right)£\ c^2_{k,0.975}$. Let $y_1 = \left\{i_1,\ldots,i_{h_1¢}\right\}$ denote the indices of the observations belonging to cluster 1, with $h_1¢$ members in this set. (If $h_1\ £\ k$, then the cardinality of the subset is too small to estimate a covariance matrix, and the sCBUI algorithm returns the empty set.)

5. We search for up to three additional clusters of cases by repeating step 4, but each time creating a matrix M_r that is identical to M_{r-1} except that all columns of M_{r-1} whose indices have been identified as a member of a pre-existing cluster are replaced by columns of zeros. For example, in looking for a second cluster, we create the matrix M_2 by setting equal to a column of zeros all columns of M_1 whose indices are members of y_1. The search for clusters terminates once we reach four clusters, or once we identify a set of nearest neighbors in step 4 where the cardinality of the set is less than or equal to *k*.
6. Let the clusters be indexed by *l*, where $1 \le l \le c$ and $c\,£\ 4$. Then the estimates of the location and dispersion matrix of each cluster is

$$\overset{(}{m}_l = \frac{1}{h_l¢}\underset{i\hat{I}\ y_l}{\mathring{a}}\ \boldsymbol{x}_i$$

$$\overset{(}{C}_l = \frac{1}{h_l¢-1}\underset{i\hat{I}\ y_l}{\mathring{a}}\left(\boldsymbol{x}_i - \overset{(}{m}_j\right)\left(\boldsymbol{x}_i - \overset{(}{m}_j\right)¢$$

   If $h_j¢£\ k$, then cluster *j* is eliminated.
7. If $c \ge 2$, we now redefine clusters as follows. For each observation $\boldsymbol{x}_i \in X$ and for each cluster *l*, we compute $p_{il} := P\left(c^2_k\ £\ D_i^2\left(\overset{(}{m}_l,\overset{(}{C}_l\right)\right)$ where $P\left(c^2_k\ £\ z\right)$ is the cumulative distribution function of a chi-square random variable with *k* degrees of freedom evaluated at *z*. When p_{il} is a small number, point $\boldsymbol{x}_i$ is in close proximity – in the Mahalanobis distance sense – to the center of cluster *l*, suggesting that this case is properly a member of cluster *l*. For each *i*, we consider only the two smallest *pil*; denote the smallest by p_{il_1} and the next smallest by p_{il_2}, with associated clusters l_1 and l_2. If $p_{il_1} -\ p_{il_2} > 0.10$, then $\boldsymbol{x}_i$ is assigned to cluster l_1; in other words, if $\boldsymbol{x}_i$ is "sufficiently closer" to the center of cluster l_1 than to the center of cluster l_2, then $\boldsymbol{x}_i$ is unambiguously a member of cluster l_1. Further, if $\boldsymbol{x}_i$ is "sufficiently close" to the center of cluster l_1, then this observation is granted "protected" status and can only be subsequently removed from cluster l_1 if cluster l_1 is absorbed into another cluster. We take "sufficiently close" to mean that $p_{il_1} < 0.80$. (If

$c = 1$, then for each observation $\boldsymbol{x_i}$ compute p_{i1}; if $p_{i1} < 0.80$, then case $\boldsymbol{x_i}$ is granted "protected" status.) Proceed to step 9.

8. The cardinality of each cluster is computed, and the cluster with the most members is denoted the "major cluster," cluster *m*. For each other cluster *l*, $l \neq m$, we compute the Mahalanobis distance from the location vector of cluster *l* to that of cluster *m*, first using dispersion matrix $\hat{C}_l$ and then using the dispersion matrix $\hat{C}_m$. In particular, we compute $p_{lm} := P\left(c_k^2 \text{ £ } D^2_{\hat{m}_m}\left(\hat{m}_l, \hat{C}_l\right)\right)$ and $p_{ml} := P\left(c_k^2 \text{ £ } D^2_{\hat{m}_l}\left(\hat{m}_m, \hat{C}_m\right)\right)$. If p_{lm} £ 0.975 – that is, if $D^2_{\hat{m}_m}\left(\hat{m}_l, \hat{C}_l\right)$£ $c^2_{k,0.975}$ – then the location vector $\hat{m}_m$ is within the 97.5 percent confidence region of cluster *l*. If p_{lm} £ 0.975, p_{ml} £ 0.975, and $p_{lm}p_{lm}$ £ 0.90, we conclude that the centers of each cluster are sufficiently close, and cluster *l* is subsumed into cluster *m*.
9. Let $y_m = \left\{i_1, \ldots, i_{h_m}\right\}$ denote the indices of the observations belonging to the major cluster. If $h_m > k$, we define

$$m^{CBUI} = \frac{1}{h_m} \mathop{\text{å}}_{i\,\text{î}\,y_m} \boldsymbol{x_i}$$

$$C^{sCBUI¢} = \frac{1}{h_m - 1} \mathop{\text{å}}_{i\,\text{î}\,y_m} \left(\boldsymbol{x_i} - m^{CBOI}\right)\left(\boldsymbol{x_i} - m^{CBOI}\right)^{¢}$$

10. UDPs or outliers consist of the union of all "protected" data points and all data points satisfying $D_i^2\left(m^{sCBUI}, C^{sCBUI¢}\right) > c^2_{k,0.975}$.
11. Finally, the dispersion matrix is adjusted with a consistency factor:

$$C^{sCBUI} = \frac{h_m/n}{P\left(c^2_{k+2} < c^2_{k,h_m/n}\right)} C^{sCBOI¢}$$

Optionally, one may refine these estimates by undertaking five concentration steps and then undertaking refinements, such as those in steps 2 and 3 in the "Final selection and refinements" subsection below. (In the sequel, our refinements of this estimate followed those of the RMVN algorithm.)

Hybrid MCD Algorithm

Introduction

The hybrid method borrows ideas from both RMVN and detMCD. With one or both methods, it shares some starts, iterated concentration steps, and reweighting steps upon final selection of a subset of data. It differs from both along a number of dimensions. There are two starts unique to the hybrid method. Since one of these is sCBUI, this hybrid method effectively uses both schemes' outliers → robust estimates, and robust estimates → outliers. While RMVN has two starts, and detMCD has six, the hybrid MCD method has eight. In contrast to both methods, initial rescaling of data is done for a strict subset of the starts, using median absolute deviation. Concentration occurs on raw (non-rescaled) data. In contrast to detMCD, final selection of the attractor makes use of the FCH distance-based test. Reweighting to produce final estimates and selection of outliers mainly follows methods developed for fMCD and detMCD. (Other options were investigated, but proved inferior.) While the hybrid method is not as fast as detMCD or RMVN, owing to its use of the sCBUI start, it is still quite fast and feasible to apply on very large data sets of high dimension, unlike fMCD and many other methods.

The hybrid method proceeds by constructing eight initial estimates $\hat{\mu}_r$ and $\hat{\Sigma}_r$ (r = 1,...,8) for the center and scatter of X. Then five concentration steps are undertaken. Then as noted above, the selection of the attractor makes use of the FCH distance-based test. Finally, estimates are reweighted and refined, and outliers determined on the basis of the final estimates.

Data standardization: Starts 1-5

The first five starts below are similar to the detMCD procedure in that these starts begin with standardized data, although unlike detMCD, standardization uses an alternative to Q_n, and subsequent concentration takes place using non-rescaled (raw) data. To standardize the data, from each data point, we subtract med$_X$, and then divide by mad$_X$. The standardized data set is denoted by Z with rows z_i (i = 1,...,n) and columns Z_j (j= 1,...,k).

Location and scatter estimates for standardized data starts

For the standardized data starts, after computing a preliminary estimate $\breve{S}_r$ of the covariance matrix of Z, the preliminary matrix immediately undergoes a sequence of operations based on eigenvalue manipulations that were introduced in Maronna and Zamar (2002). (When scatter matrices are constructed element-wise, these operations ensure that it is a positive definite [and reasonably accurate] covariance estimate.) In particular:

1. Compute the matrix E of eigenvectors of $\breve{S}_r$ and project Z onto E: $B = ZE$.
2. Estimate the (positive definite) covariance of Z by $\breve{\Sigma}_r\left(Z\right) = ELE^T$ where L, the "robust variance" vector, is given by $L = diag\left(\text{mad}_B \cdot \text{mad}_B^T\right)$.

3. To estimate the center of *Z* we sphere the data, apply the coordinate-wise median, and transform it back: $\hat{\mu}_r\left(Z\right) = \breve{\Sigma}_r^{1/2}\left(MED\left(Z\breve{\Sigma}_r^{-1/2}\right)\right)$
4. Rescale all estimates: $\hat{\mu}_r = \hat{\mu}_r\left(Z\right) + \text{med}_X$ and $\hat{\Sigma}_r = \text{mad}_X\breve{\Sigma}_r\text{mad}_X^T$.

Description of eight starts

1. The first initial scatter matrix is obtained by computing the hyperbolic tangent (sigmoid) of each column of *Z*; that is, $Y_j = \tanh\left(Z_j\right)$ for (*j*= 1,...,*k*). This is a bounded function that substantially reduces large coordinate-wise outliers. The classical correlation matrix of *Y* yields $\breve{S}_1 = corr\left(Y\right)$. Location and scatter estimates are then constructed as in the previous subsection.
2. The second initial scatter matrix is closely related to the first. As above, $Y_j = \tanh\left(Z_j\right)$ for (*j*= 1,...,*k*). Then the smallest 10thile, and largest 30thtile, of points are trimmed. Denote the subset of retained cases in *Y* by *TH*, the "trimmed hyperbolic" subset. The estimate of scatter is the classical correlation matrix of this subset, $\breve{S}_2 = S_{TH}$. Location and scatter estimates are then constructed as in the previous subsection.
3. Let R_j be the ranks of the column Z_j, and set $\breve{S}_3 = corr\left(R\right)$. This is the Spearman correlation matrix of *Z*. Location and scatter estimates are then constructed as in the previous subsection.
4. The fourth scatter matrix is based on the *spatial sign* covariance matrix (Visuri, Koivunen, and Oja, (2000)). Define $t_i = z_i/\left\|z_i\right\|$ for all *i*, where $\left\|z_i\right\| = D_i\left(\text{med}_Z, I_k\right)$, and set $\breve{S}_4 = \text{cov}\left(T\right)$. Location and scatter estimates are then constructed as in the previous subsection.
5. The fifth scatter matrix, novel to this paper, is the *rom-ρ* correlation matrix. This correlation matrix is estimated using a variant of the high-breakdown *rom-ρ* method introduced in Chakhchoukh et al. (2010); this method estimates a pairwise correlation r on the basis of a ratio of medians of products (*rom*). While a closed-form relationship between *rom* and the correlation coefficient ρ does not seem to exist, the relationship between sample estimates of *rom* and ρ is stable (see the Appendix) and can be obtained numerically. Our variant begins with scaled data *Z*; recall that each column has an ostensible mean of zero and standard deviation of 1. We compute each pairwise $\rho_{l,j}$ (for columns *l* and *j*) as

$$\hat{r}_{ij} = f\left(MED\left(Z_iZ_j\right)\right)$$

where *f* is a linear function of odd powers of its argument. (Parameter estimates for our function *f* are provided in the Appendix.) Diagonal elements of $\breve{S}_5$ are set equal to 1. *rom-ρ* has a breakdown value of 25 percent because it computes medians of products. $\hat{\mu}_5$ and $\hat{\Sigma}_5$ are constructed as in the previous subsection.

6. The sixth and seventh starts correspond to the starts used in RMVN. In particular, the sixth location and scatter estimates are the classical estimators, taken over the entire data set *X*.
7. The seventh start is the median ball estimator, which uses the classical estimators computed from the cases within the median ball. In particular, let the median ball (MB) subset consist of the $m = \lceil n/2 \rceil$ observations x_i with smallest norm, that is, the *m* cases satisfying $\|x_i\| = D_i\left(med_Z, I_k\right) \le MED\left[D_i\left(med_Z, I_k\right)\right]$. Then

   $$\hat{m}_7 = \frac{1}{m}\underset{\hat{\imath}\ MB}{\mathring{a}} \boldsymbol{x}_i \quad \text{and} \quad \hat{\text{S}}_7 = \frac{1}{m}\underset{\hat{\imath}\ MB}{\mathring{a}} (\boldsymbol{x}_i - \hat{m}_7)(\boldsymbol{x}_i - \hat{m}_7)\text{¢}$$

   Note that the BACON algorithm of Billor, Hadi, and Velleman (2000) has a similar start, but with fewer observations.
8. The eighth start consists of the sCBUI estimates, described above. We use $\left(m^{sCBUI}, C^{sCBUI\text{¢}}\right)$.

Final selection and refinements

After five concentration steps are undertaken, we obtain the concentrated estimates $\hat{\mu}_r'$ and $\hat{\Sigma}_r'$ for $r = 1,...,8$. These are termed "raw" estimates. As described in step 1 below, the final raw estimator is selected from the eight raw estimates, based on the MCD criterion, but only after imposing a necessary condition. Then the estimator is refined.

1. Covariance determinants are computed and ranked. Selection is based on the MCD criterion, but only after imposing a necessary condition taken from Olive and Hawkins (2010): For start *i* to be chosen, its location estimate must lie within the median ball, that is, $\|\mu_i'\| \le MED\left[D_i\left(\text{med}_Z, I_k\right)\right]$. If all other starts fail this test, then, as in RMVN, the median ball start is selected. Denote the selected estimates $\hat{\mu}_{raw}'$ and $\hat{\Sigma}_{raw}'$.
2. Following Cerioli (2010), two re-estimation and reweighting steps are undertaken. In the first stage, we select the $m' = \lceil n/2 \rceil$ observations x_i with the smallest distance from $\hat{\mu}_{raw}'$

that is, the cases satisfying $\|x_i\|_{raw} = D_i\left(\hat{\mu}'_{raw}, \hat{\Sigma}'_{raw}\right) \le MED\left[D_i\left(\hat{\mu}'_{raw}, \hat{\Sigma}'_{raw}\right)\right]$. Denote these cases the RAW subset. The refined raw estimate of location is $\hat{m}^{final}_{raw} = \frac{1}{m¢}\,\text{å}_{i\text{Î}\;RAW}\, x_i$, and the refined raw estimate of scatter is

$$\hat{\Sigma}^{final}_{raw} = rac\left(k,n,m'\right)\cdot S_{RAW}\ .$$

$rac(.)$ is a small-sample correction and consistency factor that makes the covariance matrix consistent for normal variables and unbiased in small samples, given by

$$rac\left(k,n,m'\right) = \frac{m'/n}{P\left(\chi^2_{k+2} < \chi^2_{k,m'/n}\right)} s_1\left(k,n,m'\right)$$

where $s_1\left(k,n,m'\right)$ is a small-sample calibration factor (see Croux and Haesbroeck (1999) and Pison, Van Aelst, and Willems (2002)).

3. A final reweighting step is performed, which increases finite sample efficiency considerably (see Pison, Van Aelst, and Willems (2002)). In this case, we select all observations satisfying $D_i\left(\hat{\mu}^{final}_{raw}, \hat{\Sigma}^{final}_{raw}\right) \le \chi^2_{k,0.975}$, the 0.975 quantile of the χ^2_k distribution. Suppose there are *m* cases satisfying this criterion; denote this subset *HY*, and its complement by *OUT*, denoting outliers. Then the reweighted estimates of location and scatter are given by

$$\hat{\mu}_{hybrid} = \frac{1}{m}\sum_{i\in HY} \boldsymbol{x_i}$$

and

$$\hat{\Sigma}_{hybrid} = \frac{rec\left(k,n,m\right)}{m}\sum_{i\in HY}\left(\boldsymbol{x_i} - \hat{\mu}_{hybrid}\right)\left(\boldsymbol{x_i} - \hat{\mu}_{hybrid}\right)^T$$

where the scaling *rec*(*k*, *n*, *m*) guarantees consistency of the reweighted estimator and improves its small-sample behavior; it is given by

$$rec\left(k,n,m\right) = \frac{m/n}{P\left(\chi^2_{k+2} < \chi^2_{k,0.975}\right)} s_2\left(k,n,m\right)$$

where $s_2\left(k,n,m\right)$ is a small-sample calibration factor (see Pison, Van Aelst, and Willems (2002) and Cerioli (2010)).

Discussion of Cerioli (2010)

We considered identifying UDPs based on the more accurate approximations given in Cerioli (2010), a study focused on controlling the false discovery rate. (Green and Martin (2014) discuss these approximations and also provide a modified approximation.) There are two relevant distributions: the distribution for cases in the subset *HY*, versus the distribution for cases in the subset *OUT*. The latter cases follow a different distribution because they were not used to estimate the scatter matrix. However, using the better approximations in Cerioli (2010) actually yielded somewhat inferior detection rates in our simulation study and led to somewhat inferior scatter matrix estimates.

Other Methods Examined

We provide a summary of the other methods here; they are described in more detail in the cited articles.

fMCD

The fMCD estimator uses the classical estimator applied to *K* = 500 randomly drawn elemental subsets of size *k*+1 as starts. The 10 starts with the smallest covariance determinants are then concentrated. Final selection is determined by minimum covariance determinant. Reweighting steps differ across implementations, but often occur as in the hybrid method. Cases satisfying $D_i^2\left(\hat{L}_{fMCD},\hat{C}_{fMCD}\right) > \chi^2_{k,0.975}$ are identified as outliers. For details, see Rousseeuw and Van Driessen (1999). Note that the simulation results of Hawkins and Olive (2002) indicate that, in large multivariate data sets, 500 could well be too small by several orders of magnitude. We will see below that fMCD fails dramatically in some cases.

detMCD

The detMCD algorithm begins by standardizing *X* by subtracting the component-wise median med$_X$ and dividing by the Q_n scale estimator (Rousseeuw and Croux, (1993)). All computations are carried out on standardized data. All six starts undertake the steps outlined in the subsection “Location and scatter estimates for standardized data starts.” Four of the starts are identical to the hybrid method: the hyperbolic tangent, the Spearman correlation matrix, the spatial sign covariance matrix, and the median ball estimator. The fifth start is derived from the ranks matrix used in the Spearman correlation estimates; in particular, normal scores are derived from a simple transformation of the ranks, and the scatter estimate is the correlation matrix of these scores. The sixth start is the raw OGK estimator based on the method of Maronna and Zamar (2002) applied to the robust covariance estimate of Gnanadesikan and Kettenring (1972), although the location estimator is the component-wise median and the dispersion estimator is Q_n. detMCD iterates concentration steps to convergence, and selection is based on MCD. Cases satisfying $D_i^2\left(L_{detMCD}, C_{detMCD}\right) > \chi^2_{k,0.975}$ are identified as outliers. For details, see Hubert, Rouusseeuw, and Verdonck (2012).

RMVN

All FCH estimators, including RMVN, use the same two starts: the classical estimator and the median ball estimator. Its large sample theory (for uncontaminated data) is as follows. Cator and Lopuhaä (2010, 2012) show that MCD is consistent provided that the MCD functional is unique. Distributions where the functional is unique are termed "unimodal" and rule out, for example, a spherically symmetric uniform distribution. FCH estimators are consistent if the data $\boldsymbol{x_1},\ldots,\boldsymbol{x_n}$ are drawn iid from a "unimodal" EC$_k$($\boldsymbol{\mu}$,$\boldsymbol{\Sigma}$, g) distribution with a nonsingular covariance matrix Cov($\boldsymbol{x}_i$) and g is continuously differentiable with a finite 4th moment (see Olive and Hawkins, (2010)).

Zhang, Olive, and Ye (2012) use five concentration steps. For FCH estimators, selection is based on MCD, but with the necessary condition that for the classical start to be chosen, its concentrated location estimate must lie within the median ball. RMVN uses several reweighting steps. First, if $\left(L_1, C_1\right)$ is the location and scatter estimate selected after concentration, then $L_2 = L_1$ and $C_2 = C_1 \cdot \left[MED\left(D_i^2\left(L_1, C_2\right)\right)\big/\chi^2_{k,0.5}\right]$. Now let $\left(L_3, C_3\right)$ be the classical estimators applied to the n_1 cases satisfying $D_i^2\left(L_2, C_2\right) \le \chi^2_{k,0.975}$. Let $q_1 = \min\left[0.5\left(0.975\right)n/n_1, 0.995\right]$, $L_4 = L_3$, and $C_4 = C_3 \cdot \left[MED\left(D_i^2\left(L_3, C_3\right)\right)\big/\chi^2_{k,q_1}\right]$. Let $\left(L_5, C_5\right)$ be the classical estimators applied to the n_2 cases satisfying $D_i^2\left(L_4, C_4\right) \le \chi^2_{k,0.975}$. Let $q_2 = \min\left[0.5\left(0.975\right)n/n_2, 0.995\right]$, $L_{RMVN} = L_5$, and $C_{RMVN} = C_5 \cdot \left[MED\left(D_i^2\left(L_5, C_5\right)\right)\big/\chi^2_{k,q_1}\right]$. Cases satisfying $D_i^2\left(L_{RMVN}, C_{RMVN}\right) > \chi^2_{k,0.975}$ are identified as outliers.

Median ball (MB)

The median ball estimator is the post-concentration location and scatter estimate $\left(L_1, C_1\right)$ corresponding to the median ball start in RMVN. Cases satisfying $D_i^2\left(L_{MB}, C_{MB}\right) > \chi^2_{k,0.975}$ are identified as outliers.

Refined median ball (RMB)

The refined median ball estimator corresponds to the RMVN estimate, with the restriction that only the median ball start is used. Cases satisfying $D_i^2\left(L_{RMB}, C_{RMB}\right) > \chi^2_{k,0.975}$ are identified as outliers.

Software

We use the Matlab implementation of detMCD and fMCD from LIBRA (available at https://wis.kuleuven.be/statdatascience/robust/software.) We make use of the official Matlab 2016b implementation of RMVN ("Olive-Hawkins method" in *robustcov*), except that we alter the Matlab code to follow the refinement details in David Olive's R implementation (available at http://lagrange.math.siu.edu/Olive/Personal.html). MB and RMB, along with the hybrid and CBUI methods, are implemented in RATS code, available from the authors.

Large-Sample Theory

As noted above, RMVN is the only method backed by large-sample theory. But two things are worth noting. First, nothing is guaranteed in small samples or when UDPs are present. Second, both of the RMVN starts are incorporated into our hybrid method, and an MCD criterion is used to select the final estimator. It is difficult to believe that a start leading to a refined estimator whose covariance determinant is *smaller* than those of the two RMVN starts would lead to *inferior* estimates; putting this differently, the hybrid estimator "should" share the same asymptotic properties as RMVN.

Simulation Evidence

We first use Monte Carlo simulation of data sets to assess reliability. The first simulation exercise examines the size of the various methods, using Gaussian data. The second simulation exercise involves clean or regular data that are contaminated with various data from a different data-generating process (yielding UDPs, or outliers); the forms for clean data and for each outlier configuration are Gaussian and taken from the literature. We compare the ability of the methods to identify the clean data versus the outliers, and to accurately estimate the mean and covariance matrix of the clean data. These tasks are not synonymous, since reasonably accurate outlier detection does not guarantee the accuracy of parameter estimates in small samples. Accuracy is enhanced by appropriate rescaling and reprocessing, once a subset of the data has been selected to be the basis of parameter estimation.

Size, or False Discovery Rate

Following Zhang, Olive, and Ye (2012), size estimation is performed by generating data sets from k-variate normal distributions, where data are iid $N_k(0,\Sigma_k)$ where $\Sigma_k = diag(1,2,3,\ldots,k)$. The estimated size of each rule is the proportion of data points falsely identified as outliers, averaged over 500 simulations. As is conventional in this literature, we use $\alpha = 0.025$. Table 1 is entitled *False Discovery Rate* because each entry indicates the fraction of cases falsely identified as outliers.

The RMB exhibits an extremely small false discovery rate; at the other end of the spectrum is the sCBUI method, which might be referred to as paranoid. (Such paranoia is useful in some contexts, as will be clear below.) The MB also has a very high false discovery rate. RMVN has an edge over detMCD and the hybrid method along this dimension. The false discovery rate of the hybrid method, if equipped with FDR control following Cerioli (2010), is 0.000 (see Appendix). But as our results in the Appendix indicate, when data are contaminated, this control over FDR comes at the cost of missing a substantial fraction of the outliers under some outlier configurations.

False Discovery Rate ($\alpha = 0.025$)

N	k	RMVN	FMCD	detMCD	Hybrid	MB	sCBUI	RMB
100	5	0.043	0.092	0.067	0.087	0.375	0.463	0.002
100	10	0.087	0.181	0.123	0.142	0.427	0.517	0.003
100	20	0.287	0.333	0.358	0.255	0.479	0.749	0.035
200	5	0.031	0.051	0.045	0.053	0.320	0.399	0.000
200	10	0.041	0.067	0.058	0.076	0.353	0.332	0.001
200	20	0.091	0.148	0.122	0.131	0.433	0.728	0.001
1000	5	0.026	0.035	0.035	0.030	0.262	0.333	0.000
1000	10	0.027	0.035	0.034	0.033	0.239	0.167	0.000
1000	20	0.030	0.038	0.037	0.040	0.246	0.148	0.000

Table 1: False Discovery Rate on Clean Data

Detection of UDPs and Accuracy of Parameter Estimates

Simulation design

A simulation study is used to assess the relative performance of the seven methods in a) identifying clean data versus UDPs or outliers, and b) accurately estimating the covariance matrix and mean of the clean data. The simulations use 500 runs, and γ is the total percentage of outliers. The parameter ψ governs the mean of the outlier distribution and, in some cases, the mean of the clean data. Four outlier configurations were taken from the literature. Each configuration specifies the distributions of both clean data and of outliers.

The first and second configurations follow Zhang, Olive, and Ye (2012) and specify clean cases as multivariate normal: $x \sim N_k\left(\mathbf{0}, diag\left(1,2,\ldots,k\right)\right)$. The first configuration specifies outliers as multivariate normal: $x \sim N_k\left(\left(0,\ldots,0,\psi\right)^T, 0.0001I_k\right)$, a near point mass at the major axis. The second configuration specifies outliers as multivariate normal: $x \sim N_k\left(\psi\mathbf{1}, diag\left(1,2,\ldots,k\right)\right)$ where $\mathbf{1} = \left(1,\ldots,1\right)^T$, a mean shift. The near point mass and mean shift outlier configurations are often used in the literature.

The third and fourth configurations are taken from García-Escudero et al. (2008). These involve two groups of outliers (rather than one), and the outlier configurations are considered quite challenging, in that the covariance matrices differ across groups and often display what the authors term “severe overlap”: The clean data are generated from a rather diffuse distribution (as is one of the groups of outliers), so that outlier groups are relatively close (in a statistical sense) to the location of the clean data. This can be seen in Figure 1, which plots the first two dimensions of a simulation with N = 1000 and $\psi = 8$ using the fourth outlier configuration. Indeed, while $\psi = 8$ in all of the simulations in García-Escudero et al. (2008), in the present study this parameter setting was often too difficult for many choices of k and N, and we had to set $\psi = 12$ or $\psi = 16$ in order to observe appreciable performance differences across methods.

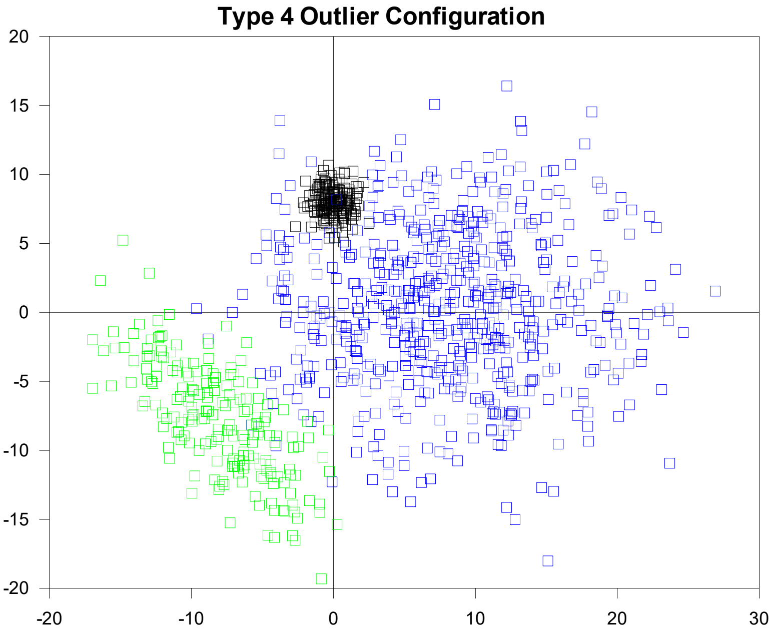


Figure 1: Type 4 Outlier Configuration, taken from García-Escudero et al. (2008)

In these configurations, clean cases are multivariate normal: $x \sim N_k\left((\psi, 0, \ldots, 0,)^T, diag(a, b, 1, \ldots, 1)\right)$; the first "tight cluster" outlier group is multivariate normal: $x \sim N_k\left((0, \psi, 0, \ldots, 0,)^T, \mathbf{I}\right)$; and the second "nonspherical" outlier group is multivariate normal: $x \sim N_k\left((-\psi, -\psi, 0, \ldots, 0,)^T, \Sigma_2\right)$ with

$$\Sigma_2 = \left(\begin{array}{cc|c} 15 & -10 & \\ -10 & 15 & \mathbf{0} \\ \hline & & \\ \multicolumn{2}{c}{\mathbf{0}} & \mathbf{I} \end{array} \right).$$

The constants *a* and *b* control the dispersion of the clean data, hence the differences between the third and fourth configurations. In the third "moderately diffuse" clean data configuration, $(a, b) = (20, 5)$, while in the fourth "very diffuse" clean data configuration, $(a, b) = (45, 30)$. Outlier groups always have equal proportions; that is, each is of proportion $\gamma/2$.

For each outlier configuration, we selected $N \in \{100, 300, 1000\}$, $k \in \{5, 10, 20, 30\}$, and various values of ψ and γ, guided by the choices in the articles cited, but with the additional goal of choosing combinations that would generate performance differences across methods. In each table or figure, each entry is the average of the particular metric over 500 simulation runs.

Identification of clean cases and outliers

To assess each method's ability to identify the clean data versus the outliers, we make use of two very straightforward metrics: the proportion of clean cases identified as clean, and the proportion of outliers identified as outliers (sometimes termed the *detection rate*). These results are presented in Tables 2 and 3. Discussion follows those tables.

Assessing the accuracy of estimates

To assess the accuracy of the estimated scatter matrices, we make use of three metrics. In each simulation run *j*, we compute the classical estimator S_j from the clean cases, and the scatter matrices $\hat{C}_{jr}$ corresponding to each method *r*, and perform two comparisons. First, we computed the sum of the absolute differences of the diagonal elements of S_j and $\hat{C}_{jr}$: $\hat{\sigma}_{jr} = \sum_{\eta=1}^{k} \left| s_{j\eta\eta} - \hat{c}_{jr\eta\eta} \right|$.

For a given simulation configuration, the average over 500 simulations is denoted $\hat{\sigma}_r$. In Figure 2, we plot $\hat{\sigma}_r$ against $\hat{\sigma}_{\text{detMCD}}$ for $r \neq \{\text{detMCD}, \text{fMCD}\}$. In interpreting the figure, it is important to note that three observations corresponding to very large $\widehat{\sigma}_{\text{detMCD}}$ had to be dropped. Inclusion of $\widehat{\sigma}_{\text{fMCD}}$ removes the information content from the figure, as these terms are frequently very large . (The full data underlying this figure are presented in the Appendix.)

While this metric is intuitive, it may not adequately summarize the importance of the differences between two scatter matrices. As pointed out by, for example, Soofi and Dadpay (2002), since many statistical techniques rely on nonlinear functions of estimated covariance matrices, it is useful to investigate error measures that include the inverse and the determinant of the estimated matrix. Hence, we also consider an alternative metric, a likelihood-ratio test of equality of two covariance matrices. To test H_0: $C = S$ for simulation j, we form the test statistic (see Rencher and Christensen (2012), pp. 260-261):

$$u' = \xi u = \xi \nu \left[\ln|S| - \ln|C| + tr\left(CS^{-1}\right) - k \right]$$

where ν is the degrees of freedom of C, and ξ is a first-order Bartlett correction, given by

$$\xi = \left[1 - \frac{1}{6\nu - 1}\left(2k + 1 - \frac{2}{k+1} \right) \right].$$

The test statistic u' is approximately χ^2 distributed with $k(k+1)/2$ degrees of freedom. It is worth noting that u is related to the eigenvalues $\lambda_1, ..., \lambda_k$ of CS^{-1}; in particular,

$$u = \nu \left[\sum_{i=1}^{k} \left(\lambda_i - \ln \lambda_i \right) - k \right].$$

We test $\hat{C}_{jr}$ = S_j for each method r, at the 10 percent level of significance, and record the fraction of simulation runs for which equality is not rejected. These results are presented in Table 4. (In the appendix, we present the corresponding results based on the test $\hat{C}_{jr} = \Sigma_0$).

For location estimates, our metric is $\sum_{\eta=1}^{k} \left| \overline{x}_{j\eta} - \hat{l}_{jr\eta} \right|$ where $\hat{L}_{jr} = \left(\hat{l}_{jr1}, ..., \hat{l}_{jrk} \right)$ is the location estimate for method r for simulation run j. These results are presented in Table 5.

Outlier type	N	k	γ	pm	RMVN	fMCD	detMCD	Hybrid	MB	sCBUI	RMB
Type 1	100	5	0.25	10	0.84	0.53	0.67	0.97	0.76	0.68	0.99
	100	5	0.25	20	0.97	0.52	0.97	0.97	0.76	0.67	0.98
	100	20	0.20	100	0.80	0.44	0.77	0.83	0.64	0.34	0.94
	100	20	0.20	4000	0.80	0.45	0.77	0.83	0.64	0.33	0.94
	300	5	0.25	20	0.98	0.76	0.97	0.99	0.82	0.77	0.99
	300	10	0.25	20	0.97	0.66	0.72	0.98	0.80	0.81	0.99
	300	20	0.20	100	0.96	0.62	0.96	0.95	0.71	0.55	0.99
	300	30	0.20	100	0.92	0.45	0.91	0.92	0.67	0.21	0.98
	300	20	0.20	4000	0.96	0.64	0.95	0.95	0.71	0.56	0.99
	1000	20	0.20	100	0.98	0.81	0.97	0.98	0.83	0.88	0.99
	1000	10	0.10	100	0.97	0.97	0.97	0.98	0.80	0.86	0.99
	1000	30	0.20	100	0.97	0.74	0.97	0.97	0.80	0.72	0.99
Type 2	100	5	0.02	10	0.95	0.91	0.93	0.92	0.63	0.55	1.00
	100	5	0.45	10	0.98	0.98	0.98	0.98	0.92	0.65	0.99
	100	10	0.25	3	0.94	0.88	0.92	0.92	0.70	0.63	0.99
	100	20	0.25	5	0.79	0.71	0.82	0.85	0.68	0.42	0.94
	100	20	0.25	10	0.83	0.73	0.82	0.87	0.68	0.49	0.94
	100	20	0.35	10	0.85	0.68	0.93	0.93	0.78	0.74	0.95
	300	5	0.25	10	0.98	0.97	0.97	0.99	0.82	0.79	0.99
	300	10	0.25	10	0.97	0.97	0.97	0.98	0.80	1.00	0.99
	300	10	0.25	3	0.98	0.97	0.97	0.98	0.80	0.84	0.99
	300	20	0.25	5	0.96	0.95	0.96	0.96	0.74	0.68	0.98
	300	20	0.25	10	0.96	0.96	0.96	0.96	0.74	0.74	0.98
	300	20	0.35	10	0.98	0.93	0.98	0.97	0.81	0.85	0.98
	1000	10	0.25	10	0.98	0.97	0.97	0.99	0.85	0.90	0.99
	1000	10	0.05	10	0.97	0.97	0.97	0.97	0.78	0.85	1.00
Type 3	100	5	0.20	8	0.97	0.95	0.96	0.96	0.72	0.56	0.99
	100	5	0.40	8	0.97	0.94	0.96	0.96	0.81	0.73	0.99
	100	5	0.40	16	0.98	0.98	0.98	0.99	0.87	0.74	0.98
	300	5	0.24	8	0.98	0.98	0.98	0.99	0.81	0.73	0.99
	300	10	0.24	8	0.98	0.97	0.97	0.98	0.79	0.82	0.99
	300	20	0.24	8	0.96	0.95	0.95	0.95	0.73	0.78	1.00
	300	20	0.40	8	0.96	0.95	0.95	0.95	0.74	0.82	0.99
	1000	5	0.40	8	0.98	0.97	0.97	0.98	0.89	0.91	0.99
	1000	5	0.40	12	0.98	0.98	0.98	1.00	0.92	0.91	0.98
	1000	5	0.40	16	0.98	0.98	0.98	1.00	0.92	0.91	0.98
	1000	10	0.40	16	0.98	0.98	0.98	1.00	0.92	0.95	0.98
	1000	5	0.20	16	0.98	0.97	0.97	0.99	0.82	0.85	0.99
	1000	10	0.20	16	0.98	0.97	0.97	0.98	0.83	0.89	0.99
	1000	20	0.40	8	0.98	0.97	0.98	0.98	0.83	0.92	0.99
	1000	20	0.40	12	0.97	0.97	0.97	0.99	0.93	0.91	0.99
	1000	20	0.40	16	0.98	0.97	0.97	0.99	0.91	0.93	0.98
Type 4	100	5	0.40	8	0.96	0.92	0.95	0.92	0.63	0.60	0.99
	100	5	0.24	8	0.97	0.93	0.95	0.93	0.68	0.50	1.00
	100	5	0.40	16	0.97	0.93	0.96	0.97	0.87	0.74	0.99
	300	5	0.24	8	0.98	0.97	0.97	0.97	0.74	0.66	1.00
	300	10	0.24	8	0.97	0.96	0.96	0.96	0.73	0.78	1.00
	300	20	0.40	8	0.95	0.93	0.94	0.93	0.66	0.83	1.00
	1000	5	0.40	8	0.98	0.96	0.97	0.95	0.72	0.75	1.00
	1000	5	0.40	12	0.97	0.96	0.96	0.97	0.77	0.90	0.99
	1000	5	0.40	16	0.97	0.96	0.97	0.98	0.92	0.92	0.99
	1000	5	0.20	16	0.98	0.98	0.98	0.99	0.82	0.80	0.99
	1000	5	0.24	8	0.98	0.97	0.97	0.97	0.77	0.76	1.00
	1000	10	0.40	16	0.98	0.97	0.97	0.99	0.92	0.94	0.99
	1000	10	0.20	16	0.98	0.97	0.97	0.98	0.83	0.92	1.00
	1000	10	0.10	16	0.98	0.97	0.97	0.98	0.80	0.93	1.00
	1000	20	0.40	8	0.97	0.97	0.97	0.97	0.80	0.92	1.00
	1000	20	0.40	16	0.98	0.97	0.97	0.98	0.83	0.81	0.99

Table 2: Percentage of Clean Cases Identified

Note that both sCBUI and MB often miss a substantial percentage of clean cases. Having said that, these methods generally outperform fMCD for the type 1 outlier configuration. Of all the methods examined, RMB has a very slight edge. In a handful of cases, the hybrid method clearly dominates detMCD or RMVN or both; but otherwise these three methods have generally a quite similar performance, and that performance is usually excellent.

	N	*k*	*γ*	pm	RMVN	FMCD	detMCD	Hybrid	MB	sCBUI	RMB
Type 1	100	5	0.25	10	0.49	0.00	0.03	0.84	1.00	1.00	0.67
	100	5	0.25	20	1.00	0.00	1.00	1.00	1.00	1.00	1.00
	100	20	0.2	100	1.00	0.00	1.00	1.00	1.00	1.00	1.00
	100	20	0.2	4000	1.00	0.02	1.00	1.00	1.00	1.00	1.00
	300	5	0.25	20	1.00	0.00	1.00	1.00	1.00	1.00	1.00
	300	10	0.25	20	0.97	0.00	0.00	1.00	1.00	1.00	1.00
	300	20	0.2	100	1.00	0.00	1.00	1.00	1.00	1.00	1.00
	300	30	0.2	100	1.00	0.00	1.00	1.00	1.00	0.91	1.00
	300	20	0.2	4000	1.00	0.06	1.00	1.00	1.00	1.00	1.00
	1000	20	0.2	100	1.00	0.00	1.00	1.00	1.00	1.00	1.00
	1000	10	0.1	100	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	1000	30	0.2	100	1.00	0.00	1.00	1.00	1.00	0.99	1.00
Type 2	100	5	0.02	10	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	100	5	0.45	10	1.00	1.00	1.00	1.00	1.00	0.79	1.00
	100	10	0.25	3	0.45	0.64	0.52	0.70	0.90	0.97	0.29
	100	20	0.25	5	0.80	0.47	0.99	0.91	1.00	1.00	0.99
	100	20	0.25	10	0.99	0.54	1.00	1.00	1.00	1.00	1.00
	100	20	0.35	10	0.87	0.37	1.00	0.95	1.00	1.00	1.00
	300	5	0.25	10	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	300	10	0.25	10	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	300	10	0.25	3	0.39	0.75	0.60	0.86	0.96	0.97	0.17
	300	20	0.25	5	1.00	0.79	1.00	1.00	1.00	1.00	1.00
	300	20	0.25	10	1.00	0.98	1.00	1.00	1.00	1.00	1.00
	300	20	0.35	10	1.00	0.09	1.00	1.00	1.00	1.00	1.00
	1000	10	0.25	10	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	1000	10	0.05	10	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Type 3	100	5	0.2	8	0.81	0.88	0.84	0.88	1.00	1.00	0.74
	100	5	0.4	8	0.50	0.54	0.52	0.62	0.84	0.99	0.49
	100	5	0.4	16	1.00	1.00	1.00	1.00	1.00	0.99	1.00
	300	5	0.24	8	0.70	0.80	0.78	0.87	1.00	1.00	0.64
	300	10	0.24	8	0.52	0.61	0.57	0.78	0.99	1.00	0.50
	300	20	0.24	8	0.35	0.48	0.40	0.61	0.97	0.99	0.28
	300	20	0.4	8	0.36	0.32	0.11	0.46	0.69	0.99	0.37
	1000	5	0.4	8	0.50	0.50	0.51	0.57	0.86	0.97	0.50
	1000	5	0.4	12	1.00	1.00	1.00	1.00	1.00	0.98	1.00
	1000	5	0.4	16	1.00	1.00	1.00	1.00	1.00	0.99	1.00
	1000	10	0.4	16	1.00	0.91	1.00	1.00	1.00	0.92	1.00
	1000	5	0.2	16	1.00	1.00	1.00	1.00	1.00	0.93	1.00
	1000	10	0.2	16	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	1000	20	0.4	8	0.29	0.42	0.12	0.51	0.59	0.98	0.42
	1000	20	0.40	12	0.65	0.51	0.41	0.98	0.98	1.00	0.59
	1000	20	0.4	16	1.00	0.52	0.53	1.00	1.00	1.00	1.00
Type 4	100	5	0.4	8	0.19	0.31	0.15	0.37	0.55	0.93	0.11
	100	5	0.24	8	0.13	0.28	0.19	0.31	0.64	0.93	0.02
	100	5	0.4	16	0.52	0.53	0.53	0.60	0.99	0.97	0.51
	300	5	0.24	8	0.06	0.18	0.13	0.29	0.58	0.91	0.00
	300	10	0.24	8	0.05	0.09	0.07	0.16	0.57	0.82	0.00
	300	20	0.24	8	0.07	0.10	0.09	0.12	0.57	0.65	0.00
	300	20	0.4	8	0.06	0.08	0.07	0.09	0.54	0.57	0.00
	1000	5	0.4	8	0.07	0.29	0.23	0.48	0.53	0.94	0.01
	1000	5	0.4	12	0.51	0.51	0.51	0.50	0.55	0.88	0.50
	1000	5	0.4	16	0.51	0.51	0.51	0.51	1.00	0.85	0.50
	1000	5	0.2	16	0.63	0.74	0.72	0.88	1.00	0.97	0.59
	1000	5	0.24	8	0.04	0.13	0.11	0.30	0.56	0.78	0.00
	1000	10	0.4	16	0.46	0.51	0.51	0.53	0.91	0.72	0.49
	1000	10	0.20	16	0.53	0.57	0.56	0.72	1.00	0.98	0.51
	1000	10	0.10	16	0.60	0.67	0.67	0.78	1.00	0.86	0.51
	1000	20	0.4	8	0.03	0.04	0.04	0.04	0.36	0.51	0.00
	1000	20	0.4	16	0.51	0.51	0.09	0.50	0.59	0.88	0.50

Table 3: Detection Rate: Percentage of Outliers Detected

While sCBUI and MB have difficulty correctly identifying clean cases, the situation is reversed with respect to correctly identifying outliers. For this task, sCBUI offers the best performance, outperforming MB, a method that in turn outperforms all other methods, with only a handful of exceptions. The hybrid method outperforms both RMVN and detMCD. Generally speaking, detMCD has a bit of a performance edge over RMVN, but occasionally it is notably worse; and generally speaking, RMVN outperforms RMB. Finally, once again it is evident that fMCD has difficulty with outlier configuration type 1.

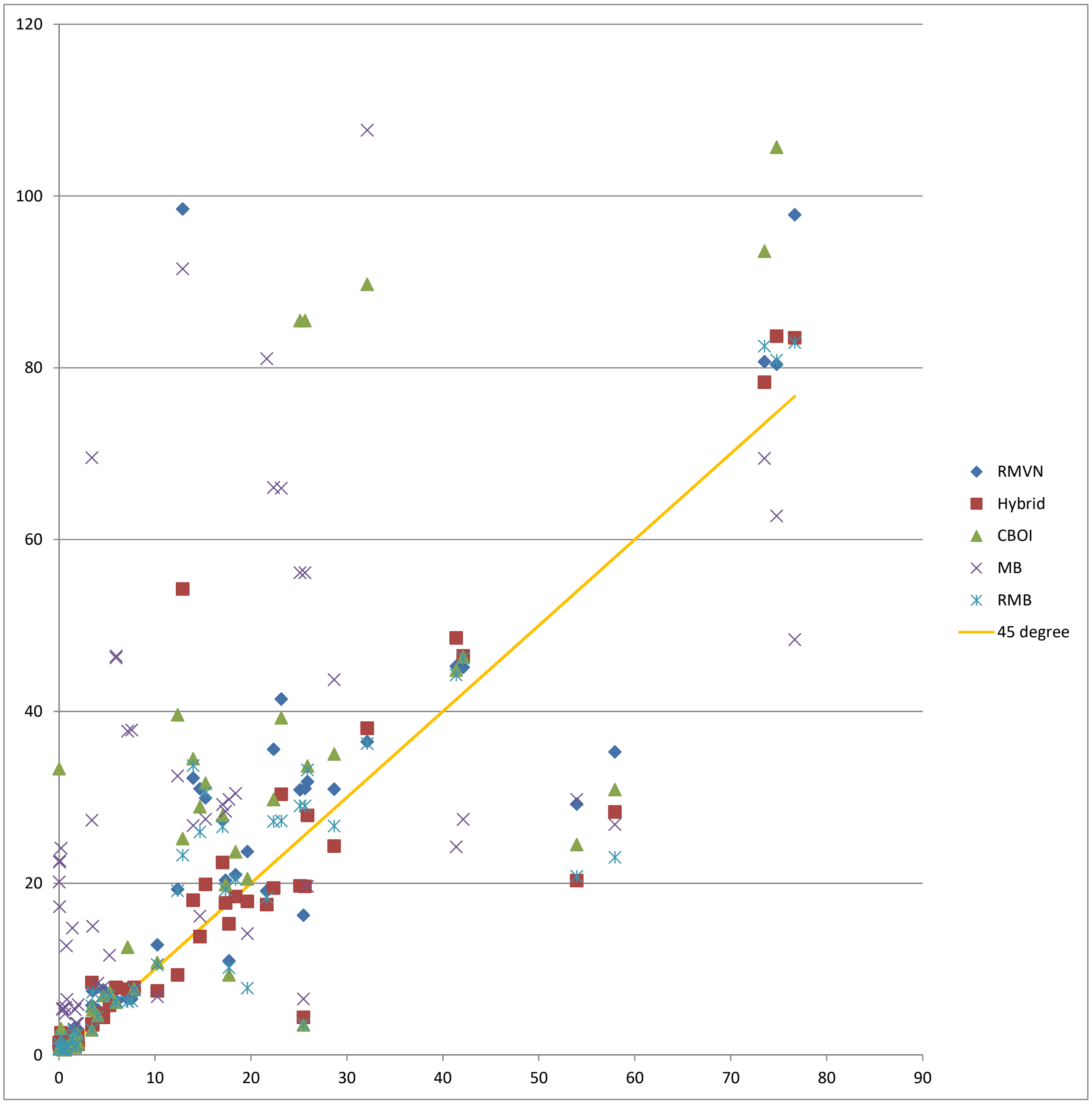


Figure 2: Absolute Differences: *C* - *S*

As noted above, for each $j \neq \{detMCD, fMCD\}$, we plot $\hat{\sigma}_{jr}$ against $\hat{\sigma}_{jdetMCD}$, for each simulation configuration. A 45° line is included for comparison purposes. Three cases in which detMCD returned enormously larger sums are dropped, and in two of the remaining cases, outliers from sCBUI are dropped. The entirety of the data is presented in Table A1 in the Appendix.

Three things are worth noting in Figure 2. First, the performance of sCBUI ("CBOI" in the legend) and MB is generally inferior to that of detMCD, with a handful of exceptions. Second, RMVN, RMB, detMCD, and the hybrid method all offer very similar performance when judged by this metric. (We emphasize again that three large detMCD outliers were dropped.) Third, fMCD cannot even be depicted, as its performance is frequently abysmal according to this metric.

H_0: $C = S$ (at $\alpha = 10\%$)

	N	k	γ	pm	RMVN	FMCD	detMCD	Hybrid	sCBUI	MB	RMB
Type 1	100	5	0.25	10	0.41	0.00	0.01	0.83	0.87	0.00	0.87
	100	5	0.25	20	0.98	0.00	0.99	0.99	0.99	0.00	0.98
	100	20	0.20	100	0.81	0.00	0.60	0.96	0.33	0.00	0.82
	100	20	0.20	4000	0.79	0.02	0.66	0.96	0.33	0.00	0.81
	300	5	0.25	20	1.00	0.00	1.00	1.00	1.00	0.00	1.00
	300	10	0.25	20	0.99	0.00	0.00	1.00	1.00	0.00	1.00
	300	20	0.20	100	1.00	0.00	1.00	1.00	1.00	0.00	1.00
	300	30	0.20	100	1.00	0.00	1.00	1.00	0.89	0.00	1.00
	300	20	0.20	4000	1.00	0.07	1.00	1.00	1.00	0.00	1.00
	1000	20	0.20	100	1.00	0.00	1.00	1.00	1.00	0.00	1.00
	1000	10	0.10	100	1.00	1.00	1.00	1.00	1.00	0.00	1.00
	1000	30	0.20	100	1.00	0.00	1.00	1.00	0.99	0.00	1.00
Type 2	100	5	0.02	10	0.95	0.55	0.83	0.91	0.97	0.01	0.97
	100	5	0.45	10	0.99	1.00	1.00	1.00	0.75	0.00	0.99
	100	10	0.25	3	0.34	0.52	0.36	0.54	0.58	0.00	0.38
	100	20	0.25	5	0.59	0.08	0.95	0.86	0.56	0.00	0.87
	100	20	0.25	10	0.85	0.22	0.98	0.99	0.69	0.00	0.87
	100	20	0.35	10	0.80	0.00	1.00	0.93	0.92	0.00	0.98
	300	5	0.25	10	1.00	1.00	1.00	1.00	1.00	0.00	1.00
	300	10	0.25	10	1.00	1.00	1.00	1.00	1.00	0.00	1.00
	300	10	0.25	3	0.20	0.71	0.43	0.69	0.32	0.00	0.18
	300	20	0.25	5	1.00	0.77	1.00	1.00	0.00	0.00	1.00
	300	20	0.25	10	1.00	0.97	1.00	1.00	0.00	0.00	1.00
	300	20	0.35	10	1.00	0.00	1.00	1.00	1.00	0.00	1.00
	1000	10	0.25	10	1.00	1.00	1.00	1.00	1.00	0.00	1.00
	1000	10	0.05	10	1.00	1.00	1.00	1.00	1.00	0.00	1.00
Type 3	100	5	0.20	8	0.60	0.75	0.70	0.66	0.67	0.00	0.67
	100	5	0.40	8	0.02	0.04	0.02	0.08	0.05	0.00	0.04
	100	5	0.40	16	0.99	1.00	1.00	1.00	0.99	0.00	0.99
	300	5	0.24	8	0.29	0.51	0.43	0.45	0.37	0.00	0.36
	300	10	0.24	8	0.04	0.35	0.17	0.37	0.05	0.00	0.07
	300	20	0.24	8	0.08	0.06	0.03	0.38	0.16	0.00	0.16
	300	20	0.40	8	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	1000	5	0.40	8	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	1000	5	0.40	12	1.00	1.00	1.00	1.00	1.00	0.00	1.00
	1000	5	0.40	16	1.00	1.00	1.00	1.00	1.00	0.00	1.00
	1000	10	0.40	16	1.00	0.82	1.00	1.00	1.00	0.00	1.00
	1000	5	0.20	16	1.00	1.00	1.00	1.00	1.00	0.00	1.00
	1000	10	0.20	16	1.00	1.00	1.00	1.00	1.00	0.00	1.00
	1000	20	0.40	8	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	1000	20	0.40	12	0.17	0.00	0.01	0.93	0.11	0.00	0.12
	1000	20	0.40	16	1.00	0.00	0.03	1.00	1.00	0.00	1.00
Type 4	100	5	0.40	8	0.11	0.12	0.12	0.07	0.07	0.00	0.10
	100	5	0.24	8	0.41	0.37	0.59	0.42	0.36	0.00	0.40
	100	5	0.40	16	0.03	0.03	0.02	0.07	0.06	0.00	0.05
	300	5	0.24	8	0.08	0.37	0.34	0.12	0.06	0.00	0.07
	300	10	0.24	8	0.28	0.54	0.47	0.49	0.22	0.00	0.27
	300	20	0.24	8	0.99	0.90	0.96	0.96	1.00	0.00	1.00
	300	20	0.40	8	0.01	0.00	0.01	0.00	0.02	0.00	0.02
	1000	5	0.40	4	0.10	0.00	0.00	0.09	0.28	0.00	0.21
	1000	5	0.40	8	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	1000	5	0.24	8	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	1000	5	0.40	12	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	1000	5	0.40	16	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	1000	5	0.20	16	0.17	0.25	0.21	0.35	0.19	0.00	0.23
	1000	10	0.40	16	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	1000	10	0.20	16	0.00	0.01	0.00	0.01	0.00	0.00	0.00
	1000	10	0.10	16	0.07	0.71	0.66	0.66	0.07	0.00	0.14
	1000	20	0.40	8	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	1000	20	0.40	16	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 4: Test of Equality of Scatter Matrices, *S* = *C*

(Percentage of Non-Rejections at 10% level)

Once again, it is evident that fMCD experiences great difficulty coping with outlier configuration 1. Outlier configuration type 4 is evidently extremely challenging for all methods. The hybrid method and detMCD are fairly evenly matched for outlier types 2 and 4. The hybrid method has a clear edge for outlier configuration three; in that case, both detMCD and fMCD can get blown away in high dimensions.

Sum of Absolute Differences: *L* versus $\overline{x}$
(expressed as percentage of median across methods)

	N	v	gamma	pm	RMVN	FMCD	detMCD	hybrid	sCBUI	MB	RMB
Type 1	100	5	0.25	10	1.18	2.31	2.01	1.00	0.29	0.43	0.29
	100	5	0.25	20	0.96	39.49	1.00	0.17	1.05	3.94	1.00
	100	20	0.2	100	0.91	10.22	1.00	0.05	1.54	1.32	0.92
	100	20	0.2	4000	0.92	358.08	1.00	0.06	1.49	1.33	0.92
	1000	20	0.2	100	1.01	61.04	0.85	0.62	1.00	4.42	1.00
	1000	10	0.1	100	1.00	0.98	0.97	0.82	1.02	4.98	1.03
	300	5	0.25	20	0.99	57.18	0.68	0.43	1.01	4.33	1.00
	300	10	0.25	20	1.00	19.67	18.59	0.39	0.79	3.16	0.76
	300	20	0.2	100	0.99	31.89	0.98	0.30	1.09	3.30	1.00
	300	30	0.2	100	0.97	15.25	1.00	0.11	2.32	2.29	0.96
	300	20	0.2	4000	1.00	1091.00	1.01	0.31	0.99	3.25	0.98
	1000	30	0.2	100	0.99	35.23	0.87	0.51	1.26	4.27	1.00
Type 2	100	5	0.02	10	1.00	1.72	1.34	0.44	1.00	3.46	0.94
	100	5	0.45	10	1.36	0.81	0.05	0.00	40.33	3.41	1.00
	100	20	0.25	5	2.00	4.40	0.75	1.00	0.91	1.11	0.70
	100	20	0.25	10	1.18	9.92	0.91	0.07	1.00	1.53	0.91
	100	20	0.35	10	3.15	13.63	0.45	1.00	0.59	1.01	0.56
	100	10	0.25	3	1.26	1.00	1.09	0.85	0.65	0.72	1.05
	1000	10	0.25	10	1.00	0.65	0.64	0.61	1.01	4.46	1.02
	1000	10	0.05	10	0.99	1.04	1.03	0.85	1.00	5.06	1.00
	300	5	0.25	10	1.00	0.76	0.73	0.44	1.01	4.33	1.00
	300	10	0.25	10	1.00	0.89	0.84	0.43	1.01	4.19	1.01
	300	10	0.25	3	1.59	0.59	0.99	1.00	1.05	0.61	1.50
	300	20	0.25	5	1.06	7.17	0.93	0.31	1.00	3.50	1.00
	300	20	0.25	10	1.00	2.48	0.93	0.24	1.00	3.63	1.00
	300	20	0.35	10	0.99	94.33	0.74	0.16	1.00	3.66	1.00
Type 3	100	5	0.2	8	1.12	0.88	1.00	2.23	0.98	1.52	0.95
	100	5	0.4	8	1.02	1.00	1.02	1.08	0.94	0.55	0.98
	100	5	0.4	16	1.00	0.49	0.31	0.14	1.32	4.42	1.06
	300	5	0.24	8	1.08	0.64	0.73	1.05	1.00	0.56	1.00
	300	10	0.24	8	1.10	0.75	0.90	1.01	1.00	0.44	1.00
	300	20	0.24	8	1.04	1.00	1.04	1.10	0.87	0.52	0.87
	300	20	0.4	8	1.01	1.03	1.10	1.00	0.96	0.84	0.95
	1000	5	0.4	8	1.02	0.98	1.02	0.90	1.00	0.33	1.00
	1000	5	0.4	12	1.00	0.17	0.19	0.25	1.05	3.04	1.05
	1000	5	0.4	16	1.00	0.10	0.09	0.23	1.01	2.93	1.01
	1000	10	0.4	16	1.00	13.98	0.17	0.31	1.00	2.92	1.00
	1000	5	0.2	16	1.00	0.74	0.73	0.69	1.02	4.81	1.03
	1000	10	0.2	16	1.00	0.77	0.76	0.69	1.04	4.64	1.05
	1000	20	0.4	8	1.04	1.01	1.14	0.99	1.00	0.98	0.99
	1000	20	0.4	12	1.00	2.11	2.20	0.31	1.06	0.16	0.98
	1000	20	0.4	16	1.00	52.78	52.19	0.38	0.99	3.08	0.99
	1000	5	0.4	4	1.04	0.99	1.00	0.00	0.00	0.00	0.00
Type 4	100	5	0.4	8	1.00	1.00	1.04	0.97	0.99	1.48	0.97
	100	5	0.24	8	1.01	1.00	1.00	0.98	1.00	1.17	0.99
	100	5	0.4	16	1.01	1.06	1.06	0.95	1.00	0.17	0.98
	300	5	0.24	8	1.03	0.94	0.97	0.96	1.01	1.22	1.00
	300	10	0.24	8	1.01	0.98	1.00	0.98	1.00	1.09	1.00
	300	20	0.4	8	1.00	1.01	1.01	0.99	0.98	1.26	0.98
	1000	5	0.4	4	0.90	1.05	1.00	1.01	0.88	1.60	0.90
	1000	5	0.4	8	1.05	0.98	1.00	0.96	1.06	1.51	0.99
	1000	5	0.24	8	1.02	0.93	0.94	0.90	1.01	1.23	1.00
	1000	5	0.4	16	1.00	1.01	1.01	0.99	1.03	0.04	1.00
	1000	5	0.2	16	1.68	0.93	1.00	1.00	1.57	0.41	1.41
	1000	10	0.1	16	1.27	0.83	0.87	1.00	1.25	0.66	1.19
	1000	10	0.2	16	1.32	0.91	0.96	1.00	1.33	0.25	1.32
	1000	10	0.4	16	1.01	1.00	1.00	0.97	1.14	0.10	1.00
	1000	20	0.4	8	1.00	1.00	1.00	1.00	0.99	1.10	0.99
	1000	20	0.4	16	1.00	1.00	1.17	1.00	1.09	0.87	1.00

Table 5: Absolute Differences between Location Estimates: L/L_{median}

Each difference is recorded relative to the median across methods for each simulation configuration.

When considering the accurate estimation of the mean, the hybrid method offers the best overall performance: Its estimate is either close to the median or else achieves the minimum deviation, with rare exceptions (such as the first configuration for the type 3 outlier configuration). Also notable is RMB, which is very consistent and *never* far above the median. It is also worth noting that detMCD is also very consistent, with one big miss, and sCBUI is also reasonably accurate (though it has two big misses). RMVN is not far behind.

In sum, simulation evidence leads us to the following conclusions about relative performance. Overall, the hybrid method offers performance that is on a par with, or modestly superior to, detMCD, which in turn has a slight edge over RMVN. Each of these three methods is attractive: Each is practical to implement, each easily outperforms classical methods, and the overall performance of these estimators appears to be comparable. Conversely, fMCD should not be considered a reliable method. If the task is outlier detection, sCBUI is the best, followed by MB. If the task is accurate scatter estimation, both hybrid and detMCD are generally good choices, though RMVN is nearly on par; MB and sCBUI offer sub-par performance for this task. If the task is the accurate estimation of location, the hybrid method or RMB is perhaps the best choice.

REAL-WORLD DATA ANALYSIS

In this section we first apply our new methods to some data sets that have been used as benchmarks in the literature. Finally, we re-examine a prominent economic study. We ask the question: Are the results in this study driven by unusual cases?

SWISS BANKNOTES DATA

Flury and Riedwyl (1988) introduced a Swiss forgeries data set that has become a standard benchmark in the field for cluster analysis and for principal components analysis (see, for example, Salibian-Barrera, Van Aelst, and Willems (2006)), and a standard example for logistic regression (see Olive 2012). But these data are not generally used to assess outlier detection methods, for reasons that will become clear. (An exception is Atkinson, Riani, and Cerioli (2004), who use it to demonstrate the forward search, starting from 20 notes believed to be genuine. This analysis finds two distinct groups of forgeries.) Flury and Riedwyl investigated whether one could use simple linear dimensions obtained from genuine versus forged 1000-franc Swiss banknotes to enable the detection of other forgeries without the use of sophisticated equipment. They obtained 200 Swiss banknotes withdrawn from circulation, which experts classified as either genuine or forged. They took six measurements on each note: length, height on left side, height on right side, width of bottom margin, width of top margin, and diagonal length of the inner frame. In this data set, the first 100 observations are believed to be genuine, while the next 100 observations were identified as forgeries. However, some of the notes in either category may have been misclassified. Ritter (2014) remarks that, although small, this data set is “not easy.” For instance, the forged notes may not have all been produced in the same manner, and, by construction, outliers comprise 50 percent of the data.

sCBUI identifies all 100 forgeries, in addition to flagging 14 notes from the genuine pool as being outliers. But this data set is “not easy.” RMVN locates just 15 forgeries (with an additional 7 notes from the genuine pool identified as outliers), and the hybrid method locates only 16 forgeries (with an additional 9 notes from the genuine pool identified as outliers).

BUXTON HEIGHT DATA

Buxton (1920, p. 232-35) gives various measurements on 591 men, including height, head length, nasal height, bigonial breadth, and cephalic index. The fourth group of observations, taken from men in proximity to the village of Levkoniko on Cyprus, consists of 88 cases; we remove one of these cases because of missing values. Five of these individuals appear to be somewhat unusual, in that their heights were recorded to be about 19mm, in conjunction with massive head lengths that exceeded five feet. These five cases are outliers with enormous leverage, as will be evident below. For illustrative purposes, we follow Olive (2017) and predict stature on the basis of an intercept and four variables: head length, nasal height, bigonial breadth, and cephalic index.

	N	Constant	Head Length	Nasal Height	Bigonial Breadth	Cephalic Index	$\bar{R}^2$
OLS: Full Data	87	1546 (8.00)	-1.12 (-58.9)	6.11 (4.01)	-0.59 (-0.54)	1.13 (0.74)	0.98
OLS: Remove Outliers	76	808 (1.70)	1.69 (1.04)	4.83 (3.10)	0.15 (0.10)	3.75 (1.60)	0.10

Table 6. OLS and Robust Regression Results, Buxton Height Data

When the five cases are included in the regression, $\bar{R}^2$ is 98 percent, suggesting an excellent fit to the data. Estimates suggest that head length is a powerful predictor; men with longer heads are evidently shorter, on average. Residuals appear to be well-behaved (see Figure 3). Masking is on display: Cases 61-65 are the tiny men outliers, and they are invisible in Figure 3.

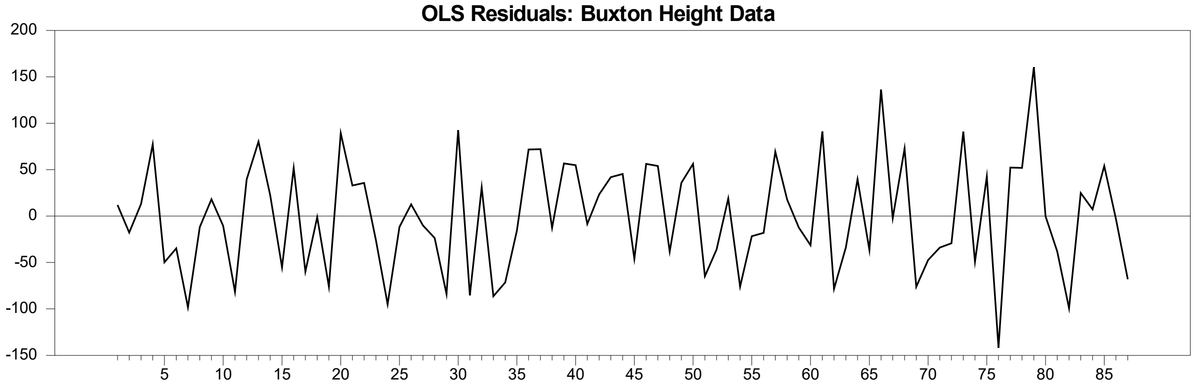


Figure 3: OLS Residuals from Buxton Height Data

Only three cases are identified as outliers according to conventional studentized residual methods (either internal or external) – and *none* of these is one of the five “tiny men” cases! Cook’s *d* fares slightly better, but still only correctly identifies two of the five tiny men cases. When these two cases are removed, regression results are nearly unaffected. Our hybrid method identifies 11 cases as outliers, including all five of the tiny men. (The other six are men with extremely short noses or narrow jaws, in conjunction with unusual height.) When these cases are removed, the estimated coefficient on head length switches sign and loses statistical significance. The regression $\bar{R}^2$ drops to 10 percent.

REEXAMINATION OF COSTINOT, OLDENSKI, AND RAUCH (2011)

In this subsection, we re-examine a prominent study in the economics literature.[10] We seek to determine whether central results in this study are sensitive to, or perhaps driven by, unusual observations. If a group of UDPs appear to play a central role in coefficient estimates, they merit further investigation.[11] Our intention is not to sully these authors. We have great respect for them, as they have demonstrated their commitment to impartial scientific analysis by making their data and code available. These researchers were undoubtedly unaware of the existence of tools like the ones presented here.

Cosinit, Oldenski, and Rauch (2011) (COR) rank 77 goods-producing industries on the basis of the routineness of tasks involved in the production process. Their subsequent regression analysis suggests that, after controlling for various other sector-level characteristics, the degree of sector routineness is a significant predictor of the sector share of intrafirm imports; specifically, the higher the routineness of the sector, the lower the share of intrafirm imports. Sector-level regression controls include capital intensity, the log of the ratio of capital to labor; skill intensity, the log of the ratio of nonproduction workers to production workers in a given industry; R&D intensity, the log of the ratio of R&D spending to sales; relationship specificity, the importance of relationship-specific investments; intermediation, the industry-specific ease of contracting out parts of the production process; and dispersion, the distribution of firm size within an industry on the basis of variation of sales by firms.

We find that the results in the paper are not robust. Our method indicates that a significant percentage of these data is drawn from a different distribution than the bulk of the data. In the baseline regressions, after removing these unusual observations using our hybrid method, we find uniformly weaker evidence that routineness plays an important role. Across specifications, coefficient magnitudes decline, and statistical significance often disappears. Furthermore, once we include all the relevant sector controls, the evidence for an important role of dispersion strengthens considerably, at the expense of the routineness and R&D intensity variables: Coefficient estimates on both are no longer statistically different from zero. We emphasize that these stark differences hinge on 7,000 observations that are drawn from a different distribution than the bulk of the data.

[10] In a previous version, we also examined Card and Krueger (1994) – where we find that the result is driven by UDPs (verifying a finding in Neumark and Wascher (2000)) – and Mian and Sufi (2014), where we find that the result is strengthened if we remove UDPs.

[11] We again draw attention to Knez and Ready (1997), who investigate the effects of outliers in the famous Fama and French (1992, 1993) studies. These authors state: “We find that the risk premium on size that was estimated by Fama and French (1992) completely disappears when the 1 percent most extreme observations are trimmed each month. We also show that the negative average of the monthly size coefficients reported by Fama and French can be entirely explained by the 16 months with the most extreme coefficients.”

Table 7 - Baseline Regressions

Model:	1		2		3		4		5	
	Original	Robust	Original	Robust	Original	Robust	Original	Robust	Original	Robust
N:	29645	20600	29645	20034	29645	19612	29645	20546	27775	20707
Routine	-0.1826***	**-0.0749****	-0.0829**	-0.0568	-0.0858**	-0.0611	-0.0903***	**-0.0895****	-0.0829**	-0.0293
	(-6.75)	(-2.243)	(-2.21)	(-1.205)	(-2.47)	(-1.266)	(-2.59)	(-2.088)	(-2.48)	(-0.652)
Ln(K/L)			0.0117	**-0.1038*****	0.0576*	**-0.1289****	0.0703*	-0.0717	0.0645*	-0.0301
			(0.38)	(-2.669)	(1.66)	(-2.017)	(1.75)	(-1.207)	(1.65)	(-0.467)
Ln(S/L)			0.0160	0.0437	0.0026	0.0361	0.0048	0.0662	-0.0242	-0.0097
			(0.42)	(1.078)	(0.08)	(0.770)	(0.13)	(1.485)	(-0.67)	(-0.213)
Ln(R&D)			0.1646***	0.0479	0.1270***	0.0547	0.1357***	0.0791	0.1105***	0.0637
			(4.22)	(1.073)	(2.88)	(1.034)	(3.06)	(1.505)	(2.70)	(1.215)
Specificity					0.0816**	-0.0419	0.0838**	-0.0090	0.0673	0.0174
					(2.17)	(-0.687)	(2.13)	(-0.154)	(1.63)	(0.250)
Intermediation							0.0324	**0.1134****	0.0151	0.012
							(0.88)	(2.587)	(0.41)	(0.220)
Dispersion									0.0730*	**0.1578*****
									(1.92)	(4.047)
Fixed effects	Ctry-year		Ctry-year		Ctry-year		Ctry-year		Ctry-year	
R^2	0.261	0.275	0.281	0.277	0.285	0.287	0.286	0.311	0.292	0.345

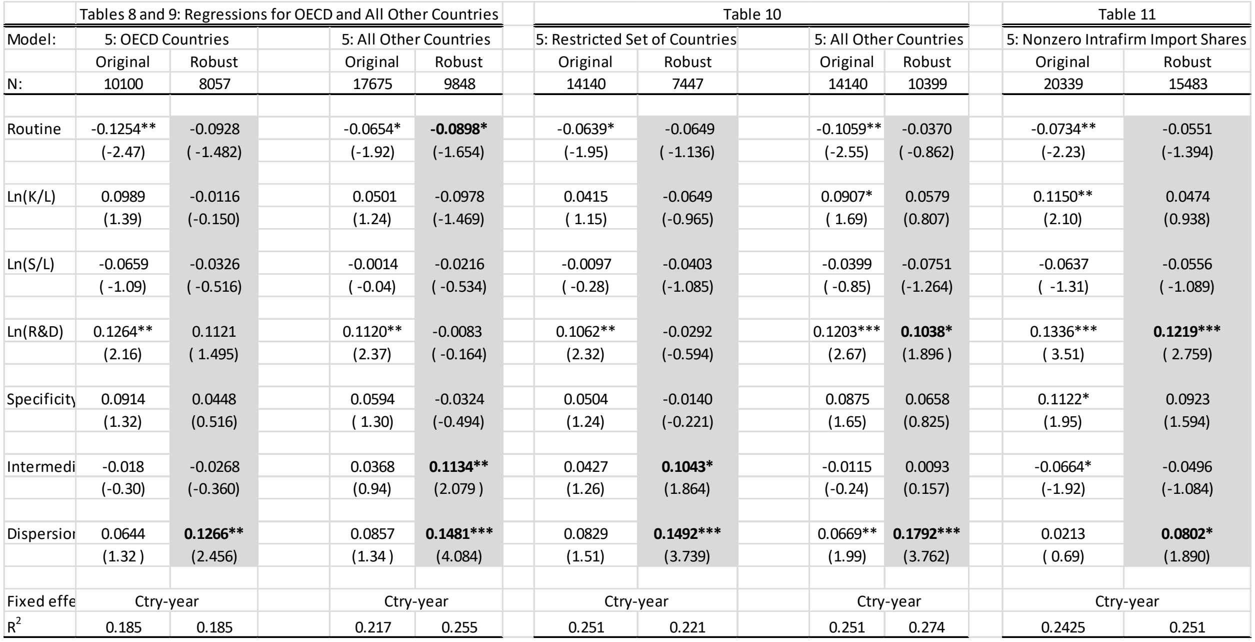
Table 7: Costinot, Oldenski, and Rauch Table 7, With and Without UDPs

To probe this result further, we repeat the robustness exercises of COR for model 5, which includes all the controls. These checks include re-running the regression on four different subgroups of countries: OECD and non-OECD countries, as well as countries where "at least two-thirds of intrafirm U.S. imports from that country are imported by U.S.-owned firms" (what the authors call the "restricted set of countries"), and countries where that share is less than two-thirds (what we will refer to as the "unrestricted set of countries").[12] After removing unusual observations, we find that routineness is a significant predictor of intrafirm imports only in the non-OECD country group, whereas dispersion remains strongly significant across all subgroups. This of course contrasts sharply with the finding of COR: that the coefficient on routineness is statistically significant across all four subgroups. Further, this finding removes support for the authors' theoretical framework, the basis of which is the behavior of intrafirm imports specifically by U.S.-owned multinationals, which the restricted set of countries is intended to capture.

Finally, as in COR, model 5 is re-run including only firms with nonzero intrafirm import share. Here too we fail to find any strong evidence in favor of the effect of routineness after removing unusual observations. Meanwhile, dispersion still plays a significant, albeit substantially smaller role, than in our previous results. This regression also suggests that this group of firms

[12] We note that Costinot, Oldenski, and Rauch (2011) do not report regression results for the unrestricted countries. We report them here, along with the results after removing data outliers using our hybrid method.

drives the positive and significant impact of R&D intensity on intrafirm imports that COR found across all of the regressions in their original work.



	Tables 8 and 9: Regressions for OECD and All Other Countries				Table 10				Table 11	
Model:	5: OECD Countries		5: All Other Countries		5: Restricted Set of Countries		5: All Other Countries		5: Nonzero Intrafirm Import Shares	
	Original	Robust	Original	Robust	Original	Robust	Original	Robust	Original	Robust
N:	10100	8057	17675	9848	14140	7447	14140	10399	20339	15483
Routine	-0.1254**	-0.0928	-0.0654*	**-0.0898***	-0.0639*	-0.0649	-0.1059**	-0.0370	-0.0734**	-0.0551
	(-2.47)	(-1.482)	(-1.92)	(-1.654)	(-1.95)	(-1.136)	(-2.55)	(-0.862)	(-2.23)	(-1.394)
Ln(K/L)	0.0989	-0.0116	0.0501	-0.0978	0.0415	-0.0649	0.0907*	0.0579	0.1150**	0.0474
	(1.39)	(-0.150)	(1.24)	(-1.469)	(1.15)	(-0.965)	(1.69)	(0.807)	(2.10)	(0.938)
Ln(S/L)	-0.0659	-0.0326	-0.0014	-0.0216	-0.0097	-0.0403	-0.0399	-0.0751	-0.0637	-0.0556
	(-1.09)	(-0.516)	(-0.04)	(-0.534)	(-0.28)	(-1.085)	(-0.85)	(-1.264)	(-1.31)	(-1.089)
Ln(R&D)	0.1264**	0.1121	0.1120**	-0.0083	0.1062**	-0.0292	0.1203***	**0.1038***	0.1336***	**0.1219*****
	(2.16)	(1.495)	(2.37)	(-0.164)	(2.32)	(-0.594)	(2.67)	(1.896)	(3.51)	(2.759)
Specificity	0.0914	0.0448	0.0594	-0.0324	0.0504	-0.0140	0.0875	0.0658	0.1122*	0.0923
	(1.32)	(0.516)	(1.30)	(-0.494)	(1.24)	(-0.221)	(1.65)	(0.825)	(1.95)	(1.594)
Intermedi	-0.018	-0.0268	0.0368	**0.1134****	0.0427	**0.1043***	-0.0115	0.0093	-0.0664*	-0.0496
	(-0.30)	(-0.360)	(0.94)	(2.079)	(1.26)	(1.864)	(-0.24)	(0.157)	(-1.92)	(-1.084)
Dispersion	0.0644	**0.1266****	0.0857	**0.1481*****	0.0829	**0.1492*****	0.0669**	**0.1792*****	0.0213	**0.0802***
	(1.32)	(2.456)	(1.34)	(4.084)	(1.51)	(3.739)	(1.99)	(3.762)	(0.69)	(1.890)
Fixed effe	Ctry-year		Ctry-year		Ctry-year		Ctry-year		Ctry-year	
R^2	0.185	0.185	0.217	0.255	0.251	0.221	0.251	0.274	0.2425	0.251

Table 8: Costinot, Oldenski, and Rauch Tables 8-11, With and Without Outliers

We again emphasize that we have great respect for these authors, and we suspect that it would not be difficult to discover that the findings of many prominent studies were driven by UDPs. This is a cautionary note for all of us. Indeed, we ourselves have not always made use of the best available methods, to see if our own results are robust!

Conclusion

We present two new methods for multivariate outlier identification and robust estimation of multivariate location and dispersion. We provide evidence indicating that these methods perform on a par with, or better than, two of the currently best available methods. We also demonstrate, by re-examining a prominent economic study, that results can be sensitive to a modest percentage of atypical cases in the data. This is information worth knowing. A finding like this will challenge the conclusions drawn from the full sample and force the researcher to investigate further, to determine what characterizes the atypical cases, and to reach deeper conclusions about the relationships being studied. We therefore suggest that researchers should routinely run a quick check to see if the results are robust along this dimension. Our tools make this straightforward to accomplish.

Many methods, such as factor analysis, directly or indirectly rely on an accurate estimate of the covariance matrix. These may benefit from a step that undertakes a robust estimation of the covariance matrix using the hybrid method.

This study has not addressed special topics related to outliers that arise in the time series context. While outliers have long been a central issue in seasonal adjustment (see, for example, Findlay et al. (1998)) and in estimating inflation trends (see, for example, Bryan and Cecchetti (1994) or Higgins and Verbrugge (2015)), they are often ignored in time series analysis more broadly – despite the fact that outliers will have the same distorting effects on classical estimators in that context. We do, however, note that both LTS-based and MM-based robust estimation procedures for vector autoregressions have been developed; see Croux and Joossens (2008) and Muler and Yohai (2013).

All of the methods in this study apply only to continuous variables. However, the attributes in a data set are often a mixture of categorical and continuous types. Categorical attributes generally take on few values, and these values may not have an ordering. This makes it difficult to define distance metrics for such data points. Work on this area has begun (see, for example, Otey, Ghoting, and Parthasarathy (2006)), but we leave the development of meaningful distance metrics in mixed-type attribute spaces for future work.

References

Atkinson, Anthony C., Marco Riani, and Andrea Cerioli. 2004. *Exploring Multivariate Data with the Forward Search*. Springer Series in Statistics. New York, NY: Springer New York. https://doi.org/10.1007/978-0-387-21840-3.

Bansal, Neeraj, and Amit Chugh. 2013. "Differentiate Clustering Approaches for Outlier Detection." *International Journal of Innovative Research in Computer and Communication Engineering* 1 (2): 193-196.

Bassett, Gilbert W. 1991. "Equivariant, Monotonic, 50% Breakdown Estimators." *The American Statistician* 45 (2): 135–37. https://doi.org/10.1080/00031305.1991.10475787.

Bhaduri, Kanishka, Bryan L. Matthews, and Chris R. Giannella. 2011. "Algorithms for Speeding up Distance-Based Outlier Detection." In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 859–867. KDD '11. San Diego, California, USA: Association for Computing Machinery. https://doi.org/10.1145/2020408.2020554.

Billor, Nedret, Ali S. Hadi, and Paul F. Velleman. 2000. "BACON: Blocked Adaptive Computationally Efficient Outlier Nominators." *Computational Statistics & Data Analysis* 34 (3): 279–98. https://doi.org/10.1016/S0167-9473(99)00101-2.

Blankmeyer, Eric. 2016. "Robust Regression When the True R^2 Is Mediocre." Texas State University. https://doi.org/10.2139/ssrn.2273737.

Bryan, Michael F., and Stephen G. Cecchetti. 1994. "Measuring Core Inflation." In *Monetary Policy*, edited by N. Gregory Mankiw, 195–219. Studies in Business Cycles. National Bureau of Economic Research. https://ideas.repec.org/h/nbr/nberch/8333.html.

Buxton, Leonard H. Dudley. 1920. "The Anthropology of Cyprus." *The Journal of the Royal Anthropological Institute of Great Britain and Ireland* 50 (January): 183–235. https://doi.org/10.2307/2843379.

Card, David, and Alan B. Krueger. 1994. "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania." *The American Economic Review* 84 (4): 772–93. https://www.jstor.org/stable/2118030.

Cator, Eric A., and Hendrik P. Lopuhaä. 2010. "Asymptotic Expansion of the Minimum Covariance Determinant Estimators." *Journal of Multivariate Analysis* 101 (10): 2372–2388. https://doi.org/10.1016/j.jmva.2010.06.009.

Cator, Eric A., and Hendrik P. Lopuhaä. 2012. "Central Limit Theorem and Influence Function for the MCD Estimators at General Multivariate Distributions." *Bernoulli* 18 (2): 520–551. https://doi.org/10.3150/11-BEJ353.

Cerioli, Andrea. 2010. "Multivariate Outlier Detection With High-Breakdown Estimators." *Journal of the American Statistical Association* 105 (489): 147–56. https://doi.org/10.1198/jasa.2009.tm09147.

Chakhchoukh, Yacine. 2010. “A New Robust Estimation Method for ARMA Models.” *IEEE Transactions on Signal Processing* 58 (7): 3512–22. https://doi.org/10.1109/TSP.2010.2046413.

Costinot, Arnaud, Lindsay Oldenski, and James Rauch. 2011. “Adaptation and the Boundary of Multinational Firms.” *The Review of Economics and Statistics* 93 (1): 298–308. https://doi.org/10.7910/DVN/ZVSJWQ.

Croux, Christophe, and Gentiane Haesbroeck. 1999. “Influence Function and Efficiency of the Minimum Covariance Determinant Scatter Matrix Estimator.” *Journal of Multivariate Analysis* 71 (2): 161-190. https://doi.org/10.1006/jmva.1999.1839.

Croux, Christophe, and Kristel Joossens. 2008. “Robust Estimation of the Vector Autoregressive Model by a Least Trimmed Squares Procedure.” In *COMPSTAT 2008*, edited by Paula Brito, 489–501. Heidelberg: Physica-Verlag HD. https://doi.org/10.1007/978-3-7908-2084-3_40.

Croux, Christophe, and Stefan Van Aelst. 2002. “Comment on ‘Nearest-Neighbor Variance Estimation (NNVE): Robust Covariance Estimation via Nearest-Neighbor Cleaning.’” *Journal of the American Statistical Association* 97 (460): 1006–9.

Fama, Eugene F., and Kenneth R. French. 1992. “The Cross-Section of Expected Stock Returns.” *Journal of Finance* 47 (2): 427–265. https://doi.org/10.1111/j.1540-6261.1992.tb04398.x.

Fama, Eugene F., and Kenneth R. French. 1993. “Common Risk Factors in the Returns on Stocks and Bonds.” *Journal of Financial Economics* 33 (1): 3–56. https://doi.org/10.1016/0304-405x(93)90023-5.

Findley, David F., Brian C. Monsell, William R. Bell, Mark C. Otto, and Bor-Chung Chen. 1998. “New Capabilities and Methods of the X-12-ARIMA Seasonal-Adjustment Program.” *Journal of Business & Economic Statistics* 16 (2): 127–77. https://doi.org/10.2307/1392565.

Flury, Bernhard, and Hans Riedwyl. 1988. *Multivariate Statistics: A Practical Approach*. London: Chapman & Hall, Ltd.

García-Escudero, Luis A., Alfonso Gordaliza, Carlos Matrán, and Agustin Mayo-Iscar. 2008. “A General Trimming Approach to Robust Cluster Analysis.” *The Annals of Statistics* 36 (3): 1324–45. https://doi.org/10.1214/07-AOS515.

Ghoting, Amol, Srinivasan Parthasarathy, and Matthew Eric Otey. 2006. “Fast Mining of Distance-Based Outliers in High-Dimensional Datasets.” In *Proceedings of the 2006 SIAM International Conference on Data Mining*, 609–13. Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611972764.70.

Gnanadesikan, Ramanathan, and Jon R. Kettenring. 1972. “Robust Estimates, Residuals, and Outlier Detection with Multiresponse Data.” Biometrics 28(1), 81–124.

Green, Christopher G., and Doug Martin. 2014. “An Extension of a Method of Hardin and Rocke, with an Application to Multivariate Outlier Detection via the IRMCD Method of

Cerioli." Working Paper. University of Washington. https://christophergreen.github.io/papers/hr05_extension.pdf.

Hadi, Ali S., A. H. M. Rahmatullah Imon, and Mark Werner. 2009. "Detection of Outliers." *Wiley Interdisciplinary Reviews: Computational Statistics* 1 (1): 57–70. https://doi.org/10.1002/wics.6.

Hampel, Frank R. 1975. "Beyond Location Parameters: Robust Concepts and Methods." *Bulletin of the International Statistical Institute* 46: 375–82.

Hampel, Frank R., Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. 1986. *Robust Statistics: The Approach Based on Influence Functions*. Wiley Series in Probability and Mathematical Statistics. New York: Wiley.

Hardin, Johanna, and David M. Rocke. 2005. "The Distribution of Robust Distances." *Journal of Computational and Graphical Statistics* 14 (4): 928–46. https://doi.org/10.1198/106186005X77685.

Hawkins, Douglas M., and David J. Olive. 2002. "Inconsistency of Resampling Algorithms for High-Breakdown Regression Estimators and a New Algorithm." *Journal of the American Statistical Association* 97 (457): 136–59. https://doi.org/10.1198/016214502753479293.

Higgins, Amy, and Randal J. Verbrugge. 2015. "Is a Nonseasonally Adjusted Median CPI a Useful Signal of Trend Inflation?" *Economic Commentary (Federal Reserve Bank of Cleveland)*, November, 1–6. https://doi.org/10.26509/frbc-ec-201513.

Hössjer, Ola. 1994. "Rank-Based Estimates in the Linear Model with High Breakdown Point." *Journal of the American Statistical Association* 89 (425): 149–58. https://doi.org/10.1080/01621459.1994.10476456.

Hubert, Mia, Peter J. Rousseeuw, and Tim Verdonck. 2012. "A Deterministic Algorithm for Robust Location and Scatter." *Journal of Computational and Graphical Statistics* 21 (3): 618–37. https://doi.org/10.1080/10618600.2012.672100.

Janson, Wesley, and Randal J. Verbrugge. 2020. "Improving Inference via Data-Based Identification of Unobserved Variables and Unusual Data Points." Manuscript in preparation, Federal Reserve Bank of Cleveland.

Knez, Peter J., and Mark J. Ready. 1997. "On The Robustness of Size and Book-to-Market in Cross-Sectional Regressions." *The Journal of Finance* 52 (4): 1355–82. https://doi.org/10.1111/j.1540-6261.1997.tb01113.x.

Maronna, Ricardo A., R. Douglas Martin, and Víctor J. Yohai. 2006. *Robust Statistics: Theory and Methods*. 1st ed. Wiley Series in Probability and Statistics. Wiley. https://doi.org/10.1002/0470010940.

Maronna, Ricardo A., and Zamar, Ruben H. 2002. "Robust Estimates of Location and Dispersion for High-Dimensional Datasets." *Technometrics* 44, 307–317. https://doi.org/10.1198/004017002188618509.

Mian, Atif, and Amir Sufi. 2014. "What Explains the 2007-2009 Drop in Employment?" Econometrica 82 (6), 2197–2223. https://doi.org/10.3982/ECTA10451.

Muler, Nora, and Victor J. Yohai. 2013. “Robust Estimation for Vector Autoregressive Models.” *Computational Statistics & Data Analysis* 65 (September): 68–79. https://doi.org/10.1016/j.csda.2012.02.011.

Müller, Ulrich K. 2007. “A Theory of Robust Long-Run Variance Estimation.” *Journal of Econometrics* 141 (2): 1331–52. https://doi.org/10.1016/j.jeconom.2007.01.019.

Neumark, David, and William Wascher. 2000. “Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania: Comment.” *American Economic Review* 90 (5): 1362–96. https://doi.org/10.1257/aer.90.5.1362.

Olive, David J. 2003. “Prediction Intervals in the Presence of Outliers.” Southern Illinois University. https://pdfs.semanticscholar.org/39ba/0af091882616d8ac1e039a4f27c9540f378a.pdf.

Olive, David J. 2008. “Applied Robust Statistics.” Southern Illinois University. 2008. http://lagrange.math.siu.edu/Olive/ol-bookp.htm.

Olive, David J. 2011. “Robust Statistics.” Southern Illinois University. 2011. http://lagrange.math.siu.edu/Olive/robbook.html.

Olive, David J., and Douglas M. Hawkins. 2010. “Robust Multivariate Location and Dispersion.” http://lagrange.math.siu.edu/Olive/pphbmld.pdf.

Olive, David J., and Douglas M. Hawkins. 2011. “Practical High Breakdown Regression.” Southern Illinois University. http://lagrange.math.siu.edu/Olive/pphbreg.pdf.

Olive, David J. 2017. *Robust Multivariate Analysis.* Springer International Publishing. https://doi.org/10.1007/978-3-319-68253-2.

Otey, Matthew Eric, Amol Ghoting, and Srinivasan Parthasarathy. 2006. “Fast Distributed Outlier Detection in Mixed-Attribute Data Sets.” *Data Mining and Knowledge Discovery* 12 (2–3): 203–28. https://doi.org/10.1007/s10618-005-0014-6.

Pamula, Rajendra, Jatindra Kumar Deka, and Sukumar Nandi. 2011. “An Outlier Detection Method Based on Clustering.” In *2011 Second International Conference on Emerging Applications of Information Technology*, 253–56. Kolkata, India: IEEE. https://doi.org/10.1109/EAIT.2011.25.

Pison, Greet, Stefan Van Aelst, and G. Willems. 2002. “Small Sample Corrections for LTS and MCD.” *Metrika* 55 (1–2): 111–23. https://doi.org/10.1007/s001840200191.

Rencher, Alvin C., and William F. Christensen. 2012. *Methods of Multivariate Analysis*, 3rd Ed. Wiley Series in Probability and Statistics. John Wiley & Sons. https://doi.org/10.1002/9781118391686.

Ritter, Gunter. 2014. *Robust Cluster Analysis and Variable Selection*. 1st ed. Chapman and Hall/CRC. https://doi.org/10.1201/b17353.

Rousseeuw, Peter J. 1984. “Least Median of Squares Regression.” *Journal of the American Statistical Association* 79 (388): 871–80. https://doi.org/10.1080/01621459.1984.10477105.

Rousseeuw, Peter J. 1985. "Multivariate Estimation with High Breakdown Point." In *Mathematical Statistics and Applications*, edited by Wilfried Grossmann, Georg Ch. Pflug, István Vincze, and Wolfgang Wertz, 283–97. Dordrecht: Springer Netherlands. https://www.researchgate.net/profile/Peter_Rousseeuw/publication/239666038_Multivariate_Estimation_With_High_Breakdown_Point/links/0deec53137b8cc68aa000000.pdf.

Rousseeuw, Peter J., and Christophe Croux. 1993. "Alternatives to Median Absolute Deviation." Journal of the American Statistical Association 88(424), 1273-1283. https://doi.org/10.1080/01621459.1993.10476408.

Rousseeuw, Peter J., and Mia Hubert. 1999. "Regression Depth." Journal of the America Statistical Association 94 (466): 388–402. https://doi.org/10.2307/2670155.

Rousseeuw, Peter J., and Annick M. Leroy. 1987. *Robust Regression and Outlier Detection*. Wiley Series in Probability and Mathematical Statistics. New York: Wiley. https://doi.org/10.1002/0471725382.

Rousseeuw, Peter J., Stefan Van Aelst, Katrien Van Driessen, and Jose A. Gulló. 2004. "Robust Multivariate Regression." *Technometrics* 46 (3): 293–305. https://doi.org/10.1198/004017004000000329.

Rousseeuw, Peter J., and Katrien Van Driessen. 1999. "A Fast Algorithm for the Minimum Covariance Determinant Estimator." *Technometrics* 41 (3): 212–23. https://doi.org/10.1080/00401706.1999.10485670.

Rousseeuw, Peter J., and Katrien Van Driessen. 2006. "Computing LTS Regression for Large Data Sets." *Data Mining and Knowledge Discovery* 12 (1): 29–45. https://doi.org/10.1007/s10618-005-0024-4.

Rousseeuw, Peter J., and Bert C. Van Zomeren. 1990. "Unmasking Multivariate Outliers and Leverage Points." Journal of the American Statistical Association 85 (411), 633-639. https://doi.org/10.1080/01621459.1990.10474920.

Salibián-Barrera, Matías, Stefan Van Aelst, and Gert Willems. 2006. "Principal Components Analysis Based on Multivariate MM Estimators With Fast and Robust Bootstrap." *Journal of the American Statistical Association* 101 (475): 1198–1211. https://doi.org/10.1198/016214506000000096.

Soofi, Ehsan S., and Ali Dadpay. 2002. "Comment on 'Nearest-Neighbor Variance Estimation (NNVE): Robust Covariance Estimation via Nearest-Neighbor Cleaning.'" *Journal of the American Statistical Association* 97 (460): 1012–14. https://www.jstor.org/stable/3085824.

Torti, Francesca, Domenico Perrotta, Anthony C. Atkinson, and Marco Riani. 2012. "Benchmark Testing of Algorithms for Very Robust Regression: FS, LMS and LTS." *Computational Statistics & Data Analysis* 56 (8): 2501–12. https://doi.org/10.1016/j.csda.2012.02.003.

Visuri, Samuli, Visa Koivunen, and Hannu Oja. 2000. "Sign and Rank Covariance Matrices." *Journal of Statistical Planning and Inference* 91 (2): 557–75. https://doi.org/10.1016/S0378-3758(00)00199-3.

Wang, Naisyin, and Adrian E. Raftery. 2002. “Nearest-Neighbor Variance Estimation (NNVE): Robust Covariance Estimation via Nearest-Neighbor Cleaning.” *Journal of the American Statistical Association* 97 (460): 994–1006. https://doi.org/10.1198/016214502388618780.

Wilks, S.S. 1962. *Mathematical Statistics*. New York: John Wiley and Sons. https://doi.org/10.1002/bimj.19640060317.

Yohai, Victor J. 1987. “High Breakdown-Point and High Efficiency Robust Estimates for Regression.” *The Annals of Statistics* 15 (2): 642–56. https://doi.org/10.1214/aos/1176350366.

Yu, Dantong, Gholamhosein Sheikholeslami, and Aidong Zhang. 2002. “FindOut : Finding Outliers in Very Large Datasets.” *Knowledge and Information Systems* 4 (4): 387–412. https://doi.org/10.1007/s101150200013.

Zähle, Henryk. 2016. “A Definition of Qualitative Robustness for General Point Estimators, and Examples.” *Journal of Multivariate Analysis* 143 (January): 12–31. https://doi.org/10.1016/j.jmva.2015.08.004.

Zaman, Asad, Peter J. Rousseeuw, and Mehmet Orhan. 2001. “Econometric Applications of High-Breakdown Robust Regression Techniques.” *Economics Letters* 71 (1): 1–8. https://doi.org/10.1016/S0165-1765(00)00404-3.

Zhang, Jianfeng, David J. Olive, and Ping Ye. 2012. “Robust Covariance Matrix Estimation with Canonical Correlation Analysis.” *International Journal of Statistics and Probability* 1 (2): 119–36. https://doi.org/10.5539/ijsp.v1n2p119.

APPENDIX

A.1 SUM OF ABSOLUTE DIFFERENCES

Sum of Absolute Differences: *C* vs. *S*											
	N	*k*	γ	pm	RMVN	FMCD	detMCD	Hybrid	SCBUI	MB	RMB
Type 1	100	5	0.25	10	16.26	27.13	25.49	4.38	3.48	6.51	3.55
	100	5	0.25	20	1.08	98.53	0.83	1.19	1.11	6.46	1.10
	100	20	0.20	100	30.86	2301.49	25.12	19.67	85.48	56.13	28.99
	100	20	0.20	4000	31.01	3494712.20	25.63	19.64	85.48	56.13	28.99
	300	5	0.25	20	0.60	89.18	0.39	0.68	0.64	5.29	0.64
	300	10	0.25	20	3.23	103.12	100.18	2.20	1.90	14.60	1.77
	300	20	0.20	100	6.51	2015.29	7.16	7.65	12.54	37.68	6.19
	300	30	0.20	100	19.12	2242.64	21.67	17.50	261.23	81.04	18.31
	300	20	0.20	4000	6.53	2898167.17	7.58	7.58	10129.50	37.84	6.29
	1000	20	0.20	100	2.91	1773.79	3.43	3.60	2.91	27.33	2.91
	1000	10	0.10	100	0.96	1.72	1.69	0.91	0.94	5.28	0.95
	1000	30	0.20	100				7.53	22.33	54.29	5.73
Type 2	100	5	0.02	10	1.23	2.23	1.81	1.23	1.24	3.69	1.22
	100	5	0.45	10	1.46	0.51	0.02	1.39	33.32	20.12	1.32
	100	20	0.25	5	41.43	63.45	23.16	30.35	39.24	65.97	27.25
	100	20	0.25	10	35.57	253.82	22.36	19.44	29.73	66.04	27.18
	100	20	0.35	10	98.48	420.41	12.89	54.25	25.19	91.51	23.26
	300	5	0.25	10	0.61	0.43	0.41	0.69	0.65	5.43	0.65
	300	10	0.25	10	1.81	1.51	1.40	2.05	1.84	14.78	1.83
	300	10	0.25	3	7.44	2.42	3.52	3.52	5.26	14.98	7.31
	300	20	0.25	5	6.37	22.43	5.95	7.86	6.14	46.24	6.21
	300	20	0.25	10	6.19	15.88	5.98	7.81	6.14	46.44	6.15
	300	20	0.35	10	5.78	437.64	3.42	8.42	5.68	69.52	5.71
	1000	10	0.25	10	1.01	0.79	0.78	1.27	0.98	12.70	0.98
	1000	10	0.05	10	0.93	2.05	2.01	1.27	1.43	5.85	
Type 3	100	5	0.20	8	5.24	3.49	4.05	4.83	4.60	8.34	4.49
	100	5	0.40	8	20.97	16.91	18.41	18.43	23.64	30.45	20.44
	100	5	0.40	16	2.00	0.35	0.23	2.57	3.06	24.07	2.11
	300	5	0.24	8	7.59	3.95	4.63	4.39	6.91	7.85	6.92
	300	10	0.24	8	12.82	8.36	10.25	7.47	10.78	6.78	10.52
	300	20	0.24	8	23.67	15.80	19.63	17.86	20.50	14.13	7.77
	300	20	0.40	8	35.29	38.59	57.94	28.29	30.89	26.83	23.01
	1000	5	0.40	8	20.33	18.32	17.37	17.70	19.82	28.35	19.21
	1000	5	0.40	12	0.81	0.10	0.10	1.43	0.80	22.61	0.80
	1000	5	0.40	16	0.76	0.06	0.06	1.44	0.77	22.45	0.77
	1000	10	0.40	16	0.68	15.90	0.08	1.09	0.70	17.26	0.70
	1000	5	0.20	16	0.64	0.71	0.69	0.86	0.66	5.53	0.66
	1000	10	0.20	16	0.59	0.64	0.62	0.69	0.58	4.79	0.58
	1000	20	0.40	8	29.20	22.62	53.97	20.29	24.48	29.74	20.79
	1000	20	0.40	12	26.03			3.94	29.98	15.29	25.87
	1000	20	0.40	16	0.69	90.57	88.18	1.22	0.71	15.30	0.71
Type 4	100	5	0.40	8	30.95	18.80	28.69	24.32	35.04	43.68	26.66
	100	5	0.24	8	27.34	14.37	17.06	22.41	27.87	29.16	26.56
	100	5	0.40	16	80.69	70.63	73.52	78.31	93.57	69.45	82.50
	300	5	0.24	8	29.92	12.21	15.28	19.86	31.63	27.44	30.53
	300	10	0.24	8	31.80	22.62	25.90	27.90	33.64	19.64	33.17
	300	20	0.40	8	45.24	40.81	41.39	48.54	44.80	24.22	44.24
	1000	5	0.40	4	10.95	19.40	17.73	15.27	9.32	29.76	10.16
	1000	5	0.40	8	19.29	7.53	12.36	9.31	39.60	32.47	19.12
	1000	5	0.24	8	32.22	12.08	14.00	18.01	34.49	26.70	33.68
	1000	5	0.40	12	36.47	31.54	32.12	38.02	89.70	107.65	36.24
	1000	5	0.40	16	80.40	74.32	74.79	83.68	105.66	62.75	80.87
	1000	5	0.20	16	30.99	13.61	14.71	13.77	28.90	16.15	25.95
	1000	10	0.40	16	97.81	76.49	76.69	83.48	209.44	48.34	82.94
	1000	10	0.20	16	43.20			28.65	43.10	12.14	42.86
	1000	10	0.10	16	17.29			11.34	17.31	6.38	16.32
	1000	20	0.40	8	45.11	42.06	42.14	46.45	46.33	27.42	46.18
	1000	20	0.40	16	82.52	77.79	224.59	84.23	112.08	160.43	80.21

Table A1: Absolute differences between scatter matrix estimates

A.2 Rom-ρ approximation

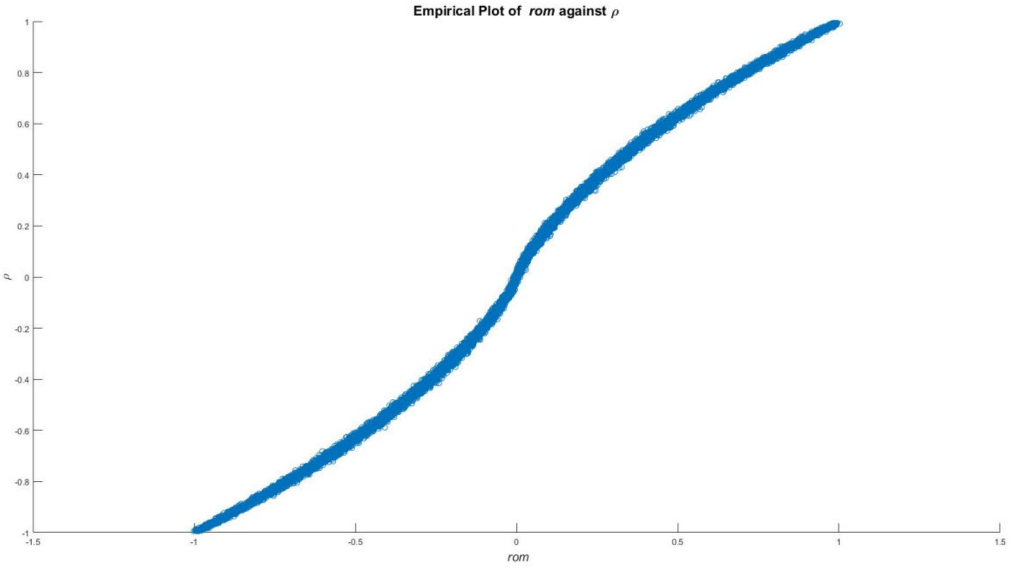


Figure A1. Plot of rom and ρ computed in each of 10,000 simulations of a zero mean bivariate Gaussian sample with 10,000 observations.

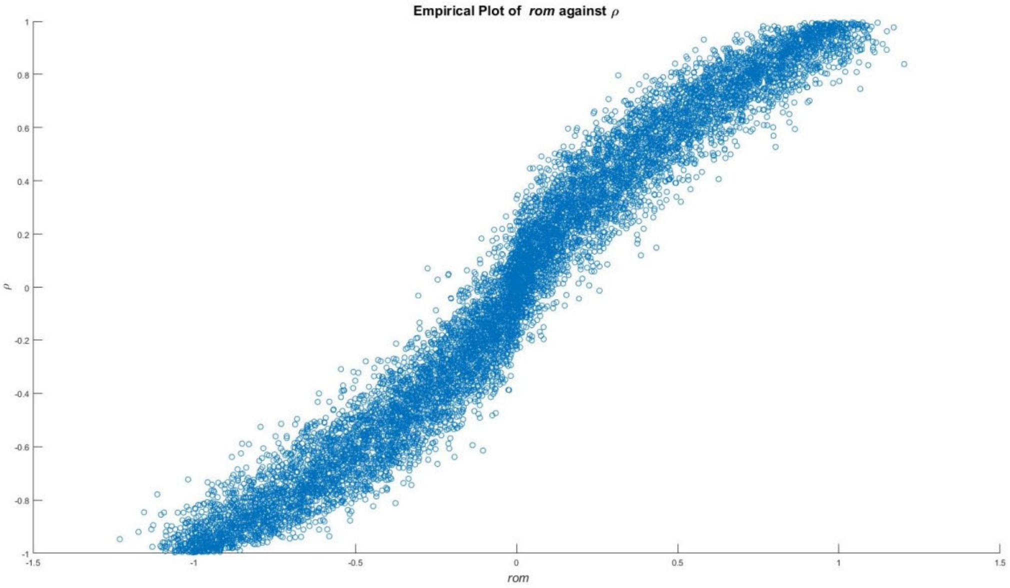


Figure A2. Plot of rom and ρ computed in each of 10,000 simulations of a zero mean bivariate Gaussian sample with 100 observations.

Let $\hat{r}_{ij} = MED\left(Z_i Z_j\right)$. Then

$$\hat{r}_{ij} = 3.6960356\,\hat{r}_{ij} - 34.0331641\,\hat{r}_{ij}^3 + 285.1035280\,\hat{r}_{ij}^5 - 768.8966090\,\hat{r}_{ij}^7$$

A.3 Hybrid Size and Power Comparison Versus Cerioli (2010) FDR Control Method

False Discovery Rate

N	v	Hybrid Method	Hybrid Method with Cerioli FDR Control
100	5	0.087	0.000
100	10	0.142	0.000
100	20	0.255	0.000
200	5	0.053	0.000
200	10	0.076	0.000
200	20	0.131	0.000
1000	5	0.030	0.000
1000	10	0.033	0.000
1000	20	0.040	0.000

Applying Cerioli (2010) FDR control method to the outlier detection in the hybrid method yields an FDR of 0.000.

Power

	N	k	γ	pm	Hybrid	Hybrid Method with Cerioli FDR Control
Type 1	100	5	0.25	10	0.84	0.18
	100	5	0.25	20	1	1
	100	20	0.2	100	1	1
	100	20	0.2	4000	1	1
	300	5	0.25	20	1	1
	300	10	0.25	20	1	0.98
	1000	20	0.2	100	1	1
	1000	10	0.1	100	1	1
	300	20	0.2	100	1	1
	300	30	0.2	100	1	1
	300	20	0.2	4000	1	1
	1000	30	0.2	100	1	1
Type 2	100	5	0.02	10	1	1
	100	5	0.45	10	1	1
	100	20	0.25	5	0.91	0.69
	100	20	0.25	10	1	1
	100	20	0.35	10	0.95	0.93
	100	10	0.25	3	0.70	0.02
	300	5	0.25	10	1	1
	300	10	0.25	10	1	1
	300	10	0.25	3	0.86	0.10
	1000	10	0.25	10	1	1
	1000	10	0.05	10	1	1
	300	20	0.25	5	1	1
	300	20	0.25	10	1	1
	300	20	0.35	10	1	1
Type 3	100	5	0.2	8	0.88	0.32
	100	5	0.4	8	0.62	0.29
	100	5	0.4	16	1	1
	300	5	0.24	8	0.87	0.53
	300	10	0.24	8	0.78	0.28
	300	20	0.24	8	0.61	0.01
	300	20	0.4	8	0.46	0.06
	1000	5	0.4	8	0.57	0.50
	1000	5	0.4	12	1	1
	1000	5	0.4	16	1	1
	1000	10	0.4	16	1	1
	1000	5	0.2	16	1	1
	1000	10	0.2	16	1	1
	1000	20	0.4	8	0.51	0.28
	1000	20	0.4	12	0.98	0.67
	1000	20	0.4	16	1	1
Type 4	100	5	0.4	8	0.37	0.01
	100	5	0.24	8	0.31	0.00
	100	5	0.4	16	0.60	0.37
	300	5	0.24	8	0.29	0.00
	300	10	0.24	8	0.16	0.00
	300	20	0.4	8	0.09	0.00
	1000	5	0.4	8	0.48	0.01
	1000	5	0.4	12	0.50	0.50
	1000	5	0.4	16	0.51	0.50
	1000	5	0.2	16	0.88	0.53
	1000	5	0.24	8	0.30	0.00
	1000	10	0.2	16	0.72	0.50
	1000	10	0.1	16	0.78	0.51
	1000	10	0.4	16	0.53	0.49
	1000	20	0.4	8	0.04	0.00
	1000	20	0.4	16	0.50	0.49
	1000		0.4	4	0.02	0.00

Applying Cerioli (2010) FDR control method to the outlier detection in the hybrid method results in a notable loss of power.

A.4 Masking and Invalid Knowledge Discovery

Aberrant or anomalous points, if they are not randomly distributed, may so influence parameter and covariance estimates that nothing seems amiss, a phenomenon known as *masking*. As noted above, masking occurs when a cluster of anomalous points effectively disguise each other, compromising inference without there being any indication that something has gone wrong. Importantly, masking can spoil inference even if the aberrant points are not extreme along any dimension; see Figure 1, an example from a bivariate data set in Rousseeuw and Leroy (1987) containing the body weight and brain weight of 28 animal species. Because the data are bivariate, in this example it is easy to detect three unusual observations, which are dinosaurs. We plot the classical OLS fitting line, and a robust OLS fitting line; the "effect" of body weight is (quite precisely) estimated to be either 0.5 or 0.75. Under masking, regression residuals appear well behaved, and widely used "leave-one-out" diagnostics are incapable of detecting the aberrant data. In this example, none of the dinosaur observations is identified as an outlier by Cook's *d*. It is easy to generate more extreme examples. For more discussion, see Appendix A.4, Rousseeuw and van Zomeren (1990), or Olive (2008).

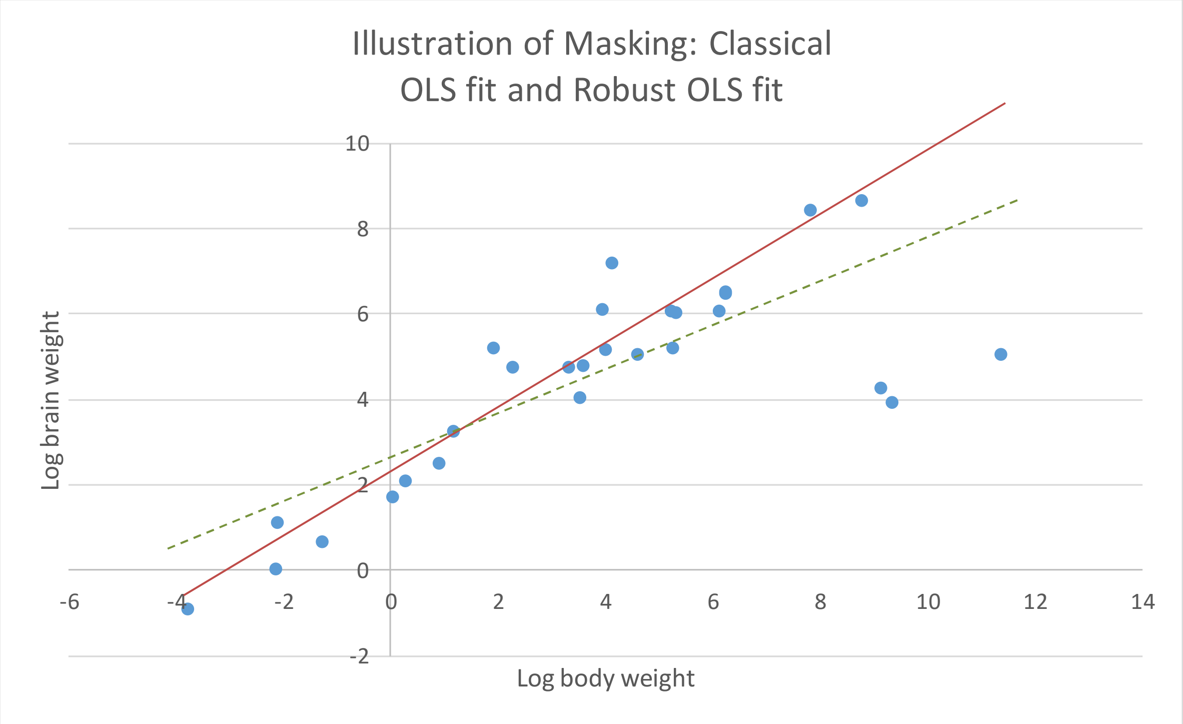


Figure 1. Log brain weight versus log body weight for 28 animals with classical OLS fit (dashed green line) and robust OLS fit (solid red line).

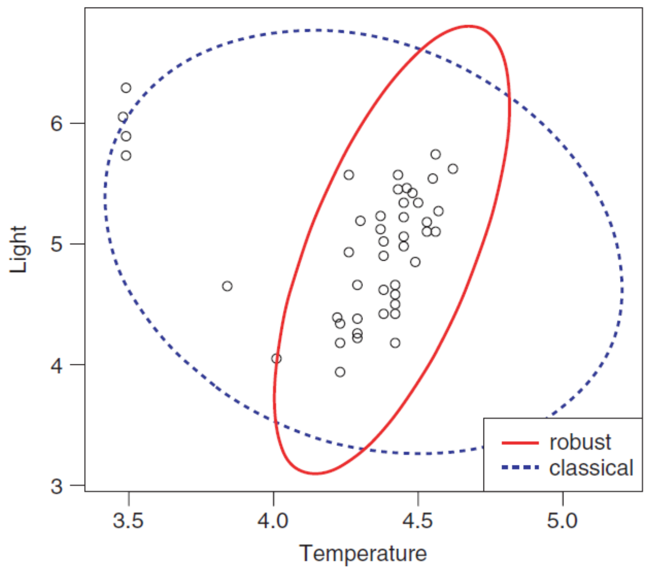


FIGURE 2 | Scatter plot of Light versus Temperature for 47 stars. The ellipses represent 99% confidence concentration. The large ellipse is based on all 47 stars. The small ellipse is obtained based on robust estimates of the location and scale.

An even more dramatic illustration of masking comes from the field of astronomy. The figure at right is taken from Hadi, Imon, and Werner (2009), in turn based on a data set presented and discussed by Rousseeuw and Leroy (1987). This example illustrates how classical statistical analysis, which assumes that the data are homogeneous and free from outliers, can even lead a researcher to a conclusion that is *opposite* of the truth. The data were taken from the Hertzsprung-Russell diagram of the star cluster CYG OB1 and consist of the measurements for 47 stars, including the logarithms of surface temperature and light intensity. The figure shows a scatter plot of these data. For simplicity, consider constructing a confidence interval for this bivariate distribution, assuming bivariate normality. The figure plots the 99th quantile of a χ^2 distribution. Two ellipses with the same confidence level are depicted. The large ellipse is computed using the classical estimates (sample mean and sample covariance matrix); it indicates two outliers. The smaller ellipse is computed using robust estimates of the mean and covariance; it indicates six outliers. In this small two-dimensional data set, it is easy to identify unusual points using a plot of the data. Four of these points are *clearly* different from the rest of the distribution. In fact, those four are all giants. Although the outliers constitute a small percentage of the data, the effect on the confidence region is dramatic. A researcher who wished to construct a linear model predicting the light emitted by a star as a function of its temperature, but who ignored the potential for outliers, would falsely conclude that the two variables are inversely related (in particular, the coefficient estimate is –0.4); but upon dropping just the four giants from the data, the coefficient estimate is positive (+2.0). These authors state: “This example underlines the idea that accurate identification of outliers before performing statistical analysis is absolutely necessary, if reliable conclusions are to be drawn...Unfortunately, because of the insidious nature of the masking effect, the identification of outliers in multivariate data is not an easy task.”