Why the Rousseeuw Yohai Paradigm is One of the Largest and Longest Running Scientific Hoaxes in History

David J. Olive [*]
Southern Illinois University

December 4, 2012

Abstract

The Rousseeuw Yohai paradigm for high breakdown multivariate statistics is based on one of the largest and longest running scientific hoaxes in history: that impractical brand name estimators can be efficiently computed by using some criterion to select a fit from a fixed number of easily computed trial fits. The bait and switch hoax is to give theory for high complexity impractical brand name estimators, but to actually use practical Fake-brand name estimators that are not backed by large sample or breakdown theory. Another hoax is to claim, without proof, that the practical Fake-brand name estimator is the brand name estimator.

KEY WORDS: robust regression, robust multivariate location and dispersion, high breakdown statistics

There has been a breakdown in the research and refereeing of multivariate "high breakdown robust statistics." The Rousseeuw Yohai paradigm is to replace the impractical brand name estimator by a practical Fake-brand name estimator that computes no more than a few thousand easily computed trial fits, but no breakdown or large sample theory is given for the Fake-brand name estimator (the "bait and switch hoax"). Most of the literature follows the Rousseeuw Yohai paradigm, using estimators like Fake-MCD, Fake-LTS, Fake-MVE, Fake-S, Fake-LMS, Fake-τ, Fake-Stahel-Donoho, Fake-Projection, Fake-MM, Fake-LTA, Fake-Constrained M, ltsreg, lmsreg, cov.mcd, cov.mve or OGK that are not backed by theory. Maronna, Martin and Yohai (2006, ch. 2, 6) and Hubert, Rousseeuw and Van Aelst (2008) provide references for the above estimators. Most of the brand name estimators were invented in papers by Rousseeuw, Yohai, Maronna or Tyler.

[*] David J. Olive is Associate Professor (E-mail: *dolive@siu.edu*), Department of Mathematics, Southern Illinois University, Carbondale, IL 62901-4408, USA.

Workers also often claim that the Fake-brand name estimator is affine equivariant, but if randomly chosen elemental starts are used, then the estimator depends on the random number seed and is not affine equivariant unless the random number seed is fixed. Hence many implementations of Fake-brand name estimators are not even affine equivariant. Run the program twice on the same data set and see if you get the same answer.

Problems with these estimators have been pointed out many times. For example, Huber and Ronchetti (2009, p. xiii, 8-9, 152-154, 196-197) suggest that high breakdown regression estimators do not provide an adequate remedy for the ill effects of outliers, that their statistical and computational properties are not adequately understood, that they "break down for all except the smallest regression problems by failing to provide a timely answer!" and that "there are no known high breakdown point estimators of regression that are demonstrably stable." Also see Stigler (2010). Woodruff and Rocke (1994, p. 889) point out that in practice the algorithm *is* the estimator (so the Fake-brand name estimator, rather than the brand name estimator, is the estimator). Rousseeuw (1993, p. 126) states that the random sampling versions of PROGRESS are *not* high breakdown algorithms.

Widely used multivariate "robust estimators" from the Rousseeuw Yohai paradigm are discussed in the JASA discussion paper Hawkins and Olive (2002) who prove that elemental concentration algorithms are zero breakdown and that elemental basic resampling estimators are zero breakdown and inconsistent. Also see Olive and Hawkins (2010, 2011). The proofs did not depend on the criterion used to pick the trial fit.

Hubert, Rousseeuw and Van Aelst (2002) reported that they appreciate this work. Maronna and Yohai (2002) correctly note that the algorithm estimators are inconsistent if the number of concentration steps is finite, but consistency is not known if the concentration is iterated to convergence. So it is not known whether Fake-MCD and Fake-LTS are consistent. These five authors ignore these results in their later work, resulting in the hoax. Note that the Maronna, Martin and Yohai (2006, p. 198-199) multivariate location and dispersion estimators that use $K = 500$ randomly chosen elemental sets and $k = 1$ concentration steps are inconsistent, and that the authors fail to cite Hawkins and Olive (2002) or Maronna and Yohai (2002).

Rousseeuw and Van Driessen (2006): "Computing LTS Regression for Large Data Sets" and Rousseeuw, Van Aelst and Hubert (1999, p. 425) claim that the LTS estimator can be computed with Fake-LTS. Hubert, Rousseeuw and Van Aelst (2008): "High Breakdown Multivariate Methods" do admit that most highly robust estimators take too long to compute, but claim that the zero breakdown Fake-MCD and Fake-LTS elemental concentration estimators can be used to efficiently compute MCD and LTS. The hoax is especially aggravating since variants of Fake-LTS were the running examples in Hawkins and Olive (2002). Rousseeuw and Van Driessen (1999): "A Fast Algorithm for the Minimum Covariance Determinant Estimator," and Hubert, Rousseeuw, and Verdonck (2012) also claim that Fake-MCD can be used to efficiently compute MCD.

The hoax can perhaps be most easily seen in Hubert, Rousseeuw and Verdonck (2012). They use $K = 6$ easily computed estimators $(T_i, \boldsymbol{C}_i)$ of multivariate location and dispersion, compute the determinant $det(\boldsymbol{C}_i)$ of each dispersion estimator and declare that they have efficiently computed the MCD estimator with the $(T_j, \boldsymbol{C}_j)$ that had the smallest determinant. However, they fail to prove that their latest Fake-MCD estimator i) is the MCD estimator, ii) is asymptotically equivalent to the MCD estimator, iii) is consistent, or iv) is high breakdown.

Van Aelst and Willems (2011) uses the bait and switch hoax, claiming that the infinite complexity S estimator and high complexity MM estimator can be computed with the Fake-S and Fake-MM estimators. Similarly, Bergesio and Yohai (2011, p. 666) also use a bait and switch hoax, replacing MCD by Fake-MCD in their WML estimator, and claim to compute $\sqrt{n}$ consistent high breakdown projection estimators for generalized linear models similar to the Fake-projection estimators for linear regression given in Maronna and Yohai (1993). The infinite complexity S estimators can not be computed since it can not be shown that the global minimum has been reached. Note that for multiple linear regression and multivariate location and dispersion, no one knows how to compute projection estimators (defined by computing all possible projections on the unit hypersphere) that have been shown to be high breakdown and consistent if the number of predictors $p > 2$. See Zuo and Lai (2011).

Most of the researchers in the field have been taken in by this obvious hoax. Hence the literature on high breakdown multivariate statistics is untrustworthy: the best papers either give large sample theory for brand name estimators that take far too long to compute, or give practical outlier resistant methods that could possibly be used as diagnostics but have not been shown to be consistent or high breakdown. As a rule of thumb, if $p > 2$ then the brand name estimators take too long to compute, so researchers who claim to be using a practical brand name estimator are actually using a Fake-brand name estimator.

Brand name estimators have absurdly high complexity:

Estimators with $O(n^4)$ or higher complexity take too long to compute and will rarely be used. The literature on estimators with $O(n^p)$ complexity typically claims that the estimator can be computed for $n = 100$ if $p = 4$, while simulations tend to use $p \leq 2$. Since estimators need to be widely used before they are trustworthy, the brand name high breakdown multivariate robust estimators are untrustworthy for $p > 2$. For $p > 4$, estimators with $O(n^p)$ or higher complexity **rank among the worst estimators ever proposed.** If $n = 100$, if the complexity is 1000 n^p, and if the computer can perform 10^7 operations per second, then the algorithm takes 10^{2p-4} seconds where 10^4 seconds is about 2.8 hours, 1 day is slightly less than 10^5 seconds, 10^6 seconds is slightly less than 2 weeks and 10^9 seconds is about 30 years. Suppose n and p are 200 and 10 respectively so that the data set is small by modern standards, and suppose the complexity is $n^p \approx 10^{23}$. A computer that could analyze one candidate solution per microsecond would take 5 billion years to evaluate the theoretical estimator.

Estimators computed by searching all "half sets," such as LTS and MVE, have exponential complexity since the Banzhaf Index, $C(n-1, 0.5(n-1)) \approx 2^n/\sqrt{2\pi(n-1)}$, is exponential in n. See Grofman (1981) and Ross (1989, p. 147). Woodruff and Rocke (1994, p. 893) note that if 1 billion subsets of size 101 could be evaluated per second, it would require 10^{33} millenia to search through all $C(200, 101)$ subsets if the sample size $n = 200$.

The fastest brand name estimator of multivariate location and dispersion that has been shown to be both consistent and high breakdown is the minimum covariance determinant (MCD) estimator with $O(n^v)$ complexity where $v = 1 + p(p+3)/2$. See Bernholt and Fischer (2004). The MVE complexity is exponential, and for $p > 2$ there may be no known method for computing S, τ, projection based, constrained M, MM, and Stahel-Donoho estimators.

The fastest brand name regression estimators that have been shown to be high breakdown and consistent are LMS and LTA with $O(n^p)$ complexity. See Bernholt (2005, 2006). The least quantile of differences and repeated median complexities are far higher, LTS complexity is exponential, and for $p > 2$ there may be no known method for computing S, τ, projection based, constrained M and MM estimators.

Some Theory and Conjectures for Fake Estimators

Some theory for the Fake-estimators actually used will be given after some notation. Let $p =$ the number of predictors. The Fake-MCD and Fake-S estimators are zero breakdown variants of the elemental concentration and elemental resampling algorithms that use K elemental fits where K is a fixed number that does not depend on the sample size n. To produce an elemental fit, randomly select h cases and compute the classical estimator $(T_i, \boldsymbol{C}_i)$ (or $T_i = \hat{\boldsymbol{\beta}}_i$ for regression) for these cases, where $h = p$ for multiple linear regression and $h = p+1$ for multivariate location and dispersion. The elemental resampling algorithm uses one of the K elemental fits as the estimator, while the elemental concentration algorithm refines the K elemental fits using all n cases. See Olive and Hawkins (2010, 2011) for more details.

Breakdown is computed by determining the smallest number of cases d_n that can be replaced by arbitrarily bad contaminated cases in order to make $\|T\|$ or $\|\hat{\boldsymbol{\beta}}\|$ arbitrarily large or to drive the smallest or largest eigenvalues of the dispersion estimator $\boldsymbol{C}$ to 0 or ∞. High breakdown estimators have $\gamma_n = d_n/n \to 0.5$ and zero breakdown estimators have $\gamma_n \to 0$ as $n \to \infty$.

A crucial observation is that the theory of the algorithm estimator depends on the theory of the trial fits, not on the estimator corresponding to the criterion. Note that if K = 1 and the classical estimator is used, computing the determinant of the sample covariance matrix does not convert the classical estimator into the MCD estimator and computing the median squared residual does not convert OLS into the LMS estimator. For another example, let $(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$ and $(\text{MED}(\boldsymbol{W}), diag(1, 3, ..., p))$ be the high breakdown trial fits. If the minimum determinant criterion is used, then the final estimator is $(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$. Although the MCD criterion is used, the algorithm estimator does not have the same properties as the MCD estimator.

Note that an estimator can not be consistent for θ unless the number of randomly selected cases goes to ∞, except in degenerate situations. The following theorem shows Fake-MCD and Fake-S are zero breakdown estimators. (If $K_n \to \infty$, then the elemental estimator is zero breakdown if $K_n = o(n)$. A necessary condition for the elemental basic resampling estimator to be consistent is $K_n \to \infty$.)

Theorem 1: a) The elemental basic resampling algorithm estimators are inconsistent. b) The elemental concentration and elemental basic resampling algorithm estimators are zero breakdown.

Proof: a) Note that you can not get a consistent estimator by using Kh randomly selected cases since the number of cases Kh needs to go to ∞ for consistency except in degenerate situations.

b) Contaminating all Kh cases in the K elemental sets shows that the breakdown value is bounded by $Kh/n \to 0$, so the estimator is zero breakdown. QED

Theorem 1 shows that the elemental basic resampling PROGRESS estimators of Rousseeuw (1984), Rousseeuw and Leroy (1987) and Rousseeuw and van Zomeren (1990) are zero breakdown and inconsistent, and thus the Rousseeuw and van Zomeren (1990, p. 649) claim that their MVEE estimator gives a good approximation to the MVE estimator is false. Yohai's two stage estimators, such as MM, need initial consistent high breakdown estimators such as LMS, MCD or MVE, but were implemented with the inconsistent zero breakdown elemental estimators such as `lmsreg`, Fake-LMS, Fake-MCD, MVEE or Fake-MVE. See Hawkins and Olive (2002, p. 157). Salibian-Barrera and Yohai (2008) still use the elemental basic resampling algorithm. You can get consistent estimators if $K_n \to \infty$ or $h_n \to \infty$ as $n \to \infty$. You can get high breakdown estimators and avoid singular starts if all $K_n = C(n, h) = O(n^h)$ elemental sets are used, but such an estimator is impractical.

Some workers claim that their elemental estimators search for sets for which the classical estimator can be computed, hence the above trivial results do not hold. (This claim does not excuse the fact that the workers fail to provide any large sample or breakdown theory for their "practical estimators.") For practical estimators, this claim is false since the estimator will not be practical if the program goes into an endless loop or searches all $O(n^p)$ elemental sets when supplied with messy data. Practical estimators may search more than the default number of K_d elemental sets, but still terminate if $K > K_d$ sets fails to produce an estimator. For example, the Rousseeuw and Leroy (1987) PROGRESS algorithm uses $K_d = 3000$ and $K \leq 30000$ elemental sets. Fake-MCD uses $K_d = 500$ and terminates very quickly.

Bali, Boente, Tyler and Wang (2011) gave possibly impressive theory for infinite complexity impractical robust projection estimators, but should have given theory for the practical Fake-projection estimator actually used. To estimate the first principal direction for principal component analysis, the Fake-projection (CR) estimator uses n projections $\boldsymbol{z}_i = \boldsymbol{w}_i/\|\boldsymbol{w}_i\|$ where $\boldsymbol{w}_i = \boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_n$. Note that for $p = 2$ one can select 360 projections through the origin and a point on the unit circle that are one degree apart. Then there is a projection that is highly correlated with any projection on the unit circle. If $p = 3$, then 360 projections are not nearly enough to adequately approximate all projections through the unit sphere. Since the surface area of a unit hypersphere is proportional to n^{p-1}, approximations rapidly get worse as p increases.

Theory for the Fake-projection (CR) estimator may be simple. Suppose the data is multivariate normal $N_p(\boldsymbol{0}, diag(p, 1, ..., 1))$. Then $\boldsymbol{\beta} = (1, 0, ..., 0)^T$ (or $-\boldsymbol{\beta}$) is the population first direction. Heuristically, assume $\hat{\boldsymbol{\mu}}_n = \boldsymbol{0}$, although in general $\hat{\boldsymbol{\mu}}_n$ should be a good $\sqrt{n}$ consistent estimator of $\boldsymbol{\mu}$ such as the coordinatewise median. Let $\boldsymbol{b}_o$ be the "best" estimated projection $\boldsymbol{z}_j$ that minimizes $\|\boldsymbol{z}_i - \boldsymbol{\beta}\|$ for $i = 1, ..., n$. "Good" projections will have a $\boldsymbol{y}_i$ that lies in one of two "hypercones" with a vertex at the origin and centered about a line through the origin and $\pm\boldsymbol{\beta}$ with radius r

at $\pm\boldsymbol{\beta}$. So for $p = 2$ the two "cones" are determined by the two lines through the origin with slopes $\pm\ r$. The probability that a randomly selected $\boldsymbol{y}_i$ falls in one of the two "hypercones" is proportional to r^{p-1}, and for $\boldsymbol{b}_o$ to be consistent for $\boldsymbol{\beta}$ need $r \to 0$, P(at least one $\boldsymbol{y}_i$ falls in "hypercone") $\to 1$ and $n \to \infty$. If these heuristics are correct, need $r \propto n^{\frac{-1}{p-1}}$ for $\|\boldsymbol{b}_o - \boldsymbol{\beta}\| = O_P(n^{\frac{1}{p-1}})$. Note that $\boldsymbol{b}_o$ is not an estimator since $\boldsymbol{\beta}$ is not known, but the rate of the "best" projection $\boldsymbol{b}_o$ gives an upper bound on the rate of the Fake-projection estimator $\boldsymbol{v}_1$ since $\|\boldsymbol{v}_1 - \boldsymbol{\beta}\| \geq \|\boldsymbol{b}_o - \boldsymbol{\beta}\|$. If the scale estimator is $\sqrt{n}$ consistent, then for a large class of elliptically contoured distributions, a conjecture is that $\|\boldsymbol{v}_1 - \boldsymbol{\beta}\| = O_P(n^{\frac{1}{2(p-1)}})$ for $p > 1$.

Alternatives to the Fake Estimators

A long standing question in Statistics is whether high breakdown multivariate statistics is a viable field of study. Are there useful high breakdown estimators of multivariate location and dispersion and multiple linear regression that are practical to compute? Can high breakdown estimators be incorporated into a practical algorithm in such a way that the algorithm estimator is consistent?

There is an alternative to the Rousseeuw Yohai paradigm. Use the estimators of Olive and Hawkins (2010, 2011) who avoid the "bait and switch error" by giving theory for the practical HBREG, FCH, RFCH and RMVN estimators actually used in the software. Good results can be obtained if intelligently selected trial fits are used. For a concentration algorithm, let a start be the initial estimator and the attractor the trial fit that results after applying $k + 1$ concentration steps to the start.

The Devlin, Gnanadesikan and Kettenring (1981) DGK estimator $(T_{k,D}, \boldsymbol{C}_{k,D}) = (T_{DGK}, \boldsymbol{C}_{DGK})$ uses the classical estimator $(T_{-1,D}, \boldsymbol{C}_{-1,D}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$ as the only start.

The Olive (2004) median ball (MB) estimator $(T_{k,M}, \boldsymbol{C}_{k,M}) = (T_{MB}, \boldsymbol{C}_{MB})$ uses $(T_{-1,M}, \boldsymbol{C}_{-1,M}) = (\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$ as the only start where MED($\boldsymbol{W}$) is the coordinatewise median. Hence $(T_{0,M}, \boldsymbol{C}_{0,M})$ is the classical estimator applied to the "half set" of data closest to MED($\boldsymbol{W}$) in Euclidean distance. For nonspherical elliptically contoured distributions, $\boldsymbol{C}_{MB}$ is a biased estimator of $c\boldsymbol{\Sigma}$. However, the bias seems to be small even for $k = 0$, and to get smaller as k increases. If the median ball estimator is iterated to convergence, we do not know whether $\boldsymbol{C}_{MB} \xrightarrow{P} c\boldsymbol{\Sigma}$.

The DGK and MB attractors can be used to define several robust estimators. Let the "median ball" be the hypersphere containing the half set of data closest to MED($\boldsymbol{W}$) in Euclidean distance. The FCH estimator uses the MB attractor if the DGK location estimator $T_{DGK} = T_{k,D}$ is outside of the median ball, and the attractor with the smallest determinant, otherwise. Let $(T_A, \boldsymbol{C}_A)$ be the attractor used. Then the estimator $(T_{FCH}, \boldsymbol{C}_{FCH})$ takes $T_{FCH} = T_A$ and

$$\boldsymbol{C}_{FCH} = \frac{\text{MED}(D_i^2(T_A, \boldsymbol{C}_A))}{\chi^2_{p,0.5}} \boldsymbol{C}_A \qquad (1)$$

where $\chi^2_{p,0.5}$ is the 50th percentile of a chi–square distribution with p degrees of freedom. The RFCH estimator uses two standard "reweight for efficiency steps" while the RMVN estimator uses a modified method for reweighting.

Reyen, Miller, and Wegman (2009) simulate the OGK and MBA estimators for $p = 100$ and n up to 50000. The OGK complexity is $O[p^3 + np^2 \log(n)]$ while that of

MBA, RMBA, FCH, RFCH and RMVN is $O[p^3 + np^2 + np\log(n)]$. These estimators are roughly 100 times faster than Fake-MCD.

The assumption below gives the class of distributions for which FCH, RFCH and RMVN have been shown to be $\sqrt{n}$ consistent. Distributions where the MCD functional is unique are called "unimodal," and rule out, for example, a spherically symmetric uniform distribution.

Assumption (E1): The $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid from a "unimodal" $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution with nonsingular covariance matrix $\text{Cov}(\boldsymbol{x}_i)$ where g is continuously differentiable with finite 4th moment: $\int (\boldsymbol{x}^T\boldsymbol{x})^2 g(\boldsymbol{x}^T\boldsymbol{x}) d\boldsymbol{x} < \infty$.

Theorem 2, Lopuhaä (1999). Suppose $(T, \boldsymbol{C})$ is a consistent affine equivariant estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ with rate n^δ where $s > 0$ and $0 < \delta \leq 0.5$. Assume (E1) holds. Then the classical estimator $(T_0, \boldsymbol{C}_0)$ applied to the cases with $D_i^2(T, \boldsymbol{C}) \leq h^2$ is a consistent affine equivariant estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ with the same rate n^δ where $a > 0$. The constant a depends on the positive constants s, h^2, p and the elliptically contoured distribution, but does not otherwise depend on the consistent start $(T, \boldsymbol{C})$.

Let $\delta = 0.5$. Applying the above theorem iteratively for a fixed number k of steps produces a sequence of estimators $(T_0, \boldsymbol{C}_0), ..., (T_k, \boldsymbol{C}_k)$ where $(T_j, \boldsymbol{C}_j)$ is a $\sqrt{n}$ consistent affine equivariant estimator of $(\boldsymbol{\mu}, a_j\boldsymbol{\Sigma})$ where the constants $a_j > 0$ depend on s, p, h and the elliptically contoured distribution, but do not otherwise depend on the consistent start $(T, \boldsymbol{C}) \equiv (T_{-1}, \boldsymbol{C}_{-1})$.

Concentration applies the classical estimator to cases with $D_i^2(T, \boldsymbol{C}) \leq D_{(c_n)}^2(T, \boldsymbol{C})$. Let

$$b = D_{0.5}^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{2}$$

be the population median of the population squared distances. Olive and Hawkins (2010) show that if $(T, \boldsymbol{C})$ is a $\sqrt{n}$ consistent affine equivariant estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ then $(T, \tilde{\boldsymbol{C}}) \equiv (T, D_{(c_n)}^2(T, \boldsymbol{C})\ \boldsymbol{C})$ is a $\sqrt{n}$ consistent affine equivariant estimator of $(\boldsymbol{\mu}, b\boldsymbol{\Sigma})$ where $b > 0$ is given by Equation (2), and that $D_i^2(T, \tilde{\boldsymbol{C}}) \leq 1$ is equivalent to $D_i^2(T, \boldsymbol{C}) \leq D_{(c_n)}^2(T, \boldsymbol{C})$). Hence Lopuhaä (1999) theory applied to $(T, \tilde{\boldsymbol{C}})$ with $h = 1$ is equivalent to theory applied to the concentration estimator using the affine equivariant estimator $(T, \boldsymbol{C}) \equiv (T_{-1}, \boldsymbol{C}_{-1})$ as the start. Since b does not depend on s, concentration produces a sequence of estimators $(T_0, \boldsymbol{C}_0), ..., (T_k, \boldsymbol{C}_k)$ where $(T_j, \boldsymbol{C}_j)$ is a $\sqrt{n}$ consistent affine equivariant estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ where the constant $a > 0$ is the same for each j. Then Olive and Hawkins (2010) show that the DGK and MCD estimators are estimating the same quantity. Note that the DGK estimator is practical to compute but has a much lower breakdown value than the impractical MCD estimator.

Theorem 3, Olive and Hawkins (2010). Assume (E1) holds. a) Then the DGK estimator and MCD estimator are $\sqrt{n}$ consistent affine equivariant estimators of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$.

b) The FCH, RFCH and RMVN estimators are $\sqrt{n}$ consistent estimators of $(\boldsymbol{\mu}, c_i\boldsymbol{\Sigma})$ for $c_1, c_2, c_3 > 0$ where $c_i = 1$ for multivariate normal data. If the clean data are in general position, then T_{FCH} is a high breakdown estimator and $\boldsymbol{C}_{FCH}$ is nonsingular even if nearly half of the cases are outliers.

Summary

Although it is obvious that theory for the practical estimator actually used should be given, this theory is not given in the high breakdown multivariate literature. Rousseeuw, Yohai, Hubert, Van Aelst and Maronna are responsible for one of the largest scientific hoaxes in history. These five authors and associate editors He, Tyler, Davies, Zamar, Zuo, and Croux are largely to blame for the breakdown in the refereeing process for high breakdown multivariate statistics. This group of 11 seems to continuously publish what should be unpublishable papers by the other authors and to block publication of papers that give theory for the estimators actually used, making Rousseeuw (1991) ironic. For PROGRESS to be made in highly outlier resistant multivariate robust statistics, journals need to remove this group from the refereeing process.

Tyler and Davies give theory for estimators with absurdly high complexity. He and Zuo make the massive inexcusable bait and switch error (not a hoax since they state that the estimator for which they gave theory takes too long to compute). Maronna and Yohai define estimators that need an initial consistent high breakdown estimator, but use an inconsistent zero breakdown elemental basic resampling estimator instead. Davies, Yohai, Maronna and Zamar write papers that only give breakdown or maximum bias theory for estimators with absurdly high complexity. Breakdown and bias properties are weaker than the property of being asymptotically unbiased. Yohai, Croux and Van Aelst give large sample or influence function theory for estimators (like MCD) with absurdly high complexity, but replace the impractical estimators by practical estimators (like Fake-MCD) that have no large sample theory.

It is inexcusable that so many authors have been taken in by the Rousseeuw Yohai hoax. Minimum standards for research require researchers to at least check that theory has been proven for the practical estimator. No breakdown or large sample theory is given for Fake-MCD or Fake-LTS: see Rousseeuw and Van Driessen (1999, 2006), and the estimators can be shown to be zero breakdown in one sentence as in Theorem 1b. Papers by Croux, Van Aelst and Yohai that give theory for a brand name estimator and then use a practical Fake-estimator in the software are more likely to fool nonexperts than work by Rousseeuw, but arguably the only theory for practical highly outlier resistant multivariate estimators worth reading is in papers by Olive, eg Olive and Hawkins (2010, 2011) and Hawkins and Olive (2002).

REFERENCES

Bali, J.L., Boente, G., Tyler, D.E. and Wang, J.L. (2011), "Robust Functional Principal Components: a Projection-Pursuit Approach," *The Annals of Statistics*, 39, 2852-2882.

Bergesio, A., and Yohai, V.J. (2011), "Projection Estimators for Generalized Linear Models," *Journal of the American Statistical Association,* 106, 661-671.

Bernholt, T. (2005), "Computing the Least Median of Squares Estimator in Time $O(n^d)$," *Proceedings of ICCSA 2005*, LNCS, 3480, 697-706.

Bernholt, T. (2006), "Robust Estimators are Hard to Compute," technical report available from (http://ls2-www.cs.uni-dortmund.de/∼bernholt/ps/tr52-05.pdf).

Bernholt, T., and Fischer, P. (2004), "The Complexity of Computing the MCD-Estimator," *Theoretical Computer Science*, 326, 383-398.

Devlin, S.J., Gnanadesikan, R., and Kettenring, J.R. (1981), "Robust Estimation of Dispersion Matrices and Principal Components," *Journal of the American Statistical Association,* 76, 354-362.

Grofman, B. (1981), "Fair Apportionment and the Banzhaf Index," *The American Mathematical Monthly,"* 88, 1-5.

Hawkins, D.M., and Olive, D.J. (2002), "Inconsistency of Resampling Algorithms for High Breakdown Regression Estimators and a New Algorithm," (with discussion), *Journal of the American Statistical Association,* 97, 136-159.

Huber, P.J., and Ronchetti, E.M. (2009), *Robust Statistics,* 2nd ed., Wiley, Hoboken, NJ.

Hubert, M., Rousseeuw, P.J., and Van Aelst, S. (2002), "Comment on 'Inconsistency of Resampling Algorithms for High Breakdown Regression and a New Algorithm' by D.M. Hawkins and D.J. Olive," *Journal of the American Statistical Association,* 97, 151-153.

Hubert, M., Rousseeuw, P.J., and Van Aelst, S. (2008), "High Breakdown Multivariate Methods," *Statistical Science*, 23, 92-119.

Hubert, M., Rousseeuw, P.J., and Verdonck, T. (2012), "A Deterministic Algorithm for Robust Location and Scatter," *Journal of Computational and Graphical Statistics,* 21, 618-637.

Lopuhaä, H.P. (1999), "Asymptotics of Reweighted Estimators of Multivariate Location and Scatter," *The Annals of Statistics,* 27, 1638-1665.

Maronna, R.A., Martin, R.D., and Yohai, V.J. (2006), *Robust Statistics: Theory and Methods*, Wiley, Hoboken, NJ.

Maronna, R.A., and Yohai, V.J. (1993), "Bias-Robust Estimates of Regression Based on Projections," *The Annals of Statistics*, 21, 965-990.

Maronna, R.A., and Yohai, V.J. (2002), "Comment on 'Inconsistency of Resampling Algorithms for High Breakdown Regression and a New Algorithm' by D.M. Hawkins and D.J. Olive," *Journal of the American Statistical Association,* 97, 154-155.

Olive, D.J. (2004), "A Resistant Estimator of Multivariate Location and Dispersion," *Computational Statistics and Data Analysis*, 46, 99-102.

Olive, D.J., and Hawkins, D.M. (2010), "Robust Multivariate Location and Dispersion," Preprint, see (www.math.siu.edu/olive/preprints.htm).

Olive, D.J., and Hawkins, D.M. (2011), "Practical High Breakdown Regression," Preprint, see (www.math.siu.edu/olive/preprints.htm).

Reyen, S.S., Miller, J.J., and Wegman, E.J. (2009), "Separating a Mixture of Two Normals with Proportional Covariances," *Metrika*, 70, 297-314.

Ross, S.M. (1989), *Introduction to Probability Models*, 4th ed., Academic Press, San Diego, CA.

Rousseeuw, P.J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association,* 79, 871-880.

Rousseeuw, P.J. (1991), "Why the Wrong Papers get Published," *Chance: New Directions for Statistics and Computing,* 4, 41-43.

Rousseeuw, P.J. (1993), "A Resampling Design for Computing High-Breakdown Regression," *Statistics and Probability Letters,* 18, 125-128.

Rousseeuw, P.J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association,* 79, 871-880.

Rousseeuw, P.J., and Leroy, A.M. (1987), *Robust Regression and Outlier Detection,* Wiley, New York, NY.

Rousseeuw, P.J., Van Aelst, S., and Hubert, M. (1999), "Rejoinder to Discussion of 'Regression Depth'," *Journal of the American Statistical Association,* 94, 419-433.

Rousseeuw, P.J., and Van Driessen, K. (1999), "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics,* 41, 212-223.

Rousseeuw, P.J., and Van Driessen, K. (2006), "Computing LTS Regression for Large Data Sets," *Data Mining and Knowledge Discovery*, 12, 29-45.

Rousseeuw, P.J., and van Zomeren, B.C. (1990), "Unmasking Multivariate Outliers and Leverage Points," *Journal of the American Statistical Association,* 85, 633-651.

Salibian-Barrera, M., and Yohai, V.J. (2008), "High Breakdown Point Robust Regression with Censored Data," *The Annals of Statistics,* 36, 118-146.

Stigler, S.M (2010), "The Changing History of Robustness," *The American Statistician,* 64, 271-281.

Van Aelst, S., and Willems, G. (2011), "Robust and Efficient One-Way MANOVA Tests," *Journal of the American Statistical Association*, 106, 706-718.

Woodruff, D.L., and Rocke, D.M. (1994), "Computable Robust Estimation of Multivariate Location and Shape in High Dimension Using Compound Estimators," *Journal of the American Statistical Association,* 89, 888-896.

Zuo, Y. and Lai, S. (2011), "Exact Computation of Bivariate Projection Depth and the Stahel-Donoho Estimator," *Computational Statistics & Data Analysis*, 55, 1173-1179.

Annotated list of papers with massive inexcusable errors:

Bai, Z.D., and He, X. (1999), "Asymptotic Distributions of the Maximal Depth Estimators for Regression and Multivariate Location," *The Annals of Statistics,* 27, 1616-1637. These estimators are impractical to compute.

Bali, J.L., Boente, G., Tyler, D.E. and Wang, J.L. (2011), "Robust Functional Principal Components: a Projection-Pursuit Approach," *The Annals of Statistics*, 39, 2852-2882. Uses Fake-projection estimator.

Cerioli, A. (2010), "Multivariate Outlier Detection with High-Breakdown Estimators," *Journal of the American Statistical Association*, 105, 147-156. Uses the zero breakdown Fake-MCD estimator.

Croux, C., Rousseeuw, P.J., and Hössjer, O. (1994), "Generalized S-Estimators," *Journal of the American Statistical Association,* 89, 1271-1281. S estimators have infinite complexity and the LQD complexity is far too high.

Danilov, M., Yohai, V.J., and Zamar, R.H. (2012), "Robust Estimation of Multivariate Location and Scatter in the Presence of Missing Data," *Journal of the American Statistical Association,* 107, 1178-1186. S estimators have infinite complexity.

Davies, P.L. (1987), “Asymptotic Behavior of S-Estimators of Multivariate Location Parameters and Dispersion Matrices,” *The Annals of Statistics,* 15, 1269-1292. S estimators have infinite complexity.

Davies, P.L., and Gather, U. (2005), “Breakdown and Groups,” *The Annals of Statistics,* (with discussion), 33, 977-1035. Breakdown is too trivial a property to warrant a paper, let alone a discussion paper.

Ferretti, N., Kelmansky, D., Yohai, V.J. and Zamar, R.H. (1999), “A Class of Locally and Globally Robust Regression Estimates,” *Journal of the American Statistical Association*, 94, 174-188. The generalized tau estimator needs an impractical initial robust MLD estimator. Fake-MCD and Fake-MVE are inadequate.

Gather, U. and Hilker, T. (1997), “A Note on Tyler’s Modification of the MAD for the Stahel-Donoho Estimator,” *The Annals of Statistics,* 25, 2024-2026. The Stahel Donoho estimator has infinite complexity for $p > 2$ (and for $p = 2$ when the paper was written).

Gervini, D., and Yohai, V.J. (2002), “A Class of Robust and Fully Efficient Regression Estimators,” *The Annals of Statistics,* 30, 583-616. REWLS needs impractical initial estimator so is impractical to compute.

He, X., Cui, H., and Simpson, D.G. (2004), “Longitudinal Data Analysis Using t-type Regression,” *Journal of Statistical Planning and Inference,* 122, 253-269. See paper below.

He, X., Simpson, D.G., and Wang, G.Y. (2000), “Breakdown Points of t-type Regression Estimators,” *Biometrika,* 87, 675-687. For the above two papers, the t type estimators need MVE as initial estimator, so are impractical to compute. Fake-MVE is inadequate.

Hubert, M., Rousseeuw, P.J., and Van Aelst, S. (2008), “High Breakdown Multivariate Methods,” *Statistical Science*, 23, 92-119. Hoax: claims MCD and LTS can be computed efficiently by Fake-MCD and Fake-LTS.

Hubert, M., Rousseeuw, P.J., and Verdonck, T. (2012), “A Deterministic Algorithm for Robust Location and Scatter,” *Journal of Computational and Graphical Statistics,* to appear. Hoax: claims MCD can be computed efficiently by Fake-MCD. Falsely claims to compute MCD by generating six easily computed estimators but provides no theory. Claim is false since MCD is impractical to compute but affine equivariant. Their 2nd Fake-MCD estimator (DET-MCD) is practical to compute but not affine equivariant. Needs to cite Devlin, Gnanadesikan, and Kettenring (1981) who invent concentration and who use concentration on the classical estimator. Needs to cite (Gnanadesikan, R., and Kettenring, J.R. (1972), “Robust Estimates, Residuals, and Outlier Detection with Multiresponse Data,” *Biometrics,* 28, 81-124.) who give a similar algorithm. See Rousseeuw and Leroy (1987, p. 254). Needs to cite Olive (2004) who suggests concentration on a few intelligently selected starts such as the classical and OGK estimators. Needs to cite Olive and Hawkins (2010) which provides theory for concentration. Followers of the Rousseeuw Yohai paradigm have been blocking the publication of my theory for concentration since 2004.

Kent, J.T., and Tyler, D.E. (1996), “Constrained M-estimation for Multivariate Location and Scatter,” *The Annals of Statistics,* 24, 1346-1370. CM for MLD has infinite complexity.

Kent, J.T., and Tyler, D.E. (2001), "Regularity and Uniqueness for Constrained M-Estimates and Redescending M-Estimates," *The Annals of Statistics,* 29, 252-265. CM estimator has infinite complexity.

Maronna, R.A., Martin, R.D., and Yohai, V.J. (2006), *Robust Statistics: Theory and Methods*, Wiley, Hoboken, NJ. Hoax: claims that infinite and high complexity estimators can be practically computed by algorithms that generate no more than a few thousand easily computed trial fits. Fails to cite Hawkins and Olive (2002) although Maronna and Yohai were discussants and admitted that concentration algorithms are inconsistent if the number of concentration steps is fixed. Needs to cite theory for concentration algorithms, but followers of the Rousseeuw Yohai paradigm have been blocking the publication of my theory for concentration since 2004.

Maronna, R.A., and Zamar, R.H. (2002), "Robust Estimates of Location and Dispersion for High-Dimensional Datasets," *Technometrics,* 44, 307-317. Present their conjectures that OGK is high breakdown and consistent as theorems, but fail to provide proofs. Proving the conjectures will need impressive large sample theory. The breakdown conjecture may be false since the initial dispersion estimator can be singular.

Rousseeuw, P.J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association,* 79, 871-880. Massively cited and interesting as an outlier diagnostic, but uses the zero breakdown inconsistent Fake-LMS estimator.

Rousseeuw, P.J., and Hubert, M. (1999), "Regression Depth," *Journal of the American Statistical Association,* 94, 388-433. Takes far too long to compute.

Rousseeuw, P.J., and Van Driessen, K. (1999), "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics,* 41, 212-223. Claims to compute MCD with the zero breakdown Fake-MCD estimator. Needs to cite Woodruff and Rocke (1994), who invented partitioning.

Rousseeuw, P.J., and Van Driessen, K. (2000), "An Algorithm for Positive-Breakdown Regression Based on Concentration Steps," in *Data Analysis: Modeling and Practical Application,* eds. Gaul, W., Opitz, O., and Schader, M., Springer-Verlag, New York, NY, 335-346. See paper below.

Rousseeuw, P.J., and Van Driessen, K. (2002), "Computing LTS Regression for Large Data Sets," *Estadistica*, 54, 163-190. See paper below.

Rousseeuw, P.J., and Van Driessen, K. (2006), "Computing LTS Regression for Large Data Sets," *Data Mining and Knowledge Discovery*, 12, 29-45. Hoax.

The above three papers seem to be about the same, and use the zero breakdown Fake-LTS estimator. A variant of this estimator was shown to be inconsistent in Hawkins and Olive (2002), and Rousseeuw was a discussant. Also (Ruppert, D. (1992), "Computing S-Estimators for Regression and Multivariate Location/Dispersion," *Journal of Computational and Graphical Statistics,* 1, 253-270.) invented concentration for LTS. Rousseeuw was a discussant but fails to cite Ruppert (1992).

Rousseeuw, P.J., Van Aelst, S., Van Driessen, K., and Agulló, J. (2004), "Robust Multivariate Regression," *Technometrics*, 46, 293-305. Uses the hoax that MCD can be computed with Fake-MCD.

Rousseeuw, P.J., and van Zomeren, B.C. (1990), "Unmasking Multivariate Outliers and Leverage Points," *Journal of the American Statistical Association,* 85, 633-651. Uses the zero breakdown inconsistent Fake-MVE estimator.

Rousseeuw, P.J., and Yohai, V.J. (1984), "Robust Regression by Means of S-Estimators," in *Robust and Nonlinear Time Series Analysis, Lecture Notes in Statistics,* eds. Franke, J., Härdle, W., and Martin, D., Springer-Verlag, NY, 26, 256-272. S estimators have infinite complexity.

Salibian-Barrera, M., Willems, G., and Zamar, R. H. (2008), "The Fast τ-Estimator of Regression," *Journal of Computational and Graphical Statistics*, 17, 659-682. The Fake-τ estimator is conjectured to have good statistical properties, but impressive large sample theory is needed.

Salibian-Barrera, M., and Yohai, V.J. (2006), "A Fast Algorithm for S-regression Estimates," *Journal of Computational Graphics and Statistics,* 15, 414-427. Fake-S estimator.

Salibian-Barrera, M., and Yohai, V.J. (2008), "High Breakdown Point Robust Regression with Censored Data," *The Annals of Statistics,* 36, 118-146. Hoax: uses zero breakdown inconsistent elemental resampling algorithm that Yohai admitted was inconsistent in the Maronna and Yohai (2002) discussion of Hawkins and Olive (2002).

Tyler, D.E. (1994), "Finite Sample Breakdown Points of Projection Based Multivariate Location and Scatter Statistics," *The Annals of Statistics,* 22, 1024-1044. Breakdown is too weak a property and projection estimators have infinite complexity for $p > 2$.

Van Aelst, S., and Willems, G. (2011), "Robust and Efficient One-Way MANOVA Tests," *Journal of the American Statistical Association*, 106, 706-718. Hoax: uses the Fake-S estimator instead of the S estimator.

Yohai, V.J. (1987), "High Breakdown-Point and High Efficiency Robust Estimates for Regression," *The Annals of Statistics,* 15, 642-656. The MM estimator needs an impractical initial estimator.

Yohai, V.J., and Zamar, R.H. (1988), "High Breakdown-Point Estimates of Regression by Means of the Minimization of an Efficient Scale," *Journal of the American Statistical Association,* 83, 406-410. The τ-estimator needs an impractical initial estimator.

Yohai, V.J., and Zamar, R.H. (1993), "A Minimax-Bias Property of the Least $\alpha-$Quantile Estimates," *The Annals of Statistics,* 21, 1824-1842. Bias is too weak of a property to deserve a paper.

Zhelonkin, M., Genton, M.G., and Ronchetti, E. (2012), "On the Robustness of Two-Stage Estimators," *Statistics & Probability Letters*, 82, 726-732. The MM estimator is impractical.