# ROBUST ESTIMATION

.

## 1. Robust Estimation

We have learned how to measure cases where data fails to conform to the assumptions necessary for ordinary least squares to be optimal. How do we deal with data that is not matching our gaussian assumptions?

To start, we will work on robust estimation of statistical measures such as the average. An understanding of these robust techniques will then help us easily construct robust regression estimators.

## 2. Robust Location And Scale

We begin by considering *location*, which we can think of as the middle or "average" of a set of numbers, and *scale* which measures how much the numbers deviate from that average.

Without a good center to any fit, it is hard to imagine it performing well.

An ancient and powerful robust metric is the *median* which is computed by sorting all available values, and then taking the middle one (if $n$ is odd) or the average of the two middle ones (if $n$ is even).

Having defined a robust average (median or $\text{Med}(\cdot)$), we can use it to form another long-used robust statistical metric, this time of scale. We construct the *median absolute deviation* as

$$MAD = \text{Med}\left\{\left|\boldsymbol{y} - \text{Med}\left\{\boldsymbol{y}\right\}\right|\right\}$$

Because our distribution may well not be normal, but is still often close to being normal, our goal is to find more general robust estimation techniques that are

- "nearly optimal" when the distribution really is normal, and
- efficient when the distribution is not quite normal

How will measure efficiency and quality when our data is not normal?

### 2.1. Distributional Mixtures.

Consider the case where "most" data points come from a clean, correct distribution $G$, but that a small proportion, $\epsilon$, are "poisoned" by data problems, counterparty axes,

or other confounding effects. The poisoned samples come from a far different distribution $H$.

Somehow, we want to correctly measure how $x$ relates to $y$ without being greatly affected by the poisoned samples. This is, of course, fairly easy if we know *which* samples came from $H$ but in reality there is usually no way to determine that. A mathematically convenient example is to assume that both $G$ and $H$ are normal, but $H$ has a much wider standard deviation, say 3 times that of $G$. Then practically all samples from $H$ would have a high residual, and we might expect a very good estimate by zeroing their influence in the *trimmed mean*

$$m_\delta = \text{mean}\left\{ y_i : \left| Q(y_i) - \frac{1}{2} \right| < \frac{1}{2} - \delta \right\}$$

The trimmed mean is clearly insensitive to any of a small subset of values veering off to infinity. However in practice, the task of trying to determine an appropriate $\delta$ relies on estimating $\epsilon$. As values of $\epsilon$ get small, the relative sample size to have any hope of doing so becomes huge.

2.2. **Robust Location.** The median can be thought of as an extreme trimmed mean, with $\delta \approx 0.5$. We see that its source of robustness is that it discards essentially *all* the data (with the exception of the middle point). On the other hand, the median can tolerate up to 50% gross errors before it's made arbitrarily large. We say that the *breakdown point* of the median is 50% while for the usual mean it is 0% and for the trimmed mean it is $\delta$. Now, returning to our picture of poisoned samples, is very hard to figure out which samples came from $G$ and which from $H$ even if we somehow knew $\epsilon$ exactly. Therefore, finding outliers is a daunting task. We have other reasons to eschew the process of trying to identify outliers, including

- There may be some information remaining in apparent outliers
- Outliers get hard to spot in highly multivariate cases
- Other effects such as minimum tick size come into play
- Rejecting the highest contributors to variance will likely cause us to significantly underestimate true variance

To better specify our poison model of bad data, we'll say that its distribution is

$$F = (1 - \epsilon)N(\mu, 1) + \epsilon(N(\mu, \tau))$$

where *tau* is an uncomfortably large value for the standard deviation. Given this distribution we can compute the variance of the mean as

$$\frac{1}{n}(1-\epsilon+\epsilon\tau^2) \tag{1}$$

and the theoretical variance of the sample median as approximately

$$\frac{\pi}{2}\left(\frac{1}{n}(1-\epsilon+\epsilon/\tau)^{-2}\right) \tag{2}$$

Let's say we have two different unbiased estimators $\dot{m}$ and $\ddot{m}$ for the average (or really any statistical quantity). Then we'll define the *relative efficiency* of the two as

$$RE(\dot{m},\ddot{m}) = \frac{\mathrm{Var}(\ddot{m})}{\mathrm{Var}(\dot{m})}$$

and if we take the limit as sample size $n \to \infty$ we can define *asymptotic relative efficiency* or *ARE* as the value of that limit. If we don't say otherwise, then we implicitly define $\ddot{m} = \mathrm{mean}(\cdot)$. For a normal distribution, we see from Formulas 1 and 2 the asymptotic relative efficiency of the median against the mean is $\frac{2}{\pi}$. For data distributed according to a Student $t$ distribution with 5 degrees of freedom, it is much better at about 96%. If we are willing to spend extra processor cycles, the Hodges-Lehmann *median paired mean estimator* or *MPM estimator*,

$$\mathrm{Med}\left\{\frac{x_i+x_j}{2}\right\}_{i,j=1}^{n}$$

has high efficiency (over 85%) at the normal and a breakdown point of 29%.

2.3. **Robust Scale.** Tukey (1960) shows that for our poisoned distribution example, things get very interesting. Contrast standard deviation with the *median absolute deviation* which we recall is defined as

$$MAD = \mathrm{Med}\left\{\left|\boldsymbol{y} - \mathrm{Med}\left\{\boldsymbol{y}\right\}\right|\right\}$$

We find that if we construct the *normalized median average deviation* (also known as *normalized MAD* or MADN) estimator

$$\dot{\sigma} = \frac{MAD}{\Phi^{-1}(3/4)}$$

we can compare it to the standard deviation to obtain

| $\epsilon$ | ARE($\dot{\sigma}$) |
|---|---|
| 0 | 0.876 |
| 0.10% | 0.948 |
| 0.20% | 1.016 |
| 1% | 1.44 |
| 5% | 2.04 |

showing that even a single poisoned value in 100 can quickly make robust measures the optimal behavior .

We can also define the *normalized interquartile range*, based on quantiles, as

$$R^Q = \frac{Q_{3/4} - Q_{1/4}}{\Phi^{-1}(3/4) - \Phi^{-1}(1/4)}$$

2.3.1. *Scale by IRPM.* In 2011/2012 Tarr, Müller and Weber proposed an estimator based on pairwise means, in the spirit of the Hodges-Lehmann location estimator. We begin by defining

$$h(x_1, x_2) = \frac{x_1 + x_2}{2}$$

and considering the set of pairwise means

$$\{h(x_i, x_j),\ 1 \le i, j \le n\}$$

of size

$$\binom{n}{2} = \frac{n(n-1)}{2}.$$

$H_n$ is the empirical distribution function of pairwise means

$$H_n(t) := \frac{2}{n(n-1)} \sum_{i<j} \mathbb{1}_{h(x_i,x_j)\le t}$$

and the general estimator is

$$P_n(\tau) = c_{n,\tau} \left[ H_n^{-1}\left(\frac{1+\tau}{2}\right) - H_n^{-1}\left(\frac{1-\tau}{2}\right) \right]$$

where $c_{n,\tau}$ is chosen to make the estimator consistent for the normal distribution. Asymptotically

$$c_\tau = \left[ H_\Phi^{-1}\left(\frac{1+\tau}{2}\right) - H_\Phi^{-1}\left(\frac{1-\tau}{2}\right) \right]^{-1},$$

taking $H_\Phi$ is the standard gaussian CDF of means,

$$H_\Phi(t) = \int_{-\infty}^{\infty} \Phi(2t - x)\Phi(x)\,\mathrm{d}x$$

More efficient scale estimates than IQR are available, however, at relatively small computation cost. In 1993, the *pairwise difference estimator* or *PDifE*

$$Q_n = c(n)\{pd(x_i, x_j)|i \le j\}_{(n)} = c(n)\{\left|x_i - x_j\right| \, |i \le j\}_{(n)}$$

of Rosseeuw and Croux became popular[1]. and $0 < \tau \le 1$. We nearly always take $\tau = 1/2$ so that we are computing the interquartile range of the pairwise distances, and we will call this estimator $P_n$, the *interquartile range of pairwise means* or *IRPM*. It has breakdown at approximately

$$1 - \sqrt{\frac{1+\tau}{2}}.$$

In the general case, it is not possible to obtain an analytic expression for $c_{n,\tau}$ at finite sample sizes $n > 4$. However, Tarr *et alia* provide a table of values of $c_{n,0.5}$ up to $n = 40$ and note that, for $n > 40$, one can use

$$c_{n,\tau} = \frac{1}{1 - 0.7/n}.$$

They also discuss the advantages of trimming or *adaptive trimming* of samples for $P_n$, where samples are trimmed, based on *preliminary* estimates $\ddot{m}$ and $\ddot{\sigma}$ of scale and location, if

$$\frac{|x_i - \ddot{m}|}{\ddot{\sigma}} > 5.$$

This trimming helps greatly with highly pathological distributions where $P_n$ otherwise tends to underperform the PDifE $Q_n$.

The IRPM is computable in $O(n \log n)$ time using the most efficient available algorithm. For a flavor of that algorithm, let $\boldsymbol{D}$ be a matrix of squared distances between vector elements of matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ (here both are just comprised of a copy of our data set). Then $\boldsymbol{D}$ has elements

$$d_{ij} = \|\boldsymbol{x}_i\|^2 + \|\boldsymbol{y}_i\|^2 - 2\boldsymbol{x}^* \cdot \boldsymbol{y}$$

so we see that

$$\boldsymbol{D} = \boldsymbol{1}\boldsymbol{X}^*\boldsymbol{X}^* + (\boldsymbol{Y}^*\boldsymbol{Y}^*\boldsymbol{1})^* - 2(\boldsymbol{X}\boldsymbol{Y}^*)^*$$

so distance computations are a matrix operation in $O(n)$ times lookups in $O(\log n)$.

---

[1]Let $C = \left(\Phi^{-1}(5/8)\sqrt{2}\right)^{-1} = 2.21914446598508$. The coefficients for $n = 2, \ldots, 9$ are $C \cdot k$ for $k = 0.400, 0.993, 0.514, 0.845, 0.612, 0.859, 0.670, 0.874$. For larger $n$, the coefficient is $Cn/(n + 1.4)$ for the odd $n$ and $Cn/(n + 3.8)$ for even $n$.

## 3. Correlation and Covariance

It is tricky to estimate correlation and covariance in a robust manner, because of the potentially different scales of the two dimensions. In particular it is hard to satisfy the homogeneity condition

$$\mathrm{Cov}(aX, bY) = ab\ \mathrm{Cov}(X, Y)$$

Gnanadesikan and Kettenring proposed we can use a difference identity to get something that works, by noticing

$$\mathrm{Cov}(x, y) = \frac{1}{4}\left(\mathrm{SD}\,(x+y)^2 - \mathrm{SD}\,(x-y)^2\right)$$

suggesting a pairwise robust correlation using a robust standard deviation estimator $\hat{\sigma}$as

$$\mathrm{RCorr}(x, y) = \frac{1}{4}\left(\hat{\sigma}\left(\frac{x}{\hat{\sigma}(x)} + \frac{y}{\hat{\sigma}(y)}\right)^2 - \hat{\sigma}\left(\frac{x}{\hat{\sigma}(x)} - \frac{y}{\hat{\sigma}(y)}\right)^2\right)$$

and obtaining the covariance from it via

$$\mathrm{RCov}(x, y) = \hat{\sigma}(x)\hat{\sigma}(y)\mathrm{RCorr}(x, y)$$

This technique is not guaranteed to give a positive definite covariance matrix, so one needs to use some sort of projection technique[2] to force one.

For full covariance matrices, is sometimes makes much more sense to "shrink" the covariance matrix in the simple manner of Ledoit and Wolf. We start with a sample covariance matrix $\boldsymbol{S}$ and a simple correlation matrix $\boldsymbol{F}$ that we consider to be reasonable a priori. We create a value for *shrinkage intensity* $\hat{\delta}$ based on correlations found in $\boldsymbol{S}$ and the distance from $\boldsymbol{F}$ in Frobenius norm, and then set our estimate to

$$\boldsymbol{\Sigma}^{\mathrm{Shrunken}} = \hat{\delta}\boldsymbol{F} + (1-\hat{\delta})\boldsymbol{S}$$

A shrunken covariance matrix will nearly always[3] be positive definite if $\boldsymbol{S}$ itself was positive definite.

### 3.1. Robust Z-Score.

A *Z-score* is a scale-normalized distance from expected location. In the gaussian case this is

$$z_i^{\mathrm{Gauss}} = \frac{y_i - \mathrm{Mean}(y_i)}{\mathrm{Std}(y_i)}$$

Now that we have robust estimates $\hat{\mu}$ and $\hat{\sigma}$ of scale an location, we need no longer restrict ourselves to the gaussian Z-score above, and can

[2] See the work by Nick Higham.

[3] It is possible to construct pathological counterexamples with negative correlations in $\boldsymbol{S}$ and/or odd formulations of $\boldsymbol{F}$.

form a robust Z-score

$$\hat{z}_i = \frac{y_i - \hat{\mu}}{\hat{\sigma}}$$

In the multivariate case, the z-score generalizes to the *Mahalanobis distance*

$$\hat{\boldsymbol{z}}_i = \sqrt{(\boldsymbol{y_i} - \hat{\boldsymbol{\mu}})^* \boldsymbol{\Sigma}^{-1} (\boldsymbol{y_i} - \hat{\boldsymbol{\mu}})}$$

Note here the inverse covariance matrix $\boldsymbol{\Sigma}^{-1}$ which we may estimate robustly.

The concept of Z-Score is particularly useful to us because its robust extensions will provide the foundation of techniques to deal with regression residuals.

## 4. Mathematics of Robustness

Let us now take a more general perspective on constructing alternative estimators, by rephrasing the process in terms of maximum likelihood estimation. We will be rewarded by easily perceiving how to extend the techniques to the case of linear models.

Our key assumption is that the data distribution is of some arbitrary shape with location and scale parameters unknown. We assume the distribution has some canonical shape $g$ which we consider to be zero-centered and of unit scale.

Based on the distribution's functional form, we will find $\mu$ and $\sigma$ that best explain the data, in the sense of having the highest probability or *maximum likelihood*.

To do so, we will find a $\mu$ such that subtracting $\mu$ from all the data maximizes our $g$-likelihood, or dividing all the data by $\sigma$ maximizes our $g$-likelihood, or both.

Let's say the true density $\zeta_\mu$ of $y$ has a location (average) parameter $\mu$. The probability of observing our data set $\boldsymbol{y}$ is the product of the individual sample probabilities

$$P = \prod_{i=1}^{n} \zeta_\mu(y_i)$$

and is maximal.

In our view, we are trying to find the $\mu$ that maximizes this overall density when $\zeta_\mu$ is just a shifted version of the canonical $g$. Such a $\mu$ would be the *maximum likelihood estimate*.

A practical difficulty is that for large sample counts $n$ this $P$ tends to be such a vanishingly small number that digital computers experience computational underflow trying to calculate it. However, we can use

any monotonic function to transform it, and maximize the transformed version instead.

The logarithm is a particularly convenient monotonic function, and using it we obtain

$$\begin{aligned}\log(P) &= \log\left(\prod_{i=1}^{n} g(y_i - \mu)\right) \\ &= \sum_{i=1}^{n} \log(g(y_i - \mu))\end{aligned}$$

Define the function $h(x) = -\log(g(x))$, and we find that we need to minimize

$$h(\boldsymbol{y}) = \sum_{i=1}^{n} h(y_i - \mu).$$

The minimum occurs, of course, when

$$\frac{\partial}{\partial \mu} h(\boldsymbol{y} - \mu) = 0$$

that is, when

$$\sum_{i=1}^{n} h'(y_i - \mu) = 0$$

Any solution to this equation is called an *M-estimate* of the median, where the "M" stands for "MLE-like". Note that a choice of $h$ is equivalent to a choice of $g$, since $g(x) = e^{-h(x)}$, so if we wish we may directly select $h$ and then simply *define* the density $g$ as its negative exponent.

Note that if we had the gaussian density

$$\Phi' = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

then $h^G = x^2/2$ and we would be minimizing a sum of squares. The derivative of $h^G$ is $x$, yielding a requirement for $\mu$ that

$$\frac{1}{n}\sum(y_i - \mu) = 0$$

recovering the usual mean $\mu = \frac{1}{n}\sum(y_i)$.

4.0.1. *Double Exponential.* Another interesting case is when $h$ is the *double exponential distribution*

$$g(x) = g^{de}(x) = \frac{1}{2} e^{-|x|}$$

so that $h^{de}(x) = |x|$. We don't bother yet with the derivative $\frac{\partial}{\partial x} h^{de}$, but just note that we are trying to find

$$\arg\min_{\mu} \sum |y_i - \mu|$$

The derivative $\frac{\partial}{\partial x} h^{de}(x)$ is a step function, flat almost everywhere, indicating that the minimum of $h^{de}$ occurs where a sign change in $\frac{\partial}{\partial x} h^{de}$ happens, i.e. as it crosses 0. Thus, the median itself is an M-estimate via the function $g^{de}$.

Another interpretation is that if we *know* our data is double-exponentially distributed, then the MLE of its average location comes from the median and not the mean.

4.1. **General Center Location.** Let's define $\psi = h'$. For any distribution $f$ of residuals that we have, we can define its version of the distribution's center

$$\mu_0 \ni \mathbb{E}_f \left[\psi(x - \mu_0)\right] = 0$$

and compute the *asymptotic variance*, a ratio of expectations

$$v = \frac{\mathbb{E}_f \left[\psi(x - \mu_0)^2\right]}{\mathbb{E}_f \left[\psi'(x - \mu_0)\right]^2}$$

For a finite sample size, the strong law of large numbers implies that an estimator is *asymptotically normal*

$$\hat{\mu} \sim N(\mu_0, \sqrt{\frac{v}{n}})$$

so it makes sense to characterize an estimator's behavior by the variance of estimates it generates. We define the *asymptotic efficiency* of $\hat{\mu}$ based on $v_0$, the asymptotic variance of the maximum likelihood estimator, as

$$\text{Eff}(\hat{\mu}) = \frac{v_0}{v}$$

4.1.1. *Huber Response.* An historically important choice of $h$ is the *Huber function*

$$h_k(x) = h(x; k) = \begin{cases} x^2 & |x| \le k \\ 2k|x| - k^2 & |x| > k \end{cases}$$

which has the usual quadratic penalty within the range of its scale parameter $k$, but only linear penalty outside $k$. The rate of increase of

penalty, which defines the *influence curve*, is then

$$\psi_k(x) = 2\begin{cases} x & |x| \le k \\ k \cdot \operatorname{sign}(x) & |x| > k \end{cases}$$

In this case we can actually compute $v$ because

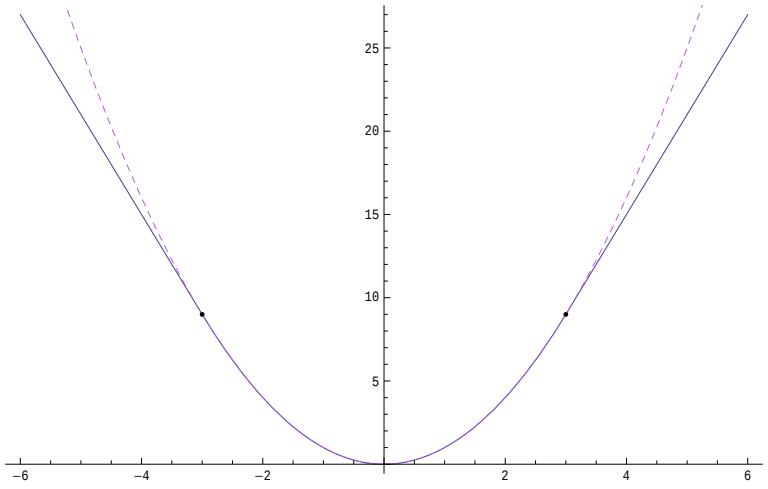


FIGURE 1. Huber function, scale parameter 3 (solid line), with standard quadratic (dashed)

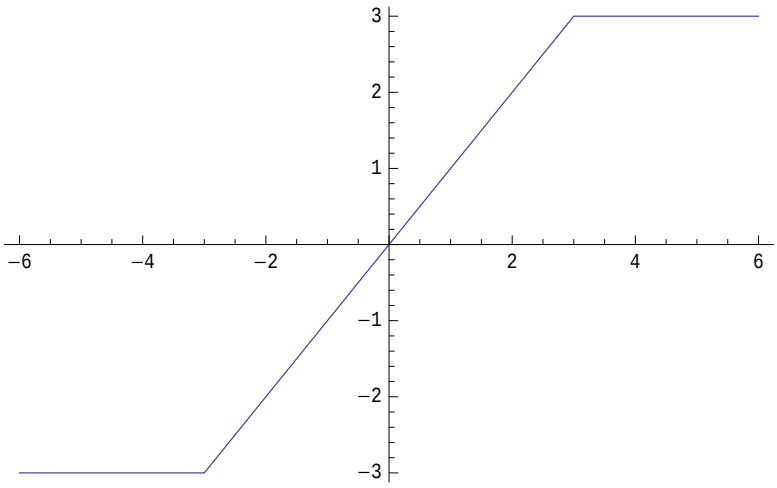


FIGURE 2. Huber Derivative $\Psi_H$

$$\mathbb{E}_f\left[\psi(x)^2\right] = 2\left(k^2(1-\Phi(k)) + \Phi(k) - 1/2 - k\Phi'(k)\right)$$

and

$$\mathbb{E}_f\left[\psi'(x)\right] = \Phi(k) - \Phi(-k)$$

4.1.2. *Tukey Response.* Consider the location MLE for the Student $t$ distribution with $\nu$ degrees of freedom, whose $h_t$ has derivative

$$\psi_t(x) \approx \frac{x}{x^2+\nu}$$

This leads us to believe that heavy tails affect the best estimator in that rather than going linear in penalty outside a given range, we may wish to cut off further contributions completely. We therefore consider the *bisquare function* by Tukey

$$b_k(x) = \begin{cases} 1 - \left(1 - \left(\frac{x}{k}\right)^2\right)^3 & |x| \le k \\ 1 & |x| > k \end{cases}$$

which has the influence curve

$$\psi_k^b(x) = \begin{cases} \frac{6}{k^2} \cdot x \left(1 - \left(\frac{x}{k}\right)^2\right)^2 & |x| \le k \\ 0 & |x| > k \end{cases}$$

Estimators whose derivative goes to zero outside some region have a

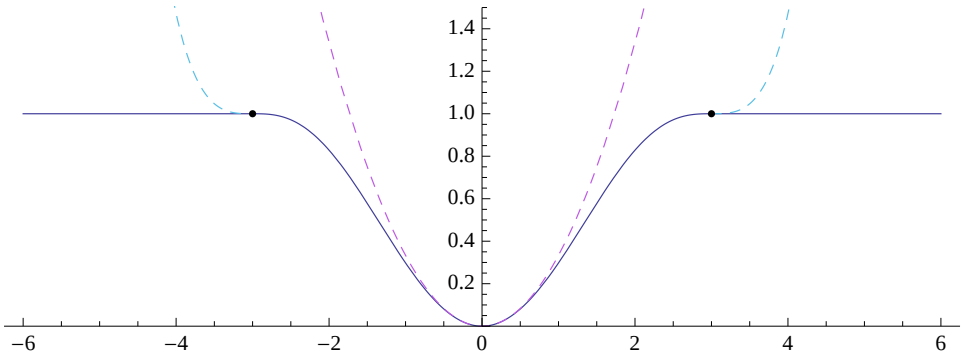


FIGURE 3. Bisquare function, scale parameter 3 (solid line), with standard quadratic and degree 6 continuation (dashed)

nontrivial breakdown point and are said to be *redescending*.

4.1.3. *Relative Weight.* Let's define a *relative weight* function

$$W(x) = \begin{cases} \psi(x)/x & x \neq 0 \\ \psi'(x) & x = 0 \end{cases}$$

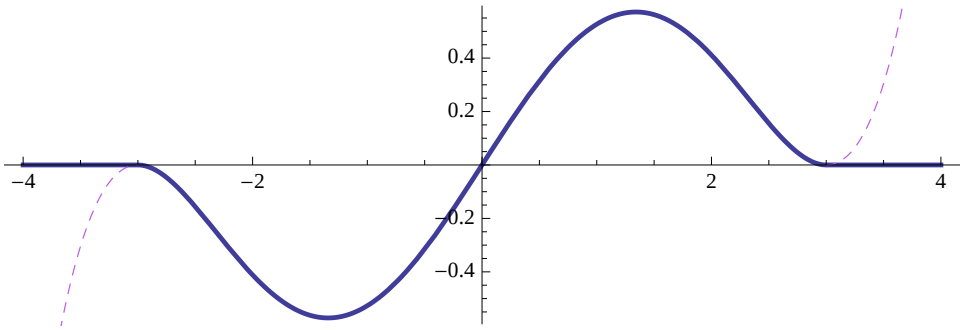


FIGURE 4. Bisquare derivative

or

$$w_i = W(y_i - \mu) \simeq \frac{\psi(y_i - \mu)}{y_i - \mu}.$$

. This function represents how strongly influence increases for a given M-estimator *relative to least-sqaures*.

With this definition our "approximate MLE" is a search for the case where

$$\sum_{i=1}^{n}(y_i - \mu)w_i = 0$$

which is also the computation

$$\hat{\mu} = \frac{\sum w_i y_i}{\sum w_i}$$

The Tukey bisquare has relative weight while the Huber function,

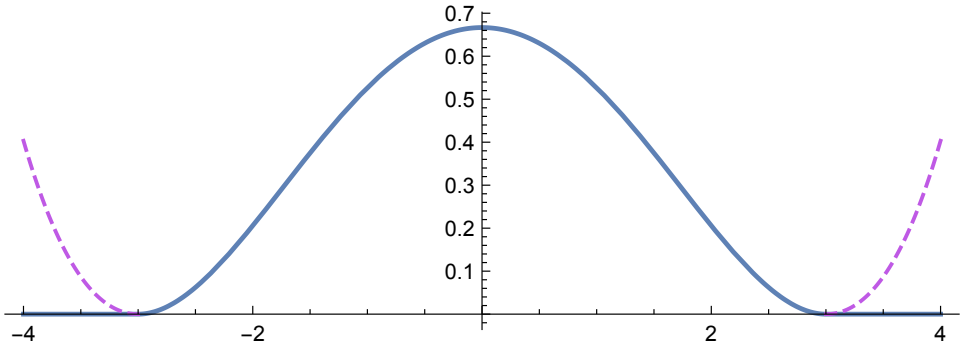


FIGURE 5. Weight of Tukey bisquare relative to Gaussian (solid), with degree 6 continuation (dashed)

which is not redescending, has relative weight that goes as $1/x$ outside the center region.

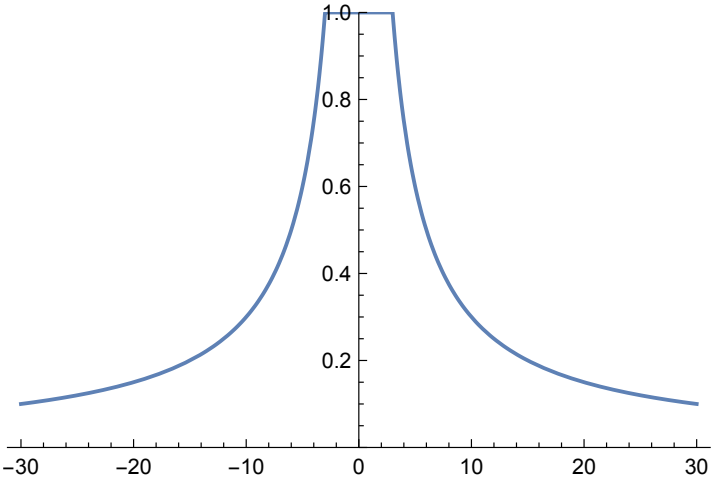


FIGURE 6. Weight of Huber function relative to Gaussian (note wide range)

We'll return to this perspective later when it is time to talk about computational efficiency.

4.2. **M-Estimation of Dispersion.** Returning to measures of dispersion, we can also see many uses for a robust estimate of variability in a data set. Let's say our data has zero location, and hypothesize that it has been multiplied by a *scale parameter* $\sigma$, so

$$y_i = \sigma t_i$$

where the $t_i$ are i.i.d. and come from a distribution $f$. The density of the $y_i$ then depends on $\sigma$, and its maximum likelihood is

$$\hat{\sigma} = \arg\max_{\sigma} \frac{1}{\sigma^n} \prod_1^n f\left(y_i/\sigma\right)$$

Define the function

$$\rho(z) = z\psi(z)$$

from the *logarithmic derivative*

$$\psi(z) = \frac{f'(z)}{f(z)}$$

Note this is different from the $\psi$ we were using for location estimates.

If the distribution $f$ were gaussian $f_G$, we would be computing (up to unimportant constants)

$$\psi_G(z) = \frac{f_G'(z)}{f_G(z)} = \frac{z \exp\left(\frac{1}{2}z^2\right)}{\exp\left(\frac{1}{2}z^2\right)} = z$$

and so

$$\rho_G(z) = z^2.$$

Finding $\hat{\sigma}$ is equivalent to solving

$$\frac{1}{n}\sum_1^n \frac{f_G'\left(y_i/\hat{\sigma}\right)}{f_G\left(y_i/\hat{\sigma}\right)} y_i/\hat{\sigma} = 1$$

or

$$\frac{1}{n}\sum_1^n \rho_G\left(y_i/\hat{\sigma}\right) = 1$$

When we have $f_G$, the standard gaussian, then recall $\rho_G(z) = z^2$ so $\hat{\sigma}$ becomes the RMS.

Any estimate satisfying an equation

$$\frac{1}{n}\sum_1^n \rho\left(y_i/\hat{\sigma}\right) = \delta$$

will be called an *M-estimate of scale*. We allow $\delta$ to differ from one since, in effect, the RMS for non-gaussian distribution is expected to differ from 1.

We mostly choose the *bisquare scale* with

$$\rho(x) = \psi_1^b(x)$$

For scale estimates we take a different definition of weight function

$$W^s(x) = \begin{cases} \psi(x)/x^2 & x \neq 0 \\ \psi''(x) & x = 0 \end{cases}$$

yielding

$$\hat{\sigma}^2 = \frac{1}{n\delta}\sum W^s(y_i/\hat{\sigma})y_i^2$$

We may view $\hat{\sigma}$ as a weighted RMS estimate. For the bisquare scale, we have

$$W^{b,s}(x) = \begin{cases} 3 - 3x^2 + x^4 & |x| \leq 1 \\ 1/x^2 & |x| > 1 \end{cases}$$

4.3. **Fitting Robust Scale With Location.** For large $n$ the approximate distribution of the location estimate

$$\hat{\mu} = \arg\min_{\mu} \sum \rho\left(\frac{y_i - \mu}{\sigma}\right)$$

is

$$v = \sigma^2 \frac{\mathbb{E}\left[\psi(x-\mu)^2/\sigma^2\right]}{\mathbb{E}\left[\psi'(x-\mu)/\sigma\right]^2}$$

An "obvious" choice for fitting is now in front of us. Compute $\hat{\sigma}$ first, then obtain $\hat{\mu}$ as

$$\hat{\mu} = \arg\min_{\mu} \sum \rho\left(\frac{y_i - \mu}{\hat{\sigma}}\right)$$

so that $\hat{\mu}$ solves

$$\sum \psi\left(\frac{y_i - \mu}{\hat{\sigma}}\right) = 0$$

It turns out the efficiency of $\hat{\mu}$ does not depend on that of $\hat{\sigma}$, but its robustness does.

4.4. **Computation.** Assume we have computed a satisfactory starting $\hat{\sigma}$ (perhaps from MADN), and wish to get a new $\hat{\mu}$. Then we can consider the dynamical system defined by iterations of

$$\Lambda(\boldsymbol{w}, \hat{\mu}) = \left\{ W\left(\frac{y_i - \mu}{\hat{\sigma}}\right), \frac{\sum w_i y_i}{\sum w_i} \right\}$$

This will have an attractor at $\hat{\mu}$, with exponential convergence in its neighborhood. Therefore we can define an algorithm with a halting condition based on incremental changes begin sufficiently small. The algorithm for scale is almost exactly the same as for location, based on

$$\sigma_{k+1} = \sqrt{\frac{1}{n\sigma_k} \sum w_i y_i^2}$$

This technique is called *iteratively reweighted least squares*.

4.4.1. *Potential Problems.* There are two major problems with this fitting procedure:

- There may be multiple minima. This can be addressed by choosing decent starting values.
- It can be computationally expensive. This is addressable by working with data subsets (data subsets are an excellent general principle).

## 5. MLE-Like estimates for Regression

Our ability to calculate scale and location in a robust manner now allows us to develop robust estimates of regression coefficients. We take the model

$$y_i = \boldsymbol{x}_i^* \boldsymbol{\beta} + u_i$$

or

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$$

Given a particular choice of $\boldsymbol{\beta}$ we form the *residuals*

$$r_i = r_i(\boldsymbol{\beta}) = y_i - \boldsymbol{x}_i^* \boldsymbol{\beta}.$$

In the common case where we desire a constant term, we will sometimes split out its coefficient $\beta_0$ and use the notation $\boldsymbol{\beta}_1$ to denote the remaining elements of the $\boldsymbol{\beta}$ vector. Our goal will be to obtain an estimate of $\boldsymbol{\beta}$ such that the residuals have zero location and minimal (scaled) influence. That is to say, we want to minimize

$$\sum \rho\left(\frac{r_i(\hat{\boldsymbol{\beta}})}{\sigma}\right)$$

for some $\rho$ function and "auxiliary" scale estimator tuned so as to make our process that estimates $\hat{\boldsymbol{\beta}}$ scale invariant.

Now, due to the influence of the independent variables, our $y_i$ are independent but *not* identically distributed. Given a scale parameter $\sigma$ we take a likelihood driver $f_0$ such the random errors $u_i$ have density

$$u_i \sim \frac{1}{\sigma} f_0\left(\frac{u}{\sigma}\right).$$

Define $\rho_0 = -\log f_0$, and $\psi_0 - \rho_0'$, and observe that a maximum likelihood estimate requires us to maximize

$$\prod f$$

which is to find

$$\arg\min_{\hat{\boldsymbol{\beta}}} \sum \rho_0\left(\frac{r_i(\hat{\boldsymbol{\beta}})}{\sigma}\right)$$

or solve

$$\sum \psi_0\left(\frac{r_i(\hat{\boldsymbol{\beta}})}{\sigma}\right) \boldsymbol{x}_i = 0$$

One example, predating Laplace's study of least squares over 200 years ago by nearly 50 years, is the choice to minimize

$$\sum |r_i|$$

which we saw before corresponds to $f_0$ being the double exponential distribution. In general a *regression M-estimate* is a solution to

$$\sum \psi\left(\frac{r_i(\hat{\boldsymbol{\beta}})}{\hat{\sigma}}\right)\boldsymbol{x}_i = 0 \tag{3}$$

for some choice of $\rho$ function, its derivative $\psi$ and a scale estimator $\hat{\sigma}$. Solutions will not necessarily be unique (just as with bisquare location estimates), leading to the potential problem of non-global minima. However if $\psi$ is monotonic, as with the Huber weights, then the solution *is* unique.

M-estimation begins by computing an L1 fit, i.e.

$$\vec{\beta}^{\mathrm{L1}} = \sum \arg\min_{\boldsymbol{\beta}} |y_i - \boldsymbol{x}_i^* \boldsymbol{\beta}|$$

which can be done quickly using Tukey's 1977 "median polish" algorithm. This gives us a set of initial residuals $r_i^{\mathrm{L1}}$ and an initial scale estimate

$$\sigma^{\mathrm{L1}} = \frac{1}{0.675}\mathrm{Med}\{|r_i| \ni r_i \neq 0\}$$

This is generically unique, having a monotonic $\psi$ from the double exponential. It therefore defines a good starting point to, say, IRLS estimates of a bisquare estimate in Equation (3). Again, if we like, we can use phased IRLS to run an algorithm that also updates our scale estimate at each iteration, solving the system

$$\sum \psi\left(\frac{r_i(\hat{\boldsymbol{\beta}})}{\hat{\sigma}}\right)\boldsymbol{x}_i = 0$$

$$\frac{1}{n}\sum \rho_{\mathrm{scale}}\left(\frac{r_i(\hat{\boldsymbol{\beta}})}{\hat{\sigma}}\right)\boldsymbol{x}_i = \delta$$

## 6. Quantile-Quantile Analysis

One way in which we can think about the type of model it might be reasonable to consider is to determine which variables have similar distributions. To check if two observed variables come from similar distributions, we can reprise our previous use of the normal QQ plot to a *quantile-quantile plot* or *QQ plot* that is empirical in both dimensions.

Even if both distributions are fat-tailed, we can at least try to determine how similar they may be.

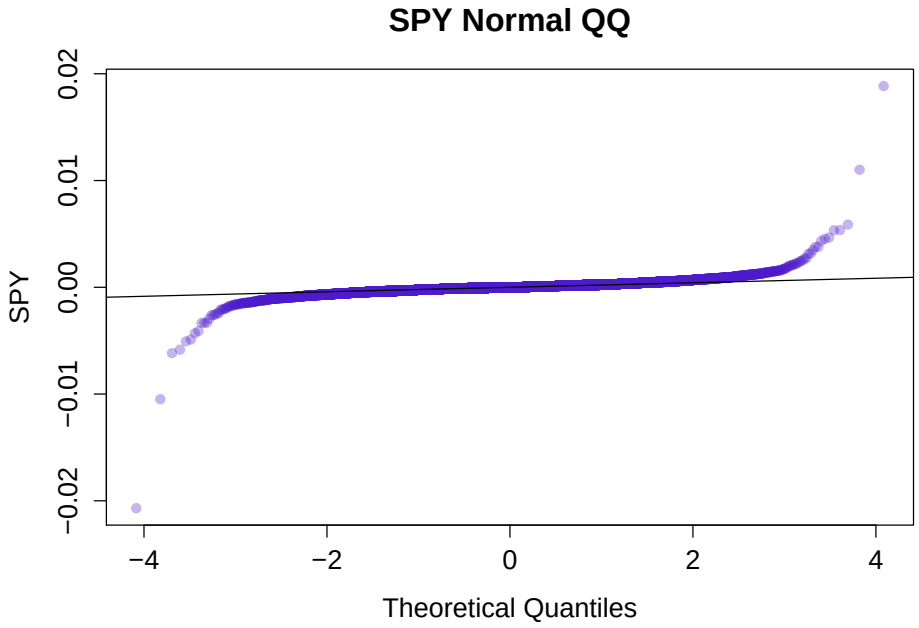


FIGURE 7. SPY has fat tails.

## 7. Outliers in All Directions

The techniques above are concocted to deal well with problematic data in the dependent variable, but not all errors occur there. The $M$-estimates are well-suited to designed experiments, but not to general market data relationships. We also may have the problem of unreliable *independent* variables. If we have rows $\boldsymbol{x}_i$ with erroneous outliers and high leverage, nothing in our previous analysis will help us mitigate their effect.

Now, we'll start with some assumptions that take into account randomness inherent to $\boldsymbol{x}$ samples. First of all, we'll assume not only that the errors $u_i$ were i.i.d. and with a *distribution* independent of the $\boldsymbol{x}_i$ but now that their actual *values* are independent of the $\boldsymbol{x}_i$ (violations would be highlighted in a scale-location plot). This assumption is good enough to ensure that least-squares estimates satisfy

$$\mathbb{E}\left(\boldsymbol{\beta}_{LS} \mid \boldsymbol{X}\right) = \boldsymbol{\beta}$$

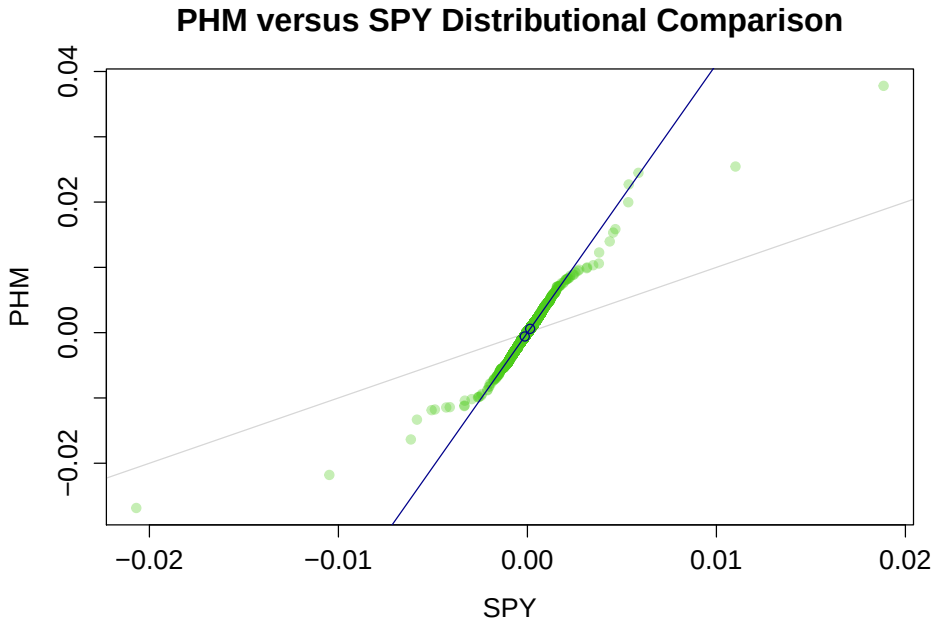


FIGURE 8. Distributional comparison.

Our $u_i$ may not be normal, but if they have variance $\sigma^2$ we define

$$\boldsymbol{V_x} = \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^*)$$

and the asymptotic covariance matrix of $\boldsymbol{\beta}$

$$\boldsymbol{C}_\beta = \sigma^2 \boldsymbol{V_x}^{-1}$$

then the distribution of $\boldsymbol{\beta}_{LS}$ approaches the multivariate normal

$$\boldsymbol{N}_p\left(\boldsymbol{\beta}, \frac{1}{n}\boldsymbol{C}_\beta\right).$$

This is particularly useful because we can use the properties of the outer product to show that covariance matrix of the intercept term $\beta_0$ with the remaining regression coefficients $\boldsymbol{\beta}_1$ asymptotically approaches

$$\sigma^2 \begin{pmatrix} 1 + \boldsymbol{\mu}_x^* \boldsymbol{C}_x^{-1} \boldsymbol{\mu}_x & \boldsymbol{\mu}_x^* \\ \boldsymbol{\mu}_x & \boldsymbol{C}_x^{-1} \end{pmatrix}$$

where $\boldsymbol{\mu}_x = \mathbb{E}(\boldsymbol{x})$ and $\boldsymbol{C}_x = \mathrm{Var}(\boldsymbol{x})$.

7.1. **MM Estimates.** When $\boldsymbol{X}$ is random, the L1 estimate $\sigma^{\mathrm{LS}}$ is no longer a good way of determining scale, because we are likely to have an $\boldsymbol{X}$ outlier with high leverage that determines the fit. Any estimate

we want to make has to have a good starting point in the parameter search algorithm, so what do we do?

We find a starting point by running a bisquare $\rho$ in the simultaneous scale-regression M-estimator. This gives us a reliable scale estimate $\hat{\sigma}$. In practice any scale estimate arising from an bounded M-scale estimator $\rho_0$

$$\min \frac{1}{n} \sum \rho_0 \left( \frac{r_i}{\hat{\sigma}} \right)$$

will do. The process can be sped up by working on subsets and/or using least-trimmed squares techniques. Taking this scale estimate, we call on another bounded $\rho$, again possibly the bisquare, and computing the M-estimate of $\hat{\beta}$ as before. This can be proven to be consistent and efficient.

7.2. **Visually Comparing Fit Quality.** Let's say we have two different model fits. By creating a *residual-residual plot* or *RR plot*, we can obtain a perspective on relative fit quality. Two equally good fits will have all points generally falling on the identity line. Deviations from the line show regions where one fit or the other handles the data better.

By limiting influence of outliers, we more or less guarantee that RR plots will recapitulate the influence function. However, a visual examination of the plot should, if the data is well-fit, not deviate otherwise.

7.3. **Multivariate Fat-tailed Parametric Distributions.** For testing various estimators, it can be useful to have a multivariate fat-tailed distribution. This can be achieved in the marginals with even a gaussian copula, but we can also construct more flexible *matrix-variate Student t* distributions

$$f^{\mathrm{St}}_{\nu,\boldsymbol{\mu},\boldsymbol{M},\boldsymbol{\Sigma},\boldsymbol{S}} = \gamma \|\boldsymbol{\Sigma}\|^{-\frac{N}{2}} \|\boldsymbol{S}\|^{-\frac{N}{2}} \cdot \left| \boldsymbol{I} + \frac{1}{\nu} \boldsymbol{S}^{-1} \left( \boldsymbol{X} - \boldsymbol{M} \right)^{*} \boldsymbol{\Sigma}^{-1} \left( \boldsymbol{X} - \boldsymbol{M} \right) \right|^{\frac{\nu+N}{2}}$$

with

$$\gamma = (\nu\pi)^{-\frac{NK}{2}} \prod_{i=0}^{K-1} \frac{\Gamma\left(\frac{\nu+N-i}{2}\right)}{\Gamma\left(\frac{\nu-i}{2}\right)}$$

These to the case of *vector-variate Student t* distributions when the matrices $\boldsymbol{X}$ and $\boldsymbol{M}$ have just one column.

Now the matrix $\boldsymbol{D}^1$ is easily computed, and it is similarly trivial to find the Cholesky decomposition $\boldsymbol{D}^{-1/2}$ of its inverse, so we obtain

$$\boldsymbol{D}^{-1/2}\boldsymbol{Q}^*\boldsymbol{\Sigma}\boldsymbol{Q}\boldsymbol{D}^{-1/2} = \boldsymbol{I}$$

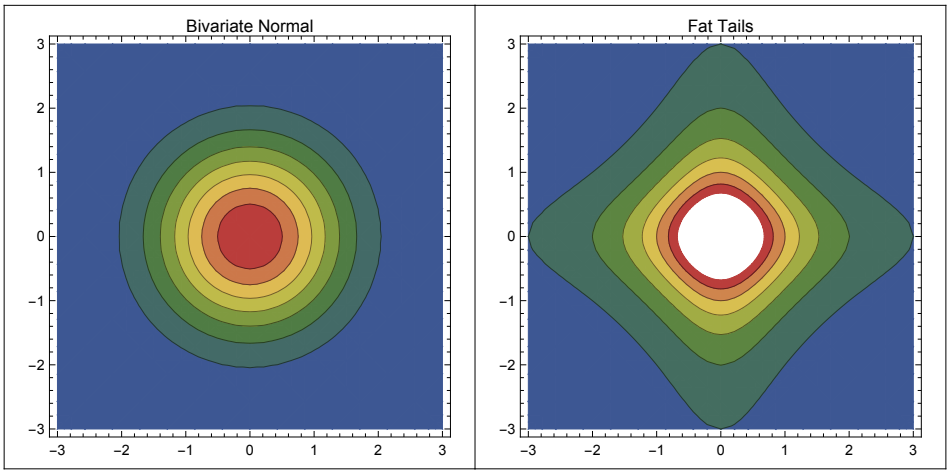


FIGURE 9. The effect of fat tails on bivariate density. Note that simulataneous tail events in the two marginal distribution remain highly unlikely despite the greater probability of extreme events in one or the other margin.

Here the covariance between any two of the $K$ columns is

$$\text{Cov}\left(\boldsymbol{X}^{(j)}, \boldsymbol{X}^{(k)}\right) = \frac{\nu}{\nu - 2} S_{jk}\boldsymbol{\Sigma}$$

and between any two rows is

$$\text{Cov}\left(\boldsymbol{X}_{(m)}, \boldsymbol{X}_{(n)}\right) = \frac{\nu}{\nu - 2}\Sigma_{mn}\boldsymbol{S}$$

## 8. Where Robust Goes Wrong

8.1. **Regions of Interest.** The tradeoff for robust estimation's ability to eliminate or reduce the influence of outliers is that we often care more about extreme values of the data than normal ones. This is most true in risk computations, where quotidian variation in portfolio value matters little either to portfolio managers or to their clients, but large losses are of great concern.

It is notorious in risk management that, at times of market stress, "all correlations go to 1". The consequences for risk modeling are profound, and largely argue *against* using robust techniques in the usual way for estimating tail risk.

This doesn't prevent us from using – indeed almost mirroring – some of the ideas behind robust estimation. For example, consider choosing a $\rho$ function such that the influence of small residuals is relatively small, while that of distant values becomes large. We have to be very careful with this sort of case because

- It is not MLE-like, and the IRLS algorithm will fail
- Problems of domination by outliers can become even worse than for least squares unless we carefully control the weights in the tail region

8.2. **Discrete residuals.** A case with discrete residuals, as when minimum tick sizes come into play, can be hard for continuous models, especially robust regressions, to work with.

8.3. **Regimes.** If we fit a least-squares model to two different economic regimes, one with high volatility and one with low volatility, then the larger residuals will mostly come from the high-volatility region. Thus the regression will perform poorly as a predictor in low-volatility cases. The opposite tends to be true for robust regressions, whose underweighting of the greater residuals in the high-volatility region will tend to cause the fit to conform to the characteristics of the low-volatility regime.

To deal with regimes, our main approach is to try to adapt to them. For example an *exponentially-weighted moving average* or *EWMA* will behave well. In return space, we can often ignore location, allowing our EWMA to specify

$$\beta^{(t)} = \lambda\beta^{(t-1)} + (1-\lambda)\frac{r_t^y}{r_t^x}$$

8.4. **Wide Variation.** Ledoit-Wolf style estimators may be faced with huge variation, causing them to converge to an implausible constant correlation matrix.

## 9. Whitening the Data

If we wish to apply a multidimensional smoothing kernel, or separate out price influences that we believe to be unrelated, it is often useful to work in a space where our data is, at least approximately, uniformly distributed in every dimension.

Assume we have a covariance matrix $\mathbf{\Sigma}$ with a reasonable condition number. For the general case we can compute a *Schur decomposition* as the unitary matrix $\boldsymbol{Q}$ such that

$$\boldsymbol{Q}^*\mathbf{\Sigma}\boldsymbol{Q} = \boldsymbol{D}$$

for some diagonal $\boldsymbol{D}$ where we find $\boldsymbol{Q}$ using Householder matrices accumulated to *Hessenberg decompositions*. Since in our case the matrix is symmetric, we can instead reduce $\boldsymbol{\Sigma}$ to *tridiagonal* form using Householder matrices to treat rows and columns simultaneously.

$$\begin{pmatrix} b_1 & c_1 & & & & & \\ a_2 & b_2 & c_2 & & & & \\ & a_3 & b_3 & c_3 & & & \\ & & & & & & \\ & & & & & & \\ & & & a_{n-2} & b_{n-2} & c_{n-2} & \\ & & & & a_{n-1} & b_{n-1} & c_{n-1} \\ & & & & & a_n & b_n \end{pmatrix}$$

Diagonalizing a tridiagonal matrix like this can be quickly performed with a set of Givens rotations.

Now if we want to apply a smoothing kernel on "dark" data $\boldsymbol{x}$. Then if we define

$$\boldsymbol{w} = \boldsymbol{D}^{-1/2}\boldsymbol{Q}^*\boldsymbol{x}$$

then $\boldsymbol{w}$ has zero mean. We can apply our kernel, and then (typically) transform back into the original space using

$$\boldsymbol{x} = \boldsymbol{Q}\boldsymbol{D}^{1/2}\boldsymbol{w}$$

## 10. IWLS Convergence Proof

A function $f$ is *convex* if $f(\lambda x + (1-\lambda y)) \leq \lambda f(x) + (1-\lambda)f(y)$ when $\alpha$ is in the unit interval

A *$\rho$-function* is a function $\rho$ such that

- $\rho(x)$ is nondecreasing in $|x|$
- $\rho(0) = 0$
- $\forall x \ni \rho(x) < \rho(\infty), \rho$ is increasing at $x$ if $x > 0$, decreasing otherwise
- $\rho$ is either unbounded or $\rho(\infty) = 1$

If $\rho(\infty) = 1$ then we say $\rho$ is a *bounded $\rho$-function* . A *$\psi$-function* is a generalized derivative of a $\rho$-function. The weight function is

$$W(x) = \psi(x)/x$$

except at zero (where we take a limit).

Consider our goal of satisfying that

$$\sum \psi\left(\frac{y_i - \mu}{\hat{\sigma}}\right) = 0$$

which is locating extrema of

$$h(\hat{\mu}) = \sum \rho\left(\frac{y_i - \mu}{\hat{\sigma}}\right)$$

Let $G(r) = \rho(\sqrt{r})$, so that with $\psi$ as the derivative of $\rho$

$$W(r) = 2G'(r^2)$$

which implies $W$ is nonincreasing if and only if $G'$ is. We will assume $W$ is nonincreasing (the algorithm can fail otherwise), hence $G$ is nonincreasing and concave.

Define

$$w_i = w_i(\mu) = W\left(\frac{(y_i - \mu_k)}{\hat{\sigma}}\right)$$

and

$$U(\mu) = \sum w_i(\mu)$$

which is never negative.

Consider the function

$$f(\mu) = \frac{\sum w_i(\mu) y_i}{\sum w_i(\mu)}$$

then we have our algorithm defined by

$$\mu_{k+1} = f(\mu_k)$$

so that

$$U(\mu_k)\mu_{k+1} = \sum w_i \mu_{k+1} = \sum w_i y_i$$

(In a more general setting, we set

$$U(\boldsymbol{\beta}) = \arg\min_{\gamma} \sum W\left(\frac{(y_i - \boldsymbol{x}_i^*\boldsymbol{\beta})}{\sigma}\right) \boldsymbol{x}_i \boldsymbol{x}_i^*$$

and

$$f(\boldsymbol{\beta}) = \arg\min_{\gamma} \sum W\left(\frac{(y_i - \boldsymbol{x}_i^*\boldsymbol{\beta})}{\sigma}\right) (y_i - \boldsymbol{x}_i^*\gamma)^2$$

which is a least-squares problem, giving the technique its IRLS name )

Now concavity of $G'$ implies

$$G'(y) \leq G(x) + G'(x)(y - x)$$

so, along with $W(2) = 2G'(r^2)$ taking $y = (y_i - \mu_{k+1})^2$ and similarly for $x$ we obtain

$$\begin{aligned} h(\hat{\mu}_{k+1}) - h(\hat{\mu}_k) &\leq \frac{1}{\sigma^2} \sum G' \left( \frac{(y_i - \mu_k)^2}{\hat{\sigma}} \right) \left( (y_i - \mu_{k+1})^2 - (y_i - \mu_k)^2 \right) \\ &= \frac{1}{2\sigma^2} \sum w_i \left( (y_i - \mu_{k+1}) + (y_i - \mu_k) \right) \left( (y_i - \mu_{k+1}) - (y_i - \mu_k) \right) \end{aligned}$$

These terms involve a difference of the $\mu_k$, and

$$(y_i - \mu_{k+1}) + (y_i - \mu_k) = 2y_i - (\mu_{k+1} + \mu_k)$$

so we get

$$\begin{aligned} h(\hat{\mu}_{k+1}) - h(\hat{\mu}_k) &= \frac{1}{2\sigma^2} (\mu_k - \mu_{k+1}) \sum w_i (2y_i - \mu_k - \mu_{k+1})) \\ &= \frac{1}{2\sigma^2} (\mu_k - \mu_{k+1}) U(\mu_k) (\mu_{k+1} - \mu_k) \\ &\leq 0 \end{aligned}$$

Therefore

$$h(\hat{\mu}_{k+1}) \leq h(\hat{\mu}_k)$$

Since $h$ is bounded below it has a limit $\hat{\mu} = \mu_\infty$.