Fooled by Correlation: Common Misinterpretations in Social "Science"

Nassim Nicholas Taleb
March 2019

***Abstract*—We present consequential mistakes in uses of correlation in social science research:**

1) **use of subsampling since (absolute) correlation is severely subadditive**
2) **misinterpretation of the informational value of correlation owing to nonlinearities,**
3) **misapplication of correlation and PCA/Factor analysis when the relationship between variables is nonlinear,**
4) **How to embody sampling error of the input variable**
5) **Intransitivity of correlation**
6) **Other similar problems mostly focused on psychometrics (IQ testing is infected by the "dead man bias")**
7) **How fat tails cause R^2 to be fake.**

We compare to the more robust entropy approaches.

Contents

I. Correlation is subaditive (in absolute value)

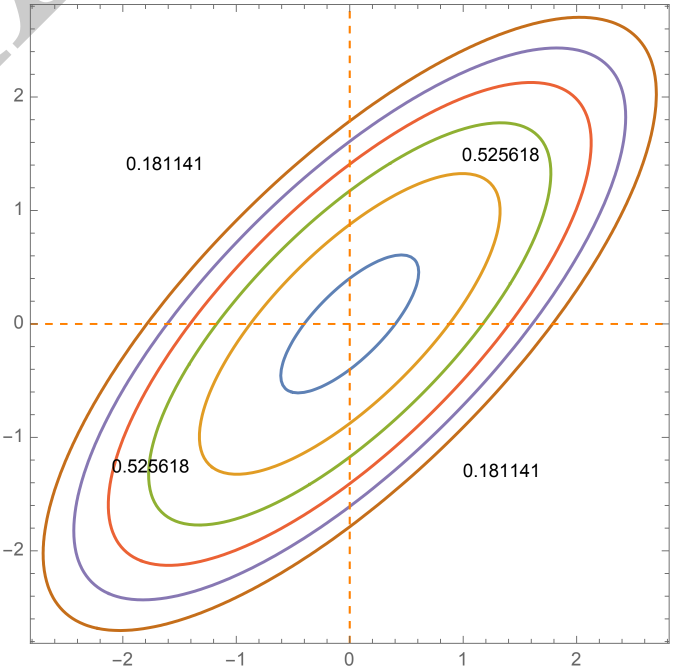

Fig. 1. Total correlation is .75, but quadrant correlations are .52 (second and fourth quadrant) and .18 (first and third). If in turn we make the "quadrants" smaller, say the 2^{nd} one into $\mathcal{Q} = (0,2),(0,2)$, correlation will be even lowe, ≈ .38 (next figure).

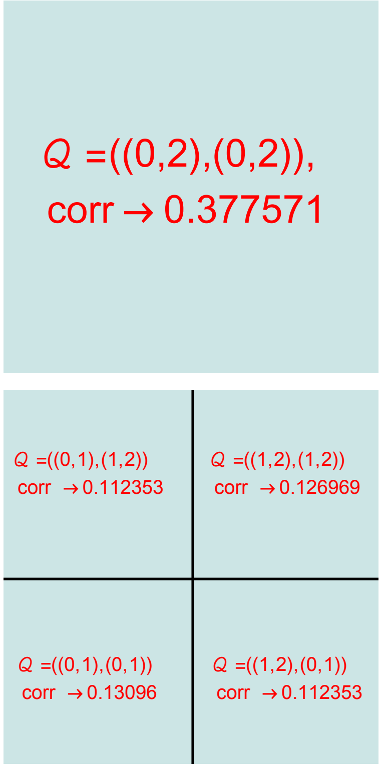

Fig. 2. Dividing the space into smaller and smaller squares yields lower correlations

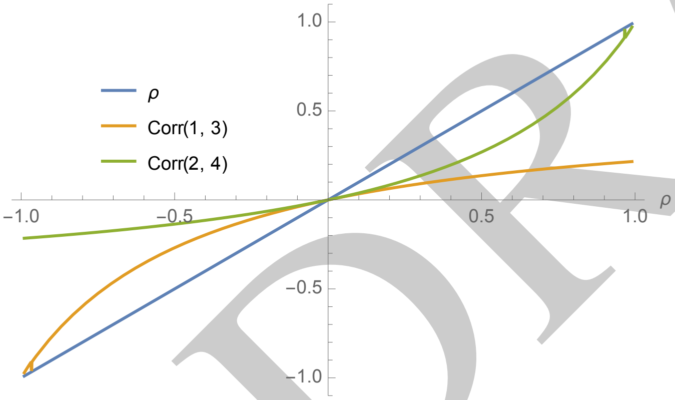

Fig. 3. Total correlation and the corresponding ones in the 4 quadrants.

Rule 1: Subadditivity

Correlation cannot be used for nonrandom subsamples.

Let X, Y be normalized random variables, Gaussian distributed with correlation ρ and pdf $f(x,y)$. If we sample *randomly* from the distribution and break it up further into *random* sub-samples, then, under adequate conditions, the expected correlation of each sub-sample should, obviously, converge to ρ.

However, should we break up the data into non random subsamples, say quadrants, octants, etc. along the x and y axes, as in Fig. 1 and measure the correlation in each square, we end up with considerably lower (probability) weighted sums of individual correlations in absolute value, with equality for $|\rho|= 0$ and, for some cases, $|\rho|= 1$.

Consider 4 equal quadrants, as in Fig. 1 the correlation is .75 but quadrant correlations have for value .52 and .18.

Let $\mathbb{1}_{x,y\in\mathcal{Q}}$ be an indicator function taking value 1 if both x and y are in a square partition $\mathcal{Q}$ and 0 otherwise. Let $\pi(\mathcal{Q})$ be the probability of being in partition $\mathcal{Q}$,

$$\pi(\mathcal{Q}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbb{1}_{x,y\in\mathcal{Q}} f(x,y)dydx.$$

μ_x the conditional mean for x when both x and y are in $\mathcal{Q}$ (and and the same for μ_y):

$$\mu_x(\mathcal{Q}) = \frac{1}{\pi(\mathcal{Q})}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\mathbb{1}_{x,y\in\mathcal{Q}} f(x,y)dydx$$

$$\mu_y(\mathcal{Q}) = \frac{1}{\pi(\mathcal{Q})}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y\mathbb{1}_{x,y\in\mathcal{Q}} f(x,y)dydx$$

$v_.$ is the conditional variance, and cov(.,.) the conditional covariance.

$$v_x(\mathcal{Q}) = \frac{1}{\pi(\mathcal{Q})}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbb{1}_{x,y\in\mathcal{Q}} f(x,y)(x-\mu_x(\rho,\mathcal{Q}))^2dydx$$

$$Cov_{x,y}(\mathcal{Q}) = \frac{1}{\pi(\mathcal{Q})}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbb{1}_{x,y\in\mathcal{Q}} f(x,y)(x-\mu_x(\mathcal{Q}))(y - \mu_y(\mathcal{Q}))dydx$$

Finally, the local correlation:

$$Corr(\mathcal{Q}) = \frac{Cov_{x,y}(\mathcal{Q})}{\sqrt{v_x(\mathcal{Q})v_y(\mathcal{Q})}}$$

Theorem 1

For all $\mathcal{Q}$ in $\mathbb{R}^2$, we have

$$|Corr(\mathcal{Q})| \leq |\rho|$$

Proof. Appendix. □

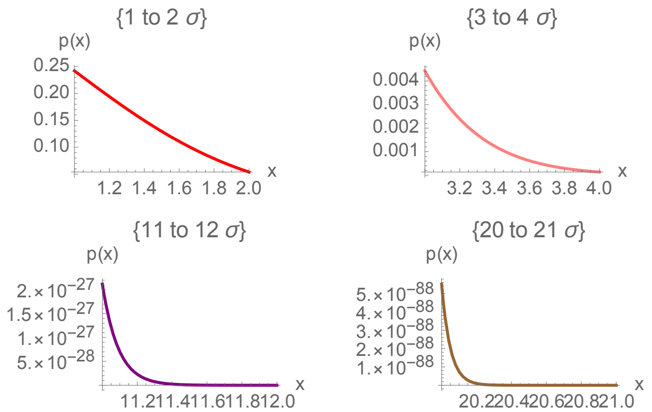

Fig. 4. As we sample in blocks in the tails separated by 1 standard deviation on the x axis, we observe a drop in standard deviation as the Gaussian distribution concentrates in the left side of the partition as we go further in the tails. Power laws have an opposite behavior.

A. Intuition via one-dimensional representations

The problem becomes much easier when we consider the behavior in lower dimensions –for Gaussian variables.

The intuition is as follows. Take a sample of X, a Normalized Gaussian random variable. Verify that the variance is 1. Divide the data into positive and negative. Each will have a conditional variance of $1-\frac{2}{\pi}=\approx 0.363$. Divide the segments further, and there will be additional drop in variance.

And, although one is programmed to think that the tail should be more volatile, it isn't so; the segments in the tail have an increasingly lower variance as one gets further away, see in Fig. 4.

Rule 2

Variance is superadditive for the subexponential class, and subadditive outside of it.

Let $p(x)$ be the density of the Normalized Gaussian, $a, b \in \mathbb{R}, a<b$

$$v(a,b)=\frac{1}{P(a,b)}\int_a^b p(x)(x-\mu(a,b))^2\,dx, \tag{1}$$

where

$$\begin{aligned} P(a,b) &= \int_{-\infty}^{\infty} p(x)\mathbb{1}_{a<x<b}\,dx \\ &= \frac{1}{2}\left(\operatorname{erf}\left(\frac{b}{\sqrt{2}}\right)-\operatorname{erf}\left(\frac{a}{\sqrt{2}}\right)\right), \end{aligned} \tag{2}$$

$$\mu(a,b)=\frac{\sqrt{\frac{2}{\pi}}e^{-\frac{a^2}{2}-\frac{b^2}{2}}\left(e^{\frac{b^2}{2}}-e^{\frac{a^2}{2}}\right)}{\operatorname{erf}\left(\frac{b}{\sqrt{2}}\right)-\operatorname{erf}\left(\frac{a}{\sqrt{2}}\right)}. \tag{3}$$

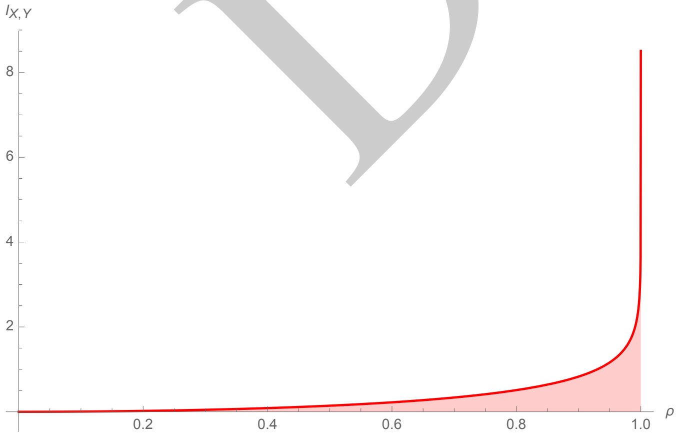

Fig. 5. Mutual Information is a nonlinear function of ρ which in fact makes it additive. Intuitively, in the Gaussian case, ρ should never be interpreted linearly: a ρ of $\frac{1}{2}$ carries ≈ 4.5 times the information of a $\rho=\frac{1}{4}$, and a ρ of $\frac{3}{4}$ 12.8 times!

B. Mutual Information is Additive

We define $I_{X,Y}$ the mutual information between r.v.s X and Y.

$$I_{X,Y}=\int_{\mathcal{D}_X}\int_{\mathcal{D}_Y} f(x,y)\log\left(\frac{f(x,y)}{f(x)f(y)}\right)\,\mathrm{d}x\,\mathrm{d}y \tag{4}$$

and of course

$$\log\frac{f(x,y)}{f(x)f(y)}=\log\frac{f(x|y)f(y)}{f(x)}=\log\frac{f(y|x)f(x)}{f(y)}$$

Theorem 2

$I_{X,Y}$ is additive across partitions of $\mathcal{D}_X$ and $\mathcal{D}_Y$.

Proof. The result is immediate. We have:
$I_{X,Y}=\mathbb{E}\left(\log f(x,y)\right)-\mathbb{E}\left(\log f(x)\right)-\mathbb{E}\left(\log f(y)\right)$. Consider the additivity of measures on subintervals. □

C. Example of Quadrants

Assume we are, as before, in a situation where X and Y follow a standardized bivariate Gaussian distribution with correlation ρ –and let's compare to the results shown in Fig. 1.

Breaking $I_{X,Y}$ in 4 quadrants:

$$I_{x<0,y\geq 0} = \frac{1}{P_{x<0,y\geq 0}}\left(-\frac{2\sqrt{1-\rho^2}\rho+\log\left(1-\rho^2\right)\cos^{-1}(\rho)}{4\pi}\right) \tag{5}$$

$$I_{x\geq 0,y\geq 0} = \frac{1}{P_{x\geq 0,y\geq 0}}\frac{2\sqrt{1-\rho^2}\rho+\log\left(1-\rho^2\right)\left(\cos^{-1}(\rho)-\pi\right)}{4\pi} \tag{6}$$

$$I_{x\geq 0,y<0} = \frac{1}{P_{x\geq 0,y<0}}\left(\frac{-2\rho\sqrt{1-\rho^2}+i\log\left(1-\rho^2\right)\cosh^{-1}(\rho)}{4\pi}\right) \tag{7}$$

$$I_{x<0,y<0} = \frac{1}{P_{x<0,y<0}}\frac{4\rho\sqrt{1-\rho^2}-\log\left(1-\rho^2\right)\left(2\sin^{-1}(\rho)+\pi\right)}{8\pi} \tag{8}$$

We can see that

$$\begin{aligned} &P_{x<0,y\geq 0}I_{x<0,y\geq 0}+P_{x\geq 0,y\geq 0}I_{x\geq 0,y\geq 0} \\ &\quad +P_{x\geq 0,y<0}I_{x\geq 0,y<0}+P_{x<0,y<0}I_{x<0,y<0} \\ &\qquad = -\frac{1}{2}\log\left(1-\rho^2\right) \end{aligned} \tag{9}$$

$$\begin{array}{cc} -\frac{2\sqrt{1-\rho^2}\rho+\log\left(1-\rho^2\right)\cos^{-1}(\rho)}{4\pi} & \frac{2\sqrt{1-\rho^2}\rho+\log\left(1-\rho^2\right)\left(\cos^{-1}(\rho)-\pi\right)}{4\pi} \\ \frac{-2\rho\sqrt{1-\rho^2}+i\log\left(1-\rho^2\right)\cosh^{-1}(\rho)}{4\pi} & \frac{4\rho\sqrt{1-\rho^2}-\log\left(1-\rho^2\right)\left(2\sin^{-1}(\rho)+\pi\right)}{8\pi} \end{array}$$

II. Rescaling: A 50% correlation doesn't mean what you think it means

What does a 50% correlation mean? Not much, which shows that perhaps much of social science has little scientific significance outside citation rings and political agendas.

Rule 3

Correlation should never be interpreted linearly without translation via some rescaling.

In [1] it has been shown that great many econometricians, while knowing their statistical equations down pat, don't get the real practical implication –all in one direction, the *fooled by randomness* one. The authors has a version of the effect in [2] as professionals and graduate students failed to realize that they interpreted mean deviation as standard deviation, therefore underestimating volatility, especially under fat tails. That 70 pct. of econometricians misinterpreted their own data is quite telling.

There are clearly some cognitive limitations, compounded by the specificity and scaling of the correlation metric. A .5 correlation is vastly inferior to, say, .7 and the information is worse that $\times\frac{5}{7}$ of the latter; there needs to be a more idiot-proof (psychologist-proof) rescaling to compare the two. Actually a .5 correlation has between .06 and .14 information if a 1 correlation conveys an information content of 1 and 0 correlation one of 0.

Clearly, it is erroneous to look at correlation without some change of metri c –rescaling –to allow for relative interpretation.

We will examine the following rescaling methods for a vector $(X,Y)^T \in \mathbb{R}^2$:

1) Conditional standard deviation $\sqrt{\mathbb{V}(X|Y)} \in [0,1]$ to accommodate distances between $\mathbb{E}(X|Y)$ and $\mathbb{E}(X)$.
2) The ϕ metric we derive here, in $[0,1]$, to accommodate distances between $\mathbb{E}(X|Y)$ and $\mathbb{E}(Y)$.
3) The more rigorous mutual information, unbounded, in $[0,\infty)$.
4) The p-Mutual information, bounded to allow for comparisons with others in $[0,1]$; $1-p$ certainty would equivalent to 1. For instant p could be $\frac{1}{99}$, with corresponding a definition of "certainty" of .99, or $\frac{1}{999}$ for other applications.

A. Variance method

The conditional mean for a multivariate Gaussian is:

$$\mathbb{E}(X|Y) = \mathbb{E}[X] + \frac{\sigma_1}{\sigma_2}\rho(y - \mathbb{E}[[Y]]) \tag{10}$$

We have a bivariate gaussian with means μ_1 and μ_2, variances σ_1^2 and σ_2^2, and correlation ρ. The joint distribution $f(x,y)$: $f(x,y) = \frac{e^{-\frac{\frac{(x-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2}}{2(1-\rho^2)}}}{2\pi\sqrt{1-\rho^2}\sigma_1\sigma_2}$ and $\mathbb{E}(X|Y) = \frac{\int_{-\infty}^{\infty} x f(x,y)\,dx}{\int_{-\infty}^{\infty} f(x,y)\,dx} = \mu_1 + \frac{\sigma_1\rho(y-\mu_2)}{\sigma_2}$.

The conditional variance for a multivariate Gaussian:

$$\mathbb{V}(X|Y) = \mathbb{E}\left(\left(X - \mathbb{E}(X|Y)\right)^2 \middle| Y\right) = \left(1-\rho^2\right)\sigma_1^2 \tag{11}$$

Which means that the expected value of X given Y is normally distributed.

$$\mathbb{E}(X|Y) \sim \mathcal{N}\Bigg(\mathbb{E}[X] + \sqrt{\frac{\mathbb{V}(X)}{\mathbb{V}(Y)}}\rho(y - \mathbb{E}[[Y]]), \left(1-\rho^2\right)\mathbb{V}(X)\Bigg) \tag{12}$$

We measure the certainty when $\mathbb{E}(X|Y)$ is degenerate at $E(X)$, which requires $\rho = 1$. Hence, for normalized variables, the rescaling metric becomes:

$$R(\rho) = 1 - \sqrt{\mathbb{V}(X|Y)} = 1 - \sqrt{(1-\rho^2)} \tag{13}$$

for Gaussian variables.

1) Drawback: The problem with such a metric is that it ignores the (normalized) distance between X and Y, ignoring the "similarity" between the two variables, focusing only on its variance given a certain information.

2) Adjusted variance method: To get more information we adjust Eq. 13 by the coefficient of similarity, using correlation as a distance. It is similar to the ϕ function but not targeted to specific intervals.

$$R(\rho)_a = |\rho|\left(1 - \sqrt{(1-\rho^2)}\right) \tag{14}$$

B. The ϕ function

Next we create a "proportion of normalized similarity" between two random variables.

Let X and Y be normalized random variables. Consider the ratio of the probability of both X and Y being in an interval $[K-\Delta, K+\Delta]$ under a correlation structure ρ, over the probability of both X and Y being in same interval assuming correlation $= 1$. The function ϕ is the "proportion of normalized similarity" for Y given X. Note, unlike with the conditional variance approach, we measure the certainty when $\mathbb{E}(X|Y)$ is degenerate at $E(Y)$ (instead of $E(X)$). Hence, for

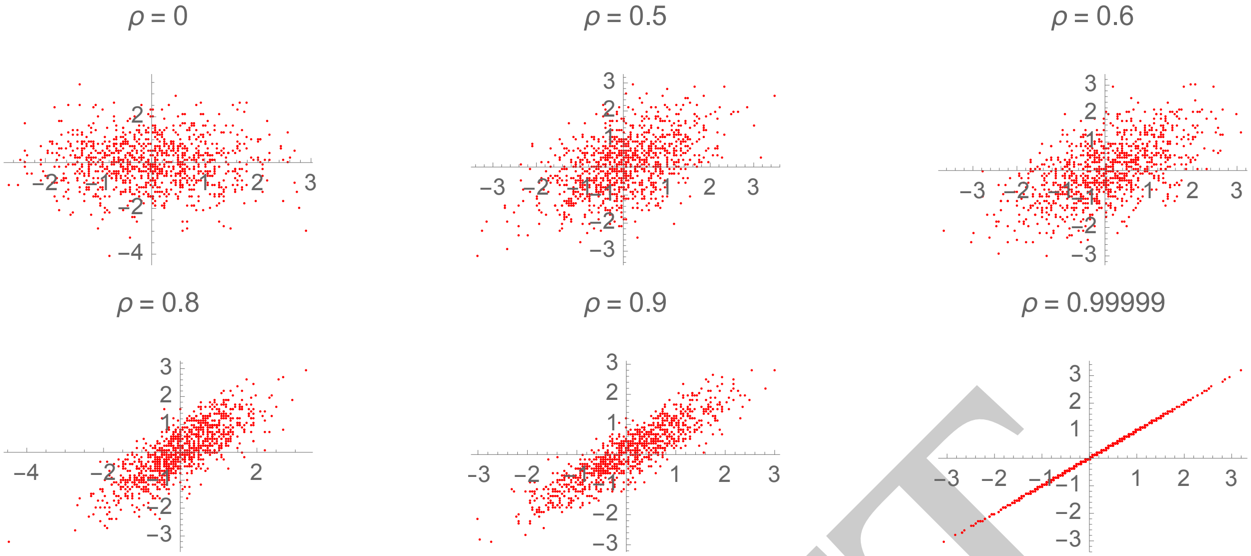

Fig. 6. One needs to translate ρ into information. See how $\rho = .5$ is much closer to 0 than to a $\rho = 1$. There are considerable differences between .9 and .99

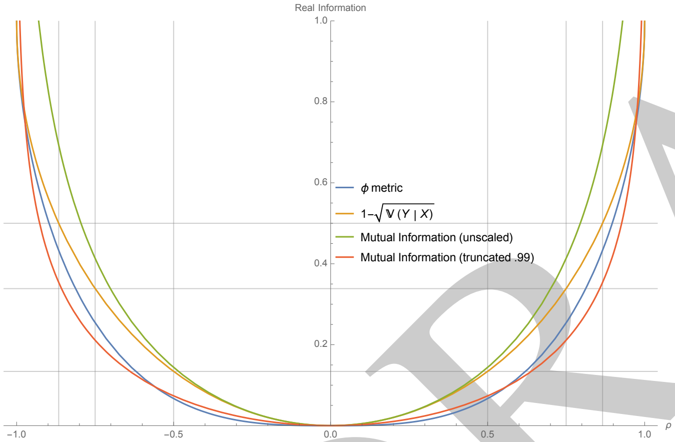

Fig. 7. Various rescaling methods,linerarizing information and putting correlation in perspective.

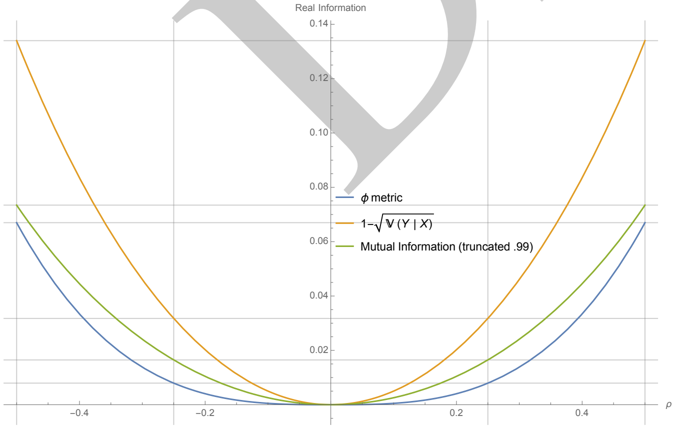

Fig. 8. Various rescaling methods, seen around $[0, |\frac{1}{2}|]$

normalized variables, the rescaling metric becomes:

$$\begin{aligned}
&\phi(\rho, K) \\
&\quad = \frac{\mathbb{P}(X \in (K-\Delta, K+\Delta) \wedge Y \in (K-\Delta, K+\Delta))|_{\rho}}{\mathbb{P}(X \in (K-\Delta, K+\Delta) \wedge Y \in (K-\Delta, K+\Delta))|_{\rho=1}} \\
&\quad = \frac{\mathbb{P}(X \in (K-\Delta, K+\Delta) \wedge Y \in (K-\Delta, K+\Delta))}{\mathbb{P}(X \in (K-\Delta, K+\Delta))}
\end{aligned}$$

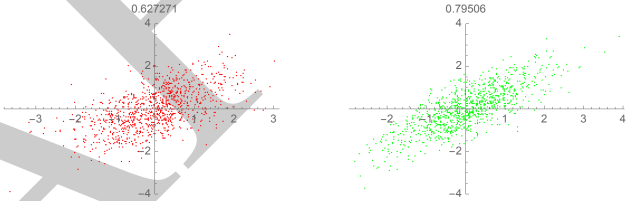

Fig. 9. Correlation for twice the mutual information.

The numerator

$$\int_{K-\Delta}^{\Delta+K} \int_{K-\Delta}^{\Delta+K} \frac{e^{-\frac{x^2-2\rho x y+y^2}{2\left(1-\rho^2\right)}}}{2\pi\sqrt{1-\rho^2}}dxdy$$

does not integrate, which necessitates numerical methods.

C. Mutual Information

As we saw above, $I_{X,Y}$ the mutual information between r.v.s X and Y and joint PDF $f(.,.)$, because of its additive properties, allows a better representation of relative correlations, via the rescaling function $-\frac{1}{2}\log\left(1-\rho^2\right)$. Such rescaling function doesn't apply in all situations and should be used as a translator in a limited way.

Mutual information is both additive and able to detect nonlinearities. In Fig.11, $I_{X,Y} > -\frac{1}{2}\log\left(1-\rho^2\right)$.

D. PCA with Mutual Information

Now one can perform information based PCA maps, if the Data is Gaussian, by rescaling substituting the performance.

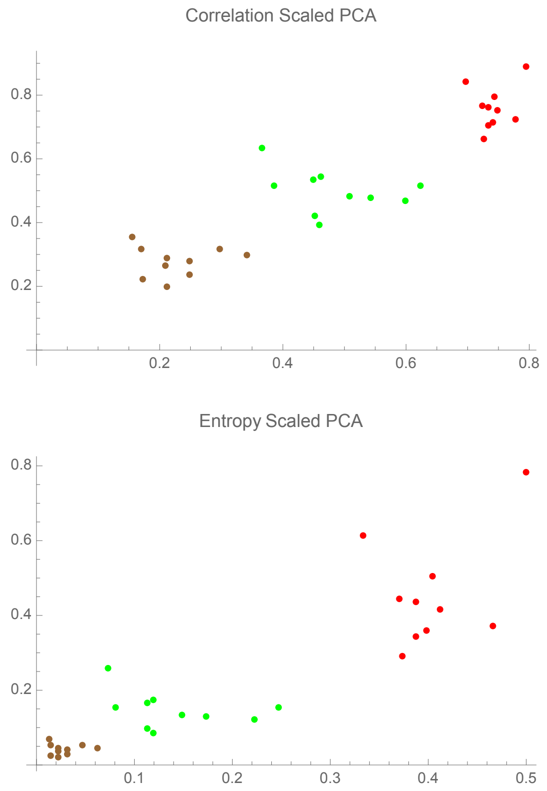

Fig. 10. Entropy rescaled principal component analysis changes the relative distances

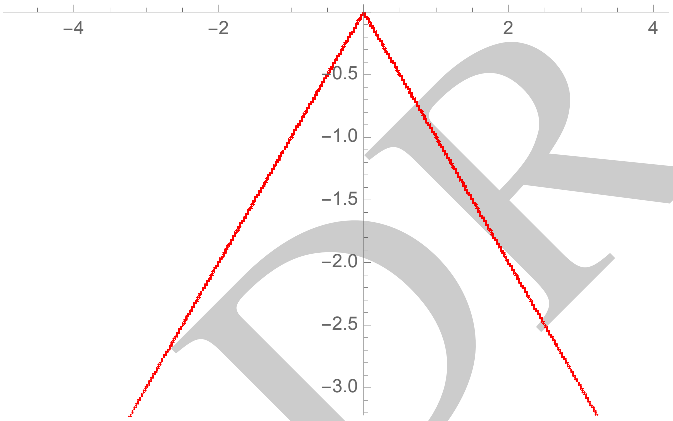

Fig. 11. The function $y = \mathbb{1}_{x \leq 0} x - \mathbb{1}_{x>0} x$. Correlation here between x and y is 0, but mutual information isn't fooled (it is maximal, or what is called infinite). Most (if not all) paradoxes of dependence with correlation disappear with mutual information.

III. Embedding Measurement Error

Next we perform a new trick for error propagation under Gaussian errors and multivariate Gaussian correlation.

Assume IQ (X) correlates with performance P (Y) with coefficient ρ. (Ignore for now the circularity). A certain individual's score, Z has a standard deviation of κ in his or her tests score. (In other words, his or her performance on test is normally distributed with mean X (a random variable) and variance κ^2) What is the covariance/correlation between the score Z and performance P, that is between Z and Y?

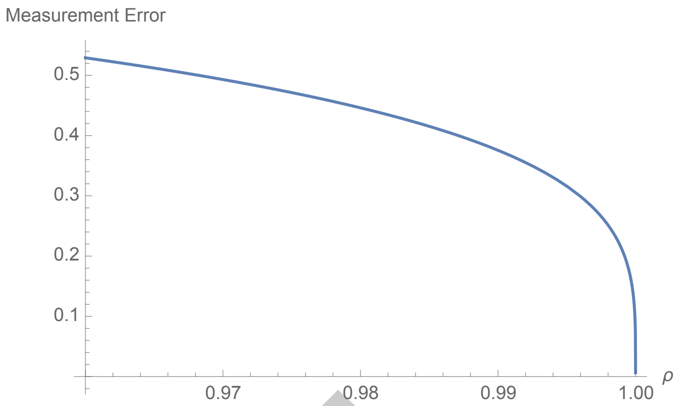

Fig. 12. Translating correlation into measurement error expressed in standard deviation. Consider that IQ testing has an 80% correlation between test and retest.

Let $g(\mu, \sigma; x)$ be the PDF of the NormalDistribution with mean μ and variance σ^2, and $f(.,.)$ the joint distribution for a multivariate Gaussian.

$$\begin{aligned}\rho' &= \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u y f(u,y) g(x,\kappa,u) du dx dy}{\sqrt{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u^2 \left(g\left(0,\sigma_1,x\right) g(x,\kappa,u)\right) dx du}} \\ &\quad \frac{1}{\sqrt{\left(\int_{-\infty}^{\infty} y^2 \left(g\left(0,\sigma_2,y\right) dy\right)}} \\ &= \frac{\rho}{\sqrt{\kappa^2+1}}\end{aligned} \tag{16}$$

Another approach. When psychometricians measure IQ (which varies for the same individual between test and retest) and correlate it to performance, the noise between individuals is embedded in the correlation (assuming of course linearity and state-independence of the correlation, which is not usually the case).

However the psychotards miss the notion of effect: for a single individual, the noise around one's IQ can vastly swamp the effect from correlation! See Fig. 12.

The other problem is that psychometricians and psychologists work with correlation, when the real product is covariance.

As we saw $\mathbb{E}(X|Y) = \mathbb{E}\Big[[X] + \frac{\sigma_Y}{\sigma_X}\rho(y - \mathbb{E}[[Y]])$, so if someone takes an IQ test and gets 1 std away from the mean, the expected result is .8 std away from the mean. Completely missed by the psychotards. But it gets worse via the transitivity problem.

IV. Transitivity of Correlations

Eugenists: I spotted another error by eugenists (a trend self-styled "race realism" found in psychology particularly in evolutionary psychology and behavioral genetics, a field that collects rejects and seems to have, on the good day, the rigor of astrology). They don't seem to have much going for them: the error below is pervasive.

They make the following inference:

(i) There is a positive correlation between genetics and IQ scores.

(ii) There is a positive correlation between IQ scores and performance.

(iii) Hence there is a positive correlation between genetics and performance.

Problem is that (iii) doesn't flow from (i) and (ii). You can have the first two correlations positive and the third one negative.

Let us organize the correlations pairwise, where $\rho_{12}, \rho_{13}, \rho_{23}$ are indexed by 1 for genes, 2 for scores, and 3 for performance. Let $\sigma^2_{(.)}$ be the respective variances. Let Σ be the covariance matrix, without specifying the distribution:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1\sigma_3 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 \\ \rho_{13}\sigma_1\sigma_3 & \rho_{23}\sigma_2\sigma_3 & \sigma_3^2 \end{pmatrix}.$$

Let us apply Sylvester' s criterion (a necessary and sufficient criterion to determine whether a Hermitian matrix is positive-semidefinite) and, using determinants, produce conditions for the positive-definite-ness of Σ. The criterion states that a Hermitian matrix M is positive-semidefinite if and only if the leading principal minors are nonnegative.

The constraint on the first principal minor is obvious (Cauchy-Schwarz):

$$\begin{vmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{vmatrix} = \sigma_1^2\sigma_2^2 - \rho_{12}^2\sigma_1^2\sigma_2^2 \geq 0,$$

so

$$-1 \leq \rho_{12} \leq 1.$$

The second constraint:

$$|\Sigma| = -\left(\rho_{12}^2 - 2\rho_{13}\rho_{23}\rho_{12} + \rho_{13}^2 + \rho_{23}^2 - 1\right)\sigma_1^2\sigma_2^2\sigma_3^2 \geq 0,$$

produces the following bounds on ρ_{13}:

$$\rho_{12}\rho_{23} - \sqrt{\left(\rho_{12}^2 - 1\right)\left(\rho_{23}^2 - 1\right)} \leq \rho_{13} \leq \sqrt{\left(\rho_{12}^2 - 1\right)\left(\rho_{23}^2 - 1\right)} + \rho_{12}\rho_{23} \quad (17)$$

Example: If we have $\rho_{12} = \frac{1}{3}$ and $\rho_{23} = \frac{1}{3}$, we get the following bound

$$-\frac{7}{9} \leq \rho_{13} \leq 1,$$

So obviously (iii) is false as the correlation can be negative.

Conditions for transitivity: We assume transitivity when we have the identity $\rho_{13} = \rho_{12}\rho_{23}$.

Consider the situation where $\rho_{12}^2 + \rho_{23}^2 = 1$. From 17:

$$\rho_{12}\rho_{23} - \sqrt{\rho_{23}^2 - \rho_{23}^4} \leq \rho_{13} \leq \rho_{12}\rho_{23} + \sqrt{\rho_{23}^2 - \rho_{23}^4} \quad (18)$$

The inequality tightens on both sides as $\rho_{12}^2 + \rho_{23}^2$ becomes greater than 1.

Background: It is remarkable that people take transitivity for granted; even maestro Terry Tao wasn't aware of it.

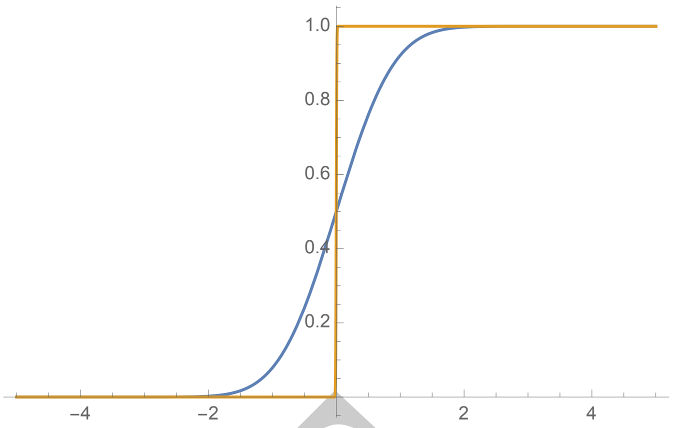

Fig. 13. Sigmoid. The Heaviside is a special case of the Gompertz curve with $c \to 0$.

V. Nonlinearities and other defects in "IQ" studies and psychometrics in general

We discuss the effect of nonlinearity in general but IQ studies and psychometric is a treasure trove of defective use of statistical metrics, especially correlation.

See *IQ is largely a pseudoscientific swindle*:
https://medium.com/incerto/iq-is-largely-a-pseudoscientific-swindle-f131c101ba39

Rule 4: Nonlinearity

One cannot use total correlation entailing X_1 and X_2 when the association between X_1 and X_2 depends in expectation on X_1 or X_2.

We note that the rule does not cover stochastic correlation or heteroskedasticity where there is no "drift".

Rule 5: Dimension reduction

One cannot use orthogonal factors or apply a principal component reduction for r.v.s $X_1, \ldots, X_n$ if for all $i \neq j$ the association between X_i and X_j depends in expectation on the level of either X_i or X_j. (The flaw infects the "g" in psychometry.)

A. Using a detector of disease as a detector of health

A metric to detect disease will masquerade as a detector of health if one uses (Pearson) correlation! Because of the nonlinearity of disease. Let us consider disease anything $+K$ STDs away.

Looking for the induced correlation of performance as a binary variable $\{0, 1\}$ for IQ $> KSTDs$, assuming everything is Gaussian.

Let $\mathbb{I}_{x>K}$ be the Heaviside Theta Function. We are looking at ρ the correlation between X and $\mathbb{I}_{x>K}$.

$$\rho = \frac{\mathbb{E}\left((x - \mathbb{E}(x))\left(\mathbb{I}_{x>K} - \mathbb{E}\left(\mathbb{I}_{x>K}\right)\right)\right)}{\sqrt{\mathbb{E}\left((x - \mathbb{E}(x))^2\right)\mathbb{E}\left(\left(\mathbb{I}_{x>K} - \mathbb{E}\left(\mathbb{I}_{x>K}\right)\right)^2\right)}} \quad (19)$$

For a Gaussian:

$$\rho = \frac{\sqrt{\frac{2}{\pi}} e^{-\frac{(K-\mu)^2}{2\sigma^2}}}{\sqrt{1 - \text{erf}\left(\frac{K-\mu}{\sqrt{2}\sigma}\right)^2}} \tag{20}$$

$\rho /. \{K \to 70, \mu \to 100, \sigma \to 15\} \approx .36$ and for $K = \mu$, $\sqrt{\frac{2}{\pi}} \approx .798$

1) Sigmoidal functions: $z(x) = e^{-b e^{-c(K+x)}}$

B. ReLu type functions (ramp payoffs)

Let us look at how correlation misrepresents of association in Fig. 11. Let $f(x) = (-R_1 + x)\, \mathbb{1}_{x<R_1}$, $R_1 \leq 0,\ x \in \mathbb{R}$ We can prove that: if for any piecewise linear function $f(.)$ such that $\rho_{-R,R_1} = 1$, $\rho_{R,R_1} = 0$, where $\rho_{.,.}$ denote piecewise correlation in $[-R, R_1)$ and $(R_1, R]$, the unconditional correlation is $\rho = \frac{2R-R_1}{R\sqrt{-3+\frac{8R}{R+R_1}}}$ where

$$\rho_{-R,R} = \frac{\int_{-R}^{R} (x-\mu_x)\left(f(x)-\mu_{f(x)}\right) dx}{\sqrt{\int_{-R}^{R} (x-\mu_x)^2\, dx \left(\int_{-R}^{R} \left(f(x)-\mu_{f(x)}\right)^2\right) dx}}$$

where $\mu_x = \frac{1}{2R}\int_{-R}^{R} x\, dx$ and $\mu_{f(x)} = \frac{1}{2R}\int_{-R}^{R} f(x)\, dx$

In the special symmetric case where $R_1 = 0$, we get $\rho = \frac{2}{\sqrt{5}} \approx .894$.

C. Dead man bias

> *QUIZ:* You administer IQ tests to 10K people, then give them a "performance test" for anything, any task. 2000 of them are dead. Dead people score 0 on IQ and 0 on performance. The rest have the IQ uncorrelated to the performance. What is the spurious correlation IQ/performance?

Answer: roughly 37%

The systematic bias comes from the fact that if you hit someone on the head with a hammer, he or she will be bad at everything. (And any test of incompetence can work there). There is no equivalent to someone suddenly becoming good at everything.

Hence all tests of competence will show some positive correlation to IQ even if they are random! And if you see a low correlation, means that the real correlation is... negative.

Assume $X, Y \sim$ Uniform Distribution[0,1] as most representative, p alive, $(1-p)$ dead (or in the clinical tails)

$$\begin{aligned}\rho = &\frac{1}{(1-p)\int_0^1 (0-\mu_x)^2\, dx + p\int_0^1 (x-\mu_x)^2\, dx}\Bigg((1\\ &-p)\int_0^1\int_0^1 (0-\mu_x)(0-\mu_y)\, dx\, dy \\ &+p\int_0^1\int_0^1 (x-\mu_x)(y-\mu_y)\, dx\, dy\Bigg)\\ =& \frac{1}{3p-4}+1\end{aligned} \tag{21}$$

D. State dependent correlation (Proof that psychometrics fail in their use of the "g")

We simplify the proof in 2D. Traditionally one writes the covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_{11}^2 & \rho\sigma_{11}\sigma_{22} \\ \rho\sigma_{11}\sigma_{22} & \sigma_{22}^2 \end{pmatrix}$$

Here we can't anymore since ρ is X dependent

$$\Sigma x = \begin{pmatrix} \sigma_{11}^2 & \sigma_{11}\sigma_{22}(fx) \\ \sigma_{11}\sigma_{22}(fx) & \sigma_{22}^2 \end{pmatrix};$$

Now the eigenvalues of the matrix Σ are also X dependent.

$$\begin{aligned}&\lambda(x)\\ &= \Bigg\{\frac{1}{2}\Bigg(-\sqrt{4\sigma_{22}^2\sigma_{11}^2 f(x)^2 + \sigma_{11}^4 - 2\sigma_{22}^2\sigma_{11}^2 + \sigma_{22}^4} + \sigma_{11}^2 \\ &\quad + \sigma_{22}^2\Bigg), \frac{1}{2}\Bigg(\sqrt{4\sigma_{22}^2\sigma_{11}^2 f(x)^2 + \sigma_{11}^4 - 2\sigma_{22}^2\sigma_{11}^2 + \sigma_{22}^4} \\ &\quad + \sigma_{11}^2 + \sigma_{22}^2\Bigg)\Bigg\}.\end{aligned} \tag{22}$$

We have the second derivative no longer flat–as we see in the graph the function is not just non constant but nonlinear;nonlinearities show in second derivative,

$$\begin{aligned}\lambda'(x) = \Bigg\{&\frac{1}{2}\Bigg(-\frac{4\sigma_{11}^2\sigma_{22}^2 f(x) f''(x)}{\sqrt{4\sigma_{22}^2\sigma_{11}^2 f(x)^2 + \sigma_{11}^4 - 2\sigma_{22}^2\sigma_{11}^2 + \sigma_{22}^4}} \\ &+ \frac{16\sigma_{11}^4\sigma_{22}^4 f(x)^2 f'(x)^2}{\left(4\sigma_{22}^2\sigma_{11}^2 f(x)^2 + \sigma_{11}^4 - 2\sigma_{22}^2\sigma_{11}^2 + \sigma_{22}^4\right)^{3/2}} \\ &- \frac{4\sigma_{11}^2\sigma_{22}^2 f'(x)^2}{\sqrt{4\sigma_{22}^2\sigma_{11}^2 f(x)^2 + \sigma_{11}^4 - 2\sigma_{22}^2\sigma_{11}^2 + \sigma_{22}^4}}\Bigg), \\ &\frac{1}{2}\Bigg(\frac{4\sigma_{11}^2\sigma_{22}^2 f(x) f''(x)}{\sqrt{4\sigma_{22}^2\sigma_{11}^2 f(x)^2 + \sigma_{11}^4 - 2\sigma_{22}^2\sigma_{11}^2 + \sigma_{22}^4}} \\ &- \frac{16\sigma_{11}^4\sigma_{22}^4 f(x)^2 f'(x)^2}{\left(4\sigma_{22}^2\sigma_{11}^2 f(x)^2 + \sigma_{11}^4 - 2\sigma_{22}^2\sigma_{11}^2 + \sigma_{22}^4\right)^{3/2}} \\ &+ \frac{4\sigma_{11}^2\sigma_{22}^2 f'(x)^2}{\sqrt{4\sigma_{22}^2\sigma_{11}^2 f(x)^2 + \sigma_{11}^4 - 2\sigma_{22}^2\sigma_{11}^2 + \sigma_{22}^4}}\Bigg)\Bigg\}\end{aligned} \tag{23}$$

implying yuuuge random terms affecting loads and the "factors".

VI. Statistical Testing of Differences Between Variables

A pervasive error: Where X and Y are two random variables, the properties of $X - Y$, say the variance, probabilities, and higher order attributes are markedly different from the difference in properties. So $\mathbb{E}(X-Y) = \mathbb{E}(X) - \mathbb{E}(Y)$ but of course, $Var(X-Y) \neq Var(X) - Var(Y)$, etc. for higher norms. It means that P-values are different, and of course the coefficient of variation ("Sharpe"). Where σ is the standard deviation of the variable (or sample):

$$\frac{\mathbb{E}(X-Y)}{\sigma(X-Y)} \neq \frac{\mathbb{E}(X)}{\sigma(X)} - \frac{\mathbb{E}(Y))}{\sigma(Y)}$$

$$\sigma(X-Y)=\sqrt{-2\rho_{12}\sigma_2\sigma_1+\sigma_1^2+\sigma_2^2}$$

In *Fooled by Randomness* (2001):

> A far more acute problem relates to the outperformance, or the comparison, between two or more persons or entities. While we are certainly fooled by randomness when it comes to a single times series, the foolishness is compounded when it comes to the comparison between, say, two people, or a person and a benchmark. Why? Because both are random. Let us do the following simple thought experiment. Take two individuals, say, a person and his brother-in-law, launched through life. Assume equal odds for each of good and bad luck. Outcomes: lucky-lucky (no difference between them), unlucky-unlucky (again, no difference), lucky- unlucky (a large difference between them), unlucky-lucky (again, a large difference).

Ten years later (2011) it was found that 50% of neuroscience papers (peer-reviewed in "prestigious journals") that compared variables got it wrong.

> In theory, a comparison of two experimental effects requires a statistical test on their difference. In practice, this comparison is often based on an incorrect procedure involving two separate tests in which researchers conclude that effects differ when one effect is significant (P < 0.05) but the other is not (P > 0.05). We reviewed 513 behavioral, systems and cognitive neuroscience articles in five top-ranking journals (Science, Nature, Nature Neuroscience, Neuron and The Journal of Neuroscience) and found that 78 used the correct procedure and 79 used the incorrect procedure. An additional analysis suggests that incorrect analyses of interactions are even more common in cellular and molecular neuroscience.

In Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E. J. (2011). Erroneous analyses of interactions in neuroscience: a problem of significance. Nature neuroscience, 14(9), 1105-1107.

Fooled by Randomness was read by many professionals (to put it mildly); the mistake is still being made. Ten years from now, they will still be making the mistake.

VII. Fat Tailed Residuals in Linear Regression Models

We mentioned in Chapter **??** that linear regression fails to inform under fat tails. Yet it is practiced. For instance, it is patent that income and wealth variables are power law distributed (with a spate of problems, see our Gini discussions in [3]). However IQ scores are Gaussian (seemingly by design). Yet people regress one on the other failing to see that it is improper.

Consider the following linear regression in which the independent and independent are of different classes:

$$Y=aX+b+\epsilon,$$

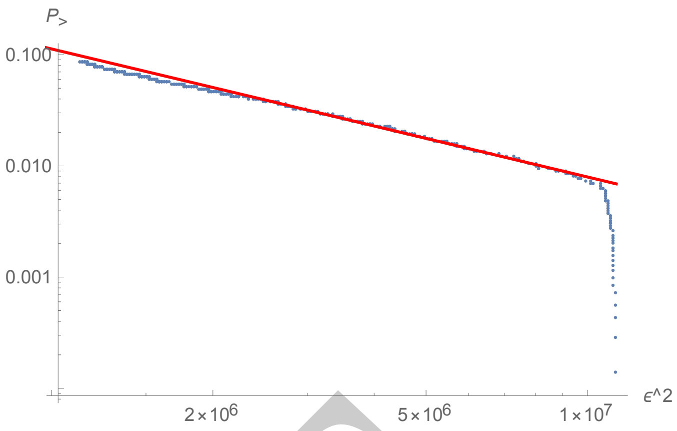

Fig. 14. The logloglot of the squared residuals ϵ^2 for the IQ-income linear regression using standard Winsconsin Longitudinal Studies (WLS) data. We notice that the income variables are winsorized. Clipping the tails creates the illusion of a high R^2. Actually, even without clipping the tail, the coefficient of determination will show much higher values owing to the small sample properties for the variance of a power law.

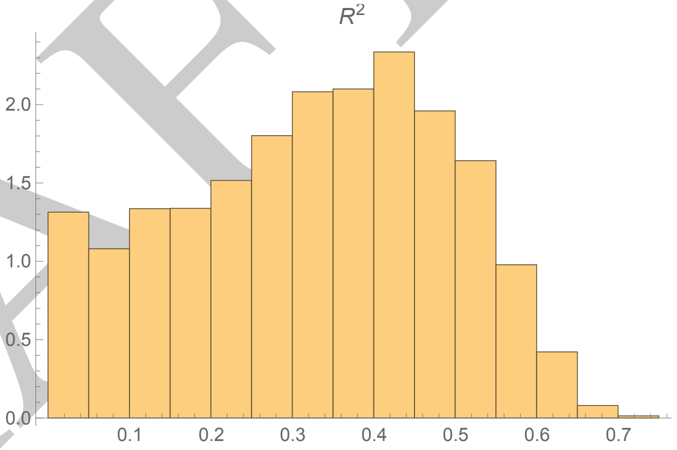

Fig. 15. An infinite variance case that shows a high R^2 in sample; but it ultimately has a value of 0. Remember that R^2 is stochastic. The problem greatly resembles that of P values in Chapter **??** owing to the complication of a metadistribution in $[0,1]$

where X is standard Gaussian ($\mathcal{N}(0,1)$) and ϵ is power law distributed, with $\mathbb{E}(\epsilon)=0$ and $\mathbb{E}(\epsilon^2)<+\infty$. There are no restrictions on the parameters.

Clearly we can compute the coefficient of determination R^2 as 1 minus the ratio of the expectation of the sum of residuals over the total squared variations, so we get the more general answer to our idiosyncratic model. Since $X\sim\mathcal{N}(0,1)$, $Y\sim\mathcal{N}(b,|a|)$, we have

$$\mathbb{E}(R^2)=1-\frac{\mathbb{E}\left(\epsilon^2\right)}{a^2+\mathbb{E}\left(\epsilon^2\right)}. \tag{24}$$

Proof. Since the expectation of the total variation of Y is $\mathbb{E}\left((aX+b-b+\epsilon)^2\right)$. □

And of course, for infinite variance:

$$\underset{E(\epsilon^2)\to+\infty}{lim}\mathbb{E}(R^2)=0.$$

We can also compute it by taking, simply, the square of the correlation between X and Y. For instance, assume the distribution for ϵ is the Student T distribution with zero mean, scale σ and tail exponent $\alpha < 2$ (as we saw earlier, we get identical results with other ones so long as we constrain the mean to be 0). Let's start by computing the correlation: the numerator is the covariance $Cov(X, Y) = \mathbb{E}\left((aX + b + \epsilon)X\right) = a$. The denominator (standard deviation for Y) becomes $\sqrt{\mathbb{E}\left(((aX + \epsilon) - a)^2\right)} = \sqrt{\frac{2\alpha a^2 - 4a^2 + \alpha\sigma^2}{\alpha - 2}}$. So

$$\mathbb{E}(R^2) = \frac{a^2(\alpha - 2)}{2(\alpha - 2)a^2 + \alpha\sigma^2} \tag{25}$$

And the Dirac the limit from above:

$$\underset{\alpha \to 2^+}{lim} \mathbb{E}(R^2) = 0.$$

We are careful here to use $\mathbb{E}(R^2)$ rather than the seemingly deterministic R^2 because it is a stochastic variable that will be extremely sample dependent. Indeed, given that *in sample* the expectation will always be finite. Even if the ϵ are Cauchy! The point is illustrated in Fig. 15. The point invalidates much studies of the relations IQ-wealth and IQ-income of the kind [4]; we can see the striking effect in Fig. 14. Given that R is bounded in $[0, 1]$, it will reach its true value very slowly, see the P-Value problem in Chapter **??**.

Property 1. *When a fat tailed random variable is regressed over a thin tailed one, the coefficient of determination R^2 will be biased higher, and requires a much larger sample size to converge (if it ever does).*

We will examine in [5] the slow convergence of power laws distributed variables under the law of large numbers (LLN): it can be as much as 10^{13} times slower than the Gaussian.

Appendix

1) Mean Deviation vs Standard Deviation: The metric standard deviation itself is a computational metric and does not map to distances as interpreted. Mean absolute deviation does.

Why the [REDACTED] did statistical science pick STD over Mean Deviation? Here is the story, with analytical derivations not seemingly available in the literature. In Huber [6]:

> There had been a dispute between Eddington and Fisher, around 1920, about the relative merits of *dn* (mean deviation) and *Sn* (standard deviation). Fisher then pointed out that for **exactly normal** observations, *Sn* is 12% more efficient than *dn*, and this seemed to settle the matter. (My emphasis)

Let us rederive and see what Fisher meant.

Let n be the number of summands: the Asymptotic Relative Efficiency (ARE) is

$$ARE = \lim_{n \to \infty} \left(\frac{\mathbb{V}(Std)}{\mathbb{E}(Std)^2} \bigg/ \frac{\mathbb{V}(Mad)}{\mathbb{E}(Mad)^2} \right)$$

Assume we are certain that X_i, the components of sample follow a Gaussian distribution, normalized to mean=0 and a standard deviation of 1.

2) Relative Standard Deviation Error: The characteristic function $\Psi_1(t)$ of the distribution of x^2: $\Psi_1(t) = \int_{-\infty}^{\infty} \frac{e^{-\frac{x^2}{2} + itx^2}}{\sqrt{2\pi}} \, dx = \frac{1}{\sqrt{1-2it}}$. With the squared deviation $z = x^2$, f, the pdf for n summands becomes:

$$\begin{aligned} f_Z(z) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-itz) \left(\frac{1}{\sqrt{1-2it}} \right)^n \, dt \\ &= \frac{2^{-\frac{n}{2}} e^{-\frac{z}{2}} z^{\frac{n}{2}-1}}{\Gamma\left(\frac{n}{2}\right)}, z \\ &> 0. \end{aligned} \tag{26}$$

Now take $y = \sqrt{z}$, $f_Y(y) = \frac{2^{1-\frac{n}{2}} e^{-\frac{z^2}{2}} z^{n-1}}{\Gamma\left(\frac{n}{2}\right)}$, $z > 0$, which corresponds to the Chi Distribution with n degrees of freedom. Integrating to get the variance: $\mathbb{V}_{std}(n) = n - \frac{2\Gamma\left(\frac{n+1}{2}\right)^2}{\Gamma\left(\frac{n}{2}\right)^2}$. And, with the mean equalling $\frac{\sqrt{2}\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}$, we get $\frac{\mathbb{V}(Std)}{\mathbb{E}(Std)^2} = \frac{n\Gamma\left(\frac{n}{2}\right)^2}{2\Gamma\left(\frac{n+1}{2}\right)^2} - 1$.

3) Relative Mean Deviation Error: Characteristic function again for $|x|$ is that of a folded Normal distribution, but let us redo it:

$\Psi_2(t) = \int_0^{\infty} \sqrt{\frac{2}{\pi}} e^{-\frac{x^2}{2} + itx} = e^{-\frac{t^2}{2}} \left(1 + i \operatorname{erfi}\left(\frac{t}{\sqrt{2}}\right)\right)$, where erfi is the imaginary error function $erf(iz)/i$.

The first moment:

$M_1 = -i \frac{\partial}{\partial t^1} \left(e^{-\frac{t^2}{2n^2}} \left(1 + i \operatorname{erfi}\left(\frac{t}{\sqrt{2}n}\right)\right) \right)^n \bigg|_{t=0} = \sqrt{\frac{2}{\pi}}$.

The second moment,

$M_2 = (-i)^2 \frac{\partial^2}{\partial t^2} \left(e^{-\frac{t^2}{2n^2}} \left(1 + i \operatorname{erfi}\left(\frac{t}{\sqrt{2}n}\right)\right) \right)^n \bigg|_{t=0} = \frac{2n + \pi - 2}{\pi n}$. Hence, $\frac{\mathbb{V}(Mad)}{\mathbb{E}(Mad)^2} = \frac{M_2 - M_1^2}{M_1^2} = \frac{\pi - 2}{2n}$.

4) Finalmente, the Asymptotic Relative Efficiency For a Gaussian:

$$\text{ARE} = \lim_{n \to \infty} \frac{n \left(\frac{n\Gamma\left(\frac{n}{2}\right)^2}{\Gamma\left(\frac{n+1}{2}\right)^2} - 2 \right)}{\pi - 2} = \frac{1}{\pi - 2} \approx .875$$

which means that the standard deviation is 12.5% more "efficient" than the mean deviation *conditional on the data being Gaussian* and these blokes bought the argument. Except that the slightest contamination blows up the ratio. Norm ℓ^2 is not appropriate for about anything.

A. Effect of Fatter Tails on the "efficiency" of STD vs MD

Consider a standard mixing model for volatility with an occasional jump with a probability p. We switch between Gaussians (keeping the mean constant and central at 0) with:

$$\mathbb{V}(x) = \begin{cases} \sigma^2(1 + a) & \text{with probability } p \\ \sigma^2 & \text{with probability } (1 - p) \end{cases}$$

For ease, a simple Monte Carlo simulation would do. Using $p = .01$ and $n = 1000$... Figure 16 shows how a=2 causes degradation. A minute presence of outliers makes MAD more "efficient" than STD. Small "outliers" of 5 standard deviations cause MAD to be five times more efficient.[1]

[1] The natural way is to center MAD around the median; we find it more informative for many of our purposes here (and, more generally, in decision theory) to center it around the mean.

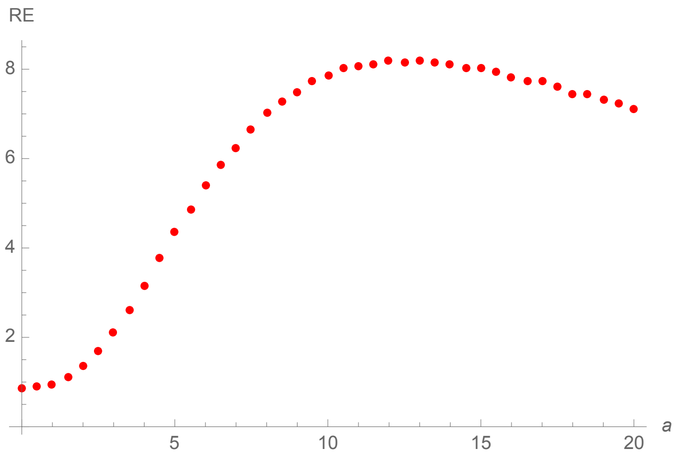

Fig. 16. A simulation of the Relative Efficiency ratio of Standard deviation over Mean deviation when injecting a jump size $\sqrt{(1+a)} \times \sigma$, as a multiple of σ the standard deviation.

References

[1] E. Soyer and R. M. Hogarth, "The illusion of predictability: How regression statistics mislead experts," *International Journal of Forecasting*, vol. 28, no. 3, pp. 695–711, 2012.

[2] D. Goldstein and N. Taleb, "We don't quite know what we are talking about when we talk about volatility," *Journal of Portfolio Management*, vol. 33, no. 4, 2007.

[3] A. Fontanari, N. N. Taleb, and P. Cirillo, "Gini estimation under infinite variance," *Physica A: Statistical Mechanics and its Applications*, vol. 502, no. 256-269, 2018.

[4] J. L. Zagorsky, "Do you have to be smart to be rich? the impact of iq on wealth, income and financial distress," *Intelligence*, vol. 35, no. 5, pp. 489–501, 2007.

[5] N. N. Taleb, *Technical Incerto, Vol 1: The Statistical Consequences of Fat Tails, Papers and Commentaries*. Monograph, 2019.

[6] P. J. Huber, *Robust Statistics*. Wiley, New York, 1981.