Dude, Where's my Alpha?

dispatches from a 'quant'

Steven E. Pav
steven@cerebellumcapital.com

Cerebellum Capital

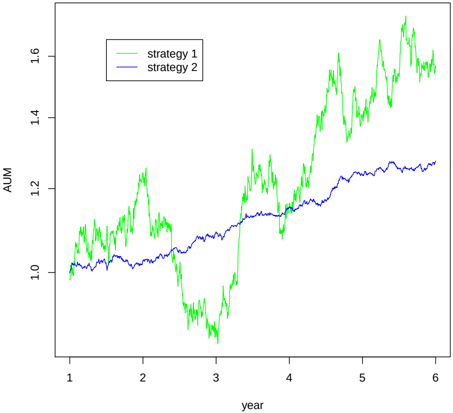
USF 2013-02-06

Stipulate the following:

- Generating ideas for trading strategies is easy.
  Generating *good* ideas is harder.
- Backtesting strategies is an engineering problem.
  Not easy, not insurmountable: lots of possible biases, missing data, some domain knowledge needed.
- If evaluating strategies were easy, a lot of smart people would be rich.

Evaluating strategies must be hard.

A Simple Example: Choose One



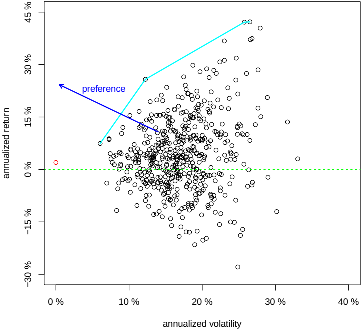
Which strategy would you rather invest in? "Slow and Steady" or "Loose Cannon"?

- Let p_t be the 'mark-to-market' (MtM) at time t.
- Compute the *returns*:

$$\text{Geometric: } l_t =_{\mathsf{df}} \log \frac{p_t}{p_{t-1}},$$
$$\text{Arithmetic: } r_t =_{\mathsf{df}} \frac{p_t}{p_{t-1}} - 1 = e^{l_t} - 1.$$

  By simple calculus: $l_t \leq r_t$ with equality at zero.
- Sequential Geometric returns are additive by telescoping.
- Contemporaneous Arithmetic returns are additive: Arithmetic return of a portfolio is the dollar-weighted average of the components' arithmetic returns. (This includes 'shorting'.)

- Let $\mu = \mathsf{E}\left[l\right], \sigma^2 = \operatorname{var}\left(l\right)$.
  Throughout, mostly assume l_t are *i.i.d.*, (unrealistic).
- For fixed μ, prefer a smaller σ; for fixed σ prefer larger μ.
  The *efficient frontier* is the set of optimal strategies under this preference:



- μ, σ are unknown; take sample estimates, $\hat{\mu}$, $\hat{\sigma}$.
- The Sharpe Ratio (SR) is the sample statistic[17]

$$\hat{\zeta} =_{\text{df}} \frac{\hat{\mu}}{\hat{\sigma}},$$

  May also include a 'risk-free' rate:

$$\hat{\zeta} = \frac{\hat{\mu} - r_0}{\hat{\sigma}}.$$

- Population analogue: Signal-to-Noise Ratio (SNR) $\zeta =_{\text{df}} \mu/\sigma$.
- Connection between SR and t-statistic: $\hat{\zeta} = t/\sqrt{n}$.
- SR is Student's original test statistic. [2] The 'Student Ratio'?

Sharpe Ratio Minutiae I

- No real standard on arithmetic vs. geometric returns. (Although using arithmetic returns looks better ...)
- The units of Sharpe are 'per square root time': $\hat{\mu}$ is '(percent) per time', $\hat{\sigma}$ is '(percent) per root time'.
- Sharpe is often quoted in annualized units, *i.e.*, $\mathsf{yr}^{-1/2}$. Avoid ambiguity and *always* include units.
- *n.b.*, $1\mathsf{yr}^{-1/2} = \frac{1}{2}\mathsf{Q}^{-1/2} = \frac{1}{\sqrt{12}}\mathsf{mo.}^{-1/2} = \frac{1}{\sqrt{253}}\mathsf{day}^{-1/2}$
- For equities quant strategies, achieved SR translates as: $1\mathsf{yr}^{-1/2} \Rightarrow$ "good", $2\mathsf{yr}^{-1/2} \Rightarrow$ "great", $3\mathsf{yr}^{-1/2} \Rightarrow$ "legend". In HFT, higher SR the norm, but large fixed costs (like r_0).

- Cantelli's Inequality gives:

$$\Pr\left\{\text{negative return on period}\right\} \leq \frac{1}{1+\zeta^2}$$

(Often more natural to consider $\zeta^2, \hat{\zeta}^2$.)

- Another view: a year-on-year loss is a "ζ sigma event". Central Limit Theorem: If returns are well behaved, enough gambles are made, annual return will be nearly normal.

$$\zeta = 1\text{yr}^{-1/2} \Rightarrow \Pr\left\{\text{down year}\right\} \approx 16\%$$

$$\zeta = 2\text{yr}^{-1/2} \Rightarrow \Pr\left\{\text{down year}\right\} \approx 2.3\%$$

$$\zeta = 3\text{yr}^{-1/2} \Rightarrow \Pr\left\{\text{down year}\right\} \approx 0.13\%$$

$$\zeta = 4\text{yr}^{-1/2} \Rightarrow \Pr\left\{\text{down year}\right\} \approx 0.0032\%$$

Diversification and Sharpe Ratio

- *Squared* SNR of independent strategies is subadditive: For k $\perp\!\!\!\perp$ strategies, with SNRs $\zeta_1, \ldots, \zeta_k$, optimal rebalancing combination has
$$\zeta_*^2 = \sum_i \zeta_i^2$$
Diminishing returns in ζ: $2\text{yr}^{-1/2} + 1\text{yr}^{-1/2} \Rightarrow 2.2\text{yr}^{-1/2}$.
- For dependent strategies, correlation is relevant.
For two strategies with correlation ρ:
$$\zeta_*^2 = \frac{\zeta_1^2 + \zeta_2^2 - 2\rho\zeta_1\zeta_2}{1 - \rho^2}$$
- In general, want anti-correlated strategies ('hedges').
Avoid strategies correlated to the 'retail alpha'.

Diversification vs Correlation

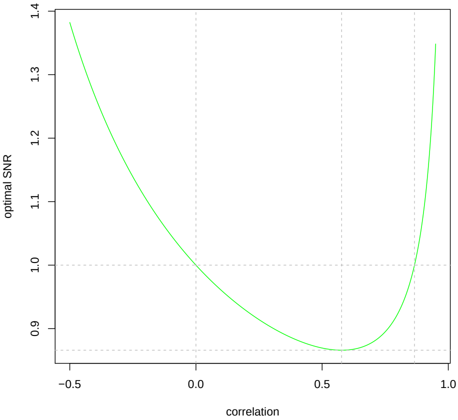


Figure: ζ_* is plotted versus ρ for $\zeta_1 = 0.5\text{yr}^{-1/2}, \zeta_2 = 0.87\text{yr}^{-1/2}$. It is possible to get *no* diversification benefit beyond best strategy.

- SR is actually a *biased* estimator of SNR:[14]

$$\mathsf{E}\left[\hat{\zeta}\right] = \sqrt{\frac{n-1}{2}}\frac{\Gamma\left((n-2)/2\right)}{\Gamma\left((n-1)/2\right)}\zeta = \frac{1}{c_4}\zeta.$$

  Though the bias is small ($\leq 0.77\%$ when $n \geq 100$).

- Standard error assuming normal returns: [21, 10, 6]

$$s.e. = \sqrt{\frac{1+\frac{\zeta^2}{2}}{n-1}} \approx \sqrt{\frac{1+\frac{\hat{\zeta}^2}{2}}{n-1}} \approx \sqrt{\frac{1}{n}}.$$

  The latter annualizes as you would expect.
  Rule of thumb: $\left|\zeta - \hat{\zeta}\right| \leq \frac{2}{\sqrt{n}}$, with probability ≈ 0.95.

- Fixes for heteroskedasticity, autocorrelation, skew. [20, 10, 15] (Typically these bias $\hat{\zeta}$ by $< 10\%$)
- No easy fix for omitted variable bias *i.e.*, "something changed in the world."
- Perhaps better to generalize to regression:

$$l_t = \alpha + \sum_i \beta_i f_{i,t} + \epsilon_t,$$

  with f_i returns from known benchmark strategies.
- Typically want no exposure to f_i because it is available via ETPs, or one believes they are zero mean or cannot be timed.
- Whole literature on regression to perform inference on α. Inference on α/σ is less common.

Relation between sample size, effect size (SNR), and rates of false positives and false negatives.
Figure out if you have enough data or enough 'alpha'.

- Hypothetical Vendor: "I have three years of historical data."
- Hypothetical Strategist: "This strategy has SNR $0.6\text{yr}^{-1/2}$."
- Hypothetical Investor: "You have one year to prove yourself."

Good approximations of form

$$n \approx \frac{\kappa}{\zeta^2},$$

with κ a function of type I and type II rates, *etc.* [20, 6]

	one sided	two sided
power $= 0.50$	2.72	3.86
power $= 0.80$	6.20	7.87

Table: Value of κ to achieve given power in t-test, $\alpha = 0.05$.

rule of thumb for Sharpe power

To test SNR > 0, with 0.05 type I rate, and 50 % power,

$$n \approx \frac{2.72}{\zeta^2}. \qquad \text{mnemonic form: } e \approx n\zeta^2$$

This rule is sobering:

- Hypothetical Vendor: "I have three years of historical data."
  Answer: Strategy must have SNR $\geq 0.95\text{yr}^{-1/2}$.
- Hypothetical Strategist: "This strategy has SNR $0.6\text{yr}^{-1/2}$."
  Answer: Need 7.6 years of data to backtest. (!)
- Hypothetical Investor: "You have one year to prove yourself."
  Answer: Strategy must have SNR $\geq 1.6\text{yr}^{-1/2}$.

Why do Bad Things Happen to Smart People? I

Figure: "The majority of deployed quant strategies are type I errors."
(Image courtesy of Automated Trader magazine.)

Why do Bad Things Happen to Smart People? II

Many explanations are offered:

- Bad backtests: inability to model costs and impact, observer effects, "backtest arbitrage," outright "time-travelling."
- Biased statistical tests. Rejecting the null for uninteresting reasons.
- Claims of "non-stationarity" ("the dog ate my alpha.")

Perhaps most important:

- False positives / Excessive type I rate.
- Sequential model overfit.

Consider the outcomes:

test \ truth	profitable	not profitable
reject null	win!	type I error
fail to reject	type II error	business as usual

Let α, β be the type I and II rates; let c_0 be incidence rate of profitable strategies. Condition on test rejection:

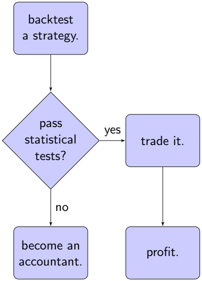
test \ truth	profitable (c_0)	not profitable ($1 - c_0$)
reject null	$c_0(1-\beta)$	$(1-c_0)\alpha$
fail to reject	$c_0\beta$	$(1-c_0)(1-\alpha)$

False Discovery Rate (FDR) is ratio of false to all positives:

$$\mathrm{FDR} =_{\mathrm{df}} \frac{(1-c_0)\alpha}{(1-c_0)\alpha + c_0(1-\beta)} \geq 1 - \frac{c_0}{\alpha} \text{ when } \beta \geq c_0.$$

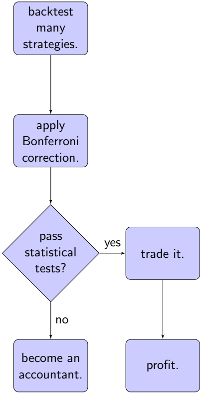
Probably $c_0 \ll 0.05$. But is $c_0 \approx 10^{-5}$? $\approx 10^{-10}$?

How it is supposed to happen:

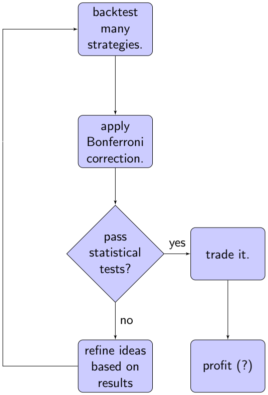


Why does Overfit Happen? II

Or even this:



How it typically happens:



This is totally broken.

Why does Overfit Happen? IV

Why do quants overfit? A toxic mix:

- Failure is not an option: cannot become an accountant.
- The incidence rate is *very* low.
- By power rules, you need *years* of data to confirm alpha.
- There is only one history to work with:
  One collects data at a rate of 1 day per day.
- Often ‘refining’ ideas while debugging code.
- Sloppy process: often not aware one is overfitting.
- Data snooping tests require good record keeping, cannot deal with sequential overfit. [22, 3, 5]

- Suppose you can invest in p different assets. (Stocks, ETPs, strategies, *etc.*) How do you choose a portfolio?
- This generalizes the discrete question, "is this asset good or not?" to the continuous "how much shall I invest in it?"
- Consider the arithmetic returns p-vector, $\boldsymbol{x}_t$. Let $\boldsymbol{\mu}$, Σ be the (vector) mean and (matrix) covariance.
- If $\hat{\boldsymbol{w}}$ is proportion of capital in each asset, return is $\hat{\boldsymbol{w}}^\top \boldsymbol{x}_t$. This has mean $\hat{\boldsymbol{w}}^\top \boldsymbol{\mu}$ and variance $\hat{\boldsymbol{w}}^\top \Sigma \hat{\boldsymbol{w}}$. The SNR is

$$\frac{\hat{\boldsymbol{w}}^\top \boldsymbol{\mu}}{\sqrt{\hat{\boldsymbol{w}}^\top \Sigma \hat{\boldsymbol{w}}}}$$

In brief, pick a portfolio to maximize SNR.

- The population parameters are unknown. Given n observations of returns, compute 'usual' estimates $\hat{\boldsymbol{\mu}}$, $\hat{\Sigma}$.
- Markowitz Portfolio maximizes the SR: [11, 12, 13]

$$\hat{\boldsymbol{w}}_* =_{\mathsf{df}} \operatorname*{argmax}_{\hat{\boldsymbol{w}}} \frac{\hat{\boldsymbol{w}}^\top \hat{\boldsymbol{\mu}}}{\sqrt{\hat{\boldsymbol{w}}^\top \hat{\Sigma} \hat{\boldsymbol{w}}}}.$$

May also incorporate r_0, but must bound risk:

$$\hat{\boldsymbol{w}}_* =_{\mathsf{df}} \operatorname*{argmax}_{\hat{\boldsymbol{w}} : \hat{\boldsymbol{w}}^\top \hat{\Sigma} \hat{\boldsymbol{w}} \le s^2} \frac{\hat{\boldsymbol{w}}^\top \hat{\boldsymbol{\mu}} - r_0}{\sqrt{\hat{\boldsymbol{w}}^\top \hat{\Sigma} \hat{\boldsymbol{w}}}}.$$

- The optimal portfolio is $\hat{\boldsymbol{w}}_* \propto \hat{\Sigma}^{-1} \hat{\boldsymbol{\mu}}$.
- The SR of this portfolio is

$$\hat{\zeta}_* = \sqrt{\hat{\boldsymbol{\mu}}^\top \hat{\Sigma}^{-1} \hat{\boldsymbol{\mu}}} - \frac{r_0}{s}.$$

Drop the r_0 for now, and consider the RV $\hat{\zeta}_*^2 = \hat{\boldsymbol{\mu}}^\top \hat{\Sigma}^{-1} \hat{\boldsymbol{\mu}}$.

- Hotelling T^2 is multivariate generalization of t: [4, 16, 19]

$$T^2 =_{\mathsf{df}} n\hat{\boldsymbol{\mu}}^\top \hat{\Sigma}^{-1}\hat{\boldsymbol{\mu}} = n\hat{\zeta}_*^2$$

(Recall that in the univariate case $t = \sqrt{n}\hat{\zeta}$.)

- For normal returns, rescaled T^2 has F-distribution:

$$\frac{(n-p)}{p(n-1)} n\hat{\zeta}_*^2 \sim F\left(n\zeta_*^2, p, n-p\right),$$

where ζ_* is SNR of *population* optimal portfolio, $\boldsymbol{\nu}_* = \Sigma^{-1}\boldsymbol{\mu}$. So $\hat{\zeta}_*$ can be used to perform inference on ζ_*.

- *n.b.* ζ_* is maximal SNR of *any* portfolio, including $\hat{\boldsymbol{w}}_*$. We expect some loss: $SNR\left(\hat{\boldsymbol{w}}_*\right) \leq SNR\left(\boldsymbol{\nu}_*\right)$. The loss depends on sample size, effect size; not well understood.

- Observe $\hat{\zeta}_*$, perform inference on ζ_*, an upper bound on SNR of $\hat{\boldsymbol{w}}_*$.
- Via F-distribution we have

$$\mathrm{E}\left[\hat{\zeta}_*^2\right] = \frac{\zeta_*^2 + c}{1 - c}, \qquad \text{where } c =_{\mathrm{df}} p/n.$$

  Gives unbiased estimator: $\mathrm{E}\left[(1-c)\hat{\zeta}_*^2 - c\right] = \zeta_*^2$.

  (Oops! unbiased estimator of ζ_*^2 can be negative!)
- CI, MLE on ζ_* via F-CDF & PDF. Shrinkage estimators.[9]
- If $\hat{\zeta}_*^2 \leq \frac{c}{1-c}$ then MLE of ζ_* is 0. [18]

rule of thumb, based on F-MLE

"If $\hat{\zeta}_*^2 < c$, don't bother!"

Example: Hotelling on S&P Sector Indices

Index data `BASI`, `INDU`, `CONG`, `HLTH`, `CONS`, `TELE`, `UTIL`, `FINA`, `TECH` from `fPortfolio::SPISECTOR`, from 2000-01-04 to 2008-10-17. ($n = 2198$ days, $p = 9$ stocks)

- Optimal in-sample Sharpe ratio is $1\text{yr}^{-1/2}$.
- MLE for ζ_* is $0.1\text{yr}^{-1/2}$. Close to the rule of thumb cutoff: $\hat{\zeta}_*^2 = 1.05\text{yr}^{-1}$ and $c = 1.04\text{yr}^{-1}$.
- 95% CI for ζ_* is $\left[0\text{yr}^{-1/2}, 1.3\text{yr}^{-1/2}\right]$

Conclusion: Markowitz portfolio on S&P sectors *not* recommended.

Approximation of Strategy Overfit I

Caricature of quant work:

- Construct strategy with free parameters θ;
- Backtest strategy for $\theta_1, \theta_2, \ldots, \theta_m$.
- Pick θ_i that maximizes SR of backtest, θ_*.
- Profit! (or not)

Q: How to estimate the SNR of θ_*?

Approximation of Strategy Overfit II

Toy Example: Moving Average Crossover:

- θ is vector of 2 window lengths; Long the instrument exactly when one moving average exceeds the other.
- Brute-force backtest for allowable window lengths.

Approximation of Strategy Overfit III

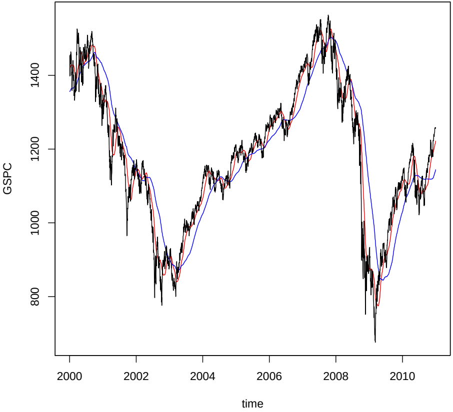


Figure: Example of 2-window MAC on GSPC.

Approximation of Strategy Overfit IV

Q: How to estimate the SNR of θ_*?
A(?): Make PCA-like linear approximation of returns vectors:

$$\left\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\right\} \approx \mathcal{K} \subset \mathcal{L} =_{\mathsf{df}} \left\{\mathsf{Y}\hat{\boldsymbol{w}} \,\middle|\, \hat{\boldsymbol{w}} \in \mathbb{R}^p\right\}$$

Use $\zeta\left(\theta_*\right) \approx \max_{\mathcal{L}} \zeta = \zeta_*$, then make inference on ζ_*.
Observe maximal strategy SR

$$\hat{\zeta}_* = \hat{\zeta}\left(\theta_*\right) =_{\mathsf{df}} \max_{1 \le i \le m} \hat{\zeta}\left(\theta_i\right),$$

use it to approximate maximal SR over $\mathcal{L}$.
You have to estimate p, by Monte Carlo under null, PCA, or SWAG method.

Overfit of Simple 2-Window MAC

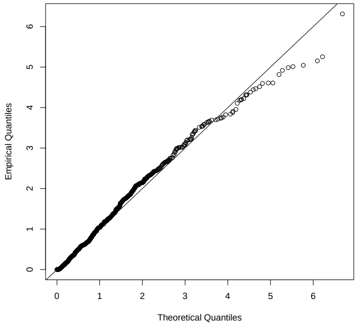


Figure: 1024 Monte Carlo sims of model selection in 2-window MAC over 2500 days, under the null (no population drift or autocorrelation). In-sample $\hat{\zeta}_*$ values transformed to F-statistics with $p = 2.075$.

Use on `GSPC` adjusted returns (2000-01-03 to 2009-12-31, 2515 days), MLE of ζ_* is $0.65\text{yr}^{-1/2}$; 95% CI is $\left(0\text{yr}^{-1/2}, 1.3\text{yr}^{-1/2}\right)$.

- I hope I did not scare you.
- Statistics can useful to avoid quant disasters.
- It is better to be lucky than good.
- We are looking for interns!

Bibliography I

Mark Britten-Jones.
The sampling error in estimates of mean-variance efficient portfolio weights.
The Journal of Finance, 54(2):655–671, 1999.

William Sealy Gosset.
The probable error of a mean.
Biometrika, 6(1):1–25, March 1908.
Originally published under the pseudonym "Student".

P. R. Hansen.
A test for superior predictive ability.
Journal of Business and Economic Statistics, 23(4), 2005.

Harold Hotelling.
The generalization of Student's ratio.
Annals of Mathematical Statistics, 2(3):360–378, 1931.

Po-Hsuan Hsu, Yu-Chin Hsu, and Chung-Ming Kuan.
Testing the predictive ability of technical analysis using a new stepwise test without data snooping bias.
2008.

N. L. Johnson and B. L. Welch.
Applications of the non-central t-distribution.
Biometrika, 31(3-4):362–389, March 1940.

Raymond Kan and Daniel R. Smith.
The distribution of the sample minimum-variance frontier.
Management Science, 54(7):1364–1380, 2008.

Bibliography II

Raymond Kan and GuoFu Zhou.
Tests of mean-variance spanning.
Annals of Economics and Finance, 13(1), 2012.

T. Kubokawa, CP Robert, and AK Saleh.
Estimation of noncentrality parameters.
Canadian Journal of Statistics, 21(1):45–57, 1993.

Andrew W. Lo.
The Statistics of Sharpe Ratios.
Financial Analysts Journal, 58(4), July/August 2002.

H. Markowitz.
Portfolio selection*.
The journal of finance, 7(1):77–91, 2012.

H.M. Markowitz.
The early history of portfolio theory: 1600-1960.
Financial Analysts Journal, pages 5–16, 1999.

H.M. Markowitz.
Foundations of portfolio theory.
The Journal of Finance, 46(2):469–477, 2012.

Robert E. Miller and Adam K. Gehr.
Sample size bias and Sharpe's performance measure: A note.
Journal of Financial and Quantitative Analysis, 13(05):943–946, 1978.

J. D. Opdyke.
Comparing sharpe ratios: So where are the p-values?
Journal of Asset Management, 8(5), 2007.

Pranab Kumar Sen.
Union-intersection principle and constrained statistical inference.
Journal of Statistical Planning and Inference, 137(11):3741–3752, 2007.
Special Issue: In Celebration of the Centennial of The Birth of Samarendra Nath Roy (1906-1964).

William F. Sharpe.
Mutual fund performance.
Journal of Business, 39:119, 1965.

M. C. Spruill.
Computation of the maximum likelihood estimate of a noncentrality parameter.
Journal of Multivariate Analysis, 18(2):216–224, 1986.

N.H. Timm.
Applied multivariate analysis: methods and case studies.
Springer Texts in Statistics. Physica-Verlag, 2002.

Gerald van Belle.
Statistical Rules of Thumb (Wiley Series in Probability and Statistics).
Wiley-Interscience, 1 edition, March 2002.

Christian Walck.
Hand-book on STATISTICAL DISTRIBUTIONS for experimentalists.
December 1996.

Halbert White.
A reality check for data snooping.
Econometrica, 68:1097–1127, 2000.

distribution	param	skew	ex. kurtosis	type I
Gaussian		0	0	0.048
Student's t	df = 10	0	1	0.039
SP500		-1	23	0.059
symmetric SP500		0	22	0.058
Tukey h	h = 0.1	0	5.5	0.057
Tukey h	h = 0.24	0	1.3e+03	0.056
Tukey h	h = 0.4		Inf	0.15
Lambert W x Gaussian	delta = -0.2	-1.2	5.7	0.05
Lambert W x Gaussian	delta = -0.4	-2.7	18	0.08
Lambert W x Gaussian	delta = -1.2	-30	5.2e+03	0.28

Table: Empirical type I rates of the test for $\zeta = 1.0$ via distribution of the Sharpe ratio are given for various distributions of returns. The empirical rates are based on 1024 simulations of three years of daily returns, with a nominal rate of $\alpha = 0.05$. Skew appears to have a much more adverse effect than kurtosis alone.

Idea: lever up until volatility "too high." Set this by a drawdown cap. Sharpe Ratio limits drawdowns.

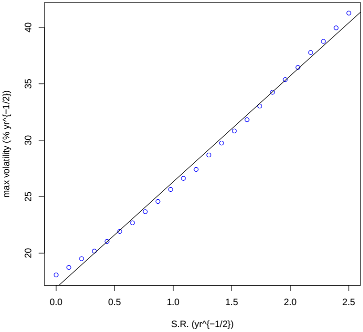


Figure: Maximum volatility (percent per root year) so that probability of a 33 % drawdown over a year is 2.5% or less. Assumes independence, homoskedasticity, mild kurtosis.

- Can generalize Hotelling statistic to get 'Spanning Tests' [1, 8]
- Let T^2_{p+q},T^2_p be Hotelling stats on full set of $p+q$ assets and subset of p assets. Do the q marginal assets 'add any value'.
- Let

$$\Delta T^2 = (n-p-1)\frac{T^2_{p+q} - T^2_p}{n-1+T^2_p} = (n-p-1)\frac{\hat{\zeta}^2_{*,p+q} - \hat{\zeta}^2_{*,p}}{1-(1/n)+\hat{\zeta}^2_{*,p}}.$$

- ΔT^2 also takes a (non-central) Hotelling distribution:

$$\frac{n-(p+q)}{q(n-p-1)}\Delta T^2 \sim F\left(q, n-(p+q), \delta\right),$$

$$\delta = (n-p-1)\frac{\zeta^2_{*,p+q} - \zeta^2_{*,p}}{1-(1/n)+\zeta^2_{*,p}}.$$

- Same inference on δ can be applied (MLE, CI).

Delta Hotelling on S&P Sector Indices

Delta Hotelling: what do `BASI`, `INDU`, `CONG`, `HLTH`, `CONS`, `TELE` add to `UTIL`, `FINA`, `TECH`?

- MLE for $\left(\zeta^2_{*,p+q} - \zeta^2_{*,p}\right)$ is $0\mathrm{yr}^{-1}$.
- 95% CI for $\left(\zeta^2_{*,p+q} - \zeta^2_{*,p}\right)$ is $\left[0\mathrm{yr}^{-1}, 0.26\mathrm{yr}^{-1}\right]$.