A *post hoc* test on the Sharpe ratio

Steven E. Pav *

November 14, 2019

Abstract

We describe a *post hoc* test for the Sharpe ratio, analogous to Tukey's test for pairwise equality of means. The test can be applied after rejection of the hypothesis that all population signal-noise ratios are equal. The test is applicable under a simple correlation structure among asset returns. Simulations indicate the test maintains nominal type I rate under a wide range of conditions and is moderately powerful under reasonable alternatives.

1 Introduction

Sharpe's "reward-to-variability ratio" was originally devised to compare the performance of mutual funds. Sharpe found it to be weakly predictive of out-of-sample performance when measured over a *decade* of returns. [14] Early research on the Sharpe ratio, as it came to be known[1], ignored its statistical nature, treating it like an observable population parameter, though this was soon remedied. [8, 4, 7] More recently, statistical procedures have been proposed to test whether the population Sharpe ratios of several assets (*e.g.*, mutual funds, ETFs, hedge funds, *etc.*) are equal. [6, 17] Here we propose a test to be used to compare pairwise differences after application of such a test.

2 The test

Suppose one has observed n *i.i.d.* samples of some k-vector $\boldsymbol{x}$, representing the returns of k different "assets." We imagine these assets to be different mutual funds, or trading strategies, ETFs, *etc.* From the sample one computes the Sharpe ratio of each asset, resulting in a k-vector, $\hat{\boldsymbol{\zeta}}$.

One natural question to ask is whether the "signal-noise ratio" (the population analogue of the Sharpe ratio) of each asset is equal. One can test the hypothesis of equal signal-noise ratio via the tests of Leung and Wong, or Wright, Yam and Yung. [6, 17] The test of Wright *et al.*, for example, uses asymptotic normality of the $\hat{\boldsymbol{\zeta}}$ to construct a statistic following a χ^2 distribution under the null. In the case where one rejects the null hypothesis of equality, one seeks a

---

[1]Although it was described over a decade earlier by A. D. Roy. [13]

post hoc test, to determine which pairs of the k assets have different signal-noise ratio.

Testing the equality of signal-noise ratios is analogous to the classical procedure for testing equality of means via ANOVA. [1] The *post hoc* procedure that classically followed a rejection of the null in ANOVA is Tukey's range test, sometimes called the "honest significant difference" (HSD) test. [16, 2] In the ANOVA, and Tukey's HSD, the quantity is assumed to have identical variance among all individuals, but potentially different means in different groups. For this reason, the variance is estimated by pooling all observations. It is unnecessary to assume equal volatility of the returns for the k different assets when testing the signal-noise ratio. While this simplifies our *post hoc* test somewhat, typically in the testing of asset returns one observes them contemporaneously, and they are generally correlated.

Tukey's HSD proceeds by computing an upper quantile on the *range* of independent normals divided by a rescaled χ variable. When the means of two individuals differ by more than this amount, one rejects the null that they are equal. Our test will perform a similar computation.

Previously the author showed that when returns are drawn from a multivariate normal distribution with correlation R, then

$$\hat{\zeta} \approx \mathcal{N}\left(\zeta, \frac{1}{n}\left(\mathsf{R} + \frac{1}{2}\operatorname{Diag}\left(\zeta\right)\left(\mathsf{R} \odot \mathsf{R}\right)\operatorname{Diag}\left(\zeta\right)\right)\right), \tag{1}$$

where ζ is the vector of signal-noise ratios and n is the sample size. [11] Note that the approximate covariance matrix here generalizes the well-known standard error of the scalar Sharpe ratio. [5, 4, 7, 10] In the case of the small signal-noise ratios likely to be encountered in practice, that approximation may be further simplified to

$$\hat{\zeta} \approx \mathcal{N}\left(\zeta, \frac{1}{n}\mathsf{R}\right). \tag{2}$$

Then under the null hypothesis that $\zeta = \zeta_0$, one observes

$$\boldsymbol{z} = \sqrt{n}\mathsf{R}^{-1/2}\left(\hat{\zeta} - \zeta_0\right) \approx \mathcal{N}\left(\boldsymbol{0}, \mathsf{I}\right), \tag{3}$$

where $\mathsf{R}^{-1/2}$ is the inverse of the (symmetric) square root of R.

As previously, we assume a simple rank-one form for the correlation matrix,

$$\mathsf{R} = \rho\left(\boldsymbol{1}\boldsymbol{1}^\top\right) + (1 - \rho)\mathsf{I}, \tag{4}$$

where $|\rho| \leq 1$. [11] Under this assumption, it is simple to show that

$$\mathsf{R}^{-1/2} = c\left(\boldsymbol{1}\boldsymbol{1}^\top\right) + (1 - \rho)^{-1/2}\mathsf{I}, \tag{5}$$

for some constant c.

Now we consider the difference in Sharpe ratios of two assets, indexed by i and j. Let $\boldsymbol{v} = \boldsymbol{e}_i - \boldsymbol{e}_j$, where $\boldsymbol{e}_i$ is the ith column of the identity matrix. From Equation 3 we have

$$\begin{aligned}
\boldsymbol{v}^\top\boldsymbol{z} &= \boldsymbol{v}^\top\sqrt{n}\mathsf{R}^{-1/2}\left(\hat{\zeta} - \zeta_0\right), \\
&= \sqrt{n}\boldsymbol{v}^\top\left[c\left(\boldsymbol{1}\boldsymbol{1}^\top\right) + (1 - \rho)^{-1/2}\mathsf{I}\right]\left(\hat{\zeta} - \zeta_0\right), \\
&= \sqrt{\frac{n}{1 - \rho}}\boldsymbol{v}^\top\hat{\zeta}.
\end{aligned}$$

Here we have used that $\boldsymbol{v}^\top \mathbf{1} = 0$ and under the null hypothesis, $\boldsymbol{\zeta}_0$ is some constant times $\mathbf{1}$. Thus

$$\hat{\zeta}_i - \hat{\zeta}_j = \sqrt{\frac{1-\rho}{n}} \left(z_i - z_j\right). \tag{6}$$

Now note that the $\boldsymbol{z}$ is distributed as a standard multivariate normal. So the *range* of $\hat{\boldsymbol{\zeta}}$, which is to say $\max_{i,j} \left(\hat{\zeta}_i - \hat{\zeta}_j\right)$, is distributed as $\sqrt{(1-\rho)/n}$ times the range of a standard k-variate normal.

To quote this as a hypothesis test,

$$\max_{i,j} \left|\hat{\zeta}_i - \hat{\zeta}_j\right| \geq HSD = q_{1-\alpha,k,\infty} \sqrt{\frac{(1-\rho)}{n}}, \tag{7}$$

with probability α, where the $q_{1-\alpha,k,l}$ is the upper α-quantile of the Tukey distribution with k and l degrees of freedom. In the R language, this quantile may be computed via the `qtukey` function. [12, 9] With $l = \infty$, the cutoff HSD is the rescaled upper α quantile of the range of k independent Gaussians. That is, $q_{1-\alpha,k,\infty}$ is the number such that

$$1 - \alpha = k \int_{-\infty}^{\infty} \phi\left(x\right) \left(\Phi\left(x + q_{1-\alpha,k,\infty}\right) - \Phi\left(x\right)\right)^{k-1} \, \mathrm{d}x.$$

We note that the approximation of Equation 1 may be too coarse for the computation of the HSD cutoff. Even if the covariance given there is approximately correct, it is likely that the distributional shape of $\hat{\boldsymbol{\zeta}}$ is far enough from multivariate normal that we cannot use Tukey's distribution for a cutoff, especially when n is small and k is large. In that case, one is tempted to heuristically compare the observed range to

$$HSD = q_{1-\alpha,k,n-1} \sqrt{\frac{(1-\rho)}{n-1}}. \tag{8}$$

The reasoning here is that we are essentially computing the range of (non-independent) t statistics, up to scaling, which is almost the same as the Tukey distribution, which is the ratio of the range of normals divided by a pooled χ variable. In our testing below we will refer to the cutoff of Equation 7 as "$df = \infty$" and the cutoff of Equation 8 as the "$df = n - 1$" cutoff.

Bonferroni Cutoff: We note that an alternative calculation provides a very similar cutoff value. Considering two assets with correlation ρ. Suppose the signal-noise ratios of the two assets are, respectively, $\zeta\left(1 + \epsilon\right)$ and ζ. The difference in Sharpe ratios can then be shown to be approximately normal: [10]

$$\left[\hat{\zeta}_1 - \hat{\zeta}_2\right] \rightsquigarrow \mathcal{N}\left(\epsilon\zeta, \frac{2}{n}\left(1 - \rho\right) + \frac{\zeta^2}{2n}\left(1 + (1 + \epsilon)^2 - 2\rho^2\left(1 + \epsilon\right)\right)\right). \tag{9}$$

Assuming that ζ^2/n will be very small for most practical work, one can compute the alternative cutoff, a "Bonferroni Cutoff," as

$$BC = \sqrt{\frac{2\left(1-\rho\right)}{n}} z_{1-\alpha/\binom{k}{2}},$$

3

where z_α is the α quantile of the standard normal distribution. This cutoff is based on a Bonferroni correction that recognizes we are performing $\binom{k}{2}$ pairwise comparison tests. The cutoff BC is typically very similar to HSD (for $df = \infty$) or slightly smaller (and is easier to compute). We note that since BC is based on a normal approximation, it may suffer from the same issues that the HSD cutoff does for small samples. However, there is hope one can compute an exact small n Bonferroni cutoff.

Arbitrary correlation structure: The test outlined above is strictly only applicable to the rank-one correlation matrix, $\mathsf{R} = \rho \left(\mathbf{1}\mathbf{1}^\top \right) + (1 - \rho)\,\mathsf{I}$. To apply the test to assets with arbitrary correlation matrices, one would like to appeal to a stochastic dominance result. For example, if one could adapt Slepian's lemma to the distribution of the *range*, then the above analysis could be applied where ρ is the smallest off-diagonal correlation, to give a test with maximum type I rate of α. However, it is not immediately clear that Slepian's lemma can be so modified. [15, 19, 18] The Bonferroni Cutoff, however, is easily adapted to this kind of worst-case analysis, however.

3 Examples

3.1 Simulations under the null

Basic Simulations We spawn 4 years of daily data (252 days per year) from 16 assets, each with signal-noise ratio of $1\mathrm{yr}^{-1/2}$. Returns are multivariate normal with correlation $\mathsf{R} = \rho \left(\mathbf{1}\mathbf{1}^\top \right) + (1 - \rho)\,\mathsf{I}$, for $\rho = 0.8$. We compute the Sharpe ratio of the simulated returns, $\hat{\boldsymbol{\zeta}}$, then compute the range $\max_{i,j} \hat{\zeta}_i - \hat{\zeta}_j$. We repeat this experiment 5,000 times. In Figure 1, we Q-Q plot these simulated ranges of the Sharpe ratio against the theoretical quantile function

$$\sqrt{\frac{1 - \rho}{n}} q_{\cdot,k,\infty}.$$

We see good agreement between theoretical and actual, with little deviance from the $y = x$ line.

We then convert these simulated ranges to p-values via the `ptukey` function in R, using the $df = \infty$ cutoff and the actual ρ. We Q-Q plot these putative p-values against a uniform law in Figure 2, and again find very good agreement. To check the tails we have transformed the p-values to $p \mapsto |2p - 1|$ and plotted in log-log scale.
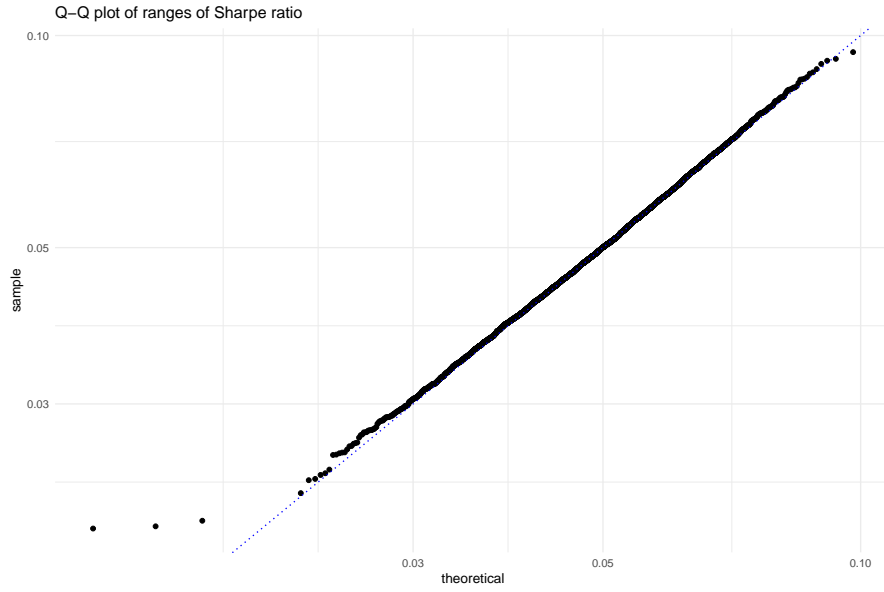
Q–Q plot of ranges of Sharpe ratio

Figure 1: The quantiles of the range of Sharpe ratio from 5,000 simulations are plotted against a transformed Tukey distribution, $\sqrt{(1-\rho)/n}\, q_{\cdot,k,\infty}$. The points show little deviation from the plotted $y = x$ line.
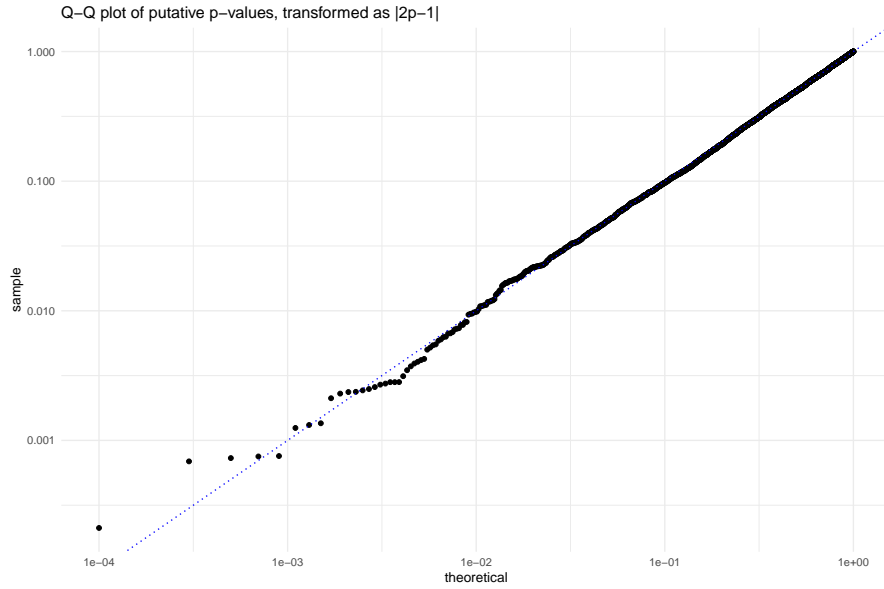

Q–Q plot of putative p–values, transformed as |2p−1|

Figure 2: The computed p-values from 5,000 simulations are plotted against a uniform law, visually confirming that the p-values are nearly uniform. Simulations use the exact ρ to compute the p-values via `ptukey`. We transform the p-values and plot $|2p - 1|$ in log-log space to emphasize the tails. The points show little deviation from the plotted $y = x$ line.

Varying n and k: Next we perform the same kind of simulations, but vary the number of days observed in each simulation, n, as well as the number of different assets, k. We let the former vary from 20 to 1,280 measured in days, and the latter vary from 8 to 32. We take $\mathsf{R} = \rho \left(\mathbf{1}\mathbf{1}^\top \right) + (1 - \rho)\,\mathsf{I}$, for $\rho = 0.8$ and set the signal-noise ratio to $1\mathrm{yr}^{-1/2}$. We assume 252 days per year for annualizing the Sharpe ratio. For each set of $50,000$ simulations, we compute the empirical type I rate at the nominal 0.05 level by comparing the range to the HSD cutoff. We tabulate rejections using both the $df = \infty$ and $df = n - 1$ cutoffs.

We plot that type I rate against n in Figure 3 for the different values of k. For the $df = \infty$ cutoff, the procedure is apparently anticonservative, yielding too many type I errors, when the sample size is small and the number of assets is large. For the $df = n - 1$ cutoff, however, the nominal type I rate is approximately achieved.
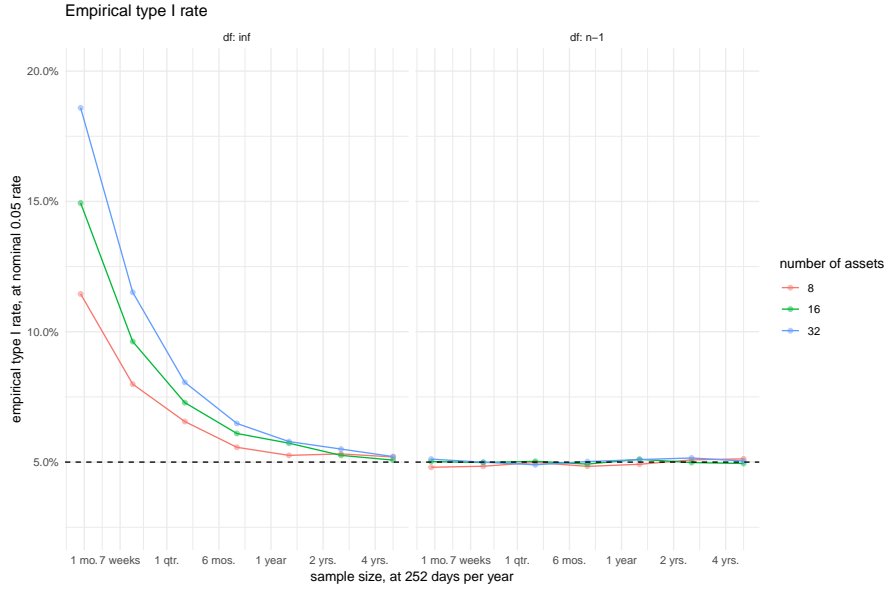

Figure 3: The empirical type I rate at the nominal 0.05 level is plotted against the number of days in each simulation, for different values of k. Simulations use the exact ρ to perform the hypothesis test. The two facets show rejection rates under the $df = n - 1$ and $df = \infty$ cutoffs. When using the $df = \infty$ cutoff, the test is anti-conservative for the "large k, small n" case, but the nominal rate is nearly achieved for the $df = n - 1$ cutoff.

Varying ρ: We next perform the same simulations under the null, with $\mathsf{R} = \rho \left(\mathbf{1}\mathbf{1}^\top \right) + (1 - \rho)\,\mathsf{I}$, but scanning through ρ. We set $n = 1,008$ days, $k = 16$, and set the signal-noise ratio to $1\mathrm{yr}^{-1/2}$. We compute the empirical type I rate at the nominal 0.05 level for each set of $50,000$ simulations. We plot that type I rate against ρ in Figure 4. For these values of n, k, the procedure achieves near nominal type I rate, and does not vary in a systematic way with ρ.
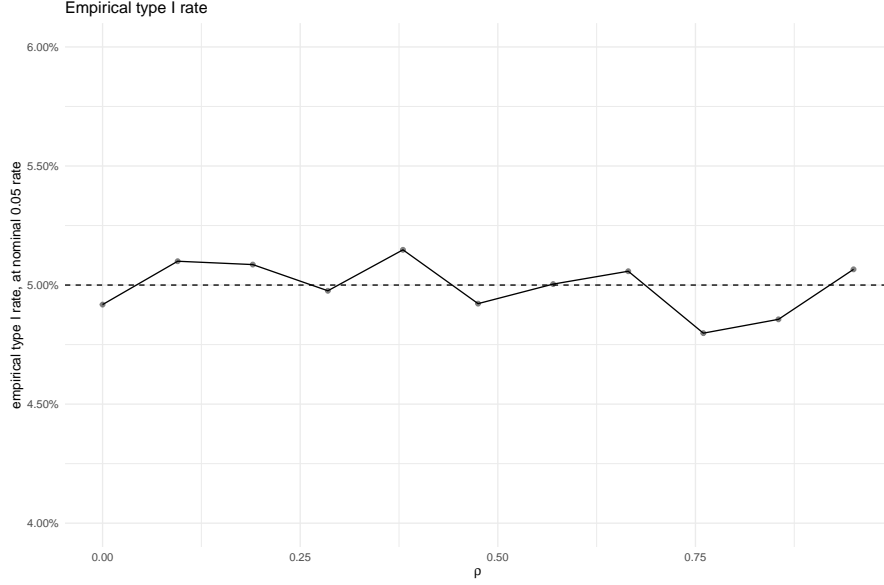
Figure 4: The empirical type I rate at the nominal 0.05 level is plotted against the correlation, ρ. Simulations use the exact ρ to perform the hypothesis test, and the $df = n - 1$ cutoff.

Feasible Estimator, Varying ρ: In the simulations above we have used the actual ρ in computing the threshold for rejection of the null. We repeat the experiments using a *feasible* test where we esimate ρ from the sample. We compute the correlation of returns, then take the median value of the upper triangle of the correlation matrix.

In the first set of simulations, the true correlation matrix follows $\mathsf{R} = \rho \left(\mathbf{1}\mathbf{1}^{\top} \right) + (1 - \rho) \, \mathsf{I}$. We set $n = 1,008$ days, $k = 16$, and set the signal-noise ratio to $1\mathrm{yr}^{-1/2}$. We compute the empirical type I rate at the nominal 0.05 level for each set of $50,000$ simulations. We plot that type I rate against ρ in Figure 5. For these values of n, k, the procedure achieves near nominal type I rate, and does not appear to suffer from having estimated the ρ. In fact the plot greatly resembles Figure 4 where we have used the actual ρ.

Feasible Estimator, Misspecified Model, Varying ρ: We repeat those simulations, estimating the ρ from the sample, but now we let the correlation matrix take an "AR(1)" structure. That is, we let $\hat{\mathsf{R}}^{i,j} = \rho^{|i-j|}$, and vary ρ. Again we have $n = 1,008$ days, $k = 16$, the signal-noise ratio is equal to $1\mathrm{yr}^{-1/2}$. We compute the empirical type I rate at the nominal 0.05 level for each set of $50,000$ simulations.

For these simulations, we also record the type I rate when the ρ is not estimated, but instead assumed to be 0. Given that $\rho = 0$ forms a kind of 'stochastic lower bound', we expect that the procedure will be anti-conservative when performed this way. Indeed we see in the plot that the empirical type I rate decreases to zero in increasing ρ. For the case where we take the median sample correlation as the estimate of ρ, the procedure is somewhat conservative
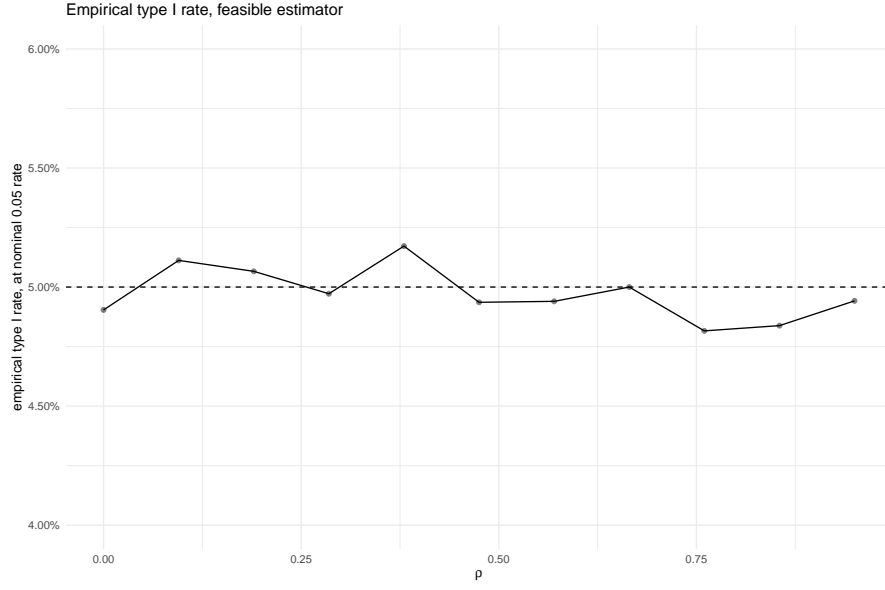
Figure 5: The empirical type I rate at the nominal 0.05 level is plotted against the correlation, ρ. Simulations use an *estimated* ρ to perform the hypothesis test, and the $df = n - 1$ cutoff.

for small ρ, then anti-conservative for large ρ. This is not surprising: for large ρ, the median element of $\hat{\mathsf{R}}$ will be fairly large, but the correlation among assets is somewhat weak. A more robust heuristic for estimating the ρ is needed.
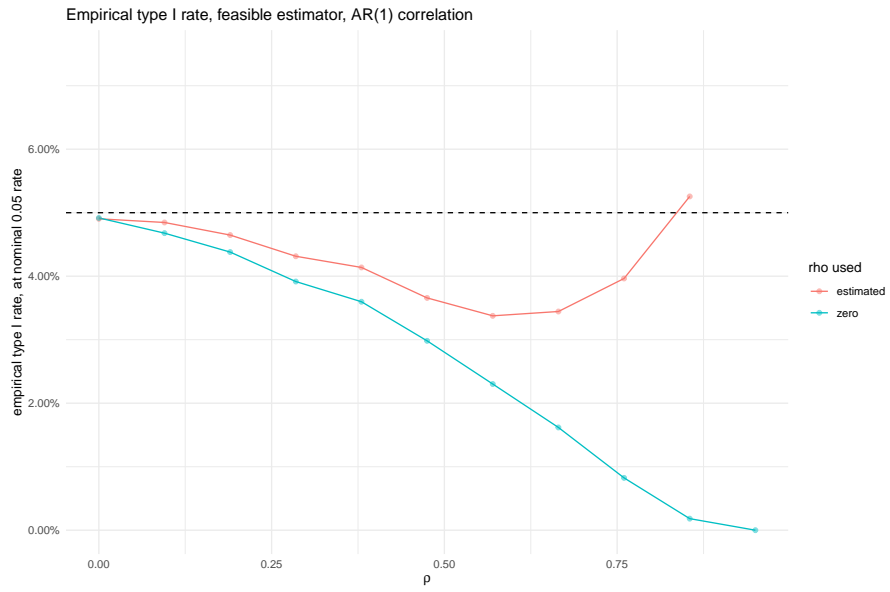
Figure 6: The empirical type I rate at the nominal 0.05 level is plotted against the correlation, ρ, for simulations where the correlation follows an AR(1) structure. Simulations use an *estimated* ρ to perform the hypothesis test, and the $df = n - 1$ cutoff. We include separate lines for the cases where ρ is estimated, and for where it is assumed to equal 0. For the estimated ρ, the procedure is conservative for small and moderate ρ, but anticonservative for ρ near 1. When $\rho = 0$ is assumed, the procedure is increasingly conservative in ρ.

3.2 Simulations under the alternative

We next perform the same simulations under the alternative. It is somewhat difficult to quantify the power of this procedure because the procedure can reject multiple nulls for a given experiment. Indeed, in the simulations under the null above, we analyzed the rate of *any* rejections for the multiple comparisons performed in a single simulation.

Under the alternative, one good: In the first set of simulations we let ζ have a single non-zero value, call it ζ, and vary that ζ. We then compute, as the 'range', the Sharpe ratio of the single good asset minus the minimum Sharpe ratio of the $k - 1$ remaining assets. Because we are only testing $k - 1$ comparisons, rather than $\binom{k}{2}$, we expect to see fewer than the nominal type I rate when $\zeta = 0$. Moreover, we are performing a one-sided test. As such it may be more natural to compare the rejection rate to $\alpha/2$.

We take $\mathsf{R} = \rho \left(\mathbf{1}\mathbf{1}^\top \right) + (1 - \rho) \, \mathsf{I}$, letting ρ vary from 0 to 0.9; we set $n = 1,008$ days, $k = 16$, and let ζ vary from $0\mathrm{yr}^{-1/2}$ to $1.5\mathrm{yr}^{-1/2}$.

We compute the rejection rate at the nominal 0.05 level for each set of $10,000$ simulations. We plot that (true) rejection rate against ζ in Figure 7. For these values of n, k, the procedure is fairly weak, only achieving power of one half for large ζ or highly correlated assets. It is not surprising that the power is increasing in ρ: one expects less spread among the assets for higher ρ, thus a true difference in signal-noise ratio is more easily detected. This same effect is visible in the paired test for equality of signal-noise ratios. [10]

Under the alternative, half good: We repeat those experiments, but set half the 16 assets to have signal-noise ratio equal to ζ, and the rest to have zero signal-noise ratio. We compute, as the 'range', the maximum Sharpe ratio of the good assets minus the minimum Sharpe ratio of the $k - 1$ remaining assets. We are effectively testing $(k/2)^2$ comparisons, Because we are only testing $k - 1$ comparisons, rather than $\binom{k}{2}$, so we expect to see fewer than the nominal type I rate when $\zeta = 0$.

As above we take $\mathsf{R} = \rho \left(\mathbf{1}\mathbf{1}^\top \right) + (1 - \rho) \, \mathsf{I}$, let ρ vary from 0 to 0.9, $n = 1,008$ days, $k = 16$, and let ζ vary from $0\mathrm{yr}^{-1/2}$ to $1.5\mathrm{yr}^{-1/2}$.

We compute the rejection rate at the nominal 0.05 level for each set of $10,000$ simulations. We plot that (true) rejection rate against ζ in Figure 8. For these values of n, k, the procedure is again fairly underpowered, with higher power for more correlated assets.
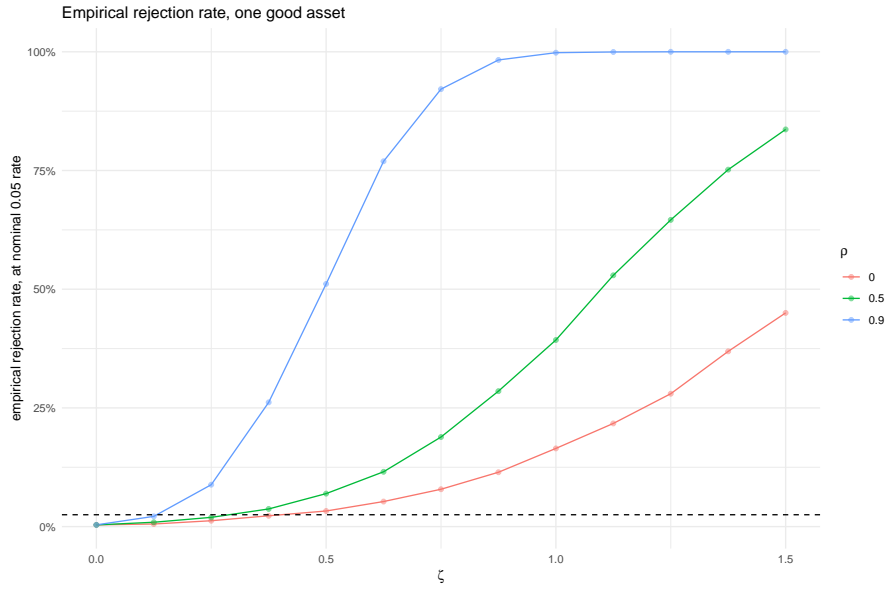
Figure 7: The empirical rejection rate at the nominal 0.05 level is plotted against the correlation, ρ. The population consists of one good asset with signal-noise ratio equal to ζ, and the remainder with zero signal-noise ratio. Rejection is based on the Sharpe ratio of the single good asset minus the minimum Sharpe ratio of the remaining assets. Simulations use the exact ρ to perform the hypothesis test, and the $df = n - 1$ cutoff. We plot a horizontal line at half the nominal type I rate, 0.025, because we are performing a one-sided test.
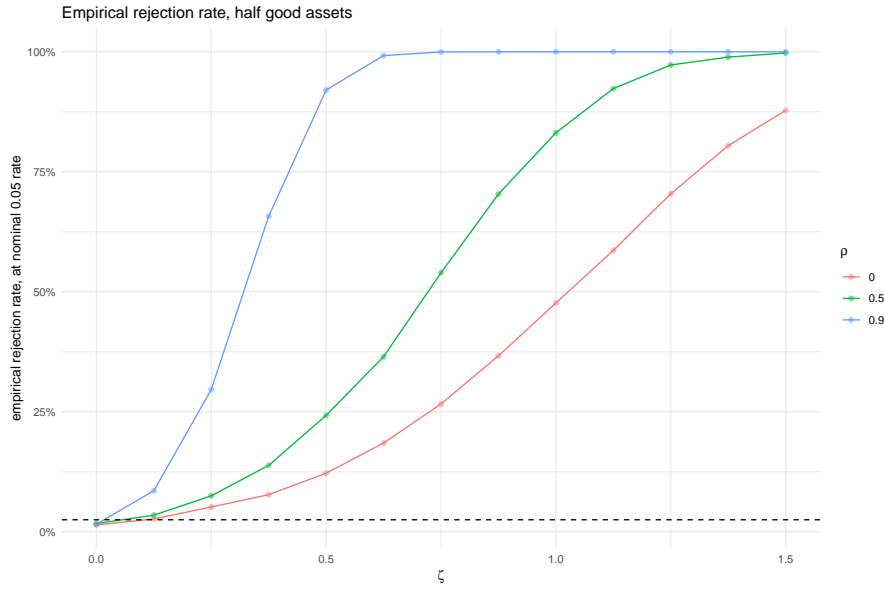
Figure 8: The empirical rejection rate at the nominal 0.05 level is plotted against the correlation, ρ. The population consists of half good assets with signal-noise ratio equal to ζ, and the remainder with zero signal-noise ratio. Rejection is based on the maximum Sharpe ratio of the good assets minus the minimum Sharpe ratio of the remaining assets. Simulations use the exact ρ to perform the hypothesis test, and the $df = n - 1$ cutoff. We plot a horizontal line at half the nominal type I rate, 0.025, because we are performing a one-sided test.

3.3 Real Assets

We now apply the technique to real asset returns.

Five Industry Portfolios: We consider the 5 industry portfolios, whose returns are computed and distributed by French. [3] The dataset consists of 1104 months of returns, from Jan 1927 to Dec 2018. The returns are highly correlated, and the correlation matrix is likely well modeled by the form $\rho \left(\mathbf{1} \mathbf{1}^{\top} \right) + (1 - \rho) \mathsf{I}$, with ρ estimated as approximately 0.8. The Sharpe ratios range from $0.485 \text{yr}^{-1/2}$ for Other to $0.667 \text{yr}^{-1/2}$ for Healthcare.

First we perform the hypothesis test of equality of signal-noise ratios, as proposed by Wright *et al.* [17] We compute a statistic of 12.2 which should be distributed as a $\chi^2 (4)$ under the null. [10] This corresponds to a p-value of 0.016, and we reject the null of equality of all signal-noise ratios.

Using the $df = n - 1$ formulation and the estimated ρ, we compute $HSD = 0.18 \text{yr}^{-1/2}$ for $\alpha = 0.05$, and narrowly reject the equality of signal-noise ratios for Other and Healthcare. In Figure 9, we plot these Sharpe ratios, along with error bars at plus and minus one HSD.
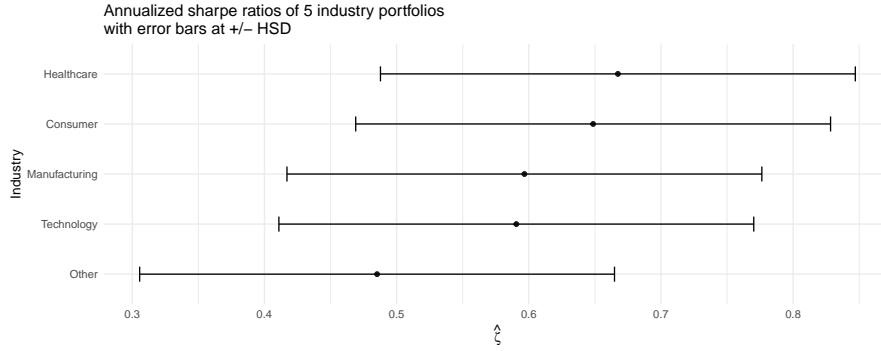

Figure 9: The annualized Sharpe ratio of French's 5 industry portfolios are plotted, as computed on monthly returns from Jan 1927 to Dec 2018. We plot error bars at $\pm HSD$ for $\alpha = 0.05$. We narrowly reject equality of the signal-noise ratio of Other and Healthcare.

Sharpe's 34 Mutual Funds: We consider the returns of the 34 mutual funds described by Sharpe in his original paper. [14] We transcribed the annualized percent return and standard deviation values from Sharpe's Table I. In his paper, Sharpe computed the "reward-to-variability ratio" of each using a fixed rate of 3%; however, we compute the Sharpe ratio without subtracting a fixed rate. The Sharpe ratios range from $0.549 \text{yr}^{-1/2}$ for Incorporated Investors to $1.087 \text{yr}^{-1/2}$ for American Business Shares.

We do not have access to the series of returns, and cannot estimate the correlation structure. Somewhat optimistically we make the wild guess $\rho = 0.85$. Based on this value, and setting $\alpha = 0.05$, we compute $HSD = 0.68 \text{yr}^{-1/2}$, and we fail to reject the null hypothesis that all signal-noise ratios are equal. In Figure 10, we plot these Sharpe ratios, along with error bars at $\pm HSD$. Given

the lack of separation of the funds, it is curious that Sharpe found correlation between the in-sample and out-of-sample Sharpe ratios of his funds. [14]
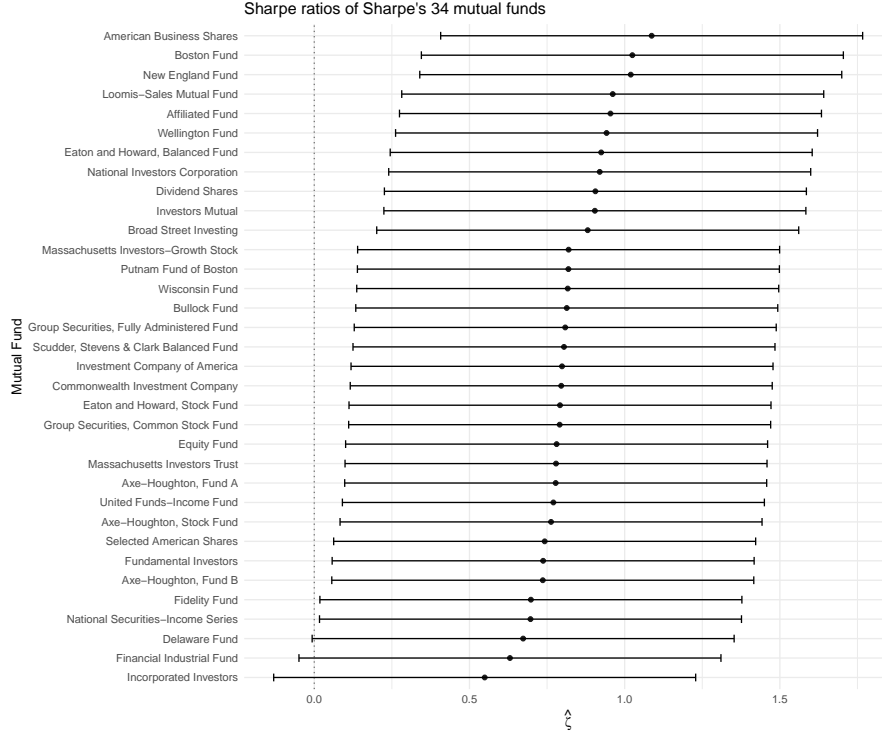

Figure 10: The annualized Sharpe ratio of Sharpe's 34 mutual funds are plotted. Returns are from the decade 1954-1963. [14] We plot error bars at $\pm HSD$. We fail to reject the nulls that all pairwise differences are equal.

4   Future Work

A number of issues remain outstanding:

1. The heuristic use of the $df = n-1$ cutoff requires theoretical justification.

2. A stochastic inequality like Slepian's lemma for ranges would allow one to apply the test using a lower bound ρ to achieve maximum type I rate.

3. Should we expect the Tukey HSD cutoff and the Bonferroni Cutoff to be nearly equal, or will one dominate the other under certain conditions?

4. Can we quantify the power of the test?

5. Though we suspect it cannot, can the power of the test be improved?

14

References

[1] L. J. Bain and M. Engelhardt. *Introduction to Probability and Mathematical Statistics*. Classic Series. Cengage Learning, 1992. ISBN 9780534380205. URL http://books.google.com/books?id=MkFRIAAACAAJ.

[2] Frank Bretz, Torsten Hothorn, and Peter Westfall. *Multiple comparisons using R*. Chapman and Hall/CRC, 2016. URL http://www.ievbras.ru/ecostat/Kiril/R/Biblio_N/R_Eng/Bretz2011.pdf.

[3] Kenneth French. 5 industry portfolios. Privately Published, 2019. URL http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library/det_5_ind_port.html.

[4] J. D. Jobson and Bob M. Korkie. Performance hypothesis testing with the Sharpe and Treynor measures. *The Journal of Finance*, 36(4):pp. 889–908, 1981. ISSN 00221082. URL http://www.jstor.org/stable/2327554.

[5] N. L. Johnson and B. L. Welch. Applications of the non-central t-distribution. *Biometrika*, 31(3-4):362–389, mar 1940. doi: 10.1093/biomet/31.3-4.362. URL http://dx.doi.org/10.1093/biomet/31.3-4.362.

[6] Pui-Lam Leung and Wing-Keung Wong. On testing the equality of multiple Sharpe ratios, with application on the evaluation of iShares. *Journal of Risk*, 10(3):15–30, 2008. URL http://www.risk.net/digital_assets/4760/v10n3a2.pdf.

[7] Andrew W. Lo. The statistics of Sharpe ratios. *Financial Analysts Journal*, 58(4), July/August 2002. URL http://ssrn.com/paper=377260.

[8] 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. doi: 10.2307/2330636. URL https://doi.org/10.2307/2330636.

[9] R. E. Odeh and J. O. Evans. Algorithm AS 70: The percentage points of the normal distribution. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 23(1):96–97, 1974. ISSN 00359254, 14679876. doi: 10.2307/2347061. URL http://www.jstor.org/stable/2347061.

[10] Steven E. Pav. A short Sharpe course. Privately Published, 2017. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3036276.

[11] Steven E. Pav. Conditional inference on the asset with maximum Sharpe ratio. Privately Published, 2019. URL http://arxiv.org/abs/1906.00573.

[12] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL http://www.R-project.org/. ISBN 3-900051-07-0.

[13] A. D. Roy. Safety first and the holding of assets. *Econometrica*, 20(3):pp. 431–449, 1952. ISSN 00129682. URL http://www.jstor.org/stable/1907413.

[14] William F. Sharpe. Mutual fund performance. *Journal of Business*, 39:119, 1965. URL http://ideas.repec.org/a/ucp/jnlbus/v39y1965p119.html.

[15] David Slepian. The one-sided barrier problem for Gaussian noise. *Bell System Technical Journal*, 41(2):463–501, 1962. doi: 10.1002/j.1538-7305.1962.tb02419.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1538-7305.1962.tb02419.x.

[16] John W. Tukey. Comparing individual means in the analysis of variance. *Biometrics*, 5(2):99–114, 1949. ISSN 0006341X, 15410420. doi: 10.2307/3001913. URL http://www.jstor.org/stable/3001913.

[17] John Alexander Wright, Sheung Chi Phillip Yam, and Siu Pang Yung. A test for the equality of multiple Sharpe ratios. *Journal of Risk*, 16(4), 2014. URL http://www.risk.net/journal-of-risk/journal/2340044/latest-issue-of-the-journal-of-risk-volume-16-number-4-2014.

[18] Chuancun Yin. Stochastic orderings of multivariate elliptical distributions, 2019. URL https://arxiv.org/abs/1910.07158.

[19] Ofer Zeitouni. GAUSSIAN FIELDS notes for lectures. unpublished notes, 2017. URL http://www.wisdom.weizmann.ac.il/~zeitouni/notesGauss.pdf.