Testing strategies based on multiple signals

Robert Novy-Marx[*]

March 2016

Abstract

Strategies selected by combining multiple signals suffer severe overfitting biases, because underlying signals are typically signed such that each predicts positive in-sample returns. As a result, "highly significant" backtested performance is easy to generate selecting stocks using combinations of randomly generated signals, which by construction have no true power. This paper analyzes t-statistic distributions for multi-signal strategies, both empirically and theoretically, to determine appropriate critical values, which can be several times standard levels. Overfitting bias also severely exacerbates the multiple testing bias that arises when investigators consider more results than they present. Combining the best k out of n candidate signals yields biases similar to those obtained using the single best of n^k candidate signals.

Keywords: Data mining, bias, inference, stock selection.

JEL classification: G11, C58.

[*] Simon Graduate School of Business, University of Rochester, 500 Joseph C. Wilson Blvd., Box 270100, Rochester, NY 14627. Email: robert.novy-marx@simon.rochester.edu.

1. Introduction

Multi-signal equity strategies, i.e., strategies that select or weight stocks on the basis of composite measures that combines multiple signals, are common in the money management industry. For example, the MSCI Quality Index, according to its fact sheet, identifies "stocks with high quality scores based on three main fundamental variables: high return on equity (ROE), stable year-over-year earnings growth, and low financial leverage." Popular "smart beta" products, such as Research Affiliates' Fundamental Indices, also rely heavily on the methodology, weighting stocks on the basis of multiple "fundamental" measures such as sales, cash flow, book value, and dividends.

Increasingly, multi-signal strategies are attracting scholarly attention. Notable examples of composite signals employed by academics for stock selection include Piotroski's (2000) F-score measure of financial strength, constructed from nine market signals; the Gompers, Ishii, and Metrick's (2003) Index, which combines 24 governance rules to proxy for shareholder rights; the Baker and Wurgler (2006) Index, which combines six signals of investor sentiment; Asness, Frazzini, and Pedersen's (2013) quality score, which combines 21 stock level characteristics; and Stambaugh and Yuan's (2015) mispricing factors, which combine 11 anomaly signals.

Unfortunately, inferences drawn from tests of these sorts of strategies are misleading, because the backtested performance of strategies selected on the basis of multiple signals is biased, often severely. The bias results from overfitting. Underlying signals are typically signed such that each predicts positive in-sample returns. That is, an aspect of the data (in-sample performance of the individual signals) is used when constructing the strategy, yielding a particular pernicious form of the data-snooping biases discussed in Lo and MacKinlay (1990).

This overfitting bias is distinct from the selection bias (or multiple testing bias) confronted by McLean and Pontiff (2016), Harvey, Liu, and Zhu (2016), and Baily and Lopez de Prado (2014). Selection bias results when the researcher employs the best performing signal from among multiple candidates, and fails to account for doing so. The overfitting bias considered here is strong even when there is no selection bias, i.e., even when a researcher employs each and every signal considered.

While the overfitting and selection biases are distinct they do interact, with the selection bias severely exacerbating the overfitting bias. In fact, the presence of the two makes the bias exponentially worse. The bias resulting from combining the best k signals from a set of n candidates is almost as bad as that from using the single best signal out of n^k candidates.

To demonstrate the severity of the overfitting bias I construct, using real stock returns, empirical distributions of backtested t-statistics for multi-signal strategies, selected on the basis of purely random signals. The sorting variables (i.e., the random signals) have no real power, but strategies based on combinations of the "signals" perform strongly. Essentially, diversifying across the recommendations of stock pickers that performed well in the past yields even better past performance, even when the recommendations just follow (or go against) the results of monkeys throwing darts at the Wall Street Journal. This strong backtested performance in no way suggests, of course, that these recommendations have any power predicting returns going forward. For some of the constructions I consider strategies usually backtest, in real data, with t-statistics in excess of five, and statistical significance at the 5% level requires t-statistics in excess of seven, despite the fact that the variable used to select stocks has no power predicting returns.

To develop intuition for the observed empirical results I derive theoretical distributions for critical t-statistics, under the null that signals are uninformative and strategy returns are normally distributed. These critical values, which have close analytic approximations, are similar to those observed in the data. Analysis of these results yields several additional intuitions. First, it suggests that the overfitting bias is severely exacerbated, at least when there is little selection bias, when more weight is put on stronger signals. That is, when researchers allow themselves more flexibility to overfit the data, in the form of freedom to weight different signals differently, then backtested performance is significantly more biased. It also suggests that when researchers constrain themselves to weight signals equally, then the optimal use of the roughly half of the signals that backtest least strongly, at least for the purpose of maximizing backtested t-statistics, is to simply ignore them. Finally, the model implies the approximate power law for the interaction of the overfitting and selection biases, that the bias that results from combining the best k out of n candidate signals yields biases almost as large as those that result from selecting the single best of n^k candidate signals. This suggested theoretical relation also holds in the data.

Note that these results do not suggest that strategy performance cannot be improved by combining multiple signals. The basic tenants of Markowitz's (1952) modern portfolio theory hold, and efficient combinations of high Sharpe ratio assets have even higher Sharpe ratios. The results do strongly suggest, however, that the marginal contribution of each individual signal should be evaluated individually. That is, while one should combine multiple signals they believe in, one should not believe in a combination of signals simply because they backtest well together.

2. Empirical results

This section considers the backtested performance of strategies selected by combining random signals. It generates empirical distributions of backtested t-statistics for a general class of multi-signal strategies, when signals are uninformative about expected returns, by considering combinations of multiple random signals millions of times. By construction, these signals have no real power, and cannot predict performance out-of-sample.

2.1. Strategy construction

2.1.1. Single signal strategies

Given a signal for stock selection, the associated strategy's returns are constructed by weighting the returns to individual stocks in proportion to the signal. That is, the associated strategy's returns are given by

$$r^i = \frac{\sum_{j=1}^N S_j^i r_j}{\sum_{j=1}^N S_j^i}, \tag{1}$$

where the superscripts correspond to the signal, the subscripts to individual stocks, and time has been suppressed for notational convenience.[1]

The strategy's active return relative to the benchmark, average stock return $r^{bmk} = \frac{1}{N}\sum_{j=1}^N r_j$ is then

$$\begin{aligned} r^i - r^{bmk} &= \frac{\sum_{j=1}^N S_j^i \left(r_j - r^{bmk}\right)}{\sum_{j=1}^N S_j^i} \\ &= \frac{\sum_{j=1}^N \left(S_j^i - S^i\right)\left(r_j - r^{bmk}\right)}{\sum_{j=1}^N S_j^i} \\ &= \frac{\sum_{j=1}^N \left(S_j^i - S^i\right) r_j}{\sum_{j=1}^N S_j^i}, \end{aligned} \tag{2}$$

where $S^i \equiv \frac{1}{N}\sum_{j=1}^N S_j^i$ denotes the average signal, and the second and third equalities follow from the definitions of r^{bmk} and S^i, respectively. The performance of this active return is the focus of this study.

[1] For expositional simplicity I focus here on strategies constructed without regard for market capitalizations (i.e., purely signal-weighted strategies). Appendix A.1 provides similar results for strategies that weight individual stocks in proportion to both their signals and their market capitalizations (i.e., signal- and capitalization-weighted strategies). The more general specification considered there embeds many common construction schemes, including standard quantile sorts, as well as the rank-weighting scheme employed by Frazzini and Pedersen (2014) to construct their betting-against-beta (BAB) factor.

Note that any results related to this active return apply equally to the performance of long/short strategies, as the active return to the long-only strategy considered here weights stocks in proportion to how far stocks' signals are from the mean signal, and is thus a long/short strategy constructed on the basis of the signal. In terms of the returns to a long/short strategy based on the signal that invests one dollar on each the long and short sides,

$$r^{i,L/S} \quad = \quad \frac{\sum_{j=1}^{N}\left(S_j^i - S^i\right) r_j}{\frac{1}{2}\sum_{j=1}^{N}\left|S_j^i - S^i\right|}, \tag{3}$$

the active returns to the long-only strategy can be written as

$$r^i - r^{bmk} \quad = \quad \left(\frac{\frac{1}{2}\sum_{j=1}^{N}\left|S_j^i - S^i\right|}{\sum_{j=1}^{N} S_j^i}\right) r^{i,L/S}. \tag{4}$$

A dollar in the long-only strategy is thus equal to a dollar in the benchmark strategy, plus a tilt toward the one dollar long/one dollar short strategy. The size of this tilt, i.e., the leverage on the dollar long/dollar short strategy, is essentially determined by the signal-to-noise ratio of the stock selection signal. If the signal is normally distributed, then the expected leverage multiplier is $\sigma_S/\mu_S/\sqrt{2\pi} \approx 0.4\sigma_S/\mu_S$. Because the average signal is diversified across the individual stocks' signals, there is little variation in the leverage multiplier, provided the population mean of the signal is not too small.

2.1.2. Implementation

Given a signal, I construct signal-weighted strategy performance using stock return data from CRSP, rebalancing annually at the beginning of each year, over the 20 year sample spanning January 1995 through December 2014. The sample length, short for academic studies but common in practitioner oriented studies, is largely irrelevant given the paper's focus on t-statistics. It has the advantage of somewhat alleviating the computational burden posed when simulating millions of strategies constructed using the entire cross section.

Individual underlying signals are randomly generated, drawn independently for each stock at the start of each year from a normal distribution with positive mean. A strategy's active return relative to the benchmark is thus a long/short strategy that weights stocks relative to individual stocks' z-scores for the signal, a construction that is common in industry. Empirically the performance of a z-score weighted long/short strategy is almost indistinguishable from the performance of a quantile sorted strategy that buys and sells stocks in the top and bottom 35% of the signal distribution. For example, the returns to value (or momentum) strategies constructed using z-score weighting, a quantile sort, and the rank-weighting procedure of Frazzini and Pedersen (2014), explain between 97.8 and 99.6% of the return variation of the others, and none of the strategies yield abnormal returns relative to either of the others that exceed six basis points per month (details provided in Appendix A.2).

2.1.3. Strategies based on multiple signals

Let $\boldsymbol{S^{\omega}} = \sum_{i=1}^{n} \omega^i S_j^i$ be a composite signal, constructed as any linear combination of any n underlying signals. Then the returns to the strategy that weights stocks on the basis of the composite signal is

$$\begin{aligned} r^{\boldsymbol{\omega}} &= \frac{\sum_{j=1}^{N} \left(\sum_{i=1}^{n} \omega^i S_j^i\right) r_j}{\sum_{j=1}^{N} \left(\sum_{i=1}^{n} \omega^i S_j^i\right)} \\ &= \frac{\sum_{i=1}^{n} \omega^i \left(\sum_{j=1}^{N} S_j^i r_j\right)}{\sum_{i=1}^{n} \omega^i \left(\sum_{j=1}^{N} S_j^i\right)} \qquad (5) \\ &= \sum_{i=1}^{N} w^i r^i, \end{aligned}$$

where $w^j \equiv \omega^j S^j / \sum_{i=1}^{N} \omega^i S^i$. That is, the returns to the strategy based on the composite signal is just a weighted average of the returns to the strategies based on the individual signals, where the weights are proportional to the individual signals' cross-sectional averages and their weights in the composite signal.[2] This establishes an exact correspondence between the performance of integrated strategies (i.e., those based on composite signals) and siloed strategies (i.e., those that allocate resources across strategies based on single signals). This correspondence is easily observed in practice. For example, over the July 1963 to December 2014 sample, the active returns to a strategy based on a 50/50 mix of stocks' z-scores for value and momentum is completely indistinguishable from the active return to a 50/50 mix of the strategies based on the individual z-scores.[3]

2.2. Best k-of-n strategy performance

What type of strategy arises naturally out of a typical research process? A researcher may have a concept or model that suggests a certain sort of stocks will have higher returns. She then investigates a set of possible empirical proxies that she thinks might signal stocks that look good on this dimension, and chooses to somehow combine the few that work the best. If an investigator considers n signals, and combines the best k of these to select stocks, the result is a best k-of-n strategy. When $k = 1$ this results in pure multiple testing, or selection, bias. This bias is relatively well understood, and interesting here primarily as a point of comparison.

[2] The weights on the single-signal strategy returns are approximately directly proportional to the individual signals' weights in the composite signal if the individual signals are identically distributed, because with a large cross section each signal's mean is close to its ex ante expected value, and thus essentially equal across identically distributed signals.

[3] Strategies here are restricted to stocks for which, at the time of portfolio formation, there are current data for both value and momentum (positive book-to-market, and stock performance over the first eleven months of the preceding year, respectively). To ensure that the strategies take only long positions, which requires that the signal is non-negative, the stock selection is done using an off-set standard score, $S^i_j \equiv z^i_j + 5$, where z^i_j is cross sectionally demeaned log book-to-market ($i =$ value) or log returns ($i =$ momentum) for stock j, scaled by the cross sectional standard deviation of the sorting variable.

At the opposite extreme, when $k = n$ the result is a pure overfitting bias. When $1 < k < n$ the result is a combination of sample selection and overfitting bias.

For strategies that combine multiple signals there is an additional issue, related to how signals are combined. In particular, will the investigator constrain herself to putting the same weight on each signal, or will she give herself the freedom to employ different weights for each signal? I consider both cases here, proxying for the latter by signal-weighting the signals, i.e., weighting each signal in proportion to the in-sample performance of the active returns of the strategy based on it alone. These correspond to the weights on the ex-post mean-variance portfolio of the individual strategies' active returns (i.e., the weights on the maximal information ratio portfolio), assuming these active returns are uncorrelated and have identical volatilities.

2.2.1. Critical t-statistics for multi-signal strategies

Because of the selection and overfitting biases that result when researchers consider more signals than they employ, and when they sign individual signals to generate positive in-sample returns, the distribution of t-statistics for a multi-signal strategy does not have a standard normal distribution. Critical values derived from that distribution consequently cannot be used to draw inferences regarding significance of performance for multi-signal strategies. Under the null that none of the signals have power predicting returns, absolute t-statistics that exceed 1.96 should be observed on only 5% of the strategies constructed from the underlying signal. There is no reason to believe, however, that large t-statistics are equally unlikely on multi-signal strategies. Figures 1 and 2 show the effects of these biases, in the special cases when only one bias is present ($k = 1$ and $k = n$), and in the general case, respectively. It does so by giving empirically observed critical 5% t-statistics derived from a large number of best k-of-n strategies (100,000 for each $\{k, n\}$ pair). These strategy returns are calculated using real stock returns, employing randomly generated, uninformative

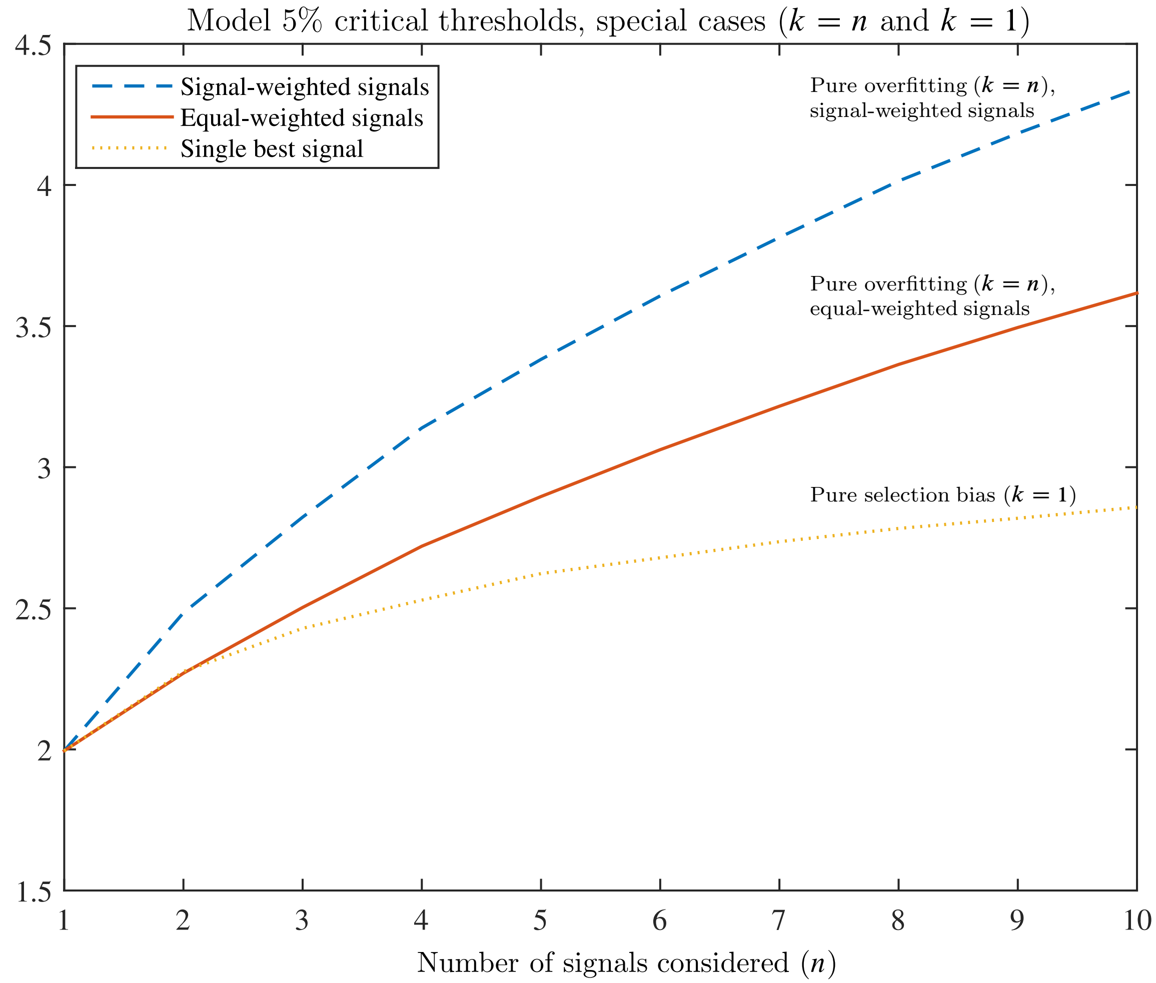


Fig. 1. Five percent critical t-statistics for the active return on best 1-of-n and best n-of-n strategies. The bottom, dotted line shows 5% critical thresholds for the pure selection bias case, when the investigator presents the strongest result from a set of n random strategies, where $n \in \{1, 2, ..., 10\}$. The middle, solid line and the top, dashed line, shows 5% critical thresholds when there is pure overfitting bias. In these cases stocks are selected by combining all n random signals, but the underlying signals are individually signed so that each predicts positive in-sample active returns. In the top, dashed line signals are signal-weighted, while in the middle, solid line signals are equal-weighted. Critical values come from generating 100,000 sets of n randomly generated signals. Returns are signal-weighted, with stocks held in proportion to the signal used for strategy construction, and rebalanced annually at the start of each year. Return data come from CRSP, and the sample covers January 1995 through December 2014.

signals.

Figure 1 shows empirical 5% critical values for the best 1-of-n strategies, which suffer from pure selection bias, and for best n-of-n strategies, which suffer from pure overfitting bias. For the pure overfitting results, it shows both the case when the composite signal is constructed by equal-weighting the individual signals and when it is constructed by signal-weighting the individual signals. The figure shows that overfitting by signal-weighting signals yields much higher critical values than overfitting by equal-weighting signals. That is, the pure overfitting problem is more acute when the investigator has more freedom to overweight good signals. But even when an investigator constrains herself to weight signals equally, and employs every signal she considers, the resulting overfitting bias is still significantly more acute than the more familiar multiple testing bias.

Figure 2 shows empirical 5% critical values for more general best k-of-n strategies, which suffer from both selection and overfitting biases. It shows these critical values when the best one to ten signals are employed, for the cases when 10, 20, 40, or 100 signals are considered, both when signals are equal-weighted (solid lines) and when signals are signal-weighted (dotted lines). The combined biases yield extreme critical values. For the best 2-of-10 strategies the 5% critical t-statistic is almost three and a half; for the best 3-of-20 strategies it exceeds four; for the best 4-of-40 strategies it is five. The figure also shows a non-monotonicity in the critical value for the equal-weighted signal case when $n = 10$, and a large divergence between the critical values of the equal-weighted signal and signal-weighted signal strategies when most the signals considered are employed. This occurs because in-sample performance is impaired by putting significant weight on poor quality signals, an effect that is considered in greater detail in the next section.

Appendix A.1 shows qualitatively similar results for strategies that weight stocks by market capitalization as well as signal.

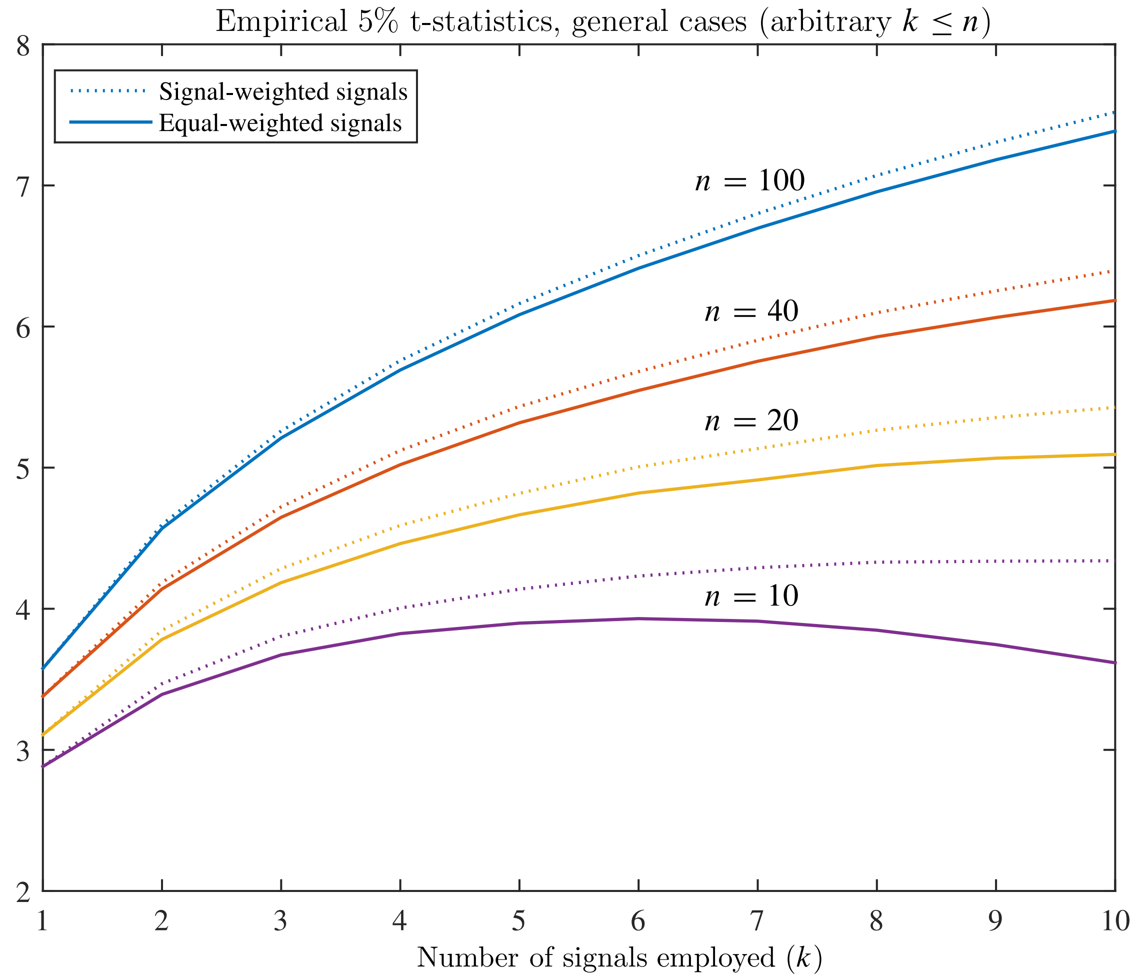


Fig. 2. Five percent critical t-statistics for best k-of-n strategies. The figure shows 5% critical thresholds for strategies selected using a signal constructed by combining the best $k = 1, 2, ..., 10$ performing signals, when the investigator considered $n \in \{10, 20, 40, 100\}$ candidate signals. Solid lines show the cases when the composite signal is constructed by equal-weighting the k best performing candidate signals, and dotted lines the cases when the composite signal is constructed by signal-weighting the signals. Critical values come from generating 100,000 sets of n randomly generated signals. Returns are signal-weighted, with stocks held in proportion to the signal used for strategy construction, and rebalanced annually at the start of each year. Return data come from CRSP, and the sample covers January 1995 through December 2014.

3. Theoretical results

To develop intuition for the results observed in the preceding section, I analyze a simplified model of strategy performance, deriving distributions of backtest t-statistics, and properties of these distributions, when stocks are selected on the basis of multiple signals.

3.1. Model

Suppose there are n underlying signals used for stock selection, and that the active returns to signal-weighted strategies are normally distributed, uncorrelated across signals, and have the same volatilities (σ). Under these assumptions, the performance of strategies based on composite signals, formed as linear combinations of the underlying signals, are relatively easy to analyze. First, the normality of returns implies that standard results from modern portfolio theory hold. Second, weights on individual stocks in a strategy are proportional to the signal used to select them, and composite signals are linear combinations of the underlying signals. By the results of section 2.1.3, therefore, a strategy selected on the basis of a composite signal is identical to a portfolio of strategies based on the individual underlying signals, held in proportion to the weights used to construct the composite signal. This allows us to apply the standard results of modern portfolio theory.

Given n signals, the ex-post Sharpe ratio of the active returns to the strategy that selects stocks based on a composite signal that puts weights $\boldsymbol{\omega} = (\omega_1, \omega_2, ..., \omega_n)$ on the individual signals is, by the results of subsection 2.1.3, essentially the same as that to the portfolio that puts weights $\boldsymbol{\omega}$ on strategies selected from the individual signals,

$$SR_{\boldsymbol{\omega}} \quad = \quad \frac{\boldsymbol{\omega}'\boldsymbol{\mu}^e}{\sqrt{\boldsymbol{\omega}'\boldsymbol{\Sigma}\,\boldsymbol{\omega}}}, \tag{6}$$

where $\boldsymbol{\mu}^e$ is the vector of realized average active returns to strategies selected using the n signals, and $\boldsymbol{\Sigma} = \sigma^2\mathbb{I}$ is the variance-covariance matrix for these active returns. The sample t-statistic estimated on the combined strategy is thus

$$t_{\boldsymbol{\omega}} = \frac{\boldsymbol{\omega}'\boldsymbol{t}}{\sqrt{\boldsymbol{\omega}'\boldsymbol{\omega}}}, \tag{7}$$

where $\boldsymbol{t}$ is the vector of t-statistics estimated on the individual strategies.

The back-tested performance of strategies formed on the basis of multiple signals consequently depends on how the signals are used. The two most common choices, equal-weighting the signals ($\boldsymbol{\omega} = \mathbb{1}/\left\|\mathbb{1}\right\|_1$) and signal-weighting the signals ($\boldsymbol{\omega} = \boldsymbol{t}/\left\|\boldsymbol{t}\right\|_1$), here correspond to well understood strategies, the minimum variance and ex-post mean-variance efficient portfolios, respectively.[4]

Without loss of generality, we may assume that the elements of $\boldsymbol{t}$ are arranged in increasing order, and letting $\mathbb{I}_{n,k}$ be the orthogonal projection onto the lower k dimensional sub-space (i.e., $(\mathbb{I}_{n,k})_{ij} = 1$ if $i = j > n-k$, and zero otherwise), we then have that the sample t-statistics for the minimum variance (i.e., equal-weighted) and mean variance efficient (i.e., signal-weighted) strategies based on the best k-of-n signals, denoted $t_{n,k}^{\text{MV}}$ and $t_{n,k}^{\text{MVE}}$, respectively, are given by

$$t_{n,k}^{\text{MV}} = \frac{\mathbb{1}'\mathbb{I}_{n,k}\boldsymbol{t}}{\sqrt{\mathbb{1}'\mathbb{I}_{n,k}\mathbb{1}}} = \frac{\left\|\mathbb{I}_{n,k}\boldsymbol{t}\right\|_1}{\sqrt{k}} = \frac{\sum_{i=1}^{k} t_{(n+1-i)}}{\sqrt{k}} \tag{8}$$

$$t_{n,k}^{\text{MVE}} = \frac{\boldsymbol{t}'\mathbb{I}_{n,k}\boldsymbol{t}}{\sqrt{\boldsymbol{t}'\mathbb{I}_{n,k}\boldsymbol{t}}} = \left\|\mathbb{I}_{n,k}\boldsymbol{t}\right\|_2 = \sqrt{\sum_{i=1}^{k} t_{(n+1-i)}^2}, \tag{9}$$

where $t_{(j)}$ denotes the j^{th} order statistic of $\{t_1, t_2, ..., t_n\}$. That is, the t-statistic for the best k-of-n strategy that equal weights signals is the L^1-norm of the vector of the

[4] The results presented here for the case when signals are equal-weighted hold more generally. When there is heterogeneity in the volatilities of strategies based on the individual signals, it then corresponding to the "risk-parity" strategy based on the multiple signals.

largest k order statistics of $\boldsymbol{t}$ divided by $\sqrt{k}$. Equivalently, it is $\sqrt{k}$ times the average t-statistic of the strategies corresponding to the signals employed. For the strategy that signal weights the signals, the t-statistic is the L^2-norm of the largest k order statistics. Strategy construction tells us that $t_{n,k}^{\text{MV}} \leq t_{n,k}^{\text{MVE}}$, and standard results for L^p-norms imply that the bound is tight if and only if the k largest order statistics are all equal.

When employing k signals from a set of n candidates, I will denote the critical threshold for $t_{n,k}^{\text{MV}}$ and $t_{n,k}^{\text{MVE}}$ at a p-value of p by $t_{n,k,p}^{*}$ and $t_{n,k,p}^{**}$, respectively.

3.2. Critical t-statistics for special cases

Before analyzing arbitrary $t_{n,k,p}^{*}$ and $t_{n,k,p}^{**}$, it is again useful to develop intuition by first considering the extreme cases. The first of these occurs when the investigator considers several signals, but only reports results for the single best performing strategy ($k = 1$), and again corresponds to pure selection bias. The properties of the critical threshold in this case are well understood, but provide a useful point of comparison. The second occurs when the investigator employs all the signals considered ($k = n$), but signs each to predict positive in-sample returns, and again corresponds to pure overfitting bias. It is free from sample-selection bias, but is biased nevertheless because the joint signal is constructed using information regarding the directionality with which each individual signal predicts in-sample returns.

3.2.1. Pure selection bias: Inference when a single signal is used

For the best 1-of-n strategies, note that the order statistics for standard uniform random variables follow beta distributions, $U_{(k)} \sim \text{Beta}(k, n+1-k)$. So for the maximal order statistic, $P\left(U_{(n)} < x\right) = x^n$, or $P\left(U_{(n)} > (1-p)^{1/n}\right) = p$, implying a critical t-statistic for rejection at the p level for the single best result from n random strategies of

$$t^{*}_{n,1,p} \;=\; t^{**}_{n,1,p} \;=\; N^{-1}\left(\left(1-\tfrac{p}{2}\right)^{1/n}\right), \tag{10}$$

where $N^{-1}(\cdot)$ is the inverse of the cumulative normal distribution.

These critical values can be interpreted by recognizing that $P\left(|\chi| > t^{*}_{n,1,p}\right) = 2N\left(-t^{*}_{n,1,p}\right) = 2\left(1-(1-p/2)^{1/n}\right) \approx p/n$. That is, to close approximation the actual p-value is n times as large as the p-value commonly claimed for an observed t-statistic. If one suspects that the observer considered 10 strategies, significance at the 5% level requires that the results appear significant, using standard tests, at the 0.5% level. This is the standard Bonferroni correction for multiple comparison when the hypothesis is that the expected returns to all n candidate strategies are zero.

3.2.2. Pure overfitting bias: Inference when all signals considered are used

When all the signals considered are used there is no selection bias, but overfitting still occurs because the signals are typically signed so that they predict high returns in-sample. In this case the distribution of the observed t-statistic for the strategy based on the signal-weighted signal (i.e., the ex post MVE combination of the n strategies) is trivial. All the signals are employed, and the t-statistics on the excess returns to the strategies based on the underlying signals come from independent standard normal draws. The t-statistic on the signal-weighted strategy is consequently distributed as a chi-distribution with n degrees of freedom,

$$t^{\mathrm{MVE}}_{n,n} \;=\; \|\boldsymbol{t}\|_2 \;\sim\; \sqrt{\chi^2_n}. \tag{11}$$

The critical t-statistic for the strategy constructed using the signal-weighted combination of n randomly selected signals comes from inverting the chi-squared distribution,

$$t^{**}_{n,n,p} \;=\; \sqrt{\Phi^{-1}_{\chi^2_n}\left(1-p\right)}, \tag{12}$$

where Φ_X denotes the cumulative distribution function for the random variable X.

The critical values for the strategies based on the equally weighted signals are more difficult, but have a simple asymptotic approximation. The mean and variance of the absolute value of the standard normal variable are $\sqrt{2/\pi}$ and $1-2/\pi$, respectively, so $\lim_{n\to\infty}\|\boldsymbol{t}\|_1 \sim N\left(n\sqrt{2/\pi}, n(1-2/\pi)\right)$, and

$$t^{\mathrm{MV}}_{n,n} \;=\; \frac{\|\boldsymbol{t}\|_1}{\sqrt{n}} \;\underset{n\to\infty}{\sim}\; \sqrt{\tfrac{2n}{\pi}}+\left(\sqrt{1-\tfrac{2}{\pi}}\right)\chi. \tag{13}$$

This implies an asymptotic critical value of

$$t^{*}_{n,n,p} \;\approx\; \left(\sqrt{\tfrac{2}{\pi}}\right)\sqrt{n}+\left(\sqrt{1-\tfrac{2}{\pi}}\right)N^{-1}(1-p)\,. \tag{14}$$

The true distribution of $t^{\mathrm{MV}}_{n,n}$ is positively skewed and has excess kurtosis, especially for small n, so the true probability that $t^{\mathrm{MV}}_{n,n}$ exceeds this critical value exceeds p. That is, the estimate is a lower bound on the true critical value. The performance of this estimator is improved, especially for small n, by replacing p with $\frac{np}{n+1}$,

$$t^{*}_{n,n,p} \;\approx\; \left(\sqrt{\tfrac{2}{\pi}}\right)\sqrt{n}+\left(\sqrt{1-\tfrac{2}{\pi}}\right)N^{-1}\left(1-\tfrac{np}{n+1}\right). \tag{15}$$

Figure 3 shows 5% critical values for these special cases. The bottom, dotted line shows $t^{*}_{1,n,5\%}$, the critical value for the best 1-of-n strategy, which suffers from pure selection bias. The top, dashed line, and the solid, middle line, show $t^{*}_{n,n,5\%}$ and $t^{**}_{n,n,5\%}$, the critical values for the best n-of-n strategies when signals are signal-weighted and equal-weighted, respectively, which suffer from pure overfitting bias. The dash-dotted

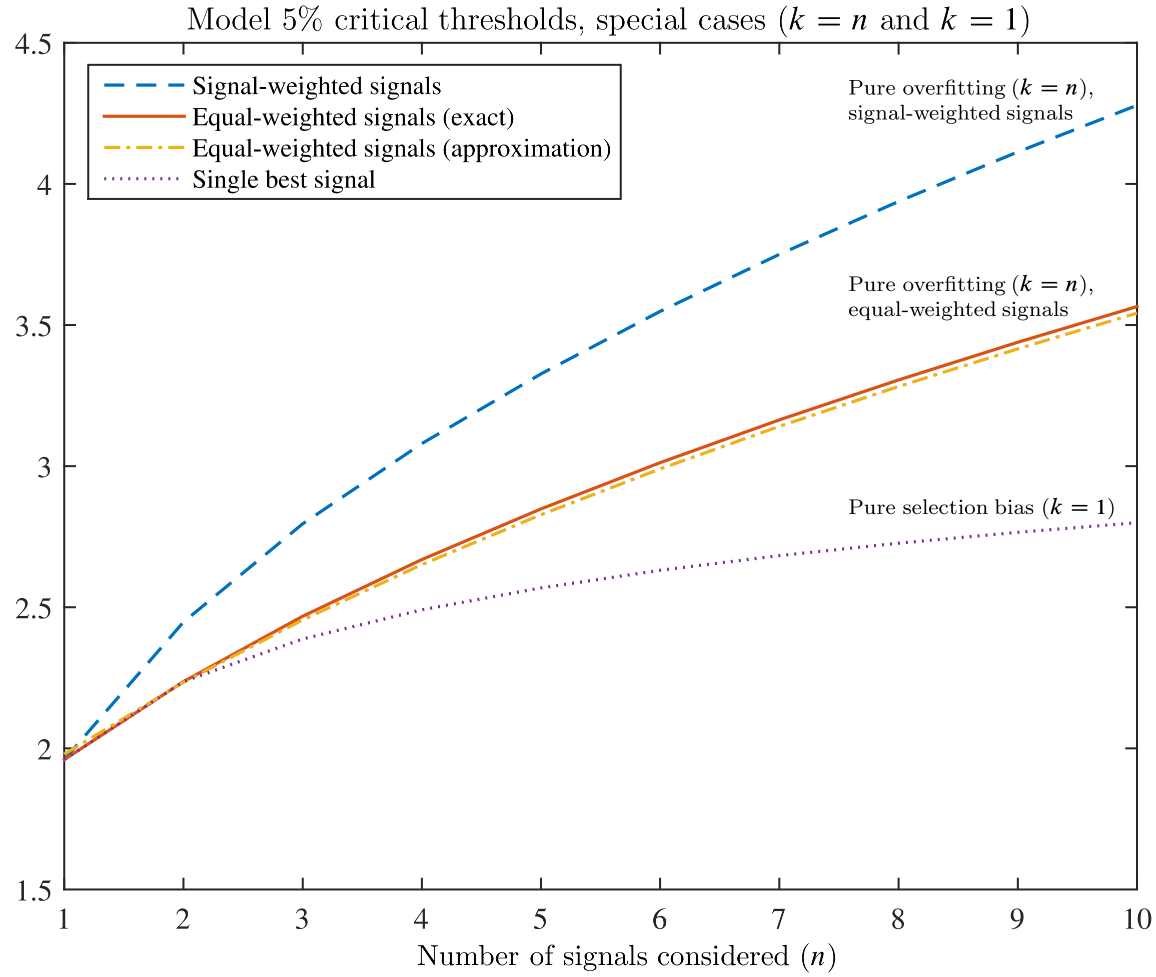


Fig. 3. Model 5% critical t-statistics for best 1-of-n and best n-of-n strategies. The bottom, dotted line shows 5% critical thresholds for the pure selection bias case, when the investigator presents the strongest result from a set of n random strategies, where $n \in \{1, 2, ..., 10\}$. The middle, solid line and the top, dashed line, show 5% critical thresholds when there is pure overfitting bias. In these cases stocks are selected by combining all n random signals, but the underlying signals are signed so that they predict positive in-sample returns. In the top, dashed line signals are signal-weighted, while in the middle, solid line signals are equal-weighted. The dot-dashed line shows the analytic approximation for the equal-weighted best n-of-n case.

line shows the analytic approximation for $t^*_{n,n,5\%}$, provided in equation (15), which closely matches the exact value. The figure shows a remarkable resemblance to the empirical distributions bootstrapped from real stock market data using random signals, provided in Figure 1. A direct comparison is provided in the next subsection.

3.3. General case: best $\boldsymbol{k}$-of-$\boldsymbol{n}$ strategies

To calculate the critical t-statistic more generally, when the selection and overfitting biases interact, note that equations (8) and (9) can be rewritten, using the fact that the top k order statistics of the uniform distribution are uniformly distributed on the interval between the next largest order statistic and one, as

$$\sqrt{k}t^{\mathrm{MV}}_{n,k} = \sum_{i=1}^{k} t_{(n+1-i)} = \sum_{i=1}^{k} \left(t_i \middle| t_i > t_{(n-k)}\right) \tag{16}$$

$$\left(t^{\mathrm{MVE}}_{n,k}\right)^2 = \sum_{i=1}^{k} t^2_{(n+1-i)} = \sum_{i=1}^{k} \left(t^2_i \middle| t_i > t_{(n-k)}\right). \tag{17}$$

The right hand sides of the previous two equations are conditional sums of independent random variables, and have distributions that can be calculated as k-fold convolutions,

$$\phi_{\sqrt{k}t^{\mathrm{MV}}_{n,k}}(x) = \int_{y=0}^{\infty} \phi_{t_{(n-k)}}(y)\phi^{*k}_{\chi|\chi>y}(x)dy \tag{18}$$

$$\phi_{\left(t^{\mathrm{MVE}}_{n,k}\right)^2}(x) = \int_{y=0}^{\infty} \phi_{t_{(n-k)}}(y)\phi^{*k}_{\chi^2|\chi>y}(x)dy, \tag{19}$$

where ϕ_X^{*k} is the k-fold convolution of the probability density function for X, and

$$\phi_{\chi|\chi>y}(x) = \frac{n(x)\mathbb{1}_{x\geq y}}{1-N(y)} \tag{20}$$

$$\phi_{\chi^2|\chi>y}(x) = \frac{\phi_{\chi^2}(x)\mathbb{1}_{x\geq y}}{1-\Phi_{\chi^2}(y)} \tag{21}$$

$$\phi_{t_{(n-k)}}(y) = 2n(y)\phi_{\text{Beta}(n-k,k+1)}\big(2N(y)-1\big), \tag{22}$$

with the last of these following from the fact that the underlying signed t-statistics have absolute standard normal distributions, $t_i \sim N^{-1}\left(\frac{1+U}{2}\right)$ where U is a standard uniform random variable, and $U_{(k)} \sim \text{Beta}(k, n+1-k)$.

The distributions of the t-statistics of interest can then be recovered using

$$\phi_{t_{n,k}^{\text{MV}}}(x) = \sqrt{k}\phi_{\sqrt{k}\phi_{n,k}^{\text{MV}}}\left(\sqrt{k}x\right) \tag{23}$$

$$\phi_{t_{n,k}^{\text{MVE}}}(x) = 2x\phi_{\left(t_{n,k}^{\text{MVE}}\right)^2}\left(x^2\right). \tag{24}$$

While these distribution cannot be expressed as closed-form analytic functions, they are easy to calculate numerically.[5] Figure 4 shows 5% critical values as a function of the number of signals employed ($k \in \{1, 2, ...10\}$), for cases in which the investigator considers ten, 20, 40, or 100 signals ($n \in \{10, 20, 40, 100\}$). The figure again shows a strong resemblance to the corresponding empirical critical values provided in Figure 2. It also shows the same non-monotonicity in the critical value for the equal-weighted signal case when $n = 10$, and a large divergence between the

[5] The required convolutions are quickly calculated using fast Fourier transforms. The fact that $\phi_{\chi^2}(x)$ is unbounded as x approaches zero introduces computational complications, especially when $k = n-1$ and there is a significant probability that some of the underlying strategies exhibit marginal in-sample performance. These complications can be avoided using the fact that the two-fold convolution of the conditional chi-squared distribution has a (bounded) analytic characterization,

$$\phi_{\chi^2|\chi>y}^{*2}(x) = \frac{e^{-x/2}\sin^{-1}\left(1-\frac{2y}{x}\right)\mathbb{1}_{x\geq 2y}}{\pi\left(1-\Phi_{\chi^2}(y)\right)^2}. \tag{25}$$

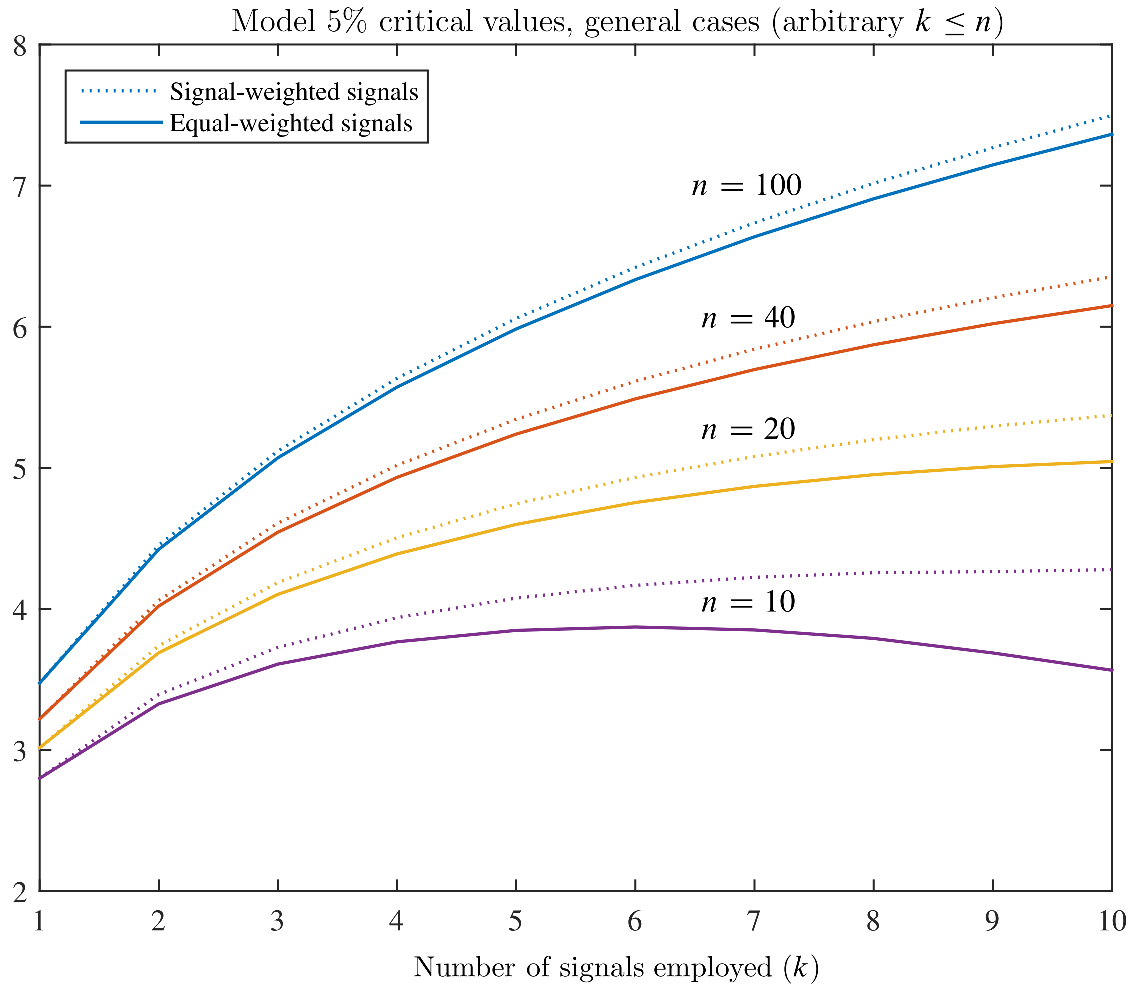


Fig. 4. Model 5% critical t-statistics for best k-of-n strategies. The figure shows 5% critical thresholds for strategies selected using a signal constructed by combining the best $k = 1, 2, ..., 10$ performing signals, when the investigator considered $n \in \{10, 20, 40, 100\}$ candidate signals. Solid lines show the cases when the composite signal is constructed by equal-weighting the k best performing candidate signals, and dotted lines the cases when the composite signal is constructed by signal-weighting the signals.

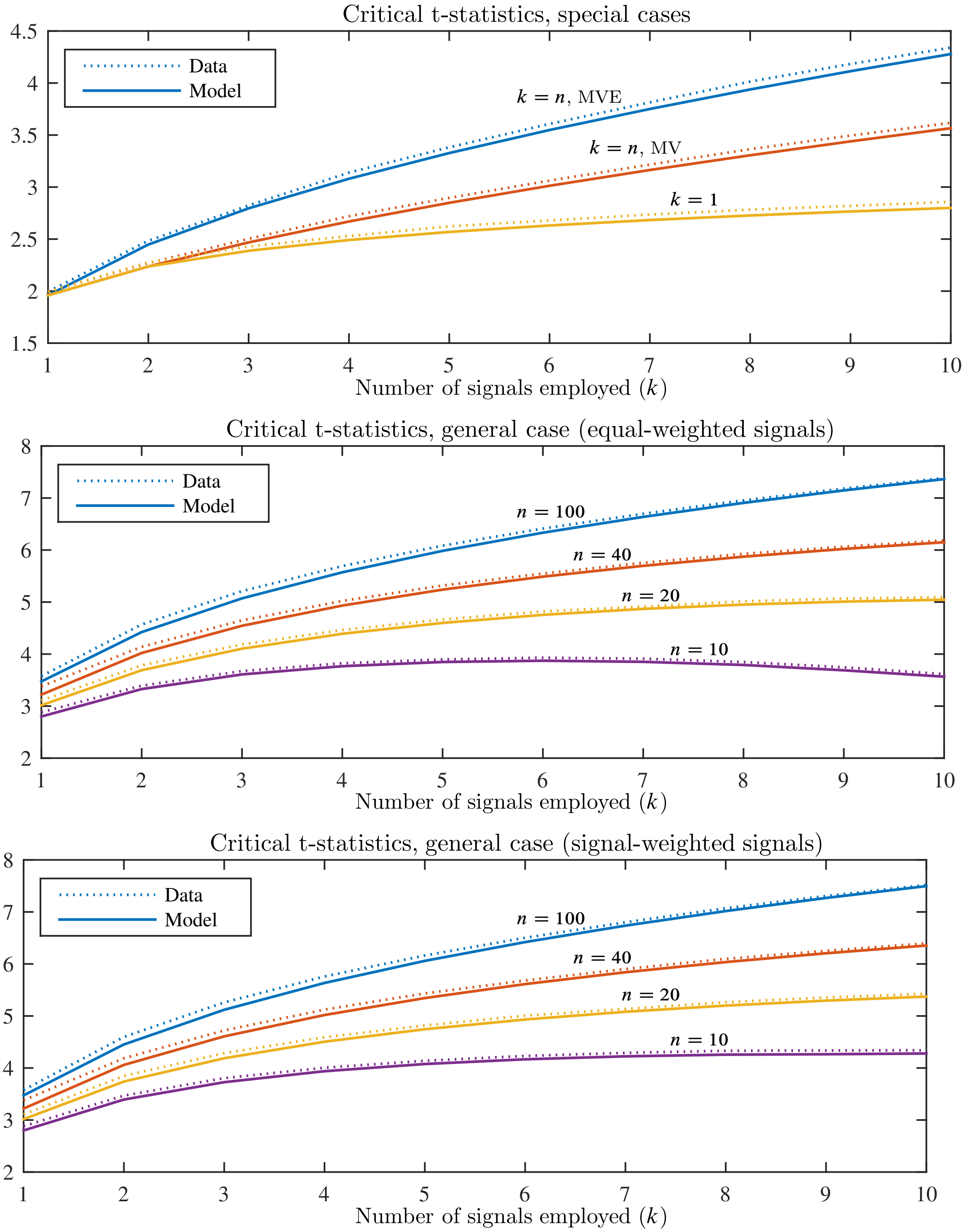


Fig. 5. Comparison of empirical and theoretic 5% critical t-statistics. Panel A shows the 5% critical t-statistics observed in the data, and those that come out of the model, in the special cases when there is pure selection bias ($k = 1$) and pure overfitting bias ($k = n$) for both signal-weighed (MVE) and equal-weighted (MV) signals. Panel B shows these critical values in the general best k-of-n case, when from one to ten signals are selected from a set of 10, 20, 40, or 100 candidate signals, and signals are combined by equal-weighting. Panel C shows the same when signals are signal-weighted.

critical values for strategies that equal- and signal-weight signals when $k \approx n$. These effects are considered in greater detail in the next subsection.

Figure 5 provides a direct comparison of the empirical critical 5% t-statistics shown in Figures 1 and 2, and the theoretic distributions derived for the model shown in Figures 3 and 4. The top panel shows the special cases, for $k = 1$, and for $k = n$ both when signals are signal-weighted (MVE) and when signals are equal-weighted (MV). The middle panel shows the cases for $k \in \{1, 2, ..., 10\}$ and $n \in \{10, 20, 40, 100\}$ when signals are equal-weighted. The bottom panel shows the same, when signals are signal-weighted. In all cases the empirical and model results closely agree, with empirical critical values slightly exceeding those derived from theory, reflecting the excess kurtosis and heteroskedasticity of real world returns that is absent from the model.

4. Critical value approximation

The previous sections have focused on critical t-statistics. This has simplified the analysis, but has largely obscured an important fact: the distribution of best k-of-n t-statistics observed in the data and derived in the model are approximately normal. This approximation is close if n is not too close to one. Figure 6 illustrates this fact, showing the distributions of t-statistics for best 3-of-20 strategies, both from the model and as observed in the data, along with a normal distribution with a mean of 3.15 and standard deviation of 0.56. The empirical density is calculated from the slope of the empirical cumulative distribution function, estimated at each point from the best fit cubic polynomial using all the t-statistics observed within one half of the point of estimation.

The near normality observed in Figure 6, as well as the success of the normal approximation for the critical t-statistic for the special case when $k = n$ and

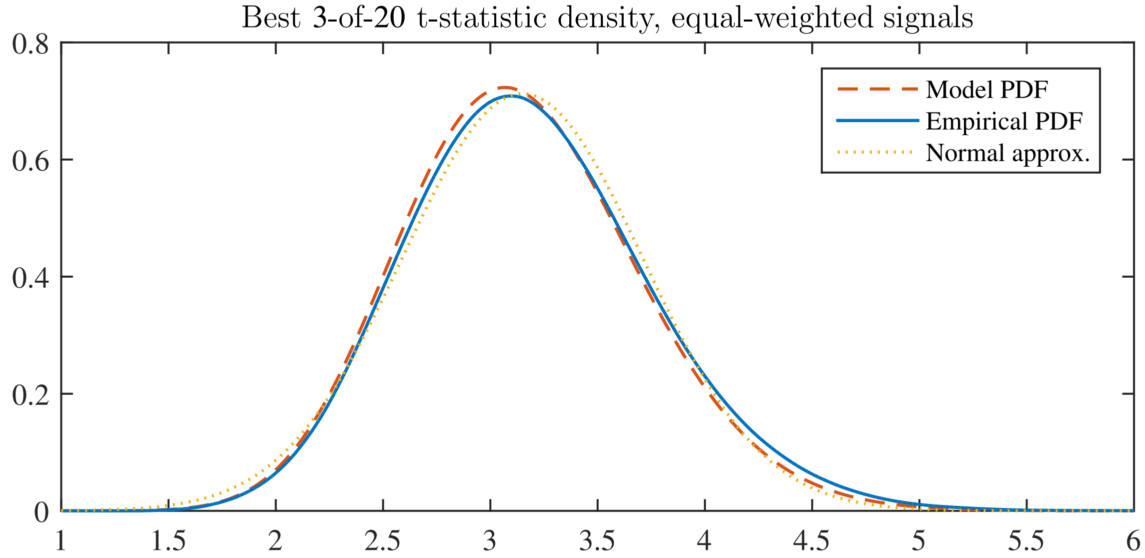


Fig. 6. Densities for best 3-of-20 strategy t-statistics, equal-weighted signals. The figure shows the distributions of t-statistics, both from the model (dashed line) and as observed in the data (solid line), together with a normal distribution with a mean of 3.15 and standard deviation of 0.56 (dotted line). The empirical density is calculated from the slope of the empirical cumulative distribution function (CDF), estimated at each point from the best fit cubic polynomial using all the t-statistics observed within one half of the point of estimation. The empirical CDF comes from the active returns to 100,000 best 3-of-20 strategies, selected on the bases of signals randomly generated at the start of each year, over the sample period January 1995 through December 2014. Stock return data come from CRSP.

signals are equal-weighted, derived in subsection 3.2, suggests a similar critical value approximation for the general cases. Appendix A.4 derives relatively simple analytic approximations, for general best k-of-n strategy critical t-statistics for any p-value, both when signals are equal-weighted and signal-weighted. These are given by

$$t^{*}_{n,k,p} \approx \mu_{t^{\mathrm{MV}}_{n,k}} + \sigma_{t^{\mathrm{MV}}_{n,k}} N^{-1}\left(1 - \tfrac{kp}{k+1}\right) \tag{26}$$

$$t^{**}_{n,k,p} \approx \sqrt{\mu_{(t^{\mathrm{MVE}}_{n,k})^2} + \sigma_{(t^{\mathrm{MVE}}_{n,k})^2} N^{-1}\left(1 - \tfrac{kp}{k+1}\right)}, \tag{27}$$

where

$$
\begin{aligned}
\mu_{t^{\mathrm{MV}}_{n,k}} &= \sqrt{k}\,\lambda_{n,k} && (28)\\
\sigma^2_{t^{\mathrm{MV}}_{n,k}} &= \Sigma_{n,k} - \lambda^2_{n,k} + \tfrac{k(n-k)\left(\lambda_{n,k}-\mu_{n,k}\right)^2}{(k+1)(n+2)} && (29)\\
\mu_{\left(t^{\mathrm{MVE}}_{n,k}\right)^2} &= k\,\Sigma_{n,k} && (30)\\
\sigma^2_{\left(t^{\mathrm{MVE}}_{n,k}\right)^2} &= k\left(\mu^3_{n,k}\lambda_{n,k} + 3\Sigma_{n,k} - \Sigma^2_{n,k}\right) + \tfrac{k^2(n-k)\left(\Sigma_{n,k}-\mu^2_{n,k}\right)^2}{(k+1)(n+2)}, && (31)
\end{aligned}
$$

and

$$
\begin{aligned}
\mu_{n,k} &\equiv N^{-1}\big(E\left[\tfrac{1}{2}\left(1+U_{(n-k)}\right)\right]\big) &&= N^{-1}\Big(1-\tfrac{k+1}{2(n+1)}\Big) && (32)\\
\lambda_{n,k} &\equiv E\left[\chi|\chi > \mu_{n,k}\right] &&= \frac{n(\mu_{n,k})}{1-N(\mu_{n,k})} = 2\left(\tfrac{n+1}{k+1}\right)n(\mu_{n,k}) && (33)\\
\Sigma_{n,k} &\equiv E\left[\chi^2|\chi > \mu_{n,k}\right] &&= 1+\mu_{n,k}\lambda_{n,k}. && (34)
\end{aligned}
$$

The variances terms in equations (26) and (27) are relatively insensitive to n and k, while the means are strongly increasing in both indices, at least for small k.

Figure 7 compares these analytic approximations to the exact critical values, calculated from equations (18) and (19) using numeric integration. Panel A shows the case when the investigator considers 100 candidate signals ($n = 100$), for the full range of the possible number of signals employed ($k = 1, 2, ..., 100$). The top, light solid line shows the exact critical values for the cases when signals are signal-weighted, while the closely tracking dotted line shows the corresponding approximation. The lower, dark lines show the same for the cases when signals are equal-weighted. Panels B and C depict similar results, when the investigator considers only 40 or 20 candidate signals.

An obvious feature of Figure 7 is the peak in $t^*_{n,k,p}$ near the middle, where $k \approx n/2$. The performance of the ex-post mean variance efficient strategy is

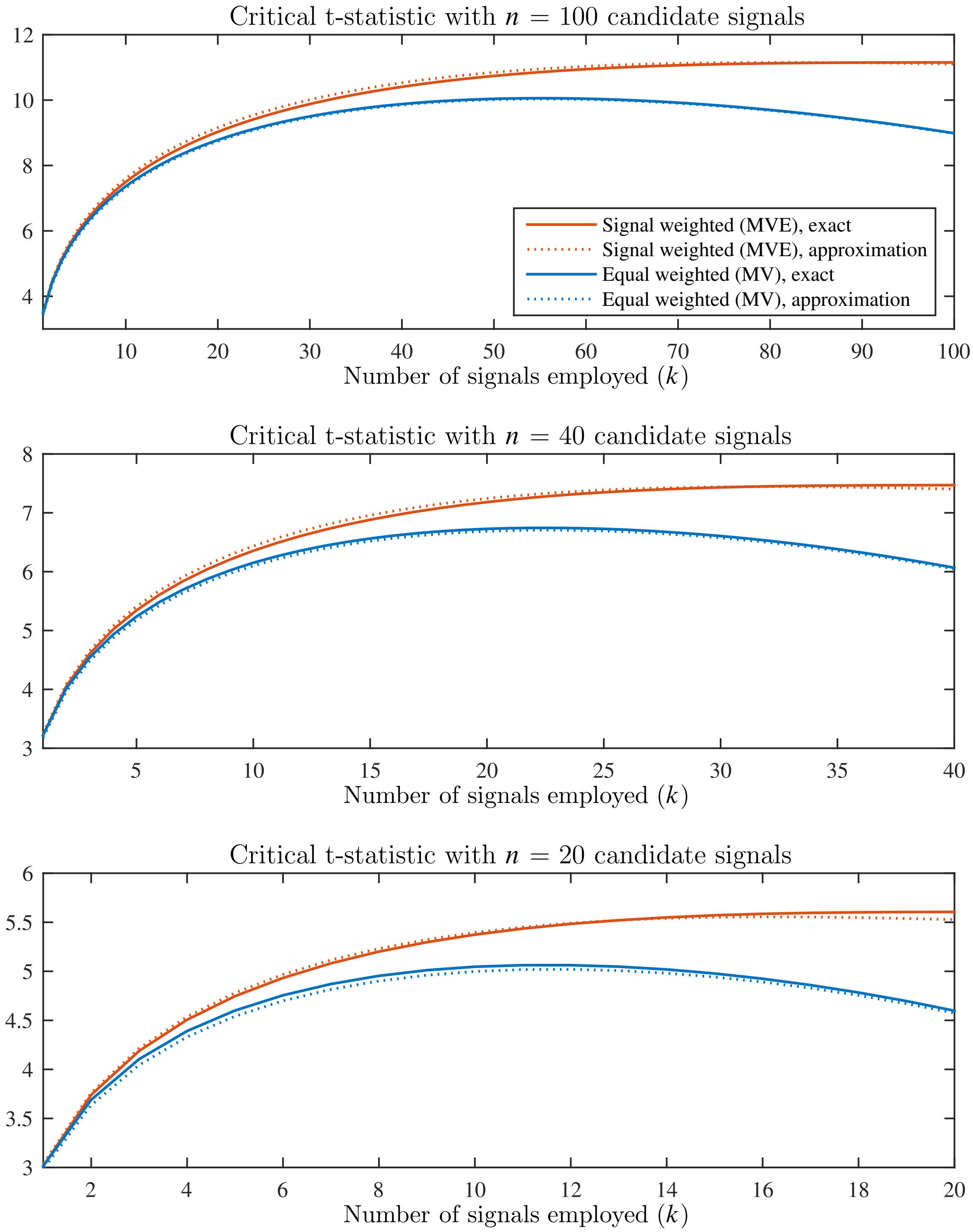


Fig. 7. Comparison of exact and approximate 5% critical t-statistics for best k-of-n strategies. The figure shows 5% critical thresholds for best k-of-n strategies. Solid lines are exact values, while dotted lines are analytic approximations. The top, lighter lines show critical values for strategies that signal-weight signals, while the lower, darker lines correspond to strategies that equal-weight signals.

always weakly improved by adding strategies to the investment opportunity set, so the critical t-statistic threshold for the signal-weighted combination of strategies, $t^{**}_{n,k,p}$, is increasing in $k \leq n$ for all n and p. The same is not true for the critical t-statistic threshold for the equal-weighted combination of strategies, $t^{*}_{n,k,p}$. With these strategies there is a tension. Increasing the number of signals used decreases strategy volatility, which tends to improve performance. At the same time, using more signals reduces average returns, as the average quality of the signals for predicting in-sample performance falls, which tends to hurt performance. Initially the first effect dominates, and performance improves with more signals, but eventually the signal quality deteriorates sufficiently that the gains from additional strategy diversification are more than offset by the loss in average returns. At that point employing additional signals is detrimental to performance. The point at which overall performance starts to deteriorate, i.e., the optimal number of signals to use to predict in-sample performance, can be approximated by noting that this occurs (abusing the infinitesimal notation) when $\frac{d}{dk}E\left[t^{\text{MV}}_{n,k}\right] = 0$. Differentiating the expected t-statistic using the right hand side of equation (8), this implies

$$\frac{E\left[t_{(n-k)}\right]}{E\left[\sum_{i=1}^{k} t_{(n+1-i)}\right]} \quad = \quad \frac{1}{2k}, \tag{35}$$

or, after rearranging and using iterated expectations, that

$$E\left[t_{(n-k)}\right] \quad = \quad E\left[\frac{E\left[\sum_{i=1}^{k} t_{(n+1-i)}\middle|t_{(n-k)}\right]}{2k}\right] \quad = \quad \frac{E\left[\lambda\left(t_{(n-k)}\right)\right]}{2}. \tag{36}$$

That is, there is no longer a benefit to using additional signals when the next signal is only half as good as the average of all the better signals already employed.

Finally, using $E\left[\lambda\left(t_{(n-k)}\right)\right] \approx \lambda\left(E\left[t_{(n-k)}\right]\right)$ and $E\left[t_{(n-k)}\right] \approx \mu_{n,k}$, the previous equation implies that $\mu_{n,k} = \lambda(\mu_{n,k})/2$, or using $\mu_{n,k} = N^{-1}\left(1 - \frac{k+1}{2(n+1)}\right)$ and letting $x^* = 0.612$ be the solution to $2x = n(x)/N(-x)$, that $\frac{k+1}{n+1} \approx 2N(-x^*) \quad = 0.541$, or simply $k \approx n/2$. Consistent with Figure 7, given n candidate signals the maximal Sharpe ratio strategy that equal-weights signals employs roughly half the signals. That is, when forming equal-weighted strategies, the optimal "use" of the worst performing half of the typical set of candidate signals is to simply ignore them. Observing that a multi-signal strategy fails to employ any poor quality signals consequently raises concerns that the investigator threw out poor performing candidates, suggesting selection bias (i.e., $n > k$) as well as overfitting bias (i.e., signing each signal so that it performs well in sample). In this case the expected t-statistic is $E[t^{\text{MV}}_{2k,k}] \approx 4n\big(N^{-1}(3/4)\big)\sqrt{k} = 1.27\sqrt{k}$, almost 60% higher that the t-statistic that would have been expected absent the selection bias, $E[t^{\text{MV}}_{k,k}] \approx \sqrt{2k/\pi} = 0.8\sqrt{k}$.

5. Pure selection bias equivalence

Another way to quantify the impact of the combined sample selection and overfitting biases, and how they interact, is to calculate the number of candidate signals an investigator would need to consider, when selecting stocks using a single signal, to get the same bias expected from a best k-of-n strategy. That is, to calculate the size of the candidate signals set $n^*_{n,k,p}$ (or $n^{**}_{n,k,p}$) implicitly defined by $t^*_{n^*_{n,k,p},1,p} = t^*_{n,k,p}$ (or $t^{**}_{n^{**}_{n,k,p},1,p} = t^{**}_{n,k,p}$). Using $t^*_{n,1,p} = t^{**}_{n,1,p} = N^{-1}\left(\left(1-\frac{p}{2}\right)^{1/n}\right)$, these imply

$$n^*_{n,k,p} = \frac{\ln\left(1-\frac{p}{2}\right)}{\ln\left(N\left(t^*_{n,k,p}\right)\right)}. \tag{37}$$

$$n^{**}_{n,k,p} = \frac{\ln\left(1-\frac{p}{2}\right)}{\ln\left(N\left(t^{**}_{n,k,p}\right)\right)}. \tag{38}$$

Table 1. Single-signal candidates for best k-of-n 5% critical threshold
The table reports the number of candidates a researcher would need to consider, when selecting the single strongest signal, to get the same 5% critical t-statistic for a best k-of-n strategy. Panel A reports the cases when the signals are equal-weighted ($n^{*}_{n,k,p}$), and Panel B the cases when the signals are signal-weighted ($n^{*}_{n,k,p}$).

Signals	Signals used (k)				
considered (n)	2	3	4	5	10
Panel A: Equal-weighted signals (minimum variance strategies)					
10	73	211	386	522	170
20	327	1,780	6,240	16,500	144,000
40	1,460	15,100	98,500	483,000	8.15×10^{7}
100	10,300	268,000	4.03×10^{6}	4.32×10^{7}	3.32×10^{11}
Panel B: Signal-weighted signals (mean-variance efficient strategies)					
10	97	357	816	1,450	3,550
20	423	2,770	11,500	34,900	884,000
40	1,790	21,700	167,000	919,000	3.17×10^{8}
100	11,800	352,000	5.98×10^{6}	7.09×10^{7}	9.16×10^{11}

Table 1 shows single-signal equivalent sample sizes, for actual sample size from ten to 100 (rows), when employing two to ten signals (columns). The table shows the pernicious interaction between the sample selection and overfitting biases. Panel A, which shows the case when multiple signals are equal-weighted, shows that using just the best three signals from 20 candidates yields a bias as bad as if the investigator had used the single best performing signal from 1,780 candidates. With five signals selected from 40 candidates, the bias is almost as bad as if the investigator had used the single best performing signal from half a million candidates. Panel B shows even stronger results when the researcher has the freedom to overweight more strongly performing signals.

A key feature of Table 1 is that the equivalent number of single signals considered ($n^{*}_{n,k,p}$ and $n^{**}_{n,k,p}$) grows quickly with the number of true candidate signals (n), and this growth rate is strongly increasing in the number of signals employed (k). To help understand this, consider the following. For each $j \in \{1, 2, ..., n^{*}\}$, let

$X_1^j, X_2^j, \ldots, X_{2n}^j$ be a sequence of independent standard normal random draws, and $S_k^j = \left(\sum_{i=1}^k X_i^j\right) / \sqrt{k}$ denote the scaled k^{th} partial sum of the j^{th} sequence. Note that S_k^j has a standard normal distribution for any j and k (independent across sequences), so is distributed like the performance of a strategy formed on a single uninformative random signal. It also corresponds, by construction, to the expected t-statistic on a strategy selected on the basis of an equal-weighted combination of the first k of $2n$ uninformative signals (the X_i^j), where the composite signal is constructed *without* signing the individual underlying signal to predict positive in-sample returns. Now how many scaled k^{th} partial sums do you expect to have to look at before observing signed performance ($|S_k^j|$) as good as you expect from a best k-of-n strategy ($t_{n,k}^{\mathrm{MV}}$)?

For any given sequence, the probability that the k^{th} partial sum is equal to the sum of the sequence's k largest (or smallest) order statistics is $P\left(\sqrt{k} S_k^j = \sum_{i=1}^k X_{(2n+1-i)}^j\right) = \binom{2n}{k}^{-1}$. On average an investigator consequently has to observe $\frac{1}{2}\binom{2n}{k}$ sequences before seeing one for which the k^{th} partial sum is comprised of k extreme order statistics. Letting $X_{(k;n)}$ denote the k^{th} order statistic of n independent draws of X, and using the fact that $|\chi|_{(n-i+1;n)} \sim \left(\chi_{(n-i+1;2n)} \middle| \chi_{(n;2n)} = 0\right) \overset{\text{approx.}}{\sim} \chi_{(n-i+1;2n)}$ for large n, because both the mean and variance of $\chi_{(n;2n)}$ are close to zero, we have that $t_{n,k}^{\mathrm{MV}} \overset{\text{approx.}}{\sim} \left(S_k^j \mid \sqrt{k} S_k^j = \sum_{i=1}^k X_{(2n+1-i)}^j\right)$. That is, in one out of every $\frac{1}{2}\binom{2n}{k}$ individual signals an investigator comes across one that looks like the scaled sum of k extreme order statistics from $2n$ standard normal draws, and this object has approximately the same distribution as the performance of a best k-of-n strategy selected using uninformative signals.

In practice this overstates the number of single-signal strategies an investigator would have to observe before seeing performance as strong as expected from a typical best k-of-n strategy. On the way to seeing a sequence for which the k^{th} partial sum is comprised of its top k order statistics, one expects to see k other sequences with k^{th} partial sums that include k of the sequence's top $k+1$ order statistics, and some of these scaled partial sums may be even larger. It does, however, suggest that the number of single signals an investigator can expect to observe before seeing best k-of-n performance is of order $\binom{n}{k}$. Using the fact that $(n/k)^k < \binom{n}{k} < n^k$, this implies an approximately affine relation between $\ln n^*_{n,k,p}$ and $\ln n$, with a slope proportional to k, the number of signals employed.

This approximate log-linear relation is immediately apparent in Figure 8, which plots, on a log-log scale, the number of single-signal candidates required to generate the best k-of-n 5% critical values, ($n^*_{n,k,5\%}$, solid lines, and $n^{**}_{n,k,5\%}$, dotted lines), as a function of the number of candidate strategies actually considered (n, from ten to 100). These are plotted for strategies constructed using the best two, three, and five signals. The figure clearly shows the approximately log-linear relation between the number of single signals required to generate the bias inherent in the best k-of-n strategy ($n^*_{n,k,5\%}$, $n^{**}_{n,k,5\%}$) and the actual number of candidate signals considered (n), with slopes roughly proportional to the number of signals employed (k).[6] That is, $n^*_{n,k,p} \approx o\left(n^k\right)$.

[6] The approximately log-linear relation can be characterized using standard results for binomial coefficients, $\log n^*_{n,k,p} = o\left(\log \binom{n}{k}\right)$ and $\log \binom{n}{k} \approx k \log\left(n/k - 1/2\right) + k - \log\sqrt{2\pi k}$. Alternatively, it can be derived using the critical value approximation considered in Section 4. For p of interest (i.e., small p), $n^*_{n,k,p} \approx \frac{p/2}{N(-t^*_{n,k,p})}$. Using the approximation that $N(-\tau) \approx n(\tau)/\tau$ for large τ and taking logs gives that $\ln n^*_{n,k,p}$ is of order $\left(t^*_{n,k,p}\right)^2$. Using the approximation for $t^*_{n,k,p}$ provided in equation (26), this implies that $\ln n^*_{n,k,p}$ is of order $k\lambda^2_{n,k}$. Finally, using $\lambda_{n,k} \approx \mu_{n,k}$ for $k \ll n$ and $\mu_{n,k} = N^{-1}\left(1 - \frac{k+1}{2(n+1)}\right)$, together with the inverse normal approximation $N^{-1}(x) \approx \sqrt{\frac{-\ln\left(4x(1-x)\right)}{\sqrt{\pi/8}}}$, this implies that $\ln n^*_{n,k,p}$ is approximately affine in $\ln n$, with a slope roughly proportional to k.

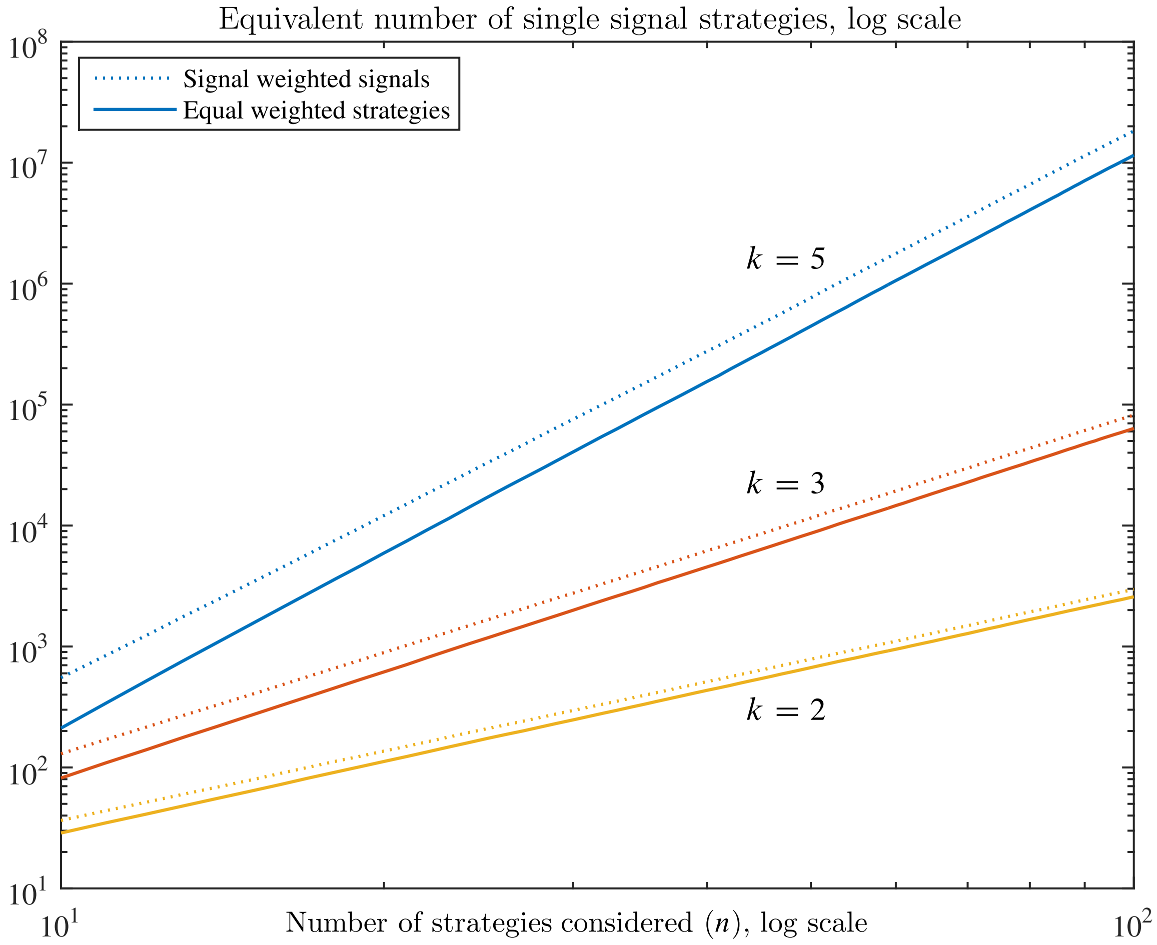


Fig. 8. Number of single candidate strategies to generate the five percent critical thresholds for best k-of-n strategies. The figure plots, using a log-log scale, the number of single signals required to generate the best k-of-n critical values at the 5% level, $t^*_{n,k,5\%}$ (equal-weighted signals, solid lines) and $t^{**}_{n,k,5\%}$ (signal-weighted signals, dotted lines), as a function of the number of candidate strategies actually considered (n, from ten to 100). Strategies are constructed using the best two, three, and five signals (k). The figure shows the expected approximately linear log-log relation, with slopes proportional to the number of signals employed.

6. Conclusion

Multi-signal strategies cannot be evaluated using conventional tests. It is easy, combining spurious, marginal signals, to generate backtested performance that looks impressive, at least when evaluated using the wrong statistics. One solution is to evaluate multi-signal strategies using different statistics, derived in this paper. Another easier potential solution is to evaluate the marginal power of each signal individually, accounting, at a minimum, for the multiple testing bias presented by the number of signals used. That is, to evaluate the alphas of strategies based on each individual signal underlying any multi-signal strategy relative to all of the others, using the Bonferroni correction to account for the fact that the investigator has considered at least as many signals as she actually employs, and may have considered many more.

This is certainly not meant to suggest that an investor should not employ multiple signals, provided that they believe in each signal individually. Signals that work well individually will work even better together. One should not, however, believe in multiple signals because they backtest well together. High back-tested multi-signal Sharpe ratios do not imply that any individual signal has any power predicting returns, or that the multi-signal strategy is likely to perform well going forward.

Finally, these results have implications when evaluating competing asset pricing models. Barillas and Shanken (2015) show that model comparison only requires examining the extent to which each model prices the others' factors. This is a results about in-sample performance evaluation, however, where observed model success simply reflects ex-post mean-variance efficiency of some combination of the model's factors. The results presented here consequently suggests that a model employing factors based on multiple signals, even if perform strongly under the criteria of Barillas and Shanken, may have little expected out of sample power pricing the cross section.

A. Appendix

A.1. Capitalization-weighted strategies

The strategy return is the signal- and capitalization-weighted average stock return (suppressing the time index),

$$r^i = \frac{\sum_{j=1}^N S_j^i m_j r_j}{\sum_{j=1}^N S_j^i m_j}. \tag{39}$$

Define the benchmark return and average signal as the capitalization-weighted average stock return and signal, respectively,

$$r^{bmk} \equiv \sum_{j=1}^N \left(\tfrac{m_j}{M}\right) r_j \tag{40}$$

$$S^i \equiv \sum_{j=1}^N \left(\tfrac{m_j}{M}\right) S_j^i, \tag{41}$$

where $M \equiv \sum_{j=1}^N m_j$ is the aggregate market capitalization multiplier for the universe the strategy trades in. Note that the variance in the average signal is much smaller than the variance in the signal for any individual stock, because it is diversified across all the stocks in the strategy's opportunity set. Using these definitions, we can write the strategy's active return relative to the benchmark as

$$\begin{aligned} r^i - r^{bmk} &= \frac{\sum_{j=1}^N S_j^i m_j \left(r_j - r^{bmk}\right)}{\sum_{j=1}^N S_j^i m_j} \\ &= \frac{\sum_{j=1}^N \left(S_j^i - S^i\right) m_j \left(r_j - r^{bmk}\right)}{\sum_{j=1}^N S_j^i m_j} \\ &= \frac{\sum_{j=1}^N \left(S_j^i - S^i\right) m_j r_j}{\sum_{j=1}^N S_j^i m_j}, \end{aligned} \tag{42}$$

where the second and third equalities follow from the definitions of r^{bmk} and S^i, respectively.

This active return may be viewed as a long/short strategy that takes positions in stocks in proportion to how far the signal is from the capitalization weighted mean. In particular, noting that the returns to a long/short strategy based on the signal that invests one dollar on each the long and short sides is

$$r^{i,L/S} = \frac{\sum_{j=1}^{N} \left(S_j^i - S^i\right) m_j r_j}{\frac{1}{2}\sum_{j=1}^{N} \left|S_j^i - S^i\right| m_j}, \tag{43}$$

the active returns to the long-only strategy can be written as

$$r^i - r^{bmk} = \left(\frac{\frac{1}{2}\sum_{j=1}^{N} \left|S_j^i - S^i\right| m_j}{\sum_{j=1}^{N} S_j^i m_j}\right) r^{i,L/S}. \tag{44}$$

That is, one dollar in the long-only strategy can be thought of as one dollar in the benchmark strategy, plus a tilt toward the one dollar long/one dollar short strategy based on the signal. The size of this tilt, i.e., the leverage on the dollar long/dollar short strategy, is determined by the signal-to-noise ratio of the stock selection signal. If the signal is normally distributed and uncorrelated with the capitalization multiplier, then the expected leverage multiplier is again $(2\pi)^{-1/2}\,\sigma_S/\mu_S \approx 0.4\,\sigma_S/\mu_S$. Provided there are a large number of stocks in the investment opportunity set, there is very little variation in realized values around this expectation.

A.1.1. Siloed vs. integrated strategies: a false dichotomy

Given any n, signals let $S^{\boldsymbol{\omega}} = \sum_{i=1}^{n} \omega^i S_j^i$ be a composite signal constructed as any linear combination of the n fundamental signals. Then

$$\begin{aligned} r^c &= \frac{\sum_{j=1}^{N} \left(\sum_{i=1}^{n} \omega^i S_j^i\right) m_j r_j}{\sum_{j=1}^{N} \left(\sum_{i=1}^{n} \omega^i S_j^i\right) m_j} \\ &= \frac{\sum_{i=1}^{n} \omega^i \left(\sum_{j=1}^{N} S_j^i m_j r_j\right)}{\sum_{i=1}^{n} \omega^i \left(\sum_{j=1}^{N} S_j^i m_j\right)} \\ &= \sum_{i=1}^{n} w^i r^i \end{aligned} \tag{45}$$

where $w^i \equiv \omega^i S^i / \sum_{j=1}^{n} \omega^j S^j$. That is, the returns to the strategy based on the composite signal is just a weighted average of the returns to the strategies based on the individual signals, where the weights are proportional to both an individual signal's capitalization-weighted average value and its weight in the composite signal. This establishes an exact correspondence between the performance of integrated strategies (i.e., those based on composite signals) and siloed strategies (i.e., those that allocate resources across single signal strategies).

A.2. Comparison of z-score weighing and quantile sorting

The long and short sides of the z-score weighted and tertile sorted strategies have 77% of their positions in common, a fact which is clearly evident in Figure A.1, which shows the relative portfolio weights on the long sides of long/short strategies based on quantile sorting, z-score weighting, and rank weighting. The non-overlapping positions of the z-score weighted and tertile sorted strategies do not significantly differ in their exposure to the underlying sorting variable. For example, a dollar in the long side of a z-score weighted value strategy is equivalent to a dollar in the equal-weighted portfolio that holds the top 35% of stocks by book-to-market, plus a 23 cent long/short position that buys both extreme and marginal value stocks (those in the top 13% of the book-to-market distribution and those in the 50-65% range, respectively) while shorting moderate value stocks (those in the 65-87% range of the book-to-market distribution). The rank-weighting scheme employed by Frazzini and Pedersen (2014) is even more similar, sharing 92% and 83% of its holding in common with the z-score weighted strategy and the tertile sorted strategy, respectively. Because the stock holdings of the three different constructions are so similar, the performance of the three strategies is almost indistinguishable, with any one of the three strategies returns explaining 98% of the return variation in any of the others.

The close correspondence between the three strategies is also reflected in Figure A.2 and Table A.1, which compare the performance of value and momentum strategies constructed using the three different weighting schemes. For each stock selection signal (book-to-market, constructed using Compustat data, or stock performance over the first eleven months of the preceding year), strategies are constructed using a standard quantile sort ($\boldsymbol{r}^{\text{tertile}}$), or as the active returns to signal-weighted strategies, where the signals employed are the z-score ($\boldsymbol{r}^{\text{z-score}}$) or percentile ranking ($\boldsymbol{r}^{\text{rank}}$) of the sorting variable.[7] Active returns are measured relative to the equal-weighted average of the applicable universe (i.e., all stocks with book-to-market or past performance data, respectively), and levered to run at the average in-sample volatility observed on the tertile sorted strategy (9.5% for value and 14.7% for then momentum strategies, respectively).

Figure A.2 shows the growth of a dollar invested in each of the three value strategies (top panel), and in each of the three momentum strategies (bottom panel), in the middle of 1963. In both cases the figure shows extremely similar performance.

[7] For the z-scores I use an inverse normal transformation, which ensures signal normality and does not necessitate dropping negative book-to-market firms. Using standard scores (i.e., cross-sectionally demeaning and scaling by the cross-sectional standard deviation) for log book-to-market and past log returns yields similar results.

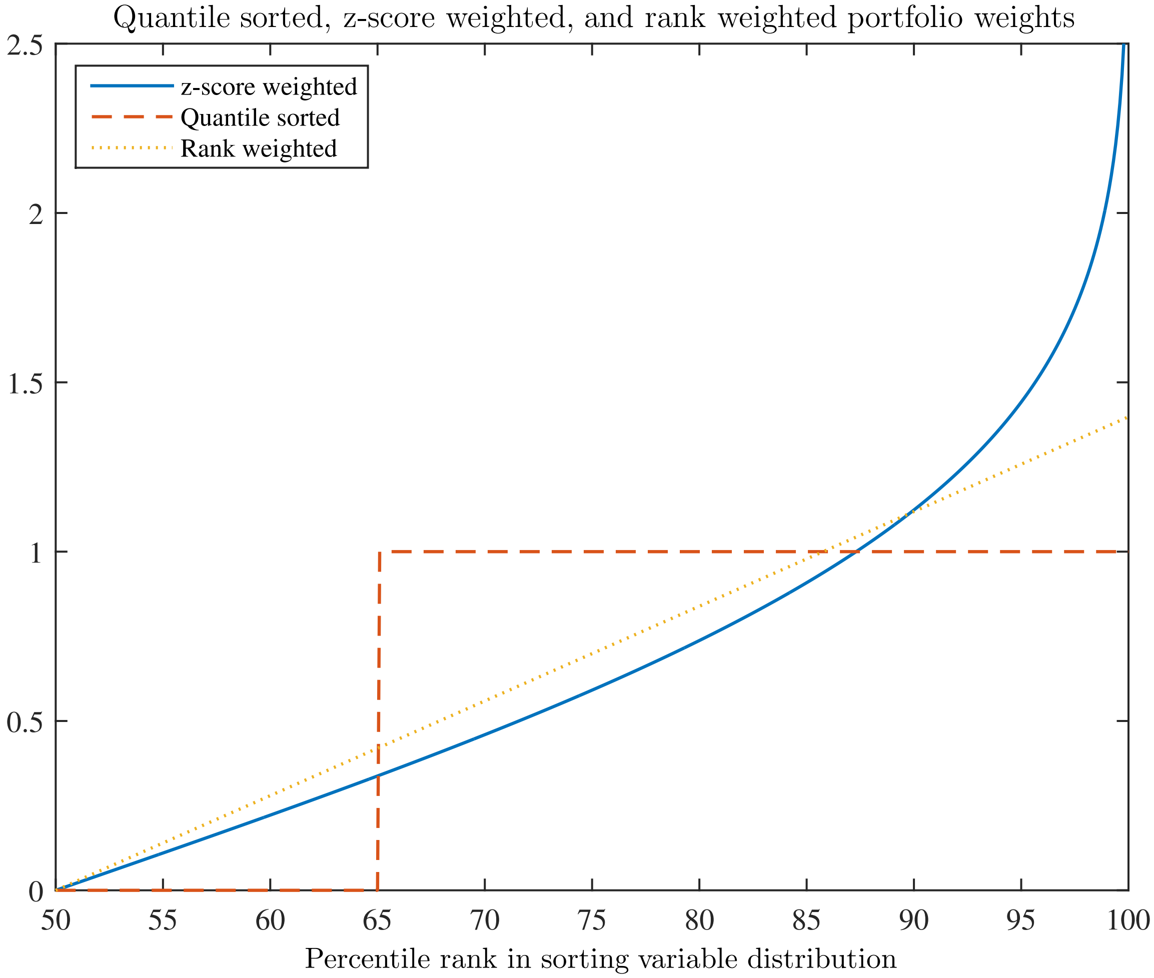
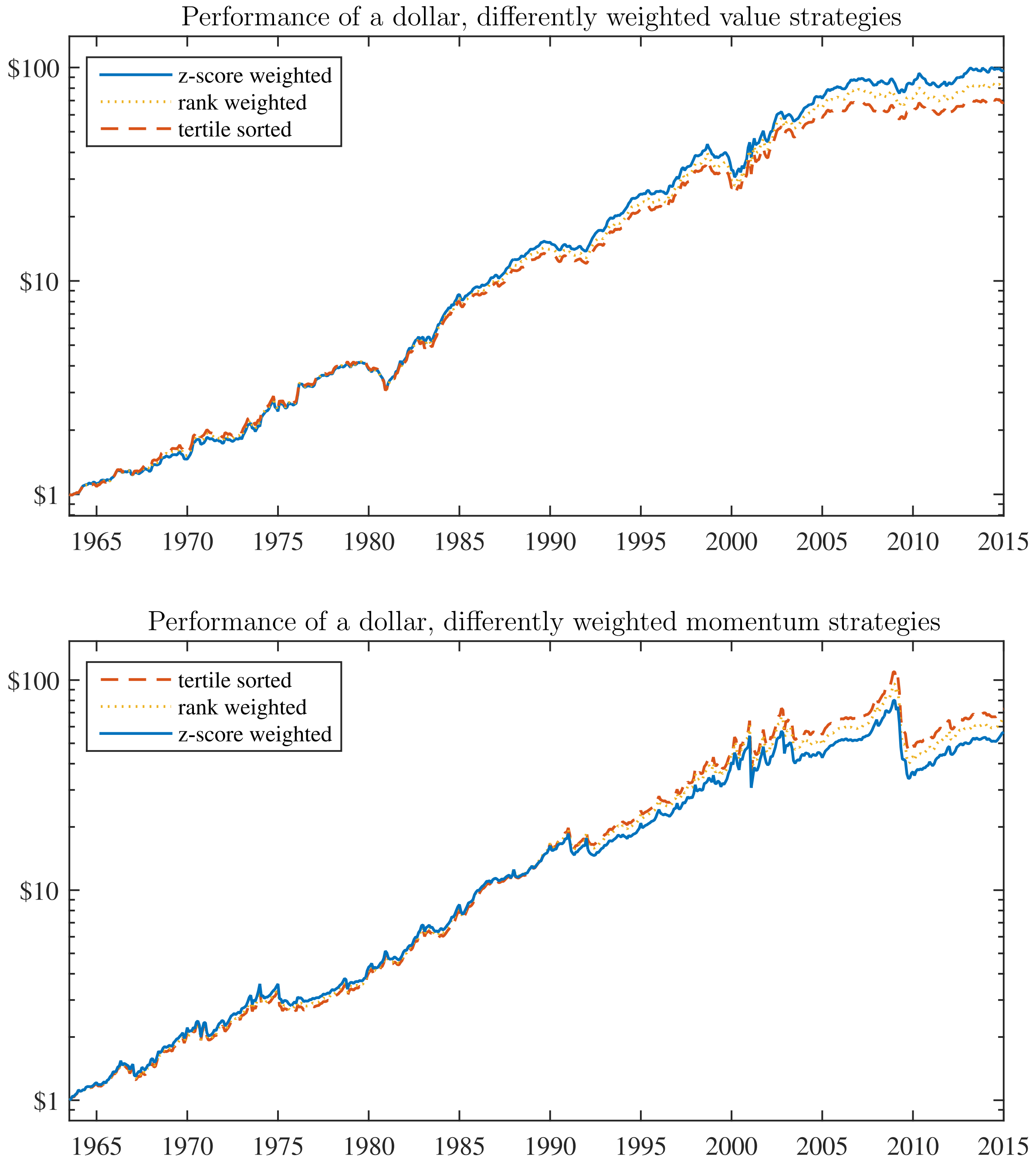


Fig. A.1. Relative portfolio weights, long sides of long/short strategies based on quantile sorting, z-score weighting, and rank weighting. The figure plots, as a function of a stock's position in the cross sectional ranking of the sorting variable, the weight on stocks in the long side of long/short strategies formed on the basis of the sorting variable. Weights are shown relative to the weight on individual stocks held in the equal-weighted quantile sorted strategy, which holds (shorts) the top (bottom) 35% of the distribution. Weights are shown for three sorting methodologies, the z-score weighted construction (dark blue solid line), the quantile sorted strategy (red dashed line), and the rank weighted construction (light yellow dotted line). Short sidde weights (not shown) are symmetric around the median (50% level) of the sorting variable distribution.



Performance of a dollar, differently weighted value strategies

Performance of a dollar, differently weighted momentum strategies

Fig. A.2. Growth of a dollar, differently constructed value and momentum strategies. The figure shows the performance of long/short strategies selected on the basis of book-to-market (top panel, rebalanced annually at the end of June) and performance over the first eleven months of the preceding month (bottom panel, rebalanced monthly). Strategies are tercile sorted (top and bottom 35%, equal weighted), or weighted in proportion to the percentile rank or z-score of the stock selection variable. Signal-weighted strategies are levered to run at the average in-sample volatility observed on the tertile sorted strategy (9.5% for value and 14.7% for then momentum strategies, respectively). Data come from CRSP and Compustat. The sample covers July 1963 through December 2014.

Table A.1. Empirical comparison of different stock-weighting schemes
The table reports results from time-series regressions of the return of value or momentum strategies (panels A and B, respectively) onto other strategies constructed using the same sorting variable but different weighting scheme. Strategies are long/short. The weighting schemes include a standard quantile sort (r^{tertile}, using top and bottom 35%), and strategies that weight stocks on the basis of how far an individual stock's signal rank or signal z-score is from the cross-sectional average (r^{rank} and $r^{\text{z-score}}$, respectively). Data come from CRSP and Compustat, and the sample runs from July 1963 through December 2014).

Independent variable	Dependent variable					
	$r^{\text{z-score}}$		r^{tertile}		r^{rank}	
Panel A: Value strategies						
α	0.06	0.03	-0.05	-0.03	0.03	-0.03
	[3.84]	[3.34]	[-2.84]	[-2.78]	[3.30]	[-2.78]
$r^{\text{z-score}}$			0.99		1.00	
			[166.6]		[308.9]	
r^{tertile}	0.99					1.00
	[166.6]					[302.2]
r^{rank}		1.00		1.00		
		[302.2]		[308.9]		
$\hat{R}^2$ (%)	97.8	99.3	97.8	99.4	99.4	99.3
Panel B: Momentum strategies						
α	-0.03	-0.02	0.05	0.02	-0.02	0.02
	[-1.65]	[-1.86]	[2.20]	[1.77]	[-1.48]	[2.16]
$r^{\text{z-score}}$			0.99		1.00	
			[201.7]		[389.0]	
r^{tertile}	0.99					1.00
	[201.7]					[377.1]
r^{rank}		1.00		1.00		
		[377.1]		[389.0]		
$\hat{R}^2$ (%)	98.5	99.6	98.5	99.6	99.6	99.6

The z-score weighted strategy slightly out performs the standard quantile sorted strategy for value, but slightly underperforms for momentum. In both cases the performance of the rank-weighted strategies lies in between.

Table A.1 quantifies these results. It shows results of time-series regressions of the returns to each of the strategies onto those of each of the other strategies constructed using the same sorting variable. Between 97.8 and 99.6% of the return variation of each strategy is explained by any of the others, and no strategy has abnormal returns relative to any of the others that exceed six basis points per month.

Appendix B show a similar correspondence between value weighted quantile sorted strategies, common in the academic literature, and the active returns to signal and capitalization weighted strategies, which weight individual stocks in proportion to both to their market capitalizations and to the z-score or rank of the sorting variable used in strategy construction.

A.3. Empirical 5% critical values, cap weighted strategies

Figures A.3 and A.4 provide results (strategy performance and best k-of-n strategy empirical 5% critical t-statistics), similar to those provided in Figure A.2 to 2 and Table A.1, for strategies that are capitalization weighted as well as signal-weighted.

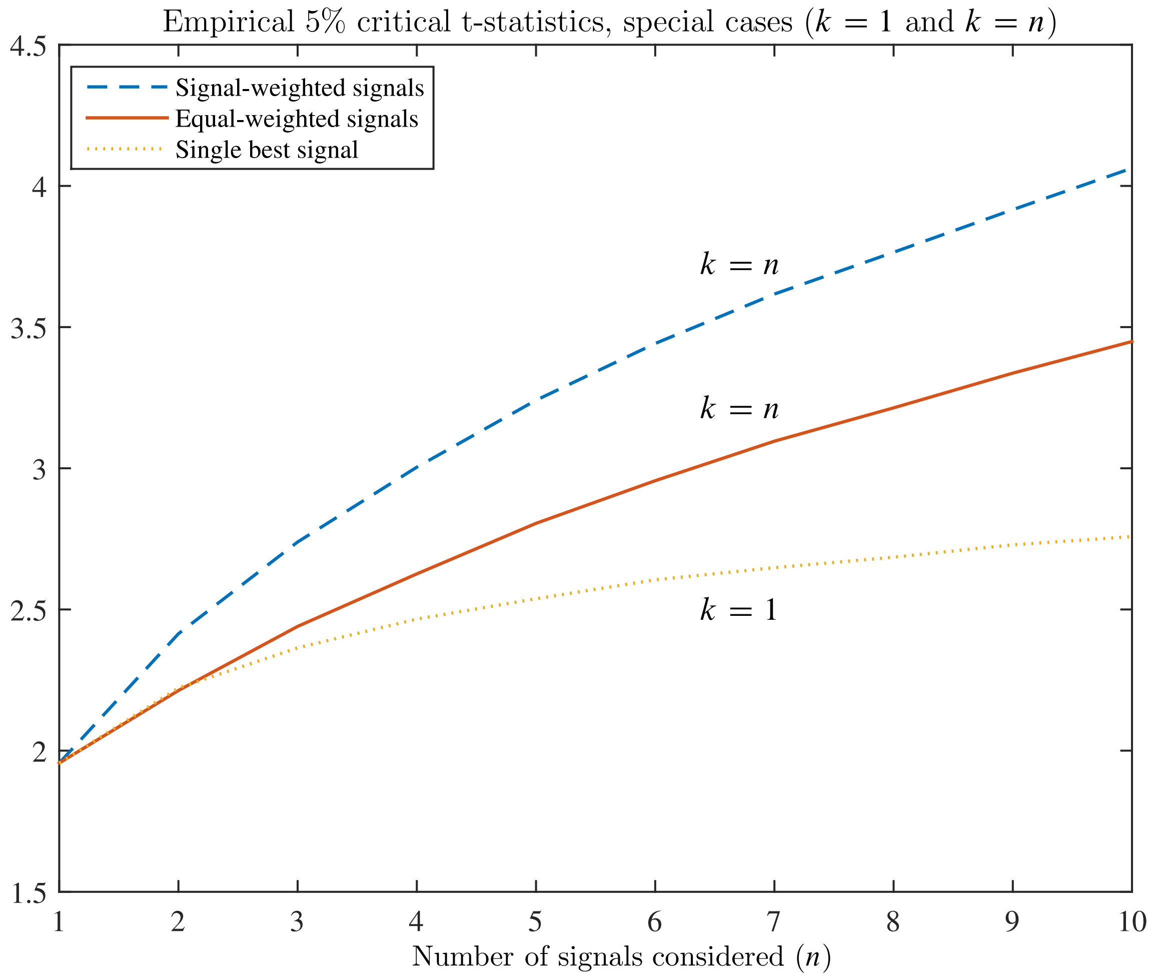


Fig. A.3. Five percent critical t-statistics for the active return on best 1-of-n and best n-of-n strategies, cap-weighted strategies. The bottom, dotted line shows 5% critical thresholds for the pure selection bias case, when the investigator presents the strongest result from a set of n random strategies, where $n \in \{1, 2, ..., 10\}$. The middle, solid line and the top, dashed line, show 5% critical thresholds when there is pure overfitting bias. In these cases stocks are selected by combining all n random signals, but the underlying signals are signed so that they predict positive in-sample active returns. In the top, dashed line signals are signal-weighted, while in the middle, solid line signals are equal-weighted. Critical values come from generating 100,000 sets of n randomly generated signals. In each strategy stocks' returns are signal- and capitalization-weighted, with stocks held in proportion to both the signal used for strategy construction and firms' market capitalizations, and rebalanced annually at the start of each year. Return data come from CRSP, and the sample covers January 1995 through December 2014.

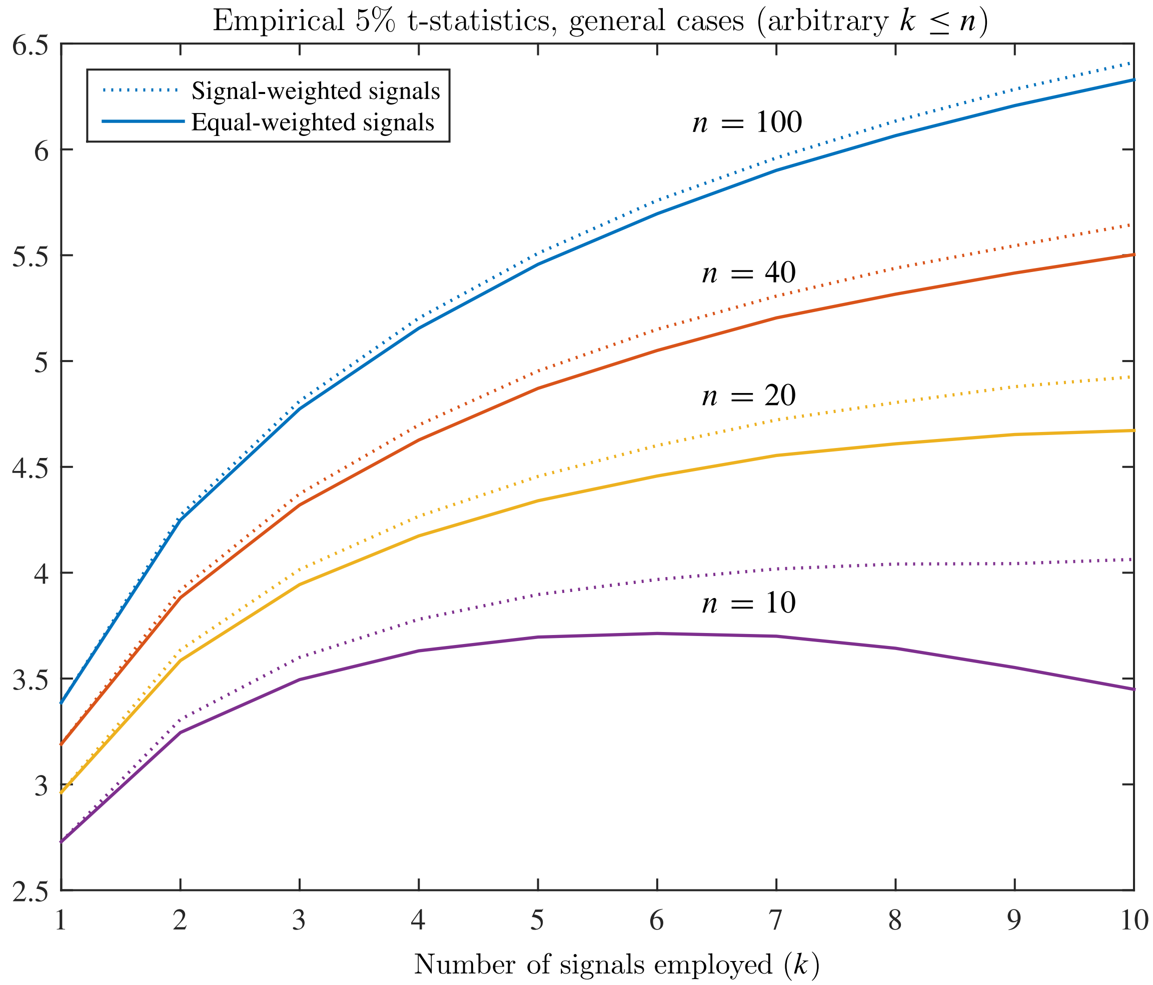


Fig. A.4. Five percent critical t-statistics for best k-of-n strategies, cap-weighted. The figure shows 5% critical thresholds for strategies selected using a signal constructed by combining the best $k = 1, 2, ..., 10$ performing signals, when the investigator considered $n \in \{10, 20, 40, 100\}$ candidate signals. Solid lines show the cases when the composite signal is constructed by equal-weighting the k best performing candidate signals, and dotted lines the cases when the composite signal is constructed by signal-weighting the signals. Critical values come from generating 100,000 sets of n randomly generated signals. In each strategy stocks' returns are signal- and capitalization-weighted, with stocks held in proportion to both the signal used for strategy construction and firms' market capitalizations, and rebalanced annually at the start of each year. Return data come from CRSP, and the sample covers January 1995 through December 2014.

A.4. Critical value approximation

Recall, from equations (8) and (9), that

$$
\begin{aligned}
t_{n,k}^{\mathrm{MV}} &= \frac{\sum_{i=1}^{k} t_{(n+1-i)}}{\sqrt{k}} && (46)\\
\left(t_{n,k}^{\mathrm{MVE}}\right)^2 &= \sum_{i=1}^{k} t_{(n+1-i)}^2 . && (47)
\end{aligned}
$$

Then the top k order statistics of the standard uniform random variable, $U_{(n)}, U_{(n-1)}, ..., U_{(n+1-k)}$, are distributed uniformly on the interval $[U_{(n-k)}, 1]$, so

$$
\sum_{i=1}^{k} t_{(n+1-i)} = k\lambda\big(t_{(n-k)}\big) + \sum_{i=1}^{k} \left(\left[t_i \middle| t_i > t_{(n-k)}\right] - \lambda\big(t_{(n-k)}\big)\right), \qquad (48)
$$

where $\lambda(x) \equiv E[\chi|\chi > x] = n(x)/N(-x)$ denotes the inverse Mill's ratio. The first term on the right hand side of the previous equation inherits the approximate normality of $U_{(n-k)}$, with mean and variance, using $t_{(n-k)} \sim N^{-1}\big(\frac{1}{2}(1 + U_{(n-k)})\big)$ and letting $\mu_{n,k} \equiv N^{-1}\big(E\left[\frac{1}{2}\left(1 + U_{(n-k)}\right)\right]\big) = N^{-1}\Big(1 - \frac{k+1}{2(n+1)}\Big)$ and $\lambda_{n,k} \equiv \lambda(\mu_{n,k})$, given by

$$
\begin{aligned}
E\left[k\lambda\big(t_{(n-k)}\big)\right] &\approx k\lambda_{n,k} && (49)\\
\mathrm{Var}\left(k\lambda\big(t_{(n-k)}\big)\right) &\approx k^2\,\mathrm{Var}\Big(\lambda'(x)\left.\right|_{\mu_{n,k}} \times \left(N^{-1}\right)'(x)\left.\right|_{E\left[\frac{1}{2}\left(1+U_{(n-k)}\right)\right]} \times \frac{U_{(n-k)}}{2}\Big)\\
&= \frac{k^2(n-k)\left(\lambda_{n,k} - \mu_{n,k}\right)^2}{(k+1)(n+2)}, && (50)
\end{aligned}
$$

where the last equality follows from $\lambda'(x) = \lambda(x)^2 - x\lambda(x)$ and $\left(N^{-1}\right)'(x) = 1/n(x)$, and because the order statistics of the standard uniform random variable have beta distributions, $U_{(n-k)} \sim \mathrm{Beta}(n-k, k+1)$, so $E\left[\frac{1}{2}\left(1 + U_{(n-k)}\right)\right] = 1 - \frac{k+1}{2(n+1)}$ and $\mathrm{Var}\left(U_{(n-k)}\right) = \frac{(n-k)(k+1)}{(n+1)^2(n+2)}$.

The second term on the right hand side of equation (48) is mean zero, converges to normality for large k by the central limit theorem, and has a variance of approximately

$$k\mathrm{Var}\left(t_i \middle| t_i > \mu_{n,k}\right) = k\left(1 + \mu_{n,k}\lambda_{n,k} - \lambda_{n,k}^2\right). \tag{51}$$

Taken together these imply $t_{n,k}^{\mathrm{MV}} \overset{\text{approx.}}{\sim} N\left(\mu_{t_{n,k}^{\mathrm{MV}}}, \sigma_{t_{n,k}^{\mathrm{MV}}}^2\right)$ where

$$\mu_{t_{n,k}^{\mathrm{MV}}} = \sqrt{k}\,\lambda_{n,k} \tag{52}$$

$$\sigma_{t_{n,k}^{\mathrm{MV}}}^2 = 1 + \mu_{n,k}\lambda_{n,k} - \lambda_{n,k}^2 + \tfrac{k(n-k)\left(\lambda_{n,k} - \mu_{n,k}\right)^2}{(k+1)(n+2)}. \tag{53}$$

Then

$$\begin{aligned} p &= P\left(t_{n,k}^{\mathrm{MV}} > t_{n,k,p}^*\right) \\ &\approx P\left(\mu_{t_{n,k}^{\mathrm{MV}}} + \sigma_{t_{n,k}^{\mathrm{MV}}}\chi > t_{n,k,p}^*\right) \\ &= 1 - N\left(\frac{t_{n,k,p}^* - \mu_{t_{n,k}^{\mathrm{MV}}}}{\sigma_{t_{n,k}^{\mathrm{MV}}}}\right), \end{aligned} \tag{54}$$

which implies

$$t_{n,k,p}^* \approx \mu_{t_{n,k}^{\mathrm{MV}}} + \sigma_{t_{n,k}^{\mathrm{MV}}} N^{-1}(1-p). \tag{55}$$

Similarly, $\left(t_{n,k}^{\mathrm{MVE}}\right)^2$ is approximately normally distributed. Its mean is

$$\begin{aligned} E\left[\left(t_{n,k}^{\mathrm{MVE}}\right)^2\right] &= E\left[E\left[\left(t_{n,k}^{\mathrm{MVE}}\right)^2 \middle| t_{(n-k)}\right]\right] \\ &\approx E\left[\left(t_{n,k}^{\mathrm{MVE}}\right)^2 \middle| t_{(n-k)} = \mu_{n,k}\right] \\ &= kE\left[t_i^2 \middle| t_i > \mu_{n,k}\right] \\ &= k\left(1 + \mu_{n,k}\lambda_{n,k}\right), \end{aligned} \tag{56}$$

where we have again used the fact that the top k order statistics of a uniform random variable are distributed jointly uniformly over the interval exceeding the next highest order statistic. Its variance is

$$\mathrm{Var}\left(\left(t_{n,k}^{\mathrm{MVE}}\right)^2\right) = \mathrm{Var}\left(E\left[\left(t_{n,k}^{\mathrm{MVE}}\right)^2 \middle| t_{(n-k)}\right]\right) + E\left[\mathrm{Var}\left(\left(t_{n,k}^{\mathrm{MVE}}\right)^2 \middle| t_{(n-k)}\right)\right]. \tag{57}$$

For the second term on the right hand side of the previous equation,

$$\begin{aligned} E\left[\mathrm{Var}\left(\left(t_{n,k}^{\mathrm{MVE}}\right)^2 \middle| t_{(n-k)}\right)\right] &\approx \mathrm{Var}\left(\left(t_{n,k}^{\mathrm{MVE}}\right)^2 \middle| t_{(n-k)} = \mu_{n,k}\right) \\ &= k\mathrm{Var}\left(t_i^2 \middle| t_i > \mu_{n,k}\right) \qquad (58) \\ &= k\left(\mu_{n,k}^3 \lambda_{n,k} + 3\left(1 + \mu_{n,k}\lambda_{n,k}\right) - \left(1 + \mu_{n,k}\lambda_{n,k}\right)^2\right), \end{aligned}$$

where the last line follows from the known conditional moments of the normal distribution.

For the first term on the right hand side of equation (57), note that

$$\begin{aligned} E\left[\left(t_{n,k}^{\mathrm{MVE}}\right)^2 \middle| t_{(n-k)}\right] &= E\left[\sum_{i=1}^{k} t_{(n+1-i)}^2 \middle| t_{(n-k)}\right] \\ &= E\left[\sum_{i=1}^{k} t_i^2 \middle| t_i > t_{(n-k)}\right] \qquad (59) \\ &= k\left(1 + t_{(n-k)}\,\lambda\left(t_{(n-k)}\right)\right). \end{aligned}$$

So

$$\mathrm{Var}\left(E\left[\left(t_{n,k}^{\mathrm{MVE}}\right)^2 \middle| t_{(n-k)}\right]\right) = k^2\mathrm{Var}\left(t_{(n-k)}\,\lambda\left(t_{(n-k)}\right)\right) \tag{60}$$

and

$$\begin{aligned}\mathrm{Var}\left(\lambda\left(t_{(n-k)}\right)\right) &\approx \mathrm{Var}\left([x\lambda(x)]'\left|_{\mu_{n,k}} \times \left(N^{-1}\right)'(x)\right|_{E\left[\frac{1}{2}\left(1+U_{(n-k)}\right)\right]} \times \tfrac{U_{(n-k)}}{2}\right) \\ &= \frac{(n-k)\left(1+\mu_{n,k}\lambda_{n,k}-\mu_{n,k}^2\right)^2}{(k+1)(n+2)}, \qquad (61)\end{aligned}$$

where the last equality follows from the results of equation (50), together with the fact that $[x\lambda(x)]' = \lambda(x)\left(1 + x\lambda(x) - x^2\right) = \lambda'(x)\left(\frac{1+x\lambda(x)^2-x^2}{\lambda(x)-x}\right)$.

Taken together these imply $\left(t_{n,k}^{\mathrm{MVE}}\right)^2 \overset{\text{approx.}}{\sim} N\left(\mu_{\left(t_{n,k}^{\mathrm{MVE}}\right)^2}, \sigma^2_{\left(t_{n,k}^{\mathrm{MVE}}\right)^2}\right)$ where

$$\begin{aligned}\mu_{\left(t_{n,k}^{\mathrm{MVE}}\right)^2} &= k^2\left(1+\mu_{n,k}\lambda_{n,k}\right) \qquad (62)\\ \sigma^2_{\left(t_{n,k}^{\mathrm{MVE}}\right)^2} &= k\left(\mu_{n,k}^3\lambda_{n,k} + 3\left(1+\mu_{n,k}\lambda_{n,k}\right) - \left(1+\mu_{n,k}\lambda_{n,k}\right)^2\right) \\ &\quad + \tfrac{k^2(n-k)\left(1+\mu_{n,k}\lambda_{n,k}-\mu_{n,k}^2\right)^2}{(k+1)(n+2)}. \qquad (63)\end{aligned}$$

Then

$$\begin{aligned}p &= P\left(t_{n,k}^{\mathrm{MVE}} > t_{n,k,p}^{**}\right) \\ &\approx P\left(\mu_{\left(t_{n,k}^{\mathrm{MVE}}\right)^2} + \sigma_{\left(t_{n,k}^{\mathrm{MVE}}\right)^2}\chi > \left(t_{n,k,p}^{**}\right)^2\right) \qquad (64)\\ &= 1 - N\left(\frac{\left(t_{n,k,p}^{**}\right)^2 - \mu_{\left(t_{n,k}^{\mathrm{MVE}}\right)^2}}{\sigma_{\left(t_{n,k}^{\mathrm{MVE}}\right)^2}}\right),\end{aligned}$$

which implies

$$t_{n,k,p}^{**} \approx \sqrt{\mu_{\left(t_{n,k}^{\mathrm{MVE}}\right)^2} + \sigma_{\left(t_{n,k}^{\mathrm{MVE}}\right)^2} N^{-1}(1-p)}. \qquad (65)$$

Estimator performance is improved, as when $k = n$, by replacing p with $\frac{np}{n+1}$, to correct for the fact that the true distribution has positive skew and excess kurtosis.

B Additional Figures and Tables

Table B.1. Comparison of signal-weighting schemes, value weighted strategies The table reports results from time-series regressions of the return of signal and capitalization weighted value or momentum strategies (panels A and B, respectively) onto other strategies constructed using the same sorting variable but different signal parameterizations (z-score, rank, or top/bottom 35% indicators). Strategies are long/short. The weighting schemes include a standard quantile sort (r^{tertile}, using top and bottom 35%), and strategies that weight stocks on the basis of how far an individual stock's market capitalization times signal (rank or z-score) are from the cross-sectional average (r^{rank} and $r^{\text{z-score}}$, respectively). Data come from CRSP and Compustat, and the sample runs from July 1963 through December 2014).

Independent variable	Dependent variable					
	$r^{\text{z-score}}$		r^{tertile}		r^{rank}	
Panel A: Value strategies						
α	-0.03	0.01	0.06	0.06	-0.04	-0.00
	[-0.62]	[0.48]	[1.49]	[1.79]	[-1.09]	[-0.22]
$r^{\text{z-score}}$			0.93		0.96	
			[65.4]		[81.2]	
r^{tertile}	0.93					0.99
	[65.4]					[202.1]
r^{rank}		0.99		0.96		
		[202.1]		[81.2]		
$\hat{R}^2$ (%)	87.4	98.5	87.4	91.4	91.4	98.5
Panel B: Momentum strategies						
α	0.06	0.02	0.07	0.07	0.02	-0.01
	[0.70]	[0.81]	[0.85]	[0.92]	[0.30]	[-0.37]
$r^{\text{z-score}}$			0.90		0.93	
			[50.0]		[64.1]	
r^{tertile}	0.90					0.99
	[50.0]					[181.8]
r^{rank}		0.99		0.93		
		[181.8]		[64.1]		
$\hat{R}^2$ (%)	80.2	98.2	80.2	87.0	87.0	98.2

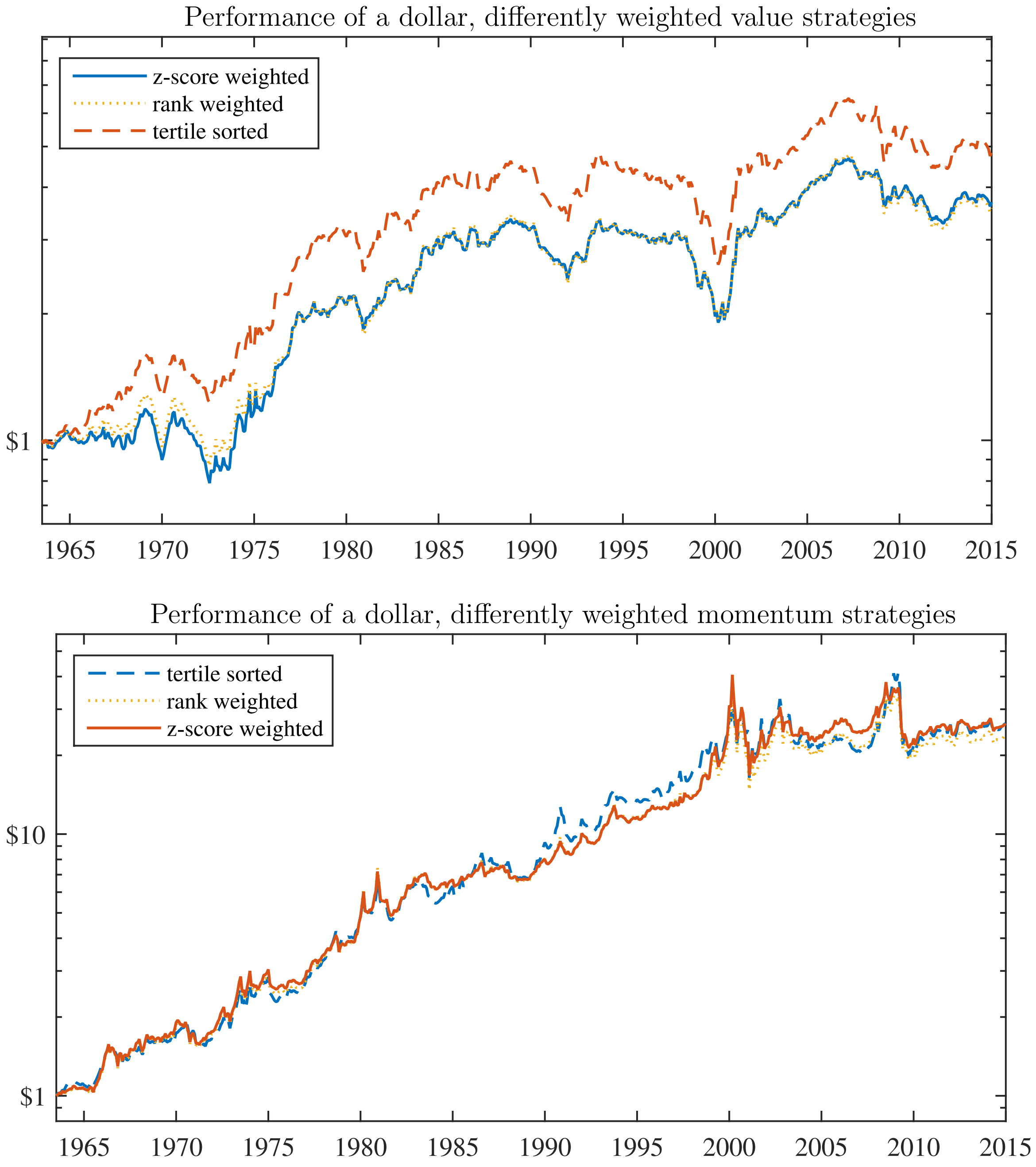


Fig. B.1. Growth of a dollar, differently constructed cap-weighted value and momentum strategies. The figure shows the performance of long/short strategies selected on the basis of book-to-market (top panel, rebalanced annually at the end of June) and performance over the first eleven months of the preceding month (bottom panel, rebalanced monthly). Strategies are tercile sorted (top and bottom 35%, value weighted), or weighted in proportion to the percentile rank or z-score of the stock selection variable and individual stock market capitalizations. Signal-weighted strategies are levered to run at the average in-sample volatility observed on the tertile sorted strategy (10.0% for value and 16.8% for then momentum strategies, respectively). Data come from CRSP and Compustat. The sample covers July 1963 through December 2014.

References

[1] Asness, Cliff, Andrea Frazzini, and Lasse H. Pedersen. 2013. “Quality minus junk.” AQR working paper.

[2] Bailey, David H., and Marcos Lopez de Prado. 2014. “The deflated Sharpe ratio: Correcting for selection bias, backtest overfitting and non-normality.” Journal of Portfolio Management 40, 94–107.

[3] Baker, Malcolm, Jeffrey Wurgler. 2006. “Investor sentiment and the cross section of stock returns.” Journal of Finance 55, 1645–1680.

[4] Barillas, Francisco, and Jay Shanken. 2015. “Which Alpha?” NBER working paper no. 21698.

[5] Frazzini, Andrea, and Lasse H. Pedersen. 2014. “Betting against beta.” Journal of Financial Economics 111, 1–25.

[6] Gompers, Paul, Joy Ishii, and Andrew Metrick. 2003. “Corporate governance and equity prices” Quarterly Journal of Economics 118, 107–156.

[7] Harvey, Campbell R., Yan Liu, and Heqing Zhu. 2016. “...and the cross-section of expected returns.” Review of Financial Studies 29, 5–68.

[8] Lo, Andrew W., and A. Craig MacKinlay. 1990. “Data-snooping biases in tests of asset pricing models.” Review of Financial Studies 3, 431–467.

[9] Markowitz, Harry. 1952. “Portfolio selection.” Journal of Finance 7, 77–91.

[10] McLean, David R., and Jeffrey Pontiff, 2016, “Does academic research destroy stock return predictability?” Journal of Finance 71, 5–32.

[11] Piotroski, Joseph D. 2000. “Value investing: The use of historical financial statement information to separate winners from losers.” Journal of Accounting Research, pp. 1–41.

[12] Stambaugh, Robert F., and Yu Yuan. 2015. “Mispricing factors.” NBER working paper no. 21533.