Backtesting strategies based on multiple signals

Robert Novy-Marx*

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. "Highly significant" backtested performance is easy to generate by selecting stocks on the basis of 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 a bias almost as large as those obtained by selecting 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 a composite measure 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; and Asness, Frazzini, and Pedersen's (2013) quality score, which combines 21 stock level characteristics.

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 form (and a particularly pernicious form) of the data-snooping biases considered generally by Lo and MacKinlay (1990).

This overfitting bias is distinct from the selection bias (or multiple testing bias) confronted by McLean and Pontiff (2013) and Harvey, Liu, and Zhu (2013). Selection

bias results when the researcher employs the best performing signal or signals 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 out of 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 empirical distributions of backtested t-statistics for multi-signal strategies, constructed on the basis of purely random signals. The sorting variable (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.

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, suggesting that the bias that results from combining the best k out of n candidate signals yields biases almost as bad as selecting the single best of n^k candidate signals.

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.

The rest of the paper is organized as follows. Section 2 provides results from a large scale bootstrapping exercise. It describes empirical distributions for backtested t-statistics for strategies, using real stock returns data, selected on the basis of combinations of fictitious, randomly generated "signals." Section 3 presents and analyzes a simplified model, in which signals are uninformative (i.e., do not predict cross sectional differences in average returns), and the returns to strategies selected on the basis of individual signals are uncorrelated and normally distributed. It derives critical t-statistics for strategies selected on the basis of combinations of signals, and uses these results to develop intuition regarding the factors that determine the magnitude of the backtesting biases. Section 4 derives the approximate power law governing the interaction between the selection and overfitting

biases, showing that backtests of a strategy based on the best k of n candidate signals are biased almost as badly as those of a strategy based on the single best of n^k candidate signals. Section 5 concludes.

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 thousands of times. By construction, these signals have no real power, and cannot predict performance out-of-sample.

2.1 Strategy construction

Given any signal, I construct a long/short strategy, rebalanced at the end of June, using return and capitalization data for individual stocks from CRSP. The weight of a stock in the long or short side of a strategy is proportional to both the signal and a capitalization multiplier. That is, stock i is held proportionally to $(S_{i,t} - S_t)m_{i,t}$, where $S_{i,t}$ and $m_{i,t}$ are the signal and capitalization multiplier for the stock i at time t, and S_t is the median signal across all stocks. This specification embeds many common construction schemes. For example, if the signal is an indicator for whether some stock characteristic is in the extreme 10% of the cross-sectional distribution, then the strategies are just long and short extreme deciles, equal weighted if the capitalization multiplier is one for all stocks and value weighted if it is market equity. If the signal is the cardinal ranking of the characteristic and the capitalization multiplier is always one, then it gives the rank-weighting scheme employed by Frazzini and Pedersen (2014) to construct their betting-against-beta (BAB)

factor.

I focus on signal-and-capitalization weighted strategies, because these strategies map most closely into the theoretical model presented in Section 3, and because the methodology is similar to that commonly employed in industry. For the capitalization multiplier I use market equity. The signals are randomly generated "z-scores," drawn independently for each stock at the end of each June from a standard normal distribution, or linear combinations of these randomly generated signals.[1] Using rank weighting, or a simple quantile sort, does not change any of the conclusions of the analysis. In fact, Appendix A shows that the performance of both signal and rank weighted strategies is essentially indistinguishable from that of strategies based on simple tercile sorts, equal weighted if the capitalization multiplier is one and value weighted if the capitalization multiplier is market equity. So, for example, the salient feature of the rank-weighting scheme employed by Frazzini and Pedersen (2014) to construct their BAB factor is not the rank-weighting, but that the scheme ignores market capitalizations. The long and short sides of the strategy are thus almost indistinguishable, before levering or unlevering, from equal-weighted portfolios holding the top or bottom 35% of stocks by estimated market betas.

2.2 A simple illustration

Before analyzing the full scope of the overfitting bias that results form combining multiple signals, it is useful to illustrate the associated problems as simply as possible. This illustration simply demonstrates that combining signals that backtest positively can yield impressive backtested results, even when none of the signals employed to construct

[1] In particular, because these strategies' portfolio weights are linear in the signals, and the composite signals are linear combinations of the underlying signals, strategies based on composite signals are linear combinations of the strategies based on individual signals. This fact is employed in Section 3, and simplifies analysis of the model.

the composite signal has real power, or even individually generates significant backtested results.

To do this, I generate 10,000 sets of 100 random signals. By construction, none of the signals have any true power to predict returns. For each set of signals I construct signal-and-cap weighted strategies, and record t-statistics for these strategies' excess returns. For each set I then combine the signals of the two to ten strategies that backtested with absolute t-statistics closest to either one, or to one and a half.[2] These signals are combined through simple addition, after first flipping the signs on strategies that generated negative excess returns. Because of the computational intensity of the procedure, I limit the analysis to a relatively short sample, covering July 1993 through the end of 2014.

Figure 1 shows the average backtested t-statistics for strategies based on combinations of from two to ten signals that all individually backtest with t-statistics of basically one (lower solid line) or one and a half (upper solid line). Despite the fact that none of the underlying signals are themselves significant, or have any true power, the figure shows that the combined signals often look highly significant in the backtests. In fact, the figure shows that the average t-statistic is almost as large as the average t-statistic of the underlying signals times the square root of the number of signals employed (dotted lines). As a result, combining just two signals with t-statistics of 1.5, or five signals with t-statistics of one, yields strategies that have average t-statistics significant at the 5% level, at least if one counterfactually assumes, as is common, that the combined strategy t-statistics have a standard normal distribution.

[2]Specifically, I select the $k = 2, 3, ..., 10$ strategies corresponding to those with consecutive order statistics with absolute backtested t-statistics with means closest to either one or 1.5.

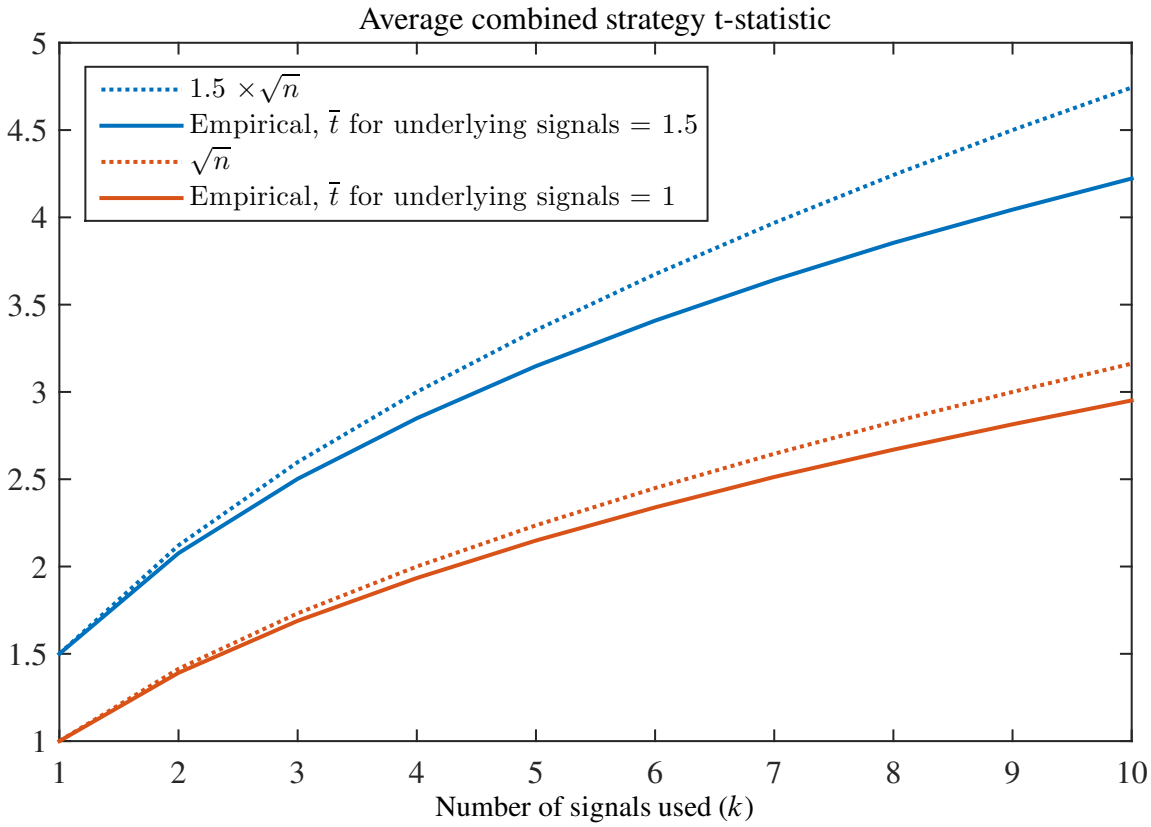

Fig. 1. Average t-statistic on strategies that combine insignificant signals. The figure shows the t-statistic on the excess return to strategies that combine from one to ten random signals, where each of the underlying signals yields strategies with t-statistics of one (lower solid line) or 1.5 (upper solid line). Strategies are signal-and-cap weighted, i.e., stocks are held in direct proportion to both market capitalization and the demeaned signal used for strategy construction. Return and capitalization data come from CRSP, and the sample covers July 1993 through December 2014. Average t-statistics are calculated over 10,000 sets of random signals. Dotted lines show the average t-statistic of the underlying signals, multiplied by the square-root of the number of signals combined.

2.3 Best *k*-of-*n* strategies

The previous subsection clearly illustrates the overfitting problem that biases backtested performance of multi-signal strategies. The construction of those strategies, however, with all the underlying signals selected to produce similar, modest performance, is highly artificial. This section analyses strategies that are more likely to arise naturally out of a real

research process. In this 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. 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 additional degrees of 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 strategies based on the individual signals. These correspond to the weights on the ex-post mean-variance portfolio of the individual strategies, assuming the strategies are uncorrelated and have identical volatilities.

2.3.1 Distribution of t-statistics from multi-signal strategies

Figure 2 shows the distribution of t-statistics for some simple best k-of-n strategies. Results are again created by generating 10,000 sets of n random signals. The figure shows results kernel smoothed with a bandwidth of 0.2.

Panel A shows the case of pure selection bias ($k = 1$), for $n \in \{1, 2, 4, 6, 10\}$. It also shows, for comparison, the density for the absolute value of the standard normal. The empirical distribution of t-statistics for best 1-of-1 strategy looks approximately normal, though with a slightly fatter tail, reflecting the excess kurtosis and heteroskedasticity

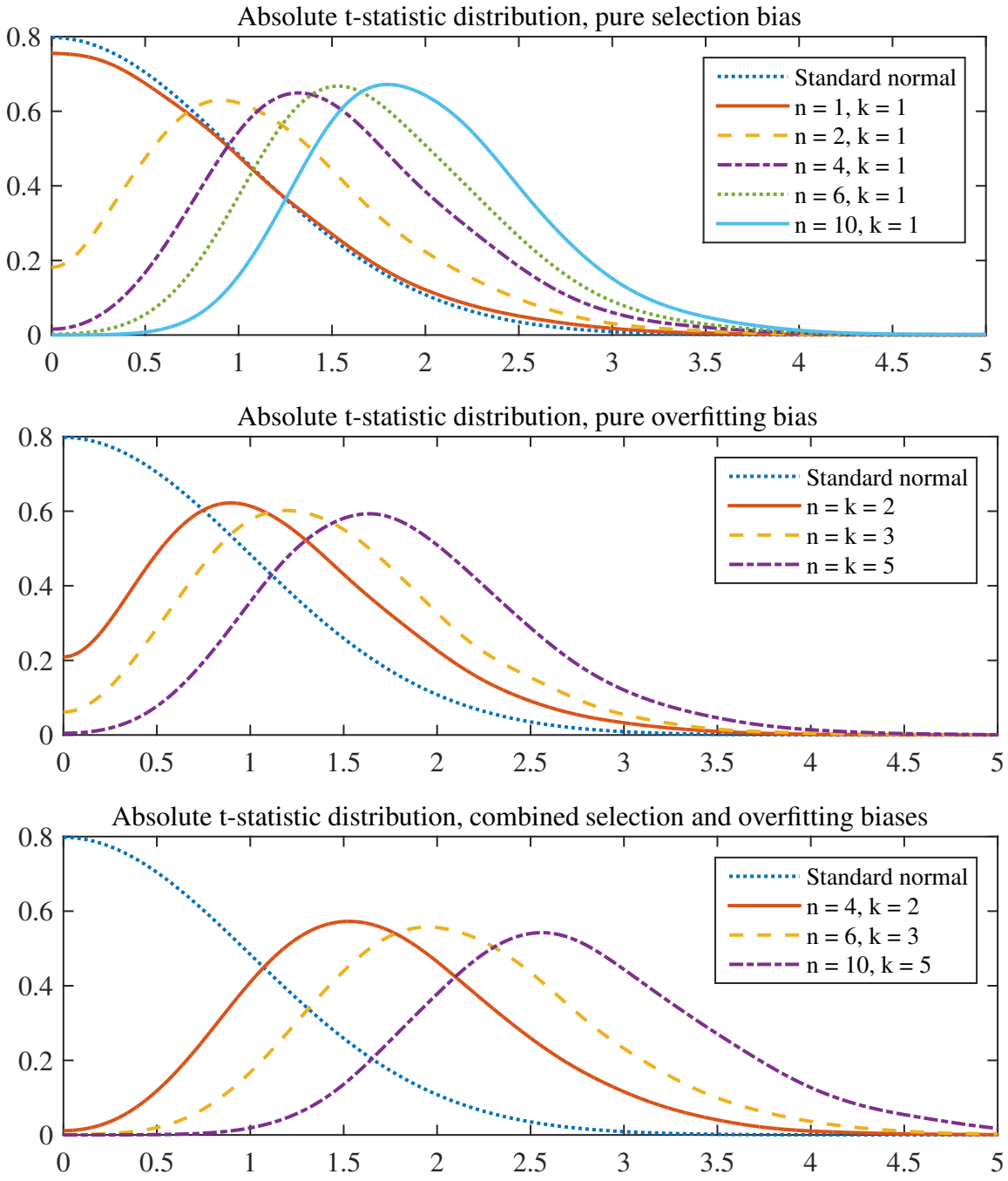

Fig. 2. Empirical t-statistic distribution for best k-of-n strategies, equal-weighted signals. Panel A shows the case of pure selection bias ($k = 1$), for $n \in \{1, 2, 4, 6, 10\}$; Panel B the case of pure overfitting bias ($k = n$), for $n \in \{2, 3, 5\}$; and Panel C the combined case, when $n = 2k$ for $n \in \{2, 3, 5\}$. Distributions are bootstrapped from 10,000 draws of n randomly generated signals, and kernel smoothed with a bandwidth of 0.2. Strategies are signal-and-cap weighted, with stocks held in direct proportion to both market capitalization and the demeaned signal used for strategy construction, and rebalanced annually, at the end of June. Return and capitalization data come from CRSP, and the sample covers July 1993 through December 2014.

observed on equity strategy returns. The densities for the cases when $n > 1$ still look approximately normal, but are shifted to the right. For $n = 2$ the mode has shifted out from zero to almost one. For $n = 10$ it is almost two.

Panel B shows the case of pure overfitting bias ($k = n$), for $n \in \{2, 3, 5\}$. Results are shown for signals that are equal weighted; signals weighting the signal yields even more extreme results (shown in Appendix B). The figure suggests that even the pure overfitting biases are more acute that the selection biases. The distribution of t-statistics for the best 5-of-5 strategy is shifted almost as far as it is for the best 1-of-10 strategy.

Panel C shows the impact of the biases, depicting the cases when the investigator considers twice as many signals as she uses ($n = 2k$), for $n \in \{2, 3, 5\}$. The effects clearly compound. More than half of the mass of the distribution for the best 3-of-6 strategy is to the right of two—so most strategies selected this way look "highly significant." Employing three signals in the selection criteria when constructing marketed indices is relatively common, suggesting that the highly significant performance of many of these indices, inferred from t-statistics greater than 1.96, is anything but.

2.3.2 Critical t-statistics for multi-signal strategies

Because the distribution of t-statistics for multi-signal strategies does not have a standard normal distribution, critical values derived from that distribution cannot be used to draw inferences regarding significance of performance for multi-signal strategies.

Figure 3 shows empirical 5% critical values for best 1-of-n strategies, which suffer from pure selection bias, and 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 by signal weighing the individual signals. It shows that the impact on the critical threshold from the sample selection and equal weighted overfitting biases are similar for very small n, but that the impact of the

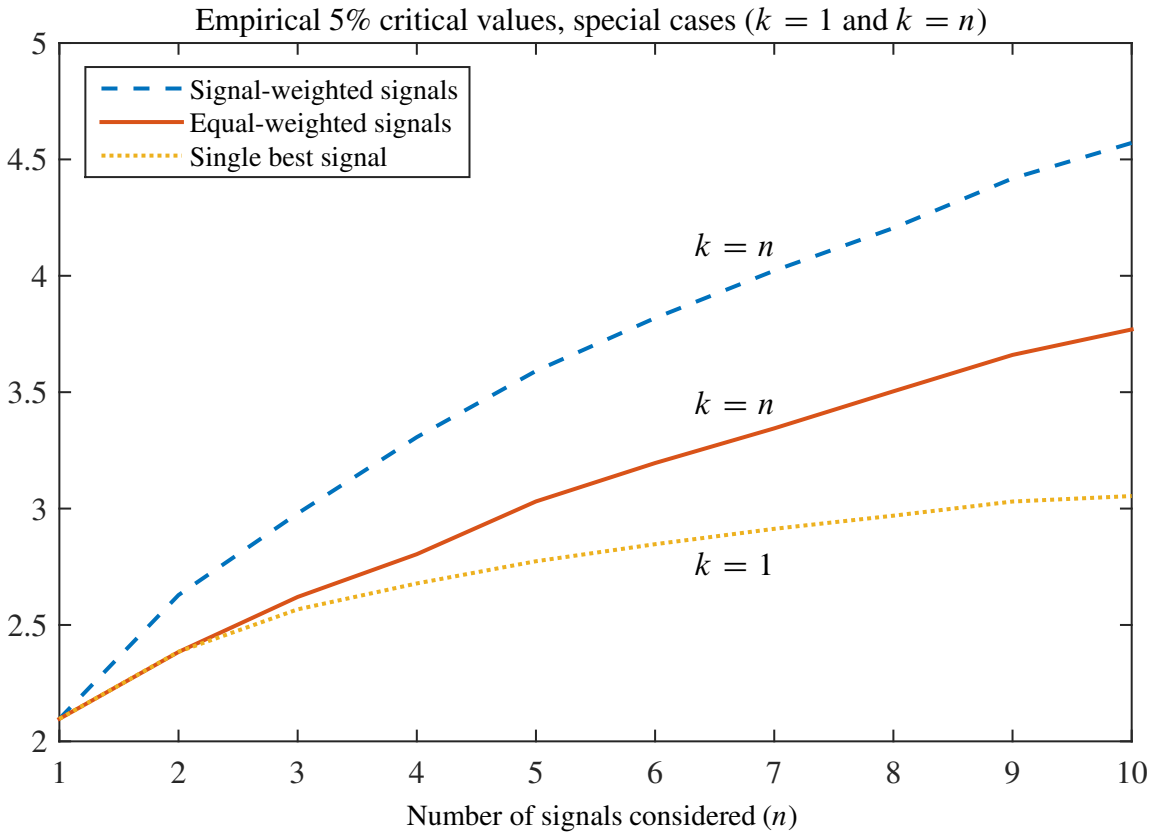

Fig. 3. Five percent critical t-statistics for best 1-of-n and best n-of-n strategies. The bottom, dotted line shows 5% critical thresholds for the pure selection bias case, when the investigator presents the strongest result from a set of n random strategies, where $n \in \{1, 2, ..., 10\}$. The middle, solid line and the top, dashed line, 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 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. Critical values come from generating 10,000 sets of n randomly generated signals. Strategies are signal-and-cap weighted, with stocks are held in direct proportion to both market capitalization and the demeaned signal used for strategy construction, and rebalanced annually, at the end of June. Return and capitalization data come from CRSP, and the sample covers July 1993 through December 2014.

overfitting bias is more acute for even moderately large n. It also shows that overfitting by signal weighting the signals yields much higher critical values than equal weighting the signals. That is, the pure overfitting problem is more acute when the investigator has more freedom to overweight good signals. The critical value of 2.1 at $n = 1$ reflects the fat tail of the distribution of t-statistics for the best 1-of-1 case, observed in Figure 2. This excess kurtosis yields a higher frequency of extreme t-statistics, pushing up the critical value relative to the one obtained under the standard normal assumption.

Figure 4 shows empirical 5% critical values for best k-of-n strategies, which suffer from both selection and overfitting biases. It shows these critical values for the cases when the best one to ten signals are employed, 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 3-of-10 strategies the 5% critical t-statistic is almost four; for the best 3-of-20 strategies it is almost five; for the best 7-of-100 strategies it is almost seven. 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.

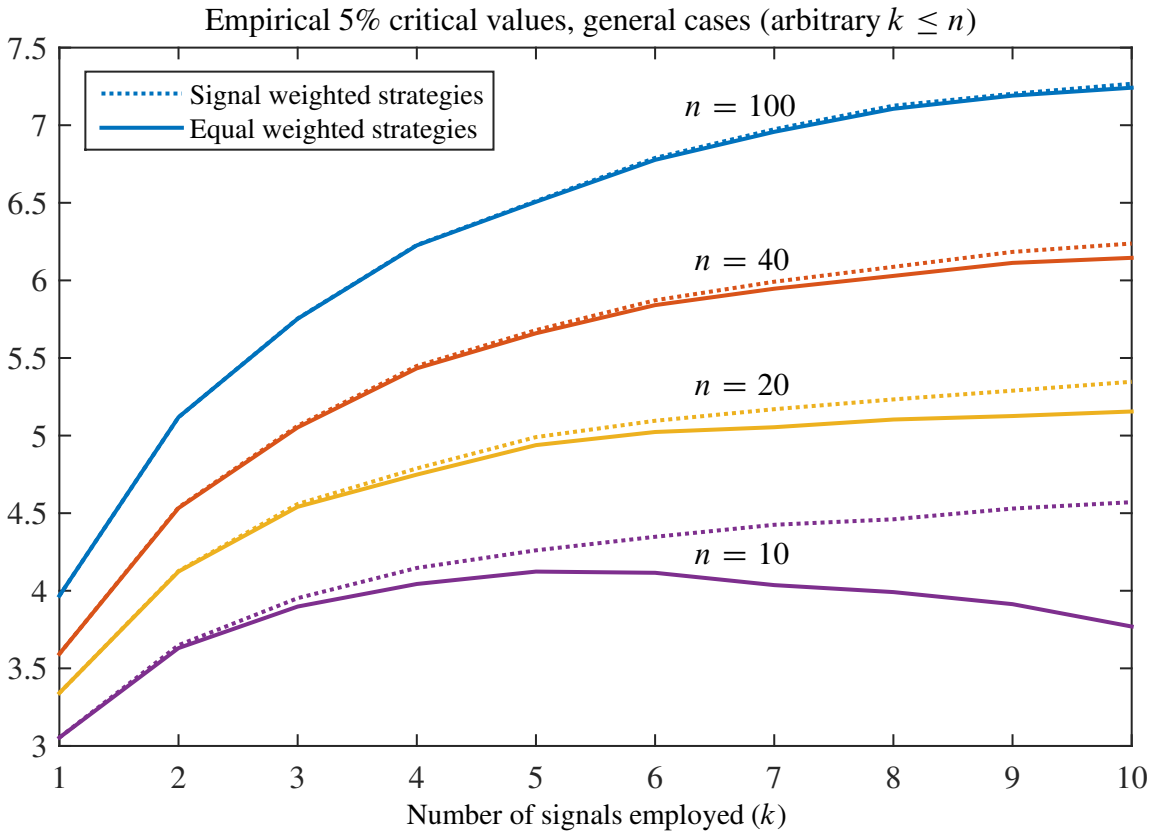

Fig. 4. Five percent critical t-statistics for best k-of-n strategies. The figure shows 5% critical thresholds for strategies selected using a signal constructed by combining the best $k = 1, 2, ..., 10$ performing signals, when the investigator considered $n \in \{10, 20, 40, 100\}$ candidate signals. 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 10,000 sets of n randomly generated signals. Strategies are signal-and-cap weighted, with stocks are held in direct proportion to both market capitalization and the demeaned signal used for strategy construction, and rebalanced annually, at the end of June. Return and capitalization data come from CRSP, and the sample covers July 1993 through December 2014.

3. Model results

To develop intuition for the results observed in the preceding section, I now 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 returns to signal-weighted strategies (i.e., long/short strategies that hold stocks in proportion to the demeaned signals) are normally distributed, uncorrelated across signals, and have the same volatilities (or are levered to have the same volatilities). 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 the 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, so a strategy selected on the basis of a composite signal is consequently identical to a portfolio of strategies based on the individual underlying signals, held in proportion to the weights used to construct the composite signal.[3]

Given n signals, the ex-post Sharpe ratio to the strategy that selects stocks based on a composite signal that puts weights $\vec{\omega}$ on the individual signals is thus the same as that to

[3]That is, the assumption that individual stock weights are proportional to the signal used to select them, together with composite signals that are linear combinations of the underlying signals, yields an exact equivalence between "siloed" and "integrated" solutions (i.e., the solution that runs multiple pure strategies side-by-side, and the one that select stocks to get exposure to all the signals simultaneously). While these assumptions are not satisfied for standard quantile sorted strategies, Appendix A shows that the performance of quantile sorted strategies is essentially indistinguishable from that of rank-weighted strategies, or from z-score weighted strategies, which do satisfy the assumptions.

the portfolio that puts weights $\vec{\omega}$ on strategies selected from the individual signals,

$$SR_{\vec{\omega}} = \frac{\vec{\omega}'\vec{\mu}^e}{\sqrt{\vec{\omega}'\Sigma\vec{\omega}}}, \tag{1}$$

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

$$\hat{t}_{\vec{\omega}} = \frac{\vec{\omega}'\vec{t}}{\sqrt{\vec{\omega}'\vec{\omega}}}, \tag{2}$$

where $\vec{t}$ 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 ($\vec{\omega} = \mathbb{1}/\left\|\mathbb{1}\right\|_1$) and signal-weighting the signals ($\vec{\omega} = \vec{t}/\left\|\vec{t}\right\|_1$), here correspond to well understood strategies, the minimum variance and ex-post mean-variance efficient portfolios, respectively.

Without loss of generality, we may assume that the elements of $\vec{t}$ 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 $t^{\text{MV}}_{n,k}$ and

$t_{n,k}^{\text{MVE}}$, respectively, are given by

$$t_{n,k}^{\text{MV}} = \frac{\mathbb{1}'\mathbb{I}_{n,k}\vec{t}}{\sqrt{\mathbb{1}'\mathbb{I}_{n,k}\mathbb{1}}} = \frac{\left\|\mathbb{I}_{n,k}\vec{t}\right\|_1}{\sqrt{k}} = \frac{\sum_{i=1}^{k} t_{(n+1-i)}}{\sqrt{k}} \tag{3}$$

$$t_{n,k}^{\text{MVE}} = \frac{\vec{t}'\mathbb{I}_{n,k}\vec{t}}{\sqrt{\vec{t}'\mathbb{I}_{n,k}\vec{t}}} = \left\|\mathbb{I}_{n,k}\vec{t}\right\|_2 = \sqrt{\sum_{i=1}^{k} t_{(n+1-i)}^2}, \tag{4}$$

there $t_{(j)}$ 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 largest k order statistics of $\vec{t}$ divided by $\sqrt{k}$. Equivalently, it is $\sqrt{k}$ times the average t-statistic of the strategies corresponding to the signals employed. This conceptualization is basically consistent with the empirical results presented in Figure 1, which showed that the backtested performance of strategies selected on an average of k signals that individually backtest with similar strength was roughly $\sqrt{k}$ times as strong as the backtested performance of strategies based on the individual signals.

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 values for special cases: Best 1-of-*n* and *n*-of-*n* strategies

Before analyzing arbitrary $t_{n,k,p}^{*}$ and $t_{n,k,p}^{**}$, it is again useful to develop some intuition by first considering the extreme cases in k. 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 returns that come from the whole sample.

3.2.1 Pure sample 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, $u_{(k)} \sim B(k, n+1-k)$. 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 draws from a standard, normal random variable of

$$t^*_{n,1,p} \;=\; t^{**}_{n,1,p} \;=\; N^{-1}\left(\left(1-\tfrac{p}{2}\right)^{1/n}\right), \tag{5}$$

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

These critical values can be interpreted by recognizing that, for small p, $\left(1-\frac{p}{2}\right)^{1/n} \approx 1-\frac{p}{2n}$. In fact, when $p < 35\%$ then for any n the difference is at least an order of magnitude smaller than p, $\left|N^{-1}\left(\left(1-\frac{p}{2}\right)^{1/n}\right) - N^{-1}\left(1-\frac{p}{2n}\right)\right| < p/10$. 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, $p \approx n \times 2\left(N(-|t|)\right)$. Put simply, the observed result is n times more likely to have occured randomly than is typically reported. 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 straegy is consequently distributed as a chi-distribution with n degrees of freedom,

$$t_{n,n}^{\text{MVE}} \quad = \quad \left\| \vec{t} \right\|_2 \quad \sim \quad \sqrt{\chi_n^2}. \tag{6}$$

The critical t-statistic for the ex-post MVE combined strategy constructed using the n randomly selected signals comes from inverting the chi-squared distribution,

$$t_{n,n,p}^{**} \quad = \quad \sqrt{\Phi_{\chi_n^2}^{-1}(1-p)}, \tag{7}$$

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

$\lim_{n\to\infty} \left\|\vec{t}\right\|_1 \sim N(n\sqrt{2/\pi}, n(1-2/\pi))$, and

$$t^{\mathrm{MV}}_{n,n} \;=\; \frac{\left\|\vec{t}\right\|_1}{\sqrt{n}} \;\underset{n\to\infty}{\sim}\; \sqrt{\tfrac{2n}{\pi}} + \left(\sqrt{1-\tfrac{2}{\pi}}\right)\chi. \tag{8}$$

This implies an asymptotic critical value of

$$t^{*}_{n,n,p} \;\approx\; \left(\sqrt{\tfrac{2}{\pi}}\right)\sqrt{n} + \left(\sqrt{1-\tfrac{2}{\pi}}\right)N^{-1}(1-p)\,. \tag{9}$$

The true distribution of $t^{\mathrm{MV}}_{n,n}$ is positively skewed and has excess kurtosis, especially when n is small, 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 $\frac{np}{n+1}$,

$$t^{*}_{n,n,p} \;\approx\; \left(\sqrt{\tfrac{2}{\pi}}\right)\sqrt{n} + \left(\sqrt{1-\tfrac{2}{\pi}}\right)N^{-1}\left(1-\tfrac{np}{n+1}\right). \tag{10}$$

Figure 5 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 line shows the analytic approximation for $t^{*}_{n,n,5\%}$, provided in equation (10), 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 3. The model critical values are uniformly roughly 7% below their empirical counterparts, for all three cases and across the range of signals employed (k), because the model lacks the excess kurtosis and heteroskedasticity observed in the data.

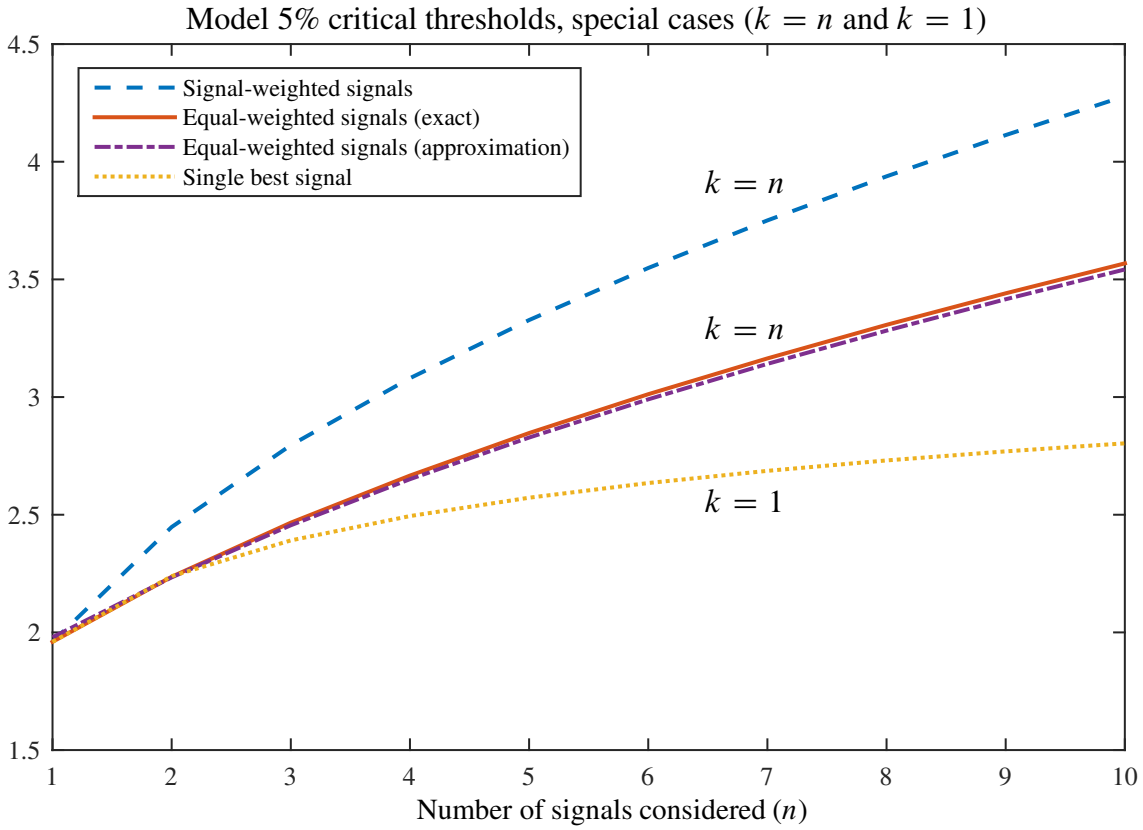

Fig. 5. Model 5% critical t-statistics for best 1-of-n and best n-of-n strategies. The bottom, dotted line shows 5% critical thresholds for the pure selection bias case, when the investigator presents the strongest result from a set of n random strategies, where $n \in \{1, 2, ..., 10\}$. The middle, 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.

3.3. General case: best *k*-of-*n* strategies

To calculate the critical t-statistic more generally, when the selection and overfitting biases interact, note that the order statistics for n draws of the standard uniform distribution are distributed uniformly on the standard n-simplex $\Delta_*^n = \{(u_1, u_2, ..., u_n) \in \mathbb{R}^n | 0 \leq u_1 \leq u_2 \leq ... \leq u_n \leq 1\}$, which has a volume of $1/n!$. Using the symmetry of the

normal distribution, the order statistics for the absolute values of n draws of the standard normal distribution is obtained by pulling the simplex back through the cumulative normal distribution, $N^{-1}\left(\frac{\mathbb{1}+\Delta_*^n}{2}\right)$, where the inverse normal is taken element-by-element. The probability that $t_{n,k}^{\text{MV}}$ and $t_{n,k}^{\text{MVE}}$ are greater than any given τ are thus

$$P[t_{n,k}^{\text{MV}} > \tau] = E\left[\mathbb{1}_{\left\{\left\|\mathbb{I}_{n,k}\boldsymbol{t}\right\|_1 \geq \sqrt{k}\tau\right\}}\right] \tag{11}$$

$$P[t_{n,k}^{\text{MV}} > \tau] = E\left[\mathbb{1}_{\left\{\left\|\mathbb{I}_{n,k}\boldsymbol{t}\right\|_2 \geq \tau\right\}}\right], \tag{12}$$

where $\boldsymbol{t} = N^{-1}\left(\frac{\mathbb{1}+\boldsymbol{u}}{2}\right)$, and $\boldsymbol{u}$ is distributed uniformly on the simplex Δ_*^n. These equations implicitly define critical thresholds for the best k-of-n strategy for any p-value, $t_{n,k,p}^* = \left\{\tau | P[t_{n,k}^{\text{MV}} > \tau] = p\right\}$ and $t_{n,k,p}^{**} = \left\{\tau | P[t_{n,k}^{\text{MVE}} > \tau] = p\right\}$.

While these do not admit analytic solutions, they are easy to calculate numerically. Figure 6 shows 5% critical values as a function k, the number of signals employed, for $k = 1, 2, ...10$, for cases in which the investigator considers ten, 20, 40, or 100 signals (i.e., $n \in \{10, 20, 40, 100\}$). The figure again shows a strong resemblance to the corresponding empirical critical values provided in Figure 4. The model shows similar, though slightly lower, critical values for the pure selection bias cases ($k = 1$), strong, but slightly weaker dependence on k for small k, and slightly high sensitivity to k for large k.[4] The figure 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 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.

[4]The model's results correspond even more closely to the empirical distribution of critical t-statistics for signal-weighted strategies constructed without regards to market capitalizations (i.e., $m_{i,t} = 1$ for all i and t). The empirical critical values for these strategies are shown in Table 11 of Appendix C.

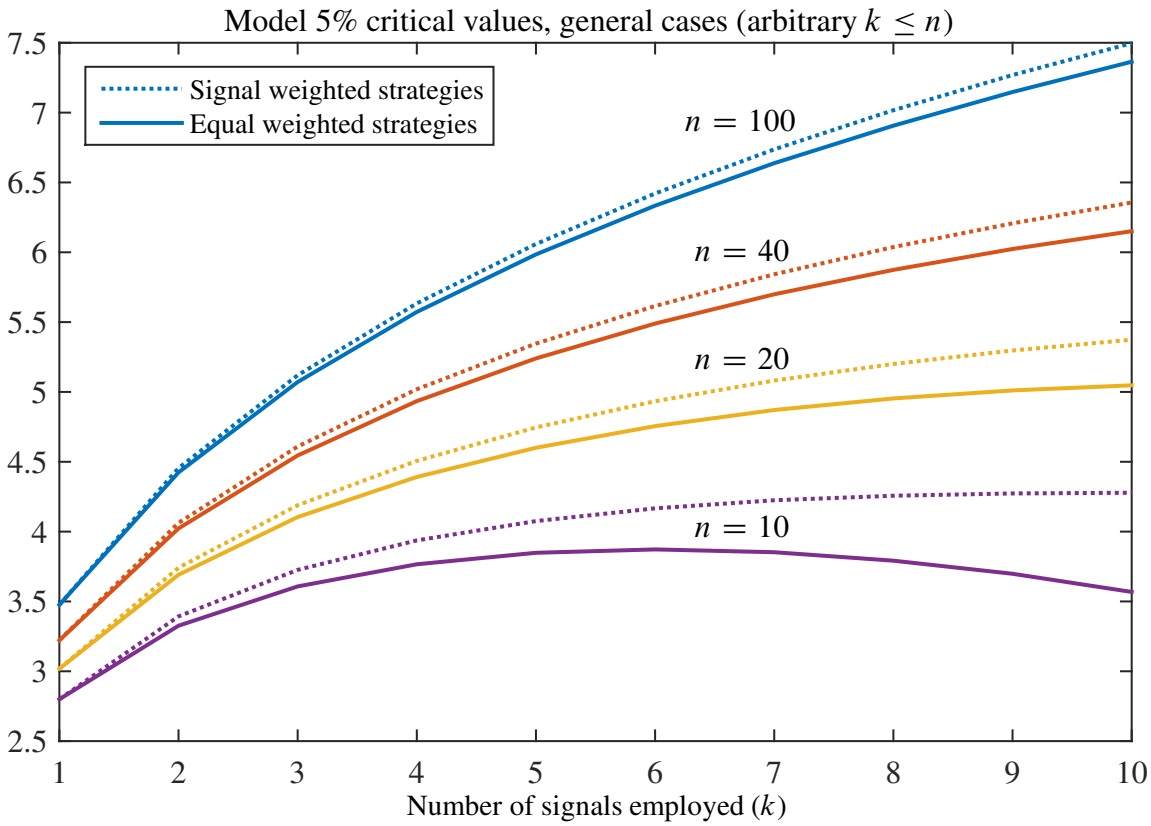

Fig. 6. Model 5% critical t-statistics for best k-of-n strategies. The figure shows 5% critical thresholds for strategies selected using a signal constructed by combining the best $k = 1, 2, ..., 10$ performing signals, when the investigator considered $n \in \{10, 20, 40, 100\}$ candidate signals. Solid 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.

3.3.1 Critical value approximation, general case

Appendix D derives relatively simple analytic approximations, given by

$$t^{*}_{n,k,p} \approx \mu_{t^{\text{MV}}_{n,k}} + \sigma_{t^{\text{MV}}_{n,k}} N^{-1}\left(1 - \tfrac{kp}{k+1}\right) \tag{13}$$

$$t^{**}_{n,k,p} \approx \sqrt{\mu_{(t^{\text{MVE}}_{n,k})^2} + \sigma_{(t^{\text{MVE}}_{n,k})^2} N^{-1}\left(1 - \tfrac{kp}{k+1}\right)}, \tag{14}$$

where

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

and

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

The variances terms in equations (13) and (14) are relatively insensitive to n and k, while the means are strongly increasing in both indices (at least for small k), consistent with the basic rightward shift in the empirical distributions observed in Figure 2.

Figure 7 compares these analytic approximations to the exact critical values, calculated 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 show 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 $k \approx n/2$.

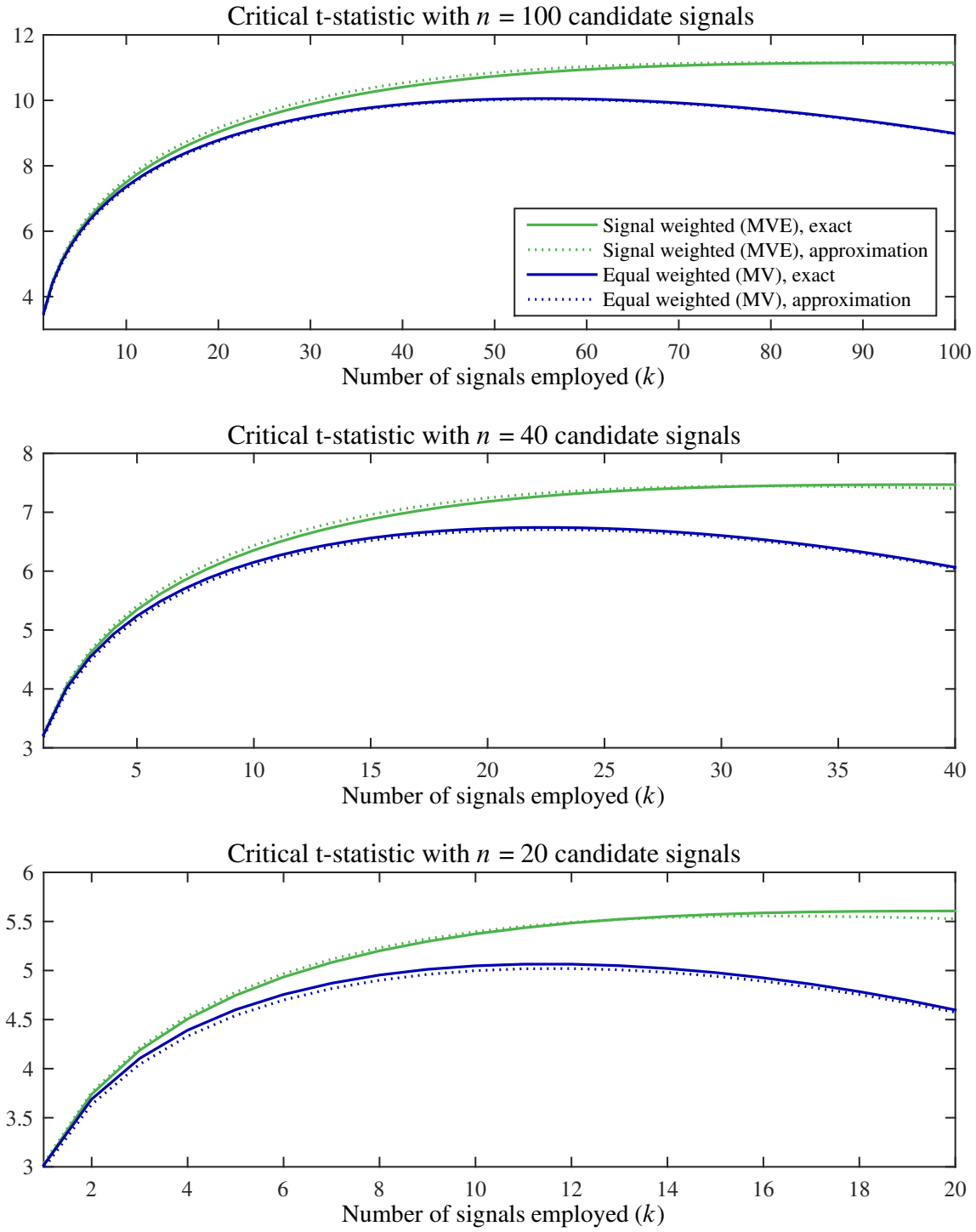

Fig. 7. Comparison of exact and approximate 5% critical t-statistics for best k-of-n strategies. The figure shows 5% critical thresholds for best k-of-n strategies. Solid lines are exact values, while dotted lines are analytic approximations. The top, lighter lines show critical values for strategies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

The performance of the ex-post mean variance efficient strategy is 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 quality of the signal 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 (3), this implies

$$\frac{E\left[t_{(n-k)}\right]}{E\left[\sum_{i=1}^{k} t_{(n+1-i)}\right]} = \frac{1}{2k}, \tag{15}$$

or, after rearranging and using iterated expectations, that

$$E\left[t_{(n-k)}\right] = E\left[\frac{E\left[\sum_{i=1}^{k} t_{(n+1-i)} \middle| t_{(n-k)}\right]}{2k}\right] = \frac{E\left[\lambda\left(t_{(n-k)}\right)\right]}{2}. \tag{16}$$

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

$$\frac{k+1}{n+1} \approx 2N(-x^*) = 0.541, \tag{17}$$

or that $k \approx n/2$. 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 $E[t^{\text{MV}}_{2k,k}] \approx 4n\left(N^{-1}(3/4)\right)\sqrt{k} = 1.27\sqrt{k}$, almost 60% higher that the t-statistic that would have been expected absent the selection bias, $E[t^{\text{MV}}_{k,k}] \approx \sqrt{2k/\pi} = 0.8\sqrt{k}$.

4 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 to get the same bias selecting stocks using a single signal. That is, given any critical value τ and p-value p, we can find n^* such that $\tau = N^{-1}\left(\left(1 - \frac{p}{2}\right)^{1/n^*}\right)$. Solving for n^*, the number of single-signal strategies the investigator would need to consider to have the same critical value, is thus

$$n^* = \frac{\ln\left(1 - \frac{p}{2}\right)}{\ln\left(N(\tau)\right)}. \tag{18}$$

Table 1. Single-signal candidates required to get 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 (i.e., reports n^* such that $t^*_{1,n^*,5\%} = t^*_{k,n,5\%}$). Panel B reports the cases when the signals are signal-weighted (i.e., reports n^* such that $t^*_{1,n^*,5\%} = t^{**}_{k,n,5\%}$).

Signals considered (n)	Signals used (k)				
	2	3	4	5	10
Panel A: Equal-weighted signals (minimum variance strategies)					
10	57	164	305	426	141
20	227	1,250	4,490	12,000	113,000
50	1,360	17,500	147,000	900,000	5.03×10^{8}
100	5,220	128,000	2.02×10^{6}	2.33×10^{7}	2.85×10^{11}
Panel B: Signal-weighted signals (mean-variance efficient strategies)					
10	73	260	614	1,102	2,690
20	277	1,810	7,680	24,400	658,000
50	1,600	23,400	223,000	1.55×10^{6}	1.70×10^{9}
100	6,020	165,000	2.88×10^{6}	3.70×10^{7}	7.78×10^{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 signal-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 1,250 candidates. With five signals selected from 50 candidates, the bias is almost as bad as if the investigator had used the single best performing signal from one million. 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^*) generally grows quickly with the number of signals employed (k). To quantify this relation, note that $n^* \approx \frac{p/2}{N(-\tau)}$. Taking logs, using the approximation $N(-\tau) \approx n(\tau)/\tau$ for large τ, gives that $\ln n^*$ is of order τ^2, or using the approximation for $t^*_{n,k,p}$ given in equation (13),

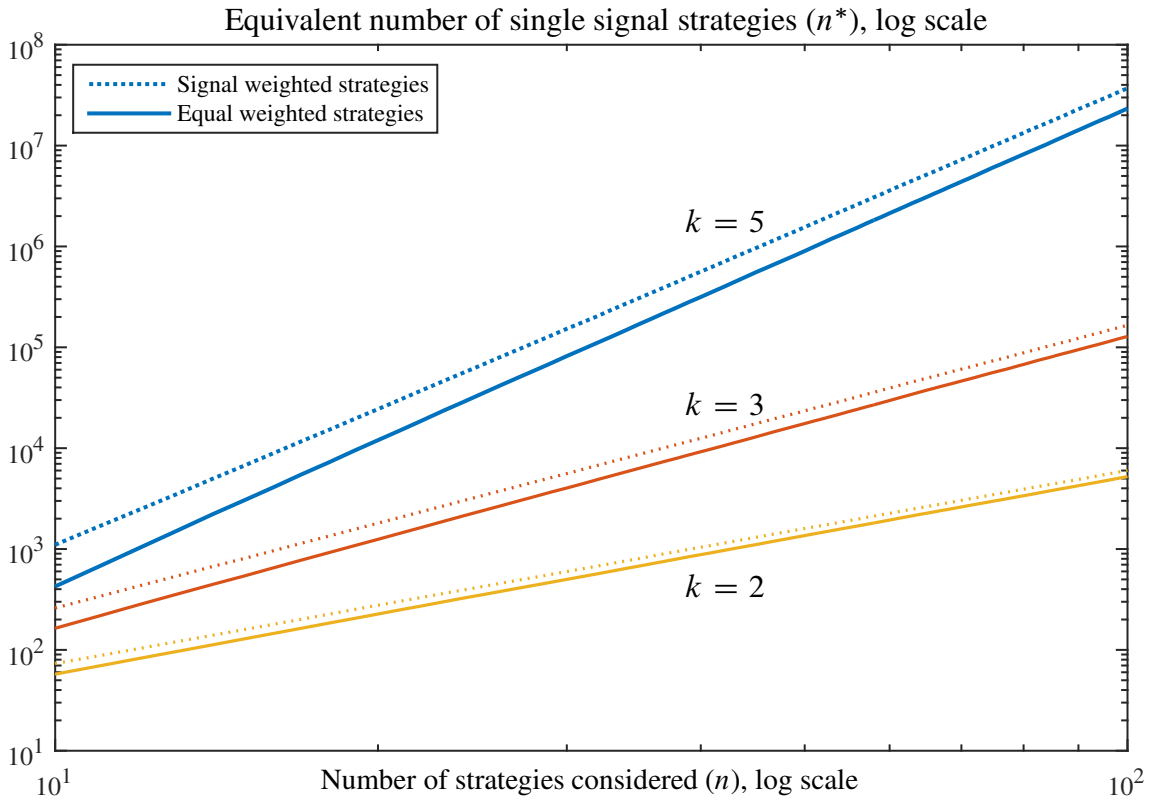

Fig. 8. Number of single candidates strategies to generate the five percent critical thresholds for best k-of-n strategies. The figure plots, using a log-log scale, the number of single signals required to generate the best k-of-n critical values at the 5% level, $t^*_{n,k,5\%}$ (equal-weighted signals, 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.

that $\ln n^*$ is of order $k\lambda^2_{n,k}$. Finally, using $\lambda_{n,k} \approx \mu_{n,k} = N^{-1}\left(1 - \frac{k+1}{2(n+1)}\right)$ for $k \gg n$ and the inverse normal approximation $N^{-1}(\rho) \approx \sqrt{\frac{-\ln(4\rho(1-\rho))}{\sqrt{\pi/8}}}$, this implies that $\ln n^*$ is approximately affine in $\ln n$, with a slope approximately proportional to k. That is, n^* is roughly proportional to n^k.

This approximate power relation is evident in Figure 8, which plots, on a log-log scale, the number of single-signal candidates required (n^*) to generate the best k-of-n 5% critical values, $t^*_{n,k,5\%}$ (equal-weighted signal, solid lines) and $t^{**}_{n,k,5\%}$ (signal-weighted

signal, dotted lines), as a function of the number of candidate strategies actually considered (n, form ten to 100). This n^* is plotted for strategies constructed using the best two, three, and five signals. As predicted, the figure shows an approximately linear log-log relation, with slopes proportional to the number of signals employed.

This approximate power law can be understood intuitively as follows. The biases that result from using the best k-of-n signals are worse than one would get by partitioning the signals into k random sub-samples of size n/k, and then using the single best signal from each sub-sample. The expected in-sample improvement in the quality of the employed signals that results from going from one candidate to n/k signals is almost as large as going from one signal to n signals, because the oversampling bias is highly convex in the candidate sample size, and the effects on n^* are multiplicative. The bias that results from the best k-of-n strategies is consequently almost as pronounced as if the investigator had used the single best signal from n^k candidate strategies.

5. Conclusion

Multi-signal strategies cannot be evaluated using conventional tests. Combining spurious, marginal signals, it is easy 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. An easier solution is to evaluate the marginal power of each signal separately.

This is not to say that one should not use multiple signals that one believes in. Signals that work well individually will work even better together. One should not, however, believe in multiple signals because they backtest well together. Backtesting well together does not imply that any of the signals, or even the combined signal, has any power.

A. Signal and rank weighted strategies

The strategies considered in Section 2 hold stocks in proportion to $(S_{i,t} - S_t)m_{i,t}$, where $S_{i,t}$ and $m_{i,t}$ are the signal and capitalization multiplier for stock i at time t, and S_t is the median signal across all stocks. This specification embeds standard quantile sorts (when $S_{i,t}$ equals an indicator that the sorting characteristic is in the top $x\%$ of the cross-sectional distribution, minus an indicator that it is in the bottom $x\%$ of the cross-sectional distribution), rank-weighted strategies (when $S_{i,t}$ is the cross-sectional rank of the sorting characteristic), and z-score weighted strategies (when $S_{i,t}$ is the z-score of the sorting characteristic). It also embeds both value-weighted and equal-weighted strategies (the former when market cap multiplier $m_{i,t}$ is a stock's market capitalization; the latter when it is one for all stocks).

To facilitate comparison between the empirical results of Section 2 and the theory presented in Section 3, I assume that the signals were uncorrelated normally distributed z-scores. Under this assumption, integrated solutions (i.e., strategies based on composite signals) are exactly equivalent to siloed solutions (i.e., to portfolios of strategies based on the individual signals). This simplifies the analysis, because it allows us to use well known results from portfolio theory. This section shows that this assumption is largely immaterial, as the performance of strategies based on quantile sorts, rank-weighting, and z-score weighting are essentially indistinguishable.

This fact should not, perhaps, be very surprising. Figure 9 shows the relative weights on the long side of long/short strategies constructed by tertile (top 35%) sorting (dotted line), rank-weighting (dashed line), and z-score weighting (solid line). The figure shows that each dollar of the long side of a z-score weighted strategy differs from the quantile portfolio that equal weights the top 35% of stocks by the sorting characteristic by only

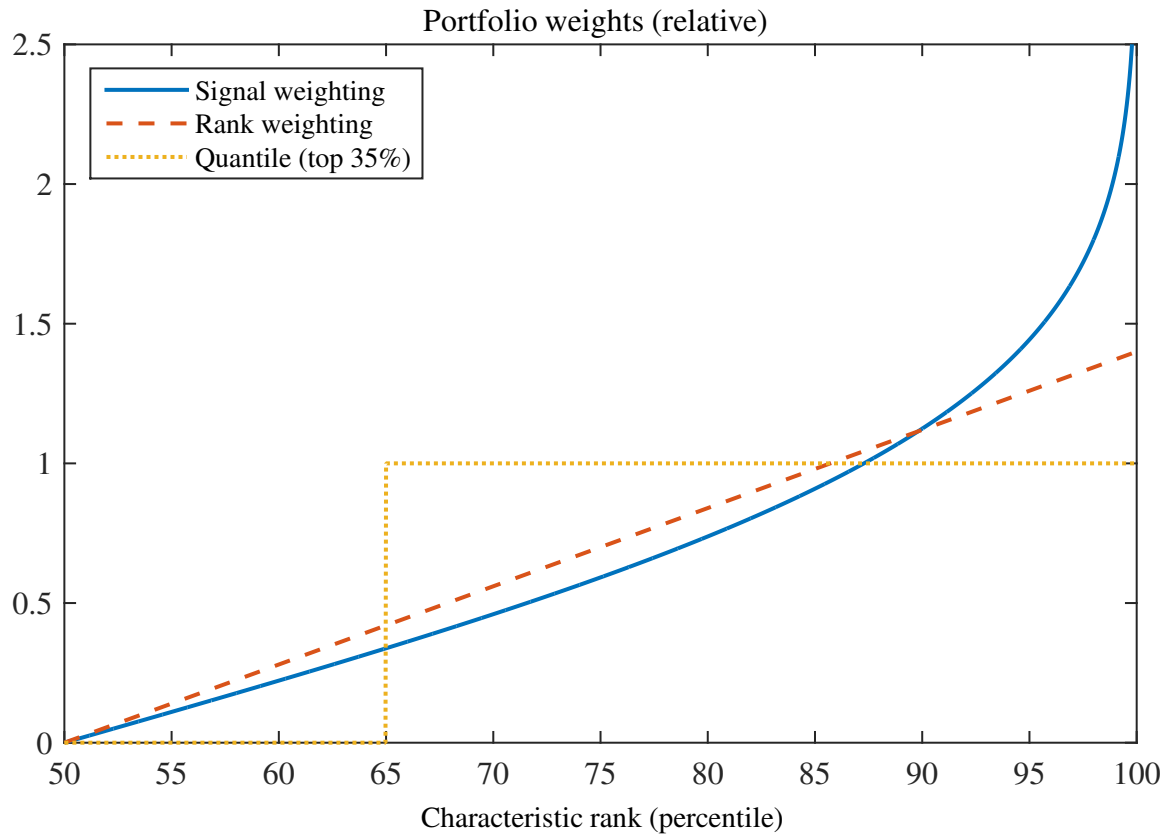

Fig. 9. Relative weights on tertile (top 35%) and signal-weighted strategies. The figure shows the relative weights on individual stocks in the long side of long/short equity strategies, as a function of the cross-sectional rank of the stocks on the sorting characteristic. Weights are given relative to the weights put on stocks in the tertile sorted (top 35%) portfolio (dotted line). The dashed line shows rank-weighted strategies, and the solid line shows z-score weighted strategies.

23 cents. The impact of this relatively modest deviation is holdings is also unclear, as relative to the quantile portfolio, the z-score weighted portfolio overweights the stocks with the most extreme and most marginal characteristic rankings, and underweights those with moderately strong characteristic rankings. The rank-weighted strategies differ even less, deviating from the quantile portfolio's holdings by only 17 cents per dollar.

These similarities are reflected in the performance of strategies constructed by quantile sorting, rank-weighting, and z-score weighting. Table 2 compares the performance of

value-minus-growth (VMG) strategies, based on book-to-market, constructed using the three different methodologies.[5] It also compares strategies that are equal and value weighted. The table provides compelling evidence that the most relevant feature of rank-weighting or z-score weighting is the fact that these weighting schemes weight without respect for capitalization, and thus look like equal weighted, not value weighted, strategies. Similarly, the most relevant feature of the rank and capitalization weighted or z-score and capitalization weighted strategies is the fact that they weight proportional to capitalization, and thus look like value weighted, not equal weighted, strategies.

Panel A regresses the returns to the rank-weighted VMG strategy on to the returns to an equal-weighted, tertile sorted VMG strategy, and onto a rank and capitalization weighted VMG strategy. It shows that the rank-weighted strategy is almost indistinguishable from the simple equal-weighted, tertile strategy, which explains 99.4 percent of its return variation. The rank-weighted strategy and the rank and capitalization weighted strategy are much less correlated, and actually do not even have a significant relation, after controlling for the performance of the equal-weighted, tertile sorted strategy.

Panels B through D tell similar stories. The z-score weighted strategy looks like an equal weighted, tertile strategy, not a z-score and capitalization weighted strategy. The rank and capitalization weighted, and z-score and capitalization weighted, strategies both look like value-weighted, tertile sorted strategies, not like rank-weighted or z-score weighted strategies.

Overall, the evidence is remarkably consistent. The use of quantile sorting, or rank or z-score weighting, has little impact on the performance of the strategies, which always look highly similar. The choice of equal-weighting or value-weighting for the capitalization multiplier, however, has a material impact on the nature of the strategies.

[5]Because of the extreme skew in book-to-market, I use normal z-scores, $z_{i,t} = N^{-1}\left((r_{i,t} - 1/2)/(\max_j\{r_{j,t}\} + 1/2)\right)$ where $r_{i,t}$ is firm i's B/M rank at time t.

Table 2. Rank weighted, normal z-score weighted, and quantile sorted strategy relations

The table reports results from time-series regressions of value-minus-growth (VMG) strategy returns onto the returns of other VMG strategies. Stocks are weighted in proportion to a demeaned signal of book-to-market (quantile sorted using the high and low 35%; rank-weighted; or normal z-score weighted, where $z_{i,t} = N^{-1}\left((r_{i,t} - 1/2)/(\max_j\{r_{j,t}\} + 1/2)\right)$ and $r_{i,t}$ is firm i's B/M rank at time t). Portfolios are rebalanced annualy, at the end of June. Data come from CRSP and Copmustat, and the sample covers July 1963 through December 2014.

	Rank weighted			
Explanatory strategies	(1)	(2)	(3)	(4)
Panel A: y = rank weighted VMG				
α	0.78	0.02	0.57	0.02
	[6.42]	[2.28]	[6.25]	[2.19]
Equal weighted VMG		1.01		1.02
		[320.4]		[236.2]
Rank and cap weighted VMG			0.70	-0.00
			[22.6]	[-0.85]
Adj.-R^2		99.4	45.3	99.4
Panel B: y = Normal z-score weighted VMG				
α	0.82	0.04	0.60	0.04
	[6.50]	[2.16]	[6.28]	[2.10]
Equal weighted VMG		1.05		1.05
		[177.0]		[130.6]
Normal z-score and cap weighted VMG			0.69	-0.00
			[22.3]	[-0.61]
Adj.-R^2		98.1	44.6	98.1
Panel C: y = rank and cap weighted VMG				
α	0.31	0.01	-0.20	-0.02
	[2.63]	[0.66]	[-2.22]	[-0.99]
Value weighted VMG		0.99		0.96
		[166.9]		[126.8]
Rank weighted VMG			0.65	0.05
			[22.6]	[6.81]
Adj.-R^2		97.8	45.3	98.0
Panel D: y = Normal z-score and cap weighted VMG				
α	0.33	0.02	-0.20	-0.02
	[2.70]	[0.97]	[-2.14]	[-0.71]
Value weighted VMG		1.04		0.99
		[132.5]		[101.0]
Signal weighted VMG			0.65	0.06
			[22.3]	[6.84]
Adj.-R^2		96.6	44.6	96.8

B. Densities for t-statistics, signal weighted strategies

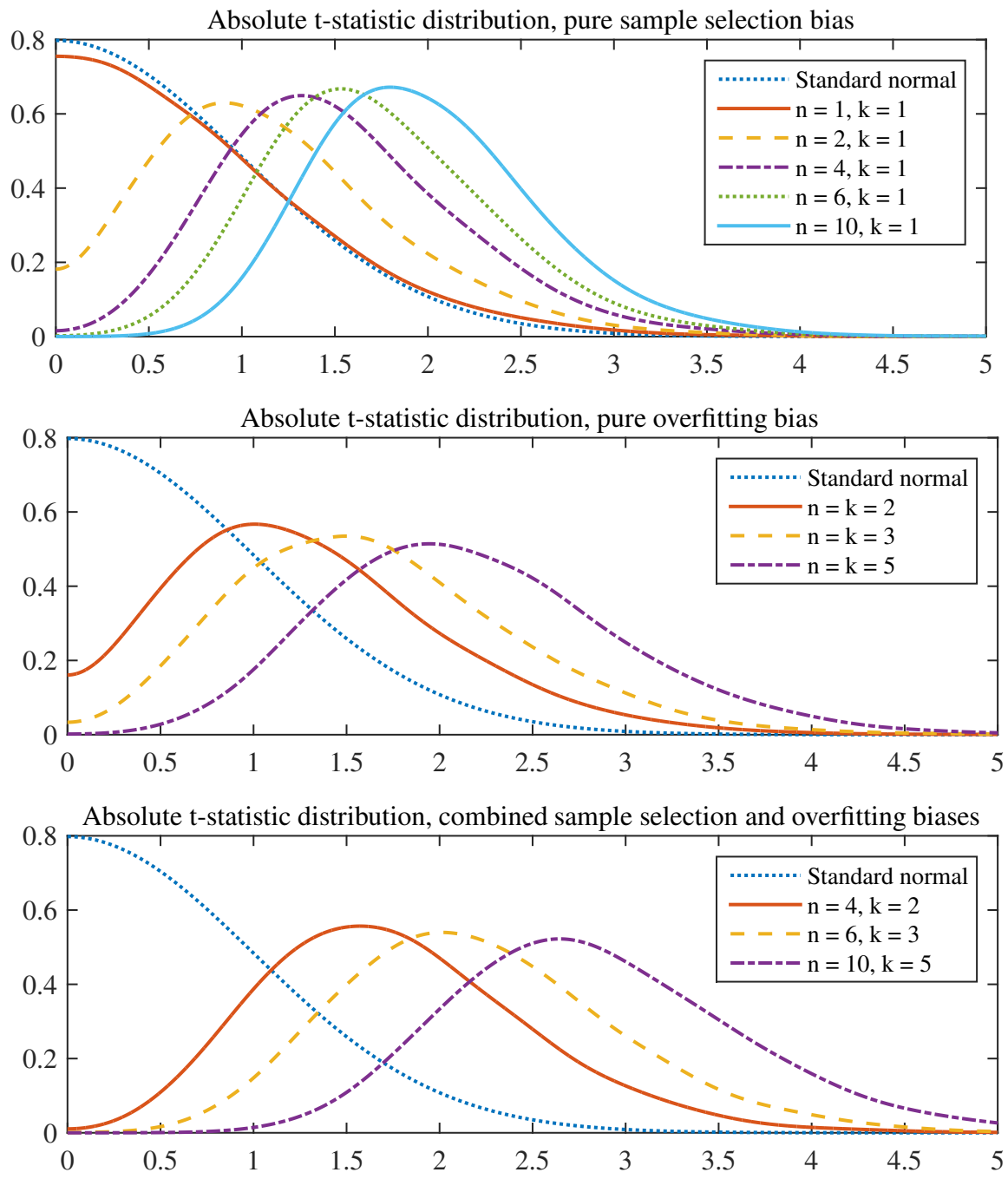

Fig. 10. Empirical t-statistic distribution for best k-of-n strategies, signal-weighted signals. Panel A shows the case of pure selection bias ($k = 1$), for $n \in \{1, 2, 4, 6, 10\}$; Panel B the case of pure overfitting bias ($k = n$), for $n \in \{2, 3, 5\}$; and Panel C the combined case, when $n = 2k$ for $n \in \{2, 3, 5\}$. Distributions are bootstrapped from 10,000 draws of n randomly generated signals, and kernel smoothed with a bandwidth of 0.2. Strategies are signal-and-cap weighted, with stocks held in proportion to both market capitalization and the demeaned signal used for strategy construction, and rebalanced annually, at the end of June. Return and capitalization data come from CRSP. The sample covers July 1993 through December 2014.

C. Best k-of-n critical values, without cap weighting

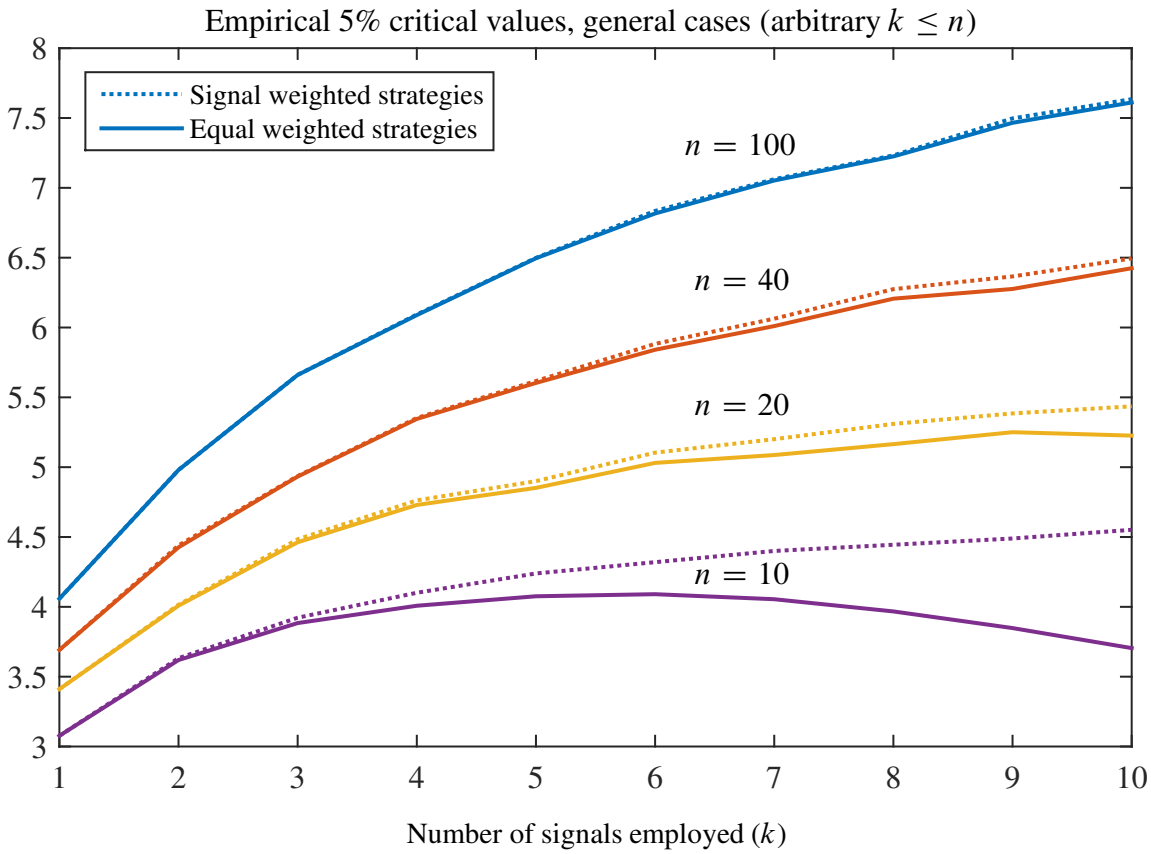

Fig. 11. Five percent critical t-statistics for best k-of-n strategies, without capitalization weighting. The figure shows 5% critical thresholds for strategies selected using a signal constructed by combining the best $k = 1, 2, ..., 10$ performing signals, when the investigator considered $n \in \{10, 20, 40, 100\}$ 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 10,000 sets of n randomly generated signals. Strategies are signal weighted, with stocks are held in direct proportion to the demeaned signal used for strategy construction, and rebalanced annually, at the end of June. Return and capitalization data come from CRSP, and the sample covers July 1993 through December 2014.

D. Critical value approximation

Recall, from equations (3) and (4), that

$$t_{n,k}^{\mathrm{MV}} = \frac{\sum_{i=1}^{k} t_{(n+1-i)}}{\sqrt{k}} \tag{19}$$

$$\left(t_{n,k}^{\mathrm{MVE}}\right)^2 = \sum_{i=1}^{k} t_{(n+1-i)}^2. \tag{20}$$

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

$$\sum_{i=1}^{k} t_{(n+1-i)} = k\lambda\left(t_{(n-k)}\right) + \sum_{i=1}^{k} \left(\left[t_i \middle| t_i > t_{(n-k)}\right] - \lambda\left(t_{(n-k)}\right)\right), \tag{21}$$

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 $t_{(n-k)} \sim N^{-1}\left(\frac{1}{2}(1 + U_{(n-k)})\right)$ and letting $\mu_{n,k} \equiv N^{-1}\left(E\left[\frac{1}{2}\left(1 + U_{(n-k)}\right)\right]\right) = N^{-1}\left(1 - \frac{k+1}{2(n+1)}\right)$ and $\lambda_{n,k} \equiv \lambda(\mu_{n,k})$, given by

$$E\left[k\lambda\left(t_{(n-k)}\right)\right] \approx k\lambda_{n,k} \tag{22}$$

$$\begin{aligned} \mathrm{Var}\left(k\lambda\left(t_{(n-k)}\right)\right) &\approx k^2 \mathrm{Var}\left(\lambda'(x)\big|_{\mu_{n,k}} \times \left(N^{-1}\right)'(x)\big|_{E\left[\frac{1}{2}\left(1+U_{(n-k)}\right)\right]} \times \frac{U_{(n-k)}}{2}\right) \\ &= \frac{k^2(n-k)\left(\lambda_{n,k} - \mu_{n,k}\right)^2}{(k+1)(n+2)}, \end{aligned} \tag{23}$$

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, $U_{(n-k)} \sim B(n-k, k+1)$, 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 (21) is mean zero, converges to normality for large k by the central limit theorem, and has a variance of approximately

$$k\,\mathrm{Var}\left(t_i \middle| t_i > \mu_{n,k}\right) \;=\; k\left(1 + \mu_{n,k}\lambda_{n,k} - \lambda_{n,k}^2\right). \tag{24}$$

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

$$\mu_{t_{n,k}^{\mathrm{MV}}} \;=\; \sqrt{k}\,\lambda_{n,k} \tag{25}$$

$$\sigma^2_{t_{n,k}^{\mathrm{MV}}} \;=\; 1 + \mu_{n,k}\lambda_{n,k} - \lambda_{n,k}^2 + \tfrac{k(n-k)\left(\lambda_{n,k}-\mu_{n,k}\right)^2}{(k+1)(n+2)}. \tag{26}$$

Then

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

which implies

$$t^*_{n,k,p} \;\approx\; \mu_{t_{n,k}^{\mathrm{MV}}} + \sigma_{t_{n,k}^{\mathrm{MV}}} N^{-1}(1-p). \tag{27}$$

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

$$\begin{aligned} E\left[\left(t_{n,k}^{\mathrm{MVE}}\right)^2\right] &= E\left[E\left[\left(t_{n,k}^{\mathrm{MVE}}\right)^2 \,\middle|\, t_{(n-k)}\right]\right] && (28)\\ &\approx E\left[\left(t_{n,k}^{\mathrm{MVE}}\right)^2 \,\middle|\, t_{(n-k)} = \mu_{n,k}\right] && (29)\\ &= kE\left[t_i^2 \,\middle|\, t_i > \mu_{n,k}\right] && (30)\\ &= k\left(1+\mu_{n,k}\lambda_{n,k}\right), && (31) \end{aligned}$$

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

$$\mathrm{Var}\left(\left(t_{n,k}^{\mathrm{MVE}}\right)^2\right) = \mathrm{Var}\left(E\left[\left(t_{n,k}^{\mathrm{MVE}}\right)^2 \,\middle|\, t_{(n-k)}\right]\right) + E\left[\mathrm{Var}\left(\left(t_{n,k}^{\mathrm{MVE}}\right)^2 \,\middle|\, t_{(n-k)}\right)\right]. \quad (32)$$

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

$$\begin{aligned} E\left[\mathrm{Var}\left(\left(t_{n,k}^{\mathrm{MVE}}\right)^2 \,\middle|\, t_{(n-k)}\right)\right] &\approx \mathrm{Var}\left(\left(t_{n,k}^{\mathrm{MVE}}\right)^2 \,\middle|\, t_{(n-k)} = \mu_{n,k}\right)\\ &= k\,\mathrm{Var}\left(t_i^2 \,\middle|\, t_i > \mu_{n,k}\right) && (33)\\ &= k\left(\mu_{n,k}^3\lambda_{n,k} + 3\left(1+\mu_{n,k}\lambda_{n,k}\right) - \left(1+\mu_{n,k}\lambda_{n,k}\right)^2\right), \end{aligned}$$

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 (32), note that

$$\begin{aligned} E\left[\left(t_{n,k}^{\mathrm{MVE}}\right)^2 \,\middle|\, t_{(n-k)}\right] &= E\left[\sum_{i=1}^{k} t_{(n+1-i)}^2 \,\middle|\, t_{(n-k)}\right] \\ &= E\left[\sum_{i=1}^{k} t_i^2 \,\middle|\, t_i > t_{(n-k)}\right] \\ &= k\left(1 + t_{(n-k)}\,\lambda\left(t_{(n-k)}\right)\right). \end{aligned} \tag{34}$$

So

$$\mathrm{Var}\left(E\left[\left(t_{n,k}^{\mathrm{MVE}}\right)^2 \,\middle|\, t_{(n-k)}\right]\right) = k^2 \mathrm{Var}\left(t_{(n-k)}\,\lambda\left(t_{(n-k)}\right)\right) \tag{35}$$

and

$$\begin{aligned} \mathrm{Var}\left(\lambda\left(t_{(n-k)}\right)\right) &\approx \mathrm{Var}\left(\left[x\lambda(x)\right]'\big|_{\mu_{n,k}} \times \left(N^{-1}\right)'(x)\,\big|_{E\left[\frac{1}{2}\left(1+U_{(n-k)}\right)\right]} \times \tfrac{U_{(n-k)}}{2}\right) \\ &= \frac{(n-k)\left(1 + \mu_{n,k}\lambda_{n,k} - \mu_{n,k}^2\right)^2}{(k+1)(n+2)}, \end{aligned} \tag{36}$$

where the last equality follows from the results of equation (23), 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 $\left(t_{n,k}^{\mathrm{MVE}}\right)^2 \overset{approx.}{\sim} N\left(\mu_{\left(t_{n,k}^{\mathrm{MVE}}\right)^2}, \sigma^2_{\left(t_{n,k}^{\mathrm{MVE}}\right)^2}\right)$ where

$$\mu_{\left(t_{n,k}^{\mathrm{MVE}}\right)^2} = k^2\left(1 + \mu_{n,k}\lambda_{n,k}\right) \tag{37}$$

$$\begin{aligned} \sigma^2_{\left(t_{n,k}^{\mathrm{MVE}}\right)^2} &= k\left(\mu_{n,k}^3\lambda_{n,k} + 3\left(1 + \mu_{n,k}\lambda_{n,k}\right) - \left(1 + \mu_{n,k}\lambda_{n,k}\right)^2\right) \\ &\quad + \frac{k^2(n-k)\left(1+\mu_{n,k}\lambda_{n,k}-\mu_{n,k}^2\right)^2}{(k+1)(n+2)}. \end{aligned} \tag{38}$$

Then

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

which implies

$$t_{n,k,p}^{**} \approx \sqrt{\mu_{\left(t_{n,k}^{\mathrm{MVE}}\right)^2} + \sigma_{\left(t_{n,k}^{\mathrm{MVE}}\right)^2} N^{-1}(1-p)}. \tag{39}$$

References

[1] Asness, Cliff, Andrea Frazzini, and Lasse H. Pedersen. 2013. "Quality Minus Junk." AQR working paper.

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

[3] Frazzini, Andrea, and Lasse H. Pedersen. 2013. "Betting against beta." Journal of Financial Economics 111, 1–25.

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

[5] Harvey, Campbell R., Yan Liu, and Heqing Zhu. 2013. "...and the cross-section of expected returns." Working paper.

[6] Lo, Andrew W., and A. Craig MacKinlay. 1990. "Data-snooping biases in tests of asset pricing models." Review of Financial Studies.

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

[8] McLean and Pontiff, 2013, "Does Academic Research Destroy Stock Return Predictability?" Working paper.

[9] 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.