Fund-of-Funds Construction by Statistical Multiple Testing Methods

Michael Wolf*
IEW, University of Zurich
CH-8006 Zurich, Switzerland
mwolf@iew.uzh.ch

Dan Wunderli†
IEW, University of Zurich
CH-8006 Zurich, Switzerland
danwunderli@iew.uzh.ch

September 2009

Abstract

Fund-of-funds (FoF) managers face the task of selecting a (relatively) small number of hedge funds from a large universe of candidate funds. We analyse whether such a selection can be successfully achieved by looking at the track records of the available funds alone, using advanced statistical techniques. In particular, at a given point in time, we determine which funds significantly outperform a given benchmark while, crucially, accounting for the fact that a large number of funds are examined at the same time. This is achieved by employing so-called multiple testing methods. Then, the equal-weighted or the global minimum variance portfolio of the outperforming funds is held for one year, after which the selection process is repeated. When backtesting this strategy on two particular hedge fund universes, we find that the resulting FoF portfolios have attractive return properties compared to the $1/N$ portfolio (that is, simply equal-weighting all the available funds) but also when compared to two investable hedge fund indices.

KEY WORDS: Bootstrap, familywise error rate, fund-of-funds, performance evaluation.

JEL CLASSIFICATION NOS: C12, C14, C22, G11.

*Research supported by the Swiss National Science Foundation (NCCR FINRISK, Module A3).

†Many thanks to Ashok Kaul, Andrea Heuson, Iwan Meier, and Chayawat Ornthalanai for helpful comments. I would also like to thank Eurekahedge Inc. for the kind support and the FMA for organizing the FMA European Conference 2008 in Prague. I gratefully acknowledge financial support from the Swiss Banking Institute.

1 The Challenge

A fund-of-funds (FoF) manager or an institutional investor faces the challenge of selecting a (relatively) small number of 'good' hedge funds from a large universe of candidate funds. We shall address the problem of fund selection from a statistical point of view. The analysis will be based solely on the track records of the individual managers. Arguably, the track record constitutes the single most important piece of information to judge the quality of a fund manager. [1] But making sense of the track records is a non-trivial task.

If we want to answer the question whether a particular fund manager is skilled based on his track record, we can use a statistical test. Such a test declares a fund manager skilled if his alpha with respect to a suitable benchmark is statistically proven to be positive 'beyond a reasonable doubt', say a doubt threshold of 5%. This doubt threshold, say 5%, is denoted by the *significance level* of the test. By design, there is only a small chance then, say 5%, that a lucky manager passes the test, that is, gets wrongly identified as skilled. [2] Importantly, this logic assumes that *only one* manager is tested. If *many* managers are tested at the same time, the small individual doubts accumulate to a large global doubt. In other words, it now becomes very likely that some lucky managers will pass the test. This is undesirable for investment purposes. In general, only skilled managers will continue to outperform, while lucky managers will not.

The following analogy might help illustrate this dilemma. Imagine a person claims to have — some, though not necessarily perfect — extrasensory perception (ESP). A possible test consists of secretly tossing a coin ten times and having the person predict the outcome of each toss. It would then be reasonable to identify the person as possessing ESP if she scores at least nine correct predictions. The logic is that somebody guessing completely at random has a chance of about 1.1% to score at least nine correct predictions. As a result, there is only a small chance that an 'ignorant' person passes the test by chance. [3] But now consider 1,000 persons taking the test at the same time (perhaps because we put out a related job ad) and assume they are all ignorant. One would expect $0.011 \times 1,000 = 11$ persons to pass the test by chance alone, that is, to get lucky. And the probability that at least one person will pass the test by chance alone, if they all guess independently of each other, is $1 - (1 - 0.011)^{1000} = 99.998\%$.

If our goal is to select the skilled managers from a large universe of candidates, we face a similar challenge as Cinderella:

"The good ones into the pot,
The bad ones into the crop."

We want to identify the skilled managers (*"The good ones into the pot"*) but exclude at the same time the lucky managers (*"The bad ones into the crop"*). But, unlike her, we must face the imperfect nature of statistical tests. [4] As a result, naïve testing, without taking the multiple evaluations into account, will allow lucky managers to creep in. This pitfall is rephrased in Grinold and Kahn (2000) in the following words:

> "The fundamental goal of performance analysis is to separate skill from luck. But, how do you tell them apart? In a population of 1,000 investment managers, about 5 percent, or 50, should have exceptional performance by chance alone. None of the successful managers will admit to being lucky; all of the unsuccessful managers will cite bad luck."

2 The Solution

We now discuss the solution to the challenge. In doing so, we first need to introduce some notation.

There are N funds in the universe and the (common) return history comprises T observations. The alpha of a given fund manager with respect to his corresponding benchmark is denoted by α_n, for $n = 1, \ldots, N$. The choice of the appropriate benchmark is up to the FoF manager, not the statistician. For example, the benchmark could simply be the riskfree rate. Or it could be a hedge fund index, comprised of funds that have a similar investment style. More generally, multi-factor benchmarks as in Kosowski et al. (2007) are also possible.

We look at individual hypotheses of the form:

$$H_n : \alpha_n \leq 0 \quad \text{vs.} \quad H_n' : \alpha_n > 0 \ . \tag{1}$$

So for each fund, the null hypothesis corresponds to a non-skilled manager (that is, his alpha is negative or zero), while the alternative corresponds to a skilled manager (that is, his alpha is positive). The two sets of non-skilled (or potentially lucky) managers and skilled managers are denoted by $\mathcal{I}$ and $\mathcal{I}'$, respectively:

$$\mathcal{I} = \{n : \alpha_n \leq 0\} \quad \text{and} \quad \mathcal{I}' = \{n : \alpha_n > 0\} \ .$$

The goal is to make individual decisions about each testing problem (1) while controlling the probability of lucky managers to pass the test by chance. A particular manager n is declared skilled by our statistical method if H_n is rejected in favor of H_n'. Depending on the (unknown) state of nature, there are two possibilities if this happens. On the one hand, if H_n is actually true, we make a mistake in the sense of declaring a non-skilled manager as skilled. Or, in the lingo of the statistician, we make a *false discovery*. On the other hand, if H_n is actually false, we correctly identified a skilled manager as skilled. Or, in the lingo of the statistician, we made a *true discovery*.

2.1 Formal Description of the Solution

For the purpose of this paper, accounting for multiple testing means that we are concerned about the possibility of even one lucky manager to pass the test or, in other words, to make even a single false discovery. [5]

Let F denote the number of false discoveries that our statistical method is going to make. Then the *familywise error rate* (FWE) is defined as the probability of making even one false discovery:

$$\text{FWE} \equiv P\{F > 0\} = P\{\text{Reject at least one } H_n \text{ with } n \in \mathcal{I}\} \ .$$

An appropriate statistical multiple testing method then ensures that this probability lies below some small, prespecified level, say 5% or 10%. Usually this level is denoted by α in the statistical literature but here we shall denote it by δ instead in order to avoid any confusion with the α's of the fund managers. Therefore, the goal is to ensure that:

$$\text{FWE} \leq \delta \ .$$

By limiting the probability that even one lucky manager passes the test, we can in turn be confident that all managers identified by the statistical method are truly skilled.

More specifically, assume $\delta = 10\%$. Then, after applying the method, we can be $1 - \delta$, or 90%, confident that all identified managers are truly skilled. As a result, with a high probability, our statistical FoF portfolio will only consist of skilled managers.

2.2 Implementation of the Solution

Implementing the solution in practice is anything but trivial. A host of statistical problems arise, among others:

- The non-normality of hedge fund returns.
- The time series nature of hedge fund returns.
- The choice of the individual performance measures: raw alpha estimate $\hat{\alpha}$ vs. t-statistic. The t-statistic is obtained by dividing the raw alpha estimate by its estimation uncertainty, which is quantified via a standard error.
- Accounting for the dependency across managers in order to improve the power of the statistical method, that is, its ability to detect skilled managers.

For each fund we compute an estimate of α_n, denoted by $\hat{\alpha}_n$, and a corresponding standard error $\hat{\sigma}_n$. [6] The 'studentized' test statistic for testing H_n vs. H_n' is then given by

$$t_n = \frac{\hat{\alpha}_n}{\hat{\sigma}_n} \ .$$

The funds are ranked according to their test statistics, that is, the fund with the largest t_n statistic is the top fund according to this ranking and so on.

Alternatively, it would be possible to rank the fund managers simply according to their non-studentized test statistics $\hat{\alpha}_n$, that is, according to the 'raw' alpha estimates. While this is actually the more common approach in the mainstream finance media, we consider it misguided. Ranking by the $\hat{\alpha}_n$ does not account for the (wildly) varying risks taken on by the various fund managers. On the other hand, ranking by the t_n does, since a larger risk will be reflected by a larger standard error $\hat{\sigma}_n$. This in the very same spirit as using the Sharpe ratio (that is, a risk-adjusted performance measure) to judge the performance of a fund manager rather than the raw excess return (that is, a not-risk-adjusted performance measure).

How to compute $\hat{\alpha}_n$ and the corresponding standard error $\hat{\sigma}_n$ depends on the given benchmark. A very general setup covering most practical applications are multi-factor benchmarks as in Kosowski et al. (2007). In such cases, $\hat{\alpha}_n$ can be computed from a

standard OLS time series regression, based on the observed fund return and factor data. But care must be taken in computing the standard error $\hat{\sigma}_n$. It would be generally wrong to simply use the standard error provided by the OLS output, since it does not properly account for the time series nature of hedge fund returns (and potentially also some of the factors). Instead one should use a HAC standard error [7] employing kernel estimation techniques; for example, see Andrews (1991) and Andrews and Monahan (1992).

Once the test statistics t_n have been obtained, it is the task of the multiple testing method to compute a cutoff value, denoted by d, from the joint track records of all managers in the investment universe and then declare those managers as skilled for which $t_n > d$. Crucially, this has to be done in a way such that the FWE is controlled. Of course, controlling a multiple testing criterion is only one side of the coin. It could be trivially achieved by never declaring any fund manager as skilled (that is, by choosing $c = \infty$). Naturally, there is also the other side of the coin. At same time, we wish to identify as many skilled managers as possible. So in the lingo of the statistician, we want to employ a multiple testing method with as much *power* as possible. The current state of the art is developed Romano and Wolf (2005) and can be summarized as follows.

It turns out that the ideal critical value d would be given by the $1 - \delta$ quantile of the following random variable:

$$\max_{1 \leq n \leq N} \frac{(\hat{\alpha}_n - \alpha_n)}{\hat{\sigma}_n} \ . \tag{2}$$

Importantly, the value of d is not only determined by the N marginal distributions of the individual statistics $(\hat{\alpha}_n - \alpha_n)/\hat{\sigma}_n$ but also by their cross-dependence structure. Such a procedure is not realistic, nevertheless, since the distribution of the random variable (2) is not known in practice. However, a consistent estimator of d, denoted by $\hat{d}$, can be obtained by a bootstrap method. Namely, $\hat{d}$ is obtained as the $1 - \delta$ quantile of the following random variable:

$$\max_{1 \leq n \leq N} \frac{(\hat{\alpha}^*_n - \hat{\alpha}_n)}{\hat{\sigma}^*_n} \ . \tag{3}$$

To this end, artificial return data are generated by an appropriate time series bootstrap mechanism. The estimator of α_n and its corresponding standard error computed from this artificial data set are denoted by $\hat{\alpha}^*_n$ and $\hat{\sigma}^*_n$, respectively. The algorithm to compute $\hat{\sigma}^*_n$ generally depends on the particular bootstrap mechanism chosen. We refer the interested reader to Romano and Wolf (2005) for the details. The *bona fide* decision rule is then to declare all funds managers as skilled for which $t_n > \hat{d}$.

The price one has to pay for replacing d by $\hat{d}$ is that control of the FWE is replaced by *asymptotic* control of the FWE:

$$\limsup_{T\to\infty} \text{FWE} \leq \delta \ .$$

However, simulation studies show that for practically relevant sample sizes T, the finite-sample control of the FWE is very satisfactory; see Romano and Wolf (2005) and Romano et al. (2008).

Remark 2.1. A key innovation of Romano and Wolf (2005) is to develop a *stepwise* method to detect as many skilled managers as possible. Instead of using a formal algorithm, it can be quite easily described in English. Assume there are $N = 100$ managers under test simultaneously and that 10 of them are detected as skilled using the procedure described above. We are left then with a smaller universe of 90 managers. The 'trick' now is to use the same formal procedure on the remaining smaller universe, which might lead to the detection of some further skilled managers.

The reason is as follows. The individual test statistics t_n will stay the same, of course. However, the critical value $\hat{d}$ in this second step will generally be smaller, since now we are looking at the maximum over 90 statistics, rather than over 100 statistics, and so the resulting $1 - \delta$ quantile will be at most as large but typically strictly smaller. So some further rejections may result. In which case we continue to play the same game in the third step and so on, until no further rejections result any more.

This more powerful stepwise method still provides asymptotic control of the FWE. □

For the empirical analysis of this paper, we use the riskfree rate as the common benchmark for all hedge funds. In this case, the corresponding alpha is simply the expected excess return of the fund (over the riskfree rate). For a given fund, $\hat{\alpha}_n$ is computed as the sample average excess return over the observed investment period. The corresponding standard error $\hat{\sigma}_n$ is a standard HAC standard error employing a kernel estimation technique. In particular, we use the method of Andrews and Monahan (1992), based on the QS (quadratic spectral) kernel.

2.3 Comparison to Related Approaches

Needless to say, we are not the first ones to suggest to carry out hedge fund selection based on the managers' track records. We lack the time and the space to discuss all previously suggested approaches in detail and so limit ourselves to two selected comparisons.

Our method will, with a high probability, only identify skilled managers. As described above, the method works in the following way. Rank the fund managers by a certain performance criterion computed from their respective track records. Then based on the chosen input parameter δ, the method selects an *a priori* random number of the top funds, which are then declared as skilled. In other words, the threshold a manager must pass is actually computed from the joint track records themselves and is therefore stochastic. Knowing the number of funds in the investment universe will not tell us how many funds will end up in the FoF portfolio until we actually jointly examine all the track records.

This is in contrast to some previous approaches that suggest to pick either an *a priori* fixed percentage or or an *a priori* fixed number of the top funds for the FoF portfolio; see Joehri and Leippold (2006) and Gregoriou et al. (2006), respectively. In discussing such approaches, we will focus on the fixed-percentage strategies; the critique would be similar for the fixed-number strategies.

The obvious question is how to pick the percentage *ex ante*? When backtesting the strategy, for a given investment universe and a given investment period, there usually will be a certain percentage leading *ex post* to a very good performance. But there is no universally 'optimal' percentage. The results will vary with the investment universe and/or the investment period. To put it in the context of non-skilled vs. skilled managers and selecting two (overly) extreme scenarios just to make the point: if all managers are non-skilled, the optimal percentage is zero; if all the managers are skilled, the optimal percentage is 100. Knowing from previous published studies that a certain percentage worked well for a certain investment universe during a certain investment period, is not overly helpful to a FoF manager faced with a different universe and a different period. In fact, such information might actually be quite misleading.

On the other hand, the use of our multiple testing methods gives the FoF manager the confidence that for his specific investment universe and investment period, the selected fund managers are all skilled. And such a selection should result in continued attractive future performance for the corresponding FoF portfolio. Whether this indeed is the case will be examined in the next section by means of some backtesting exercises. Importantly, these exercises do not require any hindsight knowledge but instead yield true 'out-of-sample' performances.

3 Investment Universes and Portfolio Construction

We use the CISDM database from `http://wrds.wharton.upenn.edu` and a customized Eurekahedge datafeed from `http://www.eurekahedge.com` to get monthly series of net-of-fees hedge fund returns.

We apply an 'observe ten years–invest one year' strategy with a three-month sell lag, moving at an annual frequency. More specifically, on October 1, of every year y, we feed 117 months of past return data into the multiple testing method. It then detects the statistically significantly skilled fund managers. We then invest in the equal-weighted portfolio of the detected hedge funds from January to December in year $y+1$. Then the procedure repeats, that is, on October 1 of year $y+1$, we already need to decide which hedge funds we want to invest in over the next year $y+2$. Given the annually moving 'observe ten years–invest one year' strategy, six investment periods from year 2000 to 2005 (for CISDM) and from year 2002 to 2007 (for Eurekahedge), respectively, are obtained.

At any given investment point in time, we are only selecting from a certain sub-universe of all funds contained in the respective database (CISDM or Eurekahedge). First, we restrict attention to funds which both have a complete 117-month return history *and* are open to investment at this point. Second, we exclude funds that (overall) lost money over this 117-month period. [8] Third, we exclude all funds that have at least one recorded monthly return exceeding 50% in absolute value. [9] Fourth, to avoid the inclusion of funds which are 'too similar' to each other, we impose that all the pairwise sample correlations over the 117-month period lie below 0.95, so some further funds might have to be excluded. [10]

In addition to the equal-weighted portfolio of the outperforming funds, we build a global minimum variance portfolio (GMV) with the outperforming funds. Specifically, given K outperforming funds over 117 months detected by our multiple testing method, we solve the following optimization problem within each 117 months window

$$\begin{aligned} \min_{\mathbf{w}} \quad & \mathbf{w}'\hat{\Sigma}\mathbf{w} \\ \text{s.t.} \quad & \mathbf{w} \geq \mathbf{0} \\ & \mathbf{w}'\mathbf{1} = \mathbf{1}, \end{aligned} \tag{4}$$

using quadratic programming methods. Since the true covariance matrix Σ is unknown, we estimate it using a suitable shrinkage estimator from the joint track records of the K funds over the last 117 months; see the Appendix for details. Optimization problem (4) returns an optimal weight $\mathbf{w}^*$ for each 117 month window. In the following year, one

then invests in the $\mathbf{w}^*$-weighted portfolio of the outperforming funds. The equal-weighted portfolio is simply the $\mathbf{w}^* = [1/K \dots 1/K]$ weighted portfolio of the outperforming funds. The rebalancing and the three months sell-lag is as before.

As pointed out before, selecting an appropriate benchmark for a given hedge fund is the task of the FoF manager, not of the statistician. Since we are 'ignorant' in this respect, we simply chose the riskfree rate as the universal benchmark. Such a choice certainly appears reasonable and may even be the natural one from certain view points. In practice, the particular riskfree rate we use is from the CRSP Risk Free Rates file. [11]

The multiple testing criterion we employ is the control of the FWE with parameter $\delta = 10\%$. So at any given point in time, we can be 90% confident that all identified managers are truly skilled.

It is then natural to ask whether there is any 'value' in our statistical technique of constructing a FoF. An obvious competitor is the $1/N$ portfolio, that is, the equal-weighted portfolio of all available hedge funds. Recent work by DeMiguel et al. (2009), in the context of building equity portfolios, shows that this simple minded portfolio is actually surprisingly difficult to outperform for statistical methods that construct portfolios based on the past return data. However, in contrast to equity investing, the $1/N$ portfolio is often not feasible for a FoF manager, given the various minimum investments of the individual funds. Hence, it is of interest to see whether statistical FoF portfolio, based on a much smaller investment universe, can do (at least) as well as the $1/N$ portfolio. So for each investment universe, we also include the $1/N$ portfolio in our study.

Remark 3.1. Having a smaller investment universe by applying a multiple testing method rather than investing in all available funds is particularly important when portfolio optimization, such as choosing the global minimum variance portfolio, is used. In this case, the smallest weight (or investment portion) will often be much smaller than the inverse of the number of funds to invest in. So the larger the number of funds, given the various minimum investments, the less feasible such a 'optimized' strategy becomes. □

Furthermore, we consider two investable hedge fund indices for comparison. The HFRX Global Investable Hedge Fund Index is from `www.hedgefundresearch.com` and the CS/Tremont Investable Hedge Fund Index from `www.hedgeindex.com`. Note that the inclusion of these indices somewhat amounts to comparing apples to oranges, since they correspond to investment universes different from both the CISDM and the Eurekahedge databases. Nevertheless it is interesting to see how our statistical FoF portfolios fare against some 'real life' competitors.

3.1 Idealistic Setup

In a first analysis, all hedge funds that have a complete return history of 192 months are part of our chosen investment universes. This is idealistic, since we will never know in January 2000, say, which funds will survive until December 2005 in order to restrict our attention to them. Nevertheless, it is also of interest to compare our statistical FoF portfolio to the $1/N$ portfolio in this context.

Remark 3.2. Constructing investment portfolios based on statistical multiple testing methods, investing in assets which are established as outperforming, is certainly not restricted to the hedge fund industry. More generally, this approach could also be applied to equities, bonds, foreign exchange, etc. The frequency of individual assets 'dying' in such alternative markets will often be much reduced compared to the hedge fund industry, or even (close to) zero. So including the results for a world without dying individual funds/assets is not only of academic interest. □

In a second step, we will make the investment setup more realistic with respect to the characteristics of the hedge fund industry and not using any future knowledge about fund survivorship.

Either way, we always impose a realistic sell lag of three months. That is, we have to decide at October 1 in year $y-1$ which funds to sell at January 1 of year y. For simplicity, we synchronize the buy decisions with the sell decisions. So on October 1 of year $y-1$, the portfolio to be held throughout year y is chosen.

Our CISDM investment universe comprises 97 hedge funds, ranging from January 1990 to December 2005. The Eurekahedge investment universe contains 61 hedge funds over the period January 1992 to December 2007. Restricting attention to the hedge funds actually open to investment throughout the 16-year period further reduces the sizes of the two universes to 91 and 54, respectively.

3.2 Realistic Setup

In the second part of our analysis, we evaluate a completely realistic strategy, both for the FoF and the $1/N$ portfolios as follows. In October of a given year, we take as the investment universe all funds that have a complete 117-month history. As before, we impose a reasonable sell lag of three months and synchronize the sell decisions with the buy decisions.

We then construct both our statistical FoF portfolios and the $1/N$ portfolio and hold them for a year. During that year, some funds might 'die' of course. Not all funds will

generally return all money to the investors. We, therefore, assume a uniform recovery rate of 90% of the investments at the time a fund closes down. [12] The recovered money is then invested in the riskfree rate for the remainder of year. Then we play the same game again next October. So in this way, the size of the investment universe K actually varies over time. Finally, we impose a disinvest-reinvest restriction, as many fund managers are not willing to tolerate a come-and-go-as-you-please behavior of investors. If we disinvest from fund n in October of any year, we are not allowed to reinvest in fund n in any of the following years anymore. [13]

The sizes of the CISDM investment universes only containing open funds are 86, 116, 160, 211, 268, 371 for the years 2000, 2001, ..., 2005, respectively. The sizes of the Eurekahedge investment universes with only open funds are 92, 118, 137, 138, 136, 119 for the years 2002, 2003, ..., 2007, respectively.

3.3 Statistical Significance of Portfolio Outperformance

Of course, we must keep in mind that any performance measures computed from a finite investment period are only sample-based estimates rather than true 'population numbers' (or *parameters* in the lingo of the statistician). So when comparing two portfolios based on a given performance measure, we cannot necessarily conclude that the portfolio with the higher sample-based estimate is indeed better. In other words, we cannot claim any statistical significance based on the sample-based estimates only. To this end, rather, we need to employ a proper statistical test.

Let us focus on the Sharpe ratio which, arguably, is the single most important performance measure. We want to establish whether the true 'underlying' Sharpe ratio of the statistical FoF portfolio is indeed larger than the one of the $1/N$ portfolio in the idealized setup. Denote these two parameters by $\mathrm{SR}_{\mathrm{FoF}}$ and $\mathrm{SR}_{1/N}$, respectively. Further, denote their difference by Δ, that is,

$$\Delta = \mathrm{SR}_{\mathrm{FoF}} - \mathrm{SR}_{1/N} \ .$$

Since we have an *a priori* belief that $\Delta > 0$ and would like to 'verify' this belief by a statistical test, we consider a one-sided test of the kind:

$$H : \Delta \leq 0 \quad \text{vs.} \quad H : \Delta > 0 \ .$$

For both investment universes, the sample-based estimates $\hat{\Delta}$ are indeed positive: for the CISDM universe, we obtain $\hat{\Delta} = 0.37 - 0.32 = 0.05$; for the Eurekahedge universe, we obtain $\hat{\Delta} = 0.37 - 0.27 = 0.10$, as reported in Table 3. But again, this does not 'prove' that the two population Δ's are also positive.

Testing for the difference between two population Sharpe ratios is a non-trival matter. The most commonly used method in the finance literature is the test of Memmel (2003), which is a corrected version of the earlier test of Jobson and Korkie (1981). Unfortunately, this test was derived using the overly strict assumptions of return data that follow a normal distribution and are additionally independent over time. At least one of these two assumptions is generally violated in practice. For hedge fund return data, typically both assumptions are violated. As a result, the test of Memmel (2003) tends to overstate the statistical evidence that is really contained in the observed data. Therefore, since we want to demonstrate that our FoF portfolios outperform the $1/N$ portfolios with respect to the Sharpe ratio, using the test of Memmel (2003) would actually be tempting. However, it would not be correct.

Ledoit and Wolf (2008) propose a bootstrap test that instead yields reliable inference in the presence of non-normal return distributions and time series effects. In other words, it gives a fair appraisal of the statistical significance actually contained in the observed data. Note that their bootstrap test is designed for two-sided hypotheses of the kind

$$H : \Delta = 0 \quad \text{vs.} \quad H' : \Delta \neq 0,$$

but it can be easily modified to apply to the one-sided case as well.

As stated, we believe that the Sharpe ratio is the single most important performance measure. Looking at measures that are not adjusted for the risk taken out by the fund manager, such as the average (excess) return can be quite misleading. Nevertheless, we can apply a statistical test to the difference between average (excess) returns as well. Again, we propose to use a bootstrap test that yields reliable inference in the presence of non-normal return distributions and time series effects. Testing for means is easier than testing for Sharpe ratios. Therefore, the test of Ledoit and Wolf (2008) can be 'simplified' in a straightforward manner to deal with means.

4 Results

The results are summarized in Tables 1 and 2 for the idealistic setup and in Tables 3 and 4 for the realistic setup, respectively. Importantly, all summary statistics are on a monthly basis, that is, they are not annualized. [14] In addition, Figures 1 and 2 provide some graphical representation of the various return distributions.

4.1 Idealistic Setup

First, we report the number of hedge funds making up the statistical FoF portfolio in each of the six annual investment periods. For the CISDM portfolio this number varies between 3 and 9, compared to a universe size of 91. For the Eureka portfolio, this number varies between 1 and 5, compared to a universe size of 54. The size of the HFRX index varies over time, always being larger than 60. The size of the CS/Tremont index is 60.

Second, we report the mean of the annual excess log returns over the six annual investment periods. We find that for both investment universes (and their slightly different respective investment periods), the statistical FoF portfolios yield a lower excess return than the $1/N$ portfolio. However, these differences are not statistically significant, as reported in Table 2.

Third, we report the mean of the 'raw' log annual returns (that is, not in excess of the riskfree rate). Not surprisingly, the comparisons are qualitatively very similar to the ones for the excess returns.

Fourth, we report the Sharpe ratios of the monthly log excess returns. As already stated, for both investment universes, our statistical FoF portfolios have a (somewhat) smaller excess return and a (much) smaller portfolio size than the $1/N$ portfolio. Typically, one would expect smaller portfolios to have less favorable Sharpe ratios than larger ones due to diversification effects. However, the opposite is the case for both investment universes, with the differences being rather large at times. This is especially remarkable in case of the Eurekahedge universe where the size of the statistical FoF portfolios ranges from 1 to 5. Statistical significance at the 10% level is only achieved in one case, though: namely for the EW-FoF portfolio with the CISDM data.

Fifth, we report the maximum drawdown over the out-of-sample investment period of $6 \cdot 12 = 72$ months. Again, for both investment universes, the statistical FoF portfolios outperform the $1/N$ portfolio, adding further evidence to the claim that multiple testing technique successfully identifies a small number of skilled managers from the large investment pool.

The boxplots in Figure 1 clearly show that the $1/N$ portfolio, despite its larger universe size, yields returns that are much more variable compared to the two statistical portfolios. In addition, portfolio optimization appears successful in the sense that the returns of GMV-FoF are somewhat less variable compared to EW-FoF.

We finally note that the statistical portfolios generally compare favorably to the investable indices as well.

4.2 Realistic Setup

First, we report the number of hedge funds making up the statistical FoF portfolio in each of the six annual investment periods. We observe that the sizes of the CISDM FoF portfolios vary between 10 and 14. The Eureka FoF portfolios contain between 9 and 21 funds. The size of the HFRX index varies over time, always being larger than 60. The size of the CS/Tremont index is 60.

Second, we report the mean of the monthly excess log returns over the six annual investment periods. We see again that the mean excess monthly returns are lower than their $1/N$ counterparts. However, these differences are not statistically significant; see Table 4.

Third, we report the mean of the 'raw' log monthly returns (that is, not in excess of the riskfree rate). Not surprisingly, the comparisons are qualitatively very similar to the ones for the excess returns.

Fourth, we report the Sharpe ratios of the monthly log excess returns. As before, the statistical portfolios yield consistently higher Sharpe rations compared to the $1/N$ portfolio, though not at a level of statistical significance.

Fifth, we report the maximum drawdown over the out-of-sample investment period of $6 \cdot 12 = 72$ months. Again, for both investment universes, the statistical FoF portfolios outperform the $1/N$ portfolio, with the differences being rather large. In fact, for both universes, the $1/N$ portfolio has the worst drawdown of all five portfolios.

The boxplots in Figure 2 clearly show that the $1/N$ portfolio, despite its larger universe size, yields returns that are much more variable compared to the two statistical portfolios. In addition, portfolio optimization appears successful in the sense that the returns of GMV-FoF are somewhat less variable compared to EW-FoF.

We finally note that the statistical portfolios generally compare favorably to the investable indices as well.

Remark 4.1. We generally fail to find statistical significance when testing for outperformance. This may not be surprising, given that it is notoriously difficult to find statistical significance in small samples of noisy financial returns (our out-of-sample period only comprises 72 months). There is, nevertheless, a clear and strong pattern. For each performance criterion (average excess return, Sharpe ratio, or maximum drawdown), there is a total of eight comparison cases (two setups, two data sets, and two statistical portfolios). In all eight cases, the story is always the same: the statistical portfolio yields a lower average excess return but outperforms the $1/N$ portfolio both in terms of the Sharpe ratio and the maximum drawdown. The latter two criteria are probably

more relevant, as most FoF managers promote their ability to manage the risk in their portfolios.

We can also ask the question whether portfolio optimization via the GMV portfolio yields further benefits. Here the results are a bit mixed, but they suggest that the general answer may be yes. In terms of the Sharpe ratio and maximum drawdown, the GMV statistical portfolio does better than the equal-weighted statistical portfolio in three cases (both data sets in the idealistic setup and the Eurekahedge data set in the realistic setup) and worse in one case (the CISDM data set in the realistic setup). In addition, the gains in the cases of outperformance are larger than the losses in the case of underperformance. □

5 Conclusions

We have studied whether it is possible to construct hedge fund portfolios with attractive return properties based on the past track records of all managers in the investment universe alone. Importantly, such a strategy must not rely on any hindsight knowledge, say about which fixed percentage of top managers for a given investment universe and investment period would have worked well.

Our approach consists of comparing each manager to a given benchmark (which could be common or be allowed to vary with managers) and then to determine which managers statistically outperform their benchmark. Such managers are deemed 'skilled' and we simply go on to hold an equal-weighted or a global-minimum-variance portfolio of all skilled managers as our FoF portfolio. This process is repeated, and the portfolios thus updated, every year.

Crucially, in determining which managers statistically outperform their benchmark, one must take into account that a large number of managers are examined at the same time. In other words, one must account for the problem of multiple comparisons (of managers against benchmark). We do this by employing some state-of-the-art statistical multiple testing methods. These methods take the non-normal return distributions and time series nature of hedge fund returns into account to properly control the chance of non-skilled managers creeping into our FoF portfolio. On the other hand, these methods are also optimized with respect to detecting as many skilled managers as possible in order to build a well-diversified portfolio.

We backtested this strategy (without using any hindsight knowledge) on two hedge fund universes. When comparing the performance of the statistical FoF portfolios to

their most natural competitor, namely the $1/N$ portfolio, we found that they deliver consistent improvements both in terms of the Sharpe ratio and the maximum monthly drawdown. The return properties are also attractive when compared to two investable hedge fund indices (based on different investment universes).

While traditional approaches to construct FoFs, such as due diligence, will remain vital, we believe that statistical selection techniques based on the past track records alone can be an attractive (and cost efficient) alternative method. Of course, there is no reason not to combine these two approaches. Indeed, while clearly beyond the scope of this paper, the combination of more complex traditional approaches with statistical selection techniques might well result in the best of both worlds.

End Notes

1. To be sure, there may be other pieces of information as well, such as the general background of the manager, his investment philosophy, the size and location of his office, etc. However, such factors are not easily quantifiable and/or available and so they will be left out for the statistical analysis.

2. Imposing a significance level of 0% is not possible, as it would imply that no manager, based on a finite track record, could ever be found skilled, no matter how impressive his track record may be.

3. Again, if we did not allow for a small chance of an ignorant person passing the test, based on a finite number of tosses, nobody could ever be declared as having ESP even if she predicts all outcomes correctly.

4. Cinderella enjoyed the help of pigeons who could perfectly tell whether a particular lentil was 'good' or 'bad'.

5. Put in the context of Cinderella, we do not want even one bad one ending up in the pot.

6. The standard error $\hat{\sigma}_n$ is an estimate of the unknown standard deviation of $\hat{\alpha}_n$.

7. HAC stands for 'heteroskedasticity and autocorrelation consistent'.

8. Since we are benchmarking against the risk-free rate always, no fund manager that lost money overall could possibly be considered outperforming.

9. The motivation here is two-fold. On the one hand, such recorded returns might simply correspond to data-entry mistakes. On the other hand, even if such returns are true, they may have a large impact on the data analysis because of their undue effect on sample means, sample standard deviations, and sample Sharpe ratios.

10. The motivation here is that sometimes 'basically the same fund' can appear under slightly different names. We implicitly take the stance that the FoF manager would only want to invest in one of such funds.

11. We employ the average rate of *ask* and *bid*.

12. Of course, recovery rates vary in practice. But this additional knowledge is not available to us. So to impose a fixed 'average rate' appears the best feasible solution.

13. The results do not change much if this disinvest-reinvest restriction is not imposed. For the sake of brevity, the results without this restriction are not reported.

14. While annualizing (excess) returns is straightforward, annualzing Sharpe ratios is not. The usual method of multiplying the monthly Sharpe rations by $\sqrt{12}$ is misleading for hedge funds due to the autocorrelation of the returns over time; see Lo (2002).

References

Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, 59:817–858.

Andrews, D. W. K. and Monahan, J. C. (1992). An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator. *Econometrica*, 60:953–966.

DeMiguel, V., Garlappi, L., and Uppal, R. (2009). Optimal versus naive diversification: How inefficient is the $1/N$ portfolio strategy? *Review of Financial Studies*, 22:1915–1953.

Gregoriou, G. N., Hübner, G., Papageorgiou, N., and Rouah, F. (2006). Simple hedge fund strategies as an alternative to funds of funds: evidence from large-cap funds. In Gregoriou, G. N., editor, *Funds of Hedge Funds*, Quantitative Finance Series, pages 117–131. Elsevier.

Grinold, R. C. and Kahn, R. N. (2000). *Active Portfolio Management*. McGraw-Hill, New York, second edition.

Jobson, J. D. and Korkie, B. M. (1981). Performance hypothesis testing with the Sharpe and Treynor measures. *Journal of Finance*, 36:889–908.

Joehri, S. and Leippold, M. (2006). Quantitative hedge fund selection for funds of funds. In Gregoriou, G. N., editor, *Funds of Hedge Funds*, Quantitative Finance Series, pages 433–454. Elsevier.

Kosowski, R., Naik, N., and Teo, M. (2007). Do hedge funds deliver alpha? a Bayesian and bootstrap analysis. *Journal of Financial Economics*, 84:229–264.

Ledoit, O. and Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, 10(5):603–621.

Ledoit, O. and Wolf, M. (2004). Honey, I shrunk the sample covariance matrix. *Journal of Portfolio Management*, 30(4):110–119.

Ledoit, O. and Wolf, M. (2008). Robust performance hypothesis testing with the Sharpe ratio. *Journal of Empirical Finance*, 15:850–859.

Lo, A. (2002). The statistics of Sharpe ratios. *Financial Analyst Journal*, 58:36–42.

Memmel, C. (2003). Performance hypothesis testing with the Sharpe Ratio. *Finance Letters*, 1:21–23.

Romano, J. P., Shaikh, A. M., and Wolf, M. (2008). Formalized data snooping based on generalized error rates. *Econometric Theory*, 24(2):404–447.

Romano, J. P. and Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282.

A Appendix

A.1 New Shrinkage Estimator for Σ

When estimating a covariance matrix based on (limited) past track records, one should not use the sample covariance matrix. This is especially true when the estimated covariance matrix is used for purposes of portfolio optimization. The intuitive reason is that the optimizer will latch on to the large estimation error contained in the sample covariance matrix and produce very unstable portfolios that often yield poor out-of-sample performance. This important point is discussed by Ledoit and Wolf (2003, 2004) who also offer a remedy. Namely, shrink the sample covariance matrix to a highly structured estimator, called the *shrinkage target*. Such an estimator will be biased, unlike the sample covariance matrix, but in return contain very little estimation error. Combining the two estimators via shrinkage will result in an optimal bias-variance trade-off.

Ledoit and Wolf (2003, 2004) suggest shrinkage targets for a universe of stocks: the single-factor model and the single-correlation model. But targets have common feature: the diagonal of the matrix is the same as the diagonal of the sample covariance matrix. As a result, only the sample covariances get shrunken/modified but not the sample variances.

We feel that such an approach is sub-optimal when dealing with hedge funds instead of stocks. Due to the wildly varying amounts of risk taken on by the various funds, already the differences between the sample variances will be overstated. It, therefore, appears useful to shrink the sample variances in addition to the sample covariances.

Therefore, we propose the *two-parameter* model as a shrinkage target. It has one common variance and one common covariance. The estimation of these two parameters is straightforward. One simply takes the average of all sample variances and the average of all sample covariances, respectively. One then is left to find a formula for the optimal shrinkage intensity. The general methodology is outlined in Ledoit and Wolf (2003) and the details are left to the reader. Computer code in the Matlab language can be downloaded for free from the following website: `http://www.iew.uzh/chairs/wolf`.

A.2 Tables and Figures

Table 1: Performance of Portfolios: Idealistic Setup

	# of hedge funds in each of the 6 years	average exc. return	average return	Sharpe ratio	maximum drawdown
CISDM data, investment period: 01/2000–12/2005.					
EW-FoF	9, 9, 3, 7, 5, 8	0.38%	0.60%	0.28	−4.22%
GMV-FoF	9, 9, 3, 7, 5, 8	0.42%	0.61%	0.59	−1.47%
$1/N$	91	0.51%	0.73%	0.20	−10.02%
HFRX Global	> 60	0.39%	0.64%	0.28	−3.92%
CS/Tremont	60	0.38%	0.60%	0.48	−2.06%
Eurekahedge data, investment period: 01/2002–12/2007.					
EW-FoF	1, 1, 1, 3, 5, 5	0.40%	0.63%	0.57	−1.89%
GMV-FoF	1, 1, 1, 3, 5, 5	0.38%	0.60%	0.64	−0.56%
$1/N$	54	0.64%	0.86%	0.31	−7.52%
HFRX Global	> 60	0.27%	0.49%	0.23	−3.57%
CS/Tremont	60	0.35%	0.57%	0.40	−2.68%

Table 2: Statistical Significance of Outperformance: Idealistic Setup

	Alternative hypothesis	i='CISDM'	i='Eureka'
j='mean excess return'	$\mu^{\text{exc}}_{\text{EW-FoF}} < \mu^{\text{exc}}_{1/N}$	$p = 0.35$	$p = 0.18$
j='Sharpe ratio'	$SR_{1/N} < SR_{\text{EW-FoF}}$	$p = 0.36$	$p = 0.09$
j='mean excess return'	$\mu^{\text{exc}}_{\text{GMV-FoF}} < \mu^{\text{exc}}_{1/N}$	$p = 0.38$	$p = 0.17$
j='Sharpe ratio'	$SR_{1/N} < SR_{\text{GMV-FoF}}$	$p = 0.20$	$p = 0.11$

Note: If a p-value is smaller than α, then the data supports the alternative hypothesis at significance level α.

Table 3: Performance of Portfolios: Realistic Setup

	# of hedge funds in each of the 6 years	average exc. return	average return	Sharpe ratio	maximum drawdown
CISDM data, investment period: 01/2000–12/2005.					
EW-FoF	10, 14, 13, 14, 10, 11	0.36%	0.58%	0.37	−1.83%
GMV-FoF	10, 14, 13, 14, 10, 11	0.20%	0.41%	0.33	−3.66%
$1/N$	86,116,160,211,268,371	0.54%	0.76%	0.32	−5.62%
HFRX Global	> 60	0.39%	0.61%	0.28	−3.92%
CS/Tremont	60	0.38%	0.60%	0.48	−2.06%
Eurekahedge data, investment period: 01/2002–12/2007.					
EW-FoF	18, 21, 21, 21, 10, 9	0.26%	0.48%	0.37	−3.55%
GMV-FoF	18, 21, 21, 21, 10, 9	0.30%	0.53%	0.67	−0.60%
$1/N$	92,118,137,138,136,119	0.46%	0.68%	0.27	−5.73%
HFRX Global	> 60	0.27%	0.49%	0.23	−3.57%
CS/Tremont	60	0.35%	0.57%	0.40	−2.68%

Table 4: Statistical Significance of Outperformance: Realistic Setup

	Alternative hypothesis	i='CISDM'	i='Eureka'
j='mean excess return'	$\mu^{\text{exc}}_{\text{EW-FoF}} < \mu^{\text{exc}}_{1/N}$	$p = 0.27$	$p = 0.17$
j='Sharpe ratio'	$SR_{1/N} < SR_{\text{EW-FoF}}$	$p = 0.34$	$p = 0.33$
j='mean excess return'	$\mu^{\text{exc}}_{\text{GMV-FoF}} < \mu^{\text{exc}}_{1/N}$	$p = 0.11$	$p = 0.26$
j='Sharpe ratio'	$SR_{1/N} < SR_{\text{GMV-FoF}}$	$p = 0.54$	$p = 0.11$

Note: If a p-value is smaller than α, then the data supports the alternative hypothesis at significance level α.

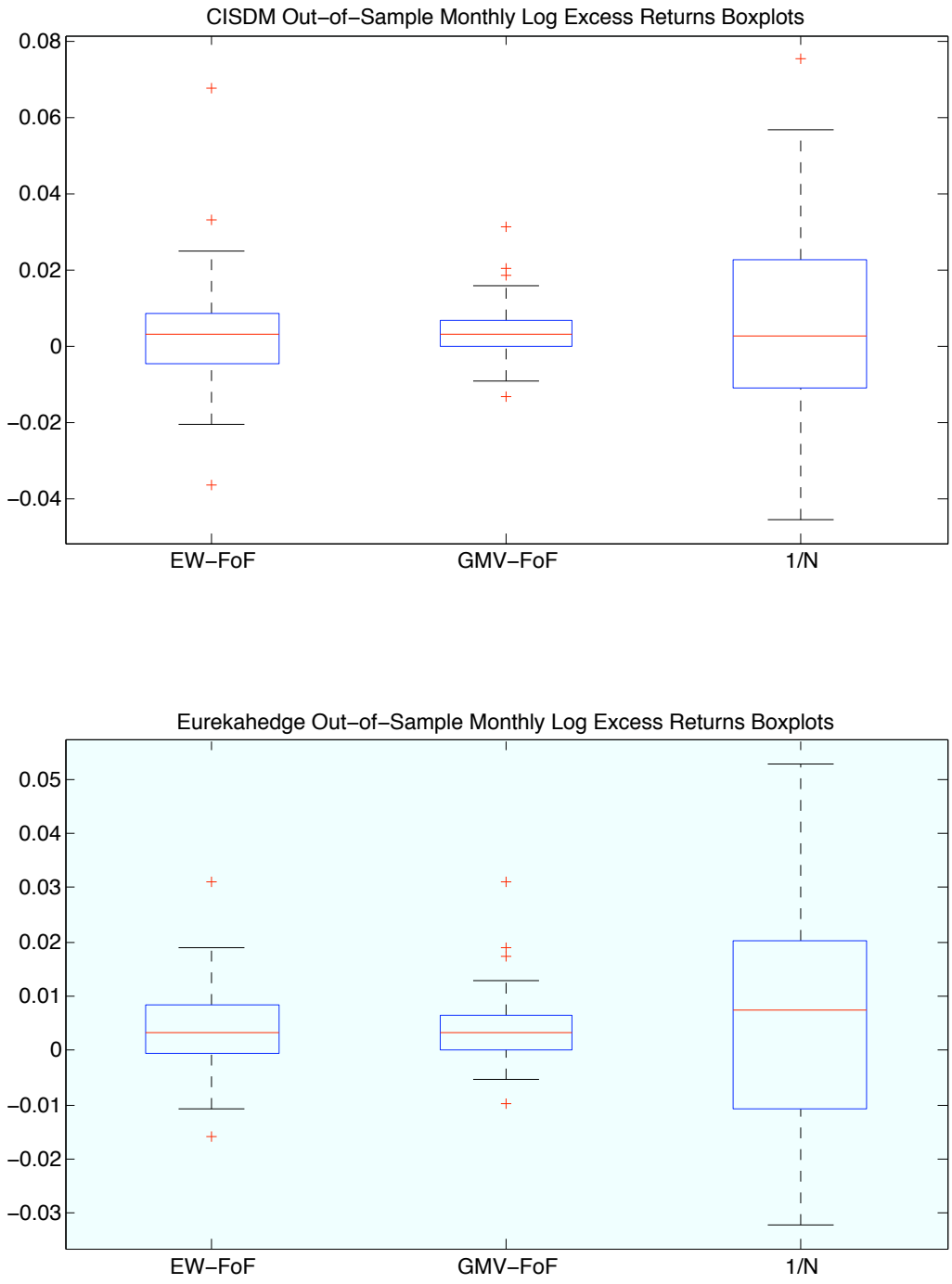

Figure 1: Box Plots of Out-of-Sample Log Excess Returns: Idealistic Setup

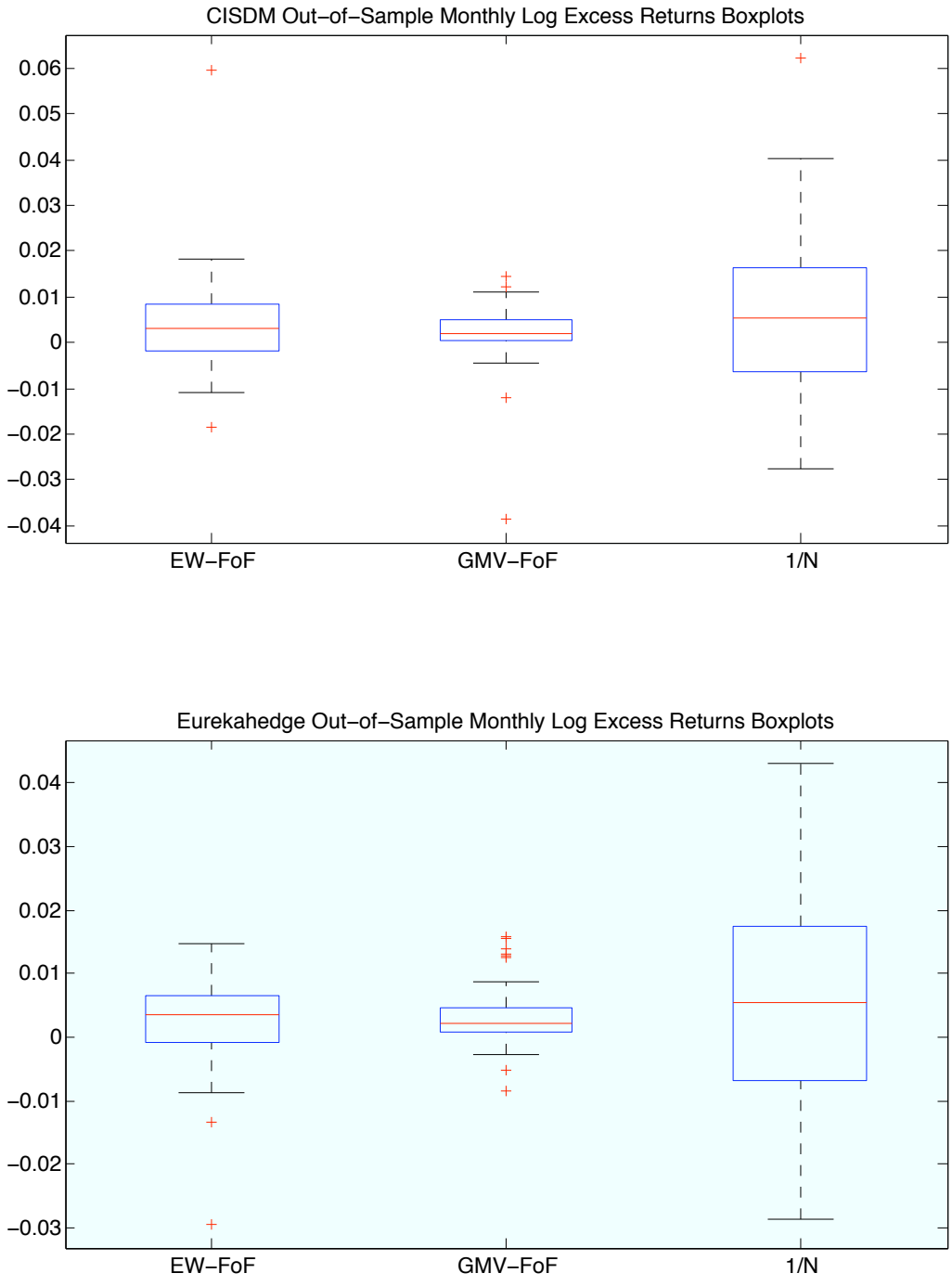

Figure 2: Box Plots of Out-of-Sample Log Excess Returns: Realistic Setup