University of Zurich

Department of Economics

Working Paper Series

ISSN 1664-7041 (print)
ISSN 1664-705X (online)

Working Paper No. 305

Robust Performance Hypothesis Testing with Smooth Functions of Population Moments

Olivier Ledoit and Michael Wolf

October 2018

Robust Performance Hypothesis Testing with Smooth Functions of Population Moments

Olivier Ledoit
Department of Economics
University of Zurich
CH-8032 Zurich, Switzerland
olivier.ledoit@econ.uzh.ch

Michael Wolf
Department of Economics
University of Zurich
CH-8032 Zurich, Switzerland
michael.wolf@econ.uzh.ch

October 2018

Abstract

Applied researchers often want to make inference for the difference of a given performance measure for two investment strategies. In this paper, we consider the class of performance measures that are smooth functions of population means of the underlying returns; this class is very rich and contains many performance measures of practical interest (such as the Sharpe ratio and the variance). Unfortunately, many of the inference procedures that have been suggested previously in the applied literature make unreasonable assumptions that do not apply to real-life return data, such as normality and independence over time. We will discuss inference procedures that are asymptotically valid under very general conditions, allowing for heavy tails and time dependence in the return data. In particular, we will promote a studentized time series bootstrap procedure. A simulation study demonstrates the improved finite-sample performance compared to existing procedures. Applications to real data are also provided.

KEY WORDS: Bootstrap, HAC inference, kurtosis, Sharpe ratio, sknewness, variance.

JEL CLASSIFICATION NOS: C12, C14, C22.

1 Introduction

Much applied financial research is concerned with the evaluation of investment strategies (such as stocks, portfolios, mutual funds, hedge funds, and technical trading rules). The single most relevant performance measure, arguably, is still the Sharpe ratio, introduced by Sharpe (1966).

There exist a host of other performance measures, too many to list them all in this paper. For purposes of tractability, we shall restrict attention to performance measures that can be expressed as smooth functions of population moments of the returns of a given investment strategy. Note here that the returns can be raw returns or they can be returns in excess of another strategy.

Allow us to use the Sharpe ratio as the perfect case in point: it is defined as the ratio of the mean over the standard deviation of the returns of an investment strategy in excess of the riskfree rate. As such, the Sharpe ratio can be expressed as a smooth function of two population moments: the first and the second moment; see Example 2.1.

Other performance measures that fall in the class of smooth functions of population moments are the variance, the mean, the skewness, and the kurtosis; see Examples 2.2–2.5.

The variance is a relevant performance measure for investment strategies that aim to implement the global minimum variance (GMV) portfolio. Such strategies are becoming ever more popular, among other reasons, because they do not require to estimate mean returns; for example, see Jagannathan and Ma (2003), Elton et al. (2006), Kempf and Memmel (2006), Garlappi et al. (2007), DeMiguel et al. (2009a), DeMiguel et al. (2009b), Frahm and Memmel (2010), Güttler and Trübenbach (2011), Scherer (2011), and Candelon et al. (2012).

The mean is rarely used on its own by sophisticated finance practitioners, since it does not account for the volatility (or risk) of an investment strategy. On the other hand, it is still a widely-used performance measure in the mainstream financial media. It can also be considered to be an appropriate performance measure for an investment strategy with (approximately) known volatility, such as money market funds.

In addition, it has been shown that investors seek portfolios that exhibit high positive skewness and low kurtosis; for example, see Harvey and Siddique (2000), Dittmar (2002), Patton (2004), and Mitton and Vorkink (2007).

The general scenario that we shall consider is as follows. There are two investment strategies under consideration and a particular performance measure has been chosen. Can it be established statistically that one strategy outperforms the other? Crucially, in designing a proper statistical inference procedure, one must account for two well-established stylized facts of financial returns: heavy tails and dependence over time.

Unfortunately, many inference procedures that have been suggested in the finance literature to compare two investment strategies with respect to a chosen performance measure are based on the assumption that returns are normally distributed and independent over time. Perhaps the most prominent example is the test for the equality of two Sharpe ratios suggested by Jobson and Korkie (1981) and its corrected version by Memmel (2003). Such inference

procedures are not reliable in the context of real-life financial data; see Section 4.

In this paper, we discuss inference procedures that are more generally valid, allowing for heavy tails and dependence over time of the returns. One possibility is to use normal-theory based on HAC standard errors. Such an approach works asymptotically but does not always have satisfactory properties in finite samples. As an improved alternative, we suggest a studentized time series bootstrap procedure. We would like to point out that much of this paper is based on the earlier works of Ledoit and Wolf (2008, 2011).

2 The General Problem

We use notation going back to Jobson and Korkie (1981) and Memmel (2003). There are two investment strategies i and n whose excess returns over a given benchmark at time t are r_{ti} and r_{tn}, respectively. The benchmark depends on the application at hand. The two most salient cases are zero (resulting in raw returns) and the riskfree rate.

A total of T return pairs $(r_{1i}, r_{1n})', \ldots, (r_{Ti}, r_{Tn})'$ are observed. It is assumed that these observations constitute a strictly stationary time series so that, in particular, the bivariate return distribution does not change over time. This distribution has mean vector μ and covariance matrix Σ given by

$$\mu := \begin{pmatrix} \mu_i \\ \mu_n \end{pmatrix} \quad \text{and} \quad \Sigma := \begin{pmatrix} \sigma_i^2 & \sigma_{in} \\ \sigma_{in} & \sigma_n^2 \end{pmatrix} .$$

Crucially, we dot assume the distribution to be normal, nor do we assume that returns $r_t := (r_{t1}, r_{tn})'$ are independent over time.

The parameter of interest is

$$\Delta := \theta_i - \theta_n \ , \tag{2.1}$$

where θ is a given performance measure. Hence, θ_i is the performance measure for the first strategy and θ_n is the performance measure for the second strategy.

We are interested in testing

$$H_0 : \Delta = 0 \quad \text{vs.} \quad H_1 : \Delta \neq 0 \ .$$

We consider the class of performance measures θ that can be expressed as a smooth function of a finite number of population moments. In particular, for $l = i, n$, let

$$\nu_l^{(m)} := \mathbb{E}(r_l^m)$$

denote the (uncentered) mth population moment of the returns of strategy l. Then, for $l = i, n$, we assume that θ_l can be expressed as

$$\theta_l = h\big(\nu_l^{(1)}, \ldots, \nu_l^{(M)}\big) \ ,$$

where $M \geq 1$ is an integer and $h : \mathbb{R}^M \to \mathbb{R}$ is a smooth function (in the sense of being one time continuously differentiable).

Example 2.1 (Sharpe ratio)**.** In this case,

$$\theta_l := \frac{\mu_l}{\sigma_l} = h\big(\nu_l^{(1)}, \nu_l^{(2)}\big),$$

where

$$h(a,b) := \frac{a}{\sqrt{b-a^2}} \ .$$

Example 2.2 (Variance)**.** As explained by Ledoit and Wolf (2011), it is advantageous to consider the log-variance, instead of the variance itself, so that

$$\theta_l := \log(\sigma_l^2) = h\big(\nu_l^{(1)}, \nu_l^{(2)}\big),$$

where

$$h(a,b) := \log(b-a^2) \ .$$

Example 2.3 (Mean)**.** In this case,

$$\theta_l := \mu_l = h\big(\nu_l^{(1)}\big),$$

where

$$h(a) := a \ .$$

Example 2.4 (Skewness)**.** In this case,

$$\theta_l := \frac{\mathbb{E}\big[(r_l-\mu_l)^3\big]}{\sigma_l^3} = h\big(\nu_l^{(1)}, \nu_l^{(2)}, \nu_l^{(3)}\big),$$

where

$$h(a,b,c) := \frac{2a^3+c-3ab}{(b-a^2)^{1.5}} \ .$$

Example 2.5 (Kurtosis)**.** In this case,

$$\theta_l := \frac{\mathbb{E}\big[(r_l-\mu_l)^4\big]}{\sigma_l^4} - 3 = h\big(\nu_l^{(1)}, \nu_l^{(2)}, \nu_l^{(3)}, \nu_l^{(4)}\big),$$

where

$$h(a,b,c,d) := \frac{-3a^4+d-4ac+6a^2b}{(b-a^2)^2} - 3 \ .$$

(Note here the usual subtraction of three in the definition of the kurtosis, so that the kurtosis of a normally-distributed random variable is zero.)

For $l = i, n$, let $\nu_l' := (\nu_l^{(1)}, \ldots, \nu_l^{(M)})$. Furthermore, let $\nu' := (\nu_i', \nu_n')$. Then the parameter of interest Δ in (2.1) can be written as

$$\Delta := f(\nu) = h(\nu_i) - h(\nu_n) = \theta_i - \theta_n \ ,$$

so that $f : \mathbb{R}^{2M} \to \mathbb{R}$ is also a smooth function, defined as

$$f(a_1, \ldots, a_M, b_1, \ldots, b_M) := h(a_1, \ldots, a_M) - h(b_1, \ldots, b_M) \ . \tag{2.2}$$

For $l = i, n$, denote the (uncentered) mth sample moment of the observed returns by

$$\hat{\nu}_l^{(m)} := \frac{1}{T} \sum_{t=1}^{T} r_{tl}^m \ .$$

Then the estimator of the parameter of interest, Δ, is given by

$$\hat{\Delta} := \hat{\theta}_i - \hat{\theta}_n \ , \tag{2.3}$$

where

$$\hat{\theta}_l := h\big(\hat{\nu}_l^{(1)}, \ldots, \hat{\nu}_l^{(M)}\big) \ . \tag{2.4}$$

For $l = i, n$, let $\hat{\nu}_l' := (\hat{\nu}_l^{(1)}, \ldots, \hat{\nu}_l^{(M)})$. Furthermore, let $\hat{\nu}' := (\hat{\nu}_i', \hat{\nu}_n')$. Then the estimator of Δ can also be expressed as

$$\hat{\Delta} := f(\hat{\nu}) \ .$$

3 Solutions

We assume that

$$\sqrt{T}(\hat{\nu} - \nu) \xrightarrow{d} N(0, \Psi) \ , \tag{3.1}$$

where Ψ is an unknown symmetric positive definite matrix of dimension $2M \times 2M$ and the symbol $\xrightarrow{d}$ denotes convergence in distribution. This relation holds under mild regularity conditions. For example, when the data are assumed to be independent and identically distributed (i.i.d.), it is sufficient to have both $\mathbb{E}(r_{1i}^{2M})$ and $\mathbb{E}(r_{1n}^{2M})$ finite. For various sets of sufficient conditions in the time series case, see White (2001), for example.

The delta method then implies that

$$\sqrt{T}(\hat{\Delta} - \Delta) \xrightarrow{d} N\big(0, \nabla' f(\nu) \Psi \nabla f(\nu)\big) \ , \tag{3.2}$$

where the $2M \times 1$ vector-valued function $\nabla f(\cdot)$ is the gradient of $f(\cdot)$.[1]

Therefore, if a consistent estimator $\hat{\Psi}$ of Ψ is available, then an asymptotic standard error[2] for $\hat{\Delta}$ is given by

$$s(\hat{\Delta}) := \sqrt{\frac{\nabla' f(\hat{\nu}) \hat{\Psi} \nabla f(\hat{\nu})}{T}} \ . \tag{3.3}$$

Given the formula (2.2) for $f(\cdot)$, it holds that

$$\nabla' f(\nu) = \big(\nabla' h(\nu_i), -\nabla' h(\nu_n)\big) \ .$$

We now give explicit formulas for the gradient $\nabla h(\cdot)$ in Examples 2.1–2.5.

[1] The convergence result (3.2) requires that $\nabla f(\nu) \neq 0$; this assumption holds in all the examples that we consider.

[2] In our terminology, "standard error" means the estimated standard deviation of an estimator rather than the standard deviation itself.

Example 3.1 (Example 2.1 continued: Sharpe ratio)**.** In this case,

$$\nabla' h(a,b) = \left(\frac{b}{(b-a^2)^{1.5}}, -\frac{1}{2}\frac{a}{(b-a^2)^{1.5}} \right) .$$

Example 3.2 (Example 2.2 continued: Variance)**.** In this case,

$$\nabla' h(a,b) = \left(-\frac{2a}{b-a^2}, \frac{1}{b-a^2} \right) .$$

Example 3.3 (Example 2.3 continued: Mean)**.** In this case,

$$\nabla h(a) = 1 .$$

Example 3.4 (Example 2.4 continued: Skewness)**.** In this case,

$$\nabla' h(a,b,c) = \left(\frac{-3b^2+3ac}{(b-a^2)^{2.5}}, \frac{-3c+3ab}{2(b-a^2)^{2.5}}, \frac{1}{(b-a^2)^{1.5}} \right)$$

Example 3.5 (Example 2.5 continued: Kurtosis)**.** In this case,

$$\nabla' h(a,b,c,d) = \left(\frac{12ab^2-12a^2c+4ad-4bc}{(b-a^2)^3}, \frac{-6a^2b+8ac-2d}{(b-a^2)^3}, -\frac{4a}{(b-a^2)^2}, \frac{1}{(b-a^2)^2} \right)$$

3.1 HAC Inference

As is well known ((Andrews, 1991), the limiting covariance matrix in (3.1) is given by

$$\Psi := \lim_{T\to\infty} \frac{1}{T} \sum_{s=1}^{T} \sum_{t=1}^{T} \mathbb{E}[y_s y_t'],$$

where

$$y_t' := (r_{ti} - \nu_i^{(1)}, \ldots, r_{ti}^M - \nu_i^{(M)}, r_{tn} - \nu_n^{(1)}, \ldots, r_{tn}^M - \nu_n^{(M)}) .$$

By change of variables, the limit can be alternatively expressed as

$$\Psi = \lim_{T\to\infty} \Psi_T, \quad \text{with} \quad \Psi_T := \sum_{j=-T+1}^{T-1} \Gamma_T(j), \quad \text{where}$$

$$\Gamma_T(j) := \begin{cases} \frac{1}{T}\sum_{t=j+1}^{T} \mathbb{E}[y_t y_{t-j}'] & \text{for } j \geq 0 \\ \frac{1}{T}\sum_{t=-j+1}^{T} \mathbb{E}[y_{t+j} y_t'] & \text{for } j < 0 \end{cases} .$$

The standard method to come up with a consistent estimator $\hat{\Psi} := \hat{\Psi}_T$ is to use heteroskedasticity and autocorrelation robust (HAC) kernel estimation; for example, see Andrews (1991) and Andrews and Monahan (1992). In practice, this involves choosing a real-valued kernel function $k(\cdot)$ and a bandwidth S_T. Apart from being a symmetric function, the kernel $k(\cdot)$ typically satisfies the three conditions $k(0) = 1$, $k(\cdot)$ is continuous at 0, and $\lim_{x\to\infty} k(x) = 0$. The kernel estimate for Ψ is then given by

$$\hat{\Psi} := \frac{T}{T-2M} \sum_{j=-T+1}^{T-1} k\left(\frac{j}{S_T}\right) \hat{\Gamma}_T(j), \quad \text{where} \tag{3.4}$$

$$\hat{\Gamma}_T(j) := \begin{cases} \frac{1}{T}\sum_{t=j+1}^{T} \hat{y}_t \hat{y}'_{t-j} & \text{for } j \geq 0 \\ \frac{1}{T}\sum_{t=-j+1}^{T} \hat{y}_{t+j} \hat{y}'_t & \text{for } j < 0 \end{cases} ,$$

where

$$\hat{y}'_t := (r_{ti} - \hat{\nu}_i^{(1)}, \ldots, r_{ti}^M - \hat{\nu}_i^{(M)}, r_{tn} - \hat{\nu}_n^{(1)}, \ldots, r_{tn}^M - \hat{\nu}_n^{(M)}) .$$

The factor $T/(T-2M)$ in (3.4) is a small-sample degrees-of-freedom adjustment that is introduced to offset the effect of the estimation of the $2M \times 1$ vector ν in the computation of the $\hat{\Gamma}_T(j)$, that is, the use of the $\hat{y}_t$ rather than the y_t.

An important feature of a kernel $k(\cdot)$ is its characteristic exponent $1 \leq q \leq \infty$, determined by the smoothness of the kernel at the origin. Note that the bigger q, the smaller is the asymptotic bias of a kernel variance estimator; on the other hand, only kernels with $q \leq 2$ yield estimates that are guaranteed to be positive semi-definite in finite samples. Most of the commonly used kernels have $q = 2$, such as the Parzen, Tukey-Hanning, and Quadratic-Spectral (QS) kernels, but exceptions do exist. For example, the Bartlett kernel has $q = 1$ and the Truncated kernel has $q = \infty$. For a broader discussion on this issue, see Andrews (1991) again.

Once a particular kernel $k(\cdot)$ has been chosen for application, one must pick the bandwidth S_T. Several automatic methods, based on various asymptotic optimality criteria, are available to this end; for example, see Andrews (1991) and Newey and West (1994).

Finally, given the kernel estimator $\hat{\Psi}$, the standard error $s(\hat{\Delta})$ is obtained as in (3.3) and then it is combined with the asymptotic normality (3.2) to make HAC inference as follows.

A two-sided p-value for the null hypothesis H_0: $\Delta = 0$ is given by

$$\hat{p} := 2\Phi\left(-\frac{|\hat{\Delta}|}{s(\hat{\Delta})}\right) ,$$

where $\Phi(\cdot)$ denotes the c.d.f. of the standard normal distribution. Alternatively, a nominal $1-\alpha$ two-sided confidence interval for Δ is given by

$$\hat{\Delta} \pm z_{1-\alpha/2}\, s(\hat{\Delta}) ,$$

where z_λ denotes the λ quantile of the standard normal distribution, that is, $\Phi(z_\lambda) = \lambda$.

It is, however, well known that such HAC inference is often liberal when samples sizes are small to moderate. This means hypothesis tests tend to reject a true null hypothesis too often compared to the nominal significance level and confidence intervals tend to undercover; for example, see Andrews (1991), Andrews and Monahan (1992), and Romano and Wolf (2006).

3.2 Bootstrap Inference

There is an extensive literature demonstrating the improved inference accuracy of the studentized bootstrap over 'standard' inference based on asymptotic normality; for example, see Hall (1992) for i.i.d. data and Lahiri (2003) for time series data. Very general results are

available for parameters of interests that are smooth functions of means, covering our scenario of interest (2.3)–(2.4).

Arguably, the regularity conditions used by Lahiri (2003, Section 6.5) in the time series case are rather strong (and too strong for most financial applications); for example, they assume $35 + \delta$ finite moments (where δ is some small number) and certain restrictions on the dependence structure.[3] However, it should be pointed out that these conditions are *sufficient* only to prove the very complex underlying mathematics but not *necessary*. Even when these conditions do not hold, the studentized bootstrap typically continues to outperform 'standard' inference; see Section 4. To avoid any confusion, it should also be pointed that these strong regularity conditions are only needed to prove the superiority of the studentized bootstrap; proving first-order validity of the bootstrap inference does not really require stronger sufficient conditions compared to 'standard' inference.

We propose to test H_0: $\Delta = 0$ by inverting a bootstrap confidence interval. That is, one constructs a two-sided bootstrap confidence interval with nominal level $1 - \alpha$ for Δ. If this interval does not contain zero, then H_0 is rejected at nominal level α. The advantage of this 'indirect' approach is that one can simply resample from the observed data. If one wanted to carry out a 'direct' bootstrap test, one would have to resample from a probability distribution that satisfies the constraint of the null hypothesis, that is, from some modified data where the two (sample) performance measures are equal; for example, see Politis et al. (1999, Section 1.8).

In particular, we propose to construct a symmetric studentized bootstrap confidence interval. To this end, the two-sided distribution function of the studentized statistic is approximated via the bootstrap as follows:

$$\mathcal{L}\left(\frac{|\hat{\Delta} - \Delta|}{s(\hat{\Delta})}\right) \approx \mathcal{L}\left(\frac{|\hat{\Delta}^* - \hat{\Delta}|}{s(\hat{\Delta}^*)}\right) . \tag{3.5}$$

In this notation, Δ is true difference between the two performance measures, $\hat{\Delta}$ is the estimated difference computed from the original data, $s(\hat{\Delta})$ is a standard error for $\hat{\Delta}$ (also computed from the original data), $\hat{\Delta}^*$ is the estimated difference computed from bootstrap data, and $s(\hat{\Delta}^*)$ is a standard error for $\hat{\Delta}^*$ (also computed from bootstrap data). Finally, $\mathcal{L}(X)$ denotes the distribution of the random variable X.

Letting $z^*_{|\cdot|,\lambda}$ denote a λ quantile of $\mathcal{L}(|\hat{\Delta}^* - \hat{\Delta}|/s(\hat{\Delta}^*))$, a bootstrap $1-\alpha$ confidence interval for Δ is then given by

$$\hat{\Delta} \pm z^*_{|\cdot|,1-\alpha}\, s(\hat{\Delta}) . \tag{3.6}$$

The point is that when data are heavy-tailed or of time series nature, then $z^*_{|\cdot|,1-\alpha}$ will typically be somewhat larger than $z_{1-\alpha/2}$ for small to moderate samples, resulting in more conservative inference compared to the HAC procedures of Section 3.1.

[3] The conditions are too lengthy to be reproduced here.

We are left to specify (i) how the bootstrap data are to be generated and (ii) how the standard errors $s(\hat{\Delta})$ and $s(\hat{\Delta}^*)$ are to be computed. To this end, it is useful to distinguish between i.i.d. data and time series data. The first case, i.i.d. data, is included mainly for completeness of the exposition. It is well known that financial returns are generally not i.i.d.. Even when the autocorrelation of the returns is negligible (which often happens with the stock and mutual fund returns), there usually exists autocorrelation of the squared returns, that is, volatility clustering. We therefore recommend to always use the bootstrap procedure for time series data in practice.

3.2.1 I.I.D. Data

To generate bootstrap data, one simply uses Efron's (1979) bootstrap, resampling individual pairs from the observed pairs $r_t = (r_{ti}, r_{tn})'$, $t = 1, \ldots, T$, with replacement. The standard error $s(\hat{\Delta})$ is computed as in (3.3). Since the data are i.i.d., one takes for $\hat{\Psi}$ here simply the sample covariance matrix of the vectors $(r_{ti}, \ldots, r_{ti}^M, r_{tn}, \ldots, r_{tn}^M)'$, $t = 1, \ldots, T$. The standard error $s(\hat{\Delta}^*)$ is computed in exactly the same fashion but from the bootstrap data instead of the original data. To be more specific, denote the tth return pair of the bootstrap sample by $r_t^* := (r_{ti}^*, r_{tn}^*)'$. Then one takes for $\hat{\Psi}^*$ the sample covariance matrix of the vectors $(r_{ti}^*, \ldots, (r_{ti}^*)^M, r_{tn}^*, \ldots, (r_{tn}^*)^M)'$, $t = 1, \ldots, T$. Furthermore, the estimator of $\nu := (\nu_i^{(1)}, \ldots, \nu_i^{(M)}, \nu_n^{(1)}, \ldots, \nu_n^{(M)})'$ based on the bootstrap data is denoted by $\hat{\nu}^* := (\nu_i^{*,(1)}, \ldots, \nu_i^{*,(M)}, \nu_n^{*,(1)}, \ldots, \nu_n^{*,(M)})'$. Finally, the bootstrap standard error for $\hat{\Delta}^*$ is given by

$$s(\hat{\Delta}^*) := \sqrt{\frac{\nabla' f(\hat{\nu}^*) \hat{\Psi}^* \nabla f(\hat{\nu}^*)}{T}} \,. \tag{3.7}$$

3.2.2 Time Series Data

The application of the studentized bootstrap is somewhat more involved when the data are of time series nature. To generate bootstrap data, we use the circular block bootstrap of Politis and Romano (1992), resampling now blocks of pairs from the observed pairs $r_t := (r_{ti}, r_{tn})'$, $t = 1, \ldots, T$, with replacement.[4] These blocks have a fixed size $b \geq 1$. The standard error $s(\hat{\Delta})$ is computed as in (3.3). The estimator $\hat{\Psi}$ is obtained via kernel estimation; in particular we propose the prewhitened QS kernel of Andrews and Monahan (1992).[5] The standard error $s(\hat{\Delta}^*)$ is the 'natural' standard error computed from the bootstrap data, making use of the special block dependence structure; see Götze and Künsch (1996) for details. To be more specific, let $m := \lfloor n/b \rfloor$, where $\lfloor \cdot \rfloor$ denotes the integer part. Again, the estimator of ν based on the bootstrap data is denoted by $\hat{\nu}^*$. Then define

$$y_t^* := (r_{ti}^* - \hat{\nu}_i^{*,(1)}, \ldots, (r_{it}^*)^M - \hat{\nu}_i^{*,(M)}, r_{tn}^* - \hat{\nu}_n^{*,(1)}, \ldots, (r_{nt}^*)^M - \hat{\nu}_n^{*,(M)})' \quad t = 1, \ldots, T \,,$$

[4] The motivation for using the circular block bootstrap instead of the moving blocks bootstrap of Künsch (1989) is to avoid the 'edge effects' of the latter; see Romano and Wolf (2006, Section 4).

[5] We have found that the prewhitened Parzen kernel, which is defined analogously, yields very similar performance.

$$\zeta_j := \frac{1}{\sqrt{b}} \sum_{t=1}^{b} y^*_{(j-1)b+t} \quad j = 1, \ldots, m \ ,$$

and

$$\hat{\Psi}^* := \frac{1}{m} \sum_{j=1}^{m} \zeta_j \zeta_j' \ .$$

With this more general definition[6] of $\hat{\Psi}^*$, the bootstrap standard error for $\hat{\Delta}^*$ is again given by formula (3.7).

An application of the studentized circular block bootstrap requires a choice of the block size b. To this end, we suggest to use a *calibration* method, a concept dating back to Loh (1987). One can think of the actual coverage level $1 - \lambda$ of a block bootstrap confidence interval as a function of the block size b, conditional on the underlying probability mechanism $\mathbb{P}$ that generated the bivariate time series of returns, the nominal confidence level $1 - \alpha$, and the sample size T. The idea is now to adjust the 'input' b in order to obtain the actual coverage level close to the desired one. Hence, one can consider the block size calibration function $g : b \to 1 - \lambda$. If $g(\cdot)$ were known, one could construct an 'optimal' confidence interval by finding $\tilde{b}$ that minimizes $|g(b) - (1 - \alpha)|$ and then use $\tilde{b}$ as the block size of the time series bootstrap; note that $|g(b) - (1 - \alpha)| = 0$ may not always have a solution.

Of course, the function $g(\cdot)$ depends on the underlying probability mechanism $\mathbb{P}$ and is therefore unknown. We now propose a bootstrap procedure to estimate it. The idea is that in principle we could simulate $g(\cdot)$ if $\mathbb{P}$ were known by generating data of size T according to $\mathbb{P}$ and by computing confidence intervals for Δ for a number of different block sizes b. This process is then repeated many times and for a given b, one estimates $g(b)$ as the fraction of the corresponding intervals that contain the true parameter. The method we propose is identical except that $\mathbb{P}$ is replaced by an estimate $\hat{\mathbb{P}}$ and that the true parameter Δ is replaced by the 'pseudo' parameter $\hat{\Delta}$.

Algorithm 3.1 (Choice of the Block Size)**.**

1. Fit a semi-parametric model $\hat{\mathbb{P}}$ to the observed data $(r_{1i}, r_{1n})', \ldots, (r_{Ti}, r_{Tn})'$.
2. Fix a selection of reasonable block sizes b.
3. Generate K pseudo sequences $(r^*_{1i}, r^*_{1n})'_k \ldots, (r^*_{Ti}, r^*_{Tn})'_k$, $k = 1, \ldots, K$, according to $\hat{\mathbb{P}}$. For each sequence, $k = 1, \ldots, K$, and for each b, compute a confidence interval $\mathrm{CI}_{k,b}$ with nominal level $1 - \alpha$ for $\hat{\Delta}$.
4. Compute $\hat{g}(b) := \#\{\hat{\Delta} \in \mathrm{CI}_{k,b}\}/K$.
5. Find the value $\tilde{b}$ that minimizes $|\hat{g}(b) - (1 - \alpha)|$.

Of course, the question remains which semi-parametric model to fit to the observed return data. When using monthly data, we recommend to simply use a VAR model in conjunction

[6] Note that for the special case $b = 1$, this definition simplifies to the sample covariance matrix of the bootstrap data in the case of i.i.d. data.

with time series bootstrapping the residuals.[7] If the data are sampled at finer intervals, such as daily data, one might want to use a bivariate GARCH model instead.

Next, one might ask what is a selection of reasonable block sizes? The answer is any selection that contains a b with $\hat{g}(b)$ very close to $1-\alpha$. If nothing else, this can always be determined by trial and error. In our experience, $\hat{g}(\cdot)$ is typically monotonically increasing in b. So if one starts with $b_{low}=1$ and a b_{up} 'sufficiently' large, one is left to specify some suitable grid between those two values. In our experience, again, for a sample size of $T=120$, the choices $b_{low}=1$ and $b_{up}=10$ usually suffice. In that case, the final selection $\{1,2,4,6,8,10\}$ should be fine, as $\hat{g}(\cdot)$ does not tend to decrease very fast in b.

Finally, how large should K be chosen in application to real data? The answer is as large as possible, given the computational resources. $K=5,000$ will certainly suffice for all practical purposes, whereas $K=1,000$ should be the lower limit.

Remark 3.1 (Computation of a p-value)**.** As outlined above, a two-sided test for H_0: $\Delta=0$ at significance level α can be carried out by constructing a bootstrap confidence interval with confidence level $1-\alpha$. The test rejects if zero is not contained in the interval. At times, it might be more desirable to obtain a p-value. In principle, such a p-value could be computed by 'trial and error' as the smallest α for which the corresponding $1-\alpha$ confidence interval does not contain zero. However, such a procedure is rather cumbersome. Fortunately, there exists a shortcut that allows for an equivalent 'direct' computation of such a p-value. Denote the original studentized test statistic by d, that is,

$$d := \frac{|\hat{\Delta}|}{s(\hat{\Delta})} \ ,$$

and denote the *centered* studentized statistic computed from the kth bootstrap sample by $\tilde{d}^{*,k}$, $k=1,\ldots,K$, that is,

$$\tilde{d}^{*,k} := \frac{|\hat{\Delta}^{*,k}-\hat{\Delta}|}{s(\hat{\Delta}^{*,k})} \ ,$$

where K is the number of bootstrap resamples. Then the p-value is computed as[8]

$$\hat{p} := \frac{\#\{\tilde{d}^{*,k} \geq d\}+1}{K+1} \ . \ \blacksquare \tag{3.8}$$

4 Simulation Study

The purpose of this section is to shed some light on the finite-sample performance of the various procedures via some (necessarily limited) simulations. We do this for the two performance

[7] At this point we opt for the stationary bootstrap of Politis and Romano (1994), since it is quite insensitive to the choice of the average block size. The motivation for time series bootstrapping the residuals is to account for some possible 'left over' non-linear dependence not captured by the linear VAR model.

[8] The addition of 1 in both the numerator and the denominator of the fraction follows the recommendation of Davison and Hinkley (1997, Section 4.2.1).

measures of most interest: the Sharpe ratio and the variance. We compute empirical rejection probabilities under the null, based on 5,000 simulations per scenario. The nominal levels considered are $\alpha = 0.01, 0.5, 0.1$. All bootstrap p-values are computed as in (3.8), employing $M = 499$. The sample size is $T = 120$ always.[9]

4.1 Competing Procedures

The following procedures are included in the study:

- **(JKM)** The test of Jobson and Korkie (1981), using the corrected version of Memmel (2003), when the performance measure is the Sharpe ratio.
- **(F)** The classic F-test for the equality of two variances when the performance measure is the variance; for example, see Mood et al. (1974, Section IX.4.4).
- **(HAC)** The HAC test of Section 3.1 based on the QS kernel with automatic bandwidth selection of Andrews (1991).
- **(HAC$_{PW}$)** The HAC test of Section 3.1 based on the prewhitened QS kernel with automatic bandwidth selection of Andrews and Monahan (1992).
- **(Boot-IID)** The bootstrap procedure of Section 3.2.1.
- **(Boot-TS)** The bootstrap procedure of Section 3.2.2. We use Algorithm 3.1 to choose a data-dependent block size from the input block sizes $b \in \{1, 2, 4, 6, 8, 10\}$. The semi-parametric model used is a VAR(1) model in conjunction with bootstrapping the residuals. For the latter we employ the stationary bootstrap of Politis and Romano (1994) with an average block size of 5.

4.2 Data Generating Processes

In all scenarios, we want the null hypothesis of equal performance measures to be true. This is easiest achieved if the two marginal return processes are identical.

It is natural to start with i.i.d. bivariate normal data with equal mean 1 and equal variance 1. The within-pair correlation is chosen as $\rho = 0.5$, which seems a reasonable number for many applications. This DGP is denoted by Normal-IID.

We then relax the strict i.i.d. normal assumption in various dimensions.

First, we keep the i.i.d. assumption but allow for heavy tails. To this end, we use bivariate t_6 data, shifted to have equal mean 1 and standardized to have common variance 1. The within-pair correlation is $\rho = 0.5$ again. This DGP is denoted by t_6-IID. Next, we consider an uncorrelated process but with correlations in the squared returns, as is typical for stock returns. The standard way to model this is via a bivariate GARCH(1,1) model. In particular, we use the bivariate *diagonal-vech* model dating back to Bollerslev et al. (1988). Let $\tilde{r}_{ti} := r_{ti} - \mu_i$, $\tilde{r}_{tn} := r_{tn} - \mu_n$, and denote by Ω_{t-1} the conditioning information available at time $t-1$. Then

[9] For example, many empirical applications use ten years of monthly data.

the diagonal-vech model is defined by

$$\begin{aligned}
\mathbb{E}(\tilde{r}_{ti}|\Omega_{t-1}) &= 0 \\
\mathbb{E}(\tilde{r}_{tn}|\Omega_{t-1}) &= 0 \\
\mathrm{Cov}(\tilde{r}_{ti}\tilde{r}_{tn}|\Omega_{t-1}) =: h_{tin} &= c_{in} + a_{in}\,\tilde{r}_{(t-1)i}\,\tilde{r}_{(t-1)n} + b_{in}\,h_{(t-1)in}\ .
\end{aligned}$$

In other words, the conditional (co)variances depend only on their own lags and the lags of the corresponding (cross)products. We use the following coefficient matrices:

$$C := \begin{pmatrix} 0.15 & 0.13 \\ 0.13 & 0.15 \end{pmatrix} \qquad A := \begin{pmatrix} 0.075 & 0.050 \\ 0.050 & 0.075 \end{pmatrix} \qquad B := \begin{pmatrix} 0.90 & 0.89 \\ 0.90 & 0.89 \end{pmatrix}$$

These matrices are inspired by the bivariate estimation results based on weekly returns on a broad U.S. market index and a broad U.K. market index.[10] However, all three diagonals are forced to be equal to get identical individual return processes; see Ledoit et al. (2003, Table 2).

The first variant of the GARCH model uses i.i.d. bivariate standard normal innovations to recursively generate the series $\tilde{r}_t := (\tilde{r}_{ti}, \tilde{r}_{tn})'$. At the end, we add a global mean, that is, $r_t := \tilde{r}_t + \mu$, where μ is chosen as $\mu := (16.5/52, 16.5/52)'$. Again this choice is inspired by the previously mentioned estimation results, forcing $\mu_i = \mu_n$ to get identical individual return processes; see Ledoit et al. (2003, Table 1). This DGP is denoted by Normal-GARCH.

The second variant of the GARCH model uses i.i.d. bivariate t_6 innovations instead (standardized to have common variance equal to 1, and covariance equal to 0).[11] Everything else is equal. This DGP is denoted by t_6-GARCH.

Finally, we also consider correlated processes. To this end, we return to the two i.i.d. DGPs Normal-IID and t_6-IID, respectively, but add some mild autocorrelation to the individual return series via an AR(1) structure with AR coefficient $\phi = 0.2$.[12] This then corresponds to a VAR(1) model with bivariate normal or (standardized) t_6 innovations. The resulting two DGPs are denoted by Normal-VAR and t_6-VAR, respectively.

4.3 Results

The results when the performance measure is the Sharpe ratio are presented in Table 1 and can be summarized as follows:

- JKM works well for i.i.d. bivariate normal data but is not robust against fat tails or time series effects, where it becomes liberal.

[10] We use estimation results based on weekly returns, since generally there are very few GARCH effects at monthly or longer return horizons. With weekly data, $T = 120$ corresponds to a data window of slightly over two years.

[11] There is ample evidence that the innovations of GARCH processes tend to have tails heavier than the normal distribution; for example, see Kuester et al. (2006) and the references therein.

[12] For example, a first-order autocorrelation around 0.2 is quite typical for monthly hedge fund returns.

- HAC inference, while asymptotically consistent, is often liberal in finite samples. This finding is consistent with many previous studies; for example, see Romano and Wolf (2006) and the references therein.
- Boot-IID works well for i.i.d. data but is liberal for time series data.
- Boot-TS works well both for i.i.d. and time series data.

The results when the performance measure is the variance are presented in Table 2 and can be summarized as follows:

- F works well for i.i.d. bivariate normal data but is not robust against fat tails or time series effects, where it becomes liberal.
- HAC inference, while asymptotically consistent, is often liberal in finite samples. This finding is consistent with many previous studies; for example, see Romano and Wolf (2006) and the references therein.
- Boot-IID works well for i.i.d. data but is liberal for time series data.
- Boot-TS works well both for i.i.d. and time series data.

Remark 4.1. We also included HAC and HAC_{PW} based on the (prewhitened) Parzen kernel instead of the (prewhitened) QS kernel in the numerical work. The results were virtually identical and are therefore not reported. Since the Parzen kernel has bounded support, whereas the QS kernel does not, it is somewhat more convenient to implement. ■

5 Empirical Applications

As a brief illustration, we consider two applications to investment funds when the performance measure is the Sharpe ratio. In each case, we want to test the null hypothesis of equality of the Sharpe ratios of the two funds being compared.

The first application deals with mutual funds. The selected funds are Fidelity (FFIDX), a 'large blend' fund, and Fidelity Aggressive Growth (FDEGX), a 'mid-cap growth' fund. The data were obtained from Yahoo! Finance.[13]

The second application deals with hedge funds. The selected funds are Coast Enhanced Income and JMG Capital Partners. The data were obtained from the CISDM database; see Romano et al. (2008, Section 9).

In both applications, we use monthly log returns in excess of the riskfree rate. The return period is 01/1994 until 12/2003, so $T = 120$. Table 3 provides some relevant summary statistics. Note that all returns are in percentages and that none of the statistics are annualized.

Table 4 presents the corresponding p-values of the five procedures considered in the simulation study. Boot-TS uses a data-dependent choice of block size based on Algorithm 3.1. The semi-parametric model is a VAR(1) model in conjunction with bootstrapping the residuals. For the latter we employ the stationary bootstrap of Politis and Romano (1994) with an

[13] We use close prices adjusted for dividends and stock splits.

average block size of 5. The nominal confidence level is $1 - \alpha = 0.95$ and the set of input block sizes is $\{1, 2, 4, 6, 8, 10\}$. The two estimated calibration functions, based on $K = 5,000$ pseudo sequences, are displayed in Figure 1. As a result, the estimated optimal block sizes are $\tilde{b} = 4$ for the mutual funds application and $\tilde{b} = 6$ for the hedge funds application.

The bootstrap p-values are computed as in (3.8), employing $M = 4,999$. In both applications, JKM results in a rejection of the null at significance level $\alpha = 0.05$, whereas HAC, HAC_{PW}, and Boot-TS do not. Not surprisingly, given the noticeable autocorrelation of hedge fund returns, the differences are more pronounced for the second application. Boot-IID results in a rejection for the mutual funds data but not for the hedge fund data. But, as discussed previously, we recommend to always use Boot-TS with financial return data.

6 Conclusion

Testing for the equality of a given performance measure of two investment strategies is an important problem in applied financial research. In this paper, we have considered the class of performance measures that can be expressed as a smooth functions of population means of the underlying returns. This class is very rich and contains, among others, the Sharpe ratio, the variance, the mean, the skewness, and the kurtosis. Unfortunately, many of the inference procedures that have been suggested previously in the applied literature make unreasonable assumptions that do not apply to real-life return data, such as normality and independence over time. As was demonstrated in simulations and empirical applications, inference based on such procedures is unreliable and can lead to erroneous findings.

We have discussed two alternative inference procedures that are asymptotically valid under very general conditions, allowing for heavy tails and time dependence in the return data. HAC inference uses kernel estimators to come up with consistent standard errors. The resulting inference works well with large samples but is often liberal for small to moderate sample sizes. In such applications, it is preferable to use a studentized time series bootstrap. Arguably, this procedure is quite complex to implement, but corresponding programming codes are freely available at `econ.uzh.ch/faculty/wolf.html`. These codes are for the Sharpe ratio and for the variance as performance measures. But they can be quite easily adapted to other performance measures (that fall in the class of smooth functions of population means).

Finally, both HAC inference and the studentized bootstrap procedure detailed in this paper could be modified to make inference for (the difference of) various refinements to the Sharpe ratio recently proposed in the literature—for example, see Ferruz and Vicente (2005) and Israelsen (2003, 2005)—as well as many other performance measures, such as the Information ratio, Jensen's alpha, or the Treynor ratio, to name just a few. Some guidance on how to do this when the performance measure is a regression coefficient—for example, in the case of Jensen's alpha—can be found in Romano and Wolf (2006).

References

Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, 59(3):817–858.

Andrews, D. W. K. and Monahan, J. C. (1992). An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator. *Econometrica*, 60(4):953–966.

Bollerslev, T., Engle, R. F., and Wooldridge, J. M. (1988). Modelling the coherence in short-run nominal exchange rates: A multivariate Generalized ARCH model. *Review of Economics and Statistics*, 72:498–505.

Candelon, B., Hurlin, C., and Tokpavi, S. (2012). Sampling error and double shrinkage estimation of minimum variance portfolios. *Journal of Empirical Finance*, 19(4):511–527.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press, Cambridge.

DeMiguel, V., Garlappi, L., Nogales, F. J., and Uppal, R. (2009a). A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. *Management Science*, 55(5):798–812.

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

Dittmar, R. (2002). Nonlinear princing kernels, kurtosis preference, and evidence from the cross section of equity returns. *Journal of Finance*, 57:369–403.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7:1–26.

Elton, E. J., Gruber, M. J., and Spitzer, J. (2006). Improved estimates of correlation coefficients and their impact on optimum portfolios. *European Financial Management*, 12(3):303–318.

Ferruz, L. and Vicente, L. (2005). Style portfolio performance: Evidence from the Spanish equity funds. *Journal of Asset Management*, 5:397–409.

Frahm, G. and Memmel, C. (2010). Dominating estimators for minimum-variance portfolios. *Journal of Econometrics*, 159(2):289–302.

Garlappi, L., Uppal, R., and Wang, T. (2007). Portfolio selection with parameter and model uncertainty: A multi-prior approach. *Review of Financial Studies*, 20:41–81.

Götze, F. and Künsch, H. R. (1996). Second order correctness of the blockwise bootstrap for stationary observations. *Annals of Statistics*, 24:1914–1933.

Güttler, A. and Trübenbach, F. (2011). Alternative objective functions for quasi-shrinkage portfolio optimization. Research Paper 10-07, European Business School. Available at SSRN: `http://ssrn.com/abstract=1576567`.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer, New York.

Harvey, C. and Siddique, A. (2000). Conditional skewness in asset princing. *Journal of Finance*, 55:1263–1295.

Israelsen, C. L. (2003). Sharpening the Sharpe ratio. *Financial Planning*, 33(1):49–51.

Israelsen, C. L. (2005). A refinement to the Sharpe ratio and Information ratio. *Journal of Asset Management*, 5:423–427.

Jagannathan, R. and Ma, T. (2003). Risk reduction in large portfolios: Why imposing the wrong constraints helps. *Journal of Finance*, 54(4):1651–1684.

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

Kempf, A. and Memmel, C. (2006). Estimating the global mininum variance portfolio. *Schmalenbach Business Review*, 58:332–348.

Kuester, K., Mittnik, S., and Paolella, M. S. (2006). Value–at–risk prediction: A comparison of alternative strategies. *Journal of Financial Econometrics*, 4:53–89.

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *Annals of Statistics*, 17:1217–1241.

Lahiri, S. N. (2003). *Resampling Methods for Dependent Data*. Springer, New York.

Ledoit, O., Santa-Clara, P., and Wolf, M. (2003). Flexible multivariate GARCH modeling with an application to international stock markets. *Review of Economics and Statistics*, 85(3):735–747.

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

Ledoit, O. and Wolf, M. (2011). Robust performance hypothesis testing with the variance. *Wilmott Magazine*, September:86–89.

Loh, W. Y. (1987). Calibrating confidence coefficients. *Journal of the American Statistical Association*, 82:155–162.

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

Mitton, T. and Vorkink, K. (2007). Equilibrium underdiversification and the preference for skewness. *Review of Financial Studies*, 20:1255–1288.

Mood, A. M., Graybill, F. A., and Boes, D. C. (1974). *Introduction to the Theory of Statistics*. McGraw-Hill, New York.

Newey, W. K. and West, K. D. (1994). Automatic lag selection in covariance matrix estimation. *Review of Economic Studies*, 61:631–653.

Patton, A. (2004). On the out-of-sample importance of skewness and asymmetric dependence for asset allocation. *Journal of Financial Econometrics*, 2:130–168.

Politis, D. N. and Romano, J. P. (1992). A circular block-resampling procedure for stationary data. In LePage, R. and Billard, L., editors, *Exploring the Limits of Bootstrap*, pages 263–270. John Wiley, New York.

Politis, D. N. and Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, 89:1303–1313.

Politis, D. N., Romano, J. P., and Wolf, M. (1999). *Subsampling*. Springer, New York.

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. (2006). Improved nonparametric confidence intervals in time series regressions. *Journal of Nonparametric Statistics*, 18(2):199–214.

Scherer, B. (2011). A note on the returns from mininum variance investing. *Journal of Empirical Finance*, 18:652–660.

Sharpe, W. F. (1966). Mutual fund performance. *Journal of Business*, 39:119–138.

White, H. L. (2001). *Asymptotic Theory for Econometricians*. Academic Press, New York, revised edition.

Table 1: Empirical rejection probabilities (in percent) for various data generating processes (DGPs) and inference procedures; see Section 4 for a description. For each DGP, the null hypothesis of equal Sharpe ratios is true and so the empirical rejection probabilities should be compared to the nominal level of the test, given by α. We consider three values of α, namely $\alpha = 1\%$, 5% and 10%. All empirical rejection probabilities are computed from 5,000 repetitions of the underlying DGP, and the same set of repetitions is shared by all inference procedures.

DGP	JKM	HAC	HAC_{PW}	Boot-IID	Boot-TS
Nominal level $\alpha = 1\%$					
Normal-IID	1.2	1.2	1.2	1.1	1.0
t_6-IID	3.5	1.9	2.1	1.4	1.3
Normal-GARCH	1.7	1.8	1.8	1.5	1.1
t_6-GARCH	1.8	2.0	2.0	1.6	1.2
Normal-VAR	2.5	2.2	1.8	2.7	1.2
t_6-VAR	6.4	2.6	2.2	1.8	1.1
Nominal level $\alpha = 5\%$					
Normal-IID	5.0	5.3	5.4	4.9	4.8
t_6-IID	10.7	6.7	6.9	5.2	5.0
Normal-GARCH	7.2	7.1	7.2	6.0	5.5
t_6-GARCH	7.4	7.7	7.5	6.9	5.7
Normal-VAR	9.5	6.9	6.1	8.5	5.0
t_6-VAR	14.5	7.9	7.3	7.3	5.1
Nominal level $\alpha = 10\%$					
Normal-IID	10.3	10.3	10.7	10.1	9.6
t_6-IID	17.9	12.4	12.5	10.3	9.9
Normal-GARCH	12.8	12.5	12.3	12.4	10.5
t_6-GARCH	13.7	13.3	13.1	13.1	11.1
Normal-VAR	15.6	12.4	10.8	15.6	9.7
t_6-VAR	22.5	13.3	12.0	13.3	9.8

Table 2: Empirical rejection probabilities (in percent) for various data generating processes (DGPs) and inference procedures; see Section 4 for a description. For each DGP, the null hypothesis of equal variances is true and so the empirical rejection probabilities should be compared to the nominal level of the test, given by α. We consider three values of α, namely $\alpha = 1\%$, 5% and 10%. All empirical rejection probabilities are computed from 5,000 repetitions of the underlying DGP, and the same set of repetitions is shared by all inference procedures.

DGP	F	HAC	HAC_{PW}	Boot-IID	Boot-TS
Nominal level $\alpha = 1\%$					
Normal-IID	0.2	1.2	1.4	0.9	0.9
t_6-IID	4.2	1.5	1.7	0.8	0.8
Normal-GARCH	0.4	1.4	1.3	1.0	0.9
t_6-GARCH	0.3	1.5	1.5	1.0	1.0
Normal-VAR	0.5	2.1	2.0	1.6	0.9
t_6-VAR	3.8	2.1	2.0	1.1	1.0
Nominal level $\alpha = 5\%$					
Normal-IID	2.4	6.1	6.1	5.1	4.9
t_6-IID	11.5	6.8	7.0	4.9	4.7
Normal-GARCH	2.1	5.4	5.5	5.0	4.8
t_6-GARCH	2.4	5.7	5.9	5.1	5.0
Normal-VAR	3.1	7.2	6.7	6.4	4.8
t_6-VAR	10.9	6.9	6.5	5.3	4.9
Nominal level $\alpha = 10\%$					
Normal-IID	5.9	11.3	11.1	10.2	9.8
t_6-IID	18.3	11.4	10.4	10.1	9.7
Normal-GARCH	5.6	10.8	11.0	10.2	10.1
t_6-GARCH	6.0	10.9	11.2	10.1	9.8
Normal-VAR	7.3	12.4	11.7	12.0	9.9
t_6-VAR	17.8	12.4	12.0	10.2	10.0

Table 3: Summary sample statistics for monthly log returns in excess of the riskfree rate: mean, standard deviation, Sharpe ratio, and first-order autocorrelation.

Fund	$\bar{r}$	s	$\widehat{Sh}$	$\hat{\phi}$
Fidelity	0.511	4.760	0.108	−0.010
Fidelity Agressive Growth	0.098	9.161	0.011	0.090
Coast Enhanced Income	0.245	0.168	1.461	0.152
JMG Capital Partners	1.228	1.211	1.014	0.435

Table 4: p-values (in percent) for various inference procedures; see Section 4 for a description. The data set 'Mutual Funds' corresponds to the top two funds of Table 2; the data set 'Hedge funds' corresponds to the bottom two funds of Table 2. All p-values are for the two-sided test of equal Sharpe ratios.

Data	JKM	HAC	HAC_{PW}	Boot-IID	Boot-TS
Mutual funds	3.9	6.3	6.7	4.4	9.2
Hedge funds	1.0	14.7	25.4	5.8	29.4

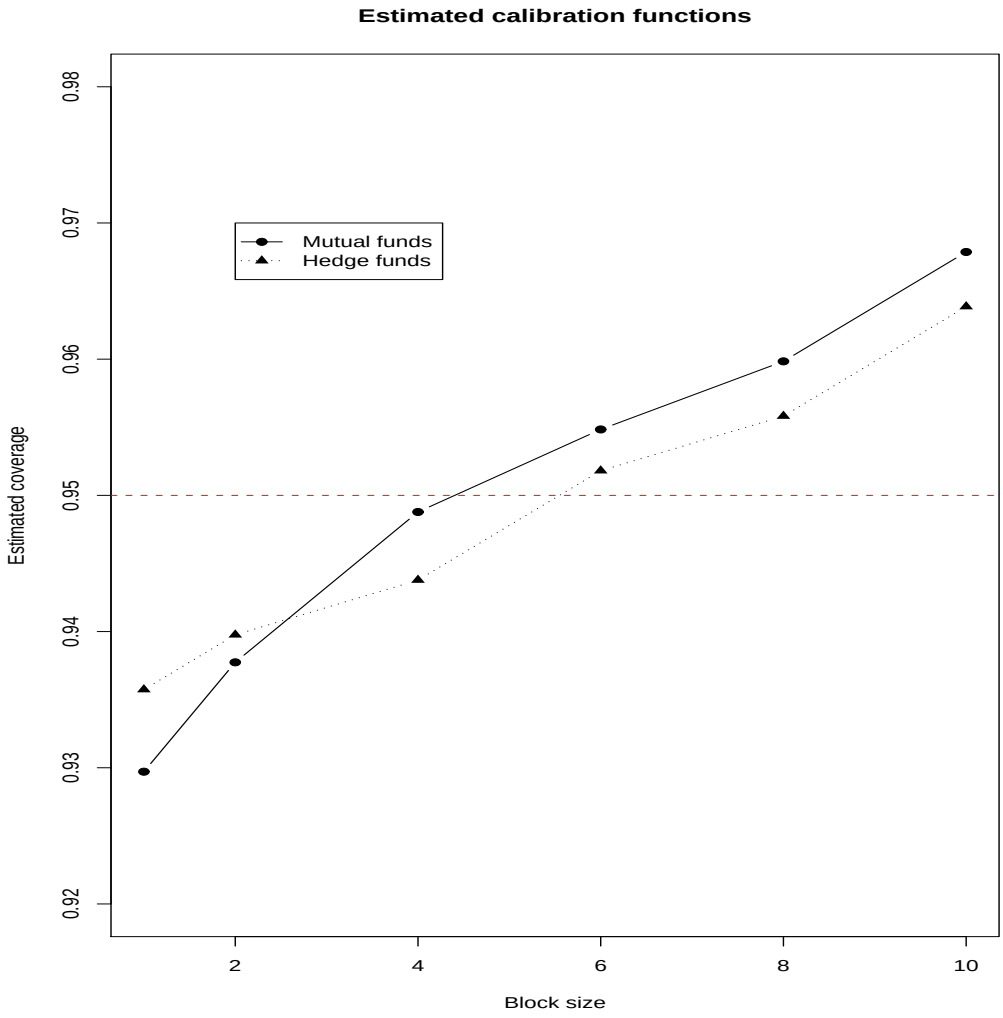

Figure 1: Estimated calibration functions for the two empirical applications. The nominal level is $1 - \alpha = 0.95$. The resulting estimated optimal block sizes are $\tilde{b} = 4$ for the mutual funds application and $\tilde{b} = 6$ for the hedge funds application.