

Robust Performance Hypothesis Testing With the Variance

Olivier Ledoit
Department of Economics
University of Zurich
CH-8006 Zurich, Switzerland
olivier.ledoit@econ.uzh.ch

Michael Wolf*
Department of Economics
University of Zurich
CH-8006 Zurich, Switzerland
michael.wolf@econ.uzh.ch

September 2011

Abstract

Applied researchers often test for the difference of the variance of two investment strategies; in particular, when the investment strategies under consideration aim to implement the global minimum variance portfolio. A popular tool to this end is the F -test for the equality of variances. Unfortunately, this test is not valid when the returns are correlated, have tails heavier than the normal distribution, or are of time series nature. Instead, we propose the use of robust inference methods. In particular, we suggest to construct a studentized time series bootstrap confidence interval for the ratio of the two variances and to declare the two variances different if the value one is not contained in the obtained interval. This approach has the advantage that one can simply resample from the observed data as opposed to some null-restricted data. A simulation study demonstrates the improved finite-sample performance compared to existing methods.

KEY WORDS: Bootstrap, HAC inference, Variance.

JEL CLASSIFICATION NOS: C12, C14, C22.

*Corresponding author; phone: +41-44 634 5096; fax: +41-44 634 4907. This research has been supported by the NCCR Finrisk project “New Methods in Theoretical and Empirical Asset Pricing”.

1 Introduction

Many applications of financial performance analysis are concerned with the comparison of the variances of two investment strategies (such as stocks, portfolios, mutual funds, hedge funds, or technical trading rules). This is of particular interest when the investment strategies aim to implement the global minimum variance (GMV) portfolio. The GMV portfolio has received much renewed interest in the recent literature; for example, see Jagannathan and Ma (2003), Kempf and Memmel (2006), Garlappi et al. (2007), Elton et al. (2008), DeMiguel et al. (2009a), DeMiguel et al. (2009b), Candelou et al. (2010), and Güttler and Trübenbach (2010).

Since the true quantities are not observable, the variances have to be estimated from historical return data and the comparison has to be based on statistical inference, such as hypothesis tests or confidence intervals. The most popular test for equality of variances is the classical F -test; for example, see Mood et al. (1974, Section IX.4.4). However, this test requires the data to come from a bivariate normal distribution with correlation zero and to be independent over time. This joint requirement is basically never met for financial returns.

In this paper, we discuss inference methods that are more generally valid. One possibility is to compute a HAC standard error¹ for the difference of the estimated variances by the methods of Andrews (1991) and Andrews and Monahan (1992), say. 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.

From a purely academic point of view, this paper can be considered a rather straightforward modification of our previous work Ledoit and Wolf (2008) which deals with the comparison of the Sharpe ratios of two investment strategies. However, we feel that not all practitioners would have the time and energy to carry out the modification on their own, in particular as far as the programming code is concerned. Furthermore, an innovative variance-stabilizing transformation plays a key role in the modification in order to obtain inference methods with improved finite-sample properties; see Remark 3.1. We, therefore, hope that the finance profession will indeed find value in our new work.

2 The Problem

We use the same notation as Jobson and Korkie (1981), Memmel (2003), and Ledoit and Wolf (2008) who study the related problem of testing for equality of two Sharpe ratios. There are two investment strategies i and n whose returns at time t are r_{ti} and r_{tn} , respectively.² A total of T return pairs $(r_{1i}, r_{1n})', \dots, (r_{Ti}, r_{Tn})'$ 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

¹In this paper, a standard error of an estimator denotes an estimate of the true standard deviation of the estimator.

²Strictly speaking, the previously mentioned works consider excess returns over a given benchmark. This more general scenario also suits our set-up by choosing the benchmark to be zero.

given by

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

The usual sample means and sample variances of the observed returns are denoted by $\hat{\mu}_i, \hat{\mu}_n$ and $\hat{\sigma}_i^2, \hat{\sigma}_n^2$, respectively. The ratio of the two variances is given by

$$\Theta = \frac{\sigma_i^2}{\sigma_n^2}$$

and its estimator is

$$\hat{\Theta} = \frac{\hat{\sigma}_i^2}{\hat{\sigma}_n^2} .$$

The hypotheses of interest are

$$H_0 : \Theta = 1 \quad \text{vs.} \quad H_1 : \Theta \neq 1 . \tag{1}$$

The classical F -test is based on the following test statistic

$$F = \frac{\hat{\sigma}_i^2}{\hat{\sigma}_n^2} .$$

Denote by F_{λ, k_1, k_2} the λ quantile of F_{k_1, k_2} , the F distribution with k_1 and k_2 degrees of freedom. The F -test rejects H_0 at significance level α if and only if (iff)

$$F < F_{\alpha/2, T-1, T-1} \quad \text{or} \quad F > F_{1-\alpha/2, T-1, T-1} .$$

Crucially, this test requires that the data come from a bivariate normal distribution with $\sigma_{in} = 0$ and be independent over time. If the data are correlated in the sense of $\sigma_{in} \neq 0$, have tails heavier than the normal distribution, or are dependent over time, the test is not valid, not even in an asymptotic sense. Since financial data exhibit generally at least one of these three violations, one should not use the F -test when testing the equality of variances of investment strategies.

3 Solutions

The exposition in this section follows closely Ledoit and Wolf (2008, Section 3).

We start by re-formulating the testing problem. Define

$$\Delta = \log(\Theta) = \log(\sigma_i^2) - \log(\sigma_n^2)$$

with sample counterpart

$$\hat{\Delta} = \log(\hat{\Theta}) = \log(\hat{\sigma}_i^2) - \log(\hat{\sigma}_n^2) .$$

Then the testing problem (1) is equivalent to the following one

$$H_0 : \Delta = 0 \quad \text{vs.} \quad H_1 : \Delta \neq 0 . \tag{2}$$

Remark 3.1 The purpose of the log-transformation is to conduce better finite-sample properties of our proposed inference methods by means of being a variance-stabilizing transformation; for example, see Efron and Tibshirani (1993, Section 12.6). The naïve approach to modifying the method of Ledoit and Wolf (2008) would be to test

$$H_0 : \sigma_i^2 - \sigma_n^2 = 0 \quad \text{vs.} \quad H_1 : \sigma_i^2 - \sigma_n^2 \neq 0$$

instead. However, this approach would lead to inference methods with inferior finite-sample properties.³

Let $\gamma_i = E(r_{1i}^2)$ and $\gamma_n = E(r_{1n}^2)$. Their sample counterparts are denoted by $\hat{\gamma}_i$ and $\hat{\gamma}_n$, respectively. Furthermore, let $v = (\mu_i, \mu_n, \gamma_i, \gamma_n)'$ and $\hat{v} = (\hat{\mu}_i, \hat{\mu}_n, \hat{\gamma}_i, \hat{\gamma}_n)'$. This allows us to write

$$\Delta = f(v) \quad \text{and} \quad \hat{\Delta} = f(\hat{v}) \tag{3}$$

with

$$f(a, b, c, d) = \log(c - a^2) - \log(d - b^2) . \tag{4}$$

We assume that

$$\sqrt{T}(\hat{v} - v) \xrightarrow{d} N(0; \Psi) , \tag{5}$$

where Ψ is an unknown symmetric positive semi-definite matrix. This relation holds under mild regularity conditions. For example, when the data are assumed i.i.d., it is sufficient to have both $E(r_{1i}^4)$ and $E(r_{1n}^4)$ finite. In the time series case it is sufficient to have finite $4 + \delta$ moments, where δ is some small positive constant, together with an appropriate mixing condition; for example, see Andrews (1991). The delta method then implies

$$\sqrt{T}(\hat{\Delta} - \Delta) \xrightarrow{d} N(0; \nabla' f(v) \Psi \nabla f(v)) \tag{6}$$

with

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

Now, if a consistent estimator $\hat{\Psi}$ of Ψ is available, then a standard error for $\hat{\Delta}$ is given by

$$s(\hat{\Delta}) = \sqrt{\frac{\nabla' f(\hat{v}) \hat{\Psi} \nabla f(\hat{v})}{T}} . \tag{7}$$

3.1 HAC Inference

As is well-known, Ψ can be consistently estimated by heteroskedasticity and autocorrelation consistent (HAC) kernel methods. For details, the reader is referred to Ledoit and Wolf (2008, Subsection 3.1). Given the kernel estimator $\hat{\Psi}$, the standard error $s(\hat{\Delta})$ is obtained as in (7) and then combined with the asymptotic normality (6) to make HAC inference as follows.

³Corresponding simulation results are not included in the paper but available from the authors upon request.

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 $1 - \alpha$ 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.

It is, however, well established 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), Romano and Wolf (2006), and Ledoit and Wolf (2008).

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; see Hall (1992) for i.i.d. data and Lahiri (2003) for time series data. Very general results are available for parameters of interests which are smooth functions of means. Our parameter of interest, Δ , fits this bill; see (3) and (4). Taking into account our specific definitions of Δ , $f(\cdot)$, and $\nabla f(\cdot)$, the actual implementation of the bootstrap inference is identical to the one of Ledoit and Wolf (2008, Section 3.2). So the reader is referred there for the details.

In particular, the test at significance level α is carried out by constructing a two-sided symmetric bootstrap confidence interval for Δ with confidence level $1 - \alpha$ and rejecting H_0 iff $\Delta_0 = 0$ is not contained in the interval. The advantage of this ‘indirect’ test by inverting a confidence interval is that one can simply resample from the observed data. A ‘direct’ test, on the other hand, would require one to bootstrap from a null distribution where the two variances are indeed equal.

This approach is equivalent to constructing a two-sided bootstrap confidence interval for Θ and rejecting H_0 iff $\Theta_0 = 1$ is not contained in the interval. Here, the bootstrap confidence interval for Θ is obtained by simply applying the exponential transformation to the two endpoints of the bootstrap confidence interval for Δ ; see Efron and Tibshirani (1993, Section 12.6).⁴

In addition to carrying out a test at fixed significance level α , it is also very easy to compute bootstrap p -values, an approach which some researchers might find more informative; see Remark 3.2 of Ledoit and Wolf (2008).

⁴The resulting bootstrap confidence interval for Θ will also be two-sided but, generally, no longer symmetric.

4 Simulation Study

The purpose of this section is to shed some light on the finite sample performance of the various methods via some (necessarily limited) simulations. 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 employing $M = 499$ resamples. The sample size is $T = 120$ always.⁵

4.1 Competing Methods

The following methods are included in the study:

- **(F)** The classical F -test.
- **(HAC)** The HAC test of Subsection 3.1 based on the QS kernel with automatic bandwidth selection of Andrews (1991).
- **(HAC_{PW})** The HAC test of Subsection 3.1 based on the prewhitened QS kernel with automatic bandwidth selection of Andrews and Monahan (1992).
- **(Boot-IID)** The bootstrap method of Subsection 3.2.1 of Ledoit and Wolf (2008).
- **(Boot-TS)** The bootstrap method of Subsection 3.2.2 of Ledoit and Wolf (2008). We use their Algorithm 3.1 to pick 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 variances to be true. This is easiest achieved if the two marginal return processes are identical.

We start with i.i.d. bivariate normal with equal variance one and within-pair correlation chosen as $\rho = 0.5$. The assumptions of normality and independence over time are gradually relaxed. In total, we consider the same six data generating processes (DGPs) as Ledoit and Wolf (2008, Section 4).

4.3 Results

The results are presented in Table 1 and summarized as follows:

- Not surprisingly, the F -test does not work for any DGP, as its joint requirement of zero-correlation bivariate normal data which are independent over time is never met.

⁵Many empirical applications use ten years of monthly data.

Depending on the DGP, the inference can be conservative or liberal, sometimes by a large amount.

- HAC inference, while asymptotically consistent, is generally liberal in finite samples. This finding is consistent with many previous studies; e.g., see Romano and Wolf (2006), Ledoit and Wolf (2008), and the references therein.
- Boot-IID works well for i.i.d. and GARCH data, but is liberal for VAR data.
- Boot-TS works well for all DGPs.

Remark 4.1 We also included HAC and HAC_{PW} based on the (prewhitened) Parzen kernel instead of the (prewhitened) QS kernel. The results were virtually identical and are therefore not reported. Since the Parzen kernel has a finite support while the QS kernel does not, it is somewhat more convenient to implement; for example, see Andrews (1991).

5 Conclusion

Testing for the equality of the variances of two investment strategies is an important tool for performance analysis; it is of particular relevance when the two strategies aim to implement the global minimum variance portfolio. A common tool is the classical F -test. Unfortunately, this test is not robust against tails heavier than the normal distribution, non-zero correlation of strategies' returns during common return periods, and time series characteristics. Since all three effects are quite common with financial returns, the F -test should not be used.

We have discussed alternative inference methods which are robust. 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 code will be made freely available at <http://www.econ.uzh.ch/faculty/wolf.html>.

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.
- Candelou, B., Hurlin, C., and Topkavi, S. (2010). Sampling error and double shrinkage estimation of minimum variance portfolios. Technical report, University of Paris Ouest Nanterre La Défense.
- 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.
- Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.
- Elton, E. J., Gruber, M. J., and Spitzer, J. F. (2008). Improved estimates of correlation and their impact on the optimum portfolios. NYU Finance Working Paper FIN-04-016, New York University. Available at SSRN: <http://ssrn.com/abstract=588924>.
- 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.
- Güttler, A. and Trübenbach, F. (2010). 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.
- 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 minimum variance portfolio. *Schmalenbach Business Review*, 58:332–348.
- Lahiri, S. N. (2003). *Resampling Methods for Dependent Data*. Springer, New York.

- Ledoit, O. and Wolf, M. (2008). Robust performance hypothesis testing with the Sharpe ratio. *Journal of Empirical Finance*, 15:850–859.
- Memmel, C. (2003). Performance hypothesis testing with the Sharpe Ratio. *Finance Letters*, 1:21–23.
- Mood, A. M., Graybill, F. A., and Boes, D. C. (1974). *Introduction to the Theory of Statistics*. McGraw-Hill, New York.
- Politis, D. N. and Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, 89:1303–1313.
- Romano, J. P. and Wolf, M. (2006). Improved nonparametric confidence intervals in time series regressions. *Journal of Nonparametric Statistics*, 18(2):199–214.

Table 1: Empirical rejection probabilities (in percent) for various data generating processes (DGPs) and inference methods; 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 methods.

| 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 |