AP 2018–04

Chair of Applied Stochastics and Risk Management

Helmut Schmidt Universität
Universität der Bundeswehr Hamburg

Faculty of Economic and Social Sciences
Department of Mathematics and Statistics

Working Paper

An Intersection-Union Test for the Sharpe Ratio

Gabriel Frahm

March 30, 2018

Please use only the latest version of the manuscript. Distribution is unlimited.

Working Paper

AP 2018–04
March 30, 2018

An Intersection-Union Test for the Sharpe Ratio

Gabriel Frahm

Helmut Schmidt University
Faculty of Economic and Social Sciences
Department of Mathematics and Statistics
Chair of Applied Stochastics and Risk Management
Holstenhofweg 85, D-22043 Hamburg, Germany

URL: www.hsu-hh.de/stochastik
Phone: +49 (0)40 6541-2791
E-mail: frahm@hsu-hh.de

Working Paper
Please use only the latest version of the manuscript. Distribution is unlimited.

Supervised by: Prof. Dr. Gabriel Frahm
Chair of Applied Stochastics and
Risk Management

URL: www.hsu-hh.de/stochastik

An Intersection-Union Test for the Sharpe Ratio

Gabriel Frahm*
Helmut Schmidt University
Department of Mathematics and Statistics
Chair of Applied Stochastics and
Risk Management

March 30, 2018

Abstract

An intersection-union test for supporting the hypothesis that a given investment strategy is optimal among a set of alternatives is presented. It compares the Sharpe ratio of the benchmark with that of each other strategy. The intersection-union test takes serial dependence into account and does not presume that asset returns are multivariate normally distributed. An empirical study based on the G–7 countries demonstrates that it is hard to find significant results due to the lack of data, which confirms a general observation in empirical finance.

Keywords: Ergodicity, Gordin's condition, heteroscedasticity, intersection-union test, Jobson-Korkie test, performance measurement, Sharpe ratio.

JEL Subject Classification: C12, G11.

*Phone: +49 40 6541-2791, e-mail: frahm@hsu-hh.de.

1. Motivation

THIS work builds upon Frahm et al. (2012), in which the authors argue why joint and multiple testing procedures should be applied in order to judge whether or not some investment strategy is optimal among a set of several alternatives. Frahm et al. (2012) can be understood as a complement to DeMiguel et al. (2009), who doubt that portfolio optimization on the basis of time-series information is worthwhile at all. Indeed, modern portfolio theory suffers from a serious drawback, namely that portfolio weights are very sensitive to estimation risk. It is well-known that portfolio optimization fails on estimating expected asset returns.

DeMiguel et al. (2009) show that well-established investment strategies are not *significantly* better than the naive strategy, i.e., the equally weighted portfolio. Of course, this does not mean that naive diversification is optimal, but we usually have not enough observations in order to prove the opposite. They highlight a general problem of empirical finance, namely that hypothesis testing is difficult due to the lack of data. This is all the more true if there is more than one (single) null hypothesis. The results reported by DeMiguel et al. (2009) are convincing, but their statistical methodology does not take the undesirable effects of joint and multiple testing into account. By contrast, the test presented in this work is designed to address those problems.

The literature provides a wide range of different investment strategies (see, e.g., Bartosz, 2012, Burgess, 2000, Conrad and Kaul, 1998, DeMiguel et al., 2009, Menkhoff et al., 2012, Shen et al., 2007, Szakmary et al., 2010, Vrugt et al., 2004, Zagrodny, 2003) and we are typically concerned with the question of whether a given investment strategy is optimal among a set of alternatives.[1] In order to validate our hypothesis, we usually compare the performance of our benchmark, e.g., its certainty equivalent or Sharpe ratio, with the performance of each other strategy that is taken into consideration. Let $d > 1$ be the number of investment strategies and $i \in \{1, 2, \ldots, d\}$ be our benchmark. We may suppose that $i = 1$ without loss of generality. Further, let $\eta = (\eta_1, \eta_2, \ldots, \eta_d) \in \mathbb{R}^d$ be a (column) vector of performance measures. Now, first of all, consider the hypotheses

$$H_{0\wedge}: \eta_1 \geq \eta \qquad \text{vs.} \qquad H_{1\wedge}: \eta_1 \ngeq \eta.$$

That is, $H_{0\wedge}$ states that our benchmark is optimal. After performing a (joint) hypothesis test, we could reject the null hypothesis $H_{0\wedge}$ in favor of the alternative hypothesis $H_{1\wedge}$. In this case, we could say that there exists *some* strategy that is better than our benchmark, but not *which* one.[2] By contrast, if we are not able to reject $H_{0\wedge}$ we must not conclude that our benchmark is optimal. A well-known method for testing the intersection of a number of single null hypotheses is studied by Roy (1953), which is called union-intersection test (Sen and Silvapulle, 2002). However, union-intersection tests are not the object of this work.

[1] A different question is whether some *asset universe* allows the investor to achieve a higher performance compared to another asset universe (Hanke and Penev, 2018).

[2] In order to identify the outperforming strategies we would have to apply a multiple test. For more details on that topic see, e.g., Frahm et al. (2012) as well as Romano and Wolf (2005).

By contrast, here I consider the following hypotheses:

$$H_{0\vee}: \eta_1 \not\geq \eta \qquad \text{vs.} \qquad H_{1\vee}: \eta_1 \geq \eta.$$

Now, the joint null hypothesis $H_{0\vee}$ asserts that our benchmark is not optimal. If we are able to reject $H_{0\vee}$, our benchmark turns out to be (significantly) optimal among *all* alternatives. By contrast, in the case in which we cannot reject the null hypothesis we must not conclude that our benchmark is outperformed by any other strategy. Applying a test for $H_{0\vee}$ might be the primary goal both in theoretical and in practical applications of portfolio theory.

The former test can be rewritten, equivalently, as

$$H_{0\wedge}: \bigwedge_{i=2}^{d} \eta_1 \geq \eta_i \qquad \text{vs.} \qquad H_{1\wedge}: \bigvee_{i=2}^{d} \eta_1 < \eta_i,$$

whereas the latter test reads

$$H_{0\vee}: \bigvee_{i=2}^{d} \eta_1 < \eta_i \qquad \text{vs.} \qquad H_{1\vee}: \bigwedge_{i=2}^{d} \eta_1 \geq \eta_i.$$

This explains the chosen symbols for the null and the alternative hypothesis. However, in the following I focus on the latter test and write only "H_0" and "H_1" for notational convenience.

The test proposed in this work is very simple: The null hypothesis is rejected if and only if we can reject each single hypothesis $H_{0i}: \eta_1 < \eta_i$ in favor of $H_{1i}: \eta_1 \geq \eta_i$. Let A_i be the event that H_{0i} is rejected. The probability that all single null hypotheses are rejected amounts to

$$\mathbb{P}\left(\bigcap_{i=2}^{d} A_i\right) \leq \bigwedge_{i=2}^{d} \mathbb{P}(A_i).$$

If H_{0i} is true for some $i \in \{2, 3, \ldots, d\}$ we must have that $\mathbb{P}(A_i) \leq \alpha_i$, where $\alpha_i \in (0,1)$ denotes the significance level of the (single) hypothesis test for H_{0i}. Under H_0 at least one single null hypothesis must be true and thus we have that

$$\bigwedge_{i=2}^{d} \mathbb{P}(A_i) \leq \bigvee_{i=2}^{d} \alpha_i \,.$$

Hence, the proposed test for H_0 has level $\alpha \in (0,1)$ if $\alpha_2, \alpha_3, \ldots, \alpha_d \leq \alpha$. The least conservative choice is $\alpha_2 = \alpha_3 = \ldots = \alpha_d = \alpha$, in which case H_0 is rejected if and only if the largest p-value of all single tests falls below α. Throughout this work, I assume that each single test has level α.

At first glance, this testing procedure might seem to suffer from a lack of power because it does not take the dependence structure of the single test statistics into account. Nonetheless, it is a likelihood-ratio test that is commonly referred to as an intersection-union test (Berger, 1997). Thus, it inherits the general asymptotic optimality properties of likelihood-ratio tests that are known from likelihood theory (see, e.g., van der Vaart, 1998, Chapter 15 and 16). Another striking feature might be the fact that the overall test has the same significance level as each

single test. This is because H_0 is rejected only if *all* single tests lead to a rejection and so we need no Bonferroni correction in order to preserve the significance level of each single test. For more details on that topic see Berger (1997) as well as Sen and Silvapulle (2002).

In this work, I present an intersection-union test in order to decide whether a given investment strategy is optimal among a set of alternative strategies. This is done with respect to the Sharpe ratio. Joint and multiple tests for the Sharpe ratio are applied also in Frahm et al. (2012) by using a stationary block-bootstrap procedure. By contrast, here I provide analytical results. I refrain from assuming that asset returns are serially independent and multivariate normally distributed. Each single test represents a (nonparametric) generalization of the Jobson-Korkie test (Jobson and Korkie, 1981, Memmel, 2003). Finally, I apply the intersection-union test to historical data.

The same problem is addressed by Ledoit and Wolf (2008) as well as Schmid and Schmidt (2009) in a bivariate setting. However, the intersection-union test presented here is motivated by a *multivariate* point of view, i.e., $d > 2$, and its primary goal is to avoid any kind of selection bias that can occur when testing a joint hypothesis. Thus, it cannot be said that the intersection-union test is "better" or "worse" than the tests proposed by Ledoit and Wolf (2008). It is hardly possible to provide any general answer to this question at all (Ledoit and Wolf, 2008, Section 4 and 5). Instead, I try to fill a gap between Frahm et al. (2012) as well as Ledoit and Wolf (2008):

(i) I derive closed-form expressions for the standard errors of the test statistics, instead of providing numerical results that have been obtained by bootstrapping, and

(ii) I do this for the case $d \geq 2$ but not (only) for $d = 2$.

2. The Intersection-Union Test

2.1. Gordin's Condition

In the following, "$X_n \to X$" denotes almost sure convergence, whereas "$X_n \rightsquigarrow X$" stands for convergence in distribution. Let $P_t > 0$ be the price of some asset or, more generally, the value of some strategy at time $t \in \mathbb{Z}$. Throughout this work, the terms "asset" and "strategy" as well as "price" and "value" are used synonymously. The asset return after Period t is defined as $R_t := P_t / P_{t-1} - 1$.[3] I assume that the return process $\{R_t\}$ is (strongly) stationary with expected return $\mu := \mathbf{E}(R_t)$ and variance $\sigma^2 := \mathbf{Var}(R_t) < \infty$. The process $\{R_t\}$ shall also be ergodic. This means that $\frac{1}{n}\sum_{t=1}^{n} f(R_t) \to \mathbf{E}\big(f(R)\big)$ for each integrable function f of R, where the random variable R has the same distribution as each component of $\{R_t\}$. This guarantees that every finite moment of R can be consistently estimated by the corresponding moment estimator. The return process is ergodic if it is mixing (Bradley, 2005). More precisely, for all $k, l = 1, 2, \ldots$, the random vector $(R_t, R_{t+1}, \ldots, R_{t+k})$ is asymptotically independent of $(R_{t-n}, R_{t-n+1}, \ldots, R_{t-n+l})$ as $n \to \infty$ (Hayashi, 2000, p. 101).

The ergodicity of $\{R_t\}$ implies that $\mu_n \to \mu$, where $\mu_n := \frac{1}{n}\sum_{t=1}^{n} R_t$ is the sample mean of $R_1, R_2, \ldots, R_n$. Put another way, the return process satisfies the Strong Law of Large Numbers. In order to preserve the Central Limit Theorem (CLT), i.e., $\sqrt{n}\,(\mu_n - \mu) \rightsquigarrow \mathcal{N}\big(0, \sigma_{\mathrm{L}}^2\big)$, we need

[3] Any capital income that occurs during Period t is considered part of P_t.

an additional requirement. This is known as Gordin's condition (Hayashi, 2000, p. 402). Let $\mathcal{H}_t := (R_t, R_{t-1}, \ldots)$ be the history of $\{R_t\}$ at time $t \in \mathbb{Z}$. It is assumed that $\mathbf{E}(R_t \mid \mathcal{H}_{t-n})$ converges in mean square to μ as $n \to \infty$ and, according to Hayashi (2000, p. 403), we must have that

$$\sum_{k=0}^{\infty} \sqrt{\mathbf{E}(\varepsilon_k^2)} < \infty$$

with $\varepsilon_k := \mathbf{E}(R_t \mid \mathcal{H}_{t-k}) - \mathbf{E}(R_t \mid \mathcal{H}_{t-k-1})$ for $k = 0, 1, \ldots$. It can be shown that $\sigma_{\mathrm{L}}^2 = \sum_{k=-\infty}^{\infty} \Gamma(k)$, where Γ is the autocovariance function of $\{R_t\}$ (Hayashi, 2000, Proposition 6.10). The number σ_{L}^2 is referred to as the large-sample variance of $\{R_t\}$, whereas σ^2 represents its stationary variance. In the following, I assume that $\tau^2 := \mathbf{Var}\big((R_t - \mu)^2\big) < \infty$ and that Gordin's condition is satisfied not only for $\{R_t\}$ but also for $\{(R_t - \mu)^2\}$.

The aforementioned requirements can easily be extended to any d-dimensional return process (Hayashi, 2000, p. 405) and applied to a broad class of standard time-series models. There exist a number of alternative criteria for the CLT, which can be found, e.g., in Brockwell and Davis (1991, p. 213) as well as Hamilton (1994, p. 195). However, to the best of my knowledge, Gordin's condition represents the most unrestrictive set of assumptions about the serial dependence structure of a stochastic process (Eagleson, 1975). In particular, it can be considered a natural generalization of the CLT for martingale difference sequences (Hayashi, 2000, p. 106).

It is worth emphasizing that the number of dimensions, d, is supposed to be *fixed*. At least, we have to assume that $n, d \to \infty$ such that $n/d \to \infty$. If n/d tends to a finite number, the CLT might become invalid and other interesting issues that are well-known from random matrix theory can arise (Frahm and Jaekel, 2015). By contrast, if the number of observations relative to the number of strategies is sufficiently large, we may expect that the CLT is satisfied under the aforementioned conditions.

I suppose, without loss of generality, that the risk-free interest rate is constantly zero. That is, I implicitly refer to asset returns in *excess* of the risk-free interest rate that can be observed at the beginning of each period. The Sharpe ratio $\eta := \mu/\sigma$ (Sharpe, 1966) is frequently used as a performance measure both in theory and in practice. In the following section, I present the intersection-union test, which can be applied in order to judge whether a given investment strategy possesses the largest Sharpe ratio among a set of alternatives. This can be done under the quite general assumptions about the return process $\{R_t\}$ mentioned above.

2.2. Asymptotic Properties of Sharpe Ratios

In this section, I present some asymptotic properties of Sharpe ratios. The reader can find the derivations in the appendix. It holds that

$$\sigma_n^2 := \frac{1}{n}\sum_{t=1}^{n}(R_t - \mu_n)^2 = \frac{1}{n}\sum_{t=1}^{n}(R_t - \mu)^2 - \underbrace{(\mu_n - \mu)^2}_{\to 0} \to \sigma^2$$

and

$$\sqrt{n}\left(\sigma_n^2 - \sigma^2\right) = \sqrt{n}\left\{\frac{1}{n}\sum_{t=1}^{n}\left[(R_t - \mu)^2 - \sigma^2\right]\right\} - \underbrace{\sqrt{n}\left(\mu_n - \mu\right)}_{\rightsquigarrow \mathcal{N}(0,\sigma_{\mathrm{L}}^2)}\underbrace{\left(\mu_n - \mu\right)}_{\to 0} \rightsquigarrow \mathcal{N}\left(0, \tau_{\mathrm{L}}^2\right).$$

This means that σ_n^2 is a consistent estimator for the stationary variance σ^2 and $\sqrt{n}\left(\sigma_n^2 - \sigma^2\right)$ is asymptotically normally distributed with large-sample variance τ_{L}^2.

For assessing the large-sample variance of $\left\{R_t\right\}$, i.e., $\sigma_{\mathrm{L}}^2 = \sum_{k=-\infty}^{\infty}\Gamma(k)$, we need to estimate the autocovariance function Γ. There are many ways to achieve this goal. Usually, one applies either heteroscedasticity-autocorrelation consistent (HAC) inference or some bootstrap procedure (Andrews, 1991, Ledoit and Wolf, 2008, Politis, 2003). A nice comparison between HAC inference and bootstrapping in the context of performance measurement can be found in Ledoit and Wolf (2008). Bootstrapping is a very powerful tool, but it can be computationally more intensive than HAC inference. Moreover, sometimes it is not clear whether or not the necessary (mathematical) conditions for the bootstrap are satisfied. The method proposed here, in some sense, bypasses the aforementioned problems. However, also HAC estimation can be somewhat obscure when it comes to choosing the right kernel and bandwidth, etc. For this reason, I keep things as simple as possible, i.e., I choose the box-kernel-type HAC-estimator

$$\sigma_{\mathrm{L}n}^2 := \Gamma_n(0) + 2\sum_{k=1}^{l}\Gamma_n(k),$$

where Γ_n is the empirical autocovariance function of $\{R_t\}$ with $l \ll n$ (Hayashi, 2000, p. 142), i.e.,

$$k \mapsto \Gamma_n(k) := \frac{1}{n}\sum_{t=k+1}^{n}\left(R_t - \mu_n\right)\left(R_{t-k} - \mu_n\right).$$

It is a stylized fact of empirical finance that $\Gamma_n(k) \approx \Gamma(k) \approx 0$ for all $k \neq 0$, i.e., asset returns are not significantly autocorrelated, and so we may expect that $\sigma_{\mathrm{L}n}^2 \approx \sigma_n^2$.

The large-sample variance of $\left\{(R_t - \mu)^2\right\}$ is τ_{L}^2, which can be estimated by

$$\tau_{\mathrm{L}n}^2 := \Pi_n(0) + 2\sum_{k=1}^{l}\Pi_n(k),$$

where Π_n is the empirical autocovariance function of $\left\{(R_t - \mu_n)^2\right\}$, i.e.,

$$k \mapsto \Pi_n(k) := \frac{1}{n}\sum_{t=k+1}^{n}\left((R_t - \mu_n)^2 - \sigma_n^2\right)\left((R_{t-k} - \mu_n)^2 - \sigma_n^2\right).$$

Typically, asset returns are conditionally heteroscedastic. This means that, in contrast to σ_{L}^2 vs. σ^2, the large-sample variance τ_{L}^2 can be significantly larger than the stationary variance τ^2.

Gordin's condition guarantees that

$$\sqrt{n}\left(\begin{bmatrix}\mu_n - \mu\\ \sigma_n^2 - \sigma^2\end{bmatrix}\right) \rightsquigarrow \mathcal{N}\left(0, \begin{bmatrix}\sigma_{\mathrm{L}}^2 & \kappa_{\mathrm{L}}\\ \kappa_{\mathrm{L}} & \tau_{\mathrm{L}}^2\end{bmatrix}\right),$$

where κ_{L} represents the large-sample covariance between R and $(R-\mu)^2$. Due to the so-called "leverage effect" (Black, 1976), we can expect that κ_{L} is negative. Moreover, we already know that $\sqrt{n}\,(\mu_n - \mu) \rightsquigarrow \mathcal{N}\big(0, \sigma_{\mathrm{L}}^2\big)$ and, by applying the delta method, we obtain

$$\sqrt{n}\,(\sigma_n - \sigma) \rightsquigarrow \mathcal{N}\left(0, \frac{\tau_{\mathrm{L}}^2}{4\sigma^2}\right),$$

which can be used in order to calculate the standard error of σ_n.

The Sharpe ratio is estimated by $\eta_n := \mu_n/\sigma_n$ and the delta method leads to

$$\sqrt{n}\big(\eta_n - \eta\big) \rightsquigarrow \mathcal{N}\left(0, \frac{\sigma_{\mathrm{L}}^2}{\sigma^2} - \frac{\eta\kappa_{\mathrm{L}}}{\sigma^3} + \frac{\eta^2\tau_{\mathrm{L}}^2}{4\sigma^4}\right).$$

Schmid and Schmidt (2009) obtain the same large-sample variance of $\{\eta_n\}$ under the assumption that the processes are strongly mixing (Bradley, 2005), but that assumption seems to be more restrictive than Gordin's condition.

To the best of my knowledge, Lo (2002) is the first who analyzes the potential impact of serial dependence when estimating the Sharpe ratio. Mertens (2002) points out that the formula for independent and identically distributed asset returns presented by Lo (2002) is implicitly based on the normal-distribution hypothesis. More precisely, he shows that the large-sample variance of $\{\eta_n\}$ is

$$1 + \frac{\eta^2}{2} - \gamma_3\eta + \frac{\gamma_4 - 3}{4}\cdot\eta^2$$

if the components of $\{R_t\}$ are independent and identically distributed, where

$$\gamma_3 := \frac{\mathbf{E}\big((R_t-\mu)^3\big)}{\sigma^3} \qquad \text{and} \qquad \gamma_4 := \frac{\mathbf{E}\big((R_t-\mu)^4\big)}{\sigma^4}$$

denote the skewness and the kurtosis of R_t, respectively. Lo (2002) implicitly presumes that $\gamma_3 = 0$ and $\gamma_4 = 3$, in which case the large-sample variance of $\{\eta_n\}$ is $1 + \eta^2/2$. Some of those results can be found also in Opdyke (2007). However, Ledoit and Wolf (2008) mention that the formula for serially dependent asset returns presented by Opdyke (2007) is wrong because it does not distinguish between large-sample and stationary (co-)variances. One purpose of this work is to clarify the aforementioned misunderstandings.

Suppose, without loss of generality, that we want to compare the Sharpe ratio of Strategy 1 with that of Strategy 2. The reader can verify in the appendix that

$$\sqrt{n}\left(\begin{bmatrix}\eta_{1n} - \eta_1\\ \eta_{2n} - \eta_2\end{bmatrix}\right) \rightsquigarrow \mathcal{N}\left(0, \begin{bmatrix}\omega_{11} & \omega_{12}\\ \omega_{21} & \omega_{22}\end{bmatrix}\right)$$

with

$$\omega_{11} = \frac{\sigma_{\text{L}1}^2}{\sigma_1^2} - \frac{\eta_1 \kappa_{\text{L}1}}{\sigma_1^3} + \frac{\eta_1^2 \tau_{\text{L}1}^2}{4\sigma_1^4}, \qquad \omega_{22} = \frac{\sigma_{\text{L}2}^2}{\sigma_2^2} - \frac{\eta_2 \kappa_{\text{L}2}}{\sigma_2^3} + \frac{\eta_2^2 \tau_{\text{L}2}^2}{4\sigma_2^4},$$

and

$$\omega_{12} = \omega_{21} = \frac{\lambda_{11}}{\sigma_1 \sigma_2} - \frac{\eta_2 \sigma_1 \lambda_{12} + \eta_1 \sigma_2 \lambda_{21}}{2\sigma_1^2 \sigma_2^2} + \frac{\eta_1 \eta_2 \lambda_{22}}{4\sigma_1^2 \sigma_2^2},$$

where

$$\begin{bmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{bmatrix}$$

is the large-sample covariance matrix of $\left(R_{1t}, (R_{1t} - \mu_1)^2\right)$ and $\left(R_{2t}, (R_{2t} - \mu_2)^2\right)$.

We conclude that

$$\sqrt{n}\left(\Delta\eta_n - \Delta\eta\right) \rightsquigarrow \mathcal{N}\left(0, \omega_{11} + \omega_{22} - 2\omega_{12}\right)$$

with $\Delta\eta_n := \eta_{1n} - \eta_{2n}$ and $\Delta\eta := \eta_1 - \eta_2$. It is worth emphasizing that the benchmark must be chosen *before* examining the Sharpe ratios. Otherwise, the entire procedure would suffer from a selection bias and then the results derived so far are no longer valid. However, this is not a serious drawback. If our choice of the benchmark is based on historical data we can simply apply the test out of sample.

As already mentioned at the end of Section 1, the given result represents a nonparametric generalization of the Jobson-Korkie test (Jobson and Korkie, 1981), which is frequently used in finance. The latter is based on the assumption that asset returns are serially independent and multivariate normally distributed. In this special case, it follows that

$$\sqrt{n}\left(\Delta\eta_n - \Delta\eta\right) \rightsquigarrow \mathcal{N}\left(0, 2\left(1 - \rho_{12}\right) + \frac{\eta_1^2 + \eta_2^2 - 2\eta_1\eta_2\rho_{12}^2}{2}\right),$$

where $\rho_{12} := \sigma_{12}/(\sigma_1\sigma_2)$ is the linear correlation coefficient between the return on Strategy 1 and the return on Strategy 2. This expression for the large-sample variance of $\{\Delta\eta_n\}$ corrects a typographical error made by Jobson and Korkie (1981) (Memmel, 2003).

2.3. Empirical Study

In order to demonstrate the intersection-union test, I consider monthly excess returns on the MSCI stock indices for the G–7 countries, i.e., Canada, France, Germany, Italy, Japan, UK and USA, from January 1970 to January 2018. The given indices are calculated on the basis of USD stock prices that are adjusted for dividends, splits, etc.[4] The sample size corresponds to $n = 577$ and the risk-free interest rate is calculated on the basis of the secondary market 3-month US treasury bill rate at the beginning of each period.[5] I choose the equally weighted portfolio (EWP) of all G–7 countries as a benchmark. This choice can be justified by the argument that investors should make use of international diversification (Jorion, 1985).

For estimating the large-sample variances, I choose the lag length $l = 12$. First of all, I show that $\Gamma_n(k) \approx 0$ for all $k \in \{1, 2, \ldots, l\}$. For this purpose, I focus on the empirical autocorrelation

[4]The total returns have been retrieved from the MSCI webpage (https://www.msci.com/end-of-day-data-country).
[5]The data have been obtained from the Federal Reserve Bank of St. Louis (https://fred.stlouisfed.org/series/TB3MS).

function, i.e., $k \mapsto \rho_n(k) := \Gamma_n(k)/\Gamma_n(0)$. Figure 1 contains the correlograms with respect to $\{R_t\}$ for the EWP and each G–7 country, where the red lines indicate the critical thresholds for the null hypothesis that the (true) autocorrelation at k is zero on the significance level $\alpha = 0.05$. Further, the reader can find the Ljung-Box Q-statistic in each plot, whose critical threshold on the significance level $\alpha = 0.05$ amounts to 21.0261. The given results confirm the general opinion that first-order autocorrelations of asset returns do not significantly differ from zero.[6] Put another way, the large-sample variances and covariances of asset returns are not significantly larger than their stationary counterparts. This picture changes substantially in Figure 2, which shows the empirical autocorrelations with respect to $\{(R_t - \mu_n)^2\}$. Now, the Ljung-Box test always leads to a rejection of the null hypothesis $H_0 : \rho(1) = \rho(2) = \ldots = \rho(12) = 0$. That is, there is a strong evidence that monthly asset returns exhibit conditional heteroscedasticity.

The following table contains the estimated large-sample variances divided by their stationary counterparts both for $\{R_t\}$ and for $\{(R_t - \mu_n)^2\}$:

	EWP	**Canada**	**France**	**Germany**	**Italy**	**Japan**	**UK**	**USA**
$\sigma^2_{\mathrm{L}n}/\sigma^2_n$	1.4987	1.0299	1.2036	1.1255	1.6913	2.1828	1.2720	1.0118
$\tau^2_{\mathrm{L}n}/\tau^2_n$	2.5962	2.7550	2.3081	2.9514	2.3707	2.8368	2.5027	2.6202

We can see that the estimates of the large-sample variance of $\{R_t\}$ do not differ very much from the stationary ones—except for Japan, where the large-sample variance seems to be more than twice the stationary variance. By contrast, the estimates of the large-sample variance of $\{(R_t - \mu_n)^2\}$ are always more than twice their stationary counterparts. Hence, it is inappropriate to ignore the serial dependence structure of monthly asset returns.

Table 1 contains the means, standard deviations, and Sharpe ratios for the EWP and the G–7 countries based on the monthly asset returns from January 1970 to January 2018. The standard errors are given in parentheses. Despite the large number of observations, the standard errors of μ_n and η_n are big compared to the corresponding estimates. This is a common problem in financial econometrics or, more specifically, in performance measurement. The last row of Table 1 contains the standard errors of the Sharpe ratios under the Jobson-Korkie assumption, i.e., that asset returns are serially independent and multivariate normally distributed. These numbers are smaller than their nonparametric counterparts and they do not vary too much. Under the Jobson-Korkie assumption, the large-sample variance of $\{\eta_n\}$ amounts to $1 + \eta^2/2 \approx 1$. Hence, the standard error of η_n is approximately $1/\sqrt{n}$, which explains why the standard errors are almost constant in the last row of Table 1.

Now, in principle, we would like to support the (alternative) hypothesis that the EWP is optimal compared to each G–7 country. Unfortunately, Table 1 shows that UK has the largest Sharpe ratio and so the EWP cannot be significantly better. Interestingly, this was not always the case. A closer inspection of the data reveals that the EWP had the largest Sharpe ratio before the financial crisis 2007–2008. However, now we have to stop our testing procedure, but for informational purposes I provide the Sharpe-ratio differences for each 7 pairs, the corresponding

[6] The only exception is Japan, where we can find a relatively large Q-statistic of 31.7637.

	EWP	**Canada**	**France**	**Germany**	**Italy**	**Japan**	**UK**	**USA**
μ_n	0.0053	0.0052	0.0062	0.0060	0.0033	0.0054	0.0052	0.0057
$\mathbf{SE}(\mu_n)$	0.0023	0.0024	0.0029	0.0028	0.0040	0.0037	0.0020	0.0026
σ_n	0.0461	0.0560	0.0640	0.0627	0.0732	0.0599	0.0436	0.0620
$\mathbf{SE}(\sigma_n)$	0.0030	0.0040	0.0037	0.0041	0.0038	0.0035	0.0028	0.0077
η_n	0.1149	0.0923	0.0971	0.0961	0.0449	0.0898	0.1202	0.0927
$\mathbf{SE}(\eta_n)$	0.0581	0.0462	0.0492	0.0479	0.0537	0.0624	0.0548	0.0508
$\mathbf{SE}_{\mathrm{JK}}(\eta_n)$	0.0419	0.0417	0.0417	0.0417	0.0417	0.0417	0.0418	0.0417

Table 1: Means, standard deviations, and Sharpe ratios for the EWP and the G–7 countries. The standard errors are given in parentheses.

	Canada	**France**	**Germany**	**Italy**	**Japan**	**UK**	**USA**
$\Delta\eta_n$	0.0226	0.0178	0.0187	0.0700	0.0251	-0.0053	0.0222
$\mathbf{SE}(\Delta\eta_n)$	0.0213	0.0317	0.0419	0.0269	0.0374	0.0381	0.0376
t	1.0635	0.5598	0.4472	2.6054	0.6718	-0.1397	0.5891
$\mathbf{SE}_{\mathrm{JK}}(\Delta\mu_n)$	0.0291	0.0227	0.0257	0.0299	0.0354	0.0290	0.0274
t_{JK}	0.7758	0.7821	0.7298	2.3420	0.7083	-0.1833	0.8089

Table 2: Sharpe ratio differences, standard errors, and t-statistics.

standard errors, and the associated t-statistics in Table 2. The reader can verify that it would have been hard to reject H_0, anyway. The problem is that *every* t-statistic must be greater than $\Phi^{-1}(1-\alpha) = 1.6449$ in order to reject H_0, but this stringent condition is fulfilled only for Italy.

The lower part of Table 2 contains the standard errors of the Sharpe ratio differences and the t-statistics that are calculated under the Jobson-Korkie assumption. Although the standard errors of η_n that are obtained under the same distributional assumption are always lower than their nonparametric counterparts (see the last row of Table 1), the same effect cannot be observed regarding $\Delta\eta_n$. The Jobson-Korkie assumption underestimates the standard errors for some indices, but it overestimates them for other indices. All in all it appears to be very difficult to compare investment strategies by historical observation because the given results are hardly ever significant if we apply a joint or a multiple hypothesis test (Frahm et al., 2012).

3. Conclusion

In portfolio optimization we are often concerned with the question of whether a given investment strategy is optimal among a set of alternatives. In this work, I presented an intersection-union test for the null hypothesis that the benchmark is suboptimal in terms of the Sharpe ratio. The proposed test can easily be implemented. Further, it accounts for serial dependence and it does not presume that asset returns are multivariate normally distributed. Thus, it is compatible with the stylized facts of empirical finance. However, an empirical study demonstrates that, in most practical applications, it is hard to reject the null hypothesis due to the lack of data.

A. Asymptotic Results

We can write $\sigma = f(\sigma^2)$ with $f\colon \sigma^2 \mapsto \sqrt{\sigma^2}$. The first derivative of f at σ^2 is $(2\sigma)^{-1}$. Hence, the asymptotic variance of $\sqrt{n}\,(\sigma_n - \sigma)$ is $\tau_{\mathrm{L}}^2\,(2\sigma)^{-2} = \tau_{\mathrm{L}}^2/(4\sigma^2)$.

Further, the Sharpe ratio can be written as $\eta = g(\mu, \sigma^2)$ with $g\colon (\mu, \sigma^2) \mapsto \mu/\sqrt{\sigma^2}$. We obtain

$$\frac{\partial g(\mu,\sigma^2)}{\partial \mu} = \frac{1}{\sigma} \qquad \text{and} \qquad \frac{\partial g(\mu,\sigma^2)}{\partial \sigma^2} = -\frac{\mu}{2\sigma^3}\,.$$

Hence, the asymptotic variance of $\sqrt{n}\big(\eta_n - \eta\big)$ reads

$$\frac{\sigma_{\mathrm{L}}^2}{\sigma^2} - 2\cdot\frac{\mu\kappa_{\mathrm{L}}}{2\sigma^4} + \frac{\mu^2\tau_{\mathrm{L}}^2}{4\sigma^6} = \frac{\sigma_{\mathrm{L}}^2}{\sigma^2} - \frac{\eta\kappa_{\mathrm{L}}}{\sigma^3} + \frac{\eta^2\tau_{\mathrm{L}}^2}{4\sigma^4}.$$

Further, if the components of $\{R_t\}$ are independent and identically distributed, we have that $\sigma_{\mathrm{L}}^2 = \sigma^2$,

$$\begin{aligned}\kappa_{\mathrm{L}} &= \mathbf{Cov}\big(R_t, (R_t-\mu)^2\big) = \mathbf{E}\big(R_t(R_t-\mu)^2\big) - \mu\sigma^2 \\ &= \mathbf{E}\big((R_t-\mu)^3\big) + \mu\sigma^2 - \mu\sigma^2 = \mathbf{E}\big((R_t-\mu)^3\big),\end{aligned}$$

and $\tau_{\mathrm{L}}^2 = \mathbf{Var}\big((R_t-\mu)^2\big) = \mathbf{E}\big((R_t-\mu)^4\big) - \sigma^4$, i.e., $\kappa_{\mathrm{L}}/\sigma^3 = \gamma_3$ and $\tau_{\mathrm{L}}^2/\sigma^4 = \gamma_4 - 1$. Thus, we conclude that

$$\frac{\sigma_{\mathrm{L}}^2}{\sigma^2} - \frac{\eta\kappa_{\mathrm{L}}}{\sigma^3} + \frac{\eta^2\tau_{\mathrm{L}}^2}{4\sigma^4} = 1 + \frac{\eta^2}{2} - \gamma_3\eta + \frac{\gamma_4 - 3}{4}\cdot\eta^2\,.$$

Now, consider the asymptotic covariance matrix of

$$\sqrt{n}\,\left(\begin{bmatrix}\eta_{1n} - \eta_1 \\ \eta_{2n} - \eta_2\end{bmatrix}\right).$$

The above result immediately leads to

$$\omega_{11} = \frac{\sigma_{\mathrm{L}1}^2}{\sigma_1^2} - \frac{\eta_1\kappa_{\mathrm{L}1}}{\sigma_1^3} + \frac{\eta_1^2\tau_{\mathrm{L}1}^2}{4\sigma_1^4} \qquad \text{and} \qquad \omega_{22} = \frac{\sigma_{\mathrm{L}2}^2}{\sigma_2^2} - \frac{\eta_2\kappa_{\mathrm{L}2}}{\sigma_2^3} + \frac{\eta_2^2\tau_{\mathrm{L}2}^2}{4\sigma_2^4}.$$

Moreover, the asymptotic covariance between $\sqrt{n}\,\big(\eta_{1n} - \eta_1\big)$ and $\sqrt{n}\,\big(\eta_{2n} - \eta_2\big)$ is

$$\begin{aligned}\omega_{12} = \omega_{21} &= \begin{bmatrix}\partial g(\mu_1,\sigma_1^2)/\partial\mu_1 \\ \partial g(\mu_1,\sigma_1^2)/\partial\sigma_1^2\end{bmatrix}' \begin{bmatrix}\lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22}\end{bmatrix} \begin{bmatrix}\partial g(\mu_2,\sigma_2^2)/\partial\mu_2 \\ \partial g(\mu_2,\sigma_2^2)/\partial\sigma_2^2\end{bmatrix} \\ &= \frac{\lambda_{11}}{\sigma_1\sigma_2} - \frac{\mu_2\lambda_{12}}{2\sigma_1\sigma_2^3} - \frac{\mu_1\lambda_{21}}{2\sigma_1^3\sigma_2} + \frac{\mu_1\mu_2\lambda_{22}}{4\sigma_1^3\sigma_2^3} = \frac{\lambda_{11}}{\sigma_1\sigma_2} - \frac{\eta_2\sigma_1\lambda_{12} + \eta_1\sigma_2\lambda_{21}}{2\sigma_1^2\sigma_2^2} + \frac{\eta_1\eta_2\lambda_{22}}{4\sigma_1^2\sigma_2^2}.\end{aligned}$$

If the asset returns are serially independent, the large-sample (co-)variances coincide with their stationary counterparts. More precisely, it holds that $\sigma_{\mathrm{L}1}^2 = \sigma_1^2$, $\sigma_{\mathrm{L}2}^2 = \sigma_2^2$, and $\lambda_{11} = \sigma_{12}$. Moreover, by using some standard results for the multivariate normal distribution (Muirhead, 1982, p. 43), we obtain $\kappa_{\mathrm{L}1} = \kappa_{\mathrm{L}2} = 0$, $\tau_{\mathrm{L}1}^2 = 2\sigma_1^4$, $\tau_{\mathrm{L}2}^2 = 2\sigma_2^4$, $\lambda_{12} = \lambda_{21} = 0$, and $\lambda_{22} = 2\sigma_{12}^2$. Thus,

we have that

$$\omega_{11} = \frac{\sigma_1^2}{\sigma_1^2} + \frac{\eta_1^2 2\sigma_1^4}{4\sigma_1^4} = 1 + \frac{\eta_1^2}{2} \qquad \text{and} \qquad \omega_{22} = \frac{\sigma_2^2}{\sigma_2^2} + \frac{\eta_2^2 2\sigma_2^4}{4\sigma_2^4} = 1 + \frac{\eta_2^2}{2}$$

as well as

$$\omega_{12} = \frac{\sigma_{12}}{\sigma_1\sigma_2} + \frac{\eta_1\eta_2 2\sigma_{12}^2}{4\sigma_1^2\sigma_2^2} = \rho_{12} + \frac{\eta_1\eta_2\rho_{12}^2}{2}.$$

This leads to the large-sample variance of $\Delta\eta_n$, i.e.,

$$\omega_{11} + \omega_{22} - 2\omega_{12} = 2\left(1 - \rho_{12}\right) + \frac{\eta_1^2 + \eta_2^2 - 2\eta_1\eta_2\rho_{12}^2}{2}\,.$$

B. Correlograms

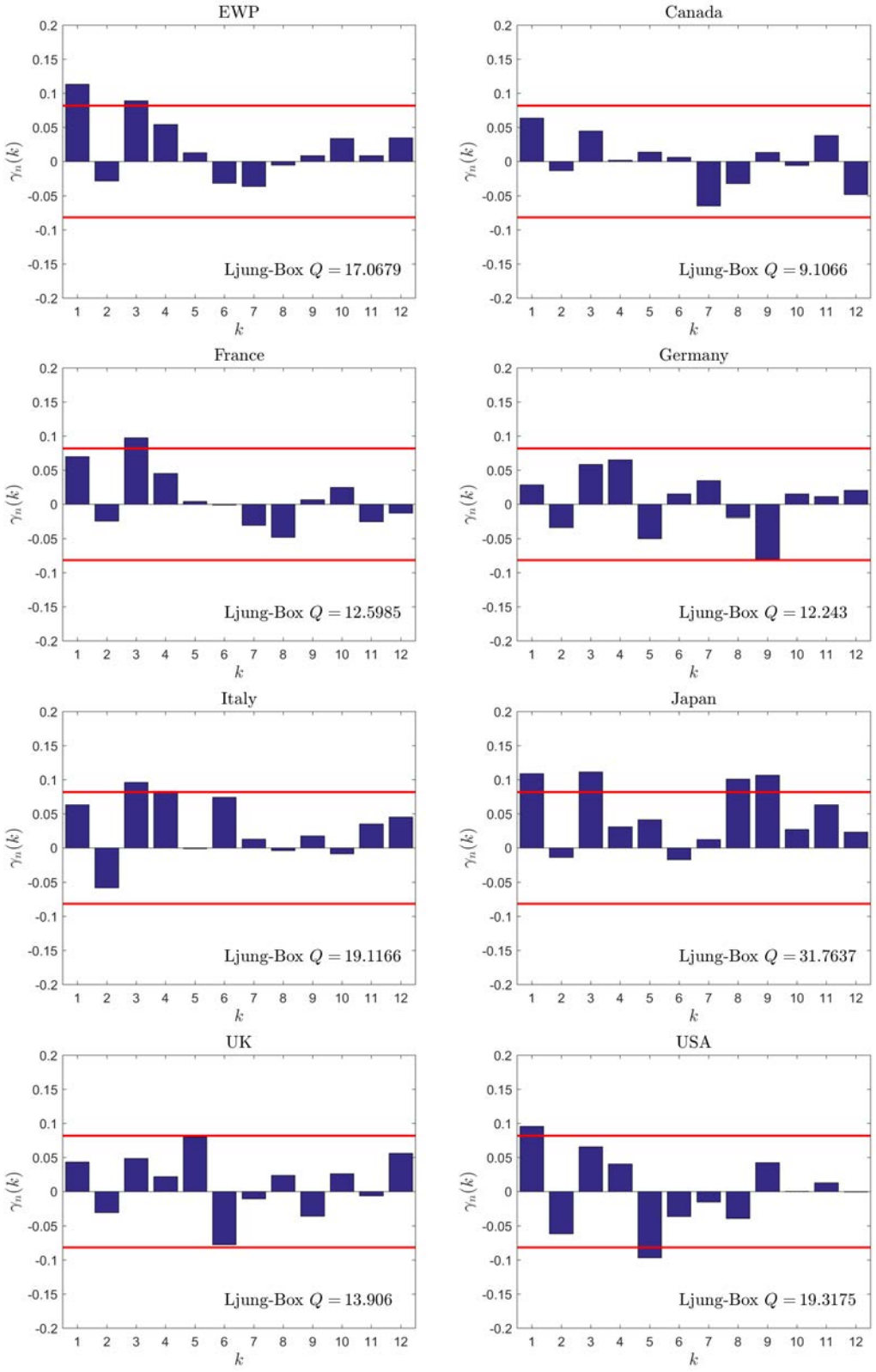

Figure 1: Correlograms with respect to $\{R_t\}$ of the EWP and each G–7 country.

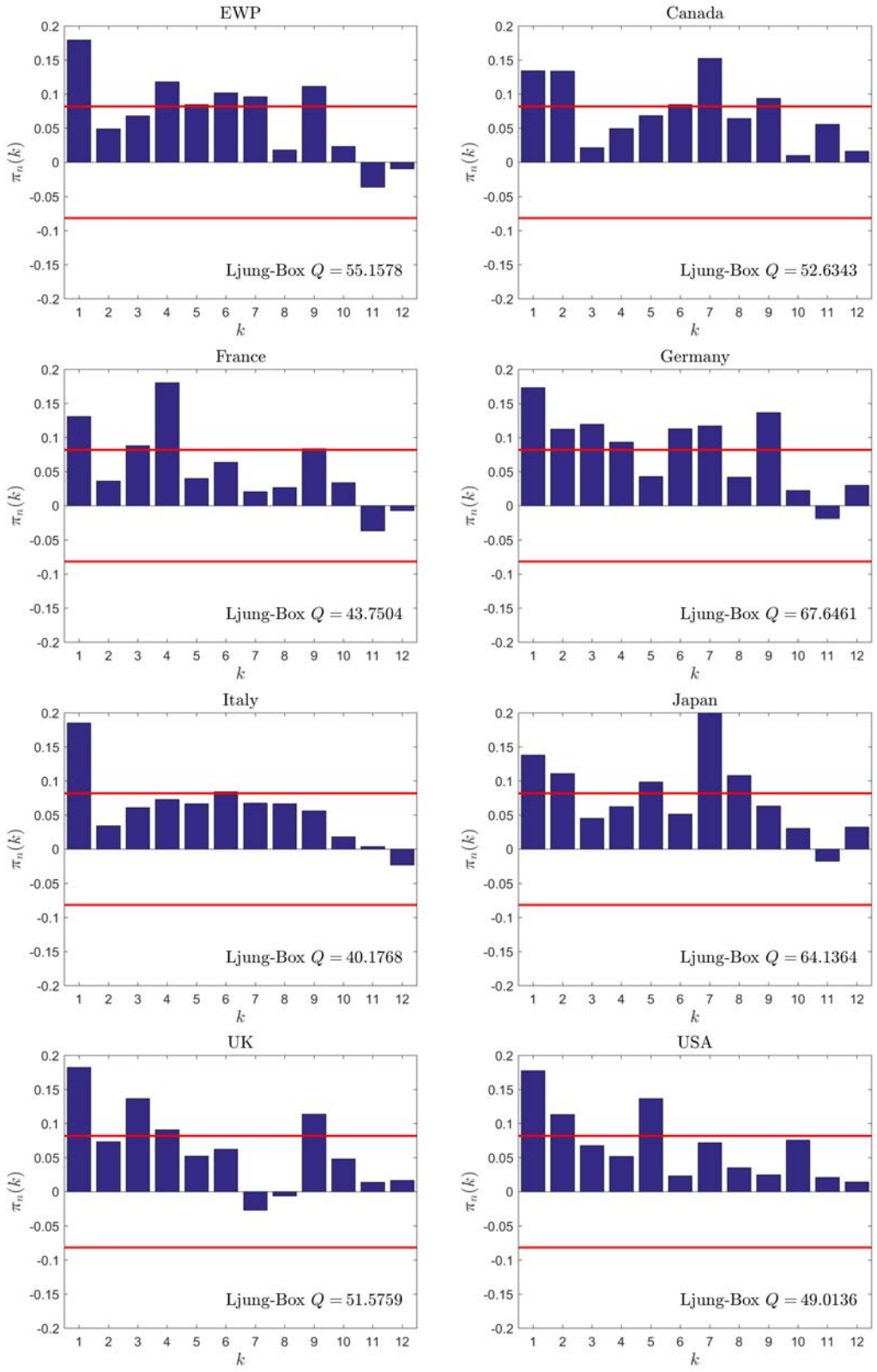

Figure 2: Correlograms with respect to $\{(R_t - \mu_n)^2\}$ of the EWP and each G–7 country.

References

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

Bartosz, S. (2012): "Downside risk approach for multi-objective portfolio optimization," in D. Klatte, H.J. Lüthi, K. Schmedders (editors), "Operations Research Proceedings 2011," Springer, pp. 191–196.

Berger, R. (1997): "Likelihood ratio tests and intersection-union tests," in S. Panchapakesan, N. Balakrishnan (editors), "Advances in Statistical Decision Theory and Applications," Birkhäuser, pp. 225–237.

Black, F. (1976): "Studies of stock price volatility changes," in "Proceedings of the Business and Economics Section of the American Statistical Association," pp. 177–181.

Bradley, R. (2005): "Basic properties of strong mixing conditions. A survey and some open questions," *Probability Surveys* **2**, pp. 107–144.

Brockwell, P., Davis, R. (1991): *Time Series: Theory and Methods*, Springer, 2nd edition.

Burgess, A. (2000): "Statistical arbitrage models of the FTSE 100," in Y. Abu-Mostafa, B. LeBaron, A. Lo, A. Weigend (editors), "Computational Finance," MIT Press, pp. 297–312.

Conrad, J., Kaul, G. (1998): "An anatomy of trading strategies," *The Review of Financial Studies* **11**, pp. 489–519.

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

Eagleson, G. (1975): "On Gordin's central limit theorem for stationary processes," *Journal of Applied Probability* **12**, pp. 176–179.

Frahm, G., Jaekel, U. (2015): "Tyler's M-estimator in high-dimensional financial-data analysis," in K. Nordhausen, S. Taskinen (editors), "Modern Nonparametric, Robust and Multivariate Methods," Chapter 17, Springer, pp. 289–305.

Frahm, G., Wickern, T., Wiechers, C. (2012): "Multiple tests for the performance of different investment strategies," *Advances in Statistical Analysis* **96**, pp. 343–383.

Hamilton, J. (1994): *Time Series Analysis*, Princeton University Press.

Hanke, M., Penev, S. (2018): "Comparing large-sample maximum Sharpe ratios and incremental variable testing," *European Journal of Operational Research* **265**, pp. 571–579.

Hayashi, F. (2000): *Econometrics*, Princeton University Press.

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

Jorion, P. (1985): "International portfolio diversification with estimation risk," *Journal of Business* **58**, pp. 259–278.

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

Lo, A. (2002): "The statistics of Sharpe ratios," *Financial Analysts Journal* **58**, pp. 36–52.

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

Menkhoff, L., Sarno, L., Schmeling, M., Schrimpf, A. (2012): "Currency momentum strategies," *Journal of Financial Economics* **106**, pp. 660–684.

Mertens, E. (2002): "Comments on variance of the iid estimator in Lo (2002)," Technical report, University of Basel.

Muirhead, R. (1982): *Aspects of Multivariate Statistical Theory*, John Wiley.

Opdyke, J. (2007): "Comparing Sharpe ratios: So where are the *p*-values?" *Journal of Asset Management* **8**, pp. 308–336.

Politis, D. (2003): "The impact of bootstrap methods on time series analysis," *Statistical Science* **18**, pp. 219–230.

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

Roy, S. (1953): "On a heuristic method of test construction and its use in multivariate analysis," *Annals of Mathematical Statistics* **24**, pp. 220–238.

Schmid, F., Schmidt, R. (2009): "Statistical inference for Sharpe's ratio," in A. Berkelaar, C. J., K. Nyholm (editors), "Interest Rate Models, Asset Allocation and Quantitative Techniques for Central Banks and Sovereign Wealth Funds," Palgrave Macmillan, pp. 337–357.

Sen, P., Silvapulle, M. (2002): "An appraisal of some aspects of statistical inference under inequality constraints," *Journal of Statistical Planning and Inference* **107**, pp. 3–43.

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

Shen, Q., Szakmary, A., Sharma, S. (2007): "An examination of momentum strategies in commodity futures markets," *Journal of Futures Markets* **27**, pp. 227–256.

Szakmary, A., Shen, Q., Sharma, S. (2010): "Trend-following trading strategies in commodity futures: a re-examination," *Journal of Banking and Finance* **34**, pp. 409–426.

van der Vaart, A. (1998): *Asymptotic Statistics*, Cambridge University Press.

Vrugt, E., Bauer, R., Molenaar, R., Steenkamp, T. (2004): "Dynamic commodity timing strategies," Technical report, SSRN, http://dx.doi.org/10.2139/ssrn.581423.

Zagrodny, D. (2003): "An optimality of change loss type strategy," *Optimization* **52**, pp. 757–772.