Testing Out-of-Sample Portfolio Performance

Ekaterina Kazak[*] Winfried Pohlmeier[†]

University of Konstanz, GSDS University of Konstanz, CoFE, RCEA

September 20, 2018

Abstract

This paper studies the quality of portfolio performance tests based on out-of-sample returns. By disentangling the components of out-of-sample performance we show that observed differences are driven to a large extent by the differences in estimation risk. Our Monte Carlo study reveals that the puzzling empirical findings of inferior performance of theoretically superior strategies mainly result from the low power of these tests. Thus our results provide an explanation why the null hypothesis of equal performance of the simple equally weighted portfolio compared to many alternatives, theoretically superior strategies cannot be rejected in many out-of-sample horse races. Our findings turn out to be robust with respect to different designs and the implementation strategies of the tests.

For the applied researcher we provide some guidance to cope with the problem of low power. In particular, we show by the means of a novel pretest-based portfolio strategy, how the information of performance tests can be used optimally.

Keywords: statistical tests, simulation, finance, bootstrapping, decision making

[*] Department of Economics, Universitätsstraße 1, D-78462 Konstanz, Germany. Phone: +49-7531-88-2204, fax: -4450, email: Ekaterina.Kazak@uni-konstanz.de.

[†] Department of Economics, Universitätsstraße 1, D-78462 Konstanz, Germany. Phone: +49-7531-88-2660, fax: -4450, email: Winfried.Pohlmeier@uni-konstanz.de.

1 Introduction

Estimation risk is a well-known issue in empirical portfolio modeling. For a given performance measure, estimation risk may cause a theoretically superior portfolio strategy to be inferior compared to simple alternatives when it comes to a comparison of the performance measures based on their estimated counterparts. The most prominent example is the equally weighted $(1/N)$ portfolio strategy, for which the null hypothesis of equal out-of-sample performance compared to a more sophisticated, theory based strategy often cannot be rejected at conventional significance levels (DeMiguel et al. (2009b)). While the literature on portfolio choice largely concentrates on robustification strategies (e.g. Brodie et al. (2009), DeMiguel and Nogales (2009), DeMiguel et al. (2009a)) in order to improve the empirical performance, comparatively little attention has been devoted to the statistical quality of performance tests to check, if a given strategy significantly outperforms an alternative one. To our knowledge this is the first paper trying to shed light on the quality of different portfolio performance tests in terms of size distortions and power.

The vast majority of portfolio performance studies either rely on the tests by Jobson and Korkie (1981) in its corrected version by Memmel (2003) or on the bootstrap approach by Ledoit and Wolf (2008). Both tests were originally developed for testing differences in the Sharpe ratios of two asset returns. However, they can also be adopted to compare other performance measures such as the variance (Ledoit and Wolf (2011)) or the certainty equivalent. Another way of testing the performance measure differences is based on the Delta method (DeMiguel et al. (2009b)), which is used to compute the standard errors of the difference in the certainty equivalents (CE) of two competing portfolio strategies. The tests by Jobson and Korkie (1981) and DeMiguel et al. (2009b) depend on the assumption of bivariate normality of the underlying process of portfolio returns. In this case the asymptotic distribution of the difference in performance measures can be written as a function of the first and the second empirical moments of the two return processes in order to obtain the asymptotic standard error. The normality assumption is relaxed in the bootstrap approach of Ledoit and Wolf (2008), which is also valid in settings, where the return distribution is heavy-tailed with heteroskedasticity and/or autocorrelation.

So far little is known about the underlying forces leading to the poor testing results if theory guided performance strategies are compared to the $1/N$-strategy. We therefore take a closer look on the stochastic nature of out-of-sample portfolio returns where the out-of-sample return distribution is a non-standard mixture distribution driven by the true underlying asset return process but also by the distribution of the estimated portfolio weights. This scenario is most common in empirical horse races comparing the out-of-sample performance of two competing portfolio strategies. We show, that the null hypothesis of equal out-of-sample performance of two alternative strategies jointly tests the sum of three different sources: (i) the performance difference implied by the underlying theoretical strategies, (ii) performance differences caused by within-sample estimation risk and (iii) performance differences due to the volatility of out-of-sample portfolio return resulting from the variance in the estimated portfolio weights. In particular, we show that the observed differences in out-of-sample performance are mainly driven by the estimation risk component, which explains that the naive and theoretically inferior $1/N$ strategy can easily outperform a theory based strategy due to the high estimation risk of the latter. Most importantly, the decomposition leads to a clear definition of the null and the alternative hypothesis in terms of the three components and allows us to assess the power of the tests at economically realistic deviations from the null.

In our Monte-Carlo simulations we generally find low power for various popular testing strategies and scenarios when the $1/N$ strategy serves as the benchmark. This explains why the hypothesis of equal performance of the $1/N$ strategy with some other strategy is often not rejected. The low power of the test for realistic (positive) performance differences under the alternative hypothesis suggests a cautious choice of the significance level and the sidedness of the test. In particular, it calls for choosing an optimal trade-off between Type I and Type II error.

Despite the lack of power, information from the tests is not useless. We show that the information provided by performance tests can be utilized within a novel pretest-based strategy where the trade-off between size and power is optimized with respect to the out-of-sample performance. Similar to shrinkage strategies, which combine a given portfolio strategy with the equally weighted portfolio by some optimality criterion (DeMiguel et al. (2009b), Frahm and Memmel (2010)), our pretest estimator uses the information about both strategies through the outcome of the performance test. However, contrary to the shrinkage approaches, our pretest strategy can be continuously updated to incorporate the most recent information in the rolling window forecasting set-up.

The paper is organized as follows. In Section 2 we take a closer look at the distributional properties of portfolio performance measures based on in-sample and out-of-sample returns and analyze the specific distributional properties of out-of-sample portfolio returns in the case of estimated portfolio weights. In Section 3 we provide a Monte-Carlo evidence on the power distortions of conventional portfolio tests. Section 4 proposes alternative strategies for the applied researcher to cope with these deficiencies and deals with the optimal choice of the significance level. In particular, we suggest various pretesting estimators to optimize the gains from portfolio selection by choosing the optimal significance level. Section 5 summarizes the main findings and gives an outlook on future research.

2 Return Distribution and Out-of-sample Performance

2.1 Measuring out-of-sample performance

In the following we argue that measuring and testing portfolio performance on out-of-sample returns needs special consideration due to the stochastic nature of the estimated portfolio weights. We first consider different within- and out-of-sample concepts of measuring portfolio performance, which give rise to a refined definition of the null-hypothesis to be tested. In particular, we will work out the crucial differences between performance measures based on the true parameters (subsection i.), within-sample (subsection ii.) and out-of-sample concepts (subsection iii.) and their corresponding limiting distributions depending on the sample size and the length of the out-of-sample evaluation horizon (subsection iv.).

In what follows assume a portfolio universe of N risky assets and a risk-free asset. Let r_t be an excess return vector at time t with mean vector $\mathrm{E}\left[r_t\right] = \mu$ and variance-covariance matrix $\mathrm{V}\left[r_t\right] = \Sigma$. Moreover, let $\omega(s) = \omega_{(s)}(\mu, \Sigma)$ be the $N \times 1$ vector of portfolio weights for strategy s, e.g. $\omega(g) = \frac{\Sigma^{-1}\iota}{\iota'\Sigma^{-1}\iota}$ for the global minimum variance portfolio (GMVP) minimizing the portfolio variance, $\omega(e) = \frac{1}{N}\iota$ for the equally weighted portfolio and $\omega(m) = \frac{1}{\gamma}\Sigma^{-1}\mu$ for the mean-variance portfolio, maximizing the certainty equivalent (CE) for a given risk aversion parameter γ.

i.) Theoretical performance

For strategy s the portfolio return at time t is given by $r_t^p(s) = \omega(s)' r_t$ with mean $\mu_p(s) = \mathrm{E}\left[r_t^p(s)\right] = \omega(s)'\mu$ and variance $\sigma_p^2(s) = \mathrm{V}\left[r_t^p(s)\right] = \omega(s)'\Sigma\,\omega(s)$. Moreover, denote the performance measure based on strategy s by $\mathcal{P}\big(\mu_p(s), \sigma_p^2(s)\big)$ with the performance difference $\Delta_0(s, \tilde{s}) = \mathcal{P}\big(\mu_p(s), \sigma_p^2(s)\big) - \mathcal{P}\big(\mu_p(\tilde{s}), \sigma_p^2(\tilde{s})\big)$, where s is said to dominate $\tilde{s}$ if $\Delta_0(s, \tilde{s}) \geq 0$.

In the following we restrict our analysis to the certainty equivalent given by $CE(\omega(s)) = \mathcal{P}\big(\mu_p(s), \sigma_p^2(s)\big) = \mu_p(s) - \frac{\gamma}{2}\,\sigma_p^2(s)$, but the general arguments put forward in testing the differences in certainty equivalents of two strategies also hold for other performance measures such as the Sharpe ratio or the portfolio variance. Moreover, the power of tests based on $\Delta_0(s, \tilde{s})$ using the CE can be analyzed under a realistic assumptions of the alternative hypothesis in terms of expected return differences of the two competing strategies.

Note that $\Delta_0(s, \tilde{s})$ is a concept based on population parameters where its domain depends on the underlying performance measure $\mathcal{P}(\cdot)$ and on the specific choice of s and $\tilde{s}$. For instance, for a test comparing the performance of the mean-variance portfolio ($s = m$) with some other strategy ($\tilde{s} \neq m$), the domain of the CE-based performance is always non-negative, $\Delta_0(m, \tilde{s}) \geq 0$. For such a case, where one strategy is dominating another strategy by definition, testing the null hypothesis implies testing on the parameter bound. For a two-sided test of the null of equal performance, $\Delta_0(m, \tilde{s}) = 0$, the parameter space under the alternative, $\Delta_0(m, \tilde{s}) \neq 0$, is inappropriately defined. Moreover, the null of equal performance is economically not very meaningful, as it can only be obtained under rather unrealistic properties of the underlying return process. Similar arguments hold for the GMVP if the performance measure is based on the portfolio variance or the tangency portfolio, if the comparison is based on two Sharpe ratios.

The role of estimation risk becomes clear by considering the performance measure evaluated at the estimated portfolio weight ($\hat{\omega}(s) = \omega_{(s)}(\hat{\mu}, \hat{\Sigma})$). For the CE this is given by

$$CE(\hat{\omega}(s)) = \hat{\omega}(s)'\mu - \frac{\gamma}{2}\,\hat{\omega}(s)'\Sigma\,\hat{\omega}(s). \tag{1}$$

For any given data generating process $CE(\omega(m)) \geq CE(\hat{\omega}(s))$ holds by definition. But due to the randomness of $CE(\hat{\omega}(s))$ none of the empirical strategies can be ordered ex-ante, i.e. $CE(\hat{\omega}(s)) \lesseqgtr CE(\hat{\omega}(\tilde{s}))$ for all s and $\tilde{s}$. Provided the estimator for the portfolio weights is unbiased, the mean of $CE(\hat{\omega}(s))$ takes the form (Cho, 2011):

$$\mathrm{E}\left[CE(\hat{\omega}(s))\right] = CE(\omega(s)) - \frac{\gamma}{2} tr\big(\Sigma\,\mathrm{V}\left[\hat{\omega}(s)\right]\big). \tag{2}$$

Due to estimation risk in the portfolio weights the mean $\mathrm{E}\left[CE(\hat{\omega}(s))\right]$ is lower than the theoretical CE. The difference increases with the degree of risk aversion γ and the estimation uncertainty reflected by the variance of the estimated portfolio weights $\mathrm{V}\left[\hat{\omega}(s)\right]$. Performance measurement based on the expected CE accounts for both the mean-variance trade-off from the financial risks and for the estimation risk. In addition, the inclusion of estimation risk permits a theoretically inferior strategy to be empirically superior due to lower estimation risk. Therefore, a two-sided hypothesis based on the expected performance differences, e.g. $H_0 : \mathrm{E}\left[CE(\hat{\omega}(s))\right] - \mathrm{E}\left[CE(\hat{\omega}(\tilde{s}))\right] = 0$, is meaningful, even if strategy s is strictly dominating $\tilde{s}$ in theory. Finally, note that a null hypothesis based on the differences in the expected empirical performance measures tests the sum of the theoretical performance differences and the differences in estimation risk.

ii.) Within-sample performance

Consider now the case when performance is measured by the estimated CE evaluated at the plug-in estimated portfolio weights, where the mean and the variance of the return vector are replaced by their sample counterparts $\hat{\mu} = \frac{1}{T}\sum_{t=1}^{T} r_t$ and $\hat{\Sigma} = \frac{1}{T-1}\sum_{t=1}^{T}\big(r_t - \hat{\mu}\big)\big(r_t - \hat{\mu}\big)'$. The number of in-sample observations (size of the estimation window) is denoted by T. For the estimated within-sample portfolio return $\hat{r}_t^p(s) = \hat{\omega}(s)' r_t$ we obtain the empirical or within-sample certainty equivalent $\widehat{\mathcal{P}}\big(\hat{\mu}_p(s), \hat{\sigma}_p^2(s)\big)$:

$$\widehat{CE}(\hat{\omega}(s)) = \hat{\mu}_p(s) - \frac{\gamma}{2}\,\hat{\sigma}_p^2(s), \tag{3}$$

where $\hat{\mu}_p(s) = \frac{1}{T}\sum_{t=1}^{T}\hat{r}_t^p(s) = \hat{\omega}(s)'\hat{\mu}$ and $\hat{\sigma}_p^2(s) = \frac{1}{T-1}\sum_{t=1}^{T}\big(\hat{r}_t^p(s) - \hat{\mu}_p(s)\big)^2 = \hat{\omega}(s)'\hat{\Sigma}\,\hat{\omega}(s)$. The estimated within-sample CE difference is defined as $\hat{\Delta}(s,\tilde{s}) = \widehat{CE}(\hat{\omega}(s)) - \widehat{CE}(\hat{\omega}(\tilde{s}))$. Since $\widehat{CE}(\hat{\omega}(s))$ is a random variable the portfolio performance difference based on the empirical concepts is also random and may be larger or smaller than zero. However, for consistent estimates of the true portfolio weights, $\widehat{CE}(\hat{\omega}(s))$ is a consistent estimator of $CE(\omega(s))$ such that $\operatorname*{plim}_{T\to\infty} \hat{\Delta}(s,\tilde{s}) = \Delta_0(s,\tilde{s})$, i.e. for large samples the problem of testing on the parameter bound for some portfolio comparisons persists.

A properly defined null and alternative hypotheses in terms of fixed parameters taking into account estimation risk could be based on the expected difference of the empirical performance measures, $\mathrm{E}\left[\widehat{\mathcal{P}}\big(\hat{\mu}_p(s), \hat{\sigma}_p^2(s)\big) - \widehat{\mathcal{P}}\big(\hat{\mu}_p(\tilde{s}), \hat{\sigma}_p^2(\tilde{s})\big)\right]$. For $\widehat{CE}(\hat{\omega}(s))$ this would be:

$$\mathrm{E}\left[\widehat{CE}(\hat{\omega}(s))\right] = \mathrm{E}\left[\hat{\omega}(s)'\hat{\mu}\right] - \frac{\gamma}{2}\,\mathrm{E}\left[\hat{\omega}(s)'\hat{\Sigma}\,\hat{\omega}(s)\right]. \tag{4}$$

Contrary to (1), which depends on the unknown first and second moments of the return process, $\widehat{CE}(\hat{\omega}(s))$ is feasible. However, its mean differs substantially from (2) due to its strong nonlinearity in $\hat{\Sigma}$ and $\hat{\mu}$. Besides the difference due to theoretical performance and estimation risk concerning $\omega(\cdot)$ the null hypothesis based on (4) includes in addition the estimation risk related to $\widehat{\mathcal{P}}$.

iii.) Out-of-sample performance

In the following we consider a typical rolling window set-up, where for period $t+1$ the out-of-sample portfolio return $\hat{r}_{t+1}^p(s)$ is based on a one-step ahead forecast of the portfolio weights $\hat{\omega}_{t+1|t}(s)$ with period $\{t-T,\ldots,t\}$ as the estimation window. We adopt the standard assumption for static models that the last available estimate $\hat{\omega}_t(s)$ is used to compute the out-of-sample return for the next period, $\hat{r}_{t+1}^p(s) = \hat{\omega}_{t+1|t}(s)'r_{t+1} = \hat{\omega}_t(s)'r_{t+1}$. Assuming independence for the return process, r_{t+1} and $\hat{\omega}_t(s)$ are independent. Population mean and variance of the out-of-sample portfolio returns (*op*) are given by

$$\begin{aligned}
\mu_{op}(s) &= \mathrm{E}\left[\hat{r}_{t+1}^p(s)\right] = \mathrm{E}\left[\hat{\omega}_t(s)\right]'\mu,\\
\sigma_{op}^2(s) &= \mathrm{V}\left[\hat{r}_{t+1}^p(s)\right] = \mathrm{E}\left[\hat{\omega}_t(s)'\Sigma\,\hat{\omega}_t(s)\right] + \mu'\,\mathrm{V}\left[\hat{\omega}_t(s)\right]\mu.
\end{aligned}$$

The variance of the out-of-sample return process, $\sigma_{op}^2(s)$, reveals the typical dual character. The first term represents the additional volatility resulting from estimation uncertainty concerning $\omega(s)$, while the second term captures the volatility resulting from the fact that the expectation of the out-of-sample return has to be estimated. Provided unbiased estimation of the portfolio weights, the theoretical out-of-sample CE takes the form:

$$CE_{op}(\hat{\omega}_t(s)) = \mu_{op}(s) - \frac{\gamma}{2}\sigma^2_{op}(s) = \mathrm{E}\left[CE(\hat{\omega}_t(s))\right] - \frac{\gamma}{2}\mu' \,\mathrm{V}\left[\hat{\omega}_t(s)\right]\mu,$$

where $\mathrm{E}\left[CE(\hat{\omega}_t(s))\right] = CE(\omega(s)) - \frac{\gamma}{2}tr\big(\Sigma\,\mathrm{V}\left[\hat{\omega}(s)\right]\big)$ is given by (2). Compared to the theoretical CE defined in i.) there is a double penalization on theoretical out-of-sample CE through the within-sample estimation risk given by the term $-\frac{\gamma}{2}tr\big(\Sigma\,\mathrm{V}\left[\hat{\omega}(s)\right]\big)$ and the out-of-sample risk given by $-\frac{\gamma}{2}\mu'\,\mathrm{V}\left[\hat{\omega}_t(s)\right]\mu$. The out-of sample difference in portfolio performance is given by

$$\Delta_{op}(s,\tilde{s}) \equiv \Delta_0(s,\tilde{s}) - \frac{\gamma}{2}\Big[tr(\Sigma\,\mathrm{V}\left[\hat{\omega}_t(s)\right]) - tr(\Sigma\,\mathrm{V}\left[\hat{\omega}_t(\tilde{s})\right])\Big] - \frac{\gamma}{2}\mu'\Big[\mathrm{V}\left[\hat{\omega}_t(s)\right] - \mathrm{V}\left[\hat{\omega}_t(\tilde{s})\right]\Big]\mu. \quad (5)$$

For the case of the equally weighted portfolio as the benchmark portfolio, $\tilde{s} = e$, $\Delta_{op}(s,\tilde{s})$ simplifies, since $\mathrm{V}\left[\hat{\omega}_t(\tilde{s})\right] = 0$. This drives the strong out-of-sample performance of the equally weighted portfolio compared to empirical portfolio strategies relying on imprecisely estimated portfolio weights. Following Kempf and Memmel (2006) and Okhrin and Schmid (2006) under normally distributed returns r_t the variance of the estimated GMVP weights can be written as:

$$\mathrm{V}\left[\hat{\omega}_t(g)\right] = \frac{1}{T-N-1}\frac{1}{\iota'\Sigma^{-1}\iota}\left(\Sigma^{-1} - \frac{\Sigma^{-1}\iota\iota'\Sigma^{-1}}{\iota'\Sigma^{-1}\iota}\right),$$

where ι denotes $N \times 1$ vector of ones, N is the number of assets and T is the in-sample estimation window length. Thus the expression for the out-of-sample CE difference for the GMVP and equally weighted portfolio simplifies to

$$\Delta_{op}(g,e) = \Delta_0(g,e) - \frac{\gamma}{2}\frac{1}{T-N-1}\frac{1}{\iota'\Sigma^{-1}\iota}\left[N - 1 + \mu'\Sigma^{-1}\mu - \frac{(\mu'\Sigma^{-1}\iota)^2}{\iota'\Sigma^{-1}\iota}\right].$$

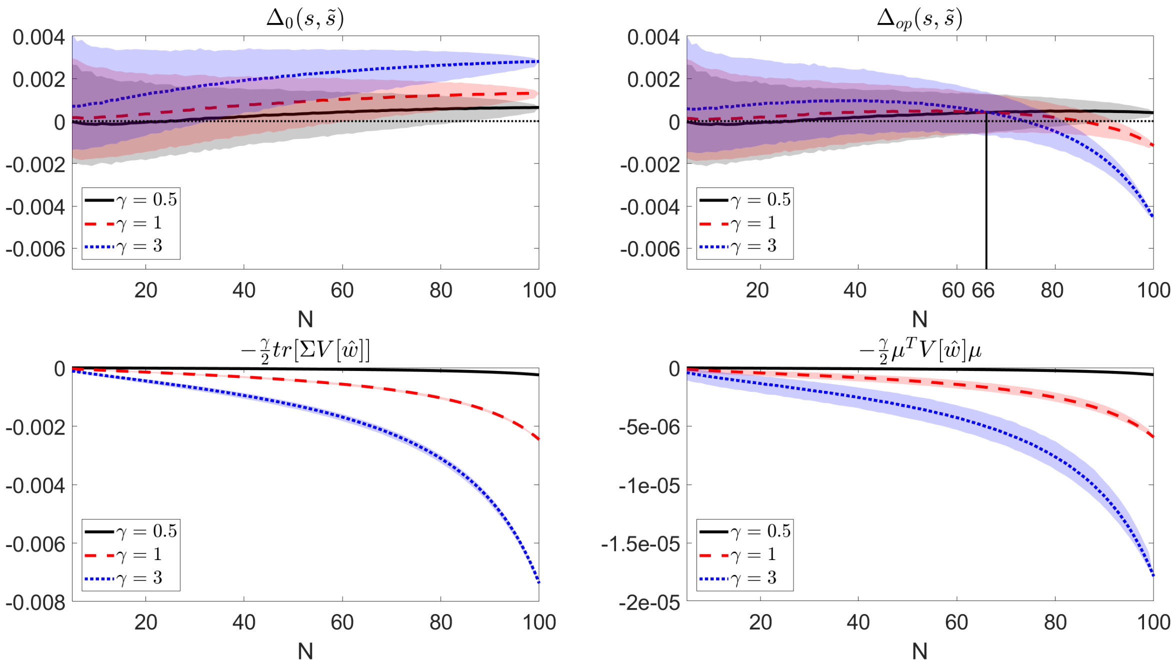
Figure 1 provides insights into the quantitative importance of the the components of the out-of-sample CE differences for the GMVP and the equally weighted portfolio for different portfolio dimensions. In particular for a given portfolio size N we randomly draw an integer index without replacement, which determines the selection of assets from a pool of a hundred assets[1]. We then compute the sample mean and covariance matrix from the K.R.French dataset of the selected assets and treat them as the theoretical μ and Σ respectively. We then compute all the

[1] http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html, monthly excess returns of 100 industry portfolios.

ingredients of the out-of-sample CE difference following (5) and plot the average of those over 5 000 replications for each N. The shaded areas around the mean curves correspond to the 95% confidence intervals formed as the 2.5% and 97.5% quantiles over the 5 000 randomly drawn portfolios.

The upper-left plot depicts the difference in theoretical CE for GMVP and equally weighted portfolio for different number of assets N and different values of the risk aversion parameter γ. Unsurprisingly, the larger the asset universe, the stronger the diversification effect, so that the dominance of the theoretical CE of the GMVP increases over the equally weighted portfolio. The dominance also increases with higher risk aversion as the diversification effect becomes more relevant with increasing γ.

Figure 1: CE and Out-of-sample CE Differences



Average CE differences over 5 000 random N out of 100 asset combinations for GMVP and equally weighted portfolio by dimension of the asset universe N for different values of the risk aversion parameter γ. Shadow regions correspond to the 95% confidence intervals over the 5 000 randomly drawn portfolios. The estimation window length T is set to 120 (10 years of monthly observations). Upper-left plot: difference in theoretical CE, $\Delta_0(s, \tilde{s})$. Lower-left plot: estimation noise penalty. Lower-right plot: out-of-sample risk penalty. Upper-right plot: overall out-of-sample CE difference as given in (5).

The upper-right plot shows the out-of-sample CE difference for the two portfolio strategies as given by (5). Compared to the theoretical CE the out-of-sample CE's are substantially smaller. The strategy yielding the highest theoretical CE ($\gamma = 3$, dotted blue line), dominates the other strategies only slightly and only if the asset universe is small. Moreover, for most γ-values the difference is negative indicating the dominance of the equally weighted portfolio. The lower panels provide more insight into the pathology of this finding. The lower-left plot depicts the size of the within-sample estimation risk (second term of (3)), which due to its relative size clearly determines the shape of the theoretical out-of-sample curves. The larger the asset universe the lower the precision of the weight estimates.

The size of risk aversion parameter plays a crucial role for the out-of-sample performance. It determines how strongly the out-of-sample CE is penalized by the presence of estimation noise. Consequently, the choice of a very low risk aversion parameter is not just a matter of economic reasoning, it seriously twists the empirical findings in favor of theory based approaches. Therefore, for the sake of scientific clarity, findings of empirical horse races should be reported for a range of γ-values. The plot on the lower-right depicts the out-of-sample risk penalty. This effect points in the same direction as the within-sample effect. However, the dimension of this effect is considerably smaller. All in all from Figure 1 we can conclude that the out-of-sample performance of tests against the equally weighted portfolio strategy are to a large extend driven by the estimation risk. The performance of the equally weighted portfolio mainly results from boiling down estimation risk to zero at a comparatively small cost of following a theoretical inferior strategy.

iv.) Empirical out-of-sample performance

The empirical out-of-sample performance measure is widely used in applied work to compare the performance of different portfolio strategies. Here $\mu_{op}(s)$ and $\sigma^2_{op}(s)$ are replaced by their sample counterparts based on data from an evaluation sample of size H:

$$\begin{aligned} \widehat{CE}_{op}(\hat{\omega}_t(s)) =& \hat{\mu}_{op}(s) - \frac{\gamma}{2}\hat{\sigma}^2_{op}(s), \qquad (6) \\ \text{where:} \qquad \hat{\mu}_{op}(s) =& \frac{1}{H}\sum_{h=1}^{H} \hat{r}^p_{t+h}(s) = \frac{1}{H}\sum_{h=1}^{H} \hat{\omega}_{t+h-1}(s)' r_{t+h}, \\ \hat{\sigma}^2_{op}(s) =& \frac{1}{H-1}\sum_{h=1}^{H} \left(\hat{r}^p_{t+h}(s) - \hat{\mu}_{op}(s)\right)^2. \end{aligned}$$

The large sample properties of $\hat{\mu}_{op}(s)$ and $\hat{\sigma}^2_{op}(s)$ depend on which type of limiting behavior is considered. Obviously, with an increasing sample size T, holding the size of evaluation window fixed, we obtain

$$\operatorname*{plim}_{\substack{T\to\infty\\ H=const.}} \hat{\mu}_{op}(s) = \omega(s)'\bar{r}_H \qquad \text{and} \qquad \operatorname*{plim}_{\substack{T\to\infty\\ H=const.}} \hat{\sigma}^2_{op}(s) = \omega(s)'\hat{\Sigma}_H\omega(s)', \tag{7}$$

where $\bar{r}_H$ and $\hat{\Sigma}_H$ denote the sample mean and the sample covariance for a sample of size H. Both estimators are converging to random variables whose variance depends on the size of the evaluation window. For $T \to \infty$ and H fixed $\widehat{CE}_{op}(\hat{\omega}_t(s))$ remains a random variable due to the sampling variation present in $\bar{r}_H$ and $\hat{\Sigma}_H$. Therefore, opting solely for a large sample size T is not really helpful, if evaluation window H is small. Moreover, due to the presence of potential structural breaks, it is also reasonable for applied researchers to base the weight estimates on more recent samples.

Alternatively, consider the limiting case for the evaluation window holding T fixed. Under the assumption of independence this yields:

$$\operatorname*{plim}_{\substack{T=const.\\ H\to\infty}} \hat{\mu}_{op}(s) = \mathrm{E}\left[\hat{r}^p_{t+h}(s)\right] = \mu_{op}(s) \qquad \text{and} \qquad \operatorname*{plim}_{\substack{T=const.\\ H\to\infty}} \hat{\sigma}^2_{op}(s) = \sigma^2_{op}(s). \tag{8}$$

Only for the case of $H \to \infty$ and $T \to \infty$ the empirical out-of-sample $\widehat{CE}_{op}(\hat{\omega}_t(s))$ converges to the theoretical $CE(\omega(s))$, as with increasing estimation sample size the estimation risk vanishes.

In order to stabilize the out-of-sample portfolio returns it is common to hold the estimated portfolio weights fixed over a certain range of the evaluation window (Golosnoy and Okhrin (2007)). By changing the weights less frequently than in the rolling window scenario assumed above, the volatility of the out-of-sample return process is reduced. However, such a strategy is somewhat problematic, if the goal is to obtain a consistent estimate of CE_{op}. This becomes obvious by considering the most extreme case when the whole out-of-sample return series is based on a single estimate of the weight vector $\hat{r}^p_{t+h}(s) = \hat{\omega}_t(s)'r_{t+h}$. Contrary to (8) the mean and the variance of the out of sample return remain random variables with

$$\operatorname*{plim}_{\substack{T=const.\\ H\to\infty}} \hat{\mu}_{op}(s) = \hat{\omega}_t(s)'\mu \qquad \text{and} \qquad \operatorname*{plim}_{\substack{T=const.\\ H\to\infty}} \hat{\sigma}^2_{op}(s) = \hat{\omega}_t(s)'\Sigma\,\hat{\omega}_t(s). \tag{9}$$

In this case the corresponding limiting CE remains a random variable, where $\hat{\omega}_t(s)$ serves as a time invariant random effect.

In the following we use the out-of sample CE difference based on (5) to define the null hypothesis of equal performance of two strategies as well as for defining the deviation from the null for our power analysis. This theoretical parameter to be tested is estimated by $\hat{\Delta}_{op}(s,\tilde{s}) = \widehat{CE}_{op}(\hat{\omega}_t(s)) - \widehat{CE}_{op}(\hat{\omega}_t(\tilde{s}))$ as defined in (6). As shown below, the bivariate out-of-sample return processes $(\hat{r}^p_{t+h}(s), \hat{r}^p_{t+h}(\tilde{s}))'$, on which $\hat{\Delta}_{op}(s,\tilde{s})$ is estimated, reveals specific properties, which deviate from the return processes usually assumed in simulation studies.

2.2 Out-of-sample return distribution

The testing procedure by Ledoit and Wolf (2008) is most frequently used to test portfolio performance. They propose to estimate the confidence bands for the difference in two portfolio performance measures $\hat{\Delta}$ and to check via the bootstrap how often the parameter value $\Delta = 0$ lies within the estimated confidence bands. The Monte-Carlo evidence provided by Ledoit and Wolf (2008) rests on the assumption that the joint distribution of the two competing return series comes from a bivariate data generating process with marginal distributions belonging to the same family. They argue that such an assumption is reasonable, if the two return series under consideration are, for instance, return series of two hedge funds (e.g. a bivariate GARCH process). However, a large fraction of the literature (DeMiguel et al. (2009b), Brodie et al. (2009), Antoine (2012)) is concerned with the out-of sample performance of portfolio strategies based on the same set of assets. As shown below, the fact that the out-of-sample returns are generated from the same underlying return distribution implies certain properties that should be accounted for in the Monte-Carlo design. In the following we will focus on the out-of sample return series of two portfolio strategies based on estimated portfolio weights. In such a scenario the out-of-sample portfolio return series follow a mixture distribution depending on the return vector, but also on the estimated portfolio weights. The distribution of the estimated portfolio weights, however, depends strongly on the portfolio strategy as well as on the estimator chosen.

We illustrate the argument by comparing the out-of-sample return series for two competing portfolio strategies s and $\tilde{s}$ whose performance is to be tested. The distribution function of the portfolio return for the same strategy based on estimated weights, $\hat{r}^p_{t+1}(s) = \hat{\omega}_t(s)'r_{t+1}$, takes the form

$$f\big(\hat{r}^p_{t+1}(s)\big) = f\big(\hat{\omega}_t(s)'r_{t+1}\big) = f\big(\hat{\omega}_t(s)'r_{t+1}|\hat{\omega}_t(s)\big)\cdot g\big(\hat{\omega}_t(s)\big), \tag{10}$$

where $g\big(\hat{\omega}_t(s)\big)$ is the marginal distribution of the portfolio weight estimator. Depending on the portfolio strategy and the chosen estimator the marginal distribution varies substantially. Assuming $r_t \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \Sigma)$ for the return process, Okhrin and Schmid (2006) show that the (plug-in) estimator for the Global Minimum Variance Portfolio (GMVP) weights follow a multivariate elliptical t-distribution. The GMVP return conditional on the estimated weights would still be normally distributed with

$$\hat{r}^p_{t+1}(s)\big|_{\hat{\omega}_t(s)} \sim \mathcal{N}\left(\hat{\omega}_t(s)'\mu, \hat{\omega}_t(s)'\Sigma\hat{\omega}_t(s)\right),$$

but the unconditional distribution (10) becomes fat tailed through the estimation uncertainty reflected by $g\big(\hat{\omega}_t(s)\big)$. Obviously for any strategy with known portfolio weights, such as the equally-weighted ($1/N$) strategy, the distribution remains normal:

$$f(r^p_{t+1}(\tilde{s})) = f\big(\omega_t(\tilde{s})'r_{t+1}\big) = f\big(\frac{1}{N}\iota' r_{t+1}\big) = f\big(\frac{1}{N}\sum_{j=1}^{N} r_{j,t+1}\big).$$

An extreme case of the out-of-sample return distribution occurs for the plug-in estimator of the tangency portfolio, where the distribution of the portfolio weight estimates has no first and second moments and follow a multivariate type of Cauchy distribution (Okhrin and Schmid, 2006)[2].

Consider now the bivariate distribution for an out-of-sample return process for strategy s based on estimated portfolio weights and strategy $\tilde{s}$ based on non-stochastic portfolio weights. The joint distribution of the portfolio returns is given by a bivariate mixture distribution of the form

$$f\big(\hat{r}^p_{t+1}(s), r^p_{t+1}(\tilde{s})\big) = f\big(\hat{\omega}_t(s)'r_{t+1}, \omega_t(\tilde{s})'r_{t+1}|\hat{\omega}_t(s)\big)\cdot g\big(\hat{\omega}_t(s)\big). \tag{11}$$

[2] Note that in the literature the use of the term tangency portfolio is not unique. Here we follow, e.g. Britten-Jones (1999) and define the tangency portfolio as the one which is tangent to the minimum-variance bound such that the weights add up one, i.e. $\omega(tan) = \frac{\Sigma^{-1}\mu}{\iota'\Sigma^{-1}\mu}$.

For the case of i.i.d. normality of the underlying return process r_t the bivariate density conditional on the estimated weights is given by

$$\left.\begin{pmatrix} \hat{r}^p_{t+1}(s) \\ r^p_{t+1}(\tilde{s}) \end{pmatrix}\right|_{\hat{\omega}_t(s)} \sim \mathcal{N}\left(\begin{bmatrix} \hat{\omega}_t(s)'\mu \\ \omega_t(\tilde{s})'\mu \end{bmatrix}, \begin{bmatrix} \hat{\omega}_t(s)'\Sigma\hat{\omega}_t(s) & \hat{\omega}_t(s)'\Sigma\,\omega_t(\tilde{s}) \\ \omega_t(\tilde{s})'\Sigma\hat{\omega}_t(s) & \omega_t(\tilde{s})'\Sigma\,\omega_t(\tilde{s}) \end{bmatrix}\right).$$

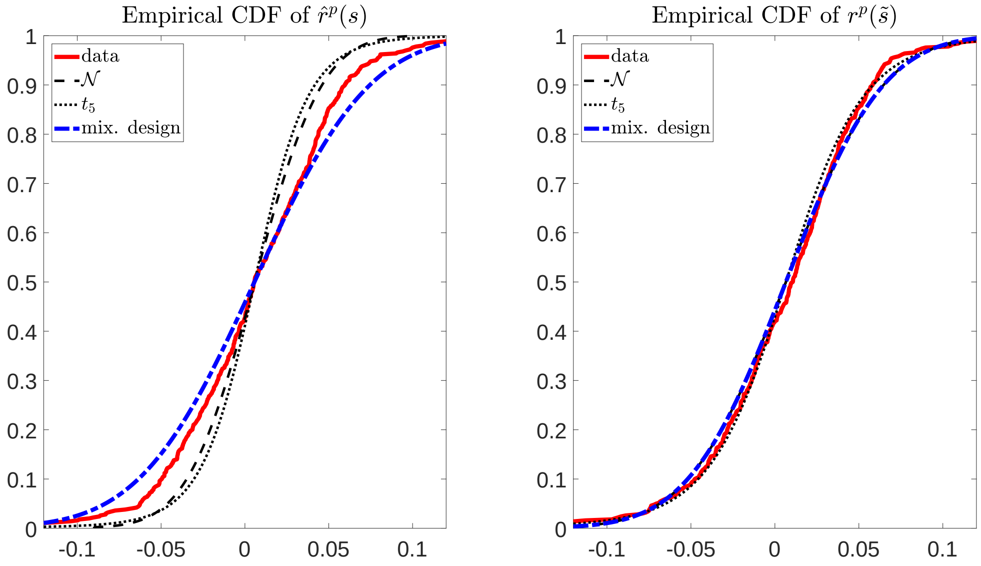
The mixture distribution (11) results from the underlying estimation strategy of the weights, provided that the weights of $\tilde{s}$ are non-stochastic. Note, that the joint unconditional distribution incorporates the time varying correlation between the out-of-sample portfolio returns trough $\hat{\omega}_t(s)$. Furthermore, the mixture distribution out-of-sample portfolio returns implies fat tails even under i.i.d. normality of the original return process r_t and without imposing any GARCH structure.

Appendix A.1 reports descriptive statistics of the monthly return data on 30 industry portfolios from K.R. French's website for the period from January 1953 until December 2015 and the properties of the out-of-sample portfolio returns based on this dataset. We consider 3 portfolio allocation strategies: equally weighted portfolio ($s = e$), Global Minimum Variance Portfolio ($s = g$), the ridge covariance matrix estimator combined with the GMVP ($s = g(\lambda)$) where the penalty parameter λ is chosen according to Ledoit and Wolf (2003)).

In Table 6 we report the average descriptive statistics of the returns in the dataset alongside the properties of the average return, which has a significant sample mean of 0.7%, is left-skewed and heavy-tailed. The GMVP return is not skewed but still heavy-tailed. Ridging the covariance matrix shrinks the GMVP weights to the $1/N$, therefore the empirical properties of the ridged GMVP return are closer to the ones of the equally weighted portfolio. Furthermore, as can be seen on Figure 5, the out-of-sample portfolio returns do not possess any significant autocorrelation or partial autocorrelation.

We now compare the proposed mixture distribution given in (11) with the simulation designs commonly used in the literature in terms of the ability to capture the properties of the empirical out-of-sample portfolio returns. Figure 2 depicts empirical cumulative distribution functions of the out-of-sample portfolio returns based on the actual data together with the average ECDF of the simulated out-of-sample returns. The left panel corresponds to the ECDF of the GMVP returns and the right panel corresponds to the returns of the equally weighted portfolio. The

Figure 2: ECDF of the out-of-sample portfolio returns: N = 30



Empirical cumulative distribution function of the out-of-sample portfolio returns for the GMVP ($\hat{r}^p(s)$) and equally weighted portfolio ($\hat{r}^p(\tilde{s})$): based on the real data (red), average ECDF of the simulated from bivariate t_5 (dash black), simulated from bivariate normal (dotted black) and simulated from the proposed mixture design (dash-dot blue). The mean and standard deviation of the simulated returns are adjusted to be the same as of the empirical portfolio returns. The number of simulations is 50 000.

black dotted lines correspond to the average CDF of the portfolio return for a given strategy when the returns are drawn from the bivariate normal distribution using the sample mean and sample covariance matrix of the actual out-of-sample portfolio returns. The black dashed lines correspond to the CDF of a bivariate t-distribution with 5 degrees of freedom, where the mean and standard deviation has been adjusted as well. The dash-dot blue lines depict the CDF for the proposed mixture design in (11). The distribution fit for the right panel corresponding to the equally weighted portfolio returns is very good for all the designs, whereas for the GMVP the proposed mixture design captures the ECDF shape better than the other designs.

In Table 7 we report empirical moments of the out-of-sample portfolio returns, computed on the real data and empirical moments of the simulated out-of-sample returns. The main difference in the simulation designs comes from the differences in the third and fourth empirical moments. If the out-of-sample returns are drawn from a bivariate normal distribution, the sample skewness and sample kurtosis tend to be smaller than the desired values. On the other hand, returns generated from the bivariate t-distribution reveal too much leptokurtosis and skewness compared to the empirical out-of-sample returns. Our conclusions are supported by the rejection rate of the two-sample Kolmogorov-Smirnov test, which compares the ECDF of the real portfolio returns with the simulated ones. The mixture design has a substantially lower rejection rate compared to the bivariate t- and normal distribution, where the null hypothesis is rejected only in 39% of the draws in comparison to a unit rejection probability for the competing simulation designs. As the proposed mixture design captures the empirical properties of the out-of-sample returns more accurately than the competing designs, we use it in our simulation study when addressing the size and power properties of portfolio performance tests.

3 Size and Power of Performance Tests

3.1 Monte Carlo Design

In the following we examine the statistical properties of the out-of-sample portfolio performance tests by means of a Monte Carlo study. We concentrate on the performance tests of different empirical strategies against the equally weighted portfolio $\omega(\tilde{s}) = \omega(e) = \frac{1}{N}\iota$. This strategy is frequently taken as a benchmark in many empirical applications and often cannot be significantly outperformed by more sophisticated portfolio strategies. As a competing strategy we first consider the plug-in estimated GMVP $\hat{\omega}(s) = \hat{\omega}(g) = \dfrac{\hat{\Sigma}^{-1}\iota}{\iota'\hat{\Sigma}^{-1}\iota}$, which is particularly appealing for a simulation study as a closed-form solution of the distribution function $f(\hat{\omega}(g))$ exists from which $\hat{\omega}(g)$ can be drawn. The second alternative strategy to be considered is the GMVP combined with the ridge estimator of the covariance matrix, $\hat{\omega}(g(\lambda)) = \dfrac{(\hat{\Sigma} + \lambda I_N)^{-1}\iota}{\iota'(\hat{\Sigma} + \lambda I_N)^{-1}\iota}$, where λ denotes the shrinkage parameter and I_N the identity matrix of dimension N. In the following we use the Ledoit and Wolf (2003) approach for choosing the optimal shrinkage intensity of the covariance matrix.

The simulation study below is conducted as follows. Under normality of the underlying return process, $r_t \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \Sigma)$, the vector of portfolio weights is drawn according to Okhrin and Schmid (2006) from the multivariate elliptical t-distribution with parameters (Σ, N, T) from which, in a second step, the bivariate vector of out-of-sample portfolio returns is drawn according to the mixture distribution in (11). The parameters (μ, Σ) for the return process are set to the sample mean and variance of the monthly return data on 30 industry portfolios of Kenneth R. French[3]. We define the expected CE difference of the two competing strategies s and $\tilde{s}$ under the

[3] http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

null according to (5). Under the null the difference in the out-of-sample CE's is a pre-specified value Δ^*. Thus the hypothesis to be tested is $\Delta_{op}(s, \tilde{s}) - \Delta^* = 0$, where typically $\Delta^* = 0$ is assumed.

When addressing the power properties of the tests, we simulate the out-of-sample returns such that under the alternative the actual CE difference exceeds the CE difference hypothesized under the null by 1% in annual terms. We control the data generating process by decreasing the volatility and the corresponding covariances of one particular asset. The goal of being able to protect a 1% annual return difference does not seem overly ambitious given that the annual out-of-sample CE of the equally weighted portfolio is around 6% for our sample.

Table 8 in Appendix A.2.2 reports the values of the expected CE difference under the null and the alternative and its annualized counterparts for a given pair of strategies, risk aversion γ and estimation window length T. For example, for $N/T = 0.01$ and $\gamma = 3$ the monthly expected CE difference between GMVP and equally weighted portfolio under the null is 0.08% and it becomes 0.16% under the alternative. These differences correspond to a 0.94% annual difference under the null and 1.94% under the alternative. Thus in monthly terms under the alternative the underlying CE difference is at least twice as large as under the null and one would expect the test to reject the null hypothesis in such a set up.

In order to check for the impact of estimation noise on the performance of the tests, we consider different N/T ratios. Figure 6 depicts the histograms of the variance of the GMVP portfolio returns for different estimation noise ratios. Clearly, the larger the ratio, the more skewed the distribution of the portfolio variance. These skewness properties translate into the distribution of the out-of-sample CE difference, scaled by $-\frac{\gamma}{2}$. Therefore, the larger the risk aversion parameter the more left-skewed the distribution of the CE difference. As we will show below, the skewness property leads to size distortions for some tests particularly if two sided hypotheses are to be tested.

The significance of the CE difference is tested by the means of the Student t-test. The standard error of the CE difference is either computed using the Delta method or the bootstrap.

The Delta method for testing the difference in CE of the two strategies $(s, \tilde{s})$ is given by:

$$\Delta_{op} = \phi(\vartheta) = (\mu_{op}(s) - \frac{\gamma}{2}\sigma^2_{op}(s)) - (\mu_{op}(\tilde{s}) - \frac{\gamma}{2}\sigma^2_{op}(\tilde{s})) \tag{12}$$

$$\sqrt{H}\big(\hat{\Delta}_{op}(s,\tilde{s}) - \Delta_{op}(s,\tilde{s})\big) \xrightarrow{d} \mathcal{N}\left(0, \frac{\partial\phi(\vartheta)'}{\partial\vartheta} V[\hat{\vartheta}] \frac{\partial\phi(\vartheta)}{\partial\vartheta}\right), \tag{13}$$

where $\vartheta = (\mu_{op}(s), \sigma^2_{op}(s), \mu_{op}(\tilde{s}), \sigma^2_{op}(\tilde{s}))'$ is the vector of the mean and variance of portfolio returns, the covariance matrix $V[\hat{\vartheta}]$ has a well-known form (DeMiguel et al. (2009b))[4].

Another way of obtaining the standard error for the CE difference is the non-parametric bootstrap, where the bivariate vector of the out-of-sample portfolio returns is sampled with replacement, resulting in the empirical distribution of the CE difference and the t-statistics. We first consider the percentile bootstrap, where the null is rejected if $\hat{\Delta}_{op}(s,\tilde{s}) - \Delta^* > q_\Delta(1-\alpha/2)$ or $\hat{\Delta}_{op}(s,\tilde{s}) - \Delta^* < q_\Delta(\alpha/2)$, where $q_\Delta(\cdot)$ is the quantile of the bootstrap distribution of $\hat{\Delta}^b_{op}(s,\tilde{s}) - \hat{\Delta}_{op}(s,\tilde{s})$, with $\hat{\Delta}^b_{op}(s,\tilde{s})$ denoting the estimated out-of-sample CE difference of the bootstrap sample. Another way of utilizing the bootstrap distribution is approximating the quantiles of the test statistic. For this case the null hypothesis is rejected if $\frac{\hat{\Delta}_{op}(s,\tilde{s}) - \Delta^*}{s.e.[\hat{\Delta}_{op}(s,\tilde{s})]} > q_t(1-\alpha/2)$ or if $\frac{\hat{\Delta}_{op}(s,\tilde{s}) - \Delta^*}{s.e.[\hat{\Delta}_{op}(s,\tilde{s})]} < q_t(\alpha/2)$, where $s.e.[\hat{\Delta}_{op}(s,\tilde{s})]$ is computed by the Delta method and $q_t(\cdot)$ denotes the bootstrap quantile of $\frac{\hat{\Delta}^b_{op}(s,\tilde{s}) - \hat{\Delta}_{op}(s,\tilde{s})}{s.e.[\hat{\Delta}^b_{op}(s,\tilde{s})]}$, where the standard error is computed by the Delta method for each bootstrap sample.

Ledoit and Wolf (2008) emphasize that the circular block-bootstrap based test accounts for the time series dependence of return series, which, for instance, may be crucial if the return series of two hedge funds are to be compared. However, a large fraction of empirical studies applying their tests to check for the out-of-sample performance of Markowitz type portfolio strategies are based on monthly return data. As shown in Figure 5, these monthly out-of-sample returns show no significant autocorrelation regardless of the underlying empirical portfolio strategy. Therefore it is not too surprising that in our simulations the results obtained for the circular block-bootstrap based tests hardly differ from the ones obtained by assuming uncorrelatedness in out-of-sample return process. For the sake of brevity, we do not report the results based on the circular block-bootstrap.

For each pair of strategies we consider different out-of-sample horizon lengths $H = 100$, 500, 1000,

[4] An explicit way of computing the standard errors of the CE difference is given in Appendix A.2.1

which govern the speed of convergence of the out-of-sample CE to its true value. The number of Monte Carlo simulations is 50 000 for all designs. For the bootstrap-based tests we use 5 000 replications. For the level of risk aversion we consider the values $\gamma = 0.5, 1, 3$ as commonly used in the literature (DeMiguel et al., 2009b; Kan and Zhou, 2007; Tu and Zhou, 2011).

3.2 Test Properties

We first consider the testing properties for the GMVP ($\omega(g)$) versus the equally weighted portfolio ($\omega(e)$) case. Table 1 reports the empirical Type I error of the tests for a given nominal significance level of 5% and for a very low N/T ratio. For instance, for a portfolio with $N = 30$ assets the ratio $N/T = 0.01$ requires an estimation window of $T = 3000$ observations. We consider this scenario mainly for illustrative purposes, since for this case the estimated GMVP weights are very close to their theoretical values, so that the sampling error is mainly due to the testing horizon.

Table 1: Empirical rejection probabilities under H_0 for GMVP vs 1/N. $\alpha = 5\%$. $N/T = 0.01$

		Two-sided			One-sided		
		Delta method	Bootstrap Percentile	Bootstrap t-statistic	Delta method	Bootstrap Percentile	Bootstrap t-statistic
H = 100	$\gamma = 0.5$	0.0539	0.0505	0.0542	0.0520	0.0503	0.0520
	$\gamma = 1.0$	0.0544	0.0508	0.0547	0.0522	0.0502	0.0523
	$\gamma = 3.0$	0.0547	0.0504	0.0549	0.0515	0.0504	0.0519
H = 500	$\gamma = 0.5$	0.0540	0.0532	0.0541	0.0514	0.0511	0.0517
	$\gamma = 1.0$	0.0536	0.0531	0.0540	0.0522	0.0518	0.0523
	$\gamma = 3.0$	0.0543	0.0538	0.0545	0.0512	0.0513	0.0514
H =1000	$\gamma = 0.5$	0.0564	0.0564	0.0566	0.0537	0.0535	0.0538
	$\gamma = 1.0$	0.0561	0.0561	0.0565	0.0530	0.0528	0.0529
	$\gamma = 3.0$	0.0563	0.0559	0.0564	0.0519	0.0525	0.0523

Figures in the table correspond to the share of Monte Carlo draws where the null hypothesis was rejected (out of 50 000 draws). H denotes the out-of-sample evaluation window length and γ denotes risk aversion coefficient.

As was previously pointed out, in the presence of negligible within-sample estimation noise the estimated out-of-sample CE for a fairly large out-of-sample evaluation horizon H equals the true out-of-sample CE. Noticeably, different testing methods, i.e. Delta method, percentile and t-bootstrap yield very similar results for different combinations of H and γ. For all tests and designs we do not find any substantial size distortions, i.e. the empirical null rejection probabilities are very close to the nominal value of 5%. The bootstrap percentile method outperforms slightly the two other methods in terms of minimizing size distortions for most of the scenarios.

Table 2: Power at 1% expected CE difference for GMVP vs 1/N. $\alpha = 5\%$. $N/T = 0.01$

		Two-sided			One-sided		
		Delta method	Bootstrap Percentile	Bootstrap t-statistic	Delta method	Bootstrap Percentile	Bootstrap t-statistic
H = 100	$\gamma = 0.5$	0.0625	0.0588	0.0630	0.0871	0.0845	0.0875
	$\gamma = 1.0$	0.0622	0.0581	0.0624	0.0851	0.0825	0.0855
	$\gamma = 3.0$	0.0618	0.0587	0.0625	0.0830	0.0818	0.0842
H = 500	$\gamma = 0.5$	0.0875	0.0868	0.0874	0.1399	0.1396	0.1404
	$\gamma = 1.0$	0.0903	0.0896	0.0906	0.1405	0.1405	0.1408
	$\gamma = 3.0$	0.0876	0.0875	0.0877	0.1368	0.1375	0.1375
H =1000	$\gamma = 0.5$	0.1280	0.1276	0.1283	0.1993	0.1992	0.1999
	$\gamma = 1.0$	0.1271	0.1267	0.1272	0.1952	0.1956	0.1960
	$\gamma = 3.0$	0.1227	0.1238	0.1229	0.1922	0.1933	0.1928

Figures in the table correspond to the share of the Monte Carlo draws where the null hypothesis was rejected (out of 50 000 draws). H denotes the out-of-sample evaluation window length and γ denotes risk aversion coefficient.

However, for this best case scenario in terms of estimation precision the picture changes when the power properties of the different tests are considered. For the analysis of the power properties we assume that the two out-of-sample CEs differ by additional annualized rate of 1%, which we regard as being not an unrealistic value. As reported in Table 2 we find substantial differences in empirical rejection probabilities under the alternative hypothesis. Generally, the power of the considered tests is low, especially for the shorter out-of-sample evaluation periods H. Our findings for this best case scenario clearly point out that the low power properties of the tests largely explain the non-rejection of null hypothesis of equal performance often found in the literature. It is important to note, that the power properties of the one-sided tests are better than the two-sided ones. However, even under unrealistic but favorable parameter constellations the tests have very high false negative rates, i.e. they are not able to report a significant difference in the CEs with probability of approximately 80%.

Increasing the N/T ratio reveals quite similar patterns. Table 10 in Appendix A.2.2 reports the null rejection probabilities for the lower $N/T = 0.1$ ratio, where we find minor size distortions. As pointed out at the beginning of this Section, the distribution of the CE difference in the presence of the estimation noise becomes left skewed due to the right skewed distribution of the GMVP portfolio variance. The left quantile of the distribution of the CE difference becomes smaller, thus the normal approximation of the Delta method overestimates the left quantile, which results in the over-rejections of the null hypothesis and the problem becomes more severe with increase in γ. Besides, with an increase in the out-of-sample horizon H the skewness becomes more pronounced, therefore, the size distortions increase with H. Noticeably, the two-sided percentile and t-bootstrap do not perform any better than the two-sided Delta method, as the bootstrap generally approximates the asymptotic distribution of the underlying test statistic, and estimation noise is the finite sample feature. Nevertheless, the size distortions in this case is not a huge problem, as for the worst case scenario of large γ and H the maximum over-rejection rate is 5%. Most importantly, the one-sided tests, i.e. the null hypothesis $\Delta_{op}(s, \tilde{s}) - \Delta^* < 0$, effectively use only the right quantile of the distribution and therefore avoid the left-skewness problem and do not suffer from the size distortions.

In Table 3 we report empirical null rejection probabilities under the alternative for $N/T = 0.1$. The power of the tests increases with the out-of-sample evaluation window and one-sided tests have on average 2-5% more power. Tables 9 and 11 in A.2.2 report the power properties of the tests under rather unrealistic CE differences, i.e. the case when under the alternative the actual CE difference exceeds the CE difference hypothesized under the null by 5% in annual terms. This for example implies 0.48% expected CE difference in monthly terms and 5.91% in annual terms for the $N/T = 0.01$ and $\gamma = 3$. Obviously, when the distance between the null and the alternative gets extremely large, the tests gain up to 90% power. Our findings strongly suggest to use the one-sided hypothesis testing in the favor of the commonly used two-sided hypothesis for two reasons: they do not use the left quantile of the distribution and have better power properties.

The Receiver Operating Characteristics (ROC) curves based on our simulations provide more insights into the optimal trade-off between Type I error (false positive rate, sensitivity) and power (true positive rate, specificity). Figure 3 depicts those for different parameter constellations for the tests based on the Delta method: The graph in the upper-left panel compares the ROC curves of one-sided tests for different out-of-sample horizons H, the upper-right panel depicts the ROC for different estimation noise ratios N/T and lower-left panel reports the ROC for two-sided and one-sided tests.

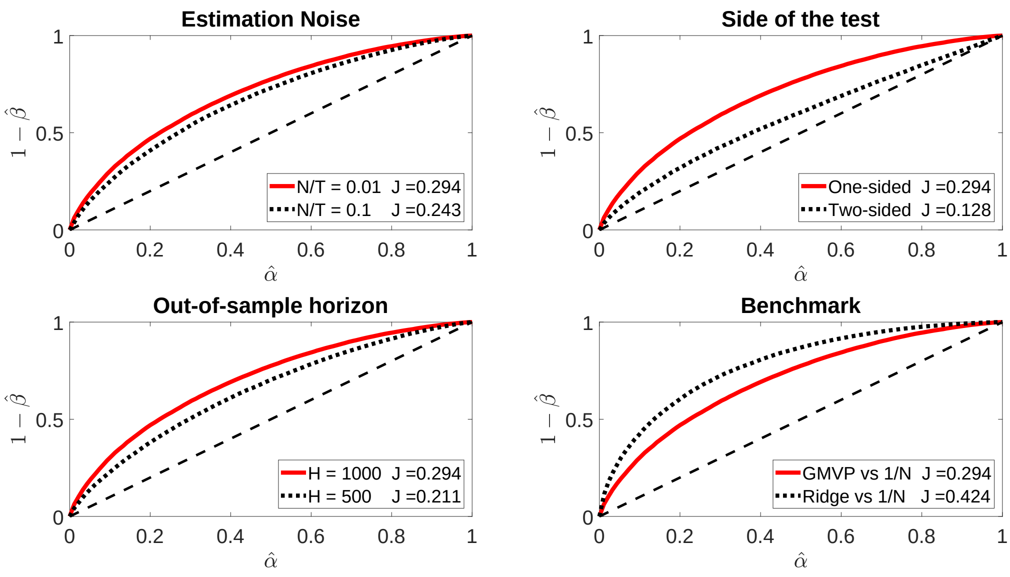
Different ROC curves can be compared by the means of Youden's J statistic, which gives the

Table 3: Power at 1% expected CE difference for GMVP vs 1/N. $\alpha = 5\%$. $N/T = 0.1$

		Two-sided			One-sided		
		Delta method	Bootstrap Percentile	Bootstrap t-statistic	Delta method	Bootstrap Percentile	Bootstrap t-statistic
H = 100	$\gamma = 0.5$	0.0637	0.0593	0.0636	0.0822	0.0783	0.0821
	$\gamma = 1.0$	0.0648	0.0608	0.0654	0.0827	0.0804	0.0833
	$\gamma = 3.0$	0.0618	0.0585	0.0623	0.0759	0.0744	0.0769
H = 500	$\gamma = 0.5$	0.0998	0.0994	0.1006	0.1422	0.1414	0.1424
	$\gamma = 1.0$	0.1000	0.0995	0.1005	0.1417	0.1409	0.1416
	$\gamma = 3.0$	0.0965	0.0955	0.0964	0.1280	0.1279	0.1279
H =1000	$\gamma = 0.5$	0.1530	0.1529	0.1535	0.2088	0.2087	0.2090
	$\gamma = 1.0$	0.1498	0.1499	0.1507	0.2055	0.2056	0.2059
	$\gamma = 3.0$	0.1430	0.1434	0.1436	0.1843	0.1847	0.1847

Figures in the table correspond to the share of the Monte Carlo draws where the null hypothesis was rejected (out of 50 000 draws). H denotes the out-of-sample evaluation window length and γ denotes risk aversion coefficient.

Figure 3: ROC for different parameter constellations



Upper-left panel: ROC of one-sided Delta method with $\gamma = 1$, $N/T = 0.01$ and different out-of-sample horizons H. Upper-right panel: ROC of one-sided Delta method with $\gamma = 1$, $H = 1000$ and different estimation noise ratios N/T. Lower-left panel: ROC of one-sided and two-sided Delta method with $\gamma = 1$ and $H = 1000$. Lower-right panel: ROC for one-sided Delta method with $\gamma = 1$, $H = 1000$, $N/T = 0.01$ for different pairs of strategies to be tested. For each test we report the Youden's J-statistic computed as the maximum difference between The ROC curve and the 45-degree line indicating the case of random guessing.

most informative combination of empirical power and empirical size of the test over the range of nominal significance levels α. Since the size distortions are negligible ($\alpha \approx \hat{\alpha}$), we can see that for conventional nominal significance levels ($\alpha = 0.01, 0.05$) the ROC-curves are only sightly above the 45-degree line, i.e. choosing the portfolio strategies based on the performance tests leads to a classification, which is only slightly superior to random classification. The best classifications in terms of Youden's J-statistic can be obtained for α-values larger than 0.2. Enlarging the size of the evaluation window by a factor of 2 (upper-left panel) or the sample size by a factor of 10 (upper-right panel) only leads to moderate improvements in the ability of the test to classify the competing strategies correctly. A stronger improvement in terms of Youden's J-statistic can be obtained by choosing a one-sided test over a two-sided one (lower-left panel).

Ridging the GMVP (lower-right panel) also yields strong improvements of the test to classify correctly. The ROC for testing the ridged GMVP against equally weighted portfolio is substantially higher than the one for the first pairs of strategies, implying the higher power for a given size of the considered test. Note that the cross-sectional correlation of returns from the tuple of strategies, $(\omega(g(\lambda)), \omega(e))$, is stronger than that of the tuple, $(\omega(g), \omega(e))$, as with increasing the shrinkage parameter $\omega(g(\lambda))$ moves closer to the case of equal weights, with $\omega(g(\lambda)) = \omega(e)$ for the ridge penalty parameter $\lambda \to \infty$. Consequently the correlation between the two estimated certainty equivalents increases, which leads to a lower standard error of the out-of-sample CE difference and improves the performance of the tests. Appendix A.2.2 provides detailed results on the size and power properties for different parameter constellations.

Summing up, the power of portfolio performance tests is rather low, independently of the testing method used. In any case the one-sided tests perform better than their two sided counterparts, i.e. they have higher power for a given size. Furthermore, the correlation between the compared out-of-sample portfolio returns plays a central role, i.e. the choice of the benchmark strategy is crucial. The equally weighted portfolio can be regarded as a particularly challenging choice, as its weights are by definition uncorrelated with the competing strategy's weights. This leads to a low correlation between the out-of-sample portfolio returns of the equally weighted portfolio and any other, estimation based strategy and consequently to a low correlation between the two corresponding out-of-sample CEs. Lower (positive) correlations between the CEs, however, lead to a larger standard error of the performance measure difference. For an applied researcher the low testing power implies that despite having a strategy that is truly superior to the benchmark, the test will not be able to reject the null of equal performance with a high probability.

4 Power-optimal Pretest Portfolios

For any classical test the choice of an appropriate significance level depends on the type of decision to be made given the outcome of the test. More specifically, in some circumstances it might be reasonable to select a lower significance level (higher probability of a Type I error) than a conventional one in order to increase the probability of rejecting the benchmark strategy. For the problem of deciding between two portfolio strategies in the presence of low power this implies assigning to the alternative a higher probability to be pursued if it is truly superior.

In the following we examine the role of the significance level for designing a pretest estimator which uses the estimates of the two strategies depending on the test outcome. Consider the true expected CE difference of two competing portfolio strategies $\Delta_{op}(s,\tilde{s})$. The goal is to choose either strategy s or strategy $\tilde{s}$ depending on the test outcome. Null and alternative hypotheses take the usual one-sided form:

$$H_0: \quad \Delta_{op}(s,\tilde{s}) \leq 0 \qquad \text{and} \qquad H_1: \quad \Delta_{op}(s,\tilde{s}) > 0. \tag{14}$$

Let the pretest estimator of the portfolio weights forecasts for $t+h$ be such that it depends either on strategy s in case the null is rejected or on $\tilde{s}$ otherwise:

$$\omega_{t+h}(s,\tilde{s}) = \mathbb{1}\left(\hat{\Delta}_{op}(s,\tilde{s}) > \Delta^*(\alpha)\right)\left(\omega_{t+h}(s) - \omega_{t+h}(\tilde{s})\right) + \omega_{t+h}(\tilde{s}), \quad h = 1,\ldots,H, \tag{15}$$

with the estimated CE difference $\hat{\Delta}_{op}(s,\tilde{s}) = \widehat{CE}_{op}(s) - \widehat{CE}_{op}(\tilde{s})$ and the critical value $\Delta^*(\alpha)$ for significance level α. Moreover, assume as a thought experiment that the test result is given at $t+h$, so that the investor knows ex-ante, which strategy will be preferred according to the test. Based on $\omega_{t+h}(s,\tilde{s})$ the corresponding pretested out-of-sample CE takes the form:

$$\widehat{CE}_{op}(s,\tilde{s}) = \mathbb{1}\left(\hat{\Delta}_{op}(s,\tilde{s}) > \Delta^*(\alpha)\right)\hat{\Delta}_{op}(s,\tilde{s}) + \widehat{CE}_{op}(\tilde{s}).$$

For a dominating strategy s the expected CE of the pretest estimator takes the form:

$$\mathrm{E}\left[\widehat{CE}_{op}(s,\tilde{s})\middle|\,\Delta_{op}(s,\tilde{s})>0\right]=\pi(\alpha)\,\mathrm{E}\left[\hat{\Delta}_{op}(s,\tilde{s})\middle|\,\hat{\Delta}_{op}(s,\tilde{s})>\Delta^{*}(\alpha)\right]+\mathrm{E}\left[\widehat{CE}_{op}(\tilde{s})\right], \quad (16)$$

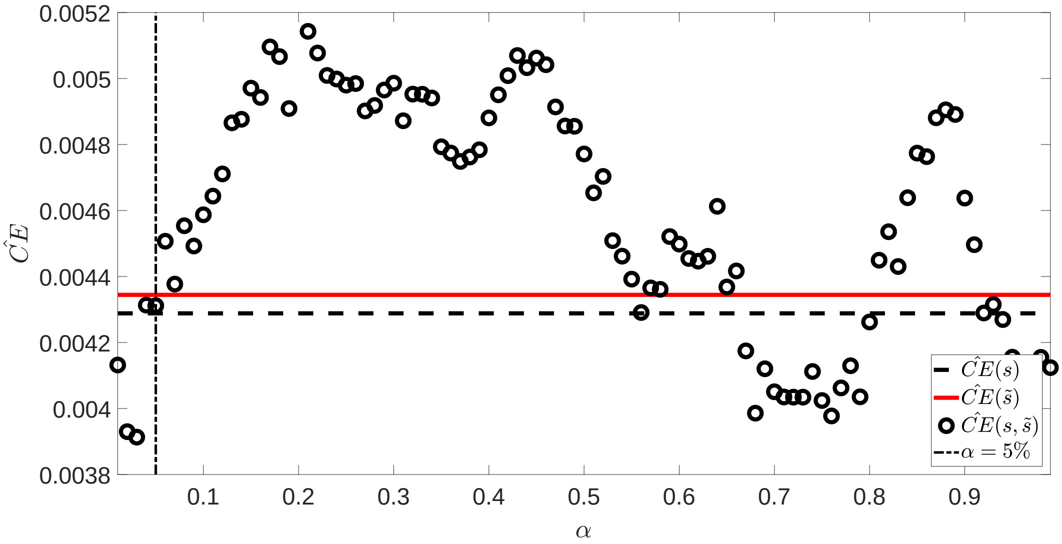
where $\pi(\alpha)$ is the power of the test. Although s is dominating $\tilde{s}$ by assumption, for a test with low power the pretest estimator hardly outperforms strategy $\tilde{s}$, because the first term on the right hand side of (16) is close to zero for conventional choices of the significance level. Moreover the dependence of the expected difference between the pretest CE, $\mathrm{E}\left[\widehat{CE}_{op}(s,\tilde{s})\middle|\,\Delta_{op}>0\right]$, and the benchmark strategy $\mathrm{E}\left[\widehat{CE}_{op}(\tilde{s})\right]$ on α is undetermined, because $\frac{\partial\,\mathrm{E}\left[\widehat{CE}_{op}(s,\tilde{s})\middle|\Delta_{op}>0\right]}{\partial\alpha}\gtrless 0$.

The pretesting idea, however, can be used to construct an empirically feasible portfolio allocation strategy. Assume that at time t an investor relies on the test in order to choose the strategy for the next period $t+1$, i.e. an investor performs a one-sided test for the CE difference $\hat{\Delta}_{op}(s,\tilde{s})=\widehat{CE}_{op}(s)-\widehat{CE}_{op}(\tilde{s})$ and chooses to use strategy s if the null of the inferior performance is rejected. However, choosing the lower significance level, i.e. increasing the probability of Type I error, and at the same time improves the power properties. Thus, there exists an optimal α for the test, which results in the higher out-of-sample CE.

For illustrative purposes consider testing the CE difference of the GMVP and the equally weighted portfolio. For a grid of α-values we perform a one-sided test and choose the GMVP strategy for the next period if the null hypothesis is rejected. Figure 4 illustrates the empirical application of such strategy. The solid red line corresponds to the out-of-sample CE if the strategy s is chosen over the whole out-of-sample horizon. The dashed black line is the corresponding out-of-sample CE for the equally weighted portfolio. The dots denote the resulting out-of-sample CE, where the choice of the next period's strategy is based on testing for different significance levels. Clearly, the exercise shows that by choosing higher α than a conventional one (values to the right of the straight line which indicates an α of 5 %) an investor can expect a higher CE. The relationship between the pretested CE and the significance level is non-monotonic, so that local optima occur.

However, as indicated above, this strategy is infeasible, as the investor has to know the outcome of the test at the same time when deciding between $\omega_t(s)$ and $\omega_t(\tilde{s})$. As a feasible solution we propose to choose the significance level α, which maximizes the in-sample CE difference. First, at time t the weights of the two competing strategies $\hat{\omega}_t(s)$ and $\hat{\omega}_t(\tilde{s})$ are

Figure 4: Pretesting strategy: illustration.



Pretesting strategy: GMVP vs $1/N$, $\gamma = 1$. K.R. French data on 5 industry portfolios with estimation window length of $T = 60$ (5 years of monthly observations), corresponding to $N/T = 0.08$ ratio. Vertical dash-dot line corresponds to the conventional $\alpha = 5\%$.

estimated based on the sample $\{t - T, \dots, t\}$. The estimated within sample CE for the strategy s is computed as

$$\widehat{CE}_{in}(s|t-T,...,t) = \hat{\omega}_t(s)'\bar{r}_t - \frac{\gamma}{2}\hat{\omega}_t(s)'\hat{\Sigma}_t\hat{\omega}_t(s), \tag{17}$$

where $\bar{r}_t$ denotes the sample mean and $\hat{\Sigma}_t$ the sample covariance matrix of the returns based on the estimation window $\{t - T, \dots, t\}$. The in-sample CE difference $\hat{\Delta}_{in}(s, \tilde{s}|t-T,...,t) = \widehat{CE}_{in}(s|t-T,...,t) - \widehat{CE}_{in}(\tilde{s}|t-T,...,t)$ can be tested against zero by the means of one-sided test as specified in (14). Analogously the in-sample pretest CE can be defined as

$$CE^*_{in}(\alpha, s, \tilde{s}|t-T,...,t) = \mathbb{1}\left(\hat{\Delta}_{in}(s,\tilde{s}|t-T,...,t) > \Delta^*(\alpha)\right)\hat{\Delta}_{in}(s,\tilde{s}|t-T,...,t) + \widehat{CE}_{in}(\tilde{s}|t-T,...,t)$$

and can be computed for the grid of α. Finally, the in-sample CE optimizing significance level α^*_t is chosen for the test, determining the strategy for the next period $t + 1$:

$$\alpha^*_{t+1} = \arg\max_{\alpha} CE^*_{in}(\alpha, s, \tilde{s}|t-T,...,t). \tag{18}$$

The above procedure is repeated with every shift of the estimation window. In practice this results in a very unstable series of significance level choices $\{\alpha^*_{t+1}, ...\alpha^*_{t+H}\}$, as the choice of α^*_{t+1}

is data driven and also depends on the instable estimates of the portfolio weights. On the other hand, the sequence of α^*'s along the rolling estimation window takes into account changes of the return process across time, e.g. volatility regimes.

In order to mitigate the instability problem we suggest two alternative ways of stabilizing this pretesting strategy. The first one is in the line of James-Stein type of shrinkage by shrinking α^*_{t+1} towards a target α_0, e.g. to the conventional 5% level

$$\alpha^s_{t+1} = (1-\lambda)\alpha^*_{t+1} + \lambda\alpha_0, \tag{19}$$

where $0 \leq \lambda \leq 1$ is the shrinkage parameter to be chosen by the investor. Thus with an increase of the shrinkage parameter α^s_{t+1} becomes more stable across time.

The second way of stabilizing the α^*-estimates is to smooth the series adaptively according to

$$\alpha^m_{t+1} = (1-\lambda)\alpha^*_{t+1} + \lambda\alpha^m_t, \tag{20}$$

where the tuning parameter λ is chosen to control the degree of smoothness. Contrary to the shrinkage method the adaptive smoothing takes into account not only the latest optimal choice but also the previous estimates with geometrically decaying weights. Both α^s_{t+1} and α^m_{t+1} require the choice of λ, which can be optimized both in-sample and out-of-sample.

We demonstrate the performance of the feasible pretesting algorithms described above for a dataset containing 100 assets in total[5], from which we form portfolios of different sizes $N = \{5, 30, 50\}$. We consider different combinations of risk aversion parameters $\gamma = \{0.5, 1, 3\}$ and estimation window sizes $T = \{60, 120\}$. The performance of the considered strategies is evaluated based on the out-of-sample CE averaged over a 1000 Monte-Carlo iterations. For each considered portfolio size we randomly draw N assets from the available dataset, compute the out-of-sample CE and repeat the procedure 1000 times. In Table 4 we report the average of the annualized out-of-sample CEs.

The results of the empirical portfolio performance evaluation reported in Table 4 are perfectly in line with the theoretical findings from previous sections. In this example the competing strategies are the GMVP and the equally weighted portfolio. Based on them the out-of-sample test-driven choice of the strategy is performed according to the three pretesting rules for the

[5] The data is taken from K.R.French website and contains monthly excess returns from 01/1953 till 12/2015.

Table 4: Average out-of-sample CE: empirical application.

	T=60			T=120		
	γ=0.5	γ=1	γ=3	γ=0.5	γ=1	γ=3
			N=5			
GMVP	0.0841	0.0755	0.0470	0.0794	0.0722	0.0427
1/N	0.0855	0.0777	0.0417	0.0814	0.0721	0.0367
In-sample	0.0892	0.0806	0.0496	0.0837	0.0756	0.0438
Shrinking	0.0876	0.0794	0.0456	0.0837	0.0749	0.0405
Smoothing	**0.0919**	**0.0861**	**0.0516**	**0.0931**	**0.0778**	**0.0455**
			N=30			
GMVP	0.0749	0.0662	0.0355	0.0787	0.0733	0.0492
1/N	0.0867	0.0784	0.0459	0.0823	0.0738	0.0407
In-sample	0.0875	0.0793	0.0468	0.0899	0.0821	0.0517
Shrinking	0.0868	0.0782	0.0460	0.0863	0.0783	0.0481
Smoothing	**0.0929**	**0.0837**	**0.0525**	**0.0903**	**0.0832**	**0.0536**
			N=50			
GMVP	0.0541	0.0335	-0.0479	0.0799	0.0730	0.0484
1/N	**0.0866**	**0.0784**	**0.0460**	0.0823	0.0737	0.0409
In-sample	0.0799	0.0646	-0.0030	**0.0965**	**0.0876**	0.0543
Shrinking	0.0851	0.0730	0.0223	0.0875	0.0803	0.0532
Smoothing	0.0842	0.0727	0.0175	0.0952	0.0865	**0.0567**

Figures in the table correspond to the annualized average out-of-sample CE over 1000 randomly formed portfolios of the specified size. T denotes the estimation window length, γ denotes risk aversion coefficient and N is the number of assets. The numbers in bold correspond to the largest CE obtained for a given γ, N, T combination. The tuning parameter λ for both shrinking and smoothing the α^* series is set to be 0.5.

significance level: (i) in-sample CE optimizing α^*_{t+1} as in (18), (ii) target shrinkage α^s_{t+1} as in (19) and (iii) adaptive smoothing α^m_{t+1} as in (20). As expected, for a given risk aversion γ the performance of the equally weighted portfolio is not changing with estimation window and portfolio sizes. Contrary to this, the GMVP performance in terms of the CE often worsens with the size of the portfolio. This particularly happens when estimation risk is severe, i.e. when the sample size is small relative to the number of parameters to be estimated. The flexible choice of the significance level α^*_{t+1} allowing for a switch between the strategies results in the higher out-of-sample CE for all three feasible methods. The only exception is the case with $N = 50/T = 60$, where the equally weighted portfolio is dominating, where the information contained in the noisy estimates of the the GMVP weights is simply not sufficient.

The superior performance of the pretesting strategies can be better understood by looking at

Table 5: Percentage of the risky strategy choice for different pretesting strategies

	T=60			T=120		
	γ=0.5	γ=1	γ=3	γ=0.5	γ=1	γ=3
			N=5			
In-sample	49.7%	53.1%	60.0%	47.6%	52.4%	62.1%
Shrinkage	24.6%	27.7%	34.2%	21.7%	25.2%	34.0%
Smoothing	34.7%	37.2%	43.3%	34.6%	38.5%	46.7%
			N=30			
In-sample	52.9%	57.1%	67.2%	49.1%	54.8%	70.0%
Shrinkage	27.1%	29.1%	37.4%	27.9%	30.9%	41.9%
Smoothing	36.7%	39.7%	47.2%	38.4%	42.7%	54.8%
			N=50			
In-sample	49.9%	53.9%	62.4%	48.8%	55.1%	69.2%
Shrinkage	29.2%	30.1%	31.9%	32.4%	34.7%	42.3%
Smoothing	35.7%	37.3%	41.0%	39.8%	44.3%	54.9%

Figures in the table correspond to the average percent of choosing GMVP over equally weighted portfolio strategy over 1000 randomly formed portfolios of the specified size. T denotes the estimation window length, γ denotes risk aversion coefficient and N is the number of assets.

Table 5 which reports the share of cases, for which the GMVP is chosen over equally weighted portfolio. Obviously, for a fixed estimation window length and portfolio size the increase in the risk aversion parameter leads to the higher share of the GMVP choices. For a given γ the shares do not seem to change much for both, different portfolio sizes and estimation window lengths.

5 Conclusions

This paper takes a closer look on the properties of the out-of-sample portfolio performance tests. The pathology of these tests shows that the contribution of estimation noise to the test statistics by far outweighs the contribution resulting from differences in the underlying theoretical portfolio strategies. Therefore the test outcomes mainly reflect differences in estimation uncertainty for the two portfolio strategies under consideration. Furthermore, our findings suggest that the discussion on the empirical performance of the competing strategies should take into account the low power of the portfolio performance tests.

We provide evidence that the out-of-sample portfolio returns generated from estimated portfolio weights in a rolling window design, follow a fat-tailed mixture distribution different from the ones conventionally assumed in finance. By means of a Monte Carlo study based on these mixture distributions we are able to show that the out-of-sample portfolio tests reveal minor size distortions, but suffer from low power regardless of the type of test applied or its specific implementation. This explains why in many out-of-sample horse races more sophisticated, theoretically founded strategies are often incapable to outperform the simple, equally weighted portfolio strategy at conventional significance levels.

Apart from this unfavorable news, our study provides some guidelines for empirical studies using out-of-sample performance tests. In order to improve the power of the test, it is advisable to use a one-sided hypothesis instead of the two-sided one. Secondly, the equally weighted portfolio is a particular tricky benchmark strategy as the estimated standard error of the CE difference is particularly large. Basing the tests on other meaningful comparisons reduces the power problem to some extent. Lastly, if possible, the size of the testing horizon matters more than the size of the estimation window, i.e. the researcher should go for the largest testing horizon feasible.

We show that testing at a conventional significance level is not a reasonable strategy for two reasons. Firstly, our ROC analysis indicates that tests based on conventional significance levels are hardly superior in choosing the better out-of-sample strategy compared to a random classification. Secondly, we show that the tests are uninformative if the goal is to select a strategy which is optimal in terms of financial performance. Based on a CE optimal choice of the significance level a simple pretest-based strategy is shown to be superior over the standalone strategies. The gain of the strategy results from its flexibility to switch between the GMVP and equal portfolio weights in different volatility regimes based on the outcome of the performance test. We show, that despite their low power, the performance tests can be used for pretest estimation of the portfolio weights which is able to outperform the underlying stand-alone strategies. Future research should be concerned with the optimal data-driven choice of the tuning parameter of the proposed strategy, e.g. based on a training sample along the lines of conventional machine learning approaches.

Our study concentrates on testing procedures usually applied to monthly portfolio data. In future work it needs to be shown whether our findings can be generalized to portfolio strategies estimated returns at higher frequencies where autocorrelation plays a major role. Future work also should investigate to what extent our findings can be generalized to other performance measures than the certainty equivalent.

6 Acknowledgments

Financial support by the German Research Foundation (DFG) through research project ”Bayesian Portfolio Regularization” and the Graduate School of Decision Science (GSDS) is gratefully acknowledged. This work was performed on the computational resource bwUniCluster funded by the Ministry of Science, Research and the Arts Baden-Württemberg and the Universities of the State of Baden-Württemberg, Germany, within the framework program bwHPC. The authors like to thank Patrik Guggenberger for helpful comments. All remaining errors are ours.

References

Antoine, B. (2012): “Portfolio Selection with Estimation Risk: A Test-Based Approach,” *Journal of Financial Econometrics*, 10, 164–197.

Britten-Jones, M. (1999): “The Sampling Error in Estimates of Mean-Variance Efficient Portfolio Weights,” *The Journal of Finance*, 54, 655–671.

Brodie, J., I. Daubechies, C. D. Mol, D. Giannone, and I. Loris (2009): “Sparse and Stable Markowitz Portfolios,” *Proceedings of the National Academy of Sciences*, 106, 12267–12272, PMID: 19617537.

Cho, D. D. (2011): "Estimation Risk in Covariance," *Journal of Asset Management*, 12, 248–259.

DeMiguel, V., L. Garlappi, F. J. Nogales, and R. Uppal (2009a): "A Generalized Approach to Portfolio Optimization: Improving Performance by Constraining Portfolio Norms," *Management Science*, 55, 798–812.

DeMiguel, V., L. Garlappi, and R. Uppal (2009b): "Optimal Versus Naive Diversification: How Inefficient is the 1/N Portfolio Strategy?" *The Review of Financial Studies*, 22, 1915–1953.

DeMiguel, V. and F. J. Nogales (2009): "Portfolio selection with robust estimation," *Operations Research*, 57, 560–577.

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

Golosnoy, V. and Y. Okhrin (2007): "Multivariate Shrinkage for Optimal Portfolio Weights," *The European Journal of Finance*, 13, 441–458.

Jobson, J. D. and B. M. Korkie (1981): "Performance Hypothesis Testing with the Sharpe and Treynor Measures," *The Journal of Finance*, 36, 889–908.

Kan, R. and G. Zhou (2007): "Optimal Portfolio Choice with Parameter Uncertainty," *Journal of Financial and Quantitative Analysis*, 42, 621–656.

Kempf, A. and C. Memmel (2006): "Estimating the Global Minimum Variance Portfolio," *Schmalenbach Business Review (SBR)*, 58, 332–348.

Ledoit, O. and M. Wolf (2003): "Improved Estimation of the Covariance Matrix of Stock Returns with an Application to Portfolio Selection," *Journal of Empirical Finance*, 10, 603–621.

——— (2008): "Robust Performance Hypothesis Testing with the Sharpe Ratio," *Journal of Empirical Finance*, 15, 850–859.

——— (2011): "Robust Performances Hypothesis Testing With the Variance," *Wilmott*, 2011, 86–89.

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

Okhrin, Y. and W. Schmid (2006): "Distributional Properties of Portfolio Weights," *Journal of Econometrics*, 134, 235–256.

Tu, J. and G. Zhou (2011): "Markowitz meets Talmud: A combination of sophisticated and naive diversification strategies," *Journal of Financial Economics*, 99, 204–215.

A Appendix

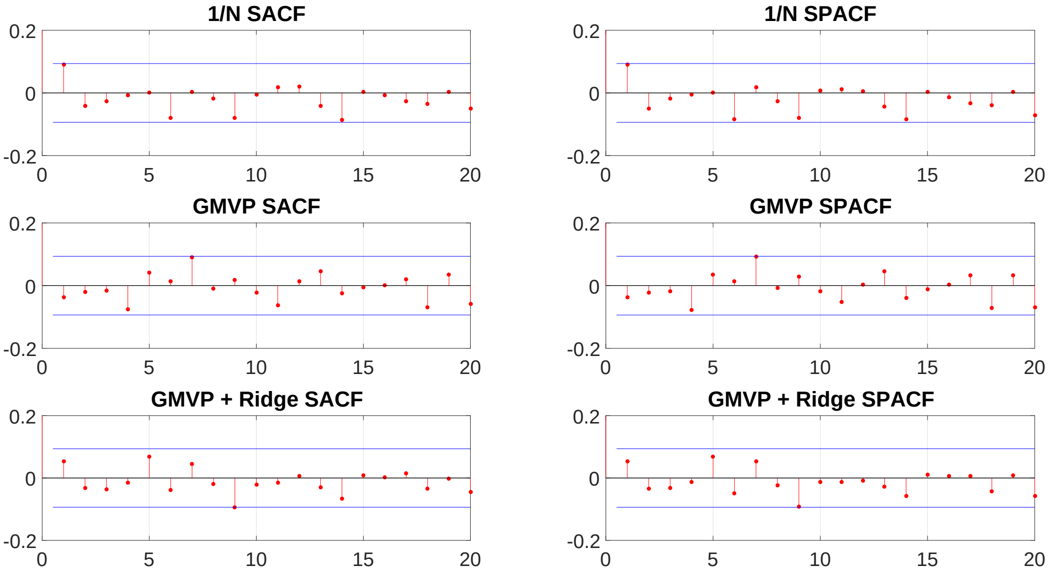
A.1 Properties of the Return Process Within and Out-of-Sample

Table 6: Descriptive statistics

	Average over all returns	Average return ($\hat{r}^p(e)$)	$\hat{r}^p(g)$	$\hat{r}^p(g(\lambda))$
$\hat{\mu}$	0.0065	0.0065	0.0056	0.0066
(p-val)		(0.0002)	(0.0010)	(0.0000)
$\hat{\sigma}$	0.0594	0.0466	0.0449	0.0380
$\hat{S}$	-0.5061	-0.5093	-0.0315	-0.5733
(p-val)		(0.0000)	(0.7343)	(0.0000)
$\hat{K}$	5.4024	5.6645	4.0207	5.8418
(p-val)		(0.0000)	(0.0000)	(0.0000)
JB	249.9073	235.9786	30.3306	272.3244
(p-val)		(0.0000)	(0.0000)	(0.0000)

Descriptive statistics of the returns in the data set and the corresponding out-of-sample portfolio returns for different portfolio allocation strategies: equally weighted portfolio $\hat{r}^p(e)$ (average return), GMVP $\hat{r}^p(g)$, GMVP combined with ridge estimator of covariance matrix $\hat{r}^p(g(\lambda))$. The Table reports the sample mean $\hat{\mu}$ and the corresponding p-value when testing the mean against zero; $\hat{\sigma}$ denotes standard deviation; skewness $\hat{S}$ is reported alongside with the p-values of the test against zero; kurtosis $\hat{K}$ is tested against 3, JB denotes the Jarque-Bera test for normality with the corresponding p-value in parenthesis. The out-of-sample portfolio returns are constructed using a rolling window with estimation window $T = 60$, where portfolio is rebalanced every month resulting in the vector of length $H = 696$. The number of assets is set to $N = 30$, shrinkage intensity $\lambda = 0.0078$.

Figure 5: Autocorrelation and Partial Autocorrelation of out-of-sample portfolio returns for different strategies.



Sample autocorrelation function (SACF), sample partial autocorrelation function (SPACF) for equally weighted portfolio ($1/N$), Global Minimum Variance Portfolio (GMVP), the ridge covariance matrix estimator combined with the GMVP (GMVP + Ridge).

Table 7: Descriptive statistics of simulated returns

	Data	Mixture design	Bivariate normal	Bivariate t
	GMVP out-of-sample portfolio return			
$\hat{\mu}$	0.0056	0.0058	0.0058	0.0058
$\hat{\sigma}$	0.0449	0.0539	0.0315	0.0315
$\hat{S}$	-0.0315	0.0003	0.0000	0.0004
$\hat{K}$	4.0207	3.5525	2.8842	7.1761
$\overline{KS}$	–	0.3870	0.9954	1.000
	1/N out-of-sample portfolio return			
$\hat{\mu}$	0.0065	0.0067	0.0067	0.0067
$\hat{\sigma}$	0.0466	0.0455	0.0455	0.0455
$\hat{S}$	-0.5093	-0.0664	0.0005	-0.0008
$\hat{K}$	5.6645	2.9994	2.9902	7.1536
$\overline{KS}$	–	0.0055	0.0540	0.1325

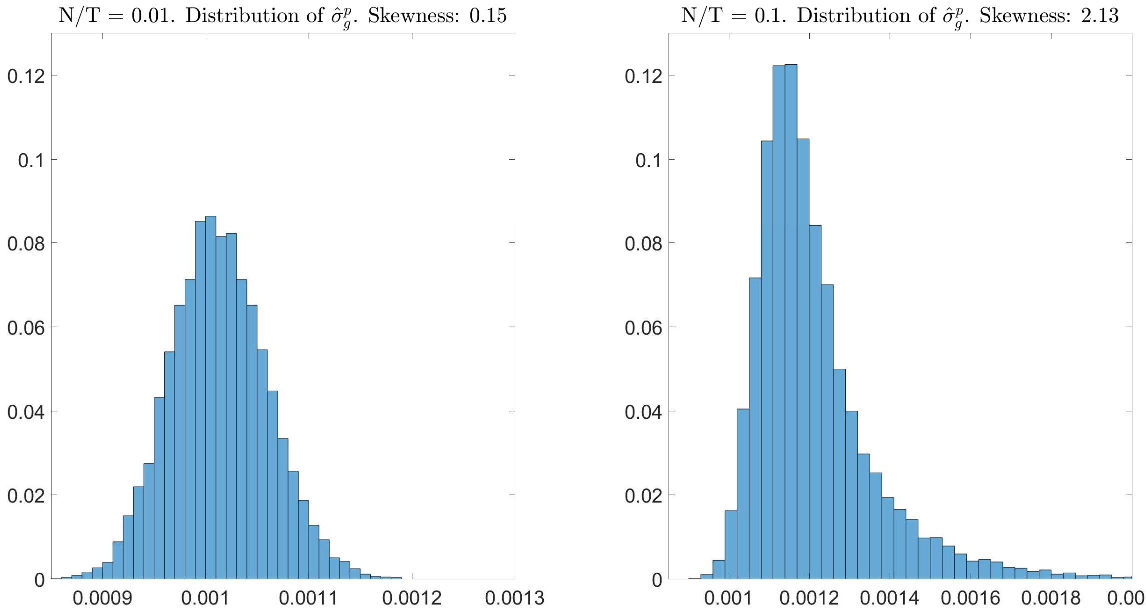
Descriptive statistics of the out-of-sample portfolio returns in the data set and the corresponding average out-of-sample portfolio returns simulated according to the proposed mixture design: sample mean $\hat{\mu}$, sample standard deviation $\hat{\sigma}$, sample skewness $\hat{S}$ and sample kurtosis $\hat{K}$. $\overline{KS}$ reports the average null rejection rate for the two-sample Kolmogorov-Smirnov test, where the ECDF of the simulated returns is tested against the ECDF of the portfolio returns computed from the real data. The number of considered assets $N = 30$. The length of the out-of-sample return vector is set to $H = 696$, estimation window $T = 60$.

Table 8: Theoretical out-of-sample CE difference under the null and alternative hypotheses

	GMVP - 1/N							
	H_0				H_1			
	N/T = 0.01		N/T = 0.1		N/T = 0.01		N/T = 0.1	
	Monthly	Annually	Monthly	Annually	Monthly	Annually	Monthly	Annually
$\gamma = 0.5$	-0.06%	-0.67%	-0.06%	-0.70%	0.03%	0.33%	0.03%	0.30%
$\gamma = 1.0$	-0.03%	-0.35%	-0.03%	-0.41%	0.05%	0.65%	0.05%	0.59%
$\gamma = 3.0$	0.08%	0.94%	0.06%	0.76%	0.16%	1.94%	0.15%	1.76%
	GMVP(ridge) - 1/N							
$\gamma = 0.5$	0.01%	0.18%	0.01%	0.15%	0.10%	1.18%	0.09%	1.15%
$\gamma = 1.0$	0.03%	0.36%	0.03%	0.30%	0.11%	1.36%	0.11%	1.30%
$\gamma = 3.0$	0.09%	1.11%	0.08%	0.93%	0.17%	2.11%	0.16%	1.93%

Figures in the table correspond to the theoretical out-of-sample CE difference, $\Delta_{op}(s, \tilde{s})$, for a given couple of strategies (in percent, risk aversion parameter γ and estimation noise ratio N/T. The number of assets: $N = 30$. Annualized differences are computed as $\Delta_{op}(\text{annual}) = (1 + \Delta_{op}(\text{monthly}))^{12} - 1$.

Figure 6: Simulated distribution of the GMVP portfolio variance



N/T = 0.01. Distribution of $\hat{\sigma}^p_g$. Skewness: 0.15

N/T = 0.1. Distribution of $\hat{\sigma}^p_g$. Skewness: 2.13

Histograms normalized to probability of the GMVP portfolio variances over 10 000 Monte-Carlo draws for different N/T ratios with $N = 30$. The out-of-sample portfolio returns are simulated according to the proposed mixture design with length of the simulated returns $H = 1000$. The description of the histograms contains the sample skewness over the 10 000 variances.

A.2 Size and Power Simulations

A.2.1 Delta Method

The standard error for the CE difference can be computed according to equation (13) with

$$V\left[\hat{\vartheta}\right] = \begin{pmatrix} \hat{\sigma}^2_{op}(s) & \hat{\sigma}_{op}(s,\tilde{s}) & 0 & 0 \\ \hat{\sigma}_{op}(s,\tilde{s}) & \hat{\sigma}^2_{op}(\tilde{s}) & 0 & 0 \\ 0 & 0 & 2\hat{\sigma}^4_{op}(s) & 2\hat{\sigma}^2_{op}(s,\tilde{s}) \\ 0 & 0 & 2\hat{\sigma}^2_{op}(s,\tilde{s}) & 2\hat{\sigma}^4_{op}(\tilde{s}) \end{pmatrix},$$

where $\hat{\sigma}_{op}(s,\tilde{s})$ denotes the sample covariance between the out-of-sample portfolio returns $\hat{r}_p(s)$ and $\hat{r}_p(\tilde{s})$. However, in the similar spirit to the Ledoit and Wolf (2008) working with the uncentered moments might be more convenient. Given the out-of-sample portfolio returns $\hat{r}_p(s), r_p(\tilde{s})$ of length H define $\hat{y} = (\hat{r}_p(s) - \bar{\hat{r}}_p(s), r_p(\tilde{s}) - \bar{r}_p(\tilde{s}), \hat{r}^2_p(s) - \overline{\hat{r}^2}_p(s), r^2_p(\tilde{s}) - \overline{r^2}_p(\tilde{s}))$, where $\bar{\hat{r}}(\cdot)$ denotes the sample average. The standard error of the out-of-sample CE difference is then computed as

$$S.E. = \sqrt{\frac{\hat{\nabla}'\hat{\Psi}\hat{\nabla}}{H}} \quad \text{with} \quad \hat{\Psi} = \frac{1}{H}\hat{y}'\hat{y} \quad \text{and} \quad \hat{\nabla} = \left[1 + \gamma \cdot \bar{\hat{r}}^p(s), -1 - \gamma \cdot \bar{r}^p(\tilde{s}), -\frac{\gamma}{2}, \frac{\gamma}{2}\right]'.$$

A.2.2 GMVP against equally weighted portfolio

Table 9: Power at 5% expected CE difference for GMVP vs 1/N. $\alpha = 5\%$. $N/T = 0.01$

		Two-sided			One-sided		
		Delta method	Bootstrap Percentile	Bootstrap t-statistic	Delta method	Bootstrap Percentile	Bootstrap t-statistic
H = 100	$\gamma = 0.5$	0.1766	0.1684	0.1764	0.2636	0.2566	0.2644
	$\gamma = 1.0$	0.1774	0.1705	0.1779	0.2657	0.2602	0.2666
	$\gamma = 3.0$	0.1810	0.1777	0.1826	0.2703	0.2684	0.2722
H = 500	$\gamma = 0.5$	0.6157	0.6141	0.6162	0.7284	0.7272	0.7279
	$\gamma = 1.0$	0.6201	0.6194	0.6208	0.7323	0.7312	0.7326
	$\gamma = 3.0$	0.6300	0.6316	0.6309	0.7409	0.7432	0.7425
H =1000	$\gamma = 0.5$	0.8892	0.8881	0.8886	0.9376	0.9372	0.9376
	$\gamma = 1.0$	0.8915	0.8913	0.8915	0.9387	0.9387	0.9387
	$\gamma = 3.0$	0.8995	0.9002	0.8994	0.9441	0.9444	0.9439

Figures in the table correspond to the share of Monte Carlo draws where the null hypothesis was rejected (out of 50 000 draws). H denotes the out-of-sample evaluation window length and γ denotes risk aversion coefficient.

Table 10: Empirical rejection probabilities under H_0 for GMVP vs 1/N. $\alpha = 5\%$. $N/T = 0.1$

		Two-sided			One-sided		
		Delta method	Bootstrap Percentile	Bootstrap t-statistic	Delta method	Bootstrap Percentile	Bootstrap t-statistic
H = 100	$\gamma = 0.5$	0.0596	0.0546	0.0591	0.0531	0.0510	0.0534
	$\gamma = 1.0$	0.0597	0.0550	0.0593	0.0527	0.0506	0.0529
	$\gamma = 3.0$	0.0607	0.0562	0.0604	0.0494	0.0481	0.0498
H = 500	$\gamma = 0.5$	0.0764	0.0757	0.0767	0.0641	0.0639	0.0644
	$\gamma = 1.0$	0.0786	0.0779	0.0786	0.0632	0.0629	0.0633
	$\gamma = 3.0$	0.0803	0.0795	0.0806	0.0547	0.0551	0.0551
H =1000	$\gamma = 0.5$	0.1030	0.1030	0.1035	0.0788	0.0788	0.0791
	$\gamma = 1.0$	0.0990	0.0992	0.0994	0.0726	0.0723	0.0725
	$\gamma = 3.0$	0.1120	0.1120	0.1124	0.0642	0.0642	0.0641

Figures in the table correspond to the share of Monte Carlo draws where the null hypothesis was rejected (out of 50 000 draws). H denotes the out-of-sample evaluation window length and γ denotes risk aversion coefficient.

Table 11: Power at 5% expected CE difference for GMVP vs 1/N. $\alpha = 5\%$. $N/T = 0.1$

		Two-sided			One-sided		
		Delta method	Bootstrap Percentile	Bootstrap t-statistic	Delta method	Bootstrap Percentile	Bootstrap t-statistic
H = 100	$\gamma = 0.5$	0.1720	0.1639	0.1724	0.2591	0.2527	0.2592
	$\gamma = 1.0$	0.1739	0.1678	0.1746	0.2608	0.2549	0.2616
	$\gamma = 3.0$	0.1724	0.1695	0.1750	0.2585	0.2564	0.2602
H = 500	$\gamma = 0.5$	0.5914	0.5899	0.5917	0.7056	0.7045	0.7058
	$\gamma = 1.0$	0.5958	0.5945	0.5957	0.7064	0.7063	0.7072
	$\gamma = 3.0$	0.5854	0.5868	0.5860	0.6973	0.6985	0.6979
H =1000	$\gamma = 0.5$	0.8661	0.8662	0.8663	0.9200	0.9198	0.9198
	$\gamma = 1.0$	0.8608	0.8610	0.8609	0.9191	0.9181	0.9184
	$\gamma = 3.0$	0.8511	0.8521	0.8515	0.9083	0.9088	0.9086

Figures in the table correspond to the share of Monte Carlo draws where the null hypothesis was rejected (out of 50 000 draws). H denotes the out-of-sample evaluation window length and γ denotes risk aversion coefficient.

Table 12: Empirical rejection probabilities under H_0 for GMVP vs 1/N. $\alpha = 5\%$. $N/T = 0.25$

		Two-sided			One-sided		
		Delta method	Bootstrap Percentile	Bootstrap t-statistic	Delta method	Bootstrap Percentile	Bootstrap t-statistic
H = 100	$\gamma = 0.5$	0.0569	0.0527	0.0566	0.0536	0.0510	0.0539
	$\gamma = 1.0$	0.0559	0.0519	0.0562	0.0531	0.0510	0.0534
	$\gamma = 3.0$	0.0566	0.0524	0.0568	0.0524	0.0510	0.0534
H = 500	$\gamma = 0.5$	0.0670	0.0660	0.0670	0.0624	0.0622	0.0627
	$\gamma = 1.0$	0.0673	0.0670	0.0680	0.0633	0.0630	0.0635
	$\gamma = 3.0$	0.0676	0.0666	0.0679	0.0632	0.0633	0.0632
H =1000	$\gamma = 0.5$	0.0817	0.0814	0.0822	0.0736	0.0738	0.0742
	$\gamma = 1.0$	0.0819	0.0818	0.0825	0.0730	0.0728	0.0731
	$\gamma = 3.0$	0.0851	0.0849	0.0854	0.0726	0.0730	0.0728

Figures in the table correspond to the share of Monte Carlo draws where the null hypothesis was rejected (out of 50 000 draws). H denotes the out-of-sample evaluation window length and γ denotes risk aversion coefficient.

Table 13: Power at 1% expected CE difference for GMVP vs 1/N. $\alpha = 5\%$. $N/T = 0.25$

		Two-sided			One-sided		
		Delta method	Bootstrap Percentile	Bootstrap t-statistic	Delta method	Bootstrap Percentile	Bootstrap t-statistic
H = 100	$\gamma = 0.5$	0.0639	0.0601	0.0640	0.0838	0.0806	0.0844
	$\gamma = 1.0$	0.0653	0.0616	0.0653	0.0832	0.0805	0.0836
	$\gamma = 3.0$	0.0607	0.0569	0.0609	0.0812	0.0795	0.0821
H = 500	$\gamma = 0.5$	0.0989	0.0985	0.0993	0.1438	0.1430	0.1438
	$\gamma = 1.0$	0.0980	0.0973	0.0982	0.1428	0.1420	0.1428
	$\gamma = 3.0$	0.0989	0.0982	0.0989	0.1422	0.1423	0.1424
H =1000	$\gamma = 0.5$	0.1449	0.1446	0.1451	0.2051	0.2048	0.2053
	$\gamma = 1.0$	0.1473	0.1471	0.1476	0.2057	0.2063	0.2065
	$\gamma = 3.0$	0.1471	0.1471	0.1471	0.2068	0.2070	0.2069

Figures in the table correspond to the share of Monte Carlo draws where the null hypothesis was rejected (out of 50 000 draws). H denotes the out-of-sample evaluation window length and γ denotes risk aversion coefficient.

Table 14: Power at 5% expected CE difference for GMVP vs 1/N. $\alpha = 5\%$. $N/T = 0.25$

		Two-sided			One-sided		
		Delta method	Bootstrap Percentile	Bootstrap t-statistic	Delta method	Bootstrap Percentile	Bootstrap t-statistic
H = 100	$\gamma = 0.5$	0.1695	0.1626	0.1700	0.2571	0.2503	0.2574
	$\gamma = 1.0$	0.1751	0.1684	0.1754	0.2601	0.2545	0.2610
	$\gamma = 3.0$	0.1766	0.1730	0.1772	0.2642	0.2619	0.2667
H = 500	$\gamma = 0.5$	0.5913	0.5901	0.5918	0.7055	0.7051	0.7060
	$\gamma = 1.0$	0.5887	0.5876	0.5889	0.7006	0.6998	0.7009
	$\gamma = 3.0$	0.5978	0.5993	0.5982	0.7082	0.7097	0.7098
H =1000	$\gamma = 0.5$	0.8605	0.8599	0.8602	0.9170	0.9166	0.9169
	$\gamma = 1.0$	0.8637	0.8634	0.8636	0.9177	0.9175	0.9178
	$\gamma = 3.0$	0.8629	0.8634	0.8630	0.9170	0.9174	0.9172

Figures in the table correspond to the share of Monte Carlo draws where the null hypothesis was rejected (out of 50 000 draws). H denotes the out-of-sample evaluation window length and γ denotes risk aversion coefficient.

Table 15: Empirical rejection probabilities under H_0 for GMVP vs 1/N. $\alpha = 5\%$. $N/T = 0.5$

		Two-sided			One-sided		
		Delta method	Bootstrap Percentile	Bootstrap t-statistic	Delta method	Bootstrap Percentile	Bootstrap t-statistic
lH = 100	$\gamma = 0.5$	0.0608	0.0569	0.0613	0.0573	0.0554	0.0576
	$\gamma = 1.0$	0.0598	0.0559	0.0602	0.0596	0.0568	0.0595
	$\gamma = 3.0$	0.0632	0.0589	0.0636	0.0607	0.0583	0.0612
lH = 500	$\gamma = 0.5$	0.0820	0.0807	0.0815	0.0728	0.0721	0.0732
	$\gamma = 1.0$	0.0836	0.0831	0.0842	0.0744	0.0737	0.0742
	$\gamma = 3.0$	0.0926	0.0924	0.0931	0.0825	0.0821	0.0827
H =1000	$\gamma = 0.5$	0.1155	0.1148	0.1156	0.0949	0.0945	0.0948
	$\gamma = 1.0$	0.1192	0.1189	0.1196	0.0959	0.0957	0.0961
	$\gamma = 3.0$	0.1338	0.1333	0.1337	0.1099	0.1097	0.1098

Figures in the table correspond to the share of Monte Carlo draws where the null hypothesis was rejected (out of 50 000 draws). H denotes the out-of-sample evaluation window length and γ denotes risk aversion coefficient.

Table 16: Power at 1% expected CE difference for GMVP vs 1/N. $\alpha = 5\%$. $N/T = 0.5$

		Two-sided			One-sided		
		Delta method	Bootstrap Percentile	Bootstrap t-statistic	Delta method	Bootstrap Percentile	Bootstrap t-statistic
H = 100	$\gamma = 0.5$	0.0664	0.0622	0.0664	0.0833	0.0800	0.0835
	$\gamma = 1.0$	0.0660	0.0611	0.0665	0.0822	0.0790	0.0823
	$\gamma = 3.0$	0.0677	0.0632	0.0681	0.0843	0.0815	0.0842
H = 500	$\gamma = 0.5$	0.1106	0.1092	0.1104	0.1465	0.1458	0.1469
	$\gamma = 1.0$	0.1097	0.1087	0.1096	0.1437	0.1430	0.1437
	$\gamma = 3.0$	0.1213	0.1199	0.1212	0.1550	0.1539	0.1549
H =1000	$\gamma = 0.5$	0.1670	0.1664	0.1668	0.2092	0.2096	0.2101
	$\gamma = 1.0$	0.1673	0.1670	0.1675	0.2068	0.2065	0.2069
	$\gamma = 3.0$	0.1855	0.1845	0.1853	0.2233	0.2231	0.2234

Figures in the table correspond to the share of Monte Carlo draws where the null hypothesis was rejected (out of 50 000 draws). H denotes the out-of-sample evaluation window length and γ denotes risk aversion coefficient.

Table 17: Power at 5% expected CE difference for GMVP vs 1/N. $\alpha = 5\%$. $N/T = 0.5$

		Two-sided			One-sided		
		Delta method	Bootstrap Percentile	Bootstrap t-statistic	Delta method	Bootstrap Percentile	Bootstrap t-statistic
H = 100	$\gamma = 0.5$	0.1613	0.1541	0.1624	0.2454	0.2381	0.2450
	$\gamma = 1.0$	0.1623	0.1556	0.1625	0.2439	0.2371	0.2449
	$\gamma = 3.0$	0.1641	0.1588	0.1649	0.2442	0.2403	0.2458
H = 500	$\gamma = 0.5$	0.5403	0.5384	0.5407	0.6492	0.6476	0.6493
	$\gamma = 1.0$	0.5353	0.5333	0.5352	0.6471	0.6461	0.6474
	$\gamma = 3.0$	0.5403	0.5394	0.5400	0.6449	0.6448	0.6455
H = 1000	$\gamma = 0.5$	0.7961	0.7946	0.7953	0.8636	0.8634	0.8639
	$\gamma = 1.0$	0.7963	0.7959	0.7965	0.8626	0.8625	0.8631
	$\gamma = 3.0$	0.7733	0.7736	0.7739	0.8414	0.8414	0.8415

Figures in the table correspond to the share of Monte Carlo draws where the null hypothesis was rejected (out of 50 000 draws). H denotes the out-of-sample evaluation window length and γ denotes risk aversion coefficient.

A.2.3 Ridged GMVP against equally weighted portfolio

Table 18: Empirical rejection probabilities under H_0 for GMVP with ridged covariance matrix vs 1/N. $\alpha = 5\%$. $N/T = 0.01$

		Two-sided			One-sided		
		Delta method	Bootstrap Percentile	Bootstrap t-statistic	Delta method	Bootstrap Percentile	Bootstrap t-statistic
H = 100	$\gamma = 0.5$	0.0559	0.0516	0.0557	0.0541	0.0523	0.0542
	$\gamma = 1.0$	0.0554	0.0517	0.0559	0.0539	0.0523	0.0544
	$\gamma = 3.0$	0.0562	0.0517	0.0564	0.0533	0.0528	0.0543
H = 500	$\gamma = 0.5$	0.0550	0.0541	0.0552	0.0529	0.0527	0.0531
	$\gamma = 1.0$	0.0541	0.0537	0.0544	0.0528	0.0528	0.0530
	$\gamma = 3.0$	0.0543	0.0534	0.0544	0.0531	0.0537	0.0536
H =1000	$\gamma = 0.5$	0.0563	0.0564	0.0569	0.0541	0.0540	0.0541
	$\gamma = 1.0$	0.0551	0.0548	0.0552	0.0543	0.0542	0.0542
	$\gamma = 3.0$	0.0585	0.0577	0.0583	0.0574	0.0582	0.0580

Figures in the table correspond to the share of Monte Carlo draws where the null hypothesis was rejected (out of 50 000 draws). H denotes the out-of-sample evaluation window length and γ denotes risk aversion coefficient.

Table 19: Power at 1% expected CE difference for GMVP with ridged covariance matrix vs 1/N. $\alpha = 5\%$. $N/T = 0.01$

		Two-sided			One-sided		
		Delta method	Bootstrap Percentile	Bootstrap t-statistic	Delta method	Bootstrap Percentile	Bootstrap t-statistic
H = 100	$\gamma = 0.5$	0.0702	0.0667	0.0705	0.1065	0.1030	0.1074
	$\gamma = 1.0$	0.0706	0.0667	0.0711	0.1032	0.1008	0.1037
	$\gamma = 3.0$	0.0849	0.0838	0.0860	0.1327	0.1331	0.1345
H = 500	$\gamma = 0.5$	0.1401	0.1393	0.1401	0.2199	0.2192	0.2201
	$\gamma = 1.0$	0.1283	0.1283	0.1289	0.2036	0.2032	0.2037
	$\gamma = 3.0$	0.2154	0.2188	0.2167	0.3189	0.3210	0.3191
H =1000	$\gamma = 0.5$	0.2349	0.2348	0.2354	0.3396	0.3390	0.3394
	$\gamma = 1.0$	0.2064	0.2061	0.2062	0.3052	0.3057	0.3057
	$\gamma = 3.0$	0.3853	0.3884	0.3858	0.5051	0.5087	0.5066

Figures in the table correspond to the share of Monte Carlo draws where the null hypothesis was rejected (out of 50 000 draws). H denotes the out-of-sample evaluation window length and γ denotes risk aversion coefficient.

Table 20: Empirical rejection probabilities under H_0 for GMVP with ridged covariance matrix vs 1/N. $\alpha = 5\%$. $N/T = 0.1$

		Two-sided			One-sided		
		Delta method	Bootstrap Percentile	Bootstrap t-statistic	Delta method	Bootstrap Percentile	Bootstrap t-statistic
H = 100	$\gamma = 0.5$	0.0607	0.0567	0.0607	0.0574	0.0552	0.0578
	$\gamma = 1.0$	0.0609	0.0567	0.0612	0.0587	0.0568	0.0590
	$\gamma = 3.0$	0.0619	0.0579	0.0621	0.0632	0.0627	0.0637
H = 500	$\gamma = 0.5$	0.0790	0.0785	0.0794	0.0718	0.0716	0.0720
	$\gamma = 1.0$	0.0784	0.0777	0.0787	0.0751	0.0750	0.0754
	$\gamma = 3.0$	0.0865	0.0858	0.0867	0.0875	0.0886	0.0881
H =1000	$\gamma = 0.5$	0.1080	0.1076	0.1084	0.0900	0.0898	0.0899
	$\gamma = 1.0$	0.1092	0.1088	0.1094	0.0959	0.0960	0.0961
	$\gamma = 3.0$	0.1213	0.1218	0.1220	0.1184	0.1197	0.1192

Figures in the table correspond to the share of Monte Carlo draws where the null hypothesis was rejected (out of 50 000 draws). H denotes the out-of-sample evaluation window length and γ denotes risk aversion coefficient.

Table 21: Power at 1% expected CE difference for GMVP with ridged covariance matrix vs 1/N. $\alpha = 5\%$. $N/T = 0.1$

		Two-sided			One-sided		
		Delta method	Bootstrap Percentile	Bootstrap t-statistic	Delta method	Bootstrap Percentile	Bootstrap t-statistic
H = 100	$\gamma = 0.5$	0.0716	0.0679	0.0721	0.1088	0.1052	0.1092
	$\gamma = 1.0$	0.0818	0.0774	0.0827	0.1250	0.1224	0.1263
	$\gamma = 3.0$	0.0878	0.0857	0.0890	0.1340	0.1339	0.1361
H = 500	$\gamma = 0.5$	0.1494	0.1487	0.1494	0.2265	0.2259	0.2271
	$\gamma = 1.0$	0.1926	0.1926	0.1931	0.2802	0.2808	0.2812
	$\gamma = 3.0$	0.2257	0.2280	0.2263	0.3190	0.3218	0.3206
H =1000	$\gamma = 0.5$	0.2451	0.2448	0.2457	0.3453	0.3456	0.3457
	$\gamma = 1.0$	0.3195	0.3192	0.3196	0.4234	0.4243	0.4240
	$\gamma = 3.0$	0.3857	0.3887	0.3868	0.4868	0.4888	0.4875

Figures in the table correspond to the share of Monte Carlo draws where the null hypothesis was rejected (out of 50 000 draws). H denotes the out-of-sample evaluation window length and γ denotes risk aversion coefficient.