Abstract

This paper exploits the idea of pretesting to choose between competing portfolio strategies. We propose an estimator for a portfolio weight vector, which optimally trades off between Type I and Type II errors when choosing the best investment strategy. Furthermore, we accommodate the idea of bagging in the portfolio testing problems, which helps to avoid sharp thresholding and reduces the amount of portfolio turnover.

Our approach borrows from both shrinkage and forecast combination literature. The portfolio weights of our strategy are weighted averages of the portfolio weights from a set of stand-alone strategies. More specifically, the weights are generated from a pseudo out-of-sample portfolio pretesting, such that they reflect the probability that a given strategy will be overall best performing. Contrary to previous approaches, the shrinkage intensity is continuously updated to incorporate the most recent information in the rolling window forecasting set-up. We show that the bagged pretest estimator performs exceptionally well, especially when combined with adaptive smoothing. The resulting strategy allows for a flexible and smooth switch between the underlying strategies and is shown to outperform the corresponding stand-alone strategies.

Keywords: pretest estimation, bagging, portfolio allocation, adaptive learning

JEL classification: C12, C52, C58, G11
1 Introduction

Since the seminal work by Markowitz (1952) finding an optimal portfolio allocation has been a great challenge in empirical finance. In particular, for realistic portfolio dimensions it has been widely accepted that due to inherent large sampling errors, the central input parameters, the mean vector and the covariance matrix of returns, cannot be estimated with sufficient precision. As a consequence the empirical portfolio weights are often unrealistic in size, suffer from extreme outliers and large fluctuations over time and thus reveal a very poor out-of-sample portfolio performance. In the last two decades a spectrum of alternative regularization strategies has been proposed, which showed some improvements in terms of out-of-sample performance compared to the plug-in estimators. Despite these improvements the naive, equally weighted (1/N) portfolio remains a strong competitor and often cannot be significantly outperformed by the sophisticated regularization approaches (e.g. DeMiguel et al., 2009). In particular, for higher dimensional and high dimensional portfolios the improvements are still far from being satisfactory. If at all, contrary to their theoretical counterparts empirical portfolios hardly improve in performance when the portfolio dimension is increased, i.e. the theoretical gains from improving the diversification are largely eliminated by the losses due to the increase in estimation noise.

In this paper we present a unifying machine learning approach to portfolio allocation, which benefits from different pathways of the modern machine learning and statistics literature including algorithmic learning, bagging, model averaging and regularization, which all as stand alone methods have been shown to contribute to a reduction of the estimation noise. The key and novel idea of our approach is to use the information from pretesting different portfolio strategies and estimators to optimize the out-of-sample portfolio performance. We show that the information from the pretest is of great value when the return series are noisy and complex (e.g. as a consequence of structural breaks) and a reshuffling of the portfolio weights has to be traded-off against increasing turn-over costs.

In the empirical portfolio literature a large class of strategies to stabilize the estimates of the portfolio weights either tries to shrink the portfolio weights directly or regularizes the estimates of the mean return vector and/or the variance-covariance matrix. Notable examples among many other contributions are Jagannathan and Ma (2003), who propose to impose a norm-constraint
directly to the portfolio optimization for stabilizing the weight estimates in small samples to account for no short sale constraints. In a similar spirit Brodie et al. (2009), Li (2015) Goto and Xu (2015) use the $\ell_1$-penalization (lasso) and the $\ell_2$-penalization (e.g. Yen, 2015) to constrain the portfolio weights. In a series of papers Ledoit and Wolf (2003), Ledoit and Wolf (2004a), Ledoit and Wolf (2017) focus on the regularization the covariance matrix as the key ingredient in portfolio optimization. Kourtis et al. (2012) propose a shrinkage approach for the inverse of covariance matrix, which can be directly used in most of the portfolio weight estimators. However, the distinction between norm constraining the portfolio weights and regularization of the moment estimates is somewhat artificial as for many cases both strategies lead to the same portfolios (DeMiguel et al., 2009; Fan et al., 2012).

An alternative class of regularization strategies has its root in the James-Stein shrinkage idea, where the weights of the shrinkage portfolio are linear combinations of different portfolio strategies, typically one with an asymptotically optimal weight vector and another one with no estimation estimation noise (e.g. the equally weighted portfolio). The weights are chosen such that a given loss function is minimized (e.g. Kan and Zhou, 2007; Frahm and Memmel, 2010; Tu and Zhou, 2011; Ao et al., 2018). As these estimators yield out-of-sample portfolio returns which are themselves also linear combinations of the standalone out-of-sample portfolio returns, on which the out-of-sample portfolio performance is evaluated, they can be regarded as forecasting combination estimators even is the loss function is a within-sample criterion.

Despite the plethora of robustification strategies no unambiguous conclusion can be drawn whether any of these approaches can outperform the equally weighted portfolio for finite samples in a general setup because the finite sample properties depend on many factors, which are only partly under the control of the investor. For instance, the trading frequency influences the comparison, as daily rebalancing usually invokes higher transaction costs compared to monthly rebalancing. Ranking of the investment strategies often changes in terms of different portfolio criteria, e.g. the best strategy according to Sharpe ratio is not necessarily the one which has the lowest investment risk. The choice of the sample size $T$, which is used for estimating portfolio weights changes the performance of estimated strategies, as regularization-based estimators require larger samples to show superior performance. Last but not least, the time period considered for the evaluation is crucial, as during the high volatility period on the market
shrinking the weights towards a more stable equally weighted portfolio is particularly appealing. Notably, for the regularization-based approaches the choice of tuning parameter plays a crucial role. However there is often no clear guidance on the choice of optimal shrinkage (regularization) intensity choice. Even asymptotically, the stochastic properties of shrinkage estimators strongly depend on the underlying assumptions on the dimensionality of the problem (e.g. under the standard assumption with T approaching infinity given a fixed N or under high dimensionality with both N and T approaching infinity). In this sense asymptotic findings are of little practical guidance for an investor, who is looking for a unified approach that is flexible enough to adapt to a given investment scenario.

In this paper we therefore address the empirical portfolio allocation problem from a different perspective. Instead of working with a single regularization strategy or a combination of strategies, we develop a flexible algorithm which optimally combines a given set of weight estimates in a data-driven and time-adaptive way with respect to a portfolio performance measure of choice.

The new methodology makes four novel contributions to the literature. Our first contribution is to use pretesting as a statistical tool to choose a superior strategy with respect to a given out-of-sample portfolio performance measure. Choosing between strategies according to the outcome of a pretest has the advantage that due to the metric of the performance test, a more sophisticated and supposedly less robust strategy is only chosen if the performance difference compared to a simple, more robust strategy is large and the estimated performance difference is estimated with sufficient precision.

In its basic form our pretest estimator is a binary rule based on the simple t-test checking if there is a significant difference in the performance of two strategies measured by the out-of-sample certainty equivalent or the Sharpe ratio, both net of the transaction costs. Based on the test outcome the investor chooses which strategy to take for the next invest period. The most commonly used performance tests which are used in the literature are the ones proposed by Ledoit and Wolf (2008) and DeMiguel et al. (2009).

Previous work by Kazak and Pohlmeier (2019), however, has shown that the existing portfolio performance tests are correctly sized, but have a very low power in realistic scenarios. Therefore the test outcomes in many empirical horse races, where sophisticated, data-driven strategies failed to significantly outperform the 1/N strategy can simply be explained by the lack of power.
of these tests given conventional significance levels. This in particular holds true, if the 1/N portfolio is taken as the benchmark portfolio.

Our second contribution is the use of information from the pretest in an optimal and time adaptive manner. As standard significance levels are basically meaningless due to the low power of the tests, we choose the actual significance level in order to optimize the trade-off between Type I and Type II error with respect to the chosen portfolio performance measure. Thus the significance level serves as the data-driven tuning parameter. In a rolling window set-up we propose to make the tuning parameter to be time adaptive. This guarantees that not only the most recent information is used for the decision but also the past behavior of the tuning parameter.

Our pretest strategy has certain similarities with James-Stein type of combination estimators mentioned above. In theory the optimal combination weights are functions of the unknown variance covariance matrix. In empirical studies the combination weights are estimated by replacing the theoretical variance covariance matrix by its sample counterpart. Therefore, the lack of precision in estimating the variance-covariance matrix directly translates into a low precision of the estimates of the combination weights particularly in higher dimensional settings. Hence, by using optimal but estimated combination weights additional estimation noise is added. This phenomenon is well documented in the literature on optimal point forecasts, where models with fixed sub-optimally chosen weights outperform models with optimal estimated weights (e.g. Timmermann, 2006).

The classical version of our pretest strategy faces similar challenges as the optimal combination estimators. Here the challenges of high dimensionality translate into the quality of statistical tests and therefore generate noisy test outcomes and pretest estimates. Therefore our third contribution is to introduce the combination of bootstrap aggregation (bagging) with pretest estimation in the portfolio context. We modify the classical pretest estimator replacing the indicator functions of test decisions with the bootstrapped probabilities. The proposed combination of different portfolio allocations weighs every strategy proportional to the probability that the null hypothesis of its inferior performance compared to the benchmark is rejected, thus the weights for each strategy are interpretable and fully data-driven.

We show that the bagging step substantially stabilizes the pretest estimator and reduces
portfolio turnover. The sources of this gain are manifold. First, bagging per se can be shown to stabilize any binary decision problem. Secondly, as pretest estimation is a special type of regularization method with sharp thresholding. Replacing the binary indicator by the bootstrapped probabilities abolishes sharp thresholding property and leads to a smooth probability distribution of the pretested portfolio weights.

As a weighting estimator for the portfolio weights our approach has superior properties compared to the stand-alone approaches. The literature on forecast combinations distinguishes between two directions of combining forecasts (e.g. Wei and Yang, 2012). One direction, searches for the optimal weights given a true data generating process. The approaches by Kan and Zhou (2007), Frahm and Memmel (2010), Tu and Zhou (2011) and others belong to this this direction and have to cope with the estimation noise resulting from the estimation of the unknown optimal combination weights. The second direction combines forecasts for adaptation. The bagged combination approach proposed in this paper follows this direction, which intends to improve forecasts than every stand-alone candidate by exploiting the concavity of the objective function (see e.g. Bonaccolto and Paterlini, 2020).

Finally, bagging reduces the impact of outliers. This is highly relevant, if one of the strategies to be considered is the tangency portfolio. As Okhrin and Schmid (2006) have shown for the case of iid-normality of the returns the plug-in portfolio weight estimates have no first and second moments and follow a multivariate type of Cauchy distribution. Their theoretical finding, which to some extend was neglected in the empirical finance literature, can explain the large outliers found in the returns of the tangency portfolio and makes the plug-in estimator of tangency portfolio basically useless in applied work. We show that bagging helps to reduce the impact of outliers and provides the opportunity to use the tangency portfolio in applied work. To our knowledge, our work is the first one which addresses the outlier problem in the the empirical portfolio literature.

Finally, our last contribution is to show in an extensive empirical study that the proposed bagged pretest estimator outperforms the underlying weight estimation strategies and other competitors. The results are shown to be robust with respect to different parameter constellations: estimation window length, portfolio dimension, choice of portfolio performance measure and trading frequency. Our empirical study supports the theoretical work from the combination
forecast literature that the cost of combining for improvement due to parameter estimation is substantially higher than that of combining for adaptation.

This paper is organized as follows. In Section 2 we use a simple motivating example which illustrates the problem of an optimal strategy choice and propose the novel bagged pretest estimator. Section 3 provides an empirical illustration of the proposed method. Section 4 summarizes the main findings and gives an outlook on future research.
2 Pretest Estimator

This section introduces the setup of a rolling window portfolio choice and illustrates the problem of sample and portfolio size instability. We then introduce the pretest estimator, which is either based on in-sample (in) or out-of-sample (os) portfolio performance comparison. Finally, Section 2.3 introduces bagged pretest estimator.

Consider a standard portfolio choice set-up with $N$ risky assets. Let $r_t$ be an excess return vector at time $t$ with mean vector $\mathbb{E}[r_t] = \mu$ and variance-covariance matrix $V[r_t] = \Sigma$. Moreover, let $\omega(s) = \omega(s)(\mu, \Sigma)$ be the $N \times 1$ vector of portfolio weights for strategy $s$, e.g. $\omega(g) = \frac{\Sigma^{-1}\mu}{\mu^\prime \Sigma^{-1} \mu}$ for the global minimum variance portfolio (GMVP) minimizing the portfolio variance, $\omega(e) = \frac{1}{N} t$ for the equally weighted portfolio and $\omega(tn) = \frac{\Sigma^{-1} \mu}{\mu^\prime \Sigma^{-1} \mu}$ for the tangency portfolio, maximizing the Sharpe ratio\(^1\). For a strategy $s$ the portfolio return at time $t$ is given by $r^p_t(s) = \omega(s)^\prime r_t$ with mean $\mu^p(s) = \mathbb{E}[r^p_t(s)] = \omega(s)^\prime \mu$ and variance $\sigma^2_p(s) = V[r^p_t(s)] = \omega(s)^\prime \Sigma \omega(s)$.

Consider a portfolio performance measure for the strategy $s$, which could be for example the certainty equivalent (CE) given by $CE(\omega(s)) = \mu^p(s) - \gamma \sigma^2_p(s)$ with $\gamma$ being the risk aversion coefficient of the investor; or a Sharpe ratio $SR(\omega(s)) = \frac{\mu^p(s)}{\sigma_p(s)}$. The CE is the return which makes the investor indifferent between investing into the risky portfolio or receiving the certainty equivalent return $U(CE(\omega(s))) = \mathbb{E}[U(r^p(s))], with $U(\cdot)$ being the utility function of the investor (Merton and Samuelson, 1992). Our analysis below concentrates on the CE as a performance measure, because of its simple return interpretation. However, our proposed pretest strategies can be easily extended to the Sharpe ratio or any other popular portfolio evaluation criterion, whenever there exists an appropriate statistical performance test, for an example consider Section 3.2. In the following strategy $s$ is said to outperform strategy $\tilde{s}$ if the difference in certainty equivalents is non-negative $\Delta_0(s, \tilde{s}) = CE(\omega(s)) - CE(\omega(\tilde{s})) \geq 0$.

For given population parameters $(\mu, \Sigma)$ the dominant strategy according to a chosen portfolio evaluation criteria is known. However, in empirical applications the first two moments of the return process have to be estimated and estimation risk has to be taken into account. Furthermore, investment decision is a dynamic process, therefore financial or forecasting risk has to be accounted for as well. In empirical work the actual performance of competing portfolio

\(^1\)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.
allocation strategies is often evaluated based on the out-of-sample certainty equivalent, which takes into account both estimation and forecasting risk (Kazak and Pohlmeier, 2019). In the following we consider a typical rolling window set-up, where for the period \( t + 1 \) the out-of-sample portfolio return \( \hat{r}_{t+1}^p(s) \) is based on a one-step forecast of the portfolio weights \( \hat{\omega}_{t+1|t}(s) \) with period \( \{ t - T, \ldots, t \} \) as an 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: \( \hat{r}_{t+1}^p(s) = \hat{\omega}_{t+1|t}(s)'r_{t+1} = \hat{\omega}_t(s)'r_{t+1} \). The estimation window is shifted one period ahead \( H \) times resulting in the \( H \times 1 \) vector of the out-of-sample portfolio returns \( \hat{r}_t^p(s) \). Different portfolio strategies are then evaluated based on the out-of-sample certainty equivalent \( \hat{CE}_{os}(\hat{\omega}(s)) \) given by:

\[
\hat{CE}_{os}(\hat{\omega}(s)) = \hat{\mu}_{os}(s) - \frac{\gamma}{2} \hat{\sigma}_{os}^2(s),
\]

where:

\[
\hat{\mu}_{os}(s) = \frac{1}{H} \sum_{h=1}^{H} \hat{r}_{t+h}^p(s) = \frac{1}{H} \sum_{h=1}^{H} \hat{\omega}_{t+h-1}(s)'r_{t+h},
\]

\[
\hat{\sigma}_{os}^2(s) = \frac{1}{H-1} \sum_{h=1}^{H} \left( \hat{r}_{t+h}^p(s) - \hat{\mu}_{os}(s) \right)^2.
\]

The driving force of the portfolio performance based on the out-of-sample CE is estimation noise, i.e. theoretically superior strategies usually do not perform well in empirical applications as the estimation error dominates the theoretical gain. In particular, out-of-sample portfolio performance is sensitive to the length of the sample \( T \) used to estimate the weights \( \hat{\omega}_t(s) \). For a given portfolio dimension \( N \), even a slight decrease in \( T \) might lead to a sharp increase in the variance of estimated portfolio weights, thus change the out-of-sample CE dramatically.

### 2.1 Motivating Example

This example demonstrates that no general statement can be made with respect to the optimal allocation strategy for a given portfolio space and estimation window length. A slight change in the estimation set-up might lead to reverse ranking of portfolio investment strategies.

Consider an investor who chooses among three different strategies and rebalances the portfolio monthly. Based on monthly excess returns of 100 industry portfolios from K.R.French database\(^2\) the investor estimates the weights of the following strategies:

\(^2\)The data is taken from K.R.French website and contains monthly excess returns from 05/1964 till 12/2015.
1. GMVP based on the plug-in sample covariance matrix estimator

\[
\hat{\omega}(g) = \frac{\hat{\Sigma}^{-1}\iota}{\iota ^{\intercal} \hat{\Sigma}^{-1} \iota },
\]

where \( \hat{\Sigma} = \frac{1}{T} \sum_{t=1}^{T} (r_t - \bar{r})(r_t - \bar{r})^{\prime} \) and \( \iota \) is an \( N \times 1 \) vector of ones.

2. Tangency portfolio with a shrunken covariance matrix

\[
\hat{\omega}(tn) = \frac{\hat{\Sigma}^{-1} \hat{\mu}}{\iota ^{\intercal} \hat{\Sigma}^{-1} \hat{\mu}},
\]

where \( \hat{\mu} = \frac{1}{T} \sum_{t=1}^{T} r_t \) and \( \hat{\Sigma} = (1 - \delta) \hat{\Sigma} + \delta I_N \) with \( \delta = \frac{0.05N}{1 + 0.05N} \), \( I_N \) - identity matrix of size \( N \), \( \hat{\Sigma} = \frac{1}{T} \sum_{t=1}^{T} (r_t - \bar{r})(r_t - \bar{r})^{\prime} \). The shrunken covariance matrix \( \hat{\Sigma} \) is defined in the similar spirit to Ledoit and Wolf (2003) with identity matrix as target.

3. Equally weighted portfolio

\[
\hat{\omega}(e) = \omega(e) = \frac{1}{N} \iota.
\]

For the portfolio evaluation the length of the out-of-sample period is set to \( H = 500 \) observations and the risk aversion parameter to \( \gamma = 1 \). Transaction costs at period \( t \) which an investor has to pay every month after portfolio rebalancing are computed as follows:

\[
TC_t(s) = c \cdot \sum_{j=1}^{N} |\hat{\omega}_{j,t+1}(s) - \hat{\omega}_{j,t}(s)|,
\]

where \( TC_t(s) \) denotes transaction costs for strategy \( s \) at period \( t \), \( \hat{\omega}_{j,t+1} \) - portfolio weight before rebalancing at \( t + 1 \) and \( c \) - cost per transaction (5 basis points, DeMiguel et al. (2009)). The out-of-sample CE is then computed based on the net portfolio returns \( \hat{r}_{t, \text{net}}(s) = \hat{r}_t^p(s) - TC_t(s) \) in the very similar way to (1). For a more general comparison for a given pair of \( (N, T) \) we randomly draw unique asset subsets of size \( N \) from the available asset space and compute the out-of-sample CE for each of the 500 random draws.

The upper-left panel of Figure 1 depicts the average annualized out-of-sample CE over 500 randomly drawn asset subsets for a grid of portfolio sizes \( N \) (x-axis) and the in-sample estimation window length \( T \) of 15 years (180 months). The equally weighted portfolio (in dotted red) does not require weight estimation and is therefore stable across the portfolio sizes \( N \). The
Figure 1: Annualized out-of-sample CE for Underlying Strategies

Lines on the plots represent the average annualized out-of-sample CE in % computed on net portfolio returns according to (1) over a 500 randomly drawn portfolios of size $N$ (x-axis). For each randomly drawn portfolio the out-of-sample CE is computed over an out-of-sample evaluation horizon of $H = 500$ observations, risk aversion parameter $\gamma = 1$. Left panels depict the average out-of-sample CE with the weights of GMVP (in dashed black) from (2), tangency portfolio (in solid blue) from (3), and equally weighted portfolio (in dotted red) from (4) computed over an estimation window length of $T = 180$ observations. Right panels plot the average out-of-sample CE of the same strategies where the weights were estimated on $T = 120$. Shaded areas on the upper row of the figure represent the 95-percentile intervals over 500 randomly drawn portfolios. The upper row of plots corresponds to the [1%, 15%] limits of the y-axis and the lower plots correspond to the [−60%, 15%] limits of the y-axis.

Out-of-sample CE of GMVP (dashed black line) decreases as the sample covariance matrix of the returns adds estimation noise with increase in $N$. The tangency portfolio (solid blue line) with shrunken covariance matrix outperforms competing strategies for all $N > 20$. The shaded areas around the lines reflect the distribution of the CE across randomly drawn portfolios. For $N > 50$ tangency portfolio combined with shrinkage outperforms $1/N$ not only in mean, but across 95% of different asset combinations. The upper-right panel depicts the performance of the same strategies, but with the estimation sample length $T = 120$. The out-of-sample CE of GMVP decreases steeper and has larger variance, as there are fewer observations available to estimate portfolio weights, thus $1/N$ outperforms the GMVP in mean for $N > 50$. Notably the equally weighted portfolio produces a larger out-of-sample CE over the 95% of 500 portfolios only for $N = 90$. The solid blue line of tangency portfolio only becomes visible by rescaling
the y-axis. As can be seen on the lower-right panel, the performance of the tangency portfolio worsens dramatically. Note the difference in scaling of the y-axes, e.g. for \( N = 20 \), \( T = 180 \) the out-of-sample CE of the tangency portfolio corresponds to 10% return in annualized terms, whereas taking \( T = 120 \) observations for the weight estimation results in \(-10\%\) annualized loss. The tangency portfolio is known to produce extremely unstable weight estimates for the smaller estimation windows, while the performance of the GMVP and the equally weighted portfolio is considerably less sensitive (Okhrin and Schmid, 2006). Remarkably, the performance of tangency portfolio in (3) depends on the shrinkage intensity for the covariance matrix, and this example illustrates the problem of shrinkage estimators performing well only for larger samples.

However, in some circumstances it is better to keep the estimation window length smaller, e.g. due to structural breaks in the financial markets only the recent information should be included in the weight estimation. Therefore in this paper we propose a data driven procedure for an optimal portfolio allocation strategy choice, which outperforms the underlying weight estimation strategies regardless of the parameter constellations.

### 2.2 Pretest as a Decision Rule

In the first part of our algorithm we develop an appropriate decision rule for choosing a portfolio allocation strategy. We use pretesting as a statistical tool helping the investor to decide between different strategies in a data-driven way. Assume for the sake of notation simplicity that the investor has to decide between two alternative strategies \( s \) and \( \tilde{s} \) and maximizes the expected utility\(^3\). The difference in the out-of-sample CE’s between the two strategies \( \Delta_{os}(s, \tilde{s}) \) is defined as

\[
\Delta_{os}(s, \tilde{s}) = CE_{os}(\hat{\omega}_t(s)) - CE_{os}(\hat{\omega}_t(\tilde{s})),
\]

\[
CE_{os}(\hat{\omega}_t(s)) = \mu_{os}(s) - \frac{\gamma}{2} \sigma_{os}^2(s), \quad \text{with}
\]

\[
\mu_{os}(s) = E[\hat{r}^p_{t+1}(s)] = E[\hat{\omega}_t(s)]' \mu,
\]

\[
\sigma_{os}^2(s) = V[\hat{r}^p_{t+1}(s)] = E[\hat{\omega}_t(s)' \Sigma \hat{\omega}_t(s)] + \mu' V[\hat{\omega}_t(s)] \mu.
\]

\(^3\)Note that all the conclusions of this section also hold for testing Sharpe ratio difference or difference in portfolio variance.
Note that the out-of-sample CE contains the term $-\gamma^2 \mu' \mathbb{V} [\hat{\omega}_t(s)] \mu$, which penalizes estimation-based strategies compared to the equally weighted portfolio, for which $\mathbb{V} [\hat{\omega}_t(e)] = 0$. Therefore, the pretest estimator is constructed based on the $\Delta_{os}(s, \tilde{s})$ to take into account estimation risk differences. The goal is to select either strategy $s$ or strategy $\tilde{s}$ depending on the test outcome. Null and alternative hypotheses take the one-sided form:

$$H_0 : \Delta_{os}(s, \tilde{s}) \leq 0 \quad \text{and} \quad H_1 : \Delta_{os}(s, \tilde{s}) > 0. \quad (6)$$

Let the pretest estimator of the portfolio weights forecasts for $t + 1$ be such that it depends either on strategy $s$ in case the null is rejected or on $\tilde{s}$ otherwise:

$$\omega_t(s, \tilde{s}, \alpha) = \mathbb{I} \{\hat{\Delta}_{os}(s, \tilde{s}) > \Delta^*(\alpha)\} (\omega_t(s) - \omega_t(\tilde{s})) + \omega_t(\tilde{s}), \quad (7)$$

with the estimated CE difference $\hat{\Delta}_{os}(s, \tilde{s}) = \hat{CE}_{os}(s) - \hat{CE}_{os}(\tilde{s})$ according to (1) and the critical value $\Delta^*(\alpha)$ for a significance level $\alpha$. In other words, the pretest estimator selects the strategy $s$ if it is significantly better than the alternative and the sensitivity of the pretest estimator depends on $\alpha$: the lower the nominal level, the stricter is the pretest rule and the greater should be the difference $\hat{\Delta}_{os}(s, \tilde{s})$ for selecting the strategy $s$ over $\tilde{s}$.

There are three major difficulties arising from applying the pretest estimator defined in (7) in practice. First of all, at time period $t$ the investor does not know the out-of-sample CE difference $\hat{\Delta}_{os}(s, \tilde{s})$ over $H$ periods ahead, therefore the test decision from (6) is unknown. Secondly, in empirical applications an investor has to decide on $\alpha$, which influences the performance of the pretest estimator, e.g. for $\alpha = 50\%$ ($\Delta^*(\alpha) = 0$) the CE of strategy $s$ has to be just slightly larger than the CE of $\tilde{s}$ in order to be selected, whereas for the commonly used levels of significance of 1% and 5% the difference $\hat{\Delta}_{os}(s, \tilde{s})$ has to be fairly large for the null rejection. Kazak and Pohlmeier (2019) show that in realistic scenarios the empirical power of the portfolio performance tests is very low, which implies that even if the strategy $s$ is truly superior, the pretest estimator is not able to select the dominating strategy with high probability. Therefore going for the conservative $\alpha$-level is not a reasonable choice, as in the presence of low power it will force the pretest estimator to select $\tilde{s}$ even in cases when $s$ is dominating. In particular, the problem of the low power calls for an optimal trade-off between Type I and Type II error. Last but not
least, instability in weight estimation leads to a large variation in out-of-sample CE difference, which translates into highly volatile test decisions in the rolling window set-up. In this paper we propose a feasible pretest estimator with a data-driven and time adaptive significance level choice in a combination with bagging, which solves the instability problem.

2.2.1 Feasible Pretest Strategy

i.) Within-sample Pretesting

We solve the first issue of unfeasible testing by constructing the pretest weight estimator based on the in-sample CE difference which is available at time $t$. First, at time $t$ the weights $\hat{\omega}_t(s)$ are estimated based on the sample $\{t-T, \ldots, t\}$. The estimated within sample CE for the strategy $s$ is computed as

$$C E_{in,t}(s) = \hat{C}E_{in}(s|t-T, \ldots, t) = \hat{\omega}_t(s)' \bar{r}_t - \frac{\gamma}{2} \hat{\omega}_t(s)' \hat{\Sigma}_t \hat{\omega}_t(s), \quad (8)$$

where $\bar{r}_t$ denotes the sample mean and $\hat{\Sigma}_t$ the sample covariance matrix of the returns based on the estimation window $\{t-T, \ldots, t\}$. The in-sample test statistic deciding between $s$ and the benchmark $\tilde{s}$ is defined as

$$t_{in,t}(s, \tilde{s}) = \frac{C E_{in,t}(s) - \hat{C}E_{in,t}(\tilde{s})}{S.E. \left[ C E_{in,t}(s) - \hat{C}E_{in,t}(\tilde{s}) \right]},$$

where the standard error for the CE difference is computed via delta method (DeMiguel et al., 2009) or an appropriate block bootstrap (Ledoit and Wolf, 2008).\footnote{Appendix A provides the details on the commutation of the standard error.} For the multivariate comparison where the investor selects between $M$ alternatives $s_1, \ldots, s_M$ and a benchmark strategy $\tilde{s}$ and for a given significance level $\alpha$ the pretest estimator based on the in-sample CE is defined as

$$\hat{\omega}_{in,t}(S, \alpha) = \hat{\omega}_{in,t}(s_1, \ldots, s_M, \tilde{s}, \alpha) = \sum_{i=1}^{M} 1(s_i, \alpha) \hat{\omega}_t(s_i) + \left(1 - \sum_{i=1}^{M} 1(s_i, \alpha)\right) \hat{\omega}_t(\tilde{s}), \quad (9)$$
where $S = \{s_1, ..., s_M, \tilde{s}\}$. The strategy selection function is defined as

$$1(s_i, \alpha) = \{ \max \{ t_{in,t}(s_1, \tilde{s}), ..., t_{in,t}(s_M, \tilde{s}), q_{\alpha} \} = t_{in,t}(s_i, \tilde{s}) \}, \quad i = 1, ..., M,$$

where $q_{\alpha}$ is the corresponding asymptotic critical value for the nominal level $\alpha$: the $(1 - \alpha)$ quantile of the standard normal distribution. In other words, the pretest estimator selects the strategy with the largest standardized difference from the benchmark $\tilde{s}$ and which at the same time crosses the threshold $q_{\alpha}$.

However, selecting the strategy based on the in-sample comparison does not necessarily provide a good out-of-sample choice: the pretest estimator as defined in (9) does not take into account transaction costs and the forecasting risk.

ii.) Out-of-sample Pretesting

We therefore suggest to use another feasible pretest estimator, which may be obtained by performing a pseudo-out-of-sample exercise, commonly used for parameter training in statistical learning. The goal of the pretest estimator is to select an optimal strategy in a data-driven way which results in the highest out-of-sample CE, taking transaction costs into account. A feasible out-of-sample pretest estimator is thus obtained by dividing the within-sample period into two parts, where the first part is used for the weight estimation and the second part is used for the computation of pseudo-out-of-sample returns and the pseudo-out-of-sample CE.\(^5\) The CE-optimal strategy is then defined as the one having the largest pseudo-out-of-sample CE net of the transaction costs. In particular, the weights for the strategy $s$ are computed based on the sample of length $T - \kappa$: $\{t - T, ..., t - \kappa\}$. The out-of-sample portfolio returns are computed in the rolling window of length $\kappa$ and the transaction costs are subtracted from the out-of-sample returns.

\(^5\)In machine learning literature the first part would be typically referred to as *training set* and the second part as a *validation set*. 
returns at each time point $t$ as in (5). The resulting pseudo-out-of-sample CE is similar to (1):

$$\hat{CE}_{os,t}(\hat{\omega}(s)) = \hat{\mu}_{os,t}(s) - \frac{\gamma}{2} \hat{\sigma}_{os,t}^2(s),$$

(10)

where:

$$\hat{\mu}_{os,t}(s) = \frac{1}{\kappa} \sum_{h=0}^{\kappa-1} \hat{r}_t^{p}(s) - \frac{1}{\kappa} \sum_{h=0}^{\kappa-1} \hat{\omega}_{t-h-1}(s)' r_{t-h},$$

$$\hat{\sigma}_{os,t}^2(s) = \frac{1}{\kappa - 1} \sum_{h=0}^{\kappa - 1} \left( \hat{r}_t^{p}(s) - \hat{\mu}_{os,t}(s) \right)^2.$$  

The out-of-sample pretest estimates of portfolio weights are computed based on the difference in the pseudo-out-of-sample CE:

$$t_{os,t}(s_i, \tilde{s}) = \frac{CE_{os,t}(s) - CE_{os,t}(\tilde{s})}{\text{S.E.} \left[ CE_{os,t}(s) - CE_{os,t}(\tilde{s}) \right]}.$$  

(11)

$$\hat{\omega}_{os,t}(S, \alpha) = \hat{\omega}_{os,t}(s_1, ..., s_M, \tilde{s}, \alpha) = \sum_{i=1}^{M} 1_{os}(s_i, \alpha) \hat{\omega}_t(s_i) + \left( 1 - \sum_{i=1}^{M} 1_{os}(s_i, \alpha) \right) \hat{\omega}_t(\tilde{s}),$$

for $i = 1, ..., M$ and $S = \{s_1, ..., s_M, \tilde{s}\}$, where the out-of-sample strategy selection function is given by

$$1_{os}(s_i, \alpha) = 1 \{ \text{max} [t_{os,t}(s_1, \tilde{s}), ..., t_{os,t}(s_M, \tilde{s}), q_{\alpha}] = t_{os,t}(s_i, \tilde{s}) \}.$$  

(13)

The pseudo-out-of-sample CE from (10) is used only for the computations of the test statistic and corresponding strategy selection function in (13), and the out-of-sample pretest estimator selects the weight estimates computed on the whole in-sample estimation window. Note, that the out-of-sample portfolio returns used in the pseudo-out-of-sample CE are net of the transaction costs, therefore by construction the pretest estimator takes into account the amount of rebalancing or turnover. Also note, that the estimation window used in (10) is reduced to $T - \kappa$ observations of the initial estimation window $T$, which results in adding more estimation noise to the problem and making the weight estimates more unstable compared to in-sample pretesting.

### 2.2.2 Significance Level Selection

So far the selection strategy estimator was defined (13) is defined for a given significance level. We now address the problem of the optimal choice of the significance level $\alpha$. From the purely
statistical perspective, increasing the nominal level in order to improve on the power of the test is not meaningful. From the investors perspective, however, there is a well-defined portfolio performance measure, such that the trade-off between the Type I and Type II error of the pretest estimator can be optimized with respect to it. We propose to select a significance level $\alpha$, which maximizes the difference in portfolio performance measure.

The proposed significance level algorithm is constructed as follows. At each period $t$ the pretest weight estimates either in-sample or out-of sample, $\hat{\omega}_{in/os,t}(S, \alpha_j)$, are computed for the grid of $(0, 0.5)$ $\alpha$-values of length $J$. For every $\hat{\omega}_{in/os,t}(S, \alpha_j)$, $j = 1, \ldots, J$ the CE of the pretest estimator is computed similarly to (8) or (10):

$$\hat{CE}_{in,t}(S, \alpha_j) = \hat{\omega}_{in,t}(S, \alpha_j)' \tilde{r}_t - \frac{\gamma}{2} \hat{\omega}_{in,t}(S, \alpha_j)' \hat{\Sigma}_t \hat{\omega}_{in,t}(S, \alpha_j),$$

$$\hat{CE}_{os,t}(S, \alpha_j) = \hat{\mu}_{os,t}(S, \alpha_j) - \frac{\gamma}{2} \hat{\sigma}_{os,t}^2(S, \alpha_j).$$

Finally, the CE-optimal significance level $\alpha^*_{t+1}$ is selected for the test, determining the strategy for the next period $t + 1$:

$$\alpha^*_{in,t+1} = \arg \max_{\alpha} \hat{CE}_{in,t}(S, \alpha_j),$$
$$\alpha^*_{os,t+1} = \arg \max_{\alpha} \hat{CE}_{os,t}(S, \alpha_j).$$

(14)

In other words, the algorithm selects the smallest significance level where the pretest estimator switches to the strategy with the highest Certainty Equivalent. The above procedure is repeated with every shift of the estimation window. Thus with the change of the estimation window the sequence of $\alpha^*_{in/os}$’s accounts for the changes of the return process across time, e.g. volatility regimes. In practice, however, this simple procedure produces rather unstable series of significance levels $\{\alpha^*_{in/os,t+1}, \ldots, \alpha^*_{in/os,t+H}\}$, as the choice of $\alpha^*_{in/os,t+1}$ is data driven and also depends on the bumpy estimates of the portfolio weights. See for example the black solid line in Figure 2 for the roughness of the series of CE-optimal significance levels based on monthly returns of $N = 10$ assets in Section 2.1.

In order to mitigate the bumpiness problem we suggest to smooth the $\alpha^*_{in/os}$-series adaptively according to

$$\alpha^s_{t+1} = (1 - \lambda)\alpha^*_{in/os,t+1} + \lambda \alpha^s_t,$$

(15)
The plotted lines depict the selected significance levels over time based on out-of-sample CE pretesting as in (13). The black line corresponds to the selected significance level according to (14); the red line corresponds to (15), where the smoothing parameter $\lambda$ was selected in a data-driven way. Analysis is based on the monthly return data from the example in Section 2.1, with $N = 10$, $T = 120$, $\gamma = 1$.

where the tuning parameter $\lambda \in (0,1)$ is selected to control the degree of smoothness. The adaptive smoothing takes into account not only the latest optimal choice $\alpha_{in/os,t+1}^*$ but also the previous estimates with geometrically decaying weights. The smoothing parameter $\lambda$ is selected via a grid search in the similar fashion as the $\alpha_{in/os,t+1}^*$: for a given couple $(\alpha_{in/os,t+1}^*, \alpha_s^*)$ the in-sample or out-of-sample CE is computed on the grid of $\lambda$’s and the optimal $\lambda$ is the one maximizing the CE of the pretest estimator. The effect of adaptive smoothing is depicted in Figure 2, which plots the evolution of the data-driven significance level choices for the out-of-sample CE pretesting over time. The black line corresponds to the significance level selection from (14), which is optimized over a grid. The red line corresponds to adaptively smoothed significance level from (15), which is still time-varying, but is considerably more stable compared to the original grid-search selected values. The selected significance level ultimately determines how large should the standardized difference between certainty equivalents of a strategy $s_i$ and $\tilde{s}$ be, for the strategy $s_i$ to be selected. Producing an instantaneous large positive difference in performance measure is not sufficient. If in the previous period the standardized difference had to be large to reject the null, it should be large in the next period as well. The significance level
selection $\alpha_{os}$ is highly sensitive to the changes in the variance of the CE difference (denominator of (11)), which given the same level of CE difference is larger in crisis periods and results in the lower test statistic. This implies that if a strategy $s_i$ slightly outperforms $\tilde{s}$, the threshold $q_\alpha$ becomes smaller and takes value around 0. If in the next month however the pseudo out-of-sample CE of $s_i$ is strongly larger than of the benchmark $\tilde{s}$ the threshold $q_\alpha$ increases. This pattern is nicely reflected in Figure 2 around the financial crisis of 2008. Over-adjusting for estimation noise leads to a rapid switch between the strategies and thus high turnover. The smoothed significance level $\alpha_{os}$ in red varies much less and fluctuates in a region between 15% and 30%, as it takes into account previous significance level selection. We therefore suggest to implement the adaptive smoothing of the significance level selection in the final version of the pretest estimator, which optimally balances Type I and Type II error of the underlying test and at the same time takes into account previous $\alpha^*$ choices. Thus, the data-driven way of adaptively smoothed significance level combined with the pretest estimator based on the feasible test-statistic provides a decision rule for a portfolio strategy for the next period, which is optimal with respect to the portfolio performance measure chosen. However, like any other pretesting strategy, which is used for model selection, our pretesting strategy of selecting a single optimal portfolio strategy for investing in the next period within a rolling window set-up implies sharp thresholding which may lead to large turnover costs, if the pretest estimator frequently switches between different strategies. It shares this undesirable property with other sharp-thresholding portfolio allocation strategies based on $\ell_1$-norm shrinkage (lassoing).

2.3 Optimal Portfolio Allocation with Statistical Learning

As a solution to the bumpiness resulting from the sharp-thresholding of the pretest strategy we propose bagging to improve the pretest estimator. In order to do so, the in-sample estimation window is bootstrapped $B$ times smoothing the strategy selection functions of the pretest estimator by the bootstrapped probabilities. In the following we discuss the pretest estimator for the pseudo-out-of-sample test statistic in (11), as it takes into account forecasting risk and transaction costs. The algorithm for the bagged in-sample pretesting can be constructed along the same lines. The computation of the bagged out-of-sample pretest estimator when choosing
between strategies $s_i$ and the benchmark strategy $\bar{s}$ can be summarized in a following algorithm:\(^6\)

\begin{center}
\textit{Bagging the pretest estimator}
\end{center}

1. At the period $t$ define the estimation window of length $T$: $\{r_{t-T}, \ldots, r_t\}$.

2. Divide the sample into two parts and compute the pseudo-out-of-sample CE net of the transaction costs according to (10) and the out-of-sample test statistic according to (11) for each pair of strategies $(s_i, \bar{s})$, $i = 1, \ldots, M$.

3. Compute for a grid of $J$ $\alpha$ values the out-of-sample pretest weight estimates according to (12) and select $\alpha_{os,t+1}^*\alpha$ which results in the largest out-of-sample CE on the grid.

4. For a grid of $\lambda$ values compute the smoothed $\alpha_{os,t+1}^*$ using the previous optimal significance level and select the one maximizing the out-of-sample CE of the pretest estimator according to (15).

For every bootstrap iteration $b = 1, \ldots, B$ repeat Steps (5) and (6):

5. Randomly sample the rows of the in-sample $T \times N$ data with replacement and repeat step (2) for the out-of-sample test statistic computation keeping the weight estimates fixed.

6. For a significance level selected in Step 4 compute the strategy selection indicators $\mathbbm{1}_b^{os}(s_i, \alpha_{os,t+1}^*)$ for every strategy $i$ and bootstrap iteration $b$.

7. The bagged probability of the strategy $s_i$ is then defined as $\hat{p}(s_i, \alpha_{os,t+1}^*) = \frac{1}{B} \sum_{b=1}^{B} \mathbbm{1}_b^{os}(s_i, \alpha_{os,t+1}^*)$ using (13).

8. Finally the bagged out-of-sample pretest weight estimator for the period $t + 1$ is defined as an average of the weights estimated on the whole sample weighted by the bootstrap probabilities:

$$\hat{\omega}_{os,t}(S, \alpha_{os,t+1}^*) = \sum_{i=1}^{M+1} \hat{p}(s_i, \alpha_{os,t+1}^*) \hat{\omega}_t(s_i).$$ \hspace{0.5cm} (16)

The proposed out-of-sample bagged pretest estimator is novel by providing the investor with a fully data driven way of an optimal portfolio allocation strategy with respect to a specific performance measure. For instance, if the investor is looking for a strategy with the highest out-of-sample Sharpe ratio, the proposed bagging algorithm can be easily adapted. Moreover, the algorithm takes into account the amount of transaction costs and is time adaptive through the significance level selection.

The bagged pretest estimator defined in (16) benefits from the two-fold averaging. In the

\(^6\)Note, that monthly returns do not posses any significant SACF or SPACF patterns, therefore the i.i.d. bootstrap of Efron (1992) is appropriate to use. For the returns of higher frequencies one should use the circular block bootstrap by Politis and Romano (1992).
first step, the selected portfolio weight estimator is a weighted average combination of the underlying portfolio strategies $\hat{\omega}_t(s_i)$. Schanbacher (2015) shows that for a concave portfolio performance measure,\(^7\) the performance of combined portfolio strategies is not worse than a combined performance of the individual strategies, i.e.

$$CE\left(\sum_{i=1}^{M+1} \hat{p}(s_i, \alpha_{os,t+1}^s) \hat{\omega}_t(s_i)\right) \geq \sum_{i=1}^{M+1} \hat{p}(s_i, \alpha_{os,t+1}^s) CE(\hat{\omega}_t(s_i)),$$

as long as $\hat{p}(s_i, \alpha_{os,t+1}^s) \in (0, 1)$, which naturally holds in our setting by construction. The second type of averaging enters the proposed pretest estimator in the bagged probabilities of rejecting the null.\(^8\) A single test decision is likely to be influenced by pronounced fluctuations in financial markets and relatively short estimation windows, i.e. financial risk and estimation uncertainty. Bootstrapping each test decision provides the investor with an insight of how unstable is the current market situation. For example, for a calm period, the bagged probability would assign a large weight to a strategy selected by the sharp pretest estimator. Whereas for the crisis period, the bootstrapped strategy choice will be more diverse and $\hat{p}(s_i, \alpha_{os,t+1}^s)$ would be further away from the $(0, 1)$ bounds, see an example in Appendix B.

Moreover, bagging is a powerful way of variance reduction for the unstable estimators, e.g. it is widely used for stabilizing classification and regression trees which are based on sharp thresholding. Bühlmann and Yu (2002) show that bagging reduces the variance of the pretest estimator for the sample mean and the same arguments applies to the strategy selection function. For a strategy $s_i$ let $F_{s_i}(\cdot)$ denote the cumulative distribution function of the out-of-sample test statistic from (11). The mean and the variance of the pretest indicator function are given by

$$E[\mathbb{1}(t_{os}(s_i, \bar{s}) > q_\alpha)] = 1 - E[\mathbb{1}(t_{os}(s_i, \bar{s}) \leq q_\alpha)] = 1 - F_{s_i}(q_\alpha),$$

$$V[\mathbb{1}(t_{os}(s_i, \bar{s}) > q_\alpha)] = (1 - F_{s_i}(q_\alpha)) F_{s_i}(q_\alpha),$$

where $q_\alpha$ denotes the critical value corresponding to the significance level $\alpha$, i.e. for a fixed significance level $\alpha = 50\%$ the threshold is exactly zero and the variance of the pretest indicator is equal to $1/4$ for a symmetric $F_{s_i}$. Note that the pretest estimator is unstable in the sense of

\(^7\)For the discussion on concavity of different performance measures see Schanbacher (2015). Certainty Equivalent is a strictly concave measure as long as different strategies generate different variances.

\(^8\)In machine learning literature such estimators are often referred to as ensemble methods.
Breiman (1996), i.e. it assumes values 0 or 1 with positive probability, even if the sample size $T \to \infty$. Following Corollary 2.1 of Bühlmann and Yu (2002) variance of the bagged pretest indicator for $\alpha = 50\%$ is given by

$$V \left[ \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}_{\alpha}(s_i, 50\%) \right] = V \left[ F_{s_i}(t_{as}(s_i, \tilde{s})) \right] \to V [U],$$

where $U$ denotes a $[0, 1]$-uniformly distributed random variable and its variance equals $1/12$, which is three times smaller than the variance of the single pretest indicator function. The variance of the indicators in the portfolio context translates into the amount of portfolio rebalancing imposed by pretesting: a smaller variance of the strategy selection functions implies less turnover. In other words using bagging helps to stabilize the problem and reduce transaction costs, as the bootstrapped probabilities smooth the transition between the strategies along the rolling window.

Figure 3: Investment Decisions over Time: Pretest vs. Bagging

Contributions of the three different underlying strategies for the pretest estimator discussed in Section 2.1 under sharp thresholding and for the bagged probability weights: first panel - GMVP as in (2), second panel - tangency portfolio as in (3) and third panel - the equally weighted benchmark $\tilde{s}$ as in (4). Black dots depict the sharp pretest test decision and red lines depict the bagged probabilities, i.e. the smoothed version of the pretest decision. Analysis is based on the monthly return data from the example in Section 2.1, with $N = 10$, $T = 120$, $\gamma = 1$ and out-of-sample CE pretesting from (11) combined with adaptively smoothed significance level (15).

Figure 3 plots the bagged probabilities (red lines) together with a non-smoothed pretest.
strategy selection functions (black dots) assigned to the competing strategies from the example in Section 2.1 over time. Pretesting is based on the out-of-sample CE from (11) combined with adaptively smoothed significance level (15). Three panels on the figure correspond to the competing portfolio strategies with the equally weighted portfolio taken as a benchmark \( \tilde{s} \). Black dots of the pretest selection functions tend to switch between different strategies rather frequently and the bagged probabilities smooth the transition between portfolio strategies over time. The proposed algorithm is time adaptive, e.g. for February 2000 the pretest estimator suggests to change the portfolio decomposition reducing the weight of the tangency strategy in favor of the \( 1/N \), anticipating the dot-com crash in March 2000. Appendix B presents a detailed numerical example around this period to illustrate the adaptivity and flexibility of the out-of-sample bagged pretest estimator. The dot-com example sheds some light on the source of ensemble methods advantage. A very unstable tangency portfolio strategy in some periods might actually perform well and produce out-of-sample CE difference which is significantly better than the stable \( 1/N \) benchmark. The proposed dynamic strategy selection adjustment utilizes valuable information from CE-optimal but very unstable strategies in calm periods and returns to the stable conservative strategies in crisis times.

2.3.1 Sequential Performance Weighting Strategy

As a competitor for the bagged pretest estimator we consider sequential relative performance weighting inspired by the approach of Shan and Yang (2009) originally proposed for the combination of quantile forecasts. The idea behind this approach is very similar to boosting, which reweighs observations using the exponential loss function of the error associated with each observation. First, at each period \( t \) the relative performance of different strategies is measured by the exponential function of portfolio performance measure. Then the time-adaptive weight \( d_{t,i} \) for the strategy \( i \) is computed according to (17), where the initial values for the weights are fixed to \( d_{0,i} = \frac{1}{M+1} \). The resulting weight estimator is denoted as \( \hat{\omega}_t^{SP}(S) \):

\[
d_{t,i} = \frac{d_{t-1,i} \exp(CE_{in/os,t}(s_i))}{\sum_{j=1}^{M+1} d_{t-1,j} \exp(CE_{in/os,t}(s_j))},
\]

\[
\hat{\omega}_t^{SP}(S) = \sum_{j=1}^{M+1} d_{t,i} \hat{\omega}_t(s_i), \quad i = 1, ..., M + 1,
\]

22
where $S = \{s_1, ..., s_M, \tilde{s}\}$ is a set of the underlying strategies and $d_{t,i}$ are the time-adaptive coefficients used to weight the relative performance of the underlying strategies. This approach is not computationally demanding and is adapting the weights with every shift of the estimation window. It is also a smooth combination of the underlying strategies, potentially requiring less rebalancing and reducing transaction costs. We propose to apply this idea to the portfolio allocation context and compare it to the performance of the bagged pretest estimator. The main difference between the two methods lies in the construction of the weights each strategy receives. Sequential performance weighting combines all the available strategies relative to their performance, whereas pretesting assigns a positive weight to a given strategy only if it performs significantly better than the benchmark.
3 Empirical Evidence

3.1 CE and monthly data

We first provide empirical evidence on the performance of our pretested portfolio strategies by elaborating the example given in Section 2.1 and consider the GMVP, tangency and the equally weighted portfolios as the underlying competing strategies. We use data on monthly excess returns from K.R. French database\(^9\) of a 100 industry portfolios. In the analysis below for a given portfolio size \(N\) we randomly draw \(N\) out of 100 unique assets and report the average portfolio performance over 500 random draws. For the pseudo out-of-sample test statistic computation in (11) we split the sample into two equal parts, i.e. \(\kappa = \frac{T}{2}\). For the bagging step the number of bootstrap iterations is set to \(B = 200\), the out-of-sample evaluation window is fixed to \(H = 500\) observations and the risk aversion parameter is set to \(\gamma = 1\).\(^{10}\)

Table 1: Out-of-sample CE for \(T = 120\).

<table>
<thead>
<tr>
<th></th>
<th>N = 10</th>
<th>N = 20</th>
<th>N = 30</th>
<th>N = 40</th>
<th>N = 50</th>
<th>N = 60</th>
<th>N = 70</th>
<th>N = 80</th>
<th>N = 90</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMVP</td>
<td>8.6%</td>
<td>8.8%</td>
<td>8.7%</td>
<td>9.0%</td>
<td>8.6%</td>
<td>8.0%</td>
<td>6.8%</td>
<td>5.6%</td>
<td>2.8%</td>
</tr>
<tr>
<td>Tangency</td>
<td>-49.0%</td>
<td>-100.0%</td>
<td>-100.0%</td>
<td>-83.7%</td>
<td>-100.0%</td>
<td>-100.0%</td>
<td>-92.3%</td>
<td>-∞</td>
<td>-93.8%</td>
</tr>
<tr>
<td>1/N</td>
<td>8.5%</td>
<td>8.6%</td>
<td>8.6%</td>
<td>8.6%</td>
<td>8.6%</td>
<td>8.6%</td>
<td>8.6%</td>
<td>8.6%</td>
<td>8.6%</td>
</tr>
<tr>
<td>(\alpha = 5%)</td>
<td><strong>9.1%</strong></td>
<td>9.2%</td>
<td>5.0%</td>
<td>8.9%</td>
<td>8.8%</td>
<td>8.3%</td>
<td>-9.8%</td>
<td>-14.6%</td>
<td>1.6%</td>
</tr>
<tr>
<td>(\alpha^t_{in})</td>
<td>8.7%</td>
<td>8.7%</td>
<td>-10.0%</td>
<td>5.4%</td>
<td>-100.0%</td>
<td>-96.4%</td>
<td>-11.7%</td>
<td>-98.9%</td>
<td>0.1%</td>
</tr>
<tr>
<td>(\alpha^t_{in} B)</td>
<td>9.0%</td>
<td>9.1%</td>
<td>9.1%</td>
<td>7.6%</td>
<td>9.0%</td>
<td>9.0%</td>
<td>8.8%</td>
<td>8.7%</td>
<td>-2.0%</td>
</tr>
<tr>
<td>(\alpha^t_{os})</td>
<td>5.1%</td>
<td>-81.0%</td>
<td>-20.5%</td>
<td>-0.3%</td>
<td>-44.8%</td>
<td>-86.8%</td>
<td>-13.9%</td>
<td>-99.9%</td>
<td>-13.5%</td>
</tr>
<tr>
<td>(\alpha^t_{os} B)</td>
<td><strong>9.1%</strong></td>
<td><strong>9.3%</strong></td>
<td><strong>9.3%</strong></td>
<td><strong>9.4%</strong></td>
<td><strong>9.3%</strong></td>
<td><strong>9.3%</strong></td>
<td><strong>9.3%</strong></td>
<td><strong>9.3%</strong></td>
<td><strong>9.3%</strong></td>
</tr>
<tr>
<td>Seq. Perf.</td>
<td><strong>9.1%</strong></td>
<td><strong>9.3%</strong></td>
<td><strong>9.3%</strong></td>
<td><strong>9.4%</strong></td>
<td><strong>9.4%</strong></td>
<td><strong>9.4%</strong></td>
<td><strong>9.2%</strong></td>
<td><strong>8.9%</strong></td>
<td><strong>8.2%</strong></td>
</tr>
</tbody>
</table>

Numbers in the table correspond to the annualized average out-of-sample CE computed on net portfolio returns over a 500 randomly drawn portfolios of size \(N\). For each randomly drawn portfolio the out-of-sample CE is computed over an evaluation horizon of \(H = 500\) observations, risk aversion parameter \(\gamma = 1\), in-sample estimation window length \(T = 120\). \(\alpha = 5\%\) denotes the pretest estimator based on the in-sample test statistic as in (9) with a fixed significance level of 5\%\(.\alpha^t_{in}\) allows for a flexible significance level selection according to (15) and \(\alpha^t_{in} B\) is the bagged version of \(\alpha^t_{in}\). \(\alpha^t_{os}\) and \(\alpha^t_{os} B\) denote the out-of-sample significance level selection and the one combined with bagging as in (16). Seq. Perf. corresponds to (18). Numbers in bold correspond to the largest CE for a given portfolio size \(N\). \(-\infty\) used for the annualized loss of tangency portfolio for \(N = 80\) as the number is too big, please refer to the lower right panel of Figure 1 for illustration.

Table 1 reports the average out-of-sample certainty equivalent net of the transaction costs for the underlying strategies (first block), pretest estimators (second block) and sequential performance weighting (last row) for the in-sample estimation window length of 10 years (120 monthly observations) with different portfolio sizes \(N\) in the corresponding columns. As depicted\(^9\) the data is taken from K.R. French website and contains monthly excess returns from 05/1964 till 12/2015.\(^{10}\) As a robustness check Appendix C.2 reports the results for a range of out-of-sample evaluation horizons \(H = 60\) and risk aversion parameters \(\gamma\). The proposed out-of-sample pretest estimator combined with bagging shows a stable performance across different designs and portfolio sizes, independently of how bad is the performance of the underlying strategies.

---

\(^9\)The data is taken from K.R. French website and contains monthly excess returns from 05/1964 till 12/2015.

\(^{10}\)As a robustness check Appendix C.2 reports the results for a range of out-of-sample evaluation horizons \(H = 60\) and risk aversion parameters \(\gamma\). The proposed out-of-sample pretest estimator combined with bagging shows a stable performance across different designs and portfolio sizes, independently of how bad is the performance of the underlying strategies.
in Figure 1 the theoretical gains from increasing the portfolio dimension are accompanied by increasing losses due to estimation noise so that the out-of-sample CE net of the transaction costs of the GMVP declines with increasing $N$. Due to the extraordinary volatility of its portfolio weight estimates the tangency portfolio yields a negative CE’s despite shrinking of the covariance matrix and a (moderate) sample size of 120 monthly observations. The comparatively strong performance of the equally weighted portfolio is not too surprising as it is profiting from the absence of any estimation noise.

The fourth row of Table 1 contains the average out-of-sample CE’s for the basic pretest estimator based on the in-sample testing given by (9) with a fixed conventional significance level of 5%. Here the pretest estimator chooses the strategy which standardized in-sample CE difference from the $1/N$ is greater than the 95% quantile of the standard normal distribution. This pretest estimator performs well for smaller asset spaces. However for $N \geq 30$ its performance deteriorates and the CE of the basic pretest estimator is no longer larger than the values of the underlying strategies, potentially suffering from the estimation noise.

Allowing for a flexible selection of the significance level based on the in-sample CE according to (15) yields quite unstable CE estimates across $N$ and is generally performing poorly. A comparison with the estimates based on the larger estimation window (see Table 3 in the Appendix for estimates based on $T = 180$ ) reveals that pre-test estimation (with or without dynamic adjustment of the significance level) requires larger sample sizes to achieve a stable performance. In contrast to this, however, the combination of pretesting with bagging ($\alpha_{mB}^s$) outperforms the underlying strategies with the exception of $N = 90$.

The next row ($\alpha_{os}^s$) reports the average out-of-sample CE for the pretest estimator with adaptively smoothed out-of-sample significance level selection, which takes into account transaction costs, but suffers from the short evaluation period of $T/2$. The out-of-sample pretest estimator is also unstable across $N$ and does not outperform the underlying strategies. However it always outperforms the worst underlying strategy. Notably, the bagged out-of-sample pretest estimator ($\alpha_{os}^sB$) performs exceptionally well, producing the largest out-of-sample CE for all portfolio sizes and always exceeds the CE of best underlying strategy. In terms of the plain average out-of-sample CE comparisons reported in Table 1, the sequential performance weighting estimator is a reasonable competitor except for $N = 90$. 

25
Note, that our portfolios of different sizes are designed such that the asset spaces for the smaller portfolios are true subsets of the larger ones. This guarantees that the theoretical CE of a larger portfolio always has to dominate the CE of a smaller portfolio. For the empirical portfolios this dominance frequently does not hold even for regularized portfolio estimates, because the increase in estimation risk resulting from a larger portfolio dimension dominates theoretical gains of a better portfolio diversification. A typical example is the sequence of CE’s for the GMVP (first row of Table 1), where on average the CE of a portfolio of 90 assets is just a third of a CE of a portfolio formed of only 10 assets. In contrast, our bagged pretest strategy insures against the pitfalls of larger portfolio dimensions.

The plain out-of-sample performance comparisons of different portfolio strategies as reported in Table 1 only provide partial insights for an investor as they are short of providing information on the distributional properties of the estimated performance measures. For instance, two portfolio allocation strategies may perform equally well when they are solely compared on the basis of their average out-of-sample performance measures. However across different asset selections, they may differ in terms of their variance and in the probability of the occurrence of outliers. Similar to the original portfolio choice problem the investor faces a risk and return trade-off concerning the choice of the estimation strategy. Figure 4 depicts the boxplots of the out-of-sample CE’s of the considered strategies across the 500 random draws of portfolios of a given size $N$. The proposed bagged out-of-sample pretest estimator shows a strong performance in terms of their precision (indicated by their inter-quartile ranges) and the number of outliers.

The boxplot of the bagged out-of-sample pretest estimator denoted by $osB$ demonstrates to what extend bagging contributes to the stabilization of the portfolio allocation strategy and how, in particular, bagging cushions against the impact of (mainly strong negative) outliers which characterize the other strategies. For instance, for $N \geq 40$ the out-of-sample CE’s of the pretest estimators despite of having very similar medians are very different from each other in terms of the number of negative outliers. Comparing boxplots $in$ with $inB$ and $os$ with $osB$ shows that bagging always helps to stabilize the out-of-sample performance of the pretest estimators and works exceptionally well in the combination with the time-adaptive out-of-sample significance level selection. In particular, the out-of-sample CE of the proposed estimator $osB$ is almost as stable as the equally weighted portfolio and at the same time it outperforms the $1/N$ benchmark.
Figure 4: Boxplots of the Out-of-sample CE for $T = 120$.

Boxplots of the out-of-sample CE computed on net portfolio returns over a 500 randomly drawn portfolios of size $N$. For each randomly drawn portfolio the out-of-sample CE is computed over an evaluation horizon of $H = 500$ observations, risk aversion parameter $\gamma = 1$, in-sample estimation window length $T = 120$. X-axes denote different ways of computing portfolio weights: $G$ from (2), $TN$ from (3), $1/N$ from (4), $5\%$ denotes the pretest estimator based on the in-sample test statistic as in (9) with a fixed significance level of $5\%$, $in$ allows for a flexible significance level selection according to (15), $inB$ is the bagged version of $in$. $os$ and $osB$ denote the out-of-sample significance level selection and the same pretest estimator combined with bagging as in (16). $SP$ corresponds to (18).

in mean and has very few outliers in comparison with the sequential performance weighting from (18).

The reduction of the outliers immediately translates into the reduction of turnover costs. Therefore the source of the superior performance of the bagged pretest estimator mainly lies in the reduction in transaction cost. This is documented in Table 2, which informs about the the average turnover for each strategy relative to $1/N$.\textsuperscript{11} As expected, the turnover of the GMVP and the tangency portfolio increases with the increase in $N$. For larger portfolio dimension the amount of rebalancing required by the GMVP can be up to 30 times more than the rebalancing of $1/N$. Bagging the pretest estimators reduces turnover and the average turnover of the proposed bagged out-of-sample pretest estimator ($\alpha_{osB}$) is closest to the turnover of the equally weighted portfolio. The average turnover of sequential performance weighting turns out to be considerably larger and steeply increases with the portfolio size $N$.

\textsuperscript{11}Turnover of a strategy $s$ is computed as $TO(s) = \frac{1}{H} \sum_{t=1}^{H} \left( \sum_{j=1}^{N} |\tilde{\omega}_{j,t+1}(s) - \tilde{\omega}_{j,t}(s)| \right)$.
Our findings from above are generally confirmed when the analysis is based on the Sharpe ratio (SR) as the underlying performance measure. Figure 7 in Appendix depicts the boxplots of the Sharpe ratios across the 500 random portfolios of size $N$. Similarly to the out-of-sample CE, bagging stabilizes the SR and reduces the number of outliers. Furthermore, the results are robust to the in-sample estimation window length. Tables 3 and 4 report the average out-of-sample CE and turnover of the considered strategies for $T = 180$. In this case, the tangency portfolio and the GMVP perform well and the proposed bagged pretest estimator is performing as good as the best underlying strategy. Figures 9 and 8 report the boxplots of the out-of-sample CE and SR for different portfolio sizes $N$ and $T = 180$, where again, the most stable strategy with less outliers is the bagged out-of-sample pretest estimator, performing very well for all randomly drawn portfolios.

### 3.2 Selection of tuning parameters with daily data

In what follows we show that the proposed bagging algorithm can be used not only for optimal balancing of the different portfolio allocation strategies contributions, it can also be adapted for selecting an optimal tuning parameter for a given estimation approach. For instance if a researcher is interested in applying a GMVP strategy, but is choosing among the available estimators of covariance matrix, this algorithm offers a data-driven and performance measure-specific way of switching between the estimators, or shrinkage parameters, in a time-adaptive way.

For our second application we consider daily returns of S&P500 constituents from January 2014.
until the end of December 2014 ($T = 251$ days) for estimation and January 2015 - December 2016 ($H = 503$ days) for evaluation periods. We consider three different versions of the covariance matrix shrinkage for the GMVP allocation in (2): (1) the plug-in sample covariance estimator; (2) the shrinkage to market estimator by Ledoit and Wolf (2004b) with tuning parameter $\delta_1$ and (3) shrinkage proportional to the portfolio dimension $\delta_2 = \frac{0.05N}{1+0.05N}$, as was used in the previous application. Bagging is adjusted for the daily data, such that we now use circular block bootstrap (Politis and Romano, 1992). We consider Sharpe ratio as the portfolio performance measure, which is used for pretesting and sequential performance weighting.

Figure 5: Boxplots of the Out-of-sample SR for $T = 251$ and Daily Data.

Each boxplot corresponds to the distribution of the out-of-sample SR computed on net portfolio returns over a 500 randomly drawn portfolios of size $N$. For each randomly drawn portfolio the out-of-sample SR is computed over an evaluation horizon of $H = 503$ observations, in-sample estimation window length $T = 251$. X-axes denote different ways of computing portfolio weights: G from (2), $\delta_1$ for GMVP with Ledoit and Wolf (2004b) shrinkage, $\delta_2 = 0.05 \cdot N$ for Shrunken GMVP similar to (3), 5% denotes the pretest estimator based on the out-of-sample test statistic as in with a fixed significance level of 5%, op and opB denote the out-of-sample significance level selection and the same pretest estimator combined with bagging as in (16). SP corresponds to (18).

Figure 5 depicts the boxplots of the out-of-sample Sharpe ratios of the considered strategies across the 500 randomly drawn portfolio of size $N$. On every graph the first three boxplots correspond to the underlying versions of the GMVP, where $\delta_2$-type of covariance matrix shrinkage converges to the equally weighted allocation with an increase in $N$. Not surprisingly for the noisy daily data the larger the asset space, the more pronounced is the dominance of the $GMVP(\delta_2)$, e.g. for $N = 90$ the intense shrinkage has the highest Sharpe ratio among the
underlying strategies for all 500 randomly drawn portfolios. Therefore, all of the proposed estimators consistently select the \( GMVP(\delta_2) \) strategy in every period for every portfolio. This feature of the portfolio choice algorithm guarantees that whenever there is a clear dominating strategy it will always be selected. On the contrary, for the smaller portfolio dimensions of up to \( N = 30 \) assets, pretest-based estimator and sequential performance weighting have some room for improvement over the underlying strategies. In this case the bagged strategy is as stable as the intense shrinkage and has less negative outliers than the sequential performance weighting. Tables 5 and 6 in the Appendix report the average out-of-sample Sharpe ratio and average turnover for the considered strategies. It follows from the reported results that the bagged pretest estimator has the lowest turnover and is therefore suggested to be the preferred way of automatizing portfolio allocation. Furthermore, whenever the underlying strategies are heterogeneous enough, in the sense that there are less profitable & less risky allocations versus more profitable & more risky strategies, combining them in the proposed data driven way yields better performance compared to the underlying strategies.
4 Conclusions

In this paper we propose a novel method to robustify empirical portfolios. The core idea is to combine existing portfolio strategies by means of machine learning using the information from out-of-sample pretesting in a data-driven and time adaptive way. Hereby our approach accounts for the varying quality of alternative portfolio strategies in terms of their performance and their estimation risk across time. The basic pretest strategy is improved by a bagging step which yields out-of-sample selection probabilities as weights for the underlying portfolio strategies. By combining out-of-sample pretesting with idea of of model averaging our portfolio strategy provides double insurance against prediction risks.

In a first application based on monthly returns data the pretest estimation approach shows a strong performance particularly when the high dimension of the asset space is large relative to the sample size. We show that the extraordinary performance of the estimator lies in the reduction of outliers and hence the reduction of turnover costs. For a second application on daily returns data we show that the approach can also be beneficial to select the tuning parameter within a given class portfolio estimators.

In the first place, we prefer our machine learning approach not to be regarded as another competitor in the plethora of alternative portfolio strategies. We rather see it as a tool to cushion the consequences of a disadvantageous choice of a stand-alone portfolio strategy. Despite its promising performance we still see chances for further improvements. For example, one path of further improvements may be the development of more powerful performance tests in the pretest stage. Thus far, the choice of the sampling window and the split of the sampling window into an estimation window and a pseudo-out-of-sample testing window are taken as given. A more refined approach could also optimize over these two tuning parameters. Finally, the novel boosting-based approach also introduced in this study, which showed a sub-optimal but still a comparatively solid performance, deserves additional refinements. For instance, it may be worth investigating whether combining the two statistical learners can yield a further improvement in portfolio allocation.
References


Tu, J. and G. Zhou (2011): “Markowitz Meets Talmud: A Combination of Sophisticated and


A Appendix

The significance of the CE difference is tested by the means of the Student t-test. The standard error of the CE difference is computed using the Delta method (the bootstrap option is also available, but is more computationally costly). The Delta method for testing the difference in CE of the two strategies \((s, \tilde{s})\) is given by:

\[
\Delta_{os} = \phi(\vartheta) = \left( \mu_{os}(s) - \gamma \sigma_{os}^2(s) \right) - \left( \mu_{os}(\tilde{s}) - \gamma \sigma_{os}^2(\tilde{s}) \right)
\]

\[
\sqrt{H} \left( \Delta_{os}(s, \tilde{s}) - \Delta_{os}(s, \tilde{s}) \right) \xrightarrow{d} N \left( 0, \frac{\partial \phi(\vartheta)'}{\partial \vartheta} V[\hat{\vartheta}] \frac{\partial \phi(\vartheta)}{\partial \vartheta} \right),
\]

where \(\vartheta = (\mu_{os}(s), \sigma_{os}^2(s), \mu_{os}(\tilde{s}), \sigma_{os}^2(\tilde{s}))'\) 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. (2009)):

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

where \(\hat{\sigma}_{os}(s, \tilde{s})\) 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 \(\hat{y} = (\hat{r}_p(s) - \bar{\hat{r}}_p(s), r_p(\tilde{s}) - \bar{r}_p(\tilde{s}), \hat{r}_p^2(s) - \bar{r}_p^2(s), r_p^2(\tilde{s}) - \bar{r}_p^2(\tilde{s}))\), where \(\bar{\hat{r}}(\cdot)\) denotes the sample average. The standard error of the out-of-sample CE difference is then computed as

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

For testing the difference in Sharpe Ratios we use the same principle as in Ledoit and Wolf (2008), where the gradient for the standard error of the SR difference has to be adjusted
accordingly:

$$S.E. = \sqrt{\frac{\hat{\nabla}'\hat{\Psi}\hat{\nabla}}{H}}$$ with $\hat{\Psi} = \frac{1}{H} \hat{y}' \hat{y}$ and

$$\hat{\nabla} = \left[\frac{\hat{r}_p^2(s)}{(\hat{r}_p^2(s) - \bar{\hat{r}}_p^2(s))^{3/2}}, \frac{\hat{r}_p^2(\tilde{s})}{(\hat{r}_p^2(\tilde{s}) - \bar{\hat{r}}_p^2(\tilde{s}))^{3/2}}, -\frac{1}{2} \frac{\bar{\hat{r}}_p(s)}{(\hat{r}_p^2(s) - \bar{\hat{r}}_p^2(s))^{3/2}}, -\frac{1}{2} \frac{\bar{\hat{r}}_p(\tilde{s})}{(\hat{r}_p^2(\tilde{s}) - \bar{\hat{r}}_p^2(\tilde{s}))^{3/2}}\right]'.

## B Dot-com bubble

To get more insight into Figure 3 consider the beginning of the year 2000. March 2000 is typically associated with a peak of a Dot-com stock market bubble and as a numerical example we report the mechanics behind the bagging pretest estimator just before March 2000.

Figure 6: Median return: Dot-com bubble.

Black line on the plot represents the median return of $N = 10$ assets over the in-sample estimation window of size $T = 120$ and out-of-sample period from February 2000 (red line). Analysis is based on the monthly return data from the example in Section 2.1.

In this example we consider the same underlying strategies (GMVP, tangency portfolio with shrunken covariance matrix and 1/N) as in Section 2.1 and the out-of-sample significance level choice with bagging from Section 2.3. In order to choose a strategy for February 2000 the estimator looks at $T = 120$ previous returns and computes the pseudo out-of-sample CE over
the last 60 observations:

\[ \hat{CE}_{os}(g) = 0.0139 \quad \hat{CE}_{os}(tn) = 0.0173 \quad \hat{CE}_{os}(e) = 0.0118. \]

The standardized difference, pseudo out-of-sample test statistic, for each strategy against equally-weighted portfolio is given by:

\[ t_{os}(g,e) = 0.2764 \quad t_{os}(tn,e) = 0.4274. \]

The optimal significance level, which takes into account the previous significance levels, \( \alpha_{os} = 0.35 \), which implies a low threshold level of \( q_{0.35} = 0.385 \), such that the out-of-sample pretest estimator suggest the tangency portfolio strategy for investment in February 2000. Repeating the testing procedure with bootstrap results in

\[ \hat{p}(g, 0.35) = 0.03 \quad \hat{p}(tn, 0.35) = 0.92 \quad \hat{p}(e, 0.35) = 0.05, \]

where the bagged probabilities are computed according to step 7 of the algorithm described in Section 2.3. For February 2000 the decision of bagging coincides with the sharp thresholding and suggests investing the majority of the portfolio according to the tangency portfolio weights.

In the next period we shift the estimation period to February 2000 and observe some abnormal return values as can be seen on Figure 6. The resulting out-of-sample CEs and test statistics are:

\[ \hat{CE}_{os}(g) = 0.0156 \quad \hat{CE}_{os}(tn) = 0.0151 \quad \hat{CE}_{os}(e) = 0.0158, \]

\[ t_{os}(g, e) = -0.191 \quad t_{os}(tn, e) = -0.0484. \]

The grid search among all significance levels, taking into account the previous level of 35%, results in a smallest significance level of 1%. The out-of-sample pretest estimator thus decides in favour of the equally weighted portfolio. However, bootstrapping the test decision gives

\[ \hat{p}(g, 0.01) = 0.01 \quad \hat{p}(tn, 0.01) = 0.3 \quad \hat{p}(e, 0.01) = 0.69, \]

where the bagged pretest estimator implies that in 30% of the cases the tangency portfolio
strategy might be the best one. The performance of the algorithm crucially depends on the significance level choice and it is therefore very important to make $\alpha$ data-driven. The example above illustrates the flexibility of the proposed bagged pretest estimator, as it takes into account outliers, historical performance of strategies and the desired portfolio performance measure.

C Monthly Data

C.1 Long evaluation horizon: $H = 500$

Figure 7: Boxplots of the Out-of-sample SR for $T = 120$.

Boxplots of the out-of-sample SR computed on net portfolio returns over a 500 randomly drawn portfolios of size $N$. For each randomly drawn portfolio the out-of-sample CE is computed over an out-of-sample evaluation horizon of $H = 500$ observations, risk aversion parameter is $\gamma = 1$, in-sample estimation window length $T = 120$. X-axes denote different ways of computing portfolio weights: $G$ from (2), $TN$ from (3), $1/N$ from (4), $5\%$ denotes the pretest estimator based on the in-sample test statistic as in (9) with a fixed significance level of $5\%$, $in$ allows for a flexible significance level choice according to (15), $inB$ is the bagged version of $in$, $os$ and $osB$ denote the out-of-sample significance level choice and the same pretest estimator combined with bagging as in (16). $SP$ corresponds to (18).
Figure 8: Boxplots of the Out-of-sample SR for $T = 180$.

Boxplots of the out-of-sample SR computed on net portfolio returns over a 500 randomly drawn portfolios of size $N$. For each randomly drawn portfolio the out-of-sample CE is computed over an out-of-sample evaluation horizon of $H = 500$ observations, risk aversion parameter is $\gamma = 1$, in-sample estimation window length $T = 180$. X-axes denote different ways of computing portfolio weights: G from (2), TN from (3), 1/N from (4), 5% denotes the pretest estimator based on the in-sample test statistic as in (9) with a fixed significance level of 5%, in allows for a flexible significance level choice according to (15), inB is the bagged version of in. os and osB denote the out-of-sample significance level choice and the same pretest estimator combined with bagging as in (16). SP corresponds to (18).

Table 3: Out-of-sample CE for $T = 180$.

<table>
<thead>
<tr>
<th></th>
<th>N = 10</th>
<th>N = 20</th>
<th>N = 30</th>
<th>N = 40</th>
<th>N = 50</th>
<th>N = 60</th>
<th>N = 70</th>
<th>N = 80</th>
<th>N = 90</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMVP</td>
<td>8.6%</td>
<td>8.5%</td>
<td>8.5%</td>
<td>8.3%</td>
<td>8.2%</td>
<td>7.9%</td>
<td>7.5%</td>
<td>7.0%</td>
<td>6.5%</td>
</tr>
<tr>
<td>Tangency</td>
<td>7.4%</td>
<td>9.4%</td>
<td>9.4%</td>
<td>9.5%</td>
<td>9.5%</td>
<td>9.5%</td>
<td>9.5%</td>
<td>9.4%</td>
<td>9.5%</td>
</tr>
<tr>
<td>1/N</td>
<td>8.5%</td>
<td>8.6%</td>
<td>8.6%</td>
<td>8.6%</td>
<td>8.6%</td>
<td>8.6%</td>
<td>8.6%</td>
<td>8.6%</td>
<td>8.6%</td>
</tr>
<tr>
<td>$\alpha = 5%$</td>
<td>9.3%</td>
<td>9.4%</td>
<td>9.4%</td>
<td>9.5%</td>
<td>9.5%</td>
<td>9.5%</td>
<td>9.5%</td>
<td>9.4%</td>
<td>9.5%</td>
</tr>
<tr>
<td>$\alpha_{in}$</td>
<td>9.3%</td>
<td>9.4%</td>
<td>9.4%</td>
<td>9.5%</td>
<td>9.5%</td>
<td>9.5%</td>
<td>9.5%</td>
<td>9.4%</td>
<td>9.5%</td>
</tr>
<tr>
<td>$\alpha_{in}^B$</td>
<td>9.2%</td>
<td>9.4%</td>
<td>9.4%</td>
<td>9.4%</td>
<td>9.4%</td>
<td>9.4%</td>
<td>9.4%</td>
<td>9.4%</td>
<td>9.4%</td>
</tr>
<tr>
<td>$\alpha_{os}$</td>
<td>8.4%</td>
<td>8.9%</td>
<td>9.1%</td>
<td>9.4%</td>
<td>9.7%</td>
<td>9.8%</td>
<td>9.8%</td>
<td>9.6%</td>
<td>9.0%</td>
</tr>
<tr>
<td>$\alpha_{os}^B$</td>
<td>9.1%</td>
<td>9.3%</td>
<td>9.3%</td>
<td>9.4%</td>
<td>9.4%</td>
<td>9.4%</td>
<td>9.4%</td>
<td>9.4%</td>
<td>9.4%</td>
</tr>
<tr>
<td>Seq. Perf.</td>
<td>9.1%</td>
<td>9.2%</td>
<td>9.3%</td>
<td>9.3%</td>
<td>9.3%</td>
<td>9.2%</td>
<td>9.2%</td>
<td>9.1%</td>
<td>9.1%</td>
</tr>
</tbody>
</table>

Figures in the table correspond to the annualized average out-of-sample CE computed on net portfolio returns over a 500 randomly drawn portfolios of size $N$. For each randomly drawn portfolio the out-of-sample CE is computed over an out-of-sample evaluation horizon of $H = 500$ observations, risk aversion parameter is $\gamma = 1$, in-sample estimation window length $T = 180$. $\alpha = 5\%$ denotes the pretest estimator based on the in-sample test statistic as in (9) with a fixed significance level of 5%, $\alpha_{in}$ allows for a flexible significance level choice according to (15) and $\alpha_{in}^B$ B is the bagged version of $\alpha_{in}$. $\alpha_{os}$ and $\alpha_{os}^B$ denote the out-of-sample significance level choice and the same pretest estimator combined with bagging as in (16). Seq. Perf. corresponds to (18).
Figure 9: Boxplots of the Out-of-sample CE for $T = 180$.

Boxplots of the out-of-sample CE computed on net portfolio returns over a 500 randomly drawn portfolios of size $N$. For each randomly drawn portfolio the out-of-sample CE is computed over an out-of-sample evaluation horizon of $H = 500$ observations, risk aversion parameter is $\gamma = 1$, in-sample estimation window length $T = 180$. X-axes denote different ways of computing portfolio weights: $G$ from (2), $TN$ from (3), $1/N$ from (4), $5\%$ denotes the pretest estimator based on the in-sample test statistic as in (9) with a fixed significance level of $5\%$, $in$ allows for a flexible significance level choice according to (15), $inB$ is the bagged version of $in$. $os$ and $osB$ denote the out-of-sample significance level choice and the same pretest estimator combined with bagging as in (16). $SP$ corresponds to (18).

Table 4: Turnover for $T = 180$.

<table>
<thead>
<tr>
<th></th>
<th>N = 10</th>
<th>N = 20</th>
<th>N = 30</th>
<th>N = 40</th>
<th>N = 50</th>
<th>N = 60</th>
<th>N = 70</th>
<th>N = 80</th>
<th>N = 90</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMVP</td>
<td>3.1</td>
<td>5.6</td>
<td>7.7</td>
<td>9.9</td>
<td>12.0</td>
<td>14.2</td>
<td>16.6</td>
<td>19.1</td>
<td>22.0</td>
</tr>
<tr>
<td>Tangency</td>
<td>2.4</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>$1/N$</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>$\alpha = 5%$</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>$\alpha^{*}_{in}$</td>
<td>1.1</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>$\alpha^{*}_{in}$</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>$\alpha^{*}_{in}$</td>
<td>2.4</td>
<td>2.8</td>
<td>3.3</td>
<td>3.8</td>
<td>3.9</td>
<td>3.8</td>
<td>3.5</td>
<td>2.5</td>
<td>1.0</td>
</tr>
<tr>
<td>$\alpha^{*}_{in}$</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Seq. Perf.</td>
<td>1.1</td>
<td>1.4</td>
<td>1.8</td>
<td>2.1</td>
<td>2.5</td>
<td>2.8</td>
<td>3.0</td>
<td>3.2</td>
<td>3.3</td>
</tr>
</tbody>
</table>

Figures in the table correspond to the average turnover over a 500 randomly drawn portfolios of size $N$ divided by the turnover of $1/N$ for every column. For each randomly drawn portfolio the average turnover is computed over an out-of-sample evaluation horizon of $H = 500$ observations, in-sample estimation window length $T = 180$. $\alpha = 5\%$ denotes the pretest estimator based on the in-sample test statistic as in (9) with a fixed significance level of $5\%$, $\alpha^{*}_{in}$ allows for a flexible significance level choice according to (15) and $\alpha^{*}_{in}$ $B$ is the bagged version of $\alpha^{*}_{in}$. $\alpha^{*}_{os}$ and $\alpha^{*}_{os}$ $B$ denote the out-of-sample significance level choice and the same pretest estimator combined with bagging as in (16). Seq. Perf. corresponds to (18).
C.2 Short evaluation horizon: $H = 60$

As a robustness check we examine the performance of the proposed out-of-sample bagging over a shorter evaluation period of $H = 60$ (5 years of monthly data) and different levels of risk aversion parameter $\gamma$. We consider two rolling window scenarios and use Certainty Equivalent as a target portfolio performance measure:


Figure 10: Out-of-sample CE for $T = 120$: crisis evaluation.

Lines on the plots represent the average out-of-sample CE computed on net portfolio returns according to (1) over 500 randomly drawn portfolios of size $N$ (x-axis). For each randomly drawn portfolio the out-of-sample CE is computed over an out-of-sample evaluation horizon of $H = 60$ observations. Different panels correspond to different risk aversion parameters $\gamma$. Panels depict the average out-of-sample CE with the weights of GMVP (in dotted black) from (2), tangency portfolio (in solid black) from (3) and shrinkage intensity $\delta = \frac{0.05 N}{1 + 0.05 N}$, and equally weighted portfolio (in dashed black) from (4) computed over an estimation window length of $T = 120$ observations. Solid red line corresponds to the pretest estimator with the out-of-sample significance level choice combined with bagging as in (16). Evaluation Horizon $H$ corresponds to a period Jan 2008 - Dec 2012.
Figure 11: Out-of-sample CE for $T = 120$: post-crisis evaluation.

Lines on the plots represent the average out-of-sample CE computed on net portfolio returns according to (1) over 500 randomly drawn portfolios of size $N$ (x-axis). For each randomly drawn portfolio the out-of-sample CE is computed over an out-of-sample evaluation horizon of $H = 60$ observations. Different panels correspond to different risk aversion parameters $\gamma$. Panels depict the average out-of-sample CE with the weights of GMVP (in dotted black) from (2), tangency portfolio (in solid black) from (3) and shrinkage intensity $\delta = \frac{0.05}{1 + 0.05N}$, and equally weighted portfolio (in dashed black) from (4) computed over an estimation window length of $T = 120$ observations. Solid red line corresponds to the pretest estimator with the out-of-sample significance level choice combined with bagging as in (16). Evaluation Horizon $H$ corresponds to a period Jan 2010 - Dec 2014.
## D Daily Data

### Table 5: Out-of-sample SR for $T = 251$ and Daily Data.

<table>
<thead>
<tr>
<th></th>
<th>N = 10</th>
<th>N = 20</th>
<th>N = 30</th>
<th>N = 40</th>
<th>N = 50</th>
<th>N = 60</th>
<th>N = 70</th>
<th>N = 80</th>
<th>N = 90</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMVP</td>
<td>0.0034</td>
<td>-0.0112</td>
<td>-0.0297</td>
<td>-0.0498</td>
<td>-0.0738</td>
<td>-0.1016</td>
<td>-0.1312</td>
<td>-0.1615</td>
<td>-0.1967</td>
</tr>
<tr>
<td>$\delta_1$</td>
<td>0.0066</td>
<td>-0.0019</td>
<td>-0.0111</td>
<td>-0.0209</td>
<td>-0.0322</td>
<td>-0.0440</td>
<td>-0.0576</td>
<td>-0.0673</td>
<td>-0.0810</td>
</tr>
<tr>
<td>$\delta_2$</td>
<td>0.0140</td>
<td>0.0149</td>
<td>0.0150</td>
<td>0.0157</td>
<td>0.0150</td>
<td>0.0152</td>
<td>0.0153</td>
<td>0.0152</td>
<td>0.0154</td>
</tr>
<tr>
<td>$\alpha = 5%$</td>
<td>0.0117</td>
<td>0.0142</td>
<td>0.0148</td>
<td>0.0156</td>
<td>0.0151</td>
<td>0.0152</td>
<td>0.0153</td>
<td>0.0152</td>
<td>0.0154</td>
</tr>
<tr>
<td>$\alpha_{os}$</td>
<td>-0.0015</td>
<td>0.0005</td>
<td>0.0032</td>
<td>0.0083</td>
<td>0.0108</td>
<td>0.0132</td>
<td>0.0149</td>
<td>0.0151</td>
<td>0.0154</td>
</tr>
<tr>
<td>$\alpha_{os}^{s}$</td>
<td>0.0140</td>
<td>0.0149</td>
<td>0.0151</td>
<td>0.0157</td>
<td>0.0151</td>
<td>0.0152</td>
<td>0.0153</td>
<td>0.0152</td>
<td>0.0154</td>
</tr>
<tr>
<td>Seq. Perf.</td>
<td>0.0092</td>
<td>0.0123</td>
<td>0.0142</td>
<td>0.0153</td>
<td>0.0147</td>
<td>0.0148</td>
<td>0.0148</td>
<td>0.0145</td>
<td>0.0145</td>
</tr>
</tbody>
</table>

Figures in the table correspond to the average out-of-sample SR computed on net portfolio returns over a 500 randomly drawn portfolios of size $N$. For each randomly drawn portfolio the out-of-sample SR is computed over an out-of-sample evaluation horizon of $H = 503$ observations, in-sample estimation window length $T = 251$. Table rows correspond to different ways of computing portfolio weights: $G$ from (2), $\delta_1$ for GMVP with Ledoit and Wolf (2004b) shrinkage, $\delta_2 = \frac{0.05N}{1+0.05N}$ for Shrunken GMVP similar to (3), 5% denotes the pretest estimator based on the out-of-sample test statistic as in with a fixed significance level of 5%, $\alpha$ and $\alpha B$ denote the out-of-sample significance level choice and the same pretest estimator combined with bagging as in (16). SP corresponds to (18).

### Table 6: Turnover for $T = 251$ and Daily Data.

<table>
<thead>
<tr>
<th></th>
<th>N = 10</th>
<th>N = 20</th>
<th>N = 30</th>
<th>N = 40</th>
<th>N = 50</th>
<th>N = 60</th>
<th>N = 70</th>
<th>N = 80</th>
<th>N = 90</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMVP</td>
<td>0.0341</td>
<td>0.0701</td>
<td>0.1112</td>
<td>0.1560</td>
<td>0.2042</td>
<td>0.2595</td>
<td>0.3173</td>
<td>0.3834</td>
<td>0.4546</td>
</tr>
<tr>
<td>$\delta_1$</td>
<td>0.0279</td>
<td>0.0527</td>
<td>0.0783</td>
<td>0.1041</td>
<td>0.1291</td>
<td>0.1557</td>
<td>0.1812</td>
<td>0.2053</td>
<td>0.2305</td>
</tr>
<tr>
<td>$\delta_2$</td>
<td>0.0085</td>
<td>0.0087</td>
<td>0.0087</td>
<td>0.0087</td>
<td>0.0088</td>
<td>0.0087</td>
<td>0.0087</td>
<td>0.0087</td>
<td>0.0087</td>
</tr>
<tr>
<td>$\alpha = 5%$</td>
<td>0.0123</td>
<td>0.0097</td>
<td>0.0090</td>
<td>0.0088</td>
<td>0.0088</td>
<td>0.0087</td>
<td>0.0087</td>
<td>0.0087</td>
<td>0.0087</td>
</tr>
<tr>
<td>$\alpha_{os}$</td>
<td>0.0383</td>
<td>0.0330</td>
<td>0.0280</td>
<td>0.0213</td>
<td>0.0162</td>
<td>0.0122</td>
<td>0.0095</td>
<td>0.0089</td>
<td>0.0087</td>
</tr>
<tr>
<td>$\alpha_{os}^{s}$</td>
<td>0.0085</td>
<td>0.0087</td>
<td>0.0087</td>
<td>0.0087</td>
<td>0.0088</td>
<td>0.0087</td>
<td>0.0087</td>
<td>0.0087</td>
<td>0.0087</td>
</tr>
<tr>
<td>Seq. Perf.</td>
<td>0.0149</td>
<td>0.0144</td>
<td>0.0117</td>
<td>0.0110</td>
<td>0.0109</td>
<td>0.0110</td>
<td>0.0112</td>
<td>0.0113</td>
<td>0.0114</td>
</tr>
</tbody>
</table>

Figures in the table correspond to the average turnover over a 500 randomly drawn portfolios of size $N$. For each randomly drawn portfolio the average turnover is computed over an out-of-sample evaluation horizon of $H = 503$ observations, in-sample estimation window length $T = 251$. Table rows correspond to different ways of computing portfolio weights: $G$ from (2), $\delta_1$ for GMVP with Ledoit and Wolf (2004b) shrinkage, $\delta_2 = \frac{0.05N}{1+0.05N}$ for Shrunken GMVP similar to (3), 5% denotes the pretest estimator based on the out-of-sample test statistic as in with a fixed significance level of 5%, $\alpha$ and $\alpha B$ denote the out-of-sample significance level choice and the same pretest estimator combined with bagging as in (16). SP corresponds to (18).