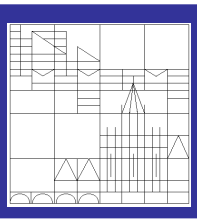
University of Konstanz
Department of Economics



Improved Portfolio Choice Using Second-Order Stochastic Dominance

James E. Hodder, Jens C. Jackwerth, and Olga Kolokolova

Working Paper Series
2010-14

http://www.wiwi.uni-konstanz.de/workingpaperseries

Improved Portfolio Choice Using Second-Order Stochastic Dominance

by

James E. Hodder
Jens Carsten Jackwerth
Olga Kolokolova

First draft: February 24, 2009
This version: November 11, 2010

James E. Hodder, University of Wisconsin-Madison, jhodder@bus.wisc.edu
Jens Jackwerth, University of Konstanz, jens.jackwerth@uni-konstanz.de
Olga Kolokolova, University of Manchester, olga.kolokolova@mbs.ac.uk

We would like to thank Olivier Scaillet, Karim Abadir, and seminar participants at Universidad Pompeu Fabra in Barcelona, seminar participants at the University of Frankfurt, and workshop participants in Königsfeld for helpful comments.

Improved Portfolio Choice Using Second-Order Stochastic Dominance

Abstract

We examine the use of second-order stochastic dominance as both a way to measure performance and also as a technique for constructing portfolios. Using in-sample data, we construct portfolios such that their second-order stochastic dominance over a typical pension fund benchmark is most probable. The empirical results based on 21 years of daily data suggest that this portfolio choice technique significantly outperforms the benchmark portfolio out-of-sample. As a preference-free technique it will also suit any risk-averse investor in e.g. a pension fund. Moreover, its out-of-sample performance across eight different measures is superior to widely discussed portfolio choice approaches such as equal weights, mean-variance, and minimum-variance methods.

Improved Portfolio Choice Using Second-Order Stochastic Dominance

1. Introduction

In this paper, we examine the use of second-order stochastic dominance as both a way to measure performance and also as a technique for constructing portfolios. An advantage of this approach is that it requires very modest assumptions about investor preferences. We shall see that using the concept of SSD in-sample allows constructing dominating portfolios also out-of-sample.

Large money managers such as pension funds currently use a variety of methods to estimate portfolio risk and performance. Typical risk measures include return standard deviation, return semi-variance, value at risk, and expected shortfall. Pure performance is often proxied by expected return, where details on risk and performance measures can be found in Levy (2006), Ch. 1. Risk-adjusted performance measures combine both risk and return using a single number. Widely-used measures include the Sharpe ratio, the Treynor ratio, and Jensen's alpha. Even with estimates of such measures in hand, there is the complex issue of ranking different return distributions. Fundamentally, that ranking should depend on investor preferences; and various assumptions have been used. Several popular approaches employ some variation of portfolio optimization within the Markowitz (1952) mean-variance framework.[1] However, the basic mean-variance criterion has well-known limitations. It is symmetric, and its theoretical justification requires either a quadratic utility function or multivariate normality of returns. It thus considers only the first two moments of the return distribution. Furthermore, the corresponding optimization procedures often result in extreme portfolio weights when using historical inputs, which contain estimation errors relative to the true underlying return distributions. And even the more sophisticated portfolio choice methods detailed in DeMiguel, Garlappi, and Uppal (2009) still require that some assumption on preferences which support a particular optimization criterion such as minimum variance, etc.

[1] Cumby and Glen (1990), for example, investigate whether US-only investors could benefit from international diversification. De Roon, Nijman, and Werker (2001) among others question whether including emerging-market securities can improve performance of portfolios otherwise invested in only developed markets. Glen and Jorion (1993) analyze whether the investors with a well-diversified international portfolio of stocks and bonds will benefit by adding currency futures to their portfolio. Han (2006) investigates the optimal portfolio allocation of a mean-variance investor with time-varying moments of return distributions. Martellini and Urošević (2006) analyze static mean-variance portfolio optimization problem with uncertain investment horizon.

This problem in ranking return distributions is particularly relevant for large pension funds such as the California Public Employees' Retirement Systems (CALPERS), the New York State Common Retirement Fund, or the State of Wisconsin Investment Board (SWIB). Such funds have large amounts of money under management that is intended to support the retirement benefits of very large numbers of individuals. Hence, these are major institutional investors that represent the interests of numerous individuals with presumably differing preferences. Frequently, a pension fund has fixed target portfolio holdings which are periodically reviewed and approved by its supervisory board. These target portfolio allocations are typically rather stable over time with occasional minor adjustments. However, there may be more frequent portfolio rebalancing to keep the portfolio weights reasonably close to the target as security prices move.

Most pension funds invest primarily in two asset classes: stocks and bonds. Some funds also diversify into real estate and other alternative investments. According to a 2008 survey of the 1000 largest pension funds in NN (2009), defined benefit funds invested 52% of their assets into stocks and 28% into bonds, around 6% in private equity, the same amount in real estate equity, 1.6% in cash, and the remaining 6.4% in various other assets.[2] A recent development has been the inclusion of hedge fund investments which are counted under "various other assets". We will later use such typical investment proportions to construct a benchmark portfolio.

We propose to rank portfolio return distributions based on second-order stochastic dominance (SSD) as a comparison criterion. If a return distribution "A" second-order stochastically dominates another distribution "B", then all risk-averse investors with increasing and concave utility function will prefer A to B. We argue that it is much more reasonable to assume all pension fund investors to be risk-averse rather than assuming that they all share identical and tightly parameterized preferences. SSD does exactly provide such tool: a dominating distribution will be preferred by all the potentially millions of risk-averse investors of a large pension fund without knowledge of their individual preferences.

Also, the SSD criterion does not focus on a limited number of moments but accounts for the complete return distribution, considering both gains and losses. The developed tests for SSD are nonparametric; and thus, no distributional assumptions are needed for their implementation. Last but not least, we find that portfolio optimization based on the SSD

[2] According to the same survey, defined contribution plans have rather similar investment objective with 47% in stock, 23% in bonds (interpreting the reported category "Stable Value" as fixed income investment), and somewhat larger portion of 10.5% in cash.

criterion results in fairly stabile portfolio weights, which overcomes a major problem for mean-variance optimization procedures.

SSD is a powerful tool for ranking distributions. It has been used, for example, to evaluate post merger stock performance (Abhyankar, Ho, and Zhao (2005)) and to analyze aggregated investors' preferences and beliefs (Post and Levy (2005)). Russell and Seo (1980) as well as De Giorgi (2005) apply this concept to a theoretical portfolio choice problem and discuss the properties of the SSD criterion compared to the mean-variance approach. They show, that the sets of mean-variance efficient portfolios and SSD efficient portfolios overlap but do not coincide. The concept of stochastic dominance was empirically applied to the portfolio choice problem by Post (2003), Kuosmanen (2004), and Kopa (2009). These authors test for stochastic dominance of a specified portfolio (the market portfolio) with respect to all other portfolios that can be constructed in a given asset span. The test procedure of Kopa (2009) additionally identifies an efficient portfolio that dominates the evaluated portfolio. Going one step further, Scaillet and Topaloglou (2005) augment the testing procedures of Post (2003) and Kuosmanen (2004) to allow for time varying return distributions and test for the SSD efficiency of the market portfolio. The main limitation of all these works is that they only analyze in-sample performance. For practical portfolio allocation problems, it is essential to establish the out-of-sample properties of SSD efficient portfolios.

Out-of-sample stochastic dominance analysis was conducted by Meyer, Li and Rose (2005). These authors consider the benefits of international portfolio diversification compared with a New Zealand-only portfolio. They use the concept of third-order stochastic dominance, arguing that second-order stochastic dominance tests lack power. Their in-sample portfolio choice, however, is still conducted using the mean-variance approach with a fixed target return.

Thus, existing empirical work on portfolio allocation using the SSD concept has been either restricted to in-sample analysis or did not rely on the SSD criterion for estimating portfolio weights themselves. In this paper, we extend the above work in several ways. We examine whether a typical pension-fund portfolio is SSD efficient or if that portfolio can be improved upon. In doing so, we consider the main asset classes in which major pension funds invest and form a corresponding benchmark portfolio. We then develop a procedure to determine the optimal in-sample portfolio based on the SSD criterion. Here, the optimal portfolio is constructed to have the highest value of a test statistic due to Davidson (2008), with further details provided below.

We then test whether this SSD-based portfolio dominate a benchmark portfolio out-of-sample. We compare the performance of our SSD-based portfolio with other competing portfolio choice approaches. The comparison alternatives include portfolios based on SSD-related risk-measures (minimum-variance, minimum-semi-variance, and minimum-shortfall), mean-variance-related portfolios (maximum Sharpe ratio, maximum Information ratio portfolio, and a portfolio with the minimum possible variance given the same in-sample mean return as the benchmark), and the equally weighted portfolio. DeMiguel, Garlappi, and Uppal (2009) found that this last equally weighted portfolio performed on par with a number of much more complicated alternative portfolio choice mechanisms. Thus, it is important to us to establish that our SSD-based portfolios also outperform the equally weighted portfolio. In the main run we evaluate performance of these portfolios with respect to a static benchmark portfolio typical to pension funds. In the robustness section, we also test these portfolios against each other and perform several other stability checks.

The analysis is conducted using non-overlapping windows. We develop a formal statistical test that allows us to document that our SSD-based portfolio choice technique significantly increases the propensity for selecting portfolios that dominate the benchmark out-of-sample. Thus, we propose an approach to improve the asset allocation of pension funds and other money managers without specifying a parameterized utility function. Such a technology can help to establish a lower bound on performance that any risk-averse investor would prefer (or at least be indifferent) when compared with a typical benchmark portfolio.

Further, the other SSD-related portfolios also dominate the benchmark while the equally weighted portfolio performs on par with the benchmark. The mean-variance-related portfolios tend to do worse than the benchmark. Our results are extremely robust to numerous checks on the benchmark, the methodology, other asset classes, and around market crises. Finally, we document in a simulation exercise that only the SSD-based method can handle realistic data which exhibits time varying distributions, estimation error, and non-normality while the competing methods are rather sensitive to deviations from ideal data, namely stationary, normally distributed returns.

In the following section, we introduce the methodology of constructing the SSD portfolio and the other competing portfolios. Section 3 introduces the data used and in section 4 in which we describe our empirical results. Section 5 covers a large number of robustness tests while Section 6 investigates with a simulation, which features of the data (time varying distributions, estimation error, and normality) matter for the performance of different portfolio choice mechanisms. Section 7 concludes.

2. Methodology

We first provide an overview of our methodological approach and then discuss the steps in more detail.

Consider a fixed benchmark-portfolio (Bench) of s assets which is held for a (yearly) time period from $t_0 - \Delta t$ to t_0. This benchmark can be viewed as a proxy for a typical portfolio allocation of a pension fund. For the same (in-sample) time period, the SSD-based portfolio (SSDBased) with the highest probability of second-order stochastically dominating Bench is constructed. This portfolio is designed so as to have the highest value of the test statistic of Davidson (2008), detailed below.

We also create several other competing portfolios using in-sample data and examine their out-of-sample performance. The first group of alternative portfolios is based on risk measures that are consistent with second-order stochastic dominance. This group, labeled SSD-related, includes the global minimum variance portfolio (MinVar), the global minimum semi-variance portfolio (MinSemivar) and the minimum shortfall portfolio (MinShortfall).

The second group of competing portfolios, labeled Mean-Variance-related, includes three mean-variance-type portfolios: a) the portfolio with the highest in-sample Sharpe ratio (MaxSharpe), b) the portfolio with the highest Information ratio (expected excess return over the benchmark divided by the standard deviation of this excess return) with respect to Bench (InformationRatio), and c) the minimum-variance portfolio which has the same mean return as Bench (MinVarBench). A practical problem with these portfolios is that they tend to have very unstable and sometimes extreme weights on individual securities due to the estimation error in the parameters, see e.g. Michaud (1989), Jorion (1992), as well as DeMiguel, Garlappi, and Uppal (2009). As a result, the Mean-variance-related methods normally exhibit poor out-of-sample performance. In response to this problem with weight instability, we include in our comparison group an equally weighted portfolio (Equal) which DeMiguel, Garlappi, and Uppal (2009) found to perform relatively well in their analysis.

The optimal weights for all these portfolios are determined using the in-sample data from $t_0 - \Delta t$ to t_0.

Next, using these in-sample-determined portfolio weights, the out-of-sample returns of all portfolios are computed for the period t_0 to $t_0+\Delta t$. The performance of the portfolios is compared with the benchmark's out-of-sample return to determine whether the portfolios dominate the benchmark in the SSD sense.

The analysis is repeated using $T=20$ non-overlapping windows. The former out-of-sample period becomes the new in-sample period for portfolio weights estimation, and SSD performance is then measured for the next out-of-sample period from $t_0+\Delta t$ to $t_0+2\Delta t$. This procedure results in 20 yearly out-of-sample periods. Finally, we test if the choice mechanism based on in-sample SSD optimization (as well as the other portfolio choice approaches considered) significantly outperforms the benchmark out-of-sample over the 20-year sequence of yearly periods.

To make sure that all our constructed portfolios are feasible choices for pension funds which could be precluded from shorting, we impose short sale constraints in the portfolio selection process. Thus, portfolio weights are restricted to be positive and to sum up to one for each of the considered portfolios.[3] All in-sample optimal portfolio weights are obtained using a grid search with steps of 0.02 for each weight. Thus, as we avoid any analytical or numerical optimization schemes, we do not need to make any parametric assumptions about return distributions and their correlation structure.[4] Such search is globally convergent and insures that we will find the maximum to within the 0.02 spacing, even in the presence of multiple local maxima.

The following sub-sections address the above steps in more detail.

2.1. Constructing portfolios using SSD

Graphically, second-order stochastic dominance (SSD) implies that two cumulative distribution functions cross but that the area under the dominating distribution is always smaller or equal to that of the dominated distribution for each threshold level z. If those cumulative distribution functions do not cross, first order stochastic dominance is observed. Figure 1 illustrates the SSD relation between two distributions A and B.

Formally, distribution A with cumulative distribution function $F_A(y)$ is said to second-order stochastically dominate another distribution B with cumulative distribution function $F_B(y)$ if, for all possible threshold levels z, the expected losses with respect to this

[3] As a robustness check, we wave the short selling restriction and allow the weights to take values from -1 to 1. This adversely influences the performance of the Mean-Variance-related portfolios; the SSD based portfolio still outperforms.

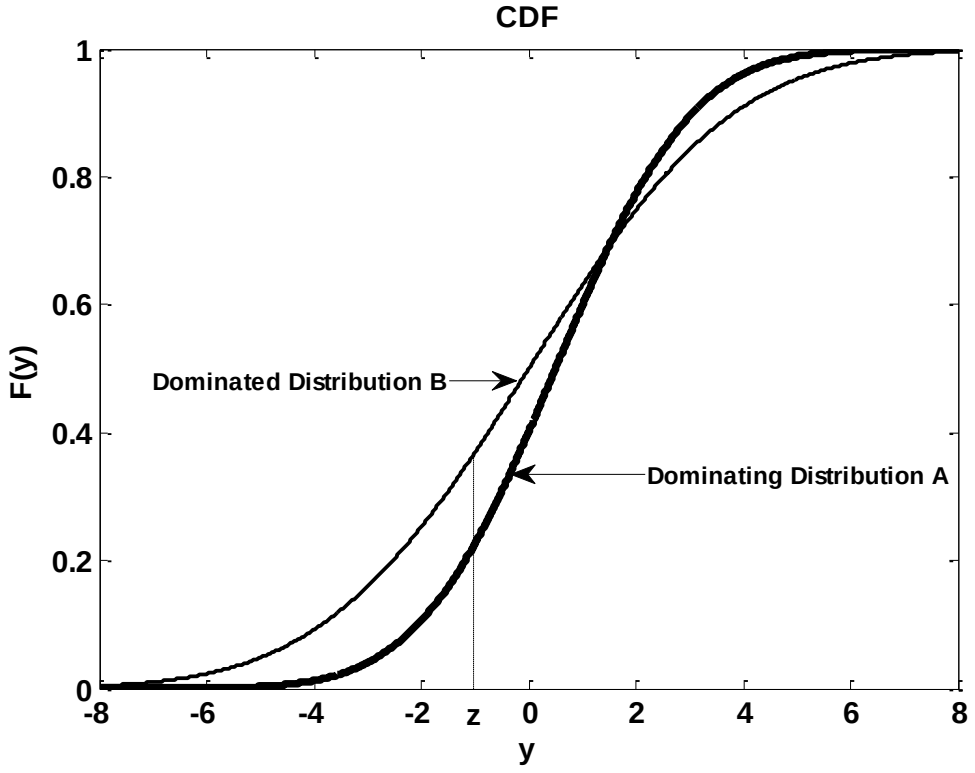
[4] Since we do not estimate distribution parameters for the (in-sample) security returns, portfolio choice approaches that utilize shrinkage techniques on the variance-covariance matrices or Bayesian priors are not directly applicable in this case. One could presumably estimate those parameters in order to implement such techniques; however DeMiguel, Garlappi, and Uppal (2009) found that several shrinkage techniques did not consistently out-perform equal weighting. Moreover, Bayesian techniques require specifying exogenous prior beliefs.

threshold in distribution A are not larger than that in distribution B with at least one strict inequality for some level of z.

$$\int_{-\infty}^{z} (z-y)dF_A(y) \leq \int_{-\infty}^{z} (z-y)dF_B(y), \ \forall z \in R \tag{1}$$

Figure 1. Example of an SSD Relation between Two Distributions

This figure plots two intersecting cumulative distribution functions characterized by the SSD relation. The area under the dominating distribution is always smaller than that of dominated distribution. On the horizontal axis, possible values y of the random variables are shown, with the vertical axis indicating values F(y) of the corresponding cumulative distribution functions.



2.1.1. Statistical tests for second-order stochastic dominance

Testing for stochastic dominance is not trivial; however, statistical tests for SSD have been developed and their properties demonstrated (see for example, Anderson (1996), Kaur, Prakasa Rao and Singh (1994), Davidson and Duclos (2000), Barrett and Donald (2003), Linton, Maasoumi and Whang (2003), Davidson (2008)). The main differences among these tests are the way the null hypothesis is formulated, the type of test statistic employed, the ability of the test to handle correlated samples, and the approach to computing p-values.

For the purpose of this paper, the most appealing test specification is the one of Davidson (2008). We rely on this test in establishing the SSD relation between different portfolio return distributions in our out-of-sample tests. We also use the test statistic of Davidson (2008) as a criterion function in constructing our SSD-based portfolio using in-sample data.

The Davidson (2008) test possesses a number of characteristics that make it superior to other SSD-test specifications. First of all, the test allows for correlated samples. This is an important limitation for most existing tests of stochastic dominance, which can deal only with uncorrelated samples. When comparing portfolios that consist of the same assets (but in different proportions), we have to consider correlated samples. Apart from Davidson (2008), the only test procedure of which we are aware that can explicitly handle correlated samples is that of Davidson and Duclos (2000).

The Davidson and Duclos (2000) test specification, however, compares distributions only at a fixed number of arbitrarily chosen points. This limitation can potentially lead to inconsistent results (see Davidson and Duclos (2000, p. 1446), as well as Barrett and Donald (2003, p. 72)). Consistency is assured only in those tests that use all available sample points, such as Kaur, Prakasa Rao, and Singh (1994) and Davidson (2008).

Additionally, the Davidson (2008) test starts with the null hypothesis of non-dominance for one distribution over another, whereas the majority of other SSD tests have as their null hypothesis dominance -- see, e.g., Anderson (1996), Davidson and Duclos (2000), plus Barrett and Donald (2003). Rejecting the null of dominance then does not imply dominance of the second distribution, since it can also happen that the test fails to rank these distributions. At the same time, rejecting the null of non-dominance delivers an unambiguous result of dominance. This formulation of the null hypothesis is also used by Kaur, Prakasa Rao and Singh (1994); however, their approach cannot cope with correlated samples.

The distribution of the Davidson (2008) test statistic under the null of non-dominance is asymptotically normal, but the p-values should be bootstrapped in small samples to assure better finite sample properties and higher power of the test.[5] We find that for 252 daily returns in one year, the asymptotic and the bootstrapped p-values nearly always correspond to

[5] Applying SSD tests to time series data, one needs to be concerned about test performance if there is time dependence in the data, such as autocorrelation in returns or GARCH effects in volatility. Unfortunately, no test so far explicitly accounts for such time-series effects. Nolte (2008) shows that the Davidson (2008) test loses power if the data are strongly serially correlated. As we will document below, serial correlation is not pronounced in the data used for the current study. Nolte also shows that the Davidson (2008) test performs well in the presence of GARCH effects. Thus, we feel comfortable using the Davidson (2008) approach.

the same significance level. The average difference in the p-values is 0.006. The largest absolute difference in the p-values for significant cases is 0.025, corresponding to a bootstrapped p-value of 0.032 compared with an asymptotic p-value of 0.057. However, for 52 weekly returns in one year, we need to bootstrap the p-values. Although the bootstrap procedure is not standard in this case, it is worked out in detail by Davidson (2008) and described in Appendix A.

2.1.2. Test statistic of Davidson (2008) and portfolio choice based on it

As the true return generating process is not known, one cannot directly compute and compare the integrals from Equation (1). Rather, one has to use their sample counterparts. Following the notation of Davidson (2008), we label the sample counterparts of the integrals from Equation (1) as $D_K^2(z)$, where K denotes the two sample distributions (A or B) that are being compared. We will refer to $D_K^2(z)$ as a dominance function:

$$D_K^2(z) = \frac{1}{N_K}\sum_{i=1}^{N_K} \max(z - y_{i,K}, 0), \tag{2}$$

where N_K is a number of observations in distribution sample *K*, $y_{i,K}$ is the i-th observation in this sample, and *z* is the threshold of interest.

In order to obtain meaningful test statistics, the set of thresholds {z} includes all unique observation from both samples $\{y_{i,A}\}$ and $\{y_{i,B}\}$ lying in the joint support of those samples such that there is at least one observation in each sample above max(z) and at least one below min(z). For more powerful tests one needs to trim the set of thresholds, a discussion which we defer until later.

In the next step, for each level of z the standardized difference between the two dominance functions is computed:

$$t(z) = \frac{D_B^2(z) - D_A^2(z)}{\left(\hat{Var}(D_A^2(z)) + \hat{Var}(D_B^2(z)) - 2\hat{Cov}(D_A^2(z), D_B^2(z))\right)^{1/2}}, \tag{3}$$

where $\hat{Var}(\cdot)$ and $\hat{Cov}(\cdot)$ are the estimated variance and covariance of the dominance functions, respectively. The precise form of these estimates is stated in Appendix B.

Second-order stochastic dominance of distribution B by distribution A implies that the quantity in Equation (3) is always non-negative, including the smallest *t(z)* value. Thus, in order to test the null hypothesis that A does not SSD B, we need to focus only on one number – the smallest value of $t(z)$. This is exactly the test statistic used by Davidson (2008):

$$t^* = \min_z t(z) . \tag{4}$$

The test statistic t^* is asymptotically normally distributed. To test for the SSD relation between two distributions, one computes the corresponding statistic t^* and determines the associated p-values either using bootstrapping or the standard normal distribution, if the sample size is large.[6] Davidson (2008) describes an appropriate bootstrap procedure for the distribution of the statistic under H0, which we summarize in Appendix A.

The larger the value of t^*, the higher the likelihood of rejecting the null; and thus, the higher is the likelihood of distribution A dominating distribution B. When constructing in-sample portfolios based on the SSD, we use the test statistic t^* as our criterion function. Under the null hypothesis, the alternative portfolio to be constructed does not dominate the benchmark portfolio. We search for a set of portfolio weights that maximizes the test statistic. For all alternative portfolios, we search for the optimal solution via a fine grid search where we vary all portfolio weights in steps of 0.02. Thus, the optimal portfolio we construct has the highest probability of rejecting the null hypothesis among all possible portfolios constructed in a given asset span.

2.2. Competing portfolios

In constructing the competing portfolios, we start with the Mean-Variance-related group of approaches and first construct two portfolios: the maximum Sharpe ratio portfolio (MaxSharpe) and the maximum Information ratio portfolio (InformationRatio). For computing MaxSharpe, we proxy for the risk-free rate using returns on the 90-day Treasury bill from Federal Reserve statistical release H.15. InformationRatio is computed by maximizing the difference in the average in-sample mean returns of the InformationRatio portfolio and Bench, scaled by the standard deviation of the tracking error between this

[6] In the current study, we use one year of daily returns for each of the portfolio choice iterations. The number of observations exceeds 250 and the asymptotic p-values are rather accurate. Thus, we reject the null of non-dominance at the 10% significance level if t^* exceeds 1.28. For the main run, we confirm that the results do not change if the bootstrapped p-values are used instead of the asymptotic ones.

portfolio and Bench. When finding the optimal weights for these portfolios, we include the short-sale constraints imposed to ensure that the portfolio is allowable for a pension fund with potential restrictions on short selling. Moreover, the short-sale constraints reduce the sensitivity of mean-variance optimization to estimation errors, outliers, and mistakes in the data – see, for example, Jagannathan and Ma (2003) who use short-sale constraints in combination with a minimum-variance portfolio.

In order to stabilize estimated weights, different approaches have been used by various authors. Kan and Zhou (2007), for example, use a mixture of mean-variance and minimum-variance portfolios. Following this path, we construct another alternative portfolio, in which the variance is minimized and the mean is restricted not to deviate from the in-sample mean of Bench by more than 1% (MinVarBench).

There are other techniques to improve mean-variance portfolio optimization. Stein (1955) plus James and Stein (1961) correct the estimated mean returns by "shrinking" them toward the mean of the global minimum-variance portfolio (Bayes-Stein shrinkage). Barry (1974) and Brown (1979) introduce a correction of the estimated variance-covariance matrix for returns based on a Bayesian diffuse prior. Pastor (2000) combines the data driven optimization with beliefs in an asset pricing model. MacKinlay and Pastor (2000) develop a missing-factor model, in which they adjusted the variance-covariance matrix for non-observed factors in an asset pricing framework. Garlappi, Uppal and Wang (2007) use a multi-prior model. All these models, however, do not necessarily perform well out-of-sample. DeMiguel, Garlappi and Uppal (2009) compare the performance of 14 different models with the naive equally-weighted scheme and find that none of the advanced models consistently outperform the simple equally weighted strategy out-of-sample based on three comparison criteria: the out-of-sample Sharpe ratio, the certainty-equivalent return for a mean-variance investor, and turnover measured as trading volume. The authors argue that the equally-weighted portfolio allocation strategy should be a natural benchmark in portfolio analysis. It is preference free, does not rely on any estimation (thus, it does not incorporate estimation errors), and it delivers a reasonable level of diversification. Following their arguments, we include the equally weighted portfolio (Equal) as a competing portfolio in our analysis. This is in line with Martellini and Ziemann (2010), who argue that estimation errors often offset the benefits of rather complicated optimal portfolio choice approaches.

As the goal of the paper is to examine out-of-sample stochastic dominance of the chosen portfolios with respect to Bench, we also construct a group of portfolios based on the risk measures consistent with SSD, such as semi-variance and expected shortfall (see, for

example, Porter (1974), Fishburn (1977), and Ogryczak and Ruszczynski (1999)). For MinSemivar, the portfolio weights are chosen to minimize the in-sample left semi-variance subject to the short-sale constraint. MinShortfall chooses weights (subject to the short-sale constraint) that minimize the expected shortfall below the 5% quantile of the in-sample portfolio returns. Following Russell and Seo (1980), who show that the minimum variance portfolio cannot be dominated in-sample and is always SSD efficient, we also include the global minimum-variance portfolio with short-sale constraints (MinVar) in our set of alternative portfolios.

2.3. Testing for significance of an increased number of dominating portfolios out-of-sample

We conduct the complete analysis for all estimation and forecast windows. That is, T=20 yearly periods of in-sample fitting for all portfolios of interest and the corresponding out-of-sample performance comparison based on the SSD criterion, where we use a significance level of 10% for the t*-statistic of Davidson (2008). There is no obvious way to aggregate 20 values of the test statistics in order to obtain a unique measure of portfolio quality. In this paper, we propose to use three relevant summary characteristics regarding out-of-sample performance: (1) the number of cases in which a given portfolio choice approach provides portfolios that dominate the benchmark out-of-sample (N^+), (2) the number of cases in which those portfolios belong to the same dominance class as the benchmark (N^0), and (3) the number of cases in which those portfolios are dominated by the benchmark (N^-).

A crucial question is whether a proposed portfolio choice mechanism significantly outperforms the benchmark out-of-sample. In order to test this, we focus on the null hypothesis of no relationship between the choice mechanism and out-of-sample dominance. We define a corresponding test statistic delta N (Δ_N) as the difference between the number of cases in which the chosen portfolio dominates the benchmark out-of-sample and the number of cases in which the chosen portfolio is dominated by the benchmark:

$$\Delta_N = N^+ - N^- \tag{5}$$

We will reject the null of no relationship if the probability of observing (under the null) a statistic larger or equal to a given Δ_N is sufficiently small. The distribution of the Δ_N

under the null is not standard and is generated using a bootstrap procedure. Having no relationship between a portfolio choice technique and future portfolio performance is equivalent to randomly picking the portfolio weights. Observed out-of-sample dominance in this case is driven purely by the random weights. In order to generate such a distribution of the Δ_N, we choose a random vector of non-negative portfolio weights, which sum up to one. Here, we use the algorithm of Rubinstein (1982) outlined in Appendix C and impose the same short-sale constraints for the bootstrapped portfolios as in our original optimization.

We undertake this procedure separately for each of the performance evaluation windows and construct hypothetical alternative (random weight) portfolios. We test for the SSD relationship between the true benchmark return distribution and the corresponding random-weight portfolio distribution in each of the performance evaluation windows. This provides the first realization of Δ_N – that is, the difference between the number of cases where the random weight portfolio dominates the benchmark and the number of cases where the benchmark dominates the random weight portfolio. We repeat the procedure 10,000 times, generating a distribution of the test statistic, which is then used for the dominance test described above. The corresponding p-value for a given level of the statistic Δ_N is computed as the share of observations in the bootstrapped distribution which are equal or larger than that level of the statistic Δ_N.

The proposed bootstrap procedure requires re-sampling of portfolio weights and not of the individual return observations. Thus, any time or cross-sectional dependence existing in the original return time series will be preserved in the bootstrapped portfolios. The SSD test of Davidson (2008) will have the same power when testing the SSD relationship between the bootstrapped portfolios and the benchmark as when the original portfolios are used.

3. The data

The majority of pension funds diversify their investment across stocks and bonds. Quite a few pension funds also invest a modest proportion of their assets into real estate. Recently, some pension funds started adding to their portfolios other, less standard, asset classes. To proxy for the last category, we use commodity investing as an additional alternative strategy and also look in the robustness section at investing into hedge funds. We approximate the performance of the four main asset classes by daily returns on the corresponding indices. The data source is Thomson Datastream.

Performance of the stock market is proxied by the total (i.e. cum dividend) return on the S&P 500 index. The data on total returns were obtained from the Datastream. We compare these returns with the hand-collected prices and dividends of the S&P 500 stocks for the period from 1989 to 2006, and find that during this period the daily returns from these two sources are virtually identical with several discrepancies on some days offset during the following days, which can possibly be attributed to counting a dividend payment one day earlier or later.

The performance of the bond market is measured as total returns on the Barclays Aggregate Bond index (former the Lehmann index). The data were also obtained from Datastream. We compare the returns from Datastream with the returns on an exchange traded fund (ETF) iShares tracking this index from September 2003 to June 2010, and find that they have virtually identical means, but the ETF is more volatile. Excluding the turbulent period of 2008-2010, the correlation coefficient of these two indices is 84%. It drops to 61% when we include the last 2 years as during the ETF suffers from larger tracking errors during the crisis period.

The real estate investment is proxied by the total return on the Datastream US real estate index. This index is based on the performance of real estate investment trusts (REITs) and constitutes a general proxy of US real estate market. We compare its performance with the returns on iShares tracking the Dow Jones REIS index. The performance of the Datastream index is perfectly aligned with performance of the iShares. The mean difference in the daily returns is 1 b.p. and the returns are highly correlated with a correlation coefficient of 98%. Large pension funds might also have exposure to real estate investment not through the trusts, but through direct ownership of commercial and residential real estate. Unfortunately, daily valuation of this kind of investment is not available.

Commodity market performance is measured by returns on the S&P-GSCI index. This is a composite index representing the monthly returns attainable in the commodity markets. It is based on unleveraged, long-only investment in commodity futures, and it is broadly diversified across the various commodities, such as energy, industrial and precious metals, agriculture, and livestock.

The risk-free rate is modeled using yields on 90-day Treasury bills from the Federal Reserve statistical release H.15. Here we assumed a dynamic trading strategy, under which a 90-day Treasury bill is purchased at time t_1 yielding y_1 per year at a price of P_1, and sold on the next trading date at time t_2 at a new yield of y_2 at a price P_2. According to the description

from the release, the yields are annualized using a 360-day year. The return over this period (r_{t_1,t_2}) is computed as

$$r_{t_1,t_2} = (P_2 - P_1) / P_1 \tag{6}$$

where

$$P_1 = \frac{1}{(1+y_1)^{90/360}} \text{ and } P_2 = \frac{1}{(1+y_2)^{(90-(t_2-t_1))/360}} \tag{7}$$

When cleaning the data, we found out that there are 29 dates at with the yields are not available. In those cases we use the yield value as of the previous day. We compare the annualized returns delivered by this strategy to the total returns on U.S. Treasury bills from 1988 to 2006 reported in the "Stock, Bonds, Bills, and Inflation 2007 Yearbook", Morningstar (2007). The yearly returns are virtually identical, the correlation coefficient exceeds 99%. Thus, we feel confident, that our trading strategy mimics the performance of 90-day Treasury bills reasonably well.

The time series of daily returns covers 21 years from January 3, 1989 to December 31, 2009 and includes 5,276 observations.[7] Although all indices used are investable through exchange traded funds (see the iShares documentation at http://de.ishares.com/global), the shorter history of these funds makes them unsuitable for the current analysis.

Descriptive statistics of the data are reported in Table 1. Panel A of Table 1 reports annual return statistics, and Panel B reports statistics based on daily returns. The daily returns on all the indices exhibit excess kurtosis and are thus not normally distributed. This fact, however, does not matter for the SSD-based portfolio choice which does not require normality. The stock, real estate, and commodity indices exhibit small negative first-order autocorrelation, while the bond index exhibits small positive autocorrelation in the daily returns. This should not introduce any problems in our optimization procedure since the levels of serial correlation in the daily returns are small (the largest in absolute value is -0.15 for the real estate index). The bootstrap test used to establish significance for the number Δ_N

[7] Our starting date is determined by the availability of all four daily time-series concurrently. In our standard run, portfolio allocations are based on daily returns; and the out-of-sample SSD relationships between the resulting portfolios are tested using straight returns. The results change only in one minor instance if the out-of-sample SSD tests are conducted using logarithmic returns.

of out-of-sample dominating portfolios (see section 2.3) does not require time-independent data and is thus also unaffected.

Table 1. Descriptive Statistics of Daily Returns on the Four Asset Classes

This table reports descriptive statistics for percentage returns on the four indices from January 3, 1989 to December 31, 2009. Panel A is based on annual percentage returns. Panel B is based on daily percentage returns. We use the S&P 500 index cum dividends to proxy for the stock market, total returns on the Barclays aggregate bond index for the bond market, Datastream U.S. real estate index for real estate investment, and S&P-GSCI index for investing in commodities. All returns are expressed in %.

	Mean	Median	Vol	Min	Max	Skewness	Kurtosis	Sharpe ratio
	Panel A: Annual Percentage Returns							
Stock	**11.20**	**10.48**	**18.18**	**-22.10**	**37.58**	**-0.31**	**2.02**	**0.38**
Bond	**9.42**	**9.72**	**6.81**	**-3.51**	**22.37**	**0.18**	**2.68**	**0.75**
Real Estate	**14.26**	**16.25**	**24.85**	**-41.85**	**65.75**	**-0.31**	**3.04**	**0.40**
Commodities	**9.79**	**18.80**	**27.23**	**-39.27**	**49.74**	**-0.43**	**1.95**	**0.20**
Risk-free	**4.34**	**4.53**	**1.98**	**1.09**	**8.35**	**0.15**	**2.64**	**0.00**
	Panel B: Daily Percentage Returns							
Stock	**0.04**	**0.06**	**1.16**	**-9.03**	**10.99**	**-0.09**	**12.31**	**0.02**
Bond	**0.03**	**0.04**	**0.30**	**-1.97**	**1.64**	**-0.19**	**5.36**	**0.06**
Real Estate	**0.05**	**0.05**	**1.63**	**-18.64**	**18.75**	**0.57**	**30.27**	**0.02**
Commodities	**0.03**	**0.03**	**1.38**	**-16.83**	**7.90**	**-0.44**	**10.68**	**0.01**
Risk-free	**0.02**	**0.01**	**0.02**	**-0.18**	**0.20**	**0.85**	**14.20**	**0.00**

Table 2 reports the correlation coefficients of the indices for yearly and daily returns. Based on yearly returns, the correlations between the indices tend to be moderate, with an exception of the correlation between the bond index and the risk-free rate of 0.49. The real estate index is negatively correlated with the risk-free rate having a correlation coefficient of -0.19. The correlations of daily returns on all indices (including the risk-free rate) tend to be smaller, with the exception of the correlation between the real estate and stock indices of 0.61 and the correlation between bond returns and the risk-free rate of 0.23.

Table 2. Correlation Coefficients of Returns on the Four Asset Classes

This table reports correlation coefficients for percentage returns on the four indices from January 3, 1989 to December 31, 2009. Panel A is based on annual percentage returns. Panel B is based on daily percentage returns. We use the S&P 500 index cum dividends to proxy for the stock market, total returns on the Barclays aggregate bond index for the bond market, Datastream U.S. real estate index for real estate investment, and S&P-GSCI index for investing in commodities.

	Bond	Real Estate	Commodities	Risk-free
	Panel A: Annual Percentage Returns			
Stock	**0.20**	**0.40**	**0.05**	**0.32**
Bond		**0.13**	**-0.08**	**0.49**
Real Estate			**-0.03**	**-0.19**
Commodities				**0.18**
	Panel B: Daily Percentage Returns			
Stock	**0.00**	**0.61**	**0.08**	**-0.05**
Bond		**-0.02**	**-0.09**	**0.23**
Real Estate			**0.06**	**-0.04**
Commodities				**-0.01**

4. Empirical results

In constructing a benchmark portfolio to represent a typical pension fund, we use portfolio weights of 50% in stocks, 30% in bonds, 10% in real estate, and 10% in commodities, in line with the above cited average allocation of the 1000 largest pension funds in 2008. The resulting portfolio has a 0.04% mean daily return and a 0.73% daily standard deviation over the entire period.

In our tests, we use one-year estimation windows and one-year forecast windows. With 21 years of data and the first year used for the initial estimation, we obtain 20 non-overlapping estimates for out-of-sample portfolio performance.[8] We investigate whether the performance of the benchmark portfolio can be improved in the SSD sense by varying portfolio weights of the four typical asset classes.

Implementing the SSD tests, we need to choose an interior interval (levels of z) in the joint support of the benchmark and the alternative portfolios on which the test statistic t^* is computed. In choosing that interval, there is a tradeoff between the power of the test and the stability of the results with respect to rare events. The more the distribution tails are trimmed,

[8] There is an implicit assumption here that funds only alter their target portfolio weights annually. This is quite realistic as changes often require approval of a supervisory board. However, these funds may well rebalance within asset classes much more frequently in response to security price changes. Since our estimation keeps all weights fixed during the year, we effectively assume that the pension funds rebalance their portfolios on a daily frequency back to the fixed weights (or weekly frequency in our extensions section).

the higher is the test's ability to rank distributions but the less informative this ranking will be regarding the tails of the distributions. For the basic set of tests, we use a 10% tail cutoff of both the largest and smallest returns of the distribution.[9] However, we investigate the results' sensitivity to the choice of a lower cutoff level in our robustness section; and our main findings remain unchanged.

4.1. Out-of-sample portfolio performance with respect to the benchmark portfolio

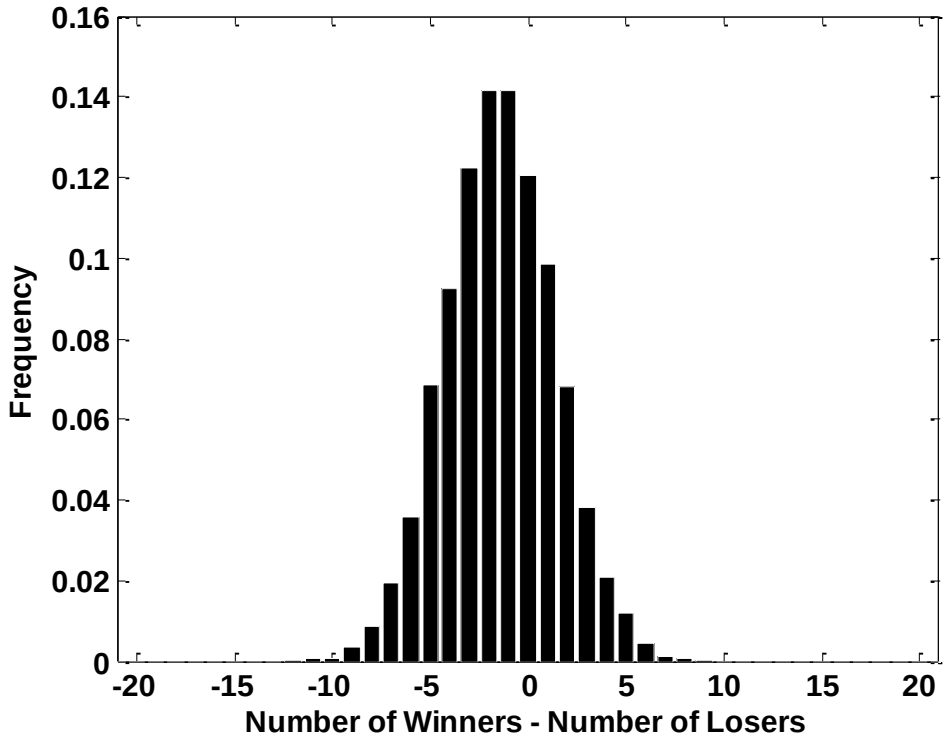
We next analyze the out-of-sample performance of a randomly chosen portfolio and the benchmark portfolio. In Figure 2 we plot the histogram of the simulated distribution of the delta N statistic Δ_N under random portfolio choice using 10,000 replications. The random portfolio performs comparable to the static benchmark portfolio, with the benchmark portfolio being slightly better. In some 63% of instances, the values of Δ_N are negative and in another 12% they are zero. This is consistent with the observation that the randomly chosen portfolio on average mimics the weights of an equally-weighted portfolio, which we will see performs reasonably well compared to the benchmark (see the subsequent discussion and results in Table 3).

Results for the out-of-sample portfolio analysis are summarized in Table 3. We use "Win" to indicate that a given portfolio dominates the benchmark out-of-sample at the 10% significance level. "Loss" indicates that a portfolio is dominated by the benchmark, and "Tie" indicates that both portfolios lie in the same dominance class. The last column of the table reports p-values from the bootstrapped distribution for the difference between the number of the out-of-sample dominating and dominated portfolios (Δ_N).

[9] Testing for stochastic dominance on a restricted (trimmed) interval goes somewhat towards the concept of almost stochastic dominance by Leshno and Levy (2002), where the distribution is said to be almost stochastic dominating if it is preferred but most (although not all) risk averse individuals.

Figure 2. Histogram of the Bootstrapped Distribution for Δ_N under Random Portfolio Choice

This figure plots the bootstrapped distribution of the Δ_N, that is, the difference between the number of dominating (winner) and dominated (loser) portfolios with respect to the benchmark, measured out-of-sample. A portfolio is said to dominate the benchmark, if the null hypothesis of non-dominance is rejected at the 10% significance level. Possible values of Δ_N are on the x-axis, with frequencies on the y-axis. The total number of (out-of-sample) periods and, thus, the maximum possible absolute value of Δ_N is 20. The sample is based on 10,000 replications.



The SSD-related group of portfolios performs admirably. SSDBased, MinVar, and MinSemivar win against the benchmark in 15 periods out of 20, and MinShortfall wins in 14 periods. None of these portfolios is dominated by Bench out-of-sample. There are, however, multiple ties, so that the alternative portfolios lie in the same dominance class as the benchmark. In terms of the p-values, all portfolios from this group significantly outperform the benchmark when compared to an uninformative random choice mechanism.

Table 3. Out-of-Sample Performance of the Alternative Portfolios

This table reports the number and percentage of the 20 forecast windows, where the considered portfolios dominate the benchmark (Win), are dominated by the benchmark (Loss), or lie in the same dominance class (Tie). The alternative portfolios are based on four asset classes: stock, bond, real estate, and commodities. The last column reports the p-values for the difference between the numbers of out-of-sample dominating and dominated portfolios.

	Out-of-Sample						
	Win		Tie		Loss		p-Values
	#	%	#	%	#	%	
SSD-related							
SSDBased	15	75	5	25	0	0	0.000
MinShortfall	14	70	6	30	0	0	0.000
MinVar	15	75	5	25	0	0	0.000
MinSemivar	15	75	5	25	0	0	0.000
Equal	5	25	12	60	3	15	0.146
Mean-Variance-related							
MinVarBench	0	0	8	40	12	60	1.000
MaxSharpe	9	45	9	45	2	10	0.002
InformationRatio	1	5	12	60	7	35	0.967

The equally weighted portfolio is a middle-performer. It wins against Bench in 5 periods out of 20 and loses in 3 periods. From the perspective of second-order stochastic dominance, the benchmark portfolio seems not to be structured any better than the equally weighted.

In contrast, the Mean-Variance-related portfolios (MinVarBench, MaxSharpe and InformationRatio) perform rather poorly, with the exception of MaxSharpe, which is the strongest portfolio within this group. It dominates Bench in 9 periods, however, it is dominated by Bench during 2 periods. InformationRatio generates out-of-sample dominance during only 1 period and loses against the benchmark in 7 periods. These results appear to be due to the unstable and extreme weights of the mean-variance optimization approach that we discussed earlier. MinVarBench is the weakest portfolio within this group; it does not win in a single period and loses in 12 periods against Bench. Its p-value of 1 in Table 3 indicates that the random choice mechanism, used in creating the bootstrap, always outperformed MinVarBench against the benchmark. MinVarBench and InformationRatio perform significantly worse than the random portfolio choice technique in terms of SSD of Bench. Such mean-variance based approaches are poor choices for any investor with an increasing and concave utility function. Moreover, those approaches might even lose against the random choice mechanism.

We next investigate more closely the time pattern of dominating portfolios generated by different methods. The SSD-related methods perform admirably in the crisis years of 2007-2009. The other methods tend to be characterized by the fact that there exist several years where the benchmark can be relatively easily beaten: 1990-1991, and broadly 2001-2003. However, there is no obvious interpretation of why the benchmark has such difficulties during those years including the post internet bubble period. The worst performing methods are characterized by altogether rare instances of second order stochastic dominance.

Superior performance of SSDBased is not surprising as the method is especially engineered for the SSD criterion. It accounts for all SSD-relevant information of the two return distributions to be compared and not only for a limited number of moments or other characteristics (e.g. shortfall). Capturing the most of the SSD-relevant information in-sample, SSDBased is able to generate out-of-sample dominance most of the time.

Table 4 reports the descriptive statistics of the optimal portfolio weights generated by the portfolio choice approaches we examined. Among the SSD-related group, all portfolios have a clear tendency to increase the bond holdings compared to stock holdings in order to minimize return volatility. However, SSDBased is the only approach within this group that also puts a substantial weight on the stock holding. Regarding time-stability of portfolio weights, the mean-variance based portfolios have the most volatile weights, which often take on the extreme values of 0 and 1.

Table 4. Optimal Portfolio Weights of the Alternative Portfolios

This table reports the descriptive statistics of the optimal portfolio weights (mean, standard deviation, minimum and maximum weights across 20 periods) delivered by the alternative portfolio choice approaches. The portfolios are based on four asset classes: stock, bond, real estate, and commodities.

	Stock Proportion				Bond Proportion			
	Mean	Vol	Min	Max	Mean	Vol	Min	Max
SSD-related								
SSDBased	0.35	0.14	0.06	0.48	0.47	0.16	0.28	0.78
MinShortfall	0.03	0.04	0.00	0.12	0.77	0.14	0.40	0.94
MinVar	0.04	0.04	0.00	0.14	0.82	0.09	0.60	0.94
MinSemivar	0.04	0.04	0.00	0.12	0.82	0.08	0.64	0.92
Equal	0.25	0.00	0.25	0.25	0.25	0.00	0.25	0.25
Mean-Variance-related								
MinVarBench	0.34	0.24	0.00	0.72	0.03	0.05	0.00	0.16
MaxSharpe	0.07	0.11	0.00	0.36	0.03	0.34	0.00	1.00
InformationRatio	0.49	0.10	0.26	0.68	0.20	0.13	0.00	0.44

Table 5 reports descriptive statistics for the returns delivered by the alternative portfolios. Panel A is based on annual returns, and Panel B is based on daily returns. In analyzing the quality of the resulting portfolios, we introduce three additional measures: certainty equivalent (CEV3), turnover (Turnover), and a share of extremes (%Extreme), which are also reported in Table 5 for annual returns.

CEV3 is defined as the inverse of the expected utility function, where we use constant relative risk aversion utility function with the risk aversion parameter γ of 3:

$$CEV = u^{-1}\left(\frac{1}{T}\sum_{t=1}^{T} u(1+r_t)\right), \tag{8}$$

$$u(1+r_t) = \frac{(1+r_t)^{1-\gamma}}{1-\gamma}, \tag{9}$$

where T is the total number of yearly periods.

Turnover serves as a proxy of transaction costs associated with the optimal portfolio rebalancing. It is computed as a time average total absolute change in all four portfolio weights:

$$Turnover = \frac{1}{T-1}\sum_{t=2}^{T}\sum_{i=1}^{4} | w_{it} - w_{it-1} |, \tag{10}$$

where w_{it} is the optimal portfolio weights of the asset class *i* in period *t*.

%Extreme is defined as a share of periods, in which at least one of four optimal portfolio weights takes an extreme value of 0 or 1.

Comparing to Bench, the SSD-based approach preserves the mean annual return while shrinking the variance by avoiding large losses (the minimum return is just -8.47% compared to -24.64% of Bench). At the same time, large gains are still possible (maximum return of SSDBased is 27.92% vs. 29.31% of Bench). It results in a higher Sharpe Ratio than for Bench (0.65 vs. 0.46). Moreover, SSDBased has the highest certainty equivalent of 8.64% than any other of the discussed portfolios including Bench.

Other portfolio choice approaches from the SSD-related group decrease the portfolio variance by even more than SSDBased, but this comes at the cost of a decline in the mean returns. These portfolios avoid large losses but also limit potential gains. They are normally less diversified: in 50 to 80% of the periods they have extreme (0 or 1) portfolio weights, as compared to only 5% of periods (1 year out of 20), during which SSDBased has extreme weights. MinShortfall, MinVar, and MinSemivar invest largely in bonds, as they characterize by the lowest variance. During the investigated period, the bond index exhibit more favorable risk-return tradeoff than stocks, resulting in very high Sharpe ratios of these portfolios. However, if we would use other bond indices with lower mean returns, e.g., 5-year treasuries, these portfolio choice approaches will still be nearly fully invested in bonds. As a consequence, their mean returns will decrease and the corresponding Sharpe ratios will be much less appealing.

The equally weighted portfolio performs rather similar to Bench, having somewhat higher mean return and standard deviation than Bench, resulting in a comparable Sharpe ratio and slightly higher certainty equivalent.

Table 5. Descriptive Statistics of Portfolio Returns

This table reports descriptive statistics of the percentage returns for different portfolio choice strategies, including the benchmark (Bench), the SSD-based portfolio (SSDBased), the minimum shortfall portfolio (MinShortfall), the minimum variance portfolio (MinVar), the minimum semi-variance (MinSemivar), the equally-weighted portfolio (Equal), the minimum variance portfolio with the mean return equal to the in-sample mean of the benchmark portfolio (MinVarBench), the maximum Sharpe ratio portfolio (MaxSharpe), and the portfolio with the maximum Information ratio relative to Bench (InformationRatio). Panel A is based on annual percentage returns. The statistics are computed using 20 yearly returns from 1989 to 2009. Panel B uses daily percentage returns for the same time period. The last three columns of the table report certainty equivalent based on power utility function with the relative risk aversion parameter of 3 (CEV3), annual portfolio turnover, and a percentage yearly periods with extreme (0 or 1) portfolio weights.

	Mean	Vol	Min	Max	Skew-ness	Kurtosis	Sharpe ratio	CEV3	Turnover	%Extreme
Panel A: Annual Returns										
Bench	9.73	12.60	-24.64	29.31	-0.82	3.98	0.46	7.26	0.00	0.00
SSD-related										
SSDBased	9.64	8.81	-8.47	27.92	0.20	3.06	0.65	8.64	0.36	0.05
MinShortfall	9.00	6.32	-1.03	21.22	0.21	2.38	0.80	8.48	0.26	0.80
MinVar	9.14	6.25	-1.51	21.71	0.25	2.59	0.83	8.63	0.19	0.60
MinSemivar	9.12	6.46	-1.49	21.82	0.37	2.49	0.80	8.59	0.23	0.50
Equal	10.08	13.08	-26.60	26.60	-1.05	4.12	0.47	7.30	0.00	0.00
Mean-Variance-related										
MinVarBench	8.59	18.28	-41.40	46.75	-0.64	4.45	0.26	2.60	0.96	0.60
MaxSharpe	9.50	10.02	-15.94	27.94	-0.85	3.88	0.56	8.01	0.93	0.85
InformationRatio	10.60	14.00	-29.46	29.69	-1.17	4.35	0.48	7.29	0.49	0.35
Panel B: Daily Returns										
Bench	0.04	0.73	-6.80	7.09	-0.30	15.36	0.03	0.03	--	--
SSD-related										
SSDBased	0.04	0.48	-3.57	2.68	-0.19	6.54	0.04	0.03	--	--
MinShortfall	0.03	0.29	-1.75	1.56	-0.17	5.19	0.07	0.03	--	--
MinVar	0.03	0.27	-1.75	1.56	-0.16	5.24	0.07	0.03	--	--
MinSemivar	0.03	0.28	-1.75	1.56	-0.17	5.35	0.07	0.03	--	--
Equal	0.04	0.75	-8.35	7.97	-0.36	21.94	0.03	0.03	--	--
Mean-Variance-related										
MinVarBench	0.03	0.97	-8.73	9.02	-0.41	13.06	0.02	0.02	--	--
MaxSharpe	0.04	0.60	-5.83	4.09	-0.49	13.30	0.04	0.03	--	--
InformationRatio	0.04	0.78	-6.46	6.76	-0.41	12.63	0.03	0.03	--	--

The Mean-Variance-related approaches all have higher standard deviations than SSDBased. Among them, MinVarBench turns out to be the most volatile with the return standard deviation of 18.28%. It also generates larger out-of-sample losses, with a minimum annual return of -41.40% vs. -24.61% for Bench. MaxSharpe has a slightly lower mean return and a lower standard deviation as compared to Bench, resulting in a higher Sharpe Ratio than that of Bench, but still smaller than that of SSDBased. InformationRatio delivers higher mean annual returns than any other of the considered portfolios but exhibits a higher variance than all portfolios but MinVarBench. The Mean-Variance-related portfolios are characterized by the highest turnover ranging from 0.49 for InformationRatio to 0.96 for MinVarBench, and rather large number of periods with extreme portfolio weights ranging from 35% for InformationRatio to 85% for MaxSharpe.

5. Robustness

In this section, we assess the stability of our results. First, we investigate whether the main patterns in our results are preserved if the benchmark portfolio composition is changed or if any of the competing portfolios becomes an out-of-sample benchmark. Second, we test the sensitivity of the results to several methodological changes, such as the length of the estimation window, using weekly instead of daily returns, the level of data trimming, and the in-sample data trimming for the alternative methods. Third, we check if the ranking of portfolio choice approaches is preserved when the asset span of the alternative portfolios is extended. Forth, we investigate the stability of the results with respect to the index choice to proxy for the asset classes. Last, we investigate the performance of the portfolio choice approaches during structural breaks, in which the estimation and forecast windows may be characterized by different return dynamics. None of the considered robustness checks changes the results substantially.

5.1. Robustness with respect to the benchmark

Here, we perform several robustness checks with respect to the benchmark. In the main run, the benchmark portfolio is static with constant portfolio weights of 30% in bonds, 50% in stock, 10% in real estate, and 10% in commodities. First, we vary these weights, keeping the benchmark static. Second, we allow small positions in a risk-free investment ranging from -5% to +10% (the benchmark portfolio is still static is this case). Last, we dynamically adjust the benchmark portfolio, such that each of the competing portfolios is

used as an out-of-sample benchmark; we thus test each portfolio against each for the out-of-sample dominance.

5.1.1. Static benchmark with different portfolio weights

The current benchmark composition is 30% in bonds, 50% in stock, 10% in real estate, and 10% in commodities. We use alternative benchmark portfolios that invest (1) 20% in bonds, 60% in stock, 10% in real estate, and 10% in commodities, (2) 35% in bonds, 50% in stock, 10% in real estate, and 5% in commodities, and (3) 20% in bonds, 60% in stock, 5% in real estate, and 15% in commodities. Additionally, we consider two alternative benchmark portfolios with weights concentrated in stocks or bonds: (4) 15% in bonds, 75% in stock, 5% in real estate, and 5% in commodities; (5) 55% in bonds, 25% in stock, 10% in real estate, and 10% in commodities. The general ranking of portfolio choice approaches does not change. The SSD-related group of portfolios always significantly outperforms the benchmark portfolio out-of-sample. Typically, SSDBased has slightly better statistical support and exhibits higher values of the delta N statistics Δ_N than its competitors. The only exception is the first scenario with Bench having 20% invested in bonds and 60% in stock. In this case, SSDBased dominate Bench in 15 of 20 periods, whereas other portfolios from the SSD-related group dominate Bench in 16 periods. Remarkably, when Bench is concentrated in stock, it becomes easier for a random portfolio to dominate it. The main mass of the bootstrapped distribution of the delta N statistics lies to the right from zero with less than 0.5% of values being negative. If, however, Bench is bond-concentrated, it becomes more difficult for a random portfolio to dominate Bench because of its low variance. The bootstrapped distribution of the delta N statistics lies in this case within negative area. The portfolio weights of SSDBased do not change much when we vary the benchmark as described.

5.1.2. Static benchmark with the risk-free investment

We now allow the standard benchmark portfolio (30% in bonds, 50% in stock, 10% in real estate, and 10% in commodities) to also have a small position in the risk-free asset. We vary the weights of the risk-free asset using -5%, -2%, 2%, 5%, and 10% weights. The holdings of the main asset classes are proportionally adjusted such that the total sum of weights equals one. The alternative portfolios are still based on the four main asset classes.

The key results of the paper do not change. The SSD-related group of portfolios always significantly outperforms the benchmark portfolio out-of-sample. SSDBased is nearly always characterized by the largest values of the delta N test statistics. Interestingly, long positions in the risk-free asset seem to positively influence the performance of Bench, whereas short positions worsen the performance. For example, the number of instances in which SSDBased dominates Bench out-of-sample decreases from 15 for no risk-free asset in Bench (Table 3) to 14 for a risk-free holding of 5%, and increases to 17 for a risk-free holding of -5%. The corresponding delta N statistics stay highly significant. Altogether, the holding of the risk-free asset does not introduce any qualitative changes compared to the results reported in Table 3.

5.1.3. Cross-comparison of the portfolios

In this section we address the performance of the portfolio choice approaches if the benchmark portfolio is changed dynamically. The in-sample optimal weights are determined as usual using Bench as a reference, if needed, the standard static benchmark. The out-of-sample tests, however, are performed with respect to a dynamically adjusted benchmark. We use each of the competing portfolios as the out-of-sample benchmark in turn, and compute the number of periods in which each of other portfolios dominate (are dominated by) this benchmark. Table 6 reports the estimation results. In each row, we report the number of winning periods of the corresponding portfolio over the benchmark indicated in the column.

Within the group of SSD-related portfolios, SSSBased is often dominated by other portfolios from this group. This is caused by these portfolios having lower variance than SSDBased, thus, when tested for SSD on an interval restricted to lie in the common support of the distributions, SSDBased has longer left tail and, thus, the null hypothesis of non-dominance of, say, MinVar over SSDBased, cannot be rejected.

All portfolios from the SSD-related group often dominate portfolios from the Mean-Variance-related group. Notably, even the equally weighted portfolio dominates MinVarBench in 15 periods without being ever dominated by this portfolio, and Equal dominates InformationRatio in eight periods while being dominated by it in only once.

Table 6. Out-of-Sample Performance with Dynamic Benchmarks

This table reports the number of the forecast windows, where the competing portfolios indicated in the rows dominate the alternative benchmarks, indicated in the columns. The portfolios are based on daily returns with four asset classes: stock, bond, real estate, and commodities.

	SSD-Related				Equal	Mean-Variance-Related		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
SSD-Related								
(1) SSDBased	--	0	0	1	5	17	4	16
(2) MinShortfall	11	--	1	3	11	18	11	14
(3) MinVar	13	3	--	3	12	17	11	13
(4) MinSemivar	12	1	1	--	12	17	10	13
(5) Equal	3	0	0	0	--	15	4	8
Mean-Variance-Related								
(6) MinVarBench	0	0	0	0	0	--	1	1
(7) MaxSharpe	5	0	0	0	8	12	--	9
(8) InformationRatio	0	0	0	0	1	9	2	--

5.2. Methodological robustness

Here, we change the methodology by considering weekly returns instead of daily returns, by varying the levels of z-interval trimming, by changing the lengths of the estimation and forecast windows, by changing significance levels in our tests, by trimming the in-sample data when using other than SSD-based portfolio choice approaches, and, finally, by allowing alternative portfolios to have small positions in the risk-free rate.

5.2.1. Weekly returns

In this sub-section, we check whether our results are an artifact of using daily returns or whether they can also be documented with weekly returns. Using weekly returns implies that the portfolios are rebalanced to their target levels each week, whereas during each week pension funds follow a buy-and-hold strategy. It also decreases the number of observations considerably. Consequently, we cannot rely on the asymptotic properties of the Davidson (2008) test in determining winning and losing distributions and we use the bootstrapped p-values instead.

As we sharply decrease the number of observations, the power of the Davidson (2008) test decreases. Consequently, it becomes more difficult to rank the portfolio return distributions according to their dominance relations. For example, the SSD-based portfolio

dominates the benchmark in 8 of 20 forecast windows based on weekly returns, compared to 15 forecast windows with daily returns. Nevertheless the results based on weekly returns are qualitatively consistent with the results for daily returns in Table 3. The SSD-related group of portfolios outperforms the benchmark out-of-sample with zero p-values. The mean-variance-related portfolios underperform. The delta N statistics of the MaxSharpe portfolio of 2 is significant only at the 10% level based on weekly returns.

5.2.2. Different levels of trimming

As described previously, the Davidson (2008) test statistic is computed using sets of z-values that lie in the joint support of the two distributions being compared. So far, we trimmed the 10% largest and 10% smallest observations from the joint support in order to assure high power of the test. To check whether tail behavior adversely influences our previous results, we now perform the analysis using smaller levels of tail trimming. Note, that the optimal portfolio weights for SSDBased are different from the main run. Given different levels of z-interval trimming, the in-sample test statistic t^* reaches its minimum at different values of portfolio weights. However, the weights are rather stable. The maximum change corresponds to the stock index, where the average weight changes from 35% for the 10% trimming to 29% for the 1% trimming. Table 7 summarizes the estimation results obtain with 1% tail trimming.

Increasing the z-interval towards the tails makes it more difficult to rank distributions based on the dominance criterion, as the tails tend to be thinner. As a result, the minimum test statistic of Davidson (2008) turns out to be smaller; making it harder to reject the null hypothesis of non-dominance. Many more portfolios are now classified as Tie. For example, the number of forecast windows where we have dominance of the SSD-based portfolio deceases from 15 with 10% tail trimming (Table 3), to 11 for 5% trimming, and to 5 with 1% trimming.

The SSDBased portfolio still significantly outperforms the benchmark out-of-sample together with other portfolios from the SSD-related group. The corresponding p-value is 0.001. The mean-variance based portfolios continue to perform poorly, with the exception of MaxSharpe, which wins in 5 periods out of 20 and loses during 1 period, still having a p-value of 0.013.

Table 7. Out-of-Sample Performance with Different Levels of z-Interval Trimming

This table reports the number and percentage of the 20 forecast windows, where the alternative portfolios dominate the benchmark (Win), are dominated by the benchmark (Loss), or lie in the same dominance class (Tie). The alternative portfolios are based on daily returns with four asset classes: stock, bond, real estate, and commodities. The last column reports p-values for the difference between the numbers of the out-of-sample dominating and dominated portfolios. The results are computed using the 1% trimming of the z-interval. The z-interval is an interval lying in the joint support of the distributions to be compared, on which the Davidson (2008) test statistic is computed.

	Out-of-Sample						
	Win		Tie		Loss		p-Values
	#	%	#	%	#	%	
SSD-related							
SSDBased	5	25	15	75	0	0	0.001
MinShortfall	6	30	14	70	0	0	0.000
MinVar	5	25	15	75	0	0	0.001
MinSemivar	7	35	13	65	0	0	0.000
Equal	2	10	18	90	0	0	0.172
Mean-Variance-related							
MinVarBench	0	0	15	75	5	25	1.000
MaxSharpe	5	25	14	70	1	5	0.013
InformationRatio	0	0	19	95	1	5	0.840

5.2.3. Changing lengths of estimation and forecast windows

Instead of using one year estimation and forecast windows, we implement the analysis based on quarterly and on two-year windows. The results only change minimally. Based on both quarterly and two-yearly horizons, the p-values of all portfolios within the SSD-related group are zeros. Mean-Variance-related portfolios perform poorly. The strongest portfolio MaxSharpe has a p-value of 0.051 at the yearly horizon, and only 0.113 at the quarterly horizon.

5.2.4. Changing significance levels for dominating portfolios

In the current analysis, an alternative portfolio is said to dominate Bench, if the null hypothesis of non-dominance can be rejected at the 10% significance level. We now change the significance level to 5% and to 1%, respectively. The results change only mildly compared to the ones reported in Table 3. The number of dominating portfolios decreases, but the ranking of the portfolio choice approaches does not change. At the 1% significance level, the difference between SSDBased and other portfolios from the SSD-related group becomes more pronounced with SSDBased dominating Bench in 10 out of 20 periods and

the second best portfolio – MinVar – dominates Bench in 7 periods. Moreover, at the 1% significance level MaxSharpe no longer significantly outperforms random portfolio choice mechanism with respect to Bench, having a p-value of 0.115.

5.2.5. In-sample trimming of other methods

While estimating optimal weights of the SSDBased portfolio we trim 10% of the in-sample data in order to compute the required test statistic. We now re-estimate optimal weights for other competing portfolios as well, based on similarly trimmed data in-sample. For example, when finding the optimal weights for MaxSharpe in-sample, for each set of weights we use only those returns, which lie within the interval [min(z), max(z)], where z is the set of return on which SSD test statistic is determined. This trimmed set is used only in-sample for finding optimal weights, and not out-of-sample, in which portfolio returns are computed based on all out-of-sample all data points available.

The estimated delta N statistics do not change for any of the competing portfolios by more than one. The only exception is MaxSharpe portfolio, for with the value of the statistics decreases from 7 reported in Table 3 to only 1, if the in-sample trimming is applied. Thus, we are confident that the superior performance of SSDBased in not driven by the in-sample trimming.

5.2.6. Relaxing short-sale constraints

The main analysis is based on portfolios optimized under short sale constraints; that is the portfolio weight cannot take values below zero. We now relax this assumption and allow individual portfolio weights to lie between -1 and 1. This change does not significantly impact the performance of the SSD-related methods, however, it harms the performance of Mean-Variance-related methods. The number of cases in which the optimal weights generated by these methods take values of -1 or 1 increases. In the standard run, MaxSharpe dominates Bench in 9 periods out of 20 and is dominated by Bench in 2 periods. After the short-sale constraints are relaxed, dominance of MaxSharpe can be documented in 8 periods, and in 5 periods MaxSharpe is dominated by Bench.

5.3. Adding the risk-free asset to the asset span

In the main run, we considered a fixed asset span of four assets (stock, bond, real estate, and commodities) for the optimal portfolio allocation. Here, we investigate the

stability of the main results when the asset span is extended by allowing positions in the risk-free asset. Such positions can stem from cash management purposes to accommodate fund flow and expenses.[10] It is also possible that a fund concerned about a market decline might decide to shift a portion of its investment into the riskless asset (rather than increasing its bond position). We explore these possibilities by allowing the portfolio weight on the risk-free asset to take on values between -5% and 10%. For each time period under consideration, the risk-free weight is optimally chosen for all the alternative portfolios except the equally weighted one. The equally weighted portfolio is now constructed in 5 different versions with the weight on the risk-free asset being -5%, -2%, 2%, 5%, and 10% respectively. The remaining weight is then equally distributed across the main asset classes.

Allowing for the risk-free investment extends the asset span and thus makes winning against the benchmark easier. We note that generally, extending the asset span for the competing methods while not changing the benchmark will tend to increase the number of wins of the competitors over the benchmark. It results in an increase in the number of instances in which the portfolios from the SSD-related group dominate Bench out-of-sample. Positive risk-free investment increases the number of instances in which the equally weighted portfolio dominates Bench. The risk-free investment does not change the qualitative performance of MaxSharpe and InformationRatio relative to Bench. Generally, the results are in line with the ones reported in Table 3.

5.4. Using different sets of indices to proxy for stock and bond asset classes

In this section we address the question of whether the reported results are driven by the particular choice of indices to proxy for four asset classes.

We repeat the main analysis, first, using the returns on the Dow Jones Industrial Average index instead of the returns on the S&P 500 index as a proxy for stock returns. The two indices are highly correlated (the correlation coefficient is 0.96), but the DJ index is characterized by slightly lower mean return and lower return standard deviation. The results change only mildly compared to the ones reported in Table 3.

[10] In normal circumstances, actual cash positions are likely to be small as pension funds attempt to stay fully invested. One large state pension fund reported in private communication a typical cash position of only 0.5%. Moreover, futures positions on equity and bond indices can be used to offset any cash build up and remain effectively fully invested.

Second, we perform a robustness check with respect to the bond index. Instead of returns on the Barclays Aggregate Bond Index, we use returns on the daily trading strategy with 5-year zero Treasury bonds, as well as returns on a similar trading strategy based on Moody's Aaa and Baa corporate bonds. The yields on these bonds are obtained from the H.15 release. Regardless the bond index used, the SSD-related group of portfolios always has a highly significant delta N statistic, whereas the performance of the MaxSharpe portfolio weakens. When 5-year Treasuries or Aaa-rated bonds are used, MaxSharpe still outperforms the random portfolio choice mechanism with respect to Bench, but the p-values increase to 0.048 and 0.027 respectively. When the Baa-rated bonds are used, which are characterized by higher return and volatility, the MaxSharpe loses significance. The corresponding p-value is 0.212.

Next we repeat the analysis using Moody's commodity index instead of the S&P-GSCI index. The commodity indices have only correlation coefficient of 24% as the composition of the Moody's index differs substantially from the S&P-GSCI index. This index consists of 12 agricultural products and 6 metals. Thus, all energy-related products are omitted in this case. Similar to the case with Baa-rated bonds, the SSD-related portfolios still significantly outperform random portfolio in terms of out-of-sample dominance over Bench, whereas MaxSharpe does not do so; its p-value increases to 0.135.

The commodity index used as a fourth, alternative asset class might not be the best proxy for alternative investments. Some large pension funds recently have started investing in hedge funds. We substitute the commodity index by the daily returns on the FHRX global hedge fund index from the 1st of April 2003 until 31st of December 2009, and repeat the analysis. In this case, we have 6, instead of 20, out-of-sample periods. The HFRX index presents performance of rather liquid hedge funds. HFR includes in this index some 6000 hedge funds with at least 24 months of performance records, having at least US$ 50 million assets under management, which are open to new investors, and are willing to trade on a transparent basis, for example, through managed accounts. These funds, thus, may not represent the universe of hedge funds, as most of hedge funds are illiquid and not that transparent. It gives, however, a rather good benchmark for a group of hedge funds, which might be of interest to highly regulated and conservative pension funds. HFR also offers investible products based on the daily HFRX indices.

The performance characteristics of the HFRX global index are rather remarkable. It has rather low mean daily return of just 0.01% and return volatility of 0.26% from 2003 to 2009, compared to the S&P500 index having 0.03% mean and 1.37% volatility for the

same period. Trying to minimize the variance, the SSD-related portfolios tend to have relative large proportions invested in this index, being on average 50%. In terms of the delta N statistics, the SSD-related portfolios dominate Bench in all 6 out-of-sample periods, whereas the best performing portfolio from the Mean-Variance-related group – MaxSharpe – dominates Bench in 3 periods and is dominated by Bench in 1 period. MaxSharpe does not significantly outperform the random portfolio choice mechanism against Bench in this case.

Last but not least, we recognize that not all pension funds invest in alternative investments altogether, and even those, which do use other investments other than stocks, bonds, and real estate, started doing so not more than a decade ago. In order to test whether the SSD approach still performs well in a rather conservative asset span, we exclude the commodity index from the analysis, and re-estimated the portfolios based on stocks, bonds, and real estate asset classes only. The benchmark in this case consists of 55% stocks, 35% bonds, and 10% real estate. It seems to be easier for SSDBased to win against the 3-asset benchmark, even though the SSDBased consists of the same three assets only. It wins now in 17 out of 20 periods (compared to 15 in Table 3), and never loses against the benchmark portfolio. Other SSD-related portfolios win in 15 periods. The next best portfolio – MaxShape – wins only in 11 periods.

5.5. Market turmoil and structural breaks

We are interested in the stability of results concerning market turmoil and related structural breaks. Gonzalo and Olmo (2008) investigate similar situations of financial distress, albeit only in-sample where we construct our tests out-of-sample. To this end, we investigate if the optimal portfolios constructed during periods prior to some specific event (associated with market turmoil) preserve their dominance during subsequent periods with adverse market dynamics. As in the standard runs, we use four asset classes with a one-year estimation window, which now ends before the event of interest. The one-year forecast window starts just prior to the event and always includes the event. We focus on four distinct events listed below.

1. Russian default 1998: The official day of the Russian default is the August 17, 1998, when the Russian government and the Central Bank of Russia announced the restructuring of ruble-denominated debt and a three month moratorium on the payment of some bank obligations. Prior to this date, however, investors' fears of

possible default led to the collapse of the Russian stock, bond, and currency markets as early as August 13, 1998. Thus, we choose the corresponding estimation and performance evaluation windows in such a way that the complete month of August 1998 is included in the latter. The estimation window is August 1, 1997 to July 31, 1998, and the forecast window is August 1, 1998 to July 31, 1999.

2. End of the internet bubble 2000: The NASDAQ composite index heavily loads on (internet) technology stock. It nearly doubled its value during 1999 and early 2000. It first dropped in value on March 13, 2000 after having reached its historical peak on March 10, 2000. We use a period of the bull market from March 1, 1999 to February 29, 2000 for the estimation of the portfolio weights, and assess the portfolio performance during the bear market from March 1, 2000 to February 28, 2001.
3. Terrorist attack of September 11, 2001: The estimation window is from September 1, 2000 to August 31, 2001, and the performance evaluation window is September 1, 2001 to August 31, 2002.
4. Financial crisis 2007-2008: The financial crisis of 2008 hit in September 2008 when several large US banks and financial firms including Lehman Brothers collapsed, leading to bankruptcies of other companies and worldwide recession. The first signs of the coming turmoil appeared, however, much earlier. In July 2007 the spread between three-month LIBOR and three-month T-bill interest rates (TED spread) that proxies for the overall credit riskiness in the economy spiked up; and on August 9, 2007 the US Federal Reserve and the European Central Bank injected $90bn into financial markets. Since we are trying to completely exclude information about the upcoming events from the estimation window, we choose July 1, 2006 to June 30, 2007 for the estimation, with the forecast window from July 1, 2007 to June 30, 2008.
5. Economic recession 2009: The estimation window covers the period of the financial crises from January 1, 2008 to December 31, 2008, and the forecast is based on a subsequent year from January 1, 2009 to December 31, 2009.

Table 8 reports estimation results based on the fixed asset span of four asset classes (stocks, bonds, real estate, and commodities). Since we only have one event in each case, we cannot compute our usual delta N test statistics of wins minus losses. Instead, the table reports the test statistic t* of Davidson (2008) and the corresponding p-value for the null hypothesis that an alternative portfolio does not dominate the benchmark out-of-sample. The hypothesis is rejected if the p-values are small.

Table 8. Portfolio Performance around Special Events

This table reports out-of-sample performance tests statistics and the corresponding p-values for different portfolio choice mechanisms, including the SSD-related group of portfolios (SSDBased, MinShortfall, MinVar, and MinSemivar), the equally weighted portfolio (Equal), and the Mean-Variance-related group of portfolios (o (MinVarBench, MaxSharpe, and InformationRatio). The portfolios are based on four asset classes: stocks, bonds, real estate, and commodities. The estimation windows include one year of daily observations prior to special events of interest (excluding the events). The performance evaluation windows cover one year after the events and include the corresponding events: (1) Russian default 1998, (2) End of the internet bubble 2000, (3) Terrorist attack of September 11, 2001, and (4) Financial crisis 2007-2008, (5) Economic recession 2009.

	(1) Russian default 1998		(2) End of the internet bubble 2000		(3) Terrorist attack of Sep. 11, 2001		(4) Financial crisis 2007-2008		(5) Economic Recession 2009	
	t-stat	p-value	t-stat	p-value	t-stat	p-value	t-stat	p-value	t-stat	p-value
SSD-related										
SSDBased	1.992	0.023	2.880	0.002	4.832	0.000	3.018	0.001	3.247	0.001
MinShortfall	1.902	0.029	4.793	0.000	4.749	0.000	2.823	0.002	3.651	0.000
MinVar	2.011	0.022	4.703	0.000	4.658	0.000	3.599	0.000	3.377	0.000
MinSemivar	1.879	0.030	4.895	0.000	4.639	0.000	2.823	0.002	3.736	0.000
Equal	0.561	0.287	3.299	0.000	2.538	0.006	0.085	0.466	-7.988	1.000
Mean-Variance-related										
MinVarBench	-2.647	0.996	-2.506	0.994	-1.223	0.889	-3.638	1.000	-5.498	1.000
MaxSharpe	2.739	0.003	-5.464	1.000	4.497	0.000	1.919	0.027	2.842	0.002
InformationRatio	-2.083	0.981	-4.239	1.000	0.402	0.344	-8.204	1.000	0.909	0.182

In all cases, the portfolios from the SSD-related group dominate the benchmark out-of-sample. All these portfolios perform relatively weaker during the year following the Russian default, for which the null of non-dominance over the benchmark portfolio can be rejected at the 5% significance level. Equal dominates the benchmark in two of five investigated periods, but not during the financial crisis of 2007-2008 and the following recession. Moreover, during year 2009, Equal is dominated by Bench. The Mean-Variance-related portfolios, as in previous tests, perform poorer around special events. MinVarBench and InformationRatio never dominate Bench around any of the special events. Moreover, InformationRatio is dominated by Bench during Financial crises 2007-2008. MaxSharpe is a relatively strong portfolio, which dominates Bench in four of five considered periods.

6. Simulation Evidence on the Performance of Different Methods Across Different Measures

In order to assess the performance of our different methods across the many performance measures, we resort to a simulation study where we concentrate on the methods SSDBased, MinVar, Equal, and MaxSharpe.[11] In particular, we are interested in uncovering which features in the data cause the different methods to perform well or not. Thus we start out with the actual data and then investigate the estimation error by bootstrapping the yearly data. Next, we look at time-variation in the distributions by forcing the portfolio optimization and the performance measurement to use data from the same distribution, thus suppressing the effect of time-variation in the distributions. Finally, we investigate the non-normality of the data by comparing our non-normal distributions to normal and skewed distributions with same means and variance-covariance matrices. As our base case we use the main asset classes stock, bond, real estate, and commodities with weights of 0.5, 0.3, 0.1, and 0.1 for the Bench portfolio.

6.1. Actual data yearly samples

In our first investigation, we take the asset performance as it occurs in the data. As always, we optimize the weights for our four methods on the input (past) data and then compute eight performance measures on the output (future) data. The measures used are mean return, return volatility, minimum return, Sharpe ratio, CEV3, Turnover, delta N (Δ_N), and the percentage of periods with extreme portfolio weights (%Extreme). Thus, we are simply showing the main results from Table 3 in a different format. We present the results in Figure 3 in the form of four spider web shaped graphs, one for each method and each spoke corresponding to a performance measure. All performance measures are being related to the minimum and maximum values across all runs in this section. This allows us to depict in the following graphs the achievement of each method in terms of a scale from 0% (minimum) to 100% (maximum). Those bounds are reported in Table 9.

[11] Bench often performs much like Equal, MinShortfall and MinSemivar tend to be similar to MinVar, MinVarBench and InformationRatio exhibit similar patterns to MaxSharpe, but are always weaker. We thus refrain from analyzing all those methods in detail.

Table 9. Minimum and Maximum Bounds for Different Performance Measures

The table reports the minimum and the maximum bound for 8 measures of portfolio quality: mean value, volatility, minimum value, Sharpe ratio, certainty equivalent for power utility function with relative risk aversion parameter of 3 (CEV3), portfolio turnover, the value of the delta N statistic, and the percentage of the extreme portfolio weights (%Extreme).

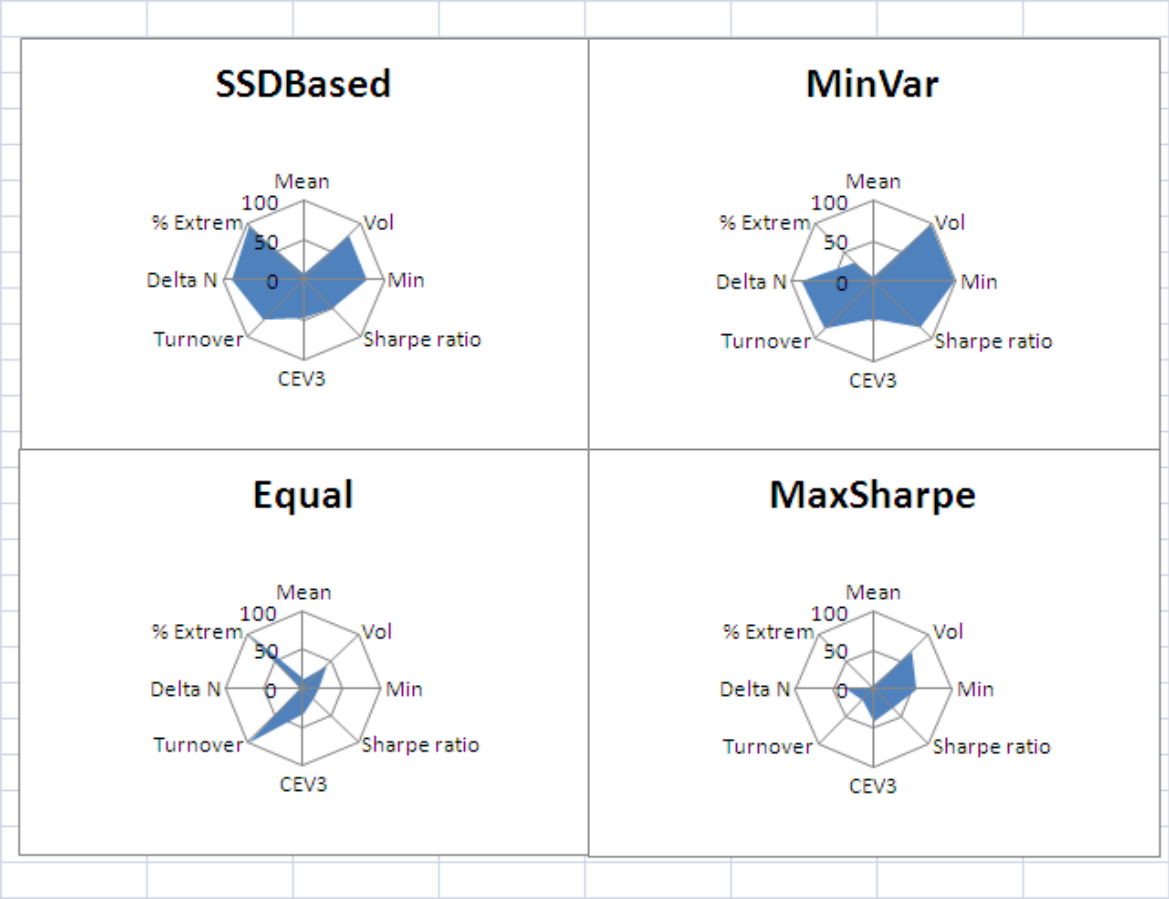
	Mean	Vol	Min	Sharpe ratio	CEV3	Turnover	Delta N	% Extreme
Min	**0.0897**	**0.0625**	**-0.3251**	**0.3683**	**0.0431**	**0.0000**	**2.0000**	**0.0000**
Max	**0.1810**	**0.1787**	**-0.0151**	**0.9369**	**0.1349**	**1.1571**	**16.8100**	**0.8875**

Now, we express each methods achievement such that 100% is always a good achievement (be that high mean or low volatility) and 0% is a poor achievement. In Figure 3 we can then appreciate the overall performance of any method by the size of the polygon in the spider web; we find that SSDBased and MinVar appear to cover wider areas than Equal and MaxSharpe.

SSDBased performs well on turnover, delta N, and the percentage of extreme weights. Mean returns are rather low but this holds for all four methods when using actual data. Performance in terms of volatility and minimum is solid while CEV3 and Sharpe ratio only give lowly performance. MinVar is a strong competitor with many dimensions with good performance. Naturally, in terms of volatility and minimum, MinVar performs well. The portfolio weights are often extreme however. It is surprising that MinVar does not do worse in terms of mean return which one would expect to be lower than for the other methods. This is explained by the bond portfolio which happens to have very high returns at low volatility. Equal has no turnover by design and thus also no probability of extreme weights. Its mean performance is on par with other methods but all other measures are not impressive. The performance of Bench is even worse and shows a shape almost identical to Equal. MaxSharpe performs very poorly on turnover and has many extreme portfolios. The performance in terms of volatility and minimum is middling at best and even on the dimension of Sharpe Ratio itself, it cannot compete well.

Figure 3. Performance of Four Methods based on the Actual Data

We present 8 performance measures, Mean, Volatility, Minimum, Sharpe ratio, CEV3, Turnover, Delta N, and % Extreme weights for the methods SSDBased, MinVar, Equal, and MaxSharpe. All performance measures are scaled so that 100 is the best performance across all runs in this section and 0 the worst. The actual data are used as historically observed.



6.2. Estimation error in bootstrapped yearly samples

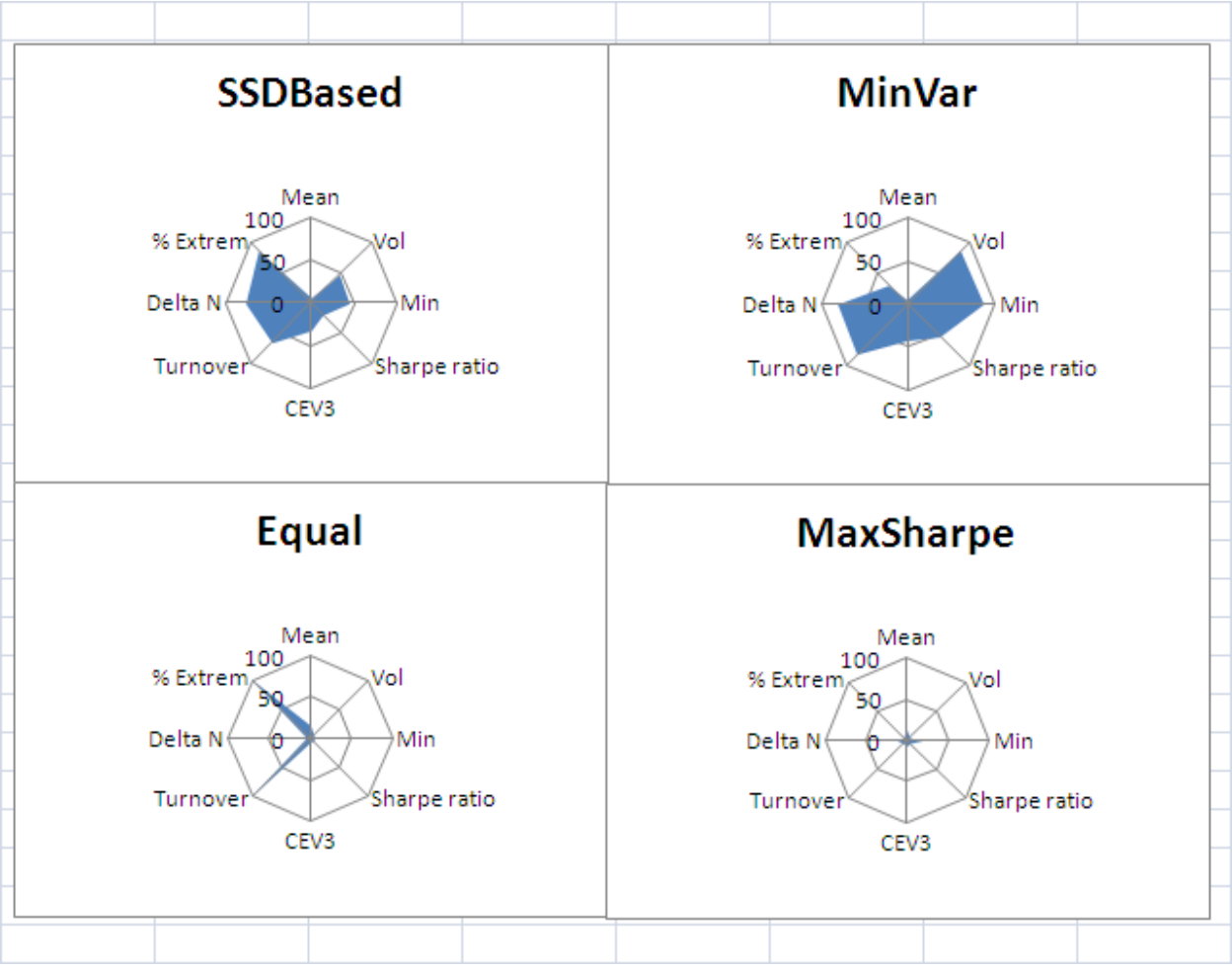
Next and more interesting, we investigate in Figure 4 the case where the input data for the portfolio choice and also the output data for the performance assessment are bootstrapped from their respective year's data, that is, output data is one year later than input data.[12] Essentially, we are now considering estimation error which comes from the fact that our yearly data is only one draw from the underlying distribution. Thus, we draw for each year's worth of data, be it input (past) or output (future) data, the same number of returns with replacement where we always draw a complete day as to not destroy the cross-sectional patterns in the original data. The bootstrap generates 100 such samples and we compute our performance measures for each of the 100 samples and average them. We see that all

[12] The intermediate cases of bootstrapping only the input or only the output data do not add much information.

methods suffer somewhat – indicating that the actual data corresponds to a draw where all methods did relatively well. SSDBased can maintain much of its good performance in terms of turnover, extreme weights, and delta N. The remaining dimensions worsen slightly. MinVar performs slightly worse on most dimensions but still impressively. Equal has perfect turnover and extreme weights scores by design which comes however at the cost of disastrous ratings on all other criteria. The Sharpe Ratio performs even worse in the bootstrapped data and is losing on all dimensions to the point where its polygon is almost invisible.

Figure 4. Performance of Four Methods based on the Bootstrapped Actual Data

We present 8 performance measures, Mean, Volatility, Minimum, Sharpe ratio, CEV3, Turnover, Delta N, and % Extreme weights for the methods SSDBased, MinVar, Equal, and MaxSharpe. All performance measures are scaled so that 100 is the best performance across all runs in this section and 0 the worst. The actual data are 100 times bootstrapped with replacement. Performance measures are averaged over the 100 bootstraps.



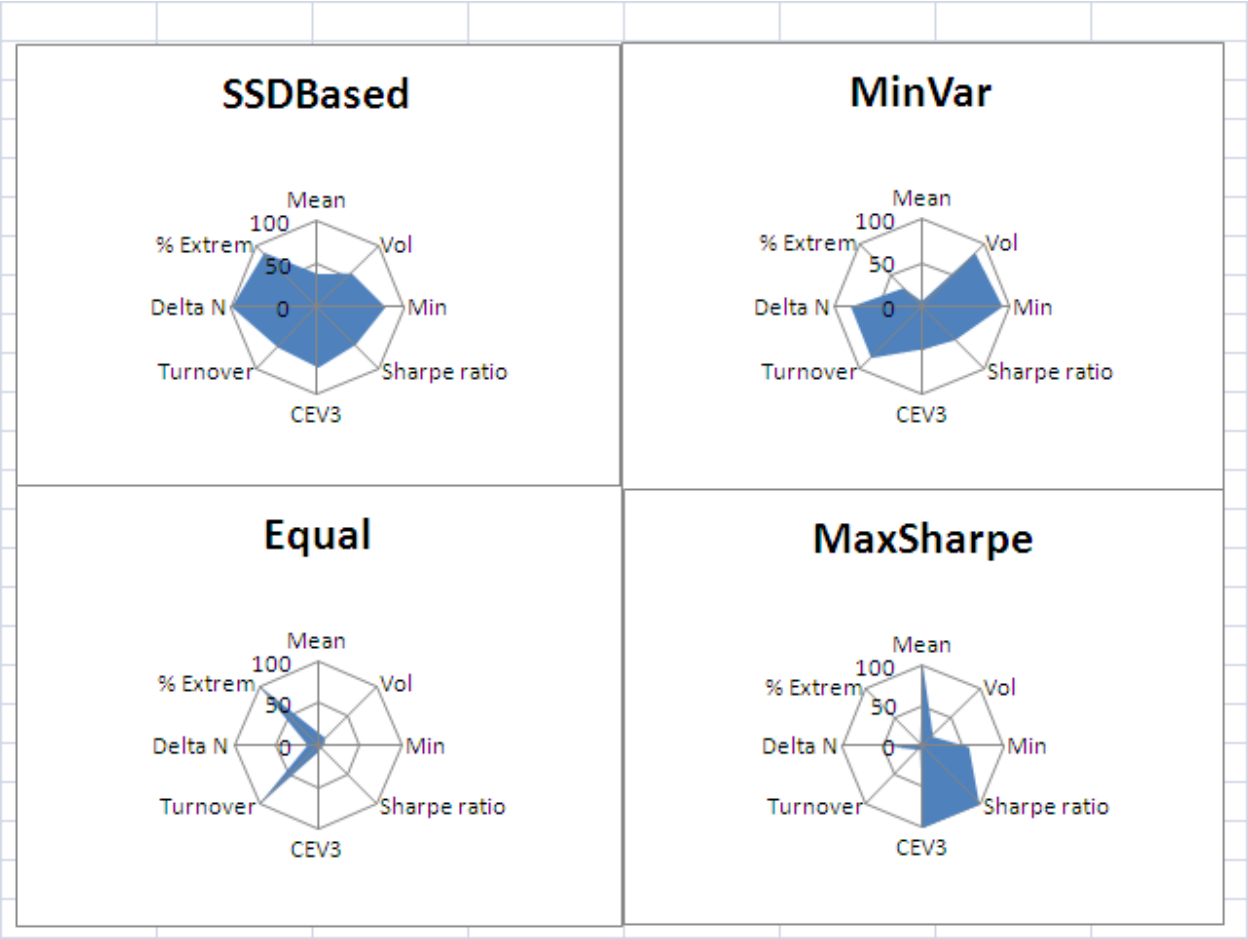
It is interesting to note what happens when we eliminate the last two years from the sample, thus creating our portfolios the last time in 2006 and assessing the performance the last time in 2007. All methods fare somewhat better once we eliminate the crisis since none of the methods is perfectly capable of taking the crash returns into account.

6.3. No time-variation in bootstrapped yearly samples where input and output come from the same year

In Figure 5, we investigate the case where both the input and the output data are bootstrapped from each year's data. The innovation is that we destroy the time variation in distributions as, for each year, the input data (for the portfolio choice) and the output data (for the performance measurement) are drawn from identical distributions. Thus, there is no time-variation in the samples, but there is still estimation error since the samples are bootstrapped from the yearly data. SSDBased improves considerably in performance while MinVar remains unchanged. The performance of Equal remains poor. The MaxSharpe performs now better in terms of mean, CEV3, Sharpe ratio, and (less so) minimum. This shows that MaxSharpe cannot handle time varying distributions very well. However, as such variation in the data from an estimation period to a performance period seems very prevalent; it makes the MaxSharpe a problematic choice for portfolio allocation.

Figure 5. Performance of Four Methods based on Bootstrapped Data where Input and Output Data are coming from the same Yearly Distribution

We present 8 performance measures, Mean, Volatility, Minimum, Sharpe ratio, CEV3, Turnover, Delta N, and % Extreme weights for the methods SSDBased, MinVar, Equal, and MaxSharpe. All performance measures are scaled so that 100 is the best performance across all runs in this section and 0 the worst. The actual data are 100 times bootstrapped with replacement. Performance measures are averaged over the 100 bootstraps. The same yearly data are used for the input (past) and output (future) data.



6.4. Normal distributions and skewed distributions

Our next variation is to continue with the setup of Section 6.3 and to draw from normal distributions with same means and variance-covariance matrices as in the yearly distributions. We find an improvement for all methods which indicates that all methods work better for normally distributed returns. Interestingly, the improvement for MaxSharpe is much larger than for the other three methods. MaxSharpe now performs much better in terms of mean, minimum, Sharpe ratio, and CEV3. Even volatility and delta N improve but the better performance of MaxSharpe depends intimately on using normally distributed returns.

Finally in this setting, we introduce left-skewed and right-skewed distributions based on the yearly normal samples. The skewness of originally bootstrapped real estate returns is about zero depending on the simulation run. We next normalize the joint distribution as above. For the left-skewed distribution, we then take away the real estate returns larger than 0.02 (some sixth of the sample) and replace them with negative returns of same size. Finally, we adjust mean and volatility to match the original values again. The resulting left-skewed distribution has skewness of about -0.3 depending on the simulation run. We similarly eliminate negative returns below -0.02 and replace them with positive returns of same size and rescale. This creates a right-skewed distribution which is a mirror image of the left-skewed distribution. There are virtually no changes as a result of introducing skewed distributions when compared to using normally distributed returns.

In conclusion, we argue that much is to be said in favor of the SSDBased method which performs well on a number of dimensions. It is the only method which delivers portfolios which are acceptable to a wide range of risk-averse investors as opposed to other methods which optimize narrowly in favor of particular utility functions. MinVar and Equal work best in situations where all assets have broadly the same means and volatilities – a situation which is not very realistic as one often allocates across stocks and bonds where the latter tend to have much tighter distributions. MinVar naturally also works well if the least risky asset has superior performance as it tends to load on that asset which happens to be the case in our data as bonds exhibit returns almost as high as the stock market but much lower volatility. The MaxSharpe method performs rather poorly unless the data are normally distributed.

7. Concluding comments

Most criteria for portfolio selection require an assumption on investor preferences or on the form of the return distribution. We propose using second-order stochastic dominance to rank portfolios, since this criterion is more general and can be applied to all situations with investors having increasing and concave utility functions. Indeed, all such risk-averse investors will prefer a second-order dominating distribution to a dominated one.

With *in-sample* analysis, it is typically possible to exploit knowledge of the data to find portfolio weights such that the resulting portfolio dominates a specified benchmark. A more interesting empirical question is whether one could find a way to determine portfolio weights using in-sample data such that the resulting portfolio dominates the benchmark *out-of-sample*.

Investigating that question, we propose an SSD-based portfolio choice approach. The portfolio weights are chosen such that the SSD test statistic of Davidson (2008) is maximized in-sample. We then test the performance of that approach out-of-sample. Using 21 years of daily returns on four asset classes (stocks, bonds, real estate, and commodities), we show that this approach significantly outperforms a benchmark portfolio out-of-sample where the benchmark is intended to proxy for a typical pension fund portfolio. Moreover, the SSD-based approach is also superior to other portfolio choice techniques, such as mean-variance-related portfolios (maximum Sharpe ratio, maximum Information ratio, and the minimum variance portfolio which matches the benchmark mean returns) and equally-weighted portfolios. There is a second group based on SSD-related portfolio choice techniques (minimum variance, minimum semi-variance, and minimum expected shortfall portfolios), which deliver dominance results similar to the SSD-based portfolio in terms of stochastic dominance over the benchmark. Together with the SSD-based portfolio, these portfolios form the best performing group. However, these alternative portfolios have lower mean returns than the SSD-based portfolio and are less diversified. While the SSD-related portfolios often dominate the benchmark and never are being dominated by the benchmark, the worst case scenarios for these portfolios are situations in which those portfolios and the benchmark lie in the same dominance class, meaning that there are some investors that might prefer one to another and other investors might reverse that choice.

We consider alternative measures of portfolio quality, such as simple return mean and volatility, minimum, Sharpe ratio, certainty equivalent, turnover, and percentage of extreme portfolio weights generated during 20 out-of-sample periods. The SSD-based portfolio

performs admirable along all these dimensions and the other SSD-related portfolios perform well, too. We also report results for the equally-weighted portfolio. In our tests, this portfolio choice alternative is inferior to the SSD-related portfolios in terms of out-of-sample dominance; however, it does improve upon the benchmark in a number of cases.

In contrast, the portfolio with the minimum variance and a mean restricted to be close to the in-sample mean of the benchmark, the maximum Sharpe ratio portfolio, and the maximum Information ratio portfolio with respect to the benchmark generally perform poorly out-of-sample and are sometimes dominated by the benchmark. This has considerable practical relevance, since portfolio choice based on the maximum Sharpe ratio appears popular in practice. The poor performance of those approaches in our tests seems due to both their ignoring higher moments and the rather unstable and extreme weights found by the in-sample optimization. They perform rather badly also with respect to portfolio return volatility and turnover. Our simulation exercise indicates that those mean-variance-related portfolio choice approaches perform nicely only if the returns are normally distributed and the distributions are not time varying. With reflection, those issues are not surprising. Nevertheless, these approaches are the only methods that actually manage to perform significantly worse than the random portfolio in terms of the second-order stochastic dominance over the benchmark portfolio.

Appendix A: Bootstrap procedure of Davidson (2008)

In this appendix, we briefly summarize the main steps of the bootstrap procedure developed in Davidson (2008). The summary is based on section 7 of Davidson (2008). The null hypothesis of the underlying test is that distribution A does not dominate distribution B. The distributions A and B are correlated, and the corresponding samples have an equal number of observations N. The observations from A and B are, thus, paired in couples ($y_{i,A}$, $y_{i,B}$).

1. The z-interval from the interior of the joint support of the distributions A and B is chosen such that there is at least one point in each sample that is above the maximum z and at least one below the minimum z.
2. The dominance functions $D_A^2(z)$ and $D_B^2(z)$ are computed for all values of z as in Equation (2). If for some z $D_A^2(z) > D_B^2(z)$, the algorithm stops and the non-dominance of A cannot be rejected.
3. The minimum test statistic t^* is computed as in Equation (4) based on $t(z)$ from Equation (3). The corresponding level of z where the minimum is attained is denoted z^*.
4. Since the observed frequencies of realized returns for each tested portfolio do not necessarily match with the probabilities of these returns under the null of non-dominance, one needs first to estimate those probabilities, and then use them to bootstrap from the observed return distribution. A relevant set of probabilities p_i for drawing each pair of observations ($y_{i,A}$, $y_{i,B}$) under the null of non-dominance of A is the solution of the following Lagrange-multiplier problem:

$$\sum_i n_i \log p_i + \lambda\left(1 - \sum_i p_i\right) - \mu\left(\sum_i p_i \left((z^* - y_{i,B})_+ - (z^* - y_{i,A})_+\right)\right), \tag{A.11}$$

where n_i is a number of pairs equal to ($y_{i,A}$, $y_{i,B}$) in the original samples A and B, n_i=1 for all i if all pairs are unique, λ is a Lagrange multiplier corresponding to a constraint that the probabilities sum to unity, μ is a Lagrange multiplier corresponding to a constraint that the dominance functions of A and B computed at z^* are equal, and $(z^* - y_{i,K})_+ = \max(z^* - y_{i,K}, 0)$, K = A, B.

5. The weighted dominance functions $\tilde{D}_K^2(z)$, K=A,B for all levels of z are constructed.

$$\tilde{D}_K^2(z) = \sum_{i=1}^{N} p_i (z - y_{i,K})_+ \qquad \text{(A.12)}$$

If $\tilde{D}_A^2(z) < \tilde{D}_B^2(z)$ for all z except of z^*, step 6 is omitted.

6. The value z^* is replaced by z_*, at which the difference $\tilde{D}_B^2(z) - \tilde{D}_A^2(z)$ is minimized. Steps 4 through 6 are repeated until the condition at step 5 is satisfied.
7. The M = 10,000 bootstrapped samples of A and B are constructed by randomly drawing with replacement the paired observation ($y_{i,A}$, $y_{i,B}$) with unequal probabilities p_i.
8. For each of M bootstrapped samples, the corresponding minimum test statistic t_j^* is computed (j =1,..,M) as in Equation (4).
9. The bootstrapped p-value is the proportion of t_j^* which are larger than the initial value t^*. The null hypothesis of non-dominance is rejected if the bootstrapped p-value is sufficiently small.

Appendix B: Variance and covariance of the dominance functions in the Davidson (2008) test

The values of the Davidson (2008) test statistic are computed for each of the chosen levels of a threshold z as shown in Equation (3). Implementing this equation requires estimation of the variances and covariance of the corresponding dominance functions. The estimates can be obtained using the original data sample as follows:

$$\hat{Var}(D_K^2(z)) = \frac{1}{N}\left(\frac{1}{N}\sum_{i=1}^{N} \max(z - y_{i,K}, 0)^2 - D_K^2(z)^2 \right),\ K = A, B, \tag{A.13}$$

$$\hat{Cov}(D_A^2(z), D_B^2(z)) = \frac{1}{N}\left(\frac{1}{N}\sum_{i=1}^{N} \max(z - y_{i,A}, 0) \cdot \max(z - y_{i,B}, 0) - D_A^2(z) \cdot D_B^2(z) \right), \tag{A.14}$$

where samples A and B are required to have the same number of observations N, and $y_{i,K}$ is the i-th observation in sample K.

Appendix C: The algorithm of Rubinstein (1982) to generate random portfolio weights

This appendix summarizes the algorithm of Rubinstein (1982) to generate random vectors uniformly distributed on the surface of a given region. The algorithm is derived from a more general case of the acceptance-rejection method, in which first the random values are drawn from a uniform distribution, and then only those are accepted, that satisfy the constraints.

For the purpose of this paper, we need to generate a vector of s random portfolio weights (w_i, $i = 1,\ldots,s$) lying between zero and one (including the edges), subject to a constraint that they sum up to unity. The algorithm proceeds in two steps.

1. Generate s random variables y_i ($i = 1,..,s$) from the exponential distribution with a mean value of one, where we use the built-in function of MATLAB.
2. Scale the generated random variables $_{yi}$ by their sum, in order to obtain the desired random portfolio weights:

$$(w_1,...,w_s) = \left(\frac{y_1}{\sum_{i=1}^{s} y_i},...,\frac{y_s}{\sum_{i=1}^{s} y_i} \right) \tag{A.10}$$

Reference

Abhyankar, A., Ho, K.-Y., Zhao, H., 2005. Long-run post merger stock performance of UK acquiring firms: A stochastic dominance perspective. Applied Financial Economics 15, 679-690.

Anderson, G., 1996. Nonparametric tests of stochastic dominance in income distributions. Econometrica 64, 1183-1193.

Barrett, G.F., Donald, S.G., 2003. Consistent tests for stochastic dominance. Econometrica 71, 71-104.

Barry, C.B., 1974. Portfolio analysis under uncertain means, variances, and covariances. Journal of Finance 29, 515-522.

Brown, S., 1979. The effect of estimation risk on capital market equilibrium. Journal of Financial and Quantitative Analysis 14, 215-220.

Cumby, R.E., Glen, J.D., 1990. Evaluating the performance of international mutual funds. Journal of Finance 45, 497-521.

Davidson, R., 2008. Testing for restricted stochastic dominance: Some further results. Working paper, McGill University.

Davidson, R., Duclos, J.-Y., 2000. Statistical inference for stochastic dominance and for the measurement of poverty and inequality. Econometrica 68, 1435-1464.

De Giorgi, E., 2005. Reward-risk portfolio selection and stochastic dominance. Journal of Banking and Finance 24, 895-926

De Roon, F.A., Nijman, T.E., Werker, B.J.M., 2001. Testing for mean-variance spanning with short sales constraints and transaction costs: The case of emerging markets. Journal of Finance 56, 721-742.

DeMiguel, V., Garlappi, L., Uppal, R., 2009. Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? Review of Financial Studies 22, 1915-1953.

Fishburn, P.C., 1977. Mean-risk analysis with risk associated with below-target returns. The American Economic Review 67, 116-126.

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

Glen, J., Jorion, P., 1993. Currency hedging for international portfolios. Journal of Finance 48, 1865-1886.

Gonzalo, J., Olmo, J., 2008. Testing downside risk efficiency under market distress. Working paper, Universidad Carlos III De Madrid.

Han, Y., 2006. Asset allocation with a high dimensional latent factor stochastic volatility model. The Review of Financial Studies 19, 237-271.

Jagannathan, R., Ma, T., 2003. Risk reduction in large portfolios: Why imposing the wrong constraints helps. Journal of Finance 58, 1651-1683.

James, W., Stein, C., 1961. Estimation with quadratic loss. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability 1, 361-379.

Jorion, P., 1992. Portfolio optimization in practice. Financial Analysts Journal 48, 68-74.

Kan, R., Zhou, G., 2007. Optimal portfolio choice with parameter uncertainty. Journal of Financial & Quantitative Analysis 42, 621-656.

Kaur, A., Prakasa Rao, B.L.S., Singh, H., 1994. Testing for second-order stochastic dominance of two distributions. Econometric Theory 10, 849-866.

Kopa, M., 2009. An efficient LP test for SSD portfolio efficiency. Working paper, Charles University in Prague.

Kuosmanen, T., 2004. Efficient diversification according to stochastic dominance criteria. Management Science 50, 1390-1406.

Leshno, M., Levy, H., 2002. Preferred by "All" And preferred by "Most" Decision makers: Almost stochastic dominance. Management Science 48, 1074-1085.

Levy, H., 2006. Stochastic dominance investment decision making under uncertainty. Springer

Linton, O.B., Maasoumi, E., Whang, Y.-J., 2003. Consistent testing for stochastic dominance under general sampling schemes. Review of Economic Studies 72, 735-765.

MacKinlay, A.C., Pastor, L., 2000. Asset pricing models: Implications for expected returns and portfolio selection. Review of Financial Studies 13, 883-916.

Markowitz, H., 1952. Portfolio selection. Journal of Finance 7, 77-91.

Martellini, L., Urošević, B., 2006. Static mean-variance analysis with uncertain time horizon. Management Science 52, 955-964.

Martellini, L., Ziemann, V., 2010. Improved estimates of higher-order comoments and implications for portfolio selection. Review of Financial Studies 23, 1467-1502.

Meyer, T.O., Li, X., Rose, L.C., 2005. Comparing mean variance tests with stochastic dominance when assessing international portfolio diversification benefits. Financial Services Review 14, 149–168.

Michaud, R.O., 1989. The Markowitz optimization enigma: Is 'optimized' optimal? Financial Analysts Journal 45, 31-42.

Morningstar, I., 2007. Stock, bonds, bills, and inflation 2007 yearbook. Morningstar, Inc., Chicago.

NN, 2009. The P&I 1000. Pensions & Investments, 37, Issue 2.

Nolte, I., 2008. Stochastic dominance tests under test. Working paper, University of Konstanz

Ogryczak, W., Ruszczynski, A., 1999. From stochastic dominance to mean-risk models: Semideviations as risk measures. European Journal of Operational Research 116, 33-50.

Pastor, L., 2000. Portfolio selection and asset pricing models. Journal of Finance 55, 179-223.

Porter, R.B., 1974. Semivariance and stochastic dominance: A comparison. The American Economic Review 64, 200-204.

Post, T., 2003. Empirical tests for stochastic dominance efficiency. Journal of Finance 58, 1905-1931.

Post, T., Levy, H., 2005. Does risk seeking drive stock prices? A stochastic dominance analysis of aggregate investor preferences and beliefs. Review of Financial Studies 18, 925-953.

Rubinstein, R.Y., 1982. Generating random vectors uniformly distributed inside and on the surface of different regions. European Journal of Operational Research 10, 205-209.

Russell, W.R., Seo, T.K., 1980. Efficient portfolios by stochastic dominance. Southern Economic Journal 46, 877-882.

Scaillet, O., Topaloglou, N.L., 2005. Testing for stochastic dominance efficiency. Working paper, HEC, University of Geneva.

Stein, C., 1955. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability 1, 197-206.