Ensembles of Portfolio Rules

Federico Nardari[1] and Rainer A. Schüssler[2]

[1] *University of Melbourne*

[2] *University of Rostock*

January 14, 2023

Abstract

We propose a framework for combining portfolio rules while mitigating the impact of estimation error. Our main goal is to integrate heterogeneous rules that previously proposed combination methods cannot accommodate, enabling researchers and investors to leverage established and ongoing advances in portfolio choice. The proposed framework relies on the (pseudo) out-of-sample returns of the considered rules, thus avoiding estimation of the PRs' return moments. The optimal combination is determined by an ensemble approach that maximizes the utility generated jointly by the candidate rules while allowing for learning about the PRs' relative performance. Based on out-of-sample evaluations of over forty years, we document substantial utility gains for our approach compared to both individual rules and previously proposed combination strategies.

Keywords: Portfolio choice; Combination of estimators; Ensemble learning; Estimation risk

JEL classifications: G11, C10

1 Introduction

Over time, many ingenious portfolio rules (PRs) have been devised, both theory based and data driven. For a cross-section of risky assets, vastly different techniques have been proposed, in large part aiming to correct the empirical shortcomings of the seminal Markowitz (1952) mean-variance (MV) framework originating from estimation uncertainty: these contributions include sophisticated shrinkage approaches (see, e.g., Barroso and Saxena 2022), strategies that exploit economic restrictions implied by asset pricing models (see, e.g., MacKinlay and Pástor 2000), volatility timing strategies (see, e.g., Kirby and Ostdiek 2012), parametric portfolio policies (see, e.g., Brandt et al. 2009; DeMiguel et al. 2020) and approaches that exploit asset characteristics using machine learning techniques (see, e.g., Freyberger et al. 2020; Gu et al. 2020; Cong et al. 2022). In addition, DeMiguel et al. (2009) and Duchin and Levy (2009) show that the naïve $1/N$ rule, which avoids estimation error by ignoring sample information, outperforms many optimization-based rules in demanding out-of-sample (OOS) settings. Similarly, many portfolio rules have been put forward for optimal market timing, i.e., the allocation between an aggregate equity portfolio and a risk-free asset. While some of these rules use macroeconomic data and financial ratios (see, e.g., Rapach et al. 2010; Ferreira and Santa-Clara 2011; Dangl and Halling 2012; Johannes et al. 2014), others rely on forward-looking information from option prices (see, e.g., Pyun 2019), or exploit long-short return anomalies in the cross-section of stocks by using machine learning methods and shrinkage techniques (Dong et al., 2022).

Each of the above mentioned PRs (and any other PR) is defined by the information set it uses and by how it maps information into asset weights for a given investment universe. Different PRs have different virtues as well as limitations, and merging heterogeneous PRs is economically and statistically motivated. Economically, combinations of PRs might diversify across their idiosyncratic risks, chief among them estimation risk, akin

to the diversification across assets for investors with a concave utility function. Further, as different PRs use different information sets and/or different methods for processing information into portfolio weights, combinations of PRs might capture complementary aspects of the return generating process, which is particularly important in light of the notoriously low signal-to-noise ratio of asset returns. Statistically, PRs can be seen as estimators. In different contexts, combinations of estimators have been shown to be theoretically appealing and to perform well in empirical applications. Overall, there are valid reasons for combining PRs rather than relying on one particular PR and dismissing all alternatives from the outset.

The extant literature has, indeed, developed several combination approaches aimed to control estimation error and, consequently, improve OOS performance. However, existing approaches are applicable to a rather specific and limited set of PRs, typically within the MV or global minimum variance (GMV) frameworks, with the addition of the $1/N$ rule (see, e.g., Kan and Zhou 2007; Tu and Zhou 2011; Lassance et al. 2022; Kan et al. 2022). Moreover, in order to determine the optimal combination, they usually rely on specific distributional assumptions with respect to the generating process of asset returns. Consequently, existing combination strategies provide a rather narrow set of tools for diversifying estimation error and, hence, for improving asset allocation performance. To the best of our knowledge, there is no utility-based optimization framework for combining an arbitrary number of PRs that rely on heterogeneous information sets and/or vastly different methods for mapping information into asset weights while, at the same time, circumventing estimation of the PRs' return moments.[1] For instance, for a cross-section of assets there is currently no utility maximization framework for combining a shrinkage-based approach such as, e.g., the method of Barroso and Saxena (2022)

[1]In our proposed approach, we treat as given the mapping between information signals and *asset* weights that each candidate PR implies. As detailed below in Section 3.1, it is straightforward to back out the asset weights implied by combining the PRs.

with, say, a volatility timing strategy and with rules based on cross-sectional characteristics. Similarly, it is not obvious how to combine market timing rules that rely on point forecasts for the equity premium with rules based on density forecasts and/or with others that exploit cross-sectional characteristics. Further, existing combination methods can be seen as candidate PRs themselves and, hence, combined with other PRs. Existing approaches cannot entertain such additional layer. Finally, there is no optimization framework for adapting the combination to changing market environments, as particular (combinations of) PRs may outperform at certain points in time, while others may shine at other points in time. All in all, the extant literature does not allow to fully exploit the relative strengths of the many solutions previously proposed to portfolio choice problems.

Our study intends to fill these gaps. In particular, our main goal is to provide an overarching optimization framework for integrating heterogeneous PRs that alternative combination methods are unable to accommodate. For developing our framework, mitigating estimation error has been an important concern.[2] Our framework can be seen as an outer layer, in which candidate PRs, irrespective of their design, can be combined. As such, our framework enables researchers and investors to comprehensively leverage existing and ongoing scientific progress in asset allocation.

The investor in our framework is endowed with power utility preferences and has access to a library of candidate PRs. In each period, they choose a combination of PRs that

[2]Lack of methods to effectively limit estimation error might be the main reason why previous research has focused on combining only two PRs. Adding more PRs to the combination does not necessarily lead to empirical gains. As pointed out by Tu and Zhou (2011), "theoretically, if the true optimal combination coefficients are known, combining more than two rules must dominate combining any subset of them. However, the true optimal combination coefficients are unknown and have to be estimated. As more rules are combined, more combination coefficients need to be estimated and the estimation errors can grow. Hence, combining more than two rules may not improve the performance."

would have maximized their pseudo OOS utility.[3] While determining the combination of PRs which optimizes the investor's utility, the framework retains many appealing features. Although previously proposed combination strategies share some of those features, no other method we are aware of possesses all of them. Specifically, our approach:

i) Relies on the pseudo OOS returns of the candidate PRs. As the optimal combination of PRs is based on OOS utility gains, nothing more than the record of the PRs' assigned asset weights and the subsequent pseudo OOS returns are needed for implementing our approach. This enables combining dissimilar PRs, for example theory-based PRs and data-driven PRs,[4] while circumventing the problem of predicting moments of the PRs' returns. As a result, our setup involves estimating fewer parameters and, hence, reduces estimation risk.[5] However, we stress that the candidate PRs themselves may or may not use estimated (conditional or unconditional) moments of asset returns to form portfolios.

ii) Is an ensemble framework. Our approach assigns combination weights based on the realized pseudo OOS utility of the combined PRs rather than based on their in-

[3]We focus on economic utility in the objective function rather than on a statistical criterion. It is well known that statistical and economic evaluation criteria are not necessarily closely related. Leitch and Tanner (1991) show that precise forecasts in terms of statistical criteria such as the root mean squared error may translate into unprofitable portfolio allocations. Cenesizoglu and Timmermann (2012) corroborate this finding in an application to equity premium forecasts, establishing only a weak relationship between economic utility measures and statistical forecast accuracy.

[4]Similar in spirit to our idea of integrating theory-based and data-driven endeavors for asset allocation, in asset pricing, Grammig et al. (2021) developed an approach to unite theory-based and data-driven approaches (seen as "divergings roads").

[5]Alternatively, one could think of using the realized PRs' returns and maximize an objective function that depends on the expected moments of those returns. However, in such a two-stage approach, the estimated moments of the PRs' returns would be needed as inputs for choosing the combination weights. For power utility preferences, one would have to estimate at least the first four moments (when using a Taylor series expansion, as it is commonly done in the literature). With an increasing number of PRs, the two-stage approach would become more and more prone to estimation error due to the proliferation of parameters. Further, in our proposed framework, down-weighting older data can be applied to economic utility, thus directly focusing on the investor's objective. Instead, in a two-stage approach, down-weighting older observations would have to be applied indirectly via the estimated moments of the PRs' returns. Likewise, in our proposed framework, regularization to prevent overfitting can be applied directly to the combination weights, whereas, in a two-stage approach, regularization would have to be applied indirectly to the estimated moments. Overall, our proposed framework appears more directly focused on the investor's objective, less prone to estimation error and more scalable to an increasing number of candidate PRs than a two-stage approach.

dividually generated utility. As an analogy to building a sports team, our combination framework does not necessarily include the individually best players, but builds the best possible team. The ensemble view automatically takes the (possibly time-varying) inter-dependencies among the PRs' OOS returns into account. These inter-dependencies include correlations and higher-order co-moments. Our approach accommodates an arbitrary number of candidate PRs.

iii) Allows for adaptive learning. In our approach, profitability (or, realized utility) in the recent past might be emphasized compared to profitability generated in the more distant past by using a weighting factor. This enables adaptive learning about the optimal combination weights and allows for rapid shifts if empirically warranted. At the level of asset returns, Farmer et al. (2022) find short stretches of predictability for (aggregate) stock returns by a given predictor that are interspersed with long periods showing no evidence of predictability. Similarly, our modeling approach is designed to capture PRs, or combinations of PRs, that are successful locally in time.

iv) Does not assume a specific data generating process (DGP) for asset returns or for the PRs' returns. To determine the combination weights, we do not invoke any assumptions regarding the return generating process.[6] Hence, our combination framework can be viewed as a controlling instance: if a candidate PR is grossly misspecified and has nothing to contribute to the ensemble, it will not be selected to be part of the combination.

We apply our combination framework to two classic portfolio choice problems. The first one involves allocating across the 50 largest US stocks at a monthly frequency, and the second one involves allocating between the S&P 500 index and treasury bills at a monthly frequency. In both applications, we entertain a pool of established and cutting-edge candidate PRs. We compile libraries of heterogeneous PRs to enhance diversification benefits and choose PRs that allow for long evaluation samples.

[6]Notice, though, that the candidate PRs may or may not make assumptions with respect to the return process.

Based on OOS evaluations of over forty years, we find substantial utility gains from combining PRs. The utility generated by our combination is either higher than that of any candidate PR or approximately as high as that of the (ex-post) best-performing candidate PR. Our combination approach appears to outperform previously proposed alternatives as well. Further, we empirically show the virtues of the ensemble framework and of adaptive learning. Combination weights change rapidly over time, documenting that different (combinations of) PRs work well at different points in time. We carry out deeper analyses to shed light on the mechanisms at work for generating utility gains and on the potential from extending the pool of candidate PRs. These analyses reveal that our proposed combination method, by maximizing utility, chooses a combination of PRs that balances predictive power of asset returns and the capability of anticipating their variance; further, utility gains increase on average with the number of combined PRs, implying further room for improvement by increasing the number of candidate PRs.

The applications presented in our empirical work are meant as illustrations of our methodological framework to establish its expediency. We do *not* advocate any particular candidate PR and stress that other researchers or investors might prefer using our approach with alternative sets of candidate PRs. We see the main contributions of our study as methodological in nature. Namely, we provide a framework that: a) allows to fully exploit the relative merits of the multitude of solutions proposed for portfolio choice problems; b) enables to assess the incremental empirical merits (or, lack thereof) of newly proposed PRs. That said, in our empirical analysis we consider state-of-the-art PRs that nest established rules (e.g., GMV, MV or 1/N) as special cases. We are not aware of any study that empirically entertains combinations of heterogeneous cutting-edge PRs.

The remainder of the paper is organized as follows. Section 2 relates our work to the literature. Section 3 lays out our methodology, and Section 4 presents our two applications. Section 5 offers some concluding remarks.

2 Relation to the Literature

Our work relates mainly to two streams in the portfolio choice literature. First, our work shares common ground with combination approaches of PRs. Along this line, Kan and Zhou (2007), Tu and Zhou (2011) and Kan et al. (2022) developed theoretically optimal combination strategies to maximize expected OOS performance under estimation risk for MV portfolio choice problems. Kan and Zhou (2007) derived an optimal three-fund rule consisting of the risk-free asset, the sample tangency portfolio, and the sample minimum-variance portfolio, that maximizes expected OOS utility. Based on the intuition that one simple method and one sophisticated method might optimize the bias-variance tradeoff, Tu and Zhou (2011) combine this three-fund portfolio (and other sophisticated PRs) with the 1/N rule. Kan et al. (2022) explore the case when there is no risk-free asset available. Lassance et al. (2022) robustify the approach by Kan et al. (2022), considering OOS utility volatility in addition to the OOS utility mean.

The optimal combining rules derived by the works cited above have gleaned valuable analytical insights into portfolio construction under estimation error, relying on the assumption that asset returns are identically and independently multivariate normally distributed. In contrast, as a data-driven combination approach, our proposed framework does not invoke any assumptions regarding the return generating process. More importantly, our approach is not restricted to PRs of particular designs, such as, e.g., MV-based rules, and allows the combination of PRs that, given their heterogeneity, could not be combined with existing methodologies. Our method is *not* a competing approach to these works, but is complementary to them: combination methods themselves represent PRs and can be included as candidate PRs in our combination approach, given that they are applicable to the investment problem at hand. We include previously suggested combination methods like those proposed by Kan et al. (2022) in our empirical analysis and show that they produce enhanced performance when combined with additional PRs.

Paye (2012) considers combinations of PRs as one possible strategy to reduce estimation risk associated with MV approximations to the economic value of PRs under general utility specifications such as power utility. He finds that the combination of estimators can substantially reduce estimation risk. Paye (2012) determines the combination weights based on a resampling approach, assuming identically and independently distributed returns, and considers equal combination weights as an alternative. Although it is not the focus of our paper, we consider, among others, PRs based on MV approximations as candidate PRs for optimizing power utility. While MV approximations might be valuable since estimations of higher moments are not available in some settings, quadratic utility that underlies MV preferences has some counter-intuitive properties such as increasing absolute risk aversion. Hence, in addition to taking preferences about higher-order moments into account, evaluating PRs based on MV in a power utility framework is also desirable due to the more intuitive properties of power utility. Pettenuzzo and Ravazzolo (2016) propose combining predictive densities based on their weighted individual past performance, while we combine PRs based on their weighted jointly generated past performance.

While approaches such as those of Tu and Zhou (2011) and Kan et al. (2022) use the covariance matrix of returns for optimizing combination weights, we propose an approach that automatically takes the joint distribution of the PRs' returns into account for determining a convex combination of weights without having to estimate co-moments. Further, our flexible combination method might unfold its full strength when combining PRs that, due to their heterogeneity, generate substantially different allocations, since it adaptively learns about time variation in the relative performance of candidate PRs and about time-varying dependencies among them, thus adjusting combination weights based on local performance.

Overall, we push the boundaries of combinations of PRs by providing an optimization

framework that (i) is designed to mitigate estimation risk by avoiding estimation of the PRs' return moments; (ii) can combine PRs of any design; (iii) incorporates additional appealing features such as adaptive learning about the combination weights, an ensemble perspective for assigning combination weights, and direct focus on the investor's utility.

Second, and related to (i) above, our work is related to approaches that directly optimize utility instead of taking a two-stage approach with the need for estimated moments of returns. These include parametric portfolio policies (Brandt et al., 2009; DeMiguel et al., 2020), a boosting approach (Nevasalmi and Nyberg, 2021), a genetic programming approach (Liu et al., 2022), a subset combination approach (Maasoumi et al., 2022) and an approach that uses deep reinforcement learning (Cong et al., 2022). The above mentioned techniques involve optimizing economic utility for specific portfolio choice problems at the individual assets level, while our approach is about maximizing utility one level up by combining PRs. It is not obvious how the above methods can be extended to the combination of PRs. Our work is complementary to these approaches as well since they can be included as candidate PRs in our framework, given that they are applicable to the investment problem at hand.

3 Methodology

3.1 Basic structure

Suppose that we have a set of M candidate PRs at our disposal, indexed as $m = 1, \ldots, M$. To set the stage for the combination, let us pin down the ingredients of our setting. For a typical point in time s, each PR assigns weights to the N assets, indexed as $n = 1, \ldots, N$,[7] based on information observed through $s-1$, the date of portfolio construction. We denote

[7]In this paper, we consider only PRs that allocate across the same investment opportunity set. However, our framework allows for PRs that allocate across different investment opportunity sets with partial or no overlap of assets. For example, one PR could allocate across different stocks, while another PR could allocate across commodities.

the (exogenously) assigned asset weights of the m-th PR for time s as $\omega_{m,s}^{(-s)}$, where $\omega_{m,s}^{(-s)}$ is an $N \times 1$ column vector, $\left(\omega_{m,s,1}^{(-s)}, \ldots, \omega_{m,s,N}^{(-s)}\right)'$. The superscript $(-s)$ indicates that information revealed at time s are not available for determining the portfolio allocation at time $s-1$.

The $N \times 1$ column vector of asset gross returns measured over the period $[s-1:s]$ (that is, one month in our applications) is indicated as $\widetilde{\mathbf{R}}_s = \left(\widetilde{R}_{s,1}, \ldots, \widetilde{R}_{s,N}\right)'$, where $\widetilde{R}_{s,n} = 1 + \widetilde{r}_{s,n}$,[8] and $\widetilde{\mathbf{r}}_s = (\widetilde{r}_{s,1}, \ldots, \widetilde{r}_{s,N})'$. Then, the pseudo OOS gross return of the m-th PR at time s can be expressed as:

$$R_{m,s} = \omega_{m,s}^{(-s)'} \widetilde{\mathbf{R}}_s. \tag{1}$$

The investor's optimization problem is to maximize the conditionally expected utility of the portfolio (gross) return $R_{p,t}$ based on information through time $t-1$ as a function of the combination weights $\{w_{m,t}\}_{m=1}^{M}$ assigned to the PRs:

$$\underset{\{w_{m,t}\}_{m=1}^{M}}{\arg\max}\ \mathbb{E}_{t-1}\left[U(R_{p,t})\right] = \mathbb{E}_{t-1}\left[U\left(\sum_{m=1}^{M} w_{m,t} R_{m,t}\right)\right], \tag{2}$$

where $U(\cdot)$ denotes utility. We treat the combination weights as constant through time t. Hence, the combination weights that maximize the investor's conditional expected utility at a given point in time are the same for all previous points in time. We can thus rewrite (2) as an unconditional optimization problem. Suppose we are at time $t-1$ and have access to a record of pseudo OOS returns generated by the PRs, spanning the interval between τ and $t-1$. Then, we can replace the expected utility in (2) with its sample counterpart, i.e., the sum of period-by-period realized utilities. The optimization

[8] Depending on the portfolio choice problem at hand, the returns may be defined as raw (total) or excess returns.

problem, then, becomes:

$$\mathbf{w}_t^* = \underset{\{w_m\}_{m=1}^M}{\arg\max} \sum_{s=\tau}^{t-1} U\left(\sum_{m=1}^M w_m R_{m,s}\right), \tag{3}$$

where $\mathbf{w}_t = (w_{1,t}, \ldots, w_{M,t})'$ and $\mathbf{w}_{\tau:t}^* = \mathbf{w}_t^*$. This unconditional formulation of the optimization problem bypasses the need for estimating (co-)moments of the PRs' returns, similarly to the framework of Brandt et al. (2009).[9]

If there is some persistence in economic states, more recent data will likely embed more relevant predictive information than older ones, since they stem from a more similar market or economic environment. To entertain such plausible economic dynamics, we allow realized joint utilities to receive different weights in the optimization. Specifically, we maximize the weighted past performance jointly generated by the PRs:

$$\mathbf{w}_t^* = \underset{\{w_m\}_{m=1}^M}{\arg\max} \sum_{s=\tau}^{t-1} \alpha^{t-1-s} \cdot U\left(\sum_{m=1}^M w_m R_{m,s}\right), \tag{4}$$

subject to

$$\sum_{m=1}^M w_m = 1; \quad w_m \geq 0, \quad m = 1, \ldots, M, \tag{5}$$

where α denotes a (fixed) forgetting factor for weighting past profitability, and the restrictions (5) impose a convex combination of the candidate PRs. By allowing for exponential down-weighting older performance (and repeating the optimization at each point in time), we select the combination weights in an adaptive manner. We, thus, include the possibility of learning about the relative strengths of the candidate PRs over specific stretches of time, enabling more rapid weight changes than in the standard unweighted formulation in (3). We will discuss the forgetting factor and the weight constraints in more detail in

[9] As discussed earlier in Section 2, while we share common ground with Brandt et al. (2009) regarding this aspect, our framework is considerably different from theirs, since our objective is to allocate combination weights across PRs rather than estimating coefficients associated with asset characteristics and mapping those coefficients into individual asset weights.

Sections (3.3) and (3.4), respectively.

For an investor with power utility preferences,[10] we can state the optimization problem (4) more specifically as:

$$\mathbf{w}_t^* = \underset{\{w_m\}_{m=1}^M}{\arg\max} \sum_{s=\tau}^{t-1} \alpha^{t-1-s} \cdot \frac{\left(\sum_{m=1}^{M} w_m R_{m,s}\right)^{1-\gamma}}{1-\gamma}, \tag{6}$$

where γ ($\gamma > 0, \gamma \neq 1$) denotes the relative risk aversion coefficient.[11] Note that utility and certainty equivalent returns are interchangeable in this optimization framework since the latter is a monotonic transformation of the former.

For certain purposes such as implementing trades and calculating transaction costs, it is necessary to know the asset weights that result from combining the PRs. With the optimized combination weights in hand, we can back out the implied weights of the N assets. These weights ω_s^* are linear combinations of the asset weights determined by the PRs (summarized in matrix $\mathbf{\Omega}_s$) and the optimized combination weights $\mathbf{w}_s^*$:

$$\underset{[N\times 1]}{\omega_s^*} = \underset{[N\times M]}{\mathbf{\Omega}_s} \cdot \underset{[M\times 1]}{\mathbf{w}_s^*}, \tag{7}$$

[10] By assuming power utility preferences at the level of combining PRs, preferences about higher-order moments and tail risk properties are taken into account. This holds also true for the case where candidate PRs in the library do not optimize power utility but, for example, use MV approximations or the allocation does not rely on any optimization framework at all. If we were to require candidate PRs to optimize power utility in the first place, we would dismiss a large portion of promising PRs from the outset. PRs that have not been designed to maximize power utility preferences might nonetheless contribute to the ensemble.

[11] We note that power utility fails to exist if the gross return approaches zero. That is, $U \longrightarrow -\infty$ if $R \longrightarrow 0$. The PRs we use in our empirical work avoid extreme returns and, hence, this is empirically not a concern. To theoretically ensure that power utility exists, we had to restrict to candidate PRs that put appropriate constraints on the asset weights.

where

$$\Omega_s = \begin{pmatrix} \omega^{(-s)}_{m=1,s,n=1} & \cdots & \omega^{(-s)}_{m=M,s,n=1} \\ \vdots & \ddots & \vdots \\ \omega^{(-s)}_{m=1,s,n=N} & \cdots & \omega^{(-s)}_{m=M,s,n=N} \end{pmatrix}.$$

The usefulness of backing out the individual assets weights can easily be seen when the positions for a certain asset implied individually by the candidate PRs are partly or fully offsetting each other. An execution desk trades the individual asset positions as implied by the combination, rather than the implied positions of different PRs on their own, thereby saving transaction costs.

3.2 Our combination framework as a stacking algorithm

Our proposed combination (6) can be classified as a stacking algorithm. Stacking is a well-studied meta-learning algorithm for combining estimators in the machine learning and statistics literature (Wolpert, 1992; LeBlanc and Tibshirani, 1996; Breiman, 1996; Yang, 2001; Van der Laan et al., 2007; Polley and Van Der Laan, 2010). Stacking algorithms were developed to minimize cross-validated risk defined by some statistical criterion. We adapt this method to maximizing cross-validated utility rather than optimizing some statistical loss criterion and extend it to accommodate exponential down-weighting of older performance for obtaining combination weights based on the local performance of (combinations of) PRs.

Stacking is an ensembling method; that is, it assesses the cross-validated risk/utility of the combined candidate estimators (here, the PRs) rather than assessing their risk/utility from a stand-alone perspective. Hence, combination weights assigned according to (6) are based on an ensemble perspective, implicitly taking time-varying inter-dependencies among PR returns into account. With power utility preferences, the entire joint distribution of PR returns is automatically used in (6) for maximizing combination weights.

Another important feature of stacking is that it uses cross-validation to avoid overfitting. Our combination is based on pseudo OOS returns. To accommodate the time-series structure of the data, standard K-fold cross-validation cannot be applied. Our approach is akin to leave-one-out cross-validation by omitting information revealed at time s for portfolio construction at time $s-1$; see (1). Figure 1 illustrates the generic mechanism of leave-one-out-cross-validation. In our context, at each point in time and for a given PR, the blue dots represent the information used for next period's portfolio allocation, and the red dots represent the implied pseudo OOS (gross) returns. Our approach relies on maximizing the utility generated by the red dots.

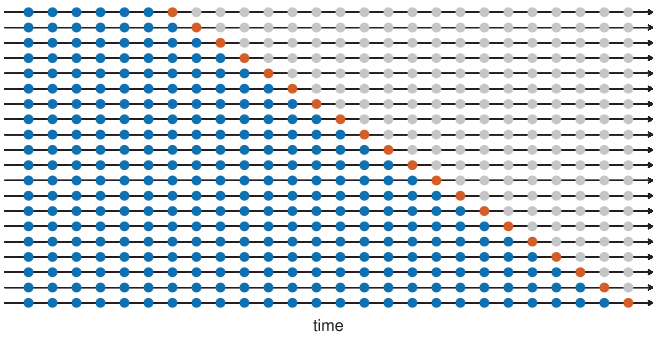

Figure 1: Schematic illustration of leave-one-out-cross-validation. The illustration is adopted from Hyndman and Athanasopoulos (2018).

If we were to include information revealed at time s for allocation at time $s-1$, the resulting returns would be in-sample returns. In such a setting, the combination would typically assign the entire weight to the PR with the highest in-sample returns. However, PRs with high in-sample returns might generate poor OOS results due to overfitting.

Stacking is a genuine combination rather than a selection method. This means that, even asymptotically, positive combination weights can be spread among different PRs rather than being assigned to the most successful PR in the library. This feature is appealing for the realistic case where none of the candidates in the library captures the true data generating process. If a single candidate PR, however, dominates any possible

combination, such PR will get the entire weight. Hence, selection is nested as a special case.

Although stacking does not impose any restrictions on the combination weights per se, convex combinations of estimators were found to provide greater stability of the final estimator (see, e.g., Breiman, 1996; Van der Laan et al., 2007).

Stacking algorithms have a strong statistical foundation. Under certain conditions, Van der Laan et al. (2007) established their asymptotic oracle performance, which means that the learning algorithm will perform asymptotically exactly as well (with respect to the defined evaluation criterion) as the best possible ex-post choice for a given dataset among the set of weighted combinations of the estimators. Beyond these theoretical results, learning algorithms based on stacking were shown to be adaptive and robust estimators for small samples in both artificial and real datasets (Wolpert, 1992; Breiman, 1996; LeBlanc and Tibshirani, 1996; Van der Laan et al., 2007; Polley and Van Der Laan, 2010). In most cases, they perform as well or even better than the ex-post best candidate estimator. As a stacking algorithm, our combination framework relies on a methodology with excellent statistical properties.

3.3 The forgetting factor

The exponential forgetting factor $\alpha \leq 1$ in (6) emphasizes the recent history of past performance. In our empirical work, we adaptively choose the value of α from the grid $\mathscr{S}_\alpha = \{0.90{:}0.01{:}1.00\}$. The lower the value of α, the more we down-weight performance in the more distant past. For example, when working with monthly data, if $\alpha = 0.99$, economic utility three years ago receives approximately 70% as much weight as economic utility last month. We take $\alpha = 0.90$ as the lower boundary of the grid since this value implies extremely fast forgetting: if $\alpha = 0.90$, utility three years ago receives only about 2% as much weight as utility last period. The effective window size is $1/(1-\alpha)$ and, hence,

10 months in case of $\alpha = 0.90$.[12] The upper bound $\alpha = 1$ implies no down-weighting of older data and, hence, standard recursive window estimation is nested as a special case.

FLEXPOOL denotes the combination where the value of α is determined in each period from the grid $\mathscr{S}_\alpha$. In the empirical analysis, we consider two additional benchmark combinations: the first one is STATPOOL, where we set $\alpha = 1$. The second one assigns equal weights to the PRs, irrespective of their past performance.

By allowing for down-weighting older performance, the combination weights can rapidly adjust to changing environments, if empirically warranted. Another advantage of the forgetting factor modeling is that changing dynamics are captured parsimoniously using only one parameter. This makes the approach less prone to estimation error than parameter-rich alternatives such as regime-switching models. For each point in time, we choose the optimal time-dependent value α_t from the grid $\mathscr{S}_\alpha$ as the one which has produced the highest utility in the past from τ^* to $t-1$:

$$\alpha_t^* = \underset{\alpha \in \mathscr{S}_\alpha}{\arg\max} \sum_{s=\tau^*}^{t-1} U\left[\left(\mathbf{w}_{t-1}^*(\alpha)\right)' \mathbf{R}_s\right], \tag{8}$$

where $\tau^* = \tau + \tau_0$, and τ_0 denotes the number of observations set aside for initial optimization of the combination weights, $\mathbf{R}_s = (R_{1,s}, R_{2,s}, \ldots, R_{M,s})'$, and $\mathbf{w}_{t-1}^*(\alpha)$ denotes the optimal combination weights according to (6), conditional on a given value of α. Note that we are using down-weighting when maximizing the combination weights in (6) for a given value of α. However, we do not use down-weighting for choosing between different values of α based on recursive evaluation in (8).[13]

Rolling windows can be seen as an ad hoc alternative to exponential down-weighting

[12] Note that $\sum_{s=0}^{\infty} \alpha^s = \frac{1}{1-\alpha}$ for $\alpha < 1$.

[13] Using recursive window estimation for choosing between different values of the forgetting factor follows, among others, Beckmann et al. (2020) and Adämmer and Schüssler (2020), and is guided by our endeavor to keep the framework as parsimonious as possible. Adding another forgetting factor for choosing between different values can easily be accomplished (Bernaciak and Griffin, 2022).

for accommodating structural breaks by allowing for rapid changes in the combination weights. Exponential down-weighting with forgetting factors estimated from the data, however, is a more sophisticated and robust choice.[14]

3.4 Weight restrictions and additional regularization

As outlined above, one motivation for imposing a convex combination of PRs is stability of the stacking algorithm. Another motivation is that we wish to ensure that any imposed restrictions on the assets weights at the level of the candidate PRs (e.g., no short sales, restrictions on sector weights, etc.) also hold at the level of the combined PRs.

Although our proposed combination approach uses pseudo OOS returns and is parsimoniously parameterized, estimation risk of the combination weights could nevertheless be a concern in finite samples. There is no guarantee that the optimized combination weights will outperform simple benchmarks such as equally weighted PRs. In our empirical work with moderate numbers of candidate PRs (five in our first application and six in our second application), estimation error of the combination weights does not appear to be a major concern.[15] In our applications, we do not impose any restrictions in addition to convex combination weights. Nevertheless, in applications with higher numbers of candidate PRs, it might be beneficial to impose additional regularization on the combination weights. For example, one could choose time-varying subsets of equally-weighted PRs. In each period, we would select a subset of $k \leq M$ PRs and assign equal weights $1/k$ to the PRs of the subset and discard the remaining PRs for this period.[16] Technically, this involves adding an ℓ_0-constraint (9) on the combination weights and enforcing the

[14] Giraitis et al. (2013) find exponential down-weighting with a data-driven forgetting factor to be the most robust choice for accommodating structural change in time series across extensive simulations and empirical exercises.

[15] In both applications we will find that utility rises on average with the number of included PRs.

[16] In the context of forecast combinations, equally-weighted subsets were found to perform well in many studies (see, e.g., Diebold and Shin, 2019; Dong et al., 2022).

non-zero weights to be equal by imposing (10):

$$\|\mathbf{w}\|_0 = k \tag{9}$$

$$w_m \in \left\{0, \frac{1}{k}\right\}, m = 1, \ldots, M, \tag{10}$$

where $\|\mathbf{w}\|_0 = \sum_{m=1}^{M} \mathbb{1}(w_m \neq 0)$ counts the number of the non-zero combination weights, and $\mathbb{1}(\cdot)$ denotes the indicator function.[17] Imposing the additional constraints (9) and (10) further mitigates estimation risk, in particular if $k \ll M$. Assigning equal weights to all PRs at disposal is nested as a special case if $k = M$. The tuning parameter k would be adaptively chosen from a grid as we do for selecting the (time-dependent) value of α.[18]

4 Empirical Work

4.1 Application to a cross-section of stocks

4.1.1 Investment universe and empirical study design

The investment universe in this application comprises the largest 50 US stocks. Their monthly excess returns are constructed from CRSP data. We use data from 1957:01 to 2020:12 and only include stocks listed in NYSE, NASDAQ, or AMEX with a share code of 10 or 11. At the beginning of a given month t, the largest 50 stocks (in terms of market value) with non-missing monthly returns in the previous 120 months constitute

[17] Although the cardinality constraint makes the problem NP-hard, state-of-the-art algorithms can solve such types of problems rapidly even for high dimensions (Bertsimas et al., 2016).

[18] A complementary robustification strategy to additional weight constraints could be using block bootstrap up to a given point in time to estimate combination weights at said time as proposed by Bonaccolto and Paterlini (2020) and Kazak and Pohlmeier (2022). While such a strategy could still generate adaptive combination weights, the possibility of rapid changes in the combination weights would be partially lost. However, block bootstrap methods could be extended to accommodate exponential down-weighting performance longer ago by introducing one additional tuning hyperparameter.

the investment universe.[19] [20] Note that the largest 50 stocks can change from month to month, that is, the investment universe is dynamic.

Each candidate PR that we entertain in our library has to assign weights to the 50 stocks at the beginning of each month. We obtain the first OOS portfolio returns in 1967:01. We set aside the first 60 OOS returns for initial optimization of the PRs' weights according to (6) and another 60 months for initial tuning of the forgetting factor α according to (8). Our first OOS evaluation takes place on January 1977. We, then, move forward and run the optimization based on an expanded sample of 61 OOS portfolio returns and choose the value of the forgetting factor also based on one additional observation. We proceed in a recursive manner and end up with an evaluation sample spanning the 1977:01 to 2020:12 period. We consider a power utility investor with a relative risk aversion of $\gamma = 3$. Our setup considers only risky assets. If we were to include a risk-free asset, this could easily be accomplished by adding a candidate PR that is represented by a vector of zeros, since the returns are defined as excess returns in this application.

4.1.2 Candidate PRs

We consider the following five candidate PRs:

- 1/N:

  This PR assigns equal weights to all assets. The 1/N rule does not exploit any sample information and, hence, avoids estimation error. It has been found to outperform a broad range of estimated optimal portfolios across many empirical datasets (DeMiguel et al., 2009; Yuan and Zhou, 2022).

- Volatility timing (VOLTIME):

[19] In the (rare) cases where the return of a stock is missing for month t, we set the excess return to zero.

[20] The choice of a rolling estimation window of 120 months follows, among others, DeMiguel et al. (2009) and Kan et al. (2022).

Kirby and Ostdiek (2012) propose a volatility timing strategy where the weights are given by:

$$\omega_{t+1,n} = \frac{(1/\widehat{\sigma}_{t+1,n})^{\eta}}{\sum_{n=1}^{N}(1/\widehat{\sigma}_{t+1,n})^{\eta}}, \quad n = 1, ..., N, \tag{11}$$

where $\widehat{\sigma}_{t+1,n}$ denotes the estimated conditional variance of the n-th risky asset for time $t+1$, using a rolling window of past returns from $t-119$ to t. This PR ignores any sample information about conditional means and covariances. The tuning parameter $\eta \geq 0$ controls timing aggressiveness. Kirby and Ostdiek (2012) consider the values $\eta = 1, 2$, and 4. We set $\eta = 4$.

- Maximizing expected OOS utility:

  Kan et al. (2022) developed theoretically optimal combination portfolios which have the highest expected OOS utility for a MV investor in a setting without a risk-free asset. Their proposed approach involves combining the GMV portfolio with a sample zero-investment portfolio, where estimation risk is taken into account to control the exposure to the sample zero-investment portfolio. If the exposure to the sample zero-investment portfolio is set to zero, the GMV is nested as a special case. The combination method proposed by Kan et al. (2022) can be applied together with refined estimates of expected returns and expected (co-)variances to form optimal portfolios by using shrinkage estimators or the single factor structure. We consider the following two specifications of their approach:[21]

  - Kan et al. (2022) combined with MacKinlay and Pástor (2000) (KWZ - MP): MacKinlay and Pástor (2000) exploit the implications of an asset pricing model with a single risk factor for estimating expected returns. By doing so, only

[21] We use estimation windows of 120 months and set the risk aversion coefficient to 3 in both PRs.

few parameters have to be estimated, reducing estimation risk.[22]

– Kan et al. (2022) combined with Ledoit and Wolf (2004) (KWZ - LW):
Ledoit and Wolf (2004) propose a shrinkage estimator of the covariance matrix involving a linear combination of the sample covariance matrix and the identity matrix.[23] [24]

- Galton-Shrinkage (GALTON):
Barroso and Saxena (2022) propose a shrinkage estimator that exploits the structure of past OOS forecast errors to correct the expected returns and expected (co-) variances as inputs for portfolio optimization. They use the cleansed inputs to compute the Galton MV portfolio whose weights are the result of simple Markowitz optimization applied to corrected inputs.[25]

The key formula for correcting the optimization inputs is:

$$\mathbf{Z}_t = \widehat{g}_0 + \widehat{g}_1 \mathbf{Z}_{t-1}, \tag{12}$$

where, for any variable $\mathbf{Z}$ of interest (mean returns, variances or pairwise correlations), $\mathbf{Z}_{t-1}$ denotes its historical estimate at time $t-1$ calculated from a rolling window of 60 observations. $\mathbf{Z}_t$ is the cleansed portfolio input for t. Fama-MacBeth regressions are used to estimate the Galton shrinkage coefficients g_0 and g_1 for the means, variances and pairwise correlations. To do so, a large estimation universe is used, comprising the record of the largest 500 US stocks at each point in time

[22] The weights for this rule are given by Equation 51 in Kan et al. (2022).
[23] The weights for this rule are given by Equation 43 in Kan et al. (2022).
[24] KWZ-MP and KWZ-LW are theory-based combination rules relying on the assumption that returns are identically and independently multivariate normally distributed. Although this assumption might not be literally true, these PRs might nevertheless contribute to the ensemble. Our data-driven framework measures to which extent these PRs provide incremental empirical value to other candidates in the pool.
[25] The weights for this rule are computed according to Equation 7 in Barroso and Saxena (2022).

to learn from.[26] For running the Fama-MacBeth regressions, we set a window of 12 ex-post realizations, and to initialize the Galton coefficients, we set aside one additional learning period of 108 months.[27] [28]

The slope coefficient in (12) controls the shrinkage intensity. At one extreme, if its estimated value equals 1, the corrected input equals the historical estimates, that is, the uncorrected estimates. At the other extreme, if its estimated value equals 0, the historical estimates are found to be completely unreliable, and the corrected inputs are set to the grand mean of the returns, variances or pairwise correlations observed up to time t.[29] Let $g_{1,mean}$, $g_{1,var}$ and $g_{1,corr}$ denote the Galton slope coefficients for the means, variances, and correlations, respectively. Different extreme values of $g_{1,mean}$, $g_{1,var}$, and $g_{1,corr}$ produce well-known strategies as special cases, namely the 1/N portfolio for $g_{1,mean} = g_{1,var} = g_{1,corr} = 0$, the sample GMV for $g_{1,mean} = 0, g_{1,var} = g_{1,corr} = 1$, and the sample Markowitz portfolio for $g_{1,mean} = g_{1,var} = g_{1,corr} = 1$.

4.1.3 Results

Table 1 reports the results for the candidate PRs and the combined PRs. Given our focus on the investor's utility, the certainty equivalent return (CER) seems to be a natural choice for measuring portfolio performance. We report (monthly) CER values[30] without transaction costs as well as with proportional transaction costs (CER^{TC}) of 20 basis

[26] Barroso and Saxena (2022) consider both larger and smaller estimation universes and find similar results.

[27] See Equations 9 to 13 in Barroso and Saxena (2022) for details relating to the estimation of the Galton coefficients.

[28] In the notation of Barroso and Saxena (2022), $H = 60$, $E = 12$ and $L = 108$.

[29] Note that we restrict the slope coefficients to lie between 0 and 1.

[30] CER values are computed over the evaluation sample from τ^{**} to T as

$$CER = \left\{ (1-\gamma) \ \frac{1}{T-\tau^{**}+1} \sum_{s=\tau^{**}}^{T} U\left(\mathbf{w}_s^* \left(\alpha_s^*\right)' \mathbf{R}_s\right) \right\}^{\frac{1}{1-\gamma}} - 1. \tag{13}$$

points (bps).[31] We further report the (monthly) Sharpe ratio without transaction costs (SR) and with proportional transaction costs (SR^{TC}) of 20 bps.[32] Avg. TO indicates the average monthly turnover.

The key findings from Table 1 can be summarized as follows. FLEXPOOL generated the highest CER value and Sharpe ratio before and after transaction costs, both compared to any candidate PR as well as compared to the alternative combination schemes. FLEXPOOL did substantially better than STATPOOL in terms of CER values and the SR. This finding illustrates the importance of emphasizing recent utility for assigning combination weights. More distant utility was substantially down-weighted throughout the sample with a forgetting factor α that fluctuated between 0.93 and 0.96 (see Figure 2). FLEXPOOL generated substantially higher utility gains than equally weighted PRs as well. Figure 3 depicts the cumulative utility differences between FLEXPOOL and equally weighted PRs, illustrating that the outperformance cumulated continuously over time rather than being due to few short-lived episodes. The utility differences are statistically significant according to the one-tailed test of Diebold and Mariano (1995) at the 5% significance level. Similarly, the differences in Sharpe ratios are statistically significant according to the one-tailed test of Ledoit and Wolf (2008) at the 5% significance level. This is quite remarkable, given that the equally weighted combination is roughly on par with the best-performing candidate PRs and, hence, provides a tough benchmark.

[31] The choice of 20 bps follows Kan et al. (2022).

[32] CER values are more appropriate as an evaluation metric than the Sharpe ratio in our power utility framework that aims to exploit time-varying investment opportunities; see, e.g., Bianchi and Guidolin (2014). Nevertheless, we report the Sharpe ratio, given its popularity in performance evaluation of asset allocation strategies.

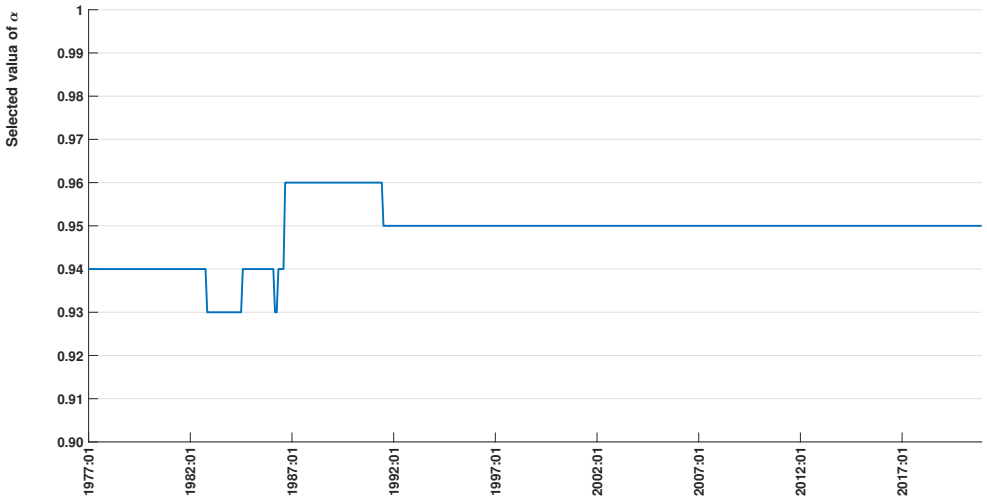

Figure 2: Evolution of the selected value of the forgetting factor α in FLEXPOOL.

Which combination weights were assigned to the different PRs and how did they change over time? Figure 4 provides answers. The subplot in the upper left corner of Figure 4 depicts the weight shares of the PRs averaged over the evaluation sample. The blue (red) bars represent the weight shares of FLEXPOOL (STATPOOL). The remaining subplots show the evolution of PR weights over time. The blue (red) lines indicate the combination weights of FLEXPOOL (STATPOOL). STATPOOL essentially splitted the combination weights between GALTON and KWZ-MP, while the weights in FLEXPOOL were broadly dispersed, with weight shares between 13.68% (GALTON) and 25.09% (KWZ-MP) over the evaluation sample. Interestingly, GALTON received the lowest average weight in FLEXPOOL, even though it was the candidate PR with the highest CER value over the evaluation sample. This result is a manifestation of the ensemble view, where (time-varying) inter-dependencies among PR returns are taken into account.

The optimal combination weights of STATPOOL are more persistent than those of FLEXPOOL, where the combination weights change rapidly and where frequently the entire weight is assigned to one candidate PR. For example, VOLTIME received a high

weight after the burst of the dotcom bubble and its aftermath as well as during the subprime crisis. The 1/N rule prevailed in the relatively tranquil period in the mid-to-late 1990s. We next conduct deeper analyses to shed more light on the mechanisms at work that produce the utility gains of FLEXPOOL.

Table 1: Summary of results.
The table reports our results for the evaluation sample from 1977:01 to 2020:12. It contains monthly CER values without transaction costs and with proportional transaction costs (CER^{TC}) of 20 bps for a power utility investor with relative risk aversion of $\gamma = 3$. As a further performance measure, the table shows the monthly Sharpe ratio before transaction costs (SR) and after proportional transaction costs of 20 bps (SR^{TC}). Avg. TO indicates the average turnover of the evaluation sample.

Candidate PRs	CER	CER^{TC}	SR	SR^{TC}	Avg. TO
1/N	0.0035	0.0033	0.1479	0.1442	0.0782
VOLTIME	0.0051	0.0049	0.1954	0.1898	0.1015
KWZ-MP	0.0046	0.0041	0.1790	0.1663	0.2464
KWZ-LW	0.0045	0.0034	0.1772	0.1469	0.5717
GALTON	0.0052	0.0046	0.2029	0.1853	0.3100
Combined PRs					
FLEXPOOL	0.0068	0.0060	0.2390	0.2168	0.4184
STATPOOL	0.0045	0.0040	0.1832	0.1681	0.2650
EQUAL WEIGHTS	0.0050	0.0046	0.1992	0.1878	0.1964

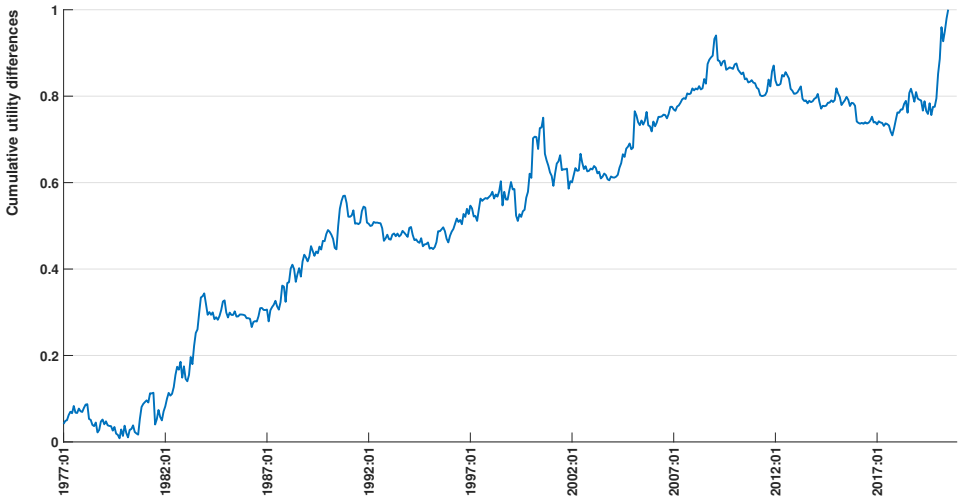

Figure 3: Cumulative utility differences between FLEXPOOL and equally weighted PRs.

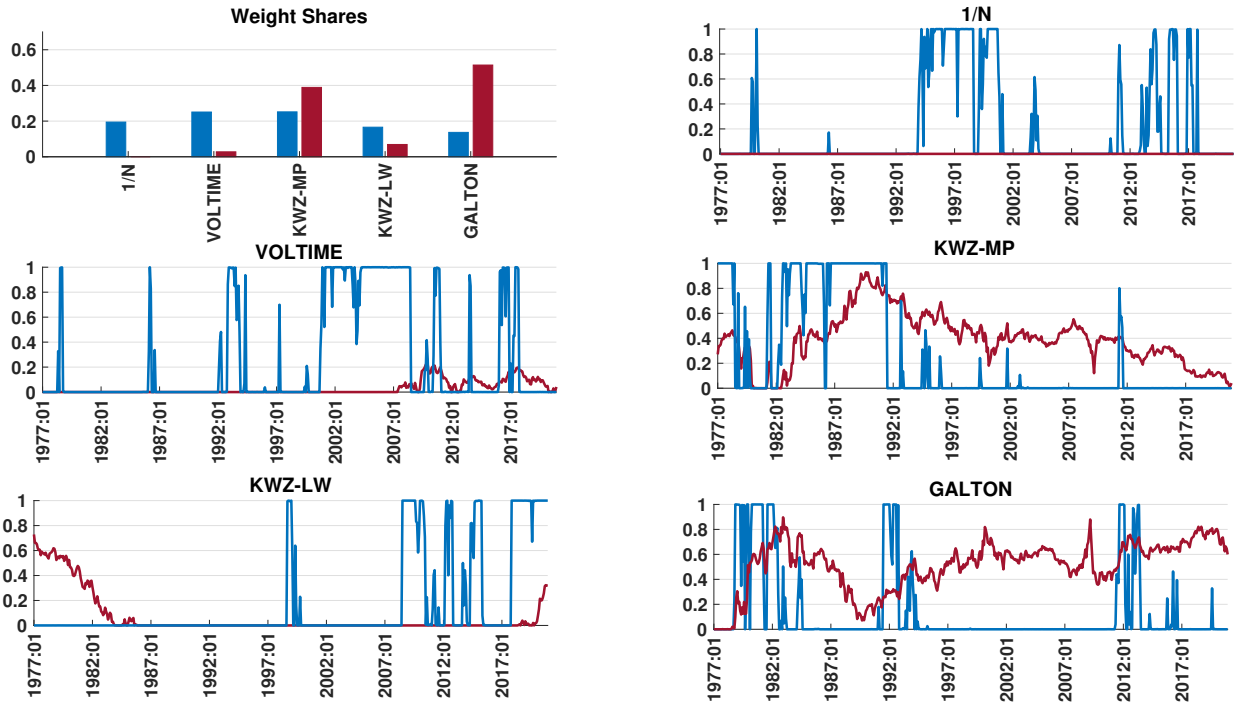

Figure 4: Combination weights.
The subplot in the upper left corner shows the weight shares of FLEXPOOL (blue bars) and STATPOOL (red bars), averaged over the evaluation sample from 1977:01 to 2020:12. The remaining subplots show the evolution of the combination weights of the candidate PRs. The blue (red) lines indicate the combination weights in FLEXPOOL (STATPOOL).

4.1.4 Deeper analyses

Relationship between the number of combined PRs and utility gains

How does the performance of FLEXPOOL depend on the number of combined PRs? So far, we have only reported the result for the case where we combine all five considered PRs. How would the results change if we were to combine subsets of two, three or four PRs? Figure 5 depicts the CER values as a function of the number of combined PRs. The blue diamonds indicate the generated CER value produced by a particular subset of combined PRs. For example, in case of the subset with two combined PRs, there are $\binom{5}{2} = 10$ possible combinations. The red square shows the average CER value for a given number of combined PRs. Figure 5 illustrates that, on average, the CER values increase with the number of combined PRs.

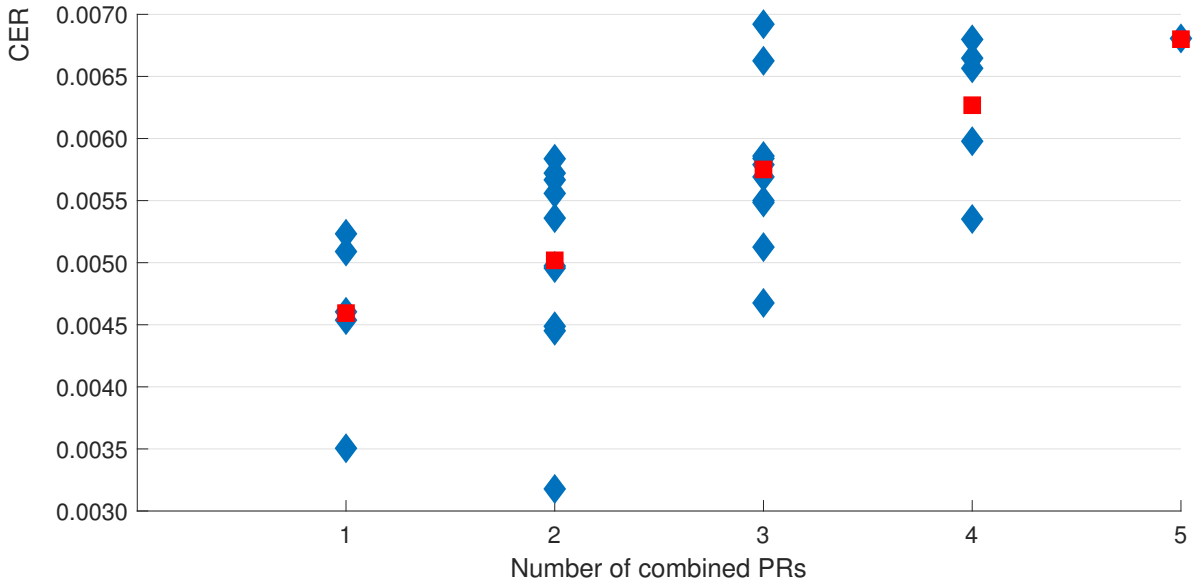

Figure 5: CER values as a function of the number of combined PRs using FLEXPOOL.
The blue diamonds indicate the generated CER values of all possible combinations for a given number of combined PRs. The red square represents the average CER value for a given number of combined PRs.

The highest CER value (0.0058) in the subset of two PRs is achieved by the combination of GALTON and KWZ-LW, the lowest performance (0.0032) is generated by

combining the 1/N rule with KWZ-LW. In the subset with three rules, the highest CER value (0.0069) is achieved by the combination of VOLTIME, KWZ-MP and KWZ-LW. The lowest performance (0.0047) among the subset with three combined rules is generated by the 1/N rule, GALTON and KWZ-MP. In the subset with four combined PRs, the highest CER (0.0068) is achieved when the 1/N rule is left out, and the lowest performance (0.0054) when GALTON is discarded. Note that the CER value for all subsets with four PRs is higher than that of the best candidate PR (0.0052).

The result that the CER values increase on average with the number of combined PRs illustrates the benefit of diversification across more than only two PRs. It also indicates that there is still room left for improving utility gains by including further PRs that contribute different aspects to the library. While the five PRs in our library differ in how they process information into asset weights, all of them rely solely on price data as information. PRs that exploit asset characteristics, e.g., by methods of Freyberger et al. (2020), Gu et al. (2020) or DeMiguel et al. (2020), could be useful extensions in the library. Similarly, including PRs that represent different factor portfolios such as momentum, growth and value appears to be of high interest as well. An extended pool of PRs might also be beneficial in terms of reducing transaction costs since trades implied by different PRs might partly offset each other.

Predictive power and risk management

Each PR, no matter how it is constructed, provides a record of asset weights and implied OOS returns. To delve deeper into the mechanisms of our combination framework, we exploit the record of asset weights by analyzing their relationship to the implied OOS returns. Concretely, we look at statistics on the predictive power and risk management skills of the candidate and combined PRs. As a proxy of predictive power, we estimate Spearman's rank correlation coefficient $\widehat{\rho}_{SP}\left(\omega^{**}, \widetilde{\widetilde{\mathbf{r}}}\right)$ as a robust correlation measure, with

$$\omega^{**} = \begin{pmatrix} \omega^*_{1977:01} \\ \vdots \\ \omega^*_{2020:12} \end{pmatrix} \text{ and } \widetilde{\widetilde{\mathbf{r}}} = \begin{pmatrix} \widetilde{\mathbf{r}}_{1977:01} \\ \vdots \\ \widetilde{\mathbf{r}}_{2020:12} \end{pmatrix},$$

where ω^{**} denotes the asset weights implied by the PRs and the combination weights computed according to (7),[33] stacked from beginning to end of the evaluation sample. With $N = 50$ assets and an evaluation sample that comprises 528 months (1977:01 to 2020:12), the vector ω^{**} has length $50 \times 528 = 26,400$. Similarly, $\widetilde{\widetilde{\mathbf{r}}}$ denotes the stacked pseudo OOS (excess) returns generated by the $N = 50$ assets.

The intuition behind the rank correlation $\widehat{\rho}_{SP}(\omega^{**}, \widetilde{\widetilde{\mathbf{r}}})$ is the following: a PR will assign a positive (negative) weight if the expected return of an asset is positive (negative). Hence, $\widehat{\rho}_{SP}\left(\omega^{**}, \widetilde{\widetilde{\mathbf{r}}}\right)$ approximates the overall predictive power of a PR. A high positive correlation indicates high predictive power.

Regarding risk management skills, proxied by a PR's capability of controlling the variance of its returns, a PR assigns a low (high) squared weight $\omega^{*,2}_{s,n}$ to the n-th asset if the expected squared return of the n-th asset is high (low) for time s. Likewise, for a pair of assets p and q ($p \neq q$), a PR takes a high (low) cross-exposure $\omega^*_{s,p}\omega^*_{s,q}$ when the product of the associated asset returns is expected to be low (high). Based on this intuition, we calculate Spearman's rank correlation $\widehat{\rho}_{SP}\left(\omega^{***}, \widetilde{\widetilde{\mathbf{r}}}\right)$, where

[33] For candidate PRs and the equally weighted benchmark combination, the optimal weights $\mathbf{w}^*_s$ in (7) are replaced by assigning the entire weight to the said candidate PR and equal weights, respectively.

$$\omega^{***} = \begin{pmatrix} \omega^*_{1977:01,n=1} & \times & \omega^*_{1977:01,n=1} \\ \vdots & \ddots & \vdots \\ \omega^*_{1977:01,n=1} & \times & \omega^*_{1977:01,n=50} \\ \vdots & \ddots & \vdots \\ \omega^*_{1977:01,n=50} & \times & \omega^*_{1977:01,n=50} \\ \vdots & \ddots & \vdots \\ \omega^*_{2020:12,n=1} & \times & \omega^*_{2020:12,n=1} \\ \vdots & \ddots & \vdots \\ \omega^*_{2020:12,n=1} & \times & \omega^*_{2020:12,n=50} \\ \vdots & \ddots & \vdots \\ \omega^*_{2020:12,n=50} & \times & \omega^*_{2020:12,n=50} \end{pmatrix} \text{ and } \tilde{\tilde{\tilde{\mathbf{r}}}} = \begin{pmatrix} \widetilde{r}_{1977:01,n=1} & \times & \widetilde{r}_{1977:01,n=1} \\ \vdots & \ddots & \vdots \\ \widetilde{r}_{1977:01,n=1} & \times & \widetilde{r}_{1977:01,n=50} \\ \vdots & \ddots & \vdots \\ \widetilde{r}_{1977:01,n=50} & \times & \widetilde{r}_{1977:01,n=50} \\ \vdots & \ddots & \vdots \\ \widetilde{r}_{2020:12,n=1} & \times & \widetilde{r}_{2020:12,n=1} \\ \vdots & \ddots & \vdots \\ \widetilde{r}_{2020:12,n=1} & \times & \widetilde{r}_{2020:12,n=50} \\ \vdots & \ddots & \vdots \\ \widetilde{r}_{2020:12,n=50} & \times & \widetilde{r}_{2020:12,n=50} \end{pmatrix}.$$

The rank correlation $\widehat{\rho}_{SP}\left(\omega^{***}, \tilde{\tilde{\tilde{\mathbf{r}}}}\right)$ approximates the ability of a PR to control the variance of the generated returns and can thus be seen as a proxy of risk management skills. The more negative the correlation is, the better are the PR's risk management skills.

Table 2 summarizes the results for predictive power and risk management skills. FLEXPOOL has by far the highest predictive power with an estimated rank correlation coefficient of 0.0162, which is different from zero at the 1% significance level.[34] Interestingly, FLEXPOOL achieves significant predictive power even though none of the candidate PRs has significant predictive power when measured over the entire evaluation sample. Key are rapidly shifting combination weights to (combinations of) PRs with predictive power which is local in time. VOLTIME exhibits by far the best risk

[34] As one might expect, the magnitudes of the correlations are fairly low, being consistent with a low degree of predictability.

management skills. When maximizing economic utility, FLEXPOOL implicitly strikes a balance between predictive power and risk management, partly sacrificing VOLTIME's risk management skills to achieve enhanced predictive power.

Table 2: Predictive power and risk management skills.
Spearman's rank correlation $\widehat{\rho}_{SP}\left(\omega^{**}, \widetilde{\widetilde{\mathbf{r}}}\right)$ approximates predictive power, and Spearman's rank correlation $\widehat{\rho}_{SP}\left(\omega^{***}, \widetilde{\widetilde{\mathbf{r}}}\right)$ approximates risk management skills. The p-values to the null that the correlation coefficient equals zero are shown underneath the correlation estimates in parentheses.

Candidate PRs	$\widehat{\rho}_{SP}\left(\omega^{**}, \widetilde{\widetilde{\mathbf{r}}}\right)$	$\widehat{\rho}_{SP}\left(\omega^{***}, \widetilde{\widetilde{\mathbf{r}}}\right)$
1/N	–	–
VOLTIME	− 0.0014 (0.8259)	−0.0525 (0.0000)
KWZ-MP	0.0042 (0.4932)	−0.0070 (0.2552)
KWZ-LW	0.0030 (0.6234)	0.0088 (0.1527)
GALTON	0.0074 (0.2283)	0.0076 (0.2162)
Combined PRs		
FLEXPOOL	0.0162 (0.0085)	−0.0141 (0.0223)
STATPOOL	0.0037 (0.5506)	0.0003 (0.9617)
EQUAL WEIGHTS	0.0064 (0.2996)	−0.0126 (0.0401)

4.2 Application to market timing

4.2.1 Investment universe and empirical study design

In this application we consider an investor endowed with power utility preferences and relative risk aversion of $\gamma = 3$ who can, in each month, allocate their wealth between the S&P 500 index and three-month US treasury bills. We restrict the weight allocated to

stocks to lie in the range $[0; 1.5]$ and therefore ensure that any considered PR adheres to these weight restrictions. Our evaluation period spans from 1977:01 to 2020:12. Each PR generates the first OOS return in 1967:01 and we use 60 months of OOS returns for the initial optimization of combination weights. We set aside another 60 observations for initial tuning of the forgetting factor α. CER values are computed based on total returns in this application.

4.2.2 Candidate PRs

We consider a diverse set of six different PRs. The first three PRs are based on strategies that exploit Bayesian predictive densities of next period's excess return y, that is, the return on the S&P 500 (including dividends) in excess of the risk-free rate r^f. Bayesian predictive densities of excess returns are attractive choices as a basis for market timing decisions, given their capability of accommodating parameter and model uncertainty as well as of using time-varying parameters (TVP) and stochastic volatility (SV). In the context of return predictability, Bayesian predictive densities have been used by, among others, Dangl and Halling (2012), Johannes et al. (2014) and Pettenuzzo and Ravazzolo (2016). While the first three PRs in our library differ with respect to specific choices that are relevant for computing the respective Bayesian predictive densities, we can present all of them in canonical form. These PRs solve the investment problem by directly maximizing the conditional expected utility of next period's wealth W_{t+1}:

$$\underset{\omega_{t+1}\in[0;1.5]}{\arg\max}\ \mathbb{E}_t\left[U(W_{t+1})|\mathscr{D}^t\right] = \tag{14}$$

$$\underset{\omega_{t+1}\in[0;1.5]}{\arg\max}\int \frac{\widetilde{R}_{t+1}^{1-\gamma}}{1-\gamma} p\left(y_{t+1}|\mathscr{D}^t\right) dy_{t+1}, \tag{15}$$

where $p\left(y_{t+1}|\mathscr{D}^t\right)$ denotes a Bayesian predictive density for the excess return y in $t+1$ based on the information set available at time t. The information set $\mathscr{D}^t$ comprises the

returns and predictors of the excess return that can be observed until t as well as the prior choices in $t = 0$. As power utility does not depend on wealth, we can set $W_t = 1$ and proceed with the gross returns in (15). Let $\widetilde{R}_{t+1}$ denote the gross total return in $t+1$, where the total return comprises the excess return y and the risk-free rate r^f. Further, let ω_{t+1} denote the weight that is allocated to the risky asset for time $t+1$. We maximize the conditional expected utility by approximating (15), based on $B = 100,000$ potential realizations $y^{(b)}_{draw,t+1}$, $b = 1, \ldots, B$, of the excess return in $t+1$ from the predictive density $p\left(y_{t+1}|\mathscr{D}^t\right)$:

$$\underset{\omega_{t+1}\in[0;1.5]}{\arg\max} \frac{1}{B}\sum_{b=1}^{B}\left\{\frac{\left[\omega_{t+1}\left(1+r^f_{t+1}+y^{(b)}_{draw,t+1}\right)+(1-\omega_{t+1})\left(1+r^f_{t+1}\right)\right]^{1-\gamma}}{1-\gamma}\right\}. \tag{16}$$

We set $\gamma = 3$.[35] To obtain a Bayesian predictive density for the excess returns, we have to impose some structure on the return generating process. We assume the dynamics of the excess return to be given by TVP regression models with the following structure:

$$y_{t+1} = \mathbf{X}_t'\theta_t + \varepsilon_{t+1}, \quad \varepsilon_{t+1} \sim \mathcal{N}\left(0, \upsilon_{t+1}\right) \tag{17}$$

$$\theta_t = \theta_{t-1} + \xi_t, \qquad \xi_t \sim \mathcal{N}\left(0, \mathbf{\Xi}_t\right), \tag{18}$$

where $\mathbf{X}_t$ denotes the vector of predictive variables observed in t. This vector contains, depending on the specific setting, a subset of twelve predictor variables from Welch

[35]Note that the risk aversion coefficient for candidate PRs could be different from the value of the risk aversion coefficient used for optimizing combination weights in (6).

and Goyal (2008).[36] Let θ_t denote the vector of (unobserved) time-varying coefficients. The observational error ε_{t+1} is assumed to be normally distributed with mean zero and (unknown) and time-varying variance υ_{t+1}. The time-varying coefficients are assumed to evolve according to a multivariate random walk without drift. We initialize the coefficients θ_0 with a diffuse conditional normal prior centered around zero.

The random shocks ξ are assumed to be multivariate normal with (unknown) and time-varying system covariance matrix $\mathbf{\Xi_t}$. Conditional on the observational variance and the system covariance, standard Bayesian methods for state-space models using the Kalman filter can be applied to estimate the coefficients θ_t and to compute the predictive distribution of the returns. The observational variance and the system covariance are, however, unknown. We use a forgetting factor approach to model their dynamics, where the value of the forgetting factor δ controls the dynamics of the coefficients, and the value of the forgetting factor κ governs the dynamics of the observational variance. If we set $\delta = 1$, all available historical observations are equally weighted in the updating process, leading to constant coefficients. If we set $\delta < 1$, older observations are exponentially down-weighted. The lower we choose the value of δ, the more we down-weight older observations.

Similarly, κ controls the dynamics of the observational variance. If we set $\kappa = 1$, we obtain a constant variance. Using a conjugate specification with an inverse-gamma prior for the observational variance and a conditional normal prior for the coefficients, along with fixed values of the forgetting factors δ and κ, we obtain a t-distributed predictive density $p\left(y_{t+1}|\mathscr{D}^t\right)$, which takes the uncertainty about the coefficients and the observa-

[36]The predictors are the dividend yield, the dividend-payout-ratio, the earnings-to-price ratio, the sum of squared daily returns on the S&P 500 index as a measure of stock variance, the book-to-market ratio, the net equity expansion, the treasury bill rate, the long-term government bond yields, the long-term government bond returns, the default return spread, the default yield spread and inflation (lagged by one additional month). We use the predictors from 1927:01 to 2020:11. We downloaded the data from Amit Goyal's homepage: `http://www.hec.unil.ch/agoyal/`. See Welch and Goyal (2008) for a more detailed description of the variables.

tional variance into account. Our PRs based on Bayesian predictive densities include the three following setups that differ with respect to the included predictors and the considered values of the forgetting factors δ and κ:

- LARGE-TVP-SV:

  This multivariate setup includes all of the twelve considered predictors from Welch and Goyal (2008) and uses Bayesian model averaging (BMA) (Raftery et al., 1997) to attach weights to the predictive densities, which are based on different specifications of the coefficients' dynamics. The dynamics are controlled by the value of the forgetting factor δ. It is chosen from the grid $\mathscr{S}_\delta = \{0.96; 0.97; 0.98; 0.99; 1.00\}$, including constant coefficients as a special case. Hence, the five individual models $\mathscr{M}_j$, $j = 1, \ldots, 5$, in this setup are defined by different values of δ. As conditional heteroskedasticity is a well-known stylized fact for asset returns, we set the forgetting factor $\kappa = 0.97$ for the observational variance, following the choice of Reuters (1996) for monthly data. A priori, we assign equal weights to the five predictive densities. After having computed the weights of the predictive densities at each point in time using Bayes' rule, asset allocation decisions can be made based upon the mixture t-distribution using the approximation (16).

- BMA-TVP-CV:

  The second setup is based on the setting proposed by Dangl and Halling (2012). With a set of twelve available predictors, there are 2^{12} different combinations of predictors that are either included in or excluded from the vector of predictors $\mathbf{X}$. The value of the forgetting factor δ for controlling the coefficients' dynamics is again chosen from the grid $\mathscr{S}_\delta = \{0.96; 0.97; 0.98; 0.99; 1.00\}$. Hence, $5 \times 2^{12} = 20,480$ different models $\mathscr{M}_j$, $j = 1, \ldots, 20,480$, defined by different subsets of included predictors and values of δ, are at disposal. Dangl and Halling (2012) use a constant variance (CV). In order to mimic their choice in this aspect, we set $\kappa = 1.00$. A

priori, we assign equal weights to the 20,480 predictive densities and update their weights using BMA.[37]

- UNIV-TVP-SV:

  Univariate TVP-SV models are common choices for modeling the dynamics of aggregate stock returns (Johannes et al., 2014; Pettenuzzo and Ravazzolo, 2016). This setup uses solely univariate (UNIV) predictive regression, including one of the twelve predictors in each of the regressions, and also considers the grid $\mathscr{S}_\delta = \{0.96; 0.97; 0.98; 0.99; 1.00\}$ for δ, κ is set to 0.97, and all predictive densities are equally weighted.

The following three PRs rely on a MV framework to form portfolios. The PRs differ in how they compute the estimated excess return $\widehat{y}_{t+1}$. The weight assigned to the S&P 500 index is computed as:

$$\underset{[0;1.5]}{\omega_{t+1}} = \frac{1}{\gamma}\left(\frac{\widehat{y}_{t+1}}{\widehat{\sigma}^2_{t+1}}\right), \tag{19}$$

where $\widehat{\sigma}^2_{t+1}$ denotes the estimate of the variance, calculated over a rolling window of 60 months. The risk aversion γ is set to 3. We consider the following PRs:

- Sum-of the-parts method (SOP):

  Imposing economic constraints, Ferreira and Santa-Clara (2011) forecast aggregate stock returns as the sum of the dividend-price ratio and the long-run historical average of earnings growth. In contrast to predictive regressions, there are no parameters to estimate and thus no estimation error.

- Combination of forecasts (CF):

  Rapach et al. (2010) propose an equally weighted combination of point forecasts,

[37] While this setup closely follows Dangl and Halling (2012), there are slight implementation differences. For example, Dangl and Halling (2012) include the cross-sectional beta premium of Polk et al. (2006) as a predictor, while we do not include it since the data are available only until 2002.

where each point forecast is based on univariate predictive regressions with constant coefficients and one of the predictors proposed in Welch and Goyal (2008). Note that we use monthly data, whereas Rapach et al. (2010) use quarterly data and consider 15 instead of 12 predictors.

- Prevailing historical mean (PHM):

  This PR uses the prevailing historical mean of the excess returns as a point forecast.

Altogether, our library of PRs contains heterogeneous asset allocation approaches. The PRs use different information sets and different ways of mapping information into asset weights. The most striking distinction between them is that some PRs rely on differently designed Bayesian predictive densities for asset allocation, while others are based on MV specifications with diverse strategies of producing point forecasts.

4.2.3 Results

Table 3 reports the results. It shows CER values without transaction costs as well as with proportional transaction costs (CER^{TC}) of 20 bps. We further report the (monthly) Sharpe ratio without transaction costs (SR) and with proportional transaction costs (SR^{TC}) of 20 bps. The R^2_{OOS}-statistic (Campbell and Thompson, 2008) compares the point forecast accuracy of a given approach to the PHM benchmark. It measures the proportional reduction in the sample mean squared forecast error compared to the prevailing historical mean benchmark. Hence, a positive R^2_{OOS}-statistic indicates that the mean square forecast error of the given approach is lower than that of the PHM. As a proxy of predictive skills, we report $\widehat{\rho}_{SP}(\omega^{**}, \mathbf{y})$, that is, Spearman's rank correlation coefficient between the weights allocated to the risky asset and the realized excess returns. Let ω^{**} denote the stacked weights of the risky asset over the evaluation sample, and $\mathbf{y}$ denotes the vector of realized excess returns over the evaluation sample. As a proxy of risk management skills, we report Spearman's rank correlation coefficient $\widehat{\rho}_{SP}(\omega^{**,2}, \mathbf{y}^2)$.

We summarize the empirical results as follows. FLEXPOOL generated the highest CER values and Sharpe ratios among the combined PRs. It exhibited both high predictive power and strong risk management skills. LARGE-TVP-SV had the strongest predictive power, while SOP had the best risk management skills. The evolution of combination weights is depicted in Figure 6. Most of the time, LARGE-TVP-SV got a high (and often even the entire) weight. SOP, however, with its strong risk management skill, was picked in three turbulent phases: in September and October 1998, after the strongly negative returns in August 1998, a time which is associated with the Russian currency crisis and Long Term Capital Management's collapse. Further, SOP was picked from 2000:12 to 2003:10, a period associated with the burst of the dotcom bubble. Lastly, SOP was chosen from 2020:04 to 2020:07 after the large drop due to the COVID-19 pandemic in March 2020. Similar to our first application, the results document FLEXPOOL's capability of automatically balancing predictive power and risk management skills of the candidate PRs when maximizing economic utility.

The value of the forgetting factor α was chosen to be 0.96 according to (8) over the entire evaluation period. The emphasis on the recent economic utility gains resulted in more rapid adjustments of the combination weights compared to STATPOOL (see Figure 6). The cumulative utility differences between FLEXPOOL and equally weighted PRs are depicted in Figure 7. They are statistically significant at the 5% level according to a one-tailed test of Diebold and Mariano (1995). Similarly, the differences in Sharpe ratios are statistically significant according to the one-tailed test of Ledoit and Wolf (2008) at the 10% significance level.

Among the candidate PRs, LARGE-TVP-SV generated by far the highest CER value and SR despite its low R^2_{OOS}-statistic of -0.1133. This result is reminiscent of the findings by Cenesizoglu and Timmermann (2012) and Leitch and Tanner (1991) that a model's point forecast accuracy and its generated value measured by an economic criterion can

Table 3: Summary of results.
The table reports our results for the evaluation sample from 1977:01 to 2020:12. It shows monthly CERs without transaction costs as well as with proportional transaction costs (CER^{TC}) of 20 bps for a power utility investor with relative risk aversion of $\gamma = 3$. We further report the (monthly) Sharpe ratio without transaction costs (SR) and with proportional transaction costs (SR^{TC}) of 20 bps. As a measure of point forecast accuracy, we report R^2_{OOS}-statistics. Predictive power and risk management skills are proxied by Spearman's rank correlations $\widehat{\rho}_{SP}(\omega^{**}, \mathbf{y})$ and $\widehat{\rho}_{SP}(\omega^{**,2}, \mathbf{y}^2)$, respectively.

	Economic Evaluation Criteria				Statistical Properties		
Candidate PRs	CER	CER^{TC}	SR	SR^{TC}	R^2_{OOS}	$\widehat{\rho}_{SP}(\omega^{**}, \mathbf{y})$	$\widehat{\rho}_{SP}(\omega^{**,2}, \mathbf{y}^2)$
LARGE-TVP-SV	0.0096	0.0090	0.2048	0.1908	−0.1133	0.1139 (0.0088)	−0.0295 (0.4989)
BMA-TVP-CV	0.0067	0.0065	0.1411	0.1363	−0.0388	0.0145 (0.7390)	− 0.0785 (0.0714)
UNIV-TVP-SV	0.0068	0.0065	0.1451	0.1394	−0.0090	0.0176 (0.6870)	− 0.0885 (0.0422)
SOP	0.0071	0.0070	0.1560	0.1522	0.0003	0.0614 (0.1592)	− 0.1213 (0.0053)
CF	0.0069	0.0068	0.1458	0.1433	0.0010	0.0143 (0.7433)	− 0.0115 (0.7928)
PHM	0.0064	0.0064	0.1389	0.1382	0.0000	−0.0256 (0.5578)	− 0.0302 (0.4880)
Combined PRs							
FLEXPOOL	0.0096	0.0091	0.2063	0.1965	−0.0502	0.0888 (0.0413)	− 0.0979 (0.0245)
STATPOOL	0.0089	0.0084	0.1929	0.1800	−0.0978	0.0929 (0.0329)	−0.0292 (0.5030)
EQUAL WEIGHTS	0.0078	0.0077	0.1672	0.1631	0.0035	0.0577 (0.1859)	−0.0844 (0.0525)

strongly diverge. Hence, the R^2_{OOS}-statistic can be a poor indicator to guide portfolio decisions. LARGE-TVP-SV overfits the data because it uses many predictors, time-varying coefficients and no shrinkage mechanism, resulting in the low R^2_{OOS}-statistic. As a complex model, LARGE-TVP-SV is strong at capturing the structure of returns with a high predictive correlation compared to shrunk models, where the signals are partly muted. However, the high variance of the forecasts based on LARGE-TVP-SV leads to a low R^2_{OOS}-statistic, which, however, is not detrimental in terms of utility since the weight restrictions on the risky asset (no short sales, up to 50% leverage) prevent excessive

portfolio weights.[38]

Similar to the results in our first application, we find that predictive power and risk management skills (positive values of $\widehat{\rho}_{SP}(\omega^{**}, \mathbf{y}^{)}$ and negative values of $\widehat{\rho}_{SP}(\omega^{**,2}, \mathbf{y}^2)$) to align well with the (ranking of the) CER values and the SRs. While our approach of directly optimizing utility at the level of PRs captures the strong economic performance of LARGE-TVP-SV, combination approaches based on statistical measures such as R^2_{OOS}-statistics would not be capable of doing so.

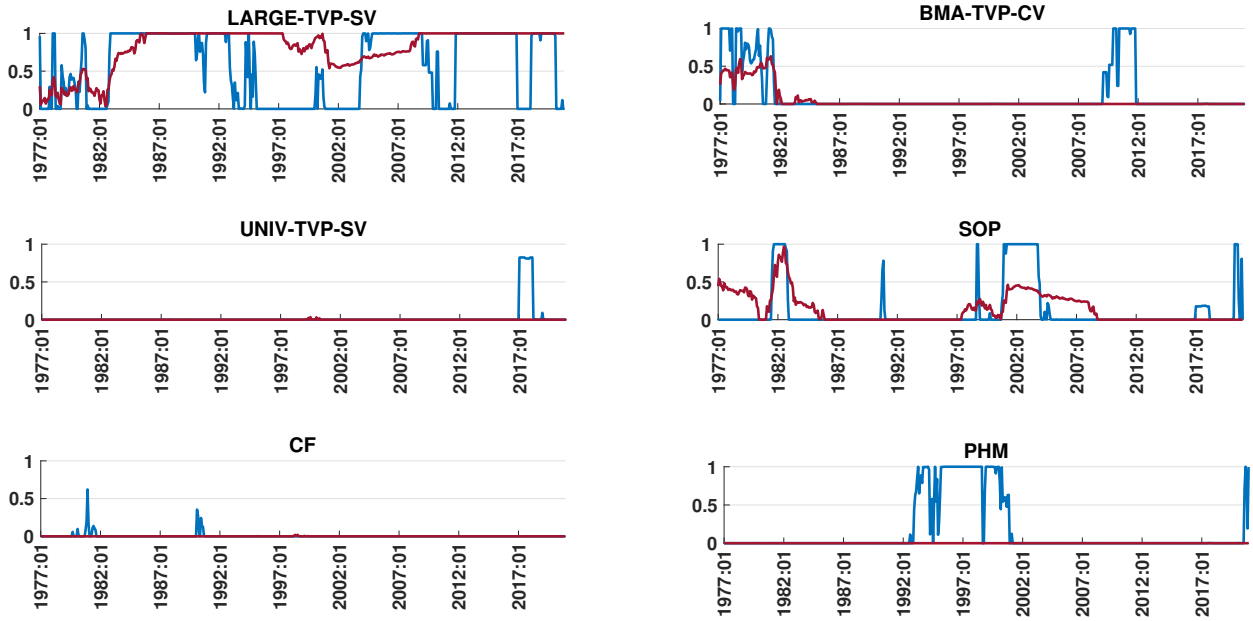

Figure 6: Evolution of combination weights.
The subplots show the evolution of the combination weights. The blue (red) lines indicate the combination weights in FLEXPOOL (STATPOOL).

[38] An alternative to imposing weight restrictions is to take advantage of double descent. That is, in settings where the number of predictors exceeds the number of observations, OOS forecasts tend to become more accurate as the number of predictors *rises*. Kelly et al. (2022) exploit this statistical phenomenon of benign overfitting for optimal market timing. However, this strategy is applicable only in particular settings. PRs based on double descent could be included as candidate rules in our framework.

Figure 7: Cumulative utility differences between FLEXPOOL and equally weighted PRs.

BMA-TVP-CV and UNIV-TVP-SV shrink the coefficients towards zero by using subsets of the predictors. As one would expect, point forecast accuracy in terms of R^2_{OOS}-statistics is higher for these approaches than in case of LARGE-TVP-SV due to their shrinkage mechanisms. However, the predictive power of BMA-TVP-CV and UNIV-TVP-SV is substantially lower as measured by the rank correlation $\widehat{\rho}_{SP}(\omega^{**}, \mathbf{y})$, and so are their CER values and SRs. Similarly, the equally weighted PRs, SOP and CF achieve decent point forecast accuracy but are clearly inferior compared to LARGE-TVP-SV in terms of CER values and SRs. PHM got temporarily high weights in the relatively calm mid-to-late 1990s. This result aligns with the finding in our first application, where 1/N got high weights during this period. Hence, it appears that simple PRs tend to be favored in tranquil periods, while flexible PRs are picked in more turbulent phases.

Figure 8 depicts the CER values as a function of the number of combined PRs.

The blue diamonds show the generated CER values produced by a particular subset of combined PRs, and the red squares indicate the average CER values for a given number of combined PRs. As was the case in our first application, the CER values increase on average as a function of the number of combined PRs.

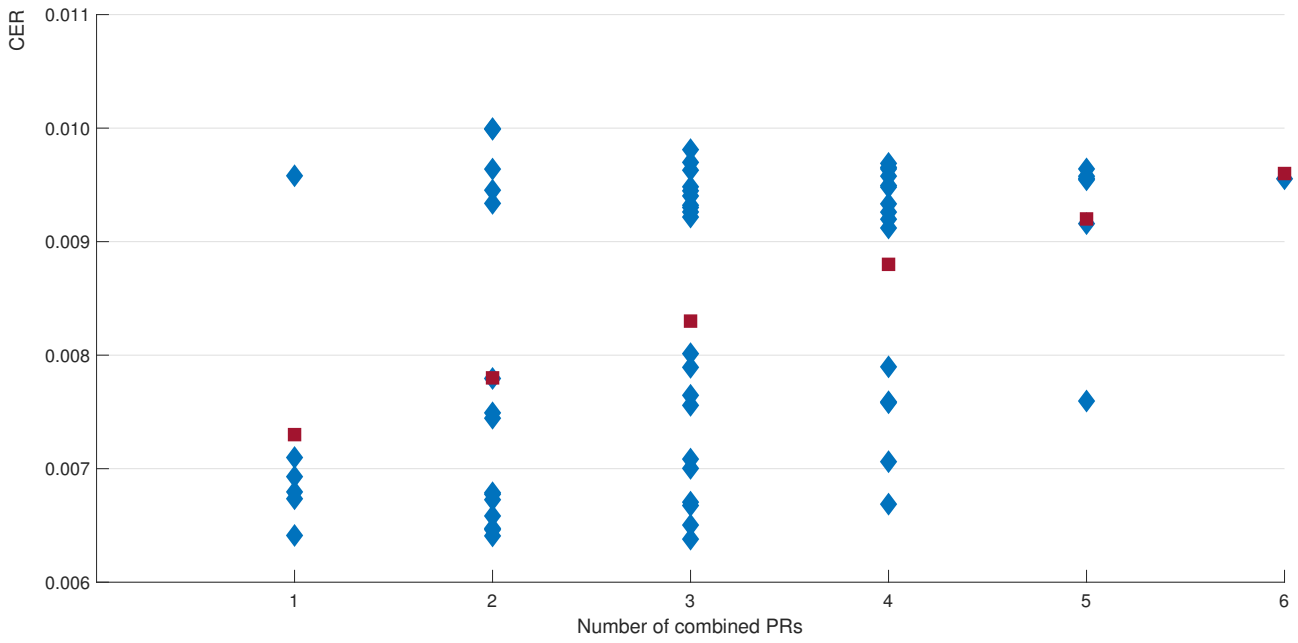

Figure 8: CER values as a function of the number of combined PRs using FLEXPOOL.
The blue diamonds indicate the CER values of all possible combinations for a given number of combined PRs. The red square represents the average CER values for a given number of combined PRs.

In addition to the results presented so far, we explored three alternative empirical settings. First, we added the buy-and-hold strategy (without leverage) as a further candidate PR to the library. It produced a CER value of 0.0073 and a Sharpe ratio of 0.1515. We found our results largely unchanged when adding the buy-and-hold strategy. The other two alternative settings used data only available over a shorter period.

We explored the utility gains from combining PRs that rely on backward-looking data and PRs that are based on forward-looking data. As a representative of a PR using forward-looking data, we chose the strategy of Pyun (2019) that provides OOS forecasts

of the equity premium based on the variance risk premium. His approach exploits the relation between the market risk premium and the price of variance risk by the variance risk exposure. The point forecasts are available from 1990:02 to 2019:12.[39] We used a rolling window of 60 months for computing the variance estimate as an additional input for the MV specification (19) and imposed the same weight restrictions (no short sales, up to 50% leverage) as in our previous analysis. We combined this PR with the LARGE-TVP-SV rule as a representative of a PR that uses backward-looking data and computed results for the evaluation period from 2000:01 to 2019:12. The PR based on the forward-looking data produced a CER value of 0.0078 and a SR of 0.2100. LARGE-TVP-SV generated a CER value of 0.0070 and a SR of 0.1955. Using FLEXPOOL, combination of both PRs marginally improved the results with a CER value of 0.0079 and a Sharpe ratio of 0.2146. As a comparison, PHM achieved a CER value of 0.0014 and a Sharpe ratio of 0.0821 over this shortened evaluation sample.

We further explored whether adding a PR based on the recently proposed approach by Dong et al. (2022) could achieve incremental value relative to LARGE-TVP-SV. Dong et al. (2022) propose a novel approach for exploiting a large number of 100 cross-sectional anomaly portfolio returns as predictors for point forecasts of aggregate excess returns. These forecasts are available from 1975:01 to 2017:12.[40] For this shortened period, we combined the strategy of Dong et al. (2022) with LARGE-TVP-SV and computed results for the evaluation sample from 1985:01 to 2017:12. Using the MV specification (19), we chose their setting where the elastic net is used as a shrinkage technique for computing expected excess returns and used a rolling window of 60 months for estimating the variance. We imposed the same weight restrictions (no short sales, up to 50% leverage) as in our previous analysis. LARGE-TVP-SV achieved a CER value of 0.0082 and the

[39]We downloaded the data from Sungjune Pyun's homepage: https://sjpyun.github.io/research.html.

[40]We downloaded the forecasts from Dave Rapach's homepage: https://sites.google.com/slu.edu/daverapach/publications.

approach of Dong et al. (2022) produced a CER value of 0.0093. The combined PRs generated a CER value of 0.0098 when using FLEXPOOL.

Overall, the results for this application indicate that flexible PRs exploiting multivariate information can be highly beneficial in terms of economic utility. In contrast, Welch and Goyal (2008) find no substantial utility gains (relative to the PHM) for any of the predictors when assessing them individually. Similarly, Goyal et al. (2021) dismiss most of the predictors advanced after Welch and Goyal (2008) in terms of economic utility based on individual evaluation. It would be interesting to revisit the economic profitability of these predictors using PRs that capture their joint impact on economic utility in multivariate setups. To this end, multivariate approaches from different domains could be combined as candidate PRs in our framework; for example, one could explore suitable machine learning techniques for this type of portfolio choice problem, e.g., the methods by Kelly et al. (2022) and Nevasalmi and Nyberg (2021), and Bayesian techniques such as the LARGE-TVP-SV candidate PR.

5 Concluding Remarks

We have introduced an ensemble framework for combining multiple PRs. Moving forward from existing approaches, the proposed combination strategy enables researchers to exploit the myriad of existing PRs in a utility maximization framework while diversifying away estimation risk and retaining many appealing properties. Our approach is capable of merging the virtues of PRs irrespectively of their design and without invoking distributional assumptions regarding the data generating process of the PRs' returns.

Two substantive applications documented the expediency of our approach. The combined PRs achieved OOS CER values that were either higher than those of any candidate PR or roughly as high as those of the ex-post best candidate PR. By taking an ensemble perspective, the candidate PR with the highest individual utility did not necessarily receive the highest weight in the combination. Rapidly shifting combination weights played an important role for enhancing OOS utility by capturing the time-varying performance of the PRs and the inter-dependencies among their (pseudo) OOS returns. Deeper analyses showed how the flexible combination balanced predictive power of asset returns and anticipating their variance. Further, the analyses revealed that utility gains on average rose with the number of candidate PRs—even without using additional regularization on the combination weights to curb estimation risk.

The overarching contribution of our study is its potential to change the way we approach portfolio choice problems: instead of striving to find a single best PR, our framework enables an extensive library of candidate PRs to contribute their strengths in an ensemble, similar to the optimization of a combination of assets. While the search for new candidate PRs relying on enhanced techniques and novel data sources will go on, our framework will also provide a tool to assess the incremental empirical merits (or, lack thereof) of newly proposed PRs.

References

Adämmer, P. and Schüssler, R. A. (2020). Forecasting the equity premium: mind the news! *Review of Finance*, 24(6):1313–1355.

Barroso, P. and Saxena, K. (2022). Lest we forget: Learn from out-of-sample forecast errors when optimizing portfolios. *The Review of Financial Studies*, 35(3):1222–1278.

Beckmann, J., Koop, G., Korobilis, D., and Schüssler, R. A. (2020). Exchange rate predictability and dynamic bayesian learning. *Journal of Applied Econometrics*, 35(4):410–421.

Bernaciak, D. and Griffin, J. E. (2022). A loss discounting framework for model averaging and selection in time series models. *arXiv preprint arXiv:2201.12045*.

Bertsimas, D., King, A., and Mazumder, R. (2016). Best subset selection via a modern optimization lens. *The Annals of Statistics*, 44(2):813–852.

Bianchi, D. and Guidolin, M. (2014). Can long-run dynamic optimal strategies outperform fixed-mix portfolios? evidence from multiple data sets. *European Journal of Operational Research*, 236(1):160–176.

Bonaccolto, G. and Paterlini, S. (2020). Developing new portfolio strategies by aggregation. *Annals of Operations Research*, 292(2):933–971.

Brandt, M. W., Santa-Clara, P., and Valkanov, R. (2009). Parametric portfolio policies: Exploiting characteristics in the cross-section of equity returns. *The Review of Financial Studies*, 22(9):3411–3447.

Breiman, L. (1996). Stacked regressions. *Machine Learning*, 24(1):49–64.

Campbell, J. Y. and Thompson, S. B. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies*, 21(4):1509–1531.

Cenesizoglu, T. and Timmermann, A. (2012). Do return prediction models add economic value? *Journal of Banking & Finance*, 36(11):2974–2987.

Cong, L. W., Tang, K., Wang, J., and Zhang, Y. (2022). Alphaportfolio: Direct construction through reinforcement learning and interpretable ai. *Social Science Research Network*, (3554486).

Dangl, T. and Halling, M. (2012). Predictive regressions with time-varying coefficients. *Journal of Financial Economics*, 106(1):157–181.

DeMiguel, V., Garlappi, L., and Uppal, R. (2009). How inefficient are simple asset allocation strategies. *Review of Financial Studies*, 22(5):1915–1953.

DeMiguel, V., Martin-Utrera, A., Nogales, F. J., and Uppal, R. (2020). A transaction-cost perspective on the multitude of firm characteristics. *The Review of Financial Studies*, 33(5):2180–2222.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 20(1):134–144.

Diebold, F. X. and Shin, M. (2019). Machine learning for regularized survey forecast combination: Partially-egalitarian lasso and its derivatives. *International Journal of Forecasting*, 35(4):1679–1691.

Dong, X., Li, Y., Rapach, D. E., and Zhou, G. (2022). Anomalies and the expected market return. *The Journal of Finance*, 77(1):639–681.

Duchin, R. and Levy, H. (2009). Markowitz versus the talmudic portfolio diversification strategies. *Journal of Portfolio Management*, 35:71–74.

Farmer, L., Schmidt, L., and Timmermann, A. (2022). Pockets of predictability. *Journal of Finance, forthcoming*.

Ferreira, M. A. and Santa-Clara, P. (2011). Forecasting stock market returns: The sum of the parts is more than the whole. *Journal of Financial Economics*, 100(3):514–537.

Freyberger, J., Neuhierl, A., and Weber, M. (2020). Dissecting characteristics nonparametrically. *The Review of Financial Studies*, 33(5):2326–2377.

Giraitis, L., Kapetanios, G., and Price, S. (2013). Adaptive forecasting in the presence of recent and ongoing structural change. *Journal of Econometrics*, 177(2):153–170.

Goyal, A., Welch, I., and Zafirov, A. (2021). A comprehensive look at the empirical performance of equity premium prediction ii. *Available at SSRN 3929119*.

Grammig, J., Hanenberg, C., Schlag, C., and Sönksen, J. (2021). Diverging roads: Theory-based vs. machine learning-implied stock risk premia. *Machine Learning-Implied Stock Risk Premia (December 17, 2021)*.

Gu, S., Kelly, B., and Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273.

Hyndman, R. J. and Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts.

Johannes, M., Korteweg, A., and Polson, N. (2014). Sequential learning, predictability, and optimal portfolio returns. *The Journal of Finance*, 69(2):611–644.

Kan, R., Wang, X., and Zhou, G. (2022). Optimal portfolio choice with estimation risk: No risk-free asset case. *Management Science*, 68(3):2047–2068.

Kan, R. and Zhou, G. (2007). Optimal portfolio choice with parameter uncertainty. *Journal of Financial and Quantitative Analysis*, 42(3):621–656.

Kazak, E. and Pohlmeier, W. (2022). Bagged pretested portfolio selection. *Journal of Business & Economic Statistics*, (just-accepted):1–35.

Kelly, B. T., Malamud, S., Zhou, K., et al. (2022). The virtue of complexity in machine learning portfolios. Technical report, Swiss Finance Institute.

Kirby, C. and Ostdiek, B. (2012). It's all in the timing: simple active portfolio strategies that outperform naive diversification. *Journal of Financial and Quantitative Analysis*, 47(2):437–467.

Lassance, N., Martin-Utrera, A., and Simaan, M. (2022). The risk of out-of-sample portfolio performance. *Available at SSRN 3855546*.

LeBlanc, M. and Tibshirani, R. (1996). Combining estimates in regression and classification. *Journal of the American Statistical Association*, 91(436):1641–1650.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411.

Ledoit, O. and Wolf, M. (2008). Robust performance hypothesis testing with the sharpe ratio. *Journal of Empirical Finance*, 15(5):850–859.

Leitch, G. and Tanner, J. E. (1991). Economic forecast evaluation: profits versus the conventional error measures. *The American Economic Review*, pages 580–590.

Liu, Y., Zhou, G., and Zhu, Y. (2022). Maximizing the sharpe ratio: A genetic programming approach. *Available at SSRN 3726609*.

Maasoumi, E., Tong, G., Wen, X., and Wu, K. (2022). Portfolio choice with subset combination of characteristics. *Available at SSRN*.

MacKinlay, A. and Pástor, L. (2000). Asset pricing models: Implications for expected returns and portfolio selection. *The Review of Financial Studies*, 13(4):883–916.

Markowitz, H. (1952). Portfolio selection. *Journal of Finance*, 7:77–91.

Nevasalmi, L. and Nyberg, H. (2021). Moving forward from predictive regressions: Boosting asset allocation decisions. *Available at SSRN 3623956*.

Paye, B. S. (2012). The economic value of estimated portfolio rules under general utility specifications. *Available at SSRN 1645419*.

Pettenuzzo, D. and Ravazzolo, F. (2016). Optimal portfolio choice under decision-based model combinations. *Journal of Applied Econometrics*, 31(7):1312–1332.

Polk, C., Thompson, S., and Vuolteenaho, T. (2006). Cross-sectional forecasts of the equity premium. *Journal of Financial Economics*, 81(1):101–141.

Polley, E. C. and Van Der Laan, M. J. (2010). Super learner in prediction.

Pyun, S. (2019). Variance risk in aggregate stock returns and time-varying return predictability. *Journal of Financial Economics*, 132(1):150–174.

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92(437):179–191.

Rapach, D. E., Strauss, J. K., and Zhou, G. (2010). Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. *The Review of Financial Studies*, 23(2):821–862.

Reuters, J. M. (1996). Riskmetrics-technical document. Technical report, Technical report, JP Morgan-Reuters.

Tu, J. and Zhou, G. (2011). Markowitz meets talmud: A combination of sophisticated and naive diversification strategies. *Journal of Financial Economics*, 99(1):204–215.

Van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1).

Welch, I. and Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, 21(4):1455–1508.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, 5(2):241–259.

Yang, Y. (2001). Adaptive regression by mixing. *Journal of the American Statistical Association*, 96(454):574–588.

Yuan, M. and Zhou, G. (2022). Why naive 1/n diversification is not so naive, and how to beat it? *Available at SSRN 3991279*.