Ensembles of Portfolio Rules

January 20, 2025

Abstract

We propose an innovative ensemble framework for combining conceptually distinct portfolio rules that existing combination methods cannot accommodate. This framework enables researchers and investors to harness established and emerging advances in portfolio choice by diversifying the idiosyncratic risks of alternative rules while leveraging their unique strengths. By maximizing the joint utility of candidate portfolio rules and dynamically adapting to their time-varying relative performance, our approach substantially enhances decision-making from an investor's perspective. Out-of-sample evaluations spanning over forty years reveal substantial utility gains in applications to a cross-section of stocks and to market timing strategies.

Keywords: Portfolio choice; Ensemble learning; Stacking

JEL classifications: G11, C10

1 Introduction

Over time, numerous portfolio rules (PRs) have been developed, offering a wide range of techniques for constructing portfolios over the cross-section of risky assets. Many of these approaches aim to address the empirical shortcomings of the seminal Mean-Variance (MV) framework proposed by Markowitz (1952). These contributions include shrinkage methods (e.g., Ledoit and Wolf 2004; Barroso and Saxena 2022), strategies leveraging the factor structures implied by asset pricing models (e.g., MacKinlay and Pástor 2000), and volatility timing strategies (e.g., Kirby and Ostdiek 2012). Additional advancements encompass parametric portfolio policies (e.g., Brandt et al. 2009; DeMiguel et al. 2020), risk-parity strategies (e.g., Maillard et al. 2010), and machine learning approaches that exploit asset characteristics (e.g., Gu et al. 2020; Cong et al. 2021; Chen et al. 2024). Moreover, DeMiguel et al. (2009), Duchin and Levy (2009), and Yuan and Zhou (2022) demonstrate that the $1/N$ rule, which avoids estimation error by disregarding sample information, often outperforms optimization-based rules in out-of-sample (OOS) scenarios.

Similarly, numerous PRs have been developed for optimal market timing—i.e., allocating between an aggregate equity portfolio and a risk-free asset. Some of these rules leverage macroeconomic data and financial ratios through combination methods (e.g., Rapach et al. 2010), a sum-of-the-parts approach (Ferreira and Santa-Clara, 2011), or Bayesian predictive models (e.g., Dangl and Halling 2012; Johannes et al. 2014; Pettenuzzo and Ravazzolo 2016). Others utilize forward-looking information embedded in option prices (e.g., Pyun 2019). Recent advancements have exploited long-short return anomalies in the cross-section of stocks, incorporating machine learning methods and shrinkage techniques (Dong et al., 2022).

Each PR, whether based on the approaches outlined above or on any other method, is defined by the

information set it utilizes and the econometric technique it employs to translate that information into asset weights for a given asset universe.[1] Fundamentally, each PR represents a specialized lens for navigating the asset universe. While PRs differ in their strengths and limitations, this diversity provides a compelling rationale for their combination.

From an economic perspective, combining PRs is similar to diversification across assets for investors with a concave utility function, as it helps mitigate the idiosyncratic risks inherent in individual PRs. These idiosyncratic risks may originate from modeling choices, as tackling an asset allocation problem involves multiple facets of uncertainty. In addition to the choice of the asset pricing rationale or portfolio theory underpinning the PR, researchers and investors must contend with at least three dimensions of uncertainty: (A) the data-generating process (DGP) assumed for asset returns, which may not align with the actual data; (B) the information set relied upon to implement a given PR, with the risk of overlooking valuable signals affecting portfolio performance; and (C) the econometric technique used to translate information into asset weights, which may fail to capture patterns critical for effective portfolio allocation and/or may be vulnerable to estimation risk (Frost and Savarino, 1986; Kan and Zhou, 2007; Tu and Zhou, 2011).[2]

From a statistical perspective, PRs can be seen as estimators, and in various contexts, combining estimators has been shown to be both theoretically sound and effective in empirical applications. Therefore, combining PRs is a robust strategy, avoiding the pitfalls of relying on a single PR in pursuit

[1] In our proposed approach, we take the asset weights implied by each candidate PR as given. As shown in Section 3.1, the asset weights resulting from the combination of PRs can be readily derived.

[2] To illustrate, consider the volatility timing strategy of Kirby and Ostdiek (2012) alongside the factor structure approach of MacKinlay and Pástor (2000): Regarding (A), Kirby and Ostdiek (2012) make no assumptions about the DGP, whereas MacKinlay and Pástor (2000) assume that asset returns follow an exact linear factor structure, but one of the factors is unobserved. For (B), Kirby and Ostdiek (2012) base their strategy on historical stock volatility, while a PR derived from MacKinlay and Pástor (2000) relies on the covariance structure and expected return estimates. Concerning (C), MacKinlay and Pástor (2000) apply an econometric framework that links expected returns to the residual covariance matrix under the factor structure assumption. In contrast, Kirby and Ostdiek (2012) adjust portfolio weights using a tuning parameter that reflects the degree of timing aggressiveness.

of a "silver bullet" while prematurely dismissing alternatives.

The existing literature has developed several combination approaches aimed specifically at controlling estimation risk and improving OOS performance. However, these approaches are generally applicable only to specific and limited sets of PRs, typically within the MV or Global Minimum Variance (GMV) framework, often including the $1/N$ rule (e.g., Kan and Zhou 2007; Tu and Zhou 2011; Kan et al. 2022; Lassance et al. 2023). To the best of our knowledge, no general utility-based optimization framework exists for combining a broad and conceptually distinct set of PRs. For example, for a cross-section of assets, there is no utility maximization framework that can integrate a shrinkage-based method such as Barroso and Saxena (2022) with a factor-structure-based strategy like MacKinlay and Pástor (2000), or with rules derived from cross-sectional characteristics (e.g., Brandt et al. 2009; Gu et al. 2020). Similarly, existing methods cannot combine market timing rules based on point forecasts for the equity premium, such as the sum-of-the-parts approach of Ferreira and Santa-Clara (2011), with Bayesian density forecast rules like Dangl and Halling (2012), or with methods exploiting cross-sectional characteristics (Dong et al., 2022). Furthermore, while proposed combination methods can themselves be considered as candidate PRs and, thus, be combined with other PRs, those existing approaches do not allow for such an additional layer. Additionally, determining the optimal combination typically relies on specific distributional assumptions about the asset return generating process. Finally, there is no optimization framework to adapt the combination to changing market environments, as certain (combinations of) PRs may outperform at certain points in time, while others may have the edge during other periods.

Overall, existing combination strategies provide a relatively narrow toolkit for diversifying across PRs and enhancing asset allocation performance. As a result, they do not fully capitalize on the complementary strengths of the diverse solutions that have been or could be proposed. The goal of this

work is to address the above-mentioned shortcomings.

The main novelty of our framework lies in its generality, enabling the combination of an arbitrary number of conceptually distinct PRs while retaining many appealing features, as detailed below in *i*) through *iv*). Importantly, our proposed framework should not be seen as a replacement for the above-mentioned combination approaches, which target specific PRs (e.g., MV or GMV). Rather, these combination rules themselves can be added to the pool of candidate PRs and combined with other differently designed rules (e.g., volatility timing or risk parity) within our ensemble. Section 5 elaborates on this key aspect and provides empirical support.

In our framework, the investor has access to a library of candidate PRs. In each period, they select a combination of PRs that would have maximized their pseudo OOS utility.[3] By optimizing the utility derived from combining PRs, our approach:

i) Relies on the pseudo OOS returns of the candidate PRs. The optimal combination of PRs in our framework is based on OOS utility gains. To implement this approach, only the assigned asset weights of the candidate PRs and their corresponding pseudo OOS returns are required. This enables the combination of conceptually distinct PRs, as each provides a weight vector for asset allocation, while avoiding the need to predict the moments of the PRs' returns.[4] Importantly, the candidate PRs themselves may or may not rely on estimated moments of asset returns in their construction. Each candidate PR is entirely free to adopt any rationale for assigning asset weights.

[3]We focus on economic utility in the objective function rather than a statistical criterion. Statistical and economic evaluation criteria are not necessarily closely aligned. Leitch and Tanner (1991) demonstrate that accurate predictions based on statistical measures, such as the root mean squared error, can lead to unprofitable portfolio allocations. Similarly, Cenesizoglu and Timmermann (2012) find only a weak relationship between economic utility measures and statistical forecast accuracy in an application to equity premium forecasts.

[4]Alternatively, one could consider using the realized PRs' returns and maximize an objective function based on their expected moments. However, this approach would require the estimated moments of the PRs' returns as inputs for determining the combination weights. With an increasing number of PRs, this approach would become more and more prone to estimation error. Further, in our proposed framework, down-weighting older data can be applied to economic utility, thus directly focusing on the investor's objective. By comparison, in this alternative approach, down-weighting older observations would have to be applied indirectly via the estimated moments of the PRs' returns.

ii) Functions as an ensemble framework. Our approach assigns combination weights based on the pseudo OOS utility generated *jointly* by the candidate PRs, rather than the utility produced by individual PRs, as is done, for example, in Pettenuzzo and Ravazzolo (2016). Similar to assembling a sports team, the framework does not necessarily select the best individual players but prioritizes building the strongest overall team. This ensemble approach inherently accounts for the time-varying interdependencies among the pseudo OOS returns of the PRs, including correlations and higher-order co-moments. Importantly, there is no conceptual limit to the number of PRs that can be included in the ensemble.

iii) Enables adaptive learning. Our approach incorporates a forgetting factor that emphasizes recent economic gains over profitability from the more distant past. This adaptive learning mechanism facilitates the dynamic adjustment of combination weights, allowing for rapid shifts when warranted by empirical evidence. At the level of asset returns, Farmer et al. (2023) document short periods of predictability for aggregate stock returns by specific predictors, interspersed with longer periods of unpredictability. Similarly, our framework is designed to identify and leverage (combinations of) PRs that perform well "locally" in time.

iv) Makes no assumptions about the DGP for assets' or PRs' returns. Our approach for determining combination weights does not rely on any assumptions about the return-generating process. However, it is important to recognize that individual candidate PRs may or may not incorporate such assumptions about the DGP for assets' returns. Even if these assumptions are not literally true, such PRs can still contribute value to the ensemble.

We apply our framework to two classic portfolio choice problems: allocating across a cross-section of U.S. stocks and allocating between the S&P 500 index and Treasury bills. To enhance diversification benefits, we construct a pool of candidate PRs, incorporating both established and emerging methods.

Using over forty years of OOS evaluations and controlling for transaction costs, we find substantial utility gains from combining PRs. The utility generated by our framework is higher than, or comparable to, that of the (ex-post) best-performing candidate PR. Moreover, our approach outperforms previously proposed combination methods.

We also empirically highlight the benefits of the ensemble framework and adaptive learning. Combination weights shift rapidly over time, demonstrating that relative performance varies importantly across periods. Analyzing the combination weights across various market and economic states reveals that simpler PRs tend to be favored during stable economic periods, while more complex PRs tend to be selected during turbulent times. In our application to a cross-section of stocks, we observe that the PR performing best individually receives the lowest overall weight in the ensemble—a reflection of the framework's emphasis on *jointly* generated utility. In our market timing application, accounting for complex return-generating processes—through the integration of multiple predictors, time-varying coefficients, and stochastic volatility—consistently generates a clear performance advantage.

Deeper analyses provide insights into the mechanisms driving utility gains and the potential for expanding the pool of candidate PRs. These analyses reveal that our method, by maximizing utility, selects PR combinations that balance predictive power for asset returns with the ability to anticipate their variance. Furthermore, average utility gains increase as more PRs are included in the pool, suggesting potential for further improvement by expanding the set of candidate PRs.

Our applications are designed to demonstrate the effectiveness of the methodological framework we propose. Importantly, we do not endorse any specific candidate PR. Instead, we emphasize that researchers and investors can adapt our framework to their preferred sets of PRs. The primary contributions of our study are, therefore, methodological. Specifically, we propose a general framework that leverages the relative strengths of various solutions to portfolio choice problems and mitigates the

idiosyncratic risks associated with relying on a single PR. By integrating PRs from diverse strands of the literature, the framework serves as a unifying tool, bridging approaches that might otherwise develop along increasingly siloed paths.

We believe this combination framework has the capacity to reshape the way portfolio choice problems will be empirically addressed going forward. Rather than focusing on identifying a single superior PR, the emphasis shifts toward discovering PRs that complement the ensemble and enhance its overall performance.

The remainder of the paper is organized as follows: Section 2 positions our approach within the existing literature. Section 3 details our methodology. Section 4 presents the empirical analysis. Section 5 highlights the relative strengths of our combination method, and Section 6 concludes and highlights promising directions for future research.

2 Relation to the literature

Our work connects to two primary strands of the portfolio choice literature. First, it aligns with combination approaches for PRs. In this context, studies such as Kan and Zhou (2007), Tu and Zhou (2011), and Kan et al. (2022) have developed strategies to optimize OOS performance under estimation risk in MV portfolio choice problems. Specifically, Kan and Zhou (2007) proposed an optimal three-fund rule—comprising the risk-free asset, the sample MV portfolio, and the sample GMV portfolio—to maximize expected OOS utility. Building on the insight that combining a simple method with a sophisticated one can balance the bias-variance trade-off, Tu and Zhou (2011) integrated this three-fund portfolio with the $1/N$ rule. Kan et al. (2022) extended this analysis to scenarios without a risk-free asset, while Lassance et al. (2023) explored combining the $1/N$ rule with the sample MV

portfolio.

The optimal combination rules proposed in the cited works have provided valuable analytical insights into portfolio construction under estimation error, operating under the assumption that asset returns are identically and independently distributed as a multivariate normal. In contrast, our proposed framework makes no assumptions about the return-generating process. More significantly, it is not constrained to specific designs of PRs. Instead, our approach accommodates the combination of conceptually distinct PRs, including those previously incompatible with existing methods due to their diversity.

While existing combination approaches primarily focus on reducing estimation risk by combining just two different PRs, our method tackles this challenge through several additional strategies. Specifically, it avoids reliance on estimates of the moments of PRs' returns, employs a parsimoniously parameterized framework based on pseudo-OOS returns rather than expected returns, and allows for additional regularization of the combination weights. Notably, the lack of effective methods for managing estimation error at the combination level appears to be a key reason why prior research has historically focused on combining only two PRs. As Tu and Zhou 2011 observe, "*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 performance.*"

In our empirical applications, we demonstrate that combining five or six PRs within our framework generally outperforms combining only two rules—while leaving ample room to integrate even more candidate PRs. However, the strength of our framework extends well beyond mitigating estimation risk.

By enabling the combination of PRs based on diverse premises—spanning modeling assumptions, information sets, and methods for translating information into asset weights—our approach offers a significantly broader and more versatile foundation for dealing with model uncertainty in portfolio choice. Consequently, our framework allows researchers and investors to fully leverage the potential of both existing and emerging advancements in asset allocation.

Importantly, our method is not a competing alternative to the aforementioned works but rather a complementary approach. Combination methods themselves can be treated as PRs and included as candidates within our ensemble, provided they align with the specific investment problem. For instance, we incorporate previously proposed combination methods, such as those in Kan et al. (2022), into our empirical analysis and demonstrate that their performance improves when combined with additional PRs. For a comprehensive conceptual and empirical discussion of how our approach relates to existing PRs, see Section 5.

Second, our work aligns with approaches that directly optimize economic utility, bypassing the two-step process that requires estimating moments of returns. Examples include parametric portfolio policies (Brandt et al., 2009; DeMiguel et al., 2020), a boosting approach (Nevasalmi and Nyberg, 2021), and a method utilizing deep reinforcement learning (Cong et al., 2021). These techniques focus on optimizing economic utility for specific portfolio choice problems at the individual asset level, whereas our approach operates at a higher level by maximizing utility through the combination of PRs. We are not aware of existing frameworks that allow to extend those approaches to the PRs' level. Furthermore, our work complements these methods, as they can be included as candidate PRs within our framework and combined with other PRs.

3 Methodology

3.1 Basic structure

Consider a set of M candidate PRs, indexed by $m = 1, \ldots, M$. At any given time s, each PR assigns weights to N assets, indexed by $n = 1, \ldots, N$, based on information observed up to $s-1$, the date of portfolio construction.[5] We denote the (exogenously) assigned asset weights of the m-th PR at time s as $\omega_{m,s}^{(-s)}$, an $N \times 1$ column vector given by $\left(\omega_{m,s,1}^{(-s)}, \ldots, \omega_{m,s,N}^{(-s)}\right)'$. The superscript $(-s)$ indicates that information revealed at time s is not used to determine the portfolio allocation at $s-1$.

The $N \times 1$ column vector of gross asset returns measured over the period $[s-1 : s]$ (e.g., one month in our applications) is denoted as $\widetilde{\mathbf{R}}_s = \left(\widetilde{R}_{s,1}, \ldots, \widetilde{R}_{s,N}\right)'$, where $\widetilde{R}_{s,n} = 1 + \widetilde{r}_{s,n}$, and $\widetilde{\mathbf{r}}_s = \left(\widetilde{r}_{s,1}, \ldots, \widetilde{r}_{s,N}\right)'$.[6] The pseudo OOS gross return for the m-th PR at time s is expressed as:

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

The investor's optimization problem is to maximize the conditional expected utility of the portfolio's (gross) return, $R_{p,t}$, based on information available up to time $t-1$. This is achieved by determining the optimal combination weights $\{w_{m,t}\}_{m=1}^{M}$ assigned to the PRs:

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

where $U(\cdot)$ represents the utility function.

[5] In this paper, we focus on PRs that allocate within the same investment opportunity set. However, our framework can accommodate PRs that allocate across different opportunity sets with partial or no overlap in assets. For example, one PR might allocate across stocks, while another PR might allocate across commodities.

[6] Depending on the specific portfolio choice problem, returns can be defined as raw (total) returns or excess returns.

The conditional expectation of the utility jointly generated by the returns of the PRs, as expressed in Equation (2), cannot generally be computed without imposing restrictive constraints on the PRs' return-generating process. In addition, several PRs, such as the $1/N$ rule, may not even generate a conditional distribution of expected returns. To address this, we technically treat the combination weights as constant over time, implying that the weights maximizing the conditional expected utility at any given moment are constant across all previous periods. As a result, we can reformulate Equation (2) as an unconditional optimization problem.

At time $t-1$, if a track record of pseudo OOS returns generated by the PRs is available over the interval $[\tau : t-1]$, we replace the expected utility in Equation (2) using its sample counterpart—the sum of period-by-period realized utilities. The optimization problem thus 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 $\mathbf{w}_t = (w_{1,t}, \ldots, w_{M,t})'$. This unconditional formulation of the optimization problem bypasses the need to estimate (co)moments of the PRs' returns. Furthermore, by replacing Equation (2) with pseudo OOS returns, we mitigate the impact of discrepancies between expected and realized returns—a frequent challenge in asset allocation.

More recent data are likely to carry greater predictive relevance than older data, as they originate from a market or economic environment more similar to the current one. To account for these plausible economic dynamics, we allow realized joint utilities to be weighted differently in the optimization process. Specifically, we maximize the weighted historical 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

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

Here, α represents a (fixed) forgetting factor used to weight past profitability, while the constraints in Equation (5) ensure a convex combination of the candidate PRs. By applying exponential down-weighting to past performance and re-optimizing at each point in time, this approach enables the adaptive selection of combination weights, even though they are technically treated as constant over the interval $[\tau : t-1]$ within each optimization at time t.

This approach allows us to account for the relative strengths of candidate PRs over specific time periods, enabling more rapid weight adjustments compared to the unweighted formulation in Equation (3). The forgetting factor and the weight constraints will be discussed in greater detail in Sections 3.4 and 3.5, respectively.

For purposes such as executing trades and calculating transaction costs, it is essential to determine the asset weights resulting from the combination of PRs. With the optimized combination weights at hand, we can derive the implied weights for the N assets. These weights, ω_s^*, are expressed as linear combinations of the asset weights assigned by the PRs (summarized in the matrix $\mathbf{\Omega}_s$) and the optimized combination weights $\mathbf{w}_s^*$:

$$\underset{[N\times 1]}{\omega_s^*} = \underset{[N\times M]}{\mathbf{\Omega}_s} \cdot \underset{[M\times 1]}{\mathbf{w}_s^*}, \tag{6}$$

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 utility of aggregate individual asset weights becomes apparent when the positions for a given asset, as implied by the candidate PRs, partially or fully offset one another. Instead of trading the positions implied by each PR individually, an execution desk trades the net positions derived from the combined PRs, thereby reducing transaction costs.

3.2 Power utility and alternative utility functions

In our empirical applications, we assume a power utility investor. By adopting power utility preferences at the level of combining PRs, the framework inherently accounts for preferences related to higher-order moments and tail risk properties. This holds true even if the candidate PRs in the library do not optimize for power utility but instead rely on MV approximations or are not based on any optimization framework. Requiring candidate PRs to optimize for power utility from the outset would exclude a large portion of promising PRs. Importantly, PRs not specifically designed to maximize power utility preferences can still make valuable contributions to the ensemble (Paye, 2012).

For an investor with power utility preferences, the optimization problem in Equation (4) can be more specifically expressed as:

$$\mathbf{w}_t^* = \underset{\{w_m\}_{m=1}^M}{\arg\max} \sum_{s=\tau}^{t-1} \alpha^{t-1-s} \cdot \frac{\left(\sum_{m=1}^{M} w_m R_{m,s}\right)^{1-\gamma}}{1-\gamma}, \tag{7}$$

where γ denotes the relative risk aversion coefficient.

Our framework accommodates alternative utility functions, allowing flexibility in its application. Specifically, it can incorporate any utility function that can be expressed as the sum of discounted

utilities. For instance, a mean-variance utility with discounting can be used:

$$\mathbf{w}_t^* = \underset{\{w_m\}_{m=1}^M}{\arg\max} \sum_{s=\tau}^{t-1} \alpha^{t-1-s} \left\{ \sum_{m=1}^{M} w_m r_{m,s} - \frac{\gamma}{2} \left(w_m r_{m,s} \right)^2 \right\}, \tag{8}$$

where $r_{m,s} = \omega_{m,s}^{(-s)'} \widetilde{\mathbf{r}}_s$. Alternatively, utility functions that emphasize downside risk can also be incorporated.

3.3 Our combination framework as a stacking algorithm

Our proposed combination framework, as expressed in Equation (7), can be classified as a stacking algorithm. Stacking is a well-established meta-learning approach for combining estimators, widely studied in the machine learning and statistics literature (Wolpert, 1992; LeBlanc and Tibshirani, 1996; Breiman, 1996; Van der Laan et al., 2007; Polley and Van Der Laan, 2010). We adapt and extend this approach to maximize a utility-based criterion within a time-series framework, incorporating exponential discounting to account for the temporal structure of the data.

Stacking is an ensemble method that evaluates the cross-validated risk or utility of the combined candidate estimators (in this case, the PRs) rather than assessing their risk or utility in isolation. As a result, the combination weights in Equation (7) are derived from an ensemble perspective, inherently capturing time-varying interdependencies among PRs' returns. Under power utility preferences, the entire joint distribution of PRs' returns is utilized in Equation (7) to optimize the combination weights.

Another important feature of stacking is its use of cross-validation to prevent overfitting. Our combination approach relies on pseudo OOS returns, which requires adjustments for the time-series structure of the data. Consequently, standard K-fold cross-validation cannot be applied directly. Instead, our approach is akin to leave-one-out cross-validation, omitting information revealed at time s

for portfolio construction at time $s-1$; see Equation (1). Figure 1 illustrates the general 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 to determine the next period's portfolio allocation, while the red dots represent the resulting pseudo OOS returns. Our method is centered on maximizing the utility derived from these red dots.

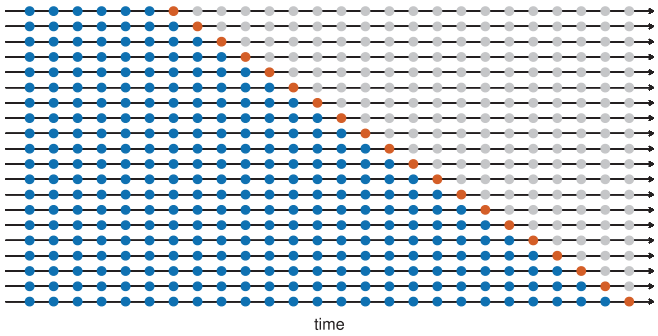

Figure 1: Schematic illustration of leave-one-out cross-validation. The illustration is adopted from Hyndman and Athanasopoulos (2018).

Including information revealed at time s when determining the allocation at time $s-1$ would result in in-sample returns. In such a scenario, the combination would likely assign all weight to the PR with the highest in-sample returns. However, PRs that perform well in-sample may exhibit poor OOS performance.

Stacking is a true combination method rather than a selection approach. This implies that, even asymptotically, positive combination weights can be distributed across multiple PRs rather than being concentrated solely on the most successful PR in the library. This characteristic is particularly appealing in the realistic scenario where none of the candidate PRs fully captures the true DGP. However, if a single candidate PR dominates all possible combinations, that PR will receive the entire weight. In this way, selection is effectively nested as a special case within the stacking framework.

Stacking algorithms are grounded in a robust statistical foundation. Under certain conditions,

Van der Laan et al. (2007) demonstrated their asymptotic oracle performance, meaning that the learning algorithm asymptotically matches the performance of the best possible ex-post choice for a given dataset, based on the defined evaluation criterion, among all weighted combinations of the estimators. Beyond these theoretical guarantees, stacking-based learning algorithms have been shown to be adaptive and robust estimators, even in small sample settings, across 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 many cases, they perform as well as, or even better than, the ex-post best candidate estimator.

By adopting a stacking algorithm, our combination framework leverages a methodology with excellent statistical properties, providing a strong theoretical justification for our approach.

3.4 The forgetting factor

The exponential forgetting factor $\alpha \leq 1$ in (7) emphasizes the recent history of past performance. In our empirical analysis, we adaptively select α from the grid $\mathscr{S}_\alpha = \{0.90{:}0.01{:}1.00\}$. The data-adaptive estimation of the forgetting factor follows Giraitis et al. (2013), who demonstrate that exponential down-weighting with a data-driven forgetting factor is a robust approach for addressing structural changes in time series, based on extensive simulations and empirical studies.

The lower α, the more heavily performance in the distant past is down-weighted. For instance, with monthly data, if $\alpha = 0.99$, economic utility three years ago retains approximately 70% of the weight given to last month's utility. We set $\alpha = 0.90$ as the lower limit of the grid, as this value implies extremely rapid forgetting: utility from three years ago receives only about 2% of the weight assigned to utility from the most recent period. The effective window size is given by $1/(1-\alpha)$,

which corresponds to 10 months for $\alpha = 0.90$.[7]

FLEXPOOL refers to the combination in which the value of α is adaptively selected in each period from the grid $\mathscr{S}_\alpha$. In our empirical analysis, we also consider two additional benchmark combinations: STATPOOL, where $\alpha = 1$, and a second approach that assigns equal weights to the PRs, irrespective of their past performance.

By enabling the down-weighting of older performance, the combination weights can be rapidly adjusted to changing environments when empirically warranted. Another advantage of the forgetting factor approach is its ability to parsimoniously capture evolving dynamics using a single parameter. This simplicity makes it less susceptible to estimation errors compared to parameter-rich alternatives like regime-switching models. At each point in time, the optimal time-dependent value α_t is selected from the grid $\mathscr{S}_\alpha$ as the one that has generated the highest utility over the interval $[\tau^* : t-1]$:

$$\alpha_t^* = \underset{\alpha \in \mathscr{S}_\alpha}{\arg\max} \sum_{s=\tau^*}^{t-1} U\left[\left(\mathbf{w}_{t-1}^*(\alpha)\right)' \mathbf{R}_s\right]. \tag{9}$$

Here, $\tau^* = \tau + \tau_0$, where τ_0 denotes the number of observations reserved for the initial optimization of the combination weights. The vector $\mathbf{R}_s = (R_{1,s}, R_{2,s}, \ldots, R_{M,s})'$ represents the PRs' returns, and $\mathbf{w}_{t-1}^*(\alpha)$ denotes the optimal combination weights derived from Equation (7), conditional on a given value of α.

It is important to note that down-weighting is applied only when maximizing the combination weights in Equation (7) for a specific value of α. The selection among different values of α, however, is based on the recursive evaluation described in Equation (9).

[7] Note that $\sum_{s=0}^{\infty} \alpha^s = \frac{1}{1-\alpha}$ for $\alpha < 1$. The upper bound $\alpha = 1$ implies no down-weighting of older data, nesting the standard recursive window estimation as a special case.

3.5 Weight restrictions and additional regularization

Although stacking imposes no inherent restrictions on the combination weights, convex combinations of estimators have been shown to enhance the stability of the final estimator (see, e.g., Breiman, 1996; Van der Laan et al., 2007). Additionally, ensuring convexity allows us to maintain any constraints on asset weights imposed at the level of the candidate PRs (e.g., no short selling, sector weight restrictions, etc.) at the level of the combined PRs as well.

Although our proposed combination approach leverages pseudo OOS returns and is parsimoniously parameterized, estimation risk at the stage of determining combination weights remains a concern in finite samples. There is no guarantee that the optimized combination weights will outperform simple benchmarks, such as equally weighted PRs. In fact, in finite samples, including more candidate PRs does not necessarily result in better performance. While adding PRs can enhance diversification potential, it also increases estimation risk due to the greater number of combination weights that need to be estimated.

The optimal number of PRs to include in the library is ultimately an empirical question, influenced by the return-risk profiles of the candidate PRs and the interdependence structure of their returns. To address the estimation risk of combination weights, our framework conveniently supports the direct application of regularization techniques to mitigate these challenges.

As a regularization strategy, an ℓ_0-constraint can be imposed on the combination weights, for example:

$$\|\mathbf{w}\|_0 = k. \tag{10}$$

Here, $\|\mathbf{w}\|_0 = \sum_{m=1}^{M} \mathbb{1}(w_m \neq 0)$ counts the number of the non-zero combination weights, where $\mathbb{1}(\cdot)$ denotes the indicator function. The tuning parameter k, with $k \leq M$, controls the size of the subset of PRs that are combined. Empirically, k can be determined in a data-adaptive manner, selecting the value that maximizes pseudo-real-time OOS performance.

To further regularize the combination weights, one could eliminate the need for estimating the combination weights altogether by imposing an additional constraint alongside (10)—that all non-zero weights are equal. This can be achieved by introducing:

$$w_m \in \left\{0, \frac{1}{k}\right\}, m = 1, \ldots, M, \tag{11}$$

The use of equally weighted subsets of PRs in constraint (10) is motivated by the empirical success of equally weighted subsets in combination approaches, particularly in forecast combinations; see, e.g., Dong et al. (2022). As an alternative for regularization, the ℓ_1-constraint can be employed instead of the ℓ_0-constraint in (10). When the ℓ_1-constraint is combined with the condition that all non-zero weights are equal (11), this approach serves as a computational shortcut to equally weighted subsets (Diebold and Shin, 2019).[8]

In principle, there is no upper limit to the number of PRs that can be included in the analysis. In our main empirical applications, which involve five to six PRs, we imposed only convex combination weights, as specified in constraint (5). On average, economic gains increased with the number of PRs included when we additionally applied constraint (10); see Figure 4 for our cross-section application

[8] A complementary robustness strategy to additional weight constraints involves using bootstrapped returns up to a given point in time to estimate combination weights at that point; see, e.g., Bonaccolto and Paterlini (2020) and Kazak and Pohlmeier (2023). While this strategy could still produce adaptive combination weights, it may reduce the flexibility to rapidly adjust the weights. Block bootstrap methods, however, could be adapted to allow for exponential down-weighting of older performance by introducing an additional tuning parameter. We leave this extension for future research.

with the largest 50 stocks and Figure 6 for the market timing application.

We also experimented with the additional weight constraints (10) and (11), selecting k in a pseudo-real-time, data-driven manner. While we did not observe any empirical gains from these additional regularization constraints in our settings, they may prove highly beneficial in applications involving a larger number of candidate PRs.

4 Empirical analysis

4.1 Application to a cross-section of stocks

4.1.1 Investment universe and empirical study design

The investment universe in this application consists of the largest 50 U.S. stocks, with monthly excess returns constructed from CRSP data. We use data spanning $1957{:}01$ to $2020{:}12$, including only stocks listed on the NYSE, NASDAQ, or AMEX with a code of 10 or 11. At the beginning of each month t, the investment universe comprises the largest 50 stocks (by market value) with non-missing monthly returns over the previous 120 months. In the rare cases where a stock's return is missing for month t, we set the excess return to zero.[9] It is important to note that the largest 50 stocks can change from month to month, making the investment universe dynamic.

Each candidate PR in our library must assign weights to the 50 stocks at the beginning of each month. The first OOS portfolio returns are obtained in $1967{:}01$. We reserve the initial 60 OOS returns for the optimization of the PR weights according to (7) and an additional 60 months for the initial tuning of the forgetting factor α as described in (9). Our first OOS evaluation takes place in January

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

1977. From that point onward, we proceed iteratively: the optimization is updated based on an extended sample of OOS portfolio returns, and the forgetting factor is chosen using an additional observation. This process is repeated recursively, resulting in an evaluation sample that spans 1977:01 to 2020:12. We consider a power utility investor with a relative risk aversion of $\gamma = 3$. Our setup focuses solely on risky assets. If we wished to include a risk-free asset, we could do so by adding a candidate PR represented by a vector of zeros, as the returns in this application are defined as excess returns.

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 rely on sample information, thereby avoiding estimation error. Empirical studies have shown that it often outperforms a wide range of estimated optimal portfolios across various data sets (DeMiguel et al., 2009; Duchin and Levy, 2009; Yuan and Zhou, 2022).

- Volatility timing (VOLTIME):

  Kirby and Ostdiek (2012) propose a volatility-timing strategy in which the weights are determined as follows:

$$\omega_{t+1,n} = \frac{\left(1/\widehat{\sigma}^2_{t+1,n}\right)^{\eta}}{\sum_{n=1}^{N}\left(1/\widehat{\sigma}^2_{t+1,n}\right)^{\eta}}, \quad n = 1, ..., N. \tag{12}$$

  Here, $\widehat{\sigma}^2_{t+1,n}$ denotes the estimated conditional variance of the n-th risky asset at time $t+1$, calculated using a rolling window of past returns from $t-119$ to t. This PR disregards any sample information regarding conditional means and covariances. The parameter η controls the

aggressiveness of the timing strategy, specifically the tilt toward the least volatile stocks. Kirby and Ostdiek (2012) consider the values $\eta = 1, 2$, and 4; we set $\eta = 4$.

- Maximizing expected OOS utility: Kan et al. (2022) develop combination portfolios that achieve the highest expected OOS utility for a MV investor in a setting without a risk-free asset. Their approach combines the GMV portfolio with a sample zero-investment portfolio, explicitly accounting for estimation risk to control the exposure to the sample zero-investment portfolio. When this exposure is set to zero, the GMV portfolio is nested as a special case. The method proposed by Kan et al. (2022) can also incorporate refined estimates of expected returns and (co-)variances, such as those obtained via shrinkage estimators or a single-factor structure, to form optimal portfolios. We consider the following two specifications of their approach:[10]

  - Kan et al. (2022) combined with MacKinlay and Pástor (2000) (KWZ - MP):

    MacKinlay and Pástor (2000) utilize the implications of an asset pricing model with a single risk factor to estimate expected returns. This approach reduces the number of parameters that need to be estimated, thereby mitigating estimation risk.[11]

  - Kan et al. (2022) combined with Ledoit and Wolf (2004) (KWZ - LW):

    Ledoit and Wolf (2004) propose a shrinkage estimator of the covariance matrix that involves a linear combination of the sample covariance matrix and the identity matrix.[12]

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

[11] The weights for this rule are given by Equation (51) in Kan et al. (2022).

[12] The weights for this rule are given by Equation (43) in Kan et al. (2022). KWZ-MP and KWZ-LW are combination rules based on the assumption that returns are identically and independently multivariate normally distributed.

- Galton-Shrinkage (GALTON):

  Barroso and Saxena (2022) propose a shrinkage estimator that leverages the structure of past OOS forecast errors to adjust the expected returns and expected (co-)variances used as inputs for portfolio optimization. The corrected inputs are then used to compute the Galton MV portfolio, whose weights result from a straightforward Markowitz optimization applied to these adjusted inputs.[13] The key formula for correcting the optimization inputs is:

$$\mathbf{Z}_t = \widehat{g}_0 + \widehat{g}_1 \mathbf{Z}_{t-1}, \tag{13}$$

  where, for each variable $\mathbf{Z}$ of interest (mean returns, variances or pairwise correlations), $\mathbf{Z}_{t-1}$ denotes its historical estimate at time $t-1$ computed using a rolling window of 60 observations. $\mathbf{Z}_t$ represents the cleansed portfolio input for t. Fama-MacBeth regressions are employed to estimate the Galton shrinkage coefficients g_0 and g_1 for the means, variances and pairwise correlations. These regressions are conducted on a large estimation universe comprising the 500 largest US stocks at each point in time, allowing for robust learning.[14]

  To run the Fama-MacBeth regressions, we use 12 ex-post realizations, and to initialize the Galton coefficients, we allocate an additional learning period of 108 months. In the notation of Barroso and Saxena (2022), $H = 60$, $E = 12$ and $L = 108$. See Equations (9) to (13) in Barroso and Saxena (2022) for details on estimating the Galton coefficients. The slope coefficient in (13) determines the degree of shrinkage. At one extreme, if its estimated value is 1, the corrected input equals the historical estimates, i.e., the uncorrected estimates. At the other extreme, if its

[13] The weights for this rule are computed according to Equation (7) in Barroso and Saxena (2022).
[14] Barroso and Saxena (2022) consider both larger and smaller estimation universes and find similar results.

estimated value is 0, the historical estimates are deemed completely unreliable, and the corrected inputs are set to the grand mean of the returns, variances, or pairwise correlations observed up to time t. Note that we restrict the slope coefficients to lie between 0 and 1. 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 $g_{1,corr}$ yield 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 Baseline results

Table 1 presents the results for both the candidate PRs and the combined PRs. Given our focus on economic utility, the certainty equivalent return (CER) is a natural choice for measuring portfolio performance. We report annualized CERs[15] without transaction costs (CER) as well as with proportional transaction costs (CER^{TC}) of 20 bps, in line with common choices in the literature; see, e.g., Kan et al. (2022). We further report the annualized Sharpe ratio without transaction costs (SR) and with proportional transaction costs (SR^{TC}) of 20 bps. As a measure of downside risk, we include the maximum drawdown after transaction costs of 20 bps (MaxDD^{TC}), given its relevance for asset managers and fiduciaries as emphasized, by van Hemert et al. (2020), among others. Finally, we report the average monthly turnover (Avg. TO).

[15] CERs are computed over the evaluation sample from τ^{**} to T as

$$CER = \left\{ \left\{ (1-\gamma) \frac{1}{T-\tau^{**}+1} \sum_{s=\tau^{**}}^{T} U\left(\mathbf{w}_s^* \left(\alpha_s^*\right)' \mathbf{R}_s\right) \right\}^{\frac{1}{1-\gamma}} - 1 \right\} \times 1,200. \tag{14}$$

The key findings from Table 1 can be summarized as follows. FLEXPOOL achieved the highest CER and SR, both before and after transaction costs, outperforming not only each candidate PR but also the alternative combination schemes. Notably, FLEXPOOL nearly doubled the CER compared to the $1/N$ rule over an OOS evaluation period of 44 years. Furthermore, FLEXPOOL outperformed STATPOOL in terms of CER, SR, and maximum drawdown. This result highlights the importance of placing greater weight on recent utility when determining combination weights. Throughout the sample, distant utility was consistently down-weighted, as reflected by a forgetting factor α that fluctuated between 0.93 and 0.96 (see Figure 2). In terms of downside risk, VOLTIME had the lowest maximum drawdown at 0.369, with FLEXPOOL being a close second at 0.379. When comparing FLEXPOOL to each candidate PR and alternative combined PRs, the outperformance in terms of CERs and SRs is statistically significant at least at the 10% level, based on one-tailed tests of Diebold and Mariano (1995) and Ledoit and Wolf (2008), respectively.

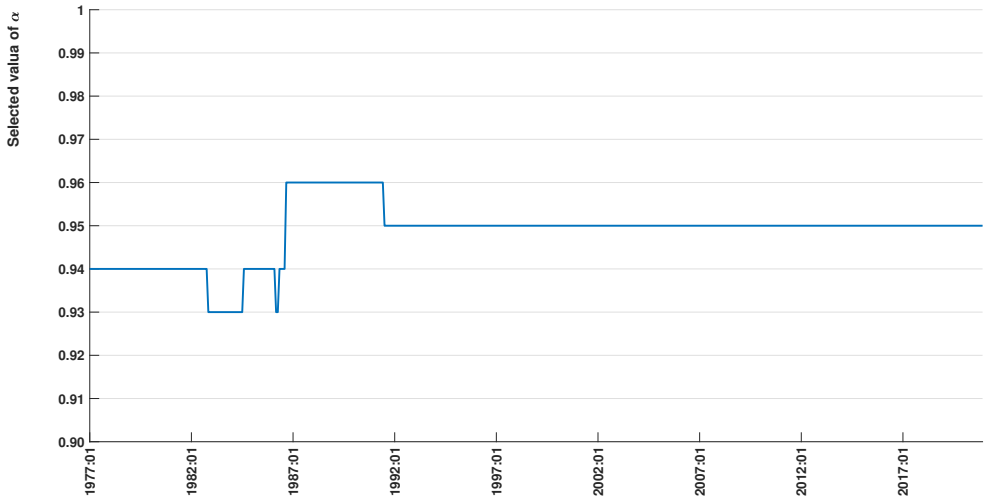

Figure 2: Evolution of the selected forgetting factor α in FLEXPOOL.

What combination weights were assigned to the different PRs, and how did they evolve over time? Figure 3 provides the answers. The subplot in the upper left corner of the figure displays the average weight shares of the PRs over the evaluation sample. The blue (red) bars represent the weight shares assigned by FLEXPOOL (STATPOOL). The remaining subplots illustrate the evolution of the PR weights over time, with the blue (red) lines showing the combination weights of FLEXPOOL (STATPOOL). STATPOOL primarily split the combination weights between GALTON and KWZ-MP, whereas the weights in FLEXPOOL were more broadly distributed, with weight shares ranging from 13.68% (GALTON) and 25.09% (KWZ-MP) over the evaluation sample. Interestingly, GALTON received the lowest average weight in FLEXPOOL, despite being the candidate PR with the highest CER over the evaluation sample. This outcome reflects the ensemble approach, which accounts for (time-varying) interdependencies among PRs' returns.

The optimal combination weights in STATPOOL are more persistent compared to those of FLEXPOOL, where the weights change rapidly and often concentrate entirely on a single candidate PR. For instance, VOLTIME received a high weight following the burst of the dotcom bubble and its aftermath, as well as during the subprime crisis. In contrast, during the relatively calm period of the mid-to-late 1990s, the $1/N$ rule dominated. Next, we conduct more in-depth analyses to better understand the mechanisms at work driving the utility gains of FLEXPOOL.

Table 1: Summary of results for a cross-section of the 50 largest stocks.
The table presents results for the evaluation sample spanning 1977:01 to 2020:12. It reports annualized CERs both without transaction costs (CER) and with proportional transaction costs (CER^{TC}) of 20 bps for a power utility investor with relative risk aversion of $\gamma = 3$. As additional performance measures, the table includes the annualized Sharpe ratio before transaction costs (SR) and after proportional transaction costs of 20 bps (SR^{TC}), the maximum drawdown based on transaction-adjusted returns (MaxDD^{TC}), and the average monthly turnover (Avg. TO).

Candidate PRs	CER	CER^{TC}	SR	SR^{TC}	MaxDD^{TC}	Avg. TO
$1/N$	4.20%	3.96%	0.512	0.499	0.541	0.078
VOLTIME	6.12%	5.88%	0.677	0.658	0.369	0.102
KWZ-MP	5.52%	4.92%	0.620	0.576	0.498	0.246
KWZ-LW	5.40%	4.08%	0.614	0.509	0.504	0.572
GALTON	6.24%	5.52%	0.703	0.642	0.461	0.310
Combined PRs						
FLEXPOOL	8.16%	7.20%	0.828	0.751	0.379	0.418
STATPOOL	5.40%	4.80%	0.635	0.582	0.476	0.265
EQUAL WEIGHTS	6.00%	5.52%	0.690	0.651	0.402	0.196

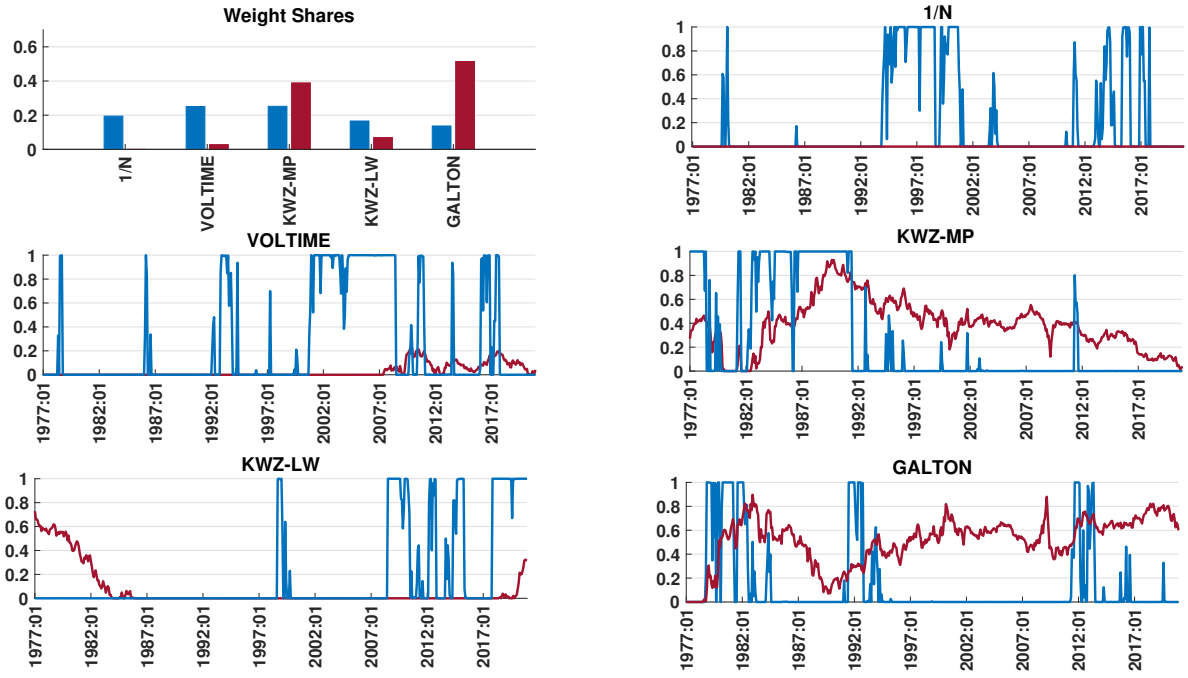

Figure 3: Combination weights.
The subplot in the upper left corner displays the average weight shares of FLEXPOOL (blue bars) and STATPOOL (red bars) over the evaluation sample from 1977:01 to 2020:12. The remaining subplots illustrate the evolution of the combination weights for the candidate PRs, with the blue lines representing FLEXPOOL and the red lines representing STATPOOL.

4.1.4 In-depth analyses

Relationship between the number of combined PRs and economic utility

How does the performance of FLEXPOOL depend on the number of PRs combined? Thus far, we have reported results only for the case where all five PRs are combined. What happens if we combine subsets of two, three, or four PRs instead? In other words, what if we set k to two, three, or four as an additional constraint in (10)?

Figure 4 displays the monthly CERs as a function of the number of PRs combined.[16] The blue diamonds represent the CERs generated by specific subsets of combined PRs. For instance, in the case of subsets with two PRs, there are $\binom{5}{2} = 10$ possible combinations. The red square indicates the average CER for a given number of combined PRs. On average, the CERs increase with the number of PRs combined.

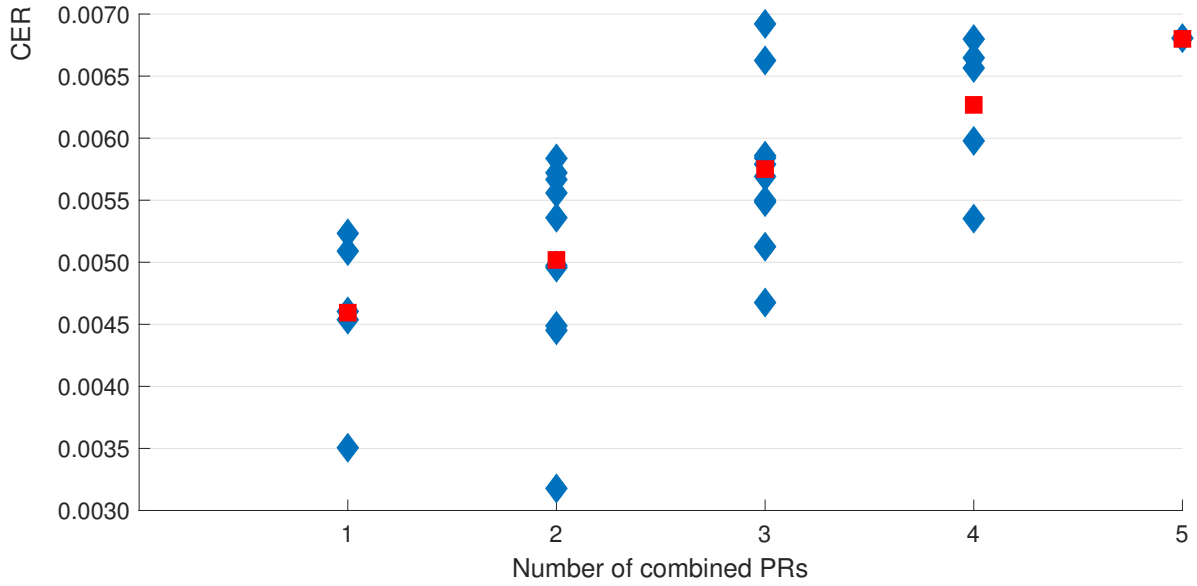

Figure 4: CERs as a function of the number of combined PRs using FLEXPOOL.
Blue diamonds represent the CERs for all possible combinations of a given number of combined PRs, while red squares indicate the average CERs for each combination size.

[16] Here, we use monthly rather than annualized CERs for graphical reasons.

The highest CER (0.0058) in the subset of two PRs is achieved by combining GALTON and KWZ-LW, while the lowest performance (0.0032) results from the combination of the $1/N$ rule with KWZ-LW. In the subset with three PRs, the highest CER (0.0069) is obtained by the combining VOLTIME, KWZ-MP, and KWZ-LW, whereas the lowest performance (0.0047) is generated by the combination of the $1/N$ rule, GALTON and KWZ-MP.

For the subset of four combined PRs, the highest CER (0.0068) is achieved by omitting the $1/N$ rule, and the lowest performance (0.0054) occurs when GALTON is omitted. Notably, the CER for all subsets with four PRs exceeds that of the best single candidate PR (0.0052). The CER for all five PRs corresponds to the reported performance of FLEXPOOL in Table 1, calculated as $0.0068 \times 1,200 = 8.16\%$.

The observation that the CER increases on average with the number of PRs combined highlights the benefits of diversification across more than just two PRs. These benefits are particularly notable given the positively correlated returns among the PRs, with empirical Pearson's correlation coefficients ranging from 0.64 (for $1/N$ and KWZ-MP) to 0.84 (for GALTON and VOLTIME).

Predictive power and risk management

Each PR, regardless of its construction, provides a record of asset weights and the corresponding implied OOS returns. To gain deeper insights into the mechanics of our combination framework, we analyze these asset weights and their relationship to the implied OOS returns. Specifically, following Frahm (2015), we examine statistics related to the predictive power and risk management of the candidate and combined PRs. As a measure of predictive power, we use Spearman's rank correlation coefficient $\widehat{\rho}_{SP}\left(\omega^{**}, \widetilde{\widetilde{\mathbf{r}}}\right)$, a robust correlation statistic, where

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

Here, ω^{**} denotes the asset weights implied by the PRs and the combination weights computed according to Equation (6), stacked from the beginning to the end of the evaluation sample. With $N = 50$ assets and an evaluation sample spanning 528 months (1977:01 to 2020:12), the vector ω^{**} has a length of $50 \times 528 = 26,400$. Similarly, $\widetilde{\widetilde{\mathbf{r}}}$ represents the stacked pseudo OOS excess returns generated by the $N = 50$ assets.[17]

The intuition behind the rank correlation $\widehat{\rho}_{SP}(\omega^{**}, \widetilde{\widetilde{\mathbf{r}}})$ is as follows: a PR assigns a positive weight to an asset if its expected return is relatively higher compared to other assets. Therefore, $\widehat{\rho}_{SP}\left(\omega^{**}, \widetilde{\widetilde{\mathbf{r}}}\right)$ serves as an approximation of the PR's overall predictive power. A high positive correlation indicates strong predictive power.

With respect to risk management, as proxied by a PR's ability to control the variance of its returns, a PR assigns a low (high) squared weight $\omega^{*,2}_{s,n}$ to the n-th asset if its expected squared return is high (low) for time s. Similarly, 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 compute Spearman's rank correlation $\widehat{\rho}_{SP}\left(\omega^{***}, \widetilde{\widetilde{\mathbf{r}}}\right)$, where

[17] For candidate PRs and the equally weighted benchmark combination, the optimal weights $\mathbf{w}^*_s$ in Equation (6) are replaced by assigning the full weight to the respective candidate PR or by assigning equal weights, respectively.

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

The rank correlation $\widehat{\rho}_{SP}\left(\omega^{***}, \widetilde{\widetilde{\widetilde{\mathbf{r}}}}\right)$ approximates a PR's ability to control the variance of the generated returns and can therefore serve as a proxy for risk management. The more negative the correlation, the better the risk management performance of the PR.

Table 2 summarizes the results for predictive power and risk management. FLEXPOOL demonstrates 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.

Interestingly, FLEXPOOL achieves significant predictive power despite none of the candidate PRs exhibiting significant predictive power when evaluated over the entire sample. The key lies in FLEXPOOL's ability to quickly adjust the combination weights toward combinations of PRs with local

predictive power.

In terms of risk management, VOLTIME outperforms all other strategies. However, in its pursuit of maximizing economic utility, FLEXPOOL implicitly balances predictive power and risk management, partially compromising VOLTIME's superior risk control to enhance utility gains.

Table 2: Predictive power and risk management.
Spearman's rank correlation $\widehat{\rho}_{SP}\left(\omega^{**}, \widetilde{\widetilde{\mathbf{r}}}\right)$ serves as an approximation of predictive power, while Spearman's rank correlation $\widehat{\rho}_{SP}\left(\omega^{***}, \widetilde{\widetilde{\mathbf{r}}}\right)$ approximates risk management. The p-values for testing the null hypothesis that the correlation coefficient equals zero are reported in parentheses below the correlation estimates.

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)

CER differences across various economic and market conditions

We conduct a time-series regression of the difference between the monthly CER achieved by FLEXPOOL and that achieved by one of the candidate PRs on a set of indicators that serve as proxies for economic and market conditions:

$$\Delta CER_t = \beta_0 + \beta_1 NegRet_t + \beta_2 Rec_t + \beta_3 HighVol_t + \beta_4 HighSent_t + e_t, \tag{15}$$

where ΔCER_t represents the difference (in bps) between the monthly CER achieved by FLEXPOOL and that achieved by one of the candidate PRs. The CERs are calculated for a power utility investor with risk aversion coefficient of $\gamma = 3$. $NegRet$ is a dummy variable that equals 1 if the S&P 500 return in a given month is negative, and 0 otherwise. Rec is a dummy variable that equals 1 if a given month falls within a recession regime as classified by the NBER, and 0 otherwise. $HighVol$ is a dummy variable that equals 1 if the realized variance (computed using daily S&P 500 returns) exceeeds the median realized variance over the entire evaluation sample (1977:01 to 2020:12), and 0 otherwise. $HighSent$ is a dummy variable that equals 1 if the investor sentiment index of Huang et al. (2015) is above its median value over the entire evaluation sample, and 0 otherwise.

Table 3 presents the estimated coefficients for the CER differences between FLEXPOOL and those achieved by each candidate PR across different economic and market states, as specified in Equation (15), with HAC-robust standard errors. A key takeaway from Table 3 is that the $1/N$ rule significantly underperforms FLEXPOOL when market returns are negative, with an average shortfall of roughly 225 bps per month. This finding aligns with Figure 3, which illustrates the evolution of the weights assigned to the candidate PRs over time and highlights that the $1/N$ rule tends to be picked during calm periods.

Table 3: Monthly CER differences (in bps) across various economic and market conditions.
The table reports the estimates of the slope coefficients for the time series regression in Equation (15) for the evaluation sample from 1977:01 to 2020:12. CERs are computed for a power utility investor with risk aversion coefficient $\gamma = 3$. ***, **, * indicates statistical significance at the 10%, 5%, and 1% level, respectively.

	NegRet	Rec	HighVol	HighSent
$1/N$	225.351^{***}	46.401	7.027	-33.825
VOLTIME	8.964	4.141	29.146	14.69
KWZ-MP	-63.523^{**}	27.01	-9.639	-15.657
KWZ-LW	6.673	8.265	-5.020	6.282
GALTON	-37.537^{**}	-1.331	-4.447	-4.920

Subsample analysis

We present subsample results (CER and CER^{TC}) for the periods before and after 2001, where the choice of the split point (January 2001) follows Avramov et al. (2023) and is justified by the decimalization that occurred in January 2001, which significantly reduced trading costs.[18]

Table 4 reports the results for CER and CER^{TC} for the pre-2001 sample (1977:01 to 2000:12) and the post-2001 sample (2001:01 to 2020:12). The results indicate that the performance of the PRs across different economic and market states is influenced by an overarching trend: economic gains tend to be smaller for most PRs in the post-2001 period.

In the pre-2001 period, KWZ-MP and GALTON were the most successful individual candidate PRs. However, using our ensemble approach, FLEXPOOL achieved higher CER (CER^{TC}) values than both of these candidate PRs, while also substantially outperforming STATPOOL and EQUAL WEIGHTS. In the post-2001 period, the relative performance of the candidate PRs changed notably compared to the pre-2001 sample. For instance, KWZ-MP and GALTON, which were the top-performing candidate PRs in the pre-2001 period, were outperformed by VOLTIME and, in particular, by KWZ-LW. FLEXPOOL

[18] Chordia et al. (2014) argue that decimalization increased liquidity and lowered trading costs, leading to greater price efficiency and reduced profitability for many trading strategies.

effectively adjusted the weights assigned to KWZ-MP and GALTON in the post-2001 period, reducing KWZ-MP's weight from 45.2% to 1.2% and GALTON's weight from over 19.5% to 6.8%. In contrast, FLEXPOOL substantially increased the weights for VOLTIME and KWZ-LW.

These patterns confirm our earlier analysis that FLEXPOOL successfully shifts weights to combinations of PRs that perform well locally over time; see our discussion on predictive power and risk management. VOLTIME's performance remains stable in both subsamples, further supporting its previously documented robustness across different economic and market states and time periods; see, e.g., Blitz and Van Vliet (2007) and Novy-Marx (2014). In the post-2001 sample, FLEXPOOL performed roughly on par with the ex-post best candidate PR, KWZ-LW, while STATPOOL and EQUAL WEIGHTS clearly underperformed KWZ-LW.

Table 4: Subsample analysis.
The table reports the subsample results (CER and CER^{TC}) for the pre-2001 period (1977:01 to 2000:12) and the post-2001 sample (2001:01 to 2020:12). ∅ Share represents the average share of a PR assigned by FLEXPOOL within each subsample.

	Pre-2001			Post-2001		
	CER	CER^{TC}	∅ Share	CER	CER^{TC}	∅ Share
Candidate PR						
$1/N$	4.80%	4.56%	0.254	3.48%	3.36%	0.123
VOLTIME	6.36%	6.12%	0.077	5.76%	5.64%	0.456
KWZ-MP	7.92%	7.08%	0.452	2.76%	2.40%	0.012
KWZ-LW	3.84%	2.16%	0.023	7.32%	6.36%	0.337
GALTON	7.80%	6.96%	0.195	4.44%	3.84%	0.068
Combined PR						
FLEXPOOL	9.00%	7.92%	x	7.20%	6.24%	x
STATPOOL	6.60%	5.76%	x	4.20%	3.72%	x
EQUAL WEIGHTS	6.72%	6.24%	x	5.16%	4.80%	x

4.2 Application to market timing

4.2.1 Investment universe and empirical study design

In this application, we consider an investor with power utility preferences and a relative risk aversion of $\gamma = 3$, who allocates their wealth monthly between the S&P 500 index and three-month U.S. Treasury bills. The weight allocated to stocks is constrained to lie within the range $[0; 1.5]$, ensuring that the PRs under consideration adhere to these weight constraints. Each PR generates its first OOS return in $1967{:}01$, and we use 60 months of OOS returns for the initial optimization of the combination weights. An additional 60 observations are reserved for the initial tuning of the forgetting factor α. Our first OOS evaluation takes place in January 1977. From that point onward, we proceed iteratively: the optimization is updated based on an extended sample of OOS portfolio returns, and the forgetting factor is chosen using an additional observation. This process is repeated recursively, resulting in an evaluation sample that spans $1977{:}01$ to $2020{:}12$.

4.2.2 Candidate PRs

We consider a diverse set of six different PRs. The key output of each PR, relevant for our ensemble approach, is the recommended asset weight for the S&P 500 index in each month. The first three PRs are based on strategies that exploit Bayesian predictive densities of the next period's excess return y, defined as 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 particularly attractive choices for market timing decisions due to their ability to accommodate parameter and model uncertainty, as well as their use of time-varying parameters (TVP) and stochastic volatility (SV). In the context of return predictability, Bayesian predictive densities have been applied by Dangl and Halling (2012), Johannes et al. (2014)

and Pettenuzzo and Ravazzolo (2016), among others.

While the first three PRs in our library differ in terms of the specific choices made for computing their respective Bayesian predictive densities, they can all be expressed in a canonical form. These PRs solve the investment problem by directly maximizing the conditional expected utility of next period's wealth W_{t+1}:

$$\underset{\omega_{t+1}\in[0;1.5]}{\arg\max}\ \mathbb{E}_t\left[U(W_{t+1})|\mathscr{D}^t\right] = \underset{\omega_{t+1}\in[0;1.5]}{\arg\max}\int \frac{\widetilde{R}_{t+1}^{1-\gamma}}{1-\gamma} p\left(y_{t+1}|\mathscr{D}^t\right) dy_{t+1}. \tag{16}$$

Here, $p\left(y_{t+1}|\mathscr{D}^t\right)$ denotes the Bayesian predictive density for the excess return y in $t+1$, based on the information set available at time t. The information set $\mathscr{D}^t$ includes the returns and predictors observable up to time t, as well as the choice of the prior at $t=0$.

Since power utility is independent of wealth, we can set $W_t = 1$ and proceed with the gross returns in (16). Let $\widetilde{R}_{t+1}$ denote the total gross return at time $t+1$, where the total return comprises the excess return y and the risk-free rate r^f. Let ω_{t+1} represent the weight allocated to the risky asset for time $t+1$. We approximate the maximization of the conditional expected utility in (16) using $B = 100,000$ potential realizations $y_{draw,t+1}^{(b)}$, $b = 1,\ldots,B$, of the excess return at $t+1$, drawn from the predictive density $p\left(y_{t+1}|\mathscr{D}^t\right)$:

$$\underset{\omega_{t+1}\in[0;1.5]}{\arg\max}\ \frac{1}{B}\sum_{b=1}^{B}\left\{\frac{\left[\omega_{t+1}\left(1+r_{t+1}^f+y_{draw,t+1}^{(b)}\right)+(1-\omega_{t+1})\left(1+r_{t+1}^f\right)\right]^{1-\gamma}}{1-\gamma}\right\}. \tag{17}$$

We set $\gamma = 3$. To derive a Bayesian predictive density for the excess returns, we impose a structure on the return-generating process. Specifically, we assume that the dynamics of the excess return follow

TVP regression models with the following structure:

$$y_{t+1} = \mathbf{X}_t'\theta_t + \varepsilon_{t+1}, \quad \varepsilon_{t+1} \sim \mathcal{N}\left(0, \upsilon_{t+1}\right) \tag{18}$$

$$\theta_t = \theta_{t-1} + \xi_t, \qquad \xi_t \sim \mathcal{N}\left(0, \boldsymbol{\Xi}_t\right), \tag{19}$$

where $\mathbf{X}_t$ denotes the vector of predictive variables observed at time t. This vector includes a subset of twelve predictor variables from Welch and Goyal (2008), depending on the specific setting.[19]

Let θ_t represent the vector of (unobserved) time-varying coefficients. The observation error ε_{t+1} is assumed to be normally distributed with mean zero and and time-varying (but unknown) variance υ_{t+1}. The time-varying coefficients are modeled to evolve according to a multivariate random walk process without drift (see, e.g., Dangl and Halling, 2012). We initialize the coefficients θ_0 using a diffuse conditional normal prior centered around zero.

The random shocks ξ are assumed to be multivariate normal with an unknown and time-varying system covariance matrix $\boldsymbol{\Xi}_\mathbf{t}$. Conditional on the observation variance and the system covariance, standard Bayesian methods for state-space models—using the Kalman filter—can be applied to estimate the coefficients and compute the predictive distribution of the returns. However, both the observation variance and the system covariance are unknown. To model their dynamics, we employ a forgetting factor approach to model their dynamics, where the forgetting factor δ governs the dynamics of the coefficients, and the forgetting factor κ governs the dynamics of the observation variance (see, e.g., Beckmann and Schüssler, 2014). If $\delta = 1$, all historical observations are equally weighted in the

[19]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). The data covers the period from 1927:01 through 2020:11 and was downloaded from Amit Goyal's homepage: `http://www.hec.unil.ch/agoyal/`. See Welch and Goyal (2008) for a more detailed description of the variables.

updating process, leading to constant coefficients. If $\delta < 1$, older observations are exponentially down-weighted. The smaller the value of δ, the more strongly older observations are down-weighted.

Similarly, κ controls the dynamics of the observation variance. If $\kappa = 1$, the observation variance remains constant. By using a conjugate specification with an inverse-gamma prior on the observation variance and a conditional normal prior on the coefficients, along with fixed values for the forgetting factors δ and κ, we obtain a t-distributed predictive density $p\left(y_{t+1}|\mathscr{D}^t\right)$. This density accounts for the uncertainty in both the coefficients and the observation variance. Our PRs, based on Bayesian predictive densities, comprise the following three setups, which differ in the selection of predictors and the specified values for the forgetting factors δ and κ:

- LARGE-TVP-SV:

  This multivariate setup incorporates all twelve considered predictors from Welch and Goyal (2008) and employs Bayesian model averaging (BMA) (Raftery et al., 1997) to assign weights to the predictive densities, which are based on different specifications of the coefficients' dynamics. The dynamics are controlled by the forgetting factor δ, chosen from the grid $\mathscr{S}_\delta = \{0.96; 0.97; 0.98; 0.99; 1.00\}$, where constant coefficients are included as a special case. Thus, the five individual models $\mathscr{M}_j$, $j = 1, \ldots, 5$, in this setup are defined by the distinct values of δ.

  Given the conditional heteroskedasticity is a well-known stylized fact of asset returns, we set the forgetting factor $\kappa = 0.97$ for the observation variance, following the RiskMetrics™ choice for monthly data (J.P.Morgan/Reuters, 1996). A priori, equal weights are assigned to the five predictive densities. At each point in time, the weights of the predictive densities are updated using Bayes' rule, and asset allocation decisions are made based on the resulting mixture

t-distribution, using the approximation in (17).

Our LARGE-TVP-SV candidate PR is motivated by the fact that "kitchen-sink"-type models for predicting the equity premium have demonstrated high utility gains in prior research, despite their poor performance based on statistical measures (Beckmann and Schüssler, 2014; Cederburg et al., 2023). We will provide further explanation for this finding in Section 4.2.3.

- BMA-TVP-CV:

  The second setup is based on the framework proposed by Dangl and Halling (2012). With twelve available predictors, there are 2^{12} different combinations of predictors that can either be included in or excluded from the vector of predictors $\mathbf{X}$. The forgetting factor δ, which controls the dynamics of the coefficients, is again chosen from the grid $\mathscr{S}_\delta = \{0.96; 0.97; 0.98; 0.99; 1.00\}$. This results in a total of $5 \times 2^{12} = 20,480$ different models $\mathscr{M}_j$, $j = 1, \ldots, 20,480$, defined by the subset of included predictors and the selected value of δ. Dangl and Halling (2012) assume a constant variance (CV). To align with their approach, we set $\kappa = 1.00$. A priori, we assign equal weights to the $20,480$ predictive densities and update these weights using BMA.[20]

- UNIV-TVP-SV:

  Univariate TVP-SV models are a common choice for capturing the dynamics of aggregate stock returns (see, e.g., Johannes et al., 2014). This setup relies on univariate (UNIV) predictive regression, where each regression includes only one of the twelve predictors. The grid $\mathscr{S}_\delta = \{0.96; 0.97; 0.98; 0.99; 1.00\}$ is used for δ, while κ is fixed at 0.97. All predictive densities are equally weighted.

[20] While this setup closely follows Dangl and Halling (2012), there are slight differences in implementation. For instance, Dangl and Halling (2012) include the cross-sectional beta premium as a predictor, which we we exclude because the data are only available through 2002.

The remaining three PRs are based on a MV framework for constructing portfolios. These PRs differ in how they estimate the excess return $\widehat{y}_{t+1}$. The weight assigned to the S&P 500 index is calculated as follows:

$$\underset{[0;1.5]}{\omega_{t+1}} = \frac{1}{\gamma}\left(\frac{\widehat{y}_{t+1}}{\widehat{\sigma}^2_{t+1}}\right). \tag{20}$$

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

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

  Imposing economic constraints, Ferreira and Santa-Clara (2011) predict aggregate stock returns as the sum of the dividend-price ratio and the long-term historical average of earnings growth.

- Combination of forecasts à la Rapach et al. (2010) (RSZ):

  Rapach et al. (2010) propose an equally weighted combination of point forecasts, where each forecast is derived from univariate predictive regressions with constant coefficients, using one of the predictors from Welch and Goyal (2008).[21]

- Prevailing Historical Mean (PHM):

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

Our library of PRs encompasses conceptually distinct approaches to market timing. These PRs utilize different information sets and employ various methods for translating that information into asset weights. The most notable distinction among them is that some rely on Bayesian predictive densities (with varying designs), while others use different strategies for generating point forecasts within a MV framework.

[21] We use monthly data, whereas Rapach et al. (2010) rely on quarterly data and consider 15 predictors instead of the 12 used here.

The PRs LARGE-TVP-SV, BMA-TVP-CV, UNIV-TVP-SV, and RSZ all rely on the predictor variables proposed in Welch and Goyal (2008) as their information set. However, these four PRs adopt markedly different econometric approaches to leverage the predictors from Welch and Goyal (2008), spanning from univariate regressions with constant or time-varying coefficients to models featuring both constant and stochastic volatility, while incorporating various forms of shrinkage—or none at all, as is the case with LARGE-TVP-SV. These methodological differences highlight the uncertainty in translating asset pricing rationales into portfolio decisions, even when the PRs agree on the information set.

SOP (Ferreira and Santa-Clara, 2011) utilizes a distinct information set compared to the four PRs discussed above. It capitalizes on the varying time-series persistence of components in its sum-of-the-parts approach, leveraging a specific empirical pattern to enhance return predictions. In contrast, PHM operates without relying on any predictors whatsoever.

The six PRs in the library, along with potentially many others, represent diverse approaches to translating asset pricing rationales into portfolio decisions. For example, time-varying coefficient models correspond to asset pricing theories that attribute fluctuations in risk premia to time-varying risk aversion, as discussed in Campbell and Cochrane (1999). However, the precise model specifications for incorporating predictors, accounting for time variation in coefficients, or determining the functional form for processing the data remain ambiguous. Our ensemble approach accommodates a broad range of possible specifications, making it well-suited to addressing the inherent uncertainty in this translation process.

4.2.3 Results

Table 5 presents the results, including annualized CERs calculated both without transaction costs (CER) and with proportional transaction costs (CER^{TC}) of 20 bps. Additionally, the table reports the annualized Sharpe ratio, both without transaction costs (SR) and with proportional transaction costs (SR^{TC}) of 20 bps. The R^2_{OOS}-statistic (Campbell and Thompson, 2008) assesses the point forecast accuracy of a given approach relative to the PHM benchmark. It quantifies the proportional reduction in the sample mean squared forecast error compared to the PHM. A positive R^2_{OOS}-statistic signifies that the mean squared forecast error of the given approach is lower than that of the PHM. As a measure of predictive power, we report $\widehat{\rho}_{SP}(\omega^{**}, \mathbf{y})$, the Spearman's rank correlation coefficient between the weights assigned to the risky asset and the realized excess returns. Here, ω^{**} denotes the stacked weights of the risky asset over the evaluation sample, and $\mathbf{y}$ represents the vector of realized excess returns over the same period. For risk management, we include $\widehat{\rho}_{SP}(\omega^{**,2}, \mathbf{y}^2)$, the Spearman's rank correlation coefficient between the squared weights of the risky asset and the squared realized excess returns.

The empirical results can be summarized as follows: FLEXPOOL achieved the highest CER and SR among the combined PRs, demonstrating both strong predictive power and robust risk management. While LARGE-TVP-SV exhibited the strongest predictive power, SOP excelled in risk management. FLEXPOOL performed similarly to the *ex-post* best candidate PR (LARGE-TVP-SV) in terms of CER and the SR. Notably, it outperformed in terms of maximum drawdown, achieving a value of 0.285, which is less than half the maximum drawdown of the PHM (0.660). During the OOS evaluation period of 44 years, the annualized outperformance against the PHM was 336 bps for FLEXPOOL against the PHM after transaction costs. When comparing FLEXPOOL to each candidate PR and

alternative combined PRs, its outperformance in terms of CERs and SRs is statistically significant at least at the 10% level, according to one-tailed tests of Diebold and Mariano (1995) and Ledoit and Wolf (2008), respectively—with the exception of LARGE-TVP-SV and STATPOOL.

Table 5: Summary of results for market timing.
The table presents results for the evaluation sample from 1977:01 to 2020:12, including annualized CERs both without transaction costs (CER) as well as with proportional transaction costs (CER^{TC}) of 20 bps for a power utility investor with a relative risk aversion of $\gamma = 3$. We also report the annualized Sharpe ratio without transaction costs (SR) and with proportional transaction costs (SR^{TC}) of 20 bps, as well as the maximum drawdown for transaction-adjusted returns (MaxDD^{TC}). To measure the accuracy of the point forecasts, we include the R^2_{OOS}-statistic. Predictive power and risk management ability are evaluated using Spearman's rank correlations: $\widehat{\rho}_{SP}(\omega^{**}, \mathbf{y})$ for predictive power and and $\widehat{\rho}_{SP}(\omega^{**,2}, \mathbf{y}^2)$ for risk management.

	Economic Evaluation Criteria					Statistical Properties		
Candidate PRs	CER	CER^{TC}	SR	SR^{TC}	MaxDD^{TC}	R^2_{OOS}	$\widehat{\rho}_{SP}(\omega^{**}, \mathbf{y})$	$\widehat{\rho}_{SP}(\omega^{**,2}, \mathbf{y}^2)$
LARGE-TVP-SV	11.52%	10.80%	0.709	0.661	0.346	−0.1133	0.1139 (0.0088)	− 0.0295 (0.4989)
BMA-TVP-CV	8.04%	7.80%	0.489	0.472	0.410	−0.0388	0.0145 (0.7390)	− 0.0787 (0.0714)
UNIV-TVP-SV	8.16%	7.84%	0.503	0.483	0.350	−0.0090	0.0176 (0.4150)	−0.0885 (0.0422)
SOP	8.52%	8.40%	0.540	0.527	0.583	0.0003	0.0614 (0.1592)	− 0.1213 (0.0530)
RSZ	8.28%	8.16%	0.505	0.496	0.646	0.0010	0.0143 (0.7433)	− 0.0115 (0.7928)
PHM	7.68%	7.56%	0.481	0.479	0.660	0.0000	− 0.0256 (0.5578)	− 0.0302 (0.4880)
Combined PRs								
FLEXPOOL	11.54%	10.92%	0.715	0.681	0.285	−0.0502	0.0888 (0.0413)	− 0.0979 (0.0245)
STATPOOL	10.68%	10.08%	0.668	0.624	0.304	−0.0978	0.0929 (0.0329)	−0.0292 (0.5030)
EQUAL WEIGHTS	9.36%	9.24%	0.579	0.565	0.450	0.0035	0.0577 (0.1859)	−0.0844 (0.0525)

Among the candidate PRs, LARGE-TVP-SV achieved by far the highest CER and SR, despite its low R^2_{OOS}-statistic of -0.1133. This result aligns with findings by Cenesizoglu and Timmermann (2012) and Leitch and Tanner (1991), who demonstrate that the point forecast accuracy of a model and its economic value can diverge significantly. As such, the R^2_{OOS}-statistic may be an unreliable indicator for guiding portfolio decisions. LARGE-TVP-SV overfits the data due to its use of numerous predictors, time-varying coefficients, and the absence of a shrinkage mechanism, which contribute to the low R^2_{OOS}-statistic. As a complex model, LARGE-TVP-SV excels at capturing return patterns with

a large and significantly positive $\widehat{\rho}_{SP}(\omega^{**}, \mathbf{y})$, offering high predictive correlations compared to shrunk models where signals are dampened. However, the high variance of its forecasts results in a low R^2_{OOS}-statistic. Notably, this low R^2_{OOS}-statistic does not adversely affect economic utility. The weight restrictions on the risky asset (i.e., no short selling and up to 50% leverage) prevent excessively volatile portfolio weights. Thus, complex methods such as LARGE-TVP-SV leverage their robust predictive signals while mitigating the drawbacks of high forecast volatility and the corresponding fluctuations in asset weights. For a broader discussion on the benefits of complexity in portfolio choice in a different context, see Kelly et al. (2024).

Similar to the findings in our first application, we observe that predictive power and risk management—indicated by positive values of $\widehat{\rho}_{SP}(\omega^{**}, \mathbf{y})$ and negative values of $\widehat{\rho}_{SP}(\omega^{**,2}, \mathbf{y}^2)$—align well with the ranking of the CERs and the SRs. While our ensemble approach, which directly optimizes utility at the PR level, successfully captures the strong economic performance of LARGE-TVP-SV, combination methods based on point forecast accuracy, such as R^2_{OOS}-statistics, would fail to do so.

The evolution of the combination weights is illustrated in Figure 5. For most of the time, LARGE-TVP-SV received a high weight, often comprising the entire allocation. However, SOP, with its strong risk management, was selected during three turbulent periods: first, in September and October 1998, following the strongly negative returns of August 1998, a period linked to the Russian currency crisis and the collapse of Long Term Capital Management. Second, SOP was selected from $2000{:}12$ to $2003{:}10$, during the aftermath of the dotcom bubble burst. Finally, SOP was picked from $2020{:}04$ to $2020{:}07$, following the sharp market drop caused by the COVID-19 pandemic in March 2020. Similar to our first application, these results highlight FLEXPOOL's ability to dynamically balance predictive power and risk management across candidate PRs while maximizing economic utility.

The forgetting factor α was set to 0.96, as determined by (9), for the entire evaluation period. The emphasis on the recent economic utility gains facilitated faster adjustments in the combination weights compared to STATPOOL (see Figure 5).

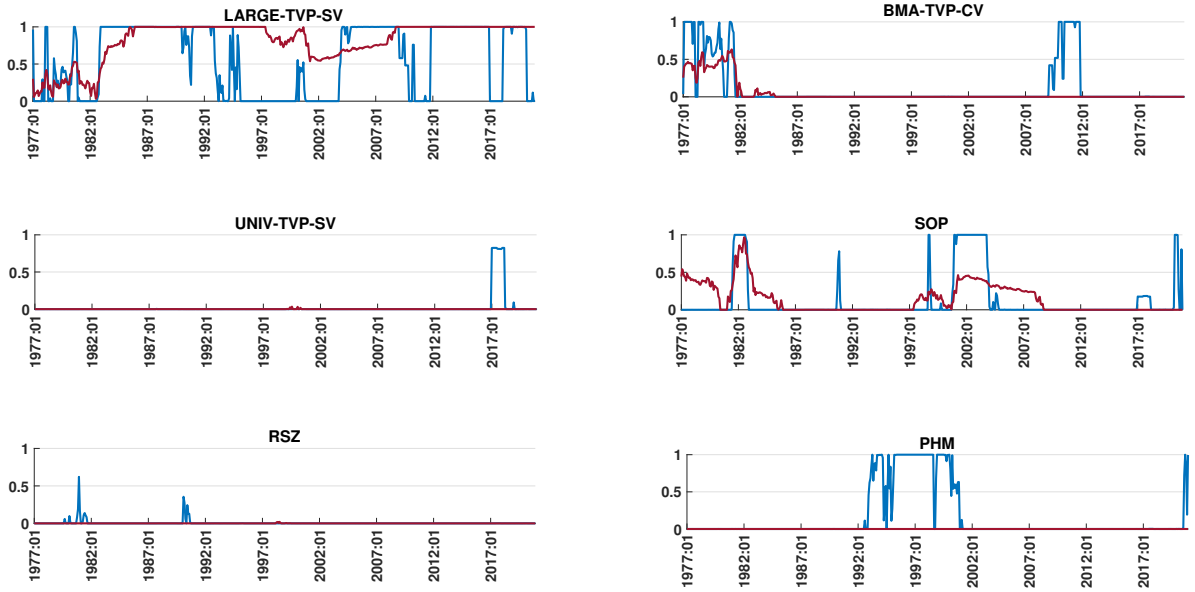

Figure 5: Evolution of combination weights.
The subplots show the evolution of the combination weights. The blue (red) lines show the combination weights in FLEXPOOL (STATPOOL).

UNIV-TVP-SV and RSZ shrink the coefficients toward zero by utilizing subsets of the predictors. As expected, these approaches achieve higher the point forecast accuracy, as measured by R^2_{OOS}-statistics, compared to LARGE-TVP-SV and FLEXPOOL. However, their predictive power, as reflected by the rank correlation $\widehat{\rho}_{SP}(\omega^{**}, \mathbf{y})$, is considerably lower, along with their CERs and SRs.

Notably, PHM received temporarily high weights during the relatively calm mid-to-late 1990s, a result consistent with our first application, where $1/N$ also received high weights during this period. This suggests a tendency for simple PRs to be favored in stable periods, while more flexible PRs are selected during turbulent times.

Figure 6 illustrates the CERs as a function of the number of combined PRs.[22] Blue diamonds represent the CERs generated by individual subsets of combined PRs, while red squares indicate the average CERs for each combination size. Consistent with the findings from our first application, the average CER increases as the number of combined PRs expands.[23]

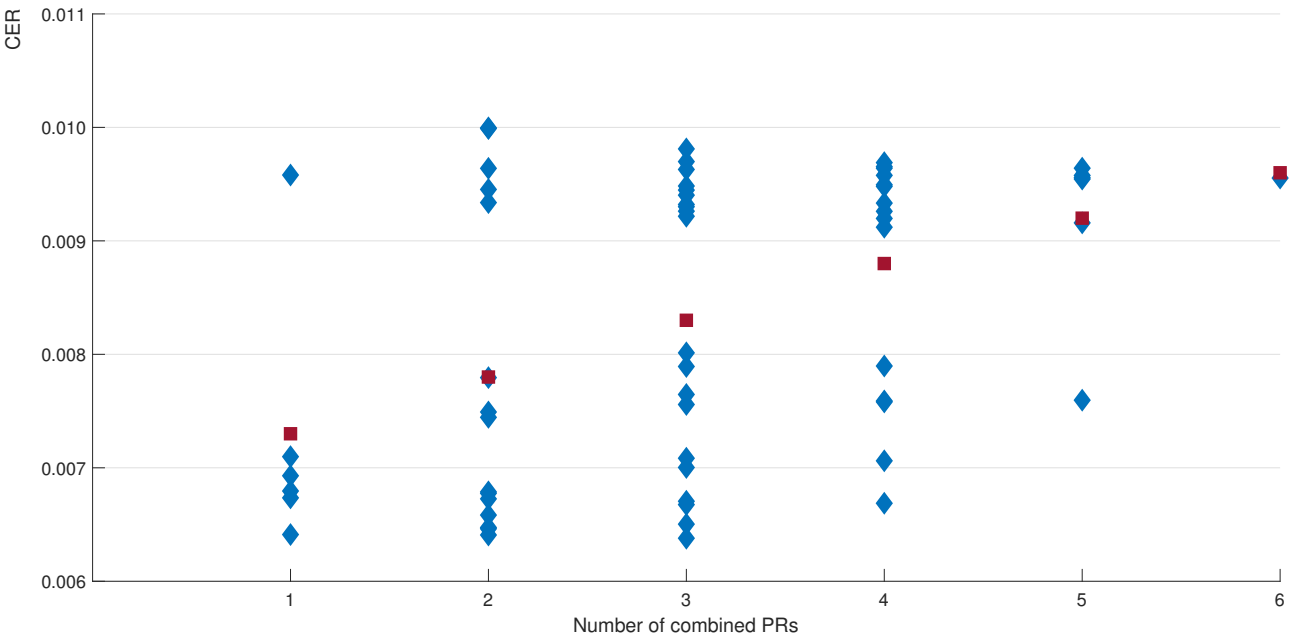

Figure 6: CERs as a function of the number of combined PRs using FLEXPOOL.
Blue diamonds represent the CERs for all possible combinations of a given number of combined PRs, while red squares indicate the average CERs for each combination size.

The results of this application highlight the critical importance of translating asset pricing rationales into portfolio choice. Predictive regressions based on combinations of univariate models, such as UNIV-TVP-SV and RSZ, yielded relatively low CERs and were very rarely selected in the combination. This finding aligns with Welch and Goyal (2008), who reported no utility gains (relative to PHM) for individual predictors. Similarly, Goyal et al. (2024) dismissed most predictors proposed after Welch

[22] Again, we use monthly rather than annualized CERs for graphical reasons.

[23] Note that the slightly higher CER within the subset of two PRs (LARGE-TVP-SV and SOP) would not have been achievable in real time, as both the forgetting factor α and the subset size k were optimized period-by-period.

and Goyal (2008) based on their lack of economic utility when evaluated individually. However, as shown in Table 5, LARGE-TVP-SV delivered strong predictive signals and played a pivotal role in the combination, demonstrating the substantial economic benefits of leveraging multiple predictors, along with time-varying coefficients and stochastic volatility. SOP and PHM, which are derived from distinct information sets, also contributed to the ensemble. SOP, with its strong risk management capabilities, was predominantly selected during recessions, while PHM was favored during expansions and periods of economic stability.

4.2.4 CER differences across various economic and market conditions

Table 6 presents the estimated coefficients for the CER differences between FLEXPOOL and those achieved by individual candidate PRs under various economic and market conditions, as specified in Equation (15). During periods of negative market returns, FLEXPOOL outperformed all candidate PRs except LARGE-TVP-SV and SOP. Specifically, FLEXPOOL outperformed PHM by approximately 225 bps per month on average during periods of negative market returns. This result is consistent with Figure 6, which illustrates that the PHM rule is predominantly selected during calm market periods.

Table 6: Monthly CER differences (in bps) across various economic and market conditions.
The table reports the coefficient estimates for the time-series regression specified in Equation (15), based on the evaluation sample from 1977:01 to 2020:12. CERs are calculated for a power utility investor with risk aversion coefficient $\gamma = 3$. ***, **, * indicate statistical significance at the 1%, 5%, and 10% level, respectively.

Candidate PR	NegRet	Rec	HighVol	HighSent
LARGE-TVP-SV	−39.950	13.505	−15.060	−2.929
BMA-TVP-CV	116.176***	62.216	−50.715	17.666
UNIV-TVP-SV	91.197***	71.958	15.349	24.713
SOP	−104.640**	116.042	−9.594	25.267
RSZ	125.335***	115.318	−28.607	−0.235
PHM	224.678***	113.957	−29.060	11.228

4.2.5 Subsample analysis

We present subsample results (CER and CER^{TC}) for the pre-2001 and post-2001 periods. Table 7 reports CER and CER^{TC} for the pre-2001 sample (1977:01 to 2000:12) and the post-2001 sample (2001:01 to 2020:12). To provide additional context and motivated by the findings of Löffler (2022), which highlight that the BUY-and-HOLD strategy tends to be a tougher benchmark than the PHM, we also report results for the BUY-and-HOLD strategy across both subsamples.

Notably, CERs were substantially higher in the pre-2001 period for all candidate PRs (including the BUY-and-HOLD strategy) as well as for the combined PRs. This suggests that, while different market conditions influenced returns, an overarching trend of a less attractive risk-return profile emerged in the post-2001 period.

In both subsamples, FLEXPOOL performed consistently well, roughly on par with LARGE-TVP-SV, the best-performing candidate PR. The relatively strong performance of LARGE-TVP-SV reinforces our findings from Section 4.2.3, where we highlighted its high predictive power as a key feature (see Table 5). The subsample analysis further demonstrates that this strong performance is not limited to short episodes.

Table 7: Subsample analysis.
The table reports the subsample results (CER and CER^{TC}) for the pre-2001 sample from 1977:01 to 2000:12 and the post-2001 sample from 2001:01 to 2020:12. ∅ Share represents the average share of a PR assigned by FLEXPOOL within each subsample.

	Pre-2001			Post-2001		
	CER	CER^{TC}	∅ Share	CER	CER^{TC}	∅ Share
Candidate PR						
LARGE-TVP-SV	13.80%	11.96%	0.490	8.76%	8.16%	0.655
BMA-TVP-CV	10.20%	9.84%	0.135	5.64%	5.40%	0.104
UNIV-TVP-SV	10.68%	10.32%	0.000	5.16%	4.80%	0.049
SOP	12.48%	12.24%	0.075	3.84%	3.72%	0.177
RSZ	12.96%	12.84%	0.010	2.76%	2.64%	0.000
PHM	11.74%	11.74%	0.289	2.76%	2.76%	0.016
Combined PR						
FLEXPOOL	14.40%	13.92%	x	8.16%	7.68%	x
STATPOOL	12.84%	12.12%	x	8.04%	7.56%	x
EQUAL WEIGHTS	12.72%	12.48%	x	5.40%	5.28%	x
BUY-and-HOLD	11.76%	11.76%	x	5.04%	5.04%	x

4.2.6 Alternative settings and additional PRs

In addition to the results presented thus far, we examined three alternative empirical settings. First, we included the Buy-and-Hold strategy as an additional candidate PR in the library. This strategy achieved an annualized CER of 8.76% and an annualized SR of 0.525. Notably, we found that adding the Buy-and-Hold strategy to the ensemble had a negligible impact on our overall results. The other two alternative settings relied on data available only for shorter time periods.

First, we assessed the utility gains from combining PRs based on backward-looking data with those utilizing forward-looking data. As a representative of a PR using forward-looking data, we selected the strategy proposed by Pyun (2019), which generates OOS forecasts of the equity premium derived from the variance risk premium. This method exploits the relationship between the market risk premium and

the price of variance risk through variance risk exposure.[24] We used a rolling window of 60 months to compute the variance estimate as an additional input to the MV specification (20) and applied the same weight restrictions as in our previous analysis. This PR was combined with the LARGE-TVP-SV rule, representing PRs based on backward-looking data, and results were computed for the evaluation period spanning from 2000:01 to 2019:12. The PR relying solely on option data achieved an annualized CER of 9.36% and an annualized SR of 0.728. LARGE-TVP-SV achieved an annualized CER of 8.40% and an annualized SR of 0.677. Using FLEXPOOL to combine the two PRs slightly improved the results, yielding an annualized CER of 9.48% and an annualized SR of 0.743. For comparison, PHM achieved an annualized CER of 1.68% and an annualized SR of 0.284 over this truncated sample.

Second, we investigated whether incorporating a PR based on the approach recently proposed by Dong et al. (2022) could enhance performance relative to LARGE-TVP-SV. Dong et al. (2022) introduce a novel method that utilizes the returns of 100 cross-sectional anomaly portfolios as predictors for point forecasts of aggregate excess returns. These forecasts are available from 1975:01 to 2017:12.[25] For this shortened period, we combined the strategy of Dong et al. (2022) with LARGE-TVP-SV and evaluated the results for the sample from 1985:01 to 2017:12. Using the MV specification (20), we selected a setting where the elastic net was employed as the shrinkage technique for estimating expected excess returns. We also used a rolling window of 60 months to compute the variance and applied the same weight restrictions (no short selling, up to 50% leverage) as in our previous analysis. LARGE-TVP-SV produced an annualized CER of 9.84%, while the approach proposed by Dong et al. (2022) achieved an annualized CER of 11.16%. The combination of both PRs using FLEXPOOL yielded an annualized CER of 11.76%, further confirming the value of complex PRs in enhancing economic utility.

[24] The point forecasts are available from 1990:02 to 2019:12. We downloaded the data from Sungjune Pyun's homepage: https://sjpyun.github.io/research.html.

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

5 The relative strengths of FLEXPOOL

As a stacking method capable of combining conceptually distinct estimators (i.e., PRs in our case), FLEXPOOL seeks to diversify their idiosyncratic risks while leveraging their individual strengths. Existing combination methods, such as those proposed by Tu and Zhou (2011) and Kan et al. (2022), are constrained in terms of both the number and type of PRs they can combine. In contrast, FLEXPOOL has the flexibility to integrate a large number of diverse PRs. This ability to combine PRs from different domains is a key strength of FLEXPOOL.

It is important to emphasize that FLEXPOOL should *not* be viewed as a substitute for combination rules grounded in a specific structure, such as those derived from asset pricing models. For example, consider the KWZ-MP rule of Kan et al. (2022), described in Section 4.1.1. This PR focuses on maximizing the expected OOS utility, with the weight vector defined as:

$$\widehat{\mathbf{w}}_{q,t} = \widehat{\mathbf{w}}_{g,t}^{MP} + \frac{g_3\left(\widehat{\psi}_t^2\right)}{\gamma}\widehat{\mathbf{w}}_{z,t}^{MP}; \tag{21}$$

see Formula 51 in Kan et al. (2022). Essentially, the KWZ-MP rule combines the GMV portfolio ($\widehat{\mathbf{w}}_{g,t}^{MP}$) with a long-short zero-investment portfolio ($\widehat{\mathbf{w}}_{z,t}^{MP}$). The parameter $g_3\left(\widehat{\psi}_t^2\right)$ is estimated using various inputs, including mean and the covariance matrix estimates. The estimation process is informed by the factor structure proposed by MacKinlay and Pástor (2000), which minimizes estimation error when combining two specific building blocks, namely ($\widehat{\mathbf{w}}_{g,t}^{MP}$) and ($\widehat{\mathbf{w}}_{z,t}^{MP}$). In contrast, FLEXPOOL estimates the parameter $g_3\left(\widehat{\psi}_t^2\right)$ parameter using a simple linear combination approach, without accounting for additional data and for the factor structure. A modified version of FLEXPOOL

optimizes the weight of the zero-investment portfolio to maximize the pseudo OOS utility.[26]

Empirically, the CER derived from the portfolio with the estimated weight for the long-short zero-investment portfolio in FLEXPOOL is slightly lower than that obtained using the KWZ-MP rule. Specifically, the KWZ-MP rule achieves an annualized CER of 5.52% before and 4.92% after transaction costs, compared to 5.28% and 4.68% for FLEXPOOL.[27] This marginal outperformance of KWZ-MP over FLEXPOOL in this specific scenario is unsurprising, given that KWZ-MP is explicitly optimized for the two PRs considered here. However, the KWZ-MP method does not accommodate additional PRs, such as VOLTIME or GALTON shrinkage, within its framework. This limitation underscores the value of our more general approach.

As demonstrated in our application to a cross-section of 50 stocks in Section 4.1, combining KWZ-MP with $1/N$, VOLTIME, KWZ-LW, and GALTON, substantially enhances performance. FLEXPOOL, in this case, surpasses KWZ-MP by roughly 228 bps in terms of annualized CER after transaction costs (see Table 1). FLEXPOOL also exhibits a notably lower maximum drawdown. Thus, FLEXPOOL should be viewed as a complementary method to existing combination rules rather than a competing approach designed to replace them.

The relative strength of FLEXPOOL in combining specialized and heterogeneous PRs ("strong learners"), rather than replacing existing combination methods, lies in its foundation within stacking algorithms. Stacking has predominantly been used in predictive tasks for cross-sectional data to combine diverse candidate estimators, such as random forests, neural networks, and polynomial linear

[26] In this configuration, FLEXPOOL was tuned by setting the GMV portfolio weight to 1 and constraining $g_3\left(\widehat{\psi}_t^2\right)$ within the range between $[0; 1]$. This ensured the resulting portfolio weights sum to 1, with the limiting cases being the GMV portfolio (if $g_3\left(\widehat{\psi}_t^2\right)$ is 0) and the plug-in portfolio (if $g_3\left(\widehat{\psi}_t^2\right)$ is 1). Although this setup slightly differs from the FLEXPOOL configuration—which uses a convex combination of candidate PRs—the optimization approach for the zero-investment portfolio weight is analogous to FLEXPOOL's method of addressing the combination problem.

[27] As in Section 4.1, the largest 50 stocks were used as the asset universe, with the same time frame applied for estimation and evaluation. The evaluation sample spans the period from January 1977 to December 2020.

models, into a single unified estimator. Typically, simple estimators, such as regression trees ("weak learners"), are aggregated (e.g., through random forests or boosting) before being included in the stacking algorithm. While the candidate estimators in an ensemble are often complex to capture hidden structures in the data, the aggregation function within stacking algorithms is usually linear. Moreover, the weights assigned to candidate estimators are constrained to be between 0 and 1 and sum to one (i.e., a convex combination).[28]

In summary, FLEXPOOL is not intended to "reinvent the wheel" by replacing existing PRs that combine different building blocks in a targeted manner, such as the KWZ-MP rule. Instead, such combination rules can be incorporated into the pool of candidate PRs and further combined within the FLEXPOOL framework. However, given the nature of stacking algorithms, relying solely on FLEXPOOL to combine a set of relatively simple PRs may not be the most effective strategy if a targeted combination approach for these PRs has already been developed.

As a second illustrative example, consider a pool of candidate PRs consisting of three simple rules: $1/N$, the (simple plug-in version of the) Markowitz portfolio, and GMV. While the GALTON shrinkage method proposed by Barroso and Saxena (2022) is not strictly a combination method, its solution to the asset allocation problem is spanned by $1/N$, the (simple plug-in version of the) Markowitz portfolio, and GMV as limiting cases (see Section 4.1.2 for details).

The GALTON shrinkage method determines the optimal degree of shrinkage for means and (co-)variances by using OOS forecast errors and shrinkage targets, requiring the estimation of several parameters from the data within the shrinkage framework. This method employs sophisticated

[28] Although nonlinear and more complex aggregation functions are theoretically possible, they have not been widely adopted. The convex combination approach has been shown to provide stability (see, e.g., Breiman, 1996; Van der Laan et al., 2007; Polley and Van Der Laan, 2010). Recall that, in the context of combining PRs, the convex combination offers an additional advantage: it ensures that any restrictions on asset weights imposed at the level of the candidate PRs (e.g., no short selling, sector weight limits, etc.) are preserved at the level of the combined PRs.

mechanisms to process data with shrinkage targets, leveraging specific information to shrink means, variances, and covariances. In contrast, FLEXPOOL uses a convex combination approach.

Table 8 compares the empirical performance of GALTON and FLEXPOOL.[29] GALTON outperformed FLEXPOOL on both pre- and post-transaction cost performance measures. However, within the framework of Barroso and Saxena (2022), there is no way to combine the GALTON rule with additional, potentially diverse PRs. When FLEXPOOL is used to combine GALTON with $1/N$, VOLTIME, KWZ-MP, and KWZ-LW, the resulting utility gains are substantially higher than when relying on GALTON alone, yielding an additional 168 bps per year in CER after transaction costs (see Table 1).[30]

The results further underscore a key principle: when specialized combination methods are available for simple rules, they should be incorporated as candidate PRs within our approach. In the absence of such methods, FLEXPOOL provides a versatile, general-purpose solution capable of delivering substantial economic gains compared to relying on any single PR, regardless of how it is selected.

[29] As in Section 4.1, the largest 50 stocks were used as the asset universe, with the same time frame applied for estimation and evaluation. The evaluation sample spanned the period from January 1977 to December 2020.

[30] One might question how the results would change if the $1/N$ rule was excluded from the pool of PRs in Table 1, given that the $1/N$ rule is also used as a building block for GALTON. Omitting $1/N$ from the pool of candidates in Section 1 has minimal impact on the results, yielding an annualized CER of 8.14% before and 7.09% after transaction costs.

Table 8: Comparison of FLEXPOOL and GALTON in a pool of simple rules.
The table presents results for the evaluation sample from 1977:01 to 2020:12, including annualized CERs both without transaction costs (CER) and with proportional transaction costs (CER^{TC}) of 20 bps for a power utility investor with a relative risk aversion of $\gamma = 3$. Additionally, the table reports the annualized Sharpe ratio before transaction costs (SR) and after proportional transaction costs of 20 bps (SR^{TC}). Avg. TO represents the average monthly turnover.

	CER	CER^{TC}	SR	SR^{TC}	Avg. TO
$1/N$	4.20%	3.96%	0.512	0.500	0.078
Markowitz	-174.36%	-438.36%	0.126	-0.494	99.031
GMV	3.84%	0.96%	0.485	0.294	1.154
GALTON	6.24%	5.52%	0.703	0.642	0.310
FLEXPOOL	4.56%	4.20%	0.524	0.496	0.108

6 Concluding remarks

We have introduced an ensemble framework for combining conceptually distinct PRs. This approach allows researchers and investors to harness the diverse range of existing PRs within a utility maximization framework, effectively diversifying away the idiosyncratic risks associated with individual PRs. Notably, our framework can seamlessly integrate PRs regardless of their design, while existing combination approaches are applicable only to specific and limited sets of PRs.

The ensemble framework we propose offers researchers and investors a robust approach to address model uncertainties regarding which asset pricing rationale, portfolio theory, information set, and econometric technique to rely on when tackling a portfolio choice problem. Rooted in the proposed methodology, the key contribution of our study lies in its potential to reshape how portfolio choice problems are addressed empirically. Rather than focusing on identifying a single best PR, our framework enables a diverse library of candidate PRs to collectively contribute their strengths, much like optimizing a portfolio of assets. While the search for innovative PRs will undoubtedly continue—driven by advancements in methodologies and data—our framework serves as a valuable

tool for assessing the incremental empirical merits (or limitations) of newly proposed PRs.

Extensive applications to a cross-section of stocks and to market timing demonstrate the effectiveness of our approach. The combined PRs consistently achieved OOS CERs after transactions costs that were either superior to those of any single candidate PR or comparable to the *ex-post* best-performing PR. Importantly, from an ensemble perspective, the PR with the highest individual CER did not always receive the highest weight in the combination. Instead, adaptive learning, reflected in rapidly adjusting combination weights, proved critical in enhancing OOS utility by capturing the time-varying relative performance of individual PRs. Detailed analyses highlighted how the flexible combination framework balances predicting the level of asset returns with anticipating their variance. Moreover, the findings indicate that, on average, utility gains increased with the number of candidate PRs, even without applying additional regularization to the combination weights to mitigate estimation risk at the combination level.

Our contribution opens several promising avenues for future research. While it is impossible for any candidate pool to be exhaustive given the vast range of potential PRs, constructing larger pools of candidate PRs could be a worthwhile endeavor. Incorporating a pre-screening step to maximize heterogeneity—using an appropriate distance measure—might further enhance the diversity and effectiveness of the pool. Our proposed ensemble framework could also prove valuable for investigating the relative merits of PRs within specific domains. For example, it could be used to evaluate whether expert-driven PRs provide additional value compared to purely algorithmic PRs.

References

Avramov, D., Cheng, S., and Metzker, L. (2023). Machine learning vs. economic restrictions: Evidence from stock return predictability. *Management Science*, 69(5):2587–2619.

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. and Schüssler, R. (2014). Forecasting equity premia using bayesian dynamic model averaging. Technical report, Center for Quantitative Economics (CQE), University of Muenster.

Blitz, D. and Van Vliet, P. (2007). The volatility effect: Lower risk without lower return. *Journal of Portfolio Management*, pages 102–113.

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 Cochrane, J. H. (1999). By force of habit: A consumption-based explanation of aggregate stock market behavior. *Journal of Political Economy*, 107(2):205–251.

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.

Cederburg, S., Johnson, T. L., and O'Doherty, M. S. (2023). On the economic significance of stock return predictability. *Review of Finance*, 27(2):619–657.

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

Chen, L., Pelger, M., and Zhu, J. (2024). Deep learning in asset pricing. *Management Science*, 70(2):714–750.

Chordia, T., Subrahmanyam, A., and Tong, Q. (2014). Have capital market anomalies attenuated in the recent era of high liquidity and trading activity? *Journal of Accounting and Economics*, 58(1):41–58.

Cong, L. W., Tang, K., Wang, J., and Zhang, Y. (2021). Alphaportfolio: Direct construction through deep reinforcement learning and interpretable ai. *Available at SSRN 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. E., Schmidt, L., and Timmermann, A. (2023). Pockets of predictability. *The Journal of Finance*, 78(3):1279–1341.

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.

Frahm, G. (2015). A theoretical foundation of portfolio resampling. *Theory and Decision*, 79(1):107–132.

Frost, P. A. and Savarino, J. E. (1986). An empirical bayes approach to efficient portfolio selection. *Journal of Financial and Quantitative Analysis*, 21(3):293–305.

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. (2024). A comprehensive 2022 look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, 37(11):3490–3557.

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

Huang, D., Jiang, F., Tu, J., and Zhou, G. (2015). Investor sentiment aligned: A powerful predictor of stock returns. *The Review of Financial Studies*, 28(3):791–837.

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.

J.P.Morgan/Reuters (1996). Riskmetrics—technical document. Technical report.

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. (2023). Bagged pretested portfolio selection. *Journal of Business & Economic Statistics*, 41(4):1116–1131.

Kelly, B., Malamud, S., and Zhou, K. (2024). The virtue of complexity in return prediction. *The Journal of Finance*, 79(1):459–503.

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., Vanderveken, R., and Vrins, F. (2023). On the combination of naive and mean-variance portfolio strategies. *Journal of Business & Economic Statistics*, pages 1–15.

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.

Löffler, G. (2022). Equity premium forecasts tend to perform worse against a buy-and-hold benchmark. *Critical Finance Review*, 11(1):65–77.

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.

Maillard, S., Roncalli, T., and Teiletche, J. (2010). On the properties of equally weighted risk contribution portfolios. *Journal of Portfolio Management*, 36:60–70.

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*.

Novy-Marx, R. (2014). Understanding defensive equity. Technical report, National Bureau of Economic Research.

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.

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.

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).

van Hemert, O., Ganz, M., Harvey, C. R., Rattray, S., Martin, E. S., and Yawitch, D. (2020). Drawdowns. *The Journal of Portfolio Management*, 46(8):34–50.

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.

Yuan, M. and Zhou, G. (2022). Why naive diversification is not so naive, and how to beat it? *Journal of Financial and Quantitative Analysis*, pages 1–32.