When Bayes-Stein Meets Machine Learning: A Generalized Approach for Portfolio Optimization *

Dimitrios Gounopoulos † Emmanouil Platanakis ‡ Gerry Tsoukalas § Haoran Wu ¶

Abstract

The Bayes-Stein model provides a framework for remedying parameter uncertainty in the Markowitz mean-variance portfolio optimization. The classical version, however, suffers from estimation errors of model components and fails to consistently outperform the naive 1/N asset allocation rule. We comprehensively investigate the drawbacks of the traditional Bayes-Stein model and develop a generalized counterpart by refining model components with various well-tailored machine learning techniques, expanding the scope and applicability of the original Bayes-Stein model. Specifically, we propose a time-dependent weighted Elastic Net (TW-ENet) approach predicting expected asset returns, a hybrid double selective clustering combination (HDS-CC) strategy calibrating shrinkage factors, and a graphical adaptive Elastic Net (GA-ENet) algorithm estimating the inverse covariance matrix. Empirical studies demonstrate that our generalized Bayes-Stein framework can always offer better out-of-sample performance than the 1/N strategy. Importantly, our study tailors existing machine learning methods considering specifics of financial issues, illustrating appealing directions for solving challenging financial problems with machine learning.

Keywords: Bayes-Stein, Portfolio optimization, 1/N, Machine learning

JEL Classification: G11, G17

*We are grateful to Alberto Martin-Utrera, Lin William Cong, Victor DeMiguel, Guofu Zhou, Bin Li (discussant), and Frederik Simon (discussant) as well as conference participants at the 2022 Chinese Finance Annual Meeting (CFAM) and the Cardiff Fintech Conference 2022, and seminar participants at Durham University, University of Aberdeen, University of Southampton and University of York for their useful comments and suggestions. This paper is the recipient of the Research Excellence Award - Financial Investment Talent Development Fund kindly awarded by the Program Committee of the 2022 Chinese Finance Annual Meeting.

†School of Management, University of Bath, D.Gounopoulos@bath.ac.uk.

‡School of Management, University of Bath, E.Platanakis@bath.ac.uk.

§University of Pennsylvania - The Wharton School & Boston University & Luohan Academy, gtsouk@wharton.upenn.edu, gerryt@bu.edu.

¶School of Management, University of Bath, hw2258@bath.ac.uk.

1 Introduction

The seminal mean-variance portfolio optimization framework of Markowitz (1952) is probably the most popular portfolio construction technique widely used by academics and the asset management industry. This model achieves an optimal risk-return trade-off when the input parameters (mean returns and covariances of returns) are known with certainty. However, in practice, vanilla mean-variance portfolios perform poorly out of sample due to parameter uncertainty and the model's extreme sensitivity to estimation errors of the input parameters (see, e.g., Chopra and Ziemba, 1993; Kan and Zhou, 2007; Kuhn et al., 2009; Lim, Shanthikumar and Vahn, 2012; Levy and Levy, 2014; Kan, Wang and Zhou, 2022). Various asset allocation schemes have been proposed to ameliorate the negative effects of estimation errors. Among them, strategies producing shrinkage estimators stand out because of their theoretical and empirical appeal in mitigating parameter uncertainty and offering gains for portfolio optimization (see Jorion, 1986; Frost and Savarino, 1986; Ledoit and Wolf, 2003, 2017).

Among the class of shrinkage estimators, the Bayes-Stein model of Jorion (1986), as one of the first endeavors embedding shrinkage models in portfolio selection[1], enjoys popularity in the literature either as the fundamental parameter estimation model of the classic Markowitz mean-variance framework, or as a benchmark model relative to alternative estimation methods (see, e.g., Board and Sutcliffe, 1994; Barroso and Saxena, 2021; Platanakis, Sutcliffe and Ye, 2021). It is founded on the pioneering James-Stein theory (Stein, 1956; Stein and James, 1961) that shrinks the maximum likelihood estimator (MLE) of a multivariate normal distribution with dimensions greater than two towards a specified target (the grand mean), delivering a mean estimator with a smaller mean squared error (MSE) than MLE. Particularly, the Bayes-Stein model derives estimators of the mean returns vector and the covariance matrix by integrating the James-Stein estimator into a Bayesian framework and assuming a suitable prior about the shrinkage factor (Jorion, 1985). It finally mitigates the estimation risk of mean returns by shrinking the sample means towards the expected return of the global minimum variance portfolio. Hence, the classical Bayes-Stein model possesses solid theoretical foundations, and it is of interest to explore it further with new perspectives.

However, the mean returns and the inverse covariance matrix[2] outputted by the conven-

[1] Another similar shrinkage estimator of mean returns is suggested by Frost and Savarino (1986), in which the historical grand average of all assets is used as the shrinkage target.

[2] In the light of Nguyen, Kuhn and Mohajerin Esfahani (2022), the covariance matrix exists in the for-

tional Bayes-Stein model still present huge estimation errors, resulting in a relatively weak out-of-sample portfolio performance. For example, it cannot consistently outperform the "naive" equally-weighted (1/N) portfolio allocation rule that does not require any optimization and equally distributes capital among different underlying assets (DeMiguel, Garlappi and Uppal, 2009). Our motivation starts from this stylized fact. Moreover, while previous studies have conducted plenty of comparative analyses of the Bayes-Stein framework to alternative models, few explicitly examine the fundamental limitations of this rather intuitive model and attempt to improve the classical shrinkage portfolio optimization framework. Besides, on the one hand, the traditional Bayes-Stein model contaminated by estimation errors of the model components (the sample means vector, the grand mean, and the shrinkage factor) fails to provide an accurate proxy for the expected returns. On the other hand, this model concentrates on the shrinkage estimator for mean returns and only provides a sample-based formulation for the covariance matrix without accounting for the financial specifics of the covariance matrix or its inverse. Thus, it is crucial to offer ways to move beyond the original Bayes-Stein framework and produce an efficient and complete counterpart.

To fill these gaps, first, we analytically identify the main drawbacks of the classical Bayes-Stein model, namely low accuracy of the mean estimator, calibration error of the shrinkage factor weighting the sample mean and grand mean, and estimation risk of the inverse covariance matrix. Based on its drawbacks, we propose a generalized framework with well-designed machine learning techniques addressing the corresponding shortcomings of the original model, thereby "revitalizing" the traditional Bayes-Stein model. Specifically, we propose a time-dependent weighted Elastic Net (TW-ENet) approach to exploit the predictability of financial asset returns for refining the core components of the Bayes-Stein model, namely the sample mean and the shrinkage target (the grand mean). Additionally, we improve the Bayes-Stein model by accommodating the calibration error of the shrinkage factor via the hybrid double selective clustering combination (HDS-CC) technique, forming a four-stage grouped Bayes-Stein shrinkage approach. Further, we propose the graphical adaptive Elastic Net (GA-ENet) method for generating a shrinkage estimator of the inverse covariance matrix to complement the Bayes-Stein model. By bringing these methodological developments together, we build a generalized Bayes-Stein framework for portfolio optimization in the era of machine learning, allowing us to expand the Bayes-Stein model's scope and applicability significantly.

mulations of optimization problems, whereas its reverse emerges in the corresponding solutions. Thus our paper focuses on the inverse covariance matrix.

Our study is related to several strands of literature. Generally, portfolio optimization falls into a "predict-then-optimize" paradigm where the outcome of moments prediction for asset returns can be used to fuel an asset allocation scheme (El Balghiti et al., 2022; Elmachtoub and Grigas, 2022). Thus, methodologies for parameter estimation and decision optimization, including traditional approaches[3] and novel machine learning techniques, can play a crucial role in the asset allocation procedure. Prior literature has leveraged various traditional methods for portfolio optimization. For instance, Anderson and Cheng (2022) propose a Bayesian-averaging heterogeneous vector autoregressive strategy incorporating plenty of return-predicting models to generate optimal portfolio choice and robust out-of-sample performance. Kan, Wang and Zhou (2022) theoretically and empirically introduce an optimal combination strategy flexible with established portfolio rules to reduce estimation risk in the input parameters. In addition, approaches based on traditional statistical or mathematical theories have also been employed in portfolio-related scenarios. For example, Giesecke et al. (2014) obtain the optimal credit swaps portfolios based on a goal program that translates the portfolio selection problem into a constrained optimization of preference-weighted portfolio moments. Sirignano, Tsoukalas and Giesecke (2016) recommend an approximate optimization approach for asset allocation in large portfolios of various risky loans. Moreover, some attempts have been made to improve other portfolio optimization frameworks, such as the Black-Litterman model (see, Bertsimas, Gupta and Paschalidis, 2012; Chen and Lim, 2020). In this paper, we seek to refine the original Bayes-Stein framework by involving some insights from traditional mathematical or statistical theories and advanced machine learning methods. For instance, we mainly employ the combination strategy to increase the stability and robustness of clustering.

In terms of machine learning methods, there is much evidence that they can shed light on solving financial problems (see, e.g., Chen, Pelger and Zhu, 2019; Avramov, Cheng and Metzker, 2022). Generally, three distinct streams of work apply machine learning tools in portfolio choice problems. The first focuses on the estimation issues of various input parameters under the traditional mean-variance paradigm, such as mean returns and the covariance matrix. For example, Ban, El Karoui and Lim (2018) adopt machine learning techniques (regularization and cross-validation) to constrain the sample variances of portfolio risk and return to reduce uncertainty. Besides, Kynigakis and Panopoulou (2022) demonstrate that returns produced by the combination of machine learning forecasting models can provide

[3]Traditional approaches refer to ones without involving machine learning, such as sophisticated econometric models.

superior benefits to asset allocation, thereby building portfolios that outperform the 1/N rule. The second stream is about the sparse portfolio selection problem by directly handling asset weights with machine learning methods. For instance, Ao, Yingying and Zheng (2019) employ the least absolute shrinkage and selection operator (LASSO) to constrain portfolio weights. Similarly, Bertsimas and Cory-Wright (2022) introduce a scalable algorithm for sparse portfolio choice by imposing a ridge regularization on asset weights. Finally, another stream of work is bypassing the mean-variance framework and constructing portfolios with assets selected by machine learning models, thereby augmenting the performance of portfolios. For instance, Cong et al. (2021) grade assets by virtue of a deep learning model (the Transformer model) and directly optimize the investment criterion (Sharpe ratio) through reinforcement learning. In contrast to existing studies on portfolio optimization with machine learning, in this paper, different highly flexible machine learning methods are tailored and implemented to take into account various specifics and limitations of the Bayes-Stein model.

In a nutshell, our study aims to address the shortcomings of the traditional Bayes-Stein model and develop a generalized counterpart that possesses desirable properties and generates superior out-of-sample portfolio performance. We achieve this aim via the novel implementation of machine learning techniques (e.g., supervised learning of regularized regression and unsupervised learning of clustering) in conjunction with statistical or mathematical approaches (e.g., combination theory). As the first study in the literature deconstructing the Bayes-Stein portfolio-optimization framework and improving it by applying machine learning techniques, we make several methodological contributions.

First, the sample-based historical mean returns vector is an essential component of the Bayes-Stein model. It also significantly impacts the grand mean (the expected return of the global minimum variance portfolio), deteriorating the Bayes-Stein mean estimator's accuracy. To deal with this problem, we suggest improving the mean estimator from the standpoint of time-series return forecasting by generating a more accurate estimator to replace the sample mean, which can then be converted into portfolio gains. Specifically, since LASSO-based methods are not designed to cater to specifics of financial time series such as structural breaks and dependence over time and across assets, we propose a TW-ENet technique that incorporates critical time-series information on financial asset returns to deal with this problem. It assigns greater weights to the more recent observations controlled by an adaptive data-driven weighting parameter that make it more suitable for the peculiarities of financial time series. Consequently, we can exploit the return predictability that the

usual Bayes-Stein model entirely ignores. Moreover, we empirically demonstrate that this approach can present a performance edge in general over other sophisticated machine learning methods in the context of portfolio management. Thus our study also offers valuable insights for the research on asset return forecasting.

Second, the shrinkage factor determines the trade-off between the sample means and the grand mean. The conventional Bayes-Stein model assigns the same shrinkage factor to all assets in the portfolio, which fails to capture feature differences between assets and implicitly jeopardizes the accuracy of the mean return estimator especially when the number of assets is relatively large. Hence, we attempt to further improve the original Bayes-Stein model from the perspective of shrinkage factor with refinements that take into account individual differences of assets. Instead of directly handling the shrinkage factor, we analytically introduce the clustering analysis (unsupervised learning) to partition assets into multiple subgroups and accordingly allocate specific shrinkage factors to these homogeneous asset subsets with indistinguishable features. Further, in order to enhance the stability and robustness of clustering, we propose a hybrid double selective clustering combination (HDS-CC) strategy with respect to the quality and diversification of clustering represented by the quality score and the normalized mutual information, respectively. Specifically, we adopt four well-established clustering algorithms, i.e., K-means, Hierarchical clustering, Spectral clustering, and Fuzzy c-means. To the best of our knowledge, we are the first to incorporate the idea of clustering combination (ensemble learning) into shrinkage models for portfolio optimization. Finally, a flexible four-stage grouped Bayes-Stein shrinkage approach based on HDS-CC is developed in our paper. This approach provides the classical shrinkage factor calibration with desirable properties, i.e. capturing individual differences of assets and retaining the time-varying feature.

Last but not least, the original Bayes-Stein model mainly focuses on the shrinkage estimator of mean returns without much attention to the inverse covariance matrix. However, in pursuing the optimal asset allocation decision under the mean-variance framework, the Bayes-Stein model still relies on the sample-based inverse covariance matrix, posing estimation risk and challenges to portfolio optimization. Thus, an accurate and reliable estimation of the inverse covariance matrix will benefit the asset allocation decision process and complement the Bayes-Stein model with a more generalized framework. We herein seek to improve the inverse covariance matrix estimation from the perspective of "sparse hedging". "hedging" implies that elements in the inverse covariance matrix of returns represent hedging relationships between assets, and "sparse" indicates it is beneficial to hedge one asset with some

of the remaining assets (Stevens, 1998; Goto and Xu, 2015). Thus, "sparse hedging" provides a financial interpretation for the inverse covariance matrix. Furthermore, achieving the sparse hedging represented by a sparse inverse covariance matrix is tantamount to Gaussian graphical modeling, where the graphical LASSO method is commonly used in the finance literature. However, issues originating from LASSO can influence the graphical LASSO as well. For example, because LASSO only arbitrarily selects one variable from a group of high-correlated variables and the number of variables selected by LASSO is bounded by the number of observations, the sparsity of the inverse covariance matrix yielded by the graphical LASSO is affected. Hence, we propose a graphical adaptive Elastic Net (GA-ENet) algorithm to estimate the sparse inverse covariance matrix. The GA-ENet approach poses adaptive weights on different elements of the inverse covariance matrix as the adaptive Elastic Net does (Zou and Zhang, 2009), leading to a theoretically superior performance than the graphical LASSO. Additionally, the newly developed GA-ENet approach extends the Bayes-Stein model, allowing it to be applied to large portfolios (the number of assets is comparable to or exceeds the sample size) where the sample-based covariance matrix is unstable or noninvertible.

In the empirical analysis, we comprehensively compare the out-of-sample performance of our generalized Bayes-Stein framework with the naive 1/N rule. Results of various datasets reveal that our generalized method results in superior asset allocation decisions with substantially higher Sharpe ratios and certainty equivalent returns, which is robust to transaction costs, risk aversion parameters, the length of estimation window, and window estimation methods. Moreover, we verify the effectiveness of different parts of our generalized Bayes-Stein model by model decomposition and confirm its combining advantages from various machine learning techniques. Furthermore, in robustness checks, we examine the effectiveness of our time-series return prediction model in the generalized Bayes-Stein framework. Comparing with other machine learning methods, i.e., the ordinary least squares post LASSO (OLS-post LASSO) approach, the combination Elastic Net (C-ENet) method, and the Random Forest technique, our proposed time-dependent weighted Elastic Net (TW-ENet) model is more effective in portfolio optimization. Notably, our empirical study shows that the classic Bayes-Stein portfolio-optimization framework can be improved through different machine learning techniques. At the same time, machine learning models can play a better role when they are adjusted according to the specifics of financial issues and modified by advanced mathematical methods. Thus, our study offers promising and appealing directions for solving challenging financial problems in the era of machine learning.

The remainder of this paper is organized as follows. Section 2 provides a description and well-grounded analysis of the conventional Bayes-Stein model, and extensively analyzes its main drawbacks. We introduce our generalized Bayes-Stein framework in detail in Section 3. In Section 4, the results of the empirical analysis are presented to substantiate the ability of the proposed model to achieve significant out-of-sample performance. Section 5 offers an overview of our robustness checks. Finally, Section 6 concludes the paper and discusses directions for future research. Appendices contain proof of propositions, datasets description, supplementary results, and robustness checks.

2 The Bayes-Stein Model and Its Limitations

Since the Bayes-Stein model stems from the cornerstone results of the James-Stein theory, for analytical tractability, we first briefly outline the statistical background and fundamentals of this model and offer a graphical representation of its portfolio management application. Then, we highlight the three major limitations of its classical version, namely the low accuracy of the mean estimator ($\hat{\boldsymbol{\mu}}_{BS}$), calibration error of the shrinkage factor (g_{BS}), and estimation risk of the inverse covariance matrix ($\hat{\boldsymbol{\Sigma}}_{BS}^{-1}$), for rationally introducing our generalized Bayes-Stein framework in Section 3.

2.1 Description

2.1.1 The James-Stein Theory

Given N independently and normally distributed variables $\boldsymbol{X} = (X_1, \ldots, X_N)'$ with unknown means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N)'$, Stein (1956) stuns the statistical community by demonstrating the inadmissibility of the sample mean estimator when $N > 2$, which is called Stein's Paradox.[4] Stein and James (1961) generalize Stein's demonstration and explicitly provide an estimator known as the James-Stein estimator $\hat{\boldsymbol{\mu}}_{JS}$ that strictly dominates the maximum likelihood estimator $\hat{\boldsymbol{\mu}}_{MLE}$ in terms of the mean squared error (MSE). For simplicity, assuming a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix the identity $\boldsymbol{I}$, $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{I})$, the James-Stein estimator with the property

[4] Inadmissibility of an estimator means that it is dominated by another estimator in terms of some criteria, e.g., total squared error loss.

$\mathbb{E}\left(\|\hat{\boldsymbol{\mu}}_{JS} - \boldsymbol{\mu}\|_2^2\right) < \mathbb{E}\left(\|\hat{\boldsymbol{\mu}}_{MLE} - \boldsymbol{\mu}\|_2^2\right)$ is defined as

$$\hat{\boldsymbol{\mu}}_{JS} = \left(1 - \frac{N-2}{\|\hat{\boldsymbol{\mu}}_{MLE}\|_2^2}\right)\hat{\boldsymbol{\mu}}_{MLE} + \frac{N-2}{\|\hat{\boldsymbol{\mu}}_{MLE}\|_2^2}\mathbf{0}, \tag{1}$$

where $N > 2$ is the number of estimated variables, and the vector $\mathbf{0}$ represents a special case of the shrinkage target, indicating the sample means vector is shrunk toward $\mathbf{0}$.

Furthermore, the James-Stein estimator can be interpreted from an empirical Bayes perspective, offering a general form for $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$:

$$\hat{\boldsymbol{\mu}}_{JS} = (1 - g_{JS})\,\hat{\boldsymbol{\mu}}_{MLE} + g_{JS}\hat{\mu}_{target}\mathbf{1}, \tag{2}$$

where $\hat{\mu}_{target}$ represents the shrinkage target, or prior information, with an arbitrary choice such as zero or the average of all sample means. The scalar $0 \leq g_{JS} \leq 1$ indicates the shrinkage factor, determining the weighting of the maximum likelihood estimator and the shrinkage target. When g_{JS} equals 0, it means no shrinkage, whereas 1 represents full shrinkage. g_{JS} is calculated as follows:

$$g_{JS} = \min\left\{1, \frac{N-2}{T} \cdot \frac{1}{(\hat{\boldsymbol{\mu}}_{MLE} - \hat{\mu}_{target}\mathbf{1})'\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\mu}}_{MLE} - \hat{\mu}_{target}\mathbf{1})}\right\}, \tag{3}$$

where T denotes the time horizon of historical observations.

2.1.2 The Bayes-Stein Model

The Bayes-Stein model is derived from the influential James-Stein theory and the empirical Bayes framework in the context of portfolio selection. A general form of the estimated mean returns vector $\hat{\boldsymbol{\mu}}_{BS}$ of the Bayes-Stein model is given by

$$\hat{\boldsymbol{\mu}}_{BS} = (1 - g_{BS})\,\hat{\boldsymbol{\mu}}_{S} + g_{BS}\mu_G\mathbf{1}, \tag{4}$$

where $\hat{\boldsymbol{\mu}}_S$ represents the vector of sample mean returns, μ_G is the target estimator that refers to the mean return of the global minimum variance portfolio (GMV), and $\mathbf{1}$ is an $N \times 1$ vector of ones. g_{BS}, $0 \leq g_{BS} \leq 1$, indicates the shrinkage factor (or shrinkage intensity),

calculated by a closed-form formula:

$$g_{BS} = \frac{N+2}{(N+2) + T\left(\hat{\boldsymbol{\mu}}_S - \mu_G \mathbf{1}\right)' \boldsymbol{\Sigma}^{-1} \left(\hat{\boldsymbol{\mu}}_S - \mu_G \mathbf{1}\right)}, \tag{5}$$

where $N > 2$ is the number of assets. As the covariance matrix $\boldsymbol{\Sigma}$ of asset returns is unknown in practice, it is replaced with $\frac{T-1}{T-N-2}\boldsymbol{S}$, denoted by $\hat{\boldsymbol{\Sigma}}_S$, where $\boldsymbol{S}$ is the sample covariance matrix.[5]

The James-Stein and Bayes-Stein mean estimators shrink the sample means towards the grand mean, but their shrinkage target selection differs. Theoretically, from the Bayesian perspective, the grand mean represents the prior information about the assets' mean returns, and the true mean is expected to vary around it. Therefore, the resulting expected returns by the Bayes-Stein model can effectively mitigate the estimation error of the sample means by shrinking them towards the target (the mean return of GMV). Furthermore, the Bayes-Stein mean estimator can effectively ease the sensitivity of the parameters of the mean-variance framework by reducing the undesirable influence of outliers of historical returns.

In addition, the covariance matrix of the Bayes-Stein model is estimated as follows:

$$\hat{\boldsymbol{\Sigma}}_{BS} = \left(\frac{T+\varphi+1}{T+\varphi}\right)\hat{\boldsymbol{\Sigma}}_S + \frac{\varphi}{T(T+\varphi+1)}\frac{\mathbf{1}\mathbf{1}'}{\mathbf{1}'\hat{\boldsymbol{\Sigma}}_S^{-1}\mathbf{1}}, \tag{6}$$

where

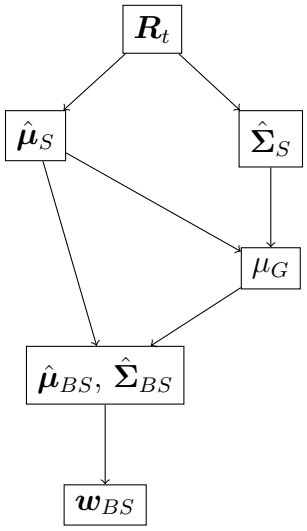
$$\varphi = \frac{N+2}{\left(\hat{\boldsymbol{\mu}}_S - \mu_G \mathbf{1}\right)' \hat{\boldsymbol{\Sigma}}_S^{-1} \left(\hat{\boldsymbol{\mu}}_S - \mu_G \mathbf{1}\right)}. \tag{7}$$

2.1.3 Graphical Representation of the Bayes-Stein Model

There are four steps to complete the optimal asset allocation when integrating the Bayes-Stein model into the mean-variance framework for portfolio selection. The graphical representation is shown in Figure 1.

[5] In this paper, we call $\hat{\boldsymbol{\Sigma}}_S$ the sample-based covariance matrix and $\boldsymbol{S}$ the sample covariance matrix.

Figure 1: Graphical Representation of the Bayes-Stein Model



First, given the historical asset returns $\boldsymbol{R}_t$, we compute the sample mean estimator $\hat{\boldsymbol{\mu}}_S$ and the sample-based covariance matrix $\hat{\boldsymbol{\Sigma}}_S$. Second, we use these two parameters to generate μ_G, the mean return of the global minimum variance (GMV) portfolio whose objective function is given as follows:

$$\min_{\boldsymbol{w}_G} \frac{1}{2} \boldsymbol{w}_G' \hat{\boldsymbol{\Sigma}}_S \boldsymbol{w}_G, \tag{8}$$

where $\boldsymbol{w}_G$ is the optimal weights of GMV.

As a result, μ_G is calculated by $\hat{\boldsymbol{\mu}}_S' \frac{\hat{\boldsymbol{\Sigma}}_S^{-1}\mathbf{1}}{\mathbf{1}'\hat{\boldsymbol{\Sigma}}_S^{-1}\mathbf{1}}$. Third, through (4) and (6), we acquire the Bayes-Stein estimates $\hat{\boldsymbol{\mu}}_{BS}$ and $\hat{\boldsymbol{\Sigma}}_{BS}$. In the end, we involve the results of the third step into the mean-variance framework[6] to generate the optimal asset weights, denoted by the vector $\boldsymbol{w}_{BS}$, by maximizing the following quadratic utility function:

$$\max_{\boldsymbol{w}_{BS}} \left\{ \boldsymbol{w}'_{BS} \hat{\boldsymbol{\mu}}_{BS} - \frac{\lambda}{2} \boldsymbol{w}'_{BS} \hat{\boldsymbol{\Sigma}}_{BS} \boldsymbol{w}_{BS} \right\}, \tag{9}$$

where λ represents the risk aversion parameter.

[6]For simplicity, we do not add any constrains here.

2.2 Limitations of the Bayes-Stein Model

The Bayes-Stein model has been widely used in the portfolio management literature as the benchmark, or the fundamental parameter estimation model of the classic Markowitz mean-variance framework. Nevertheless, few studies delve into it. In this section, we highlight three shortcomings of the original Bayes-Stein model. Note that, for easier understanding, we accordingly conduct some empirical studies to support our statements.[7]

2.2.1 Low Accuracy of the Mean Estimator $\hat{\boldsymbol{\mu}}_{BS}$

Theoretically, portfolio selection is about optimal allocation of weights across assets. Hence various techniques for estimating the input parameters of the mean-variance framework are evaluated by the out-of-sample portfolio performance. Since the accuracy of the mean estimator is crucial to portfolio gains, we first interpret the Bayes-Stein mean estimator by the bias-variance trade-off shown in the following proposition and further assess it from the perspective of forecasting performance.

Proposition 1: By bias-variance decomposition, the mean squared error (MSE) of an estimator $\hat{\boldsymbol{\mu}}$ of the mean returns vector $\boldsymbol{\mu}$ is determined by

$$\mathbb{E}\left[\|\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}\|_2^2\right] = \|\mathbb{E}[\hat{\boldsymbol{\mu}}]-\boldsymbol{\mu}\|_2^2 + \mathbb{E}\left[\|\hat{\boldsymbol{\mu}}-\mathbb{E}[\hat{\boldsymbol{\mu}}]\|_2^2\right], \tag{10}$$

where $\|\mathbb{E}[\hat{\boldsymbol{\mu}}]-\boldsymbol{\mu}\|_2^2$ and $\mathbb{E}\left[\|\hat{\boldsymbol{\mu}}-\mathbb{E}[\hat{\boldsymbol{\mu}}]\|_2^2\right]$ denote bias's square and variance of the estimator $\hat{\boldsymbol{\mu}}$, respectively.

It is well established that a multivariate normal distribution's sample mean (the maximum likelihood estimator) is unbiased, but has high variance. When the sample mean is combined with a shrinkage target, its variance is reduced, whereas its bias increases (Tu and Zhou, 2011). Therefore, the Bayes-Stein model achieves an optimal bias-variance trade-off by shrinking the maximum likelihood estimator towards the grand mean, resulting in a shrinkage mean estimator that dominates the sample mean.

However, though the Bayes-Stein mean estimator can outperform the sample mean in terms of the mean squared forecasting error and out-of-sample portfolio performance, it still suffers from errors caused by two sample-based components in the model, namely the sample

[7]For brevity, herein, we present representative results. The complete results of other datasets used in this paper are reported in Appendix C of the Online Supplementary Appendix.

mean $\hat{\boldsymbol{\mu}}_S$ and the grand mean μ_G, which may lead to its inferiority relative to alternative mean estimators.

To examine the accuracy of different mean estimators, we compute their mean squared forecasting error (MSFE). To illustrate, Table 1 displays the MSFE of the sample mean and the Bayes-Stein mean estimator for the 10 industry portfolios.[8] The results indicate that the Bayes-Stein mean estimator cannot substantially reduce the overall estimation error compared with the sample mean.

Table 1: Mean Squared Forecasting Error of the Mean Return Estimators

Table 1 reports the out-of-sample mean squared forecasting error of the monthly sample mean and Bayes-Stein mean estimator for the 10 industry portfolios. The estimation process is based on a 20-year expanding window with the initial data period from July 1963 to June 1983. The out-of-sample period covers from July 1983 to December 2021.

Asset	Sampe Mean ($\times 10^4$)	Bayes-Stein Shrinkage Mean($\times 10^4$)
Consumer Nondurables	16.86	16.91
Consumer Durables	55.92	55.86
Manufacturing	24.50	24.49
Energy	37.65	37.66
HiTec	44.75	44.72
Telcom	25.10	25.07
Shops	24.03	24.04
Health	21.19	21.22
Utilities	15.52	15.47
Other	26.94	26.93
Sum	292.46	292.37

Moreover, in terms of the forecasting performance of the Bayes-Stein, we can also find some arguments from previous portfolio optimization literature. For example, Jorion (1991) finds that the Bayes-Stein mean estimator fails to outperform the CAPM-based estimator in forecasting asset returns. Similarly, Craig MacKinlay and Pástor (2000) conclude that factor-based asset pricing models are more precise in estimating the expected returns, compared with the Jame-Stein mean estimator. Barroso and Saxena (2021) also report the weak performance of the Bayes-Stein shrinkage strategy in terms of the root mean squared

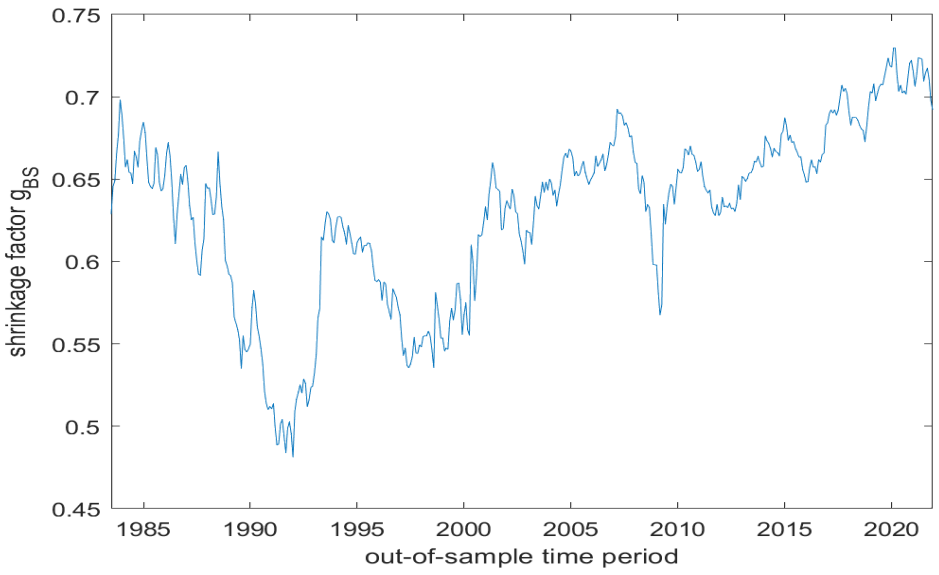
[8]This data is extracted from the website of Ken French. Appendix B of the Online Supplementary Appendix gives a detailed description of the 10 industry portfolios.

forecasting error (RMSFE). Therefore, improving the accuracy of the mean estimator by sophisticated extensions over the Bayes-Stein model will shed some light on the portfolio selection.

2.2.2 Calibration Error of the Shrinkage Factor g_{BS}

The shrinkage factor determines the optimal trade-off between the sample estimator and the shrinkage target. In existing studies, it is usually calculated by solving a heuristic quadratic minimization loss function that is the aggregate of every component's estimation loss (see, e.g., DeMiguel, Martin-Utrera and Nogales, 2013; Kircher and Rösch, 2021). In terms of the Bayes-Stein mean estimator, the shrinkage factor g_{BS} minimizes the total squared error loss of all individual mean estimators of assets. As shown in Figure 2, we can get a fixed shrinkage factor across all assets in each out-of-sample time point.

Figure 2: Shrinkage Factor for 10 Industry Portfolios



As a result, the mean returns obtained via the Bayes-Stein model only dominate the sample means under total squared error. However, this does not imply that each component of the Bayes-Stein shrinkage estimator of the mean returns can outperform the corresponding element in the sample means vector. For example, in Table 1, some individual members of the Bayes-Stein means vector, such as Consumer Nondurables and Shops, even have more estimation risk than the sample mean. Going further, we can conclude that the Bayes-Stein

shrinkage factor is generated without considering the difference between individual assets, which may produce calibration errors for the shrinkage factor. Unfortunately, albeit existing, studies pointing out this issue are scarce.

2.2.3 Estimation Risk of the Inverse Covariance Matrix $\hat{\boldsymbol{\Sigma}}_{BS}^{-1}$

The solution to (9) is $\boldsymbol{w}_{BS} = (1/\lambda)\hat{\boldsymbol{\Sigma}}_{BS}^{-1}\hat{\boldsymbol{\mu}}_{BS}$, with $1-\mathbf{1}'\boldsymbol{w}_{BS}$ invested in the risk-free asset. As a result, the vector of normalized weights ($\boldsymbol{w}$) invested in the risky assets is given as follows:

$$\boldsymbol{w} = \frac{\hat{\boldsymbol{\Sigma}}_{BS}^{-1}\hat{\boldsymbol{\mu}}_{BS}}{\mathbf{1}\hat{\boldsymbol{\Sigma}}_{BS}^{-1}\hat{\boldsymbol{\mu}}_{BS}}. \tag{11}$$

Importantly, (11) suggests that the mean-variance optimal weights are determined by the inverse covariance matrix $\hat{\boldsymbol{\Sigma}}_{BS}^{-1}$ and the mean returns vector $\hat{\boldsymbol{\mu}}_{BS}$. In the classical Bayes-Stein model, $\hat{\boldsymbol{\Sigma}}_{BS}^{-1}$ is proportional to the sample inverse covariance matrix, casting some estimation risk on the estimated inverse covariance matrix, which also influences the final asset allocation decision.

Commonly, the Frobenius norm is used to measure estimation errors of the sample inverse covariance matrix $\boldsymbol{S}^{-1}$ and the Bayes-Stein inverse covariance matrix $\hat{\boldsymbol{\Sigma}}_{BS}^{-1}$, denoted by e, given as follows:

$$e = \|\boldsymbol{\Sigma}^{-1} - \hat{\boldsymbol{\Sigma}}^{-1}\|_F = \left\{\text{trace}\left((\boldsymbol{\Sigma}^{-1} - \hat{\boldsymbol{\Sigma}}^{-1})(\boldsymbol{\Sigma}^{-1} - \hat{\boldsymbol{\Sigma}}^{-1})'\right)\right\}^{1/2}, \tag{12}$$

where $\boldsymbol{\Sigma}^{-1}$ denotes the actual inverse covariance matrix and $\hat{\boldsymbol{\Sigma}}^{-1}$ is its estimate.

As it is hard to obtain the true inverse covariance matrix[9], we use the condition numbers of $\boldsymbol{S}^{-1}$ and $\hat{\boldsymbol{\Sigma}}_{BS}^{-1}$ to demonstrate their estimation risk. In Table 2, we report the results of some commonly used datasets.[10] Results indicate that the sample inverse covariance matrix and Bayes-Stein inverse covariance matrix are ill-conditioned in all datasets. The situation worsens when the number of assets becomes large.

[9] Shi et al. (2019) get the true inverse covariance matrix by generating asset excess returns simulated with the Fama-French 3-factor model.

[10] The detailed description about the datasets is presented in Section 4 and Appendix B of the Online Supplementary Appendix.

Table 2: Condition Numbers for Inverse Covariance Matrix Estimates

Table 2 reports the mean and standard deviation of condition numbers of the sample inverse covariance matrix ($\boldsymbol{S}^{-1}$) and the Bayes-Stein inverse covariance matrix ($\hat{\boldsymbol{\Sigma}}_{BS}^{-1}$) during the out-of-sample period over three random datasets (Ind10, FF25, and FF100BM), to demonstrate the estimation risk of the Bayes-Stein inverse covariance matrix. The estimation process is based on a 20-year expanding window with the initial data period from July 1963 to June 1983. The out-of-sample period covers from July 1983 to December 2021.

	Ind10		FF25		FF100BM	
	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
$\boldsymbol{S}^{-1}$	114.9	21.2	969.47	110.37	3018.5	840.7
$\hat{\boldsymbol{\Sigma}}_{BS}^{-1}$	115	21.2	969.78	110.52	3019	840.9

3 Generalized Bayes-Stein Framework with Machine Learning

This section compares the classical Bayes-Stein model and its generalized counterpart we develop, followed by an elaboration of the generalized Bayes-Stein framework and how it works with machine learning. Specifically, we suggest a time-dependent weighted Elastic Net (TW-ENet) approach from the perspective of time-series return forecasting to replace the sample means vector in the Bayes-Stein model, a four-stage grouped Bayes-Stein shrinkage strategy based on the hybrid double selective clustering combination (HDS-CC) method to improve the measurement of shrinkage factors, and a graphical adaptive Elastic Net (GA-ENet) algorithm to estimate the inverse covariance matrix. As a result, a holistic generalized Bayes-Stein framework integrating these proposed approaches is established.

3.1 Overview

In principle, both the standard Bayes-Stein model and the generalized version concern the input parameters of the Markowitz mean-variance framework, namely the expected returns and the covariance matrix or its inverse, but they differ in their model construction process and specific components. Overall, our generalized Bayes-Stein model incorporates insights from machine learning methods to replace the three branches of the original mean estimator, including the sample means vector $\hat{\boldsymbol{\mu}}_S$, the grand mean μ_G, and the shrinkage factor g_{BS},

and generates a refined inverse covariance matrix $\hat{\boldsymbol{\Sigma}}_{GBS}^{-1}$. A comparison is presented in Table 3. We will go through it in depth now.

Table 3: Comparison of the Classical and Generalized Bayes-Stein Framework

Model Traits	Bayes-Stein Model	Generalized Bayes-Stein Framework
Sample Mean	$\hat{\boldsymbol{\mu}}_S$, simple average of historical asset returns	$\hat{\boldsymbol{\mu}}_F$, expected returns predicted by various time-series predictors via the TW-ENet approach
Grand Mean	μ_G, mean return of GMV, depends on $\hat{\boldsymbol{\mu}}_S$	μ_{GF}, mean return of GMV depends on $\hat{\boldsymbol{\mu}}_F$
Shrinkage Factor	same across all assets	same across assets of the same subgroup generated by clustering ensemble
Mean Estimator	$\hat{\boldsymbol{\mu}}_{BS}$, low accuracy	$\hat{\boldsymbol{\mu}}_{GBS}$, relatively high accuracy
Inverse Covariance Matrix	$\hat{\boldsymbol{\Sigma}}_{BS}^{-1}$, sample-based	$\hat{\boldsymbol{\Sigma}}_{GBS}^{-1}$, a sparse shrinkage estimator produced by the GA-ENet

The sample mean ($\hat{\boldsymbol{\mu}}_S$, $\hat{\boldsymbol{\mu}}_F$) and the grand mean (μ_G, μ_{GF}): In terms of the mean estimator, we have empirically demonstrated that the ability of the original Bayes-Stein mean estimator used to predict the expected asset returns remains questionable.[11] Since the Bayes-Stein mean estimator strongly relies on three components, namely the sample mean $\hat{\boldsymbol{\mu}}_S$, the grand mean μ_G, and the shrinkage factor g_{BS}, accurate and reliable input parameters will enhance the accuracy of the mean estimator. We first consider the sample mean and the grand mean. The classical Bayes-Stein model derives from purely statistical principles, which only uses historical asset returns and ignores other important financial time-series information that might help explore the predictability of asset returns.

In our extended version of the Bayes-Stein model, we build a time-series return forecasting model to create a more accurate return prediction to replace the sample mean and update the grand mean, resulting in a significantly improved shrinkage mean estimator. Fur-

[11] If not especially specified, in this paper, the expected asset returns actually refer to the expected excess returns over the risk-free rate.

thermore, in the age of big data, forecasting returns by typical econometric approaches (e.g., linear regression models and the autoregressive moving average model) is challenging due to the existence of numerous predictors and dynamic changes in financial time-series data (Lettau and Pelger, 2020; Martin and Nagel, 2022). Hence, we propose a data-driven time-dependent weighted Elastic Net (TW-ENet) technique for forecasting asset returns, which accommodates the characteristics of financial time-series data, such as high dimensionality, structural breaks, and dependence over time and across assets.

The shrinkage factor (g_{BS}): Consider now the shrinkage factor. Mathematically, (5) indicates that the Bayes-Stein shrinkage factor g_{BS} depends on three time-varying parameters, viz. the sample mean $\hat{\boldsymbol{\mu}}_S$, the grand mean μ_G, and the inverse covariance matrix $\boldsymbol{\Sigma}^{-1}$. Hence the calculated g_{BS} is also time-varying (as shown in Figure 2), which means it is a function of data samples:

$$g_{BS}(t) = f\left(\mathbf{R}(t)\right), \tag{13}$$

where $\mathbf{R}(t)$ denotes asset returns of different time periods.

So, the conventional Bayes-Stein model and its generalized counterpart enjoy the shrinkage factor's time-varying feature. However, the classical version assigns the same shrinkage factor to all assets in the portfolio regardless of their differences. In contrast, the generalized framework considers individual differences in assets, which means the shrinkage factor is not only time-varying but also asset-varying. Specifically, it uses a hybrid double selective clustering combination (HDS-CC) scheme to partition assets into distinct subsets with different shrinkage factors, thereby mitigating the calibration error of the shrinkage factor by offering gains to the conventional Bayes-Stein model.

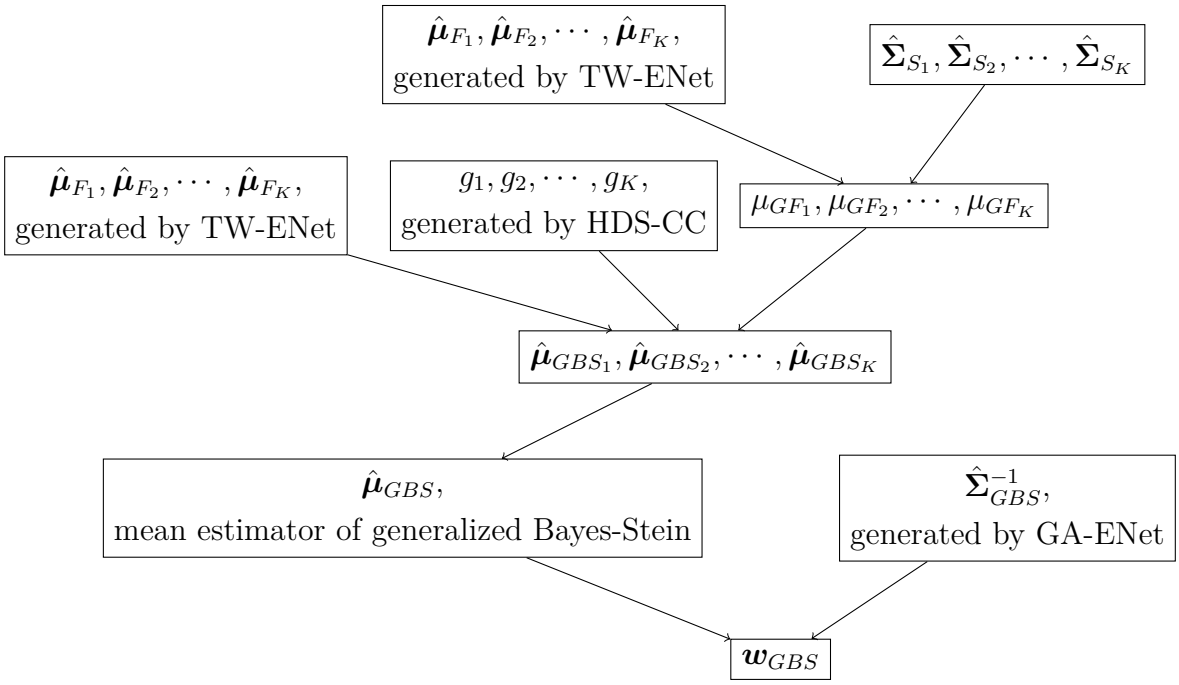
The inverse covariance matrix ($\hat{\boldsymbol{\Sigma}}_{BS}^{-1}$, $\hat{\boldsymbol{\Sigma}}_{GBS}^{-1}$): Finally, the inverse covariance matrix is a core ingredient in transforming assets' mean returns into optimal weights of the mean-variance portfolio. The original Bayes-Stein model calculates the sample-based covariance matrix first and then obtains its reverse, which is unstable and suffers from serious estimation errors. Moreover, the estimation error of the sample covariance matrix and its inverse is closely related to N/T, where N is the number of assets and T is the sample size (Kan and Zhou, 2007). The estimation error will significantly rise when N/T grows somewhat high (but less than 1). On the other hand, when $N/T > 1$ (large portfolios), it even raises the problem of the sample covariance matrix's irreversibility.

By contrast, the generalized Bayes-Stein model directly deals with estimating the in-

verse covariance matrix and suggests a sparse estimator for the inverse covariance matrix. Specifically, it applies our proposed graphical adaptive Elastic Net (GA-ENet) approach to meliorate the sparse inverse covariance matrix estimation. As a result, the generalized model improves the inverse covariance matrix estimation and complements the standard Bayes-Stein model by generating a shrinkage estimator for the inverse covariance matrix, making it more flexible and capable of constructing large portfolios.

By and large, the generalized Bayes-Stein framework attempts to address critical drawbacks of the classical Bayes-Stein model, thereby providing refined input parameters to the traditional portfolio optimization framework. As a result, the plug-in mean-variance strategy can benefit from improved parameter estimation and takes asset allocation decisions yielding better out-of-sample portfolio performance. Its graphical representation is presented in Figure 3, where the superscript $1, 2, \cdots, K$ denotes the index of different asset groups yielded by the clustering analysis in HDS-CC. $g_1, g_2, \cdots, g_K$ denote the specific shrinkage factor for the corresponding asset group. $\hat{\boldsymbol{\mu}}_{F_1}, \hat{\boldsymbol{\mu}}_{F_2}, \cdots, \hat{\boldsymbol{\mu}}_{F_K}$, and $\mu_{GF_1}, \mu_{GF_2}, \cdots, \mu_{GF_K}$ are constructed with the TW-ENet approach, and represent the upgraded sample mean and grand mean of the different asset groups, respectively. Consequently, the optimal asset weights vector $\boldsymbol{w}_{GBS}$ is produced based on the generalized Bayes-Stein framework.

Figure 3: Graphical Representation of the Generalized Bayes-Stein Framework



With this graph in mind, we can proceed to elaborate on the generalized Bayes-Stein framework in the era of machine learning.

3.2 Time-series Return Forecasting

First, the generalized Bayes-Stein model seeks to improve the classical Bayes-Stein mean estimator by substituting the sample mean components with the results of time-series return forecasting. Essentially, the return forecasting process amounts to extracting efficient information from several predictors, expressed by a conventional predictive regression model with D predictors:

$$r_t = \beta_0 + \sum_{d=1}^{D} \beta_d x_{d,t-1} + \epsilon_t, \tag{14}$$

where r_t is the expected excess return at time t, and $x_{d,t-1}$ is the d^{th} predictor variable at time $t-1$. ϵ_t represents a zero-mean disturbance term.

In the light of Welch and Goyal (2008), traditional linear regression models exhibit poor out-of-sample performance in the high-dimensional financial time series space due to multicollinearity and overfitting, which makes machine learning approaches suitable for asset returns prediction. Moreover, among various machine learning methods for returns forecasting, the least absolute shrinkage and selection operator (LASSO) and its variants (e.g, adaptive LASSO and Elastic Net) stand out owing to their predictive accuracy and model interpretability. They can perform variable selection and parameter estimation simultaneously (Tibshirani, 1996) and have been widely used to explore the predictability of asset returns (see, e.g., Freyberger, Neuhierl and Weber, 2020; Dong et al., 2022). The objective function of a popular LASSO-based approach, Elastic Net (Zou and Hastie, 2005), is given as follows:

$$\underset{\beta_0,\ldots,\beta_D \in \mathbb{R}}{\arg\min} \left[\frac{1}{2T} \sum_{t=1}^{T} \left(r_t - \beta_0 - \sum_{d=1}^{D} \beta_d x_{d,t-1} \right)^2 + \tau \left(\rho \sum_{d=1}^{D} |\beta_d| + \frac{1}{2}(1-\rho) \sum_{d=1}^{D} \beta_d^2 \right) \right], \tag{15}$$

where the parameter ρ represents a compromise between ridge ($\rho = 0$) and LASSO ($\rho = 1$). The tuning parameter τ controls the overall penalty strength.

However, existing studies that leverage LASSO-based methods for return prediction and

portfolio optimization fail to account for the specifics of financial time series. For example, LASSO-based regression models are vulnerable to structural breaks referring to abrupt time series changes, leading to models' unreliability and huge forecasting errors (Pesaran, Pick and Pranovich, 2013). In the light of of Wang, Hao and Wu (2021), more attention should be paid to more recent observations in the presence of structural breaks. Thus, we tailor the Elastic Net approach by providing observations with time-dependent weights to improve forecast accuracy. Specifically, we propose a time-dependent weighted Elastic Net (TW-ENet) approach to forecast expected returns for replacing the sample mean returns in the classical Bayes-Stein model, given as follows:

$$\underset{\beta_0,\ldots,\beta_D\in\mathbb{R}}{\arg\min}\left[\frac{1}{2T}\sum_{t=1}^{T}w_t\left(r_t-\beta_0-\sum_{d=1}^{D}\beta_d x_{d,t-1}\right)^2+\tau\left(\rho\sum_{d=1}^{D}|\beta_d|+\frac{1}{2}(1-\rho)\sum_{d=1}^{D}\beta_d^2\right)\right], \quad (16)$$

where $w_t = t^\delta$ represents the time-dependent exponential weight controlled by a positive hyperparameter δ.

Since t increases with time moving in historical time series and δ is greater than 0, the exponential function, $w_t = t^\delta$, imposes greater weights on more recent observations. Additionally, our TW-ENet approach is equivalent to the classical Elastic Net method when δ is set to 0. Moreover, because we take market components (e.g., industry portfolios and factor-sorted portfolios) as portfolio assets, following Kong et al. (2011) and Rapach et al. (2019), we use various commonly used economic variables[12] and lagged components returns[13] as independent variables.

To employ the TW-ENet approach, we equally divide the sample up to time T into the training sample $(1,\ldots,T/2)$ and validation sample $(T/2+1,\ldots,T)$ in each estimation window, and then we proceed with the following three steps:

Step I - estimation in the training sample: For L weight parameter candidates[14], denoted by $\delta_1,\ldots,\delta_L$, we apply (16) to the training sample. Further, for each candidate, following the study of Rapach and Zhou (2020), we set $\rho = 0.5$ and use the corrected Akaike information criterion (AICc) to choose the optimal L1 penalizing parameter τ_*, given as

[12] In previous studies, they are regarded as efficient predictors for equity returns (Welch and Goyal, 2008). See Appendix B of the Online Supplementary Appendix for details.

[13] We employ first-order lagged returns of portfolio assets as a set of predictors. Thus dependence over time and across assets in financial time series is also considered.

[14] In our practice, the hyperparameter δ is set from 0 to 10 with a fixed step of 0.1.

follows:

$$\tau_* = \underset{\tau_1,\ldots,\tau_M}{\arg\min} \left[N_S * \log(RSS/N_S) + 2 * df * N_S/(N_S - df) \right], \tag{17}$$

where RSS and df represent the residual sum of squares and degrees of freedom in the corresponding penalized linear regression model, respectively. N_S denotes the number of observations in the training sample and $\tau_1, \ldots, \tau_M$ denotes the set of τ.

Step II - weight parameter tuning in the validation sample: For each weight parameter candidate in $\delta_1, \ldots, \delta_L$, we employ the model estimated in the first step on the validation sample to predict expected asset returns. Then we pick out the optimal weight parameter δ_* by minimizing the weighted square prediction error [15], given as follows:

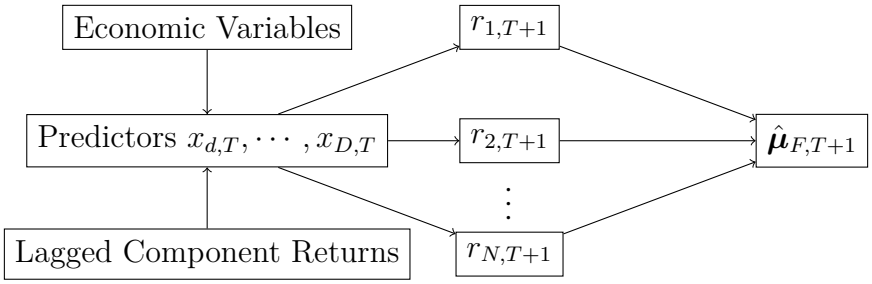
$$\delta_* = \underset{\delta_1,\ldots,\delta_L}{\arg\min} \left[\frac{\sum_{t=T/2+1}^{T} w_{t-T/2}^{T/2} \left(r_t - \beta_0 - \sum_{d=1}^{D} \beta_d x_{d,t-1} \right)^2}{\sum_{t=T/2+1}^{T} w_{t-T/2}^{T/2}} \right]. \tag{18}$$

Step III - estimation in the total sample and forecasting: We apply δ_* and (16) on the total sample $1, 2, \ldots, T$. Likewise, the optimal penalizing parameter for τ is obtained according to the AICc criteria. As a result, we can employ the estimated model to forecast the expected asset return at time $T + 1$.

After conducting our time-series return forecasting for all assets in the portfolio, we obtain the expected asset returns vector $\hat{\boldsymbol{\mu}}_F$, thereby replacing related components in the Bayes-Stein model. A graphical representation of this approach is presented in Figure 4, where critical financial time-series information about asset returns, including economic variables and lagged component returns, are combined as predictors for asset returns and inputs of our TW-ENet approach, and $r_{1,T+1}, r_{2,T+1}, \cdots, r_{N,T+1}$ and $\hat{\boldsymbol{\mu}}_{F,T+1}$ denote individual return forecasting for N assets and expected asset returns vector at time $T + 1$, respectively.

[15] This error function is borrowed from Wang, Hao and Wu (2021).

Figure 4: Graphical Representation of the Time-series Return Forecasting



3.3 Improved Calibration of the Shrinkage Factor g_{BS}

Now we consider to further refine the Bayes-Stein mean estimator by improving the calibration of the shrinkage factor from the standpoint of grouping assets based on their different characteristics. Specifically, we propose an advanced clustering combination strategy - the hybrid double selective clustering combination (HDS-CC) technique, to capture assets' individual differences and provide shrinkage factors customized to distinct asset groups. Finally, we establish a four-stage grouped Bayes-Stein shrinkage method. To introduce our approach, we first discuss the motivation behind involving clustering analysis and clustering combination in the Bayes-Stein model for improving the shrinkage factor, and then we elaborate on our proposed strategy.

In the light of Mynbayeva, Lamb and Zhao (2022), when applying shrinkage estimators to homogeneous asset subsets with indistinguishable mean returns or variances, one can create more robust portfolios than the vanilla Markowitz optimization and outperform the 1/N rule. Inspired by their work, a valuable endeavor for improving the shrinkage factor is partitioning assets into distinct subsets with homogeneous characteristics and allocating specific shrinkage factors to them. Additionally, asset returns exhibit grouped heterogeneity (Cong et al., 2022) and assets with similar idiosyncrasies are prone to appear co-movement and exhibit similar price behaviors (Herskovic et al., 2016), which also signifies the necessity of dividing assets before employing the Bayes-Stein model. Therefore, we herein advocate using clustering analysis, a type of unsupervised machine learning that can divide a collection of assets into subgroups based on their features. That is to say, assets with a high degree of similarity are partitioned into one group, where assets are more likely to have similar performance. As a result, generated shrinkage factors for multiple asset subsets can capture individual differences of assets and, simultaneously, retain the time-varying feature.

Furthermore, though many studies have exploited clustering algorithms for financial time series analysis or portfolio optimization (e.g., Tola et al., 2008; Dias, Vermunt and Ramos, 2015), the choice of clustering methods in their studies is arbitrary. Moreover, a single clustering algorithm cannot guarantee stability and robustness. So, we further introduce a clustering combination technique, which gathers multiple base clustering results to provide a consensus output, effectively increasing the accuracy, stability, robustness, and consistency of clustering (Strehl and Ghosh, 2002).[16] To the best of our knowledge, our study is the first to incorporate the idea of clustering combination into shrinkage models for portfolio selection.

General clustering combination: Typically, the clustering combination technique encompasses two steps. In the first step, we run different clustering algorithms, or the same algorithm with different initialization or parameters, to generate multiple clustering results, also known as base clusterings. Let $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\} \subset \boldsymbol{R}^p$ represent the data set of N assets in a p-dimension feature space. In this paper, we build the feature space by return-based characteristics of assets, namely historical mean returns, volatility, skewness, and kurtosis. A specific clustering algorithm will partition the data set into K clusters, which can be represented as a label factor vector denoted by $\boldsymbol{\lambda} \in \boldsymbol{N}^N$ (Strehl and Ghosh, 2002). The label factor vector divides the asset component $\boldsymbol{x}_i$ into the k^{th} cluster, where $k \in \{1, 2, \ldots, K\}$. Figure 5 gives an example of the label factor vectors for four different base clusterings of four assets partitioned into up to three clusters.

Figure 5: Base Clusterings of 4 Assets

	$\lambda^{(1)}$	$\lambda^{(2)}$	$\lambda^{(3)}$	$\lambda^{(4)}$
x_1	1	1	1	1
x_2	2	1	1	2
x_3	2	2	2	3
x_4	3	2	3	3

In the second step, the base results of the first step are combined to yield the final clusters by clustering combination methods.[17] Generally, the clustering combination function Γ can

[16] Clustering combination technique is also called cluster ensembles or consensus clustering, introduced by Strehl and Ghosh (2002) (with over 5600 Google citations as of December 2022).

[17] Various clustering combination methods have been proposed in the computer science field, which is out of the scope of this paper.

be defined as:

$$\Gamma : \left\{ \boldsymbol{\lambda}^{(b)} \mid b \in \{1, \ldots, D\} \right\} \rightarrow \boldsymbol{\lambda}^{*} \tag{19}$$

where $\boldsymbol{\lambda}^{(b)}$ denotes the label factor vector of the b^{th} base clustering results, $\boldsymbol{\lambda}^{*}$ represents the final label factor vector, and D is the number of base clusterings.

Further, Topchy, Jain and Punch (2005) conclude that the accuracy and diversity of base clusterings are crucial to the performance of the clustering combination. Therefore, in contrast to the existing studies about applications of general clustering ensemble, we employ the same clustering algorithm with different initialization and other clustering algorithms to improve the diversity and accuracy of base clusterings. Specifically, we propose a hybrid double selective clustering combination (HDS-CC) technique to obtain final subgroups of assets, thereby forming a four-stage grouped Bayes-Stein shrinkage approach.

Stage I - enhanced base clusterings with the same algorithm: It has been well-noticed that some clustering algorithms are susceptible to randomly selected initial parameters. For example, the outputs of K-means clustering are vulnerable to different initial centers. This situation is also seen in other clustering models, such as Spectral clustering and the Gaussian mixture model (GMM). To refine the results of such methods, we propose the following double selective clustering combination algorithm concerning the quality and diversification of base clusterings. Note that it is flexible and applicable to simultaneously enhance base clusterings of several clustering methods in this stage. Additionally, for simplicity, we only employ K-means in Stage I.[18]

First, we select base clusterings of high intra-cluster similarity and low inter-cluster similarity, and discard the remaining to maintain clustering quality. To mathematically measure the performance of base clusterings, we introduce the Calinski-Harabasz score[19] denoted by CH, given as follows:

$$CH = \frac{\sum_{k=1}^{K} n_k \left\| c_k - c \right\|^2}{\sum_{k=1}^{K} \sum_{i=1}^{n_k} \left\| \boldsymbol{x}_{i,k} - c_k \right\|^2} \cdot \frac{N-K}{K-1}, \tag{20}$$

where K is the number of clusters of a base clustering, N is the total number of assets, c is the global centroid of all assets, n_k and c_k are the number of assets and centroid in the k^{th}

[18] In this stage, we run K-means many times to produce base clusterings.

[19] There are various clustering quality metrics such as Silhouette coefficient, Davies-Bouldin index, etc. Since they play similar roles, we only use the Calinski-Harabasz score for simplicity.

cluster, respectively, and $\boldsymbol{x}_{i,k}$ represents the i^{th} asset in the k^{th} cluster.

After obtaining quality scores of all base clusterings, we can get a quality set $\mathcal{S}_{quality}$ with a base clusterings by a threshold, denoted by $\left\{\boldsymbol{\lambda}^{(1)}, \ldots, \boldsymbol{\lambda}^{(a)}\right\}$. We call this process "first filtering".

Second, to assess the diversification of base clustering, Strehl and Ghosh (2002) suggest using mutual information and normalizing it ranging from 0 to 1, which can measure the statistical information shared by two label factor vectors of base clusterings, given in the following proposition.

Proposition 2: Denoting two label factor vectors by $\boldsymbol{\lambda}^{(p)}$ and $\boldsymbol{\lambda}^{(q)}$, their normalized mutual information $\Phi^{(\mathrm{NMI})}\left(\boldsymbol{\lambda}^{(p)}, \boldsymbol{\lambda}^{(q)}\right)$ is determined by

$$\Phi^{(\mathrm{NMI})}\left(\boldsymbol{\lambda}^{(p)}, \boldsymbol{\lambda}^{(q)}\right) = \frac{\sum_{h=1}^{k^{(p)}} \sum_{f=1}^{k^{(q)}} n_{h,f} \log\left(\frac{N \cdot n_{h,f}}{n_h^{(p)} n_f^{(q)}}\right)}{\sqrt{\left(\sum_{h=1}^{k^{(p)}} n_h^{(p)} \log \frac{n_h^{(p)}}{N}\right)\left(\sum_{f=1}^{k^{(q)}} n_f^{(q)} \log \frac{n_f^{(q)}}{N}\right)}}, \tag{21}$$

where $n_h^{(p)}$ and $n_f^{(q)}$ represent the number of assets in the cluster C_h of $\boldsymbol{\lambda}^{(p)}$ and cluster C_f of $\boldsymbol{\lambda}^{(q)}$, respectively. $n_{h,f}$ denotes the number of assets appearing in the cluster C_h and C_f simultaneously. N is the number of all assets.

Thus, the average normalized mutual information $m^{(q)}$ of the base clustering $\boldsymbol{\lambda}^{(q)}$ can be computed by

$$m^{(q)} = \frac{1}{a-1} \sum_{p=1, p \neq q}^{a} \Phi^{\mathrm{NMI}}\left(\boldsymbol{\lambda}^{(p)}, \boldsymbol{\lambda}^{(q)}\right), \quad q = 1, \ldots, a. \tag{22}$$

Following Zhou and Tang (2006), the weight for $\boldsymbol{\lambda}^{(q)}$ is defined as:

$$w^{(q)} = \frac{m^{(q)}}{\sum_{q=1}^{a} m^{(q)}}. \tag{23}$$

By setting a threshold $\frac{1}{a}$, selected clusterings form a combination set $\mathcal{S}_{combination}$ (this

process is called "second filtering"), determined by

$$\mathcal{S}_{combination} = \left\{ q \mid w^{(q)} \geq \frac{1}{a}, 1 \leq q \leq a \right\}. \tag{24}$$

Finally, the final label λ_i of asset i is obtained by weighted voting:

$$\lambda_i = \underset{l \in \{1,\ldots,K\}}{\arg\max} \sum_{q \in \mathcal{S}_{combination}} w^{(q)} \cdot \mathbb{I}\left(\lambda_i^{(q)} = l\right). \tag{25}$$

Consequently, we produce an enhanced base clustering of K-means. Here is the illustration of the double selective clustering combination algorithm.

Algorithm: double selective clustering combination algorithm

Input: B base clusterings of N assets, quality score threshold s_{thr}

Output: final clustering $\boldsymbol{\lambda}^*$

for $b = 1$ *to* B **do**
  calculate the quality score CH_b
"first filtering": build the quality set $\mathcal{S}_{quality} = \{b \mid CH_b \geq s_{thr}, 1 \leq b \leq B\}$, with a base clusterings
for $q \in \mathcal{S}_{quality}$ **do**
  calculate the average normalized mutual information $m^{(q)}$ for the base clustering $\boldsymbol{\lambda}^{(q)}$
for $q \in \mathcal{S}_{quality}$ **do**
  compute the weight $w^{(q)}$ of the base clustering $\boldsymbol{\lambda}^{(q)}$
"second filtering": build the combination set $\mathcal{S}_{combination} = \left\{ q \mid w^{(q)} \geq \frac{1}{a}, 1 \leq q \leq a \right\}$
for $i = 1$ *to* N **do**
  obtain the final label λ_i of asset i by weighted voting with the combination set $\mathcal{S}_{combination}$
return final clustering $\boldsymbol{\lambda}^*$

Stage II - base clusterings with different algorithms: In Stage II, we apply three other well-established clustering methods as the base algorithms, namely Hierarchical clustering, Spectral clustering, and Fuzzy c-means.[20] As a result, we can generate three

[20]Since clustering combination is flexible on clustering algorithms, we apply three algorithms for simplicity.

algorithm-specific base clusterings in this stage.

Stage III - clustering combination and grouped shrinkage: In Stage III, we combine four base clusterings yielded by previous stages to gain final results via the double selective combination technique proposed in Stage I. Since we produce an enhanced base clustering by the same algorithm in Stage I and ordinary base clusterings with different algorithms in Stage II, and apply the selective combination technique twice [21], this methodology is "hybrid". It turns out that assets are divided into clusters with high intra-cluster similarity and low inter-cluster similarity. So assets that belong to the same group are allocated the same shrinkage factor, whereas assets that belong to different groups are assigned different shrinkage factors. The Bayes-Stein mean estimator of an individual group k is given as follows:

$$\hat{\boldsymbol{\mu}}_{GBS_k} = (1 - g_k)\, \hat{\boldsymbol{\mu}}_{F_k} + g_k \mu_{GF_k} \mathbf{1}, \tag{26}$$

where $k \in \{1, \ldots, K\}$ denotes the index of clustering groups and g_k represents the corresponding grouped shrinkage factor. $\hat{\boldsymbol{\mu}}_{F_k}$ and μ_{GF_k} are produced by the methodology of Section 3.2.

Stage IV - grouped shrinkage mean estimators aggregating: Finally, in Stage IV, the estimated mean vectors of all subgroups are aggregated to generate a mean returns vector of all assets, given as follows:

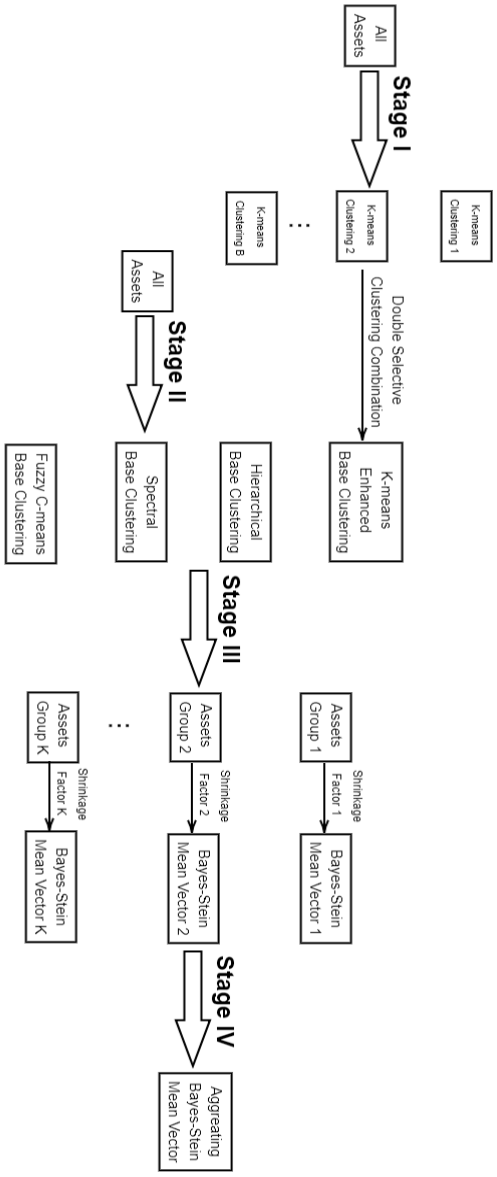
$$\hat{\boldsymbol{\mu}}_{GBS} = \text{aggregating}\left(\hat{\boldsymbol{\mu}}_{GBS_k}\right), \quad k \in \{1, \ldots, K\}. \tag{27}$$

For illustration, Figure 6 depicts the workflow of the grouped Bayes-Stein shrinkage approach. Compared with the conventional scheme for calculating the shrinkage factor in the Bayes-Stein model, the advantages of the grouped Bayes-Stein shrinkage approach are twofold. First, the hybrid double selective clustering combination (HDS-CC) strategy produces highly accurate and robust clusters concerning the time-series characteristics of assets (historical mean returns, volatility, skewness, and kurtosis). Therefore, separately applying the Bayes-Stein shrinkage to different groups is more reasonable. Hence, the grouped Bayes-Stein shrinkage approach generates representative shrinkage factors appropriate to each cluster of assets, which effectively accounts for individual differences between groups of assets. Second, the grouped Bayes-Stein shrinkage approach still employs the conven-

[21] The number of base clustering in Stage III is limited, so we discard the "filtering" process in this stage.

tional mathematical method of the Bayes-Stein model to determine the shrinkage factor for different subgroups of assets. This strategy takes full advantage of the interpretability and accuracy of the conventional mathematical approach and retains the shrinkage factor's time-varying feature.

Figure 6: Grouped Bayes-Stein Shrinkage Approach



3.4 Improved Estimation of the Inverse Covariance Matrix $\hat{\mathbf{\Sigma}}_{BS}^{-1}$

The evidence shown in Section 2 demonstrates that the traditional Bayes-Stein model calculates the sample-based covariance matrix first and then its reverse, which brings estimation errors that impair portfolio performance. This section provides an intuitive and far more direct method for improving the inverse covariance matrix estimation. To introduce our technique, we discuss the asset allocation interpretation of the sparse inverse covariance matrix and its Gaussian graphical modeling meaning, followed by the presentation of the graphical adaptive Elastic Net method.

3.4.1 Gaussian Graphical Modeling of the Inverse Covariance Matrix

The proceeding part explains why we seek to obtain a sparse estimator of the inverse covariance matrix. First, we decompose the covariance matrix $\mathbf{\Sigma}$ of N assets as follows:

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \boldsymbol{m}' \\ \boldsymbol{m} & \boldsymbol{M} \end{bmatrix}, \tag{28}$$

where the first element of $\mathbf{\Sigma}$, σ_1^2, is the variance of the first asset. $\boldsymbol{m}$ and its transpose $\boldsymbol{m}'$ are an $N-1 \times 1$ column vector and an $1 \times N-1$ row vector, respectively, indicating the first asset's covariances with each other $N-1$ assets. The square matrix $\boldsymbol{M}$ represents the covariance matrix of the remaining $N-1$ assets. Similarly, we can decompose the covariance matrix with of all other $N-1$ assets. Further, the inverse covariance matrix $\mathbf{\Sigma}^{-1}$ is given as follows:

$$\mathbf{\Sigma}^{-1} = \begin{bmatrix} (\sigma_1^2 - \mathbf{m}'\mathbf{M}^{-1}\mathbf{m})^{-1} & -(\sigma_1^2 - \mathbf{m}'\mathbf{M}^{-1}\mathbf{m})^{-1}\,\mathbf{m}'\mathbf{M}^{-1} \\ -(\sigma_1^2 - \mathbf{m}'\mathbf{M}^{-1}\mathbf{m})^{-1}\,\mathbf{M}^{-1}\mathbf{m} & \mathbf{M}^{-1} + (\sigma_1^2 - \mathbf{m}'\mathbf{M}^{-1}\mathbf{m})^{-1}\,\mathbf{M}^{-1}\mathbf{m}\mathbf{m}'\mathbf{M}^{-1} \end{bmatrix}, \tag{29}$$

In the light of Stevens (1998), the first row of the inverse covariance matrix $\mathbf{\Sigma}^{-1}$ is closely linked to a multivariate regression function that regresses the return of the first asset on those of all other assets:

$$R_{1,t} = a_1 + \sum_{n=2}^{N} \beta_{1n} R_{n,t} + \epsilon_{1,t}, \quad t = 1, \ldots, T, \tag{30}$$

where the regression coefficients can be indicated by an $1 \times N-1$ row vector $\boldsymbol{\beta_1} = (\beta_{12}, \ldots, \beta_{1N})$,

equal to $\mathbf{m}'\mathbf{M}^{-1}$ (see also Coqueret and Guida, 2020). In addition, we can further analyze the relationship between $\epsilon_{1,t}$ and $\mathbf{\Sigma}^{-1}$ according to the following proposition.

Proposition 3: The variance of $\epsilon_{1,t}$, denoted by $\sigma^2_{\epsilon_1}$, is equivalent to $\sigma_1^2 - \mathbf{m}'\mathbf{M}^{-1}\mathbf{m}$.

Consequently, the first row of $\mathbf{\Sigma}^{-1}$ can be expressed as follows:

$$\begin{bmatrix} \frac{1}{\sigma^2_{\epsilon_1}} & -\frac{\beta_{12}}{\sigma^2_{\epsilon_1}} & \cdots & -\frac{\beta_{1N}}{\sigma^2_{\epsilon_1}} \end{bmatrix}. \tag{31}$$

As a result, the inverse covariance matrix $\mathbf{\Sigma}^{-1}$ is determined by:

$$\begin{bmatrix} \frac{1}{\sigma^2_{\epsilon_1}} & -\frac{\beta_{12}}{\sigma^2_{\epsilon_1}} & \cdots & -\frac{\beta_{1N}}{\sigma^2_{\epsilon_1}} \\ -\frac{\beta_{21}}{\sigma^2_{\epsilon_2}} & \frac{1}{\sigma^2_{\epsilon_2}} & \cdots & -\frac{\beta_{2N}}{\sigma^2_{\epsilon_2}} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ -\frac{\beta_{N1}}{\sigma^2_{\epsilon_N}} & -\frac{\beta_{N2}}{\sigma^2_{\epsilon_N}} & \cdots & \frac{1}{\sigma^2_{\epsilon_N}} \end{bmatrix}. \tag{32}$$

Moreover, the optimal weights of a vanilla mean-variance portfolio are calculated by $\frac{1}{\lambda}\mathbf{\Sigma}^{-1}\boldsymbol{\mu}$, so the holdings of the first asset hinge on $\frac{1}{\lambda\sigma^2_{\epsilon_1}}\left(\mu_1 - \sum_{n=2}^{N}\beta_{1n}\mu_n\right)$. Hence, the i^{th} row of the inverse covariance matrix implies the hedging relationship of the i^{th} asset with the remaining assets, where the other $N-1$ assets hedge the i^{th} asset in the portfolio.

Additionally, Goto and Xu (2015) demonstrate that constraining the number of assets for hedging can make it more reliable and stable, thereby offering significant out-of-sample portfolio performance. This finding is called "sparse hedging", represented by a sparse inverse covariance matrix, in which some off-diagonal elements are zero. Moreover, each off-diagonal element of the inverse covariance matrix prescribes the partial correlation of a pair of assets, so the underlying structure of the inverse covariance matrix is sparse when some assets intend to be conditionally independent. Therefore, it is desirable to impose sparsity restrictions on the inverse covariance matrix to improve its estimation.

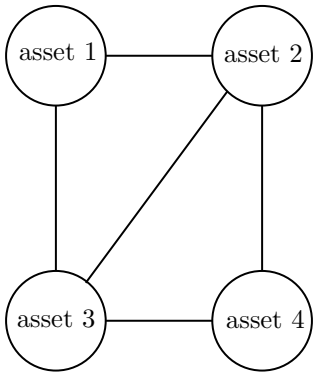
Furthermore, the sparse inverse covariance matrix estimation is akin to Gaussian graphical modeling. Assuming returns of N risky assets $\boldsymbol{R}_t$ are independent and identically distributed (i.i.d) and follow a multivariate normal distribution $\boldsymbol{R}_t \sim \text{NID}(\boldsymbol{\mu}, \mathbf{\Sigma})$ where $t = 1, \ldots, T$, the sparse structure of $\mathbf{\Sigma}^{-1}$ can be illustrated by an undirected graph with

nodes and edges, in which the nodes are the N risky assets, and the connecting edges between matching pairs of assets represent partial correlation that are non-zero off-diagonal elements of the inverse covariance matrix. The absence of a link between two nodes implies their conditional independence given the remaining assets. The objective of graphical modeling is to identify the edges representing the non-zero off-diagonal elements of the inverse covariance matrix, which can be defined as follows:

$$E = \{(i,j) : \text{ asset } i \text{ and } j \text{ are dependent given remaining assets}, 1 \leq i,j \leq N\}. \tag{33}$$

For illustration, a four-asset inverse covariance matrix is presented in Figure 7, where four graph nodes are composed of all assets, and five edges between nodes represent non-zero off-diagonal elements of the inverse covariance matrix. There is no edge between assets one and four because they are conditionally independent.

Figure 7: Undirected Graph of Four Assets



Further, to achieve Gaussian graphical modeling, a fascinating endeavour is to employ regularization on the maximum likelihood estimation of the inverse covariance matrix, indicated in the following proposition.

Proposition 4: Denoting the inverse covariance matrix by $\boldsymbol{\Theta}$ (also called the precision matrix) and the sample covariance matrix by $\boldsymbol{S}$, the log-likelihood function L is determined by

$$L \propto \{\log|\boldsymbol{\Theta}| - \text{trace}\,(\boldsymbol{S}\boldsymbol{\Theta})\}, \tag{34}$$

where $\log|\boldsymbol{\Theta}|$ is the log value of the matrix determinant, and $\text{trace}(\boldsymbol{S\Theta})$ indicates the matrix trace.

Among regularization techniques, the L1, or LASSO penalty, is most widely used in the portfolio management literature. For instance, Goto and Xu (2015) introduce the L1-penalty to generate a sparse estimation of the precision matrix for the mean-variance portfolio optimization, where the specific algorithm, graphical LASSO, is applied. The mathematical formulation is given by

$$\hat{\boldsymbol{\Theta}}_G = \arg\max_{\boldsymbol{\Theta}} \{\log|\boldsymbol{\Theta}| - \text{trace}(\boldsymbol{S\Theta}) - \alpha\|\boldsymbol{\Theta}\|_1\}, \tag{35}$$

where $\|\boldsymbol{\Theta}\|_1$ is L1 norm (the summation of the absolute values of the elements of $\boldsymbol{\Theta}$), and α represents a non-negative regularization parameter.

Generally, graphical LASSO leverages LASSO's capability to achieve sparse graphical modeling of the inverse covariance matrix by taking advantage of the L1 penalty. However, the conventional LASSO approach has two main limitations in practice (Zou and Hastie, 2005). First, it arbitrarily picks a single variable from a group of high-correlated variables instead of selecting them all. In addition, the number of variables selected by LASSO is bounded by the number of observations. Consequently, the drawbacks inherent in LASSO are transformed into the graphical LASSO algorithm when conducting Gaussian graphical modeling of the inverse covariance matrix. The restrictions mentioned above regarding LASSO in the context of inverse covariance matrix estimation of asset returns equate to restricting the number of edges between the respective pairs of assets in the related graph and selecting only a few from grouped assets. Moreover, as the variables chosen by the conventional LASSO technique are not stable over time, the structure of the Gaussian graphical model generated by the graphical LASSO algorithm is unstable, harming the model interpretation. Therefore, it is necessary to improve the graphical LASSO algorithm further to refine the inverse covariance matrix estimation, which is reported in the next section.

3.4.2 A Graphical Adaptive Elastic Net Algorithm for the Inverse Covariance Matrix Estimation

Using LASSO variants to remedy issues of the graphical LASSO algorithm is natural and applicable, thereby forming a better strategy for graphical modeling. However, despite improvements over LASSO, they have so far received limited attention from the portfolio

management literature.

Besides, among LASSO variants, adaptive Elastic Net is an attractive alternative for variable selection and parameter estimation with the oracle property (Zou and Zhang, 2009). Herein we propose the graphical adaptive Elastic Net algorithm for the sparse inverse covariance matrix estimation. To the best of our knowledge, we are the first to involve the adaptive Elastic Net in Gaussian graphical modeling. Moreover, on top of solving the problems of the graphical LASSO, desirable properties of the graphical adaptive elastic net algorithm can also bring other benefits to the process of graphical modeling. For example, this method enables different shrinkage to elements of the inverse covariance matrix, which is more appropriate for the different partial correlations between assets.

Essentially, adaptive Elastic Net is a blend of the adaptive LASSO and Elastic Net. By integrating it into the inverse covariance matrix estimation, the proposed graphical adaptive Elastic Net (GA-ENet) approach is given as follows:

$$\hat{\boldsymbol{\Theta}}_G = \arg\max_{\boldsymbol{\Theta}} \left\{ \log|\boldsymbol{\Theta}| - \text{trace}(\boldsymbol{S}\boldsymbol{\Theta}) - \phi_1 \sum_{i=1}^{N} \sum_{j=1;j\neq i}^{N} \omega_{ij} |\theta_{ij}| - \phi_2 \sum_{i=1}^{N} \sum_{j=1;j\neq i}^{N} \theta_{ij}^2 \right\}, \tag{36}$$

where ϕ_1 and ϕ_2 denotes the non-negative L1 penalizing parameter (LASSO) and L2 penalizing parameter (ridge), respectively. ω_{ij} represents the non-negative weighting factor (adaptive lasso) for θ_{ij}. Following the study of Zou and Zhang (2009), ω_{ij} is determined by $\left|-\beta_{i,j}/\sigma_{\epsilon_i}^2\right|^{-\psi}$ with $\psi > 0$, where $-\beta_{i,j}/\sigma_{\epsilon_i}^2$ is obtained from (32) by the corresponding ordinary least squares (OLS) hedging regression.

From (36), it is clear that our GA-ENet approach is equivalent to the graphical Elastic Net method when all elements in the inverse covariance matrix are imposed the same penalty, and is equal to the classical graphical LASSO when we further set ϕ_2 to 0. Further, the block coordinate descent scheme implemented in graphical LASSO suffices to solve (36).

Following the study of Friedman, Hastie and Tibshirani (2008), letting $\boldsymbol{W}$ be the estimate of $\boldsymbol{\Sigma}$, we decompose $\boldsymbol{W}$, $\boldsymbol{\Theta}$, $\boldsymbol{S}$ as follows:

$$\boldsymbol{W} = \begin{pmatrix} \boldsymbol{W}_{11} & \boldsymbol{w}_{12} \\ \boldsymbol{w}_{12}' & w_{22} \end{pmatrix}, \quad \boldsymbol{\Theta} = \begin{pmatrix} \boldsymbol{\Theta}_{11} & \boldsymbol{\theta}_{12} \\ \boldsymbol{\theta}_{12}' & \theta_{22} \end{pmatrix}, \quad \boldsymbol{S} = \begin{pmatrix} \boldsymbol{S}_{11} & \boldsymbol{s}_{12} \\ \boldsymbol{s}_{12}' & s_{22} \end{pmatrix}. \tag{37}$$

Moreover, the sub-gradient condition of (36) is:

$$\boldsymbol{\Theta}^{-1} - \boldsymbol{S} - \phi_1 \boldsymbol{\omega}\boldsymbol{\Gamma} - 2\phi_2\boldsymbol{\Theta} = \boldsymbol{0}, \tag{38}$$

where $\boldsymbol{\Gamma}$ is a $N \times N$ matrix with elements $\gamma_{ij} = \text{sign}(\theta_{ij})$ if $\theta_{ij} \neq 0$, else $\gamma_{ij} \in [-1, 1]$ if $\theta_{ij} = 0$. $\boldsymbol{\omega}$'s elements are ω_{ij}.

Then the upper right block of (38) is:

$$\boldsymbol{w}_{12} - \boldsymbol{s}_{12} - \phi_1 \boldsymbol{\omega}_{12} \boldsymbol{\Gamma}_{12} - 2\phi_2 \boldsymbol{\theta}_{12} = \mathbf{0}. \tag{39}$$

Since $\boldsymbol{W}\boldsymbol{\Theta} = \boldsymbol{I}$, we can get the following expression:

$$\begin{pmatrix} \boldsymbol{W}_{11} & \boldsymbol{w}_{12} \\ \boldsymbol{w}'_{12} & w_{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Theta}_{11} & \boldsymbol{\theta}_{12} \\ \boldsymbol{\theta}'_{12} & \theta_{22} \end{pmatrix} = \begin{pmatrix} \boldsymbol{I} & \mathbf{0} \\ \mathbf{0}' & 1 \end{pmatrix}. \tag{40}$$

Then we have:

$$\begin{cases} \boldsymbol{W}_{11}\boldsymbol{\Theta}_{11} + \boldsymbol{w}_{12}\boldsymbol{\theta}'_{12} = \boldsymbol{I}, \\ \boldsymbol{W}_{11}\boldsymbol{\theta}_{12} + \boldsymbol{w}_{12}\theta_{22} = 0, \\ \boldsymbol{w}'_{12}\boldsymbol{\Theta}_{11} + w_{22}\boldsymbol{\theta}'_{12} = 0', \\ \boldsymbol{w}'_{12}\boldsymbol{\theta}_{12} + w_{22}\theta_{22} = 1. \end{cases} \tag{41}$$

From (41), $\boldsymbol{w}_{12} = -\boldsymbol{W}_{11}\frac{\boldsymbol{\theta}_{12}}{\theta_{22}}$. Let $\boldsymbol{b} = -\frac{\boldsymbol{\theta}_{12}}{\theta_{22}}$. Substituting $\boldsymbol{w}_{12}$ and $\boldsymbol{b}$ into (39), we have:

$$\boldsymbol{W}_{11}\boldsymbol{b} - \boldsymbol{s}_{12} - \phi_1 \boldsymbol{\omega}_{12} \boldsymbol{\Gamma}_{12} + 2\phi_2 \boldsymbol{b}\theta_{22} = \mathbf{0}. \tag{42}$$

(42) is the same as the form of the sub-gradient of the penalized loss function for a linear regression model, which means here we can apply the method that addresses the penalized linear regression. Thus, the coordinate descent strategy is employed to obtain the optimal $\boldsymbol{b}$. Letting $\boldsymbol{V} = \boldsymbol{W}_{11}$ and $\boldsymbol{u} = \boldsymbol{s}_{12}$, the updating form is given as follows:

$$b_j = \begin{cases} \frac{u_j - \sum_{k \neq j}^{N-1} V_{jk} b_k - \phi_1 \omega_j}{V_{jj} + 2\phi_2 \theta_{22}}, & u_j - \sum_{k \neq j}^{N-1} V_{jk}\beta_k > \phi_1 \omega_j, \\ 0, & -\phi_1 \omega_j \leq u_j - \sum_{k \neq j}^{N-1} V_{jk}\beta_k \leq \phi_1 \omega_j, \\ \frac{u_j - \sum_{k \neq j}^{N-1} V_{jk} b_k + \phi_1 \omega_j}{V_{jj} + 2\phi_2 \theta_{22}}, & u_j - \sum_{k \neq j}^{N-1} V_{jk}\beta_k < -\phi_1 \omega_j. \end{cases} \tag{43}$$

Consequently, after each block-update[22], we can get the optimal $\boldsymbol{b}$, thereby updating

[22]The block-update represents the elements update of a partitioned matrix also referred as a block matrix.

$\boldsymbol{w}_{12} = \boldsymbol{W}_{11}\boldsymbol{b}$, $\theta_{22} = \frac{1}{w_{22}-\boldsymbol{w}_{12}'\boldsymbol{b}}$, $\boldsymbol{\theta}_{12} = -\boldsymbol{b}\theta_{22}$. Further, after N times of block-update, the wholly updated $\boldsymbol{W}$ and $\boldsymbol{\Theta}$ are generated. By checking convergence (the absolute value of the largest difference between elements of two subsequent updates of $\boldsymbol{W}$ is smaller than a predefined threshold value), we can confirm whether the block coordinate descent strategy proceeds. Here is a description of the graphical adaptive Elastic Net (GA-ENet) algorithm.

Algorithm: graphical adaptive Elastic Net

Input: Sample covariance matrix $\boldsymbol{S}$ calculated by historical samples of asset returns, threshold h, penalizing parameters λ_1 and λ_2, weighting factor matrix $\boldsymbol{\omega}$ obtained by row-by-row hedging regressions

Output: precision matrix Θ

1 Initial value: $\boldsymbol{W} = \boldsymbol{S}$, $\boldsymbol{\Theta} = \boldsymbol{S}^{-1}$

2 **for** $j = 1, 2, \ldots, N, 1, 2, \ldots N, \ldots$ **do**

3 Iteratively update $\boldsymbol{W}$ and $\boldsymbol{\Theta}$ with the block coordinate descent algorithm until convergence.

4 **return** the final $\boldsymbol{\Theta}$

4 Empirical Studies

In this section, we empirically compare our generalized Bayes-Stein framework's out-of-sample performance to that of the 1/N asset allocation rule on various datasets.

4.1 Data and Models

Following DeMiguel, Garlappi and Uppal (2009), among others, we employ similar datasets containing monthly value-weighted excess returns (over the 1-month T-bill return). They are listed in Table 4 and described in Appendix B of the Online Supplementary Appendix.

Table 4: Data Description

Data	N	Time	Abbreviation
10 industry portfolios	10	1963.06-2021.12	Ind10
20 size and book-to-market portfolios and the US equity MKT	21	1963.06-2021.12	FF21
20 size and book-to-market portfolios and the MKT, SMB, and HML portfolios	23	1963.06-2021.12	FF23
20 size and book-to-market portfolios and the MKT, SMB, HML, and UMD portfolios	24	1963.06-2021.12	FF24
25 size and book-to-market portfolios	25	1963.06-2021.12	FF25
100 size and book-to-market portfolios	100	1963.06-2021.12	FF100BM
100 size and operating profitability portfolios	100	1963.07-2021.12	FF100OP
100 size and investment portfolios	100	1963.07-2021.12	FF100INV

Upon these datasets, we construct monthly rebalanced portfolios for a mean-variance investor who allocates capital between risky assets and the risk-free rate in the pursuit of expected utility maximization.[23] In particular, we impose non-short-sale constraints and variance-based constraints[24] (VBCs) of Levy and Levy (2014) on the asset weights to lessen further the adverse effects of estimation errors from the input parameters. These constraints are commonly utilized by previous studies (see, e.g., Board and Sutcliffe, 1994; Platanakis, Sutcliffe and Ye, 2021).

Additionally, the monthly portfolio rebalancing process leads to changes in asset weights during the out-of-sample period and incurs transaction costs. As a result, we evaluate the portfolio performance taking into account these expenses. Following DeMiguel, Martín-Utrera and Nogales (2015), we add a penalty term for transaction costs and set the overall objective function integrating our generalized Bayes-Stein model as follows:

$$\max_{\boldsymbol{w}_{GBS}} \left\{ \boldsymbol{w}'_{GBS}\hat{\boldsymbol{\mu}}_{GBS} - \frac{\lambda}{2}\boldsymbol{w}'_{GBS}\hat{\boldsymbol{\Sigma}}_{GBS}\boldsymbol{w}_{GBS} - \frac{\delta}{2}\Delta\boldsymbol{w}'_{GBS}\hat{\boldsymbol{\Sigma}}_{S}\Delta\boldsymbol{w}_{GBS} \right\}, \tag{44}$$

[23] Denoting N as the number of risky assets and w_i as the weight of risky asset i, $1 - \sum_{i=1}^{N} w_i$ is the portfolio weight allocated to the risk-free asset (1-month T-bill).

[24] $\left|w_i - \frac{1}{N}\right| \frac{\sigma_i}{\bar{\sigma}} \leq \eta, \forall i$, where σ_i denotes the standard deviation of risky asset i and $\bar{\sigma}$ is the average standard deviation of all risky assets. Following Platanakis, Sutcliffe and Ye (2021), we set the VBCs parameter η to 0.15.

where $\boldsymbol{w}_{GBS}$ indicates the portfolio weights vector, and $\hat{\boldsymbol{\mu}}_{GBS}$ and $\hat{\boldsymbol{\Sigma}}_{GBS}$ represent the mean returns vector and the covariance matrix, produced by our generalized Bayes-Stein model, respectively. λ denotes the risk aversion coefficient, and we use $\lambda = 1$, $\lambda = 3$, and $\lambda = 5$ to represent aggressive, moderate, and conservative investors, respectively.[25] Besides, the term $\frac{\delta}{2}\Delta\boldsymbol{w}'_{GBS}\hat{\boldsymbol{\Sigma}}_S\Delta\boldsymbol{w}_{GBS}$ represents a quadratic transaction cost, where δ is the transaction cost parameter[26] and $\Delta\boldsymbol{w}_{GBS}$ is the asset weights change vector. For example, the weights change vector at time t is calculated by

$$\Delta\boldsymbol{w}_{GBS,t} = \boldsymbol{w}_{GBS,t} - \boldsymbol{w}^{+}_{GBS,t-1}, \tag{45}$$

where $\boldsymbol{w}^{+}_{GBS,t-1}$ denotes the asset weights vector at the end of $t-1$. So, the transaction cost taking out of the portfolio return at time t is measured by $\delta\Delta\boldsymbol{w}'_{GBS,t}\hat{\boldsymbol{\Sigma}}_S\Delta\boldsymbol{w}_{GBS,t}$.

Moreover, to facilitate the implementation of the proposed methods in our generalized Bayes-Stein framework, we employ an expanding estimation window[27], where models are built on monthly data available up to month t for computing the input parameters and corresponding optimal portfolios at month $t+1$. Note that the estimation of mean returns is slightly different from that of the covariance matrix. This is because we need one more month of asset returns as the predictors in our TW-ENet approach's design. For illustration, we use the data samples from June 1963 to June 1983 to estimate portfolio inputs and determine optimal asset weights in the first out-of-sample time point, July 1983. In this case, we need data from June 1963 to June 1983 to estimate the expected asset returns via the TW-ENet approach, while from July 1963 to June 1983 to estimate the covariance matrix, or its inverse. This estimation and the monthly portfolio rebalancing process are repeated until the end of the entire data sample. Ultimately, we can obtain monthly out-of-sample portfolio excess returns of our generalized Bayes-Stein framework.

In a similar vein, we also rebalance our benchmark, the "naive" equally-weighted (1/N) asset allocation rule that does not need any parameter estimation and optimization, and equally distributes capital among different underlying assets.

[25] Results of other risk aversion parameters are presented in Appendix C of the Online Supplementary Appendix.

[26] No transaction costs exist when $\delta = 0$. Following DeMiguel, Martín-Utrera and Nogales (2015), we set $\delta = 3 \times 10^{-7}$ in the main paper, while other values are set for robustness checks.

[27] We use 20-year and 40-year expanding estimation windows in this paper.

4.2 Performance Measures

We apply two different measures to analyze the out-of-sample performance of various portfolio strategies. First, we compare these models' out-of-sample performance using the annualized out-of-sample Sharpe ratio (SR), calculated by

$$SR = \mu_p/\sigma_p, \tag{46}$$

where μ_p and σ_p denote the annualized out-of-sample expected portfolio excess returns and the annualized out-of-sample standard deviation of portfolio excess returns, respectively.

Second, we compute the annualized certainty equivalent return (CER) as follows:

$$CER = \mu_p - \frac{\lambda}{2}{\sigma_p}^2, \tag{47}$$

where λ represents the investor's risk aversion.

Note that we employ the same approach as DeMiguel, Garlappi and Uppal (2009) to test the statistical difference of Sharpe ratios and certainty equivalent returns between our generalized Bayes-Stein framework and the 1/N rule.

4.3 Empirical Results

To evaluate the out-of-sample portfolio performance of our generalized Bayes-Stein framework and its benchmark (1/N rule), we first exhibit their Sharpe ratios and certainty equivalent returns without accounting for transaction costs in Table 5, and then report the results considering transaction costs in Table 6. Moreover, to analyze the influence of the initial expanding window length or the out-of-sample periods on the asset allocation models, we use different initial expanding windows that lead to different out-of-sample periods when presenting results.

For the case of no transaction costs, our generalized Bayes-Stein model outperforms the 1/N scheme in terms of Sharpe ratios and certainty equivalent returns in most datasets except for Ind10 where their performances are comparable. More importantly, the performance metrics differences between the two asset allocation models are substantial, and in general, our generalized Bayes-Stein model surpasses the 1/N rule by more than 30%. For example, in the dataset FF100INV under a 20-year expanding estimation window, our generalized Bayes-Stein model attains the highest Sharpe ratio of about 0.74, which is about 35% higher

than the 1/N rule with a Sharpe ratio of about 0.55. Notably, once the number of risky assets increases to 100, our generalized Bayes-Stein model performs exceptionally well and is much better than 1/N. In these large portfolios (FF100BM, FF100OP, and FF100INV), the performance of the 1/N rule may be polluted by the high number of assets, which implies the necessity and effectiveness of our improved shrinkage factor calibration approach based on clustering analysis. Moreover, the edge of our generalized Bayes-Stein framework over 1/N is consistent in both the 20-year (the out-of-sample period covers from July 1983 to December 2021) and 40-year (the out-of-sample period covers from July 2003 to December 2021) expanding estimation windows and robust to the values of risk aversion. Furthermore, by examining the performance metrics' statistical difference between our generalized Bayes-Stein framework and 1/N, we find that our generalized Bayes-Stein framework is significantly superior (at least the 10% significance level) regardless of the performance metrics in most cases.

Moreover, we further assess the out-of-sample performance of our generalized Bayes-Stein framework after taking into account transaction costs when measuring performance. Likewise, in comparison with the 1/N strategy, our generalized Bayes-Stein framework can significantly surpass it in most cases. Intuitively, transaction costs imposed on portfolio returns will hurt performance. However, since we involve a transaction cost penalty term in the objective function, the obtained optimal portfolio weights change accordingly, even leading to results contrary to our intuition in some cases (i.e., better performance post-transaction costs compared to pre-transaction costs), which is in line with the results shown in Table 6.

Furthermore, our generalized Bayes-Stein model is a holistic framework that integrates various well-designed machine learning techniques to enhance multiple aspects of the conventional Bayes-Stein model. To verify the usefulness of different parts, we decompose our generalized Bayes-Stein model into two sub-models, one that only includes the TW-ENet approach (model 1) and another that does not have the TW-ENet method (model 2). Model 1 means we only modify the sample mean component ($\hat{\boldsymbol{\mu}}_S$) of the classical Bayes-Stein model from a return forecasting perspective, whereas model 2 only involves our techniques for the shrinkage factor calibration and the inverse covariance matrix estimation. Table 7 reports the corresponding results.[28] We find that, largely, both model 1 and model 2 can still significantly outperform the 1/N rule in some cases. Hence, the machine learning methods we develop each effectively boosts the original Bayes-Stein model. Most importantly, when

[28] Additional results are reported in Appendix C of the Online Supplementary Appendix.

combining all methods (establishing the generalized Bayes-Stein model with models 1 and 2), we can attain a comprehensive model that provides more stable and superior performance in all datasets.[29]

Therefore, though it is conceived hard to defeat the 1/N strategy (Ming and Zhou, 2022), our empirical studies offer strong evidence supporting the conclusion that our generalized Bayes-Stein portfolio optimization framework can improve the original Bayes-Stein model and defeat the 1/N rule. To examine the robustness of our findings, we conduct a further extension of this study by robustness checks in Section 5.

[29]Though the metric value of model 1 is slightly higher in some cases such as FF100BM and FF100INV, the generalized Bayes-Stein model is more stable in most datasets.

Table 5: SRs and CERs without TCs for Different Estimation Windows

Table 5 reports the out-of-sample yearly Sharpe ratios (SRs) and certainty equivalent returns (CERs) of the generalized Bayes-Stein framework (GBS) and the naive equal weighted scheme (1/N) under various datasets (Ind10, FF21, FF23, FF24, FF25, FF100BM, FF100OP, and FF100INV) for different risk aversion parameters ($\lambda = 1, 3, 5$) without considering transaction costs (TCs). The results of the 20-year (the out-of-sample period covers from July 1983 to December 2021) and 40-year (the out-of-sample period covers from July 2003 to December 2021) expanding estimation windows are exhibited in Panel A and Panel B, respectively. The results of significance tests of the performance differences (GBS vs 1/N) are put in the parenthesis, where *, **, and *** indicate a statistical significance at the 10%, 5%, and 1%-level, respectively.

Panel A: SRs and CERs without TCs for a 20-year expanding estimation window

Dataset	SRs						CERs					
	$\lambda = 1$		$\lambda = 3$		$\lambda = 5$		$\lambda = 1$		$\lambda = 3$		$\lambda = 5$	
	GBS	1/N	GBS	1/N	GBS	1/N	GBS	1/N	GBS	1/N	GBS	1/N
Ind10	0.6719	0.6356	0.6789	0.6356	0.6753	0.6356	0.0883 (*)	0.0756	0.0667	0.0580	0.0455	0.0403
FF21	0.6205 (*)	0.5226	0.6224 (*)	0.5226	0.6270 (*)	0.5226	0.0902 (*)	0.0760	0.0620 (*)	0.0455	0.0354 (**)	0.0150
FF23	0.6467 (**)	0.5050	0.6448 (**)	0.5050	0.6392 (**)	0.5050	0.0904 (**)	0.0688	0.0645 (**)	0.0425	0.0387 (**)	0.0161
FF24	0.6510 (**)	0.5155	0.6603 (**)	0.5155	0.6727 (**)	0.5155	0.0886 (**)	0.0676	0.0658 (**)	0.0438	0.0446 (**)	0.0200
FF25	0.7122 (***)	0.5497	0.7112 (***)	0.5497	0.7121 (***)	0.5497	0.1066 (***)	0.0783	0.0773 (***)	0.0500	0.0492 (***)	0.0217
FF100BM	0.7317 (***)	0.3816	0.7229 (***)	0.3816	0.7178 (***)	0.3816	0.1185 (***)	0.0518	0.0824 (***)	0.0206	0.0473 (***)	-0.0105
FF100OP	0.6797	0.5635	0.6772 (***)	0.5635	0.6722 (***)	0.5635	0.1104 (***)	0.0829	0.0743 (***)	0.0526	0.0383 (**)	0.0223
FF100INV	0.7401 (***)	0.5470	0.7378 (***)	0.5470	0.7359 (***)	0.5470	0.1172 (***)	0.0796	0.0842 (***)	0.0497	0.0515 (***)	0.0198

Panel B: SRs and CERs without TCs for a 40-year expanding estimation window

Dataset	SRs						CERs					
	$\lambda = 1$		$\lambda = 3$		$\lambda = 5$		$\lambda = 1$		$\lambda = 3$		$\lambda = 5$	
	GBS	1/N	GBS	1/N	GBS	1/N	GBS	1/N	GBS	1/N	GBS	1/N
Ind10	0.7866	0.7518	0.7892	0.7518	0.7911	0.7518	0.1104 (**)	0.0915	0.0863	0.0736	0.0626	0.0558
FF21	0.6882	0.6238	0.6985	0.6238	0.7167	0.6238	0.1025	0.0965	0.0750	0.0638	0.0501	0.0311
FF23	0.7300	0.6049	0.7417	0.6049	0.7441	0.6049	0.1084	0.0882	0.0818	0.0594	0.0547	0.0307
FF24	0.6862	0.6124	0.6848	0.6124	0.7006	0.6124	0.0994	0.0854	0.0717	0.0597	0.0480	0.0339
FF25	0.8089 (***)	0.6322	0.8176 (***)	0.6322	0.8274 (***)	0.6322	0.127 (***)	0.0955	0.097 (***)	0.0647	0.0683 (***)	0.0340
FF100BM	0.8779 (***)	0.5169	0.8593 (***)	0.5169	0.8575 (***)	0.5169	0.1529 (***)	0.0783	0.1108 (***)	0.0443	0.0721 (***)	0.0104
FF100OP	0.7550 (**)	0.6552	0.7562 (**)	0.6552	0.7539 (**)	0.6552	0.1226 (**)	0.1016	0.0885 (**)	0.0693	0.0540 (**)	0.0370
FF100INV	0.8022 (**)	0.6508	0.7977 (**)	0.6508	0.7949 (**)	0.6508	0.1265 (**)	0.1001	0.0942 (**)	0.0683	0.0624 (**)	0.0366

Table 6: SRs and CERs with TCs for Different Estimation Windows

Table 6 reports the out-of-sample yearly Sharpe ratios (SRs) and certainty equivalent returns (CERs) of the generalized Bayes-Stein framework (GBS) and the naive equal weighted scheme (1/N) under various datasets (Ind10, FF21, FF23, FF24, FF25, FF100BM, FF100OP, and FF100INV) for different risk aversion parameters ($\lambda = 1, 3, 5$) considering transaction costs (TCs). The results of the 20-year (the out-of-sample period covers from July 1983 to December 2021) and 40-year (the out-of-sample period covers from July 2003 to December 2021) expanding estimation windows are exhibited in Panel A and Panel B, respectively. The results of significance tests of the performance differences (GBS vs 1/N) are put in the parenthesis, where *, **, and *** indicate a statistical significance at the 10%, 5%, and 1%-level, respectively.

Panel A: SRs and CERs with TCs for a 20-year expanding estimation window

Dataset	SRs						CERs					
	$\lambda = 1$		$\lambda = 3$		$\lambda = 5$		$\lambda = 1$		$\lambda = 3$		$\lambda = 5$	
	GBS	1/N	GBS	1/N	GBS	1/N	GBS	1/N	GBS	1/N	GBS	1/N
Ind10	0.6664	0.6356	0.6604	0.6356	0.6548	0.6356	0.0892 (*)	0.0756	0.0648	0.058	0.0424	0.0403
FF21	0.6453 (**)	0.5226	0.6544 (**)	0.5226	0.6525 (**)	0.5226	0.0956 (**)	0.076	0.0677 (**)	0.0455	0.0392 (**)	0.015
FF23	0.6529 (**)	0.5050	0.6490 (**)	0.5050	0.6421 (**)	0.5050	0.0928 (**)	0.0688	0.0656 (**)	0.0425	0.0388 (**)	0.0161
FF24	0.6375 (*)	0.5155	0.6504 (**)	0.5155	0.6626 (**)	0.5155	0.0875 (**)	0.0676	0.0646 (**)	0.0438	0.0430 (**)	0.0200
FF25	0.7067 (***)	0.5497	0.7027 (***)	0.5497	0.7049 (***)	0.5497	0.1078 (***)	0.0783	0.0765 (***)	0.0500	0.0475 (***)	0.0217
FF100BM	0.7096 (***)	0.3816	0.7057 (***)	0.3816	0.7065 (***)	0.3816	0.1152 (***)	0.0518	0.0793 (***)	0.0206	0.0453 (***)	-0.0105
FF100OP	0.6614 (**)	0.5635	0.6644 (**)	0.5635	0.6612 (**)	0.5635	0.1075 (***)	0.0829	0.0719 (**)	0.0526	0.0362 (**)	0.0224
FF100INV	0.7242 (***)	0.5470	0.7240 (***)	0.5470	0.7259 (***)	0.5470	0.1138 (***)	0.0796	0.0816 (***)	0.0497	0.0498 (***)	0.0198

Panel B: SRs and CERs with TCs for a 40-year expanding estimation window

Dataset	SRs						CERs					
	$\lambda = 1$		$\lambda = 3$		$\lambda = 5$		$\lambda = 1$		$\lambda = 3$		$\lambda = 5$	
	GBS	1/N	GBS	1/N	GBS	1/N	GBS	1/N	GBS	1/N	GBS	1/N
Ind10	0.7613	0.7518	0.7531	0.7518	0.7524	0.7518	0.1091 (*)	0.0915	0.0819	0.0736	0.0565	0.0558
FF21	0.7102	0.6238	0.7274	0.6238	0.7413	0.6238	0.1088	0.0965	0.0808	0.0638	0.0537	0.0311
FF23	0.7476 (*)	0.6049	0.7580 (*)	0.6049	0.7612 (*)	0.6049	0.1138 (*)	0.0882	0.0856 (*)	0.0594	0.0573 (*)	0.0307
FF24	0.6834	0.6124	0.6855	0.6124	0.7025	0.6124	0.1007	0.0854	0.0724	0.0597	0.0481	0.0339
FF25	0.7802 (***)	0.6322	0.7871 (***)	0.6322	0.8004 (***)	0.6322	0.1253 (***)	0.0955	0.0930 (***)	0.0647	0.0633 (***)	0.0340
FF100BM	0.8529 (***)	0.5169	0.8567 (***)	0.5169	0.8569 (***)	0.5169	0.1493 (***)	0.0783	0.1105 (***)	0.0443	0.0720 (***)	0.0104
FF100OP	0.7387 (*)	0.6552	0.7435 (*)	0.6552	0.7470 (*)	0.6552	0.1206 (**)	0.1016	0.0864 (*)	0.0693	0.0526 (*)	0.0370
FF100INV	0.8028 (**)	0.6508	0.8025 (**)	0.6508	0.809 (***)	0.6508	0.1252 (**)	0.1001	0.0945 (**)	0.0683	0.0650 (***)	0.0366

Table 7: SRs and CERs with TCs for Model Decomposition

The generalized Bayes-Stein model (GBS) is decomposed into two sub-models, one that only includes the TW-ENet approach (model 1) and another that does not have the TW-ENet method (model 2). Table 7 reports the out-of-sample yearly Sharpe ratios (SRs) and certainty equivalent returns (CERs) of the two sub-models and the naive equal weighted scheme (1/N) under various datasets (Ind10, FF21, FF23, FF24, FF25, FF100BM, FF100OP, and FF100INV) for different risk aversion parameters ($\lambda = 1, 3, 5$) considering transaction costs (TCs), which are exhibited in Panel A and Panel B, respectively. The results are obtained under a 20-year expanding estimation window (the out-of-sample period covers from July 1983 to December 2021). Significance tests of the performance differences (GBS-model 1 vs 1/N, GBS-model 2 vs 1/N) are put in the parenthesis, where *, **, and *** indicate a statistical significance at the 10%, 5%, and 1%-level, respectively.

Panel A: SRs and CERs of GBS-model 1

Dataset	SRs						CERs					
	$\lambda = 1$		$\lambda = 3$		$\lambda = 5$		$\lambda = 1$		$\lambda = 3$		$\lambda = 5$	
	GBS-model 1	1/N	GBS-model 1	1/N	GBS-model 1	1/N	GBS-model 1	1/N	GBS-model 1	1/N	GBS-model 1	1/N
Ind10	0.6693	0.6356	0.6456	0.6356	0.6295	0.6356	0.0890 (*)	0.0756	0.0608	0.0580	0.0396	0.0403
FF21	0.6288 (*)	0.5226	0.6196	0.5226	0.6170	0.5226	0.0899	0.0760	0.0600	0.0455	0.0377 (*)	0.0150
FF23	0.5870	0.5050	0.5642	0.5050	0.5493	0.5050	0.0794	0.0688	0.0501	0.0425	0.0293	0.0161
FF24	0.5997	0.5155	0.6578 (*)	0.5155	0.6346	0.5155	0.0784	0.0676	0.0607	0.0438	0.0402 (*)	0.0200
FF25	0.6789 (***)	0.5497	0.693 (***)	0.5497	0.6952 (**)	0.5497	0.1005 (***)	0.0783	0.0723 (**)	0.0500	0.0483 (**)	0.0217
FF100BM	0.7183 (***)	0.3816	0.7311 (***)	0.3816	0.7299 (***)	0.3816	0.1161 (***)	0.0518	0.0827 (***)	0.0206	0.0526 (***)	-0.0105
FF100OP	0.6593 (**)	0.5635	0.6952 (***)	0.5635	0.6699 (**)	0.5635	0.1061 (***)	0.0829	0.0754 (***)	0.0526	0.0439 (**)	0.0224
FF100INV	0.7418 (***)	0.5470	0.7388 (***)	0.5470	0.7414 (***)	0.5470	0.1156 (***)	0.0796	0.0825 (***)	0.0497	0.0549 (***)	0.0198

Panel B: SRs and CERs of GBS-model 2

Dataset	SRs						CERs					
	$\lambda = 1$		$\lambda = 3$		$\lambda = 5$		$\lambda = 1$		$\lambda = 3$		$\lambda = 5$	
	GBS-model 2	1/N	GBS-model 2	1/N	GBS-model 2	1/N	GBS-model 2	1/N	GBS-model 2	1/N	GBS-model 2	1/N
Ind10	0.6362	0.6356	0.6385	0.6356	0.6474	0.6356	0.0790	0.0756	0.0598	0.058	0.0417	0.0403
FF21	0.5823 (*)	0.5226	0.5867 (**)	0.5226	0.5883 (***)	0.5226	0.0917 (**)	0.0760	0.0571 (**)	0.0455	0.0244 (**)	0.015
FF23	0.5721 (**)	0.5050	0.5804 (***)	0.5050	0.5825 (***)	0.5050	0.0900 (***)	0.0688	0.0560 (**)	0.0425	0.0232	0.0161
FF24	0.5942 (**)	0.5155	0.5937 (***)	0.5155	0.5928 (***)	0.5155	0.0943 (***)	0.0676	0.0585 (**)	0.0438	0.0254	0.0200
FF25	0.5900	0.5497	0.5904	0.5497	0.5903 (*)	0.5497	0.0932 (**)	0.0783	0.0578	0.0500	0.0252	0.0217
FF100BM	0.5768 (***)	0.3816	0.5766 (***)	0.3816	0.5748 (***)	0.3816	0.0882 (***)	0.0518	0.0552 (***)	0.0206	0.0225 (***)	-0.0105
FF100OP	0.5039	0.5635	0.5288	0.5635	0.5401	0.5635	0.0752	0.0829	0.0466	0.0526	0.0167	0.0224
FF100INV	0.5744	0.5470	0.5875 (**)	0.5470	0.5925 (***)	0.5470	0.0867 (**)	0.0796	0.0570 (**)	0.0497	0.0266 (**)	0.0198

5 Robustness Checks

In this section, we test the generalized Bayes-Stein framework's robustness and general applicability regarding time-series return forecasting techniques and the portfolio construction design, as shown in Table 8. The detailed robustness check results are presented in Appendix D of the Online Supplementary Appendix.

Table 8: Robustness Checks Description

Number	Description
1	Check alternative time-series return forecasting techniques integrated into the generalized Bayes-Stein framework
2	Check different transaction cost parameters
3	Check a 30-year expanding estimation window
4	Check a 20-year rolling estimation window

First, in the generalized Bayes-Stein framework, we enhance the sample and grand mean by exploiting the predictability of asset returns with the help of the TW-ENet approach. As various machine learning methods have been introduced to forecast expected returns, it is necessary to validate the effectiveness or superiority of our TW-ENet in the generalized Bayes-Stein framework. For simplicity, we use the ordinary least squares post LASSO approach (OLS-post LASSO), the combination Elastic Net method (C-ENet), and the Random Forest method as benchmarks. C-ENet combines individual forecasts selected by Elastic Net, and OLS-post Lasso conducts OLS estimation on predictors selected by the LASSO. They are commonly used for return prediction of the aggregate market or market components in existing literature (see, e.g., Rapach et al., 2019; Dong et al., 2022; Hou et al., 2022). Notably, the robustness check results verify the TE-ENet approach's comparable performance to other well-established methods and reflect the flexibility of our generalized Bayes-Stein framework. This framework can easily integrate other advanced machine learning methods.

Moreover, we conduct additional studies concerning specific designs of our portfolio construction. First, we further apply alternative transaction cost estimates to robustness of our empirical findings. Second, we employ a 30-year expanding window. Finally, as mentioned in Section 4, we provide the empirical results based on the expanding window estimation in the main paper. To analyze the impact of window estimation methods, we use a 20-year rolling estimation window for robustness checks. Importantly, according to the results in Appendix D of the Online Supplementary Appendix, our generalized Bayes-Stein framework can still beat the classical Bayes-Stein model and the 1/N rule under these

different portfolio construction designs, which, to some extent, shows the stability of our generalized Bayes-Stein framework.

6 Conclusions

This paper sets out to improve the classical Bayes-Stein portfolio optimization model that performs poorly out of sample with well-designed machine learning techniques. To this end, we comprehensively analyze the drawbacks inherent in the original Bayes-Stein model with theoretical and empirical evidence and suggest corresponding machine learning methods concerning these key stylized facts. Our established holistic generalized Bayes-Stein framework, integrating a wide variety of novel machine learning approaches we develop, enjoys several desirable properties. First, model components of the original Bayes-Stein model (the sample means vector and the grand mean) are upgraded by superior expected asset returns estimation produced by the time-dependent weighted Elastic Net (TW-ENet) approach. Second, this framework covers the individual differences of portfolio assets by ameliorating the shrinkage factor measurement with a four-stage Bayes-Stein shrinkage approach based on a clustering ensemble method. Third, it extends the traditional Bayes-Stein model with a shrinkage estimator of the inverse covariance matrix yielded by a graphical adaptive Elastic Net (GA-ENet) approach. Most importantly, we find that these properties of the generalized Bayes-Stein framework can be converted into portfolio gains, validated by better out-of-sample performance compared with the 1/N portfolio allocation rule.

In general, therefore, it seems that our generalized Bayes-Stein framework can provide a promising solution for optimal asset allocation, the central concern of institutional investors. Moreover, we seek to refine and revitalize the traditional financial model by leveraging well-tailored explainable machine learning techniques, generating portfolios with superior out-of-sample performance. So, from a practical point of view, our study offers a new angle for portfolio optimization. Instead of directly applying machine learning models off the shelf, we carefully modify them concerning the specifics of the generalized Bayes-Stein model, which may provide new insights for academics or real-world investors involving machine learning in investment decisions.

However, some research limitations or directions are left for future research. First, the asset scope for portfolio construction is limited in this study. We do not exploit and validate our proposed generalized Bayes-Stein framework in a more practical investment environment, e.g., the portfolio selection problem involving concrete stocks or other asset classes. Second,

to our knowledge, this paper is the first to propose or introduce some novel machine learning methods (e.g., the clustering ensemble strategy) in related finance literature. Hence, further exploring these methods in other financial scenarios is meaningful. Third, considering their lack of interpretability, we do not suggest other more advanced machine learning methods in this paper (e.g., deep learning). Thus it is interesting to go back to the question covered in this paper in the future with the evolution of machine learning.

References

Anderson, E. and Cheng, A.r., 2022. Portfolio choices with many big models. *Management science*, 68(1), pp.690–715.

Ao, M., Yingying, L. and Zheng, X., 2019. Approaching mean-variance efficiency for large portfolios. *The review of financial studies*, 32(7), pp.2890–2919.

Avramov, D., Cheng, S. and Metzker, L., 2022. Machine learning vs. economic restrictions: Evidence from stock return predictability. *Management science*.

Ban, G.Y., El Karoui, N. and Lim, A.E., 2018. Machine learning and portfolio optimization. *Management science*, 64(3), pp.1136–1154.

Barroso, P. and Saxena, K., 2021. Lest we forget: using out-of-sample forecast errors in portfolio optimization. *The review of financial studies*, 35(3), pp.1222–1278.

Bertsimas, D. and Cory-Wright, R., 2022. A scalable algorithm for sparse portfolio selection. *Informs journal on computing*.

Bertsimas, D., Gupta, V. and Paschalidis, I.C., 2012. Inverse optimization: A new perspective on the black-litterman model. *Operations research*, 60(6), pp.1389–1403.

Board, J.L.G. and Sutcliffe, C.M.S., 1994. Estimation methods in portfolio selection and the effectiveness of short sales restrictions: Uk evidence. *Management science*, 40(4), pp.516–534.

Chen, L., Pelger, M. and Zhu, J., 2019. Deep learning in asset pricing. *arxiv preprint arxiv:1904.00745*.

Chen, S.D. and Lim, A.E., 2020. A generalized black–litterman model. *Operations research*, 68(2), pp.381–410.

Chopra, V.K. and Ziemba, W.T., 1993. The effect of errors in means, variances, and covariances on optimal portfolio choice. *The journal of portfolio management*, 19(2), pp.6–11.

Cong, L.W., Feng, G., He, J. and Li, J., 2022. Uncommon factors for bayesian asset clusters. *Available at ssrn 4219905*.

Cong, L.W., Tang, K., Wang, J. and Zhang, Y., 2021. Alphaportfolio: Direct construction through deep reinforcement learning and interpretable ai. *Ssrn electronic journal. https://doi.org/10.2139/ssrn*, 3554486.

Coqueret, G. and Guida, T., 2020. *Machine learning for factor investing: R version*. Chapman and Hall/CRC.

Craig MacKinlay, A. and Pástor, L., 2000. Asset pricing models: Implications for expected returns and portfolio selection. *The review of financial studies*, 13(4), pp.883–916.

DeMiguel, V., Garlappi, L. and Uppal, R., 2009. Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *The review of financial studies*, 22(5), pp.1915–1953.

DeMiguel, V., Martin-Utrera, A. and Nogales, F.J., 2013. Size matters: Optimal calibration of shrinkage estimators for portfolio selection. *Journal of banking & finance*, 37(8), pp.3018–3034.

DeMiguel, V., Martín-Utrera, A. and Nogales, F.J., 2015. Parameter uncertainty in multiperiod portfolio optimization with transaction costs. *Journal of financial and quantitative analysis*, 50(6), pp.1443–1471.

Dias, J.G., Vermunt, J.K. and Ramos, S., 2015. Clustering financial time series: New insights from an extended hidden markov model. *European journal of operational research*, 243(3), pp.852–864.

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

El Balghiti, O., Elmachtoub, A.N., Grigas, P. and Tewari, A., 2022. Generalization bounds in the

predict-then-optimize framework. *Mathematics of operations research*.

Elmachtoub, A.N. and Grigas, P., 2022. Smart "predict, then optimize". *Management science*, 68(1), pp.9–26.

Freyberger, J., Neuhierl, A. and Weber, M., 2020. Dissecting characteristics nonparametrically. *The review of financial studies*, 33(5), pp.2326–2377.

Friedman, J., Hastie, T. and Tibshirani, R., 2008. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3), pp.432–441.

Frost, P.A. and Savarino, J.E., 1986. An empirical bayes approach to efficient portfolio selection. *Journal of financial and quantitative analysis*, 21(3), pp.293–305.

Giesecke, K., Kim, B., Kim, J. and Tsoukalas, G., 2014. Optimal credit swap portfolios. *Management science*, 60(9), pp.2291–2307.

Goto, S. and Xu, Y., 2015. Improving mean variance optimization through sparse hedging restrictions. *Journal of financial and quantitative analysis*, 50(6), pp.1415–1441.

Herskovic, B., Kelly, B., Lustig, H. and Van Nieuwerburgh, S., 2016. The common factor in idiosyncratic volatility: Quantitative asset pricing implications. *Journal of financial economics*, 119(2), pp.249–283.

Hou, A.J., Platanakis, E., Ye, X. and Zhou, G., 2022. A model-based commodity risk measure on commodity and stock market returns. *2022 university of rochester conference in econometrics, paris december finance meeting*.

Jorion, P., 1985. International portfolio diversification with estimation risk. *Journal of business*, pp.259–278.

Jorion, P., 1986. Bayes-stein estimation for portfolio analysis. *Journal of financial and quantitative analysis*, 21(3), pp.279–292.

Jorion, P., 1991. Bayesian and capm estimators of the means: Implications for portfolio selection. *Journal of banking & finance*, 15(3), pp.717–727.

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

Kan, R. and Zhou, G.F., 2007. Optimal portfolio choice with parameter uncertainty. *Journal of financial and quantitative analysis*, 42(3), pp.621–656.

Kircher, F. and Rösch, D., 2021. A shrinkage approach for sharpe ratio optimal portfolios with estimation risks. *Journal of banking & finance*, 133, p.106281.

Kong, A., Rapach, D.E., Strauss, J.K. and Zhou, G., 2011. Predicting market components out of sample: asset allocation implications. *The journal of portfolio management*, 37(4), pp.29–41.

Kuhn, D., Parpas, P., Rustem, B. and Fonseca, R., 2009. Dynamic mean-variance portfolio analysis under model risk. *J. comput. finance*, 12(91115), p.7.

Kynigakis, I. and Panopoulou, E., 2022. Does model complexity add value to asset allocation? evidence from machine learning forecasting models. *Journal of applied econometrics*, 37(3), pp.603–639.

Ledoit, O. and Wolf, M., 2003. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of empirical finance*, 10(5), pp.603–621.

Ledoit, O. and Wolf, M., 2017. Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets goldilocks. *The review of financial studies*, 30(12), pp.4349–4388.

Lettau, M. and Pelger, M., 2020. Factors that fit the time series and cross-section of stock returns. *The review of financial studies*, 33(5), pp.2274–2325.

Levy, H. and Levy, M., 2014. The benefits of differential variance-based constraints in portfolio

optimization. *European journal of operational research*, 234(2), pp.372–381.

Lim, A.E., Shanthikumar, J.G. and Vahn, G.Y., 2012. Robust portfolio choice with learning in the framework of regret: Single-period case. *Management science*, 58(9), pp.1732–1746.

Markowitz, H., 1952. Portfolio selection. *The journal of finance*, 7(1), pp.77–91.

Martin, I.W. and Nagel, S., 2022. Market efficiency in the age of big data. *Journal of financial economics*, 145(1), pp.154–177.

Ming, Y. and Zhou, G., 2022. Why naive 1/n diversification is not so naive, and how to beat it? *Available at ssrn*.

Mynbayeva, E., Lamb, J.D. and Zhao, Y., 2022. Why estimation alone causes markowitz portfolio selection to fail and what we might do about it. *European journal of operational research*, 301(2), pp.694–707.

Nguyen, V.A., Kuhn, D. and Mohajerin Esfahani, P., 2022. Distributionally robust inverse covariance estimation: The wasserstein shrinkage estimator. *Operations research*, 70(1), pp.490–515.

Pesaran, M.H., Pick, A. and Pranovich, M., 2013. Optimal forecasts in the presence of structural breaks. *Journal of econometrics*, 177(2), pp.134–152.

Platanakis, E., Sutcliffe, C. and Ye, X.X., 2021. Horses for courses: Mean-variance for asset allocation and 1/n for stock selection. *European journal of operational research*, 288(1), pp.302–317.

Rapach, D.E., Strauss, J.K., Tu, J. and Zhou, G., 2019. Industry return predictability: A machine learning approach. *The journal of financial data science*, 1(3), pp.9–28.

Rapach, D.E. and Zhou, G., 2020. Time-series and cross-sectional stock return forecasting: New machine learning methods. *Machine learning for asset management: New developments and financial applications*, pp.1–33.

Shi, F., Shu, L., Yang, A. and He, F., 2019. Improving minimum-variance portfolios by alleviating overdispersion of eigenvalues. *Journal of financial and quantitative analysis*, 55(8), pp.2700–2731.

Sirignano, J.A., Tsoukalas, G. and Giesecke, K., 2016. Large-scale loan portfolio selection. *Operations research*, 64(6), pp.1239–1255.

Stein, C., 1956. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the third berkeley symposium on mathematical statistics and probability*. pp.197–206.

Stein, C. and James, W., 1961. Estimation with quadratic loss. *Proc. 4th berkeley symp. mathematical statistics probability*. vol. 1, pp.361–379.

Stevens, G.V., 1998. On the inverse of the covariance matrix in portfolio analysis. *The journal of finance*, 53(5), pp.1821–1827.

Strehl, A. and Ghosh, J., 2002. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec), pp.583–617.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. *Journal of the royal statistical society: Series b (methodological)*, 58(1), pp.267–288.

Tola, V., Lillo, F., Gallegati, M. and Mantegna, R.N., 2008. Cluster analysis for portfolio optimization. *Journal of economic dynamics and control*, 32(1), pp.235–258.

Topchy, A., Jain, A.K. and Punch, W., 2005. Clustering ensembles: Models of consensus and weak partitions. *Ieee transactions on pattern analysis and machine intelligence*, 27(12), pp.1866–1881.

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

Wang, Y., Hao, X. and Wu, C., 2021. Forecasting stock returns: A time-dependent weighted least

squares approach. *Journal of financial markets*, 53, p.100568.

Welch, I. and Goyal, A., 2008. A comprehensive look at the empirical performance of equity premium prediction. *The review of financial studies*, 21(4), pp.1455–1508.

Zhou, Z.H. and Tang, W., 2006. Clusterer ensemble. *Knowledge-based systems*, 19(1), pp.77–83.

Zou, H. and Hastie, T., 2005. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series b (statistical methodology)*, 67(2), pp.301–320.

Zou, H. and Zhang, H.H., 2009. On the adaptive elastic-net with a diverging number of parameters. *Annals of statistics*, 37(4), p.1733.

Online Supplementary Appendix for "When Bayes-Stein Meets Machine Learning: A Generalized Approach for Portfolio Optimization"

Dimitrios Gounopoulos[1], Emmanouil Platanakis[1], Gerry Tsoukalas[2], Haoran Wu[1]

[1]University of Bath, School of Management

[2]University of Pennsylvania - The Wharton School & Boston University & Luohan Academy

This online supplementary appendix presents proof of propositions introduced in Section 3 of the paper. It also describes various datasets used in the empirical studies, including datasets taken as underlying assets for portfolio construction and many economic variables used as predictors for expected asset returns. Moreover, it supplements additional results not exhibited in the main paper, such as a complete investigation of the limitation of the original Bayes-Stein model, portfolio performance for different risk aversion parameters, and model decomposition under different expanding estimation windows. Further, it reports the results of robustness checks about our generalized Bayes-Stein portfolio optimization mode conducted in this paper.

Contents

A Proof of Propositions

To rationally develop our generalized Bayes-Stein framework in Section 3, we suggest four crucial propositions in this paper. In this appendix, we provide their proof.

A.1 Proof of Proposition 1

To prove this proposition, we simply prove the bias-variance decomposition of the mean squared error between the mean estimator and the true mean of an individual asset as follows:

$$\begin{aligned}
\mathbb{E}\left((\hat{\mu}-\mu)^2\right) &= \mathbb{E}\left((\hat{\mu}-\mathbb{E}(\hat{\mu})+\mathbb{E}(\hat{\mu})-\mu)^2\right) \\
&= \mathbb{E}\left((\mathbb{E}(\hat{\mu})-\mu)^2\right)+\mathbb{E}\left((\hat{\mu}-\mathbb{E}(\hat{\mu}))^2\right)+2\mathbb{E}((\mathbb{E}(\hat{\mu})-\mu)(\hat{\mu}-\mathbb{E}(\hat{\mu})) \\
&= \mathbb{E}\left((\mathbb{E}(\hat{\mu})-\mu)^2\right)+\mathbb{E}\left((\hat{\mu}-\mathbb{E}(\hat{\mu}))^2\right)+2(\mathbb{E}(\hat{\mu})-\mu)(\mathbb{E}(\hat{\mu})-\mathbb{E}(\hat{\mu})) \\
&= \mathbb{E}\left((\mathbb{E}(\hat{\mu})-\mu)^2\right)+\mathbb{E}\left((\hat{\mu}-\mathbb{E}(\hat{\mu}))^2\right),
\end{aligned} \tag{A.1}$$

where $\mathbb{E}\left((\mathbb{E}(\hat{\mu})-\mu)^2\right)$ and $\mathbb{E}\left((\hat{\mu}-\mathbb{E}(\hat{\mu}))^2\right)$ denote the bias and variance of the individual mean return estimator, respectively. As a result, we can obtain the bias-variance decomposition of the vector $\hat{\boldsymbol{\mu}}$ for generality, as shown in Proposition 1.

A.2 Proof of Proposition 2

For a standard multivariate regression model, denoted by $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, the statistical measure R^2 is given as follows:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{\boldsymbol{\epsilon}'\boldsymbol{\epsilon}}{T\sigma_Y^2} = 1 - \frac{\mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}}{T\sigma_Y^2} = 1 - \frac{\mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\hat{\boldsymbol{\beta}}}{T\sigma_Y^2}. \tag{A.2}$$

Moreover, the regression of (**??**) can be rewritten as follows:

$$\mathbf{R}_1 = a_1\mathbf{1} + \mathbf{R}_{-1}\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}_1, \tag{A.3}$$

where $\mathbf{R}_{-1}$ represents returns of all assets except the first one, and $\mathbf{1}$ denotes a $T \times 1$ vector of ones (T is the number of observations).

Here we can get $\mathbf{X}\hat{\boldsymbol{\beta}} = a_1\mathbf{1} + \mathbf{R}_{-1}\boldsymbol{\beta}_1 = a_1\mathbf{1} + \mathbf{R}_{-1}\mathbf{M}^{-1}\mathbf{m}$. Substituting it into the equation

of R^2, we have:

$$T\sigma_1^2 R^2 = T\sigma_1^2 - \mathbf{R}_1'\mathbf{R}_1 + \hat{a_1}\mathbf{1}'\mathbf{R}_1 + \mathbf{R}_1'\mathbf{R}_{-1}\mathbf{M}^{-1}\mathbf{m}. \tag{A.4}$$

Let $\tilde{\boldsymbol{R}}_1$ and $\tilde{\boldsymbol{R}}_{-1}$ be the demeaned $\boldsymbol{R}_1$ and $\boldsymbol{R}_{-1}$. We can get the result as follows:

$$\begin{aligned} T\left(1-R^2\right)\sigma_1^2 &= \mathbf{R}_1'\mathbf{R}_1 - \hat{a_1}\mathbf{1}'\mathbf{R}_1 - \left(\tilde{\mathbf{R}}_1 + \frac{\mathbf{1}\mathbf{1}'}{T}\mathbf{R}_1\right)'\left(\tilde{\mathbf{R}}_{-1} + \frac{\mathbf{1}\mathbf{1}'}{T}\mathbf{R}_{-1}\right)\mathbf{M}^{-1}\mathbf{m} \\ T\left(1-R^2\right)\sigma_1^2 &= \mathbf{R}_1'\mathbf{R}_1 - \hat{a_1}\mathbf{1}'\mathbf{R}_1 - T\mathbf{m}'\mathbf{M}^{-1}\mathbf{m} - \mathbf{R}_1'\frac{\mathbf{1}\mathbf{1}'}{T}\mathbf{R}_{-1}\mathbf{M}^{-1}\mathbf{m}. \end{aligned} \tag{A.5}$$

Since $\hat{a}_1 = \frac{\mathbf{1}'}{T}\left(\mathbf{R}_1 - \mathbf{R}_{-1}\mathbf{M}^{-1}\mathbf{m}\right)$, we further obtain:

$$\begin{aligned} T\left(1-R^2\right)\sigma_1^2 &= \mathbf{R}_1'\mathbf{R}_1 - \frac{\left(\mathbf{1}_T'\mathbf{R}_1\right)^2}{T} - T\mathbf{m}'\mathbf{M}^{-1}\mathbf{m} \\ \left(1-R^2\right)\sigma_1^2 &= \sigma_1^2 - \mathbf{m}'\mathbf{M}^{-1}\mathbf{m}. \end{aligned} \tag{A.6}$$

A.3 Proof of Proposition 3

Denoting two label factor vectors by $\boldsymbol{\lambda}^{(p)}$ and $\boldsymbol{\lambda}^{(q)}$, their normalized mutual information $\Phi^{(\mathrm{NMI})}\left(\boldsymbol{\lambda}^{(p)}, \boldsymbol{\lambda}^{(q)}\right)$ is given as follows:

$$\Phi^{(\mathrm{NMI})}\left(\boldsymbol{\lambda}^{(p)}, \boldsymbol{\lambda}^{(q)}\right) = \frac{I(\boldsymbol{\lambda}^{(p)}, \boldsymbol{\lambda}^{(q)})}{\sqrt{H(\boldsymbol{\lambda}^{(p)})H(\boldsymbol{\lambda}^{(q)})}}, \tag{A.7}$$

where $I(\boldsymbol{\lambda}^{(p)}, \boldsymbol{\lambda}^{(q)})$ represents the mutual information between $\boldsymbol{\lambda}^{(p)}$ and $\boldsymbol{\lambda}^{(q)}$, and $H(\boldsymbol{\lambda}^{(p)})$ and $H(\boldsymbol{\lambda}^{(q)})$ are their entropy (a measurement that quantifies uncertainty).

Since $H(X) = -\sum_{x\in\mathcal{X}} p(x)\log p(x)$ and $I(X,Y) = \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} P_{(X,Y)}(x,y)\log\left(\frac{P_{(X,Y)}(x,y)}{P_X(x)P_Y(y)}\right)$,

we can get the following:

$$
\begin{aligned}
H(\boldsymbol{\lambda}^{(p)}) &= -\sum_{h=1}^{k^{(p)}} \frac{n_h^{(p)}}{N} \log \frac{n_h^{(p)}}{N} \\
H(\boldsymbol{\lambda}^{(q)}) &= -\sum_{f=1}^{k^{(q)}} \frac{n_f^{(q)}}{N} \log \frac{n_f^{(q)}}{N} \\
I(\boldsymbol{\lambda}^{(p)}, \boldsymbol{\lambda}^{(q)}) &= \sum_{h=1}^{k^{(p)}} \sum_{f=1}^{k^{(q)}} \frac{n_{h,f}}{N^2} \log \left(\frac{N \cdot n_{h,f}}{n_h^{(p)} n_f^{(q)}} \right),
\end{aligned}
\tag{A.8}
$$

where $n_h^{(p)}$ and $n_f^{(q)}$ represent the number of assets in the cluster C_h of $\boldsymbol{\lambda}^{(p)}$ and cluster C_f of $\boldsymbol{\lambda}^{(q)}$, respectively. $n_{h,f}$ denotes the number of assets appearing in the cluster C_h and C_f simultaneously. N is the number of all assets.

As a result, we obtain the final normalized mutual information $\Phi^{(\mathrm{NMI})}\left(\boldsymbol{\lambda}^{(p)}, \boldsymbol{\lambda}^{(q)}\right)$, given as follows:

$$
\Phi^{(\mathrm{NMI})}\left(\boldsymbol{\lambda}^{(p)}, \boldsymbol{\lambda}^{(q)}\right) = \frac{\sum_{h=1}^{k^{(p)}} \sum_{f=1}^{k^{(q)}} n_{h,f} \log \left(\frac{N \cdot n_{h,f}}{n_h^{(p)} n_f^{(q)}} \right)}{\sqrt{\left(\sum_{h=1}^{k^{(p)}} n_h^{(p)} \log \frac{n_h^{(p)}}{N} \right) \left(\sum_{f=1}^{k^{(q)}} n_f^{(q)} \log \frac{n_f^{(q)}}{N} \right)}}. \tag{A.9}
$$

A.4 Proof of Proposition 4

N Risky asset returns over a given time horizon T, assuming they are independent and identically distributed (i.i.d), follow a multivariate normal distribution:

$$
\boldsymbol{R}_t \sim \mathrm{NID}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad t = 1, \ldots, T. \tag{A.10}
$$

The probability density function (pdf) of a multivariate normal distribution is given by

$$
f(\boldsymbol{R}) = \frac{1}{\sqrt{(2\pi)^N |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2} (\boldsymbol{R} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{R} - \boldsymbol{\mu}) \right). \tag{A.11}
$$

Then We can obtain the likelihood function by

$$\prod_{t=1}^{T} f\left(\boldsymbol{R}_t\right) = \prod_{t=1}^{T} \frac{1}{\sqrt{(2\pi)^N |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\boldsymbol{R_t} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{R_t} - \boldsymbol{\mu})\right). \tag{A.12}$$

So the log-likelihood function is:

$$\log\left(\prod_{t=1}^{T} f\left(\boldsymbol{R}_t\right)\right) = \sum_{t=1}^{T}\left(\log\left((2\pi)^{\frac{-N}{2}}\right) + \log\left(|\boldsymbol{\Sigma}|^{\frac{-1}{2}}\right) - \frac{1}{2}\left(\boldsymbol{R}_t - \boldsymbol{\mu}\right)'\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{R}_t - \boldsymbol{\mu}\right)\right). \tag{A.13}$$

Then we have:

$$\log\left(\prod_{t=1}^{T} f\left(\boldsymbol{R}_t\right)\right) \propto \sum_{t=1}^{T}\left(-\frac{1}{2}\log\left(|\boldsymbol{\Sigma}|\right) - \frac{1}{2}\left(\boldsymbol{R}_t - \boldsymbol{\mu}\right)'\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{R}_t - \boldsymbol{\mu}\right)\right). \tag{A.14}$$

Denote the inverse covariance matrix by $\boldsymbol{\Theta}$ (as well as $\boldsymbol{\Sigma}^{-1}$), and it is also called a precision matrix. Let $\hat{\boldsymbol{\Sigma}}_S$ be the sample covariance matrix. Consequently, we can obtain the following result:

$$\begin{aligned} &-\frac{T}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{t=1}^{T}\left(\boldsymbol{R}_t - \boldsymbol{\mu}\right)'\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{R}_t - \boldsymbol{\mu}\right) \\ =&-\frac{T}{2}\log\left|\boldsymbol{\Theta}^{-1}\right| - \frac{1}{2}\sum_{t=1}^{T}\left(\boldsymbol{R}_t - \boldsymbol{\mu}\right)'\boldsymbol{\Theta}\left(\boldsymbol{R}_t - \boldsymbol{\mu}\right). \end{aligned} \tag{A.15}$$

Given the scalar $y'Ay = \text{trace}\left(y'Ay\right)$, trace(ABC) = trace(CAB), $\sum_t \text{trace}\left(A_tB\right) = \text{trace}\left(\sum_t A_tB\right)$ and $\sum_t A_tB = \left(\sum_t A_t\right)B$, the log-likelihood function is determined by

$$
\begin{aligned}
&-\frac{T}{2}\log\left|\boldsymbol{\Theta}^{-1}\right|-\frac{1}{2}\sum_{t=1}^{T}\left(\boldsymbol{R}_t-\boldsymbol{\mu}\right)'\boldsymbol{\Theta}\left(\boldsymbol{R}_t-\boldsymbol{\mu}\right)\\
=&-\frac{T}{2}\log\left|\boldsymbol{\Theta}\right|^{-1}-\frac{1}{2}\sum_{t=1}^{T}\operatorname{trace}\left(\left(\boldsymbol{R}_t-\boldsymbol{\mu}\right)'\boldsymbol{\Theta}\left(\boldsymbol{R}_t-\boldsymbol{\mu}\right)\right)\\
=&\frac{T}{2}\log\left|\boldsymbol{\Theta}\right|-\frac{1}{2}\sum_{t=1}^{T}\operatorname{trace}\left(\left(\boldsymbol{R}_t-\boldsymbol{\mu}\right)\left(\boldsymbol{R}_t-\boldsymbol{\mu}\right)'\boldsymbol{\Theta}\right)\\
=&\frac{T}{2}\log\left|\boldsymbol{\Theta}\right|-\frac{T}{2}\operatorname{trace}\left(\boldsymbol{S}\boldsymbol{\Theta}\right),
\end{aligned}
\tag{A.16}
$$

where $\boldsymbol{S}=\frac{1}{T}\sum_{t=1}^{T}\left(\boldsymbol{R}_t-\boldsymbol{\mu}\right)\left(\boldsymbol{R}_t-\boldsymbol{\mu}\right)'$ denotes the sample covariance matrix.

B Description of the Empirical Datasets

To validate the effectiveness and robustness of our generalized Bayes-Stein framework, we conduct multiple empirical studies over different datasets as portfolio assets. In particular, to implement the TW-ENet approach in our generalized Bayes-Stein framework, we need a set of commonly used economic variables as inputs (independent variables) of this method. This appendix discusses these datasets in detail.

B.1 Asset Datasets Description

The original empirical datasets mainly comprise two types: industry portfolios and Fama-French factor-sorted portfolios. We extract their monthly value-weighted returns from Ken French's website (https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/). As a result, excess returns over the 1-month T-bill return can be obtained by data processing. Data details are given as follows.

Industry Portfolios: We mainly use the 10 industry portfolios. It has 10 components: Consumer Durables, Manufacturing, Energy, Hi-Tech Business Equipment, Telephone and Television Transmission, Shops, Health, Utilities, Other.

Fama-French Factor-sorted Portfolios: Following the study of DeMiguel, Garlappi and Uppal (2009), we also use similar Fama-French portfolios for evaluating the out-of-sample performance of our proposed generalized Bayes-Stein framework. The following data sets are included:

- 20 size and book-to-market portfolios and the US equity MKT: the 25 Fama-French portfolios formed on size and book-to-market ratio excluding the five portfolios containing the largest firms, the market factor (the US equity MKT).
- 20 size and book-to-market portfolios and the MKT, SMB, and HML portfolios: the 25 Fama-French portfolios formed on size and book-to-market ratio excluding the five portfolios containing the largest firms, Fama-French three factors (the MKT, SMB, and HML portfolios).
- 20 size and book-to-market portfolios and the MKT, SMB, and HML portfolios: the 25 Fama-French portfolios formed on size and book-to-market ratio excluding the five portfolios containing the largest firms, Fama-French four factors (the MKT, SMB, HML, and UMD portfolios).

- 25 Portfolios Formed on Size and Book-to-Market: The portfolios are the intersections of 5 portfolios formed on size (market equity, ME) and 5 on the ratio of book equity to market equity (BE/ME).
- 100 Portfolios Formed on Size and Book-to-Market: The portfolios are the intersections of 10 portfolios formed on size (market equity, ME) and 10 on the ratio of book equity to market equity (BE/ME).
- 100 Portfolios Formed on Size and Operating Profitability: The portfolios are the intersections of 10 portfolios formed on size (market equity, ME) and 10 on profitability (OP).
- 100 Portfolios Formed on Size and Investment: The portfolios are the intersections of 10 portfolios formed on size (market equity, ME) and 10 on investment (INV).

B.2 Economic Variables Description

We adopt 14 commonly used economic variables of Welch and Goyal (2008) and lagged components returns to predict expected returns of various industry portfolios and Fama-French portfolios. For example, in 10 industry portfolios, we use the 14 economic variables and excess returns of all ten industry components at time T as predictors to forecast the return of one of 10 industries at time $T + 1$. The 14 economic variables are extracted from Amit Goyal's website (https://sites.google.com/view/agoyal145), given as follows:

- Dividend-price ratio (log): the difference between the log of dividends paid on the S&P 500 index and the log of prices (S&P 500 Index), where dividends are measured using a one-year moving sum.
- Dividend yield (log): the difference between the log of dividends and the log of lagged prices.
- Earnings-price ratio (log): the difference between the log of earnings on the S&P 500 Index and the log of prices, where earnings are measured using a one-year moving sum.
- Dividend payout ratio (log): the difference between the log of dividends and the log of earnings on the S&P 500 Index.

- Book-to-market ratio: ratio of book value to market value for the Dow Jones Industrial Average.
- Default return spread: the difference between long-term corporate bond and long-term government bond returns.
- Long-term yield: long-term government bond yield.
- Long-term return: return on long-term government bonds.
- Term spread: the difference between the long-term yield and the Treasury bill rate.
- Treasury bill rate: the interest rate on a three-month Treasury bill (secondary market).
- Default yield spread: the difference between BAA- and AAA-rated corporate bond yields.
- Stock variance: sum of squared daily returns on the S&P 500 Index.
- Inflation: calculated from the CPI (all urban consumers).
- Net equity expansion: ratio of 12-month moving sums of net issues by NYSE-listed stocks to the total end-of-year market capitalization of NYSE stocks.

C Additional Results

In the main paper, for brevity, we only exhibit representative results to back up our arguments. In this appendix, we provide additional results mentioned in the paper, such as an additional investigation of the classical Bayes-Stein model on other datasets and the portfolio performance of our generalized Bayes-Stein framework under other investor risk aversion parameters, model decomposition under different expanding estimation windows.

C.1 Limitations of the Bayes-Stein Model

Section 2.2 of the main paper demonstrates the drawbacks of the traditional Bayes-Stein model on the 10 industry portfolios. Herein we provide supplementary results on other datasets. Since we mainly use industry portfolios and similar Fama-French factor-sorted portfolios as assets in this paper, we offer additional results about the mean estimator on FF21 (20 size and book-to-market portfolios and the US equity MKT) for brevity in Table A.1, and the shrinkage factor and the inverse covariance matrix on the datasets not covered in Section 2.2 in Tables A.2 and A.3.

According to Table A.1, we can find that the overall performance difference between the Bayes-Stein means estimator and the sample mean is minor, and the sample means even surpass the Bayes-Stein means estimator in some individual components (e.g., ME4 BM1), which is consistent with our findings in the main paper. Moreover, Table A.2 reconfirms that the shrinkage factors of the Bayes-Stein model are time-varying and asset-varying, which implies that it is necessary to improve the classical shrinkage factor by considering the differences between assets. Additionally, Table A.3 indicates that the Bayes-Stein model does not enhance the inverse covariance matrix estimation, which is consistent with our arguments in the main paper.

Table A.1: Mean Squared Forecasting Error of the Mean Return Estimators

Table A.1 reports the out-of-sample mean squared forecasting error (MSFE, $\times 10^4$) of the monthly sample mean and Bayes-Stein mean estimator for FF21 (20 size and book-to-market portfolios and the US equity MKT). Asset name abbreviations are in line with ones of the Fama-French website. The estimation process is based on a 20-year expanding window with the initial data period from July 1963 to June 1983. The out-of-sample period covers from July 1983 to December 2021.

Asset	Sample Mean	Bayes-Stein
SMALL LoBM	61.51	61.39
ME1 BM2	47.31	47.28
ME1 BM3	32.47	32.43
ME1 BM4	31.05	31.03
SMALL HiBM	36.77	36.73
ME2 BM1	49.49	49.45
ME2 BM2	34.20	34.21
ME2 BM3	27.68	27.69
ME2 BM4	27.33	27.33
ME2 BM5	38.52	38.49
ME3 BM1	42.13	42.14
ME3 BM2	29.31	29.34
ME3 BM3	23.99	24.00
ME3 BM4	26.22	26.24
ME3 BM5	33.51	33.53
ME4 BM1	33.78	33.82
ME4 BM2	25.37	25.42
ME4 BM3	25.18	25.20
ME4 BM4	25.09	25.11
ME4 BM5	32.94	32.96
Mkt-RF	20.01	20.00
Sum	703.85	703.79

Table A.2: Descriptive Statistics of Shrinkage Factors

Table A.2 presents the descriptive statistics of Bayes-Stein shrinkage factors, including max, min, mean, and standard deviation (SD), over different datasets (FF21, FF23, FF24, FF25, FF100BM, FF100OP, FF100INV). The estimation process is based on a 20-year expanding window with the initial data period from July 1963 to June 1983. The out-of-sample period covers from July 1983 to December 2021.

	FF21	FF23	FF24	FF25	FF100BM	FF100OP	FF100INV
Max	0.20	0.19	0.14	0.49	0.44	0.45	0.37
Min	0.11	0.09	0.07	0.25	0.25	0.37	0.27
Mean	0.13	0.11	0.09	0.33	0.32	0.42	0.30
SD	0.02	0.02	0.01	0.07	0.06	0.02	0.03

Table A.3: Condition Numbers of Inverse Covariance Matrix Estimates

Table A.3 exhibits the mean and standard deviation of condition numbers of the sample inverse covariance matrix ($\boldsymbol{S}^{-1}$) and the Bayes-Stein inverse covariance matrix ($\hat{\boldsymbol{\Sigma}}_{BS}^{-1}$) during the out-of-sample period over additional datasets (FF21, FF23, FF24, FF100OP, FF100INV), to demonstrate the estimation risk of the Bayes-Stein inverse covariance matrix. The estimation process is based on a 20-year expanding window with the initial data period from July 1963 to June 1983. The out-of-sample period covers from July 1983 to December 2021.

	FF21		FF23		FF24		FF100OP		FF100INV	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
$\boldsymbol{S}^{-1}$	891.44	95.95	3442.20	761.34	3469.70	766.16	2815.10	1040.40	2985.00	940.08
$\hat{\boldsymbol{\Sigma}}_{BS}^{-1}$	891.53	95.98	3441.30	760.81	3469.10	765.81	2815.80	1040.80	2985.60	940.38

C.2 Portfolio Performance of Other Risk Aversion Parameters

In the main paper, we only present the results for risk aversion parameters of $\lambda = 1$, $\lambda = 3$, and $\lambda = 5$ to represent investors who are aggressive, moderate, and conservative to risks, respectively. Here we report additional results for $\lambda = 2$ and $\lambda = 4$. Note that portfolio performance is measured after transaction costs with a transaction cost parameter of $\delta = 3 \times 10^{-7}$. Importantly, we are convinced that the findings in the main paper still hold after examining the results in Table A.4.

Table A.4: Portfolio Performance for Other Risk Aversion Parameters

Table A.4 reports the out-of-sample yearly Sharpe ratios (SRs) and certainty equivalent returns (CERs) of the generalized Bayes-Stein framework (GBS) and the naive equal weighted scheme (1/N) under various datasets (Ind10, FF21, FF23, FF24, FF25, FF100BM, FF100OP, and FF100INV) for other risk aversion parameters ($\lambda = 2$ and $\lambda = 4$) considering transaction costs (TCs). The results of the 20-year (the out-of-sample period covers from July 1983 to December 2021) and 40-year (the out-of-sample period covers from July 2003 to December 2021) expanding estimation windows are exhibited in Panel A and Panel B, respectively. The results of significance tests of the performance differences (GBS vs 1/N) are put in the parenthesis, where *, **, and *** indicate a statistical significance at the 10%, 5%, and 1%-level, respectively.

Panel A: SRs and CERs with TCs for a 20-year expanding estimation window

Dataset	SRs				CERs			
	$\lambda = 2$		$\lambda = 4$		$\lambda = 2$		$\lambda = 4$	
	GBS	1/N	GBS	1/N	GBS	1/N	GBS	1/N
Ind10	0.6647	0.6356	0.6559	0.6356	0.0769	0.0668	0.0531	0.0491
FF21	0.6517 (**)	0.5226	0.6543 (**)	0.5226	0.0818 (**)	0.0608	0.0535 (**)	0.0303
FF23	0.6510 (**)	0.5050	0.6459 (**)	0.5050	0.0791 (**)	0.0556	0.0521 (**)	0.0293
FF24	0.6438 (**)	0.5155	0.6543 (**)	0.5155	0.0759 (**)	0.0557	0.0533 (**)	0.0319
FF25	0.7042 (***)	0.5497	0.7035 (***)	0.5497	0.0920 (***)	0.0642	0.0618 (***)	0.0359
FF100BM	0.7072 (***)	0.3816	0.7055 (***)	0.3816	0.097 (***)	0.0362	0.0621 (***)	0.0051
FF100OP	0.6640 (**)	0.5635	0.6629 (**)	0.5635	0.0898 (***)	0.0678	0.0540 (**)	0.0375
FF100INV	0.7237 (***)	0.5470	0.7249 (***)	0.5470	0.0976 (***)	0.0647	0.0657 (***)	0.0348

Panel B: SRs and CERs with TCs for a 40-year expanding estimation window

Dataset	SRs				CERs			
	$\lambda = 2$		$\lambda = 4$		$\lambda = 2$		$\lambda = 4$	
	GBS	1/N	GBS	1/N	GBS	1/N	GBS	1/N
Ind10	0.7551	0.7518	0.7521	0.7518	0.0952	0.0825	0.0690	0.0647
FF21	0.7196	0.6238	0.7349	0.6238	0.0947	0.0801	0.0672	0.0474
FF23	0.7542 (*)	0.6049	0.7595 (*)	0.6049	0.0999 (*)	0.0738	0.0713 (*)	0.0451
FF24	0.6814	0.6124	0.6946	0.6124	0.0859	0.0725	0.0601	0.0468
FF25	0.7819 (***)	0.6322	0.7925 (***)	0.6322	0.1088 (***)	0.0801	0.0778 (***)	0.0494
FF100BM	0.8562 (***)	0.5169	0.8565 (***)	0.5169	0.1300 (***)	0.0613	0.0910 (***)	0.0273
FF100OP	0.7405 (*)	0.6552	0.7453 (*)	0.6552	0.1033 (**)	0.0855	0.0694 (*)	0.0532
FF100INV	0.8009 (**)	0.6508	0.8060 (**)	0.6508	0.1095 (**)	0.0842	0.0798 (**)	0.0525

C.3 Portfolio Performance of Model Decomposition

We divide our generalized Bayes-Stein model into two sub-models for a more comprehensive analysis. Model 1 includes only the TW-ENet approach, while Model 2 does not. In the

main paper, we evaluated the portfolio performance of these two models under a 20-year expanding estimation window, considering transaction costs. We do this due to the representativeness of this portfolio construction design. Here we give additional results, including without transaction costs and a 40-year expanding estimation window with transaction costs, in Tables A.5 and A.6. Results demonstrate that our generalized Bayes-Stein model incorporates the merits of these two sub-models and offers more stable and superior out-of-sample performance than the 1/N asset allocation rule.

Table A.5: SRs and CERs without TCs for Model Decomposition

The generalized Bayes-Stein model (GBS) is decomposed into two sub-models, one that only includes the TW-ENet approach (model 1) and another that does not have the TW-ENet method (model 2). Table A.5 reports the out-of-sample yearly Sharpe ratios (SRs) and certainty equivalent returns (CERs) of the two sub-models and the naive equal weighted scheme (1/N) under various datasets (Ind10, FF21, FF23, FF24, FF25, FF100BM, FF100OP, and FF100INV) for different risk aversion parameters ($\lambda = 1, 3, 5$) without considering transaction costs (TCs), which are exhibited in Panel A and Panel B, respectively. The results are obtained under a 20-year expanding estimation window (the out-of-sample period covers from July 1983 to December 2021). Significance tests of the performance differences (GBS-model 1 vs 1/N, GBS-model 2 vs 1/N) are put in the parenthesis, where *, **, and *** indicate a statistical significance at the 10%, 5%, and 1%-level, respectively.

Panel A: SRs and CERs of GBS-model 1

Dataset	SRs						CERs					
	$\lambda = 1$		$\lambda = 3$		$\lambda = 5$		$\lambda = 1$		$\lambda = 3$		$\lambda = 5$	
	GBS-model 1	1/N	GBS-model 1	1/N	GBS-model 1	1/N	GBS-model 1	1/N	GBS-model 1	1/N	GBS-model 1	1/N
Ind10	0.6718	0.6356	0.6515	0.6356	0.6187	0.6356	0.0883 (*)	0.0756	0.0616	0.0580	0.0383	0.0403
FF21	0.6391 (**)	0.5226	0.6956 (***)	0.5226	0.7042 (**)	0.5226	0.093 (*)	0.0760	0.0719 (**)	0.0455	0.0496 (***)	0.0150
FF23	0.6349 (**)	0.5050	0.6944 (**)	0.5050	0.6981 (**)	0.5050	0.088 (*)	0.0688	0.0684 (**)	0.0425	0.0484 (**)	0.0161
FF24	0.6695 (**)	0.5155	0.7077 (**)	0.5155	0.7217 (**)	0.5155	0.0896 (**)	0.0676	0.0674 (**)	0.0438	0.0502 (**)	0.0200
FF25	0.6967 (***)	0.5497	0.7089 (***)	0.5497	0.6905 (**)	0.5497	0.1047 (***)	0.0783	0.0752 (***)	0.0500	0.0476 (***)	0.0217
FF100BM	0.7183 (***)	0.3816	0.7311 (***)	0.3816	0.7299 (***)	0.3816	0.1161 (***)	0.0518	0.0827 (***)	0.0206	0.0526 (***)	-0.0105
FF100OP	0.6593 (**)	0.5635	0.6952 (***)	0.5635	0.6699 (**)	0.5635	0.1061 (***)	0.0829	0.0754 (***)	0.0526	0.0439 (**)	0.0224
FF100INV	0.7418 (***)	0.5470	0.7388 (***)	0.5470	0.7414 (***)	0.5470	0.1156 (***)	0.0796	0.0825 (***)	0.0497	0.0549 (***)	0.0198

Panel B: SRs and CERs of GBS-model 2

Dataset	SRs						CERs					
	$\lambda = 1$		$\lambda = 3$		$\lambda = 5$		$\lambda = 1$		$\lambda = 3$		$\lambda = 5$	
	GBS-model 2	1/N	GBS-model 2	1/N	GBS-model 2	1/N	GBS-model 2	1/N	GBS-model 2	1/N	GBS-model 2	1/N
Ind10	0.6362	0.6356	0.6385	0.6356	0.6474	0.6356	0.0790	0.0756	0.0598	0.0580	0.0417	0.0403
FF21	0.5799 (*)	0.5226	0.5832 (**)	0.5226	0.5838 (**)	0.5226	0.0913 (**)	0.0760	0.0565 (**)	0.0455	0.0235 (**)	0.0150
FF23	0.5829 (**)	0.5050	0.5876 (***)	0.5050	0.5885 (***)	0.5050	0.0919 (***)	0.0688	0.0573 (***)	0.0425	0.0244 (*)	0.0161
FF24	0.5930 (**)	0.5155	0.5927 (***)	0.5155	0.5950 (***)	0.5155	0.0941 (***)	0.0676	0.0583 (**)	0.0438	0.0259	0.0200
FF25	0.5900	0.5497	0.5904	0.5497	0.5903 (*)	0.5497	0.0932 (**)	0.0783	0.0578	0.0500	0.0252	0.0217
FF100BM	0.5852 (***)	0.3816	0.5814 (***)	0.3816	0.5777 (***)	0.3816	0.0895 (***)	0.0518	0.056 (***)	0.0206	0.0232 (***)	-0.0105
FF100OP	0.5232	0.5635	0.5414	0.5635	0.5430	0.5635	0.0789	0.0829	0.0488	0.0526	0.0172	0.0224
FF100INV	0.5837 (*)	0.5470	0.5912 (**)	0.5470	0.5946 (***)	0.5470	0.0883 (**)	0.0796	0.0577 (**)	0.0497	0.0269 (**)	0.0198

Table A.6: SRs and CERs with TCs for Model Decomposition

The generalized Bayes-Stein model (GBS) is decomposed into two sub-models, one that only includes the TW-ENet approach (model 1) and another that does not have the TW-ENet method (model 2). Table A.6 reports the out-of-sample yearly Sharpe ratios (SRs) and certainty equivalent returns (CERs) of the two sub-models and the naive equal weighted scheme (1/N) under various datasets (Ind10, FF21, FF23, FF24, FF25, FF100BM, FF100OP, and FF100INV) for different risk aversion parameters ($\lambda = 1, 3, 5$) considering transaction costs (TCs), which are exhibited in Panel A and Panel B, respectively. The results are obtained under a 40-year expanding estimation window (the out-of-sample period covers from July 2003 to December 2021). Significance tests of the performance differences (GBS-model 1 vs 1/N, GBS-model 2 vs 1/N) are put in the parenthesis, where *, **, and *** indicate a statistical significance at the 10%, 5%, and 1%-level, respectively.

Panel A: SRs and CERs of GBS-model 1

Dataset	SRs						CERs					
	$\lambda = 1$		$\lambda = 3$		$\lambda = 5$		$\lambda = 1$		$\lambda = 3$		$\lambda = 5$	
	GBS-model 1	1/N	GBS-model 1	1/N	GBS-model 1	1/N	GBS-model 1	1/N	GBS-model 1	1/N	GBS-model 1	1/N
Ind10	0.7666	0.7518	0.7818	0.7518	0.7879	0.7518	0.1097 (*)	0.0915	0.0853	0.0736	0.0615	0.0558
FF21	0.7061	0.6238	0.7610	0.6238	0.7835 (*)	0.6238	0.1064	0.0965	0.0848	0.0638	0.0614 (*)	0.0311
FF23	0.7407	0.6049	0.7655 (*)	0.6049	0.7497	0.6049	0.1112 (*)	0.0882	0.0844 (*)	0.0594	0.0562	0.0307
FF24	0.6741	0.6124	0.7229	0.6124	0.7235	0.6124	0.0971	0.0854	0.0742	0.0597	0.0521	0.0339
FF25	0.7965 (***)	0.6322	0.8039 (***)	0.6322	0.7598 (*)	0.6322	0.1266 (***)	0.0955	0.0938 (***)	0.0647	0.0576 (**)	0.0340
FF100BM	0.8761 (***)	0.5169	0.8913 (***)	0.5169	0.8588 (***)	0.5169	0.1530 (***)	0.0783	0.1150 (***)	0.0443	0.0737 (***)	0.0104
FF100OP	0.7217	0.6552	0.7249	0.6552	0.7069	0.6552	0.1167 (*)	0.1016	0.0819	0.0693	0.0483	0.0370
FF100INV	0.8164 (***)	0.6508	0.8165 (***)	0.6508	0.8571 (***)	0.6508	0.1262 (**)	0.1001	0.0953 (**)	0.0683	0.0730 (***)	0.0366

Panel B: SRs and CERs of GBS-model 2

Dataset	SRs						CERs					
	$\lambda = 1$		$\lambda = 3$		$\lambda = 5$		$\lambda = 1$		$\lambda = 3$		$\lambda = 5$	
	GBS-model 2	1/N	GBS-model 2	1/N	GBS-model 2	1/N	GBS-model 2	1/N	GBS-model 2	1/N	GBS-model 2	1/N
Ind10	0.7814	0.7518	0.7779	0.7518	0.7794	0.7518	0.0949	0.0915	0.0772	0.0736	0.0596	0.0558
FF21	0.5791	0.6238	0.5924	0.6238	0.6027	0.6238	0.0987	0.0965	0.0585	0.0638	0.0211	0.0311
FF23	0.5627	0.6049	0.5732	0.6049	0.5838	0.6049	0.0961	0.0882	0.0545	0.0594	0.0167	0.0307
FF24	0.5657	0.6124	0.5762	0.6124	0.5880	0.6124	0.0968	0.0854	0.0551	0.0597	0.0178	0.0339
FF25	0.5714	0.6322	0.5833	0.6322	0.593	0.6322	0.0977	0.0955	0.0566	0.0647	0.0191	0.0340
FF100BM	0.6133 (*)	0.5169	0.6125 (**)	0.5169	0.6118 (***)	0.5169	0.1025 (**)	0.0783	0.0624 (**)	0.0443	0.0243 (**)	0.0104
FF100OP	0.6303	0.6552	0.6440	0.6552	0.6388	0.6552	0.1026	0.1016	0.0680	0.0693	0.0319	0.0370
FF100INV	0.6168	0.6508	0.6277	0.6508	0.6328	0.6508	0.0987	0.1001	0.0650	0.0683	0.0310	0.0366

D Robustness Checks

In this appendix, we report the results of all robustness checks described in Section 5 about our generalized Bayes-Stein framework, including alternative time series forecasting methods and different portfolio construction designs.

D.1 Alternative Time-series Return Forecasting Methods

We integrate some well-established machine learning methods into our generalized Bayes-Stein framework to examine the robustness of our proposed TW-ENet approach, including the OLS post-LASSO approach, the combination Elastic Net (C-ENet) method, and the random forest technique.[1]

To describe these methods, we consider a classical predictive regression model with a set of candidate predictors:

$$r_t = \beta_0 + \sum_{d=1}^{D} \beta_d x_{d,t-1} + \epsilon_t, \tag{A.17}$$

where r_t is the expected excess return at time t and $x_{d,t-1}$ is the d^{th} predictor variable at time $t-1$. ϵ_t represents a zero-mean disturbance term. D denotes the number of predictors.

The LASSO method of Tibshirani (1996) can set numerous coefficients to zero and refine the above function by

$$\underset{\beta_0,\ldots,\beta_D \in \mathbb{R}}{\arg\min} \left[\frac{1}{2T} \sum_{t=1}^{T} \left(r_t - \beta_0 - \sum_{d=1}^{D} \beta_d x_{d,t-1} \right)^2 + \rho \sum_{d=1}^{D} |\beta_d| \right], \tag{A.18}$$

where ρ represents the regularization parameter.

Based on the LASSO-selected predictors, we further apply the traditional ordinary least squares (OLS) regression. The whole process is called OLS post-LASSO estimation and has been employed in previous literature to forecast the expected returns of industry assets (see, Rapach et al., 2019).

In terms of the C-ENet approach, it aims to improve the forecast combination method by leveraging the ability of Elastic Net to select individual forecasts (Rapach and Zhou, 2020;

[1] We only briefly introduce the OLS post-LASSO and C-ENet since the random forest method is very common in forecasting literature.

Dong et al., 2022). Denoting individual forecasts of D predictors by $\hat{r}_{d,t}$, where $d = 1, \ldots, D$, C-ENet can identify necessary forecast components by

$$\underset{\theta_0,\ldots,\theta_D \in \mathbb{R}}{\arg\min} \frac{1}{2L} \sum_{t=T+1}^{T+L} \left(r_t - \theta_0 - \sum_{d=1}^{D} \theta_d \hat{r}_{d,t} \right)^2 + \lambda P, \tag{A.19}$$

where

$$P = \rho \sum_{d=1}^{D} |\theta_d| + \frac{1}{2}(1-\rho) \sum_{d=1}^{D} \theta_d^2, \tag{A.20}$$

the time between $T+1$ and $T+L$ is the training period, θ_d denotes the combination weights of individual forecasts, ρ represents a compromise between ridge ($\rho = 0$) and LASSO ($\rho = 1$), and λ controls the overall penalty strength.

Consequently, the individual forecasts selected by (A.19) are averaged to yield the final forecasting results.

Finally, the Sharpe ratios and certainty equivalent returns of applying different machine learning methods in our generalized Bayes-Stein framework are presented in Table A.7. According to the results, our TW-ENet is relatively inferior to one or two of these three Benchmark methods in some datasets. For example, in Ind10, the performance of the combination Elastic Net (C-ENet) method is best, whereas, in FF24, the OLS post-LASSO (OPL) approach is best. Moreover, in large portfolios (FF100BM, FF100OP, and FF100INV), the performance of random forest (RF) is comparable to our TW-ENet. However, our TW-ENet approach's performance is stable in all datasets, which verifies its overall robustness over other machine learning methods. Therefore, the robustness check results prove the TE-ENet approach's comparable performance to other well-established methods and reflect the flexibility of our generalized Bayes-Stein framework. As a result, this framework can easily integrate other advanced machine learning methods.

D.2 Portfolio Performance of Other Transaction Cost Parameters

Following DeMiguel, Martín-Utrera and Nogales (2015), we apply other transaction cost parameters ($\delta = 3 \times 10^{-6}$ and $\delta = 3 \times 10^{-8}$) to evaluate further the impacts of transaction costs on our generalized Bayes-Stein framework. The results are presented in Table A.8 and A.9, respectively. Consistent with the findings in the main paper, our generalized

Bayes-Stein framework can still provide better out-of-sample portfolio performance than 1/N, demonstrating our model's insensitivity to transaction cost parameters.

D.3 Portfolio Performance of Other Expanding Window

Different initial expanding window lengths result in different out-of-sample periods. For example, in this paper, the 20-year window means the out-of-sample period from July 1983 to December 2021, and the 40-year window represents the out-of-sample period from July 2003 to December 2021. To provide more evidence supporting our generalized Bayes-Stein framework's superiority relative to the 1/N strategy, we also apply a 30-year expanding window (indicating an out-of-sample period from July 1993 to December 2021), shown in Table A.10. Note that here we apply a high transaction cost parameter of $\delta = 3 \times 10^{-6}$ due to its stringency on the asset allocation model. Likewise, our generalized Bayes-Stein model can surpass the 1/N rule under the 30-year expanding window.

D.4 Rolling Window Estimation

We offer the empirical results based on the expanding window estimation in the main paper. Here we provide results via a 20-year rolling window estimation, as shown in Table A.11. Though we implement a different design, the results are consistent with the ones of the main paper. Therefore, the performance of our generalized Bayes-Stein model is robust under different types of estimation windows.

Table A.7: Portfolio Performance of Different Machine Learning Methods

Table A.7 reports the out-of-sample yearly Sharpe ratios (SRs, Panel A) and certainty equivalent returns (CERs, Panel B) of different machine learning methods for time-series return forecasting in the generalized Bayes-Stein framework, including our TW-ENet approach, the OLS post-LASSO (OPL), the combination Elastic Net (C-ENet), and random forest (RF). The performance metrics are measured under different risk aversion parameters ($\lambda = 1, 3, 5$) and transaction costs (TCs, a transaction cost parameter of $\delta = 3 \times 10^{-7}$), based on a 20-year expanding window (the out-of-sample period covers from July 1983 to December 2021).

Panel A: SRs

Dataset	$\lambda = 1$				$\lambda = 3$				$\lambda = 5$			
	TW-ENet	OPL	C-ENet	RF	TW-ENet	OPL	C-ENet	RF	TW-ENet	OPL	C-ENet	RF
Ind10	0.6664	0.6539	0.6935	0.5971	0.6604	0.6641	0.6736	0.5904	0.6548	0.6671	0.6697	0.5945
FF21	0.6453	0.6677	0.6041	0.6598	0.6544	0.6591	0.6027	0.6610	0.6525	0.6572	0.6030	0.6665
FF23	0.6529	0.6722	0.6006	0.6446	0.6490	0.6784	0.5985	0.6554	0.6421	0.6878	0.5988	0.6710
FF24	0.6375	0.7317	0.6091	0.7071	0.6504	0.7502	0.5999	0.7198	0.6626	0.7617	0.5975	0.7261
FF25	0.7067	0.6901	0.6155	0.6666	0.7027	0.6808	0.6102	0.6675	0.7049	0.6790	0.6078	0.6624
FF100BM	0.7096	0.6145	0.5369	0.6947	0.7057	0.6097	0.5370	0.6984	0.7065	0.6067	0.5372	0.6935
FF100OP	0.6614	0.7075	0.5538	0.6750	0.6644	0.7133	0.5611	0.6833	0.6612	0.7228	0.5653	0.6853
FF100INV	0.7242	0.7220	0.6134	0.7264	0.7240	0.7283	0.6256	0.7306	0.7259	0.7368	0.6294	0.7318

Panel B: CERs

Dataset	$\lambda = 1$				$\lambda = 3$				$\lambda = 5$			
	TW-ENet	OPL	C-ENet	RF	TW-ENet	OPL	C-ENet	RF	TW-ENet	OPL	C-ENet	RF
Ind10	0.0892	0.0844	0.0906	0.0808	0.0648	0.0643	0.0663	0.0553	0.0424	0.0443	0.0446	0.0328
FF21	0.0956	0.0970	0.0938	0.1035	0.0677	0.0679	0.0599	0.0704	0.0392	0.0406	0.0278	0.0394
FF23	0.0928	0.0955	0.0932	0.0971	0.0656	0.0701	0.0592	0.0684	0.0388	0.0462	0.0270	0.0418
FF24	0.0875	0.1018	0.0945	0.1065	0.0646	0.0797	0.0594	0.0787	0.0430	0.0580	0.0271	0.0516
FF25	0.1078	0.1001	0.0950	0.1038	0.0765	0.0713	0.0611	0.0714	0.0475	0.0443	0.0293	0.0388
FF100BM	0.1152	0.0987	0.0829	0.1178	0.0793	0.0617	0.0480	0.0793	0.0453	0.0252	0.0142	0.0403
FF100OP	0.1075	0.1133	0.0859	0.1143	0.0719	0.0805	0.0525	0.0764	0.0362	0.0483	0.0196	0.0388
FF100INV	0.1138	0.1128	0.0936	0.1201	0.0816	0.0819	0.0638	0.0845	0.0498	0.0522	0.0329	0.0492

Table A.8: SRs and CERs of the Transaction Cost Parameter of $\delta = 3 \times 10^{-8}$

Table A.8 reports the out-of-sample yearly Sharpe ratios (SRs) and certainty equivalent returns (CERs) of the generalized Bayes-Stein framework (GBS) and the naive equal weighted scheme (1/N) under various datasets (Ind10, FF21, FF23, FF24, FF25, FF100BM, FF100OP, and FF100INV) for different risk aversion parameters ($\lambda = 1, 3, 5$) considering transaction costs (TCs, a transaction cost parameter of $\delta = 3 \times 10^{-8}$). The results of the 20-year (the out-of-sample period covers from July 1983 to December 2021) and 40-year (the out-of-sample period covers from July 2003 to December 2021) expanding estimation windows are exhibited in Panel A and Panel B, respectively. The results of significance tests of the performance differences (GBS vs 1/N) are put in the parenthesis, where *, **, and *** indicate a statistical significance at the 10%, 5%, and 1%-level, respectively.

Panel A: SRs and CERs with TCs for a 20-year expanding estimation window

Dataset	SRs						CERs					
	$\lambda = 1$		$\lambda = 3$		$\lambda = 5$		$\lambda = 1$		$\lambda = 3$		$\lambda = 5$	
	GBS	1/N	GBS	1/N	GBS	1/N	GBS	1/N	GBS	1/N	GBS	1/N
Ind10	0.6664	0.6356	0.6605	0.6356	0.6547	0.6356	0.0892 (*)	0.0756	0.0648	0.058	0.0424	0.0403
FF21	0.6443 (**)	0.5226	0.6537 (**)	0.5226	0.6532 (**)	0.5226	0.0954 (**)	0.0760	0.0676 (**)	0.0455	0.0393 (**)	0.0150
FF23	0.6510 (**)	0.5050	0.6478 (**)	0.5050	0.6423 (**)	0.5050	0.0928 (**)	0.0688	0.0655 (**)	0.0425	0.0388 (**)	0.0161
FF24	0.6389 (*)	0.5155	0.6468 (**)	0.5155	0.6604 (**)	0.5155	0.0878 (**)	0.0676	0.0641 (**)	0.0438	0.0426 (**)	0.0200
FF25	0.7072 (***)	0.5497	0.7057 (***)	0.5497	0.7070 (***)	0.5497	0.1077 (***)	0.0783	0.0770 (***)	0.0500	0.0479 (***)	0.0217
FF100BM	0.7072 (***)	0.3816	0.7042 (***)	0.3816	0.7060 (***)	0.3816	0.1148 (***)	0.0518	0.0791 (***)	0.0206	0.0452 (***)	-0.0105
FF100OP	0.6585 (**)	0.5635	0.6616 (**)	0.5635	0.6588 (**)	0.5635	0.1069 (***)	0.0829	0.0714 (**)	0.0526	0.0357 (*)	0.0224
FF100INV	0.7244 (***)	0.5470	0.7240 (***)	0.5470	0.7261 (***)	0.5470	0.1138 (***)	0.0796	0.0816 (***)	0.0497	0.0498 (***)	0.0198

Panel B: SRs and CERs with TCs for a 40-year expanding estimation window

Dataset	SRs						CERs					
	$\lambda = 1$		$\lambda = 3$		$\lambda = 5$		$\lambda = 1$		$\lambda = 3$		$\lambda = 5$	
	GBS	1/N	GBS	1/N	GBS	1/N	GBS	1/N	GBS	1/N	GBS	1/N
Ind10	0.7613	0.7518	0.7531	0.7518	0.7524	0.7518	0.1091 (*)	0.0915	0.0819	0.0736	0.0565	0.0558
FF21	0.7104	0.6238	0.7285	0.6238	0.7432	0.6238	0.1088	0.0965	0.0810	0.0638	0.0540	0.0311
FF23	0.7426 (*)	0.6049	0.7522 (*)	0.6049	0.7558 (*)	0.6049	0.1130 (*)	0.0882	0.0846 (*)	0.0594	0.0563 (*)	0.0307
FF24	0.6761	0.6124	0.6827	0.6124	0.7008	0.6124	0.0994	0.0854	0.0719	0.0597	0.0479	0.0339
FF25	0.7795 (***)	0.6322	0.7873 (***)	0.6322	0.7989 (***)	0.6322	0.1252 (***)	0.0955	0.0931 (***)	0.0647	0.0630 (***)	0.0340
FF100BM	0.8551 (***)	0.5169	0.8582 (***)	0.5169	0.8577 (***)	0.5169	0.1500 (***)	0.0783	0.1108 (***)	0.0443	0.0721 (***)	0.0104
FF100OP	0.7396 (*)	0.6552	0.7435 (*)	0.6552	0.7468 (*)	0.6552	0.1208 (**)	0.1016	0.0864 (*)	0.0693	0.0526 (*)	0.0370
FF100INV	0.8037 (**)	0.6508	0.8024 (**)	0.6508	0.8086 (***)	0.6508	0.1253 (**)	0.1001	0.0945 (**)	0.0683	0.065 (***)	0.0366

Table A.9: SRs and CERs of the Transaction Cost Parameter of $\delta = 3 \times 10^{-6}$

Table A.9 reports the out-of-sample yearly Sharpe ratios (SRs) and certainty equivalent returns (CERs) of the generalized Bayes-Stein framework (GBS) and the naive equal weighted scheme (1/N) under various datasets (Ind10, FF21, FF23, FF24, FF25, FF100BM, FF100OP, and FF100INV) for different risk aversion parameters ($\lambda = 1, 3, 5$) considering transaction costs (TCs, a transaction cost parameter of $\delta = 3 \times 10^{-6}$). The results of the 20-year (the out-of-sample period covers from July 1983 to December 2021) and 40-year (the out-of-sample period covers from July 2003 to December 2021) expanding estimation windows are exhibited in Panel A and Panel B, respectively. The results of significance tests of the performance differences (GBS vs 1/N) are put in the parenthesis, where *, **, and *** indicate a statistical significance at the 10%, 5%, and 1%-level, respectively.

Panel A: SRs and CERs with TCs for a 20-year expanding estimation window

Dataset	SRs						CERs					
	$\lambda = 1$		$\lambda = 3$		$\lambda = 5$		$\lambda = 1$		$\lambda = 3$		$\lambda = 5$	
	GBS	1/N	GBS	1/N	GBS	1/N	GBS	1/N	GBS	1/N	GBS	1/N
Ind10	0.6664	0.6356	0.6604	0.6356	0.6547	0.6356	0.0892 (*)	0.0756	0.0648	0.0580	0.0424	0.0403
FF21	0.6446 (**)	0.5226	0.6541 (**)	0.5226	0.6532 (**)	0.5226	0.0955 (**)	0.0760	0.0677 (**)	0.0455	0.0393 (**)	0.0150
FF23	0.6489 (**)	0.5050	0.6489 (**)	0.5050	0.6432 (**)	0.5050	0.0924 (**)	0.0688	0.0656 (**)	0.0425	0.0390 (**)	0.0161
FF24	0.6344 (*)	0.5155	0.6462 (**)	0.5155	0.6608 (**)	0.5155	0.0874 (**)	0.0676	0.0641 (**)	0.0438	0.0426 (**)	0.0200
FF25	0.7070 (***)	0.5497	0.7031 (***)	0.5497	0.7050 (***)	0.5497	0.1079 (***)	0.0783	0.0766 (***)	0.0500	0.0475 (***)	0.0217
FF100BM	0.7079 (***)	0.3816	0.7039 (***)	0.3816	0.7047 (***)	0.3816	0.1148 (***)	0.0518	0.079 (***)	0.0206	0.0449 (***)	-0.0105
FF100OP	0.6588 (**)	0.5635	0.6620 (**)	0.5635	0.6594 (**)	0.5635	0.107 (***)	0.0829	0.0715 (**)	0.0526	0.0358 (**)	0.0224
FF100INV	0.7257 (***)	0.5470	0.7251 (***)	0.5470	0.7268 (***)	0.5470	0.114 (***)	0.0796	0.0818 (***)	0.0497	0.0500 (***)	0.0198

Panel B: SRs and CERs with TCs for a 40-year expanding estimation window

Dataset	SRs						CERs					
	$\lambda = 1$		$\lambda = 3$		$\lambda = 5$		$\lambda = 1$		$\lambda = 3$		$\lambda = 5$	
	GBS	1/N	GBS	1/N	GBS	1/N	GBS	1/N	GBS	1/N	GBS	1/N
Ind10	0.7613	0.7518	0.7531	0.7518	0.7524	0.7518	0.1091 (*)	0.0915	0.0819	0.0736	0.0565	0.0558
FF21	0.7110	0.6238	0.7289	0.6238	0.7410	0.6238	0.1089	0.0965	0.0811	0.0638	0.0536	0.0311
FF23	0.7458 (*)	0.6049	0.7558 (*)	0.6049	0.7608 (*)	0.6049	0.1136 (*)	0.0882	0.0853 (*)	0.0594	0.0572 (*)	0.0307
FF24	0.6780	0.6124	0.6824	0.6124	0.7004	0.6124	0.0997	0.0854	0.0718	0.0597	0.0478	0.0339
FF25	0.7802 (***)	0.6322	0.7869 (***)	0.6322	0.7995 (***)	0.6322	0.1253 (***)	0.0955	0.0930 (***)	0.0647	0.0631 (***)	0.0340
FF100BM	0.8543 (***)	0.5169	0.8581 (***)	0.5169	0.8582 (***)	0.5169	0.1500 (***)	0.0783	0.1108 (***)	0.0443	0.0722 (***)	0.0104
FF100OP	0.7374 (*)	0.6552	0.7428 (*)	0.6552	0.7461 (*)	0.6552	0.1203 (**)	0.1016	0.0862 (*)	0.0693	0.0524 (*)	0.0370
FF100INV	0.8035 (**)	0.6508	0.8027 (**)	0.6508	0.8089 (**)	0.6508	0.1253 (**)	0.1001	0.0945 (**)	0.0683	0.065 (***)	0.0366

Table A.10: SRs and CERs of a 30-year Expanding Window

Table A.10 reports the out-of-sample yearly Sharpe ratios (SRs) and certainty equivalent returns (CERs) of the generalized Bayes-Stein framework (GBS) and the naive equal weighted scheme (1/N) under various datasets (Ind10, FF21, FF23, FF24, FF25, FF100BM, FF100OP, and FF100INV) for different risk aversion parameters ($\lambda = 1, 3, 5$) considering transaction costs (TCs, a transaction cost parameter of $\delta = 3 \times 10^{-6}$). The results are based on a 30-year expanding estimation window (the out-of-sample period covers from July 1993 to December 2021). Significance tests of the performance differences (GBS vs 1/N) are put in the parenthesis, where *, **, and *** indicate a statistical significance at the 10%, 5%, and 1%-level, respectively.

Dataset	SRs						CERs					
	$\lambda = 1$		$\lambda = 3$		$\lambda = 5$		$\lambda = 1$		$\lambda = 3$		$\lambda = 5$	
	GBS	1/N	GBS	1/N	GBS	1/N	GBS	1/N	GBS	1/N	GBS	1/N
Ind10	0.6975	0.6685	0.6866	0.6685	0.6803	0.6685	0.0969 (**)	0.0788	0.0703	0.0618	0.0456	0.0447
FF21	0.6961 (**)	0.5749	0.7102 (**)	0.5749	0.7161 (**)	0.5749	0.1054 (*)	0.0864	0.0775 (**)	0.0548	0.0496 (**)	0.0232
FF23	0.7320 (**)	0.5597	0.7269 (**)	0.5597	0.7211 (**)	0.5597	0.1078 (**)	0.0790	0.0790 (**)	0.0516	0.0511 (**)	0.0241
FF24	0.6867	0.5725	0.7003 (*)	0.5725	0.7206 (*)	0.5725	0.0980 (*)	0.0776	0.0735 (*)	0.0529	0.0515 (*)	0.0282
FF25	0.7395 (***)	0.5919	0.7432 (***)	0.5919	0.7502 (***)	0.5919	0.1152 (***)	0.0865	0.0841 (***)	0.0573	0.0549 (***)	0.0282
FF100BM	0.7440 (***)	0.3679	0.7429 (***)	0.3679	0.7458 (***)	0.3679	0.1247 (***)	0.0500	0.0872 (***)	0.0177	0.0515 (***)	-0.0146
FF100OP	0.7579 (***)	0.6230	0.7606 (***)	0.6230	0.7568 (***)	0.6230	0.1255 (***)	0.0936	0.0899 (***)	0.0631	0.0542 (**)	0.0325
FF100INV	0.7951 (***)	0.6376	0.7926 (***)	0.6376	0.7920 (***)	0.6376	0.1255 (***)	0.0958	0.0935 (***)	0.0655	0.0618 (***)	0.0352

Table A.11: Out-of-sample Portfolio Performance of the Rolling Window

Table A.11 reports the out-of-sample yearly Sharpe ratios (SRs) and certainty equivalent returns (CERs) of the generalized Bayes-Stein framework (GBS) and the naive equal weighted scheme (1/N) under various datasets (Ind10, FF21, FF23, FF24, FF25, FF100BM, FF100OP, and FF100INV) for different risk aversion parameters ($\lambda = 1, 3, 5$) considering transaction costs (TCs, a transaction cost parameter of $\delta = 3 \times 10^{-6}$). The results are based on a 20-year rolling estimation window (the out-of-sample period covers from July 1983 to December 2021). Significance tests of the performance differences (GBS vs 1/N) are put in the parenthesis, where *, **, and *** indicate a statistical significance at the 10%, 5%, and 1%-level, respectively.

Dataset	SRs						CERs					
	$\lambda = 1$		$\lambda = 3$		$\lambda = 5$		$\lambda = 1$		$\lambda = 3$		$\lambda = 5$	
	GBS	1/N	GBS	1/N	GBS	1/N	GBS	1/N	GBS	1/N	GBS	1/N
Ind10	0.6713	0.6356	0.6683	0.6356	0.6657	0.6356	0.0889 (*)	0.0756	0.0654	0.0580	0.0441	0.0403
FF21	0.6467 (**)	0.5226	0.6571 (**)	0.5226	0.6543 (**)	0.5226	0.0960 (**)	0.0760	0.0682 (**)	0.0455	0.0395 (**)	0.0150
FF23	0.6530 (**)	0.5050	0.6542 (**)	0.5050	0.6467 (**)	0.5050	0.0931 (**)	0.0688	0.0664 (**)	0.0425	0.0395 (**)	0.0161
FF24	0.6301 (*)	0.5155	0.6465 (**)	0.5155	0.6585 (**)	0.5155	0.0868 (**)	0.0676	0.0641 (**)	0.0438	0.0423 (**)	0.0200
FF25	0.7111 (***)	0.5497	0.7055 (***)	0.5497	0.7084 (***)	0.5497	0.1086 (***)	0.0783	0.0770 (***)	0.0500	0.0481 (***)	0.0217
FF100BM	0.7258 (***)	0.3816	0.7233 (***)	0.3816	0.7238 (***)	0.3816	0.1176 (***)	0.0518	0.0823 (***)	0.0206	0.0487 (***)	-0.0105
FF100OP	0.6618 (**)	0.5635	0.6676 (**)	0.5635	0.6664 (**)	0.5635	0.1073 (***)	0.0829	0.0725 (***)	0.0526	0.0373 (**)	0.0224
FF100INV	0.7373 (***)	0.5470	0.7351 (***)	0.5470	0.7354 (***)	0.5470	0.1156 (***)	0.0796	0.0833 (***)	0.0497	0.0518 (***)	0.0198

References

DeMiguel, V., Garlappi, L. and Uppal, R., 2009. Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *The review of financial studies*, 22(5), pp.1915–1953.

DeMiguel, V., Martín-Utrera, A. and Nogales, F.J., 2015. Parameter uncertainty in multiperiod portfolio optimization with transaction costs. *Journal of financial and quantitative analysis*, 50(6), pp.1443–1471.

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

Rapach, D.E., Strauss, J.K., Tu, J. and Zhou, G., 2019. Industry return predictability: A machine learning approach. *The journal of financial data science*, 1(3), pp.9–28.

Rapach, D.E. and Zhou, G., 2020. Time-series and cross-sectional stock return forecasting: New machine learning methods. *Machine learning for asset management: New developments and financial applications*, pp.1–33.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. *Journal of the royal statistical society: Series b (methodological)*, 58(1), pp.267–288.

Welch, I. and Goyal, A., 2008. A comprehensive look at the empirical performance of equity premium prediction. *The review of financial studies*, 21(4), pp.1455–1508.