informs
http://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–30
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Distributionally Robust Mean-Variance Portfolio Selection with Wasserstein Distances

Jose Blanchet,[a] Lin Chen,[b] Xun Yu Zhou[b]

[a] Department of Management Science and Engineering, Stanford University, Stanford, California 94305; [b] Department of Industrial Engineering and Operations Research, Columbia University, New York, New York 10027

Contact: jblanche@stanford.edu, https://orcid.org/0000-0001-5895-0912 (JB); lc3110@columbia.edu (LC); xz2574@columbia.edu, https://orcid.org/0000-0001-9908-5697 (XYZ)

Received: July 1, 2018
Revised: October 14, 2019; May 15, 2020; August 20, 2020
Accepted: September 7, 2020
Published Online in Articles in Advance: ■■■■ ■■, 2021

https://doi.org/10.1287/mnsc.2021.4155

Abstract. We revisit Markowitz's mean-variance portfolio selection model by considering a distributionally robust version, in which the region of distributional uncertainty is around the empirical measure and the discrepancy between probability measures is dictated by the Wasserstein distance. We reduce this problem into an empirical variance minimization problem with an additional regularization term. Moreover, we extend the recently developed inference methodology to our setting in order to select the size of the distributional uncertainty as well as the associated robust target return rate in a data-driven way. Finally, we report extensive back-testing results on S&P 500 that compare the performance of our model with those of several well-known models including the Fama–French and Black–Litterman models.

History: Accepted by David Simchi-Levi, finance.
Funding: Material in this paper is based upon work supported by the Air Force Office of Scientific Research [Grant FA9550-20-1-0397]. Additional support is gratefully acknowledged from the National Science Foundation [Grants 1915967, 1820942, and 1838576].
Supplemental Material: The online appendix is available at https://doi.org/10.1287/mnsc.2021.4155.

Keywords: mean-variance portfolio selection • robust model • Wasserstein distance • robust Wasserstein profile inference

1. Introduction

We study data-driven mean–variance portfolio selection with model uncertainty (or ambiguity). The classical one-period Markowitz mean–variance model (Markowitz 1952) is to choose a portfolio weighting vector $\phi \in \mathbb{R}^d$ (all the vectors in this paper are by convention columns) among d stocks to maximize the risk-adjusted expected return. The precise formulation is[1]

$$\min_{\phi \in \mathbb{R}^d} \{\phi^\top \mathrm{Var}_{\mathbb{P}^*}(R)\phi : \phi^\top 1 = 1, \phi^\top \mathbb{E}_{\mathbb{P}^*}(R) = \rho\}, \qquad (1)$$

where R is the d-dimensional vector of random returns of the stocks; $\mathbb{P}^*$ is the probability measure underlying the distribution of R; $\mathbb{E}_{\mathbb{P}^*}$ and $\mathrm{Var}_{\mathbb{P}^*}$ are, respectively, the expectation and variance under $\mathbb{P}^*$; and ρ is the targeted expected return of the portfolio.

It is well known that this model has a major drawback when applied in practice. On one hand, its solutions are very sensitive to the underlying parameters, namely, the mean and the covariance matrix of the stocks. On the other hand, $\mathbb{E}_{\mathbb{P}^*}$ is unknown in practice, so one has to resort to the empirical versions of the mean and the covariance matrix instead, which are usually significantly deviated from the true ones (especially the mean because of the notorious "mean-blur" problem).

This motivates the development of the "robust" formulation of the Markowitz model, which recognizes and tries to account for the impact of the (potentially significant) discrepancies between $\mathbb{P}^*$ and its empirical version. This idea originates in the robust control approach from control theory (see, for example, Petersen et al. 2000). Hansen and Sargent (2008) give a systematic account on applications of robust control to economic models. There is also a rich literature on robustification of portfolio choice. The paper by Lobo and Boyd (2000) is among the first to provide a worst-case robust analysis with respect to the second-order moment uncertainty within the Markowitz framework. Goldfarb and Iyengar (2003) consider a robust Markowitz problem with the uncertainty set based on vector/matrix distance. Pflug and Wozabal (2007) present a Markowitz model with distributional robustness based on a Wasserstein distance, a metric measuring the discrepancy between two probability measures that we also apply in this paper.[2] Nevertheless, their formulation involves an additional value-at-risk type of constraint that leads to a much more complex optimization problem. More importantly, their choice of the uncertainty size is exogenous, and no guidance for optimally selecting the size is given.

Esfahani and Kuhn (2018) provide representations for the worst-case expectations in a Wasserstein-based

Q:1

ambiguity set centered at the empirical measure and then apply their results to portfolio selection using different risk measures, leading to models different from the Markowitz model. An important difference in our approach, which is related to the work in Blanchet et al. (2016), as we discuss, is that we focus on the order-two Wasserstein distance. This is important because, as a result of the quadratic nature of the variance objective that we consider, applying an uncertainty set based on Wasserstein of order one could potentially lead to arbitrarily large variances. We also note that the work in Esfahani and Kuhn (2018) provides guidance to choose the uncertainty size, δ. But this choice of the uncertainty size deteriorates substantially with an increase in the dimension of the underlying portfolio. So, as we elaborate, we employ an approach similar to that proposed in Blanchet et al. (2016), which must be adapted and extended to our setting. Our current work relates to the broad literature on distributionally robust optimization (DRO). In addition to Esfahani and Kuhn (2018), related duality results for Wasserstein DRO formulations in which the probability model appears linearly in the objective function are studied in Zhao and Guan (2018), Esfahani and Kuhn (2018), and Gao and Kleywegt (2016). A general result (with conditions that match the standard assumptions of the general optimal transport theory) is given in Blanchet and Murthy (2019). These results are not directly applicable to our setting because our objective function is not linear, but quadratic in the underlying probability model, so the techniques need to be adapted.

Also, in connection to robust portfolio optimization, we mention the work in Delage and Ye (2010), which constructs uncertainty regions involving means and covariances of the return vector. The paper of Wozabal (2012) also considers a robust portfolio model with risk constraints based on expected shortfall, resulting in an optimization problem that requires solving multiple convex problems. These papers do not consider an optimal choice of the size of the uncertainty sets as we do here.

Then, there is a number of papers that address a wide range of optimization techniques (such as interior-point methods, conic programming, and linear matrix inequalities) in solving robust portfolio selection problems. A selection of these includes Halldorsson and Tutuncu (2003), Costa and Paiva (2002), and Ghaoui et al. (2003). We obtain formulations that can be solved with basically any standard convex optimization software. Finally, we mention the work of Goh and Sim (2010) and Wiesemann et al. (2014), who investigate different forms of distributional ambiguity sets, as well as those of Hu and Hong (2013) and Jiang and Guan (2016), who study distributional robust formulations based on the Kullback–Leibler divergences. It is worth noting that the Kullback–Leibler divergence-based formulation is popular in economics (see Hansen and Sargent 2008). Our formulation focuses on the use of Wasserstein ambiguity sets because of the intuitive out-of-sample exploration induced by the Wasserstein distance and because of the regularization interpretation which, as we see, results from our formulation.

Precisely, in this paper, we are interested in studying a distributionally robust mean–variance (DRMV) model given by

$$\min_{\phi\in\mathcal{F}_{\delta,\bar{\alpha}}(n)} \max_{\mathbb{P}\in\mathcal{U}_\delta(\mathbb{P}_n)} \{\phi^\top \mathrm{Var}_{\mathbb{P}}(R)\phi\}, \tag{2}$$

where $\mathbb{P}_n$ is the empirical probability derived from historical information of the sample size n; $\mathcal{U}_\delta(\mathbb{P}_n) := \{\mathbb{P} : D_c(\mathbb{P},\mathbb{P}_n) \le \delta\}$ is the ambiguity set;

$$\mathcal{F}_{\delta,\bar{\alpha}}(n) = \left\{\phi : \phi^\top 1 = 1, \min_{\mathbb{P}\in\mathcal{U}_\delta(\mathbb{P}_n)} [\mathbb{E}_{\mathbb{P}}(\phi^\top R)] \ge \bar{\alpha}\right\},$$

is the feasible region of portfolios; $\mathbb{E}_{\mathbb{P}}$ and $\mathrm{Var}_{\mathbb{P}}(R)$ denote, respectively, the mean and the covariance matrix under $\mathbb{P}$; and $D_c(\cdot,\cdot)$ is a notion of discrepancy between two probability measures based on a suitably defined Wasserstein distance.[3]

Intuitively, Formulation (2) introduces an artificial adversary $\mathbb{P}$ (whose problem is that of the inner maximization) as a tool to account for the impact of the model uncertainty around the empirical distribution. There are two key parameters, δ and $\bar{\alpha}$, in this formulation, and they need to be carefully chosen. The parameter δ can be interpreted as the power given to the adversary: the larger the value of δ, the more power is given. If δ is too large relative to the evidence (i.e., the size of n), then the portfolio selection tends to be unnecessarily conservative. On the other hand, $\bar{\alpha}$ can be regarded as the lowest acceptable target return given the ambiguity set. Naturally, the choice of $\bar{\alpha}$ should be based on the original target ρ given in (1), but one also needs to take into account the size of the distributional uncertainty, δ. Using $\bar{\alpha} = \rho$ tends to generate portfolios that are too aggressive; it is more sensible to choose $\bar{\alpha} < \rho$ in a way such that $\rho - \bar{\alpha}$ is naturally informed by δ.

This paper makes three main contributions. First, we show that (2) is equivalent to an (explicitly formulated) nonrobust minimization problem in terms of the empirical probability measure in which a proper penalty term or "regularization term" is added to the objective function. The explicit regularization term that is derived from (2) is given in Theorem 1. This connects (and contrasts) to the *directly introduced* use of regularization in variance minimization techniques widely employed in both the statistics/machine learning literature and practice to, among others, address the issue of overfitting. Indeed, practitioners who use

mean–variance portfolio selection models often introduce regularization penalties, inspired by Lasso, in order to enhance the sparsity so as to include fewer stocks in their portfolios. Our use of Wasserstein distance to model distributional uncertainty naturally gives rise to a regularization term, suggesting an alternative, yet theoretical, justification and interpretation for its use in practice. Our result shows that our robust strategies are able to enhance out-of-sample performance with basically the same level of computational tractability as the standard mean–variance selection.

Our second main contribution provides guidance on the choice of the size of the ambiguity set, δ, as well as that of the worst mean return target, $\bar{\alpha}$. This is accomplished by adapting and extending the robust Wasserstein profile inference (RWPI) framework, recently introduced and developed by Blanchet et al. (2016), to our setting in a data-driven way that combines optimization principles and basic statistical theory under suitable mixing conditions on historical data.

The last contribution empirically compares the performance of our DRMV strategies with those of several well-known and well-practiced models, including the classical Markowitz, Fama–French, and Black–Litterman models. We also compare our strategies with those of another robust model, the one put forward by Goldfarb and Iyengar (2003) in which robustness is based on vector/matrix distance. All these models (including ours) are static, single-period ones, whereas in practice, a stock market is highly dynamic. In our empirical experiments, we implement them in the same rolling horizon fashion to account for the market dynamics. It should be noted that our theory applies to only a one-period model, and the numerical implementation to the multiperiod market is heuristic based on rolling horizons. Finally, we also include in our comparison precommitted optimal strategies based on a well-calibrated, nonrobust continuous-time model in Cui et al. (2012). The experiments are carried out on S&P 500 for the backtesting period 2000–2017 with the prior 10 years as the training period. Our experiments show that DRMV compares favorably against all other models in achieving no worse (far better against most models) average returns and much lower variabilities. This, we believe, is another important insight that we can draw from this research and that merits further investigation.

We should acknowledge, however, that our approach fails when the number of stocks d is not small compared with the sample size n because the inference method for choosing δ and $\bar{\alpha}$ theoretically relies on some central limit theorems that require n to asymptotically approach infinity. Moreover, when $d > n$, both the empirical probability measure and the plug-in portfolios are not consistent. The application in our mind is asset allocation in which relatively small portfolios (i.e., those having dozens of assets) are desired. Indeed, one of the main goals of the regularization technique is to reduce the number of stocks in a portfolio. In our empirical study, we have 100 stocks and only 108 data points, and our algorithm achieves good performance.

Ao et al. (2018) also provide a theoretical justification on the L^1-norm regularization by proving that the regularized portfolios asymptotically achieve the optimal mean and variance when the size of the portfolio approaches infinity. However, their result relies on the normality of return distribution, whereas we only require the data to be stationary, which, in particular, allows heavy tail distributions commonly seen in financial data. Ao et al. (2018) estimate the regularization coefficient—δ in our setting—using the cross-validation technique, and we use the statistical inference to derive this parameter in a data-driven way.[4] Moreover, all the results in Ao et al. (2018) seems to work only with the L^1-norm, whereas our results can be applied to any L^p-norm by choosing a proper Wasserstein distance. A significant contribution of Ao et al. (2018), on the other hand, is that it allows a large size of the underlying portfolio even though it must be of the same order of the sample data size.

The rest of the paper is organized as follows: In Section 2, we formulate the DRMV model and present necessary preliminaries. Section 3 demonstrates the tractability of our DRMV model after a series of transformations, and Section 4 studies the choices of distributional uncertainty size and the worst return level. Then, in Section 5, we report the empirical performance of our strategies against those of several other models. Concluding remarks are given in Section 6. Technical proofs of our results are given in appendices at the end of the paper.

2. Model Formulation

In this section, we formulate our distributionally robust Markowitz model while reviewing some useful concepts.

Let $\mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ be the space of all Borel probability measures supported on $\mathbb{R}^d \times \mathbb{R}^d$. A given element $\pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ can be assumed to be the joint distribution of a random vector (U, V), where $U \in \mathbb{R}^d$ and $V \in \mathbb{R}^d$. We use π_U and π_V to denote the marginal distributions of U and V under π. In particular, $\pi_U(A) = \pi(A \times \mathbb{R}^d)$ and $\pi_V(A) = \pi(\mathbb{R}^d \times A)$ for every Borel set $A \subset \mathbb{R}^d$.

We start with a "cost" function $c : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty]$, which we assume to be lower semicontinuous and such that $c(u, u) = 0$ for any $u \in \mathbb{R}^d$. For a given such cost function c, we introduce $D_c(\cdot, \cdot)$ representing

some "discrepancy" between two probability measures as follows:

$$D_c(\mathbb{P},\mathbb{Q}) := \inf\{\mathbb{E}_\pi[c(U,V)] : \pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d), \pi_U = \mathbb{P}, \pi_V = \mathbb{Q}\},$$

where $\mathbb{P}$ and $\mathbb{Q}$ are two probability measures supported on $\mathbb{R}^d$. This can be interpreted as the optimal (minimal) transportation cost (also known as the optimal transport discrepancy or the Wasserstein discrepancy) of moving the mass from $\mathbb{P}$ into the mass of $\mathbb{Q}$ under a cost $c(x,y)$ per unit of mass transported from x to y.

If, for a given $p > 0$, $c^{1/p}(\cdot,\cdot)$ is a metric, then so is $D_c^{1/p}(\cdot,\cdot)$; see Villani (2003). Such a metric $D_c^{1/p}(\cdot,\cdot)$ is known as a Wasserstein distance of order p. Most of the times in this paper, we choose the following cost function:

$$c(u,v) = \|u - v\|_q^2,$$

where $q \geq 1$ is fixed (which leads to a Wasserstein distance of order two).[5]

Recall that R is the d-dimensional vector of random returns of the d stocks. Let $\mathbb{P}_n$ be the empirical probability measure on $\mathbb{R}^d$ with a sample size n, that is,

$$\mathbb{P}_n(dx) = \frac{1}{n}\sum_{i=1}^n \delta_{R_i}(dx),$$

where R_i $(i = 1,2,\ldots,n)$ are realizations of R and $\delta_{R_i}(\cdot)$ is the indicator function. Define the ambiguity set as

$$\mathcal{U}_\delta(\mathbb{P}_n) = \{\mathbb{P} : D_c(\mathbb{P},\mathbb{P}_n) \leq \delta\},$$

and the feasible region of portfolios as

$$\mathcal{F}_{\delta,\bar{\alpha}}(n) = \left\{\phi \in \mathbb{R}^d : \phi^\top 1 = 1, \min_{\mathbb{P}\in\mathcal{U}_\delta(\mathbb{P}_n)} [\mathbb{E}_{\mathbb{P}}(\phi^\top R)] \geq \bar{\alpha}\right\}.$$

The DRMV approach then consists in choosing a portfolio $\phi \in \mathcal{F}_{\delta,\bar{\alpha}}(n)$, which achieves the optimal min-max value in (2).

3. Transformations, Duality, and Regularization

Problem (2) appears, in principle, very complex. First of all, the inner maximization problem is over a set of probability measures, which renders it an infinite dimensional optimization problem. Second, it is not clear whether the outer minimization problem, although finite dimensional, is convex. Therefore, (2) at its outset seems computationally insurmountable. In this section, we reformulate (2), through a series of transformations and a duality argument, as an equivalent problem that is computationally tractable.

The following theorem, whose proof is relegated to Appendix A, is the main result of the paper, which states that (2) is equivalent to a nonrobust portfolio selection problem in terms of the empirical measure $\mathbb{P}_n$ with an additional regularization term.

Theorem 1. *The primal formulation given in (2) is equivalent to the following dual problem:*

$$\begin{cases} \min\limits_{\phi\in\mathbb{R}^d} & \sqrt{\phi^\top \mathrm{Var}_{\mathbb{P}_n}(R)\phi} + \sqrt{\delta}\|\phi\|_p, \\ \text{subject to} & \phi^\top 1 = 1,\ \mathbb{E}_{\mathbb{P}_n}(\phi^\top R) \geq \bar{\alpha} + \sqrt{\delta}\|\phi\|_p, \end{cases} \tag{3}$$

in the sense that the two problems have the same optimal solutions and optimal value.

It is not hard to verify that the mapping $\phi \to \sqrt{\phi^\top \mathrm{Var}_{\mathbb{P}_n}(R)\phi} + \sqrt{\delta}\|\phi\|_p$ is convex, and the feasible region of (3) is clearly convex. So (3) and, therefore, (2) are both convex optimization problems. As such, they are tractable optimization problems.

Problem (3) has an additional term, $\sqrt{\delta}\|\phi\|_p$, in its objective function. In the asset management industry, fund managers using a mean–variance portfolio selection model often add a "penalty" or regularization term—in the form of $k\|\phi\|$, where $\|\cdot\|$ is an appropriately chosen norm—in order to enhance the sparsity of the vector as a way to include fewer stocks in the portfolios and to address the issue of overfitting.[6] Here, we provide *interpretability* of this regularization technique (which is based on experience or heuristics) by a well-established robustification idea backed by precise rationality principles; see, for example, Delage et al. (2019). Moreover, the parameter δ that reflects the level of regularization is also endogenously informed by data as we show in the next section.

Theorem 1 has another interesting interpretation related to transactions costs. Introducing the Lagrange multiplier to the second constraint in Problem (3), the latter is equivalent to

$$\begin{cases} \min\limits_{\phi\in\mathbb{R}^d} & \gamma\sqrt{\phi^\top \mathrm{Var}_{\mathbb{P}_n}(R)\phi} - \mathbb{E}_{\mathbb{P}_n}(\phi^\top R) + k\|\phi\|_p, \\ \text{subject to} & \phi^\top 1 = 1 \end{cases} \tag{4}$$

for some constants $\gamma > 0$ and $k \geq 0$. Following Olivares-Nadal and DeMiguel (2018), we can regard the term in the objective function, $k\|\phi\|_p$, as a transaction cost term. Therefore, (4) is a classical mean–variance model with transaction costs.[7] With this interpretation, Theorem 1 yields that a mean–variance model with transaction costs in the form of (4) is equivalent to a *distributionally* robust mean–variance model in the form of (2). This result is related to one of the results in Olivares-Nadal and DeMiguel (2018), which states that a mean–variance portfolio problem with L^p-norm transaction costs is equivalent to a robust optimization problem. However, there are important differences between the two results. The robust problem in Olivares-Nadal and

DeMiguel (2018) has an *ellipsoidal* uncertainty set around the sample mean *only*, and there is no robustification on the variance. The distributionally robust model in our paper is the most comprehensive one because it robustifies not only the mean but also the variance and indeed the whole distribution. Methodologically, the result of Olivares-Nadal and DeMiguel (2018) follows directly (and indeed trivially) from what is essentially the Legendre transformation or the convex conjugate (see the online companion of Olivares-Nadal and DeMiguel 2018). This implication is actually well documented in earlier papers, such as Bertsimas et al. (2004) and, in a portfolio selection context, Gotoh and Takeda (2011), which shows that any norm constraint can be turned into a robust constraint associated with the return vector. In contrast, the proof of Theorem 1 is much more involved and subtle because of the distributional constraint (see Appendix A). In turn, the distributional formulation is key in providing a natural statistical approach to selecting the size of the uncertainty. These constitute one of the main methodological contributions of this paper.

As indicated earlier, it should be noted that Theorem 1 does not directly follow from any of the strong duality results mentioned in the introduction. This is because the portfolio variance in the objective function (2) is *not* a linear function of the probability measure. A related work, Gao et al. (2017) proves only an asymptotic equivalence to regularization. On the other hand, we have the exact equivalence between (2) and a regularized optimization problem given in Theorem 1.

To conclude this section, we note that, although the cost function is chosen as $c(u,v) = \|u-v\|_q^2$ in the study here, our result (Theorem 1) actually holds for any cost function of the form $c(u,v) = \|u-v\|^2$, where $\|\cdot\|$ is any given norm with a suitable dual. More precisely, define the dual norm as $\|x\|_* = \sup_{\|z\|=1}|x^\top z|$. Then, the primal distributionally robust model under this alternative cost function is equivalent to the following dual problem:

$$\min_{\phi\in\mathcal{F}_{\delta,\bar{\alpha}}(n)}\left(\sqrt{\phi^\top \mathrm{Var}_{\mathbb{P}_n}(R)\phi}+\sqrt{\delta}\|\phi\|_*\right)^2,$$

where the feasible region is modified as

$$\mathcal{F}_{\delta,\bar{\alpha}}(n)=\left\{\phi\in\mathbb{R}^d:\phi^\top 1=1,\mathbb{E}_{\mathbb{P}_n}(\phi^\top R)\geq\bar{\alpha}+\sqrt{\delta}\|\phi\|_*\right\}.$$

For example, consider a norm as $\|x\| = (x^T\Sigma x)^{1/2}$, where Σ is a strictly positive definite matrix. Then, $\|x\|_* = (x^T\Sigma^{-1}x)^{1/2}$. Interested readers may refer to Blanchet and Kang (2017) for discussions on some other interesting norms.

4. Choice of Model Parameters

There are two key parameters, δ and $\bar{\alpha}$, in Formulation (2), the choice of which is not only curious in theory, but also crucial in practical implementation and for the success of our algorithm. The idea is that the choice of these parameters should be informed by the data (i.e., in a data-driven way) based on some statistical principles rather than being arbitrarily exogenous. Specifically, we define the distributional uncertainty region just large enough so that the correct optimal portfolio (the one that we would apply if the underlying distribution was known) becomes a plausible choice with a sufficiently high confidence level. Once this is determined, then we determine the feasible set of portfolios just large enough so that the correct optimal portfolio is feasible with adequately high confidence.

We need to impose several technical/statistical assumptions.

Assumption 1. *The underlying return time series* $(R_k : k\geq 0)$ *is a stationary, ergodic process satisfying* $\mathbb{E}_{\mathbb{P}^*}\left(\|R_k\|_2^4\right)<\infty$ *for each* $k\geq 0$. *Moreover, for each measurable* $g(\cdot)$ *such that* $|g(x)|\leq c(1+\|x\|_2^2)$ *for some* $c>0$, *the limit*

$$\Upsilon_g := \lim_{n\to\infty}\mathrm{Var}_{\mathbb{P}^*}\left(n^{-1/2}\sum_{k=1}^n g(R_k)\right)$$

exists and the central limit theorem holds:

$$n^{1/2}\left[\mathbb{E}_{\mathbb{P}_n}(g(R))-\mathbb{E}_{\mathbb{P}^*}(g(R))\right]\Rightarrow N\left(0,\Upsilon_g\right),$$

where (and henceforth) "⇒" denotes weak convergence.

Assumption 2. *For any matrix* $\Lambda\in\mathbb{R}^{d\times d}$ *and any vector* $\zeta\in\mathbb{R}^d$ *such that either* $\Lambda\neq 0$ *or* $\zeta\neq 0$,

$$\mathbb{P}^*(\|\Lambda R+\zeta\|_2>0)>0.$$

Assumption 3. *The classical model* (1) *has a unique solution* ϕ^*. *Moreover,* $\mathrm{Var}_{\mathbb{P}^*}[R]$ *is positive definite.*

Assumption 1 is standard for most time series models (after removing seasonality). Assumption 2 holds, assuming, for example, that R has a density. Assumption 3 is a technical assumption that can be relaxed, but then the evaluation of the optimal choice of δ becomes more cumbersome as we explain.

4.1. Choice of δ

The choice of the uncertainty size δ is crucial. If δ is too large, then there is too much model ambiguity, and the available data becomes less relevant. In this case, the resulting optimal portfolios tend to be just equal allocations. If δ is too small, then the effect of robustification is negligible. Therefore, the choice of δ

should *not* be exogenously specified; rather, it should be endogenously informed by the data.

Theorem 1 actually suggests an appropriate order of $\delta = \delta_n$ (here, n is the size of the available return time series data) in terms of n. Because the differences between the optimal standard deviation by solving (1) and that obtained by solving the empirical version of (1) are of order $O(n^{-1/2})$, it follows from Theorem 1 that any choice of δ_n in the order of $o(n^{-1})$ would be too small. Hence, an "optimal" order of δ_n should be of order $O(n^{-1})$.

In order to choose an appropriate δ_n, here we follow the idea behind the RWPI approach introduced in Blanchet et al. (2016). Intuitively, δ should be chosen such that the set $\mathcal{U}_\delta(\mathbb{P}_n) = \{\mathbb{P} : D_c(\mathbb{P}, \mathbb{P}_n) \le \delta\}$ contains all the probability measures that are plausible variations of the data represented by $\mathbb{P}_n$. Denote by $\mathcal{Q}(\mathbb{P})$ the classical Markowitz portfolio selection problem with target return ρ, assuming that $\mathbb{P}$ is the underlying model:

$$\begin{aligned} &\min_{1^\top \phi = 1} \ \phi^\top \mathbb{E}_{\mathbb{P}}[RR^\top]\phi \\ &\text{subject to} \quad \phi^\top \mathbb{E}_{\mathbb{P}}[R] = \rho, \end{aligned} \tag{5}$$

and by $\phi_{\mathbb{P}}$ a solution to $\mathcal{Q}(\mathbb{P})$ and $\Phi_{\mathbb{P}}$ the set of all such solutions. According to Assumption 3, we have $\Phi_{\mathbb{P}^*} = \{\phi^*\}$ for some portfolio ϕ^*. Therefore, there exist (unique) Lagrange multipliers λ_1^* and λ_2^* such that

$$\begin{aligned} 2\mathbb{E}_{\mathbb{P}^*}(RR^\top)\phi^* - \lambda_1^* \mathbb{E}_{\mathbb{P}^*}[R] - \lambda_2^* 1 = 0, \\ (\phi^*)^\top \mathbb{E}_{\mathbb{P}^*}[R] - \rho = 0. \end{aligned} \tag{6}$$

Now, when δ is suitably chosen so that $\mathcal{U}_\delta(\mathbb{P}_n)$ constitutes the models that are plausible variations of $\mathbb{P}_n$, any $\phi_{\mathbb{P}}$ with $\mathbb{P} \in \mathcal{U}_\delta(\mathbb{P}_n)$ is a plausible estimate of ϕ^*. This intuition motivates the definition of the following set

$$\Lambda_\delta(\mathbb{P}_n) = \cup_{\mathbb{P} \in \mathcal{U}_\delta(\mathbb{P}_n)} \Phi_{\mathbb{P}},$$

which corresponds to all the plausible estimates of ϕ^*. As a result, $\Lambda_\delta(\mathbb{P}_n)$ is a natural confidence region for ϕ^*, and therefore, δ should be chosen as the smallest number δ_n^* such that ϕ^* belongs to this region with a given confidence level. Namely,

$$\delta_n^* = \min\{\delta > 0 : \mathbb{P}^*(\phi^* \in \Lambda_\delta(\mathbb{P}_n)) \ge 1 - \delta_0\},$$

where $1 - \delta_0$ is a user-defined confidence level (typically 95%).

However, by the mere definition, it is difficult to compute δ_n^*. We now provide a simpler representation for δ_n^* via an auxiliary function called the robust Wasserstein profile (RWP) function. To this end, first observe that any $\phi \in \Lambda_\delta(\mathbb{P}_n)$ if and only if there exist $\mathbb{P} \in \mathcal{U}_\delta(\mathbb{P}_n)$ and $\lambda_1, \lambda_2 \in (-\infty, \infty)$ such that

$$\begin{aligned} 2\mathbb{E}_{\mathbb{P}}(RR^\top)\phi - \lambda_1 \mathbb{E}_{\mathbb{P}}[R] - \lambda_2 1 = 0, \\ \phi^T \mathbb{E}_{\mathbb{P}}(R) - \rho = 0. \end{aligned}$$

From these two equations, multiplying the first equation by ϕ, substituting the expression in the second equation and noting that $\phi \cdot 1 = 1$, we obtain

$$\lambda_2 = 2(\phi)^\top \mathbb{E}_{\mathbb{P}}(RR^\top)\phi - \lambda_1 \rho.$$

We now define the following RWP function

$$\begin{aligned} &\bar{\mathcal{R}}_n(\phi, \lambda_1, \Sigma, \mu) \\ &:= \inf\left\{ D_c(\mathbb{P}, \mathbb{P}_n) : \begin{cases} 2\Sigma\phi - \lambda_1 \mu = \left(2(\phi)^\top \Sigma\phi - \lambda_1 \mu \cdot \phi\right)1 \\ \mu = \mathbb{E}_{\mathbb{P}}[R], \Sigma = \mathbb{E}_{\mathbb{P}}(RR^\top) \end{cases} \right\}, \end{aligned}$$

for $(\phi, \lambda_1, \Sigma, \mu) \in \mathbb{R}^d \times \mathbb{R} \times \mathcal{S}_+^{d\times d} \times \mathbb{R}^d$, where $\mathcal{S}_+^{d\times d}$ is the set of all the symmetric positive semidefinite matrices, and we convent that $\inf \emptyset := +\infty$. Moreover, define

$$\bar{\mathcal{R}}_n^*(\phi^*) := \inf_{\Sigma \in \mathcal{S}_+^{d\times d}, \mu \in \mathbb{R}^d, \lambda_1 \in \mathbb{R}} \bar{\mathcal{R}}_n(\phi^*, \lambda_1, \Sigma, \mu).$$

It follows directly from the definitions that

$$\phi^* \in \Lambda_\delta(\mathbb{P}_n) \Rightarrow \bar{\mathcal{R}}_n^*(\phi^*) \le \delta, \tag{7}$$

and for any given $\epsilon > 0$,

$$\bar{\mathcal{R}}_n^*(\phi^*) \le \delta + \epsilon \Rightarrow \phi^* \in \Lambda_\delta(\mathbb{P}_n). \tag{8}$$

Let us define

$$\tilde{\delta}_n^* = \inf\{\delta > 0 : \mathbb{P}^*(\bar{\mathcal{R}}_n^*(\phi^*) \le \delta) \ge 1 - \delta_0\}.$$

It follows from (7) and (8) that $\tilde{\delta}_n^* \le \delta_n^* \le \tilde{\delta}_n^* + \epsilon$. As $\epsilon > 0$ is arbitrary, we obtain

$$\delta_n^* = \tilde{\delta}_n^* = \inf\{\delta : \mathbb{P}^*(\bar{\mathcal{R}}_n^*(\phi^*) \le \delta) \ge 1 - \delta_0\}.$$

In other words, δ_n^* is the quantile corresponding to the $1 - \delta_0$ percentile of the distribution of $\bar{\mathcal{R}}_n^*(\phi^*)$.[8]

Still, even under Assumption 3, the statistic $\bar{\mathcal{R}}_n^*(\phi^*)$ is somewhat cumbersome to work with as it is derived from solving a minimization problem in terms of the mean and variance. So, instead, we define an alternative statistic involving only the empirical mean and variance while producing an upper bound of $\delta = \delta_n$, which still preserves the target rate of convergence to zero as $n \to \infty$ (which, as we argue, should be of order $O(n^{-1})$).

Denote $\Sigma_n = \mathbb{E}_{\mathbb{P}_n}(RR^\top)$, and let λ_1^* be the Lagrange multiplier in (6). Set

$$\mu_n = \rho 1 + 2\left(\Sigma_n \phi^* - \phi^{*T} \Sigma_n \phi^* 1\right)/\lambda_1^*. \tag{9}$$

Define

$$\mathcal{R}_n(\Sigma_n, \mu_n) := \bar{\mathcal{R}}_n(\phi^*, \lambda_1^*, \Sigma_n, \mu_n).$$

It is clear that

$$\mathcal{R}_n(\Sigma_n, \mu_n) \ge \bar{\mathcal{R}}_n^*(\phi^*).$$

Therefore,

$$\mathcal{R}_n(\Sigma_n, \mu_n) \le \delta \Rightarrow \bar{\mathcal{R}}_n^*(\phi^*) \le \delta,$$

and consequently,

$$\bar{\delta}_n^* = \inf\{\delta \geq 0 : \mathbb{P}^*(\mathcal{R}_n(\Sigma_n, \mu_n) \leq \delta) \geq 1 - \delta_0\} \geq \delta_n^*. \tag{10}$$

Moreover, because of the choice of Σ_n and μ_n, we have

$$\mathcal{R}_n(\Sigma_n, \mu_n) = \inf\{\mathcal{D}_c(\mathbb{P}, \mathbb{P}_n) : \mathbb{E}_{\mathbb{P}}[RR^\top] = \Sigma_n, \mathbb{E}_{\mathbb{P}}[R] = \mu_n\}.$$

The next result shows $\bar{\delta}_n^* = O(n^{-1})$ as $n \to \infty$.

Theorem 2. *Assume* Assumptions 1 *and* 2 *hold and write* $\mu_* = \mathbb{E}_{\mathbb{P}^*}(R)$ *and* $\Sigma_* = \mathbb{E}_{\mathbb{P}^*}(RR^\top)$. *Define* $g(x) = x + 2\left(xx^\top \cdot \phi^* - \phi^{*T} xx^\top \phi^* 1\right)/\lambda_1^*$. *Then,*

$$n\mathcal{R}_n(\Sigma_n, \mu_n) \Rightarrow L_0 := \sup_{\bar{\lambda} \in \mathbb{R}^d} \left(\bar{\lambda}^\top Z - \inf_{\bar{\Lambda} \in \mathbb{R}^{d \times d}} \mathbb{E}_{\mathbb{P}^*}\left[\|\bar{\Lambda} R + \bar{\lambda}\|_p^2 \right] \right),$$

where $Z \sim N\left(0, \Upsilon_g\right)$. *Moreover, if* $p = 2$, *then*

$$L_0 = \frac{\|Z\|_2^2}{4\left(1 - \mu_*^\top \Sigma_*^{-1} \mu_*\right)}.$$

A proof of Theorem 2 is provided in Appendix B.

Note that L_0 has an explicit expression when $p = 2$. When $p \neq 2$, using the inequalities that $\|x\|_p^2 \geq \|x\|_2^2$ if $p < 2$ and $d^{(\frac{1}{2}-\frac{1}{p})}\|x\|_p^2 \geq \|x\|_2^2$ if $p > 2$, we can find a stochastic upper bound of L_0 that can be explicitly expressed. In that case, we can obtain $\bar{\delta}_n^*$ in exactly the same way, namely, first compute the $1 - \delta_0$ quantile of L_0 and then let $\bar{\delta}_n^*$ be such quantile multiplied by $1/n$. The distribution of L_0 can be calibrated using a natural plug-in estimator, leading to an asymptotically equivalent estimator of $\bar{\delta}_n^*$. The validity of this type of (plug-in) approach is explained in the next section in the context of choosing $\bar{\alpha}$, but the principle applies directly in the setting of L_0 as well. In simple words, whenever an asymptotic limiting distribution depends continuously on various parameters and consistent estimators are available for those parameters, then consistent plug-in estimators can be safely used still preserving exactly the same asymptotic distributions. The details of this approach are investigated in proposition 2 of Blanchet et al. (2019), and the performance of such plug-in estimators are tested empirically in Section 5 of this paper.

Q:9

4.2. Choice of $\bar{\alpha}$

Once δ has been chosen, the next step is to choose $\bar{\alpha}$. The idea is to select $\bar{\alpha}$ just large enough to make sure that we do not rule out the inclusion $\phi^* \in \mathcal{F}_{\delta,\bar{\alpha}}(n)$ with a given confidence level chosen by the user, and ϕ^* is the optimal solution to (1). It is equivalent to choosing v_0, where

$$\bar{\alpha} = \rho - \sqrt{\delta}\|\phi^*\|_p v_0.$$

Therefore, it follows from Proposition A.1 that $\phi^* \in \mathcal{F}_{\delta,\bar{\alpha}}(n)$ if and only if

$$(\phi^*)^\top \mathbb{E}_{\mathbb{P}_n}(R) - \sqrt{\delta}\,\|\phi^*\|_p \geq \rho - \sqrt{\delta}\,\|\phi^*\|_p v_0.$$

However, $\rho = (\phi^*)^\top \mathbb{E}_{\mathbb{P}^*}(R)$, so the previous inequality holds if and only if

$$(\phi^*)^\top \left[\mathbb{E}_{\mathbb{P}_n}(R) - \mathbb{E}_{\mathbb{P}^*}(R)\right] \geq \|\phi^*\|_p \sqrt{\delta}\,(1 - v_0). \tag{11}$$

Hence, we can choose $\sqrt{\delta}\,(1 - v_0) < 0$ sufficiently negative so that the previous inequality holds with a specified confidence level. We hope to choose a v_0 such that ϕ^* satisfies (11) with confidence level $1 - \epsilon$. This can be achieved asymptotically by a central limit theorem as the following result indicates.

Proposition 1. *Suppose that Assumptions* 1 *and* 3 *hold and let* $\{\phi_n^*\}_{n=1}^\infty$ *be any consistent sequence of estimators of* ϕ^* *in the sense that* $\phi_n^* \to \phi^*$ *in probability as* $n \to \infty$. *Then,*

$$n^{1/2}\left\{\frac{(\phi_n^*)^\top \left[\mathbb{E}_{\mathbb{P}_n}(R) - \mathbb{E}_{\mathbb{P}^*}(R)\right]}{\|\phi_n^*\|_p}\right\} \Rightarrow N\left(0, \Upsilon_{\phi^*}\right)$$

as $n \to \infty$ *where*

$$\Upsilon_{\phi^*} := \lim_{n \to \infty} \mathrm{Var}_{\mathbb{P}^*}\left(n^{-1/2} \sum_{k=1}^n (\phi^*)^\top R_k / \|\phi^*\|_p \right).$$

A proof of this proposition is delayed to Appendix C.

Using the previous result, we can estimate v_0 asymptotically. Let ϕ_n denote the optimal solution of problem $\mathcal{Q}(\mathbb{P}_n)$. We know that ϕ_n converges to ϕ^* in probability. So we choose a v_0 such that the following inequality holds with confidence level $1 - \epsilon$,

$$\frac{1}{\|\phi_n\|_p}(\phi_n)^\top \left[\mathbb{E}_{\mathbb{P}_n}(R) - \mathbb{E}_{\mathbb{P}^*}(R)\right] \geq \sqrt{\delta}(1 - v_0). \tag{12}$$

According to Proposition 1 the left-hand side of (12) is approximately normally distributed, and thus, we can choose its $1 - \epsilon$ quantile and consequently decide the value of $v_0 > 1$.

For the reader's convenience, we present a simple "menu" for estimating δ and $\bar{\alpha}$.

1. Choose the target return rate ρ.
2. Collect return data $\{R_i\}_{i=1}^n$.
3. Use the sample mean $\mu_n = \mathbb{E}_{\mathbb{P}_n}(R)$ and the sample second-moment matrix $\Sigma_n = \mathbb{E}_{\mathbb{P}_n}(RR^\top)$ to approximate μ_* and Σ_*, respectively, appearing in Theorem 2.
4. Use the solution ϕ_n, which is the solution to problem $\mathcal{Q}(\mathbb{P}_n)$ (see (5)), to approximate ϕ^* in Theorem 2.

5. Apply Theorem 2 and (10) to determine $\delta = \bar{\delta}^*$ with the 95% confidence level.

6. Choose v_0' based on the 95% quantile according to (12) and Proposition 1 and, consequently, obtain $\bar{\alpha}$. Choose v_0'' such that the equation $(\phi_n)^\top \mathbb{E}_{\mathbb{P}_n}(R) - \sqrt{\delta}\|\phi_n\|_p = \rho - \sqrt{\delta}\|\phi_n\|_p v_0''$ holds. Then, set $v_0 = max(v_0', v_0'')$.

To conclude this section, we note that the choices of δ and α are separate, each with a given confidence level. Jointly, the chosen (δ, α) may have a completely different confidence level. It remains an interesting problem to develop a joint way of choosing the two parameters to ensure a *given, fixed* confidence level.

5. Empirical Performance and Comparisons

In this section, we report the results of our back-testing experiments on S&P 500 constituents that compare the performance of our DRMV portfolios with those of the portfolios based on the following models: classical (nonrobust) single-period Markowitz, continuous-time Markowitz, Fama–French, Black–Litterman, robust Goldfarb–Iyengar, Olivares-Nadal–DeMiguel, and an equally weighted portfolio. The first four models are well known and are widely used in practice, the fifth one is an alternative robust model not based on distributional uncertainty, and the sixth one has transaction costs turned into an equivalent robust model that has an ellipsoidal uncertainty around the mean. The equally weighted strategy is actually an extreme outcome of the DRMV model when the uncertainty size $\delta = \infty$.

5.1. Experiment Design and Data Preparation

We back-tested for the period January 2000–December 2016 with the training (estimation) period being January 1991–December 1999 (i.e., the previous 10 years).[9] All the stock monthly price data was obtained from the database of Wharton Business School. At the beginning of 2000, we *randomly* chose 100 stocks from the constituents of S&P 500 that have at least 10 years' historical price data available.[10] The basic period is set to be one year in all the single-period models involved with the target annual mean return rate fixed to be $\rho = 10\%$ when applicable. Then, we used the training period to estimate the out-of-sample parameters, namely, the mean and the variance, to construct the optimal strategies of the various models tested.

5.1.1. DRMV Model. Let us first describe in detail the construction of the DRMV strategy for the selected 100 stocks. We generated this 17-year-long strategy in an (overlapping) rolling horizon fashion with each horizon being a month. Specifically, on the first trading day of January 2000, we solved our DRMV model to obtain a portfolio, denoted as ϕ_R. In doing so, we set $p = q = 2$ and $\rho = 10\%$ and obtained the parameters, δ and $\bar{\alpha}$, using the menu at the end of Section 4. We then substituted δ and $\bar{\alpha}$ in the optimization problem described in Theorem 1 to obtain ϕ_R.

We kept ϕ_R until only the first trading day of February 2000. At that point we reestimated the parameters δ and $\bar{\alpha}$ using the *immediate* previous 10-year (namely February 1991–January 2000) price data, resolved the DRMV model, and generated a new portfolio ϕ_R for February 2000, the second month in our back-testing period. We repeated the same steps for all the subsequent months.

If, at the beginning of a month, some stocks in our portfolio had been removed out of the S&P 500 during the previous month, then we would also remove them from our portfolio, replace them by the same number of stocks that were randomly picked from S&P 500 (yet having at least 10 years' historical data), and then rebalance based on our DRMV model. We still denoted by ϕ_R the overall portfolio for the 17-year period and kept track of the wealth process that had been updated at the end of each month.

In what follows, we describe the implementations of the other models, mentioned at the beginning of this section, under comparison. Except for the continuous-time Markowitz model, all the rest are single-period models, so we applied the same monthly rolling horizon approach to build the respective strategies. Moreover, for these single-period models, whenever there were stocks dropped from S&P 500, we would replace them with exactly the same set of stocks as in the DRMV model so as to maintain consistency across different models. The case of the continuous-time model is slightly more complicated, and we explain how we deal with these issues separately.

5.1.2. Single-Period and Continuous-Time Markovitz Models. For the single-period Markovitz model, for each period (month) we used the sample mean and covariance matrix of the immediate previous 10-year return data to estimate the corresponding parameters in Problem (1). Then, we generated the optimal portfolio of the classical Markowitz model, ϕ_M, by setting $\rho = 10\%$ and solving Problem (1) on exactly the same rolling horizon basis as the DRMV model.

The continuous-time Markowitz mean–variance model is based on Cui et al. (2012) in which portfolios are constructed on risky stocks only. This setting is consistent with ours.[11] It is assumed that the stock price process follows correlated time-inhomogeneous Black–Scholes dynamics. Let $\{X(t) : t \in [0, T]\}$ be the wealth process (also called an admissible wealth process) under any given admissible portfolio. The mean–variance problem is

$$\text{Minimize } \text{Var}(X(T)) \tag{13}$$

subject to

$$\{X(t): t \in [0,T]\} \text{ is admissible, } X(0) = x_0, \ \mathbb{E}[X(T)] = z,$$

where z is a given parameter representing the expected payoff at the end of the investment horizon, T. An optimal strategy is given explicitly in theorem 1.1 of Cui et al. (2012), which gives the portfolio at each given time $t \in [0,T]$ as a function of the wealth, a couple of auxiliary feedback processes, and the estimates (at time t) of the (time-inhomogeneous) diffusion and drift coefficients.

In theory, a continuous-time model requires *continuous* rebalancing all the time. Naturally, it is not possible (indeed not necessary) in practice or in our empirical implementation. In our experiments, we set $T = 1$ (year) and $z = 1.1x_0$ corresponding to an annual expected return $\rho = 10\%$, and we rebalance only monthly (instead of continuously). The one-year period is consistent with the other models under comparison. Therefore, on the first trading day of January 2000, we estimated all the necessary parameters/coefficients based on the previous 10-year data and then applied the explicit formula for the optimal portfolio, denoted as ϕ_C, given by theorem 1.1 of Cui et al. (2012). On the first trading day of February 2000, we applied the same formula but with updated estimates of the parameters/coefficients based on the immediate previous 10-year data. This way we have constructed a strategy for the whole year of 2000. For all the subsequent years, we repeat the same procedure to generate a 17-year-long strategy ϕ_C and the corresponding wealth process.

It is important to note that this model does not have an explicit no-bankruptcy constraint (i.e., it does not rule out the possibility that the wealth process may go negative during $[0,T]$). Indeed, as we see in the discussions, this model led to bankruptcy in *all* of our numerical experiments for portfolios of 100 stocks.[12] Once a bankruptcy happened, we then considered it game over and kept the zero wealth until December 2016.[13]

5.1.3. Fama–French Model. The celebrated Fama–French model helps estimate the covariance matrix when the number of stocks d is close to or greater than the sample size n. In this section, we follow the approach proposed by Fan et al. (2011) to implement the Fama–French model. Specifically, we assume the stock returns follow the factor model

$$\mathbf{r} = \mathbf{B}f + \mathbf{u}, \tag{14}$$

where $\mathbf{r} = (R_1,\ldots,R_d)^\top$ is the random vector of stock returns; $\mathbf{f} = (f_1,f_2,f_3)^\top$ consists of the three factors of the Fama–French model (i.e., Rm-Rf, SMB, and HML), respectively; $\mathbf{B} = (b_1,\ldots,b_d)^\top$ is the vector of factor loading; and $\mathbf{u} = (u_1,\ldots,u_d)^\top$ is the vector of uncorrelated errors.

Let $(\mathbf{f}_1,\mathbf{r}_1),\ldots,(\mathbf{f}_n,\mathbf{r}_n)$ be n independent and identically distributed samples of $(\mathbf{f},\mathbf{r})$. For notational simplicity, we define

$$X = (\mathbf{f}_1,\ldots,\mathbf{f}_n),\ \mathbf{R} = (\mathbf{r}_1,\ldots,\mathbf{r}_n) \text{ and } \mathbf{U} = (\mathbf{u}_1,\ldots,\mathbf{u}_n),$$

where $\mathbf{u}_i, i = 1,\ldots,n$ are the corresponding error vectors. Under Model (14), the covariance matrix Σ of stock returns satisfies

$$\Sigma = \mathbf{B}\text{cov}(\mathbf{f})\mathbf{B}^\top + \Sigma_{\mathbf{u}}, \tag{15}$$

where $\Sigma_{\mathbf{u}}$ is the covariance matrix of $\mathbf{u}$.

Then, we estimate Σ using a substitution estimator

$$\hat{\Sigma} = \hat{\mathbf{B}}\hat{\text{cov}}(\mathbf{f})\hat{\mathbf{B}}^\top + \hat{\Sigma}^\tau_{\mathbf{u}}, \tag{16}$$

where $\hat{\mathbf{B}} = YX^\top(XX^\top)^{-1}$, $\hat{\text{cov}}(\mathbf{f}) = \frac{1}{n-1}XX^\top - \frac{1}{n(n-1)}X11^\top X^\top$, and $\hat{\Sigma}^\tau_{\mathbf{u}}$ is obtained by applying the same adaptive thresholding approach in Fan et al. (2011) on $\hat{\Sigma}_{\mathbf{u}}$ with $\hat{\Sigma}_{\mathbf{u}} = \hat{\text{cov}}(\hat{\mathbf{U}})$ being the sample covariance matrix of $\hat{\mathbf{U}}$ with $\hat{\mathbf{U}} = \mathbf{Y} - \hat{\mathbf{B}}\mathbf{X}$. $\hat{\Sigma}^\tau_{\mathbf{u}}$ is obtained by applying the same adaptive thresholding approach mentioned in Fan et al. (2011) on $\hat{\Sigma}_{\mathbf{u}}$. In that paper, they used the threshold $\omega = C * 3 * \sqrt{\frac{\log d}{n}}$, where $C = 0.1$. However, in our experiment, the value $C = 0.1$ leads to bankruptcy in all the cases. So we tune this parameter and find the optimal $C = 0.01$. Q: 10

Finally, we have a substitution estimator

$$\hat{\mu} = \hat{\mathbf{B}}\bar{\mathbf{f}} \tag{17}$$

of the mean vector μ, where $\bar{\mathbf{f}} = \frac{1}{n}\sum_{i=1}^n \mathbf{f}_i$.

In implementing the Fama–French model, we first downloaded the monthly data of the three factors[14] from Kenneth French's personal website.[15] Then, on the first trading day of each month during January 2000–December 2016, we used its immediate prior 10-year history returns and factors data to estimate the covariance matrix and mean vector using (16) and (17), respectively. We used these estimates for all the randomly chosen 100 stocks as the mean vector and solve the single-period classical Markowitz model with $\rho = 10\%$. The generated portfolio was denoted as ϕ_F. This process was then repeated in the subsequent months on a rolling horizon basis.

5.1.4. Black–Litterman Model. The Black–Litterman model was developed to address the mean-blur problem, namely, the fact that compared with variance, it is much more difficult to estimate within a workable accuracy the expected returns of stocks purely based on sample means. The model estimates the stock returns by the market portfolio while keeping the sample covariance matrix and feeds them into the classical Markowitz model to obtain the optimal strategies.

To implement this model, on the first trading day of each month during January 2000–December 2016, we calculated the implied returns of all the S&P 500 constituent stocks having at least 10 years' historical data by using the following formula:

$$R_{implied} = \lambda \Sigma \phi_{market},$$

where $\lambda = 3.07$, Σ is the sample covariance matrix of the previous 10 years' returns of these stocks and ϕ_{market} is the corresponding market portfolio (i.e., ϕ_{market} is a vector whose components add up to one and are proportional to the capitalizations of the S&P 500 constituents having at least 10 years' historical data) at the closing prices of the previous trading day; see Idzorek (2002).[16] Then, we picked from $R_{implied}$ the implied returns of the 100 stocks that had been randomly chosen. We input these returns and the sample covariance matrix into the classical Markovitz model with $\rho = 10\%$ to obtain the portfolio ϕ_B. This process was repeated in subsequent months on a rolling horizon fashion.

5.1.5. Goldfarb–Iyengar Robust Model. Goldfarb and Iyengar (2003) consider the following robust Markovitz problem with a factor model for the return rate:

$$\begin{array}{ll} \text{Minimize}_{\ \phi} & \max\limits_{\mathbf{V}\in S_v,\ \mathbf{D}\in S_d} \text{Var}[\mathbf{r}^\top \phi] \\ \text{subject to} & \min\limits_{\boldsymbol{\mu}\in S_m} \mathbb{E}[\mathbf{r}^\top \phi] \geq \rho,\ \mathbf{1}^\top \phi = 1, \end{array}$$

where

$$\begin{aligned} &\mathbf{r} = \mu + \mathbf{V}^\top \mathbf{f} + \epsilon,\ \epsilon \sim N(0, \mathbf{D}), \\ &S_v = \{V : V = V_0 + W,\ \|W_i\|_g \leq \rho_i, i = 1, \ldots, n\} \end{aligned}$$

with W_i being the ith column of W and $\|w\|_g = \sqrt{w^\top G w}$ for some positive definite matrix G,

$$\begin{aligned} &S_d = \{D : D = diag(d), d_i \in [d_i^{min}, d_i^{max}], i = 1, \ldots, n\}, \\ &S_v = \{V : V = V_0 + W, \|W_i\|_g \leq \lambda_i, i = 1, \ldots, n\}, \end{aligned}$$

and

$$S_m = \{\mu : \mu = \mu_0 + \xi, |\xi_i| \leq \gamma_i, i = 1, \ldots, n\}.$$

So the uncertainty set of this model is based on vector/matrix distance as opposed to our uncertainty set that is defined through the Wasserstein distance between probability measures.

In implementing this model, we followed the instructions in section 7.2 of Goldfarb and Iyengar (2003). Specifically, we calculated the 10 years' sample returns $\mathbf{r}$ of the chosen 100 stocks. Then, we used the top five principal components of $\mathbf{r}$ together with the return data of DJA, NDX, SPC, RUT, and TYX to be the factor vector $\mathbf{f}$. By choosing the confidence threshold ω to be 95%, we estimated μ_0, $\mathbf{V}_0$, σ_i^2, γ_i, G, and λ_i. With the target annual return $\rho = 10\%$ and plugging all the parameters from the preceding steps, we used SeDuMi to solve the second-order cone programming formulation (problem (32) in Goldfarb and Iyengar 2003) to obtain the portfolio ϕ_G for each month on a rolling horizon basis, starting from January 2000.

5.1.6. Olivares-Nadal–DeMiguel Model. Olivares-Nadal and DeMiguel (2018) examine, in the context of a mean–variance model, the equivalence between certain transaction costs and ellipsoidal robustification around the mean. They then devise a data-driven approach to portfolio selection by treating the transaction costs as a regularization term to be calibrated. Their numerical experiments test for different variants of their model, but the minimum-variance portfolios (MVPs) based on quadratic transaction costs have the overall best performance in terms of the Sharpe ratio (see table 1 of Olivares-Nadal and DeMiguel 2018) with which we compare in our setting.

The MVP model is

$$\begin{array}{ll} \text{Minimize}_{\ \phi} & \phi^\top \Sigma \phi \\ \text{subject to} & \mathbf{1}^\top \phi = 1, \end{array} \tag{18}$$

where $\phi \in \mathbb{R}^d$ is the portfolio weight vector and $\Sigma \in \mathbb{R}^{d\times d}$ the estimated covariance matrix of asset returns. Adding a quadratic transaction cost, Olivares-Nadal and DeMiguel (2018) consider the following model:

$$\begin{array}{ll} \text{Minimize}_{\ \phi} & \phi^\top \Sigma \phi + \kappa \|\Sigma^{\frac{1}{2}}(\phi - \phi_0)\|_2^2 \\ \text{subject to} & \mathbf{1}^\top \phi = 1 \end{array}, \tag{19}$$

where $\kappa \in \mathbb{R}$ is the transaction cost parameter and $\phi_0 \in \mathbb{R}^d$ is the starting portfolio. By the conjugate representation of the L^2-norm, the connection of (19) to a robust model is straightforward.

It can be shown that the optimal solution ϕ_{ODM} of (19) is

$$\phi_{ODM} = \frac{1}{1+\kappa}\phi_M + \frac{\kappa}{1+\kappa}\phi_0, \tag{20}$$

where ϕ_M solves (18).

In Model (19), one needs to calibrate the parameter κ. Olivares-Nadal and DeMiguel (2018) do this by calibrating the trading volume τ (i.e., $\|\phi - \phi_0\|_1 \leq \tau$). They use 10-fold cross-validation to select the best τ. Specifically, they divide the empirical returns into 10 intervals. For each j from 1 to 10, they remove the jth interval and use the remaining returns to estimate parameters and obtain the corresponding portfolio. Then, they evaluate the portfolio on the jth interval. After completing this process for each of the 10 intervals, they compute the variance of the out-of-sample returns for different τ from the set $\{0\%, 0.5\%, 1\%, 2.5\%, 5\%, 10\%\}$ and then choose the τ that corresponds to the portfolio with the smallest variance.

Now, we show how to infer the value of κ from that of τ once the latter has been calibrated. By (20), the

trading volume can be expressed as

$$\|\phi_{ODM} - \phi_0\|_1 = \frac{1}{1+\kappa}\|\phi_M - \phi_0\|_1.$$

To determine κ, we equate the trading volume to τ, that is,

$$\|\phi_{ODM} - \phi_0\|_1 = \frac{1}{1+\kappa}\|\phi_M - \phi_0\|_1 = \tau.$$

This leads to

$$\kappa = \left(\frac{\|\phi_M - \phi_0\|_1}{\tau} - 1\right)_+,$$

where $(x)_+ := \max(x, 0)$.[17]

In implementing this model, we followed the instructions in section 3.1 of Olivares-Nadal and DeMiguel (2018). Specifically, we calculated the 10 years' sample monthly return $\mathbf{r}$ of the chosen 100 stocks. In the first month (January 2000), because we did not have the value of ϕ_0, we set $\kappa = 0$ and generated the optimal portfolio ϕ_{ODM} of (19) for that month. In any subsequent month, we took the portfolio of the previous month as ϕ_0, applied the aforementioned cross-validation method on the 10 years' sample return $\mathbf{r}$ to select the best τ, and plugged the corresponding κ into (19) to generate the optimal portfolio ϕ_{ODM}.

5.2. Comparisons

Assume that the initial wealth at the start of the backtesting period (i.e., January 2000) is one. For each randomly selected set of 100 stocks, we generate the wealth process for the period 2000–2016 under each of the seven models as described in the previous sections as well as that under equal weighting. Then, we repeat the experiments on 100 such sets of 100 stocks and obtain the *average* realized wealth process for each model. These processes, along with that of S&P 500 (normalized to start from one at the start of the testing period), are plotted in Figure 1.

Graphically, Figure 1 is "corrupted" because of the extreme behavior of the continuous-time Markowitz model. Its average performance went through the roof initially and then quickly dived to zero (all 100 experiments ended up in bankruptcy). So the continuous-time Markowitz is an extremely volatile model. This may be explained as follows. The dynamic strategies incorporate considerable feedback effects, which are computed assuming the underlying model is correct. As such, model misspecifications are *compounded* precisely because of feedback effects. The inclusion of feedback in the optimal dynamic policy, in outputs that are close to typical realizations of the underlying assumed model, results in highly profitable portfolios. On the other hand, even moderate discrepancies from the underlying model dynamics might lead to relatively poor performance. As a consequence, the dynamic model exhibits significantly high variability than the static rolling-horizon robust counterpart.

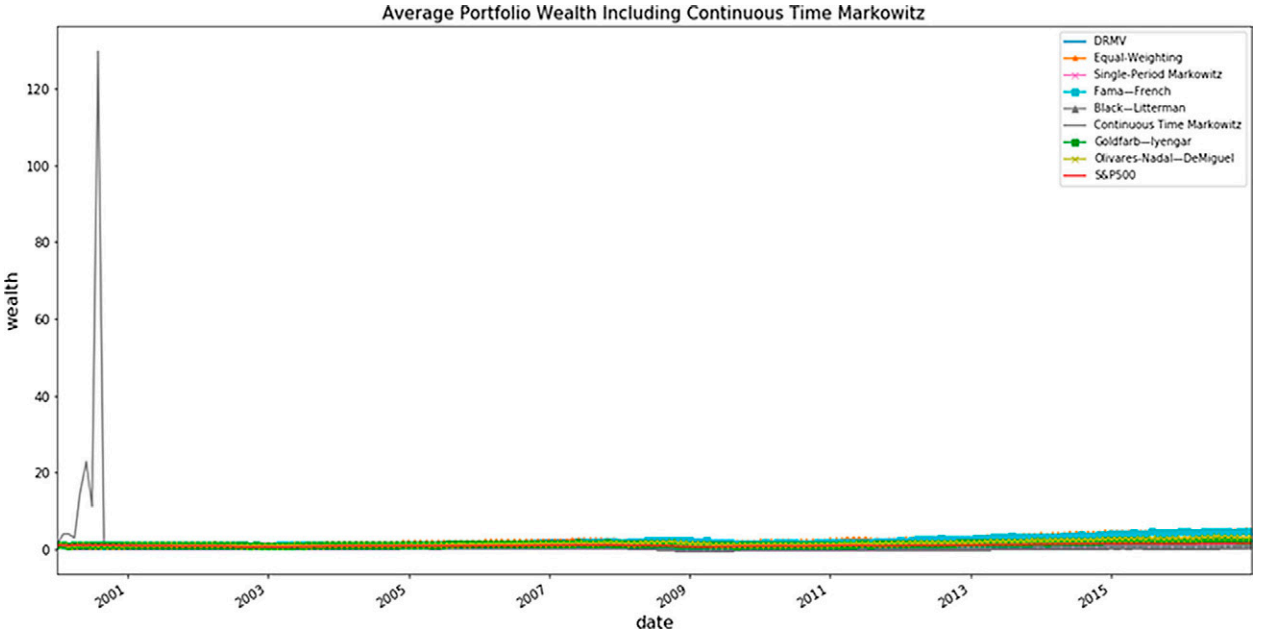
In order to be able to visualize the comparison among other portfolios, it is necessary to remove the continuous-time Markowitz from Figure 1, resulting in Figure 2. It is evident that all seven models, except Black–Litterman, (almost) uniformly and substantially outperform S&P 500 during the 17-year period. In terms of the final realized wealth, of the seven models, DRMV, equal weighting, and Fama–French lead by a substantial margin. The second-tier league includes Olivares-Nadal–DeMiguel and Goldfarb–Iyengar. The classical single-period Markowitz lags behind but still manages to outperform the market most of the time.

The average performances of DRMV and equal weighting are close although the former beats the latter most of the time. This is no surprise as the latter can be regarded as an extreme case of the distributionally robust model when the uncertainty size $\delta = \infty$, whereas the former has a nearly optimal choice of δ informed by the data.[18] We can study more closely the variability of the performances and the overall return–risk efficiency of the two models by examining their histograms of annualized returns (i.e., the distributions of the annualized returns of the 100 experiments) and those of Sharpe ratios. These are plotted in Figures 3 and 4, respectively. In both figures, DRMV is more shifted to the right than equal weighting, indicating the former outperforms the latter in the two criteria. Moreover, the two are almost equally concentrated, suggesting that both strategies have stable performance. We can also compare the histograms of kurtosis of the two; see Figure 5. There are no statistically significant differences between the two: most return distributions under both strategies are platykurtic (i.e., the kurtosis values are less than three), implying there are fewer extreme outliers than the standard normal. Overall, we can conclude that both DRMV and equal-weighting are robust and stable, but the former is superior to the latter in terms of the return–risk efficiency.

Similarly, we compare the respective histograms between DRMV and Fama–French; see Figures 6–8. DRMV has a much more concentrated return histogram, indicating a significantly more robust performance; a much more right-shifted Sharpe ratio histogram; and a more left-shifted kurtosis histogram, implying fewer extreme returns. We can, therefore, conclude that DRMV compares favorably with Fama–French in all the key metrics reported. However, it is interesting to note that DRMV utilizes only the price data, whereas Fama–French requires additional fundamental information on the companies concerned.

Likewise, DRMV outperforms the Olivares-Nadal–DeMiguel model (Olivares-Nadal and DeMiguel 2018) based on all the histogram comparisons;

Figure 1. Wealth Processes of All Portfolios (Including Continuous Time Markowitz) and S&P 500 from January 2000 to December 2016



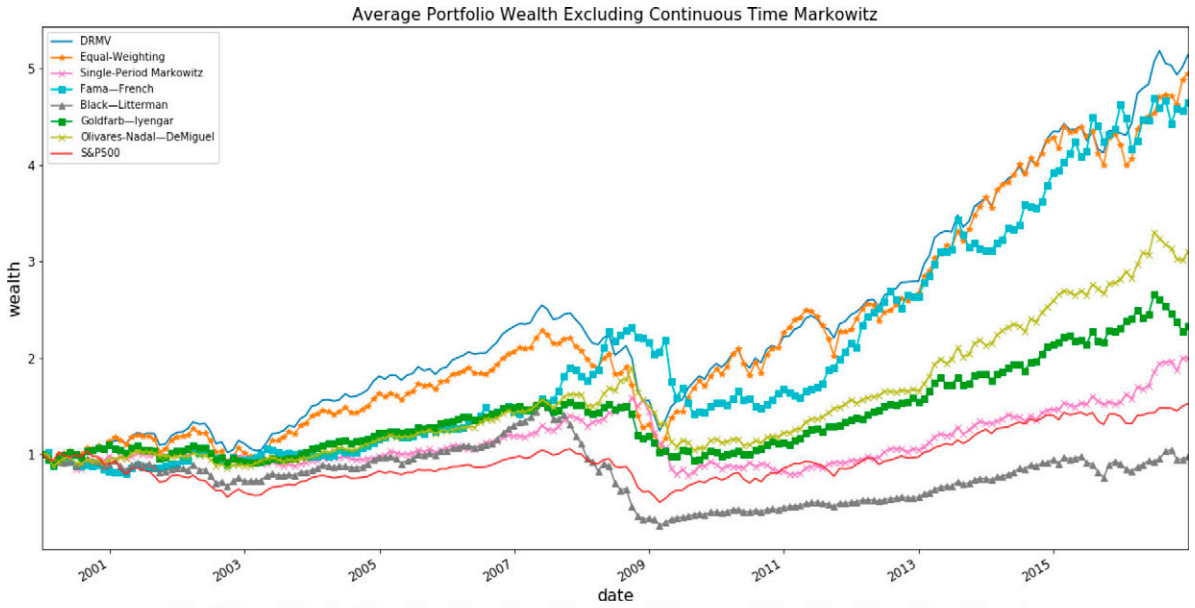
Notes. All the portfolios except S&P 500 consist of 100 stocks, and the averages are calculated over 100 numerical experiments. The *x*-axis indicates the time in months (from 1 to 204), and the *y*-axis indicates the portfolio wealth. Initial wealth is set to be one.

Q: 17

see Figures 9–11. This suggests that, among other things, the inference method for selecting the uncertainty/regularization size is advantageous compared with the cross-validation.

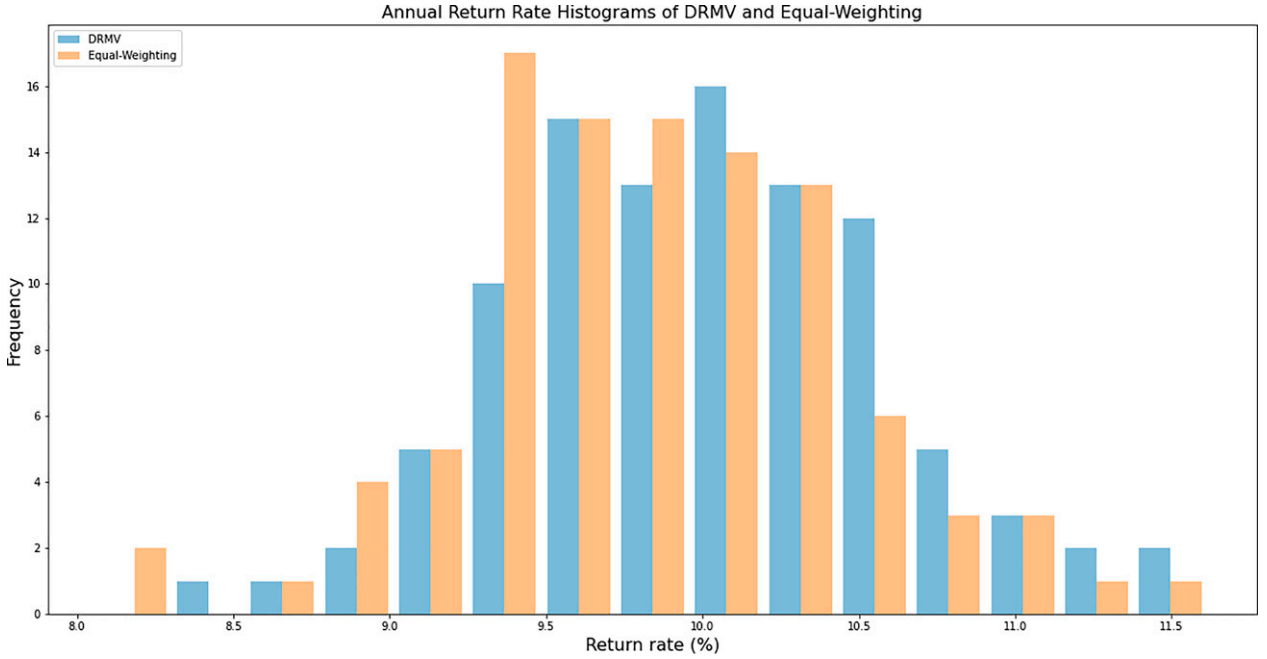
As for the robust portfolio model by Goldfarb and Iyengar (2003), we notice that it has a reasonably concentrated return histogram (Figure 12), indicating its robustness. However, DRMV's returns are not only more concentrated but also distributed more to the right than Goldfarb–Iyengar's. Together with the Sharpe ratio histogram (Figure 13), the kurtosis histogram (Figure 14), and the average wealth comparison

Figure 2. Wealth Processes of All Portfolios (Excluding Continuous-Time Markowitz) and S&P 500 from January 2000 to December 2016



Notes. All the portfolios except S&P 500 consist of 100 stocks, and the averages are calculated over 100 numerical experiments. The *x*-axis indicates the time in months (from 1 to 204), and the *y*-axis indicates the portfolio wealth. Initial wealth is set to be one.

Figure 3. Histograms of the Annualized Returns of the 100 Different Experiments on DRMV (Blue) and Equal-Weighting (Orange) Portfolios

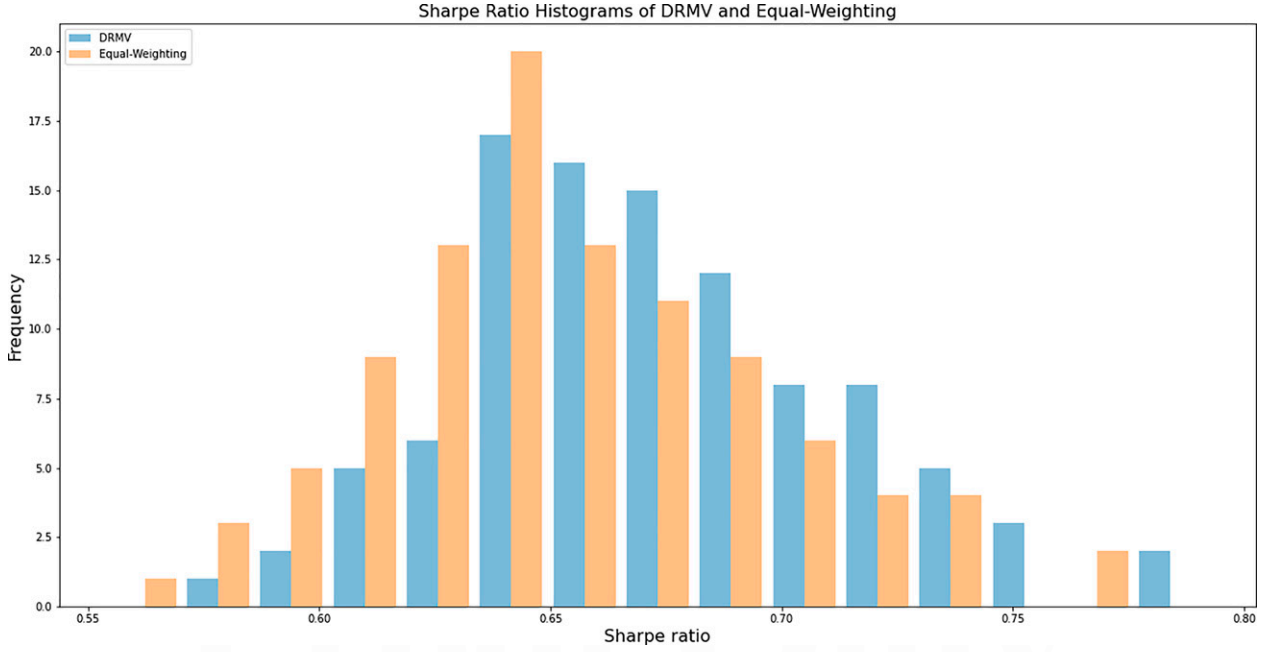


Note. The x-axis represents the annualized returns, and the y-axis represents the numbers of returns.

(Figure 2), it is clear that our uncertainty set formulation based on Wasserstein distance is a significant improvement to the matrix/vector distance.

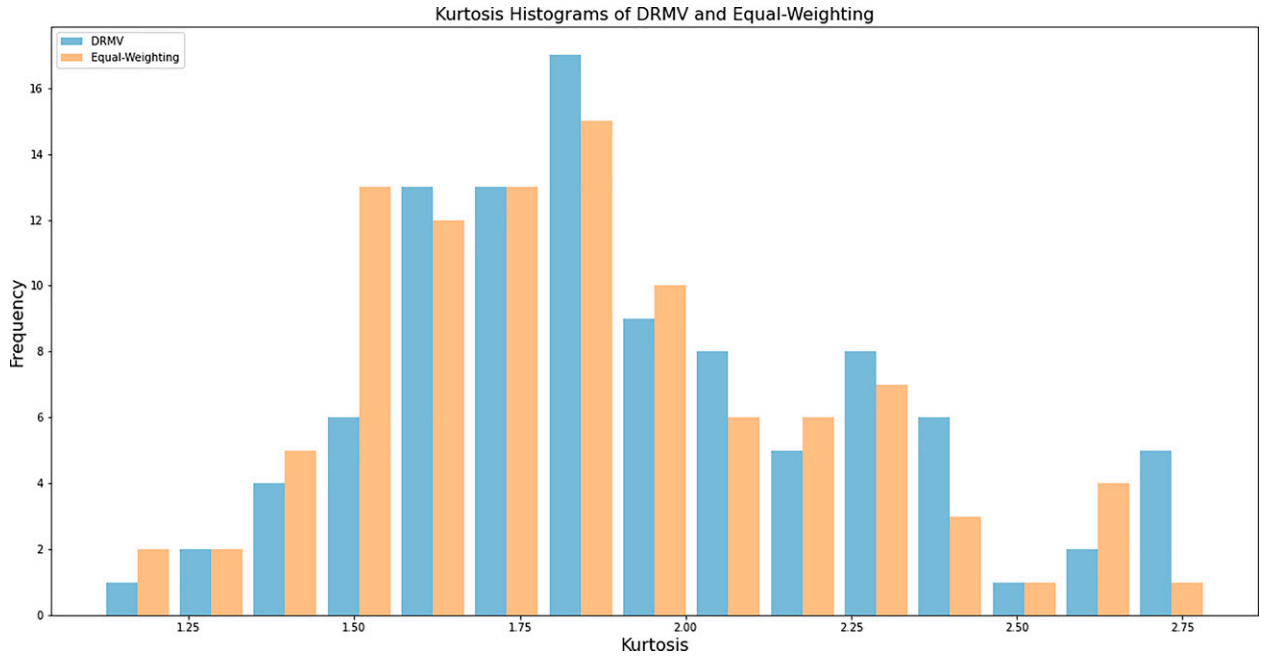
Finally, we provide comparisons of histograms between DRMV and single-period Markowitz and between DRMV and Black–Litterman; see Figures 15–20.

Figure 4. Histograms of the Sharpe Ratio of the 100 Different Experiments on DRMV (Blue) and Equal-Weighting (Orange) Portfolios



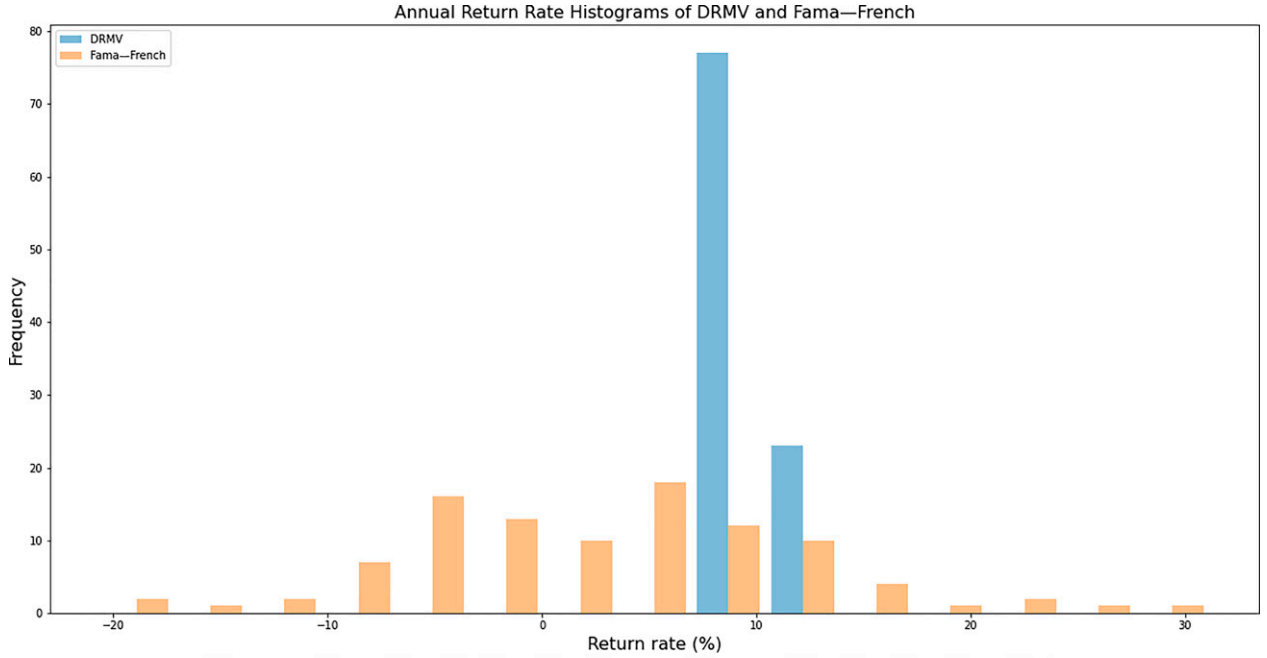
Note. The x-axis represents the Sharpe ratio, and the y-axis represents the numbers of Sharpe ratios.

Figure 5. Histograms of the Kurtosises of the 100 Different Experiments on DRMV (Blue) and Equal-Weighting (Orange) Portfolios



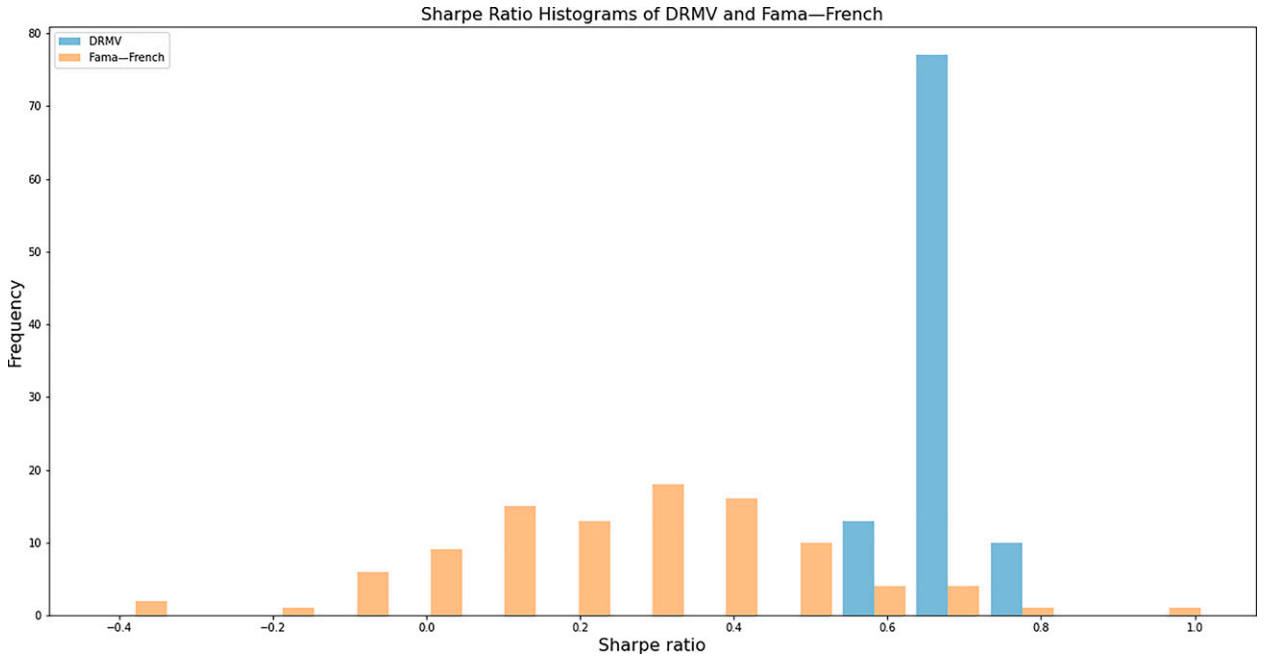
Note. The *x*-axis represents the kurtosis, and the *y*-axis represents the numbers of kurtoses.

Figure 6. Histograms of the Annualized Returns of the 100 Different Experiments on DRMV (Blue) and Fama-French (Orange) Portfolios



Note. The *x*-axis represents the annualized returns, and the *y*-axis represents the numbers of returns.

Figure 7. Histograms of the Sharpe Ratio of the 100 Different Experiments on DRMV (Blue) and Fama-French (Orange) Portfolios



Note. The x-axis represents the Sharpe ratio, and the y-axis represents the numbers of Sharpe ratios.

Clearly, DRMV has far superior performance in all the metrics.

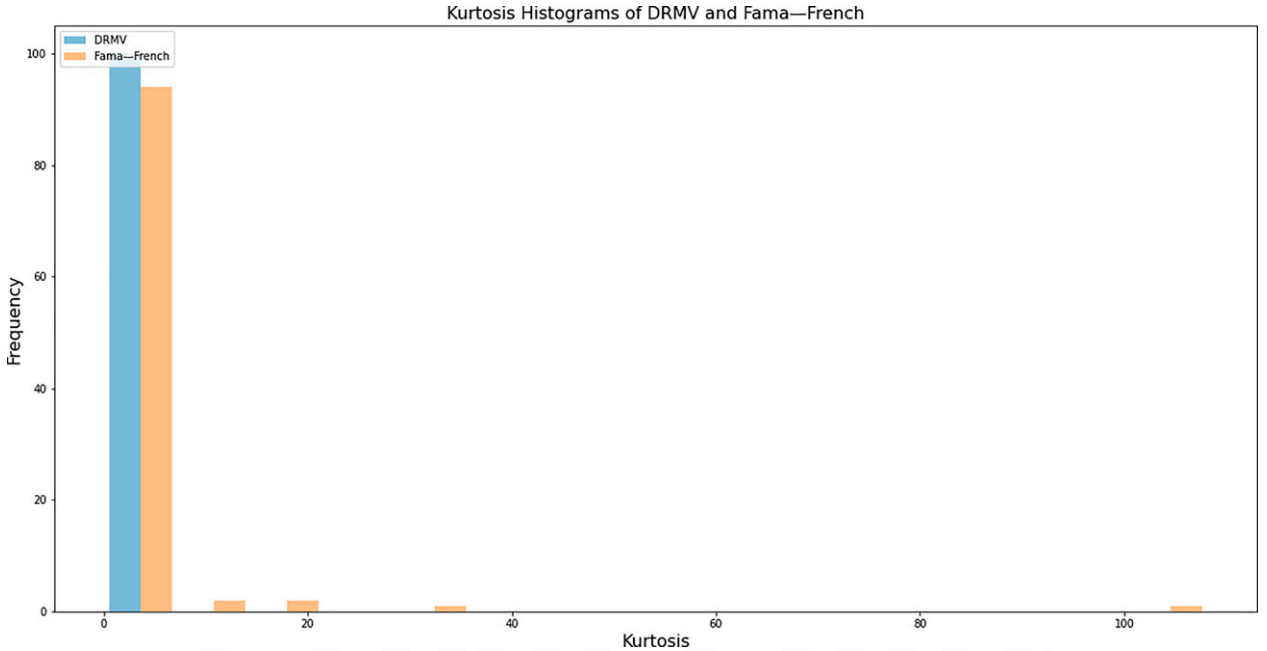
5.3. Discussions

In this section, we offer discussions on various issues related to our empirical experiments.

5.3.1. Wasserstein Order $p = 1$.

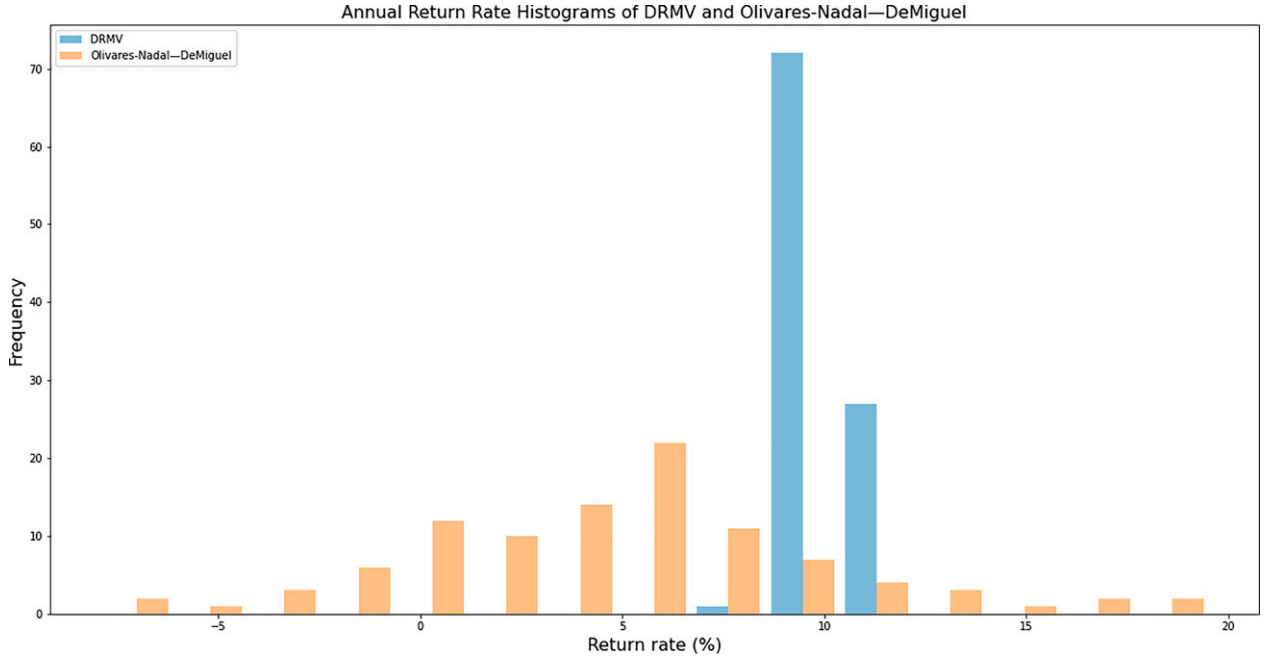
In all the previously reported experiments, we set the order of the Wasserstein distance to be $p = 2$. We have also tried $p = 1$ (and, hence, $q = \infty$) and find that the performance of the resulting strategies becomes very volatile. So we recommend using $p = 2$.

Figure 8. Histograms of the Kurtoses of the 100 Different Experiments on DRMV (Blue) and Fama-French (Orange) Portfolios



Note. The x-axis represents the kurtosis, and the y-axis represents the numbers of kurtoses.

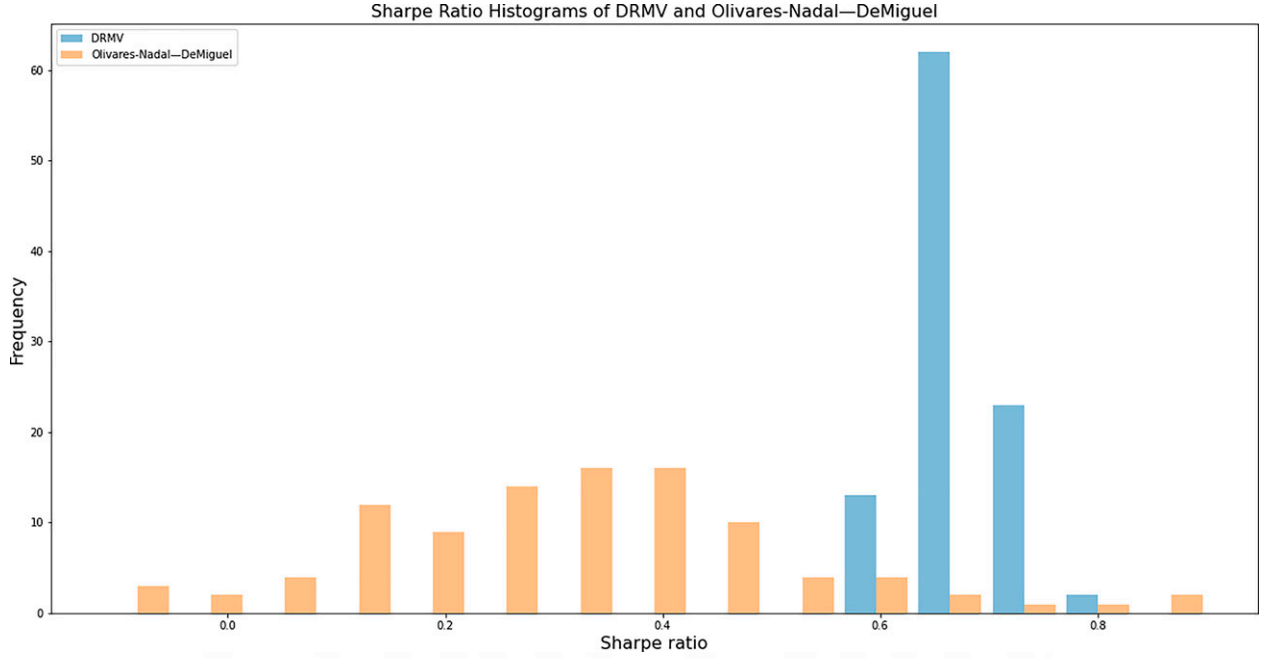
Figure 9. Histograms of the Annualized Returns of the 100 Different Experiments on DRMV (Blue) and Olivares-Nadal–DeMiguel (Orange) Portfolios



Note. The x-axis represents the annualized returns, and the y-axis represents the numbers of returns.

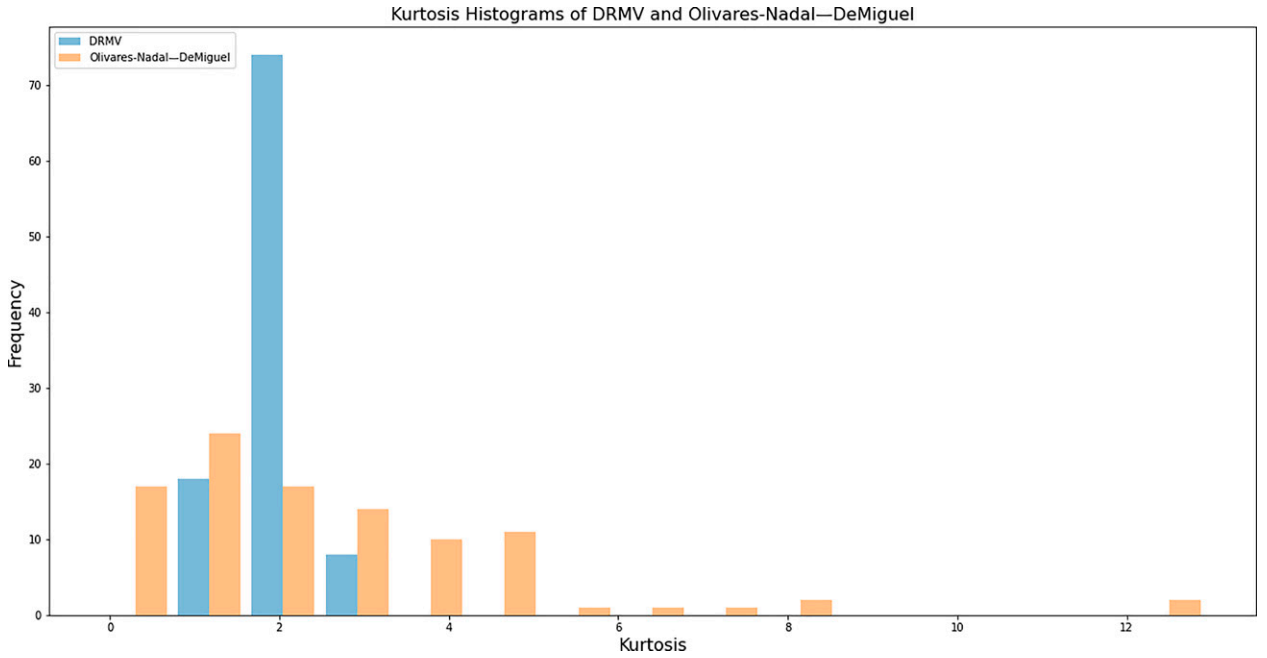
5.3.2. Different Targeted Returns. We also tested different (plausible/reasonable) values of the targeted return $\rho = 5\%, 15\%, 20\%$ in addition to $\rho = 10\%$, and find that DRMV maintains the same outperformance with respect to other models, and indeed, some other models become worse under higher targets. Figure 21 plots DRMV's average wealth processes under these four values of ρ when $d = 100$. They are almost identical,

Figure 10. Histograms of the Sharpe Ratios of the 100 Different Experiments on DRMV (Blue) and Olivares-Nadal–DeMiguel (Orange) Portfolios



Note. The x-axis represents the Sharpe ratios, and the y-axis represents the numbers of Sharpe ratios.

Figure 11. Histograms of the Kurtoses of the 100 Different Experiments on DRMV (Blue) and Olivares-Nadal–DeMiguel (Orange) Portfolios



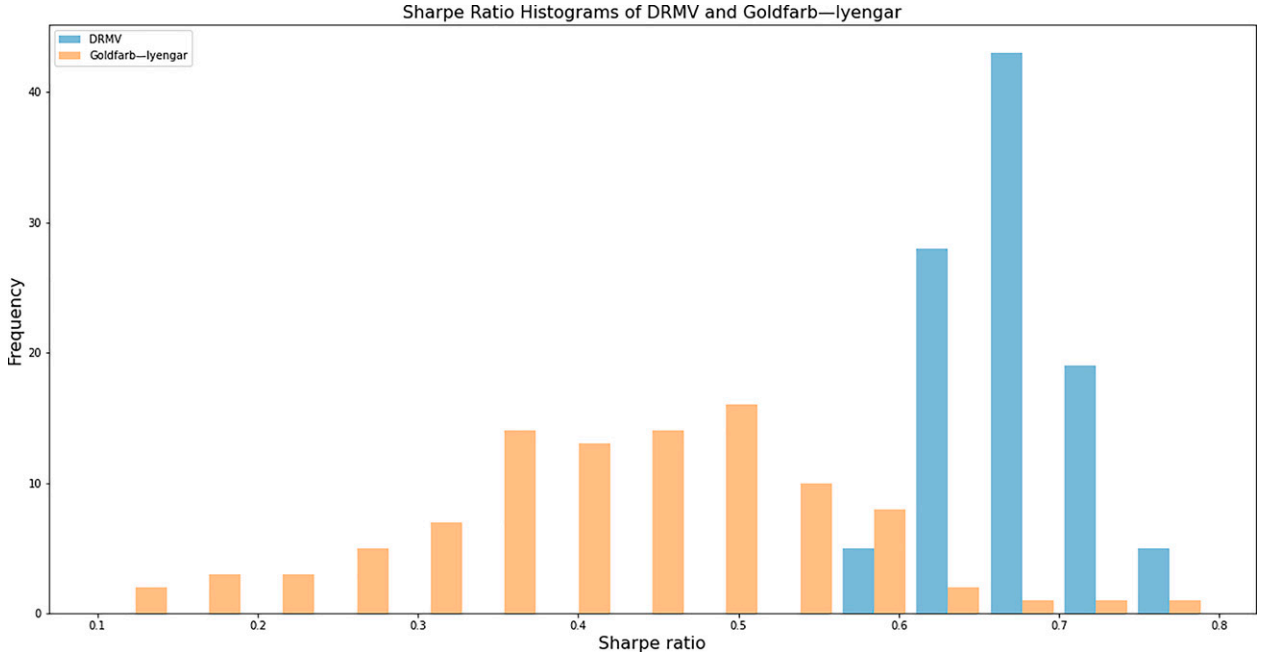
Note. The x-axis represents the kurtosis, and the y-axis represents the numbers of kurtoses.

so one could probably see only one plot in Figure 21. Hence, the average performance is very robust with different ρ's. This, in turn, suggests that the choice of a specific value of ρ is unimportant for DRMV so long as it is in the reasonable range of $[5\%, 20\%]$, thereby releasing us from the tuning and calibration of this parameter.

Figure 12. Histograms of the Annualized Returns of the 100 Different Experiments on DRMV (Blue) and Goldfarb–Iyengar (Orange) Portfolios

Notes. There are two experiments in which Goldfarb–Iyengar went bankrupt, which are not included in this histogram. The x-axis represents the annualized returns, and the y-axis represents the numbers of returns.

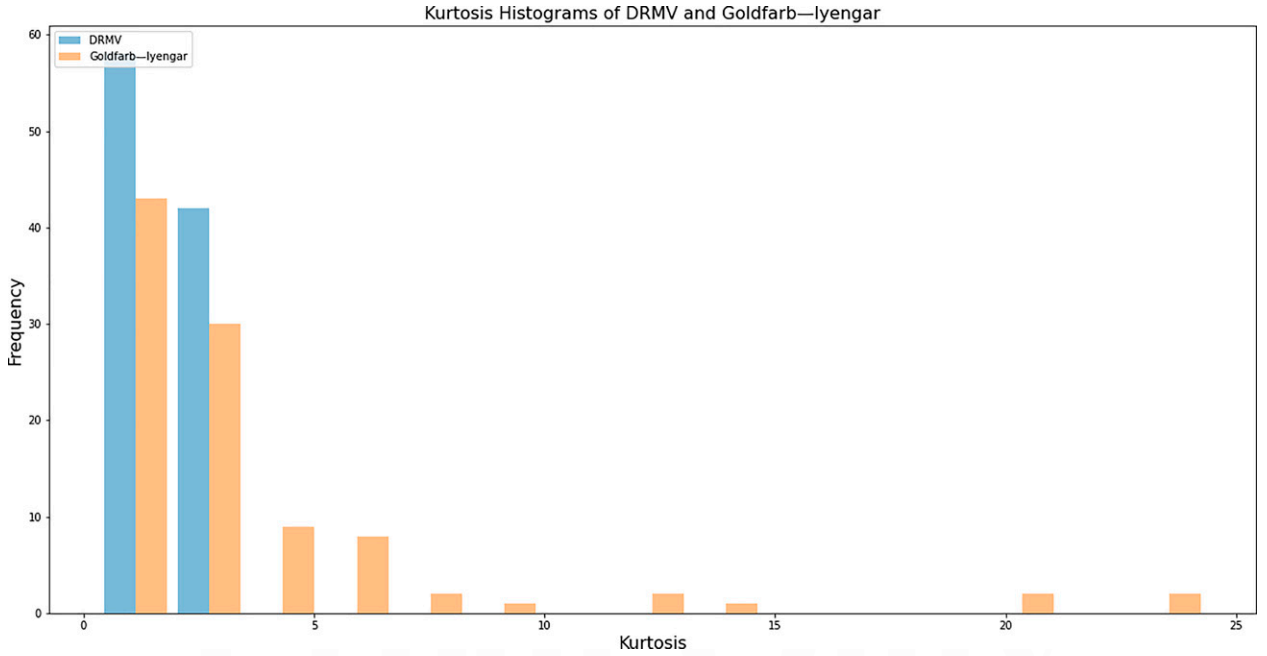
Figure 13. Histograms of the Sharpe Ratios of the 100 Different Experiments on DRMV (Blue) and Goldfarb–Iyengar (Orange) Portfolios



Note. The x-axis represents the Sharpe ratios, and the y-axis represents the numbers of Sharpe ratios.

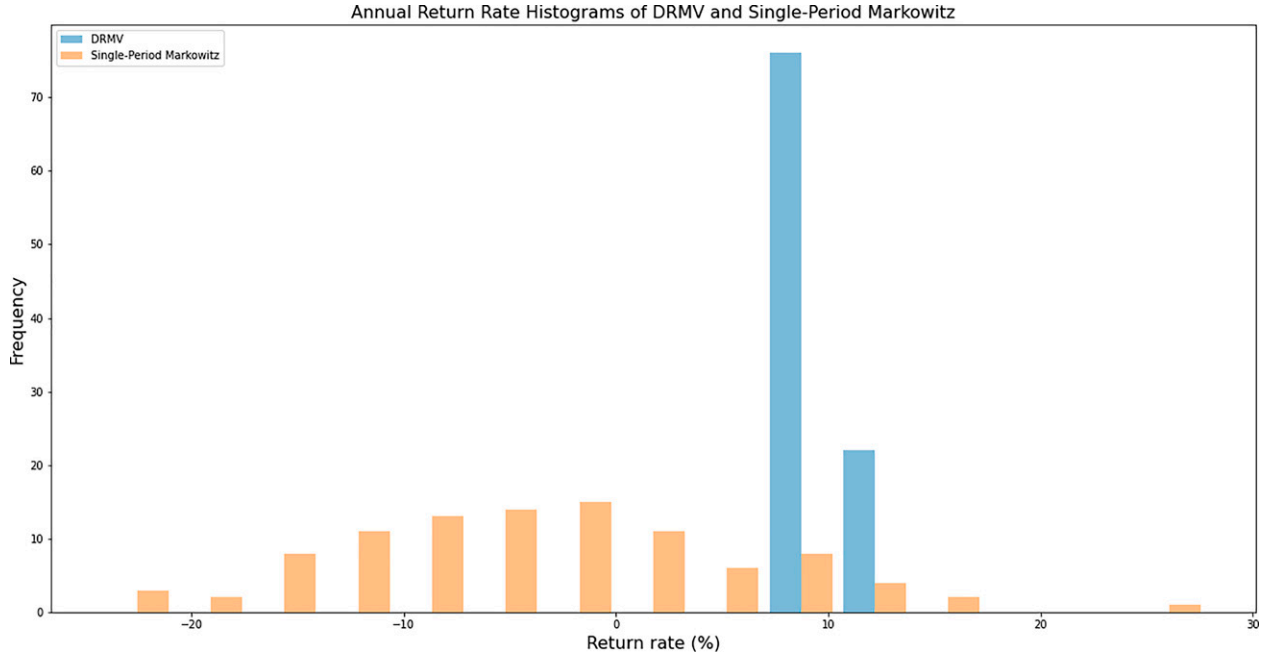
5.3.3. Turnover Rates. Fabozzi et al. (2007) observe empirically that robust portfolios have low turnover rates. This phenomenon is justified theoretically by the stipulation that robust models are equivalent to nonrobust models with transaction costs (see the discussion in Section 3), whereas transaction costs in general discourage active trades. We support this with our distributionally robust strategies. Figure 22 shows the histograms of the turnover rates (including buy and sell) with 100 experiments for $d = 100$ stocks

Figure 14. Histograms of the Kurtoses of the 100 Different Experiments on DRMV (Blue) and Goldfarb–Iyengar (Orange) Portfolios



Note. The x-axis represents the kurtosis, and the y-axis represents the numbers of kurtoses.

Figure 15. Histograms of the Annualized Returns of the 100 Different Experiments on DRMV (Blue) and Single-Period Markowitz (Orange) Portfolios

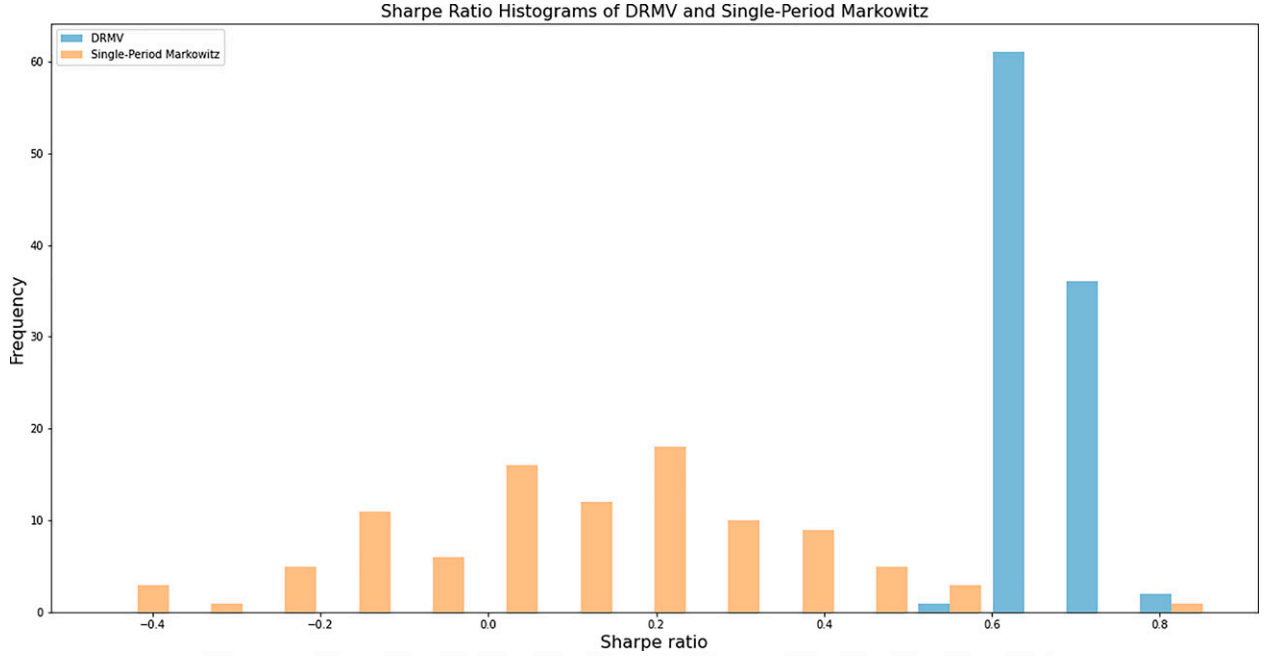


Note. The *x*-axis represents the annualized returns, and the *y*-axis represents the numbers of returns.

for both DRMV and the equally weighted portfolio. We include the latter as it is known to have low turnover and is a special instance of distributionally robust portfolios.

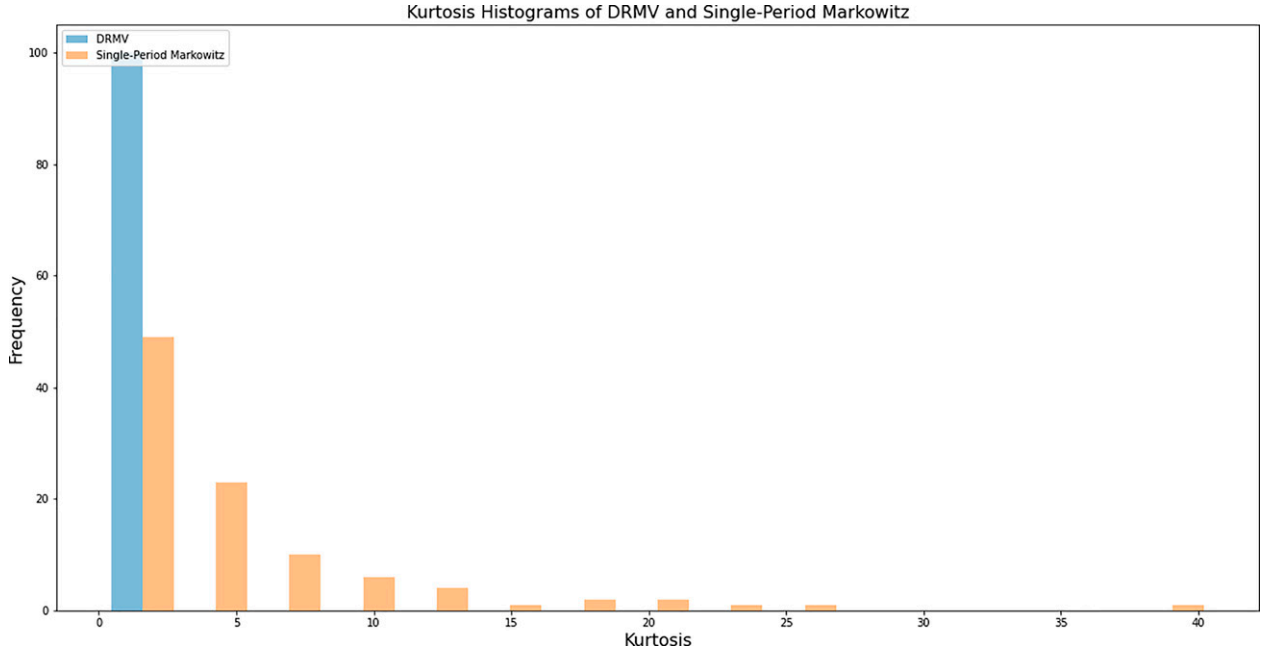
Clearly, our result reconciles with the finding of Fabozzi et al. (2007). Indeed, the two histograms are almost identical, and most of the monthly turnover rates are lower than 10%, which is considered to be

Figure 16. Histograms of the Sharpe Ratios of the 100 Different Experiments on DRMV (Blue) and Single-Period Markowitz (Orange) Portfolios



Note. The *x*-axis represents the Sharpe ratios, and the *y*-axis represents the numbers of Sharpe ratios.

Figure 17. Histograms of the Kurtoses of the 100 Different Experiments on DRMV (Blue) and Single-Period Markowitz (Orange) Portfolios

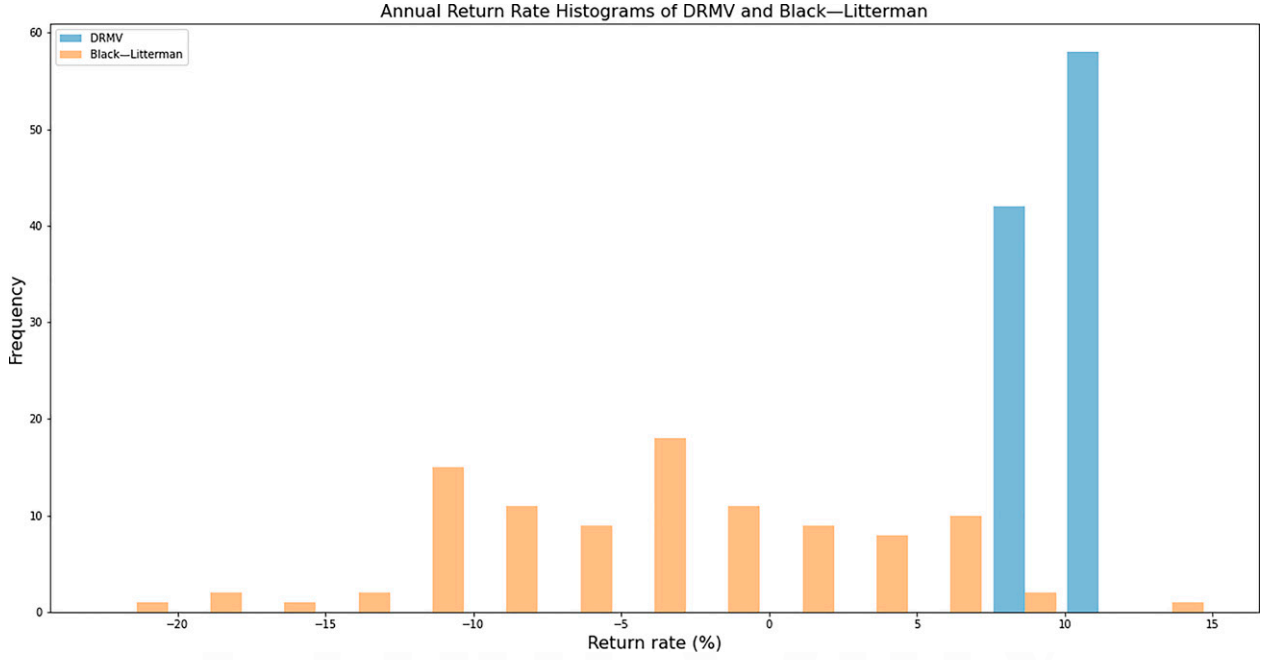


Note. The *x*-axis represents the kurtosis, and the *y*-axis represents the numbers of kurtoses.

very good and reasonable in practice. If we take the average monthly turnover rate to be 10% (definitely an upper bound for the average), then the annual turnover rate is around only 120%.

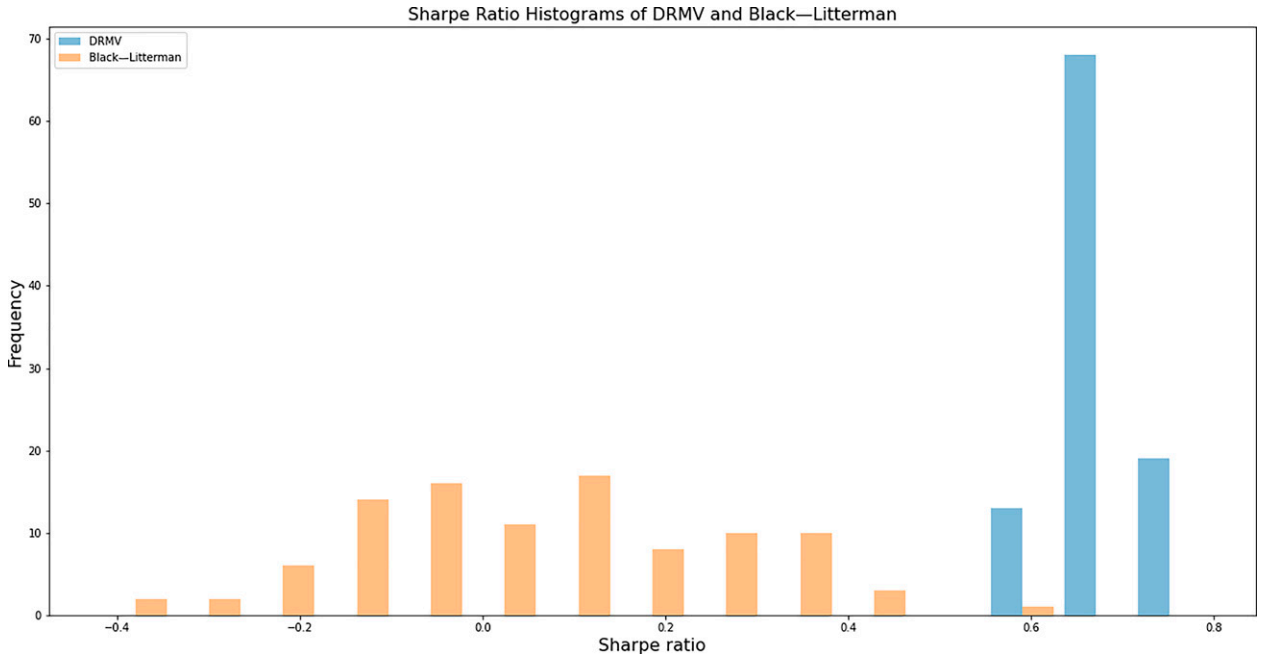
5.3.4. Shrinkage Estimators. In all the experiments reported so far, sample covariance was used for the single-period Markovitz model as well as the Black–Litterman model. Upon the recommendation of

Figure 18. Histograms of the Annualized Returns of the 100 Different Experiments on DRMV (Blue) and Black–Litterman (Orange) Portfolios



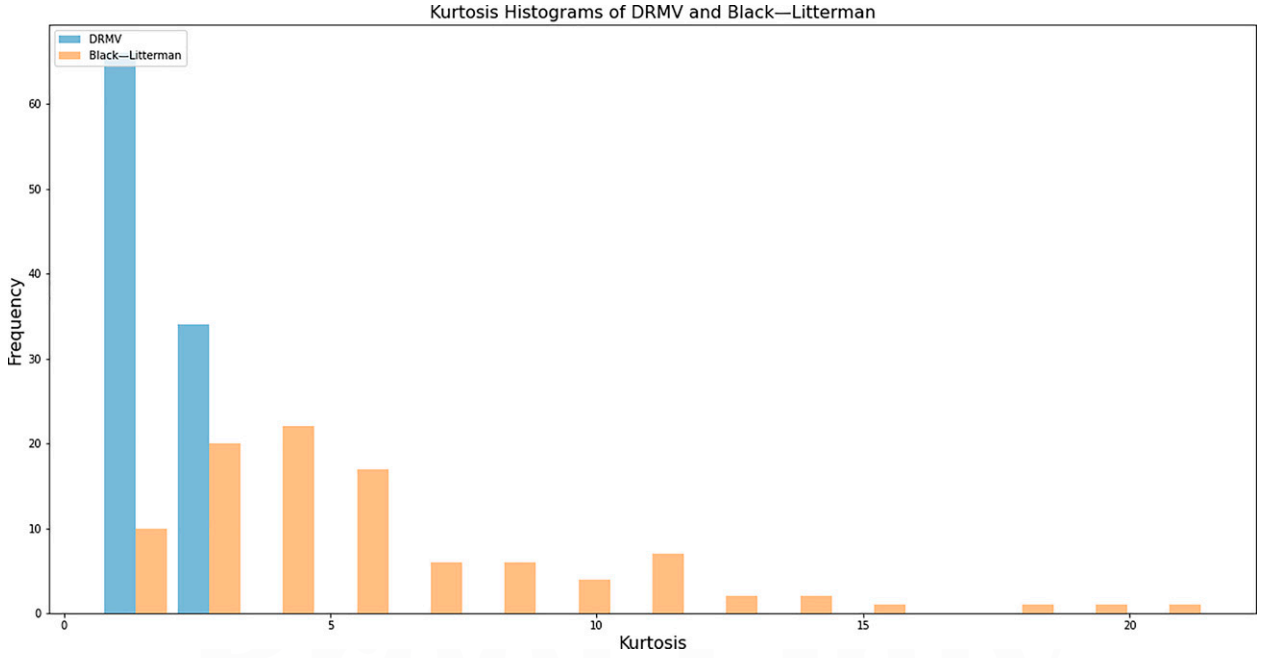
Note. The *x*-axis represents the annualized returns, and the *y*-axis represents the numbers of returns.

Figure 19. Histograms of the Sharpe Ratios of the 100 Different Experiments on DRMV (Blue) and Black–Litterman (Orange) Portfolios



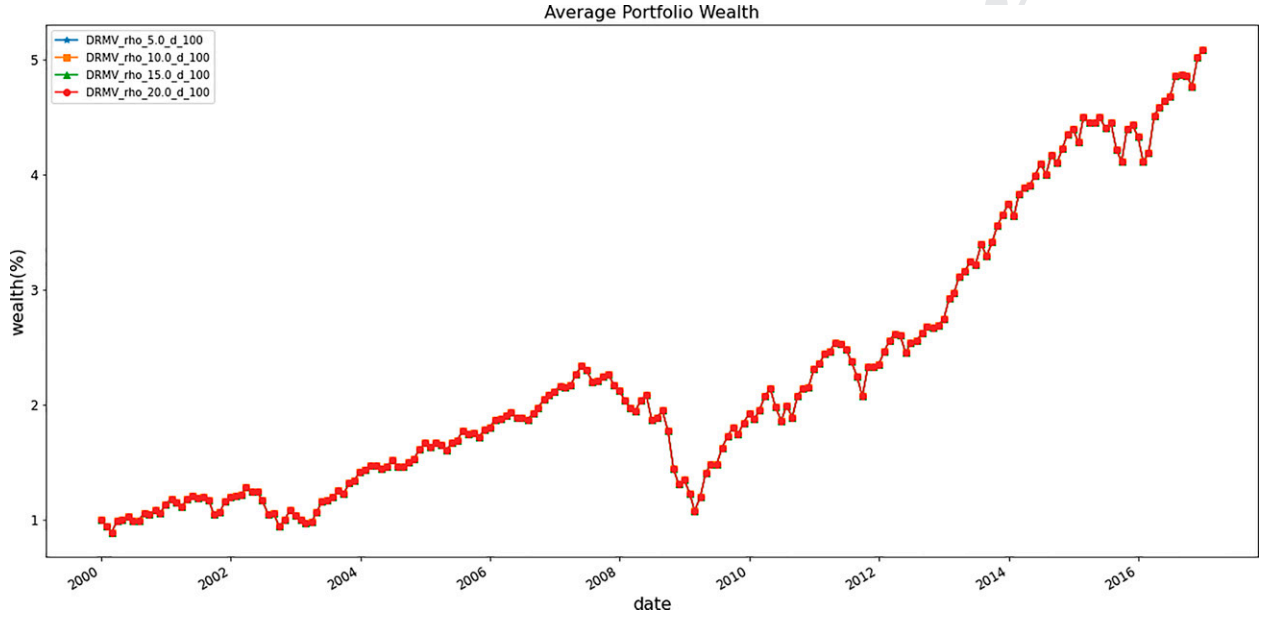
Note. The *x*-axis represents the Sharpe ratios, and the *y*-axis represents the numbers of Sharpe ratios.

Figure 20. Histograms of the kurtoses of the 100 Different Experiments on DRMV (Blue) and Black–Litterman (Orange) Portfolios



Note. The *x*-axis represents the kurtosis, and the *y*-axis represents the numbers of kurtoses.

Figure 21. DRMV's Average Wealth Processes from January 2000 to December 2016 with Different Values of ρ



Notes. The averages are calculated over 100 numerical experiments. The x-axis indicates the time, and the y-axis indicates the portfolio wealth. Initial wealth is set to be one.

one of the referees of an earlier version of the paper, we tested using the shrinkage covariance matrices for these two models.

We used the shrinkage estimator of Ollila and Raninen (2018):

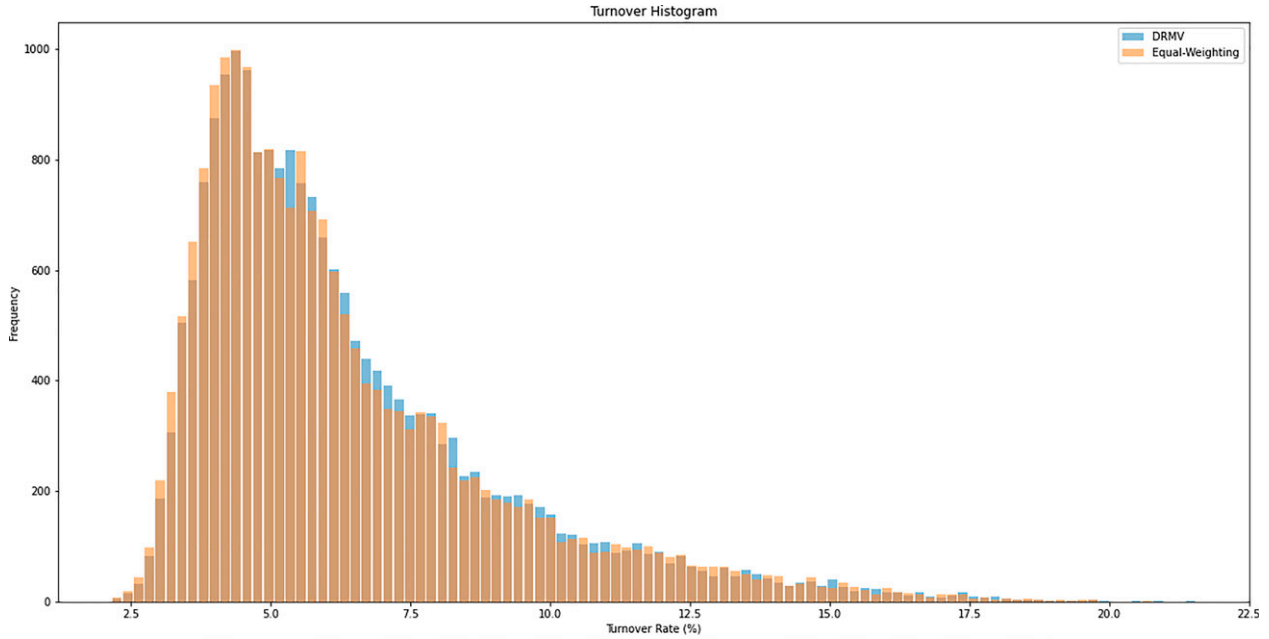
$$\Sigma_{\alpha,\beta} = \beta\Sigma_n + \alpha I_n, \tag{21}$$

where Σ_n is the sample covariance matrix and I_n is the $n \times n$ identity matrix. The parameters α and β are estimated by the following:

$$\hat{\alpha} = (1-\hat{\beta})\hat{\eta},$$

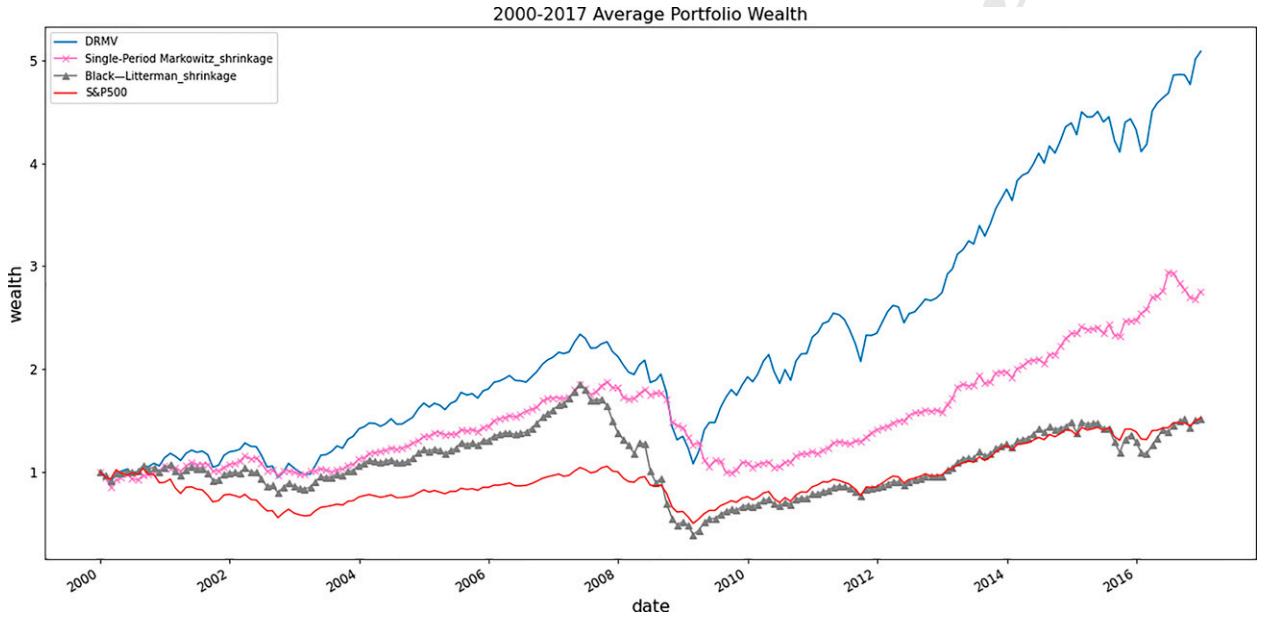
$$\hat{\beta} = \frac{(\hat{\gamma}-1)}{(\hat{\gamma}-1)+\hat{\kappa}(2\hat{\gamma}+d)/n+(\hat{\gamma}+d)/(n-1)},$$

Figure 22. Histograms of the DRMV (Blue) and Equally Weighted Portfolio (Orange) Monthly Turnover Rates of 100 Experiments



Note. The x-axis represents the turnover rate (in percentage), and the y-axis represents the numbers of turnover rates.

Figure 23. Portfolios' Average Wealth Processes Using Shrinkage Estimators from January 2000 to December 2016



Notes. All the portfolios except S&P 500 consist of 100 stocks, and the averages are calculated over 100 numerical experiments. The x-axis indicates the time in months (from 1 to 204), and the y-axis indicates the portfolio wealth. Initial wealth is set to be one.

where d is the number of stocks, $\hat{\eta} = \frac{tr(\Sigma_n)}{d}$, $\hat{\gamma} = \frac{nd}{n-1}\left[tr\left(\Sigma_{sgn}^2\right) - \frac{1}{n}\right]$ with $\Sigma_{sgn} = \frac{1}{n}\sum_{i=1}^{n} \frac{(R_i-\hat{\mu})(R_i-\hat{\mu})^T}{\|R_i-\hat{\mu}\|}$ and $\hat{\mu} = \arg\min_{\mu} \sum_{i=1}^{n} \|R_i - \mu\|$, $\hat{\kappa} = \max\left(-\frac{2}{d+2}, \frac{1}{3d}\sum_{j=1}^{d}\hat{K}_j\right)$ with $\hat{K}_j = \frac{n-1}{(n-2)(n-3)}[(n+1)\hat{k}_j + 6]$ and $\hat{k}_j = \frac{m_j^{(4)}}{(m_j^{(2)})^2} - 3$, $m_j^{(q)}$ denoting the qth order sample moment.

The two models, single-period Markowitz and Black–Litterman, are both improved with the shrinkage estimators. Figure 23 presents the comparison of average wealth processes between the two models, DRMV and S&P 500. Notably, the Black–Litterman model now outperforms S&P 500 most of the time as opposed to that without shrinkage (see Figure 2). However, DRMV still dominates the two models. Histograms of return distributions and Sharpe ratio distributions show the superiority of DRMV over the other models similar to those without shrinkage.[19]

6. Concluding Remarks

We provide a data-driven distributionally robust theory for Markowitz's mean–variance portfolio selection. The robust model can be solved via a nonrobust one based on the empirical probability measure with an additional regularization term. The size of the distributional uncertainty region is not exogenously given; rather, it is informed by the return data in a scheme that we have developed in this paper.

Our results may be generalized in different directions. We chose the l_q norm in defining our Wasserstein distance because of its popularity in regularization, but other transportation costs can be used. For example, one may consider the type of transportation cost related to adaptive regularization that is studied by Blanchet et al. (2017) or the one related to industry cluster as in Blanchet and Kang (2017). Another significant direction is a dynamic (discrete- or continuous-time) version of the DRMV model.

Appendix A. Proof of Theorem 1

The first step is to show that the feasible region over ϕ in the outer minimization part of Problem (2) can be explicitly evaluated. This is given in the following proposition.

Proposition A.1. *For $c(u,v) = \|u-v\|_q^2$, $q \geq 1$, we have* **Q: 11**

$$\min_{\mathbb{P}\in\mathcal{U}_\delta(\mathbb{P}_n)} \mathbb{E}_{\mathbb{P}}(\phi^\top R) = \mathbb{E}_{\mathbb{P}_n}(\phi^\top R) - \sqrt{\delta}\|\phi\|_p, \tag{A.1}$$

where p satisfies $1/p + 1/q = 1$.

Proof. We consider the following problem

$$\min_{\mathbb{P}\in D_c(\mathbb{P},\mathbb{P}_n)\leq\delta} \phi^\top \mathbb{E}_{\mathbb{P}}[R] \tag{A.2}$$

or, equivalently,

$$-\max_{\mathbb{P}\in D_c(\mathbb{P},\mathbb{P}_n)\leq\delta} \mathbb{E}_{\mathbb{P}}[(-\phi)^\top R]. \tag{A.3}$$

By checking Slater's condition and using proposition 4 of Blanchet et al. (2016), we obtain the dual problem:

$$\max_{\mathbb{P}\in D_c(\mathbb{P},\mathbb{P}_n)\le\delta}\mathbb{E}_{\mathbb{P}}[(-\phi)^\top R]=\inf_{\lambda\ge 0}\left[\lambda\delta+\frac{1}{n}\sum_{i=1}^{n}\Phi_\lambda(R_i)\right], \tag{A.4}$$

where

$$\begin{aligned}\Phi_\lambda(R_i)&=\sup_u\{h(u)-\lambda c(u,R_i)\}\\&=\sup_u\left\{(-\phi^\top)u-\lambda\|u-R_i\|_q^2\right\}\\&=\sup_\Delta\left\{(-\phi^\top)(\Delta+R_i)-\lambda\|\Delta\|_q^2\right\}\\&=\sup_\Delta\left\{(-\phi^\top)\Delta-\lambda\|\Delta\|_q^2\right\}-\phi^\top R_i\\&=\sup_\Delta\left\{\|\phi\|_p\|\Delta\|_q-\lambda\|\Delta\|_q^2\right\}-\phi^\top R_i\\&=\frac{\|\phi\|_p^2}{4\lambda}-\phi^\top R_i.\end{aligned}$$

Thus, (A.4) becomes

$$\begin{aligned}\max_{\mathbb{P}\in D_c(\mathbb{P},\mathbb{P}_n)\le\delta}\mathbb{E}_{\mathbb{P}}[(-\phi)^\top R]&=\inf_{\lambda\ge 0}\left\{\lambda\delta+\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\|\phi\|_p^2}{4\lambda}-\phi^\top R_i\right]\right\}\\&=\inf_{\lambda\ge 0}\left\{\lambda\delta+\frac{\|\phi\|_p^2}{4\lambda}-\phi^\top\mathbb{E}_{\mathbb{P}_n}[R]\right\}\\&=\sqrt{\delta}\|\phi\|_p-\phi^\top\mathbb{E}_{\mathbb{P}_n}[R]\end{aligned}$$

or

$$\min_{\mathbb{P}\in D_c(\mathbb{P},\mathbb{P}_n)\le\delta}\phi^\top\mathbb{E}_{\mathbb{P}}[R]=\phi^\top\mathbb{E}_{\mathbb{P}_n}[R]-\sqrt{\delta}\|\phi\|_p. \tag{A.5}$$

■

Therefore, the feasible region can be rewritten as

$$\mathcal{F}_{\delta,\bar\alpha}(n)=\left\{\phi\in\mathbb{R}^d:\phi^\top 1=1,\mathbb{E}_{\mathbb{P}_n}(\phi^\top R)\ge\bar\alpha+\sqrt{\delta}\|\phi\|_p\right\},$$

which can now be seen as clearly convex.

Next, by fixing $\mathbb{E}_{\mathbb{P}}(\phi^\top R)=\alpha\ge\bar\alpha$ in the inner maximization part of Problem (2), we obtain the following equivalent formulation:

$$\min_{\phi\in\mathcal{F}_{\delta,\bar\alpha}}\left\{\max_{\alpha\ge\bar\alpha}\left[\max_{\mathbb{P}\in\mathcal{U}_\delta(\mathbb{P}_n),\mathbb{E}_{\mathbb{P}}(\phi^\top R)=\alpha}\{\phi^\top\mathbb{E}_{\mathbb{P}}(RR^\top)\phi\}-\alpha^2\right]\right\}. \tag{A.6}$$

Introducing $\mathbb{E}_{\mathbb{P}}(\phi^\top R)=\alpha$ is useful because the innermost maximization problem in the preceding is now linear in $\mathbb{P}$. So let us concentrate on the problem

$$\max_{\mathbb{P}\in\mathcal{U}_\delta(\mathbb{P}_n),\,\mathbb{E}_{\mathbb{P}}(\phi^\top R)=\alpha}\phi^\top\mathbb{E}_{\mathbb{P}}(RR^\top)\phi. \tag{A.7}$$

The following proposition solves this problem in terms of a general cost function c.

Proposition A.2. *For any cost function c that is lower semicontinuous and nonnegative, the optimal value function of* Problem (A.7) *is given by*

$$\inf_{\lambda_1\ge 0,\,\lambda_2}\left[\frac{1}{n}\sum_{i=1}^{n}\Phi(R_i)+\lambda_1\delta+\lambda_2\alpha\right], \tag{A.8}$$

where

$$\Phi(R_i):=\sup_u\left[(\phi^\top u)^2-\lambda_1 c(u,R_i)-\lambda_2\phi^\top u\right].$$

Proof. The proof is based on a duality argument. Introducing a slack random variable $S\equiv v$, where v is a deterministic number. Then, we can recast Problem (A.7) as

$$\max\{\mathbb{E}_{\mathbb{P}}[(U^\top\phi)^2]:\mathbb{E}_\pi[c(U,R)+S]=\delta,\pi_R=\mathbb{P}_n,\pi(S=v)=1, \tag{A.9}$$

$$\mathbb{E}_\pi[U^\top\phi]=\alpha,\pi\in\mathcal{P}(\mathcal{R}^m\times\mathcal{R}^m\times\mathcal{R}_+)\}. \tag{A.10}$$

Define

$$\Omega:=\{(u,r,s):c(u,r)<\infty,s\ge 0,r\in\{R_1,\dots,R_n\}\},$$

and let

$$f(u,r,s)=\begin{bmatrix}1_{r=R_1}(u,r,s)\\\cdots\\1_{r=R_n}(u,r,s)\\\phi^\top u\\1_{s=v}(u,r,s)\\c(u,r)+s\end{bmatrix}\quad\text{and}\quad q=\begin{bmatrix}\frac{1}{n}\\\cdots\\\frac{1}{n}\\\alpha\\1\\\delta\end{bmatrix}. \tag{A.11}$$

Thus, (A.9) can be written as

$$\max\{\mathbb{E}_\pi[(U^\top\phi)^2]:\mathbb{E}_\pi[f(U,R,S)]=q,\pi\in\mathcal{P}_\Omega\}. \tag{A.12}$$

Let $f_0=\mathbf{1}_\Omega,\tilde f=(f_0,f),\tilde q=(1,q)$, $\mathcal{Q}_{\tilde f}:=\{\int\tilde f(x)d\mu(x):\mu\in\mathcal{M}_\Omega^+\}$, where $\mathcal{M}_\Omega^+$ denotes the set of nonnegative measures on Ω. If $\phi\ne 0$, then it is easy to see that $\tilde q$ lies in the interior of $\mathcal{Q}_{\tilde f}$. By proposition 6 in Blanchet et al. (2016), the optimal value of Problem (A.12) equals that of its dual problem, that is,

$$\begin{aligned}&\max\{\mathbb{E}_\pi[(U^\top\phi)^2]:\mathbb{E}_\pi[f(U,R,S)]=q,\pi\in\mathcal{P}_\Omega\}\\&=\inf_{a=(a_0,\dots,a_{n+3})\in A}\left\{a_0+\frac{1}{n}\sum_{i=1}^{n}a_i+\alpha a_{n+1}+a_{n+2}+\delta a_{n+3}\right\},\end{aligned} \tag{A.13}$$

where

$$\begin{aligned}A:=\Bigg\{&a=(a_0,\dots,a_{n+3})\in\mathbb{R}^{n+4}:a_0+\frac{1}{n}\sum_{i=1}^{n}a_i1_{r=R_i}(u,r,s)+a_{n+1}\phi^\top u\\&+a_{n+2}1_{s=v}(u,r,s)+a_{n+3}[c(u,r)+s]\ge(\phi^\top u)^2,\ \forall(u,r,s)\in\Omega\Bigg\}.\end{aligned}$$

From the definition of A, replacing $r=R_i$, we obtain that the inequality

$$a_0+a_i+a_{n+2}\ge\sup_{(u,s)\in\Omega}\left\{(\phi^\top u)^2-a_{n+3}[c(u,R_i)+s]-a_{n+1}\phi^\top u\right\} \tag{A.14}$$

holds for each $i\in\{1,\dots,n\}$. It follows directly that

$$\sup_{(u,s)\in\Omega}\left\{(\phi^\top u)^2-a_{n+3}[c(u,R_i)+s]-a_{n+1}\phi^\top u\right\} \tag{A.15}$$

$$= \begin{cases} +\infty, & \text{if } a_{n+3} < 0 \\ \sup_u \left\{(\phi^\top u)^2 - a_{n+3}c(u,R_i) - a_{n+1}\phi^\top u\right\}, & \text{if } a_{n+3} \geq 0. \end{cases} \tag{A.16}$$

Thus, the dual problem can be expressed as

$$\inf\left\{a_0 + \frac{1}{n}\sum_{i=1}^n a_i + \alpha a_{n+1} + a_{n+2} + \delta a_{n+3} : a_{n+3} \geq 0, a_0 + a_i + a_{n+2} \geq \sup_u \left\{(\phi^\top u)^2 - a_{n+3}c(u,R_i) - a_{n+1}\phi^\top u\right\}\right\}, \tag{A.17}$$

which can be transformed into

$$\inf_{a_{n+3}\geq 0}\left\{\frac{1}{n}\sum_{i=1}^n \Phi(R_i) + \alpha a_{n+1} + \delta a_{n+3}\right\}, \tag{A.18}$$

with

$$\Phi(R_i) := \sup_u \left\{(\phi^\top u)^2 - a_{n+3}c(u,R_i) - a_{n+1}\phi^\top u\right\}.$$

Using λ_1 to replace a_{n+3} and λ_2 to replace a_{n+1}, the dual problem becomes

$$\inf_{\lambda_1\geq 0}\left\{\frac{1}{n}\sum_{i=1}^n \Phi(R_i) + \lambda_2\alpha + \lambda_1\delta\right\} \tag{A.19}$$

where

$$\Phi(R_i) := \sup_u \left\{(\phi^\top u)^2 - \lambda_1 c(u,R_i) - \lambda_2\phi^\top u\right\}.$$

■

Thanks to this proposition, we are able to reduce the inner (infinite dimensional) optimization problem in (2) into a two-dimensional optimization problem in terms of λ_1 and λ_2, which can be further simplified if the cost function c has additional structure. We make this statement precise in the case of a quadratic l_q cost.

Proposition A.3. *Let $c(u,v) = \|u-v\|_q^2$ with $q \geq 1$ and $1/p + 1/q = 1$. If $(\alpha - \phi^\top \mathbb{E}_{\mathbb{P}_n}[R])^2 - \delta\|\phi\|_p^2 \leq 0$, then the value of (A.7) is equal to*

$$h(\alpha,\phi) := \mathbb{E}_{\mathbb{P}_n}\left[(\phi^\top R)^2\right] + 2(\alpha - \phi^\top \mathbb{E}_{\mathbb{P}_n}[R])\phi^\top \mathbb{E}_{\mathbb{P}_n}[R] + \delta\|\phi\|_p^2 + 2\sqrt{\delta\|\phi\|_p^2 - (\alpha - \phi^\top \mathbb{E}_{\mathbb{P}_n}[R])^2}\sqrt{\phi^\top \mathrm{Var}_{\mathbb{P}_n}(R)\phi}.$$

Proof. Writing $\Delta := u - R_i$, we have

$$\begin{aligned}
\Phi(R_i) &= \sup_u \left\{(\phi^\top u)^2 - \lambda_1 c(u,R_i) - \lambda_2\phi^\top u\right\} \\
&= \sup_u \left\{(\phi^\top u)^2 - \lambda_1\|u-R_i\|_q^2 - \lambda_2\phi^\top u\right\} \\
&= \sup_\Delta \left\{(\phi^\top(\Delta+R_i))^2 - \lambda_1\|\Delta\|_q^2 - \lambda_2\phi^\top(R_i+\Delta)\right\} \\
&= \sup_\Delta \left\{(\phi^\top R_i)^2 + (\phi^\top\Delta)^2 + 2(\phi^\top R_i)(\phi^\top\Delta) - \lambda_1\|\Delta\|_q^2 - \lambda_2\phi^\top(R_i+\Delta)\right\} \\
&= (\phi^\top R_i)^2 - \lambda_2\phi^\top R_i + \sup_\Delta \left\{(\phi^\top\Delta)^2 + 2(\phi^\top R_i)(\phi^\top\Delta) - \lambda_1\|\Delta\|_q^2 - \lambda_2\phi^\top\Delta\right\} \\
&= (\phi^\top R_i)^2 - \lambda_2\phi^\top R_i + \sup_\Delta \left\{(\|\phi\|_p^2 - \lambda_1)\|\Delta\|_q^2 + |2(R_i^\top\phi) - \lambda_2|(\|\phi\|_p\|\Delta\|_q)\right\}.
\end{aligned}$$

We can consider four cases: (1) $\|\phi\|_p^2 > \lambda_1$, $\Phi(R_i) = +\infty$; (2) $\|\phi\|_p^2 = \lambda_1$, $2R_i^\top\phi \neq \lambda_2$, $\Phi(R_i) = +\infty$; (3) $\|\phi\|_p^2 = \lambda_1$, $2R_i^\top\phi = \lambda_2$, $\Phi(R_i) = 0$; (4) $\|\phi\|_p^2 < \lambda_1$, $\Phi(R_i) = (\phi^\top R_i)^2 - \lambda_2\phi^\top R_i + \frac{(2R_i^\top\phi - \lambda_2)^2\|\phi\|_p^2}{4(\lambda_1 - \|\phi\|_p^2)}$.

For any of the first three cases, the value of $\frac{1}{n}\sum_{i=1}^n \Phi(R_i)$ is $+\infty$. Hence, only the fourth case is nontrivial. In this case, Problem (A.8) is transformed into

$$\begin{aligned}
&\inf_{\lambda_1\geq 0,\lambda_2}\left[\frac{1}{n}\sum_{i=1}^n \Phi(R_i) + \lambda_2\alpha + \lambda_1\delta\right] \\
&= \inf_{\lambda_1\geq\|\phi\|_p^2,\lambda_2}\left\{\frac{1}{n}\sum_{i=1}^n\left[(\phi^\top R_i)^2 - \lambda_2\phi^\top R_i + \frac{(2R_i^\top\phi - \lambda_2)^2\|\phi\|_p^2}{4(\lambda_1 - \|\phi\|_p^2)}\right] + \lambda_2\alpha + \lambda_1\delta\right\}
\end{aligned} \tag{A.20}$$

Define

$$H = \frac{1}{n}\sum_{i=1}^n\left[(\phi^\top R_i)^2 - \lambda_2\phi^\top R_i + \frac{(2R_i^\top\phi - \lambda_2)^2\|\phi\|_p^2}{4\left(\lambda_1 - \|\phi\|_p^2\right)}\right] + \lambda_2\alpha + \lambda_1\delta.$$

Taking a partial derivative with respect to λ_2 and setting it to be zero, we get

$$\frac{\partial H}{\partial\lambda_2} = \alpha - \frac{1}{n}\sum_{i=1}^n\left[\phi^\top R_i + \frac{(2\phi^\top R_i - \lambda_2)\|\phi\|_p^2}{2(\lambda_1 - \|\phi\|_p^2)}\right] = 0,$$

which implies (note that $\phi^\top 1 = 1$ guarantees that $\|\phi\|_p^2 > 0$)

$$\lambda_2 = 2\alpha - 2C\frac{\lambda_1}{\|\phi\|_p^2}, \tag{A.21}$$

where $C := \alpha - \phi^\top\mathbb{E}_{\mathbb{P}_n}[R]$. Moreover, λ_2 is optimal because

$$\frac{\partial^2 H}{\partial\lambda_2^2} = \frac{\|\phi\|_p^2}{2(\lambda_1 - \|\phi\|_p^2)} > 0. \tag{A.22}$$

We plug (A.21) into (A.20) and obtain

$$\begin{aligned}
&\inf_{\lambda_1\geq 0,\lambda_2}\left[\frac{1}{n}\sum_{i=1}^n \Phi(R_i) + \lambda_2\alpha + \lambda_1\delta\right] \\
&= \frac{1}{n}\sum_{i=1}^n(\phi^\top R_i)^2 + \inf_{\lambda_1\geq\|\phi\|_p^2,\lambda_2} \left\{\frac{1}{n}\sum_{i=1}^n\left[-\lambda_2\phi^\top R_i + \frac{(2R_i^\top\phi - \lambda_2)^2\|\phi\|_p^2}{4(\lambda_1 - \|\phi\|_p^2)}\right] + \lambda_2\alpha + \lambda_1\delta\right\} \\
&= \frac{1}{n}\sum_{i=1}^n(\phi^\top R_i)^2 + \inf_{\lambda_1\geq\|\phi\|_p^2,\lambda_2} \left\{\frac{1}{n}\sum_{i=1}^n\left[\frac{(2R_i^\top\phi - \lambda_2)^2\|\phi\|_p^2}{4(\lambda_1 - \|\phi\|_p^2)}\right] + \lambda_2 C + \lambda_1\delta\right\} \\
&= \frac{1}{n}\sum_{i=1}^n(\phi^\top R_i)^2 + \inf_{\lambda_1\geq\|\phi\|_p^2} \left\{\frac{1}{n}\sum_{i=1}^n\left[\frac{\left(2R_i^\top\phi - 2\alpha + 2C\frac{\lambda_1}{\|\phi\|_p^2}\right)^2\|\phi\|_p^2}{4(\lambda_1 - \|\phi\|_p^2)}\right] + \left(2\alpha - 2C\frac{\lambda_1}{\|\phi\|_p^2}\right)C + \lambda_1\delta\right\}.
\end{aligned}$$

Writing $\lambda_1 = \kappa + \|\phi\|_p^2$, we have

$$\begin{aligned}
&\inf_{\lambda_1\geq 0,\lambda_2}\left[\frac{1}{n}\sum_{i=1}^n \Phi(R_i)+\lambda_2\alpha+\lambda_1\delta\right]=\frac{1}{n}\sum_{i=1}^n(\phi^\top R_i)^2\\
&\quad+\inf_{\kappa\geq 0}\left\{\frac{1}{n}\sum_{i=1}^n\left[\frac{\left(R_i^\top\phi-\alpha+C\frac{k+\|\phi\|_p^2}{\|\phi\|_p^2}\right)^2 N}{\kappa}\right]+\left(2\alpha-2C\frac{\kappa+\|\phi\|_p^2}{\|\phi\|_p^2}\right)C+(\kappa+\|\phi\|_p^2)\delta\right\}\\
&=\frac{1}{n}\sum_{i=1}^n(\phi^\top R_i)^2+\inf_{\kappa\geq 0}\left\{\frac{C_1^2}{\|\phi\|_p^2}k+2\|\phi\|_p^2(\phi^\top\bar{W}-\alpha+C)\right.\\
&\quad\left.+\frac{1}{n}\sum_{i=1}^n\frac{(R_i^\top\phi-\alpha+C)^2\|\phi\|_p^2}{\kappa}+2\alpha C-2C^2+\kappa\left(\delta-\frac{2C^2}{\|\phi\|_p^2}\right)+\|\phi\|_p^2\delta\right\}\\
&=\frac{1}{n}\sum_{i=1}^n(\phi^\top R_i)^2+2\alpha C-2C^2+|\phi\|_p^2\delta+\inf_{\kappa\geq 0}\\
&\quad\left\{\frac{1}{n}\sum_{i=1}^n\frac{(R_i^\top\phi-\mathbb{E}_{\mathbb{P}_n}[R]\phi)^2\|\phi\|_p^2}{\kappa}+\kappa\left(\delta-\frac{C^2}{\|\phi\|_p^2}\right)\right\}.
\end{aligned}$$

If $\delta - C^2/\|\phi\|_p^2 < 0$, then the optimal value of the preceding problem is $-\infty$, which means that the primal Problem (A.7) is not feasible. If $\delta - C^2/\|\phi\|_p^2 \geq 0$, then

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n(\phi^\top R_i)^2+2\alpha C-2C^2+|\phi\|_p^2\delta\\
&\quad+\inf_{\kappa\geq 0}\left\{\frac{1}{n}\sum_{i=1}^n\frac{(R_i^\top\phi-\mathbb{E}_{\mathbb{P}_n}[R]\phi)^2\|\phi\|_p^2}{\kappa}+\kappa\left(\delta-\frac{C^2}{\|\phi\|_p^2}\right)\right\}\\
&=\frac{1}{n}\sum_{i=1}^n(\phi^\top R_i)^2+2(\alpha-\phi^\top\mathbb{E}_{\mathbb{P}_n}[R])\phi^\top\mathbb{E}_{\mathbb{P}_n}[R]+\delta\|\phi\|_p^2\\
&\quad+2\sqrt{\delta\|\phi\|_p^2-(\alpha-\phi^\top\mathbb{E}_{\mathbb{P}_n}[R])^2}\\
&\quad\sqrt{\frac{1}{n}\phi^\top\sum_{i=1}^n(R_i-\mathbb{E}_{\mathbb{P}_n}[R])(R_i-\mathbb{E}_{\mathbb{P}_n}[R])^\top\phi}\\
&=\frac{1}{n}\sum_{i=1}^n(\phi^\top R_i)^2+2(\alpha-\phi^\top\mathbb{E}_{\mathbb{P}_n}[R])\phi^\top\mathbb{E}_{\mathbb{P}_n}[R]+\delta\|\phi\|_p^2\\
&\quad+2\sqrt{\delta\|\phi\|_p^2-(\alpha-\phi^\top\mathbb{E}_{\mathbb{P}_n}[R])^2}\sqrt{\phi^\top\mathrm{Var}_{\mathbb{P}_n}[R]\phi}.
\end{aligned}$$

Thus, Problem (A.7) can be written as

$$\begin{aligned}
&\min_{\phi}\ \frac{1}{n}\sum_{i=1}^n(\phi^\top R_i)^2+2(\alpha-\phi^\top\mathbb{E}_{\mathbb{P}_n}[R])\phi^\top\mathbb{E}_{\mathbb{P}_n}[R]+\delta\|\phi\|_p^2\\
&+2\sqrt{\delta\|\phi\|_p^2-(\alpha-\phi^\top\mathbb{E}_{\mathbb{P}_n}[R])^2}\sqrt{\phi^\top\mathrm{Var}_{\mathbb{P}_n}[R]\phi},
\end{aligned}$$

subject to $1^\top\phi = 1$ and $(\alpha-\phi^\top\mathbb{E}_{\mathbb{P}_n}[R])^2-\delta\|\phi\|_p^2\leq 0$. $\quad\nu$

The condition $(\alpha-\phi^\top\mathbb{E}_{\mathbb{P}_n}[R])^2-\delta\|\phi\|_p^2\leq 0$ is to make sure that (A.7) is feasible, failing which the optimal value $h(\alpha,\phi) = -\infty$. Proposition A.3 ultimately leads to the following main result of the paper, one that transforms (2) into a nonrobust portfolio selection problem in terms of the empirical measure $\mathbb{P}_n$ with an additional regularization term.

We are now ready to prove Theorem 1.

Proof of Theorem 1. Note that

$$\begin{aligned}
&h(\alpha,\phi)-\alpha^2\\
&=\mathbb{E}_{\mathbb{P}_n}\left[(\phi^\top R)^2\right]+2(\alpha-\phi^\top\mathbb{E}_{\mathbb{P}_n}[R])\phi^\top\mathbb{E}_{\mathbb{P}_n}[R]-\alpha^2+\delta\|\phi\|_p^2\\
&\quad+2\sqrt{\delta\|\phi\|_p^2-(\alpha-\phi^\top\mathbb{E}_{\mathbb{P}_n}[R])^2}\sqrt{\phi^\top\mathrm{Var}_{\mathbb{P}_n}(R)\phi}\\
&=\mathbb{E}_{\mathbb{P}_n}\left[(\phi^\top R)^2\right]+2\alpha\phi^\top\mathbb{E}_{\mathbb{P}_n}[R]\\
&\quad-(\phi^\top\mathbb{E}_{\mathbb{P}_n}[R])^2-\alpha^2-(\phi^\top\mathbb{E}_{\mathbb{P}_n}[R])^2+\delta\|\phi\|_p^2\\
&\quad+2\sqrt{\delta\|\phi\|_p^2-(\alpha-\phi^\top\mathbb{E}_{\mathbb{P}_n}[R])^2}\sqrt{\phi^\top\mathrm{Var}_{\mathbb{P}_n}(R)\phi}\\
&=\phi^\top\mathrm{Var}_{\mathbb{P}_n}(R)\phi+\{\delta\|\phi\|_p^2-(\alpha-\phi^\top\mathbb{E}_{\mathbb{P}_n}[R])^2\}\\
&\quad+2\sqrt{\delta\|\phi\|_p^2-(\alpha-\phi^\top\mathbb{E}_{\mathbb{P}_n}[R])^2}\sqrt{\phi^\top\mathrm{Var}_{\mathbb{P}_n}(R)\phi}\\
&=\left(\sqrt{\phi^\top\mathrm{Var}_{\mathbb{P}_n}(R)\phi}+\sqrt{\delta\|\phi\|_p^2-(\alpha-\phi^\top\mathbb{E}_{\mathbb{P}_n}[R])^2}\right)^2.
\end{aligned}$$

Therefore, it follows from Proposition A.3 that

$$\max_{\alpha\geq\bar{\alpha},(\alpha-\phi^\top\mathbb{E}_{\mathbb{P}_n}[R])^2-\delta\|\phi\|_p^2\leq 0}\left[h(\alpha,\phi)-\alpha^2\right]=\left(\sqrt{\phi^\top\mathrm{Var}_{\mathbb{P}_n}(R)\,\phi}+\sqrt{\delta}\|\phi\|_p\right)^2,$$

with the optimal $\alpha_{opt} = \phi^\top\mathbb{E}_{\mathbb{P}_n}[R]\geq\bar{\alpha}$. This concludes the proof.

Appendix B. Proof of Theorem 2

Define

$$h_0(R,\Sigma) = RR^\top-\Sigma \text{ and } \mathrm{h}_1(\mathrm{R},\mu) = \mathrm{R}-\mu.$$

Then, by proposition 1 of Blanchet et al. (2016), we have that, for any given μ and Σ,

$$\mathcal{R}_n(\Sigma,\mu)=\sup_{\Lambda\in\mathbb{R}^{d\times d},\,\lambda\in\mathbb{R}^d}\left\{-\mathbb{E}_{\mathbb{P}_n}\left[\sup_{u\in\mathbb{R}^d}\left\{Tr(\Lambda h_0(u,\Sigma))+\lambda^\top h_1(u,\mu)-\|u-R\|_q^2\right\}\right]\right\}.$$

Observe that

$$\begin{aligned}
&\sup_{u\in\mathbb{R}^d}\left\{Tr(\Lambda h_0(u,\Sigma))+\lambda^\top h_1(u,\mu)-\|u-R\|_q^2\right\}\\
&=\sup_{\Delta\in\mathbb{R}^d}\left\{Tr(\Lambda h_0(\Delta+R,\Sigma))+\lambda^\top h_1(\Delta+R,\mu)-\|\Delta\|_q^2\right\}\\
&=\sup_{\Delta\in\mathbb{R}^d}\left\{Tr(\Lambda[h_0(\Delta+R,\Sigma)-h_0(R,\Sigma)])+\lambda^\top\Delta-\|\Delta\|_q^2\right\}\\
&\quad+Tr(\Lambda h_0(R,\Sigma))+\lambda^\top h_1(R,\mu).
\end{aligned}$$

Moreover, let us write

$$Tr(\Lambda[h_0(\Delta+R,\Sigma)-h_0(R,\Sigma)])=\int_0^1\frac{d}{dt}Tr(\Lambda h_0(R+t\Delta))dt.$$

However,

$$\begin{aligned}
\frac{d}{dt}Tr(\Lambda h_0(R+t\Delta))&=2Tr(\Lambda(R+t\Delta)\Delta^\top)\\
&=2Tr(\Lambda R\Delta^\top)+2t\Delta^\top\Lambda\Delta.
\end{aligned}$$

Furthermore,

$$\mathbb{E}_{\mathbb{P}_n}[Tr(\Lambda h_0(R,\Sigma))]\big|_{\Sigma=\Sigma_n}=0. \tag{B.1}$$

So, we deduce

$$\begin{aligned}
\mathcal{R}_n(\Sigma_n,\mu)=\sup_{\lambda\in\mathbb{R}^d}&\left\{-\mathbb{E}_{\mathbb{P}_n}[\lambda^\top(R-\mu)]\right.\\
&\left.+\sup_{\Lambda\in\mathbb{R}^{d\times d}}\left(-\mathbb{E}_{\mathbb{P}_n}\left[\sup_{\Delta}\left\{2Tr(\Lambda R\Delta^\top)+\Delta^\top\Lambda\Delta+\lambda^\top\Delta-\|\Delta\|_q^2\right\}\right]\right)\right\}.
\end{aligned}$$

Introduce the scaling $\Delta = \bar{\Delta}/n^{1/2}$ and $\bar{\lambda} = \lambda n^{1/2}$ and $\bar{\Lambda} = \Lambda n^{1/2}$. Then, we obtain

$$\begin{aligned} &n\mathcal{R}_n(\Sigma_n, \mu_n) \\ &= \sup_{\bar{\lambda}\in\mathbb{R}^d}\left\{-n^{-1/2}\sum_{i=1}^n \bar{\lambda}^\top(R_i-\mu_n) + \sup_{\bar{\Lambda}\in\mathbb{R}^{d\times d}} \right. \\ &\quad \left.\left(-\mathbb{E}_{\mathbb{P}_n}\left[\sup_{\bar{\Delta}}\left\{2Tr\left(\bar{\Lambda}R\bar{\Delta}^\top\right) + \bar{\Delta}^\top\bar{\Lambda}\bar{\Delta}/n^{1/2} + \bar{\lambda}^\top\bar{\Delta} - \|\bar{\Delta}\|_q^2\right\}\right]\right)\right\}. \end{aligned}$$

In the proof of proposition 3 in Blanchet et al. (2016), under Assumption 2, a technique is introduced to show that $\bar{\Delta}$ and $\bar{\lambda}$ can be restricted to compact sets with high probability, and therefore, the term $\bar{\Delta}^\top\bar{\Lambda}\bar{\Delta}/n^{1/2}$ is asymptotically negligible. On the other hand,

$$\begin{aligned} &\sup_{\Delta}\left\{2Tr\left(\bar{\Delta}^\top\bar{\Lambda}R\right) + \bar{\Delta}^\top\bar{\lambda} - \|\bar{\Delta}\|_q^2\right\} \\ &= \sup_{\Delta}\left\{2\|\bar{\Lambda}R+\bar{\lambda}\|_p\|\bar{\Delta}\|_q - \|\bar{\Delta}\|_q^2\right\} = \|\bar{\Lambda}R+\bar{\lambda}\|_p^2. \end{aligned}$$

Therefore, if

$$n^{-1/2}\sum_{i=1}^n (R_i-\mu_n) \Rightarrow -Z$$

for some Z (to be characterized momentarily), then we conclude that

$$\mathcal{R}_n(\Sigma_n,\mu_n) \Rightarrow L_0 = \sup_{\bar{\lambda}\in\mathbb{R}^d}\left\{\bar{\lambda}^\top Z - \inf_{\bar{\Lambda}\in\mathbb{R}^{d\times d}}\mathbb{E}_{\mathbb{P}^*}\left[\|\bar{\Lambda}R+\bar{\lambda}\|_p^2\right]\right\}.$$

If $p=2$, then we have

$$\mathbb{E}_{\mathbb{P}^*}\left[\|\bar{\Lambda}R+\bar{\lambda}\|_2^2\right] = \sum_i \mathbb{E}_{\mathbb{P}^*}\left(\bar{\Lambda}_{i\cdot}\cdot R + \bar{\lambda}_i\right)^2.$$

So, taking derivative with respect to the ith row, $\bar{\Lambda}_{i\cdot}$, of the matrix $\bar{\Lambda}$, $\bar{\Lambda}_{i\cdot}$, we obtain

$$\begin{aligned} \nabla_{\bar{\Lambda}_{i\cdot}}\mathbb{E}_{\mathbb{P}^*}\left[\|\bar{\Lambda}R+\bar{\lambda}\|_2^2\right] &= 2\mathbb{E}_{\mathbb{P}^*}\left((R^\top\bar{\Lambda}_{i\cdot}+\bar{\lambda}_i)R\right) \\ &= 2\mathbb{E}_{\mathbb{P}^*}\left(R^\top\bar{\Lambda}_{i\cdot}R\right) + 2\bar{\lambda}_i\mathbb{E}_{\mathbb{P}^*}(R) = 0. \end{aligned} \quad \text{(B.2)}$$

Writing

$$\mu_* = \mathbb{E}_{\mathbb{P}^*}(R) \text{ and } \Sigma_* = \mathbb{E}_{\mathbb{P}^*}\left(RR^\top\right),$$

we obtain

$$\Sigma_*\bar{\Lambda}_{i\cdot} = -\bar{\lambda}_i\mu_*.$$

Solving this equation yields $\bar{\Lambda}_{i\cdot} = -\bar{\lambda}_i\Sigma_*^{-1}\mu_*$. Therefore,

$$\mathbb{E}_{\mathbb{P}^*}\left(\bar{\Lambda}_{i\cdot}R+\bar{\lambda}_i\right)^2 = \bar{\Lambda}_{i\cdot}^\top\Sigma_*\bar{\Lambda}_{i\cdot} + 2\bar{\lambda}_i\bar{\Lambda}_{i\cdot}^\top\mu_* + \bar{\lambda}_i^2 = \bar{\lambda}_i^2\left(1-\mu_*^\top\Sigma_*^{-1}\mu_*\right).$$

By Assumption 3, $\mathrm{Var}_{\mathbb{P}^*}(R)$ is positive definite and, hence, invertible. It then follows from the Sherman–Morrison formula that

$$\begin{aligned} (\Sigma_*)^{-1} &= (\mathrm{Var}_{\mathbb{P}^*}(R) + \mu_*\mu_*^T)^{-1} \\ &= \mathrm{Var}_{\mathbb{P}^*}(R)^{-1} - \frac{\mathrm{Var}_{\mathbb{P}^*}(R)^{-1}\mu_*\mu_*^T\mathrm{Var}_{\mathbb{P}^*}(R)^{-1}}{1+\mu_*^T\mathrm{Var}_{\mathbb{P}^*}(R)^{-1}\mu_*}. \end{aligned}$$

So

$$\begin{aligned} \mu_*^T(\Sigma_*)^{-1}\mu_* &= \mu_*^T\mathrm{Var}_{\mathbb{P}^*}(R)^{-1}\mu_* - \frac{(\mu_*^T\mathrm{Var}_{\mathbb{P}^*}(R)^{-1}\mu_*)^2}{1+\mu_*^T\mathrm{Var}_{\mathbb{P}^*}(R)^{-1}\mu_*} \\ &= \frac{\mu_*^T\mathrm{Var}_{\mathbb{P}^*}(R)^{-1}\mu_*}{1+\mu_*^T\mathrm{Var}_{\mathbb{P}^*}(R)^{-1}\mu_*} < 1, \end{aligned}$$

leading to

$$\begin{aligned} L_0 &= \sup_{\bar{\lambda}\in\mathbb{R}^d}\left\{\bar{\lambda}^\top Z - \inf_{\bar{\Lambda}\in\mathbb{R}^{d\times d}}\mathbb{E}_{\mathbb{P}^*}[\|\bar{\Lambda}R+\bar{\lambda}\|_2^2]\right\} \\ &= \sup_{\bar{\lambda}\in\mathbb{R}^d}\left\{\bar{\lambda}^\top Z - \|\bar{\lambda}\|_2^2\left(1-\mu_*^\top\Sigma_*^{-1}\mu_*\right)\right\} \\ &= \frac{\|Z\|_2^2}{4\left(1-\mu_*^\top\Sigma_*^{-1}\mu_*\right)}. \end{aligned}$$

It remains to identify Z. Observe that

$$\begin{aligned} \mu_n &= \rho 1 + 2\left(\Sigma_n\phi^* - \phi^{*T}\Sigma_n\phi^*1\right)/\lambda_1^* \\ &= \rho 1 + 2\left(\Sigma_*\phi^* - \phi^{*T}\Sigma_*\phi^*1\right)/\lambda_1^* \\ &\quad + 2\left(H_n\phi^* - \phi^{*T}H_n\phi^*1\right)/\lambda_1^* \\ &= \mu_* + 2\left(H_n\phi^* - \phi^{*T}H_n\phi^*1\right)/\lambda_1^*, \end{aligned}$$

where $H_n := \Sigma_n - \Sigma_*$. By Assumption 1, we have

$$\begin{aligned} &n^{-1/2}\sum_{i=1}^n(R_i-\mu_*) \Rightarrow Z_0 \sim N(0,\Upsilon_{g_1}), \\ &n^{1/2}H_n \Rightarrow Y \sim N(0,\Upsilon_{g_2}). \end{aligned}$$

Thus,

$$\begin{aligned} &n^{-1/2}\sum_{i=1}^n\bar{\lambda}^\top(R_i-\mu_*) + 2n^{1/2}\bar{\lambda}^\top\left(H_n\phi^* - \phi^{*T}H_n\phi^*1\right)/\lambda_1^* \\ &\Rightarrow \bar{\lambda}^\top Z = \bar{\lambda}^\top(Z_0+Z_1), \end{aligned}$$

where

$$Z_1 := 2\left(Y\phi^* - \phi^{*T}Y\phi^*1\right)/\lambda_1^*.$$

Appendix C. Proof of Proposition 1

Note that

$$\begin{aligned} n^{1/2}\left\{\frac{(\phi_n^*)^\top[\mathbb{E}_{\mathbb{P}_n}(R)-\mathbb{E}_{\mathbb{P}^*}(R)]}{\|\phi_n^*\|_p}\right\} &= n^{1/2}\left\{\frac{(\phi_n^*-\phi^*)^\top[\mathbb{E}_{\mathbb{P}_n}(R)-\mathbb{E}_{\mathbb{P}^*}(R)]}{\|\phi_n^*\|_p}\right\} \\ &\quad + n^{1/2}\left\{\frac{(\phi^*)^\top[\mathbb{E}_{\mathbb{P}_n}(R)-\mathbb{E}_{\mathbb{P}^*}(R)]}{\|\phi^*\|_p}\cdot\frac{\|\phi^*\|_p}{\|\phi_n^*\|_p}\right\}. \end{aligned}$$

By the standard central limit theorem and the fact that $\phi_n^* \to \phi^*$ in probability, we conclude

$$n^{1/2}\left\{\frac{(\phi_n^*)^\top[\mathbb{E}_{\mathbb{P}_n}(R)-\mathbb{E}_{\mathbb{P}^*}(R)]}{\|\phi_n^*\|_p} - \frac{(\phi^*)^\top[\mathbb{E}_{\mathbb{P}_n}(R)-\mathbb{E}_{\mathbb{P}^*}(R)]}{\|\phi^*\|_p}\right\} \Rightarrow 0$$

as $n\to\infty$. However, again by the central limit theorem, we have

$$n^{1/2}\left\{\frac{(\phi^*)^\top[\mathbb{E}_{\mathbb{P}_n}(R)-\mathbb{E}_{\mathbb{P}^*}(R)]}{\|\phi^*\|_p}\right\} \Rightarrow N\left(0,\Upsilon_{\phi^*}\right),$$

which yields the desired result.

Endnotes

[1] There are several mathematically equivalent formulations of the original mean–variance model.

[2] A Wasserstein distance is the optimal value for a specific optimal transport problem. The notion was first formulated by Monge (1781) and its theory developed by Kantorovich (1942). It is widely used in the study of DRO problems.

[3] Recent work by Blanchet et al. (2016) shows that a similar definition of discrepancy in some other models recovers exactly certain well-known machine learning algorithms, such as square-root Lasso and support vector machines.

[4] Cross-validation, although a standard technique, is generally data-intensive and time-consuming. In the rolling horizon setting, for instance, one has to assume the parameter δ to be constant during the horizon and estimate it. In contrast, in our data-driven approach, the parameter δ changes over the horizon in a way that is sensitive to the variability in the data.

[5] Different cost functions can be used, resulting in different regularization penalties as we discuss at the end of Section 3 and in Section 6.

[6] In practice, it is not desirable to include, in the case of S&P 500 stocks, for example, all the 500 stocks in one's portfolios even though one of the key implications of the mean–variance model is diversification. From a practical perspective, including too many stocks is costly and prone to mismanagement. Therefore, adding a proper regularization term not only reduces overfitting, but also helps achieve a balance between diversification and manageability.

[7] Strictly speaking, (4) is a mean–standard deviation model, which is equivalent to the mean–variance one.

[8] Herein, the analysis is under Assumption 3. If $\Phi_{\mathbb{P}^*}$ contained more than just one element, then there would be several possible options to formulate an optimization problem for choosing δ. For example, we may choose δ as the smallest uncertainty size such that $\Phi_{\mathbb{P}^*} \subset \Lambda_\delta(\mathbb{P}_n)$ with probability $1-\delta_0$, in which case we would need to study $\sup_{\phi^* \in \Phi_{\mathbb{P}^*}} \bar{\mathcal{R}}_n^*(\phi^*)$.

[9] We chose the period 2000–2016 for our back-testing for a reason: the market was overall very volatile during this period, experiencing two major crashes: the dotcom bubble burst and the subprime financial crisis, followed by a long bull run until this day of writing (February 2018). We were particularly interested to see how robust our DRMV strategies would have been when sailing through such a bumpy journey.

[10] In theory, we should have included *all* the constituents of S&P 500 in our portfolios. However, that would be computationally inefficient and practically (almost) infeasible for most of the models under testing (e.g., the original Markowitz model). Therefore, it is desirable to choose a small subset of stocks based on which to apply various models. This "stock selection" is an ultimately important part of the overall portfolio management. In this paper, however, we aim to test the performances of "stock allocation" (namely, to allocate wealth among the stocks that have been *already* selected in order to achieve the best risk-adjusted return) of these models. That is why we randomly selected the small subset of stocks in order to focus on the part of the *stock allocation*. On the other hand, the requirement that the selected stocks have at least 10 years' price data is due to the length of the training period.

[11] There is an extensive literature on continuous-time Markowitz models; however, to our best knowledge, all the other existing models include a risk-free asset.

[12] We also tested for portfolios with 20 stocks and observed bankruptcy in more than half of our experiments. On the other hand, although the other six single-period models have no explicit no-bankruptcy constraint either, a total of only two instances of bankruptcy occurred in our experiments.

[13] Bielecki et al. (2005) solve a continuous-time mean–variance model with the no-bankruptcy constraint. However, there is a risk-free asset in that model. To have a fair comparison with the other models in which there is no risk-free account available, it is proper to choose the model of Cui et al. (2012) in our experiments. We are not aware of a work on continuous-time Markowitz models without a risk-free asset and with bankruptcy prohibition.

[14] We assume that the factors have been processed according to the available papers (Fama and French 1992, 1993).

[15] See http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

[16] Here, we include only those having at least 10 years' historical data to be consistent with the other models.

[17] Here, if $\|\phi_M - \phi_0\|_1 \le \tau$, then ϕ_M satisfies the trading volume constraint, and hence, itself is the optimal portfolio. In this case, $\kappa = 0$.

[18] The good performance of the equally weighted portfolio is well documented in the literature; see, for example, DeMiguel et al. (2009).

[19] As those histograms are similar, we do not present them here.

References

Ao M, Li Y, Zheng X (2018) Approaching mean-variance efficiency for large portfolios. *Rev. Financial Stud.* 32(7):2890–2919.

Bertsimas D, Pachamanova D, Sim M (2004) Robust linear optimization under general norms. *Oper. Res. Lett.* 32(6):510–516. Q: 12

Bielecki T, Jin H, Pliska S, Zhou X (2005) Continuous-time mean–variance portfolio selection with bankruptcy prohibition. *Math. Finance* 15(2):213–244.

Blanchet J, Kang Y (2017) Distributionally robust groupwise regularization estimator. Preprint, submitted May 11, https://arxiv.org/abs/1705.04241.

Blanchet J, Murthy K (2019) Quantifying distributional model risk via optimal transport. *Math. Oper. Res.* 44(2):565–600.

Blanchet J, Kang Y, Murthy K (2016) Robust Wasserstein profile inference and applications to machine learning. Preprint, submitted October 18, https://arxiv.org/abs/1610.05627.

Blanchet J, Murthy K, Si N (2019) Confidence regions in Wasserstein distributionally robust estimation. Preprint, submitted June 4, https://arxiv.org/abs/1906.01614.

Blanchet J, Kang Y, Zhang F, Murthy K (2017) Data-driven optimal transport cost selection for distributionally robust optimization. Preprint, submitted May 19, https://arxiv.org/abs/1705.07152.

Costa O, Paiva A (2002) Robust portfolio selection using linear-matrix inequalities. *J. Econom. Dynamic Control* 26(6):889–909.

Cui X, Gao J, Li D (2012) Continuous-time mean-variance portfolio selection with finite transactions. Zhang T, Zhou X, eds. *Stochastic Analysis and Applications to Finance*, Interdisciplinary Mathematical Sciences, vol. 13 (World Scientific), 77–98.

Delage E, Ye Y (2010) Data-driven distributionally robust optimization under moment uncertainty with application to data-driven problems. *Oper. Res.* 58(3):595–612. Q: 13

Delage E, Kuhn D, Wiesemann W (2019) Dice-sion–making under uncertainty: When can a random decision reduce risk? *Management Sci.* 65(7):3282–3301.

DeMiguel V, Garlappi L, Uppal R (2009) Optimal vs. naive diversification: How inefficient is the $1/n$ portfolio strategy? *Rev. Financial Stud.* 22(5):1915–1953.

Esfahani P, Kuhn D (2018) Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Math. Programming* 171(1–2):115–166.

Fabozzi F, Kolm P, Pachamanova D, Focardi S (2007) Robust portfolio optimization. *J. Portfolio Management* 33(3):40–48.

Fama EF, French KR (1992) The cross-section of expected stock returns. *J. Finance* 47(2):427–465.

Fama EF, French KR (1993) Common risk factors in the returns on stocks and bonds. *J. Financial Econom.* 33(1):3–56.

Fan J, Liao Y, Mincheva M (2011) High-dimensional covariance matrix estimation in approximate factor models. *Ann. Statist.* 39(6):3320–3356.

Gao R, Kleywegt A (2016) Distributionally robust stochastic optimization with wasserstein distance. Preprint, submitted April 8, https://arxiv.org/abs/1604.02199.

Gao R, Chen X, Kleywegt A (2017) Wasserstein distributional robustness and regularization in statistical learning. Preprint, submitted December 17, https://arxiv.org/abs/1712.06050.

Ghaoui L, Oks M, Oustry F (2003) Worst-case value-at-risk and robust portfolio optimization: A conic programming approach. *Oper. Res.* 51(4):543–556.

Goh J, Sim M (2010) Distributionally robust optimization and its tractable approximations. *Oper. Res.* 58(4):595–612.

Goldfarb D, Iyengar G (2003) Robust portfolio selection problems. *Math. Oper. Res.* 28(1):1–38.

Gotoh J, Takeda A (2011) On the role of norm constraints in portfolio selection. *Comput. Management Sci.* 8:323–353.

Halldorsson B, Tutuncu R (2003) An interior-point method for a class of saddle-point problems. *J. Optim. Theory Appl.* 116:559–590.

Hansen L, Sargent T (2008) *Robustness* (Princeton University Press, Princeton, NJ).

Hu Z, Hong L (2013) Kullback-Leibler divergence constrained distributionally robust optimization.

Idzorek T (2002) A step-by-step guide to the Black-Litterman model. Technical report.

Q: 14

Jiang R, Guan Y (2016) Data-drive chance constrained stochastic program. *Math. Programming* 158:291–327.

Q: 15

Kantorovich L (1942) On the transfer of masses. (in Russian) *Comptes Rendus (Doklady) de I'Académie des Sciences de I'URSS* 37(7–8): 227–229.

Lobo M, Boyd S (2000) The worst-case risk of a portfolio. https://web.stanford.edu/~boyd/papers/pdf/risk_bnd.pdf.

Markowitz H (1952) Portfolio selection. *J. Finance* 7(1):77–91.

Q: 16

Monge G (1781) *Mémoire sur la théorie des déblais et des remblais* (Histoire de l'Académie Royale des Sciences de Paris, Paris).

Olivares-Nadal A, DeMiguel V (2018) Technical note—A robust perspective on transaction costs in portfolio optimization. *Oper. Res.* 66(3):733–739.

Ollila E, Raninen E (2018) Optimal shrinkage covariance matrix estimation under random sampling from elliptical distribution. Preprint, submitted August 30, https://arxiv.org/abs/1808.10188.

Petersen I, James M, Dupuis P (2000) Minimax optimal control of stochastic uncertain systems with relative entropy constraints. *IEEE Trans. Automatic Control* 45(3):398–412.

Pflug G, Wozabal D (2007) Ambiguity in portfolio selection. *Quant. Finance* 7(4):435–442.

Villani C (2003) *Topics in Optimal Transportation.* Graduate Studies in Mathematics, vol. 58 (American Mathematics Society, Providence, RI).

Wiesemann W, Kuhn D, Sim M (2014) Distributionally robust convex optimization. *Oper. Res.* 62(6):1358–1376.

Wozabal D (2012) A framework for optimization under ambiguity. *Ann. Oper. Res.* 193:21–47.

Zhao C, Guan Y (2018) Data-driven risk-averse stochastic optimization with Wasserstein metric. *Oper. Res. Lett.* 46(2):262–267.