Machine Learning and Portfolio Optimization

Gah-Yi Ban*
Management Science & Operations, London Business School, Regent's Park, London, NW1 4SA, United Kingdom.
gban@london.edu

Noureddine El Karoui
Department of Statistics, University of California, Berkeley, CA 94720. nkaroui@stat.berkeley.edu

Andrew E.B. Lim
Department of Decision Sciences & Department of Finance, National University of Singapore Business School, Singapore, 119245. andrewlim@nus.edu.sg

The portfolio optimization model has limited impact in practice due to estimation issues when applied with real data. To address this, we adapt two machine learning methods, regularization and cross-validation, for portfolio optimization. First, we introduce *performance-based regularization* (PBR), where the idea is to constrain the sample variances of the estimated portfolio risk and return, which steers the solution towards one associated with less estimation error in the performance. We consider PBR for both mean-variance and mean-CVaR problems. For the mean-variance problem, PBR introduces a quartic polynomial constraint, for which we make two convex approximations: one based on rank-1 approximation and another based on a convex quadratic approximation. The rank-1 approximation PBR adds a bias to the optimal allocation, and the convex quadratic approximation PBR shrinks the sample covariance matrix. For the mean-CVaR problem, the PBR model is a combinatorial optimization problem, but we prove its convex relaxation, a QCQP, is essentially tight. We show that the PBR models can be cast as robust optimization problems with novel uncertainty sets and establish asymptotic optimality of both Sample Average Approximation (SAA) and PBR solutions and the corresponding efficient frontiers. To calibrate the right hand sides of the PBR constraints, we develop new, performance-based k-fold cross-validation algorithms. Using these algorithms, we carry out an extensive empirical investigation of PBR against SAA, as well as L_1 and L_2 regularizations and the equally-weighted portfolio. We find that PBR dominates all other benchmarks for two out of three of Fama-French data sets.

Key words: machine learning, portfolio optimization, regularization, risk measures, robust optimization
History: First submitted: June 24, 2013. This version: August 31, 2016.

1. Introduction

Regularization is a technique that is commonly used to control the stability of a wide range of problems. Its origins trace back to the 1960s, when it was introduced to deal with ill-posed linear

* formerly: Gah-Yi Vahn

operator problems. A linear operator problem is one of finding $x \in X$ that satisfies $Ax = b$, where A is a linear operator from a normed space X to a normed space Y, and $b \in Y$ is a predetermined constant. The linear operator problem is ill-posed if small deviations in b, perhaps due to noise, result in large deviations in the corresponding solution. Specifically, if b changes to b_δ, $||b_\delta - b|| < \delta$, then finding x that minimizes the functional $R(x) = ||Ax - b_\delta||^2$ does not guarantee a good approximation to the desired solution even if δ tends to zero. Tikhonov (1963), Ivanov (1962) and Phillips (1962) discovered that if instead of minimizing $R(x)$, the most obvious choice, one minimizes the *regularized* functional

$$R^*(x) = ||Ax - b_\delta||_2^2 + \gamma(\delta)P(x),$$

where $P(x)$ is some functional and $\gamma(\delta)$ is an appropriately chosen constant, then one obtains a sequence of solutions that does converge to the desired one as δ tends to zero. Regularization theory thus shows that whereas the self-evident method of minimizing $R(x)$ does not work, the non-self-evident method of minimizing $R^*(x)$ does.

Regularization has particularly been made known in recent years through its adoption in classification, regression and density estimation problems. The reader may be most familiar with its recent popularity in the high-dimensional regression literature [see, for example, Candes and Tao (2007) and Belloni and Chernozhukov (2013)]:

$$\min_{\beta \in \mathbb{R}^p} ||\mathbf{y} - \mathbf{X}\beta||_2 + \lambda P(\beta), \tag{1}$$

where $P(\beta) = ||\beta||_1$, $\mathbf{y} = [y_1, \ldots, y_n] \in \mathbb{R}^n$ is the data on the observable, $\mathbf{X} = [X_1, \ldots, X_n] \in \mathbb{R}^{n \times p}$ is the vector of covariates, $\beta \in \mathbb{R}^p$ is the regression coefficient that best fits the linear model $y = X\beta$, and $\lambda > 0$ is a parameter that controls the sparsity of the solution. The regression model (1) with $P(\beta) = ||\beta||_1$ is known as the Lasso model, used in high-dimensional applications where sparsity of the solution β is desirable for interpretability and recovery purposes when p is large. Another common model is the Tikhonov regularization function $P(\beta) = ||\beta||_2$, which deals with issues that arise when the data matrix $\mathbf{X}$ is ill-conditioned or singular.

In this paper, we consider regularizing the data-driven portfolio optimization problem, not for sparsity or numerical stability as in Lasso or ridge regression, but for the purpose of improving the out-of-sample performance of the solution. The portfolio optimization model we consider is:

$$\begin{aligned} w_0 = \underset{w \in \mathbb{R}^p}{\operatorname{argmin}} \; & Risk(w^\top X) \\ s.t. \quad & w^\top 1_p = 1 \\ & (w^\top \mu = R), \end{aligned} \tag{PO}$$

where $w \in \mathbb{R}^p$ is the investor's holding on p different assets, $X \in \mathbb{R}^p$ denotes the relative return on the p assets, $\mu = \mathbb{E}X$ is the mean return vector and $Risk : \mathbb{R} \to \mathbb{R}$ is some measure of risk. The

investor's wealth is normalized to 1, so $w^\top 1_p = 1$, where 1_p denotes $p \times 1$ vector of ones, and $w^\top \mu = R$ is the target return constraint, which we may or may not consider[1], hence denoted in parentheses. Note that shortselling, i.e., having $w < 0$, is allowed in this model. Setting $Risk(w^\top X) = w^\top \Sigma w$ we recover the classical model of Markowitz (1952) and setting $Risk(w^\top X) = CVaR(-w^\top X; \beta)$, where $\beta \in (0.5, 1)$ and

$$CVaR(-w^\top X; \beta) := \min_{\alpha} \ \alpha + \frac{1}{1-\beta}\mathbb{E}(-w^\top X - \alpha)^+, \tag{2}$$

we recover the Conditional Value-at-Risk[2] (CVaR) formulation of Rockafellar and Uryasev (2000).

In practice, one does not know the true distribution of X, but has access to past data: $\mathsf{X} = [X_1, \ldots, X_n]$. Assuming that these are iid, the standard data-driven approach to solving (PO) is to solve:

$$\begin{array}{ll} \hat{w}_n = \underset{w \in \mathbb{R}^p}{\operatorname{argmin}} & \widehat{Risk}_n(w^\top X) \\ s.t. & w^\top 1_p = 1 \\ & (w^\top \hat{\mu}_n = R), \end{array} \tag{SAA}$$

where $\widehat{Risk}_n(w^\top X)$ is the sample average estimate of the Risk function and $\hat{\mu}_n = n^{-1}\sum_{i=1}^n X_i$ is the sample average return vector. This approach is befittingly known as the Sample Average Approximation (SAA) method in the stochastic programming literature [see Shapiro et al. (2009) for a general overview].

As is the case with ill-posed linear operator problems (which includes regression problems), the solution to the SAA approach can be highly unstable. For the portfolio optimization problem, the fact that the SAA allocation is highly unreliable is well-documented [see Frankfurter et al. (1971), Frost and Savarino (1986, 1988b), Michaud (1989), Best and Grauer (1991), Chopra and Ziemba (1993), Broadie (1993) for the Markowitz problem and Lim et al. (2011) for the mean-CVaR problem], and has limited the wide-spread adoption of the model in practice, despite the conferral of a Nobel Prize to Markowitz in 1990 for his seminal work.

In this paper, we propose *performance-based regularization* (PBR) to improve upon the performance of the SAA approach to the data-driven portfolio allocation problem (SAA). The idea is to constrain the sample variances of estimated quantities in a problem; for portfolio optimization they are the estimated portfolio risk $\widehat{Risk}_n(w^\top X)$ and the estimated portfolio mean $w^\top \hat{\mu}_n$. The goal of PBR is to steer the solution towards one that is associated with less estimation error in the performance. The overall effect is to reduce the chance that a solution is chosen by misleadingly high in-sample performance. Performance-based regularization is thus philosophically

[1] There is empirical evidence that ignoring the mean return constraint yields better solutions [see Jorion (1985)].

[2] also known as expected shortfall [Acerbi and Tasche (2002)]

different from Tikhonov regularization (whose purpose is stability of the solution) and Lasso regularization (whose purpose is sparsity) but is natural to the portfolio optimization problem where the ultimate goal is the *out-of-sample* performance of the decision made.

We make four major contributions in this paper. Firstly, we propose and analyze new portfolio optimization models by introducing performance-based regularization to the mean-variance and mean-CVaR problems. This is an important conceptual development that extends the current literature on portfolio optimization. For the mean-variance problem, the PBR model involves a quartic polynomial constraint. Determining whether such a model is convex or not is an NP-hard problem, so we consider two convex approximations, one based on a rank-1 approximation and one based on the best convex quadratic approximation. We then investigate the two approximation models and analytically characterize the effect of PBR on the solution. In the rank-1 approximation model, PBR adds a bias to the optimal allocation directly, whereas in the quadratic approximation case, PBR is equivalent to shrinking the sample covariance matrix. For the mean-CVaR problem, the PBR model is a combinatorial optimization problem, but we prove its convex relaxation, a quadratically constrained quadratic program (QCQP), is tight, hence can be efficiently solved.

Secondly, we show that the PBR portfolio models can be cast as robust optimization problems, introducing uncertainty sets that are new to the literature. The PBR constraint on the mean return uncertainty is equivalent to the a constraint where the portfolio return is required to be robust to all possible values of the mean vector falling within an ellipsoid, centred about the true mean. This is a well-known result in robust optimization [see Ben-Tal et al. (2009)]. However, the robust counterparts of the PBR constraint on the risk have structures that have not been considered before. The robust counterparts are somewhat related to constraining estimation error in the portfolio risk, however the robust models do not enjoy the same intuitive interpretation of the original PBR formulations. We thus not only link PBR with novel robust models, but also justify the original PBR formulation in its own right, as it is motivated by the intuitive idea of cutting off solutions associated with high in-sample estimation errors, whereas the equivalent robust constraint does not necessarily enjoy intuitive interpretation.

Thirdly, we prove that the SAA and PBR solutions are asymptotically optimal under the very mild assumption that the true solutions be well-separated (i.e., identifiable). This is an important and necessary result because data-driven decisions that are not asymptotically optimal as the number of stationary observations increases is nonsensical. We also show that the corresponding performance of the SAA and PBR solutions converge to the true optimal solutions. To the best of our knowledge, this is the first paper that analytically proves the asymptotic optimality of the solutions to estimated portfolio optimization problems for general underlying return distributions [see Jobson and Korkie (1980) for asymptotic analysis when the returns are multivariate normal.].

Finally, we make an extensive empirical study of the PBR method against SAA as well as other benchmarks, including L_1 and L_2 regularizations and the equally-weighted portfolio of DeMiguel et al. (2009b). We use the five, ten and forty-nine industry data sets from French (2015). To calibrate the constraint right-hand side (rhs) of PBR and standard regularization models, we also develop a new, performance-based extension of the k-fold cross-validation algorithm. The two key differences between our algorithm and standard k-fold cross-validation are that the search boundaries for the PBR constraint rhs need to be set carefully in order to avoid infeasibility and having no effect, and that we validate by computing the Sharpe ratio (the main performance metric for investment in practice) as opposed to the mean squared error. In sum, we find that for the five and ten industry data sets, the PBR method improves upon SAA, in terms of the out-of-sample Sharpe ratio (annualized) with statistical significance at 5% and 10% respectively for both the mean-variance and mean-CVaR problems. Also for these data sets, PBR dominates standard L_1 and L_2 regularizations, as well as the equally weighted portfolio of DeMiguel et al. (2009b). The results for the forty-nine industry portfolio data set are inconclusive, with none of the strategies considered being statistically significantly different from the SAA result. We attribute this to the high-dimensionality effect [see Ledoit and Wolf (2004) and El Karoui (2010)], and leave studies of mitigating for the dimensionality to future work [Ban and Chen (2016)].

1.1. Survey of literature

As mentioned in the beginning, Tikhonov (1963), Ivanov (1962) and Phillips (1962) first introduced the notion of regularization for ill-posed linear operator problems. For details on the historical development and use of regularization in statistical problems, Vapnik (2000) is a classic text; for a more recent illustrations of the technique we refer the reader to Hastie et al. (2009).

The more conventional regularization models have been investigated for the Markowitz problem by Chopra (1993), Frost and Savarino (1988a), Jagannathan and Ma (2003), DeMiguel et al. (2009a), and for the mean-CVaR problem by Gotoh and Takeda (2010). Specifically, Chopra (1993), Frost and Savarino (1988a) and Jagannathan and Ma (2003) consider imposing a no shortsale constraint on the portfolio weights (i.e., require portfolio weights to be non-negative). DeMiguel et al. (2009a) generalizes this further by considering L_1, L_2 and A-norm regularizations, and shows that the no shortsale constraint is a special case of L_1 regularized portfolio. Our PBR model for the Markowitz problem extends this literature by considering performance-motivated regularization constraints. The actual PBR model is non-convex so we consider two convex approximations, the first being an extra affine constraint on the portfolio weights, and the second being a constraint on a particular A-norm of the vector of portfolio weights. The first corresponds to adding a bias to the SAA solution and the second corresponds to shrinking the sample covariance matrix in a specific

way. Analogously, Gotoh and Takeda (2010) considers L_1 and L_2 norms for the data-driven mean-CVaR problem; our work also extends this literature. In Sec. 5, we show that PBR out-performs the standard regularization techniques in terms of the out-of-sample Sharpe ratio.

The PBR models add a new perspective on recent developments in robust portfolio optimization that construct uncertainty sets from data [Delage and Ye (2010), Goldfarb and Iyengar (2003)]. While the PBR constraint on the portfolio mean is equivalent to the mean uncertainty constraint considered in Delage and Ye (2010), the PBR constraint on the portfolio variance for the mean-variance problem leads to a new uncertainty set which is different from Delage and Ye (2010). The main difference is that Delage and Ye (2010) considers an uncertainty set for the sample covariance matrix separately from the decision, whereas PBR considers protecting against estimation errors in the portfolio variance, thereby considering both the decision and the covariance matrix together. The difference is detailed in Appendix B. Goldfarb and Iyengar (2003) also takes the approach of directly modelling the uncertainty set of the covariance matrix, although it is different from Delage and Ye (2010) and also from our work because it starts from a factor model of asset returns and assumes that the returns are multivariate noramally distributed, whereas both Delage and Ye (2010) and our work are based on a nonparametric, distribution-free setting.

Finally, Gotoh et al. (2015) shows that a large class of distributionally robust empirical optimization problems with uncertainty sets defined in terms of ϕ-divergence are asymptotically equivalent to PBR problems. We note however that the class of models studied in Gotoh et al. (2015) does not include CVaR.

Notations. Throughout the paper, we denote convergence in probability by $\xrightarrow{P}$.

2. Motivation: Fragility of SAA in Portfolio Optimization

In this paper, we consider two risk functions: the variance of the portfolio, and the Conditional Value-at-Risk. In the former case, the problem is the classical Markowitz model of portfolio optimization, which is

$$\begin{aligned} w_{MV} = \underset{w\in\mathbb{R}^p}{\operatorname{argmin}}\ & w^\top \Sigma w \\ s.t.\quad & w^\top 1_p = 1 \\ & (w^\top \mu = R), \end{aligned} \tag{MV-true}$$

where μ and Σ are respectively the mean and the covariance matrix of X, the relative stock return, and where the target return constraint $(w^\top \mu = R)$ may or may not be imposed.

Given data $\mathsf{X} = [X_1, X_2, \ldots, X_n]$, the SAA approach to the problem is

$$\begin{aligned} \hat{w}_{n,MV} = \underset{w\in\mathbb{R}^p}{\operatorname{argmin}}\ & w^\top \hat{\Sigma}_n w \\ s.t.\quad & w^\top 1_p = 1 \\ & (w^\top \hat{\mu}_n = R), \end{aligned} \tag{MV-SAA}$$

where $\hat{\mu}_n$ and $\hat{\Sigma}_n$ are respectively the sample mean and the sample covariance matrix of X.

In the latter case, we have a mean-Conditional Value-at-Risk (CVaR) portfolio optimization model. Specifically, the investor wants to pick a portfolio that minimizes the CVaR of the portfolio loss at level $100(1-\beta)\%$, for some $\beta \in (0.5, 1)$, while reaching an expected return R:

$$\begin{array}{rll} w_{CV} = \underset{w\in\mathbb{R}^p}{\operatorname{argmin}} & CVaR(-w^\top X;\beta) & \\ s.t. & w^\top 1_p = 1 & \\ & (w^\top \mu = R), & \end{array} \tag{CV-true}$$

where

$$CVaR(-w^\top X;\beta) := \min_{\alpha} \ \alpha + \frac{1}{1-\beta}\mathbb{E}(-w^\top X - \alpha)^+,$$

as in Rockafellar and Uryasev (2000).

The SAA approach to the problem is to solve

$$\begin{array}{rll} \hat{w}_{n,CV} = \underset{w\in\mathbb{R}^p}{\operatorname{argmin}} & \widehat{CVaR}_n(-w^\top \mathsf{X};\beta) & \\ s.t. & w^\top 1_p = 1 & \\ & (w^\top \hat{\mu}_n = R), & \end{array} \tag{CV-SAA}$$

where

$$\widehat{CVaR}_n(-w^\top \mathsf{X};\beta) := \min_{\alpha\in\mathbb{R}} \ \alpha + \frac{1}{n(1-\beta)}\sum_{i=1}^{n}(-w^\top X_i - \alpha)^+,$$

is the sample average estimator for $CVaR(-w^\top X;\beta)$.

Asymptotically, as the number of observations n goes to infinity, we can show that the SAA solutions $\hat{w}_{n,MV}$ and $\hat{w}_{n,CV}$ converge in probability to w_{MV} and w_{CV} respectively [see Sec. 4 for details]. In practice, however, the investor has a limited number of relevant (i.e., stationary) observations [Jegadeesh and Titman (1993), Lo and MacKinlay (1990), DeMiguel et al. (2014)]. Solving (MV-SAA) and (CV-SAA) with finite amount of stationary data can yield highly unreliable solutions [Lim et al. (2011)]. Let us illustrate this point by a simulation experiment for (CV-SAA). There are $p = 10$ stocks, with daily returns following a Gaussian distribution[3]: $X \sim \mathcal{N}(\mu_{sim}, \Sigma_{sim})$, and the investor has $n = 250$ iid observations of X. The experimental procedure is as follows:

- Simulate 250 historical observations from $\mathcal{N}(\mu_{sim}, \Sigma_{sim})$.
- Solve (CV-SAA) with $\beta = 0.95$ and some return level R to find an instance of $\hat{w}_{n,CV}$.
- Plot the realized return $\hat{w}_{n,CV}^\top \mu_{sim}$ versus realized risk $CVaR(-\hat{w}_{n,CV}^\top X;\beta)$; this corresponds to one grey point in Fig. (1).
- Repeat for different values of R to obtain a sample efficient frontier.
- Repeat many times to get a distribution of the sample efficient frontier.

The result of the experiment is summarized in Fig. (1). The smooth curve corresponds to the population efficient frontier. Each of the grey dots corresponds to a solution instance of (CV-SAA). There are two noteworthy observations: the solutions $\hat{w}_{n,CV}$ are sub-optimal, and they are highly variable. For instance, for a daily return of 0.1%, the CVaR ranges from 1.3% to 4%.

[3] the parameters are the sample mean and covariance matrix of data from 500 daily returns of 10 different US stocks from Jan 2009– Jan 2011

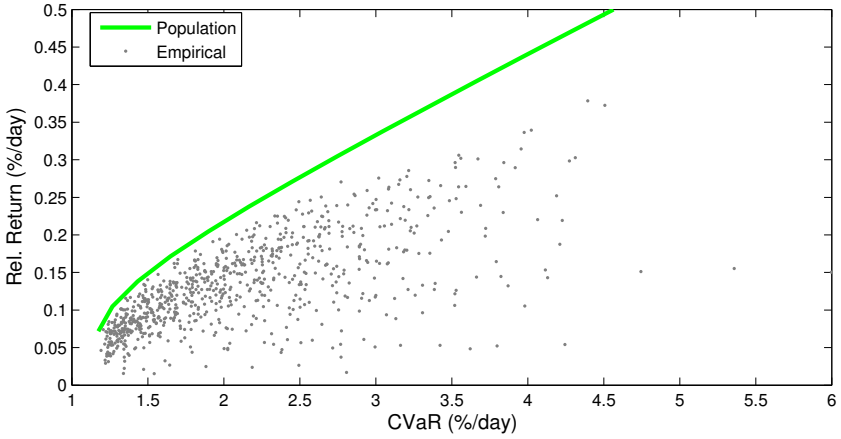

Figure 1 Distribution of realized daily return (%/day) vs. daily risk (%/day) of SAA solutions $\hat{w}_{n,CV}$ for the target return range $0.107 - 0.430$ %/day. Green line represents the population frontier, i.e., the efficient frontier corresponding to solving (CV-true).

3. Performance-based regularization

We now introduce *performance-based regularization* (PBR) to improve upon (SAA). The PBR model is:

$$\begin{aligned} \hat{w}_{n,PBR} = \underset{w\in\mathbb{R}^p}{\operatorname{argmin}} \ & \widehat{Risk}_n(w^\top X) \\ s.t. \quad & w^\top 1_p = 1 \\ & (w^\top \hat{\mu}_n = R) \\ & SVar(\widehat{Risk}_n(w^\top X)) \le U_1 \\ & (SVar(w^\top \hat{\mu}_n) \le U_2), \end{aligned} \tag{PBR}$$

where $SVar(\cdot)$ is the sample variance operator and U_1 and U_2 are parameters that control the degree of regularization.

The motivation behind the model (PBR) is intuitive and straight-forward: for a fixed portfolio w, the point estimate $\widehat{Risk}_n(w^\top X)$ of the objective has a confidence interval around it, which is approximately equal to the sample standard deviation of the estimator $\widehat{Risk}_n(w^\top X)$. As w varies, the error associated with the point estimate varies, as the confidence interval is a function of w. The PBR constraint $SVar(\widehat{Risk}_n(w^\top X)) \le U_1$ dictates that any solution w that is associated with a large estimation error of the objective function be removed from consideration, which is sensible since such a decision would be highly unreliable. A similar interpretation can be made for the second PBR constraint $SVar(w^\top \hat{\mu}_n) \le U_2$. A schematic of the PBR model is shown in Fig. 2.

Another intuition for PBR is obtained via Chebyshev's inequality. Chebyshev's inequality tells us that, for all $\delta > 0$, for some random variable Y,

$$\mathbb{P}(|Y - \mathbb{E}Y| \ge \delta) \le \frac{Var(Y)}{\delta^2}.$$

Thus letting Y equal $\widehat{Risk}_n(w^\top X)$ or $w^\top \hat{\mu}_n$, we see that constraining their sample variances has the effect of constraining the probability that the estimated portfolio risk and return deviate from the

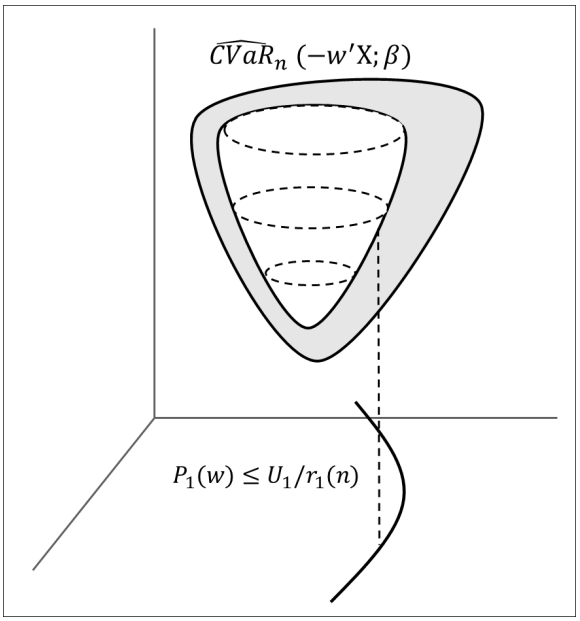

Figure 2 A schematic of PBR on the objective only. The objective function estimated with data is associated with an error (indicated by the grey shading), which depends on the position in the solution space. The PBR constraint cuts out solutions which are associated with large estimation errors of the objective.

true portfolio risk and return by more than a certain amount. In other words, the PBR constraints squeeze the SAA problem (SAA) closer to the true problem (PO) with some probability.

3.1. PBR for Mean-Variance Portfolio Optimization

The PBR model for the mean-variance problem is:

$$\begin{aligned} \hat{w}_{n,MV} = \underset{w\in\mathbb{R}^p}{\operatorname{argmin}}\ & w^\top \hat{\Sigma}_n w \\ s.t.\quad & w^\top 1_p = 1 \\ & (w^\top \hat{\mu}_n = R) \\ & SVar(w^\top \hat{\Sigma}_n w) \le U. \end{aligned} \tag{mv-PBR}$$

Note that we do not regularize the mean constraint as the sample variance of $w^\top \hat{\mu}_n$ is precisely $w^\top \hat{\Sigma}_n w$, which is already captured by the objective.

The following proposition characterizes the sample variance of the sample variance of the portfolio, $SVar(w^\top \hat{\Sigma}_n w)$:

Proposition 1. *The sample variance of the sample variance of the portfolio, $SVar(w^\top \hat{\Sigma}_n w)$ is given by:*

$$SVar(w^\top \hat{\Sigma}_n w) = \Sigma_{i=1}^p \Sigma_{j=1}^p \Sigma_{k=1}^p \Sigma_{l=1}^p w_i w_j w_k w_l \hat{Q}_{ijkl}, \tag{3}$$

where

$$\hat{Q}_{ijkl} = \frac{1}{n}(\hat{\mu}_{4,ijkl} - \hat{\sigma}_{ij}^2\hat{\sigma}_{kl}^2) + \frac{1}{n(n-1)}(\hat{\sigma}_{ik}^2\hat{\sigma}_{jl}^2 + \hat{\sigma}_{il}^2\hat{\sigma}_{jk}^2),$$

where $\hat{\mu}_{4,ijkl}$ is the sample average estimator for $\mu_{4,ijkl}$, the fourth central moment of the elements of X given by

$$\mu_{4,ijkl} = \mathbb{E}[(X_i-\mu_i)(X_j-\mu_j)(X_k-\mu_k)(X_l-\mu_l)]$$

and $\hat{\sigma}^2_{ij}$ is the sample average estimator for σ^2_{ij}, the covariance of the elements of X given by

$$\sigma^2_{ij} = \mathbb{E}[(X_i-\mu_i)(X_j-\mu_j)].$$

Proof. See Appendix A.1.

The PBR constraint of (mv-PBR) is thus a quartic polynomial in the decision vector w. Determining whether a general quartic function is convex is an NP-hard problem [Ahmadi et al. (2013)], hence it is not clear at the outset whether $SVar(w^\top \hat{\Sigma}_n w)$ is a convex function in w, and thus (mv-PBR) a convex problem. We thus consider two convex approximations of (mv-PBR).

3.1.1. Rank-1 approximation of (mv-PBR) Here we make a rank-1 approximation of the quartic polynomial constraint:

$$(w^\top \hat{\alpha})^4 \approx \sum_{ijkl} w_i w_j w_k w_l \hat{Q}_{ijkl},$$

by matching up the diagonals, i.e., $\hat{\alpha}$ is given by

$$\hat{\alpha}_i = \sqrt[4]{\hat{Q}_{iiii}} = \sqrt[4]{\frac{1}{n}\hat{\mu}_{4,iiii} - \frac{n-3}{n(n-1)}(\hat{\sigma}^2_{ii})^2}. \tag{4}$$

We thus obtain the following convex approximation of (mv-PBR):

$$\begin{aligned} \hat{w}_{n,PBR1} = \underset{w\in\mathbb{R}^p}{\operatorname{argmin}}\ & w^\top \hat{\Sigma}_n w \\ s.t.\quad & w^\top 1_p = 1 \\ & (w^\top \hat{\mu}_n = R) \\ & w^\top \hat{\alpha} \le \sqrt[4]{U}, \end{aligned} \tag{mv-PBR-1}$$

where $\hat{\alpha}$ is given in (4).

We can state the effect of PBR constraint as in (mv-PBR-1) on the SAA solution explicitly as follows.

Proposition 2. *The solution to (mv-PBR-1) with the mean constraint $w^\top \hat{\mu}_n = R$ is given by*

$$\hat{w}_{n,PBR1} = \hat{w}_{n,MV} - \frac{1}{2}\lambda^* \hat{\Sigma}_n^{-1}(\beta_1 1_p + \beta_2 \hat{\mu}_n + \hat{\alpha}), \tag{5}$$

where $\hat{w}_{n,MV}$ is the SAA solution, λ^* is the optimal Lagrange multiplier for the PBR constraint $w^\top \alpha \le \sqrt[4]{U}$ and

$$\begin{aligned} \beta_1 &= \frac{\hat{\alpha}^\top \hat{\Sigma}_n^{-1}\hat{\mu}_n . \hat{\mu}_n^\top \hat{\Sigma}_n^{-1} 1_p - \hat{\alpha}^\top \hat{\Sigma}_n^{-1} 1_p . \hat{\mu}_n^\top \hat{\Sigma}_n^{-1}\hat{\mu}_n}{1_p^\top \hat{\Sigma}_n^{-1} 1_p . \hat{\mu}_n^\top \hat{\Sigma}_n^{-1}\hat{\mu}_n - (\hat{\mu}_n^\top \hat{\Sigma}_n^{-1} 1_p)^2}, \\ \beta_2 &= \frac{\hat{\alpha}^\top \hat{\Sigma}_n^{-1}\hat{\mu}_n . 1_p^\top \hat{\Sigma}_n^{-1} 1_p - \hat{\alpha}^\top \hat{\Sigma}_n^{-1} 1_p . 1_p^\top \hat{\Sigma}_n^{-1}\hat{\mu}_n}{1_p^\top \hat{\Sigma}_n^{-1} 1_p . \hat{\mu}_n^\top \hat{\Sigma}_n^{-1}\hat{\mu}_n - (\hat{\mu}_n^\top \hat{\Sigma}_n^{-1} 1_p)^2}. \end{aligned}$$

The solution to (mv-PBR-1) without the mean constraint is given by

$$\hat{w}_{n,PBR1} = \hat{w}_{n,MV} - \frac{1}{2}\lambda^* \hat{\Sigma}_n^{-1}(\beta 1_p + \hat{\alpha}), \tag{6}$$

where $\hat{w}_{n,MV}$ is the SAA solution, λ^* is the optimal Lagrange multiplier for the PBR constraint $w^\top \alpha \leq \sqrt[4]{U}$ and

$$\beta = -\frac{1_p^\top \hat{\Sigma}_n^{-1}\hat{\alpha}}{1_p^\top \hat{\Sigma}_n^{-1} 1_p}.$$

Remark. The effect of rank-1 approximation PBR on the Markowitz problem is thus to tilt the optimal portfolio, by an amount scaled by λ^*, towards a direction that depends on the (approximated) fourth moment of the asset returns.

Proof. See Appendix A.2

3.1.2. Best convex quadratic approximation of (mv-PBR) We also consider a convex quadratic approximation of the quartic polynomial constraint:

$$(w^\top A w)^2 \approx \sum_{ijkl} w_i w_j w_k w_l \hat{Q}_{ijkl},$$

where A is a positive semidefinite (PSD) matrix. Expanding the left-hand side (lhs), we get

$$\sum_{ijkl} w_i w_j w_k w_l A_{ij} A_{kl}.$$

Let us require the elements of A to be as close as possible to the pair-wise terms in Q, i.e., $A_{ij}^2 \approx \hat{Q}_{ijij}$. Then the best PSD matrix A that approximates $\hat{Q}$ in this way is given by solving the following semidefinite program (PSD):

$$A^* = \underset{A \succeq 0}{\operatorname{argmin}} \; ||A - Q_2||_F \tag{Q approx}$$

where $||\cdot||_F$ denotes the Frobenius norm, and where Q_2 is a matrix such that its ij-th element equals $\hat{Q}_{ijij}$. We thus obtain the following convex quadratic approximation of (mv-PBR):

$$\begin{aligned} \hat{w}_{n,PBR2} = \underset{w \in \mathbb{R}^p}{\operatorname{argmin}} \; & w^\top \hat{\Sigma}_n w \\ s.t. \quad & w^\top 1_p = 1 \\ & (w^\top \hat{\mu}_n = R), \\ & w^\top A^* w \leq \sqrt{U}. \end{aligned} \tag{mv-PBR-2}$$

We can state the effect of PBR constraint as in (mv-PBR-2) on the SAA solution explicitly as follows.

PROPOSITION 3. *The solution to (mv-PBR-2) with the mean constraint $w^\top \hat{\mu}_n = R$ is given by*

$$\hat{w}_{n,PBR2} = -\frac{1}{2}\tilde{\Sigma}_n(\lambda^*)^{-1}(\nu_1^*(\lambda^*)1_p + \nu_2^*(\lambda^*)\hat{\mu}_n), \tag{7}$$

where $\tilde{\Sigma}_n(\lambda^*) = \hat{\Sigma}_n + \lambda^* A^*$, λ^* is the optimal Lagrange multiplier for the PBR constraint $w^\top A^* w \leq \sqrt{U}$ and

$$\nu_1^*(\lambda) = 2\frac{R\hat{\mu}_n^\top \tilde{\Sigma}^{-1}1_p - \hat{\mu}_n^\top \tilde{\Sigma}^{-1}\hat{\mu}_n}{1_p^\top \tilde{\Sigma}^{-1}1_p.\hat{\mu}_n^\top \tilde{\Sigma}^{-1}\hat{\mu}_n - (\hat{\mu}_n^\top \tilde{\Sigma}^{-1}1_p)^2},$$
$$\nu_2^*(\lambda) = 2\frac{-R1_p^\top \tilde{\Sigma}^{-1}1_p + \hat{\mu}_n^\top \tilde{\Sigma}^{-1}1_p}{1_p^\top \tilde{\Sigma}^{-1}1_p.\hat{\mu}_n^\top \tilde{\Sigma}^{-1}\hat{\mu}_n - (\hat{\mu}_n^\top \tilde{\Sigma}^{-1}1_p)^2}.$$

The solution to (mv-PBR-2) without the mean constraint is given by

$$\hat{w}_{n,PBR2} = \frac{\tilde{\Sigma}_n(\lambda^*)^{-1}1_p}{1_p^\top \tilde{\Sigma}_n(\lambda^*)^{-1}1_p}, \tag{8}$$

where $\tilde{\Sigma}_n(\lambda^*) = \hat{\Sigma}_n + \lambda^* A^*$ and λ^* is the optimal Lagrange multiplier for the PBR constraint $w^\top A^* w \leq \sqrt{U}$, as before.

Proof. See Appendix A.3.

For both mean-constrained and unconstrained cases, notice that the solution depends on λ only through the matrix $\tilde{\Sigma}_n(\lambda^*) = \hat{\Sigma}_n + \lambda^* A^*$. We thus retrieve the unregularized SAA solution $\hat{w}_{n,MV}$ when λ is set to zero. Thus the PSD approximation to (mv-PBR) is equivalent to using a different estimator for the covariance matrix than the sample covariance matrix $\hat{\Sigma}_n$. Clearly, $\tilde{\Sigma}_n(\lambda^*)$ adds a bias to the sample covariance matrix estimator. It is well-known that adding some bias to a standard estimator can be beneficial, and such estimators are known as shrinkage estimators. Haff (1980) and Ledoit and Wolf (2004) have explored this idea for the sample covariance matrix by shrinking the sample covariance matrix towards the identity matrix, and have shown superior properties of the shrunken estimator. In contrast, our PBR model shrinks the sample covariance matrix towards a direction that is approximately equal to the variance of the sample covariance matrix. Conversely, DeMiguel et al. (2009a) showed that using the shrinkage estimator for the covariance matrix as in Ledoit and Wolf (2004) is equivalent to L_2 regularization; and in Sec. 5 we compare the two methods.

3.2. PBR for Mean-CVaR Portfolio Optimization

The PBR model for the mean-CVaR problem is:

$$\begin{array}{ll} \min\limits_{w\in\mathbb{R}^p} & \widehat{CVaR}_n(-w^\top \mathsf{X};\beta) \\ s.t. & w^\top 1_p = 1 \\ & (w^\top \hat{\mu}_n = R) \\ & SVar(\widehat{CVaR}_n(-w^\top \mathsf{X};\beta)) \leq U_1 \\ & (SVar(w^\top \hat{\mu}_n) \leq U_2). \end{array} \tag{cv-PBR}$$

The variance of $w^\top \hat{\mu}_n$ is given by

$$Var(w^\top \hat{\mu}_n) = \frac{1}{n^2}\sum_{i=1}^{n} Var(w^\top X_i) = \frac{1}{n} w^\top \Sigma w,$$

hence $SVaR(w^\top \hat{\mu}_n) = n^{-1} w^\top \hat{\Sigma}_n w$. The variance of $\widehat{CVaR}_n(-w^\top \mathsf{X}; \beta)$ is given by the following proposition.

PROPOSITION 4. *Suppose* $\mathsf{X} = [X_1, \ldots, X_n] \overset{iid}{\sim} F$, *where* F *is absolutely continuous with twice continuously differentiable pdf. Then*

$$Var[\widehat{CVaR}_n(-w^\top \mathsf{X}; \beta)] = \frac{1}{n(1-\beta)^2} Var[(-w^\top X - \alpha_\beta(w))^+] + O(n^{-2}),$$

where

$$\alpha_\beta(w) = \inf\{\alpha : P(-w^\top X \geq \alpha) \leq 1-\beta\},$$

the Value-at-Risk (VaR) of the portfolio w *at level* β.

Proof. See Appendix A.4.

Thus, the sample variance of $\widehat{CVaR}_n(-w^\top \mathsf{X}; \beta)$ is, to first order,

$$SVar[\widehat{CVaR}_n(-w^\top \mathsf{X}; \beta)] = \frac{1}{n(1-\beta)^2} z^\top \Omega_n z,$$

where $\Omega_n = \frac{1}{n-1}[I_n - n^{-1} 1_n 1_n^\top]$, I_n being the $n \times n$ identity matrix, and $z_i = \max(0, -w^\top X_i - \alpha)$ for $i = 1, \ldots, n$.

Incorporating the above formulas for the sample variances, (cv-PBR) can be written as:

$$\begin{aligned} \min_{\alpha, w, z} \quad & \alpha + \frac{1}{n(1-\beta)} \sum_{i=1}^{n} z_i \\ s.t. \quad & w^\top 1_p = 1 \\ & (w^\top \hat{\mu}_n = R) \\ & \frac{1}{n(1-\beta)^2} z^\top \Omega_n z \leq U_1 \\ & z_i = \max(0, -w^\top X_i - \alpha),\ i = 1, \ldots, n \\ & \frac{1}{n} w^\top \hat{\Sigma}_n w \leq U_2 \end{aligned} \qquad \text{(cv-PBR}')$$

(cv-PBR′) is non-convex due to the cutoff variables $z_i = \max(0, -w^\top X_i - \alpha)$, $i = 1, \ldots, n$. Without the regularization constraint $[n(1-\beta)^2]^{-1} z^\top \Omega_n z \leq U_1$, one can solve the problem by relaxing the non-convex constraint $z_i = \max(0, -w^\top X_i - \alpha)$ to $z_i \geq 0$ and $z_i \geq -w^\top X_i - \alpha$. However, $z^\top \Omega_n z$ is not a monotone function of z hence it is not clear at the outset whether one can employ such a relaxation trick for the regularized problem.

(cv-PBR′) is a combinatorial optimization problem because one can solve it by considering all possible combinations of $\lfloor n(1-\beta) \rfloor$ out of n observations that contribute to the worst $(1-\beta)$ of the

portfolio loss (which determines the non-zero elements of z), then finding the portfolio weights that solve the problem based on these observations alone. Clearly, this is an impractical strategy; for example, there are 34220 possible combinations to consider for a modest number of observations $n=60$ (5 years of monthly data) and $\beta=0.95$.

However, it turns out that relaxing $z_i=\max(0,-w^\top X_i-\alpha)$, $i=1,\ldots,n$ does result in a tight convex relaxation. The resulting problem is a quadratically-constrained quadratic program (QCQP) which can be solved efficiently. Before formally stating this result, let us first introduce the convex relaxation of (cv-PBR$'$):

$$\begin{array}{rll} \min\limits_{\alpha,w,z} & \alpha+\dfrac{1}{n(1-\beta)}\displaystyle\sum_{i=1}^n z_i & \\ s.t. & w^\top 1_p=1 & (\nu_1)\\ & (w^\top\hat{\mu}_n=R) & (\nu_2)\\ & \dfrac{1}{n(1-\beta)^2}z^\top\Omega_n z\le U_1 & (\lambda_1)\\ & z_i\ge 0\ i=1,\ldots,n & (\eta_1)\\ & z_i\ge -w^\top X_i-\alpha,\ i=1,\ldots,n & (\eta_2)\\ & \dfrac{1}{n}w^\top\hat{\Sigma}_n w\le U_2 & (\lambda_2) \end{array} \tag{cv-relax}$$

and its dual (where the dual variables correspond to the primal constraints as indicated above):

$$\begin{array}{ll} \max\limits_{\nu_1,\nu_2,\lambda_1,\lambda_2,\eta_1,\eta_2} & g(\nu_1,\nu_2,\eta_1,\eta_2,\lambda_1,\lambda_2)\\ s.t. & \eta_2^\top 1_n=1\\ & \lambda_1\ge 0,\lambda_2\ge 0\\ & \eta_1\ge 0,\eta_2\ge 0 \end{array} \tag{cv-relax-d}$$

where

$$\begin{aligned} g(\nu_1,\nu_2,\lambda_1,\lambda_2,\eta_1,\eta_2)=&-\frac{n}{2\lambda_1}(\nu_1 1_p+\nu_2\hat{\mu}_n-\mathsf{X}\eta_2)^\top\hat{\Sigma}_n^{-1}(\nu_1 1_p+\nu_2\hat{\mu}_n-\mathsf{X}\eta_2)\\ &-\frac{n(1-\beta)^2}{2\lambda_2}(\eta_1+\eta_2)^\top\Omega_n^\dagger(\eta_1+\eta_2)+\nu_1+R\nu_2-U_1\lambda_1-U_2\lambda_2, \end{aligned}$$

and $\Omega_n^\dagger$ is the Moore-Penrose pseudo inverse of the singular matrix Ω_n.

We now state the result that (cv-PBR$'$) is a tractable optimization problem because its convex relaxation is essentially tight.

THEOREM 1. *Let $(\alpha^*,w^*,z^*,\lambda_1^*,\lambda_2^*,\eta_1^*,\eta_2^*)$ be the primal-dual optimal point of (cv-PBR$'$) and (cv-relax-d). If $\eta_2^*\neq 1_n/n$, then (α^*,w^*,z^*) is an optimal point of (cv-PBR$'$). Otherwise, if $\eta_2^*=1_n/n$, we can find the optimal solution to (cv-PBR$'$) by solving (cv-relax-d) with an additional constraint $\eta_2^\top 1_n\ge\delta$, where δ is any constant $0<\delta<1$.*

Proof. See Appendix A.5.

Remark. Theorem 1 shows that one can solve (cv-PBR′) via at most two steps. The first step is to solve (cv-PBR′); if the dual variables corresponding to the constraints $z_i \geq -w^\top X_i - \alpha,\ i=1,\ldots,n$ are all equal to $1/n$, then we solve (cv-relax-d) with an additional constraint $\eta_2^\top 1_n \geq \delta$, where δ is any constant $0 < \delta \ll 1$, otherwise the relaxed solution is feasible for the original problem hence optimal. For the record, all problem instances solved in the numerical section Sec. 5 were solved in a single step.

3.3. Robust Counterparts of PBR models

In this section, we show that the three PBR portfolio optimization models can be transformed into robust optimization (RO) problems.

Proposition 5. *The convex approximations to the Markowitz PBR problem (mv-PBR) has the following robust counterpart representation:*

$$\begin{array}{rll} \hat{w}_{n,PBR1} = \underset{w\mathbb{R}^p}{\text{argmin}} & w^\top \hat{\Sigma}_n w & \\ s.t. & w^\top 1_p = 1 & \\ & (w^\top \hat{\mu}_n = R) & \\ & \max\limits_{u\in\mathcal{U}} w^\top u \leq \sqrt[4]{U}, & \end{array} \tag{mv-PBR-RO}$$

where $\mathcal{U}$ is the ellipsoid

$$\mathcal{U} = \{u \in \mathbb{R}^p \mid u^\top P^\dagger u \leq 1,\ (I - PP^\dagger)u = 0\},$$

with $P = \alpha\alpha^\top$ for (mv-PBR-1) and $P = A^*$ for (mv-PBR-2), and $P^\dagger$ denoting the Moore-Penrose pseudoinverse of the matrix P (which equals the inverse if P is invertible, which is the case for $P = A^*$).

Proof. See Appendix A.6.

Proposition 6. *The the mean-CVaR PBR problem (cv-PBR) has the following robust counterpart representation:*

$$\begin{array}{rll} \min\limits_{\alpha,w,z} & \alpha + \dfrac{1}{n(1-\beta)}\displaystyle\sum_{i=1}^n z_i & \\ & s.t. \quad w^\top 1_p = 1 & \\ & (w^\top \hat{\mu}_n = R) & \\ & \max\limits_{u\in\mathcal{U}_1} z^\top u \leq \sqrt{U_1} & \\ z_i = \max(0, -w^\top X_i - \alpha),\ i=1,\ldots,n & & \\ (\max\limits_{\tilde{\mu}\in\mathcal{U}_2} w^\top(\tilde{\mu} - \mu) \leq \sqrt{U_2}) & & \end{array} \tag{cv-PBR-RO}$$

where $\mathcal{U}_1$ is the ellipsoid

$$\mathcal{U}_1 = \{\mu \in \mathbb{R}^p \mid (\tilde{\mu} - \mu)^\top \hat{\Sigma}_n^{-1} (\tilde{\mu} - \mu) \leq 1\},$$

and $\mathcal{U}_2$ *is the ellipsoid*

$$\mathcal{U}_2 = \{u \in \mathbb{R}^n \mid u^\top \Omega_n^\dagger u \leq 1,\ 1_p^\top u = 0\},$$

where $\Omega_n^\dagger$ *is the Moore-Penrose pseudoinverse of the matrix* Ω_n.

Proof. One can follow similar steps to the proof of Proposition 5.

While the PBR constraint on the portfolio mean is equivalent to the mean uncertainty constraint considered in Delage and Ye (2010), the PBR constraint on the portfolio variance for the mean-variance problem leads to a new uncertainty set which is different from Delage and Ye (2010). The main difference is that Delage and Ye (2010) considers an uncertainty set for the sample covariance matrix separately from the decision, whereas PBR considers protecting against estimation errors in the portfolio variance, thereby considering both the decision and the covariance matrix together. The difference is detailed in Appendix B.

4. Asymptotic Optimality of SAA and PBR solutions

In this section, we show that the SAA solution $\hat{w}_n$ and the PBR solutions are asymptotically optimal under the mild condition that the true solution be well-separated (i.e., identifiable). In other words, we show that the SAA solution converges in probability to the true optimal w_0 as the number of observations n tends to infinity. We then show that the performances of the estimated solutions also converge to that of the true optimal, i.e., the return-risk frontiers corresponding to $\hat{w}_n$ and $\hat{w}_{n,PBR}$ converge to the efficient frontier of w_0.

For ease of exposition and analysis, we will work with the following transformation of the original problem:

$$\min_{\theta=(\alpha,v)\in\mathbb{R}\times\mathbb{R}^{p-1}} M(\theta) = \min_{\theta=(\alpha,v)\in\mathbb{R}\times\mathbb{R}^{p-1}} \mathbb{E}[m_\theta(X)] \tag{PO$'$}$$

where we have re-parameterized w to $w = w_1 + Lv$, where $L = [0_{(p-1)\times 1}, I_{(p-1)\times(p-1)}]^\top$, $v = [w_2, \ldots, w_p]^\top$ and $w_1 = [1 - v^\top 1_{(p-1)}, 0_{1\times(p-1)}]^\top$, and

$$m_\theta(x) = ((w_1 + Lv)^\top x - (w_1 + Lv)^\top \mu)^2 - \lambda_0 (w_1 + Lv)^\top x, \tag{9}$$

for the mean-variance problem (MV-true), and

$$m_\theta(x) = \alpha + \frac{1}{1-\beta} z_\theta(x) - \lambda_0 (w_1 + Lv)^\top x, \tag{10}$$

for the mean-CVaR problem (CV-true), where $z_\theta(x) = \max(0, -(w_1 + Lv)^\top x - \alpha)$. In other words, we have transformed (PO) to a global optimization problem, where $\lambda_0 > 0$ determines the investor's utility on the return. Without loss of generality, we restrict the problem (PO') to optimizing over a compact subset Θ of $\mathbb{R} \times \mathbb{R}^{p-1}$.

We now prove asymptotic optimality of the SAA solution to the mean-variance and mean-CVaR problems.

THEOREM 2 **(Asymptotic Optimality of SAA solution of mean-variance problem)**. *Consider (PO′) with $m_\theta(\cdot)$ as in (9). Denote the solution by θ_{MV}. Suppose, for all $\epsilon > 0$,*

$$\sup_{\theta\in\Theta}\{M(\theta):\ d(\theta,\theta_{MV}) \geq \epsilon\} < M(\theta_{MV}). \qquad (*)$$

Then, as n tends to infinity,

$$\hat{\theta}_{n,MV} \xrightarrow{P} \theta_{MV},$$

where $\hat{\theta}_{n,MV}$ is the solution to the SAA problem

$$\min_{\theta\in\Theta}\ M_n(\theta) = \min_{\theta\in\Theta}\ \frac{1}{n}\sum_{i=1}^{n}((w_1+Lv)^\top X_i - (w_1+Lv)^\top\hat{\mu}_n)^2 - \lambda_0(w_1+Lv)^\top X_i.$$

THEOREM 3 **(Asymptotic Optimality of SAA solution of mean-CVaR problem)**. *Consider (PO′) with $m_\theta(\cdot)$ as in (10). Denote the solution by θ_{CV}. Suppose, for all $\epsilon > 0$,*

$$\sup_{\theta\in\Theta}\{M(\theta):\ d(\theta,\theta_{CV}) \geq \epsilon\} < M(\theta_{CV}) \qquad (**)$$

Then, as n tends to infinity,

$$\hat{\theta}_{n,CV} \xrightarrow{P} \theta_{CV},$$

where $\hat{\theta}_{n,CV}$ is the solution to the SAA problem

$$\min_{\theta\in\Theta}\ M_n(\theta) = \min_{\theta\in\Theta}\ \alpha + \frac{1}{n}\sum_{i=1}^{n}\frac{1}{1-\beta}z_\theta(X_i) - \lambda_0(w_1+Lv)^\top X_i.$$

Sketch of the proofs of Theorems 2 and 3. Theorems 2 and 3 are statements about the asymptotic consistency of estimated quantities $\hat{\theta}_{n,MV}$ and $\hat{\theta}_{n,CV}$ to their true respective quantities θ_{MV} and θ_{CV}. While proving (statistical) convergence of sample average-type estimators for independent samples is straight-forward, proving convergence of *solutions* of estimated optimization problems is more involved.

In mathematical statistics, the question of whether solutions of estimated optimization problems converge arises in the context of maximum likelihood estimation, whose study goes back as far as seminal works of Fisher (1922) and Fisher (1925). Huber initiated a systematic study of M-estimators (where "M" stands for maximization; i.e., estimators that arise as solutions to maximization problems) with Huber (1967), which subsequently led to asymptotic results that apply to more general settings (e.g., non-differentiable objective functions) that rely on the theory of empirical processes. Van der Vaart (2000) gives a clean, unified treatment of the main results in the theory of M-estimation, and we align our proof to the setup laid out in this book.

In particular, Van der Vaart (2000) gives general conditions under which the solution of a static optimization problem estimated from data converges to the true value as the sample size grows.

In words, the conditions correspond to the near-optimality of the estimator (for the estimated problem), that the true parameter value be well-defined, and that the estimated objective function converge uniformly to the true objective function over the domain. The first condition is satisfied because we assume $\hat{\theta}_{n,MV}$ and $\hat{\theta}_{n,CV}$ are optimal for the estimated problem. The second condition is an identifiability condition, which we assume holds for our problems via (*) and (**). This is a mild criterion that is necessary for statistical inference; e.g., it suffices that θ_{MV} (resp. θ_{CV}) be unique, Θ compact and $M(\cdot)$ continuous.

The third and final condition is the uniform convergence of the estimated objective function $M_n(\cdot)$ to its true value $M(\cdot)$. This is not a straight-forward result, especially if the objective function is not differentiable, which is the case for the mean-CVaR problem. Showing uniform convergence for such functions requires intricate arguments that involve bracketing numbers (see Chapter 19 of Van der Vaart (2000)). The proofs of Theorems 2 and 3 can be found in Appendix C.

4.1. Asymptotic optimality of PBR solutions

Let us now consider the PBR portfolio optimization problem. With similar global transformation as above, the PBR problem becomes

$$\hat{\theta}_{n,PBR} = \underset{\theta=(\alpha,v)\in\mathbb{R}\times\mathbb{R}^{p-1}}{\operatorname{argmin}} M_n(\theta;\lambda_1,\lambda_2) \tag{PBR$'$}$$

where

$$M_n(\theta;\lambda_1,\lambda_2) = w^\top \hat{\Sigma}_n w - \lambda_0 w^\top \hat{\mu}_n + \lambda_1 w^\top \alpha, \tag{11}$$

where α is as in (4), for the mean-variance problem (mv-PBR-1),

$$M_n(\theta;\lambda_1,\lambda_2) = w^\top \hat{\Sigma}_n w - \lambda_0 w^\top \hat{\mu}_n + \lambda_1 w^\top A^* w, \tag{12}$$

where A^* is as in (Q approx), for the mean-variance problem (mv-PBR-2), and

$$M_n(\theta;\lambda_1,\lambda_2) = \frac{1}{n}\sum_{i=1}^n m_\theta(X_i) + \frac{\lambda_1}{n} w^\top \hat{\Sigma}_n w + \frac{\lambda_2}{n(n-1)(1-\beta)^2}\sum_{i=1}^n \left(z_\theta(X_i) - \frac{1}{n}\sum_{j=1}^n z_\theta(X_j)\right)^2, \tag{13}$$

for the mean-CVaR problem (cv-PBR). Note $\lambda_1, \lambda_2 \geq 0$ are parameters that control the degree of regularization; they play the same role as U_1 and U_2 in the original problem formulation.

We now prove asymptotic optimality of the PBR solutions.

THEOREM 4. *Assume (*) and (**). Then, as n tends to infinity,*

$$\hat{\theta}_{n,PBR}(\lambda_1,\lambda_2) \xrightarrow{P} \theta_0,$$

where $\hat{\theta}_{n,PBR}(\lambda_1,\lambda_2)$ are minimizers of $M_n(\theta,\lambda_1,\lambda_2)$ equal to (11), (12) and (13), and θ_0 is the corresponding true solution.

The following result is an immediate consequence of Theorem 4, by the Continuous Mapping Theorem.

COROLLARY 1 **(Convergence of performance of PBR solutions)**. *Assume the same setting as Theorem 4. Then the performance of the PBR solution also converges to the true performance of the true optimal solution, i.e.,*

$$|\hat{w}_{n,PBR}^{\top}\mu - w_0^{\top}\mu| \xrightarrow{P} 0$$

and

$$|Risk(\hat{w}_{n,PBR}^{\top}X;\beta) - Risk(w_0^{\top}X;\beta)| \xrightarrow{P} 0,$$

as n tends to infinity, where $\hat{w}_{n,PBR}$ is the portfolio allocation corresponding to $\hat{\theta}_{n,PBR}$.

5. Results on Empirical Data

In this section, we compare the PBR method against a number of key benchmarks on three data sets: the five, ten and forty-nine industry portfolios from Ken French's Website, which report monthly excess returns over the 90-day nominal US T-bill. We take the most recent 20 years of data, covering the period from January 1994 to December 2013. Our computations are done on a rolling horizon basis, with the first 10 years of observations used as training data ($N_{train} = 120$) and the last 10 years of observations used as test data ($N_{test} = 120$). All computations were carried out on MATLAB2013a with the solver MOSEK and CVX, a package for specifying and solving convex programs Grant and Boyd (2014, 2008) on a Dell Precision T7600 workstation with two Intel Xeon E5-2643 processors, each of which has 4 cores, and 32.0 GB of RAM.

5.1. Portfolio strategies considered for the mean-variance problem

We compute the out-of-sample performances of the following eight portfolio allocation strategies:

1. **SAA:** solving the sample average approximation problem (MV-SAA).

2. **PBR (rank-1):** solving the rank-1 approximation problem (mv-PBR-1). The rhs of the PBR constraint, $\sqrt[4]{U}$, is calibrated using the out-of-sample performance-based k-cross validation algorithm (OOS-PBCV) which we explain in detail in Sec. 5.4.

3. **PBR (PSD):** solving the convex quadratic approximation problem (mv-PBR-2). The rhs of the PBR constraint, $\sqrt[2]{U}$, calibrated using OOS-PBCV.

4. **NS:** solving the problem (MV-SAA) with the no short-selling constraint $w \geq 0$, as in Jagannathan and Ma (2003).

5. **L1 regularization:** solving the SAA problem (MV-SAA) with the extra constraint $||w||_1 \leq U$, where U is also calibrated using OOS-PBCV.

6. **L2 regularization:** solving the SAA problem (MV-SAA) with the extra constraint $||w||_2 \leq U$, where U is also calibrated using OOS-PBCV.

7. **Min. Variance:** Solving the above (SAA, PBR (rank-1), PBR (PSD), NS, L1, L2) for the global minimum variance problem, which is (MV-true) without the mean return constraint. We do this because the difficulty in estimating the mean return is a well-known problem [Merton (1980)] and some recent works in the Markowitz literature have shown that removing the mean constraint altogether can yield better results [Jagannathan and Ma (2003)].

8. **Equally-weighted portfolio:** DeMiguel et al. (2009b) has shown that the naive strategy of equally dividing up the total wealth (i.e., investing in a portfolio w with $w_i = 1/p$ for $i = 1, \ldots, p$) performs very well relative to a number of benchmarks for the data-driven mean-variance problem. We include this as a benchmark.

5.2. Portfolio strategies considered for the mean-CVaR problem

We compute the out-of-sample performances of the following eight portfolio allocation strategies:

1. **SAA:** solving the sample average approximation problem (CV-SAA).

2. **PBR only on the objective:** solving the problem (cv-PBR) with no regularization of the mean constraint, i.e., $U_2 = \infty$. The rhs of the objective regularization constraint, U_1, is calibrated using the out-of-sample performance-based k-cross validation algorithm (OOS-PBCV) which we explain in detail in Sec. 5.4.

3. **PBR only on the constraint:** solving the problem (cv-PBR) with no regularization of the objective function, i.e., $U_1 = \infty$. The rhs of the mean regularization constraint, U_2, is calibrated using OOS-PBCV.

4. **PBR on both the objective and the constraint:** solving the problem (cv-PBR). Both regularization parameters U_1 and U_2 are calibrated using OOS-PBCV.

5. **L1 regularization:** solving the sample average approximation problem (cv-PBR) with the extra constraint $||w||_1 \leq U$, where U is also calibrated using OOS-PBCV.

6. **L2 regularization:** solving the sample average approximation problem (cv-PBR) with the extra constraint $||w||_2 \leq U$, where U is also calibrated using OOS-PBCV.

7. **Equally-weighted portfolio**: DeMiguel et al. (2009b) has shown that the naive strategy of equally dividing up the total wealth (i.e., investing in a portfolio w with $w_i = 1/p$ for $i = 1, \ldots, p$) performs very well relative to a number of benchmarks for the data-driven mean-variance problem. We include this as a benchmark.

8. **Global minimum CVaR portfolio:** solving the sample average approximation problem (CV-SAA) without the target mean return constraint $w^\top \hat{\mu}_n = R$. We do this because the difficulty in estimating the mean return is a well-known problem [Merton (1980)] and some recent works in

the Markowitz literature has shown that removing the mean constraint altogether can yield better results [Jagannathan and Ma (2003)]. Thus as an analogy to the global minimum variance problem we consider the global minimum CVaR problem.

5.3. Evaluation Methodology

We evaluate the various portfolio allocation models on a rolling-horizon basis. In other words, we evaluate the portfolio weights on the first N_{train} asset return observations (the "training data") then compute its return on the $(N_{train}+1)$-th observation. We then roll the window by one period, evaluate the portfolio weights on the 2nd to $N_{train}+1$-th return observations, then compute its return on the $N_{train}+2$-th observation, and so on, until we have rolled forward N_{test} number of times. Let us generically call the optimal portfolio weights solved over N_{test} number of times $\hat{w}_1,\ldots,\hat{w}_{N_{test}}\in\mathbb{R}^p$ and the asset returns $X_1,\ldots,X_{N_{test}}\in\mathbb{R}^p$. Also define

$$\hat{\mu}_{test} := \frac{1}{N_{test}}\sum_{t}^{N_{test}} \hat{w}_t^\top X_t$$
$$\hat{\sigma}^2_{test} := \frac{1}{N_{test}-1}\sum_{t}^{N_{test}} (\hat{w}_t^\top X_t - \hat{\mu}_{test})^2,$$

i.e., the out-of-sample mean and variance of the portfolio returns.

We report the following performance metrics:

- **Sharpe Ratio**: we compute annualized Sharpe ratio as

$$\text{Sharpe ratio} = \frac{\hat{\mu}_{test}}{\hat{\sigma}_{test}} \tag{14}$$

- **Turnover**: the portfolio turn over, averaged over the testing period, is given by

$$Turnover = \frac{1}{N_{test}}\sum_{t=1}^{N_{test}}\sum_{j=1}^{p} |\hat{w}_{t+1,j} - \hat{w}_{t,j}|^+. \tag{15}$$

For further details on these performance measures we refer the reader to DeMiguel et al. (2009b).

5.4. Calibration algorithm for U: performance-based k-fold cross-validation

One important question in solving (PBR) is how to choose the right hand side of the regularization constraints, U_1 and U_2. If they are set too small, the problem is infeasible, and if set too large, regularization has no effect and we retrieve the SAA solution. Ideally, we want to choose U_1 and U_2 so that it constrains the SAA problem just enough to maximize the out-of-sample performance. Obviously, one cannot use the actual test data set to calibrate U_1 and U_2, and we need to calibrate them on the training data set via a cross-validation (CV) method.

A common CV technique used in statistics is the k-fold CV. It works by splitting the training data set into k equally-sized bins, training the statistical model on every possible combination of

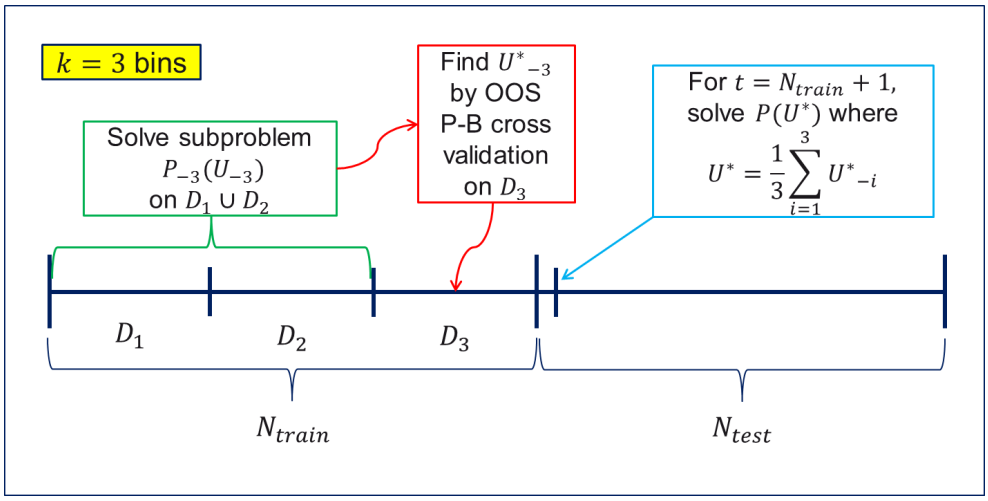

Figure 3 A schematic explaining the out-of-sample performance-based k-cross validation (OOS-PBCV) algorithm used to calibrate the constraint rhs, U, for the case $k = 3$. The training data set is split into k bins, and the optimal U for the entire training data set is found by averaging the best U found for each subset of the training data.

$k-1$ bins and then validating on the remaining bin. Any parameter that needs to be tuned is tuned via the prediction accuracy on the validation data set.

Here we develop a performance-based k-fold CV method to find U_1 and U_2 that maximize the out-of-sample Sharpe ratio on the validation data set. The two key differences between our algorithm and the standard k-fold CV is that (i) the search boundaries for U_1 and U_2 need to be set carefully in order to avoid infeasibility and having no effect, and (ii) we validate by computing the Sharpe ratio (the main performance metric for investment in practice) as opposed to some measure of error.

For simplicity, we explain the algorithm for the case of having just one regularization constraint on the objective. We thus omit the subscript and refer to the rhs by U instead of U_1. Generalization to the two-dimensional case is straight-forward. Figure 3 displays a schematic explaining the main parts of the algorithm, for the case $k = 3$. Let $D = [X_1, \ldots, X_{N_{train}}] \in \mathbb{R}^{p \times N_{train}}$ be the training data set of stock returns. This is split into k equally sized bins, $D_1, D_2, \ldots, D_k$. Let $P_{-i}(U_{-i})$ denote the PBR problem solved on the data set $D \backslash D_i$ with rhs $U = U_{-i}$. We find the optimal U, denoted by U^*, on the whole data set D by the following steps:

1. Set a search boundary for U_{-i}, $[\underline{U}_{-i}, \bar{U}_{-i}]$.
2. Solve $P_{-i}(U_{-i})$ on $D \backslash D_i$ starting at $U_{-i} = \bar{U}_{-i}$, computing the Sharpe ratio of the solution on D_i, then repeating the process with progressively smaller U_{-i} via a descent algorithm. Find $U^*_{-i} \in [\underline{U}_{-i}, \bar{U}_{-i}]$ by a stopping criterion.
3. Average over the k results to get $U^* = \frac{1}{k}\sum_{i=1}^{k} U^*_{-i}$.

We elaborate on these three parts of the CV algorithm below.

1. Set a search boundary for U_{-i}, $[\underline{U}_{-i}, \bar{U}_{-i}]$. As previously mentioned, setting the correct search boundary for U_{-i} is very important. We require the boundary for the i-th subproblem to be contained within the allowable range for the problem on the entire data set, i.e., $[\underline{U}_{-i}, \bar{U}_{-i}] \subset [\underline{U}, \bar{U}]$. This is because if we solve the PBR problem on the whole training data set with $U > \bar{U}$ then PBR will not have any effect, and likewise if we solve the PBR problem with $U < \underline{U}$, then the problem will be infeasible.

The upper bound on U is given by $\bar{U} = \widehat{Risk}_n(-\hat{w}_n^\top \mathsf{X})$, recalling that $\hat{w}_n$ is the SAA solution. In other words, the upper bound is set to be the value of the PBR penalty if the penalty were not imposed. To find $\underline{U}$, the minimum possible PBR parameter, we solve

$$\begin{array}{rrcl} & \underline{U} & = & \min\limits_{w} \quad w^\top \alpha \\ s.t. & & & w^\top 1_p = 1 \\ & & & w^\top \hat{\mu}_n = R \end{array} \tag{U-min-mv1}$$

for (mv-PBR-1),

$$\begin{array}{rrcl} & \underline{U} & = & \min\limits_{w} \quad w^\top A^* w \\ s.t. & & & w^\top 1_p = 1 \\ & & & w^\top \hat{\mu}_n = R \end{array} \tag{U-min-mv2}$$

for (mv-PBR-2), and

$$\begin{array}{rrcl} & \underline{U} & = & \min\limits_{w,z} \quad z^\top \Omega_n z \\ s.t. & & & w^\top \hat{\mu}_n = R \\ & & & w^\top 1_p = 1 \\ & & & z_i \geq -w^\top X_i - \alpha, \quad i = 1, \ldots, n. \\ & & & z_i \geq 0, \quad i = 1, \ldots, n. \end{array} \tag{U-min-cv}$$

for (cv-PBR).

To find the upper bound on the subproblem, $\bar{U}_{-i}$, we compute (SAA) on dataset $D \backslash D_i$ for $\hat{w}_{-i}$, then set

$$\bar{U}_{-i} = \min[\bar{U}, \widehat{Risk}_n(-\hat{w}_{-i}^\top \mathsf{X})].$$

To find $\underline{U}_{-i}$, we first solve

$$\begin{array}{rrcl} & \underline{U}_{tmp} & = & \min\limits_{w} \quad w^\top \alpha \\ s.t. & & & w^\top 1_p = 1 \\ & & & w^\top \hat{\mu}_{-i} = R \end{array}$$

for (mv-PBR-1), where $\hat{\mu}_{-i}$ the sample mean computed on $D \backslash D_i$,

$$\begin{array}{rrcl} & \underline{U}_{tmp} & = & \min\limits_{w} \quad w^\top A^* w \\ s.t. & & & w^\top 1_p = 1 \\ & & & w^\top \hat{\mu}_{-i} = R \end{array}$$

for (mv-PBR-2), and

$$\begin{aligned} \underline{U}_{tmp} \quad = \quad & \min_{w,z} \quad z^\top \Omega_{-i} z \\ s.t. \quad & w^\top \hat{\mu}_{-i} = R \\ & w^\top 1_p = 1 \\ & z_i \geq -w^\top X_i - \alpha, \quad i \in C \backslash C_i. \\ & z_i \geq 0, \quad i \in C \backslash C_i, \end{aligned}$$

for (cv-PBR), where Ω_{-i} is the sample variance operator computed on $D \backslash D_i$, and C and C_i are sets of labels of the elements in D and D_i respectively. We then set

$$\underline{U}_{-i} = \max[\underline{U}_{tmp}, \underline{U}].$$

The pseudocode for this part of the CV algorithm is shown in Algorithm 1.

2. Finding $U^*_{-i} \in [\underline{U}_{-i}, \bar{U}_{-i}]$. To find the optimal parameter for the i-th subproblem that *maximizes the out-of-sample Sharpe ratio*, we employ a backtracking line search algorithm [see Chapter 9.2. of Boyd and Vandenberghe (2004)], which is a simple yet effective descent algorithm. We start at the maximum $\bar{U}_{-i}$ determined in the previous step and descend by step size $t\Delta U := t(\bar{U}_{-i} - \underline{U}_{-i})/Div$, where Div a preset granularity parameter, t is a parameter that equals 1 initially then is backtracked at rate β, a parameter chosen between 0 and 1, until the stopping criterion

$$Sharpe(U - t\Delta U) < Sharpe(U) + \alpha t \Delta U \frac{dSharpe(U)}{dU}$$

is met.

Computing $dSharpe(U)/dU$, the marginal change in the out-of-sample Sharpe ratio with change in U is slightly tricky, as we do not know how the out-of-sample Sharpe ratio depends on U analytically. Nevertheless, we can compute it numerically by employing the chain rule:

$$\frac{dSharpe(U)}{dU} = \nabla_{\hat{w}^*} Sharpe(\hat{w}^*(U))^\top \left[\frac{d\hat{w}^*(U)}{dU} \right],$$

where $\hat{w}^*(U)$ is the optimal PBR solution when the rhs is set to U. The first quantity, $\nabla_{\hat{w}^*} Sharpe(\hat{w}^*(U))$, can be computed explicitly, as we know the formula for the Sharpe ratio as a function of w. Suppressing the dependency of w on U, we have:

$$\nabla_w Sharpe(w) = \frac{(w^\top \Sigma w)\mu - (w^\top \mu)\Sigma w}{(w^\top \Sigma w)^{3/2}}.$$

The second quantity $dw^*(U)/dU$ is the marginal change in the optimal solution $\hat{w}^*$ as the rhs U changes. We approximate this by solving (PBR) with $(1 - bit)U$, where $0 < bit \ll 1$ is a predetermined parameter, then computing

$$\frac{d\hat{w}^*(U)}{dU} \approx \frac{\hat{w}^*(U) - \hat{w}^*((1 - bit)U)}{bit \times U},$$

Out-of-Sample Performance-Based k-Cross Validation (OOS-PBCV)

Initialize

Choose no. of bins k

Solve PBR(U) on D_{train} to get $\hat{w}_{train}$; set $\overline{U} = (\hat{w}_{train})^\top \hat{\Sigma} \hat{w}_{train}$

Solve U-min-mv1(U) [U-min-mv2(U) or U-min-cv(U)] on D_{train} to get $\hat{w}_{Umin}$; set $\underline{U} = (\hat{w}_{Umin})^\top \hat{\Sigma} \hat{w}_{Umin}$

Divide up D_{train} randomly into k equal bins, D^b_{train}, $b = 1, \ldots, k$

Let D^{-b}_{train} denote the training data minus the b-th bin

for $b \leftarrow 1$ **to** k **do**

  Solve PBR(U) on D^{-b}_{train} to get $\hat{w}_{-b}$; set $\overline{U}_{-b} = (\hat{w}_{-b})^\top \hat{\Sigma} \hat{w}_{-b}$

  if $\overline{U}_{-b} < \underline{U}$ **then** $U^*_{-b} = \underline{U}$ and terminate

    else Solve U-min-mv1(U) [U-min-mv2(U) or U-min-cv(U)] on D^{-b}_{train} to get $\hat{w}^{-b}_{Umin}$; set $\underline{U}_{-b} = (\hat{w}^{-b}_{Umin})^\top \hat{\Sigma} \hat{w}^{-b}_{Umin}$;

  end

  if $\underline{U}_{-b} > \overline{U}$ **then** $U^*_{-b} = \overline{U}$ and terminate

    else Compare and update boundaries:

    $\overline{U}_{-b} = \min(\overline{U}_{-b}, \overline{U})$

    $\underline{U}_{-b} = \max(\underline{U}_{-b}, \underline{U})$

    Run **(OOS-PBSD)** with boundaries $[\underline{U}_{-b}, \overline{U}_{-b}]$ to find U^*_{-b} ;

  end

end

Return $U^* = \frac{1}{k} \sum_{i=1}^{k} U^*_{-i}$.

Algorithm 1: A pseudo code for the out-of-sample performance-based k-cross validation algorithm (OOS-PBCV).

where $\hat{w}^*((1-bit)U)$ is the new optimal allocation when the PBR constraint rhs is set to $(1-bit)U$.

The pseudocode for this part of the CV algorithm is shown in Algorithm 2.

In our computations, we used the parameters $\alpha = 0.4, \beta = 0.9, Div = 5, bit = 0.05$ and considered $k = 2$ and $k = 3$ bins. It took on average approximately 2 seconds to solve one problem instance for all problem sizes and bin numbers considered in this paper.

5.5. Discussion of Results: mean-variance problem

Out-of-sample Sharpe ratio Table 1 reports the out-of-sample Sharpe ratios of the eight strategies listed in Sec. 5.1. For $p = 5$, the rank-1 approximation PBR performs the best, with a Sharpe ratio of 1.3551, followed by best convex quadratic approximation PBR (1.2052), then SAA (1.1573). For this data set, standard regularizations (L_1, L_2 and no short-selling) and the equally-weighted portfolio all perform below these strategies. Similarly, for $p = 10$, the rank-1 approximation PBR performs the best, with a Sharpe ratio of 1.2112, followed by best convex

Out-of-Sample Performance-Based Steepest Descent (OOS-PBSD)

Initialize

Choose backtracking parameters $\alpha \in (0, 0.5), \beta \in (0, 1)$

Choose stepsize Div

Choose perturbation size $bit \in (0, 0.5)$

for $b \leftarrow 1$ **to** k **do**

Set $U = \overline{U}_{-b}$, $\Delta U := t(\bar{U}_{-b} - \underline{U}_{-b})/Div$, $t = 1$

Compute

$$\frac{dSharpe(U)}{dU} = \nabla_w Sharpe(\hat{w}_{-b}(U))^\top \left[\frac{d\hat{w}_{-b}(U)}{dU}\right],$$

where

$$\nabla_w Sharpe(\hat{w}_{-b}(U)) = \frac{((\hat{w}_{-b})^\top \Sigma_{-b} \hat{w}_{-b})\mu_{-b} - (\hat{w}'_{-b}\mu)\Sigma_{-b}\hat{w}_{-b}}{((\hat{w}_{-b})^\top \Sigma_{-b} \hat{w}_{-b})^{3/2}}$$

$$\frac{d\hat{w}_{-b}(U)}{dU} = \frac{\hat{w}_{-b}(U) - \hat{w}_{-b}((1-bit)U)}{bit \times U}$$

while

$$Sharpe(U - t\Delta U) < Sharpe(U) + \alpha t \Delta U \frac{dSharpe(U)}{dU}$$

do

$t = \beta t$

end

end

Return $U^*_{-b} = U - t\Delta U$.

Algorithm 2: A pseudo code for the out-of-sample performance-based steepest descent algorithm **(OOS-PBSD)**, which is a subroutine of (OOS-PBCV).

quadratic approximation PBR (1.1696), then SAA (1.1357); the other strategies again relatively under perform.

The $p = 41$ data set yields results that are quite different from those of $p = 5$ and $p = 10$, evidencing that dimensionality (i.e., the number of assets) is a significant factor in its own right (this has been observed in other studies, e.g., Jagannathan and Ma (2003) and El Karoui (2010), Karoui (2013).). While we could rank the strategies by their average out-of-sample performances, they are statistically indistinguishable at the 5% level from the SAA method (all p-values are quite large, the smallest being 0.3178). Hence we cannot make any meaningful conclusions for this data set, and we leave the study of regularizing for dimensionality to future work.

From the perspective of an investor looking at the results of Table 2, the take-away is clear: focus on a small number of assets (the Fama-French 5 Industry portfolio) and optimize using the PBR method on both the objective and mean constraints to achieve the highest Sharpe ratio.

Portfolio turnover Table 3 reports the out-of-sample Sharpe ratios of the eight strategies listed in Sec. 5.1. For obvious reasons, the equally-weighted portfolio achieves the smallest portfolio turnover. For all three data sets, we find that the two PBR approximations generally have greater portfolio turnovers than SAA, whereas the standard regularization methods (L_1, L_2 and no short-selling) have lower turnovers than SAA. This is reflective of the fact that standard regularization is by design a *solution* stabilizer, whereas PBR is not.

5.6. Discussion of Results: mean-CVaR problem

Out-of-sample Sharpe ratio Table 2 reports the out-of-sample Sharpe ratios of the eight strategies listed in Sec. 5.2. For $p = 5$ and $p = 10$ data sets, we find that PBR on both the objective and the constraint dominate the SAA solution. For example, the best Sharpe ratio for $p = 5$ for the SAA method is achieved by setting a return target of $R = 0.08$, yielding a Sharpe ratio of 1.2487, whereas the best PBR result for the same data set and target return has a Sharpe ratio of 1.2715, the difference of which is statistically significant at the 5% level (the exact p-value is 0.0453). Likewise, for $p = 10$, the best SAA Sharpe ratio of 1.0346 is dominated by the best PBR Sharpe ratio of 1.1506. This difference is statistically significant at the 10% level (the exact p-value is 0.0607). Also for $p = 5$ and $p = 10$ data sets, the PBR method consistently dominates both L_1 and L_2 regularizations across all problem target returns and choice of the number of bins used for cross validation. In addition, both the equally-weighted portfolio and the global minimum CVaR portfolios underperform SAA, hence also PBR on these data sets.

The $p = 41$ data set yields results that are quite different from those of $p = 5$ and $p = 10$, signaling that dimensionality is an important parameter in its own right. First of all, the highest Sharpe ratio of all strategies across all target return levels and choice of bins is achieved by the equally-weighted portfolio, with 0.6297. Secondly, all regularizations — PBR, L_1 and L_2 — yield results that are statistically indistinguishable from the SAA method (all p-values are quite large, the smallest being 0.6249). Hence we cannot make any meaningful conclusions for this data set, and we leave the study of regularizing for dimensionality to future work.

Lastly, let us comment on the effects of PBR on the objective and the mean estimations separately. The question that comes to mind is whether one constraint dominates the other; i.e., whether PBR on the objective only consistently dominates PBR on the mean, or vice versa. The answer is a yes, but the exact relationship depends on the data set: for $p = 5$ and $p = 10$, the Sharpe ratios of PBR on CVaR is better than that of PBR on the mean for each target return (and taking

the best of the two bin results), whereas for $p = 41$, the opposite is true. This pattern seems to indicate that for a smaller number of assets, CVaR estimation is more of an issue whereas mean estimation is more problematic for a larger number of assets.

Portfolio turnover Table 3 reports the out-of-sample Sharpe ratios of the eight strategies listed in Sec. 5.2. For obvious reasons, the equally-weighted portfolio achieves the smallest portfolio turnover. For the $p = 5$ data set, the PBR method is consistently lower than SAA, L_1 and L_2 regularization methods for each target return level and across the two bins sizes considered. The opposite is true for $p = 10$ or $p = 41$ however, with PBR having consistently higher turnovers than SAA, L_1 and L_2 regularization methods for each target return level and across the two bins sizes considered. Global minimum variance portfolios have turnovers greater than the equally-weighted portfolio but generally less than the SAA method.

6. Conclusion

We introduced performance-based regularization and performance-based cross-validation for the portfolio optimization problem and investigated them in detail. The PBR models constrain sample variances of estimated quantities in the problem, namely the portfolio risk and return. The PBR models are shown to have equivalent robust counterparts, with new, non-trivial robust constraints for the portfolio risk. We have shown that PBR with performance-based cross-validation is highly effective at improving the finite-sample performance of the data-driven portfolio decision compared to SAA as well as other benchmarks known benchmarks in the literature. We conclude that PBR is a promising modeling paradigm for handling uncertainty, and worthy of further study to generalize to other decision problems.

Table 1 Sharpe Ratios for empirical data for the mean-variance problem.

	FF 5 Industry p=5		FF 10 Industry p=10		FF 49 Industry p=41 (-8 assets with missing data)	
Mean-Variance R=0.04						
SAA	1.1459		1.1332		0.4744	
	2 bins	3 bins	2 bins	3 bins	2 bins	3 bins
PBR (rank-1)	1.2603	1.3254	1.1868	1.2098	0.4344	**0.4712**
	(0.0411)	(0.0286)	(0.0643)	(0.0509)	(0.5848)	(0.5386)
PBR (PSD)	1.1836	1.1831	1.1543	1.1678	0.4776	0.4825
	(0.0743)	(0.071)	(0.0891)	(0.0816)	(0.5593)	(0.5391)
NS	**1.0023**		0.9968		0.7345	
	(0.1404)		(0.1437)		(0.2977)	
L1	1.0136	1.0386	1.1185	1.1175	0.5419	0.5211
	(0.1568)	(0.1396)	(0.1008)	(0.1017)	(0.5044)	(0.5216)
L2	0.9711	1.0268	1.0579	1.0699	0.6672	0.6009
	(0.1781)	(0.1452)	(0.1482)	(0.1280)	(0.3950)	(0.4455)
Mean-Variance R=0.06						
SAA	1.1535		**1.1357**		0.4468	
	2 bins	3 bins	2 bins	3 bins	2 bins	3 bins
PBR (rank-1)	1.2945	1.3362	1.1870	**1.2112**	0.4011	0.4515
	(0.0297)	(0.0244)	(0.0629)	(0.0503)	(0.6136)	(0.5530)
PBR (PSD)	1.1912	**1.2052**	1.1532	**1.1696**	0.4585	0.4587
	(0.0689)	(0.0638)	(0.0898)	(0.0809)	(0.5757)	(0.5598)
NS	0.9853		0.9699		0.7124	
	(0.1422)		(0.1537)		(0.3247)	
L1	0.9963	1.0198	1.0902	1.1010	0.4991	0.4941
	(0.1535)	(0.1394)	(0.1124)	(0.1101)	(0.5490)	(0.5448)
L2	0.9713	1.0265	1.0642	1.0755	0.6313	0.5701
	(0.1735)	(0.1425)	(0.1425)	(0.1238)	(0.4250)	(0.4696)
Markowitz R=0.08						
SAA	**1.1573**		1.1225		0.4253	
	2 bins	3 bins	2 bins	3 bins	2 bins	3 bins
PBR (rank-1)	1.3286	**1.3551**	1.1743	1.2018	0.3927	0.4253
	(0.0223)	(0.0208)	(0.0668)	(0.0510)	(0.6142)	(0.5778)
PBR (PSD)	1.1813	1.1952	1.1467	1.1575	0.4477	0.4366
	(0.0648)	(0.0614)	(0.0893)	(0.0844)	(0.5852)	(0.5804)
NS	0.9664		0.9405		0.6600	
	(0.1514)		(0.1577)		(0.3790)	
L1	0.9225	0.9965	1.0318	1.0779	0.4770	0.4930
	(0.1857)	(0.1403)	(0.1332)	(0.1181)	(0.5649)	(0.5379)
L2	0.9703	**1.0284**	1.0671	**1.0776**	0.6098	0.5522
	(0.1649)	(0.1398)	(0.1398)	(0.1209)	(0.4369)	(0.4785)
Min. Variance						
SAA	1.1454		1.1331		**0.4816**	
	2 bins	3 bins	2 bins	3 bins	2 bins	3 bins
PBR (rank-1)	1.2580	1.3269	1.1922	1.2086	0.4409	0.4683
	(0.0420)	(0.0288)	(0.0603)	(0.0505)	(0.5795)	(0.5472)
PBR (PSD)	1.1883	1.1882	1.154	1.1657	**0.4942**	0.4903
	(0.0710)	(0.0693)	(0.0892)	(0.0823)	(0.5400)	(0.5322)
NS	1.0022		**1.0012**		**0.7347**	
	(0.1405)		(0.1447)		(0.3178)	
L1	1.0321	**1.0546**	**1.1199**	1.1111	**0.5424**	0.5260
	(0.1455)	(0.1286)	(0.1000)	(0.1026)	(0.5017)	(0.5151)
L2	0.9945	1.0140	1.0543	1.0760	**0.6886**	0.6204
	(0.1632)	(0.1472)	(0.1488)	(0.1236)	(0.3761)	(0.4276)
Equal	**0.6617**		**0.7019**		**0.6297**	

This table reports the annualized out-of-sample Sharpe ratios of solutions to the mean-variance problem solved with the methods described in Sec. 5.1 for three different data sets for target returns $R = 0.04, 0.06, 0.08$. For each data set, the highest Sharpe ratio attained by each strategy is highlighted in boldface. To set the degree of regularization, we use the performance-based k-fold cross validation algorithm detailed in Sec. 5.4, with $k = 2$ and 3 bins. In parentheses we report the p-values of tests of differences from the SAA method. We also report the Sharpe ratio of the equally-weighted portfolio.

Table 2 Sharpe Ratios for empirical data for the mean-CVaR problem.

	FF 5 Industry p=5		FF 10 Industry p=10		FF 49 Industry p=41 (-8 assets with missing data)	
Mean-CVaR R=0.04						
SAA	1.2137		1.0321		**0.3657**	
	2 bins	3 bins	2 bins	3 bins	2 bins	3 bins
PBR (CVaR only)	1.2113	1.1733	1.0506	1.1381	0.1304	0.1304
	(0.0554)	(0.0674)	(0.0638)	(0.0312)	(0.7908)	(0.7908)
PBR (mean only)	1.2089	1.1802	1.0994	1.0519	0.2732	0.3682
	(0.0746)	(0.0790)	(0.1051)	(0.1338)	(0.7518)	(0.6454)
PBR (both)	1.2439	1.2073	1.1112	1.1422	**0.3607**	0.2247
	(0.0513)	(0.0601)	(0.0691)	(0.0648)	(0.7054)	(0.7667)
L1	1.0112	**1.0754**	0.9254	0.9741	0.4048	0.4642
	(0.1497)	(0.1366)	(0.2293)	(0.1880)	(0.6874)	(0.6242)
L2	0.9650	1.0636	1.0031	0.9835	**0.3982**	0.3586
	(0.1780)	(0.1287)	(0.1512)	(0.1598)	(0.7087)	(0.6878)
Mean-CVaR R=0.06						
SAA	1.2179		1.0321		**0.3657**	
	2 bins	3 bins	2 bins	3 bins	2 bins	3 bins
PBR (CVaR only)	1.2223	1.2063	1.0518	1.1451	0.1265	0.1300
	(0.0503)	(0.0527)	(0.0633)	(0.0294)	(0.7920)	(0.7909)
PBR (mean only)	1.2205	1.1902	1.0988	1.0466	0.2704	0.3771
	(0.0699)	(0.0746)	(0.1053)	(0.1358)	(0.7531)	(0.6359)
PBR (both)	1.2450	1.2043	1.1122	**1.1506**	0.3503	0.2267
	(0.0504)	(0.0581)	(0.0686)	(0.0607)	(0.7102)	(0.7656)
L1	0.9404	1.0464	0.9276	0.9746	0.3888	0.4635
	(0.1812)	(0.1395)	(0.2282)	(0.1887)	(0.7001)	(0.6249)
L2	0.9271	1.0627	1.0146	0.9794	0.3842	0.3571
	(0.1977)	(0.1286)	(0.1432)	(0.1621)	(0.7175)	(0.6886)
Mean-CVaR R=0.08						
SAA	**1.2487**		**1.0346**		**0.3657**	
	2 bins	3 bins	2 bins	3 bins	2 bins	3 bins
PBR (CVaR only)	1.2493	1.2098	1.0551	1.1433	0.1304	0.1304
	(0.0434)	(0.0462)	(0.0579)	(0.0323)	(0.7908)	(0.7908)
PBR (mean only)	1.2480	1.2088	1.0987	1.0470	0.2675	0.3738
	(0.0591)	(0.0693)	(0.1053)	(0.1384)	(0.7541)	(0.6391)
PBR (both)	**1.2715**	1.2198	1.1122	1.1449	0.2656	0.2285
	(0.0453)	(0.0544)	(0.0664)	(0.0639)	(0.7618)	(0.7647)
L1	0.8921	0.9836	0.9416	**1.0087**	0.3855	**0.4872**
	(0.1964)	(0.1572)	(0.2122)	(0.1645)	(0.7008)	(0.6128)
L2	0.9367	**1.0801**	**1.0278**	0.9947	0.3784	0.3588
	(0.1989)	(0.1179)	(0.1323)	(0.1530)	(0.7177)	(0.6870)
Global min. CVaR	**1.2137**		**1.0321**		**0.3657**	
Equal	**0.6617**		**0.7019**		**0.6297**	

This table reports the annualized out-of-sample Sharpe ratios of the solutions to the mean-CVaR problem solved with SAA, PBR with regularization of the objective ("CVaR only"), the constraint ("mean only") and both the objective and the constraint ("both"), L1 and L2 regularization constraints for three different data sets and for target returns $R = 0.04, 0.06, 0.08$. For each data set, the highest Sharpe ratio attained by each strategy is highlighted in boldface. To set the degree of regularization, we use the performance-based k-fold cross validation algorithm detailed in Sec. 5.4, with $k = 2$ and 3 bins. In parentheses we report the p-values of tests of differences from the SAA method. We also report the Sharpe ratio of the equally-weighted portfolio and the solution to the global minimum CVaR problem (no mean constraint).

Table 3 Turnovers for empirical data for the mean-variance problem.

	FF 5 Industry p=5		FF 10 Industry p=10		FF 49 Industry p=41 (-8 assets with missing data)	
Mean-Variance R=0.04						
SAA	0.0935		0.1325		0.5188	
	2 bins	3 bins	2 bins	3 bins	2 bins	3 bins
PBR (rank-1)	0.1213	0.1292	0.1746	0.1851	0.5442	0.6611
PBR (PSD)	0.1002	0.0988	0.1415	0.1523	0.5201	0.4999
NS	0.0391		0.0544		0.0833	
L1	0.0986	0.0848	0.1158	0.1208	0.5167	0.4453
L2	0.1171	0.0901	0.1255	0.1071	0.4704	0.4079
Mean-Variance R=0.06						
SAA	0.1034		0.1339		0.5289	
	2 bins	3 bins	2 bins	3 bins	2 bins	3 bins
PBR (rank-1)	0.1397	0.1357	0.1741	0.1841	0.5646	0.6427
PBR (PSD)	0.1132	0.1086	0.1442	0.1513	0.5301	0.5042
NS	0.0417		0.0711		0.0859	
L1	0.1206	0.0963	0.1256	0.1205	0.4992	0.4439
L2	0.1267	0.0992	0.1379	0.1121	0.4809	0.4110
Mean-Variance R=0.08						
SAA	0.1288		0.1475		0.5434	
	2 bins	3 bins	2 bins	3 bins	2 bins	3 bins
PBR (rank-1)	0.1775	0.1504	0.1894	0.1959	0.5721	0.5434
PBR (PSD)	0.1147	0.1344	0.1689	0.1547	0.5414	0.5204
NS	0.0511		0.0965		0.1122	
L1	0.1476	0.1246	0.1480	0.1392	0.5064	0.4567
L2	0.1582	0.1241	0.1470	0.1229	0.5118	0.4200
Min. Variance						
SAA	0.1034		0.1325		0.5146	
	2 bins	3 bins	2 bins	3 bins	2 bins	3 bins
PBR (rank-1)	0.1245	0.1311	0.1756	0.1807	0.5393	0.6065
PBR (PSD)	0.1221	0.1182	0.1609	0.1682	0.5138	0.5022
NS	0.0391		0.0524		0.0835	
L1	0.0995	0.0886	0.1138	0.1219	0.4956	0.4435
L2	0.1213	0.0910	0.1255	0.1061	0.4575	0.4070
Equal	0.0427		0.0382		0.0483	

This table reports the portfolio turnovers (defined in Eq. 15) of the solutions to the mean-variance problem solved with the methods described in Sec. 5.1 for three different data sets and for target returns $R = 0.04, 0.06, 0.08$. We also report the turnovers of the equally-weighted portfolio.

Table 4 Turnovers for empirical data for the mean-CVaR problem.

Sharpe Ratios	FF 5 Industry p=5		FF 10 Industry p=10		FF 49 Industry p=41 (-8 assets with missing data)	
Mean-CVaR R=0.04						
SAA	0.2857		0.3534		1.6833	
	2 bins	3 bins	2 bins	3 bins	2 bins	3 bins
PBR (CVaR only)	0.1834	0.1985	0.4049	0.5586	1.7773	1.7773
PBR (mean only)	0.1230	0.1274	0.3104	0.2731	1.3173	1.5023
PBR (both)	0.1387	0.1388	0.3700	0.3682	1.7492	1.4158
L1	0.1992	0.1581	0.3415	0.2722	1.5158	1.3731
L2	0.1565	0.1469	0.2288	0.2270	1.2192	1.1217
Mean-CVaR R=0.06						
SAA	0.2909		0.3534		1.6833	
	2 bins	3 bins	2 bins	3 bins	2 bins	3 bins
PBR (CVaR only)	0.1918	0.2074	0.4071	0.5616	1.7922	1.7778
PBR (mean only)	0.1364	0.1414	0.3100	0.2729	1.3188	1.5147
PBR (both)	0.1519	0.1526	0.3724	0.3672	1.7610	1.4170
L1	0.2263	0.1754	0.3498	0.2723	1.5232	1.3730
L2	0.1842	0.1532	0.2407	0.2401	1.2374	1.1220
Mean-CVaR R=0.08						
SAA	0.2980		0.3615		1.6833	
	2 bins	3 bins	2 bins	3 bins	2 bins	3 bins
PBR (CVaR only)	0.2148	0.2242	0.4486	0.6472	1.7775	1.7775
PBR (mean only)	0.1517	0.1575	0.3066	0.2827	1.3228	1.5038
PBR (both)	0.1693	0.1681	0.4099	0.4034	1.6887	1.4190
L1	0.3100	0.2395	0.3628	0.3042	1.5370	1.3731
L2	0.2451	0.1835	0.2588	0.2633	1.1774	1.1227
Global min. CVaR	0.2857		0.3534		1.6833	
Equal	0.0427		0.0382		0.0483	

This table reports the portfolio turnovers (defined in Eq. 15) of the solutions to the mean-CVaR problem solved with SAA, PBR with regularization of the objective ("CVaR only"), the constraint ("mean only") and both the objective and the constraint ("both"), L1 and L2 regularization constraints for three different data sets and for target returns $R = 0.04, 0.06, 0.08$. We also report the turnovers of the equally-weighted portfolio and the solution to the global minimum CVaR problem (no mean constraint).

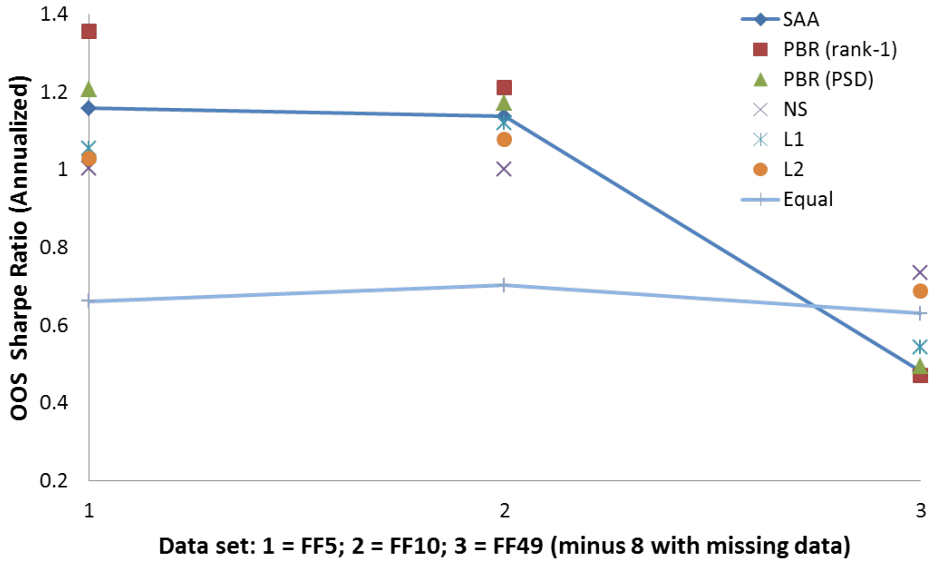

Figure 4 The out-of-sample Sharpe ratios (annualized) for the strategies considered for the mean-variance problem, for three data sets. Detailed results are in Table 1.

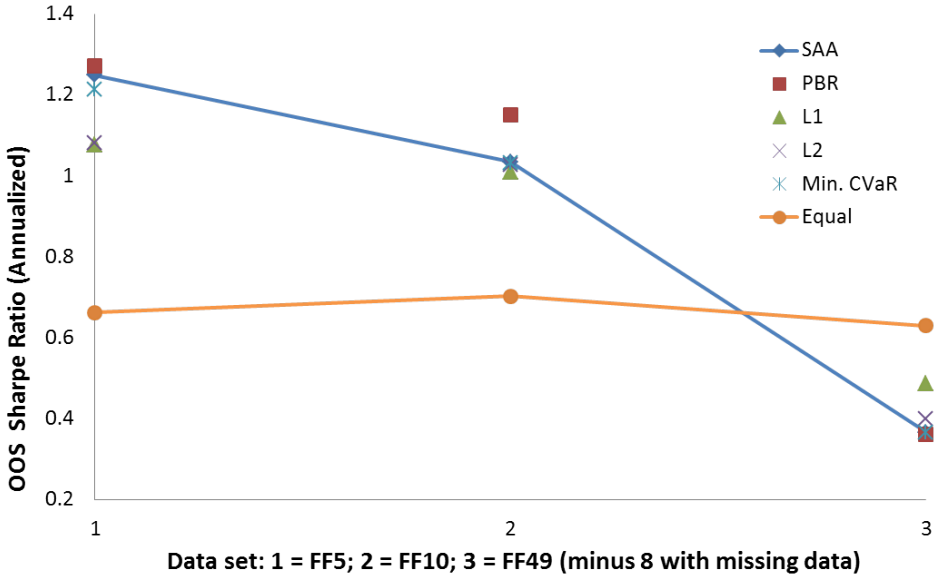

Figure 5 The out-of-sample Sharpe ratios (annualized) for the strategies considered for the mean-CVaR problem, for three data sets. Detailed results are in Table 2.

Online Appendix: Machine Learning & Portfolio Optimization

Gah-Yi Ban, Noureddine El Karoui, Andrew E.B. Lim

Appendix A: Proofs of results in Sec. 3

A.1. Proof of Proposition 1

Let $\mathbf{X}^1,\ldots,\mathbf{X}^n$ be n iid observations of the vector $\mathbf{X}$. We wish to compute the covariance of the sample covariance matrix $\mathbf{S}$, whose elements are:

$$\mathbf{S}^2_{ij} = \left[\frac{1}{n}\sum_{k=1}^{n}(X_i^k - \bar{X}_i)(X_j^k - \bar{X}_j)\right], \tag{16}$$

where $\bar{X}_i = \frac{1}{n}\sum_{k=1}^n X_i^k$ is the sample mean of the i-th element of $\mathbf{X}$.

Transformation into a U-statistic For convenience, we transform (16) into a U-statistic:

$$\mathbf{S}^2_{ij} = \frac{1}{\binom{n}{2}}\sum_{(k,l)}\frac{1}{2}(X_i^k - X_i^l)(X_j^k - X_j^l)$$

Then the variance of the ij-th element of the sample covariance matrix is given by

$$\begin{aligned}
Cov(\mathbf{S}^2_{ij},\mathbf{S}^2_{kl}) &= \mathbb{E}\left[\frac{1}{\binom{n}{2}}\sum_{(p,q)}\frac{1}{2}(X_i^p - X_i^q)(X_j^p - X_j^q) - \sigma^2_{ij}\right]\left[\frac{1}{\binom{n}{2}}\sum_{(r,s)}\frac{1}{2}(X_k^r - X_k^s)(X_l^r - X_l^s) - \sigma^2_{kl}\right] \\
&= \frac{1}{\binom{n}{2}^2}\mathbb{E}\left\{\sum_{(p,q)}\sum_{(r,s)}[\frac{1}{2}(X_i^p - X_i^q)(X_j^p - X_j^q) - \sigma^2_{ij}][\frac{1}{2}(X_k^r - X_k^s)(X_l^r - X_l^s) - \sigma^2_{kl}]\right\}.
\end{aligned}$$

There are three cases:

1. $|(k,l)\cap(k',l')| = 0$ then by independence the product is zero.
2. $|(k,l)\cap(k',l')| = 1$, in other words, we have $n(n-1)(n-2)$ terms of the form

$$[\frac{1}{2}(X_i^p - X_i^q)(X_j^p - X_j^q) - \sigma^2_{ij}][\frac{1}{2}(X_k^p - X_k^s)(X_l^p - X_l^s) - \sigma^2_{kl}]$$

Now

$$\begin{aligned}
&[\frac{1}{2}(X_i^p - X_i^q)(X_j^p - X_j^q) - \sigma^2_{ij}][\frac{1}{2}(X_k^p - X_k^s)(X_l^p - X_l^s) - \sigma^2_{kl}] \\
&= \frac{1}{4}[X_i^pX_j^p - X_i^pX_j^q - X_i^qX_j^p][X_k^pX_l^p - X_k^pX_l^s - X_k^sX_l^p] \\
&+ \frac{1}{4}X_i^qX_j^q[X_k^pX_l^p - X_k^pX_l^s - X_k^sX_l^p] + \frac{1}{4}X_k^sX_l^s[X_i^pX_j^p - X_i^pX_j^q - X_i^qX_j^p] \\
&+ \frac{1}{4}X_i^qX_j^qX_k^sX_l^s \\
&- \frac{1}{2}\sigma^2_{kl}[X_i^pX_j^p - X_i^pX_j^q - X_i^qX_j^p + X_i^qX_j^q] - \frac{1}{2}\sigma^2_{ij}[X_k^pX_l^p - X_k^pX_l^s - X_k^sX_l^p + X_k^sX_l^s] + \sigma^2_{ij}\sigma^2_{kl},
\end{aligned}$$

so taking expectations and simplifying, we get

$$\begin{aligned}&\mathbb{E}[\frac{1}{2}(X_i^p - X_i^q)(X_j^p - X_j^q) - \sigma_{ij}^2][\frac{1}{2}(X_k^p - X_k^s)(X_l^p - X_l^s) - \sigma_{kl}^2]\\&= \frac{1}{4}(\mu_{4,ijkl} - \sigma_{ij}^2\sigma_{kl}^2).\end{aligned}$$

3. $|(k,l) \cap (k',l')| = 2$, in other words, we have $\binom{n}{2}$ terms of the form

$$[\frac{1}{2}(X_i^p - X_i^q)(X_j^p - X_j^q) - \sigma_{ij}^2][\frac{1}{2}(X_k^p - X_k^q)(X_l^p - X_l^q) - \sigma_{kl}^2]$$

Now

$$\begin{aligned}&[\frac{1}{2}(X_i^p - X_i^q)(X_j^p - X_j^q) - \sigma_{ij}^2][\frac{1}{2}(X_k^p - X_k^q)(X_l^p - X_l^q) - \sigma_{kl}^2]\\&= \frac{1}{4}[X_i^pX_j^pX_k^pX_l^p + X_i^qX_j^qX_k^qX_l^q]\\&- \frac{1}{4}[X_i^pX_j^pX_k^pX_l^q + X_i^pX_j^pX_k^qX_l^p + X_i^pX_j^qX_k^pX_l^p + X_i^qX_j^pX_k^pX_l^p]\\&- \frac{1}{4}[X_i^pX_j^qX_k^qX_l^q + X_i^qX_j^pX_k^qX_l^q + X_i^qX_j^qX_k^pX_l^q + X_i^qX_j^qX_k^qX_l^p]\\&+ \frac{1}{4}[X_i^pX_j^pX_k^qX_l^q + X_i^pX_j^qX_k^pX_l^q + X_i^pX_j^qX_k^qX_l^p + X_i^qX_j^pX_k^pX_l^q + X_i^qX_j^pX_k^qX_l^p + X_i^qX_j^qX_k^pX_l^p]\\&- \frac{1}{2}\sigma_{kl}^2[X_i^pX_j^p - X_i^pX_j^q - X_i^qX_j^p + X_i^qX_j^q] - \frac{1}{2}\sigma_{ij}^2[X_k^pX_l^p - X_k^pX_l^q - X_k^qX_l^p + X_k^qX_l^q] + \sigma_{ij}^2\sigma_{kl}^2,\end{aligned}$$

so taking expectations and simplifying, we get

$$\begin{aligned}&\mathbb{E}[\frac{1}{2}(X_i^p - X_i^q)(X_j^p - X_j^q) - \sigma_{ij}^2][\frac{1}{2}(X_k^p - X_k^q)(X_l^p - X_l^q) - \sigma_{kl}^2]\\&= \frac{1}{2}(\mu_{4,ijkl} - \sigma_{ij}^2\sigma_{kl}^2 + \sigma_{ik}^2\sigma_{jl}^2 + \sigma_{il}^2\sigma_{jk}^2).\end{aligned}$$

Putting it all together,

$$Cov(\mathbf{S}_{ij}^2, \mathbf{S}_{kl}^2) = \frac{1}{n}(\mu_{4,ijkl} - \sigma_{ij}^2\sigma_{kl}^2) + \frac{1}{n(n-1)}(\sigma_{ik}^2\sigma_{jl}^2 + \sigma_{il}^2\sigma_{jk}^2). \tag{17}$$

A.2. Proof of Proposition 2

Let us start with (mv-PBR-1) with the mean constraint. The Lagrangian is:

$$\begin{aligned}\mathcal{L}(w;\nu_1,\nu_2,\lambda) &= w^\top\hat{\Sigma}_n w + \nu_1(w^\top 1_p - 1) + \nu_2(w^\top\hat{\mu}_n - R) + \lambda(w^\top\hat{\alpha} - \sqrt[4]{U})\\&= w^\top\hat{\Sigma}_n w + w^\top(\nu_1 1_p + \nu_2\hat{\mu}_n + \lambda\hat{\alpha}) - \nu_1 - \nu_2 R - \lambda\sqrt[4]{U}\end{aligned}$$

The first order condition (FOC) gives:

$$w^* = -\frac{1}{2}\hat{\Sigma}_n^{-1}(\nu_1 1_p + \nu_2\hat{\mu}_n + \lambda\hat{\alpha}) \tag{18}$$

The Lagrangian dual function is:

$$\begin{aligned} g(\nu_1,\nu_2,\lambda) &= \inf_w \mathcal{L}(w;\nu_1,\nu_2,\lambda) \\ &= -\frac{1}{4}(\nu_1 1_p + \nu_2\hat{\mu}_n + \lambda\hat{\alpha})^\top \hat{\Sigma}_n^{-1}(\nu_1 1_p + \nu_2\hat{\mu}_n + \lambda\hat{\alpha}) - \nu_1 - \nu_2 R - \lambda\sqrt[4]{U} \end{aligned}$$

At optimality, g is maximized over $(\nu_1,\nu_2,\lambda)\in\mathbb{R}^p\times\mathbb{R}^p\times\mathbb{R}_+$. We will maximize g over (ν_1,ν_2) first. The first order conditions give:

$$\begin{aligned} \frac{dg(\nu_1,\nu_2,\lambda)}{d\nu_1} &= -\frac{1}{2}(\nu_2\hat{\mu}_n + \lambda\hat{\alpha})^\top\hat{\Sigma}_n^{-1}1_p - \frac{1}{2}\nu_1 1_p^\top\hat{\Sigma}_n^{-1}1_p - 1 = 0 \\ \frac{dg(\nu_1,\nu_2,\lambda)}{d\nu_2} &= -\frac{1}{2}(\nu_1 1_p + \lambda\hat{\alpha})^\top\hat{\Sigma}_n^{-1}\hat{\mu}_n - \frac{1}{2}\nu_2\hat{\mu}_n^\top\hat{\Sigma}_n^{-1}\hat{\mu}_n - R = 0. \end{aligned}$$

Solving simultaneously, we have

$$\begin{aligned} \nu_1^* &= 2\frac{\hat{\mu}_n^\top\hat{\Sigma}_n^{-1}\hat{\mu}_n - R\hat{\mu}_n^\top\hat{\Sigma}_n^{-1}1_p}{1_p^\top\hat{\Sigma}_n^{-1}1_p.\hat{\mu}_n^\top\hat{\Sigma}_n^{-1}\hat{\mu}_n - (\hat{\mu}_n^\top\hat{\Sigma}_n^{-1}1_p)^2} + \lambda\frac{\hat{\alpha}^\top\hat{\Sigma}_n^{-1}\hat{\mu}_n.\hat{\mu}_n^\top\hat{\Sigma}_n^{-1}1_p - \hat{\alpha}^\top\hat{\Sigma}_n^{-1}1_p.\hat{\mu}_n^\top\hat{\Sigma}_n^{-1}\hat{\mu}_n}{1_p^\top\hat{\Sigma}_n^{-1}1_p.\hat{\mu}_n^\top\hat{\Sigma}_n^{-1}\hat{\mu}_n - (\hat{\mu}_n^\top\hat{\Sigma}_n^{-1}1_p)^2} = \nu_1^0 + \lambda\beta_1 \\ \nu_2^* &= 2\frac{1_p^\top\hat{\Sigma}_n^{-1}\hat{\mu}_n - R1_p^\top\hat{\Sigma}_n^{-1}1_p}{1_p^\top\hat{\Sigma}_n^{-1}1_p.\hat{\mu}_n^\top\hat{\Sigma}_n^{-1}\hat{\mu}_n - (\hat{\mu}_n^\top\hat{\Sigma}_n^{-1}1_p)^2} + \lambda\frac{\hat{\alpha}^\top\hat{\Sigma}_n^{-1}\hat{\mu}_n.1_p^\top\hat{\Sigma}_n^{-1}1_p - \hat{\alpha}^\top\hat{\Sigma}_n^{-1}1_p.1_p^\top\hat{\Sigma}_n^{-1}\hat{\mu}_n}{1_p^\top\hat{\Sigma}_n^{-1}1_p.\hat{\mu}_n^\top\hat{\Sigma}_n^{-1}\hat{\mu}_n - (\hat{\mu}_n^\top\hat{\Sigma}_n^{-1}1_p)^2} = \nu_2^0 + \lambda\beta_2, \end{aligned}$$

where ν_1^0,ν_2^0 are optimal dual variables for the original mean-variance problem. Hence, the optimal portfolio becomes

$$\begin{aligned} \hat{w}_{n,PBR1} &= -\frac{1}{2}\hat{\Sigma}_n^{-1}((\nu_1^0+\lambda\beta_1)1_p + (\nu_2^0+\lambda\beta_2)\hat{\mu}_n + \lambda\hat{\alpha}) \\ &= -\frac{1}{2}\hat{\Sigma}_n^{-1}(\nu_1^0 1_p + \nu_2^0\hat{\mu}_n) - \frac{1}{2}\lambda\hat{\Sigma}_n^{-1}(\beta_1 1_p + \beta_2\hat{\mu}_n + \hat{\alpha}) \\ &= \hat{w}_{n,MV} - \frac{1}{2}\lambda\hat{\Sigma}_n^{-1}(\beta_1 1_p + \beta_2\hat{\mu}_n + \hat{\alpha}), \end{aligned}$$

where $\hat{w}_{n,MV}$ is the optimal portfolio of the SAA mean-variance problem without PBR.

For the problem without the mean constraint, we follow similar steps. The Lagrangian is:

$$\begin{aligned} \mathcal{L}(w;\nu,\lambda) &= w^\top\hat{\Sigma}_n w + \nu(w^\top 1_p - 1) + \lambda(w^\top\hat{\alpha} - \sqrt[4]{U}) \\ &= w^\top\hat{\Sigma}_n w + w^\top(\nu 1_p + \lambda\hat{\alpha}) - \nu - \lambda\sqrt[4]{U} \end{aligned}$$

The FOC gives:

$$w^* = -\frac{1}{2}\hat{\Sigma}_n^{-1}(\nu 1_p + \lambda\hat{\alpha}) \tag{19}$$

The Lagrangian dual function is:

$$\begin{aligned} g(\nu,\lambda) &= \inf_w \mathcal{L}(w;\nu,\lambda) \\ &= -\frac{1}{4}(\nu 1_p + \lambda\hat{\alpha})^\top\hat{\Sigma}_n^{-1}(\nu 1_p + \lambda\hat{\alpha}) - \nu - \lambda\sqrt[4]{U}, \end{aligned}$$

and

$$\frac{dg(\nu,\lambda)}{d\nu} = -\frac{1}{2}\lambda\hat{\alpha}^\top\hat{\Sigma}_n^{-1}1_p - \frac{1}{2}\nu 1_p^\top\hat{\Sigma}_n^{-1}1_p - 1 = 0$$

Solving for ν, we get

$$\nu^* = -\frac{2}{1_p^\top\hat{\Sigma}_n^{-1}1_p} - \lambda\frac{1_p^\top\hat{\Sigma}_n^{-1}\hat{\alpha}}{1_p^\top\hat{\Sigma}_n^{-1}1_p} = \nu^0 + \lambda\beta$$

where ν^0 is the optimal dual variables for the SAA mean-variance problem without the mean constraint. Hence, the optimal portfolio becomes

$$\begin{aligned}\hat{w}_{n,PBR1} &= -\frac{1}{2}\hat{\Sigma}_n^{-1}((\nu^0+\lambda\beta)1_p + \lambda\hat{\alpha}) \\ &= -\frac{1}{2}\nu^0\hat{\Sigma}_n^{-1}1_p - \frac{1}{2}\lambda\hat{\Sigma}_n^{-1}(\beta 1_p + \hat{\alpha}) \\ &= \hat{w}_{n,MV} - \frac{1}{2}\lambda\hat{\Sigma}_n^{-1}(\beta 1_p + \hat{\alpha})\end{aligned}$$

where $\hat{w}_{n,MV}$ is the optimal portfolio of the SAA mean-variance problem without PBR. □

A.3. Proof of Proposition 3

Let us start with (mv-PBR-2) with the mean constraint. The Lagrangian is:

$$\begin{aligned}\mathcal{L}(w;\nu_1,\nu_2,\lambda) &= w^\top\hat{\Sigma}_n w + \nu_1(w^\top 1_p - 1) + \nu_2(w^\top\hat{\mu}_n - R) + \lambda(w^\top A^* w - \sqrt{U}) \\ &= w^\top(\hat{\Sigma}_n + \lambda A^*)w + w^\top(\nu_1 1_p + \nu_2\hat{\mu}_n) - \nu_1 - \nu_2 R - \lambda\sqrt{U}\end{aligned}$$

The FOC gives:

$$w^* = -\frac{1}{2}(\hat{\Sigma}_n + \lambda A^*)^{-1}(\nu_1 1_p + \nu_2\hat{\mu}_n) \tag{20}$$

The Lagrangian dual function is:

$$\begin{aligned}g(\nu_1,\nu_2,\lambda) &= \inf_w \mathcal{L}(w;\nu_1,\nu_2,\lambda) \\ &= -\frac{1}{4}(\nu_1 1_p + \nu_2\hat{\mu}_n)^\top(\hat{\Sigma}_n + \lambda A^*)^{-1}(\nu_1 1_p + \nu_2\hat{\mu}_n) - \nu_1 - \nu_2 R - \lambda\sqrt{U}\end{aligned}$$

At optimality, g is maximized over $(\nu_1,\nu_2,\lambda) \in \mathbb{R}^p \times \mathbb{R}^p \times \mathbb{R}_+$. We will maximize g over (ν_1,ν_2) first. The first order conditions give:

$$\begin{aligned}\frac{dg(\nu_1,\nu_2,\lambda)}{d\nu_1} &= -\frac{1}{2}\nu_2\hat{\mu}_n^\top(\hat{\Sigma}_n + \lambda A^*)^{-1}1_p - \frac{1}{2}\nu_1 1_p^\top(\hat{\Sigma}_n + \lambda A^*)^{-1}1_p - 1 = 0 \\ \frac{dg(\nu_1,\nu_2,\lambda)}{d\nu_2} &= -\frac{1}{2}\nu_1 1_p^\top(\hat{\Sigma}_n + \lambda A^*)^{-1}\hat{\mu}_n - \frac{1}{2}\nu_2\hat{\mu}_n^\top(\hat{\Sigma}_n + \lambda A^*)^{-1}\hat{\mu}_n - R = 0.\end{aligned}$$

Solving simultaneously, we have

$$\nu_1^* = 2\frac{R\hat{\mu}_n^\top\tilde{\Sigma}^{-1}1_p - \hat{\mu}_n^\top\tilde{\Sigma}^{-1}\hat{\mu}_n}{1_p^\top\tilde{\Sigma}^{-1}1_p.\hat{\mu}_n^\top\tilde{\Sigma}^{-1}\hat{\mu}_n - (\hat{\mu}_n^\top\tilde{\Sigma}^{-1}1_p)^2},$$
$$\nu_2^* = 2\frac{-R1_p^\top\tilde{\Sigma}^{-1}1_p + \hat{\mu}_n^\top\tilde{\Sigma}^{-1}1_p}{1_p^\top\tilde{\Sigma}^{-1}1_p.\hat{\mu}_n^\top\tilde{\Sigma}^{-1}\hat{\mu}_n - (\hat{\mu}_n^\top\tilde{\Sigma}^{-1}1_p)^2},$$

where $\tilde{\Sigma} = \hat{\Sigma}_n + \lambda^* A^*$. Hence, the optimal portfolio becomes

$$\hat{w}_{n,PBR2} = -\frac{1}{2}(\hat{\Sigma}_n + \lambda^* A^*)^{-1}(\nu_1^* 1_p + \nu_2^*\hat{\mu}_n),$$

which equals $\hat{w}_{n,MV}$ when $\lambda^* = 0$.

For the problem without the mean constraint, we follow similar steps to arrive at the result. □

A.4. Proof of Proposition 4

Setting. Let $\mathsf{L} = [L_1, \ldots, L_n]$ be n iid observations (of portfolio losses) from a distribution F which is absolutely continuous, has a twice continuously differentiable pdf and a finite second moment.

Let us derive an expression for the variance of $\widehat{CVaR}_n(L;\beta)$ introduced in Eq. (2) of Sec. 2.1. First, let us define a closely related estimator:

Definition 1 (**Type 1 CVaR estimator**). For $\beta \in (0.5, 1)$, we define Type 1 CVaR estimator to be

$$\widehat{CV1}_n(\mathsf{L};\beta) := \min_{\alpha\in\mathbb{R}} \ (1-\varepsilon_n)\alpha + \frac{1}{n - \lceil n\beta\rceil + 1}\sum_{i=1}^{n}(L_i - \alpha)^+,$$

where ε_n is some constant satisfying $0 < \varepsilon_n < (n - \lceil n\beta\rceil + 1)^{-1}$ and $\varepsilon_n = O(n^{-2})$.
We now show the minimizer in the definition of $\widehat{CV1}_n(\mathsf{L};\beta)$ is given by $\alpha^* = L_{(\lceil n\beta\rceil)}$.

Lemma 1. *The solution $\alpha^* = L_{(\lceil n\beta\rceil)}$ is the unique minimizer in the one-dimensional optimization problem*

$$\min_{\alpha\in\mathbb{R}} \ \left\{ G_n(\alpha) := (1-\varepsilon_n)\alpha + \frac{1}{n - \lceil n\beta\rceil + 1}\sum_{i=1}^{n}(L_i - \alpha)^+ \right\},$$

where ε_n is some constant satisfying $0 < \varepsilon_n < (n - \lceil n\beta\rceil + 1)^{-1}$ and $\varepsilon_n = O(n^{-2})$.

Proof. The expression to be minimized is a piecewise linear convex function with nodes at $L_1, \ldots, L_n$. We show that $G_n(\alpha)$ has gradients of opposite signs about a single point, $L_{(\lceil n\beta\rceil)}$, hence this point must be the unique optimal solution. Now consider, for $m \in \{-\lceil n\beta\rceil + 1, \ldots, n - \lceil n\beta\rceil\}$:

$$\begin{aligned}\Delta(m) &= G_n(L_{(\lceil n\beta\rceil + m + 1)}) - G_n(L_{(\lceil n\beta\rceil + m)}) \\ &= (1-\varepsilon_n)(L_{(\lceil n\beta\rceil + m + 1)} - L_{(\lceil n\beta\rceil + m)}) - \frac{1}{n - \lceil n\beta\rceil + 1}A,\end{aligned}$$

where

$$\begin{aligned}A &= \sum_{i=1}^{n}\left[(L_i - L_{(\lceil n\beta\rceil + m + 1)})^+ - (L_i - L_{(\lceil n\beta\rceil + m)})^+\right] \\ &= (n - \lceil n\beta\rceil - m)(L_{(\lceil n\beta\rceil + m + 1)} - L_{(\lceil n\beta\rceil + m)}).\end{aligned}$$

Thus

$$\Delta(m) = \left(L_{(\lceil n\beta\rceil+m+1)} - L_{(\lceil n\beta\rceil+m)}\right)\left((1-\varepsilon_n) - \frac{n-\lceil n\beta\rceil-m}{n-\lceil n\beta\rceil+1}\right).$$

Now $\Delta(0) > 0$ since $(L_{(\lceil n\beta\rceil+1)} - L_{(\lceil n\beta\rceil)}) > 0$ and $(1-\varepsilon_n) > (n-\lceil n\beta\rceil)(n-\lceil n\beta\rceil+1)^{-1}$ by the restriction on ε_n, and $\Delta(-1) < 0$ since $(L_{(\lceil n\beta\rceil)} - L_{(\lceil n\beta\rceil-1)}) > 0$ and $(1-\varepsilon_n) < 1$ again by the choice of ε_n. Thus $G_n(\alpha)$ has a unique minimum at $\alpha^* = L_{(\lceil n\beta\rceil)}$. □

Now consider the following CVaR estimator, expressed without the minimization:

Definition 2 (**Type 2 CVaR estimator**). For $\beta \in (0.5, 1)$, we define Type 2 CVaR estimator to be

$$\widehat{CV2}_n(\mathsf{L};\beta) := \frac{1}{n-\lceil n\beta\rceil+1}\sum_{i=1}^{n} L_i 1(L_i \geq \hat{\alpha}_n(\beta)),$$

where $\hat{\alpha}_n(\beta) := L_{(\lceil n\beta\rceil)}$, the $\lceil n\beta\rceil$-th order statistic of the sample $L_1, \ldots, L_n$.

The Type 2 CVaR estimator is an intuitive representation of CVaR because it is precisely the sample average of the top $(1-\beta)$ portion of the losses. Another advantage of the Type 2 CVaR estimator is that one can write down an explicit expression for its variance. For the rest of this subsection, our goal is to show that the Type 2 CVaR estimator is approximately equal to the Type 1 CVaR estimator, which is in turn approximately equal to the actual CVaR estimator we use in mean-CVaR portfolio optimization. The proof of Proposition 4 then follows.

Lemma 2. *Type 1 and Type 2 CVaR estimators are related by*

$$\widehat{CV2}_n(\mathsf{L};\beta) = \widehat{CV1}_n(\mathsf{L};\beta) + \varepsilon_n L_{(\lceil n\beta\rceil)}.$$

Proof. Rewriting Type 2 CVaR estimator:

$$\begin{aligned}
\widehat{CV2}_n(\mathsf{L};\beta) &= \frac{1}{n-\lceil n\beta\rceil+1}\sum_{i=1}^{n} L_i 1(L_i \geq L_{(\lceil n\beta\rceil)}) \\
&= L_{(\lceil n\beta\rceil)} + \frac{1}{n-\lceil n\beta\rceil+1}\sum_{i=1}^{n}(L_i - L_{(\lceil n\beta\rceil)})1(L_i \geq L_{(\lceil n\beta\rceil)}) \\
&= \widehat{CV1}_n(\mathsf{L};\beta) + \varepsilon_n L_{(\lceil n\beta\rceil)},
\end{aligned}$$

where the final equality is due to Lemma 1. □

We now prove Proposition 4.

Let α_n^1 and α_n be the minimizers in the definition of $\widehat{CV1}_n(\mathsf{L};\beta)$ and $\widehat{CVaR}_n(\mathsf{L};\beta)$ respectively. We can show, by elementary arguments,

$$|\widehat{CVaR}_n(\mathsf{L};\beta) - \widehat{CV1}_n(\mathsf{L};\beta)| \leq \varepsilon_n(\alpha_n^1 \vee \alpha_n).$$

Hence we have

$$\begin{aligned}\widehat{CVaR}_n(\mathsf{L};\beta) &= \widehat{CV1}_n(\mathsf{L};\beta) + O_p(\varepsilon_n) \\ &= \frac{1}{n-\lceil n\beta\rceil+1}\sum_{i=1}^{n} L_i 1(L_i \geq L_{(\lceil n\beta\rceil)}) + O_p(\varepsilon_n) \quad \text{by Lemma 2,}\end{aligned}$$

which implies

$$Var[\widehat{CVaR}_n(\mathsf{L};\beta)] = \frac{1}{n(1-\beta)^2} Var[L_i 1(L_i \geq L_{(\lceil n\beta\rceil)})] + O(n^{-2}),$$

where the $O(n^{-2})$ error comes from having approximated $n-\lceil n\beta\rceil+1$ by $n(1-\beta)$ in the denominator and since $\varepsilon_n = O(n^{-2})$. $\square$

A.5. Proof of Theorem 1

Before proving Theorem 1, we first show the following proposition.

PROPOSITION 7. *Consider the optimization problem*

$$\begin{array}{ll}\min\limits_{z\in\mathbb{R}^n} & z^\top 1_n \\ s.t. & z_i \geq 0 \ \forall\ i \\ & z_i \geq c_i \ \forall\ i \\ & z^\top \Omega_n z \leq f\end{array} \tag{21}$$

where $c \in \mathbb{R}^n$ is some constant vector, $f > 0$ is a constant scalar and $\Omega_n = (n-1)^{-1}(I_n - n^{-1}1_n 1_n^\top)$, the sample covariance operator. Suppose (21) is feasible with an optimal solution (z^*). Let $S_1(z) := \{1 \leq i \leq n : \ z_i = 0\}$, $S_2(z) := \{1 \leq i \leq n : \ z_i = c_i\}$ and $V(z) := S_1^c(z) \cap S_2^c(z) = \{1 \leq i \leq n : \ z_i > \max(0, c_i)\}$. Then, at the optimal solution z^*, we cannot have both $S_1(z^*)$ and $V(z^*)$ nonempty simultaneously.

Proof. Problem (21) is a convex optimization problem because Ω_n is a positive semidefinite matrix. The problem is also strictly feasible, since $z_0 = 2\max_i\{c_i\}1_n$ is a strictly feasible point: clearly, $z_{0,i} > \max\{0, c_i\}\ \forall\ i$ and $z_0^\top \Omega_n z_0 = 0 < f$ as 1_n is orthogonal to Ω_n. Thus Slater's condition for strong duality holds, and we can derive properties of the optimal solution by examining the KKT conditions.

The Lagrangian is

$$\mathcal{L}(z, \eta_1, \eta_2, \lambda) = \lambda z^\top \Omega_n z + (1_n - \eta_1 - \eta_2)^\top z + \eta_2^\top c - \lambda f$$

The KKT conditions are

- Primal feasibility
- Dual feasibility: $\eta_1^*, \eta_2^* \geq 0$ component-wise and $\lambda^* \geq 0$
- Complementary slackness:

$z_i^* \eta_{1,i}^* = 0\ \forall\ i$, $(z_i^* - c_i)\eta_{2,i}^* = 0\ \forall\ i$ and $\lambda^*[(z^*)^\top \Omega_n z^* - f] = 0$

- First Order Condition:

$$\nabla_{z^*}\mathcal{L} = 2\lambda\Omega_n z^* + (1_n - \eta_1^* - \eta_2^*) = 0 \tag{22a}$$

By substituting for Ω_n, (22a) can be written as

$$\frac{2\lambda}{n-1}\left(z^* - \frac{1}{n}(1_n^\top z^*)1_n\right) = -1_n + \eta_1^* + \eta_2^*. \tag{23}$$

Suppose $S_1(z^*) \neq \emptyset$ at the optimal primal-dual point $(z^*, \eta_1^*, \eta_2^*, \lambda^*)$. Then $\exists\ i_0 \in S_1(z^*)$ such that $z_{i_0}^* = 0$. The i_0-th component of (23) gives

$$-\frac{2\lambda^*}{n(n-1)}(1_n^\top z^*) = -1 + \eta_{1,i_0}^* + \eta_{2,i_0}^*. \tag{24}$$

Now suppose $V(z^*) \neq \emptyset$ at the optimal primal-dual point $(z^*, \eta_1^*, \eta_2^*, \lambda^*)$. Then $\exists\ j_0 \in V(z^*)$ such that $z_{j_0}^* > \max(0, c_i)$, which implies $\eta_{1,j_0}^* = 0$ and $\eta_{2,j_0}^* = 0$ by complementary slackness. The j_0-th component of (23) gives

$$\frac{2\lambda^*}{n-1}\left(z_{j_0}^* - \frac{1}{n}(1_n^\top z^*)\right) = -1, \tag{25}$$

which implies $\lambda^* > 0$ since λ^* cannot equal zero.

Now suppose $S_1(z^*)$ and $V(z^*)$ are both nonempty. Combining (24) and (25), we arrive at the necessary condition

$$\frac{2\lambda^*}{n-1} z_{j_0}^* = -\eta_{1,i_0}^* - \eta_{2,i_0}^*.$$

which is clearly a contradiction since $lhs > 0$ whereas $rhs \leq 0$. Hence $S_1(z^*)$ and $V(z^*)$ cannot both be nonempty, and the result follows. □

We now prove Theorem 1.

Clearly, (cv-relax) is a relaxation of (cv-PBR′): the components of the variable z in (cv-relax) are relaxations of $\max(0, -w^\top X_i - \alpha)$. Thus the two problem formulations are equivalent if at optimum, $z_i = \max(0, -w^\top X_i - \alpha)\ \forall\ i = 1, \ldots, n$ for (cv-relax).

Let $(\alpha^*, w^*, z^*, \nu_1^*, \nu_2^*, \eta_1^*, \eta_2^*, \lambda_1^*, \lambda_2^*)$ be the primal-dual optimal point for (cv-relax) and (cv-relax-d). Our aim is to show that $V(z^*)$, the set of indices for which $z_i^* > \max(0, -w^\top X_i - \alpha)$, is empty. Suppose the contrary. Then by Proposition 7, $S_1(z^*)$, the set of indices for which $z_i^* = 0$, is empty. This means $z_i^* > 0\ \forall\ i$ and $\eta_{1,i}^* = 0\ \forall\ i$ by complementary slackness.

Now consider the sub-problem for a fixed η_2 in the dual problem (cv-relax-d):

$$\max_{\eta_1 : \eta_1 \geq 0} -(\eta_1 + \eta_2)\Omega_n^\dagger(\eta_1 + \eta_2). \tag{26}$$

As 1_n is orthogonal to $\Omega_n^\dagger$, and $\Omega_n^\dagger$ is positive semidefinite, the optimal solution is of the form $\eta_1 = a1_n - \eta_2$, where a is any constant such that $a \geq \max_i(\eta_{2,i})$, with a corresponding optimal objective 0. Hence, bearing in mind the constraints $\eta_2 \geq 0$ and $\eta_2^\top 1_n = 1$ in (cv-relax-d), $\eta_1 = 0$ is one of the optimal solutions iff $\eta_2^* = 1_n/n$. Thus if $\eta_2^* \neq 1_n/n$, we get a contradiction. Otherwise, we can force the dual problem to find a solution with $\eta_1 \neq 0$ by introducing an additional constraint $\eta_1^\top 1_n \geq \delta$ for some constant $0 < \delta < 1$ in the dual problem (cv-relax-d). □

A.6. Proof of Proposition 5

The lhs of the RO constraint in (mv-PBR-RO) is equivalent to the convex optimization problem

$$\begin{array}{ll}\min\limits_{u\in\mathbb{R}^p:(I-PP^\dagger)u=0} & -w^\top u \\ s.t. & u^\top P^{-\dagger}u\leq 1\end{array} \qquad \text{(RO constraint)}$$

The Lagrangian is

$$L(u,\lambda)=-w^\top u+\lambda u^\top P^\dagger u-\lambda,$$

with

$$\nabla_u L(u,\lambda)=-w+2\lambda P^\dagger u,$$

which equals zero for $P^\dagger u^*=\frac{1}{2\lambda}w$, i.e., when

$$u^*=u^*(\omega)=\frac{1}{2\lambda}Pw+(I-PP^\dagger)\omega,$$

for arbitrary $\omega\in\mathbb{R}^p$. However, the condition $(I-PP^\dagger)u=0$ implies

$$\frac{1}{2\lambda}(I-PP^\dagger)Pw+(I-PP^\dagger)^2\omega=\frac{1}{2\lambda}(P-PP^\dagger P)w+(I-PP^\dagger)^2\omega=(I-PP^\dagger)^2\omega=0,$$

since $PP^\dagger P=P$. If P is invertible, the above is trivially satisfied because $(I-PP^\dagger)$ is the zero matrix, and if P is not invertible then ω is restricted to be orthogonal to $(I-PP^\dagger)$. In either case, $(I-PP^\dagger)\omega=0$, and $u^*=\frac{1}{2\lambda}Pw$.

Thus the dual function is

$$g(\lambda)=L(u^*(\omega),\lambda)=-\frac{1}{4\lambda}w^\top Pw-\lambda,$$

with

$$\frac{dg(\lambda)}{d\lambda}=\frac{1}{4\lambda^2}w^\top Pw-1,$$

which equals zero for $\lambda^*=\frac{1}{2}\sqrt{w^\top Pw}$. Substituting this value into the dual function, we get $-\sqrt{w^\top Pw}$. Thus the RO constraint is equal to

$$\sqrt{w^\top Pw}\leq\sqrt[4]{U},$$

and substituting $P=\alpha\alpha^\top$ for (mv-PBR-1) and $P=A^*$ for (mv-PBR-2) we obtain the PBR constraints. □

Appendix B: PBR is different from known robust optimization models

In this section, we show that the PBR constraints on the portfolio risk are not equivalent to known robust optimization constraints such as those found in Delage and Ye (2010).

Case I: mean-variance problem. Consider the set

$$A(U,\gamma) := \{w : \max_{||\Sigma-\hat{\Sigma}_n||\leq\gamma} w^\top \Sigma w \leq U,\ w^\top 1_p = 1\}$$

and the set

$$B(U') := \{w : Svar(w^\top \hat{\Sigma}_n w) \leq U',\ w^\top 1_p = 1\}.$$

We will show that no constants (U,γ) and U' can make the two sets equivalent. For ease of exposition, let us consider the single asset $(p=1)$ case. Clearly, the wealth sum constraint implies $w=1$ in this case.

The robust constraint of set $A(U,\gamma)$ is then given by

$$\begin{aligned} &\max_{|\sigma^2-\hat{\sigma}_n^2|\leq\gamma} \sigma^2 \leq U \\ \iff &\max_{\hat{\sigma}_n^2-\gamma\leq\sigma^2\leq\hat{\sigma}_n^2+\gamma} \sigma^2 \leq U \\ &\iff \hat{\sigma}_n^2 \leq U-\gamma \end{aligned}$$

Thus

$$A(U,\gamma) = \begin{cases} \{1\} & \text{if } \hat{\sigma}_n^2 \leq U-\gamma \\ \emptyset & \text{otherwise} \end{cases}$$

Now the PBR constraint in set $B(U')$ is given by, using the result from Proposition 1,

$$Svar(w^2\hat{\sigma}_n^2) = w^4\left[\frac{1}{n}(\hat{\mu}_{4,n}-\hat{\sigma}_n^4) + \frac{2}{n(n-1)\hat{\sigma}_n^4}\right] \leq U',$$

where $\hat{\mu}_{4,n}$ is the sample estimate for the fourth central moment of the asset return distribution. When $w=1$, the above equals

$$\frac{1}{n}\hat{\mu_{4,n}} + \frac{3-n}{n(n-1)\hat{\sigma}_n^4} \leq U',$$

thus

$$B(U') = \begin{cases} \{1\} & \text{if } \frac{1}{n}\hat{\mu}_{4,n} + \frac{3-n}{n(n-1)}\hat{\sigma}_n^4 \leq U' \\ \emptyset & \text{otherwise.} \end{cases}$$

It is thus clear that no choice of (U,γ) and U' would make the two sets equivalent, unless there is a particular relationship between the sample variance and the fourth central moment of the asset return distribution. It is also clear that the main difference between the two sets is that the PBR

set $B(U')$ involves fourth moments of the asset return distribution, because it penalizes uncertainty in the risk estimation, whereas the standard robust constraint only involves the second moment of the asset return distribution because the robust protection is against the range of the second moment directly, rather than the whole risk function.

Case II: mean-CVaR problem. Consider the set

$$A(U,\gamma) := \left\{ w : \max_{||\mathbf{q}-\hat{q}_n||\le\gamma} \alpha + \frac{1}{1-\beta}\sum_{i=1}^{n} q_i(-w_i-\alpha)^+ \le U,\ w^\top 1_p = 1 \right\},$$

where $\hat{q}_{i,n} = 1/n$, the empirical measure, and the norm on the measure $\mathbf{q}$ is the total variational distance, and the set

$$\begin{aligned} B(U') &:= \left\{ w : \frac{1}{n(1-\beta)^2} z^\top \Omega_n z \le U',\ w^\top 1_p = 1,\ z_i = (-w_i-\alpha)^+, i = 1,\dots,n \right\} \\ &= \left\{ w : \frac{1}{n(n-1)(1-\beta)^2} \sum_{i=1}^{n} (z_i - \bar{z}_n) \le U',\ w^\top 1_p = 1,\ z_i = (-w_i-\alpha)^+, i = 1,\dots,n \right\} \end{aligned}$$

where $\bar{z}_n$ is the sample mean of $z_1,\dots,z_n$.

As before, we will show that no constants (U,γ) and U' can make the two sets equivalent. By the equivalence of the total variational distance to the 1-norm for a discrete distribution,

$$||\mathbf{q} - \hat{\mathbf{q}}_n|| = \frac{1}{2}\sum_{i=1}^{n} \left| q_i - \frac{1}{n} \right|,$$

and since the term multiplying q_i's are all non-negative, the optimal $\mathbf{q}$, gives all weight to the largest $(-w_i-\alpha)^+$ term, i.e.,

$$\max_{||\mathbf{q}-\hat{\mathbf{q}}_n||\le\gamma} \alpha + \frac{1}{1-\beta}\sum_{i=1}^{n} q_i(-w^\top X_i - \alpha)^+ = \alpha + \frac{1}{1-\beta}\left(\frac{1}{n} + 2\gamma\right) \max_i\ (-w_i-\alpha)^+,$$

and so

$$A(U,\gamma) = \left\{ w : \alpha + \frac{1}{1-\beta}\left(\frac{1}{n} + 2\gamma\right) \max_i\ (-w_i-\alpha)^+ \le U,\ w^\top 1_p = 1 \right\}.$$

Clearly, no choice of (U,γ) and U' can make $A(U,\gamma)$ equivalent to $B(U')$.

Appendix C: Proofs of results in Sec. 4

We prove Theorem 3 first then Theorem 2.

Proof of Theorem 3. The theory of M-estimation concerns the following scenario. Consider the parametric function $m_\theta : \mathcal{X} \mapsto \bar{\mathbb{R}}$, where θ is a parameter chosen from Θ, and $\mathcal{X}$ is a subset of the Euclidean space. We are interested in finding the parameter θ^* that maximizes (for minimization, we can use $-m_\theta$ instead) the expected value of this function $M(\theta) = \mathbb{E}m_\theta(X)$, where X is drawn from the probability space $(\Omega,\mathcal{F},P)$. In the absence of the true distributional knowledge, but in the presence of iid observations $X_1,\dots,X_n$, one can estimate θ^* by minimizing instead the

empirical function $M_n(\theta) = n^{-1}\sum_{i=1}^n m_\theta(X_i)$. Of central importance is whether the solution (or, a near-optimal solution) to the empirical problem is consistent, i.e., whether it converges to the true optimal as the number of observations tend to infinity. The following theorem provides sufficient conditions for asymptotic optimality. □

THEOREM 5 **(Theorem 5.7 in Van der Vaart (2000))**. *Let M, M_n, Θ, θ^* be as defined in the paragraph above, and let $\hat{\theta}_n$ be a near-optimal maximizer of M_n, i.e.,*

$$M_n(\hat{\theta}_n) \geq M_n(\theta^*) - o_P(1).$$

If

1. $\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{P} 0$, *and*
2. *For all $\varepsilon > 0$, $\sup_{\theta\in\Theta}\{M(\theta) : d(\theta,\theta^*) \geq \varepsilon\} < M(\theta^*)$,*

then $\hat{\theta}_n \xrightarrow{P} \theta^*$ as $n \to \infty$.

Note 2 is true if M is continuous and θ^* is unique. Also, the theorem does not require that the estimated solution $\hat{\theta}_n$ be unique in any way; it holds for any sequence of estimated solutions.

We thus need to show the uniform convergence of $M_n(\cdot)$ to $M(\cdot)$. By Theorem 19.4 of Van der Vaart (2000), it suffices to show the function class $\mathcal{F} = \{m_\theta : \theta \in \Theta\}$ has a finite bracketing number $N_{[\,]}(\epsilon, \mathcal{F}, L_1(P))$ for every $\epsilon > 0$. Without loss of generality, let us assume $\Theta = [-K, K]^p$, where K is a large positive scalar[4].

One class of functions with a finite bracketing number is the *Lipschitz* class of functions, which we define below.

DEFINITION 3 (LIPSCHITZ CLASS). Consider a class of measurable functions $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$, $f_\theta : \mathcal{X} \to \mathbb{R}$, under some probability measure P. We say $\mathcal{F}$ is a *Lipschitz class* about $\theta_0 \in \Theta$ if $\theta \mapsto f_\theta(x)$ is differentiable at θ_0 for P-almost every x with derivative $\dot{f}_{\theta_0}(x)$ and such that, for every θ_1 and θ_2 in a neighborhood of θ_0, there exists a measurable function $\dot{f}$ with $\mathbb{E}[\dot{f}^2(X_1)] < \infty$ such that

$$|f_{\theta_1}(x) - f_{\theta_2}(x)| \leq \dot{f}(x)||\theta_1 - \theta_2||_2.$$

Example 19.7 of Van der Vaart (2000) shows that if $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ is a class of measurable functions with bounded $\Theta \subset \mathbb{R}^d$ and $\mathcal{F}$ is Lipschitz about $\theta_0 \in \Theta$ then for every $0 < \varepsilon < diam(\Theta)$, there exists C such that

$$N_{[\,]}(\varepsilon\sqrt{\mathbb{E}(|\dot{f}(X)|^2)}, \mathcal{F}, L_2(P)) \leq C\left(\frac{diam(\Theta)}{\varepsilon}\right)^d, \tag{27}$$

[4] This is equivalent to assuming that all our problems are feasible; the exact value of K need not be known for the proofs to go through

i.e., has a finite bracketing number for all $\varepsilon > 0$.

Going back to our problem, $\theta \mapsto m_\theta(x) = \alpha + (1-\beta)^{-1}(-\alpha - w_1^\top x - v^\top L^\top x)^+$ is clearly differentiable at θ_{CV} for all $x \in \mathbb{R}^p$. Furthermore,

$$\nabla_\theta m_\theta(x) = \begin{bmatrix} -1 \\ -L^\top x \end{bmatrix} I(x),$$

where $I(x) := \mathbb{I}(-\alpha - w_1^\top x - v^\top L^\top x \geq 0)$, hence

$$\dot{m}(x) := \max(1, ||L^\top x||_\infty) \tag{28}$$

is an upper bound on $||\nabla_\theta m_\theta(x)||_\infty$ and is independent of θ. Thus $|m_{\theta_1}(x) - m_{\theta_2}(x)| \leq \dot{m}(x)||\theta_1 - \theta_2||_2$ for all $\theta_1, \theta_2 \in [-K, K]^p$, and together with Assumption 2, $\mathcal{F}$ is a Lipschitz class, and we have our conclusion. □

Proof of Theorem 2. Here, we show the uniform convergence of $M_n(\cdot)$ to $M(\cdot)$ directly.

$$\begin{aligned} |M_n(\theta) - M(\theta)| &= \left| w^\top \hat{\Sigma}_n w - \lambda_0 w^\top \hat{\mu}_n - (w^\top \Sigma w - \lambda_0 w^\top \mu) \right| \\ &\leq \left| w^\top (\hat{\Sigma}_n - \Sigma) w \right| + \lambda_0 \left| w^\top (\hat{\mu}_n - \mu) \right| \\ &\leq K^2 ||\hat{\Sigma}_n - \Sigma||_{op} + K\lambda_0 \sum_{i=1}^p |\hat{\mu}_{n,i} - \mu_i| , \end{aligned}$$

where $||\cdot||_{op}$ is the operator norm of a matrix. It is thus clear that the above converges to zero (uniformly) as n tends to infinity, by the operator norm consistency of the sample covariance matrix and the consistency of the sample mean. □

Proof of Theorem 4. **Case I:** $M_n(\theta, \lambda_1, \lambda_2)$ equal to (11).

Following on from the proof of Theorem 2, uniform convergence of $M_n(\theta, \lambda_1, \lambda_2)$ to $M(\theta)$.

$$\begin{aligned} |M_n(\theta) - M(\theta)| &= \left| w^\top \hat{\Sigma}_n w - \lambda_0 w^\top \hat{\mu}_n + \lambda_1 w^\top \alpha - (w^\top \Sigma w - \lambda_0 w^\top \mu) \right| \\ &\leq \left| w^\top (\hat{\Sigma}_n - \Sigma) w \right| + \lambda_0 \left| w^\top (\hat{\mu}_n - \mu) \right| + \lambda_1 |w^\top \alpha| \\ &\leq K^2 ||\hat{\Sigma}_n - \Sigma||_{op} + K\lambda_0 \sum_{i=1}^p |\hat{\mu}_{n,i} - \mu_i| + K||\hat{\alpha}||_\infty \lambda_1 O\left(\frac{1}{n^{1/4}}\right), \end{aligned}$$

where $||\cdot||_{op}$ is the operator norm of a matrix. The first two terms converge to zero as n tends to infinity by the same reasoning as in the proof of Theorem 2, and the last term clearly tends to zero.

Case II: $M_n(\theta, \lambda_1, \lambda_2)$ equal to (12).

Similar to Case I, it suffices to show the uniform convergence of $M_n(\theta, \lambda_1, \lambda_2)$ to $M(\theta)$.

$$\begin{aligned} |M_n(\theta) - M(\theta)| &= \left| w^\top \hat{\Sigma}_n w - \lambda_0 w^\top \hat{\mu}_n + \lambda_1 w^\top A^* w - (w^\top \Sigma w - \lambda_0 w^\top \mu) \right| \\ &\leq \left| w^\top (\hat{\Sigma}_n - \Sigma) w \right| + \lambda_0 \left| w^\top (\hat{\mu}_n - \mu) \right| + \lambda_1 |w^\top A^* w| \\ &\leq K^2 ||\hat{\Sigma}_n - \Sigma||_{op} + K\lambda_0 \sum_{i=1}^p |\hat{\mu}_{n,i} - \mu_i| + K^2 ||A^*||_2^2 \lambda_1 O\left(\frac{1}{n^{1/2}}\right), \end{aligned}$$

where $||\cdot||_{op}$ is the operator norm of a matrix. The first two terms converge to zero as n tends to infinity by the same reasoning as in the proof of Theorem 2, and the last term clearly tends to zero.

Case III: $M_n(\theta, \lambda_1, \lambda_2)$ equal to (13).

Here, it suffices to show the uniform convergence of the extra PBR terms to zero, as we know from Theorem 3 that the SAA part of the objective $M_n(\theta, \lambda_1, \lambda_2)$ converges uniformly to $M(\theta)$. The PBR part is:

$$\begin{aligned}
&\left| \frac{\lambda_1}{n} w^\top \hat{\Sigma}_n w + \frac{\lambda_2}{(n-1)(1-\beta)^2} \sum_{i=1}^{n} \left(z_\theta(X_i) - \frac{1}{n} \sum_{j=1}^{n} z_\theta(X_j) \right)^2 \right| \\
&\leq \frac{\lambda_1}{n} \left| w^\top \hat{\Sigma}_n w \right| + \frac{\lambda_2}{n(n-1)(1-\beta)^2} \left| \sum_{i=1}^{n} \left(z_\theta(X_i) - \frac{1}{n} \sum_{j=1}^{n} z_\theta(X_j) \right)^2 \right| \\
&= O\left(\frac{1}{n} \right),
\end{aligned}$$

which clearly tends to zero as n tends to infinity. □

References

Acerbi, Carlo, Dirk Tasche. 2002. Expected shortfall: a natural coherent alternative to value at risk. *Economic notes* **31**(2) 379–388.

Ahmadi, Amir A., Alex Olshevsky, Pablo A. Parrilo, John N. Tsitsiklis. 2013. Np-hardness of deciding convexity of quartic polynomials and related problems. *Mathematical Programming* **137**(1-2) 453–476.

Ban, Gah-Yi, Chritopher J. Chen. 2016. Portfolio optimization in high dimensions: Aggregate then optimize. Working Paper.

Belloni, Alexandre, Victor Chernozhukov. 2013. Least squares after model selection in high-dimensional sparse models. *Bernoulli* **19**(2) 521–547. URL `http://dx.doi.org/10.3150/11-BEJ410`.

Ben-Tal, Aharon, Laurent El Ghaoui, Arkadi Nemirovski. 2009. *Robust optimization*. Princeton University Press.

Best, Michael J., Robert R. Grauer. 1991. On the sensitivity of mean-variance-efficient portfolios to changes in asset means: some analytical and computational results. *Review of Financial Studies* **4**(2) 315–342.

Boyd, S.P., L. Vandenberghe. 2004. *Convex optimization*. Cambridge University Press.

Broadie, Mark. 1993. Computing efficient frontiers using estimated parameters. *Annals of Operations Research* **45**(1) 21–58.

Candes, Emmanuel, Terence Tao. 2007. The dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics* **35**(6) 2313–2351.

Chopra, Vijay K., William T. Ziemba. 1993. The effect of errors in means, variances, and covariances on optimal portfolio choice. *The Journal of Portfolio Management* **19**(2) 6–11.

Chopra, V.K. 1993. Improving Optimization. *The Journal of Investing* **2**(3) 51–59.

Delage, Erick, Yinyu Ye. 2010. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations research* **58**(3) 595–612.

DeMiguel, V., L. Garlappi, F.J. Nogales, R. Uppal. 2009a. A Generalized Approach to Portfolio Optimization: Improving Performance by Constraining Portfolio Norms. *Management Science* **55**(5) 798.

DeMiguel, Victor, Lorenzo Garlappi, Raman Uppal. 2009b. Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *Review of Financial Studies* **22**(5) 1915–1953.

DeMiguel, Victor, Francisco J Nogales, Raman Uppal. 2014. Stock return serial dependence and out-of-sample portfolio performance. *Review of Financial Studies* **27**(4) 1031–1073.

El Karoui, Noureddine. 2010. High-dimensionality effects in the markowitz problem and other quadratic programs with linear constraints: Risk underestimation. *The Annals of Statistics* **38**(6) 3487–3566.

Fisher, Ronald Aylmer. 1922. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 309–368.

Fisher, Ronald Aylmer. 1925. Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 22. Cambridge University Press, 700–725.

Frankfurter, George M., Herbert E. Phillips, John P. Seagle. 1971. Portfolio selection: The effects of uncertain means, variances and covariances. *Journal of Financial and Quantitative Analysis* **6**(5) 1251–1262.

French, K.R. 2015. Data library. *Online Data Library* URL `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`.

Frost, P.A., J.E. Savarino. 1988a. For better performance: Constrain portfolio weights. *The Journal of Portfolio Management* **15**(1) 29–34.

Frost, Peter A., James E. Savarino. 1986. An empirical bayes approach to efficient portfolio selection. *Journal of Financial and Quantitative Analysis* **21**(03) 293–305.

Frost, Peter A., James E. Savarino. 1988b. For better performance: Constrain portfolio weights. *The Journal of Portfolio Management* **15**(1) 29–34.

Goldfarb, D., G. Iyengar. 2003. Robust Portfolio Selection Problems. *Mathematics of Operations Research* **28**(1) 1–38.

Gotoh, J., M.J. Kim, A.E.B. Lim. 2015. Robust empirical optimization is almost the same as mean-variance optimization. Working Paper.

Gotoh, Jun-ya, Akiko Takeda. 2010. On the role of norm constraints in portfolio selection. *Computational Management Science* **8** 323–353. URL `http://dx.doi.org/10.1007/s10287-011-0130-2`.

Grant, Michael, Stephen Boyd. 2008. Graph implementations for nonsmooth convex programs. V. Blondel, S. Boyd, H. Kimura, eds., *Recent Advances in Learning and Control*. Lecture Notes in Control and Information Sciences, Springer-Verlag Limited, 95–110. `http://stanford.edu/~boyd/graph_dcp.html`.

Grant, Michael, Stephen Boyd. 2014. CVX: Matlab software for disciplined convex programming, version 2.1. `http://cvxr.com/cvx`.

Haff, LR. 1980. Empirical bayes estimation of the multivariate normal covariance matrix. *The Annals of Statistics* **8**(3) 586–597.

Hastie, T., R. Tibshirani, J. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.*. 2nd ed. Springer-Verlag.

Huber, Peter J. 1967. The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1. 221–233.

Ivanov, Valentin K. 1962. On linear problems which are not well-posed. *Soviet Math. Dokl*, vol. 3. 981–983.

Jagannathan, Ravi, Tongshu Ma. 2003. Risk reduction in large portfolios: Why imposing the wrong constraints helps. *The Journal of Finance* **58**(4) 1651–1684.

Jegadeesh, Narasimhan, Sheridan Titman. 1993. Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of finance* **48**(1) 65–91.

Jobson, J.D., B. Korkie. 1980. Estimation for Markowitz Efficient Portfolios. *Journal of the American Statistical Association* **75**(371) 544–554.

Jorion, Philippe. 1985. International portfolio diversification with estimation risk. *Journal of Business* 259–278.

Karoui, Noureddine El. 2013. On the realized risk of high-dimensional markowitz portfolios. *SIAM Journal on Financial Mathematics* **4**(1) 737–783.

Ledoit, Olivier, Michael Wolf. 2004. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis* **88**(2) 365–411.

Lim, A.E.B., J.G. Shanthikumar, G.-Y. Vahn. 2011. Conditional Value-at-Risk in portfolio optimization: coherent but fragile. *Operations Research Letters* **39**(3) 163 – 171.

Lo, Andrew W, Archie Craig MacKinlay. 1990. When are contrarian profits due to stock market overreaction? *Review of Financial studies* **3**(2) 175–205.

Markowitz, Harry. 1952. Portfolio selection*. *The Journal of Finance* **7**(1) 77–91.

Merton, Robert C. 1980. On estimating the expected return on the market: An exploratory investigation. *Journal of Financial Economics* **8**(4) 323–361.

Michaud, R.O. 1989. The Markowitz optimization enigma: Is optimized optimal? *Financial Analysts Journal* **45**(1) 3142.

Phillips, David L. 1962. A technique for the numerical solution of certain integral equations of the first kind. *Journal of the ACM* **9**(1) 84–97.

Rockafellar, R.T., S. Uryasev. 2000. Optimization of conditional value-at-risk. *Journal of Risk* **2** 21–41.

Shapiro, Alexander, Darinka Dentcheva, Andrzej Ruszczyński. 2009. *Lectures on stochastic programming: modeling and theory*, vol. 9. Society for Industrial and Applied Mathematics.

Tikhonov, Andrey. 1963. Solution of incorrectly formulated problems and the regularization method. *Soviet Math. Dokl.*, vol. 5. 1035.

Van der Vaart, A.W. 2000. *Asymptotic statistics*. Cambridge University Press.

Vapnik, Vladimir. 2000. *The nature of statistical learning theory*. Springer.