We modify two popular methods in machine learning, regularization and cross-validation, for the portfolio optimization problem. First, we introduce performance-based regularization (PBR), where the idea is to constrain the sample variances of the estimated portfolio risk and return. The goal of PBR is to steer the solution towards one associated with less estimation error in the performance. We consider PBR for mean-variance and mean-CVaR portfolio optimization problems. For the mean-variance problem, PBR introduces a quartic polynomial constraint, from which we make two convex approximations; one based on rank-1 approximation and another based on the best convex quadratic approximation. We then analytically show that, 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 is essentially tight. We show that the PBR models can be cast as robust optimization problems and establish asymptotic optimality of both SAA and PBR solutions, and show this extends to the corresponding efficient frontiers. To calibrate the right hand sides of the PBR constraints, we develop a new, performance-based $k$-fold cross-validation algorithm. Using this algorithm, we carry out an extensive empirical investigation of PBR against SAA, as well as other methods, including $L_1$, $L_2$ regularization and the equally-weighted portfolio on three different data sets. We find that PBR dominates all other benchmarks in the literature for two widely-used data sets with statistical significance.

Key words: machine learning; portfolio optimization; stochastic optimization; finance

History: First version: June 15, 2013. This version: June 24, 2016.
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 \( b + \delta \), \( ||b + \delta - b|| < \delta \), then finding \( x \) that minimizes the functional \( R(x) = ||Ax - b||^2 \) does not guarantee a good approximation to the desired solution even if \( \delta \) 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
\[
R^*(x) = ||Ax - b||^2 + \gamma(\delta)P(x),
\]
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 \( \delta \) 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, Candès and Tao (2007) and Belloni and Chernozhukov (2009)]:
\[
\min_{\beta \in \mathbb{R}^p} ||y - X\beta||_2 + \lambda P(\beta),
\]
where \( P(\beta) = ||\beta||_1 \), \( y = [y_1, \ldots, y_n] \in \mathbb{R}^n \) is the data on the observable, \( X = [X_1, \ldots, X_n] \in \mathbb{R}^{n \times p} \) 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 \( \beta \) 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 \( X \) 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:
\[
\begin{align*}
  w_0 &= \arg\min_{w \in \mathbb{R}^p} Risk(w^\top X) \\
  &\text{s.t.} \quad w^\top \mathbf{1}_p = 1 \\
  &\quad (w^\top \mu = R),
\end{align*}
\]
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 \( \text{Risk}(w^\top X) = w^\top \Sigma w \) we recover the classical model of Markowitz (1952) and setting \( \text{Risk}(w^\top X) = \text{CVaR}(-w^\top X; \beta) \), where \( \beta \in (0.5, 1) \) and

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

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: \( X = [X_1, \ldots, X_n] \). Assuming that these are iid, the standard data-driven approach to solving (PO) is to solve:

\[
\hat{w}_n = \arg\min_{w \in \mathbb{R}^p} \hat{\text{Risk}}_n(w^\top X) \quad \text{s.t.} \quad w^\top 1_p = 1 \quad (w^\top \hat{\mu}_n = R),
\]

where \( \hat{\text{Risk}}_n(w^\top X) \) 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 \( \hat{\text{Risk}}_n(w^\top X) \) 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 \textit{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. 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 [Ben-Tal et al. (2009)]. The robust counterpart of the PBR constraint on the risk, however, does not have a natural interpretation. It is loosely related to constraining estimation error in the portfolio risk, however the robust model does not enjoy the same intuitive interpretation of the original PBR formulation. We thus not only link PBR with robust optimization, 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 the same natural insight.

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. 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 proportional transaction costs) 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.

2. Survey of literature

As mentioned in the introduction, 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. 6, we show that PBR out-performs the standard regularization techniques in terms of the out-of-sample Sharpe ratio.

Finally, 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. 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 normally distributed, whereas both Delage and Ye (2010) and our work are based on a nonparametric setting.

**Notations.** Throughout the paper, we denote convergence in probability by $P_n \rightarrow$. 

3. 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

$$w_{MV} = \arg\min_{w \in \mathbb{R}^p} w'\Sigma w$$

s.t. $w'1_p = 1$

(MV-true)

$$w = \arg\min_{w \in \mathbb{R}^p} CVaR_\beta(-w'X)$$

s.t. $w'1_p = 1$

(CV-true)

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

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

$$\hat{w}_{n,MV} = \arg\min_{w \in \mathbb{R}^p} \hat{w}'\hat{\Sigma}_n w$$

s.t. $\hat{w}'1_p = 1$

$$(\hat{w}'\hat{\mu}_n = R)$$

(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 $R$:

$$w_{CV} = \arg\min_{w \in \mathbb{R}^p} CVaR_\beta(-w'X)$$

s.t. $w'1_p = 1$

$$(w'\mu = R)$$

(CV-true)
where
\[
CVaR(-w^T X; \beta) := \min_{\alpha} \alpha + \frac{1}{1-\beta} E(-w^T X - \alpha)^+,
\]
as in Rockafellar and Uryasev (2000).

The SAA approach to the problem is to solve
\[
\hat{w}_{n,CV} = \arg\min_{w \in \mathbb{R}^p} \hat{CVaR}_n(-w^T X; \beta)
\]
\[\text{s.t.} \quad w^T 1_p = 1 \quad (w^T \hat{\mu}_n = R), \tag{CV-SAA}\]

where
\[
\hat{CVaR}_n(-w^T X; \beta) := \min_{\alpha \in \mathbb{R}} \alpha + \frac{1}{n(1-\beta)} \sum_{i=1}^n (-w^T X_i - \alpha)^+,
\]
is the sample average estimator for $CVaR(-w^T X; \beta)$.

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. 5 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 $X \sim \mathcal{N}(\mu_{sim}, \Sigma_{sim})$, 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 $\hat{w}_{n,CV}^T \mu_{sim}$ versus realized risk $CVaR(-\hat{w}_{n,CV}^T X; \beta)$; 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 green 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
4. Performance-based regularization

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

$$\hat{w}_{n,PBR} = \arg\min_{w \in \mathbb{R}^p} \widehat{\text{Risk}}_n(w^T X)$$

s.t. $w^T 1_p = 1$

$$\text{SVar}(\widehat{\text{Risk}}_n(w^T X)) \leq U_1$$

$$(\text{SVar}(w^T \hat{\mu}_n) \leq U_2),$$

where $\text{SVar}(\cdot)$ 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 $\widehat{\text{Risk}}_n(w^T X)$ of the objective has a confidence interval around it, which is approximately equal to the sample standard deviation of the estimator $\widehat{\text{Risk}}_n(w^T X)$. As $w$ varies, the error associated with the point estimate varies, as the confidence interval is a function of $w$. The PBR constraint $\text{SVar}(\widehat{\text{Risk}}_n(w^T X)) \leq U_1$ 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 $\text{SVar}(w^T \hat{\mu}_n) \leq U_2$. 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 $Y$,

$$\mathbb{P}(|Y - \mathbb{E}Y| \geq \delta) \leq \frac{\text{Var}(Y)}{\delta^2}.$$ 

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

4.1. PBR for Mean-Variance Portfolio Optimization

The PBR model for the mean-variance problem is:

\[
\hat{w}_{n,MV} = \arg\min_{w \in \mathbb{R}^p} \ w^\top \hat{\Sigma}_n w \\
\text{s.t.} \quad w^\top 1_p = 1 \quad (w^\top \hat{\mu}_n = R) \\
SVar(w^\top \hat{\Sigma}_n w) \leq U. \tag{mv-PBR}
\]

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

\[
\hat{Q}_{ijkl} = \frac{1}{n} (\hat{\mu}_{4,ijkl} - \hat{\sigma}_{ij}^2 \hat{\sigma}_{kl}^2) + \frac{1}{n(n-1)} (\hat{\sigma}_{il}^2 \hat{\sigma}_{jl}^2 + \hat{\sigma}_{il}^2 \hat{\sigma}_{jk}^2).
\]
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 \( \hat{\sigma}_{ij}^2 \) is the sample average estimator for \( \sigma_{ij}^2 \), the covariance of the elements of \( X \) given by

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

Proof. See Appendix B.

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).

4.1.1. 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

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

(4)

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

\[
\hat{w}_{n,PBR1} = \arg\min_{w \in \mathbb{R}^p} w^\top \hat{\Sigma}_n w \\
\text{s.t.} \quad w^\top 1_p = 1 \\
\quad w^\top \hat{\mu}_n = R \\
\quad w^\top \hat{\alpha} \leq \sqrt{U},
\]  

(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

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

(5)

where \( \hat{w}_{n,MV} \) is the SAA solution, \( \lambda^* \) is the optimal Lagrange multiplier for the PBR constraint \( w^\top \alpha \leq \sqrt{U} \) and

\[
\beta_1 = \frac{\hat{\alpha}^\top \hat{\Sigma}_n^{-1} \hat{\mu}_n \cdot \hat{\mu}_n \cdot \hat{\Sigma}_n^{-1} 1_p - \hat{\alpha}^\top \hat{\Sigma}_n^{-1} 1_p \cdot \hat{\Sigma}_n^{-1} \hat{\mu}_n}{1_p^\top \hat{\Sigma}_n^{-1} 1_p \cdot \hat{\mu}_n \cdot \hat{\Sigma}_n^{-1} \hat{\mu}_n - (\hat{\mu}_n \cdot \hat{\Sigma}_n^{-1} 1_p)^2},
\]

\[
\beta_2 = \frac{\hat{\alpha}^\top \hat{\Sigma}_n^{-1} \hat{\mu}_n \cdot 1_p \cdot \hat{\Sigma}_n^{-1} 1_p - \hat{\alpha}^\top \hat{\Sigma}_n^{-1} 1_p \cdot \hat{\Sigma}_n^{-1} \hat{\mu}_n}{1_p^\top \hat{\Sigma}_n^{-1} 1_p \cdot \hat{\mu}_n \cdot \hat{\Sigma}_n^{-1} \hat{\mu}_n - (\hat{\mu}_n \cdot \hat{\Sigma}_n^{-1} 1_p)^2}.
\]
The solution to (mv-PBR-1) without the mean constraint is given by

\[
\hat{w}_{n,PBR} = \hat{w}_{n,MV} - \frac{1}{2} \Lambda^* \hat{\Sigma}^{-1}(\beta 1_p + \hat{\alpha}),
\]

where \(\hat{w}_{n,MV}\) is the SAA solution, \(\Lambda^*\) is the optimal Lagrange multiplier for the PBR constraint \(w^\top \alpha \leq \hat{\nu} \hat{U}\) and

\[
\beta = -\frac{1^\top_p \hat{\Sigma}^{-1} \hat{\alpha}}{1^\top_p \hat{\Sigma}^{-1} 1_p}.
\]

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

**Proof of Proposition 2** Let us start with (mv-PBR-1) with the mean constraint. The Lagrangian is:

\[
\mathcal{L}(w; \nu_1, \nu_2, \lambda) = w^\top \hat{\Sigma} w + \nu_1 (w^\top 1_p - 1) + \nu_2 (w^\top \hat{\mu}_n - \hat{\nu}) + \lambda (w^\top \hat{\alpha} - \hat{\nu} \hat{U})
\]

\[
= w^\top \hat{\Sigma} w + w^\top (\nu_1 1_p + \nu_2 \hat{\mu}_n + \lambda \hat{\alpha}) - \nu_1 - \nu_2 R - \lambda \hat{\nu} \hat{U}
\]

The first order condition (FOC) gives:

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

The Lagrangian dual function is:

\[
g(\nu_1, \nu_2, \lambda) = \inf_w \mathcal{L}(w; \nu_1, \nu_2, \lambda)
\]

\[
= -\frac{1}{4} (\nu_1 1_p + \nu_2 \hat{\mu}_n + \lambda \hat{\alpha})^\top \hat{\Sigma}^{-1} (\nu_1 1_p + \nu_2 \hat{\mu}_n + \lambda \hat{\alpha}) - \nu_1 - \nu_2 R - \lambda \hat{\nu} \hat{U}
\]

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 \((\nu_1, \nu_2)\) first. The first order conditions give:

\[
\frac{dg(\nu_1, \nu_2, \lambda)}{d\nu_1} = -\frac{1}{2} (\nu_2 \hat{\mu}_n + \lambda \hat{\alpha})^\top \hat{\Sigma}^{-1} 1_p - \frac{1}{2} \nu_1 1_p^\top \hat{\Sigma}^{-1} 1_p - 1 = 0
\]

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

Solving simultaneously, we have

\[
\nu_1^* = 2 \frac{\hat{\mu}_n^\top \hat{\Sigma}^{-1} \hat{\mu}_n - R \hat{\mu}_n^\top \hat{\Sigma}^{-1} 1_p}{1_p^\top \hat{\Sigma}^{-1} 1_p - \hat{\mu}_n^\top \hat{\Sigma}^{-1} \hat{\mu}_n - (\hat{\mu}_n^\top \hat{\Sigma}^{-1} 1_p)^2} + \lambda \hat{\alpha}^\top \hat{\Sigma}^{-1} \hat{\mu}_n - \hat{\alpha}^\top \hat{\Sigma}^{-1} 1_p, \\
\nu_2^* = 2 \frac{\hat{\mu}_n^\top \hat{\Sigma}^{-1} \hat{\mu}_n - R \hat{\mu}_n^\top \hat{\Sigma}^{-1} 1_p}{1_p^\top \hat{\Sigma}^{-1} 1_p - \hat{\mu}_n^\top \hat{\Sigma}^{-1} \hat{\mu}_n - (\hat{\mu}_n^\top \hat{\Sigma}^{-1} 1_p)^2} + \lambda \hat{\alpha}^\top \hat{\Sigma}^{-1} \hat{\mu}_n - \hat{\alpha}^\top \hat{\Sigma}^{-1} 1_p.
where \( \nu_0^1, \nu_0^2 \) are optimal dual variables for the original mean-variance problem. Hence, the optimal portfolio becomes

\[
\hat{w}_{n,PBR} = -\frac{1}{2} \hat{\Sigma}_n^{-1} ((\nu_0^1 + \lambda \beta_1) 1_p + (\nu_0^2 + \lambda \beta_2) \bar{\mu}_n + \lambda \hat{\alpha}) \\
= -\frac{1}{2} \hat{\Sigma}_n^{-1} (\nu_0^1 1_p + \nu_0^2 \bar{\mu}_n) - \frac{1}{2} \lambda \hat{\Sigma}_n^{-1} (\beta_1 1_p + \beta_2 \bar{\mu}_n + \hat{\alpha}) \\
= \hat{w}_{n,MV} - \frac{1}{2} \lambda \hat{\Sigma}_n^{-1} (\beta_1 1_p + \beta_2 \bar{\mu}_n + \hat{\alpha}),
\]

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:

\[
\mathcal{L}(w; \nu, \lambda) = w^\top \hat{\Sigma}_n w + \nu (w^\top 1_p - 1) + \lambda (w^\top \hat{\alpha} - \frac{U}{4}) \\
= w^\top \hat{\Sigma}_n w + w^\top (\nu 1_p + \lambda \hat{\alpha}) - \nu - \lambda \frac{U}{4}
\]

The FOC gives:

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

The Lagrangian dual function is:

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

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 \( \nu \), 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 \( \nu^0 \) is the optimal dual variables for the SAA mean-variance problem without the mean constraint. Hence, the optimal portfolio becomes

\[
\hat{w}_{n,PBR1} = -\frac{1}{2} \hat{\Sigma}_n^{-1} ((\nu^0 + \lambda \beta) 1_p + \lambda \hat{\alpha}) \\
= -\frac{1}{2} \nu^0 \hat{\Sigma}_n^{-1} 1_p - \frac{1}{2} \lambda \hat{\Sigma}_n^{-1} (\beta 1_p + \hat{\alpha}) \\
= \hat{w}_{n,MV} - \frac{1}{2} \lambda \hat{\Sigma}_n^{-1} (\beta 1_p + \hat{\alpha})
\]

where \( \hat{w}_{n,MV} \) is the optimal portfolio of the SAA mean-variance problem without PBR. \( \square \)
4.1.2. 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):

\[
A^* = \arg\min_{A \succeq 0} \|A - Q^2\|_F \quad \text{(Q approx)}
\]

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):

\[
\hat{\mathbf{w}}_{n,PBR2} = \arg\min_{w \in \mathbb{R}^p} w^\top \hat{\Sigma}_n w \quad \text{s.t.} \quad w^\top \mathbf{1}_p = 1, \quad (w^\top \hat{\mu}_n = R), \quad w^\top A^* w \leq \sqrt{U}. \quad \text{(mv-PBR-2)}
\]

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

\[
\hat{\mathbf{w}}_{n,PBR2} = -\frac{1}{2} \hat{\Sigma}_n (\lambda^*)^{-1} (\nu_1^*(\lambda^*) \mathbf{1}_p + \nu_2^*(\lambda^*) \hat{\mu}_n), \quad \text{(9)}
\]

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

\[
\nu_1^*(\lambda) = 2 \frac{R \hat{\mu}_n^\top \hat{\Sigma}_n^{-1} \mathbf{1}_p - \hat{\mu}_n^\top \hat{\Sigma}_n^{-1} \hat{\mu}_n}{1_p^\top \hat{\Sigma}_n^{-1} \mathbf{1}_p \hat{\mu}_n^\top \hat{\Sigma}_n^{-1} \hat{\mu}_n - (\hat{\mu}_n^\top \hat{\Sigma}_n^{-1} \mathbf{1}_p)^2},
\]

\[
\nu_2^*(\lambda) = 2 \frac{-R \hat{\mu}_n^\top \hat{\Sigma}_n^{-1} \mathbf{1}_p + \hat{\mu}_n^\top \hat{\Sigma}_n^{-1} \hat{\mu}_n - (\hat{\mu}_n^\top \hat{\Sigma}_n^{-1} \mathbf{1}_p)^2}{1_p^\top \hat{\Sigma}_n^{-1} \mathbf{1}_p \hat{\mu}_n^\top \hat{\Sigma}_n^{-1} \hat{\mu}_n - (\hat{\mu}_n^\top \hat{\Sigma}_n^{-1} \mathbf{1}_p)^2}.
\]

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

\[
\hat{\mathbf{w}}_{n,PBR2} = \frac{\hat{\Sigma}_n(\lambda^*)^{-1} \mathbf{1}_p}{1_p^\top \hat{\Sigma}_n(\lambda^*)^{-1} \mathbf{1}_p}, \quad \text{(10)}
\]

where \(\hat{\Sigma}_n(\lambda^*) = \hat{\Sigma}_n + \lambda^* A^*\) and \(\lambda^*\) is the optimal Lagrange multiplier for the PBR constraint \(w^\top A^* w \leq \sqrt{U}\), as before.
For both mean-constrained and unconstrained cases, notice that the solution depends on \( \lambda \) only through the matrix \( \hat{\Sigma}_n(\lambda^*) = \hat{\Sigma}_n + \lambda^* A^* \). We thus retrieve the unregularized SAA solution \( \hat{w}_{n,MV} \) when \( \lambda \) is set to zero. Thus the PSD approximation to \((\text{mv-PBR})\) is equivalent to using a different estimator for the covariance matrix than the sample covariance matrix \( \hat{\Sigma}_n \). Clearly, \( \hat{\Sigma}_n(\lambda^*) \) 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. 6 we compare the two methods.

**Proof of Proposition 3**  
Let us start with \((\text{mv-PBR-2})\) with the mean constraint. The Lagrangian is:

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

\[
= w^\top (\hat{\Sigma}_n + \lambda A^*) w + w^\top (\nu_1 1_p + \nu_2 \hat{\mu}_n) - \nu_1 - \nu_2 R - \lambda \sqrt{U}
\]

The FOC gives:

\[
w^* = -\frac{1}{2} (\hat{\Sigma}_n + \lambda A^*)^{-1} (\nu_1 1_p + \nu_2 \hat{\mu}_n)
\]  \( (11) \)

The Lagrangian dual function is:

\[
g(\nu_1, \nu_2, \lambda) = \inf_w \mathcal{L}(w; \nu_1, \nu_2, \lambda)
\]

\[
= -\frac{1}{4} (\nu_1 1_p + \nu_2 \hat{\mu}_n)^\top (\hat{\Sigma}_n + \lambda A^*)^{-1} (\nu_1 1_p + \nu_2 \hat{\mu}_n) - \nu_1 - \nu_2 R - \lambda \sqrt{U}
\]

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 \((\nu_1, \nu_2)\) first. The first order conditions give:

\[
\frac{dg(\nu_1, \nu_2, \lambda)}{d\nu_1} = -\frac{1}{2} \nu_2 \hat{\mu}_n^\top (\hat{\Sigma}_n + \lambda A^*)^{-1} 1_p - \frac{1}{2} \nu_1 1_p^\top (\hat{\Sigma}_n + \lambda A^*)^{-1} 1_p - 1 = 0
\]

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

Solving simultaneously, we have

\[
\nu_1^* = 2 \frac{R \hat{\mu}_n^\top \hat{\Sigma}_n^{-1} 1_p - \hat{\mu}_n^\top \hat{\Sigma}_n^{-1} \hat{\mu}_n}{1_p^\top \hat{\Sigma}_n^{-1} 1_p \hat{\mu}_n^\top \hat{\Sigma}_n^{-1} \hat{\mu}_n - \big(\hat{\mu}_n^\top \hat{\Sigma}_n^{-1} 1_p\big)^2,}
\]

\[
\nu_2^* = 2 \frac{-R \hat{\mu}_n^\top \hat{\Sigma}_n^{-1} \hat{\mu}_n - \hat{\mu}_n^\top \hat{\Sigma}_n^{-1} 1_p}{1_p^\top \hat{\Sigma}_n^{-1} 1_p \hat{\mu}_n^\top \hat{\Sigma}_n^{-1} \hat{\mu}_n - \big(\hat{\mu}_n^\top \hat{\Sigma}_n^{-1} 1_p\big)^2,}
\]
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. □

4.2. PBR for Mean-CVaR Portfolio Optimization

The PBR model for the mean-CVaR problem is:

$$\min_{w \in \mathbb{R}^p} CVaR_n(-w^\top X; \beta)$$

s.t. $w^\top 1_p = 1$ \hfill (cv-PBR)

$$SVaR(CVaR_n(-w^\top X; \beta)) \leq U_1$$

$$SVaR(w^\top \hat{\mu}_n) \leq U_2.$$ 

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 $CVaR_n(-w^\top X; \beta)$ is given by the following proposition.

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

$$Var(CVaR_n(-w^\top X; \beta)) = \frac{1}{n(1-\beta)^2} Var((w^\top X - \alpha_\beta(w))^+) + O(n^{-2}),$$

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 $\beta$.

**Proof.** See Appendix B.

Thus, the sample variance of $CVaR_n(-w^\top X; \beta)$ is, to first order,

$$SVaR(CVaR_n(-w^\top X; \beta)) = \frac{1}{n(1-\beta)^2} z^\top \Omega_n z,$$

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:

$$\begin{align*}
\min_{\alpha, w, z} & \quad \alpha + \frac{1}{n(1 - \beta)} \sum_{i=1}^{n} z_i \\
\text{s.t.} & \quad w^\top 1_p = 1 \quad (\nu_1) \\
& \quad (w^\top \hat{\mu}_n = R) \quad (\nu_2) \\
& \quad \frac{1}{n(1 - \beta)^2} z^\top \Omega_n z \leq U_1 \quad (\lambda_1) \\
& \quad z_i \geq 0 \quad i = 1, \ldots, n \quad (\eta_1) \\
& \quad z_i \geq -w^\top X_i - \alpha, \ i = 1, \ldots, n \quad (\eta_2) \\
& \quad \frac{1}{n} w^\top \hat{\Sigma}_n w \leq U_2 \quad (\lambda_2)
\end{align*}$$

(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 $[n(1 - \beta)]^{-1} z^\top \Omega_n z \leq U_1$, 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 $[n(1 - \beta)]$ 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'):

$$\begin{align*}
\min_{\alpha, w, z} & \quad \alpha + \frac{1}{n(1 - \beta)} \sum_{i=1}^{n} z_i \\
\text{s.t.} & \quad w^\top 1_p = 1 \quad (\nu_1) \\
& \quad (w^\top \hat{\mu}_n = R) \quad (\nu_2) \\
& \quad \frac{1}{n(1 - \beta)^2} z^\top \Omega_n z \leq U_1 \quad (\lambda_1) \\
& \quad z_i \geq 0 \quad i = 1, \ldots, n \quad (\eta_1) \\
& \quad z_i \geq -w^\top X_i - \alpha, \ i = 1, \ldots, n \quad (\eta_2) \\
& \quad \frac{1}{n} w^\top \hat{\Sigma}_n w \leq U_2 \quad (\lambda_2)
\end{align*}$$

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

$$\begin{align*}
\max_{\nu_1, \nu_2, \lambda_1, \lambda_2, \eta_1, \eta_2} & \quad g(\nu_1, \nu_2, \eta_1, \eta_2, \lambda_1, \lambda_2) \\
\text{s.t.} & \quad \eta_2 1_n = 1 \quad (\nu_1) \\
& \quad \lambda_1 \geq 0, \lambda_2 \geq 0 \quad (\nu_2) \\
& \quad \eta_1 \geq 0, \eta_2 \geq 0 \quad (\lambda_2)
\end{align*}$$

(cv-relax-d)
where

\[
g(\nu_1, \nu_2, \lambda_1, \lambda_2, \eta_1, \eta_2) = -\frac{n}{2\lambda_1} (\nu_1 \lambda_1 + \nu_2 \hat{\mu}_n - X \eta_2)^\top \hat{\Sigma}_n^{-1} (\nu_1 \lambda_1 + \nu_2 \hat{\mu}_n - X \eta_2) - \frac{n(1 - \beta)^2}{2\lambda_2} (\eta_1 + \eta_2)^\top \Omega_n^\top (\eta_1 + \eta_2) + \nu_1 + R \nu_2 - U_1 \lambda_1 - U_2 \lambda_2,
\]

and \(\Omega_n^\top\) is the Moore-Penrose pseudo inverse of the singular matrix \(\Omega_n\).

We now state the result that \((\text{cv-PBR}')\) 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 \((\text{cv-PBR}')\) and \((\text{cv-relax-d})\). If \(\eta_2^* \neq 1_n/n\), then \((\alpha^*, w^*, z^*)\) is an optimal point of \((\text{cv-PBR}')\). Otherwise, if \(\eta_2^* = 1_n/n\), we can find the optimal solution to \((\text{cv-PBR}')\) by solving \((\text{cv-relax-d})\) with an additional constraint \(\eta_2^\top 1_n \geq \delta\), where \(\delta\) is any constant \(0 < \delta \ll 1\).

**Proof.** See Appendix C.

**Remark.** Theorem 1 shows that one can solve \((\text{cv-PBR}')\) via at most two steps. The first step is to solve \((\text{cv-PBR}')\); 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 \((\text{cv-relax-d})\) with an additional constraint \(\eta_2^\top 1_n \geq \delta\), where \(\delta\) 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. 6 were solved in a single step.

### 4.3. 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 \((\text{mv-PBR})\) has the following robust counterpart representation:

\[
\hat{w}_{n,PBR1} = \arg\min_{\hat{w} \in \mathbb{R}^p} \hat{w}^\top \hat{\Sigma}_n \hat{w}
\]

\[
s.t. \quad \hat{w}^\top 1_p = 1
\]

\[
(w^\top \hat{\mu}_n = R)
\]

\[
\max_{u \in \mathcal{U}} w^\top u \leq \sqrt{U},
\]

\[
(\text{mv-PBR-RO})
\]

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 \((\text{mv-PBR-1})\) and \(P = A^*\) for \((\text{mv-PBR-2})\), 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. The left hand side (lhs) of the RO constraint in (mv-PBR-RO) is equivalent to the convex optimization problem

\[
\min_{u \in \mathbb{R}^p : (I - PP^\dagger)u = 0} -w^\top u \\
\text{s.t. } u^\top P^{-1} u \leq 1
\]

(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 \( \omega \) 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

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

and substituting \( P = \alpha \alpha^\top \) for (mv-PBR-1) and \( P = A^* \) for (mv-PBR-2) we obtain the PBR constraints. \( \square \)
Proposition 6. The mean-CVaR PBR problem (cv-PBR) has the following robust counterpart representation:

\[
\min_{\alpha, w, z} \quad \alpha + \frac{1}{n(1-\beta)} \sum_{i=1}^{n} z_i \\
\text{s.t.} \quad w^\top 1_p = 1 \\
\quad \quad (w^\top \bar{\mu}_n = R) \\
\quad \quad \max_{\mu \in U_1} z^\top u \leq \sqrt{U_1} \\
\quad \quad z_i = \max(0, -w^\top x_i - \alpha), i = 1, \ldots, n \\
\quad \quad \max_{\bar{\mu} \in U_2} w^\top (\bar{\mu} - \mu) \leq \sqrt{U_2} \\
\text{(cv-PBR-RO)}
\]

where \( U_1 \) is the ellipsoid

\[
U_1 = \{ \mu \in \mathbb{R}^p \mid (\bar{\mu} - \mu)^\top \Sigma_n^{-1}(\bar{\mu} - \mu) \leq 1 \},
\]

and \( U_2 \) is the ellipsoid

\[
U_2 = \{ u \in \mathbb{R}^n \mid u^\top \Omega_n^1 u \leq 1, \ 1^\top u = 0 \},
\]

where \( \Omega_n^1 \) is the Moore-Penrose pseudoinverse of the matrix \( \Omega_n \).

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

5. 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:

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

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

\[
m_\theta (x) = ((w_1 + Lv)^\top x - (w_1 + Lv)^\top \mu)^2 - \lambda_0(w_1 + Lv)^\top x, \quad (12)
\]

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

\[
m_\theta (x) = \alpha + \frac{1}{1-\beta} z_\theta (x) - \lambda_0(w_1 + Lv)^\top x, \quad (13)
\]
for the mean-CVaR problem (CV-true), where \( z_\theta(x) = \max(0, -(w_1 + Lv)^T x - \alpha) \). 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 \( \Theta \) 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 (12). Denote the solution by \( \theta_{MV} \). Suppose, for all \( \epsilon > 0 \),

\[
\sup_{\theta \in \Theta} \{ M(\theta) : d(\theta, \theta_{MV}) \geq \epsilon \} < M(\theta_{MV}).
\]

(\*) 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

\[
\min_{\theta \in \Theta} M_n(\theta) = \min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} ((w_1 + Lv)^T X_i - (w_1 + Lv)^T \bar{\mu}_n)^2 - \lambda_0 (w_1 + Lv)^T X_i.
\]

**Theorem 3 (Asymptotic Optimality of SAA solution of mean-CVaR problem).** Consider (PO') with \( m_\theta(\cdot) \) as in (13). Denote the solution by \( \theta_{CV} \). Suppose, for all \( \epsilon > 0 \),

\[
\sup_{\theta \in \Theta} \{ M(\theta) : d(\theta, \theta_{CV}) \geq \epsilon \} < M(\theta_{CV}).
\]

(**) 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

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

**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 \( \theta_{MV} \) and \( \theta_{CV} \). 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  \( \theta_{MV} \) (resp.  \( \theta_{CV} \)) be unique,  \( \Theta \) 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 D.

5.1. Asymptotic optimality of PBR solutions

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

\[
\hat{\theta}_{n,PBR} = \arg\min_{\theta=(\alpha,v) \in \mathbb{R} \times \mathbb{R}^{p-1}} M_n(\theta; \lambda_1, \lambda_2) \tag{PBR'}
\]

where

\[
M_n(\theta; \lambda_1, \lambda_2) = w^T \hat{\Sigma}_n w - \lambda_0 w^T \hat{\mu}_n + \lambda_1 w^T \alpha,
\]

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

\[
M_n(\theta; \lambda_1, \lambda_2) = w^T \hat{\Sigma}_n w - \lambda_0 w^T \hat{\mu}_n + \lambda_1 w^T A^* w,
\]

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

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

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 (14), (15) and (16), and $\theta_0$ 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 $\hat{\theta}_{n,PBR}$.

6. 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.

6.1. 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, $\sqrt{U}$, is calibrated using the out-of-sample performance-based $k$-cross validation algorithm (OOS-PBCV) which we explain in detail in Sec. 6.4.
3. **PBR (PSD)**: solving the convex quadratic approximation problem (mv-PBR-2). The rhs of the PBR constraint, $\sqrt{U}$, 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.

### 6.2. 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. 6.4.

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.

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

$$\hat{\mu}_{\text{test}} := \frac{1}{N_{\text{test}}} \sum_{t} \hat{w}_t^\top X_t$$

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

i.e., the out-of-sample mean and variance of the portfolio returns. We can include a proportional transaction cost of $c$ to the out-of-sample portfolio returns as follows:

$$\text{Return}_t = \frac{W_{t+1}}{W_t} - 1,$$

where $W_t$ is the wealth at period $t$ net of transaction cost given by

$$W_{t+1} = W_t (1 + r_{t+1}^\top \hat{w}_t) \left(1 - c \sum_{j=1}^p |\hat{w}_{t+1,j} - \hat{w}_{t+1,j}|\right),$$

where $\hat{w}_{t,j}$ is the portfolio weight in asset $j$ at time $t$, $\hat{w}_{t+1,j}$ is the portfolio weight before rebalancing at $t + 1$ and $\hat{w}_{t+1,j}$ is the desired portfolio weight at time $t + 1$ after rebalancing. Hence the out-of-sample portfolio mean return and variance with proportional transaction costs are given by

$$\hat{\mu}_{\text{test,tc}} := \frac{1}{N_{\text{test}}} \sum_{t} \text{Return}_t$$

$$\hat{\sigma}_{\text{test,tc}}^2 := \frac{1}{N_{\text{test}} - 1} \sum_{t} (\text{Return}_t - \hat{\mu}_{\text{test,tc}})^2,$$

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

We report the following performance metrics:
• **Sharpe Ratio** (with proportional transaction costs): we compute annualized Sharpe ratio with proportional transaction cost as

\[
\text{Sharpe ratio (with trans. cost)} = \frac{\hat{\mu}_{\text{test,tc}}}{\hat{\sigma}_{\text{test,tc}}}
\]  \hspace{1cm} (17)

• **Turnover**: the portfolio turnover, averaged over the testing period, is given by

\[
\text{Turnover} = \frac{1}{N_{\text{test}}} \sum_{t=1}^{N_{\text{test}}} \sum_{j=1}^{p} |\hat{\omega}_{t+1,j} - \hat{\omega}_{t,j}|^+.
\]  \hspace{1cm} (18)

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

### 6.4. Calibration algorithm for $U$: performance-based $k$-fold cross-validation

![Figure 3](image)

A schematic explaining the out-of-sample performance-based $k$-fold 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.

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 $k - 1$ 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 $D = [X_1, \ldots, X_{N_{\text{train}}}] \in \mathbb{R}^{p \times N_{\text{train}}}$ 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 $D \setminus D_i$ 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 $U_{-i}$, $[U_{-i}, \bar{U}_{-i}]$.
2. Solve $P_{-i}(U_{-i})$ on $D \setminus D_i$ 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 $U_{-i}^* \in [U_{-i}, \bar{U}_{-i}]$ 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 $U_{-i}$, $[U_{-i}, \bar{U}_{-i}]$**. 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., $[U_{-i}, \bar{U}_{-i}] \subseteq [U, \bar{U}]$. 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 $U < U^*$, then the problem will be infeasible.

The upper bound on $U$ is given by $\bar{U} = \hat{\text{Risk}}_{n}(-\hat{w}_{n}^\top X)$, 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 $\bar{U}$, the minimum possible PBR parameter, we solve

$$
\begin{align*}
\bar{U} &= \min_{w} \quad w^\top \alpha \\
\text{s.t.} \quad w^\top 1_p &= 1 \\
\quad w^\top \hat{\mu}_n &= R
\end{align*}
$$

for (mv-PBR-1),

$$
\begin{align*}
\bar{U} &= \min_{w} \quad w^\top A^* w \\
\text{s.t.} \quad w^\top 1_p &= 1 \\
\quad w^\top \hat{\mu}_n &= R
\end{align*}
$$

for (U-min1)
for (mv-PBR-2), and
\[
U = \min_{w,z} z^\top \Omega_n z
\]
\[s.t. \quad w^\top \hat{\mu}_n = R, \quad w^\top 1_p = 1, \quad z_i \geq -w^\top X_i - \alpha, \quad i = 1, \ldots, n. \quad (U\text{-min-cv})
\]
for (cv-PBR).

To find the upper bound on the subproblem, \(\bar{U}_{-i}\), we compute (SAA) on dataset \(D \setminus D_i\) for \(\hat{w}_{-i}\), then set
\[
\bar{U}_{-i} = \min[\bar{U}, \widehat{\text{Risk}}_n(\hat{w}_{-i}^\top X)].
\]

To find \(\underline{U}_{-i}\), we first solve
\[
\underline{U}_{tmp} = \min_{w} w^\top \alpha
\]
\[s.t. \quad w^\top 1_p = 1, \quad w^\top \hat{\mu}_{-i} = R \]
for (mv-PBR-1), where \(\hat{\mu}_{-i}\) the sample mean computed on \(D \setminus D_i\),
\[
\underline{U}_{tmp} = \min_{\hat{w}} w^\top A^\top w
\]
\[s.t. \quad w^\top 1_p = 1, \quad w^\top \hat{\mu}_{-i} = R \]
for (mv-PBR-2), and
\[
\underline{U}_{tmp} = \min_{w,z} z^\top \Omega_{-i} z
\]
\[s.t. \quad w^\top \hat{\mu}_{-i} = R, \quad w^\top 1_p = 1, \quad z_i \geq -w^\top X_i - \alpha, \quad i \in C \setminus C_i, \quad z_i \geq 0, \quad i \in C \setminus C_i, \]
for (cv-PBR), where \(\Omega_{-i}\) is the sample variance operator computed on \(D \setminus D_i\), and \(C\) and \(C_i\) are sets of labels of the elements in \(D\) and \(D_i\) respectively. We then set
\[
\underline{U}_{-i} = \max\{\underline{U}_{tmp}, U\}.
\]

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 \(t\Delta U := t(\bar{U}_{-i} - \underline{U}_{-i})/\text{Div}\), where \(\text{Div}\) a preset granularity parameter, \(t\) is a parameter that equals 1 initially then is backtracked at rate \(\beta\), a parameter chosen between 0 and 1, until the stopping criterion
\[
\text{Sharpe}(U - t\Delta U) < \text{Sharpe}(U) + \alpha t\Delta U \frac{d\text{Sharpe}(U)}{dU}
\]
Computing $d\text{Sharpe}(U)/dU$, 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:

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

where $\hat{w}^*(U)$ is the optimal PBR solution when the rhs is set to $U$. The first quantity, $\nabla_{\hat{w}^*}\text{Sharpe}(\hat{w}^*(U))$, 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:

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

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 $(1 - \text{bit})U$, where $0 < \text{bit} \ll 1$ is a predetermined parameter, then computing

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

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

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

In our computations, we used the parameters $\alpha = 0.4, \beta = 0.9, Div = 5, \text{bit} = 0.05$ 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.

### 6.5. 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. 6.1. 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 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,
Out-of-Sample Performance-Based $k$-Cross Validation (OOS-PBCV)

Initialize

Choose no. of bins $k$

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

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

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

Let $D_{\text{train}}^b$ denote the training data minus the $b$-th bin

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

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

if $\hat{U}_{-b} < \hat{U}$ then $U_{-b} = \hat{U}$ and terminate

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

if $\hat{U}_{-b} > \hat{U}$ then $U_{-b} = \hat{U}$ and terminate

else Compare and update boundaries:

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

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

Run (OOS-PBSO) with boundaries $[\hat{U}_{-b}, \hat{U}_{-b}]$ 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).

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. 6.1. 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.
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 $U = \bar{U} - b$, $\Delta U := t(\bar{U} - \bar{U} - b) / Div$, $t = 1$
  Compute
  $$\frac{d\text{Sharpe}(U)}{dU} = \nabla_w \text{Sharpe}(\hat{w} - b(U))^\top \left[ \frac{d\hat{w}_b(U)}{dU} \right],$$
  where
  $$\nabla_w \text{Sharpe}(\hat{w} - b(U)) = \frac{((\hat{w} - b)^\top \Sigma_{-b} \hat{w} - b) \mu_{-b} - (\hat{w}_b^\top \mu) \Sigma_{-b} \hat{w} - b}{((\hat{w} - b)^\top \Sigma_{-b} \hat{w} - b)^{3/2}}$$

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

  while
  $$\text{Sharpe}(U - t\Delta U) < \text{Sharpe}(U) + \alpha t \Delta U \frac{d\text{Sharpe}(U)}{dU}$$
  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).

6.6. 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. 6.2. 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. 6.2. 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.

### 7. Conclusion

We introduced performance-based regularization for the portfolio optimization problem and investigated it in detail. The idea is to 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 is highly effective at improving the finite-sample performance of the portfolio decision compared to SAA as well as other benchmarks in the portfolio optimization 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 (with transaction costs) for empirical data for the mean-variance problem.

<table>
<thead>
<tr>
<th></th>
<th>FF 5 Industry</th>
<th>FF 10 Industry</th>
<th>FF 49 Industry</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p=5</td>
<td>p=10</td>
<td>p=41</td>
</tr>
<tr>
<td></td>
<td>(-8 assets with missing data)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean-Variance R=0.04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAA</td>
<td>1.1459</td>
<td>1.1332</td>
<td>0.4744</td>
</tr>
<tr>
<td></td>
<td>2 bins 3 bins</td>
<td>2 bins 3 bins</td>
<td>2 bins 3 bins</td>
</tr>
<tr>
<td>PBR (rank-1)</td>
<td>1.2603</td>
<td>1.3254</td>
<td>1.1868 1.2098</td>
</tr>
<tr>
<td></td>
<td>(0.0411) 0.0286</td>
<td>(0.0643) 0.0509</td>
<td>(0.5848) 0.5386</td>
</tr>
<tr>
<td>PBR (PSD)</td>
<td>1.1836 1.1831</td>
<td>1.1543 1.1678</td>
<td>0.4776 0.4825</td>
</tr>
<tr>
<td></td>
<td>(0.0743) 0.071</td>
<td>(0.0891) 0.0816</td>
<td>(0.5593) 0.5391</td>
</tr>
<tr>
<td>NS</td>
<td>1.0023</td>
<td>0.9968</td>
<td>0.7345</td>
</tr>
<tr>
<td></td>
<td>(0.1404) 0.1437</td>
<td>(0.2977)</td>
<td></td>
</tr>
<tr>
<td>L1</td>
<td>1.0136 1.0386</td>
<td>1.1185 1.1175</td>
<td>0.5419 0.5211</td>
</tr>
<tr>
<td></td>
<td>(0.1568) 0.1396</td>
<td>(0.1008) 0.1017</td>
<td>(0.5044) 0.5216</td>
</tr>
<tr>
<td>L2</td>
<td>0.9711 1.0268</td>
<td>1.0579 1.0699</td>
<td>0.6672 0.6009</td>
</tr>
<tr>
<td></td>
<td>(0.1781) 0.1452</td>
<td>(0.1482) 0.1280</td>
<td>(0.3950) 0.4455</td>
</tr>
<tr>
<td>Mean-Variance R=0.06</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAA</td>
<td>1.1535</td>
<td><strong>1.1357</strong></td>
<td>0.4468</td>
</tr>
<tr>
<td></td>
<td>2 bins 3 bins</td>
<td>2 bins 3 bins</td>
<td>2 bins 3 bins</td>
</tr>
<tr>
<td>PBR (rank-1)</td>
<td>1.2945 1.3362</td>
<td>1.1870 <strong>1.2112</strong></td>
<td>0.4011 0.4515</td>
</tr>
<tr>
<td></td>
<td>(0.0297) 0.0244</td>
<td>(0.0629) 0.0503</td>
<td>(0.6136) 0.5530</td>
</tr>
<tr>
<td>PBR (PSD)</td>
<td>1.1912 <strong>1.2052</strong></td>
<td>1.1532 <strong>1.1696</strong></td>
<td>0.4585 0.4587</td>
</tr>
<tr>
<td></td>
<td>(0.0689) 0.0638</td>
<td>(0.0898) 0.0809</td>
<td>(0.5757) 0.5598</td>
</tr>
<tr>
<td>NS</td>
<td><strong>0.9853</strong></td>
<td>0.9699</td>
<td>0.7124</td>
</tr>
<tr>
<td></td>
<td>(0.1422) 0.1537</td>
<td>(0.3247)</td>
<td></td>
</tr>
<tr>
<td>L1</td>
<td><strong>0.9963</strong></td>
<td>1.0092 1.1010</td>
<td>0.4991 0.4941</td>
</tr>
<tr>
<td></td>
<td>(0.1535) 0.1394</td>
<td>(0.1124) 0.1101</td>
<td>(0.5490) 0.5448</td>
</tr>
<tr>
<td>L2</td>
<td>0.9713 1.0265</td>
<td>1.0642 1.0755</td>
<td>0.6913 0.5701</td>
</tr>
<tr>
<td></td>
<td>(0.1735) 0.1425</td>
<td>(0.1425) 0.1238</td>
<td>(0.4250) 0.4696</td>
</tr>
<tr>
<td>Mean-Variance R=0.08</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAA</td>
<td><strong>1.1573</strong></td>
<td>1.1225</td>
<td>0.4253</td>
</tr>
<tr>
<td></td>
<td>2 bins 3 bins</td>
<td>2 bins 3 bins</td>
<td>2 bins 3 bins</td>
</tr>
<tr>
<td>PBR (rank-1)</td>
<td>1.3286 <strong>1.3551</strong></td>
<td>1.1743 1.2018</td>
<td>0.3927 0.4253</td>
</tr>
<tr>
<td></td>
<td>(0.0223) 0.0208</td>
<td>(0.0668) 0.0510</td>
<td>(0.6142) 0.5778</td>
</tr>
<tr>
<td>PBR (PSD)</td>
<td>1.1813 1.1952</td>
<td>1.1467 1.1575</td>
<td>0.4477 0.4366</td>
</tr>
<tr>
<td></td>
<td>(0.0648) 0.0614</td>
<td>(0.0893) 0.0844</td>
<td>(0.5852) 0.5804</td>
</tr>
<tr>
<td>NS</td>
<td>0.9664</td>
<td>0.9405</td>
<td>0.6600</td>
</tr>
<tr>
<td></td>
<td>(0.1514) 0.1577</td>
<td>(0.3790)</td>
<td></td>
</tr>
<tr>
<td>L1</td>
<td><strong>0.9225</strong></td>
<td>0.9965 1.0318 1.0779</td>
<td>0.4770 0.4930</td>
</tr>
<tr>
<td></td>
<td>(0.1857) 0.1403</td>
<td>(0.1332) 0.1181</td>
<td>(0.5649) 0.5379</td>
</tr>
<tr>
<td>L2</td>
<td><strong>0.9703</strong></td>
<td><strong>1.0284</strong></td>
<td>1.0671 <strong>1.0776</strong></td>
</tr>
<tr>
<td></td>
<td>(0.1649) 0.1398</td>
<td>(0.1398) 0.1209</td>
<td>(0.4369) 0.4785</td>
</tr>
<tr>
<td>Min. Variance</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAA</td>
<td>1.1454</td>
<td>1.1331</td>
<td><strong>0.4816</strong></td>
</tr>
<tr>
<td></td>
<td>2 bins 3 bins</td>
<td>2 bins 3 bins</td>
<td>2 bins 3 bins</td>
</tr>
<tr>
<td>PBR (rank-1)</td>
<td>1.2580 1.3269</td>
<td>1.1922 1.2086</td>
<td>0.4409 0.4683</td>
</tr>
<tr>
<td></td>
<td>(0.0420) 0.0288</td>
<td>(0.0603) 0.0505</td>
<td>(0.5795) 0.5472</td>
</tr>
<tr>
<td>PBR (PSD)</td>
<td>1.1883 1.1882</td>
<td>1.154 1.1657</td>
<td><strong>0.4942</strong> 0.4903</td>
</tr>
<tr>
<td></td>
<td>(0.0710) 0.0693</td>
<td>(0.0892) 0.0823</td>
<td>(0.5400) 0.5322</td>
</tr>
<tr>
<td>NS</td>
<td><strong>1.0022</strong></td>
<td><strong>1.0012</strong></td>
<td><strong>0.7347</strong></td>
</tr>
<tr>
<td></td>
<td>(0.1405) (0.1447)</td>
<td>(0.3178)</td>
<td></td>
</tr>
<tr>
<td>L1</td>
<td><strong>1.0321</strong></td>
<td><strong>1.0546</strong></td>
<td><strong>1.1199</strong> 1.1111</td>
</tr>
<tr>
<td></td>
<td>(0.1455) 0.1286</td>
<td>(0.1000) 0.1026</td>
<td>(0.5017) 0.5151</td>
</tr>
<tr>
<td>L2</td>
<td>0.9945 1.0140</td>
<td>1.0543 1.0760</td>
<td><strong>0.6886</strong> 0.6204</td>
</tr>
<tr>
<td></td>
<td>(0.1632) 0.1472</td>
<td>(0.1488) 0.1236</td>
<td>(0.3761) 0.4276</td>
</tr>
<tr>
<td>Equal</td>
<td>0.6617</td>
<td>0.7019</td>
<td>0.6297</td>
</tr>
</tbody>
</table>

This table reports the annualized out-of-sample Sharpe ratios (with proportional transaction cost of 50 basis points per transaction) of solutions to the mean-variance problem solved with the methods described in Sec. 6.1 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. 6.4, 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>
<thead>
<tr>
<th></th>
<th>FF 5 Industry</th>
<th>FF 10 Industry</th>
<th>FF 49 Industry</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p=5</td>
<td>p=10</td>
<td>p=41</td>
</tr>
<tr>
<td></td>
<td>(-8 assets with missing data)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean-CVaR R=0.04</td>
<td>1.2137</td>
<td>1.0321</td>
<td><strong>0.3657</strong></td>
</tr>
<tr>
<td>SAA</td>
<td>2 bins</td>
<td>3 bins</td>
<td>2 bins</td>
</tr>
<tr>
<td></td>
<td>0.12113</td>
<td>1.1733</td>
<td>1.0506</td>
</tr>
<tr>
<td></td>
<td>(0.0554)</td>
<td>(0.0674)</td>
<td>(0.0638)</td>
</tr>
<tr>
<td></td>
<td>1.2089</td>
<td>1.1802</td>
<td>1.0994</td>
</tr>
<tr>
<td></td>
<td>(0.0746)</td>
<td>(0.0790)</td>
<td>(0.1051)</td>
</tr>
<tr>
<td></td>
<td>1.2439</td>
<td>1.2073</td>
<td>1.1112</td>
</tr>
<tr>
<td></td>
<td>(0.0513)</td>
<td>(0.0601)</td>
<td>(0.0691)</td>
</tr>
<tr>
<td></td>
<td>1.0112</td>
<td><strong>1.0754</strong></td>
<td>0.9254</td>
</tr>
<tr>
<td></td>
<td>(0.1497)</td>
<td>(0.1366)</td>
<td>(0.2293)</td>
</tr>
<tr>
<td></td>
<td>0.9650</td>
<td>1.0636</td>
<td>1.0031</td>
</tr>
<tr>
<td></td>
<td>(0.1780)</td>
<td>(0.1287)</td>
<td>(0.1512)</td>
</tr>
<tr>
<td></td>
<td>0.3657</td>
<td>0.2247</td>
<td>0.3513</td>
</tr>
<tr>
<td></td>
<td>L1</td>
<td>1.012</td>
<td>1.0754</td>
</tr>
<tr>
<td></td>
<td>0.9650</td>
<td>1.0636</td>
<td>1.0031</td>
</tr>
<tr>
<td></td>
<td>(0.1780)</td>
<td>(0.1287)</td>
<td>(0.1512)</td>
</tr>
<tr>
<td></td>
<td>0.3513</td>
<td>0.2247</td>
<td>0.3513</td>
</tr>
<tr>
<td>Mean-CVaR R=0.06</td>
<td>1.2137</td>
<td>1.0321</td>
<td><strong>0.3657</strong></td>
</tr>
<tr>
<td>SAA</td>
<td>2 bins</td>
<td>3 bins</td>
<td>2 bins</td>
</tr>
<tr>
<td></td>
<td>1.2233</td>
<td>1.2063</td>
<td>1.0518</td>
</tr>
<tr>
<td></td>
<td>(0.0503)</td>
<td>(0.0527)</td>
<td>(0.0633)</td>
</tr>
<tr>
<td></td>
<td>1.2205</td>
<td>1.1902</td>
<td>1.0988</td>
</tr>
<tr>
<td></td>
<td>(0.0699)</td>
<td>(0.0746)</td>
<td>(0.1053)</td>
</tr>
<tr>
<td></td>
<td>1.2450</td>
<td>1.2043</td>
<td>1.1122</td>
</tr>
<tr>
<td></td>
<td>(0.0504)</td>
<td>(0.0581)</td>
<td>(0.0686)</td>
</tr>
<tr>
<td></td>
<td>0.9404</td>
<td>1.0464</td>
<td>0.9276</td>
</tr>
<tr>
<td></td>
<td>(0.1812)</td>
<td>(0.1395)</td>
<td>(0.2282)</td>
</tr>
<tr>
<td></td>
<td>0.9271</td>
<td>1.0627</td>
<td>1.0146</td>
</tr>
<tr>
<td></td>
<td>(0.1977)</td>
<td>(0.1286)</td>
<td>(0.1432)</td>
</tr>
<tr>
<td></td>
<td>0.3513</td>
<td>0.2247</td>
<td>0.3513</td>
</tr>
<tr>
<td>Mean-CVaR R=0.08</td>
<td>1.2487</td>
<td>1.0346</td>
<td><strong>0.3657</strong></td>
</tr>
<tr>
<td>SAA</td>
<td>2 bins</td>
<td>3 bins</td>
<td>2 bins</td>
</tr>
<tr>
<td></td>
<td>1.2493</td>
<td>1.2098</td>
<td>1.0551</td>
</tr>
<tr>
<td></td>
<td>(0.0434)</td>
<td>(0.0462)</td>
<td>(0.0579)</td>
</tr>
<tr>
<td></td>
<td>1.2480</td>
<td>1.2088</td>
<td>1.0987</td>
</tr>
<tr>
<td></td>
<td>(0.0591)</td>
<td>(0.0603)</td>
<td>(0.1053)</td>
</tr>
<tr>
<td></td>
<td><strong>1.2715</strong></td>
<td>1.2198</td>
<td>1.1122</td>
</tr>
<tr>
<td></td>
<td>(0.0453)</td>
<td>(0.0545)</td>
<td>(0.0664)</td>
</tr>
<tr>
<td></td>
<td>0.8921</td>
<td>0.9836</td>
<td>0.9416</td>
</tr>
<tr>
<td></td>
<td>(0.1964)</td>
<td>(0.1572)</td>
<td>(0.2122)</td>
</tr>
<tr>
<td></td>
<td>0.9367</td>
<td><strong>1.0801</strong></td>
<td><strong>1.0278</strong></td>
</tr>
<tr>
<td></td>
<td>(0.1989)</td>
<td>(0.1119)</td>
<td>(0.1323)</td>
</tr>
<tr>
<td>Global min. CVaR</td>
<td><strong>1.2137</strong></td>
<td><strong>1.0321</strong></td>
<td><strong>0.3657</strong></td>
</tr>
<tr>
<td>Equal</td>
<td>0.6617</td>
<td>0.7019</td>
<td><strong>0.6297</strong></td>
</tr>
</tbody>
</table>

This table reports the annualized out-of-sample Sharpe ratios (with proportional transaction cost of 50 basis points per transaction) 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. 6.4, 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 (with transaction costs) for empirical data for the mean-variance problem.

<table>
<thead>
<tr>
<th></th>
<th>FF 5 Industry</th>
<th>FF 10 Industry</th>
<th>FF 49 Industry</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p=5</td>
<td>p=10</td>
<td>p=41</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-8 assets with missing data)</td>
<td></td>
</tr>
<tr>
<td>Mean-Variance</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R=0.04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAA</td>
<td>0.0935</td>
<td>0.1325</td>
<td>0.5188</td>
</tr>
<tr>
<td>PBR (rank-1)</td>
<td>0.1213</td>
<td>0.1746</td>
<td>0.5442</td>
</tr>
<tr>
<td>PBR (PSD)</td>
<td>0.1002</td>
<td>0.1415</td>
<td>0.5201</td>
</tr>
<tr>
<td>NS</td>
<td>0.0391</td>
<td>0.0544</td>
<td>0.0833</td>
</tr>
<tr>
<td>L1</td>
<td>0.0986</td>
<td>0.1158</td>
<td>0.5167</td>
</tr>
<tr>
<td>L2</td>
<td>0.1171</td>
<td>0.1255</td>
<td>0.4704</td>
</tr>
<tr>
<td>Mean-Variance</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R=0.06</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAA</td>
<td>0.1034</td>
<td>0.1339</td>
<td>0.5289</td>
</tr>
<tr>
<td>PBR (rank-1)</td>
<td>0.1397</td>
<td>0.1741</td>
<td>0.5646</td>
</tr>
<tr>
<td>PBR (PSD)</td>
<td>0.1132</td>
<td>0.1442</td>
<td>0.5301</td>
</tr>
<tr>
<td>NS</td>
<td>0.0417</td>
<td>0.0711</td>
<td>0.0859</td>
</tr>
<tr>
<td>L1</td>
<td>0.1206</td>
<td>0.1256</td>
<td>0.4992</td>
</tr>
<tr>
<td>L2</td>
<td>0.1267</td>
<td>0.1379</td>
<td>0.4809</td>
</tr>
<tr>
<td>Mean-Variance</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R=0.08</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAA</td>
<td>0.1288</td>
<td>0.1475</td>
<td>0.5434</td>
</tr>
<tr>
<td>PBR (rank-1)</td>
<td>0.1775</td>
<td>0.1894</td>
<td>0.5721</td>
</tr>
<tr>
<td>PBR (PSD)</td>
<td>0.1147</td>
<td>0.1689</td>
<td>0.5414</td>
</tr>
<tr>
<td>NS</td>
<td>0.0511</td>
<td>0.0965</td>
<td>0.1122</td>
</tr>
<tr>
<td>L1</td>
<td>0.1476</td>
<td>0.1480</td>
<td>0.5064</td>
</tr>
<tr>
<td>L2</td>
<td>0.1582</td>
<td>0.1470</td>
<td>0.5118</td>
</tr>
<tr>
<td>Min. Variance</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAA</td>
<td>0.1034</td>
<td>0.1325</td>
<td>0.5146</td>
</tr>
<tr>
<td>PBR (rank-1)</td>
<td>0.1245</td>
<td>0.1756</td>
<td>0.5393</td>
</tr>
<tr>
<td>PBR (PSD)</td>
<td>0.1221</td>
<td>0.1609</td>
<td>0.5138</td>
</tr>
<tr>
<td>NS</td>
<td>0.0391</td>
<td>0.0524</td>
<td>0.0835</td>
</tr>
<tr>
<td>L1</td>
<td>0.0995</td>
<td>0.1138</td>
<td>0.4956</td>
</tr>
<tr>
<td>L2</td>
<td>0.1213</td>
<td>0.1255</td>
<td>0.4575</td>
</tr>
<tr>
<td>Equal</td>
<td>0.0427</td>
<td>0.0382</td>
<td>0.0483</td>
</tr>
</tbody>
</table>

This table reports the portfolio turnovers (defined in Eq. 18) of the solutions to the mean-variance problem solved with the methods described in Sec. 6.1 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.

Appendix A: Covariance of Sample Covariance Matrix

Let \( X_1, \ldots, X^n \) be \( n \) iid observations of the vector \( X \). We wish to compute the covariance of the sample covariance matrix \( S \), whose elements are:

\[
S_{ij} = \frac{1}{n} \sum_{k=1}^{n} (X^k_i - \bar{X}_i)(X^k_j - \bar{X}_j),
\]

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

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

\[
S_{ij} = \frac{1}{\binom{n}{2}} \sum_{k \neq j}^{n} \frac{1}{2} (X^k_i - X^k_j)(X^k_j - X^j_i)
\]
Table 4  Turnovers (with transaction costs) for empirical data for the mean-CVaR problem.

<table>
<thead>
<tr>
<th>Sharpe Ratios</th>
<th>FF 5 Industry</th>
<th>FF 10 Industry</th>
<th>FF 49 Industry</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p=5</td>
<td>p=10</td>
<td>p=41</td>
</tr>
<tr>
<td></td>
<td>(-8 assets with missing data)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean-CVaR R=0.04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAA</td>
<td>0.2857</td>
<td>0.3534</td>
<td>1.6833</td>
</tr>
<tr>
<td>PBR (CVaR only)</td>
<td>0.1834</td>
<td>0.1985</td>
<td>0.4049</td>
</tr>
<tr>
<td>PBR (mean only)</td>
<td>0.1230</td>
<td>0.1274</td>
<td>0.3104</td>
</tr>
<tr>
<td>L1</td>
<td>0.1992</td>
<td>0.1581</td>
<td>0.3415</td>
</tr>
<tr>
<td>L2</td>
<td>0.1565</td>
<td>0.1469</td>
<td>0.2288</td>
</tr>
<tr>
<td>Mean-CVaR R=0.06</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAA</td>
<td>0.2909</td>
<td>0.3534</td>
<td>1.6833</td>
</tr>
<tr>
<td>PBR (CVaR only)</td>
<td>0.1918</td>
<td>0.2074</td>
<td>0.4071</td>
</tr>
<tr>
<td>PBR (mean only)</td>
<td>0.1364</td>
<td>0.1414</td>
<td>0.3100</td>
</tr>
<tr>
<td>L1</td>
<td>0.2263</td>
<td>0.1754</td>
<td>0.3498</td>
</tr>
<tr>
<td>L2</td>
<td>0.1842</td>
<td>0.1532</td>
<td>0.2407</td>
</tr>
<tr>
<td>Mean-CVaR R=0.08</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAA</td>
<td>0.2980</td>
<td>0.3615</td>
<td>1.6833</td>
</tr>
<tr>
<td>PBR (CVaR only)</td>
<td>0.2148</td>
<td>0.2242</td>
<td>0.4486</td>
</tr>
<tr>
<td>PBR (mean only)</td>
<td>0.1519</td>
<td>0.1575</td>
<td>0.3066</td>
</tr>
<tr>
<td>L1</td>
<td>0.3100</td>
<td>0.2395</td>
<td>0.3628</td>
</tr>
<tr>
<td>L2</td>
<td>0.2451</td>
<td>0.1835</td>
<td>0.2588</td>
</tr>
<tr>
<td>Global min. CVaR</td>
<td>0.2857</td>
<td>0.3534</td>
<td>1.6833</td>
</tr>
<tr>
<td>Equal</td>
<td>0.0427</td>
<td>0.0382</td>
<td>0.0483</td>
</tr>
</tbody>
</table>

This table reports the portfolio turnovers (defined in Eq. 18) 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).

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

$$\text{Cov}(S_{ij}, S_{kl}) = \frac{1}{(n-2)} \sum_{(p,q)} \frac{1}{2} (X^i_p - X^i_q)(X^j_p - X^j_q) - \sigma^2_{ij} \frac{1}{(r,s)} \frac{1}{2} (X^k_r - X^k_s)(X^l_r - X^l_s) - \sigma^2_{kl}$$

$$= \frac{1}{(n-2)^2} \left[ \sum_{(p,q)} \sum_{(r,s)} \left( \frac{1}{2} (X^i_p - X^i_q)(X^j_p - X^j_q) - \sigma^2_{ij} \right) \left( \frac{1}{2} (X^k_r - X^k_s)(X^l_r - X^l_s) - \sigma^2_{kl} \right) \right]$$

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

$$\frac{1}{2} (X^i_p - X^i_q)(X^j_p - X^j_q) - \sigma^2_{ij} \frac{1}{2} (X^k_r - X^k_s)(X^l_r - X^l_s) - \sigma^2_{kl}$$

Now

$$\frac{1}{2} (X^i_p - X^i_q)(X^j_p - X^j_q) - \sigma^2_{ij} \frac{1}{2} (X^k_r - X^k_s)(X^l_r - X^l_s) - \sigma^2_{kl}$$
Figure 4  The out-of-sample Sharpe ratios (annualized, with proportional transaction costs) for the strategies considered for the mean-variance problem, for three data sets. Detailed results are in Table 1.

\[
\frac{1}{4} \left[ X_p^t X_p^f - X_p^t X_p^f - X_p^t X_p^f \right] [X_k^p X_k^f - X_k^p X_k^f - X_k^p X_k^f] \\
+ \frac{1}{4} X_k^t X_k^f [X_k^p X_k^f - X_k^p X_k^f] \\
+ \frac{1}{4} X_k^t X_k^f [X_k^p X_k^f - X_k^p X_k^f] \\
- \frac{1}{2} \sigma_{ij}^2 [X_i^p X_i^f - X_i^p X_i^f + X_i^p X_i^f] \\
- \frac{1}{2} \sigma_{ij}^2 [X_j^p X_j^f - X_j^p X_j^f + X_j^p X_j^f] \\
- \frac{1}{2} \sigma_{kl}^2 [X_k^p X_k^f - X_k^p X_k^f + X_k^p X_k^f] + \sigma_{ij}^2 \sigma_{kl}^2
\]

Figure 5  The out-of-sample Sharpe ratios (annualized, with proportional transaction costs) for the strategies considered for the mean-CVaR problem, for three data sets. Detailed results are in Table 2.
so taking expectations and simplifying, we get
\[
\mathbb{E}[\frac{1}{2}(X^p - X^q)(X^p - X^q) - \sigma^2_{ij}] = \frac{1}{4}(\mu_{\text{ijkl}} - \sigma_{ij}^2\sigma_{kl}^2).
\]

3. \|(k,l) \cap (k',l')\| = 2, in other words, we have \((\frac{\|\|}{2})\) terms of the form
\[
[\frac{1}{2}(X^p - X^q)(X^p - X^q) - \sigma^2_{ij}] = \frac{1}{4}(\mu_{\text{ijkl}} - \sigma_{ij}^2\sigma_{kl}^2).
\]

Now
\[
[\frac{1}{2}(X^p - X^q)(X^p - X^q) - \sigma^2_{ij}] = \frac{1}{4}[X^p X^q X^p X^q + X^p X^q X^q X^p + X^q X^p X^q X^p + X^q X^q X^p X^q + X^q X^q X^q X^q]
\]
\[
\frac{1}{4}[X^p X^q X^q X^q + X^p X^q X^q X^q + X^q X^q X^q X^q + X^q X^q X^q X^q] + \frac{1}{4}[X^p X^q X^q X^q + X^p X^q X^q X^q + X^q X^q X^q X^q + X^q X^q X^q X^q]
\]
\[
- \frac{1}{2}\sigma^2_{ij}[X^p X^q X^q - X^q X^q - X^q X^q + X^q X^q] - \frac{1}{2}\sigma^2_{ij}[X^p X^q X^q - X^q X^q - X^q X^q + X^q X^q] + \sigma^2_{ij}\sigma_{kl}^2,
\]
so taking expectations and simplifying, we get
\[
\mathbb{E}[\frac{1}{2}(X^p - X^q)(X^p - X^q) - \sigma^2_{ij}] = \frac{1}{4}(\mu_{\text{ijkl}} - \sigma_{ij}^2\sigma_{kl}^2).
\]

Putting it all together,
\[
\text{Cov}(s^2_{ij}, s^2_{kl}) = \frac{1}{n}(\mu_{\text{ijkl}} - \sigma^2_{ij}\sigma^2_{kl}) + \frac{1}{n(n-1)}(\sigma^2_{ij}\sigma^2_{kl}).
\]

Appendix B: Variance of the CVaR estimator

Setting. Let \(L = [L_1, \ldots, L_n]\) 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.

In this section, we derive an expression for the variance of \(\hat{\text{CVaR}}_\alpha(L; \beta)\) introduced in Eq. (3) 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
\[
\hat{\text{CVaR}}_\alpha(L; \beta) := \min_{\alpha \in \mathbb{R}} (1 - \varepsilon_n)\alpha + \frac{1}{n - [n\beta] + 1} \sum_{i=1}^{n} (L_i - \alpha)^+,\]
where \(\varepsilon_n\) is some constant satisfying \(0 < \varepsilon_n < (n - [n\beta] + 1)^{-1}\) and \(\varepsilon_n = O(n^{-2})\).

We now show the minimizer in the definition of \(\hat{\text{CVaR}}_\alpha(L; \beta)\) is given by \(\alpha^* = L_{([n\beta])}\).

**Proposition 7.** The solution \(\alpha^* = L_{([n\beta])}\) is the unique minimizer in the one-dimensional optimization problem

\[
\min_{\alpha \in \mathbb{R}} \left\{ G_n(\alpha) := (1 - \varepsilon_n)\alpha + \frac{1}{n - [n\beta] + 1} \sum_{i=1}^{n} (L_i - \alpha)^+ \right\},
\]
where \(\varepsilon_n\) is some constant satisfying \(0 < \varepsilon_n < (n - [n\beta] + 1)^{-1}\) and \(\varepsilon_n = O(n^{-2})\).
Thus where the final equality is due to Proposition 7. The Type 2 CVaR estimator is an intuitive representation of CVaR because it is precisely the sample average

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

\[
\hat{\text{CV}}^2_n(L; \beta) := \frac{1}{n - \lceil n\beta \rceil + 1} \sum_{i=1}^{n} L_i 1(L_i \geq \hat{\alpha}_n(\beta)),
\]

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 section, 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 Lemma 4 then follows.

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

\[
\hat{\text{CV}}^2_n(L; \beta) = \text{CV}^1_n(L; \beta) + \varepsilon_n L_{(\lceil n\beta \rceil)}.
\]

**Proof.** Rewriting Type 2 CVaR estimator:

\[
\hat{\text{CV}}^2_n(L; \beta) = \frac{1}{n - \lceil n\beta \rceil + 1} \sum_{i=1}^{n} L_i 1(L_i \geq L_{(\lceil n\beta \rceil)})
\]

\[
= L_{(\lceil n\beta \rceil)} + \frac{1}{n - \lceil n\beta \rceil + 1} \sum_{i=1}^{n} (L_i - L_{(\lceil n\beta \rceil)}) 1(L_i \geq L_{(\lceil n\beta \rceil)})
\]

\[= \text{CV}^1_n(L; \beta) + \varepsilon_n L_{(\lceil n\beta \rceil)},\]

where the final equality is due to Proposition 7.
B.1. Proof of Proposition 4

Let \( \alpha_n^1 \) and \( \alpha_n \) be the minimizers in the definition of \( \hat{CVaR}_n(L; \beta) \) and \( CVaR_n(L; \beta) \) respectively. We can show, by elementary arguments,

\[
|CVaR_n(L; \beta) - \hat{CVaR}_n(L; \beta)| \leq \varepsilon_n (\alpha_n^1 \lor \alpha_n).
\]

Hence we have

\[
CVaR_n(L; \beta) = \hat{CVaR}_n(L; \beta) + O_p(\varepsilon_n)
\]

\[
= \frac{1}{n - \lfloor n\beta \rfloor + 1} \sum_{i=1}^n L_i 1(L_i \geq L_{(\lfloor n\beta \rfloor)}) + O_p(\varepsilon_n) \quad \text{by Lemma 1},
\]

which implies

\[
Var[CVaR_n(L; \beta)] = \frac{1}{n(1 - \beta)^2} Var[L_i 1(L_i \geq L_{(\lfloor n\beta \rfloor)})] + O(n^{-2}),
\]

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

Appendix C: Proof of Theorem 1

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

**Proposition 8.** Consider the optimization problem

\[
\min_{z \in \mathbb{R}^n} \quad z^\top 1_n
\]

\[
\text{s.t.} \quad z_i \geq 0 \quad \forall \ i
\]

\[
\quad z_i \geq c_i \quad \forall \ i
\]

\[
\quad z^\top \Omega_n z \leq f
\]

where \( c \in \mathbb{R}^n \) is some constant vector, \( f > 0 \) is a constant scalar and \( \Omega_n = (n-1)^{-1}(1_n - n^{-1}1_n1_n^\top) \), the sample covariance operator. Suppose (20) 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 \( V(z) := S_1(z) \cap S_2(z) = \{1 \leq i \leq n : z_i > \max(0, c_i)\} \). Then, at the optimal solution \( z^* \), we cannot have both \( S_1(z^*) \) and \( V(z^*) \) nonempty simultaneously.

**Proof.** Problem (20) is a convex optimization problem because \( \Omega_n \) is a positive semidefinite matrix. The problem is also strictly feasible, since \( z_0 = 2 \max\{c_i\}1_n \) is a strictly feasible point: clearly, \( z_0 > \max\{0, c_i\} \forall \ i \) and \( z_0^\top \Omega_n z_0 = 0 < f \) as \( 1_n \) is orthogonal to \( \Omega_n \). 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**:

\[
z^*_i \eta_1^* = 0 \quad \forall \ i, \quad (z^*_i - c_i) \eta_2^* = 0 \quad \forall \ i \quad \text{and} \quad \lambda^* [(z^*)^\top \Omega_n z^* - f] = 0
\]
First Order Condition:
\[ \nabla_z \mathcal{L} = 2\lambda \Omega_n z^* + (1_n - \eta_1^* - \eta_2^*) = 0 \]  
(21a)

By substituting for \( \Omega_n \), (21a) can be written as
\[ \frac{2\lambda}{n-1} \left( z^* - \frac{1}{n} (1_n^\top z^*) 1_n \right) = -1_n + \eta_1^* + \eta_2^*. \]  
(22)

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 (22) gives
\[ - \frac{2\lambda^*}{n(n - 1)} (1_n^\top z^*) = -1 + \eta_{1,i_0}^* + \eta_{2,i_0}^*. \]  
(23)

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 (22) gives
\[ - \frac{2\lambda^*}{n-1} \left( z_{j_0}^* - \frac{1}{n} (1_n^\top z^*) \right) = -1, \]  
(24)

which implies \( \lambda^* > 0 \) since \( \lambda^* \) cannot equal zero.

Now suppose \( S_1(z^*) \) and \( V(z^*) \) are both nonempty. Combining (23) and (24), we arrive at the necessary condition
\[ \frac{2\lambda^*}{n-1} z_{j_0}^* = -\eta_{1,j_0}^* - \eta_{2,j_0}^*, \]
which is clearly a contradiction since \( \text{lhs} > 0 \) whereas \( \text{rhs} \leq 0 \). Hence \( S_1(z^*) \) and \( V(z^*) \) cannot both be nonempty, and the result follows.

**Proof of Theorem 1**

Clearly, (cv-relax) is a relaxation of (cv-PBR'): the components of the variable \( z \) in (cv-relax) are relaxations of \( \max(0,-w^\top X_i - \alpha) \) for \( i = 1, \ldots, n \) 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 8, \( S_1(z^*) \), 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 \( \eta_2 \) in the dual problem (cv-relax-d):
\[ \max_{\eta_1 \geq 0} - (\eta_1 + \eta_2) \Omega_1^1 (\eta_1 + \eta_2). \]  
(25)

As \( \Omega_0 \) is orthogonal to \( \Omega_1 \), and \( \Omega_1^1 \) is positive semidefinite, the optimal solution is of the form \( \eta_1 = a 1_n - \eta_2 \), 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 \( \eta_1^\top 1_n = 1 \) 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 \( 0 < \delta \ll 1 \) in the dual problem (cv-relax-d).
Appendix D: Proofs of results in Sec. 5

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 \( m_{\theta} : X \mapsto \mathbb{R} \), where \( \theta \) is a parameter chosen from \( \Theta \), and \( X \) is a subset of the Euclidean space. We are interested in finding the parameter \( \theta^* \) 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 \( X_1, \ldots, X_n \), one can estimate \( \theta^* \) 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 \( M, M_n, \Theta, \theta^* \) 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 \( \theta^* \) 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 \( N[\cdot](\varepsilon, \mathcal{F}, L_2(P)) \) for every \( \varepsilon > 0 \). 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 \( \mathcal{F} = \{ f_{\theta} : \theta \in \Theta \}, f_{\theta} : X \to \mathbb{R} \), 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 \( \theta_0 \) for \( P \)-almost every \( x \) with derivative \( \dot{f}_{\theta_0}(x) \) and such that, for every \( \theta_1 \) and \( \theta_2 \) in a neighborhood of \( \theta_0 \), there exists a measurable function \( \dot{f} \) with \( \mathbb{E}[\dot{f}^2(X_i)] < \infty \) 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

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

\(^4\) 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, $\theta \mapsto m_\theta(x) = \alpha + (1 - \beta)^{-1}(-\alpha - w_1^T x - v^T L^+ x)^+$ is clearly differentiable at $\theta_{CV}$ for all $x \in \mathbb{R}^p$. Furthermore,

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

where $I(x) := I(-\alpha - w_1^T x - v^T L^+ x \geq 0)$, hence

$$\hat{m}(x) := \max(1, ||L^+ x||_\infty)$$

is an upper bound on $||\nabla_{\theta} m_\theta(x)||_\infty$ and is independent of $\theta$. Thus $|m_\theta(x) - m_\theta(x)| \leq \hat{m}(x)||\theta_1 - \theta_2||_2$ 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.

$\square$

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

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

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. $\square$

**Proof of Theorem 4.** **Case I:** $M_n(\theta, \lambda_1, \lambda_2)$ equal to (14).

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

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

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 (15).

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

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

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 (16).
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:

\[
\| \frac{\lambda_1}{n} w^T \hat{\Sigma}_n w + \frac{\lambda_2}{(n-1)(1-\beta)^2} \sum_{i=1}^{n} \left( z_\theta(X_i) - \frac{1}{n} \sum_{j=1}^{n} z_\theta(X_j) \right)^2 \| \leq \frac{\lambda_1}{n} \| w^T \hat{\Sigma}_n w \| + \frac{\lambda_2}{n(n-1)(1-\beta)^2} \left( \sum_{i=1}^{n} \left( z_\theta(X_i) - \frac{1}{n} \sum_{j=1}^{n} z_\theta(X_j) \right)^2 \right) ^{1/2} = O \left( \frac{1}{n} \right),
\]

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

References


