Performance-based regularization in mean-CVaR portfolio optimization

Gah-Yi Vahn
gvahn@london.edu

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

Andrew E.B. Lim
Department of Decision Sciences & Department of Finance, NUS Business School, Singapore, 119245, and
Department of Industrial of Engineering & Operations Research, University of California, Berkeley, CA 94720.
andrewlim@nus.edu.sg, lim@ieor.berkeley.edu

Regularization is a technique widely used to improve the stability of solutions to statistical problems. We propose a new regularization concept, performance-based regularization (PBR), for data-driven stochastic optimization. The goal is to improve upon Sample Average Approximation (SAA) in finite-sample performance while maintaining minimal assumptions about the data. We apply PBR to mean-CVaR portfolio optimization, where we penalize portfolios with large variability in the constraint and objective estimations, which constrains the probabilities that the estimations deviate from the respective true values. This results in a combinatorial optimization problem, but we prove its convex relaxation is tight. We prove PBR is asymptotically optimal, and derive its first-order behavior by extending the theory of M-estimators. To calibrate the constraint right-hand side, we develop a new, performance-based $k$-fold cross-validation algorithm. An extensive empirical investigation demonstrates that PBR can improve upon SAA and standard regularization methods in the out-of-sample Sharpe ratio with statistical significance.

Key words: portfolio optimization; sample average approximation, Conditional Value-at-Risk; regularization, stochastic optimization


1. Introduction

The method of regularization was introduced in the 1960s to deal with ill-posed linear 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 $R(x) = ||Ax - b_\delta||^2$ 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 been adopted in many areas of science. In particular it is widely used in statistical problems of classification, regression and density estimation. The reader may be most familiar with its usage in regression as

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

where $y = [y_1, \ldots, y_n] \in \mathbb{R}^n$ is the data on the observable, $X = [X_1, \ldots, X_n] \in \mathbb{R}^{n \times p}$ is the vector of covariates, $\lambda > 0$ is a pre-determined scalar and $\beta \in \mathbb{R}^p$ is the regression coefficient that best fits the linear model $y = X\beta$. Ridge regression corresponds to using the Tikhonov regularization function $P(\beta) = ||\beta||_2$ and Lasso regression corresponds to setting $P(\beta) = ||\beta||_1$. The purpose of the Tikhonov regularization is to deal with the case when the data matrix $X$ is ill-conditioned or singular, which poses a problem as the regression solution requires the inverse of $X^\top X$. On the other hand, the purpose of Lasso regression is to result in a sparse coefficient vector, which is desirable when the size of the feature space is large (which is the case with “big data”) and a sparse model with high interpretability is needed.

Unlike statistical problems, stochastic optimization problems of interest to the Operations Research/Management Science community are performance-driven, where the goal is to achieve a good performance of some economically/operationally meaningful objective. When the decision maker is endowed with finite historical data, a common approach to finding a decision is to optimize a Sample Average Approximation (SAA) of the problem. As is the case with ill-posed linear operator problems (including important problems in statistics), the solution to the SAA approach can be highly unstable and result in highly unreliable out-of-sample performance. Perhaps the best known example of this is the data-driven portfolio optimization problem. The portfolio optimization problem was first devised by Markowitz (1952), and has garnered much attention since, leading to a conferral of a Nobel Prize to Markowitz in 1990. However, the adoption of Markowitz’ mean-variance model has been limited in practice because errors associated with estimating the mean and the variance render the resulting portfolio allocation highly unreliable. A series of papers have documented various aspects of this issue [Frankfurter et al. (1971), Frost and Savarino (1986, 1988b), Michaud (1989), Best and Grauer (1991), Chopra and Ziemba (1993), Broadie (1993)].
We thus propose a new regularization concept, performance-based regularization (PBR), whose purpose is to improve upon the self-evident SAA approach to stochastic optimization when one has a limited number of stationary observations to solve the problem with. Performance-based regularization is philosophically different from Tikhonov regularization (whose purpose is stability of the solution) and Lasso regularization (whose purpose is sparsity) but is natural to optimization problems where the ultimate goal is the out-of-sample performance of the decision made with a finite amount of data.

In this paper, we illustrate the PBR approach for a contemporary version of the Markowitz problem, the data-driven mean-Conditional Value-at-Risk (CVaR) problem. We focus on this problem rather than the classical one, for the following reasons. Firstly, CVaR has received much attention in recent years as a financial risk measure. Intuitively, CVaR is a more sensible measure than variance because CVaR is a one-sided statistic whereas variance is two-sided. For example, a portfolio position associated with a large upside variability with low downside variability would incorrectly be deemed as being risky if one uses variance as the risk measure, whereas this would not be the case if CVaR is used as the risk measure. From a theoretical perspective, CVaR is desirable because it satisfies the four coherence axioms of [Artzner et al. (1999), Acerbi and Tasche (2001)], whereas variance (and Value-at-Risk) fails to do so. In addition, we were motivated to improve upon the data-driven mean-CVaR problem over the Markowitz one because the work of Lim et al. (2011) demonstrated that the finite-sample performance issue is more problematic for the mean-CVaR problem than for the Markowitz problem, hence making it a more interesting problem to study in detail. We do, however, leave the investigation of PBR for the Markowitz problem for future work.

We make four major contributions in this paper. First, we introduce a new notion of regularization, whose purpose is to improve the finite-sample performance of data-driven stochastic optimization. This is an important conceptual development that extends the classical notion of regularization to make it relevant to problems of interest to the Operations Research/Management Science community. Classical regularization techniques do not address finite-sample performance of the solution; they were developed to increase the numerical stability or sparsity of the solution itself. Additionally, most studies of the SAA approach in stochastic optimization focus primarily on establishing asymptotic optimality (consistency) of the method and rarely evaluate its finite-sample performance. This is despite the fact that in practice, it is unlikely that one has access to a large number of stationary data to achieve optimality with high probability. As far as we are aware, this is the first systematic attempt to improve upon the finite-sample performance of SAA.

Second, we demonstrate PBR in detail for the mean-CVaR portfolio optimization problem, a problem for which the fragility of the SAA solution in finite samples has been well documented [Lim
et al. (2011)]. In particular, we add regularization constraints to the SAA problem that capture the variability in the estimation of the portfolio mean and CVaR. This has a probabilistic interpretation in that the model discards solutions whose mean and CVaR estimations deviate too much from their population values with high probability. The PBR model of the mean-CVaR problem is a combinatorial optimization problem, but we prove its convex relaxation, a quadratically constrained quadratic program (QCQP), is tight, hence can be efficiently solved. We further show that the PBR model is distinctly different from chance optimization, the greatest difference being that PBR is a nonparametric method whereas chance optimization is parametric.

Third, we establish that the PBR solution is asymptotically optimal and characterize its first-order behavior. Although PBR is developed to improve upon the finite-sample performance, asymptotic optimality is an important requirement as this guarantees the convergence of the solution as the number of stationary observation increases. In fact, asymptotic optimality is the minimal requirement for any data-driven decision paradigm as a method that fails to satisfy this would be nonsensical. An analytic characterization of the first-order behavior is also important as this leads us to universal insights. An analysis of the PBR solution shows that the first-order effect of the regularization function is through its Hajek projection; this has the practical implication that one does not need to consider two different regularization functions that have the same Hajek projection. The analysis requires a non-trivial extension of the theory of M-estimators as the regularization constraints introduce sample averages of permutation-symmetric functions that are not iid. To the best of our knowledge this is the first paper to derive the first-order behavior of a sampling-based stochastic optimization problem where the objective consists of sample average of non-iid terms.

Lastly, we make an extensive empirical study of the PBR method against SAA as well as standard regularization methods on the five, ten and forty-nine industry portfolios from Ken French’s website. 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 right-hand sides (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. Using the new calibration algorithm, we compare the PBR method against five different benchmarks: (i) SAA (ii) $L_1$ regularization (iii) $L_2$ regularization (iv) equally-weighted portfolio (v) global minimum CVaR portfolio, where the mean constraint is removed from the problem, as well as against the PBR method applied only to the objective and only on the mean constraint. We find that for the five and ten industry portfolios, the PBR method improves upon SAA in terms of the out-of-sample Sharpe ratio (annualized, with proportional transaction costs) by 1.8% and 11.2% with statistical
significance at 5% and 10% respectively. We also find that the PBR method dominates all the other strategies, and that regularizing the objective has a greater effect than regularizing the mean constraint. 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. Comparing across data sets, however, we find that the highest Sharpe ratio is achieved by the PBR method on the five industry portfolio.

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 modern illustration of the technique we refer the reader to Hastie et al. (2009) and references therein.

In this work we propose a new regularization method to improve upon the SAA approach to stochastic optimization in terms of the finite-sample performance. The limitations of the SAA approach in finite samples are rarely discussed in the stochastic optimization literature, however, perhaps because the SAA approach makes use of data in an intuitive manner (by replacing expected values in the optimization with sample averages) and because one can show, under fairly general conditions, that the SAA solution is consistent, i.e. asymptotically optimal. For a discussion of the main results and implications of the SAA approach, we refer the reader to Shapiro et al. (2009) and references therein.

On the other hand, the limitations of the SAA approach in portfolio optimization are well-known in finance, although the terminology “SAA” tends not to appear explicitly. Frankfurter et al. (1971) is perhaps the first to identify the limitations of the SAA approach to the classical Markowitz (mean-variance) portfolio optimization problem. There have since been many suggestions for mitigating this issue for the Markowitz problem; two main approaches are robust optimization [Goldfarb and Iyengar (2003)] and standard regularization [Chopra (1993), Frost and Savarino (1988a), Jagannathan and Ma (2003), DeMiguel et al. (2009a)]. The robust optimization approach takes the source of uncertainty (e.g. the asset returns, or its distribution), specifies an uncertainty set about the source, and minimizes the worst-case return-risk problem over this uncertainty set. The standard regularization approach solves the SAA Markowitz problem, but with a constraint on the size of the solution as measured by the $L_1$ or $L_2$ norm, which parallels Tikhonov and Lasso regularization discussed in the introduction. Both robust optimization and standard regularization approaches have been studied for the mean-CVaR problem; Gotoh and Shinozaki (2010) and Zhu and Fukushima (2009) show implementations of the robust optimization approach when the source
of uncertainty is, respectively, the return vector and the return distribution, and Gotô and Takeda (2010) demonstrates implementation of standard regularization.

**Notations.** Throughout the paper, we denote convergence in probability by \( P \xrightarrow{} \) and in distribution by \( \Rightarrow \). The notation \( X \overset{d}= Y \) for two random variables \( X \) and \( Y \) means they have the same distribution, and the symbol \( X \sim D \) is used to indicate that the random variable \( X \) follows some standard distribution \( D \).

3. Fragility of SAA in mean-CVaR portfolio optimization

An investor is to choose a portfolio \( w \in \mathbb{R}^p \) on \( p \) different assets. Her wealth is normalized to 1, so \( w^\top 1_p = 1 \), where \( 1_p \) denotes \( p \times 1 \) vector of ones. The returns of the \( p \) assets are denoted by \( X \), a \( p \times 1 \) random vector, which follows some absolutely continuous distribution \( F \) with twice continuously differentiable probability density function (pdf) and finite mean \( \mu \) and covariance \( \Sigma \). The investor wants to pick a portfolio that minimizes the CVaR of the portfolio loss at level \( 100(1 - \beta)\% \), for some \( \beta \in (0, 1) \), while reaching an expected return \( R \). That is, she wants to solve the following problem:

\[
\begin{align*}
\hat{w}_n &= \arg\min_{w} \text{CVaR}(\mathbf{-w}^\top X; \beta) \\
\text{s.t.} \quad w^\top \mu &= R \\
&\quad w^\top 1_p = 1,
\end{align*}
\]

(\text{CVaR-pop})

where

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

as in Rockafellar and Uryasev (2000).

In reality, the investor does not know the distribution \( F \). We assume the investor observes \( n \) iid realizations of asset returns, \( X = [X_1, \ldots, X_n] \in \mathbb{R}^{p \times n} \). Then the most straight-forward approach is to solve the following Sample Average Approximation (SAA) problem:

\[
\hat{w}_n = \arg\min_{w} \text{CVaR}_n(\mathbf{-w}^\top X; \beta) \\
\text{s.t.} \quad w^\top \hat{\mu}_n &= R \\
&\quad w^\top 1_p = 1,
\]

(\text{CVaR-SAA})

where \( \hat{\mu}_n = n^{-1} \sum_{i=1}^{n} X_i \) is the sample mean of the observed returns and

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

asymptotically, as the number of observations \( n \) goes to infinity, \( \hat{w}_n \) converges in probability to \( w_0 \) [see Sec. 5 for details]. In practice, however, the investor has a limited number of relevant observations. If, for example, there are \( n = 250 \) iid daily observations, and the investor wishes to control the top 5% of the losses, then there are only \( 250 \times 0.05 = 12.5 \) points to estimate the portfolio.
CVaR at level $\beta = 0.95$. For stock returns, $n = 250$ iid daily observations is rather optimistic; there is ample empirical evidence that suggests daily returns are non-stationary over this period of time. Even for time scales with more evidence for stationarity (e.g. bi-weekly/monthly), the property tends to last for no more than 5 years. See McNeil et al. (2005) and references therein for a discussion of stationarity in stock return data.

As a result, solving (CVaR-SAA) using real data results in highly unreliable solutions. Let us illustrate this point, assuming an ideal market scenario. There are $p = 10$ stocks, with daily returns following a Gaussian distribution:

$$X \sim \mathcal{N}(\mu_{\text{sim}}, \Sigma_{\text{sim}})$$

and the investor has $n = 250$ iid observations of $X$. In the following, we conduct an experiment similar to those found in Lim et al. (2011), to evaluate the performance and reliability of solving (CVaR-SAA) under this ideal scenario. Briefly, the experimental procedure is as follows:

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

The result of the experiment is summarized in Fig. (1). The green curve corresponds to the population efficient frontier. Each of the grey dots corresponds to a solution instance of (CVaR-SAA). There are two noteworthy observations: the solutions $\hat{w}_n$ 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%.

4. Performance-based regularization

4.1. Performance-based regularization

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

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

s.t. $w^\top \hat{\mu}_n = R$

$$w^\top 1_p = 1$$

$$r_1(n) P_1(w) \leq U_1$$

$$r_2(n) P_2(w) \leq U_2$$

where $P_1, P_2 : \mathbb{R}^p \to \mathbb{R}$ are regularization functions that characterize the uncertainty associated with $CVaR_n(-w^\top X; \beta)$ and $w^\top \hat{\mu}_n$, respectively, and where $r_1(n), r_2(n)$ control the decay of the regularization constraints. The idea is to regularize decisions $w$ for which the uncertainty about

\footnote{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}
the true values $w^\top \mu$ and $CVaR(-w^\top X; \beta)$ is large, while ensuring the solution still be consistent (asymptotically optimal) as the number of observations $n$ tends to infinity.

What, then, are appropriate regularization functions? Recall that we are trying to find solutions that yield efficient frontiers that are closer to the population efficient frontier. Thus the variances of $w^\top \hat{\mu}_n$ and $\hat{CVaR}_n(-w^\top X; \beta)$ make appropriate regularization functions, as they characterize the deviation from the respective population values. The variance of $w^\top \hat{\mu}_n$ is given by

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

and an approximate expression for the variance of $\hat{CVaR}_n(-w^\top X; \beta)$ is given by the following proposition.

PROPOSITION 1. Suppose $X = [X_1, \ldots, X_n]^{iid} \sim F$, where $F$ is absolutely continuous with twice continuously differentiable pdf. Then

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

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 A.
Of course, we do not know the true variances, so we contend with sample variances of the estimators \( w^\top \hat{\mu}_n \) and \( \hat{CVaR}_n(-w^\top X; \beta) \). That is, we consider the following regularization functions:

\[
\begin{align*}
\mathcal{P}_1(w) &= \frac{1}{n} w^\top \hat{\Sigma}_n w, \quad \text{where } \hat{\Sigma}_n = \text{Cov}(X), \\
\mathcal{P}_2(w) &= \frac{1}{n(1 - \beta)^2} z^\top \Omega_n z, \quad \text{where} \\
\Omega_n &= \frac{1}{n-1} [I_n - n^{-1} 1_n 1_n^\top], \quad I_n = n \times n \text{ identity matrix}, \quad \text{and} \\
z_i &= \max(0, -w^\top X_i - \alpha) \text{ for } i = 1, \ldots, n,
\end{align*}
\]

For the rest of this paper, we investigate the PBR method with the above choice of regularization functions. The motivation behind this choice is intuitive and straightforward: for a fixed portfolio \( w \), the point estimates \( \hat{CVaR}_n(-w^\top X; \beta) \) and \( w^\top \hat{\mu}_n \) have confidence intervals around them which are approximately equal to the respective sample standard deviations. Now as \( w \) varies, the error associated with the point estimate varies, as the confidence interval is a function of \( w \). The PBR constraints dictates that any solution \( w \) that is associated with a large estimation error of the objective or the mean constraint be removed from consideration, which is reasonable since such a solution would be highly unreliable. In other words, imposing the PBR constraints squeezes the SAA problem (CVaR-SAA) closer to the true problem (CVaR-pop) in a probabilistic sense. A schematic of the PBR model is shown in Fig. 2.

We can further give a probabilistic interpretation of regularization constraints. By Chebyshev’s inequality, the sample variance regularization constraint is an upper bound on the probability that the estimated values deviate from the true value by more than some level \( t \):

\[
\frac{1}{n} w^\top \Sigma w \leq t^2 \epsilon \\
\implies P(|w^\top \hat{\mu}_n - w^\top \mu| > t) \leq \epsilon
\]

That is, for a fixed level \( \epsilon \) and \( n \), there is a one-to-one mapping between the parameter \( U_1/r_1(n) \) of the regularization constraint \( r_1(n) w^\top \hat{\Sigma}_n w/n \leq U_1 \) and the parameter \( t \) of the probabilistic constraint. The sample variance regularization on \( \hat{CVaR}_n(-w^\top X; \beta) \) has a similar interpretation.

The PBR problem with sample variance penalty functions is:

\[
(\hat{\alpha}_n^r, \hat{w}_n^r, z_n^r) = \arg\min_{\alpha, w, z} \alpha + \frac{1}{n(1 - \beta)} \sum_{i=1}^{n} z_i \\
\text{s.t. } w^\top \hat{\mu}_n = R \\

\begin{align*}
\frac{r_2(n)}{n(1 - \beta)^2} z^\top \Omega_n z &\leq U_2 \\
\frac{r_1(n)}{n} w^\top \hat{\Sigma}_n w &\leq U_1 \\
z_i &= \max(0, -w^\top X_i - \alpha), \quad i = 1, \ldots, n.
\end{align*}
\]
Figure 2  A schematic of PBR on the objective only. The objective function estimated with data is associated with an error (indicated in grey), which depends on the position in the solution space. The PBR constraint cuts out solutions which are associated with large estimation errors of the objective.

(CVaR-PBR) is clearly non-convex due to the cutoff variables $z_i = \max(0, -w^TX_i - \alpha)$, $i = 1, \ldots, n$. Without the regularization constraint on $z$, one can solve the problem by relaxing the non-convex constraint to $z_i \geq 0$ and $z_i \geq -w^TX_i - \alpha$. However, $z^T\Omega_n z$ is not a monotone function of $z$ hence it is not clear at the outset whether one can employ the same relaxation trick for the regularized problem.

(CVaR-PBR) is a combinatorial optimization problem because one can solve it by considering all possible combinations of $\lfloor n(1 - \beta) \rfloor$ out of $n$ observations that contribute to the worst $(1 - \beta)$ of the portfolio loss (which determines the non-zero elements of $z$), then finding the portfolio weights that solve the problem based on these observations alone. Clearly, this is an impractical strategy; for example, there are 34220 possible combinations to consider for a modest number of observations $n = 60$ (5 years of monthly data) and $\beta = 0.95$. However, it turns out that relaxing $z_i = \max(0, -w^TX_i - \alpha)$, $i = 1, \ldots, n$ as before results in a tight convex relaxation. The resulting problem is a quadratically-constrained quadratic program (QCQP) which can be solved efficiently.

Before stating the result, let us first introduce the convex relaxation of (CVaR-PBR):

$$\min_{\alpha, w, z} \quad \alpha + \frac{1}{n(1 - \beta)} \sum_{i=1}^{n} z_i$$
\begin{align*}
s.t. \quad \begin{array}{l}
  w^\top \hat{\mu}_n = R \quad (\nu_1) \\
  w^\top 1_p = 1 \quad (\nu_2) \\
  \frac{r_1(n)}{n} w^\top \hat{\Sigma}_n w \leq U_1 \quad (\lambda_1) \\
  \frac{r_2(n)}{n(1 - \beta)} z^\top \Omega_n z \leq U_2 \quad (\lambda_2) \\
  z_i \geq 0, i = 1, \ldots, n \quad (\eta_1) \\
  z_i \geq -w^\top X_i - \alpha, i = 1, \ldots, n \quad (\eta_2)
\end{array}
\tag{CVaR-relax}
\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^\top 1_n = 1 \\
& \lambda_1 \geq 0, \lambda_2 \geq 0 \\
& \eta_1 \geq 0, \eta_2 \geq 0
\tag{CVaR-relax-d}
\end{align*}

where

\[
g(\nu_1, \nu_2, \lambda_1, \lambda_2, \eta_1, \eta_2) = -\frac{n/r_1(n)}{2\lambda_1} (\nu_1 \hat{\mu}_n + \nu_2 1_p - X\eta_2)^\top \hat{\Sigma}_n^{-1} (\nu_1 \hat{\mu}_n + \nu_2 1_p - X\eta_2) \\
-\frac{n(1 - \beta)^2/r_2(n)}{2\lambda_2} (\eta_1 + \eta_2)^\top \Omega_n^\dagger (\eta_1 + \eta_2) + R\nu_1 + \nu_2 - U_1 \lambda_1 - U_2 \lambda_2,
\]

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

We now state the main result for this section — that (CVaR-relax) 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 (CVaR-relax) and (CVaR-relax-d). If \(\eta_2^* \neq 1_n/n\), then \((\alpha^*, w^*, z^*)\) is an optimal point of (CVaR-PBR). Otherwise, if \(\eta_2^* = 1_n/n\), we can find the optimal solution to (CVaR-PBR) by solving (CVaR-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 B.

**Remark.** Theorem 1 shows that one can solve (CVaR-PBR) via at most two steps. The first step is to solve (CVaR-relax); 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 (CVaR-dual) with an additional constraint, otherwise the relaxed solution is feasible for the original problem hence optimal. For the record, all problem instances solved in this paper were solved in a single step.

**4.2. Comparison with Chance Optimization**

The PBR approach to dealing with data uncertainty, especially with its probabilistic interpretation, is perhaps reminiscent of chance optimization [see Chapter 4 of Shapiro et al. (2009) for an overview of this topic]. The two approaches are fundamentally different, however, and here we make the
distinction clear. The chance optimization approach to the stochastic problem (CVaR-pop) is to first standardize it
\[
\min_w c \\
\text{s.t. } \bar{\text{CVaR}}_n(-w^\top X; \beta) \leq c \quad (a) \\
w^\top \hat{\mu}_n = R \quad (b) \\
w^\top 1_p = 1, \quad (c)
\]
then require the random constraints (a) and (b) to hold with at least some pre-specified probability. Thus the chance optimization model is given by
\[
\min_w c \\
\text{s.t. } P(\bar{\text{CVaR}}_n(-w^\top X; \beta) \leq c; \ w^\top \hat{\mu}_n \geq R) \geq t \\
w^\top 1_p = 1, \quad (\text{Chance-1})
\]
if one considers the random constraints jointly, otherwise
\[
\min_w c \\
\text{s.t. } P(\bar{\text{CVaR}}_n(-w^\top X; \beta) \leq c) \geq t_1 \\
P(w^\top \hat{\mu}_n \geq R) \geq t_2 \\
w^\top 1_p = 1, \quad (\text{Chance-2})
\]
if one considers the random constraint separately, where \(0 < t, t_1, t_2 < 1\) are pre-specified levels.

PBR is fundamentally different from either chance optimization models (Chance-1) or (Chance-2) because the effect of PBR is to makes any randomness in the problem closer to its population value with high probability, whereas chance optimization requires the random constraints themselves be satisfied with high probability. In addition, PBR is a non-parametric method in that no distributional assumptions are made about the uncertainty; we are simply penalizing the sample variance of the random estimators in the problem. The probabilistic interpretation persists regardless of the population distribution because Chebyshev’s inequality relates the probabilistic difference between the estimated and population values as a function of the variance. On the other hand, one must make distributional assumptions about the data in order to solve the chance optimization models (Chance-1) or (Chance-2)\(^3\).

5. Consistency and first-order asymptotics of PBR solution
In this section, we characterize the theoretical behavior of the PBR solution \(\hat{w}_n^r\). We show that \(\hat{w}_n^r\), just as the SAA solution \(\hat{w}_n\), is asymptotically optimal (consistent), i.e. converges in probability to \(w_0\) as the number of observations \(n\) goes to infinity, with appropriate decaying of the regularization constraints. More importantly, we show that the performance of \(\hat{w}_n^r\) converges to the true performance of \(w_0\), in that the return-risk frontier corresponding to \(\hat{w}_n^r\) converges to the population

\(^3\)Even then, the resulting problem is more often than not computationally intractable, and further simplifications and/or approximations may be needed.
efficient frontier. Furthermore, we derive the first-order behavior of $\hat{w}_n^\epsilon$ by a non-trivial extension of asymptotic analysis of M-estimators, which leads to the insight that the first-order effect of the regularization functions is through their Hajek projections.

**Notations.** We make use of stochastic little-o and big-O notations: for a given sequence of random variables $R_n$, $X_n \equiv o_P(R_n)$ means $X_n = Y_n R_n$ where $Y_n \xrightarrow{P} 0$, and $X_n = O_P(R_n)$ means $X_n = Y_n R_n$ where $Y_n = O_P(1)$, i.e. for every $\varepsilon > 0$ there exists a constant $M$ such that $\sup_n P(|Y_n| > M) < \varepsilon$.

**Measurability Issues.** We also encounter quantities that may not be measurable (e.g. supremum over uncountable families of measurable functions). We note that whenever the “probability” of such quantities are written down, we actually mean the outer probability. For further details, see Appendix C of Pollard (1984).

For ease of exposition and analysis, we derive theoretical results for the PBR problem transformed to a global optimization problem. Although we could have formulated the mean-CVaR problem as a global optimization problem from the beginning, we chose to keep to the formulation of Rockafellar and Uryasev (2000), which apart from being more familiar in the literature is more amenable for computation (e.g. in choosing the degree of regularization $U_1$ and $U_2$). In this section we work with the globalized formulation which is more amenable to analysis and leads to important theoretical insights regardless. We obtain the global problem by dualizing the mean and regularization constraints and eliminating the non-random constraint $w^\top 1_p = 1$:

$$\hat{\theta}_n(\lambda_1, \lambda_2) = \min_{\theta = (\alpha, \nu) \in \mathbb{R} \times \mathbb{R}^{p-1}} M_n(\theta; \lambda_1, \lambda_2)$$

(CVaR-dual)

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

$$M_n(\theta; \lambda_1, \lambda_2) = \frac{1}{n} \sum_{i=1}^n m_\theta(X_i) + \frac{r_1(n) \lambda_1}{n} w^\top \Sigma_n w + \frac{r_2(n) \lambda_2}{(n-1)(1-\beta)^2} \sum_{i=1}^n \left( z_\theta(X_i) - \frac{1}{n} \sum_{j=1}^n z_\theta(X_j) \right)^2,$$

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

where $z_\theta(x) = \max(0, -w^\top x - \alpha)$ and $\lambda_0 > 0, \lambda_1, \lambda_2 \geq 0$ are pre-determined constants.

The corresponding population problem is

$$\theta_0(\lambda_1, \lambda_2) = \arg\min_{\theta \in \mathbb{R}^p} M(\theta; \lambda_1, \lambda_2) = E[M_n(\theta; \lambda_1, \lambda_2)].$$

(5)

In what follows, we assume $M(\theta; \lambda_1, \lambda_2)$ has a unique minimizer $\theta_0(\lambda_1, \lambda_2)$, and that $\hat{\theta}_n(\lambda_1, \lambda_2)$ is a near-minimizer of $M_n(\theta; \lambda_1, \lambda_2)$, i.e.

$$M_n(\hat{\theta}_n; \lambda_1, \lambda_2) < \inf_{\theta \in \mathbb{R}^p} M_n(\theta; \lambda_1, \lambda_2) + o_P(1).$$

(6)
Before we proceed, we make a note that the objective Eq. (3) can be reformulated as a U-process. Let us first define empirical process and U-process.

**Definition 1.** Let $X_1, \ldots, X_n$ be iid random vectors from $\mathcal{X}$. For a measurable function $f : \mathcal{X} \to \mathbb{R}$, the *empirical process at $f$* is

$$G_n f := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [f(X_i) - \mathbb{E}f(X_1)],$$

and for a measurable function $g : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, the *U-process at $g$* is

$$U_n g := \frac{\sqrt{n}}{\binom{n}{2}} \sum_{i \neq j} [g(X_i, X_j) - \mathbb{E}X_1, X_2 g(X_1, X_2)].$$

Observe that a U-process is the sample average of identically distributed but non-independent, permutation-symmetric functions. We can rewrite the objective of Eq. (3) as a U-process according to

$$M_n(\theta; \lambda_1, \lambda_2) = \frac{1}{\binom{n}{2}} \sum_{1 \leq i, j \leq n, i \neq j} m^U_{(\theta, \lambda_1, \lambda_2)}(X_i, X_j),$$

where

$$m^U_{(\theta, \lambda_1, \lambda_2)}(x_i, x_j) := \frac{1}{2} [m_{\theta}(x_i) + m_{\theta}(x_j)] + \lambda_1 \frac{1}{2} [(w_1 + L^T z) (x_i - x_j)]^2 + \lambda_2 \frac{1}{2} [(w_1 + L^T z) (x_i - x_j)]^2. \quad (8)$$

This is a key observation as we draw upon literature in U-processes for our results in this section.

**5.1. Consistency of the PBR solution and its performance**

Let us now prove consistency of $\hat{\theta}_n(\lambda_1, \lambda_2)$ for fixed $\lambda_1, \lambda_2 \geq 0$. The intuition behind the proof is as follows: if $M(\theta; \lambda_1, \lambda_2)$ is well-behaved such that for every $\varepsilon > 0$ there exists $\eta > 0$ such that $\|\theta_n(\lambda_1, \lambda_2) - \theta_0(\lambda_1, \lambda_2)\| > \varepsilon \implies M(\hat{\theta}_n; \lambda_1, \lambda_2) - M(\theta_0; \lambda_1, \lambda_2) > \eta$, then consistency follows from showing that the probability of the event $\{M(\hat{\theta}_n; \lambda_1, \lambda_2) - M(\theta_0; \lambda_1, \lambda_2) > \eta\}$ goes to zero for all $\varepsilon > 0$. In the proof, we show that

$$0 \leq M(\hat{\theta}_n; \lambda_1, \lambda_2) - M(\theta_0; \lambda_1, \lambda_2) \leq - M_n(\hat{\theta}_n; \lambda_1, \lambda_2) - M(\theta_0; \lambda_1, \lambda_2) + o_P(1),$$

hence the result follows by showing Uniform Law of Large Numbers (ULLN) for $M_n(\theta; \lambda_1, \lambda_2)$:

$$\sup_{\theta \in \mathbb{R}^p} |M_n(\theta; \lambda_1, \lambda_2) - M(\theta; \lambda_1, \lambda_2)| \xrightarrow{P} 0. \quad (9)$$

Now if the domain of $\theta$, $\mathbb{R}^p$, was replaced by a finite set, then WLLN goes through (in fact, Strong Law of Large Numbers also) and Eq. (9) holds. Thus we could consider approximating the class of functions $\mathcal{F} = \{m^U_{(\theta, \lambda_1, \lambda_2)} : \theta \in \mathbb{R}^p\}$ by a finite class of functions $\mathcal{F}_\varepsilon$ such that every $f_\theta \in \mathcal{F}$ is within $\varepsilon$ distance away from some function in $\mathcal{F}_\varepsilon$, for every $\varepsilon > 0$. Since the ULLN applies for
the finite class \( \mathcal{F}_\varepsilon \), we can conclude ULLN for the original class \( \mathcal{F} \) if the approximation error can be made arbitrarily small for every \( \varepsilon > 0 \).

ULLN has been extensively studied in the statistics and empirical processes literature and one of the standard approaches to showing ULLN is through bracketing numbers. Given two functions \( l, u \), the bracket \([l, u]\) is the set of all functions \( g \) with \( l \leq g \leq u \). An \( \varepsilon \)-bracket in \( L_r(P) \) is a bracket \([l, u]\) with \( \mathbb{E}_P(u - l)^r < \varepsilon^r \), and the bracketing number \( N_1(\varepsilon, \mathcal{F}, L_r(P)) \) is the minimum number of \( \varepsilon \)-brackets needed to cover \( \mathcal{F} \). Having a finite bracketing number \( N_1(\varepsilon, \mathcal{F}, L_r(P)) < \infty \) for every \( \varepsilon > 0 \) means one can find a finite approximation to \( \mathcal{F} \) with \( \varepsilon \)-accuracy for all \( \varepsilon > 0 \), and ULLN holds for such \( \mathcal{F} \) [Theorem 19.4 Van der Vaart (2000)].

There are certainly known sufficient conditions for finite bracketing numbers. For our problem, if we can replace \( \mathbb{R}^p \) with a compact set, we can show \( \mathcal{F} \) is a Lipschitz class of functions (defined in the next paragraph), which is known to have finite \( N_1(\varepsilon, \mathcal{F}, L_r(P)) \) for every \( \varepsilon > 0 \). Now for all practical purposes, we need only consider a compact subset of \( \mathbb{R}^p \), \([-K, K]^p \) where \( K \) is appropriately large enough, because the elements of \( \theta = (\alpha, v) \) are only meaningful if bounded in size (\( \alpha \) is the Value-at-Risk of the portfolio \( w = w_1 + L v \)). Hence for the rest of this section we assume a \( K \) exists such that \( \hat{\theta}_n \in [-K, K]^p \) for all \( n \) and \( \theta_0 \in [-K, K]^p \).

**Definition 2 (Lipschitz Class).** Consider a class of measurable functions \( \mathcal{F} = \{ f_\theta : \theta \in \Theta \} \), \( f_\theta : \mathcal{X} \to \mathbb{R} \), under some probability measure \( P \). We say \( \mathcal{F} \) is a Lipschitz class about \( \theta_0 \in \Theta \) if \( \theta \mapsto f_\theta(x) \) is differentiable at \( \theta_0 \) for \( P \)-almost every \( x \) with derivative \( f'_{\theta}(x) \) and such that, for every \( \theta_1 \) and \( \theta_2 \) in a neighborhood of \( \theta_0 \), there exists a measurable function \( \hat{f} \) with \( \mathbb{E}[\hat{f}^2(X_1)] < \infty \) such that

\[
|f_{\theta_1}(x) - f_{\theta_2}(x)| \leq \hat{f}(x)||\theta_1 - \theta_2||_2.
\]

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 \( 0 < \varepsilon < \text{diam}(\Theta) \), there exists \( C \) such that

\[
N_1(\varepsilon \sqrt{\mathbb{E}(|\hat{f}(X)|^2)}, \mathcal{F}, L_2(P)) \leq C \left( \frac{\text{diam}(\Theta)}{\varepsilon} \right)^d,
\]

i.e. has a finite bracketing number for all \( \varepsilon > 0 \). This result is needed in proving consistency in the following.

**Theorem 2.** Let \( r_1(n) = r_2(n) = 1 \). For fixed \( \lambda_1, \lambda_2 \geq 0 \), let \( \hat{\theta}_n(\lambda_1, \lambda_2) \) be a near-minimizer of \( M_n(\theta; \lambda_1, \lambda_2) \) as in Eq. (6), and let \( \theta_0(\lambda_1, \lambda_2) \) be the unique minimizer of \( M(\theta; \lambda_1, \lambda_2) \). Also let

\[
\mathcal{F}_1 = \{ m_\theta : \theta \in [-K, K]^p \}, \quad \mathcal{F}_2 = \{ m^L_{\theta; \lambda_1, \lambda_2} : \theta \in [-K, K]^p \},
\]

Theorem 2. Let \( r_1(n) = r_2(n) = 1 \). For fixed \( \lambda_1, \lambda_2 \geq 0 \), let \( \hat{\theta}_n(\lambda_1, \lambda_2) \) be a near-minimizer of \( M_n(\theta; \lambda_1, \lambda_2) \) as in Eq. (6), and let \( \theta_0(\lambda_1, \lambda_2) \) be the unique minimizer of \( M(\theta; \lambda_1, \lambda_2) \). Also let

\[
\mathcal{F}_1 = \{ m_\theta : \theta \in [-K, K]^p \}, \quad \mathcal{F}_2 = \{ m^L_{\theta; \lambda_1, \lambda_2} : \theta \in [-K, K]^p \},
\]
where $m_\theta$ and $m_{U(\theta;\lambda_1,\lambda_2)}$ are defined in Eqs. (4) and (8). Suppose the following:

**Assumption 1.** $\theta \mapsto M(\theta;\lambda_1,\lambda_2)$ is continuous and $\liminf |\theta| \to \pm \infty M(\theta;\lambda_1,\lambda_2) > M(\theta_0;\lambda_1,\lambda_2)$.

**Assumption 2.** $X_1, \ldots, X_n$ are iid continuous random vectors with finite second moment.

Then

$$\left| \hat{\theta}_n(\lambda_1,\lambda_2) - \theta_0(\lambda_1,\lambda_2) \right|_2 \overset{P}{\to} 0.$$ 

**Proof.** See Appendix C.1.

**Remark on Assumptions 1 and 2.** Assumption 1 makes sure the solution of the population optimization problem is well-defined. For absolutely continuous $X$ under consideration (recall the Setup in Sec. 3), (CVaR-pop) is convex so is well-defined in this sense. Assumption 2 is a standard requirement on the moments of $X$ for asymptotic normality to hold.

The following result is an immediate consequence of Theorem 2.

**Corollary 1 (Consistency of the PBR solution and its performance).** If $r_1(n) = o(1)$ and $r_2(n) = o(1)$, then

$$\left| \hat{w}_n - w_0 \right|_2 \overset{P}{\to} 0,$$

i.e. the PBR solution converges to the true optimal with appropriate decaying of the regularization constraints. Moreover, by the Continuous Mapping Theorem, the performance of the PBR solution is also consistent in that the true mean and true CVaR of $\hat{w}_n$ converge to the population optimal values. Formally,

$$\left| \hat{w}_n^\top \mu - w_0^\top \mu \right| \overset{P}{\to} 0$$

and

$$\left| \text{CVaR}(\hat{w}_n^\top X; \beta) - \text{CVaR}(w_0^\top X; \beta) \right| \overset{P}{\to} 0.$$ 

### 5.2. First-order effect of regularization constraints

We now derive the first-order behavior of $\hat{w}_n$ by analyzing the first-order asymptotics of $\hat{\theta}_n(\lambda_1,\lambda_2)$ for $r_1(n) = r_2(n) = 1$. Note that we need to set $r_1(n)$ and $r_2(n)$ to constant values in order to “see” the effect of regularization.

The first-order behavior of $\hat{\theta}_n(0,0)$ is a straightforward application of known results about M-estimators for Lipschitz class of objective functions [e.g. Theorem 5.23 of Van der Vaart (2000)]. The first-order asymptotics of $\hat{\theta}_n(\lambda_1,\lambda_2)$ when $\lambda_1, \lambda_2$ are not both zero does not follow straightforwardly from M-estimation results however because $M_n(\theta;\lambda_1,\lambda_2)$ is a sample average of identically distributed but non-independent terms.

Nevertheless, as mentioned earlier in this section, stochastic processes of the form $\theta \mapsto M_n(\theta;\lambda_1,\lambda_2)$ are known as U-processes, and we can decompose them pointwise into a sum of iid random variables and a component which is $o_P(1/\sqrt{n})$ [Hoeffding (1948)]:

$$M_n(\theta;\lambda_1,\lambda_2) = \frac{1}{n} \sum_{i=1}^n m_{(\theta;\lambda_1,\lambda_2)}(X_i) + E_n(\theta;\lambda_1,\lambda_2),$$

where $m_{(\theta;\lambda_1,\lambda_2)}(X_i)$ are defined in Eqs. (4) and (8).
where \( m_{(\theta;\lambda_1,\lambda_2)}(X_i) = 2E[X]\{m_{(\theta;\lambda_1,\lambda_2)}(X_i, X_j) - \mathbb{E}_{X_1, X_2}[m_{(\theta;\lambda_1,\lambda_2)}(X_i, X_2)]\} \) and \( E_n(\theta;\lambda_1,\lambda_2) = o_p(1/\sqrt{n}) \). Hence we conjecture \( |R_n^U(\hat{\theta}_n;\lambda_1,\lambda_2)| \rightarrow 0 \), where

\[
R_n^U(\theta;\lambda_1,\lambda_2) = \sqrt{n}(\theta - \theta_0) - [\nabla_{\theta_0}^2 \mathbb{E}m_{(\theta;\lambda_1,\lambda_2)}(X_i)]^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} m_{(\theta;\lambda_1,\lambda_2)}(X_i).
\]

Now \( \hat{\theta}_n \) changes with every \( n \) so pointwise convergent result above is not sufficient. We need to show uniform probabilistic convergence of \( R_n^U(\theta;\lambda_1,\lambda_2) \), and implicitly of \( E_n(\theta;\lambda_1,\lambda_2) \), which is generally not straight-forward. Uniform convergence requires showing the convergence of quantities such as \( \sup_{t \in T} |X_n(t)| \) for some stochastic process \( \{X_n(t) : t \in T\} \), thus we introduce the notion of weak convergence of stochastic processes.

**Definition 3 (Weak Convergence of a Stochastic Process).** A sequence of \( X_n : \Omega_n \rightarrow \ell^\infty(T) \) converges weakly to a tight random element \( X \) iff both of the following conditions hold:

1. Finite approximation: the sequence \( (X_n(t_1), \ldots, X_n(t_k)) \) converges in distribution in \( \mathbb{R}^k \) for every finite set of points \( t_1, \ldots, t_k \) in \( T \).

2. Maximal inequality: for every \( \varepsilon, \eta > 0 \) there exists a partition of \( T \) into finitely many sets \( T_1, \ldots, T_k \) such that

\[
\limsup_{n \rightarrow \infty} \mathbb{P} \left[ \sup_{s, t \in T_i} |X_n(s) - X_n(t)| \geq \varepsilon \right] \leq \eta.
\]

The point at the end of this is, as taking the supremum is a continuous map in the topology of \( \ell^\infty(T) \), weak convergence of \( X_n(\cdot) \) to \( X(\cdot) \) would allow us to conclude \( \sup_{t \in T} |X_n(t)| \rightarrow \sup_{t \in T} |X(t)| \).

Regarding empirical processes, we say a class of measurable functions \( \mathcal{F} \) is \( P \)-Donsker if \( \{G_n f : f \in \mathcal{F}\} \) converges weakly to a tight random element in \( \ell^\infty(\mathcal{F}) \). This property is related to the bracketing numbers introduced in Sec. 5.1: a class \( \mathcal{F} \) is \( P \)-Donsker if \( \varepsilon \log [\mathbb{N}_1(\varepsilon, \mathcal{F}, L_2(P))] \rightarrow 0 \) as \( \varepsilon \rightarrow 0 \) [due to Donsker; see Theorem 19.5 of Van der Vaart (2000)]. Many sufficient conditions for the weak convergence of \( \{\cup_n f : f \in \mathcal{F}\} \) are provided in Arcones and Gine (1993), and we make use of one in our derivation of the first-order asymptotics of \( \hat{\theta}_n(\lambda_1,\lambda_2) \) below.

**Theorem 3 (First-order effect of regularization).** Let \( r_1(n) = r_2(n) = 1 \) and fix \( \lambda_1, \lambda_2 \geq 0 \), \( \lambda_1, \lambda_2 \) not both zero, and let

\[
\hat{m}_{(\theta;\lambda_1,\lambda_2)}^U(x) = \nabla_{\theta} m_{(\theta;\lambda_1,\lambda_2)}^U(x)|_{\theta = \theta(\lambda_1,\lambda_2)}, \text{ for } x \in \mathbb{R}^p.
\]

Further assume Assumption 1 of Theorem 2 and

Assumption 2. \( \mathbb{E}_{X_1, X_2}[m_{(\theta;\lambda_1,\lambda_2)}^U(X_1, X_2)^2] < \infty \).

\(^4\) A random element is a generalization of a random variable. Let \( (\Omega, \mathcal{G}, P) \) be a probability space and \( \mathcal{D} \) a metric space. Then the \( \mathcal{G} \)-measurable map \( X : \Omega \rightarrow \mathcal{D} \) is called a random element.
Assumption 3. $\theta \mapsto M(\theta; \lambda_1, \lambda_2)$ admits a second-order Taylor expansion at its point of minimum $\theta_0(\lambda_1, \lambda_2)$ with nonsingular symmetric second derivative matrix $V_{\theta_0}(\lambda_1, \lambda_2)$.

Then

$$\sqrt{n}(\hat{\theta}_n(\lambda_1, \lambda_2) - \theta_0(\lambda_1, \lambda_2)) = -V_{\theta_0}^{-1}(\lambda_1, \lambda_2) \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{m}_{1, \lambda_1, \lambda_2}(X_i) + o_p(1)$$

where

$$\hat{m}_{1, \lambda_1, \lambda_2}(X_i) = 2\mathbb{E}_{X_2}[\hat{m}^U_{(\theta, \lambda_1, \lambda_2)}(X_1, X_2)] - \mathbb{E}_{X_1, X_2}[\hat{m}^U_{(\theta, \lambda_1, \lambda_2)}(X_1, X_2)]$$

is the first-order term in the Hoeffding decomposition of $M_n(\theta; \lambda_1, \lambda_2)$, also known as the Hajek projection.

Proof. See Appendix C.2.

Remark – Assumptions 2’ and 3. Assumption 2’ is a requirement on the moments of $X$ for asymptotic normality to hold for the regularized solution. For the mean-CVaR problem, Assumption 2’ equates to the existence of the fourth moment of $X$, which is obvious since the regularized problem involves asymptotics of functions that depend on first and second moments of $X$. Assumption 3 is a further requirement on the distribution of $X$ so that the population problem is sufficiently smooth at its optimum. Note these assumptions are all of the distribution of $X$, not of the problem structure. Thus Theorems 2 and 3 hold for the mean-CVaR problem for sufficiently well-behaved distributions of $X$, such as the elliptical family of distributions (which includes the Gaussian distribution).

In Appendix C.3, we provide a verification of Theorem 3 by investigating the case $X \sim \mathcal{N}(\mu, \Sigma)$ in detail.

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 ($N_{\text{train}} = 120$) and the last 10 years of observations used as test data ($N_{\text{test}} = 120$). 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

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

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

2. **PBR only on the objective**: solving the problem (3) 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. 6.3.

3. **PBR only on the constraint**: solving the problem (3) 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 (CVaR-PBR). Both regularization parameters $U_1$ and $U_2$ are calibrated using OOS-PBCV.

5. **L1 regularization**: solving the sample average approximation problem (CVaR-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 (CVaR-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 (CVaR-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.

6.2. 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

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

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

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

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

where \( \hat{w}_{t,j} \) is the portfolio weight in asset \( j \) at time \( t \), \( \hat{w}_{t+},j \) 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

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

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

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

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

- **Variance with transaction cost:** we report annualized out-of-sample portfolio variance net of transaction cost \( \hat{\sigma}^2_{\text{test},tc} \).

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

\[
\text{Turnover} = \frac{1}{N_{\text{test}}} \sum_{t=1}^{N_{\text{test}}} \sum_{j=1}^{p} |\hat{w}_{t+1,j} - \hat{w}_{t+},j|^+.
\]

For further details on these performance measures we refer the reader to DeMiguel et al. (2009b).
Figure 3  A schematic explaining the out-of-sample performance-based $k$-cross validation (OOS-PBCV) algorithm used to calibrate the constraint rhs, $U$, for the case $k = 3$. The training data set is split into $k$ bins, and the optimal $U$ for the entire training data set is found by averaging the best $U$ found for each subset of the training data.

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

One important question in solving (CVaR-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 $U_i$. 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 $U_{-i}$, $[\underline{U}_{-i}, \bar{U}_{-i}]$.

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 $U^*_{-i} \in [\underline{U}_{-i}, \bar{U}_{-i}]$ by a stopping criterion.

3. Average over the $k$ results to get $U^* = \frac{1}{k} \sum_{i=1}^{k} U^*_{-i}$.

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

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

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

$$U = \min_{w,z} z^\top \Omega_n z \quad \text{s.t.} \quad \begin{align*}
    w^\top \hat{\mu}_n &= R \\
    w^\top 1_p &= 1 \\
    z_i &\geq -w^\top X_i - \alpha, \quad i = 1, \ldots, n. \\
    z_i &\geq 0, \quad i = 1, \ldots, n.
\end{align*} \quad (U\text{-min})$$

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

$$\bar{U}_{-i} = \min[\hat{U}, \hat{CVaR}_n(-\hat{w}^\top_{-i} X; \beta)].$$

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

$$U_{mp} = \min_{w,z} z^\top \Omega_{-i} z \quad \text{s.t.} \quad \begin{align*}
    w^\top \hat{\mu}_{-i} &= R \\
    w^\top 1_p &= 1 \\
    z_i &\geq -w^\top X_i - \alpha, \quad i \in C \setminus C_i. \\
    z_i &\geq 0, \quad i \in C \setminus C_i,
\end{align*}$$
where $\Omega_{-i}$ is the sample variance operator and $\hat{\mu}_{-i}$ the sample mean 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

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

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

2. Finding $U^*_{-i} \in [\hat{U}_{-i}, \bar{U}_{-i}]$. 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 $t\Delta U := t(\bar{U}_{-i} - U_{-i})/\text{Div}$, 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

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

is met.

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:

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

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

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

The second quantity $d\hat{w}^*(U)/dU$ is the marginal change in the optimal solution $\hat{w}^*$ as the rhs $U$ changes. We approximate this by solving $\text{(CVaR-PBR)}$ 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 $\alpha = 0.4, \beta = 0.9, \text{Div} = 5, \text{bit} = 0.05$ 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.
Out-of-Sample Performance-Based k-Cross Validation (OOS-PBCV)

Initialize
Choose no. of bins $k$

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

Solve U-min(U) on $D_{\text{train}}$ to get $\hat{w}_{\text{Umin}}$; set $U = (\hat{w}_{\text{Umin}})\Sigma \hat{w}_{\text{Umin}}$

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

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

for $b \leftarrow 1$ to $k$
do
Solve CVaR-PBR(U) on $D^b_{\text{train}}$ to get $\hat{w}^b_{\text{train}}$; set $\U^b = (\hat{w}^b_{\text{train}})\Sigma \hat{w}^b_{\text{train}}$

if $\U^b < U$ then $U^*_b = U$ and terminate
\elself solve U-min(U) on $D^b_{\text{train}}$ to get $\hat{w}^b_{\text{Umin}}$; set $U^b = (\hat{w}^b_{\text{Umin}})\Sigma \hat{w}^b_{\text{Umin}}$
end

if $U^b > U$ then $U^*_b = U$ and terminate
\elself compare and update boundaries:
$$\U^b = \min(\U^b, U)$$
$$U^b = \max(U^b, U)$$
\end

Run (OOS-PBSD) with boundaries $[U^b, U^b]$ to find $U^*_b$
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).

6.4. Discussion of Results

Out-of-sample Sharpe ratio Table 1 reports the out-of-sample Sharpe ratios of the eight strategies listed in Sec. 6.1. For $p = 5$ 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. For example, the PBR Sharpe ratios for $p = 5$ are 1.2439 and 1.2073 respectively for $k = 2$ and $k = 3$ bins; whereas the Sharpe ratios for $L_1$ regularization is 1.0112 and 1.0754 and for $L_2$ regularization is 0.9650 and 1.0636 respectively for $k = 2$ and $k = 3$ bins. In addition, both
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 $U = U_{-b}$, $\Delta U := t(U_{-b} - U_{-b})/Div$, $t = 1$
  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
  \[
  \nabla_w \text{Sharpe}(\hat{w}_{-b}(U)) = \frac{((\hat{w}_{-b})^\top \Sigma_{-b} \hat{w}_{-b}) \mu_{-b} - (\hat{w}_{-b} \mu) \Sigma_{-b} \hat{w}_{-b}}{((\hat{w}_{-b})^\top \Sigma_{-b} \hat{w}_{-b})^{3/2}}
  \]
  \[
  \frac{d\hat{w}_{-b}(U)}{dU} = \hat{w}_{-b}(U) - \hat{w}_{-b}((1 - bit)U) / bit \times U
  \]
  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).

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, except that there is a lot of noise and because of this, one is better off investing in an equally-weighted portfolio than optimizing. This observation — that portfolio allocation via optimization is meaningless on a
noise data set — is similar to other studies that have looked at the data-driven Markowitz problem Frankfurter et al. (1971), Frost and Savarino (1986), Frost and Savarino (1988b), Michaud (1989), Best and Grauer (1991), Chopra and Ziemba (1993), Broadie (1993) and DeMiguel et al. (2009b). This perhaps calls for investigation into the effects of dimensionality and how to handle them effectively in asset allocation problems, and we leave these questions open for future work.

From the perspective of an investor looking at the results of Table 1, 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.

Finally, let us analyze 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.

**Out-of-sample variance** Table 2 reports the out-of-sample Sharpe ratios of the eight strategies listed in Sec. 6.1. Interestingly, for \( p = 5 \) and 10 data sets, the variances of the PBR method are consistently larger than those of the SAA method, for each target return level, for both bins sizes considered. This indicates that the dominance of PBR over SAA in the Sharpe ratio is attributed to a greater out-of-sample excess return. Another striking observation is that the equally-weighted portfolio is associated with the highest variances for all data sets; this indicates that the optimization mechanism is a variance-reduction procedure. Although we are not minimizing the variance directly in the mean-CVaR (or the global minimum CVaR) problem, this indicates that CVaR minimization indirectly also controls the variance. We certainly know this is the case for normally distributed returns because CVaR in this case is a linear combination of the mean and the standard deviation, but because CVaR is also a measure of spread, it must be intimately linked to variance for other distributions as well.

Comparing across data sets, we see that \( p = 41 \) variances are substantially greater than \( p = 5 \) or \( p = 10 \) cases, again reinforcing the presence of greater noise in this data set.

**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 the \( p = 5 \) data set, the PBR method has 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.

Comparing across data sets, we see that $p = 41$ turnovers are substantially greater than $p = 5$ or $p = 10$ cases. As frequent, large portfolio rebalancing is generally not possible and/or practical, this is yet another argument against using an optimization strategy when the number of assets being considered is large.

7. Conclusion

We introduced a new notion of regularization, performance-based regularization, and investigated it in detail for the mean-CVaR portfolio optimization problem. Our empirical results show that PBR improves upon the SAA approach, standard $L_1$ and $L_2$ regularizations, the equally-weighted portfolio and the SAA approach minus the mean constraint in terms of the out-of-sample Sharpe ratio. As PBR makes the same minimal assumptions about the data while improving finite-sample performance, it is a promising new approach to data-driven optimization.

Appendix A: 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 $\hat{\text{CVaR}}_n(L; \beta)$ introduced in Eq. (2) of Sec. 2.1. First, let us define a closely related estimator:

**Definition 4 (Type 1 CVaR estimator).** For $\beta \in (0, 1)$, we define Type 1 CVaR estimator to be

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

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

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

**Proposition 2.** The solution $\alpha^* = L_{\lceil (n\beta) \rceil}$ is the unique minimizer in the one-dimensional optimization problem

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

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

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

$$\Delta(m) = G_n(L_{\lceil (n\beta) + m + 1 \rceil}) - G_n(L_{\lceil (n\beta) + m \rceil}) = (1 - \varepsilon_n)(L_{\lceil (n\beta) + m + 1 \rceil} - L_{\lceil (n\beta) + m \rceil}) = \frac{1}{n - \lceil n\beta \rceil + 1} A,$$
Table 1: Sharpe Ratios (with transaction costs) for empirical data.

<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>Mean-CVaR R=0.04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAA</td>
<td>1.2137</td>
<td>1.0321</td>
<td>0.3657</td>
</tr>
<tr>
<td></td>
<td>2 bins</td>
<td>3 bins</td>
<td>3 bins</td>
</tr>
<tr>
<td>PBR (CVaR only)</td>
<td>1.2113</td>
<td>1.1733</td>
<td>0.5657</td>
</tr>
<tr>
<td></td>
<td>(0.0554, 0.0674)</td>
<td>(0.0638, 0.0312)</td>
<td>(0.7908, 0.7908)</td>
</tr>
<tr>
<td>PBR (mean only)</td>
<td>1.2089</td>
<td>1.1802</td>
<td>0.5519</td>
</tr>
<tr>
<td></td>
<td>(0.0746, 0.0790)</td>
<td>(0.1051, 0.1338)</td>
<td>(0.7518, 0.6454)</td>
</tr>
<tr>
<td>PBR (both)</td>
<td>1.2439</td>
<td>1.2073</td>
<td>1.1122</td>
</tr>
<tr>
<td></td>
<td>1.0754</td>
<td>0.9254</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0513, 0.0601)</td>
<td>(0.0691, 0.0648)</td>
<td>(0.7054, 0.7667)</td>
</tr>
<tr>
<td>L1</td>
<td>1.0112</td>
<td>1.0754</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.1497, 0.1366)</td>
<td>(0.2293, 0.0380)</td>
<td>(0.6874, 0.6242)</td>
</tr>
<tr>
<td>L2</td>
<td>0.9650</td>
<td>1.0636</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.1780, 0.1287)</td>
<td>(0.1512, 0.0158)</td>
<td>(0.7087, 0.6878)</td>
</tr>
<tr>
<td>Mean-CVaR R=0.06</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAA</td>
<td>1.2179</td>
<td>1.0321</td>
<td>0.3657</td>
</tr>
<tr>
<td></td>
<td>2 bins</td>
<td>3 bins</td>
<td>3 bins</td>
</tr>
<tr>
<td>PBR (CVaR only)</td>
<td>1.2223</td>
<td>1.2063</td>
<td>0.5657</td>
</tr>
<tr>
<td></td>
<td>(0.0503, 0.0527)</td>
<td>(0.0633, 0.0294)</td>
<td>(0.7920, 0.7909)</td>
</tr>
<tr>
<td>PBR (mean only)</td>
<td>1.2055</td>
<td>1.1902</td>
<td>0.5657</td>
</tr>
<tr>
<td></td>
<td>(0.0699, 0.0746)</td>
<td>(0.1053, 0.1358)</td>
<td>(0.7531, 0.6359)</td>
</tr>
<tr>
<td>PBR (both)</td>
<td>1.2450</td>
<td>1.2043</td>
<td>1.1122</td>
</tr>
<tr>
<td></td>
<td>1.1506</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0504, 0.0581)</td>
<td>(0.0686, 0.0607)</td>
<td>(0.7102, 0.7656)</td>
</tr>
<tr>
<td>L1</td>
<td>0.9404</td>
<td>1.0464</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.1812, 0.1395)</td>
<td>(0.2282, 0.0380)</td>
<td>(0.7061, 0.6242)</td>
</tr>
<tr>
<td>L2</td>
<td>0.9271</td>
<td>1.0627</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.1977, 0.1286)</td>
<td>(0.1432, 0.1521)</td>
<td>(0.7175, 0.6886)</td>
</tr>
<tr>
<td>Mean-CVaR R=0.08</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAA</td>
<td>1.2487</td>
<td>1.0346</td>
<td>0.3657</td>
</tr>
<tr>
<td></td>
<td>2 bins</td>
<td>3 bins</td>
<td>3 bins</td>
</tr>
<tr>
<td>PBR (CVaR only)</td>
<td>1.2493</td>
<td>1.2098</td>
<td>0.5657</td>
</tr>
<tr>
<td></td>
<td>(0.0434, 0.0462)</td>
<td>(0.0579, 0.0323)</td>
<td>(0.7908, 0.7908)</td>
</tr>
<tr>
<td>PBR (mean only)</td>
<td>1.2480</td>
<td>1.2088</td>
<td>0.5657</td>
</tr>
<tr>
<td></td>
<td>(0.0591, 0.0693)</td>
<td>(0.1053, 0.1384)</td>
<td>(0.7541, 0.6391)</td>
</tr>
<tr>
<td>PBR (both)</td>
<td>1.2715</td>
<td>1.2198</td>
<td>1.1122</td>
</tr>
<tr>
<td></td>
<td>1.1087</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0453, 0.0544)</td>
<td>(0.0664, 0.0639)</td>
<td>(0.7618, 0.7647)</td>
</tr>
<tr>
<td>L1</td>
<td>0.8921</td>
<td>0.9363</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.1964, 0.1572)</td>
<td>(0.2122, 0.1645)</td>
<td>(0.7008, 0.6128)</td>
</tr>
<tr>
<td>L2</td>
<td>0.9367</td>
<td>1.0801</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.1989, 0.1179)</td>
<td>(0.1323, 0.1530)</td>
<td>(0.7177, 0.6870)</td>
</tr>
</tbody>
</table>

| Equal         | 0.6617        | 0.7019        | 0.6297        |
| Global min. CVaR | 1.2137        | 1.0321        | 0.3657        |

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. 6.3, with $k = 2$ and 3 bins. In parentheses we report the $p$-values of the difference between the Sharpe ratio of each asset allocation strategy 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 2 Variances for empirical data.

<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>Mean-CVaR R=0.04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAA</td>
<td>0.1130</td>
<td>0.1098</td>
<td>0.1595</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 (CVaR only)</td>
<td>0.1154 0.1159</td>
<td>0.1200 0.1228</td>
<td>0.1781 0.1781</td>
</tr>
<tr>
<td>PBR (mean only)</td>
<td>0.1116 0.1130</td>
<td>0.1077 0.1075</td>
<td>0.1449 0.1543</td>
</tr>
<tr>
<td>PBR (both)</td>
<td>0.1138 0.1145</td>
<td>0.1134 0.1125</td>
<td>0.1384 0.1541</td>
</tr>
<tr>
<td>L1</td>
<td>0.1171 0.1132</td>
<td>0.1081 0.1080</td>
<td>0.1331 0.1352</td>
</tr>
<tr>
<td>L2</td>
<td>0.1181 0.1153</td>
<td>0.1097 0.1102</td>
<td>0.1287 0.1451</td>
</tr>
<tr>
<td>Mean-CVaR R=0.06</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAA</td>
<td>0.1115 0.1098</td>
<td>0.1098 0.1595</td>
<td></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 (CVaR only)</td>
<td>0.1142 0.1151</td>
<td>0.1200 0.1228</td>
<td>0.1779 0.1781</td>
</tr>
<tr>
<td>PBR (mean only)</td>
<td>0.1103 0.1116</td>
<td>0.1077 0.1077</td>
<td>0.1451 0.1543</td>
</tr>
<tr>
<td>PBR (both)</td>
<td>0.1125 0.1138</td>
<td>0.1134 0.1127</td>
<td>0.1395 0.1541</td>
</tr>
<tr>
<td>L1</td>
<td>0.1187 0.1140</td>
<td>0.1080 0.1079</td>
<td>0.1331 0.1352</td>
</tr>
<tr>
<td>L2</td>
<td>0.1181 0.1140</td>
<td>0.1097 0.1103</td>
<td>0.1290 0.1453</td>
</tr>
<tr>
<td>Mean-CVaR R=0.08</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAA</td>
<td>0.1109 0.1098</td>
<td>0.1098 0.1595</td>
<td></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 (CVaR only)</td>
<td>0.1140 0.1164</td>
<td>0.1209 0.1219</td>
<td>0.1780 0.1780</td>
</tr>
<tr>
<td>PBR (mean only)</td>
<td>0.1104 0.1112</td>
<td>0.1077 0.1074</td>
<td>0.1455 0.1544</td>
</tr>
<tr>
<td>PBR (both)</td>
<td>0.1119 0.1135</td>
<td>0.1138 0.1125</td>
<td>0.1416 0.1541</td>
</tr>
<tr>
<td>L1</td>
<td>0.1221 0.1176</td>
<td>0.1084 0.1076</td>
<td>0.1337 0.1329</td>
</tr>
<tr>
<td>L2</td>
<td>0.1172 0.1141</td>
<td>0.1099 0.1101</td>
<td>0.1302 0.1453</td>
</tr>
<tr>
<td>Equal</td>
<td>0.1434 0.1473</td>
<td>0.1473 0.1794</td>
<td></td>
</tr>
<tr>
<td>Global min. CVaR</td>
<td>0.1130</td>
<td>0.1098</td>
<td>0.1595</td>
</tr>
</tbody>
</table>

This table reports the annualized out-of-sample variances 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 variances of the equally-weighted portfolio and the solution to the global minimum CVaR problem (no mean constraint).

where

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

$$= (n - \lceil n\beta \rceil - m)(L_{\lceil (n\beta) + m+1 \rceil} - L_{\lceil (n\beta) + m \rceil}).$$

Thus

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

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

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

**Definition 5 (Type 2 CVaR estimator).** For $\beta \in (0.5, 1)$, we define Type 2 CVaR estimator to be

$$\hat{CV}_2 n(L; \beta) := \frac{1}{n - \lceil n\beta \rceil + 1} \sum_{i=1}^{n} L_i 1(L_i \geq \hat{\alpha}_n(\beta)).$$
Table 3 Turnovers (with transaction costs) for empirical data.

<table>
<thead>
<tr>
<th>Sharpe Ratios</th>
<th>FF 5 Industry p=5</th>
<th>FF 10 Industry p=10</th>
<th>FF 49 Industry p=41</th>
</tr>
</thead>
<tbody>
<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>2 bins</td>
<td>3 bins</td>
<td>2 bins</td>
<td>3 bins</td>
</tr>
<tr>
<td>PBR (CVaR only)</td>
<td>0.1834</td>
<td>0.1985</td>
<td>0.4049</td>
</tr>
<tr>
<td></td>
<td>0.4049</td>
<td>0.5586</td>
<td>1.7773</td>
</tr>
<tr>
<td>PBR (mean only)</td>
<td>0.1230</td>
<td>0.1274</td>
<td>0.3104</td>
</tr>
<tr>
<td></td>
<td>0.3104</td>
<td>0.2731</td>
<td>1.3173</td>
</tr>
<tr>
<td>PBR (both)</td>
<td>0.1387</td>
<td>0.1388</td>
<td>0.3700</td>
</tr>
<tr>
<td></td>
<td>0.3700</td>
<td>0.3682</td>
<td>1.7492</td>
</tr>
<tr>
<td>L1</td>
<td>0.1992</td>
<td>0.1581</td>
<td>0.3415</td>
</tr>
<tr>
<td></td>
<td>0.3415</td>
<td>0.2722</td>
<td>1.5158</td>
</tr>
<tr>
<td>L2</td>
<td>0.1565</td>
<td>0.1469</td>
<td>0.2288</td>
</tr>
<tr>
<td></td>
<td>0.2288</td>
<td>0.2270</td>
<td>1.2192</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>2 bins</td>
<td>3 bins</td>
<td>2 bins</td>
<td>3 bins</td>
</tr>
<tr>
<td>PBR (CVaR only)</td>
<td>0.1918</td>
<td>0.2074</td>
<td>0.4071</td>
</tr>
<tr>
<td></td>
<td>0.4071</td>
<td>0.5616</td>
<td>1.7922</td>
</tr>
<tr>
<td>PBR (mean only)</td>
<td>0.1364</td>
<td>0.1414</td>
<td>0.3100</td>
</tr>
<tr>
<td></td>
<td>0.3100</td>
<td>0.2729</td>
<td>1.3188</td>
</tr>
<tr>
<td>PBR (both)</td>
<td>0.1519</td>
<td>0.1526</td>
<td>0.3724</td>
</tr>
<tr>
<td></td>
<td>0.3724</td>
<td>0.3672</td>
<td>1.7610</td>
</tr>
<tr>
<td>L1</td>
<td>0.2263</td>
<td>0.1754</td>
<td>0.3498</td>
</tr>
<tr>
<td></td>
<td>0.3498</td>
<td>0.2723</td>
<td>1.5232</td>
</tr>
<tr>
<td>L2</td>
<td>0.1842</td>
<td>0.1532</td>
<td>0.2407</td>
</tr>
<tr>
<td></td>
<td>0.2407</td>
<td>0.2401</td>
<td>1.2374</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>2 bins</td>
<td>3 bins</td>
<td>2 bins</td>
<td>3 bins</td>
</tr>
<tr>
<td>PBR (CVaR only)</td>
<td>0.2148</td>
<td>0.2242</td>
<td>0.4486</td>
</tr>
<tr>
<td></td>
<td>0.4486</td>
<td>0.6472</td>
<td>1.7775</td>
</tr>
<tr>
<td>PBR (mean only)</td>
<td>0.1517</td>
<td>0.1575</td>
<td>0.3066</td>
</tr>
<tr>
<td></td>
<td>0.3066</td>
<td>0.2827</td>
<td>1.3228</td>
</tr>
<tr>
<td>PBR (both)</td>
<td>0.1693</td>
<td>0.1681</td>
<td>0.4099</td>
</tr>
<tr>
<td></td>
<td>0.4099</td>
<td>0.4034</td>
<td>1.6887</td>
</tr>
<tr>
<td>L1</td>
<td>0.3100</td>
<td>0.2395</td>
<td>0.3628</td>
</tr>
<tr>
<td></td>
<td>0.3628</td>
<td>0.3042</td>
<td>1.5370</td>
</tr>
<tr>
<td>L2</td>
<td>0.2451</td>
<td>0.1835</td>
<td>0.2588</td>
</tr>
<tr>
<td></td>
<td>0.2588</td>
<td>0.2633</td>
<td>1.1747</td>
</tr>
</tbody>
</table>

This table reports the portfolio turnovers (defined in Eq. 11) 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).

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 1 then follows.

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

\[
\hat{CV}^2_n(L; \beta) = CV_1_n(L; \beta) + \varepsilon_n L_{\lceil n\beta \rceil}.
\]

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

\[
\hat{CV}^2_n(L; \beta) = \frac{1}{n - \lfloor n\beta \rfloor} + \sum_{i=1}^{\lfloor n\beta \rfloor} L_i 1(L_i \geq L_{\lceil n\beta \rceil})
\]
and \( z \) derive properties of the optimal solution by examining the KKT conditions. Hence we have shown, by elementary arguments, that the problem is also strictly feasible, since \( \text{KKT} \) conditions are both empty, (ii) the Lagrangian is

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

where the final equality is due to Proposition 2.

**A.1. Proof of Proposition 1**

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

\[
|\overline{CV\!aR}_n(L; \beta) - \overline{CV\!1}_n(L; \beta)| \leq \varepsilon_n (\alpha_n^1 \cup \alpha_n).
\]

Hence we have

\[
\overline{CV\!aR}_n(L; \beta) = \overline{CV\!1}_n(L; \beta) + O_\varepsilon(\varepsilon_n)
\]

which implies

\[
\text{Var}[\overline{CV\!aR}_n(L; \beta)] = \frac{1}{n(1 - \beta)^2} \text{Var}[L_i 1(L_i \geq L_{(\lceil n\beta \rceil)})] + O(n^{-2}),
\]

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

**Appendix B: Proof of Theorem 1**

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

**Proposition 3.** Consider the optimization problem

\[
\begin{array}{ll}
\min & z^\top 1_n \\
\text{s.t.} & z_i \geq 0 \quad \forall i \\
& z_i \geq c_i \quad \forall i \\
& z^\top \Omega_n z \leq f
\end{array}
\]

(12)

where \( c \in \mathbb{R}^n \) is some constant vector, \( f > 0 \) is a constant scalar and \( \Omega_n = (n-1)^{-1}(I_n - n^{-1}1_n1_n^\top) \), the sample covariance operator. Suppose (12) 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 \( V(z) := S_1 \cap S_2 \) (i.e. \( V(z) \) is the set of indices for which \( z_i > \max(0, c_i) \)). Then the optimal solution \( z^* \) falls into one of three cases: (i) \( S_2(z^*) = \{1, 2, \ldots, n\} \) (i.e. \( S_1(z^*) \) and \( V(z^*) \) are both empty), (ii) \( S_1(z^*) \) is nonempty and \( V(z^*) \) is empty, or (iii) \( S_1(z^*) \) is empty and \( V(z^*) \) is nonempty.

**Proof.** Problem (12) is a convex optimization problem because \( \Omega_n \) is a positive semidefinite matrix. The problem is also strictly feasible, since \( z_0 = 2 \max_i \{c_i\} 1_n \) is a strictly feasible point: clearly, \( z_0 > \max\{0, c_i\} \) \( \forall i \) 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:
\[ z_i^* \eta_1^* i = 0 \quad \forall \ i, \ (z_i^* - c_i) \eta_2^* i = 0 \quad \forall \ i \] and \( \lambda^*[ (z^*)^\top \Omega_n z^* - f] = 0 \)
• First Order Condition:
\[
\nabla_{z^*} \mathcal{L} = 2\lambda^* \Omega_n z^* + (1_n - \eta_1^* - \eta_2^*) = 0 \quad (13a)
\]

By substituting for \( \Omega_n \), (13a) can be written as
\[
\frac{2\lambda^*}{n-1} \left( z^* - \frac{1}{n} (1_n^\top z^*) 1_n \right) = -1 + \eta_1^* + \eta_2^*. \quad (14)
\]

Suppose \( S_i(z^*) \neq \emptyset \) at the optimal primal-dual point \((z^*, \eta_1^*, \eta_2^*, \lambda^*)\). Then \( \exists i_0 \in S_i(z^*) \) such that \( z_{i_0}^* = 0 \).
The \( i_0 \)-th component of (14) gives
\[
-\frac{2\lambda^*}{n-1} (1_n^\top z^*) = -1 + \eta_1^*_{i_0} + \eta_2^*_{i_0}. \quad (15)
\]

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 (14) gives
\[
\frac{2\lambda^*}{n-1} \left( z_{j_0}^* - \frac{1}{n} (1_n^\top z^*) \right) = -1, \quad (16)
\]
which implies \( \lambda^* > 0 \) since \( \lambda^* \) cannot equal zero.

Now suppose \( S_i(z^*) \) and \( V(z^*) \) are both nonempty. Combining (15) and (16), 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 \( LHS > 0 \) whereas \( RHS \leq 0 \). Hence \( S_i(z^*) \) and \( V(z^*) \) cannot both be nonempty, and the result follows.

**Proof of Theorem 1**

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

Let \((\alpha^*, w^*, z^*, \nu_1^*, \nu_2^*, \eta_1^*, \eta_2^*, \lambda_1^*, \lambda_2^*)\) be the primal-dual optimal point for (CVaR-relax) and (CVaR-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 3, \( S_i(z^*) \), the set of indices for which \( z_i^* = 0 \), is empty. This means \( z_i^* > 0 \quad \forall \ i \) and \( \eta_1^*_{i_0} = 0 \quad \forall \ i \) by complementary slackness.

Now consider the sub-problem for a fixed \( \eta_2 \) in the dual problem (CVaR-relax-d):
\[
\max_{\eta_1, \eta_1 \geq 0} \left( \eta_1 + \eta_2 \right) \Omega_1 (\eta_1 + \eta_2). \quad (17)
\]
As \( 1_n \) is orthogonal to \( \Omega_1 \), and \( \Omega_1 \) is positive semidefinite, the optimal solution is of the form \( \eta_1 = a 1_n - \eta_2 \), where \( a \geq \max(\eta_2, 0) \), with a corresponding optimal objective 0. Hence, bearing in mind the constraints \( \eta_2 \geq 0 \) and \( \eta_2^\top 1_n = 1 \) in (CVaR-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 (CVaR-relax-d).
Appendix C: Proofs of results in Sec. 5

C.1. Proof of Theorem 2

By uniqueness of $\theta_0(\lambda_1, \lambda_2)$ and Assumption 1 (and compactness arguments), for every $\varepsilon > 0$, there exists $\eta > 0$ such that

$$||\hat{\theta}_n(\lambda_1, \lambda_2) - \theta_0(\lambda_1, \lambda_2)||_2 > \varepsilon \implies M(\hat{\theta}_n; \lambda_1, \lambda_2) - M(\theta_0; \lambda_1, \lambda_2) > \eta.$$ 

Thus if we can show the probability of the event $\{M(\hat{\theta}_n; \lambda_1, \lambda_2) - M(\theta_0; \lambda_1, \lambda_2) > \eta\}$ goes to zero for every $\varepsilon > 0$, then we have consistency.

We also have

$$M_n(\hat{\theta}_n; \lambda_1, \lambda_2) \leq M_n(\theta_0; \lambda_1, \lambda_2) + o_P(1) = M(\theta_0; \lambda_1, \lambda_2) + o_P(1),$$

the first inequality because $\hat{\theta}_n(\lambda_1, \lambda_2)$ is a near-minimizer of $M_n$, and the second equality by the Weak Law of Large Numbers (WLLN) on $M_n(\theta_0; \lambda_1, \lambda_2)$.

Thus

$$0 \leq M(\hat{\theta}_n; \lambda_1, \lambda_2) - M(\theta_0; \lambda_1, \lambda_2) = [M(\hat{\theta}_n; \lambda_1, \lambda_2) - M_n(\hat{\theta}_n; \lambda_1, \lambda_2)] + [M_n(\hat{\theta}_n; \lambda_1, \lambda_2) - M_n(\theta_0; \lambda_1, \lambda_2)] + [M_n(\theta_0; \lambda_1, \lambda_2) - M(\theta_0; \lambda_1, \lambda_2)] \leq M(\hat{\theta}_n; \lambda_1, \lambda_2) - M_n(\hat{\theta}_n; \lambda_1, \lambda_2) + o_P(1),$$

because the second term in $[ ]$ is $o_P(1)$ by $(*)$, and the last term in $[ ]$ is $o_P(1)$ by WLLN. We are left to prove $|M_n(\hat{\theta}_n; \lambda_1, \lambda_2) - M(\hat{\theta}_n; \lambda_1, \lambda_2)| \xrightarrow{p} 0$. At first glance, one may consider invoking the WLLN again. However, as $\hat{\theta}_n(\lambda_1, \lambda_2)$ is a random sequence of vectors that changes for every $n$, we cannot apply the WLLN which is a pointwise result (i.e. for each fixed $\theta \in \Theta$), and we need to appeal to the stronger ULLN.

**Case I:** $\lambda_1 = \lambda_2 = 0$. To show ULLN for the original objective, we show that $\mathcal{F}_1$ is a Lipschitz class of functions, hence $N_{\lfloor \varepsilon \rfloor}(\varepsilon, \mathcal{F}_1, L_r(P))$ for every $\varepsilon > 0$. Now $\theta \mapsto m_\theta(x) = \alpha + (1 - \beta)^{-1}(\alpha - w_0^\top x - v^\top L^\top x)\geq 0$, hence

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

where $I(x) := 1(-\alpha - w_0^\top x - v^\top L^\top x \geq 0)$, hence

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

is an upper bound on $||\nabla_\theta m_\theta(x)||_\infty$ and is independent of $\theta$. Thus $|m_{\theta_1}(x) - m_{\theta_2}(x)| \leq \hat{m}(x)||\theta_1 - \theta_2||_2$ for all $\theta_1, \theta_2 \in [-K, K]^{1+p}$, and together with Assumption 2, $\mathcal{F}_1$ is a Lipschitz class.

**Case II:** $\lambda_1 \geq 0, \lambda_2 \geq 0, \lambda_1, \lambda_2$ not both zero. Corollary 3.5 in Arcones and Gine (1993) says that ULLN also holds for the regularized objective if $N_{\lfloor \varepsilon \rfloor}(\varepsilon, \mathcal{F}_2, L_2(P \times P)) < \infty$ for every $\varepsilon > 0$. Let us now show that $\mathcal{F}_2$ is also a Lipschitz class of functions. Again, it is clear that

$$\theta \mapsto m^U(\theta;\lambda_1, \lambda_2)(x_1, x_2) = \frac{1}{2} [m_\theta(x_1) + m_\theta(x_2)] + \frac{\lambda_1}{2} [(w_1 + Lv)^\top (x_1 - x_2)]^2 + \frac{\lambda_2}{2} (z_\theta(x_1) - z_\theta(x_2))^2$$

is Lipschitz.
is differentiable at $\theta_0$ for all $(x_1, x_2) \in \mathbb{R}^p \times \mathbb{R}^p$. Also for all $\theta \in [-K, K]^{1+p}$,

$$
\nabla_{\theta} \frac{\lambda_1}{2}((w_1 + Lv)^T (x_1 - x_2))^2 = \lambda_1(x_1 - x_2)(x_1 - x_2)^T (w_1 + Lv)
$$

$$
\implies \|\nabla_{\theta} \frac{\lambda_1}{2}((w_1 + Lv)^T (x_1 - x_2))^2\|_\infty \leq \lambda_1 |x_1 - x_2|_\infty^2 |w_1 + Lv|_\infty \leq \lambda_1 C(|x_1 - x_2|_\infty^2)
$$

for some constant $C(K)$ dependent on $K$, and

$$
\nabla_{\theta} \frac{\lambda_2}{2}(z_0(x_1) - z_0(x_2))^2 = \lambda_2(z_0(x_1) - z_0(x_2)) \left[-I(x_1) + I(x_2) \right]
$$

$$
\implies \|\nabla_{\theta} \frac{\lambda_2}{2}(z_0(x_1) - z_0(x_2))^2\|_\infty \leq \lambda_2 C(K)(|x_1|_\infty + |x_2|_\infty)(\hat{m}(x_1) + \hat{m}(x_2))
$$

for some constant $C(K)$ dependent on $K$, hence

$$
\hat{m}_{(\lambda_1, \lambda_2)}^{U}(x_1, x_2) := \frac{1}{2}(\hat{m}(x_1) + \hat{m}(x_2)) + \lambda_1 C(K)|x_1 - x_2|_\infty^2 + \lambda_2 C(K)(|x_1|_\infty + |x_2|_\infty)(\hat{m}(x_1) + \hat{m}(x_2))
$$

(19)

is an upper bound on $\|\nabla_{\theta} \hat{m}_{(\theta, \lambda_1, \lambda_2)}^{U}(x_1, x_2)\|_\infty$ that is independent of $\theta$. Thus

$$
|\hat{m}_{(\theta_1, \lambda_1, \lambda_2)}^{U}(x_1, x_2) - \hat{m}_{(\theta_2, \lambda_1, \lambda_2)}^{U}(x_1, x_2)| \leq \hat{m}_{(\lambda_1, \lambda_2)}^{U}(x_1, x_2)||\theta_1 - \theta_2||_2,
$$

and together with Assumption 2', $\mathcal{F}_2$ is a Lipschitz class.

C.2. Proof of Theorem 3

In what follows, we suppress the dependence on $\lambda_1, \lambda_2$ for notational convenience.

The proof parallels the proof of Theorem 5.23 of Van der Vaart (2000). Let us assume for now that

1. For every given random sequence $h_n$ that is bounded in probability,

$$
\mathbb{U}_n[\sqrt{n}(\hat{m}_{(\theta_0, h_n)}^{U} - m_{(\theta_0)}^{U}) - h_n^T \hat{m}_{(\theta_0)}^{U}] \overset{p}{\to} 0.
$$

(1*)

and

2. $\sqrt{n}(\hat{\theta}_n - \theta_0) = O_P(1)$.

Since $\theta \mapsto M(\theta)$ is twice-differentiable, and $\nabla_{\theta} M(\theta)|_{\theta = \theta_0} = 0$ by first-order condition, we can rewrite Eq. (1*) to get

$$
n^{-1/2} \sum_{i \neq j} |m_{(\theta_0, h_n)}^{U}(X_i, X_j) - m_{(\theta_0)}^{U}(X_i, X_j)| = \frac{1}{2} h_n^T V_{\theta_0} h_n + h_n^T \mathbb{U}_n[\hat{m}_{(\theta_0)}^{U}] + o_p(1)
$$

$$
= \frac{1}{2} h_n^T V_{\theta_0} h_n + h_n^T \mathbb{G}_n[\hat{m}_{(\theta_0)}^{1}] + o_p(1),
$$

where we use the fact, from Hoeffding decomposition,

$$
\mathbb{U}_n[\hat{m}_{(\theta_0)}^{U}] = \sqrt{n} \sum_{i \neq j} |\hat{m}_{(\theta_0)}^{U}(X_i, X_j) - E_{X_1, X_2}[\hat{m}_{(\theta_0)}^{U}(X_1, X_2)]|
$$

$$
= \frac{1}{\sqrt{n}} \sum_{i = 1}^{n} |\hat{m}_{(\theta_0)}^{1}(X_i) - E_{X_1}[\hat{m}_{(\theta_0)}^{1}(X_1)]| + o_p(1) = \mathbb{G}_n[\hat{m}_{(\theta_0)}^{1}] + o_p(1),
$$

(19)
with \( \hat{m}_b \) as in the statement of the theorem.

The above statement is valid for both \( \hat{h}_n = \sqrt{n}(\hat{\theta}_n - \theta_0) \) and for \( \tilde{h}_n = -V_{\theta_0}^{-1}G_n\tilde{m}_b \). Upon substitution, we obtain

\[
\begin{align*}
\frac{n}{2} \sum_{i \neq j} [m_{\theta_0+h_n}^{U}(X_i, X_j) - m_{\theta_0}^{U}(X_i, X_j)] &= \frac{1}{2} \tilde{h}_n^T V_{\theta_0} \tilde{h}_n + \bar{h}_{\theta_0}^T G_n [\hat{m}_b] + o_p(1) \\
&\leq n \left( \frac{n}{2} \right)^{-1} \sum_{i \neq j} [m_{\theta_0+h_n}^{U}(X_i, X_j) - m_{\theta_0}^{U}(X_i, X_j)] \\
&= \frac{-1}{2} G_n [\hat{m}_b] \tilde{h}_n^T V_{\theta_0}^{-1} G_n [\hat{m}_b] + o_p(1)
\end{align*}
\]

where the inequality is from the definition of \( \bar{h}_n = \theta_0 + \tilde{h}_n/\sqrt{n} \) as a near-minimizer.

Taking the difference and completing the square, we get

\[
\frac{1}{2} \tilde{h}_n^T V_{\theta_0} \tilde{h}_n + \bar{h}_{\theta_0}^T G_n [\hat{m}_b] + o_p(1) \leq 0,
\]

and because \( V_{\theta_0} \) is nonsingular, the quadratic form on the left must converge to zero in probability. The same must be true for \( ||\tilde{h}_n + V_{\theta_0}^{-1}G_n \hat{m}_b ||_2 \).

To complete the proof, we need to show (*) and \( \sqrt{n}(\hat{\theta}_n - \theta_0) = O_p(1) \) hold.

**Proof of (*).**

Let \( f_b := \sqrt{n}(m_{\theta_0+h/\sqrt{n}}^{U} - m_{\theta_0}^{U}) - h^T \hat{m}_b^{U} \). As we are considering only sequences \( h_n \) that are bounded in probability, it suffices to show \( \sup_{h:||h||_2 \leq 1} |U_n[f_b]| \) goes to zero in probability. Again by Hoeffding decomposition, for any given random sequence \( h_n \) that is bounded in probability, \( U_n[f_h] = G_n[f_h] + E_n(h_n) \), where \( f_b \) is the first term in the Hoeffding decomposition of \( U_n[f_h] \) given by

\[
\begin{align*}
f^1_b &= \sqrt{n}(m_{\theta_0+h/\sqrt{n}}^{U} - m_{\theta_0}^{U}) - h^T \hat{m}_b^{U}, \\
m^1_b(x_1) &= 2E_{X_2}[m^{U}_b(x_1, X_2)] - E_{X_1, X_2}[m^{U}_b(X_1, X_2)],
\end{align*}
\]

and \( \hat{m}_b \) as defined in the statement of the theorem. According to Lemma 19.31 in Van der Vaart (2000), if \( \mathcal{F} := \{ m_b^1 : \theta \in [-K, K]^{1+r} \} \) is a Lipschitz class of functions,

\[
\sup_{h:||h||_2 \leq 1} \|G_n[f^1_b]\| \overset{P}{\to} 0.
\]

Now by Assumption 3, \( \theta \to m_b^1(x) \) is differentiable at \( \theta_0 \) for all \( x \in R \). Further, by triangle inequality,

\[
\begin{align*}
|m_b^1(x) - m_b^1(x)| &\leq 2E_{X_2}[m^{U}_b(x, X_2) - m^{U}_b(x, X_2)] + E_{X_1, X_2}[m^{U}_b(X_1, X_2) - m^{U}_b(X_1, X_2)] \\
&\leq m^1(x)||\theta_1 - \theta_2||_2,
\end{align*}
\]

where \( m^1(x) = (2E_{X_2}[\hat{m}^{U}(x, X_2)] + E_{X_1, X_2}(\hat{m}^{U}(X_1, X_2)), \hat{m}^{U} \) as in Eq. (19). Since \( X_1 \)'s have finite fourth moment, \( E[m^1(X_1)^2] < \infty \) and thus \( \mathcal{F} \) is a Lipschitz class.

Now we are left to show \( \sup_{h:||h||_2 \leq 1} |E_n(h)| \overset{P}{\to} 0 \) Let \( \mathcal{F}_h := \{ f_h : ||h||_2 \leq 1 \} \). According to Theorem 4.6 of Arcones and Gine (1993), \( \sup_{h:||h||_2 \leq 1} |E_n(h)| \overset{P}{\to} 0 \) if \( \mathcal{F}_h \) has a finite, integrable envelope function and both \( \mathcal{F}_h \) and \( \mathcal{F}_h^c := \{ f_h^c : ||h||_2 \leq 1 \} \) are Lipschitz classes about \( h = 0 \). \( \mathcal{F}_h \) has a finite, integrable envelope function

\[
\begin{align*}
\mathcal{F}(x_1, x_2) &= \hat{m}^{U}(x_1, x_2) + ||m_{\theta_0}(x_1, x_2)||_2 < \infty \text{ due to Assumption 2' and the Lipschitz property of } m_b^U.
\end{align*}
\]

\[
\begin{align*}
|f_h| &\leq \sqrt{n}(m_{\theta_0+h/\sqrt{n}}^{U} - m_{\theta_0}^{U}) - h^T \hat{m}_b \\
&\leq (\hat{m}^{U} + ||m_{\theta_0}||_2)||h||_2.
\end{align*}
\]
It is now straight-forward to check that $\mathcal{F}_h$ is a Lipschitz class about $h = 0$, and $\mathcal{F}_h'$ also, because it inherits the key properties from $\mathcal{F}_h$.

**Proof of** $\sqrt{n}(\hat{\theta}_n - \theta_0) = O_P(1)$.

The proof of $\sqrt{n}(\hat{\theta}_n(0, 0) - \theta_0(0, 0)) = O_P(1)$ can be found in Theorem 5.52 and Corollary 5.53 of Van der Vaart (2000), and is a standard M-estimation result. In essence, Theorem 5.52 shows that, under some regularity conditions, $P(\sqrt{n}\|\hat{\theta}_n - \theta_0\| > \alpha)$ can be bounded by $P(\|G_n[m_0]\| > \alpha'/\sqrt{n}) = P(\sqrt{n}|M_0(\theta) - M(\theta)| > \alpha')$, which is shown to go to zero via some maximal inequalities. Corollary 5.53 shows that the Lipschitz condition on $\{m_\theta : \theta \in [-K, K]^{1+r}\}$ is sufficient to satisfy the regularity conditions of the theorem.

We can extend Theorem 5.52 to show $\sqrt{n}(\hat{\theta}_n(\lambda_1, \lambda_2) - \theta_0(\lambda_1, \lambda_2))$, $\lambda_1, \lambda_2 \geq 0$ not both zero, by bounding $P(\sqrt{n}\|\hat{\theta}_n(\lambda_1, \lambda_2) - \theta_0(\lambda_1, \lambda_2)\| > \alpha)$ by

$$P(\|U_n[m_0]\| > \alpha') \leq P(\|G_n[m_0]\| + |E_n'(\theta)| > \alpha'),$$

where $E_n'$ is the remainder term after first-order projection of the U-process $U_n[m_0]$. It remains to show that for every sufficiently small $\delta > 0$,

$$\sup_{\|\theta - \theta_0\| \leq \delta} |E_n'(\theta)| \xrightarrow{P} 0,$$

(20)

which can be proven using the same reasoning for $\sup_{h, \|h\| \leq 1} |E_n(h)| \xrightarrow{P} 0$ in the proof of (*).

**Corollary 2.** Assume the same setting as Theorem 3. Then

$$\sqrt{n}(\hat{\theta}_n(\lambda_1, \lambda_2) - \theta_0(\lambda_1, \lambda_2)) \Rightarrow N(0, \Sigma_{\theta_0}(\lambda_1, \lambda_2)),$$

(21)

where $\Sigma_{\theta_0}(\lambda_1, \lambda_2) = A^{-1}BA^{-1}$,

$$A = A_{\theta_0}(\lambda_1, \lambda_2) = \nabla^2_{\theta} \mathbb{E}[m_{(\theta, \lambda_1, \lambda_2)}'(X_1)\bigg|_{\theta = \theta_0}$$

$$= \nabla^2_{\theta} \left[ \frac{1}{1 - \beta} E_{\theta}(X_1) + \lambda_1 w^\top \Sigma w + \lambda_2 Var(z_\theta(X_1)) \right] \bigg|_{\theta = \theta_0},$$

$$B = B_{\theta_0}(\lambda_1, \lambda_2) = \mathbb{E}[\nabla_{\theta_0} m_{(\theta, \lambda_1, \lambda_2)}(X_1)\nabla_{\theta_0} m_{(\theta, \lambda_1, \lambda_2)}(X_1)^\top]$$

where

$$\nabla_{\theta} m_{(\theta, \lambda_1, \lambda_2)}(x) = \left[ 1 - \frac{1}{1 - \beta} I + 2\lambda_2 E[(z_\theta(X) - \mathbb{E}z_\theta(X))(I + E\mathbb{I})] \right. $$$$ - \frac{1}{1 - \beta} L^\top X \mathbb{I} - \lambda_0 L^\top X + 2\lambda_1 L^\top (X - \mu)(X - \mu)^\top w + 2\lambda_2 E[(z_\theta(X) - \mathbb{E}z_\theta(X))(L^\top X \mathbb{I} + E L^\top X \mathbb{I})] \bigg)^\top,$$

and $\mathbb{I} = I(z_\theta(X)) \geq 0$.

**Remarks.**

1. For asymptotics of $\hat{w}_n(\lambda_1, \lambda_2)$ we have

$$\sqrt{n}(\hat{w}_n(\lambda_1, \lambda_2) - w_0(\lambda_1, \lambda_2)) \Rightarrow N(0, \Sigma_{w_0}(\lambda_1, \lambda_2)),$$

(22)

where $\Sigma_{w_0}(\lambda_1, \lambda_2) = (0_p, L)\Sigma_{\theta_0}(\lambda_1, \lambda_2)(0_p, L)^\top$.  

2. Setting $\lambda_1, \lambda_2 = 0$, we get back the unregularized mean-CVaR problem.

3. Asymptotic distribution of the efficient frontier.

With Eq. (22), we can state the distribution of the true efficient frontier — that is, the distribution of $\hat{w}_n(\lambda_1, \lambda_2)\top \mu$ and $g(\hat{w}_n(\lambda_1, \lambda_2)) := CVaR(-\hat{w}_n(\lambda_1, \lambda_2)\top X_{n+1}; \beta)$, where $X_{n+1} \sim F$, independent of $X_1, \ldots, X_n$. For the portfolio mean, we have

$$\sqrt{n}(\hat{w}_n(\lambda_1, \lambda_2)\top \mu - w_0(\lambda_1, \lambda_2)\top \mu) \Rightarrow N(0, \sigma_0^2),$$

and for the true CVaR, by the delta Method

$$\sqrt{n}(g(\hat{w}_n(\lambda_1, \lambda_2)) - g(w_0(\lambda_1, \lambda_2))) \Rightarrow N\left(0, \frac{\beta}{1-\beta} \frac{(1-\beta)^2}{2\sigma_0^2} \right).$$

C.3. Example: Verification of theory when $X \sim \mathcal{N}(\mu, \Sigma)$

In the following, we provide the detailed computation of $\Sigma_{\hat{w}_0}(0,0)$ for the unregularized solution $\hat{\theta}_n(0,0)$ when $X \sim \mathcal{N}(\mu, \Sigma)$.

**Lemma 2.** Suppose $X \sim \mathcal{N}(\mu, \Sigma)$. Then

$$z_{\hat{w}_0}(X) = -w_0^\top X - \alpha_0 \sim \sigma_D N(-\Phi^{-1}(\beta), 1),$$

and

$$p_0 = f_{-w_0^\top X}(0) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp \left\{-\frac{1}{2\sigma_0^2} (\Phi^{-1}(\beta))^2 \right\},$$

where $\sigma_0 = \sqrt{w_0^\top \Sigma w_0}$. Then $\Sigma_{\hat{w}_0}(0,0) = A_0^{-1} B_0 A_0^{-1}$, where $A_0$, $B_0$ are symmetric matrices with

$$A_0(1,1) = \frac{p_0}{1-\beta},$$

$$A_0(j, l) = \frac{p_0}{(1-\beta)} \mathbb{E}[L_j^\top X L_l^\top X | z_{\hat{w}_0}(X) = 0] \quad \text{for} \quad 2 \leq j, l \leq p,$$

$$A_0(1, j) = -\frac{p_0}{(1-\beta)} \mathbb{E}[L_j^\top X | z_{\hat{w}_0}(X) = 0] \quad \text{for} \quad 2 \leq j \leq p,$$

where $L_j$ is the $j$-th column of $L$, and

$$B_0(1,1) = \frac{\beta}{1-\beta},$$

$$B_0(j, l) = \lambda_0^2 (L_j^\top \Sigma L_l + L_j^\top \mu L_l^\top \mu) + \frac{1}{2(1-\beta)} \left( \frac{1}{1-\beta} + 2\lambda_0 \right) \mathbb{E}[L_j^\top X L_l^\top X | z_{\hat{w}_0}(X) \geq 0] \quad \text{for} \quad 2 \leq j, l \leq p,$$

$$B_0(1, j) = 0 \quad \text{for} \quad 2 \leq j \leq p.$$

**Proof.** This is a straight-forward application of Corollary 2 for the case $X \sim \mathcal{N}(\mu, \Sigma)$.

Let us now compare the asymptotic results derived above with simulations with finite number of observations. Consider 5 assets, a range of observations ($n = 250, 500, 1000, 2000$) and $X \sim \mathcal{N}(\mu_{sim}, \Sigma_{sim})$, where the model parameters are the same as the model parameters of the first five assets used in Sec. 3. For simulations, we solve the mean-CVaR problem with dualized mean constraint:

$$\min_w \overline{CVaR}_n(-w^\top X; \beta) - \lambda_0 w^\top \hat{\mu}_n$$

$$\text{s.t.} \quad w^\top I_p = 1,$$

and follow steps similar to Sec. 3.

In Fig. 4, we summarize the empirical frontiers by plotting their averages and indicating $\pm 1/2$ standard deviation error bars, in both true mean (vertical) and true risk estimations (horizontal) in grey. The population frontier is also plotted, and is shown in green, and the theoretical $\pm 1/2$ standard deviations of mean and risk estimations are juxtaposed with the empirical error bars in red. We make a couple of observations:
Figure 4  Comparison of theoretical (red) and simulated (grey) distributions of the sample efficient frontier when $X \sim \mathcal{N}(\mu_{\text{sim}}, \Sigma_{\text{sim}})$ for increasing number of observations $n = [250, 500, 1000, 2000]$. The error bars indicate $\pm 1/2 \text{ std errors}$ in the mean and CVaR. Green is the whole population efficient frontier, and blue indicates the portion that corresponds to the target return range considered for the simulations. Observe that the asymptotic variance calculated theoretically (red bars) approach the simulated variance (grey bars) with increasing $n$.

1. With increasing $n$, the theoretical error bars approach the simulated ones, as expected.
2. The theory seems to better predict the mean estimation error (vertical) better than the risk estimation error (horizontal). With finite $n$, the mean estimation error, which is computed using Eq. (1), depends only on one approximate quantity $\Sigma_{w_0}(0, 0)$, whereas the risk estimation error, computed using Eq. (24), depends on $\Sigma_{w_0}(0, 0)$ and $w_0$. Although $\hat{w}_n$ is a consistent estimator of $w_0$ asymptotically, with finite $n$ the difference does play a role, as shown by the relative inaccuracy of the horizontal error bars compared to the vertical ones. The finite sample bias also explains the gap in the positions of the population and simulated efficient frontiers.

Let us now derive asymptotic properties of the regularized solution $\hat{\theta}_n(\lambda_1, \lambda_2)$, $\lambda_1, \lambda_2 \geq 0$, when $X \sim \mathcal{N}(\mu, \Sigma)$. First, we show that when $X \sim \mathcal{N}(\mu, \Sigma)$, regularizing the variance of CVaR estimation is redundant if one regularizes the sample variance of the mean.

**Lemma 3.** Suppose $X \sim \mathcal{N}(\mu, \Sigma)$ and let $z_\theta(X) = -\alpha - w^T X$. Then $z_\theta(X) \sim \mathcal{N}(\mu_1, \sigma_1^2)$ where $\mu_1 = -\sigma_1 \Phi^{-1}(\beta)$, $\sigma_1^2 = w^T \Sigma w$, and

$$\text{Var} \left[ \max(z_\theta(X), 0) \right] = C(\beta) \sigma_1^2,$$
where \( C(\beta) \) is a constant that only depends on \( \beta \). Thus regularizing the sample variance of CVaR via \( P_2(w) = \sqrt{\text{Var}_n[z_n(X), 0]} \leq U_2 \) is redundant if one regularizes the sample variance of the mean via \( P_1(w) = w^\top \hat{\Sigma}_n w = \hat{\sigma}_{1,n}^2 \leq U_1 \).

**Proof.** Straight-forward calculations show
\[
\text{Var}[\max(z_n(X), 0)] = \left\{ (|\Phi^{-1}(\beta)|)^2 + 1 \right\} (1 - \beta) - 3\Phi^{-1}(\beta)f_{z_0}[\Phi^{-1}(\beta)] \sigma_1^2,
\]
where \( f_{z_0} \) is the pdf of the standard normal random variable \( Z_0 \).

The implication now is that when \( X \sim \mathcal{N}(\mu, \Sigma) \), we need only consider \( \lambda_1 \geq 0, \lambda_2 = 0 \) to characterize the asymptotic properties of the regularized solution, which we describe below.

**Lemma 4.** Suppose \( X \sim \mathcal{N}(\mu, \Sigma) \). Then
\[
\Sigma_{\beta_0}(\lambda_1, 0) = A_{\lambda_1}^{-1} B_1 A_{\lambda_1}^{-1},
\]
where \( A_1, B_1 \) are symmetric matrices with
\[
A_1(1, 1) = A_0(1, 1) \quad A_1(j, l) = A_0(j, l) + \lambda_1 L_j^\top \Sigma L_l \quad \text{for } 2 \leq j, l \leq p \quad A_1(1, j) = A_0(1, j) \quad \text{for } 2 \leq j \leq p
\]
where \( L_j \) is the \( j \)-th column of \( L \), and
\[
B_1(1, 1) = B_0(1, 1) \quad B_1(j, l) = B_0(j, l) + \lambda_1 \mathbb{E}[b_{0,j} b_{1,l} + b_{0,l} b_{1,j} + \lambda_1 b_{1,j} b_{1,l}] \quad \text{for } 2 \leq j, l \leq p \quad B_1(1, j) = B_0(1, j) + \lambda_1 \mathbb{E}[b_{0,j} b_{1,j}] \quad \text{for } 2 \leq j \leq p
\]
where for \( 2 \leq j, l \leq p \),
\[
\mathbb{E}[b_{0,j} b_{1,l}] = \frac{2}{1 - \beta} \mathbb{E}[L_j^\top X L_l^\top (X - \mu) w^\top (X - \mu) \mathbb{I}(z_n(X) \geq 0)] - 2\lambda_0 L_j^\top \mu L_l^\top \Sigma w
\]
\[
\mathbb{E}[b_{1,j} b_{1,l}] = 4\mathbb{E}[L_j^\top (X - \mu)(X - \mu)^\top L_l^\top (X - \mu)(X - \mu)^\top w]
\]
\[
\mathbb{E}[b_{0,j} b_{1,l}] = 2L_j^\top \Sigma w - \frac{2}{1 - \beta} \mathbb{E}[L_j^\top (X - \mu) w^\top (X - \mu) \mathbb{I}(z_n(X) \geq 0)].
\]

**Proof.** This is a straight-forward application of Corollary 2 for the case \( X \sim \mathcal{N}(\mu, \Sigma) \).

The PBR method with only a regularization on the mean estimation is a linear combination of the empirical mean-CVaR problem (CVaR-SAA) and the SAA Markowitz problem
\[
\hat{\omega}_n^M = \arg\min \ w^\top \hat{\Sigma}_n w \\
\text{s.t. } \ w^\top \hat{\mu}_n = R \\
\ w^\top 1_p = 1, \quad (\text{Mark-SAA})
\]
because the regularization is precisely the portfolio variance estimate \( w^\top \hat{\Sigma}_n w \). In particular, the single-regularization PBR problem approaches (Mark-SAA) with increasing \( \lambda_1 \). In Figure 5, we plot 1 std of \( w_n^\top \hat{\Sigma}_n = \lambda_1, 0 \) of \( \mu \) and \( CVaR(-w_n^\top \hat{\Sigma}_n 1_p; \beta) \) for the single-regularization problem as \( \lambda_1 \) is increased, for different values of \( \lambda_0 \), computed using Lemma 4. Observe that the asymptotic standard deviations for both portfolio mean and CVaR decrease with increasing \( \lambda_1 \), uniformly in \( \lambda_0 \). Now when \( X \) has a Gaussian distribution, the population solution of the mean-CVaR problem and the Markowitz problem are the same (Rockafellar and Uryasev (2000), De Giorgi (2002)). As sample mean and covariance are more stable estimators than sample mean and CVaR, the SAA Markowitz solution is more stable than the SAA mean-CVaR solution, and our theoretical result summarized by Fig. 5 is consistent with this fact.
C.4. Computation of key statistics

Given the distribution for $X$, both $A_0 = A_{\theta_0}(0,0)$ and $B_0 = B_{\theta_0}(0,0)$ are computable. The lemma below computes the key quantities that constitute $A_0$ and $B_0$ when $X \sim \mathcal{N}(\mu, \Sigma)$.

**Lemma 5.** Suppose $X \sim \mathcal{N}(\mu, \Sigma)$, and $z_\theta(X) = -\alpha - w^T X \sim \mathcal{N}(\mu_1, \sigma_1^2)$, where $\mu_1 = -\sigma_1 \Phi^{-1}(\beta)$ and $\sigma_1^2 = w^T \Sigma w$. Then

\begin{align*}
p_0 &= P(z_\theta(X) = 0) = \frac{1}{\sqrt{2\pi} \sigma_1} \exp\left(-\frac{1}{2} \Phi^{-1}(\beta)^2\right) \quad (25) \\
\mathbb{E}[\max(z_\theta(X),0)] &= \frac{\sigma_1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} \Phi^{-1}(\beta)^2\right) - \sigma_1 (1 - \beta) \Phi^{-1}(\beta) \quad (26) \\
\mathbb{E}[L_j^T X | Z_1 \geq 0] &= (1 - \beta) (L_j^T \mu - \Phi^{-1}(\beta) \frac{L_j^T \Sigma w}{\sigma_1}) - \frac{L_j^T \Sigma w}{\sigma_1} \mathbb{E}[\max(z_\theta(X),0)] \quad (27) \\
\mathbb{E}[L_j^T X | Z_1 = 0] &= L_j^T \mu - \Phi^{-1}(\beta) \frac{L_j^T \Sigma w}{\sigma_1} \quad (28) \\
\mathbb{E}[L_j^T X L_i^T X | Z_1 \geq 0] &= \frac{1}{4} (g(\mu_1, (L_i + L_j)^T \mu, \mu_1, \sigma_1, \sigma_2, -(L_j + L_i)^T \Sigma w_1) \\
&\quad - g(\mu_1, (L_j - L_i)^T \mu, \mu_1, \sigma_1, \sigma_2, -(L_j - L_i)^T \Sigma w_1) \quad (29) \\
\mathbb{E}[L_j^T X L_i^T X | Z_1 = 0] &= \frac{1}{4} (h(\mu_1, (L_i + L_j)^T \mu, \mu_1, \sigma_1, \sigma_2, -(L_j + L_i)^T \Sigma w_1) \\
&\quad - h(\mu_1, (L_j - L_i)^T \mu, \mu_1, \sigma_1, \sigma_2, -(L_j - L_i)^T \Sigma w_1)) \quad (30)
\end{align*}

where

\begin{align*}
g(\mu_1, \mu_2, \sigma_1, \sigma_2, \sigma_{12}) &= (1 - \beta) [\mu_2^2 + \sigma_2^2] + p_{012} \left[-\Phi^{-1}(\beta) \frac{\sigma_{12}}{\sigma_1} + 2\mu_2 \right] \\
h(\mu_1, \mu_2, \sigma_1, \sigma_2, \sigma_{12}) &= (\mu_2 + \frac{\sigma_{12}}{\sigma_1} \Phi^{-1}(\beta))^2 + \sigma_2^2 - \frac{\sigma_{12}^2}{\sigma_1^2}.
\end{align*}
Proof. Equations (25) and (26) are straightforward, for Eqs. (27)–(30), we use the fact that if \( Z_1 \sim N(\mu_1, \sigma_1) \) and \( Z_2 \sim N(\mu_2, \sigma_2) \),
\[
Z_2 | Z_1 = N(\mu_2 + \sigma_{12} / \sigma_1^2 (Z_1 - \mu_1), \sigma_2^2 - \sigma_{12}^2 / \sigma_1^2),
\]
(31)
where \( \sigma_{12} = \text{Cov}(Z_1, Z_2) \).

- Terms involving only \( L_j^\top X \).

Note that from (31), \( E[Z_2 | Z_1 = 0] = \mu_2 - \frac{\sigma_{12}}{\sigma_1^2}\mu_1 \). Let \( Z_2 = L_j^\top X \), and recall that \( E(L_j^\top X) = L_j^\top \mu \) and \( E(Z_1) = -\sigma_1 \Phi^{-1}(\beta) \). Also, note that \( \sigma_{12} = -L_j^\top \Sigma w \). After some algebra, we get (28).

Since we know the distribution of \( Z_2 | Z_1 \), we have
\[
E[Z_2 I(Z_1 \geq 0)] = E[I(Z_1 \geq 0)(\mu_2 + \frac{\sigma_{12}}{\sigma_1^2}(Z_1 - \mu_1))]
\]
\[
= (1 - \beta)(\mu_2 - \frac{\sigma_{12}}{\sigma_1^2}\mu_1) + \frac{\sigma_{12}}{\sigma_1^2}E[Z_1 I(Z_1 \geq 0)]
\]
\[
= (1 - \beta)(L_j^\top \mu - \Phi^{-1}(\beta)\frac{L_j^\top \Sigma w}{\sigma_1}) - \frac{L_j^\top \Sigma w}{\sigma_1}E[\max(Z_1, 0)]
\]

- Terms involving \( L_j^\top X L_i^\top X \).

To compute \( E[L_j^\top X L_i^\top X | Z_1 \geq 0] \) and \( E[L_j^\top X L_i^\top X | Z_1 = 0] \), first note that
\[
E[L_j^\top X L_i^\top X I(Z_1 \geq 0)] = \frac{1}{4}E \left[ \left( (L_j^\top X + L_i^\top X)^2 - (L_j^\top X - L_i^\top X)^2 \right)^2 | I(Z_1 \geq 0) \right].
\]
and similarly
\[
E[L_j^\top X L_i^\top X | Z_1 = 0] = \frac{1}{4}E \left[ \left( (L_j^\top X + L_i^\top X)^2 - (L_j^\top X - L_i^\top X)^2 \right)^2 | I(Z_1 = 0) \right].
\]
Hence it is sufficient to first find expressions for \( E[Z_2^2 I(Z_1 \geq 0)] \) and \( E[Z_2^2 | Z_1 = 0] \) for some normal \( Z_2 \), then apply the resulting formulae to \( Z_2 = (L_j \pm L_i)^\top X \). This results in \( \mu_2 = (L_j \pm L_i)^\top \mu \), \( \sigma_{12} = -(L_i \pm L_i)^\top \Sigma w \) and \( \sigma_2^2 = (L_j \pm L_i)^\top \Sigma (L_j \pm L_i) \).

From tower property and the conditional distribution of \( Z_2 | Z_1 \),
\[
E[Z_2^2 I(Z_1 \geq 0)] = E[I(Z_1 \geq 0) \left( (\mu_2 + \frac{\sigma_{12}}{\sigma_1^2}(Z_1 - \mu_1))^2 + \sigma_2^2 - \frac{\sigma_{12}^2}{\sigma_1^4} \right)].
\]
By simple computations,
\[
E[(Z_1 - \mu_1)^2 | Z_1 \geq 0] = \frac{\sigma_1}{\sqrt{2\pi}} \exp(-\mu_1^2/(2\sigma_1^2)) = \sigma_1^2 f_{Z_1}(0) = \sigma_1^2 p_0 \quad \text{and}
\]
\[
E[(Z_1 - \mu_1)^2 | Z_1 = 0] = \sigma_1^2 (\mu_1 p_0 + (1 - \beta)).
\]
Now \( \mu_1 / \sigma_1 = -\Phi^{-1}(\beta) \), and
\[
E[Z_2^2 I(Z_1 \geq 0)] = (1 - \beta) \left[ \mu_2^2 + \sigma_2^2 \right] + p_0 \left[ \mu_1 \frac{\sigma_{12}^2}{\sigma_1^4} + 2\sigma_{12}^2 \mu_2 \right]
\]
\[
= (1 - \beta) \left[ \mu_2^2 + \sigma_2^2 \right] + p_0 \sigma_{12} \left[ -\Phi^{-1}(\beta) \frac{\sigma_{12}}{\sigma_1} + 2\mu_2 \right]
\]
\[
:= g(\mu_1, \mu_2, \sigma_1, \sigma_2, \sigma_{12})
\]
Similarly,
\[
E[Z_2^2 | Z_1 = 0] = (\mu_2 - \frac{\sigma_{12}}{\sigma_1^2}\mu_1)^2 + \sigma_2^2 - \frac{\sigma_{12}^2}{\sigma_1^4} = (\mu_2 + \frac{\sigma_{12} \Phi^{-1}(\beta)}{\sigma_1})^2 + \sigma_2^2 - \frac{\sigma_{12}^2}{\sigma_1^4} := h(\mu_1, \mu_2, \sigma_1, \sigma_2, \sigma_{12})
\]
References


