Integrating prediction in mean-variance portfolio optimization

Andrew Butler and Roy H. Kwon
University of Toronto
Department of Mechanical and Industrial Engineering

September 20, 2021

Abstract

Many problems in quantitative finance involve both predictive forecasting and decision-based optimization. Traditionally, predictive models are optimized with unique prediction-based objectives and constraints, and are therefore unaware of how those predictions will ultimately be used in the context of their final decision-based optimization. We present a stochastic optimization framework for integrating regression based predictive models in a mean-variance portfolio optimization setting. Closed-form analytical solutions are provided for the unconstrained and equality constrained case. For the general inequality constrained case, we make use of recent advances in neural-network architecture for efficient optimization of batch quadratic-programs. To our knowledge, this is the first rigorous study of integrating prediction in a mean-variance portfolio optimization setting. We present several historical simulations using global futures data and demonstrate the benefits of the integrated approach in comparison to the decoupled alternative.

Keywords: Data driven stochastic-programming, Regression, Mean-variance optimization, Empirical risk minimization, Differentiable neural networks

1 Introduction

Many problems in quantitative finance can be characterized by the following elements:

1. A sample data set $\mathbf{Y} = \{\mathbf{y}^{(1)}, ..., \mathbf{y}^{(m)}\} = \{\mathbf{y}^{(i)}\}_{i=1}^m$ of uncertain quantities of interest, $\mathbf{y}^{(i)} \in \mathbb{R}^{d_y}$, such as asset returns.
2. A decision, $\mathbf{z} \in \mathbb{R}^{d_z}$, often constrained to some feasible region $\mathbb{Z} \subseteq \mathbb{R}^{d_z}$.
3. A *nominal objective (cost) function*, $c\colon \mathbb{R}^{d_z} \times \mathbb{R}^{d_y} \to \mathbb{R}$, to be minimized over decision variable $\mathbf{z} \in \mathbb{Z}$ in the context of the observed realization $\mathbf{y}^{(i)}$.

For example, in portfolio management we are often presented with the following problem: for a particular observation of asset returns, $\mathbf{y}^{(i)}$, the objective is to construct a vector of assets weights, $\mathbf{z}^*(\mathbf{y}^{(i)})$, that minimizes the nominal cost, $c(\mathbf{z}, \mathbf{y}^{(i)})$ and adheres to the constraint set $\mathbb{Z}$. A common choice for nominal cost is the Markowitz mean-variance quadratic objective [35], with typical constraints being that the weights be non-negative and sum to one. Of course, the realization of asset returns, $\{\mathbf{y}^{(i)}\}_{i=1}^m$, are not directly observable at the decision time and instead must be estimated through associated auxiliary data $\mathbf{X} = \{\mathbf{x}^{(1)}, ..., \mathbf{x}^{(m)}\}$, of covariates of $\mathbf{Y}$. Let $f \colon \mathbb{R}^{d_x} \times \mathbb{R}^{d_\theta} \to \mathbb{R}^{d_y}$ denote the $\boldsymbol{\theta}$-parameterized prediction model for estimating $\hat{\mathbf{y}}$. In this paper, we consider regression based prediction models of the form:

$$\hat{\mathbf{y}}^{(i)} = f(\mathbf{x}^{(i)}, \boldsymbol{\theta}) = \boldsymbol{\theta}^T \, \mathbf{x}^{(i)},$$

with regression coefficient matrix $\boldsymbol{\theta} \in \mathbb{R}^{d_\theta}$.

In most applications, estimating $\hat{\mathbf{y}}^{(i)}$ requires solving a second, independent *prediction optimization problem*, with its own unique objective function (e.g. least-squares or maximum likelihood) and feasible region Θ. Continuing with the example above; in order to generate mean-variance efficient portfolios we must supply, at a minimum, an estimate of expected asset returns and covariances. A prototypical framework would first estimate the conditional expectations of asset returns and covariances by ordinary least-squares (OLS) regression and then 'plug-in' those estimates to a mean-variance quadratic program (see for example Goldfarb and Iyengar [24], Clarke et al. [16] Chen et al. [14]).

While the focus of operations research (OR) is on optimal decision making, machine learning (ML), on the other hand, focuses on providing optimal predictions. As exemplified above, OR-based optimization and ML-based predictions are often decoupled processes; first predict, then optimize. Indeed, at first glance, a 'predict then optimize' approach seems reasonable, if not optimal. That is to say, if a predictive model provides perfect forecast accuracy ($\hat{\mathbf{y}}^{(i)} = \mathbf{y}^{(i)}$), then generating optimal decisions amounts to solving a deterministic nominal optimization problem.

While a perfect predictive model would lead to optimal decision making, in reality, all predictive models do make some error. As such, an inefficiency exists in the 'predict then optimize' paradigm: predictive models are optimized with unique *prediction-based* objectives and constraints, and thus are unaware of how those predictions will ultimately be used in the context of the nominal *decision-based* optimization.

Recently there has been a growing body of research on data-driven decision making and the relative merits of decoupled versus integrated predictive decision-making. In this paper, we follow the work of Donti et al. [17], Elmachtoub and Grigas [19] and others, and propose the use of an integrated prediction and optimization framework with direct applications to mean-variance portfolio optimization. The key distinction is that under the integrated setting, the model parameters, $\boldsymbol{\theta}$, are estimated in order to yield the 'best' decisions, not to provide the 'best' predictions. Specifically, we select $\boldsymbol{\theta}$ such that the resulting optimal decision policy, $\mathbf{z}^*(\mathbf{x}, \boldsymbol{\theta})$, produces the lowest average realized nominal cost:

$$\begin{aligned} \underset{\boldsymbol{\theta} \in \Theta}{\text{minimize}} \quad & \mathbb{E}[c(\mathbf{z}^*(\mathbf{x}, \boldsymbol{\theta}), \mathbf{y})] \\ \text{subject to} \quad & \mathbf{z}^*(\mathbf{x}, \boldsymbol{\theta}) = \underset{\mathbf{z} \in \mathbb{Z}}{\operatorname{argmin}} \, c(\mathbf{z}, f(\mathbf{x}, \boldsymbol{\theta})). \end{aligned} \tag{1}$$

We denote the objective function of the integrated prediction and optimization (IPO) problem as $L(\boldsymbol{\theta}) = \mathbb{E}[c(\mathbf{z}^*(\mathbf{x}, \boldsymbol{\theta}), \mathbf{y})]$, where $\mathbb{E}$ is the expectation operator. Solving the IPO problem over decision variables $(\boldsymbol{\theta}, \mathbf{z})$ is challenging for several reasons. First, even in the case where the nominal program is convex, the resulting integrated program is likely not convex in $\boldsymbol{\theta}$ and therefore we have no guarantee that a particular local solution is globally optimal. Secondly, in the case where $L(\boldsymbol{\theta})$ is differentiable, computing the gradient, $\nabla_{\boldsymbol{\theta}} L$, remains difficult as it requires differentiation through the argmin operator. Furthermore, solving program (1) through iterative descent methods can be computationally demanding as at each iteration we must solve several instances of $\mathbf{z}^*(\mathbf{x}^{(i)}, \boldsymbol{\theta})$.

In this paper we address the aforementioned challenges and provide an efficient framework for integrating regression predictions into a mean-variance portfolio optimization setting. To our knowledge this is the first rigorous study of a general and direct method for integrating prediction in a mean-variance portfolio optimization setting, with experimentation using real asset price data. The remainder of the paper is outlined as follows. We first review the growing body of literature in the field of integrated predictive decision-making and summarize our primary contributions. In Section 2 we formalize the IPO problem in a general setting. In Section 3 we describe the classical mean-variance portfolio optimization problem and provide the integrated formulation. We present the current state-of-the-art approach for solving general integrated 'predict then optimize' problems in which the nominal program contains linear equality and inequality constraints. We follow the work of Amos and Kolter [2] and describe a first-order method in which the gradient of the nominal program with respect to the prediction model parameters is obtained by implicit differentiation. We then consider several special instances of the IPO mean-variance optimization problem. In particular, we demonstrate that when the nominal MVO program is either unconstrained or contains only linear equality constraints then the IPO problem can be recast as a convex quadratic program and solved analytically. We discuss the sampling distribution properties of the optimal IPO regression coefficients and demonstrate that the IPO solution explicitly minimizes the tracking-error to ex-post optimal mean-variance portfolios.

In Section 4.1 we perform several simulation studies using synthetically generated data. We first compare the IPO and OLS models in an unconstrained and linear equality constrained setting and demonstrate the resilience of the IPO model to estimation error in both the mean and covariance. In Sections 4.2 we discuss the computational challenges of the current state-of-the art solution based on implicit differentiation and first-order gradient descent. We demonstrate the computational advantage of the closed-form IPO solution, which guarantees optimality and is approximately an order of magnitude faster than the state-of-the-art iterative method. In Section 4.3 we revisit the more general IPO program whereby the nominal MVO program contains inequality constraints. A simulation study demonstrates that the analytical IPO solution, with inequality constraints removed, can provide superior out-of-sample performance over a wide range of problem parameterizations. Finally in Section 5 we conclude with a simulation study using global futures data and demonstrate that the IPO methodology can provide lower realized costs and improved economic outcomes in comparison to the 'predict, then optimize' alternative.

1.1 Existing Literature

In recent years there has been a growing body of research on data-driven decision making and the relative merits of decoupled versus integrated predictive decision-making. For example, Ban and Rudin [4] present a direct empirical risk minimization approach using nonparametric kernel regression as the core prediction method. They consider a data-driven newsvendor problem and demonstrate that their approach outperforms the best-practice benchmark when evaluated out-of-sample. More recently, Kannan et al. [33] present three frameworks for integrating machine learning prediction models within a stochastic optimization setting. Their primary contribution is in using the out-of-sample residuals from leave-one-out prediction models to generate scenarios which are then optimized in the context of a sample average approximation nominal program. Their frameworks are flexible and accommodate parametric and nonparametric prediction models, for which they derive convergence rates and finite sample guarantees.

Bertsimas and Kallus [6] present a general framework for optimizing a conditional stochastic approximation program whereby the conditional density is estimated through a variety of parametric and nonparametric machine learning methods. They generate locally optimal decision policies within the context of the nominal optimization problem and consider the setting where the decision policy affects subsequent realizations of the uncertainty variable. They also consider an empirical risk minimization framework for generating predictive prescriptions and discuss the relative trade-offs of such an approach.

Recently, Elmachtoub and Grigas [19] proposed replacing the prediction-based loss function with a convex surrogate loss function that optimizes prediction variables based on the decision error induced by the prediction, as measured by the nominal objective function. They demonstrate that their 'smart predict, then optimize' (SPO) loss function attains Fisher consistency with the popular least-squares loss function and show through example that optimizing predictions in the context of nominal objectives and constraints can lead to improved overall decision error. The SPO loss function however is limited to linear nominal objective functions, and despite convexity can be computationally demanding due to repeated evaluation of the nominal program.

Our approach is most similar to, and is largely inspired by, the work of Amos and Kolter [2] and Donti et al. [17]. Indeed, a fundamental challenge in optimizing the integrated problem (1) through gradient-descent is in computing the gradient of the objective function with respect to the prediction model parameters. Specifically, while computing the gradient, $\partial L/\partial \mathbf{z}^*$, is relatively straightforward, computing the Jacobian, $\partial \mathbf{z}^* / \partial \boldsymbol{\theta}$, is complicated by virtue of the argmin operator. Amos and Kolter [2], however, demonstrate that for general quadratic programs, the solution to the system of equations provided by the KKT conditions at optimality, $\mathbf{z}^*$, provide the necessary ingredients for computing the desired gradient, $\partial L/\partial \boldsymbol{\theta}$, and therefore it is not necessary to calculate the Jacobian, $\partial \mathbf{z}^* / \partial \boldsymbol{\theta}$, explicitly. Furthermore, Amos and Kolter [2], and more recently Agrawal et al. [1], propose an efficient framework for embedding differentiable optimization problems as layers in an end-to-end trainable neural network. Such an approach is advantageous as it allows for use of efficient backpropogation technology, GPU computing and the plethora of robust stochastic gradient descent optimization routines.

Donti et al. [17] present a direct application of the aforementioned work and propose an end-to-end stochastic programming approach for estimating the parameters of probability density functions in the context of their final task-based loss function. They consider applications from power scheduling and battery storage and focus specifically on parametric models in a stochastic programming setting. They demonstrate that their task-based end-to-end approach can result in lower realized costs in comparison to traditional maximum likelihood estimation and a black-box neural network.

1.2 Main Contributions

While our methodology follows closely to that of Donti et al. [17] and Elmachtoub and Grigas [19], in this paper we provide several notable differences and extensions. First, whereas Donti et al. [17] considers the parametric density estimation problem in a stochastic optimization setting, we consider a regression based estimation with deterministic nominal optimization problem. Secondly, we consider applications from portfolio optimization; specifically the case where the nominal program is a convex mean-variance quadratic program. Moreover, we demonstrate that in the case where the solution to the nominal program is linear in the parameter $\boldsymbol{\theta}$, then we can efficiently make use of second-order Hessian information and optimize the integrated problem analytically, or more generally through standard quadratic programming.

The IPO program presented in Equation (1) is the most direct and perhaps simplest expression of an integrated prediction and optimization problem. To our knowledge this is the first rigorous study of these equations in a mean-variance portfolio optimization setting. While Elmachtoub and Grigas [19] provide an application to a mean-variance optimization (MVO) problem, their framework requires optimization of a customized and more complicated SPO surrogate loss function. Moreover, the SPO framework can only support optimization problems with a linear objective and therefore the integration is limited to prediction models that remain linear across all relevant problem variables.

In contrast, the simplicity of our IPO formulation lends itself to a very practical and approachable integrated prediction and portfolio optimization solution. We demonstrate that, in many cases, the integrated problem is no more complicated than a least-squares problem, which can be readily solved through standard quadratic programming. We consider the MVO problem under several constraint settings using both univariate and multivariate regression for the prediction model. Under special circumstances, discussed below, the IPO program emits a closed-form analytical solution and thus obviates the need for a more computationally demanding iterative optimization algorithm. Furthermore, unlike many machine-learning applications, the required action of the inverse of the Hessian of the IPO solution is computationally tractable in most practical settings. This is discussed in more detail in Section 3.2.

In this paper, asset mean returns are estimated through linear regression, whereas the asset covariance matrices are estimated by traditional weighted moving average approach [5, 44]. This is supported by the observation that asset mean returns are both nonstationary and heterogeneous and are therefore likely to be dependent on auxiliary feature data [20, 28, 37], whereas variance and covariances are typically much more stable and exhibit strong autocorrelation effects [8, 18, 42]. Moreover, Chopra and Ziemba [15] and Best and Grauer [7] report that MVO portfolio weights are an order of magnitude more sensitive to the estimate of asset mean returns compared to estimates of asset covariances.

The choice of linear regression model is deliberate and motivated by the long established history of regression forecasting in the financial literature (see for example, [21, 22, 23]). Indeed, asset returns are often characterized as time-varying and reactive, and typically exhibit extremely low signal-to-noise ratios (SNRs) [30]. As a result, low variance models, like simple linear regression, tend to generalize out-of-sample and are often preferred over models of higher complexity. That said, the general IPO framework, discussed in Section 3.1, can fully support linear and nonlinear prediction models for all variables of the nominal problem, including estimates of expected returns, expected covariances and constraint variables.

Numerical experiments are performed on both synthetically generated data and a universe of 24 global futures markets, with daily return data starting in March 1986 and extending through December 2020. From an experimental standpoint, our goal is to expand the understanding of IPO methods from both a computational and performance expectation perspective. Our experimental results on real asset price data provides proof of concept of the IPO approach, thus filling a notable gap in this fast moving field. In summary, our analytical and experimental contributions are as follows:

1. For the case where the nominal program is either unconstrained or contains only linear equality

constraints then the integrated program can be recast as an unconstrained quadratic program. We provide the necessary conditions for convexity and provide analytical solutions for the optimal IPO coefficients, $\boldsymbol{\theta}^*$. We provide conditions for which $\boldsymbol{\theta}^*$ is an unbiased estimator of $\boldsymbol{\theta}$ and derive the analytical expression for the variance. We demonstrate that the IPO coefficients explicitly minimize the tracking error to the unconstrained ex-post optimal MVO portfolio and provide the equivalent minimum-tracking error optimization program.

2. We conduct a simulation study based on synthetically generated data and compare the out-of-sample performance of the IPO and OLS models under varying degrees of estimation error in both the mean and covariance. We demonstrate that, in general, the IPO models produce consistently lower out-of-sample MVO costs, even when the underlying generating process for asset returns is linear in the auxiliary variables. Specifically, we show that, under such circumstances, and even when the asset return SNRs are relatively large, the estimation error in the covariance matrix can result in OLS models that generate sub-optimal out-of-sample MVO costs. The IPO model, on the other hand, corrects for estimation error in the sample covariance and provides consistently lower out-of-sample MVO costs in both low and moderate SNR regimes.

3. We conduct a simulation study based on synthetically generated data and demonstrate the computational advantage of the analytical IPO solution over that of the current state-of-the art method, based on implicit differentiation and gradient descent. We demonstrate that the compute time required for determining the IPO coefficients analytically is comparable to that of the equivalent least-squares problem, and is on average a full order of magnitude faster than the corresponding iterative method.

4. We discuss the computational complexity of the IPO framework as a function of the number training observations and number assets in the portfolio. We conduct a simulation study based on synthetically generated data and consider linear inequality constrained nominal MVO program under varying degrees of model misspecification. We approximate the non-convex problem with the analytical IPO solution whereby the inequality constraints are ignored. We demonstrate the computational and performance advantage of the analytical IPO solution, which is on average 100x - 1000x times faster than the current state-of-the-art method and produces solutions with lower out-of-sample variance and, in many instances, improved MVO costs.

5. We perform a simulation study using global futures data, considering both unconstrained and constrained nominal MVO programs and univariate and multivariate regression models. Out-of-sample numerical results demonstrate that the IPO model can provide lower realized cost and superior economic performance in comparison to the traditional OLS 'predict then optimize' approach.

2 Methodology

We begin by formally defining the traditional 'predict, then optimize' framework which will help draw the distinction to the IPO alternative. We assume that we have a nominal cost function $c\colon \mathbb{R}^{d_z} \times \mathbb{R}^{d_y} \to \mathbb{R}$, which takes as input a decision vector $\mathbf{z} \in \mathbb{R}^{d_z}$ constrained to a feasible region $\mathbb{Z} \subseteq \mathbb{R}^{d_z}$, and a quantity of interest $\mathbf{y}^{(i)} \in \mathbb{R}^{d_y}$. If the realization $\mathbf{y}^{(i)}$ is known then generating optimal decisions amounts to solving the following deterministic optimization problem:

$$\underset{\mathbf{z}\in\mathbb{Z}}{\text{minimize}}\, c(\mathbf{z}, \mathbf{y}^{(i)}), \tag{2}$$

with optimal solution:

$$\mathbf{z}^*(\mathbf{y}^{(i)}) = \underset{\mathbf{z}\in\mathbb{Z}}{\operatorname{argmin}}\, c(\mathbf{z}, \mathbf{y}^{(i)}). \tag{3}$$

In reality, we do not know the true value of $\mathbf{y}^{(i)}$ at decision time. We therefore estimate the values $\mathbf{y}^{(i)}$ through associated auxiliary variables $\mathbf{x}^{(i)} \in \mathbb{R}^{d_x}$. We define a prediction model by the function $f\colon \mathbb{R}^{d_x} \times \mathbb{R}^{d_\theta} \to \mathbb{R}^{d_y}$, where f belongs to the class of regression models $\mathbb{F} = \{f \mid f(\mathbf{x}, \boldsymbol{\theta}) = \boldsymbol{\theta}^T \mathbf{x}\}$ and the parameter $\boldsymbol{\theta}$ is constrained to the feasible region $\Theta \subseteq \mathbb{R}^{d_\theta}$. For a particular $\boldsymbol{\theta}$, we define the estimate $\hat{\mathbf{y}}^{(i)}$ as:

$$\hat{\mathbf{y}}^{(i)} = f(\mathbf{x}^{(i)}, \boldsymbol{\theta}) = \boldsymbol{\theta}^T \mathbf{x}^{(i)}\,. \tag{4}$$

In traditional parameter estimation, $\boldsymbol{\theta}$ would be chosen in order to minimize a prediction-based loss function $\ell\colon \mathbb{R}^{d_y} \times \mathbb{R}^{d_y} \to \mathbb{R}$, such as least-squares or negative log-likelihood. Specifically, given training data set $D = \{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})\}_{i=1}^m$ we choose $\hat{\boldsymbol{\theta}}$ such that:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}}\, \mathbb{E}_D[\ell(f(\mathbf{x}^{(i)}, \boldsymbol{\theta}), \mathbf{y}^{(i)})], \tag{5}$$

where $\mathbb{E}_D$ denotes the expectation operator with respect to the training distribution D. Once $\hat{\boldsymbol{\theta}}$ has been determined, a traditional 'predict, then optimize' framework, would then 'plug-in' the estimate, $\hat{\mathbf{y}}^{(i)}$, into program (2) in order to generate the optimal decisions $\mathbf{z}^*(\hat{\mathbf{y}}^{(i)})$.

In this paper, we are interested in optimizing $\boldsymbol{\theta}$ in the context of the nominal cost, c, and its constraints, $\mathbb{Z}$, in order to minimize the average realized nominal cost of the policy $\mathbf{z}^*(\mathbf{x}^{(i)}, \boldsymbol{\theta})$ induced by this parameterization. For a fixed instantiation $(\mathbf{x}^{(i)}, \boldsymbol{\theta})$, we solve the nominal cost program (6) in order to determine the optimal action $\mathbf{z}^*(\mathbf{x}^{(i)}, \boldsymbol{\theta})$ corresponding to our observed input $\mathbf{x}^{(i)}$. Specifically, we solve:

$$\underset{\mathbf{z}\in\mathbb{Z}}{\text{minimize}}\, c(\mathbf{z}, f(\mathbf{x}^{(i)}, \boldsymbol{\theta})), \tag{6}$$

with optimal solution given by:

$$\mathbf{z}^*(\mathbf{x}^{(i)}, \boldsymbol{\theta}) = \underset{\mathbf{z}\in\mathbb{Z}}{\operatorname{argmin}}\, c(\mathbf{z}, f(\mathbf{x}^{(i)}, \boldsymbol{\theta})). \tag{7}$$

Our objective is to therefore choose $\boldsymbol{\theta}$ in order to minimize the average realized nominal cost induced by the decision policy $\mathbf{z}^*(\mathbf{x}^{(i)}, \boldsymbol{\theta})$. The resulting integrated prediction-optimization problem is a stochastic optimization problem, presented in program (8).

$$\begin{aligned} \underset{\boldsymbol{\theta}\in\Theta}{\text{minimize}} \quad & \mathbb{E}_D[c(\mathbf{z}^*(\mathbf{x}, \boldsymbol{\theta}), \mathbf{y})] \\ \text{subject to} \quad & \mathbf{z}^*(\mathbf{x}, \boldsymbol{\theta}) = \underset{\mathbf{z}\in\mathbb{Z}}{\operatorname{argmin}}\, c(\mathbf{z}, f(\mathbf{x}, \boldsymbol{\theta})). \end{aligned} \tag{8}$$

In practice, we are typically presented with discrete observations $D = \{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})\}_{i=1}^{m}$ and therefore we can approximate the expectation by its sample average approximation [40]. The full IPO problem is presented in discrete form in program (9):

$$\begin{aligned} \underset{\boldsymbol{\theta} \in \Theta}{\text{minimize}} \quad & \frac{1}{m} \sum_{i=1}^{m} c(\mathbf{z}^*(\mathbf{x}^{(i)}, \boldsymbol{\theta}), \mathbf{y}^{(i)}) \\ \text{subject to} \quad & \mathbf{z}^*(\mathbf{x}^{(i)}, \boldsymbol{\theta}) = \underset{\mathbf{z} \in \mathbb{Z}}{\operatorname{argmin}}\, c(\mathbf{z}, f(\mathbf{x}^{(i)}, \boldsymbol{\theta})) \quad \forall i \in 1, ..., m. \end{aligned} \tag{9}$$

Note that program (9) resembles the well documented empirical risk minimization [43], where the objective is to choose $\boldsymbol{\theta}$ in order to minimize the average empirical nominal cost. The notable difference, of course, is that for a fixed instantiation $\boldsymbol{\theta}$, the decision policy, $\mathbf{z}^*(\mathbf{x}^{(i)}, \boldsymbol{\theta})$, is optimal in the context of its nominal cost program, which assumes predictions of the form $\hat{\mathbf{y}}^{(i)} = f(\mathbf{x}^{(i)}, \boldsymbol{\theta})$.

The IPO formulation results in a complicated dependency of the model parameters, $\boldsymbol{\theta}$, on the optimized values, $\mathbf{z}^*(\mathbf{x}^{(i)}, \boldsymbol{\theta})$, connected through the argmin function. This problem is frequently encountered in modern machine learning applications and there are several approaches for overcoming this challenge. For example, the problem can be structured as a bi-level optimization problem in which, under special constraint cases, an analytical solution for the gradient exists [26]. Alternatively, heuristic methods, such as the one provided by Sinha et al. [41], propose structuring the problem as a bi-level optimization and use kriging (Gaussian process models) to approximate the mapping to the lower-level optimization problem.

In Section 3, we discuss this relationship more fully in the context of mean-variance optimization. We first consider the general problem whereby the nominal program contains both linear equality and inequality constraints. We follow Amos and Kolter [2] and Donti et al. [17] and provide the current state-of-the-art formulation that re-structures the problem as an end-to-end trainable neural network with a differentiable quadratic programming layer and optimizes the prediction model parameters by gradient descent. We then consider several special cases, in which an analytical solution to the nominal program exists, and discuss in detail the theoretical properties and practical benefits of the IPO estimator.

3 IPO: Mean-Variance Optimization

We begin with a brief introduction to mean-variance portfolio optimization. The fundamental objective of portfolio optimization is the determination of next period's optimal asset allocation under conditions of uncertainty. Markowitz [35], a pioneer of Modern Portfolio Theory, proposed that investors' preferences for return and risk are characterized by quadratic utility: mean-variance optimization (MVO).

We consider a universe of d_z assets and denote the matrix of (excess) return observations as $\mathbf{Y} = [\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, ..., \mathbf{y}^{(m)}] \in \mathbb{R}^{m \times d_z}$ with $m > d_z$. Let $\mathbf{V}^{(i)} \in \mathbb{R}^{d_z \times d_z}$ denote the time-varying symmetric positive definite covariance matrix of asset returns. We define the portfolio $\mathbf{z} \in \mathbb{R}^{d_z}$, where the element, $\mathbf{z}_j$, denotes the proportion of total capital invested in the j^{th} asset. The mean variance nominal cost function at time i is presented in equation (10).

$$c(\mathbf{z}, \mathbf{y}^{(i)}) = -\, \mathbf{z}^T \, \mathbf{y}^{(i)} + \frac{\delta}{2} \, \mathbf{z}^T \, \mathbf{V}^{(i)} \, \mathbf{z} \tag{10}$$

Therefore, for a particular return observation $\mathbf{y}^{(i)}$, the optimal portfolio weights, $\mathbf{z}^*(\mathbf{y}^{(i)})$, is given by the solution to the following convex quadratic program:

$$\mathbf{z}^*(\mathbf{y}^{(i)}) = \underset{\mathbf{z}\in\mathbb{Z}}{\operatorname{argmin}} - \mathbf{z}^T\,\mathbf{y}^{(i)} + \frac{\delta}{2}\,\mathbf{z}^T\,\mathbf{V}^{(i)}\,\mathbf{z} \tag{11}$$

where $\delta \in \mathbb{R}_+$ is a risk-aversion parameter that controls the trade-off between minimizing variance and maximizing return.

In reality, we do not know the value $\mathbf{y}^{(i)}$ or $\mathbf{V}^{(i)}$ at decision time. In this paper we choose to estimate the time-varying covariance matrix using a traditional weighted moving average approach and denote the covariance estimate as $\hat{\mathbf{V}}^{(i)}$. Asset returns are modelled according to the following linear model:

$$\mathbf{y}^{(i)} = \mathbf{P}\,\mathrm{diag}(\mathbf{x}^{(i)})\,\boldsymbol{\theta} + \boldsymbol{\epsilon}^{(i)} \tag{12}$$

with residuals $\boldsymbol{\epsilon}^{(i)} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Here $\mathrm{diag}(\cdot)$ denotes the usual diagonal operator and $\mathbf{P} \in \mathbb{R}^{d_y \times d_x}$ controls the regression design with each element $\mathbf{P}_{jk} \in \{0, 1\}$. In particular, we assume that each asset has its own, perhaps unique, set of auxiliary variables. For example, if the auxiliary variables represent price-to-earnings (P/E) and debt-to-equity (D/E) ratios for each asset under consideration, then it would be unrealistic to model a particular asset's return as a function of all available P/E and D/E ratios. Indeed, doing so would almost certainly lead to model overfit. Instead, we choose to model asset j's return as a linear function of the P/E and D/E ratios relevant to asset j, specifically:

$$\hat{\mathbf{y}}_{\mathrm{j}}^{(i)} = \boldsymbol{\theta}^T_{\mathbf{a}(j)}\,\mathbf{x}_{\mathbf{a}(\mathrm{j})}{}^{(i)}, \tag{13}$$

where $\mathbf{a}(j)$ denotes the indices of the auxiliary variables relevant to asset j. Therefore,

$$\mathbf{P}_{jk} = \begin{cases} 1, & \text{if } k \in \mathbf{a}(j) \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

and for observation i, the regression estimate of asset expected returns is given by:

$$\hat{\mathbf{y}}^{(i)} = f(\mathbf{x}^{(i)}, \boldsymbol{\theta}) = \mathbf{P}\,\mathrm{diag}(\mathbf{x}^{(i)})\,\boldsymbol{\theta}\,. \tag{15}$$

Finally, under the estimation hypothesis, the mean-variance nominal cost function has the following form:

$$c(\mathbf{z}, \hat{\mathbf{y}}^{(i)}) = -\,\mathbf{z}^T\,\hat{\mathbf{y}}^{(i)} + \frac{\delta}{2}\,\mathbf{z}^T\,\hat{\mathbf{V}}^{(i)}\,\mathbf{z} = -\,\mathbf{z}^T\,\mathbf{P}\,\mathrm{diag}(\mathbf{x}^{(i)})\,\boldsymbol{\theta} + \frac{\delta}{2}\,\mathbf{z}^T\,\hat{\mathbf{V}}^{(i)}\,\mathbf{z} \tag{16}$$

In the IPO framework, the objective is to choose $\boldsymbol{\theta}$ in order to minimize the average realized nominal cost induced by the optimal decision policies $\{\mathbf{z}^*(\mathbf{x}^{(i)}, \boldsymbol{\theta})\}_{i=1}^m$. Substituting the estimated and realized costs into program (9) gives the full IPO program in discrete form, presented in program (17):

$$\begin{aligned} \underset{\boldsymbol{\theta}\in\Theta}{\text{minimize}} \quad & \frac{1}{m}\sum_{i=1}^{m}\Big(-\,\mathbf{z}^*(\mathbf{x}^{(i)}, \boldsymbol{\theta})^T\,\mathbf{y}^{(i)} + \frac{\delta}{2}\,\mathbf{z}^*(\mathbf{x}^{(i)}, \boldsymbol{\theta})^T\,\mathbf{V}^{(i)}\,\mathbf{z}^*(\mathbf{x}^{(i)}, \boldsymbol{\theta})\Big) \\ \text{subject to} \quad & \mathbf{z}^*(\mathbf{x}^{(i)}, \boldsymbol{\theta}) = \underset{\mathbf{z}\in\mathbb{Z}}{\operatorname{argmin}} - \mathbf{z}^T\,\mathbf{P}\,\mathrm{diag}(\mathbf{x}^{(i)})\,\boldsymbol{\theta} + \frac{\delta}{2}\,\mathbf{z}^T\,\hat{\mathbf{V}}^{(i)}\,\mathbf{z} \quad \forall i = 1, ..., m. \end{aligned} \tag{17}$$

and as before the objective function $L(\boldsymbol{\theta})$ is given by:

$$L(\boldsymbol{\theta}) = \frac{1}{m} \sum_{i=1}^{m} \Big(- \mathbf{z}^*(\mathbf{x}^{(i)}, \boldsymbol{\theta})^T \, \mathbf{y}^{(i)} + \frac{\delta}{2} \, \mathbf{z}^*(\mathbf{x}^{(i)}, \boldsymbol{\theta})^T \, \mathbf{V}^{(i)} \, \mathbf{z}^*(\mathbf{x}^{(i)}, \boldsymbol{\theta}) \Big). \tag{18}$$

Note at this point we have not described the feasible region, $\mathbb{Z}$, of the nominal program. In the following subsections we discuss the general case where $\mathbb{Z}$ describes a set of linear equality and inequality constraints and formalize the current state-of-the-art neural-network framework. We then discuss several special cases in which an analytical solution to the nominal MVO problem is possible and derive the relevant theory.

3.1 Current state-of-the-art methodology

We begin with the general case whereby the feasible region of the nominal program is defined by both linear equality and inequality constraints. Specifically we have the following nominal program:

$$\begin{aligned} \underset{\mathbf{z}}{\text{minimize}} \quad & - \mathbf{z}^T \, \hat{\mathbf{y}}^{(i)} + \frac{\delta}{2} \, \mathbf{z}^T \, \hat{\mathbf{V}}^{(i)} \, \mathbf{z} \\ \text{subject to} \quad & \mathbf{A} \, \mathbf{z} = \mathbf{b} \\ & \mathbf{G} \, \mathbf{z} \leq \mathbf{h} \end{aligned} \tag{19}$$

where $\mathbf{A} \in \mathbb{R}^{d_{\text{eq}} \times d_{\mathbf{z}}}$, $\mathbf{b} \in \mathbb{R}^{d_{\text{eq}}}$ and $\mathbf{G} \in \mathbb{R}^{d_{\text{iq}} \times d_{\mathbf{z}}}$, $\mathbf{h} \in \mathbb{R}^{d_{\text{iq}}}$ describe the linear equality and inequality constraints, respectively. In general, there is no known analytical solution to Program (19) and instead the solution, $\mathbf{z}^*$ is obtained through iterative interior point methods.

Furthermore, because of the inequality constraints, the IPO objective, $L(\boldsymbol{\theta})$, is generally not a convex function of $\boldsymbol{\theta}$. Nonetheless, we can solve for locally optimal solutions by applying (stochastic) gradient descent methods. From the multivariate chain-rule, the gradient, $\nabla_\theta L$ can be expressed as:

$$\nabla_\theta L = \frac{\partial L}{\partial \, \mathbf{z}^*} \frac{\partial \, \mathbf{z}^*}{\partial \, \boldsymbol{\theta}}. \tag{20}$$

In our particular case, the nominal cost function, c, is smooth and twice differentiable in the decision vector $\mathbf{z}$ and therefore it is relatively straightforward to compute the gradient $\partial L / \partial \, \mathbf{z}^*$. The Jacobian, $\partial \, \mathbf{z}^* / \partial \, \boldsymbol{\theta}$, requires differentiation through the argmin operator. We will see shortly, however, that rather than forming the Jacobian directly, we can instead compute $\partial L / \partial \, \boldsymbol{\theta}$ by implicit differentiation of the system of linear equations provided by the Karush–Kuhn–Tucker (KKT) conditions at the optimal solution $\mathbf{z}^*$ to program (19).

We follow the work of Amos and Kolter [2] and begin by first writing the Lagrangian of program (19):

$$\mathcal{L}(\mathbf{z}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = - \mathbf{z}^T \, \mathbf{y} + \frac{\delta}{2} \, \mathbf{z}^T \, \mathbf{V} \, \mathbf{z} + \boldsymbol{\lambda}^T (\mathbf{G} \, \mathbf{z} - \mathbf{h}) + \boldsymbol{\nu}^T (\mathbf{A} \, \mathbf{z} - \mathbf{b}), \tag{21}$$

where $\boldsymbol{\lambda} \in \mathbb{R}^{d_{iq}}$ and $\boldsymbol{\nu} \in \mathbb{R}^{d_{eq}}$ are the dual variables of the inequality and equality constraints, respectively. For compactness, we have temporarily dropped the index notation. The well-known KKT conditions for stationarity, primal feasibility, and complementary slackness are given by equations (22).

$$
\begin{aligned}
-\mathbf{y}+\frac{\delta}{2}\mathbf{V}\,\mathbf{z}^{*}+\mathbf{G}^{T}\boldsymbol{\lambda}^{*T}+\mathbf{A}^{T}\boldsymbol{\nu}^{*}&=0\\
(\mathbf{G}\,\mathbf{z}*-\mathbf{h})&\leq 0\\
\boldsymbol{\lambda}^{*}&\geq 0\\
\boldsymbol{\lambda}^{*}\cdot(\mathbf{G}\,\mathbf{z}^{*}-\mathbf{h})&=0\\
\mathbf{A}\,\mathbf{z}^{*}&=\mathbf{b}
\end{aligned}
\tag{22}
$$

Taking the differentials of these conditions gives the following system of equations:

$$
\begin{bmatrix}\delta\,\mathbf{V} & \mathbf{G}^{T} & \mathbf{A}^{T}\\ \mathrm{diag}(\boldsymbol{\lambda}^{*})\,\mathbf{G} & \mathrm{diag}(\mathbf{G}\,\mathbf{z}^{*}-\mathbf{h}) & 0\\ \mathbf{A} & 0 & 0\end{bmatrix}\begin{bmatrix}\mathbf{dz}\\ \mathbf{d}\,\boldsymbol{\lambda}\\ \mathbf{d}\,\boldsymbol{\nu}\end{bmatrix}=-\begin{bmatrix}\delta\mathbf{dV}\,\mathbf{z}^{*}-\mathbf{dy}+\mathbf{dG}^{T}\,\boldsymbol{\lambda}^{*}+\mathbf{dA}^{T}\,\boldsymbol{\nu}*\\ \mathrm{diag}(\boldsymbol{\lambda}^{*})\mathbf{dG}\,\mathbf{z}^{*}-\mathrm{diag}(\boldsymbol{\lambda}^{*})\mathbf{dh}\\ \mathbf{dA}\,\mathbf{z}^{*}-\mathbf{db}\end{bmatrix}.
\tag{23}
$$

We make two important observations about the system of equations (23). The first, is that the left side matrix gives the optimality conditions of the convex quadratic problem, which, when solving by interior-point methods, must be factorized in order to obtain the solution to the nominal program [12]. Secondly, the right side gives the differentials of the relevant functions at the achieved solution with respect to any of the input parameters. In particular, we seek to compute the Jacobian $\partial\,\mathbf{z}^{*}/\partial\,\mathbf{y}$. As explained by Amos and Kolter [2], the Jacobian $\partial\,\mathbf{z}^{*}/\partial\,\mathbf{y}$ is obtained by letting $\mathbf{dy}=\mathbf{I}$ (setting all other differential terms to zero) and solving the system of equations for $\mathbf{dz}$. From a computation standpoint, the required Jacobian is therefore effectively obtained 'for free' upon factorization of the left matrix when obtaining the solution, $\mathbf{z}^{*}$, to the nominal program.

In practice, however, we never explicitly form the right-side Jacobian matrix and compute $\partial\,\mathbf{z}^{*}/\partial\,\mathbf{y}$ directly. We instead follow the work of Amos and Kolter [2] and Agrawal et al. [1] and treat the nominal quadratic program as a differentiable layer in a neural-network. The IPO equivalent neural-network structure is depicted in Figure 1. In the forward pass, the input layer takes the auxiliary variables $\mathbf{x}^{(i)}$ and passes them to a simple linear layer to produce the estimates, $\hat{\mathbf{y}}^{(i)}$. The predictions are then passed to a differentiable quadratic programming layer which, for a given input $\hat{\mathbf{y}}^{(i)}$, solves the nominal program and returns the optimal portfolio weights $\mathbf{z}^{*}(\mathbf{x}^{(i)},\boldsymbol{\theta})$. At this point, we cache the factorized left-hand side matrix at the optimal solution, which will be invoked during backpropagation. Finally, the quality of the portfolio weights are evaluated by the cost function, $c(\mathbf{z}^{*}(\mathbf{x},\boldsymbol{\theta}),\mathbf{y}^{(i)})$ in the context of the true $\mathbf{y}^{(i)}$ values.

The backpropagation algorithm computes the partial derivatives in the regular fashion in order to generate the gradient of the realized cost with respect to the regression coefficients. In the case of the QP layer, we are given as input the derivative of the realized cost with respect to the optimal weights: $\partial c/\partial\,\mathbf{z}^{*}$. Following Amos and Kolter [2], we apply the implicit function theorem and compute $\partial c/\partial\,\mathbf{y}$ by multiplying the backward pass vector by the inverse of the transposed left-hand-side matrix, as shown in equation (24). Recall, we have cached the factorized left-side matrix and thus the gradient $\partial c/\partial\,\mathbf{y}$ is effectively provided at minimal additional computation cost.

$$
\begin{bmatrix}\bar{\mathbf{d}}_{\mathbf{z}}\\ \bar{\mathbf{d}}_{\boldsymbol{\lambda}}\\ \bar{\mathbf{d}}_{\boldsymbol{\nu}}\end{bmatrix}=\begin{bmatrix}\frac{\partial c}{\partial\,\mathbf{y}}\\ -\\ -\end{bmatrix}=-\begin{bmatrix}\delta\,\mathbf{V} & \mathbf{G}^{T}\,\mathrm{diag}(\boldsymbol{\lambda}^{*}) & \mathbf{A}^{T}\\ \mathbf{G} & \mathrm{diag}(\mathbf{G}\,\mathbf{z}^{*}-\mathbf{h}) & 0\\ \mathbf{A} & 0 & 0\end{bmatrix}^{-1}\begin{bmatrix}\left(\frac{\partial c}{\partial\,\mathbf{z}^{*}}\right)^{T}\\ 0\\ 0\end{bmatrix}.
\tag{24}
$$

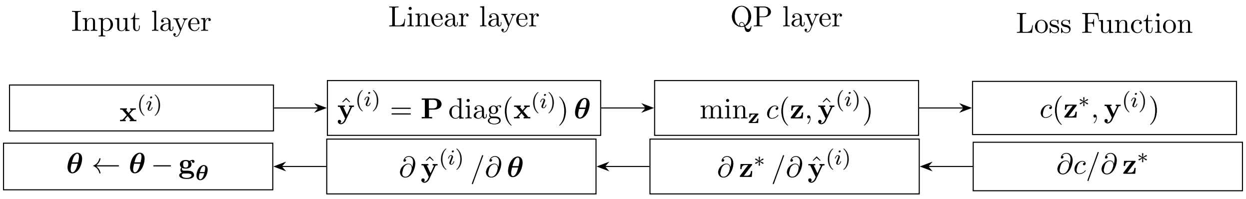


Figure 1: IPO program represented as an end-to-end neural-network with predictive linear layer, differentiable quadratic programming layer and realized nominal cost loss function.

When the nominal program contains linear inequality constraints, then we search for a locally optimal solution using stochastic gradient descent (SGD). In this case, the descent direction, $\mathbf{g}_{\boldsymbol{\theta}}$, at each iteration attempts to approximate the gradient, $\nabla_\theta L$, and is given by :

$$\mathbf{g}_\theta = \sum_{i \in B} \Big(\frac{\partial c}{\partial\,\boldsymbol{\theta}}\Big)_{|(\mathbf{z}^*(\mathbf{x}^{(i)}, \boldsymbol{\theta}), \mathbf{y}^{(i)})} \approx \nabla_\theta L$$

where in standard stochastic gradient descent, B represents a randomly drawn sample batch.

Lastly, Equation (24) allows for efficient computation of the gradients with respect to any of the relevant problem variables. While in this paper we are solely concerned with linear prediction models for the vectors of expected returns, $\mathbf{y}^{(i)}$, we note that the IPO framework can easily support integrated optimization of prediction model parameters (linear or otherwise) for the remaining problem variables. For the reader's interest, we state the gradients for all other problem variables and refer the reader to Amos and Kolter [2] for their derivation.

$$\frac{\partial c}{\partial\,\mathbf{V}} = \frac{1}{2}\Big(\bar{\mathbf{d}}_{\mathbf{z}}\,\mathbf{z}^{*T} + \mathbf{z}^*\,\bar{\mathbf{d}}_{\mathbf{z}}^T\Big) \qquad \frac{\partial c}{\partial\,\mathbf{y}} = \bar{\mathbf{d}}_{\mathbf{z}}$$

$$\frac{\partial c}{\partial\,\mathbf{A}} = \bar{\mathbf{d}}_{\boldsymbol{\nu}}\,\mathbf{z}^{*T} + \boldsymbol{\nu}^*\,\bar{\mathbf{d}}_{\mathbf{z}}^T \qquad \frac{\partial c}{\partial\,\mathbf{b}} = -\bar{\mathbf{d}}_{\boldsymbol{\nu}}$$

$$\frac{\partial c}{\partial\,\mathbf{G}} = \mathrm{diag}(\boldsymbol{\lambda}^*)\Big(\bar{\mathbf{d}}_{\boldsymbol{\lambda}}\,\mathbf{z}^{*T} + \boldsymbol{\lambda}^*\,\bar{\mathbf{d}}_{\mathbf{z}}^T\Big) \qquad \frac{\partial c}{\partial\,\mathbf{h}} = -\,\mathrm{diag}(\boldsymbol{\lambda}^*)\bar{\mathbf{d}}_{\boldsymbol{\lambda}}$$

3.2 Special case 1: $\mathbb{Z} = \mathbb{R}^{d_z}$

We consider the case where the nominal MVO program is unconstrained ($\mathbb{Z} = \mathbb{R}^{d_z}$) and therefore an analytical solution is given by Equation (25).

$$\begin{aligned} \mathbf{z}^*(\mathbf{x}^{(i)}, \boldsymbol{\theta}) &= \frac{1}{\delta}\,\hat{\mathbf{V}}^{(i)^{-1}}\,\hat{\mathbf{y}}^{(i)} \\ &= \frac{1}{\delta}\,\hat{\mathbf{V}}^{(i)^{-1}}\,\mathbf{P}\,\mathrm{diag}(\mathbf{x}^{(i)})\,\boldsymbol{\theta} \end{aligned} \tag{25}$$

Proposition 1. *Let $\mathbb{Z} = \mathbb{R}^{d_z}$ and $\Theta = \mathbb{R}^{d_\theta}$. We define*

$$\mathbf{d}(\mathbf{x}, \mathbf{y}) = \frac{1}{m\delta}\sum_{i=1}^{m}\Big(\,\mathrm{diag}(\mathbf{x}^{(i)})\,\mathbf{P}^T\,\hat{\mathbf{V}}^{(i)^{-1}}\,\mathbf{y}^{(i)}\Big) \tag{26}$$

and

$$\mathbf{H}(\mathbf{x}) = \frac{1}{m\delta} \sum_{i=1}^{m} \Big(\operatorname{diag}(\mathbf{x}^{(i)})\, \mathbf{P}^T\, \hat{\mathbf{V}}^{(i)^{-1}}\, \mathbf{V}^{(i)}\, \hat{\mathbf{V}}^{(i)^{-1}}\, \mathbf{P} \operatorname{diag}(\mathbf{x}^{(i)}) \Big). \tag{27}$$

Then the IPO program (17) *is an unconstrained quadratic program (QP) given by:*

$$\underset{\boldsymbol{\theta} \in \Theta}{\operatorname{minimize}} \frac{1}{2}\, \boldsymbol{\theta}^T\, \mathbf{H}(\mathbf{x})\, \boldsymbol{\theta} - \boldsymbol{\theta}^T\, \mathbf{d}(\mathbf{x}, \mathbf{y}). \tag{28}$$

Furthermore, if there exists an $\mathbf{x}^{(i)}$ *such that* ${\mathbf{x}_{\mathrm{j}}}^{(i)} \neq 0 \quad \forall j \in 1, ..., d_x$ *then* $\mathbf{H}(\mathbf{x}) \succ 0$ *and therefore program* (28) *is an unconstrained convex quadratic program with unique minimum:*

$$\boldsymbol{\theta}^* = \mathbf{H}(\mathbf{x})^{-1}\, \mathbf{d}(\mathbf{x}, \mathbf{y}). \tag{29}$$

All proofs are provided in Appendix A. We make a few important observations. The first, is that for the realistic case where there exists an $\mathbf{x}^{(i)}$ such that each ${\mathbf{x}_{\mathrm{j}}}^{(i)}$ are not exactly zero, then the optimal IPO regression coefficients, $\boldsymbol{\theta}^*$, are unique. Furthermore, we observe that the solution is independent of the risk-aversion parameter. This is intuitive, as when the nominal program is unconstrained, then the risk-aversion parameter simply controls the scale of the resulting solutions $\mathbf{z}^*(\mathbf{x}^{(i)}, \boldsymbol{\theta})$.

We note that the solution presented in Equation (29) requires the action of the inverse of the Hessian: $\mathbf{H}(\mathbf{x})$. In many applications of machine learning, such as computer vision or statistical meta-modelling, it is difficult, if not impossible, to solve the inverse problem without customized algorithms or prior knowledge of the data (see for example Jones and Taylor [32], Ranjan et al. [39], Ongie et al. [38]). In many cases, the dimension of the relevant Hessian is either too large for both forward-mapping and inversion in reasonable compute time or is computationally unstable due to near-singularity. In our IPO framework, we fortunately do not encounter these technical difficulties surrounding the action of the inverse. In most practical settings, the dimension of the Hessian matrix, is on the order of 10 or 100, whereas the number of observations, m, is on the order of 1000 or 10000. The Hessian is therefore likely to be computationally stable and the action of the inverse is computationally tractable. This is validated numerically in Section 4.1 and we demonstrate the computational advantage of the analytical solution over the iterative descent method.

Furthermore, while outside of the scope of the current paper, we note that under the QP formulation (28), it is trivial to incorporate both regularization and constraints on $\boldsymbol{\theta}$. This is demonstrated by Program (30):

$$\begin{aligned} \underset{\mathbf{z}}{\operatorname{minimize}} \quad & \frac{1}{2}\, \boldsymbol{\theta}^T\, \mathbf{H}(\mathbf{x})\, \boldsymbol{\theta} - \boldsymbol{\theta}^T\, \mathbf{d}(\mathbf{x}, \mathbf{y}) + \Omega(|\, \boldsymbol{\theta}\, |) \\ \text{subject to} \quad & \mathbf{A}_{\boldsymbol{\theta}}\, \boldsymbol{\theta} = \mathbf{b}_{\boldsymbol{\theta}} \\ & \mathbf{G}_{\boldsymbol{\theta}}\, \boldsymbol{\theta} \leq \mathbf{h}_{\boldsymbol{\theta}} \end{aligned} \tag{30}$$

where $\Omega \colon \mathbb{R}^{d_\theta} \to \mathbb{R}$ is a convex regularization function. In most cases, Program (30) can be solved to optimality by standard quadratic programming techniques, whereas the incorporation of regularization and constraints in the current state-of-the-art solution is structurally more challenging.

We now discuss the properties of sampling distribution of the IPO parameter estimate, $\boldsymbol{\theta}^*$, and derive an estimate of the variance, $\operatorname{Var}(\boldsymbol{\theta}^*)$. Recall, from Equation (12) we have: $\mathbf{y}^{(i)} \sim \mathcal{N}(\mathbf{P} \operatorname{diag}(\mathbf{x}^{(i)})\, \boldsymbol{\theta}, \boldsymbol{\Sigma})$.

Proposition 2. *Let*

$$\mathbf{d_u}(\mathbf{x}) = \frac{1}{m\delta} \sum_{i=1}^{m} \Big(\mathrm{diag}(\mathbf{x}^{(i)})\, \mathbf{P}^T \, \hat{\mathbf{V}}^{(i)^{-1}} \, \mathbf{P} \, \mathrm{diag}(\mathbf{x}^{(i)}) \Big), \tag{31}$$

then the optimal IPO estimate, $\boldsymbol{\theta}^*$, is a biased estimate of $\boldsymbol{\theta}$ with bias $\mathbf{H}(\mathbf{x})^{-1}\, \mathbf{d_u}(\mathbf{x})$.

Corollary 1. *Let $\boldsymbol{\theta}^*_{\mathbf{u}} = \mathbf{d_u}(\mathbf{x})^{-1}\, \mathbf{H}(\mathbf{x})\, \boldsymbol{\theta}^*$. Then $\boldsymbol{\theta}^*_{\mathbf{u}}$ is an unbiased estimator of $\boldsymbol{\theta}$.*

Corollary 2. *Let $\hat{\mathbf{V}}^{(i)} = \mathbf{V}^{(i)}\, \forall i \in \{1, ..., m\}$. Then $\boldsymbol{\theta}^*$ is an unbiased estimator of $\boldsymbol{\theta}$.*

We observe from Proposition 2 that differences, or estimation errors, between $\hat{\mathbf{V}}^{(i)}$ and $\mathbf{V}^{(i)}$, make $\boldsymbol{\theta}^*$ a biased estimator in $\boldsymbol{\theta}$. In particular, the bias can be corrected by left multiplication of $\boldsymbol{\theta}^*$ by $\mathbf{d_u}(\mathbf{x})^{-1}\, \mathbf{H}(\mathbf{x})$. This observation leads to Corollary 2, which shows that when the estimation error in the covariance is zero then $\boldsymbol{\theta}^*$ is an unbiased estimator of $\boldsymbol{\theta}$. Moreover, unlike the OLS estimate, $\hat{\boldsymbol{\theta}}$, the IPO estimate, $\boldsymbol{\theta}^*$, will correct for estimation error in the sample covariance in the (likely) event that the estimation error is nonzero. This is validated numerically and discussed in more detail in in Section 4.1.

Proposition 3. *Let $\{\mathbf{y}^{(i)}\}_{i=1}^{m}$ be independent random variables with $\mathbf{y}^{(i)} \sim \mathcal{N}(\mathbf{P}\, \mathrm{diag}(\mathbf{x}^{(i)})\, \boldsymbol{\theta}, \boldsymbol{\Sigma})$. Let $\hat{\boldsymbol{\Sigma}}$ be an unbiased estimate of the sample covariance of residuals, given by:*

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{m-1} \sum_{i=1}^{m} \Big(\mathbf{y}^{(i)} - \mathbf{P}\, \mathrm{diag}(\mathbf{x}^{(i)})\, \boldsymbol{\theta} \Big)^2. \tag{32}$$

Let

$$\mathbf{M} = \frac{1}{\delta^2 m^2} \Big(\mathrm{diag}(\mathbf{x}^{(i)})\, \mathbf{P}^T \, \hat{\mathbf{V}}^{(i)^{-1}} \, \hat{\boldsymbol{\Sigma}}\, \hat{\mathbf{V}}^{(i)^{-1}} \, \mathbf{P} \, \mathrm{diag}(\mathbf{x}^{(i)}) \Big), \tag{33}$$

then the variance, $\mathrm{Var}(\boldsymbol{\theta}^*)$, is given by:

$$\mathrm{Var}(\boldsymbol{\theta}^*) = \mathbf{H}(\mathbf{x})^{-1}\, \mathbf{M}\, \mathbf{H}(\mathbf{x})^{-1} \tag{34}$$

We conclude this section by providing an alternative, and perhaps more intuitive expression of the optimal IPO coefficients derived from portfolio tracking-error optimization. Let $\|\cdot\|_{\mathbf{V}}$ denote the elliptic norm with respect to the symmetric positive definite matrix $\mathbf{V}$, defined as:

$$\|\mathbf{w}\|_{\mathbf{V}} = \sqrt{\mathbf{w}^T\, \mathbf{V}\, \mathbf{w}}. \tag{35}$$

More specifically, $\|\mathbf{z}^{(1)} - \mathbf{z}^{(2)}\|^2_{\mathbf{V}}$ measures the tracking-error between two portfolio weights with respect to the covariance $\mathbf{V}$.

Proposition 4. *Let $\mathbf{z}^*(\mathbf{y}^{(i)})$ and $\mathbf{z}^*(\mathbf{x}^{(i)}, \boldsymbol{\theta})$ be as defined in Equation (11) and Equation (25), respectively. Then the optimal IPO coefficients, $\boldsymbol{\theta}^*$, minimizes the average tracking error between $\mathbf{z}^*(\mathbf{y}^{(i)})$ and $\mathbf{z}^*(\mathbf{x}^{(i)}, \boldsymbol{\theta})$ with respect to the realized covariance $\mathbf{V}^{(i)}$:*

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta} \in \Theta}{\mathrm{argmin}}\, \frac{1}{2m} \sum_{i=1}^{m} \|\mathbf{z}^*(\mathbf{x}^{(i)}, \boldsymbol{\theta}) - \mathbf{z}^*(\mathbf{y}^{(i)})\|^2_{\mathbf{V}^{(i)}} \tag{36}$$

Indeed, Proposition 4 states that the IPO coefficients $\boldsymbol{\theta}^*$ minimizes the average tracking error between the estimated optimal weights, $\mathbf{z}^*(\mathbf{x}^{(i)}, \boldsymbol{\theta})$ and the ex-post optimal weight $\mathbf{z}^*(\mathbf{y}^{(i)})$.

3.3 Special Case 2: $\mathbb{Z} = \{\mathbf{A}\,\mathbf{z} = \mathbf{b}\}$

We now consider the case where the nominal program is constrained by a set of linear equality constraints:

$$\mathbb{Z} = \{\mathbf{A}\,\mathbf{z} = \mathbf{b}\},$$

where $\mathbf{A} \in \mathbb{R}^{d_{\mathrm{eq}} \times d_z}$ and $\mathbf{b} \in \mathbb{R}^{d_{\mathrm{eq}}}$. We assume the non-trivial case where $\mathbf{A}$ is not full rank. Let the columns of $\mathbf{F}$ form a basis for the nullspace of $\mathbf{A}$ defined as:

$$\mathrm{Null}(\mathbf{A}) = \{\mathbf{z} \in \mathbb{R}^{d_z} \,|\, \mathbf{A}\,\mathbf{z} = 0\}.$$

Let $\mathbf{z}_0$ be a particular element of $\mathbb{Z}$. It follows that $\forall\ \mathbf{w} \in \mathbb{R}^{d_{\mathbf{z}} - d_n}$ then $\mathbf{z} = \mathbf{F}\,\mathbf{w} + \mathbf{z}_0$ is also an element of $\mathbb{Z}$, where $d_n = \mathrm{nullity}(\mathbf{A})$. We follow Boyd and Vandenberghe [12] and recast the nominal program as an unconstrained convex quadratic program:

$$\min_{\mathbf{w}} c(\mathbf{F}\,\mathbf{w} + \mathbf{z}_0, \hat{\mathbf{y}}^{(i)}), \tag{37}$$

with unique global minimum:

$$\mathbf{w}^*(\hat{\mathbf{y}}^{(i)}) = \frac{1}{\delta}(\mathbf{F}^T\,\hat{\mathbf{V}}^{(i)}\,\mathbf{F})^{-1}\,\mathbf{F}^T\left(\hat{\mathbf{y}}^{(i)} - \delta\,\hat{\mathbf{V}}^{(i)}\,\mathbf{z}_0\right) \tag{38}$$

The solution to the nominal MVO Program (11) is then given by:

$$\begin{aligned}\mathbf{z}^*(\mathbf{x}^{(i)}, \boldsymbol{\theta}) &= \frac{1}{\delta}\,\mathbf{F}(\mathbf{F}^T\,\hat{\mathbf{V}}^{(i)}\,\mathbf{F})^{-1}\,\mathbf{F}^T(\mathbf{P}\,\mathrm{diag}(\mathbf{x}^{(i)})\,\boldsymbol{\theta} - \delta\,\hat{\mathbf{V}}^{(i)}\,\mathbf{z}_0) + \mathbf{z}_0 \\ &= \frac{1}{\delta}\,\mathbf{F}(\mathbf{F}^T\,\hat{\mathbf{V}}^{(i)}\,\mathbf{F})^{-1}\,\mathbf{F}^T\,\mathbf{P}\,\mathrm{diag}(\mathbf{x}^{(i)})\,\boldsymbol{\theta} + (\mathbf{I} - \mathbf{F}(\mathbf{F}^T\,\hat{\mathbf{V}}^{(i)}\,\mathbf{F})^{-1}\,\mathbf{F}^T\,\hat{\mathbf{V}}^{(i)})\,\mathbf{z}_0\end{aligned} \tag{39}$$

Proposition 5. *Let $\mathbb{Z} = \{\mathbf{A}\,\mathbf{z} = \mathbf{b}\}$ and $\Theta = \mathbb{R}^{d_\theta}$. Define*

$$\mathbf{d}_{eq}(\mathbf{x}, \mathbf{y}) = \frac{1}{m\delta}\sum_{i=1}^{m}\Big(\mathrm{diag}(\mathbf{x}^{(i)})\,\mathbf{P}^T\,\mathbf{F}(\mathbf{F}^T\,\hat{\mathbf{V}}^{(i)}\,\mathbf{F})^{-1}\,\mathbf{F}^T(\mathbf{y}^{(i)} - \mathbf{V}^{(i)}(\mathbf{I} - \mathbf{F}(\mathbf{F}^T\,\hat{\mathbf{V}}^{(i)}\,\mathbf{F})^{-1}\,\mathbf{F}^T\,\hat{\mathbf{V}}^{(i)})\,\mathbf{z}_0)\Big) \tag{40}$$

and

$$\mathbf{H}_{eq}(\mathbf{x}) = \frac{1}{m\delta}\sum_{i=1}^{m}\Big(\mathrm{diag}(\mathbf{x}^{(i)})\,\mathbf{P}^T\,\mathbf{F}(\mathbf{F}^T\,\hat{\mathbf{V}}^{(i)}\,\mathbf{F})^{-1}\,\mathbf{F}^T\,\mathbf{V}^{(i)}\,\mathbf{F}(\mathbf{F}^T\,\hat{\mathbf{V}}^{(i)}\,\mathbf{F})^{-1}\,\mathbf{F}^T\,\mathbf{P}\,\mathrm{diag}(\mathbf{x}^{(i)})\Big). \tag{41}$$

Then the IPO program (17) is an unconstrained quadratic program given by:

$$\underset{\boldsymbol{\theta} \in \Theta}{\mathrm{minimize}}\ \frac{1}{2}\,\boldsymbol{\theta}^T\,\mathbf{H}_{eq}(\mathbf{x})\,\boldsymbol{\theta} - \boldsymbol{\theta}^T\,\mathbf{d}_{eq}(\mathbf{x}, \mathbf{y}). \tag{42}$$

Furthermore, if there exists an $\mathbf{x}^{(i)}$ such that ${\mathbf{x}_{\mathrm{j}}}^{(i)} \neq 0 \quad \forall j \in 1, ..., d_x$ then $\mathbf{H}_{eq}(\mathbf{x}) \succ 0$ and therefore program (42) is an unconstrained convex quadratic program with unique minimum:

$$\boldsymbol{\theta}^*_{eq} = \mathbf{H}_{eq}(\mathbf{x})^{-1}\,\mathbf{d}_{eq}(\mathbf{x}, \mathbf{y}). \tag{43}$$

As before we briefly discuss the properties of the sampling distribution of the equality constrained IPO parameter estimate, $\boldsymbol{\theta}^*_{\mathrm{eq}}$, and derive an estimate of the variance, $\mathrm{Var}\,(\boldsymbol{\theta}^*_{\mathrm{eq}})$.

Proposition 6. *Let*

$$\mathbf{d_e}(\mathbf{x}) = \frac{1}{m\delta} \sum_{i=1}^{m} \Big(\mathrm{diag}(\mathbf{x}^{(i)})\,\mathbf{P}^T\,\mathbf{F}(\mathbf{F}^T\,\hat{\mathbf{V}}^{(i)}\,\mathbf{F})^{-1}\,\mathbf{F}^T\,\mathbf{P}\,\mathrm{diag}(\mathbf{x}^{(i)}) \Big), \tag{44}$$

then the optimal IPO estimate, $\boldsymbol{\theta}^*_{eq}$, is a biased estimate of $\boldsymbol{\theta}$ with bias $\mathbf{H}_{eq}(\mathbf{x})^{-1}\,\mathbf{d_e}(\mathbf{x})$.

Corollary 3. *Let $\hat{\mathbf{V}}^{(i)} = \mathbf{V}^{(i)}\,\forall i \in \{1, ..., m\}$. Then $\boldsymbol{\theta}^*_{eq}$ is an unbiased estimator of $\boldsymbol{\theta}$.*

Again we observe from Proposition 6 that in general $\boldsymbol{\theta}^*_{\mathrm{eq}}$ a biased estimator of $\boldsymbol{\theta}$. In particular, the bias can be corrected by left multiplication of $\boldsymbol{\theta}^*_{\mathrm{eq}}$ by $\mathbf{d_e}(\mathbf{x})^{-1}\,\mathbf{H}_{\mathrm{eq}}(\mathbf{x})$. Furthermore when the estimation error in the covariance is zero then $\boldsymbol{\theta}^*_{\mathrm{eq}}$ is an unbiased estimator of $\boldsymbol{\theta}$.

Proposition 7. *Let $\{\mathbf{y}^{(i)}\}_{i=1}^m$ be independent random variables with $\mathbf{y}^{(i)} \sim \mathcal{N}(\mathbf{P}\,\mathrm{diag}(\mathbf{x}^{(i)})\,\boldsymbol{\theta}, \boldsymbol{\Sigma})$. Let*

$$\mathbf{M}_{eq} = \frac{1}{\delta^2 m^2} \sum_{i=1}^{m} \Big(\mathrm{diag}(\mathbf{x}^{(i)})\,\mathbf{P}^T\,\mathbf{F}(\mathbf{F}^T\,\hat{\mathbf{V}}^{(i)}\,\mathbf{F})^{-1}\,\mathbf{F}^T\,\hat{\boldsymbol{\Sigma}}\,\mathbf{F}(\mathbf{F}^T\,\hat{\mathbf{V}}^{(i)}\,\mathbf{F})^{-1}\,\mathbf{F}^T\,\mathbf{P}\,\mathrm{diag}(\mathbf{x}^{(i)}) \Big). \tag{45}$$

then the variance, $\mathrm{Var}(\boldsymbol{\theta}^*_{eq})$, is given by:

$$\mathrm{Var}(\boldsymbol{\theta}^*_{eq}) = \mathbf{H}_{eq}(\mathbf{x})^{-1}\,\mathbf{M}_{eq}\,\mathbf{H}_{eq}(\mathbf{x})^{-1} \tag{46}$$

As before, we conclude this subsection with the following proposition that states that the IPO coefficients $\boldsymbol{\theta}^*$ minimizes the average tracking error between the estimated optimal weights, $\mathbf{z}^*(\mathbf{x}^{(i)}, \boldsymbol{\theta})$ and the ex-post optimal weight $\mathbf{z}^*(\mathbf{y}^{(i)})$.

Proposition 8. *Let $\mathbf{z}^*(\mathbf{y}^{(i)})$ and $\mathbf{z}^*(\mathbf{x}^{(i)}, \boldsymbol{\theta})$ be as defined in Equation (11) and Equation (25), respectively. Then the optimal IPO coefficients, $\boldsymbol{\theta}^*_{eq}$, minimizes the average tracking error between $\mathbf{z}^*(\mathbf{y}^{(i)})$ and $\mathbf{z}^*(\mathbf{x}^{(i)}, \boldsymbol{\theta})$ with respect to the realized covariance $\mathbf{V}^{(i)}$:*

$$\begin{aligned} \boldsymbol{\theta}^*_{\mathrm{eq}} = \underset{\boldsymbol{\theta}}{\mathrm{argmin}} \quad & \frac{1}{2m} \sum_{i=1}^{m} \| \mathbf{z}^*(\mathbf{x}^{(i)}, \boldsymbol{\theta}) - \mathbf{z}^*(\mathbf{y}^{(i)}) \|^2_{\mathbf{V}^{(i)}} \\ \text{subject to} \quad & \mathbf{A}\,\mathbf{z}^*(\mathbf{x}^{(i)}, \boldsymbol{\theta}) = \mathbf{b} \\ \text{subject to} \quad & \mathbf{A}\,\mathbf{z}^*(\mathbf{y}^{(i)}) = \mathbf{b} \end{aligned} \tag{47}$$

4 Simulated experiments

4.1 Simulation 1: estimation error in $\hat{\mathbf{V}}$

Elmachtoub and Grigas [19] consider the integration of predictive forecasting with downstream optimization problems that have linear cost functions. Their simulated experiments demonstrate that the benefit of the 'smart predict, then optimize' (SPO) framework increases as the amount of model misspecification increases. Specifically, model misspecification is introduced by synthetically generating cost vectors that are polynomial functions of the simulated auxiliary data and modelling the relationship as though it is linear. In particular, they demonstrate that a linear forecasting model trained with SPO can outperform traditional prediction models, such as OLS and random forest, and the amount of outperformance increases as the degree of nonlinearity in the ground truth increases.

Here, we demonstrate that, for a mean-variance nominal decision program, the IPO model can provide lower out-of-sample MVO costs in comparison to a traditional OLS-based 'predict, then optimize' model, even when the underlying ground truth is *linear* in the auxiliary variables. In particular, we demonstrate that the OLS model is vulnerable to estimation error in $\hat{\mathbf{V}}^{(i)}$, resulting in sub-optimal decision making and increasing out-of-sample MVO costs as estimation error in $\hat{\mathbf{V}}^{(i)}$ increases. The IPO model, on the other hand, explicitly corrects for estimation error in the covariance matrix. The simulated experiment below demonstrates that the IPO model consistently outperforms the OLS model in terms of minimizing the out-of-sample MVO cost. Moreover, the outperformance is shown to be consistent over a wide range of signal-to-noise ratios (SNRs) and asset correlation assumptions commonly observed in financial forecasting. In general we observe that the benefit of the IPO model increases as the estimation error in $\hat{\mathbf{V}}^{(i)}$ increases, even when the underlying ground truth is linear in the auxiliary variables.

We follow an experimental design similar to Hastie et al. [27]. Asset returns are assumed to be normally distributed, $\mathbf{y}^{(i)} \sim \mathcal{N}(\operatorname{diag}(\mathbf{x}^{(i)})\,\boldsymbol{\theta}_0, \mathbf{V})$ where $\mathbf{V} \in \mathbb{R}^{d_z \times d_z}$ has entry (j,k) equal to $\sigma^2 \rho^{|j-k|}$, and $\sigma = 0.0125$ (20% annualized). Asset mean returns are modelled according to univariate model of the form:

$$\mathbf{y}^{(i)} = \operatorname{diag}(\mathbf{x}^{(i)})\,\boldsymbol{\theta}_0 + \tau\,\boldsymbol{\epsilon}^{(i)},$$

where auxiliary data, $\mathbf{x}^{(i)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I_{d_x}})$, and residuals, $\boldsymbol{\epsilon}^{(i)} \sim \mathcal{N}(\mathbf{0}, \mathbf{V})$. The scalar value τ controls the SNR level, where $\mathrm{SNR} = \operatorname{Var}(f(\mathbf{x}, \boldsymbol{\theta}_0)) / \operatorname{Var}(\boldsymbol{\epsilon})$. We consider asset correlation values in the range of: $\rho \in \{0, 0.25, 0.5, 0.75\}$, and SNR values: $\mathrm{SNR} \in \{0.001, 0.002, 0.003, 0.004, 0.005, 0.01, 0.05, 0.10\}$. Note that it may appear that these SNR values are extremely low. However, in most applications of asset return forecasting, the SNRs are typically found to be much less than 1%. Indeed, a moderate sized universe (25 assets) with each asset having SNRs of 1% can generate annualized Sharpe ratios in the low double digits - which is extremely rare - and SNRs of 10% are extremely unlikely at a daily trading frequency.

We introduce estimation error in $\hat{\mathbf{V}}^{(i)}$ by varying the sample size, $s = \mathrm{res} * d_z$, used for estimation. We set the number of assets, $d_z = 10$, and consider resolutions, $\mathrm{res} \in \{5, 10, 20\}$, thus giving covariance sample sizes of $s \in \{50, 100, 200\}$. In all experiments we set the risk aversion parameter $\delta = 1$.

The simulation process can be described as follows:

1. Generate the ground truth coefficients: $\boldsymbol{\theta}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I_{d_\theta}})$.

2. Generate auxiliary variables: $\mathbf{x}^{(i)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I_{d_x}})$.

3. Generate 2000 return observations: $\mathbf{y}^{(i)} \sim \mathcal{N}(\operatorname{diag}(\mathbf{x}^{(i)})\,\boldsymbol{\theta}, \tau\,\boldsymbol{\epsilon}^{(i)})$, where τ is chosen to meet the desired SNR.

4. Divide the sample data into two equally sized disjoint data sets: in-sample and out-of-sample.

5. Generate estimates $\hat{\mathbf{V}}^{(i)}$ using the chosen sample size, s.

6. Estimate the optimal OLS and IPO coefficients on the in-sample data.

7. Generate the optimal out-of-sample MVO decisions, $\mathbf{z}^*(\boldsymbol{\theta}, \mathbf{x}^{(i)})$, using the covariance estimates $\hat{\mathbf{V}}^{(i)}$ and corresponding optimal regression coefficients for predicting $\hat{\mathbf{y}}^{(i)}$.

8. Evaluate several performance metrics (described below) on the out-of-sample data.

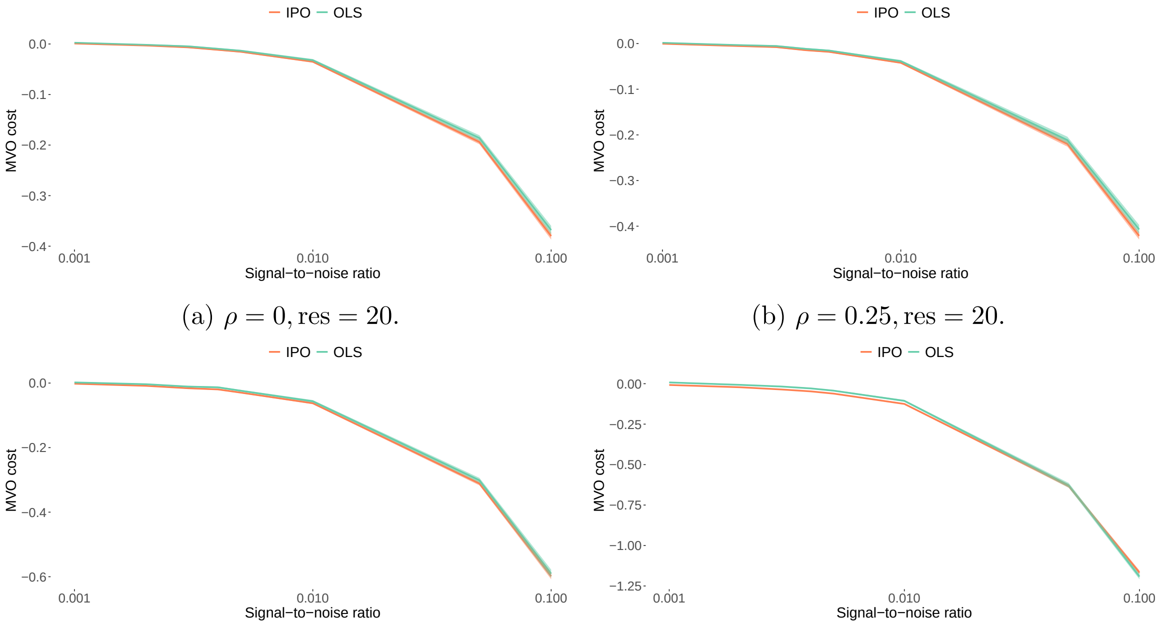
9. Repeat steps 1-8 a total of 100 times and average the results.

Performance metrics: Let $\boldsymbol{\theta}_0$ be the ground truth and $\boldsymbol{\theta}$ denote an estimated (OLS or IPO) regression coefficient. Let $\mathbf{V}$ denote the true asset covariance and let $\{\mathbf{y}^{(i)}\}_{i=1}^m$ denote the realized return observations.

- **MVO Cost**: Let $\mathbf{z}^*(\boldsymbol{\theta}, \mathbf{x}^{(i)})$ be as defined in Equation (7). The out-of-sample MVO cost is then given by:
$$c(\mathbf{z}^*(\boldsymbol{\theta}, \mathbf{x}^{(i)}), \mathbf{y}^{(i)}) = -\,\mathbf{z}^*(\boldsymbol{\theta}, \mathbf{x}^{(i)})^T\, \mathbf{y}^{(i)} + \frac{\delta}{2}\, \mathbf{z}^*(\boldsymbol{\theta}, \mathbf{x}^{(i)})^T\, \mathbf{V}\, \mathbf{z}^*(\boldsymbol{\theta}, \mathbf{x}^{(i)}) \tag{48}$$

- **Proportion of variance explained:** a measure of return prediction accuracy on defined as:
$$\mathrm{PVE}(\boldsymbol{\theta}) = 1 - \mathbb{E}[(\mathbf{y}^{(i)} - \mathrm{diag}(\mathbf{x}^{(i)})\, \boldsymbol{\theta})^2] / \,\mathrm{Var}\,(\mathbf{y}^{(i)}).$$

We consider the case where the nominal MVO program contains equality constraints. In particular, we enforce that the sum of the weights must be equal to one: $\mathbb{Z} = \{\mathbf{z}^T\, \mathbf{1} = 1\}$. Figures 2 and 3 report the average and 95%-ile range of the out-of-sample MVO costs and PVE values, respectively, as a function of the SNR. Here, the covariance resolution is set to 20 and therefore the expected estimation error in $\hat{\mathbf{V}}^{(i)}$ is relatively low. As a result, we observe that the difference in both out-of-sample MVO cost and PVE is negligible, with the IPO model producing marginally lower MVO costs and the OLS model producing marginally higher PVE, as expected. Observe that even in the most optimistic case where estimation error in $\hat{\mathbf{V}}^{(i)}$ is low and the ground truth relationship is linear, there is no adverse repercussions in using the IPO model. Furthermore, in order to effectively eliminate estimation error we require a covariance resolution on the order of 20; which in practical terms implies that for a 100 asset portfolio we require a sample size of 2000 return observations. In many forecasting applications a sample size of this magnitude would be impractical and would potentially interfere with the observed time-varying dependency of asset volatilities and correlations [20, 8, 10, 9, 11]. Furthermore, as estimation error increases, we observe that for the majority of relevant SNRs, the IPO model produces a lower realized out-of-sample MVO cost. In particular Figures 4 and 6 demonstrate a statistically significant reduction in out-of-sample MVO costs as the covariance resolution decreases to 10 and 5, respectively. Interestingly, Figures 5 and 7 demonstrate that the IPO model produces lower realized MVO costs, despite providing lower average prediction accuracy, as measured by PVE. Note that this finding is consistent with the results presented in [19]. Finally, we observe in Figures 4 and 6 that the benefit of the IPO model is greatest when the ground truth correlation values, ρ, are closest to zero. Indeed this is intuitive as the extent of covariance estimation error in both magnitude and sign is largest when correlation values approach zero.



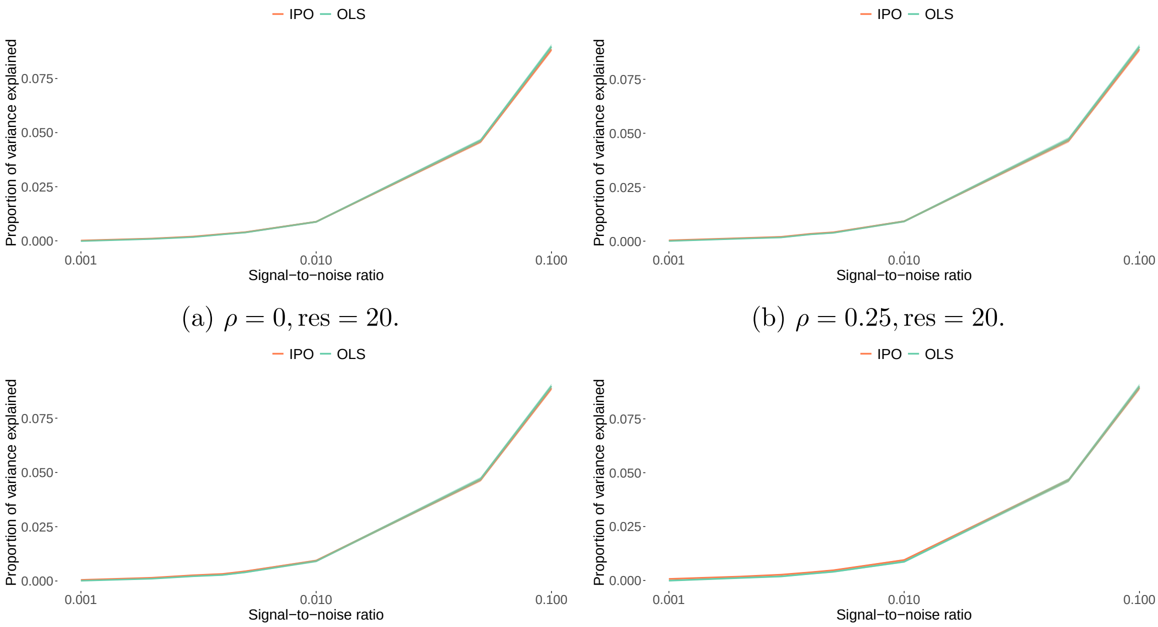
(a) $\rho = 0, \mathrm{res} = 20$.

(b) $\rho = 0.25, \mathrm{res} = 20$.

(c) $\rho = 0.50, \mathrm{res} = 20$.

(d) $\rho = 0.75, \mathrm{res} = 20$.

Figure 2: Out-of-sample MVO cost for IPO and OLS as of function of return signal-to-noise ratios.



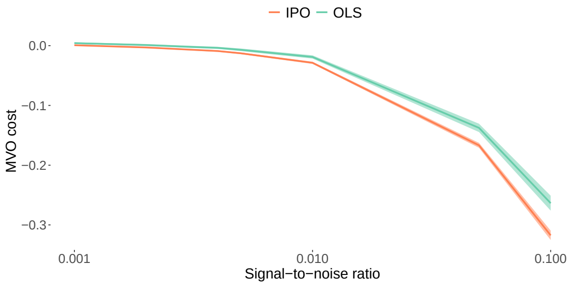
(a) $\rho = 0, \mathrm{res} = 20$.

(b) $\rho = 0.25, \mathrm{res} = 20$.

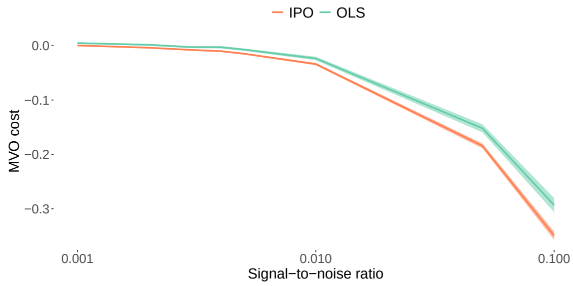
(c) $\rho = 0.50, \mathrm{res} = 20$.

(d) $\rho = 0.75, \mathrm{res} = 20$.

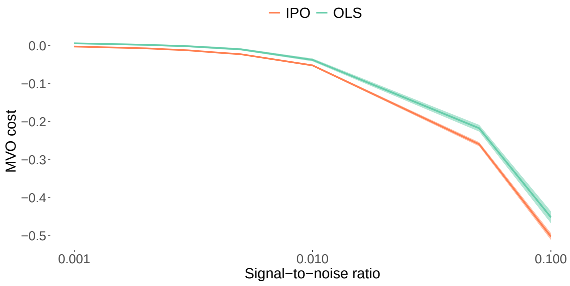
Figure 3: Out-of-sample PVE for IPO and OLS as of function of return signal-to-noise ratios.



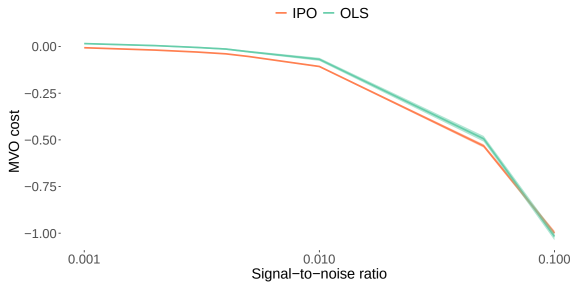
(a) $\rho = 0, \mathrm{res} = 10.$



(b) $\rho = 0.25, \mathrm{res} = 10.$

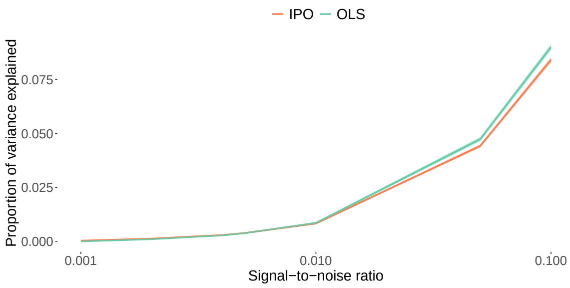


(c) $\rho = 0.50, \mathrm{res} = 10.$

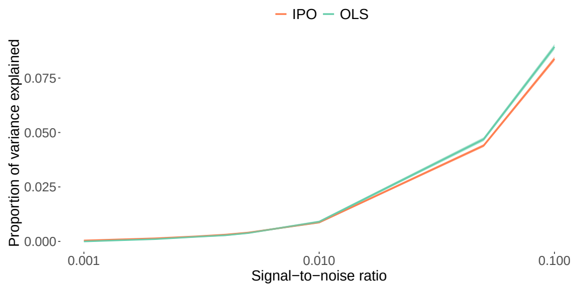


(d) $\rho = 0.75, \mathrm{res} = 10.$

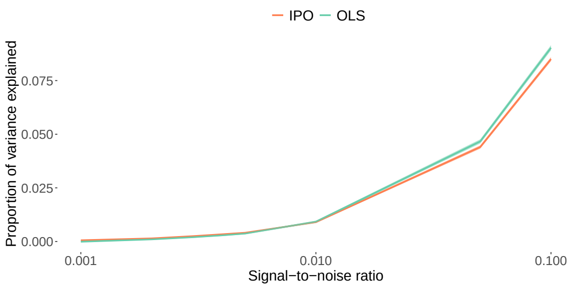
Figure 4: Out-of-sample MVO cost for IPO and OLS as of function of return signal-to-noise ratios.



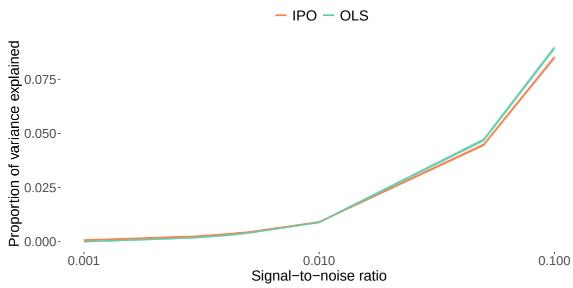
(a) $\rho = 0, \mathrm{res} = 10.$



(b) $\rho = 0.25, \mathrm{res} = 10.$

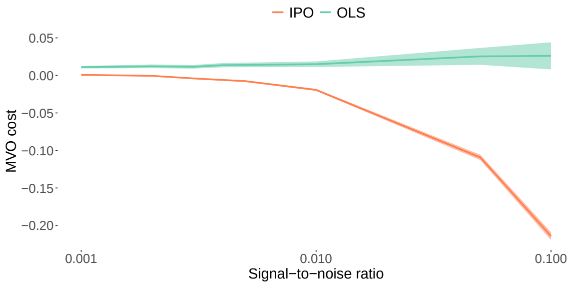


(c) $\rho = 0.50, \mathrm{res} = 10.$

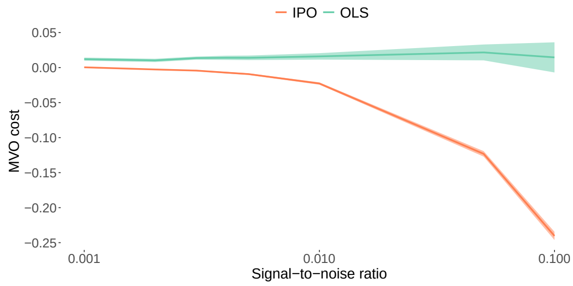


(d) $\rho = 0.75, \mathrm{res} = 10.$

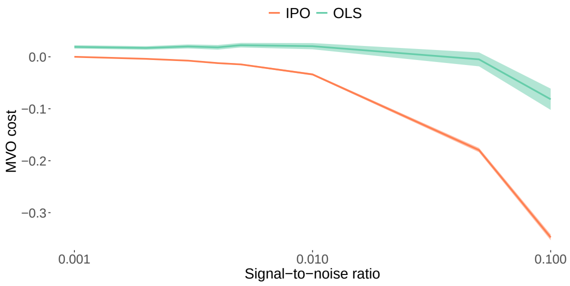
Figure 5: Out-of-sample PVE for IPO and OLS as of function of return signal-to-noise ratios.



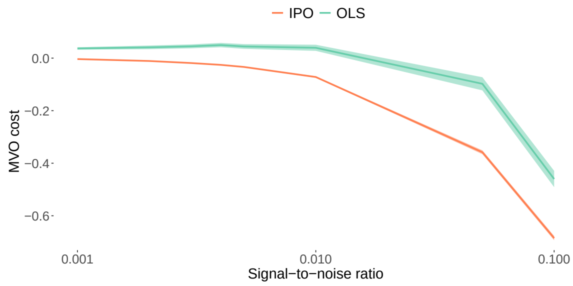
(a) $\rho = 0, \mathrm{res} = 5$.



(b) $\rho = 0.25, \mathrm{res} = 5$.

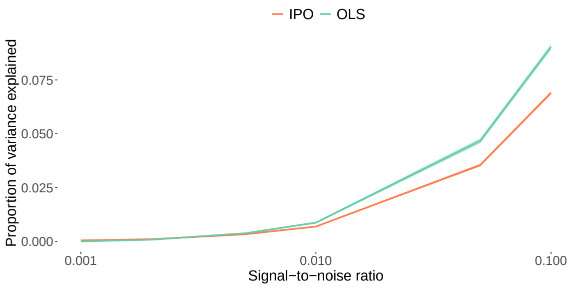


(c) $\rho = 0.50, \mathrm{res} = 5$.

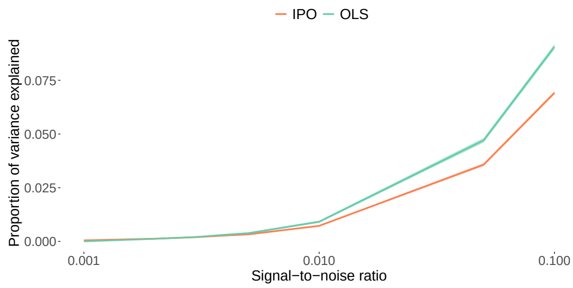


(d) $\rho = 0.75, \mathrm{res} = 5$.

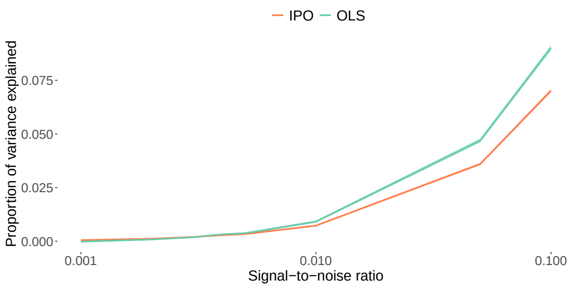
Figure 6: Out-of-sample MVO cost for IPO and OLS as of function of return signal-to-noise ratios.



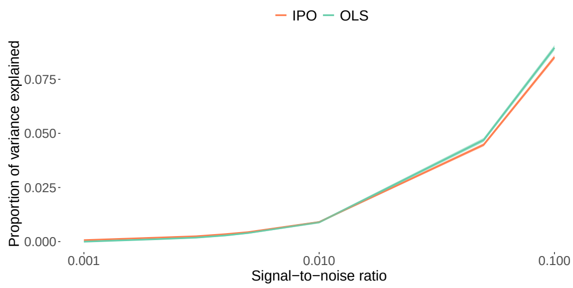
(a) $\rho = 0, \mathrm{res} = 5$.



(b) $\rho = 0.25, \mathrm{res} = 5$.



(c) $\rho = 0.50, \mathrm{res} = 5$.



(d) $\rho = 0.75, \mathrm{res} = 5$.

Figure 7: Out-of-sample PVE cost for IPO and OLS as of function of return signal-to-noise ratios.

4.2 Simulation 2: computational efficiency

Here, we compare the computational efficiency of the analytical IPO solution with the current state-of-the-art method based on implicit differentiation and iterative gradient descent, from here on denoted as IPO-GRAD. Note that the IPO-GRAD implementation is optimized such that the matrix factorization (Equation (23)), required to compute gradient, is performed once at the initialization of the algorithm. The IPO-GRAD coefficients are initialized by drawing from the standard normal distribution and the algorithm terminates when $\|\partial L/\partial \boldsymbol{\theta}\| < 10^{-6}$.

We generate synthetic asset returns, following the procedure outlined in Section 4.1, with $\rho = 0$, $\text{SNR} = 0.005$ and varying the number of assets, $d_z \in \{25, 50, 100, 250\}$. Each asset is assumed to have 3 unique auxiliary features, and therefore $d_\theta = 3d_z$. Tables 1 and 2, report the time, in seconds, taken by each method to compute the optimal regression coefficients for the unconstrained and equality constrained cases, respectively. For the IPO-GRAD method we also report the number of iterations of gradient descent. For each portfolio size, we report the average and 95%-ile range over 100 instances of simulated data. Observe that for problems with 100 or fewer assets, the computation time required to compute the optimal IPO coefficients analytically is comparable to the computation time required to compute the optimal OLS coefficients. In contrast, the IPO-GRAD method typically requires over 100 iterations of gradient descent and is anywhere from 10x - 1000x slower than the corresponding IPO method. We note that for problems of larger scale, the analytical IPO solution remains tractable and is on average 6x faster than the IPO-GRAD method.

No. Assets	OLS	IPO	IPO-GRAD	Iterations
25	0.029	0.071	4.333	178
	(0.028,0.032)	(0.07,0.08)	(3.966,5.076)	(164,210)
50	0.247	0.429	6.557	186
	(0.209,0.253)	(0.342,0.447)	(6.032,7.278)	(173,207)
100	0.545	1.7	17.642	200
	(0.491,0.638)	(1.495,1.837)	(16.03,21.301)	(183,247)
250	2.89	17.961	123.975	208.5
	(2.75,3.335)	(17.546,18.092)	(114.008,165.094)	(193,279)

Table 1: Time in seconds for computing the optimal OLS, IPO and IPO-GRAD coefficients for an unconstrained MVO problem. Results are averaged over 100 instances of simulated data.

No. Assets	OLS	IPO	IPO-GRAD	Iterations
25	0.029	0.088	4.664	176
	(0.028,0.032)	(0.085,0.094)	(4.333,5.587)	(163,211)
50	0.247	0.473	7.389	188
	(0.171,0.259)	(0.383,0.543)	(6.696,8.227)	(172,208)
100	0.549	2.025	19.711	200
	(0.492,0.669)	(1.855,2.161)	(18.034,23.449)	(183,241)
250	2.815	22.378	129.348	208
	(2.71,3.315)	(21.8,22.511)	(119.684,174.607)	(193,280)

Table 2: Time in seconds for computing the optimal OLS, IPO and IPO-GRAD coefficients for an equality constrained MVO problem. Results are averaged over 100 instances of simulated data.

4.3 Simulation 3: inequality constrained IPO

We now consider the more general case whereby the feasible region of the nominal program is defined by inequality constraints. In general, an analytical solution to the IPO Program (17) in the presence of lower-level inequality constraints is not possible. Furthermore, Program (17) is not convex in $\boldsymbol{\theta}$. As a result, the current state-of-the-art approach (IPO-GRAD), described in Section 3.1, is recommended in order to obtain locally optimal solutions.

The IPO-GRAD solution, however, is challenging for several reasons. First, in contrast to the traditional OLS approach, estimating the IPO coefficients by iterative methods can be computationally expensive; in particular as the number of assets, d_z, becomes large. Specifically, the IPO-GRAD framework requires solving, at each iteration of gradient descent, at most m constrained quadratic programs, where m is the total number of training observations. The time complexity therefore scales linearly with the number of training observations, m, and the total number of gradient descent iterations, n. Convex quadratic programs, however, are known to be solvable by interior-point methods in polynomial time, with worst-case time complexity on the order of $\mathcal{O}(d_z^3)$ [25]. Therefore, the worst-case time complexity for IPO framework is on the order of $\mathcal{O}(mnd_z^3)$. In practice, however, quadratic programs are typically solved with much fewer iterations than their worst-case bound [12]. Nonetheless, most real-world applications involve on the order of $10,000$ training observations and portfolio sizes on the order of 10 or 100. Therefore, for assets managers that construct portfolios from a very large pool of assets, estimating prediction model parameters by IPO-GRAD can be computationally burdensome and improving the efficiency of the integrated framework is therefore an open and interesting area of future research. Secondly, because the inequality constrained IPO problem is not convex, we have no guarantee that any particular local solution is globally optimal. Furthermore, estimating $\mathrm{Var}(\boldsymbol{\theta})$ and computing confidence intervals by standard parametric methods are challenged by the lack of convexity.

As a heuristic, we are interested in determining the out-of-sample efficacy of the analytical IPO solutions, presented in Sections 3.2 - 3.3, applied to the inequality constrained problem. Specifically, we compute the IPO optimal coefficients analytically by dropping the inequality constraints in the lower-level MVO problem. The realized policy, $\mathbf{z}^*(\mathbf{x}^{(i)}, \boldsymbol{\theta})$, however, enforces the inequality constraints in the out-of-sample evaluation period.

We generate synthetic asset returns, following the procedure outlined in Section 4.1, with $\rho = 0$, $\text{SNR} = 0.005$ and $d_z = 10$. Each asset is assumed to have 3 unique auxiliary features ($d_\theta = 3d_z$). The inequality constraints are standard box-constraints of the form:

$$-\gamma \leq \mathbf{z}_j \leq \gamma, \quad \forall j \in \{1, ..., d_z\},$$

and we consider several values of $\gamma \in \{0.05, 0.10, 0.25, 0.50, 0.75, 1, 2, 5, 10\}$. We also vary the risk aversion parameter $\delta \in \{1, 5, 10, 25\}$. Finally, asset mean returns are generated according to linear and nonlinear polynomial models of the form:

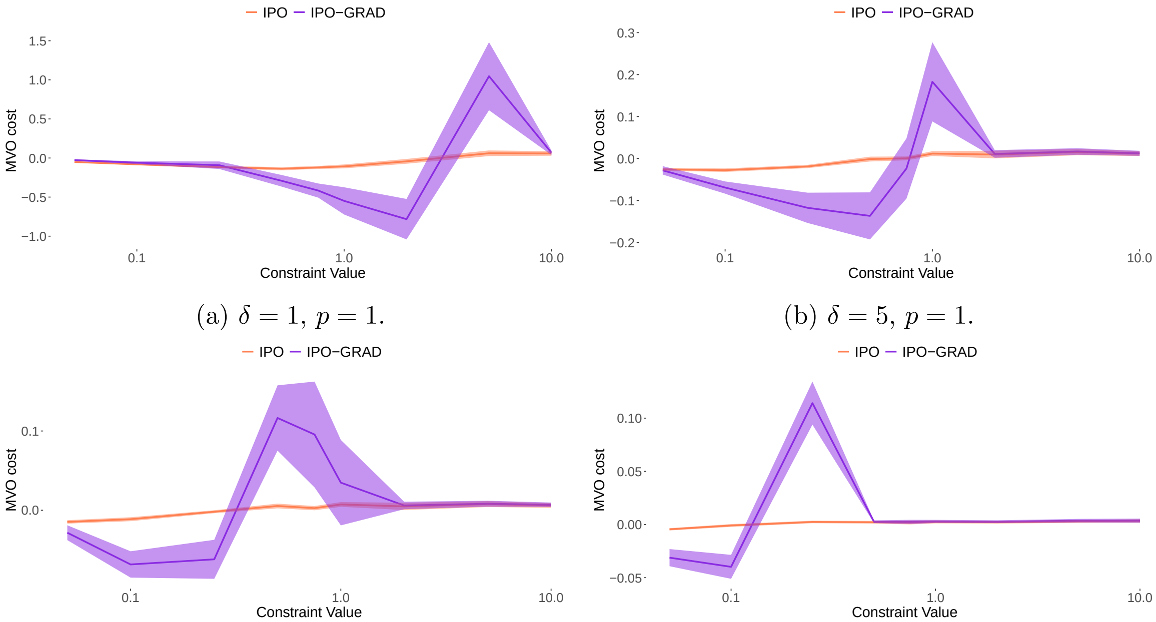
$$\mathbf{y}^{(i)} = \sum_{q=1}^{p} \mathbf{P} \operatorname{diag}(\overset{q}{\mathbf{x}^{(i)}}) \, \boldsymbol{\theta}_q + \tau \, \boldsymbol{\epsilon}^{(i)},$$

with $p \in \{1, 2, 4\}$.

Figures 8 - 10 compare the out-of-sample MVO cost for the IPO and IPO-GRAD methods as of function of the box constraint value, γ, and risk-aversion parameter, δ. For each value of γ, δ and p, we report the mean and 95%-ile range over 30 instances of simulated data. First, we would expect the out-of-sample performance of the IPO and IPO-GRAD methods to converge as γ increases. Furthermore as δ, increases, the point (along γ) at which the two solutions converge will naturally decrease. This effect is purely a consequence of the inequality constraints being non-active when either γ and/or δ are sufficiently large.

In Figure 8 asset returns are generated according to a linear ground truth model ($p = 1$). In all cases we observe that the IPO-GRAD does provide improved out-of-sample MVO costs when γ is sufficiently small ($\gamma < 0.5$). However, for moderate and large values of γ, the IPO-GRAD method provides no improvement in out-of-sample MVO costs in comparison to the IPO method. Furthermore, in Figures 9 and 10, asset returns are generated according to a quadratic ($p = 2$) and quartic ($p = 4$) ground truth model, respectively. We observe that over practically every value of γ and δ, the IPO method provides an equivalent, if not improved, out-of-sample MVO costs in comparison to the IPO-GRAD method. We note that, while not explicitly shown here, the IPO-GRAD method produces lower in-sample (training) MVO costs over every experiment instance, and is potentially overfitting the training data. Moreover, we note that in all experiments, the variance of the out-of-sample MVO costs generated by the IPO-GRAD method is substantially larger than that of the IPO method. The lack of convexity and uniqueness of solution in the IPO-GRAD formulation, along with the likelihood of model overfit, provides a potential explanation for this effect.

Finally, Table 3 reports the average time (in seconds) and 95%-ile range, taken by each method to compute the optimal regression coefficients. The results are averaged over all 360 instances of simulated data. For the IPO-GRAD method we also report the number of iterations of gradient descent. We observe that the IPO-GRAD method typically requires around 60 iterations of gradient descent and is on average 100x - 1000x slower than the corresponding IPO method. Note that the computation times reported here are for a relatively small portfolio and, given the computational complexity described above, we would expect the IPO method to provide an even larger computational advantage on medium and large sized portfolios. We therefore conclude that in the presence of inequality constraints, the IPO heuristic is a compelling alternative to the more computationally expensive IPO-GRAD solution.



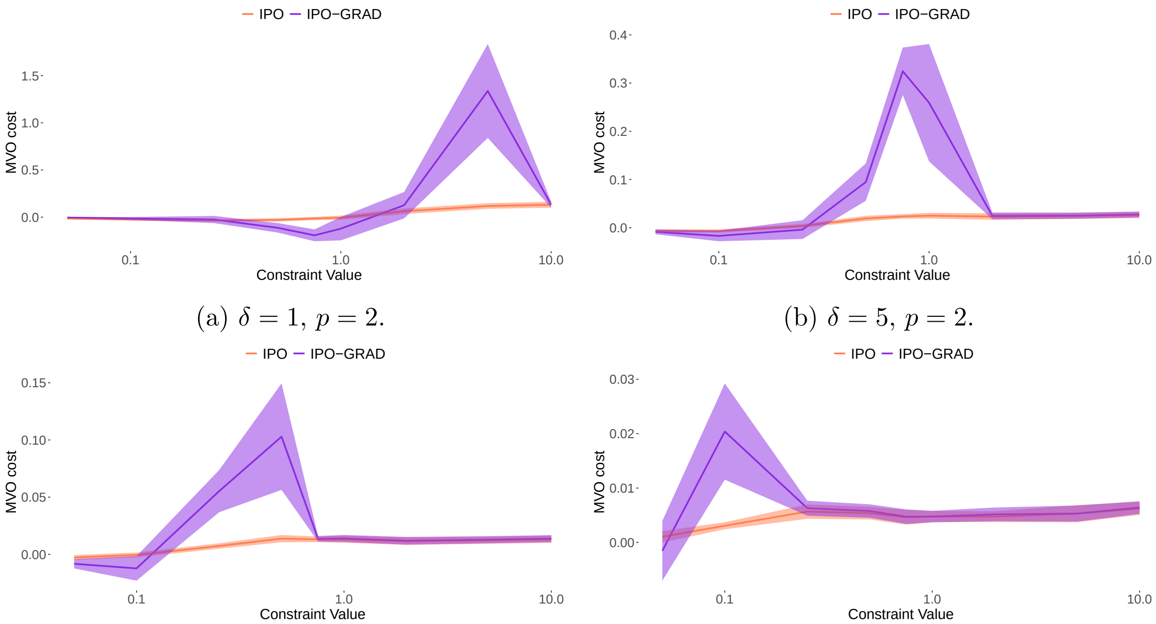
(a) $\delta = 1$, $p = 1$.

(b) $\delta = 5$, $p = 1$.

(c) $\delta = 10$, $p = 1$.

(d) $\delta = 25$, $p = 1$.

Figure 8: Out-of-sample MVO costs as of function of box constraint value with $p = 1$.



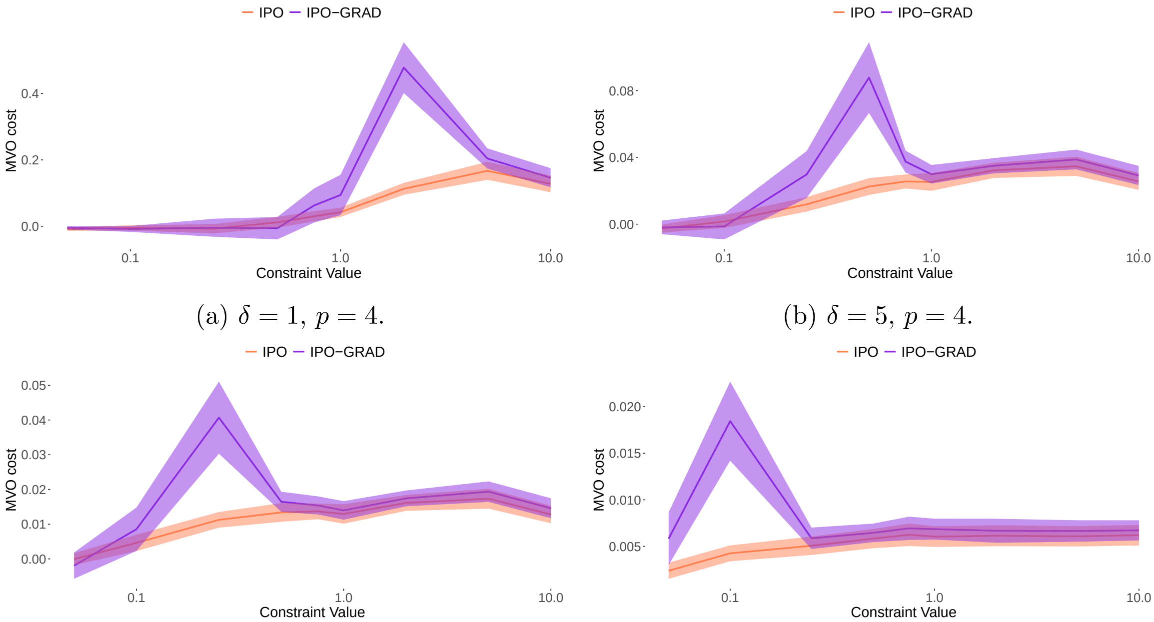
(a) $\delta = 1$, $p = 2$.

(b) $\delta = 5$, $p = 2$.

(c) $\delta = 10$, $p = 2$.

(d) $\delta = 25$, $p = 2$.

Figure 9: Out-of-sample MVO costs as of function of box constraint value with $p = 2$.



(a) $\delta = 1$, $p = 4$.

(b) $\delta = 5$, $p = 4$.

(c) $\delta = 10$, $p = 4$.

(d) $\delta = 25$, $p = 4$.

Figure 10: Out-of-sample MVO costs as of function of box constraint value with $p = 4$.

IPO	IPO-GRAD	Iterations
0.023	14.389	60
(0.022,0.024)	(6.9417,22.5718)	(29,94)

Table 3: Time in seconds for computing the optimal IPO and IPO-GRAD coefficients for an inequality constrained MVO problem. Results are averaged over 360 instances of simulated data.

5 Computational experiments

Experiment Setup:
We consider an asset universe of 24 commodity futures markets, described in Table 5. The daily price data is given from March 1986 through December 2020, and is provided by Commodity Systems Inc. Futures contracts are rolled on, at most, a monthly basis in order to remain invested in the most liquid contract, as measured by open-interest and volume. Arithmetic returns are computed directly from the price data.

In each experiment we follow Zumbach [44] and estimate the covariance matrix using an exponential moving average with a decay rate of 0.94. We consider both univariate and multivariate prediction models. The auxiliary feature, $\{\mathbf{x}^{(i)}\}$, for univariate models is the 252-day average return, or trend, for each market. The feature therefore represents a measure of the well-documented 'trend' factor, popular to many Commodity Trading Advisors (CTAs) and Hedge Funds (see for example [3], [13],[36]). The auxiliary feature for multivariate models is the 252-day trend and the carry for each market. We follow Koijen et al. [34] and define the carry as the expected convenience yield, or cost, for holding that commodity, and is estimated by the percent difference in price between the two futures contracts closest to expiry.

As we will see below, the majority of the IPO and OLS regression coefficients are not statistically significant at an individual market level. Indeed this is common and well document in many applications of financial forecasting (see for example [29, 36]). The lack of statistical significance may be indicative of low signal-to-noise levels and/or forecasting model misspecification - conditions that are likely favourable for the IPO model. Furthermore, the absence of statistical significance does not prohibit the development of profitable portfolio level trading strategies and indeed we observe in Table 4 that the auxiliary features are statistically significant at the $95\% - ile$ level when evaluated at an aggregate level across all markets.

Feature	Coefficient	Std. Error	T-Statistic	P-Value
Carry	0.3300	0.1654	1.9953	0.0460
Trend	0.0942	0.0324	2.9101	0.0036

Table 4: Univariate regression coefficients and t-statistic summary aggregated across all available markets.

Each day we form the optimal portfolio weight, $\mathbf{z}^*(\mathbf{x}^{(i)}, \boldsymbol{\theta})$ at the close of day i, and assume execution at the following close, $i+1$. In each experiment, described below, we consider two methods for estimating asset returns:

1. **OLS:** ordinary-least squares method, with prediction coefficients, $\hat{\boldsymbol{\theta}}$.

2. **IPO:** integrated prediction and optimization method, where $\boldsymbol{\theta}^*$ is determined by the IPO optimization framework described in Section 3.

We consider 6 experiments:

1. Unconstrained MVO program with univariate regression.

2. Unconstrained MVO program with multivariate regression.

3. Equality constrained MVO program with univariate regression.

4. Equality constrained MVO program with multivariate regression.

5. Inequality constrained MVO program with univariate regression.

6. Inequality constrained MVO program with multivariate regression.

The equality constrained MVO programs are market-neutral: $\mathbb{Z} = \{\mathbf{z}^T \mathbf{1} = 0\}$, whereas the inequality constrained MVO programs are both market-neutral and include lower bound and upper bound market constraints:

$$\mathbb{Z} = \{\mathbf{z}^T \mathbf{1} = 0, -0.125 \leq \mathbf{z} \leq 0.125\}.$$

Note that the results and discussion for the equality constrained MVO are very similar to the that of unconstrained MVO and can be found in Appendix B. In order to provide realistic annualized volatilities in the $10\% - 20\%$ range, we fix the risk-aversion parameter to $\delta = 50$. All experiments start in January 2000 and end in December 2020. For each experiment, the first 14 years (March 1986 through December 1999) is used to perform the initial parameter estimation. Thereafter, we apply a walk-forward training and testing methodology. The optimal regressions coefficients are updated every 2 years using all available data for parameter estimation and the optimal policy, $\mathbf{z}^*(\mathbf{x}^{(i)}, \boldsymbol{\theta})$, is then applied to the next out-of-sample 2 year segment. All performance is gross of trading costs and in excess of the risk-free rate.

Each model is evaluated on absolute and relative terms, with a focus on out-of-sample MVO cost and out-of-sample Sharpe ratio cost, provided by Equation (49).

$$c_{\mathrm{MVO}}(\mathbf{z}, \mathbf{y}) = -\mu(\mathbf{z}, \mathbf{y}) + \frac{\delta}{2}\sigma^2(\mathbf{z}, \mathbf{y}), \quad \text{and} \quad c_{\mathrm{SR}}(\mathbf{z}, \mathbf{y}) = -\frac{\mu(\mathbf{z}, \mathbf{y})}{\sigma(\mathbf{z}, \mathbf{y})} \tag{49}$$

where

$$\mu(\mathbf{z}, \mathbf{y}) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{z}^{T^{(i)}}\, \mathbf{y}^{(i)} \quad \text{and} \quad \sigma^2(\mathbf{z}, \mathbf{y}) = \frac{1}{m}\sum_{i=1}^{m} (\mathbf{z}^{T^{(i)}}\, \mathbf{y}^{(i)} - \mu(\mathbf{z}, \mathbf{y}))^2,$$

denote the mean and variance of realized daily returns. To quantify the magnitude and consistency of observed performance metrics, and to ensure our results are robust to potential outliers in the out-of-sample periods, scatterplots are created by bootstrapping the out-of-sample distribution using 1000 samples as follows:

1. For each $k \in \{1, 2, ..., 1000\}$, sample, without replacement, a batch, B_k, with $|B_k| = 252$ observations (1 year) from the out-of-sample period.

2. For each model, compute the average realized MVO and Sharpe ratio costs over the sample, using Equation (49).

Note, in each sample draw we use the same observation dates across both methodologies in order to fairly compare the realized nominal costs over the resulting sample. We report the dominance ratio (DR), which we define as the proportion of samples for which the realized cost of the IPO model is less than that of the OLS model.

Our experiments should be interpreted as a proof-of-concept, rather than a fully comprehensive financial study. That said, we believe that the results presented below provide compelling evidence for using IPO for estimating regression coefficients. In general, the IPO models exhibit lower out-of-sample MVO costs and improved economic outcomes in comparison to the traditional OLS-based 'predict, then optimize' approach.

Lastly, all experiments were conducted on an Apple Mac Pro computer (2.7 GHz 12-Core Intel Xeon E5,128 GB 1066 MHz DDR3 RAM) running macOS 'Catalina'. The software was written using the R programming language (version 4.0.0) and torch (version 0.2.0).

Asset Class		Market (Symbol)	
Energy	WTI crude (CL)	Heating oil (HO)	Gasoil (QS)
	RBOB gasoline (XB)		
Grain	Bean oil (BO)	Corn (C)	KC Wheat (KW)
	Soybean (S)	Soy meal (SM)	Wheat (W)
Livestock	Feeder cattle (FC)	Live cattle (LC)	Lean hogs (LH)
Metal	Gold (GC)	Copper (HG)	Palladium (PA)
	Platinum (PL)	Silver (SI)	
Soft	Cocoa (CC)	Cotton (CT)	Robusta Coffee (DF)
	Coffee (KC)	Canola (RS)	Sugar (SB)

Table 5: Futures market universe. Symbols follow Bloomberg market symbology. Data is provided by Commodity Systems Inc (CSI).

5.1 Experiment 1: $\mathbb{Z}$ unconstrained, $f(\mathbf{x}, \boldsymbol{\theta})$ univariate

Economic performance metrics and average out-of-sample MVO costs are provided in Table 6 for the time period of 2000-01-01 to 2020-12-31 for the unconstrained MVO portfolios with univariate prediction models. Equity growth charts for the same time period are provided in Figure 11. We first observe that the IPO model provides higher absolute and risk-adjusted performance, as measured by the MVO cost and Sharpe ratio. Indeed the IPO model produces an out-of-sample MVO cost that is approximately 50% lower and a Sharpe ratio that is approximately 100% larger than that of the OLS model. Furthermore, the IPO models provide more conservative risk metrics, as measured by portfolio volatility, value-at-risk (VaR), and average drawdown (Avg DD). These results are highly encouraging for the IPO model.

In Figure 12 we compare the realized MVO and Sharpe ratio costs across 1000 out-of-sample realizations. In general we observe that the IPO model exhibits consistently lower MVO costs and generally higher Sharpe ratios than the OLS model. In Figure 12(a) we report a dominance ratio of 97% meaning that the IPO model realizes a lower MVO cost in 97% of samples in comparison to the OLS model. Figure 12(b) reports a dominance ratio of 68%.

In Figure 13 we report the estimated univariate regression coefficients and ± 1 standard error bar for the last out-of-sample data fold. As stated earlier, it is clear that the majority of the IPO and OLS regression coefficients are not statistically significant at an individual market basis. Note that for some markets, the IPO model provides very different regression coefficients, in both magnitude and sign, compared to the OLS coefficients. In particular we observe that, with the exception of Cocoa (CC), all IPO regression coefficients are positive. In contrast, 33% of OLS coefficients are negative.

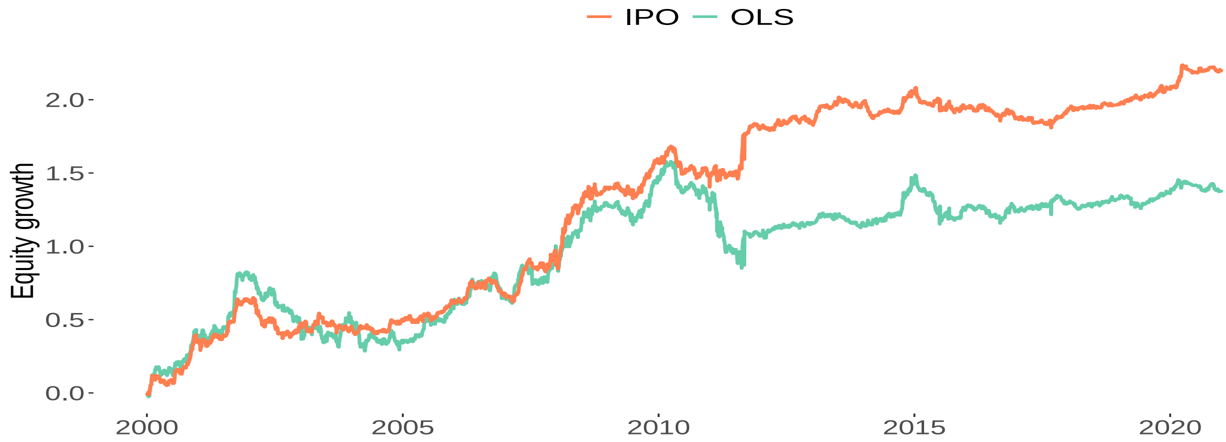
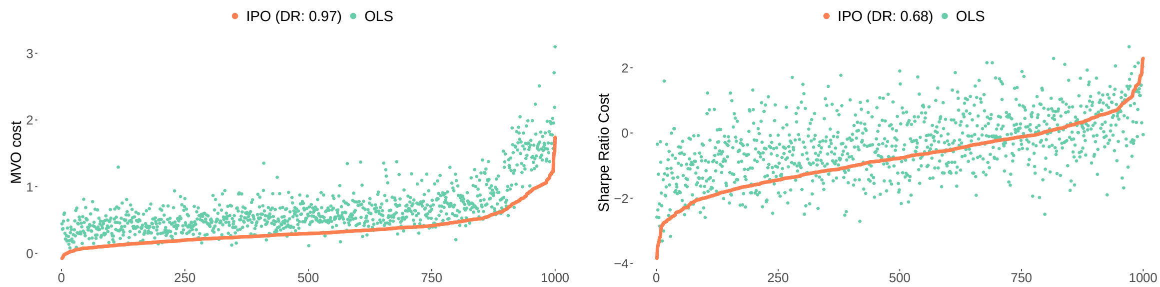


Figure 11: Out-of-sample log-equity growth for the unconstrained nominal mean-variance program and multivariate IPO and OLS prediction model.

	Annual Return	Sharpe Ratio	Volatility	Avg Drawdown	Value at Risk	MVO Cost
IPO	0.1026	0.7593	0.1352	-0.0275	-0.0107	0.3544
OLS	0.0644	0.3735	0.1725	-0.0426	-0.0142	0.6792

Table 6: Out-of-sample MVO costs and economic performance metrics for unconstrained mean-variance portfolios with univariate IPO and OLS prediction models.



(a) Out-of-sample MVO cost. (b) Out-of-sample Sharpe ratio cost.

Figure 12: Realized out-of-sample MVO and Sharpe ratio costs for the unconstrained nominal mean-variance program and univariate IPO and OLS prediction models.

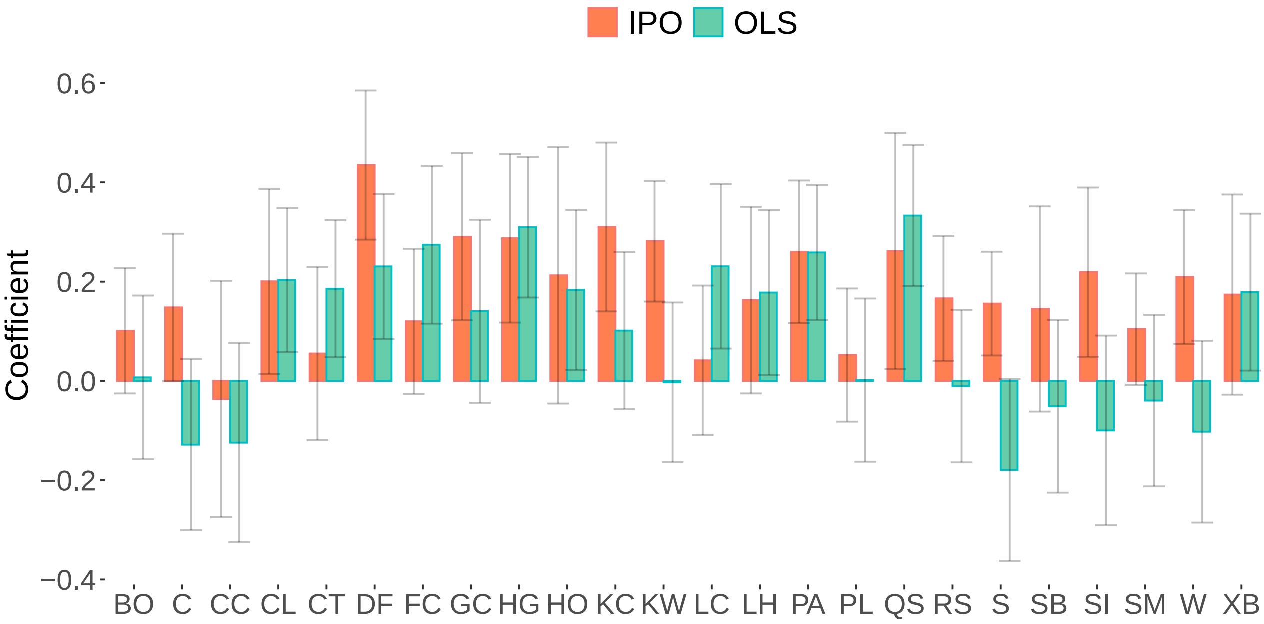


Figure 13: Optimal IPO and OLS regression coefficients for the unconstrained nominal mean-variance program and univariate prediction model.

5.2 Experiment 2: $\mathbb{Z}$ unconstrained, $f(\mathbf{x}, \boldsymbol{\theta})$ multivariate

Economic performance metrics and average out-of-sample MVO costs are provided in Table 7 for the time period of 2000-01-01 to 2020-12-31 for the unconstrained MVO portfolios with multivariate prediction models. Equity growth charts for the same time period are provided in Figure 14. Again we observe that the IPO model provides higher absolute and risk-adjusted performance and in general more conservative risk metrics. The IPO model produces an out-of-sample MVO cost that is approximately 50% lower and a Sharpe ratio that is approximately 100% larger than that of the OLS model. In Figure 15 we compare the realized MVO and Sharpe ratio costs across 1000 out-of-sample realizations. Again we observe that the IPO model exhibits consistently lower MVO costs with a dominance ratio of 99% and generally lower Sharpe ratio costs with a dominance ratio of 65%.

In Figure 16 we report the estimated regression coefficients and ± 1 standard error bar for the last out-of-sample data fold. As before, the majority of the IPO and OLS regression coefficients are not statistically significant at an individual market basis. Figures 16 (a) and 16 (b) report the estimated regression coefficients for the Carry and Trend auxiliary features, respectively. Again we observe that the IPO model provides very different regression coefficients, in both magnitude and sign, compared to the OLS coefficients. Observe that in the multivariate regression model, 50% of the OLS Trend coefficients are negative. In contrast, the IPO model has only 3 (12.5%) negative coefficients: Cocoa (CC), Live Cattle (LC) and Platinum (PL). Furthermore, in many cases such as Feeder Cattle (FC) and Soymeal (SM), the OLS coefficients are relatively large (> 0.30) whereas the corresponding IPO coefficients are effectively zero. Lastly note that the magnitude of the coefficients is approximately 10x larger than the corresponding trend coefficients and is a result of the carry auxiliary feature values being approximately an order of magnitude smaller.

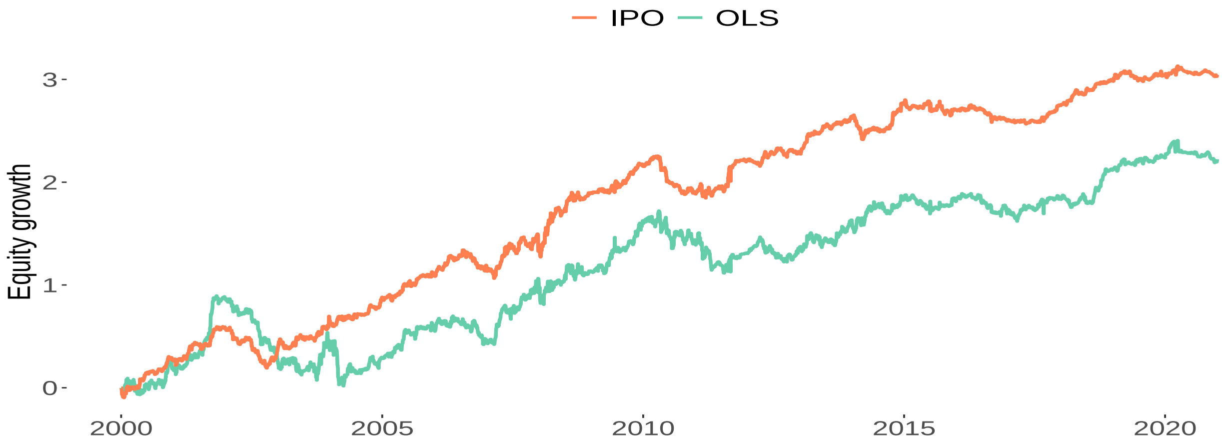
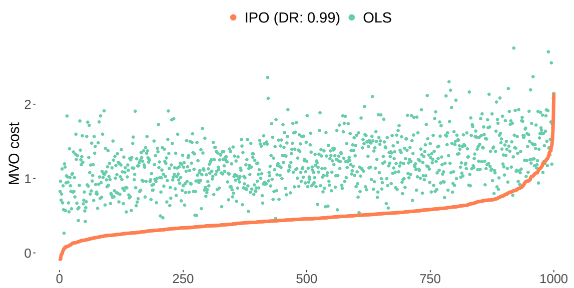


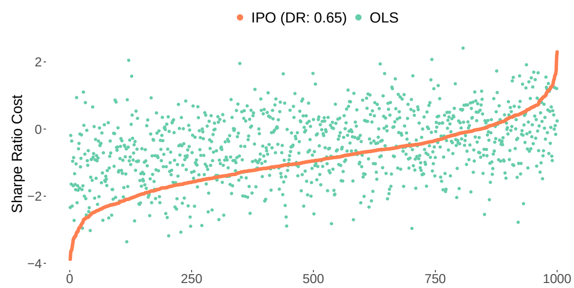
Figure 14: Out-of-sample log-equity growth for the unconstrained nominal mean-variance program and multivariate IPO and OLS prediction model.

	Annual Return	Sharpe Ratio	Volatility	Avg Drawdown	Value at Risk	MVO Cost
IPO	0.1416	0.8835	0.1603	-0.0294	-0.0138	0.5004
OLS	0.1034	0.4477	0.2310	-0.0438	-0.0208	1.2308

Table 7: Out-of-sample MVO costs and economic performance metrics for unconstrained mean-variance portfolios with multivariate IPO and OLS prediction models.

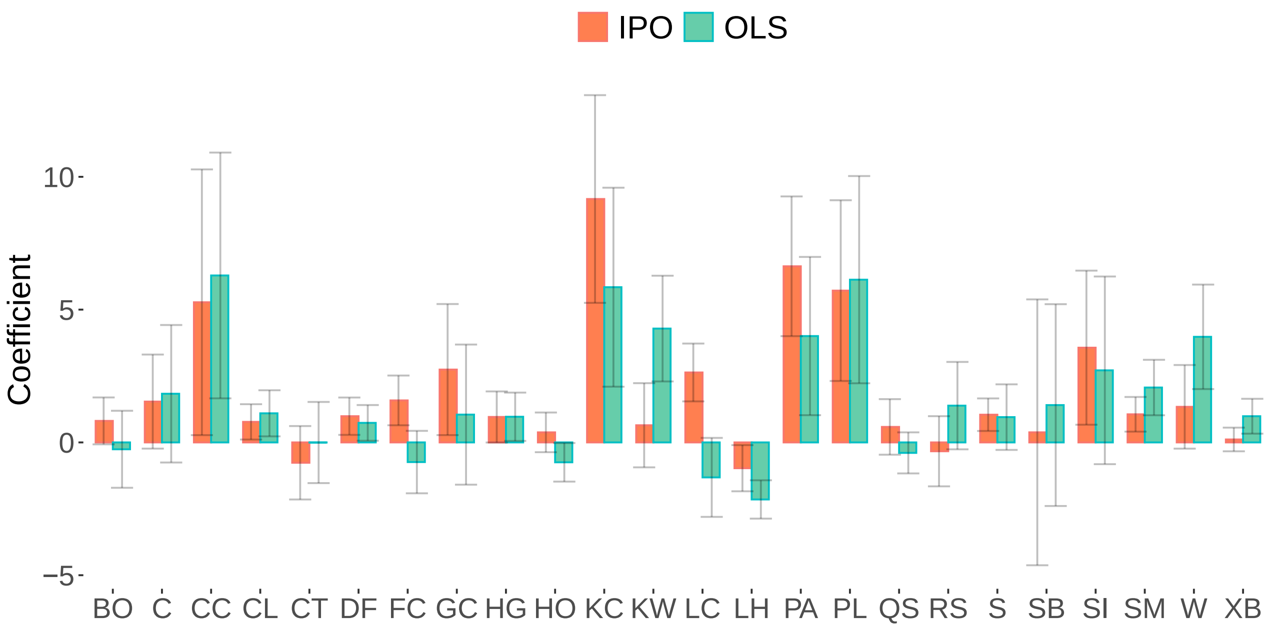


(a) Out-of-sample MVO cost.

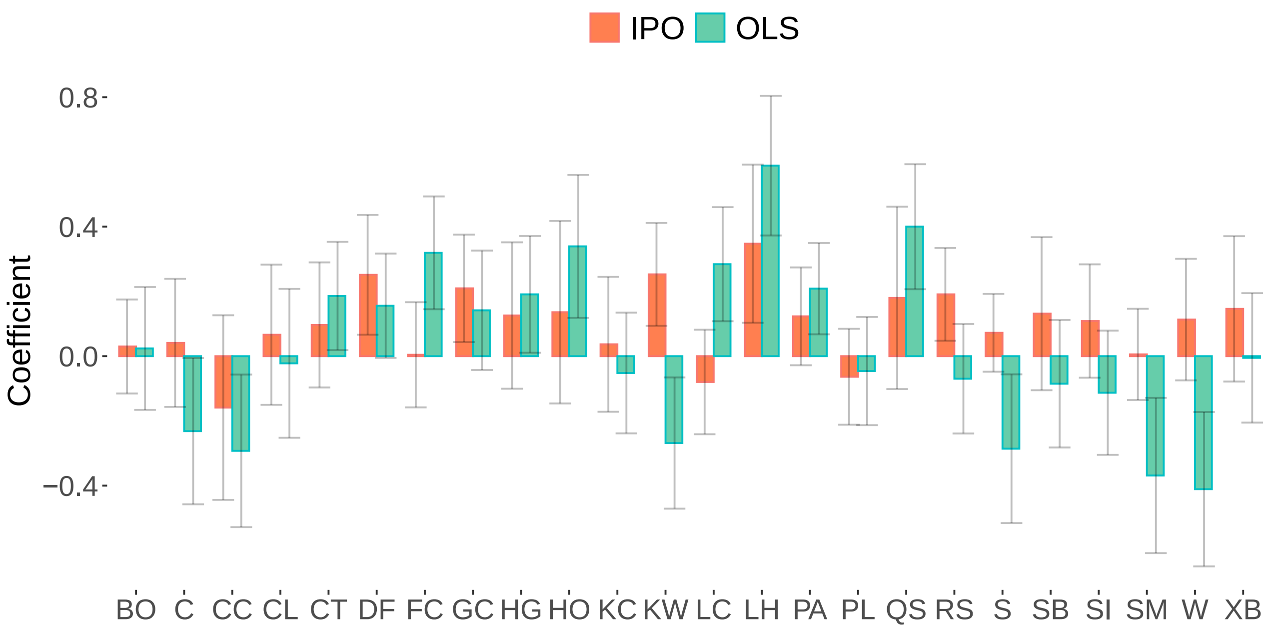


(b) Out-of-sample Sharpe ratio cost.

Figure 15: Realized out-of-sample MVO and Sharpe ratio costs for the unconstrained nominal mean-variance program and multivariate IPO and OLS prediction models.



(a) Auxiliary feature: Carry.



(b) Auxiliary feature: Trend.

Figure 16: Optimal IPO and OLS regression coefficients for the unconstrained nominal mean-variance program and multivariate prediction model.

5.3 Experiment 5: $\mathbb{Z} = \{Az = b, Gz \leq h\}$, $f(\mathbf{x}, \boldsymbol{\theta})$ univariate

Economic performance metrics and average out-of-sample MVO costs are provided in Table 8 for the time period of 2000-01-01 to 2020-12-31 for the constrained MVO portfolios with univariate prediction models. Equity growth charts for the same time period are provided in Figure 17. First, observe that the annual returns, risk and MVO costs are substantially smaller in the presence of portfolio constraints. Indeed this is consistent with the fact that box constraints are themselves a form of portfolio model regularization [31]. Nonetheless, we observe that the IPO model produces an out-of-sample MVO cost that is approximately 50% lower and a Sharpe ratio that is approximately 85% larger than that of the OLS model. In Figure 21 we compare the realized MVO and Sharpe ratio costs across 1000 bootstrapped sample realizations. Again we observe that the IPO model produces lower MVO and Sharpe ratio costs on average. Observe, however, that the dominance ratios are more modest, with values in the 60%-70% range. This result is intuitive and we would expect the out-of-sample performance of the two models to converge as the portfolio constraints become more strict. Indeed the IPO and OLS model would yield identical results in the limit where the portfolio constraints define a single weight, irrespective of the mean and covariance estimation. Lastly, in Figure 19 we report the estimated univariate regression coefficients and ± 1 standard error bar for the last out-of-sample data fold. Recall that the IPO coefficients are obtained by first dropping the inequality constraints and then solving analytically for $\boldsymbol{\theta}^*$ by Equation (43). The observations and differences between the optimal IPO and OLS coefficients are similar to those discussed in Section 5.1.

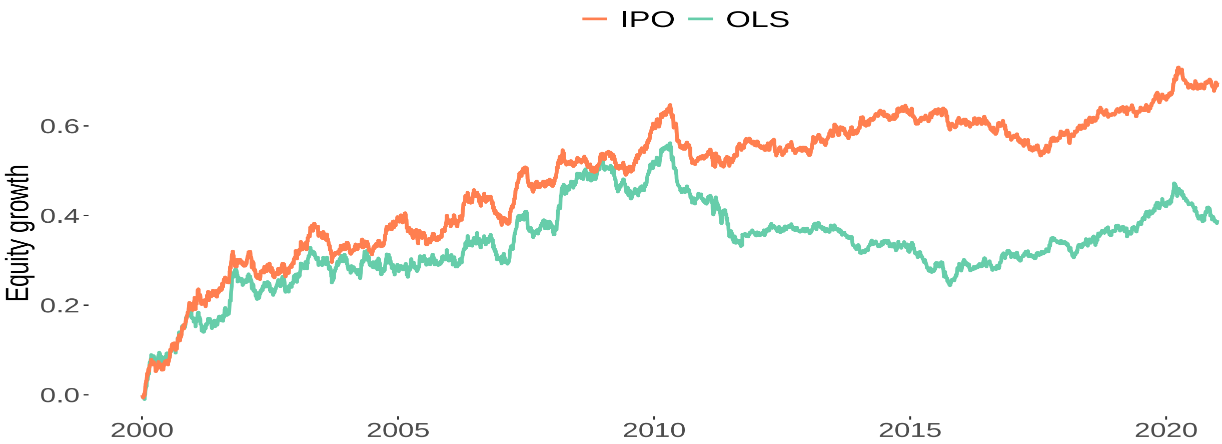
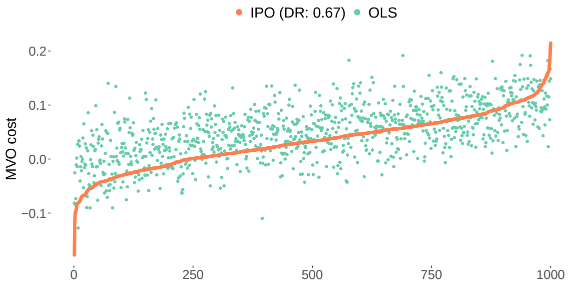


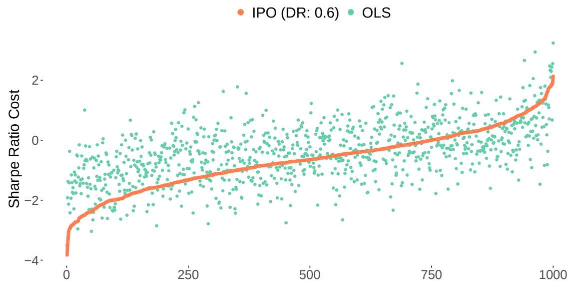
Figure 17: Out-of-sample log-equity growth for the inequality constrained nominal mean-variance program and multivariate IPO and OLS prediction model.

	Annual Return	Sharpe Ratio	Volatility	Avg Drawdown	Value at Risk	MVO Cost
IPO	0.0324	0.6310	0.0513	-0.0116	-0.0052	0.0335
OLS	0.0181	0.3421	0.0529	-0.0174	-0.0053	0.0520

Table 8: Out-of-sample MVO costs and economic performance metrics for inequality constrained mean-variance portfolios with univariate IPO and OLS prediction models.



(a) Out-of-sample MVO cost.



(b) Out-of-sample Sharpe ratio cost.

Figure 18: Realized out-of-sample MVO and Sharpe ratio costs for the inequality constrained nominal mean-variance program and univariate IPO and OLS prediction models.

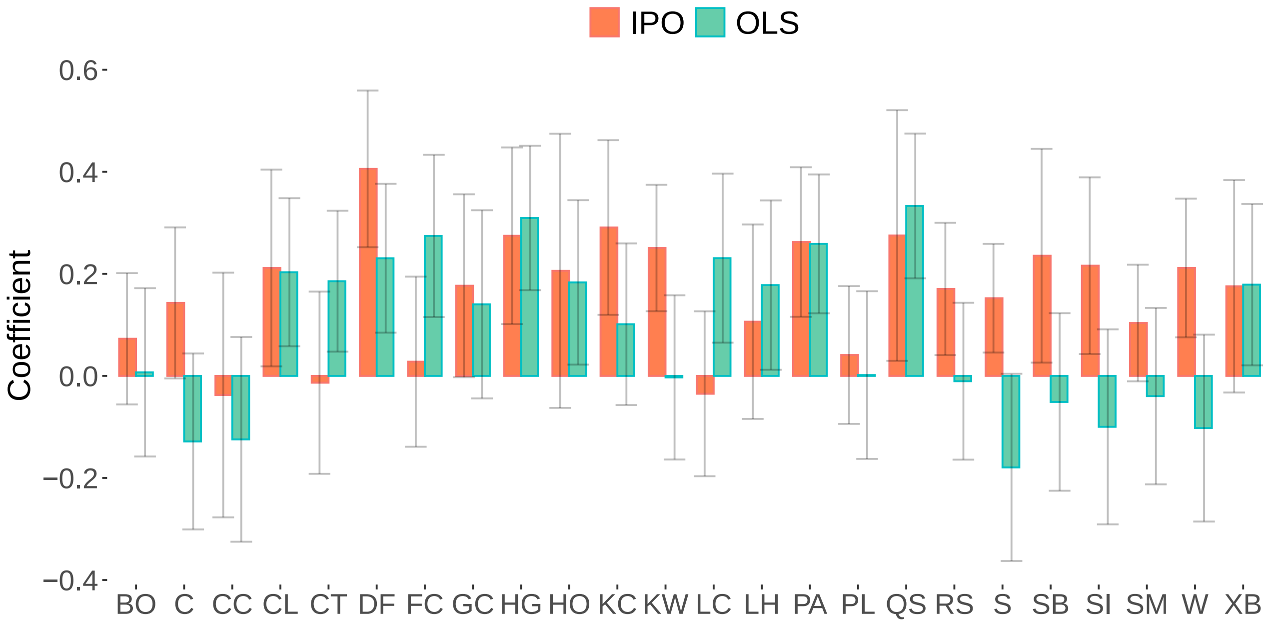


Figure 19: Optimal IPO and OLS regression coefficients for the equality constrained nominal mean-variance program and univariate prediction model.

5.4 Experiment 6: $\mathbb{Z} = \{Az = b, Gz \leq h\}$, $f(\mathbf{x}, \boldsymbol{\theta})$ multivariate

Economic performance metrics and average out-of-sample MVO costs are provided in Table 9 for the time period of 2000-01-01 to 2020-12-31 for the inequality constrained MVO portfolios with multivariate prediction models. Equity growth charts for the same time period are provided in Figure 20. Once again we observe that the IPO model provides modestly higher absolute and risk-adjusted performance and in general more conservative risk metrics. The IPO model produces an out-of-sample MVO cost that is approximately 60% lower and a Sharpe ratio that is approximately 25% larger than that of the OLS model. In Figure 21 we compare the realized MVO and Sharpe ratio costs across 1000 out-of-sample realizations. Again we observe more modest dominance ratios with values in the 55%-65% range. We observe that the IPO model provides a modest improvement to performance in comparison to the OLS model; a likely result of lower prediction model misspecification and improved portfolio regularization by virtue of the box constraints. The estimated regression coefficients are provided in Figure 22 and the findings are similar to those described in Section 5.2.

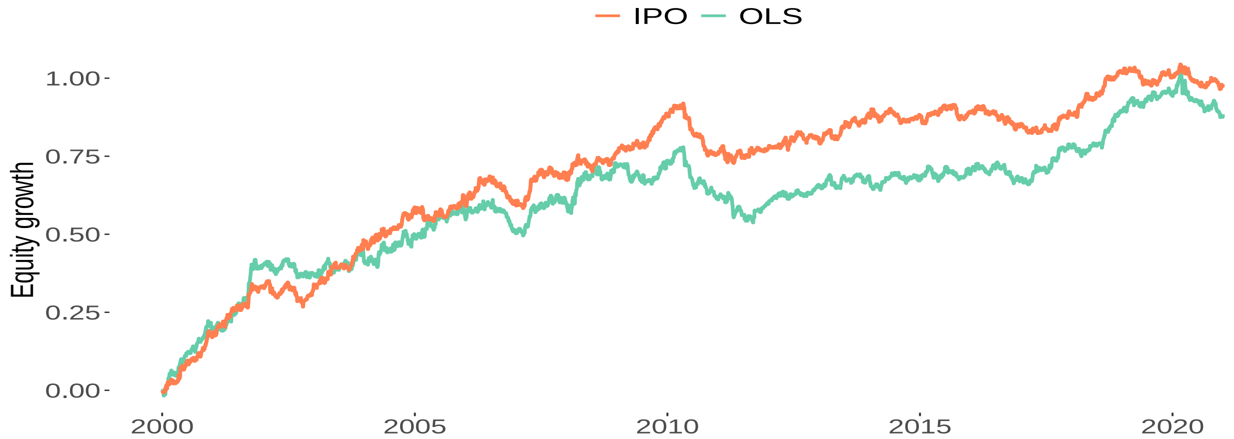
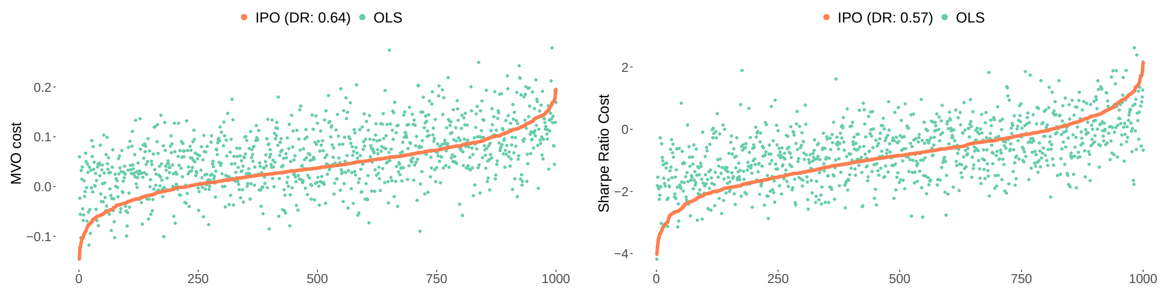


Figure 20: Out-of-sample log-equity growth for the inequality constrained nominal mean-variance program and multivariate IPO and OLS prediction model.

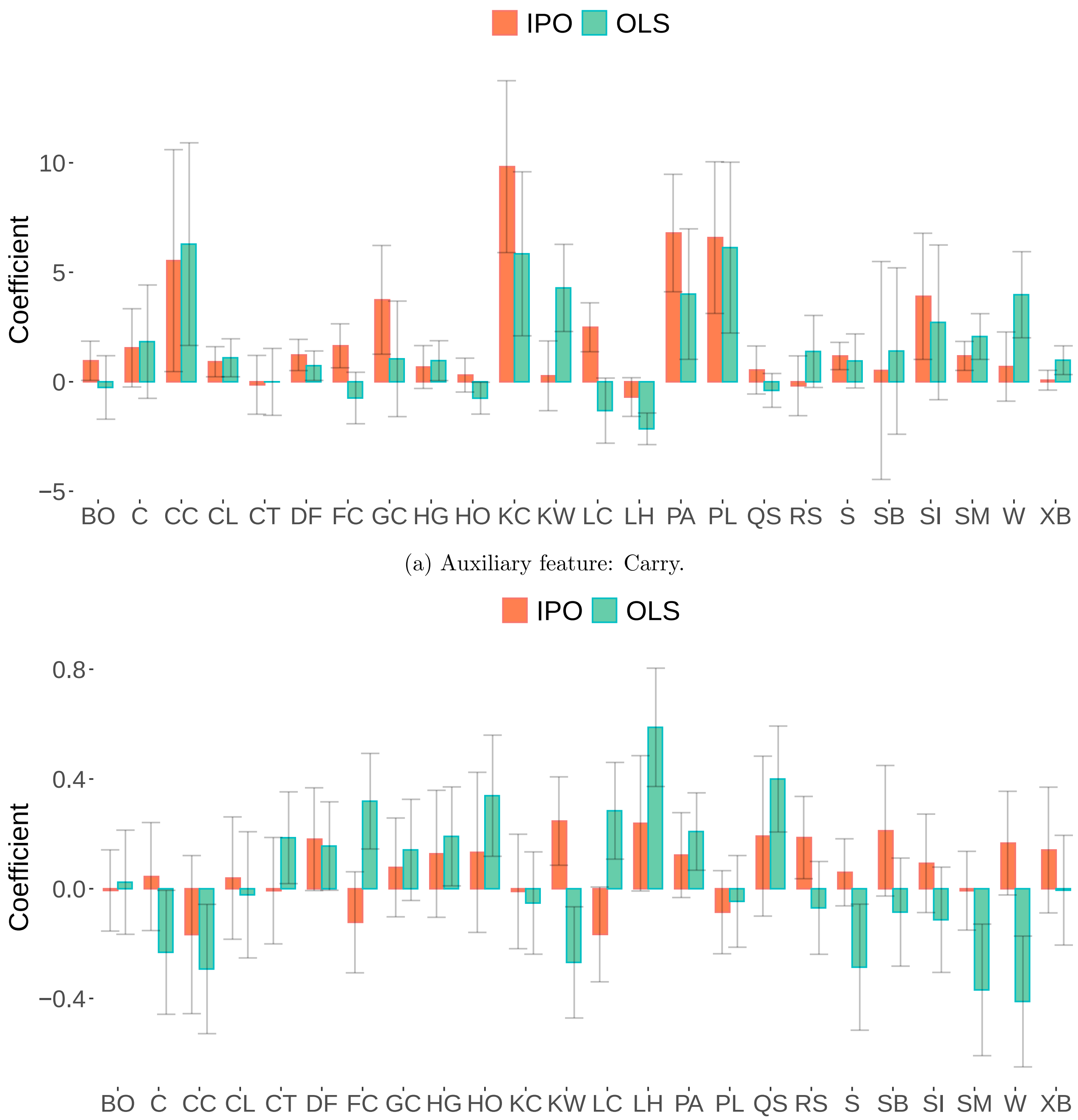
	Annual Return	Sharpe Ratio	Volatility	Avg Drawdown	Value at Risk	MVO Cost
IPO	0.0456	0.7937	0.0574	-0.0119	-0.0057	0.0369
OLS	0.0411	0.6488	0.0634	-0.0145	-0.0063	0.0593

Table 9: Out-of-sample MVO costs and economic performance metrics for inequality constrained mean-variance portfolios with multivariate IPO and OLS prediction models.



(a) Out-of-sample MVO cost.
(b) Out-of-sample Sharpe ratio cost.

Figure 21: Realized out-of-sample MVO and Sharpe ratio costs for the inequality constrained nominal mean-variance program and multivariate IPO and OLS prediction models.



(a) Auxiliary feature: Carry.

(b) Auxiliary feature: Trend.

Figure 22: Optimal IPO and OLS regression coefficients for the equality constrained nominal mean-variance program and multivariate prediction model.

6 Conclusion and future work

In this paper we proposed an integrated prediction and optimization (IPO) framework for optimizing regression coefficients in the context of a mean-variance portfolio optimization. We structured the problem as a stochastic program where, for a fixed instantiation of regression coefficients, we solve a series of deterministic nominal mean-variance optimization programs. We investigated the IPO framework under both univariate and multivariate regression settings and considered the nominal program under various forms of constraints. In a general setting, we presented the current state-of-the-art approach (IPO-GRAD) and restructured the IPO problem as a neural network with a differentiable quadratic programming layer. Where possible, we provided closed-form analytical solutions for the optimal IPO regression coefficients, $\boldsymbol{\theta}^*$, and the sufficient conditions for uniqueness. We described the sampling distribution properties of $\boldsymbol{\theta}^*$ and provided the conditions for which $\boldsymbol{\theta}^*$ is an unbiased estimator of $\boldsymbol{\theta}$ and provided the expression for the variance.

Extensive numerical simulations demonstrate the computational and performance advantage of the analytical IPO methodology. We demonstrated that, over a wide range of realistic signal-to-noise ratios, the IPO model outperforms the OLS model in terms of minimizing out-of-sample MVO costs. This is true even when the underlying 'ground-truth' return generating process is linear in the auxiliary variables. We demonstrated, for a wide range of portfolio sizes, the computational advantage of computing the IPO coefficients analytically, which is on average 10x-1000x faster than the IPO-GRAD methodology. We briefly discussed the computational complexity of the IPO-GRAD methodology and proposed a heuristic which drops the inequality constraints during parameter estimation and invokes the analytical IPO solution. We find that in many instances the IPO-GRAD model overfits the training data, whereas the analytical IPO model produces solutions with lower out-of-sample variance, and in many cases, improved out-of-sample MVO costs. We concluded with several experiments using global futures data, under various forms of constraints and prediction model specifications. Out-of-sample results demonstrate that the IPO model provided lower realized MVO costs and superior economic performance in comparison to the traditional OLS 'predict then optimize' approach.

In the presence of general inequality constraints we determined that the current state-of-the-art IPO model is computationally burdensome and has a tendency to overfit the training data. We believe that methods for regularizing both the prediction and the nominal optimization program, as well as methods for choosing the 'best' feature subsets are an interesting area of future research and may lead to approaches that improve upon our heuristic IPO solution. Future work also includes incorporating other forms of prediction models into the IPO framework as well as exploring methods for performing the more difficult joint prediction of asset returns and covariances.

References

[1] Akshay Agrawal, Brandon Amos, Shane Barratt, Stephen Boyd, Steven Diamond, and J. Zico Kolter. Differentiable convex optimization layers. In *Advances in Neural Information Processing Systems*, volume 32, pages 9562–9574. Curran Associates, Inc., 2019.

[2] Brandon Amos and J. Zico Kolter. Optnet: Differentiable optimization as a layer in neural networks. *CoRR*, abs/1703.00443, 2017. URL `http://arxiv.org/abs/1703.00443`.

[3] Nick Baltas and Robert Kosowski. Momentum strategies in futures markets and trend-following funds, 2012.

[4] Gah-Yi Ban and Cynthia Rudin. The big data newsvendor: Practical insights from machine learning. *Operations Research*, 67(1):90–108, 2019.

[5] L. Bauwens, S. Laurent, and J. Rombouts. Multivariate garch models: A survey. *Econometrics eJournal*, 2003.

[6] Dimitris Bertsimas and Nathan Kallus. From predictive to prescriptive analytics. *Management Science*, 66(3):1025–1044, 2020.

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

[8] Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307 – 327, 1986. ISSN 0304-4076. doi: https://doi.org/10.1016/0304-4076(86)90063-1. URL `http://www.sciencedirect.com/science/article/pii/0304407686900631`.

[9] Tim Bollerslev. Modelling the coherence in short-run nominal exchange rates: A multivariate generalized arch model. *The Review of Economics and Statistics*, 72(3):498–505, 1990. ISSN 00346535, 15309142. URL `http://www.jstor.org/stable/2109358`.

[10] Tim Bollerslev, Robert F. Engle, and Jeffrey M. Wooldridge. A capital asset pricing model with time-varying covariances. *Journal of Political Economy*, 96(1):116–131, 1988.

[11] Tim Bollerslev, Robert Engle, and Daniel B. Nelson. Arch models. In R. F. Engle and D. McFadden, editors, *Handbook of Econometrics*, volume 4, pages 2961–3031. Elsevier, 1 edition, 1994.

[12] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[13] Benjamin Bruder, Tung-Lam Dao, Jean-Charles Richard, and Thierry Roncalli. Trend Filtering Methods for Momentum Strategies. *Lyxor Asset Management*, 2011.

[14] Binbin Chen, Shih-Feng Huang, and Guangming Pan. High dimensional mean variance optimization through factor analysis. *Journal of Multivariate Analysis*, 133:140 – 159, 2015. ISSN 0047-259X.

[15] Vijay Kumar. Chopra and William T. Ziemba. The effect of errors in means, variances, and covariances on optimal portfolio choice. *The Journal of Portfolio Management*, 19 (2):6–11, 1993. ISSN 0095-4918. doi: 10.3905/jpm.1993.409440. URL `https://jpm.pm-research.com/content/19/2/6`.

[16] Roger G Clarke, Harindra de Silva, and Robert Murdock. A factor approach to asset allocation. *The Journal of Portfolio Management*, 32(1):10–21, 2005.

[17] Priya L. Donti, Brandon Amos, and J. Zico Kolter. Task-based End-to-end Model Learning. *CoRR*, abs/1703.04529, 2017. URL `http://arxiv.org/abs/1703.04529`.

[18] Holger Drees and Catalin Starica. A simple non-stationary model for stock returns. 2002.

[19] Adam N. Elmachtoub and Paul Grigas. Smart "predict, then optimize". *arXiv*, 2020.

[20] Robert F. Engle. Autoregressive conditional heteroskedasticity with estimates of the variance of uk inflation. *Econometrica*, 50(1):987–1008, 1982.

[21] Eugene F. Fama and Kenneth R. French. The cross section of expected stock returns. *Journal of Financial Finance*, 47(2), 1992.

[22] Eugene F. Fama and Kenneth R. French. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(3):3–56, 1993.

[23] Eugene F. Fama and Kenneth R. French. A five-factor asset pricing model. *Journal of Financial Economics*, 116(1):1 – 22, 2015. ISSN 0304-405X. doi: https://doi.org/10.1016/j.jfineco.2014.10.010. URL `http://www.sciencedirect.com/science/article/pii/S0304405X14002323`.

[24] D. Goldfarb and G. Iyengar. Robust portfolio selection problems. *Mathematics of Operations Research*, 28(1):1–38, 2003.

[25] D. Goldfarb and Shucheng Liu. An o(n3l) primal interior point algorithm for convex quadratic programming. *Mathematical Programming*, 49:325–340, 1991.

[26] Stephen Gould, Basura Fernando, Anoop Cherian, Peter Anderson, Rodrigo Santa Cruz, and Edison Guo. On differentiating parameterized argmin and argmax problems with application to bi-level optimization. *CoRR*, abs/1607.05447, 2016. URL `http://arxiv.org/abs/1607.05447`.

[27] Trevor Hastie, Robert Tibshirani, and Ryan J. Tibshirani. Extended comparisons of best subset selection, forward stepwise selection, and the lasso. 2017.

[28] Der-Ann Hsu, Robert B. Miller, and Dean W. Wichern. On the stable paretian behavior of stock-market prices. *Journal of the American Statistical Association*, 69(345):108–113, 1974.

[29] Dashan Huang, Jiangyuan Li, Liyao Wang, and Guofu Zhou. Time series momentum: Is it there? *Journal of Financial Economics*, 135(3):774–794, 2020. ISSN 0304-405X. doi: https://doi.org/10.1016/j.jfineco.2019.08.004. URL `https://www.sciencedirect.com/science/article/pii/S0304405X19301953`.

[30] Ronen Israel, Bryan Kelly, and Tobias Moskowitz. Can machines learn finance. *Journal Of Investment Management*, 18, 11 2020.

[31] Ravi Jagannathan and Tongshu Ma. Risk reduction in large portfolios: Why imposing the wrong constraints helps. *The Journal of Finance*, 58(4):1651–1683, 2003.

[32] A. G. Jones and C. Taylor. Solving inverse problems in computer vision by scale space reconstruction. In *MVA*, 1994.

[33] R. Kannan, G. Bayraksan, and James R. Luedtke. Data-driven sample average approximation with covariate information. 2020.

[34] Ralph S.J. Koijen, Tobias J. Moskowitz, Lasse Heje Pedersen, and Evert B. Vrugt. Carry. *Journal of Financial Economics*, 127(2):197–225, 2018.

[35] H. Markowitz. Portfolio selection. *Journal of Finance*, 7(1):77–91, 1952.

[36] Tobias J. Moskowitz, Yao Hua Ooi, and Lasse Pedersen. Time series momentum. *Journal of Financial Economics*, 104(2):228–250, 2012.

[37] R.R. Officer. A time series examination of the market factor of the new york stock exchange. *University of Chicago PhD dissertation*, 1971.

[38] Gregory Ongie, Ajil Jalal, Christopher A. Metzler, Richard G. Baraniuk, Alexandros G. Dimakis, and Rebecca Willett. Deep learning techniques for inverse problems in imaging, 2020.

[39] P. Ranjan, R. Haynes, and R. Karsten. A Computationally Stable Approach to Gaussian Process Interpolation of Deterministic Computer Simulation Data. *Technomectrics*, 52(4):366–378, 2011.

[40] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczynski. *Lectures on Stochastic Programming*. Society for Industrial and Applied Mathematics, 2009. doi: 10.1137/1.9780898718751.

[41] Ankur Sinha, Tanmay Khandait, and Raja Mohanty. A gradient-based bilevel optimization approach for tuning hyperparameters in machine learning, 2020.

[42] Catalin Starica and Clive Granger. Nonstationarities in stock returns. *The Review of Economics and Statistics*, 87(3):503–522, 2005.

[43] V. Vapnik. Principles of risk minimization for learning theory. In J. Moody, S. Hanson, and R. P. Lippmann, editors, *Advances in Neural Information Processing Systems*, volume 4, pages 831–838. Morgan-Kaufmann, 1992.

[44] G. Zumbach. The riskmetrics 2006 methodology. 2007.

A Appendix

We begin with the following proposition that will become useful later.

Proposition 9. *Let* $\mathbf{V} \in \mathbb{R}^{m\times m}$ *be a symmetric positive definite matrix. Let* $\mathbf{B} \in \mathbb{R}^{\mathbf{m}\times\mathbf{n}}$ *and consider the quadratic form* $\mathbf{A} = \mathbf{B^T}\,\mathbf{V}\,\mathbf{B}$*. Then* $\mathbf{A}$ *is a symmetric positive definite matrix if* $\mathbf{B}$ *has full column rank.*

Proof. The symmetry of $\mathbf{A}$ follows directly from the definition. To prove positive definiteness, let $\mathbf{x} \in \mathbb{R}^n$ be a non-zero vector and consider the quadratic form $\mathbf{x}^T \mathbf{A}\, \mathbf{x}$:

$$\mathbf{x}^T \mathbf{A}\, \mathbf{x} = \mathbf{x}^T \mathbf{B}^{\mathbf{T}} \mathbf{V}\, \mathbf{B}\, \mathbf{x} = \mathbf{y}^{\mathbf{T}} \mathbf{V}\, \mathbf{y}\,.$$

Clearly $\mathbf{y}^T \mathbf{V}\, \mathbf{y} > 0$ for all $\mathbf{y} \neq 0$ and $\mathbf{y}^T \mathbf{V}\, \mathbf{y} = 0 \iff \mathbf{B}\, \mathbf{x} = \mathbf{0}$. But $\mathbf{B}$ has full column rank and therefore the only solution to $\mathbf{B}\, \mathbf{x} = \mathbf{0}$ is the trivial solution $\mathbf{x} = 0$. It follows then that $\mathbf{x}^T \mathbf{B}^{\mathbf{T}} \mathbf{V}\, \mathbf{B}\, \mathbf{x} > \mathbf{0}$ and therefore $\mathbf{A}$ is positive definite. □

A.1 Proof of Proposition 1

Let $\mathbb{Z} = \mathbb{R}^{d_z}$, then the solution to the nominal MVO Program (11) is given by:

$$\mathbf{z}^*(\mathbf{x}^{(i)}, \boldsymbol{\theta}) = \frac{1}{\delta}\, \hat{\mathbf{V}}^{(i)^{-1}}\, \hat{\mathbf{y}}^{(i)} = \frac{1}{\delta}\, \hat{\mathbf{V}}^{(i)^{-1}}\, \mathbf{P}\, \mathrm{diag}(\mathbf{x}^{(i)})\, \boldsymbol{\theta}\,. \tag{50}$$

Direct substitution of (50) into Equation (18) yields the following quadratic objective:

$$L(\boldsymbol{\theta}) = \frac{1}{2}\, \boldsymbol{\theta}^T\, \mathbf{H}(\mathbf{x})\, \boldsymbol{\theta} - \boldsymbol{\theta}^T\, \mathbf{d}(\mathbf{x}, \mathbf{y}) \tag{51}$$

where

$$\mathbf{d}(\mathbf{x}, \mathbf{y}) = \frac{1}{m\delta} \sum_{i=1}^{m} \Big(\mathrm{diag}(\mathbf{x}^{(i)})\, \mathbf{P}^T\, \hat{\mathbf{V}}^{(i)^{-1}}\, \mathbf{y}^{(i)} \Big) \tag{52}$$

and

$$\mathbf{H}(\mathbf{x}) = \frac{1}{m\delta} \sum_{i=1}^{m} \Big(\mathrm{diag}(\mathbf{x}^{(i)})\, \mathbf{P}^T\, \hat{\mathbf{V}}^{(i)^{-1}}\, \mathbf{V}^{(i)}\, \hat{\mathbf{V}}^{(i)^{-1}}\, \mathbf{P}\, \mathrm{diag}(\mathbf{x}^{(i)}) \Big). \tag{53}$$

Applying Proposition 9 it follows then that if there exists an $\mathbf{x}^{(i)}$ such that ${\mathbf{x}_{\mathrm{j}}}^{(i)} \neq 0 \quad \forall j \in 1, ..., d_x$ then $\mathbf{H}(\mathbf{x}, \mathbf{y}) \succ 0$ and therefore (54) is a convex quadratic function.

$$\underset{\boldsymbol{\theta} \in \Theta}{\text{minimize}} \quad \frac{1}{2}\, \boldsymbol{\theta}^T\, \mathbf{H}(\mathbf{x})\, \boldsymbol{\theta} - \boldsymbol{\theta}^T\, \mathbf{d}(\mathbf{x}, \mathbf{y}). \tag{54}$$

In the absence of constraints on $\boldsymbol{\theta}$, then the first-order conditions are necessary and sufficient for optimality, with optimal IPO coefficients given by:

$$\boldsymbol{\theta}^* = \mathbf{H}(\mathbf{x})^{-1}\, \mathbf{d}(\mathbf{x}, \mathbf{y}) \tag{55}$$

A.2 Proof of Proposition 2

Let $\boldsymbol{\theta}^*$ and $\mathbf{d}_{\mathbf{u}}(\mathbf{x})$ be as defined by Equation (29) and Equation (31), respectively. It follows then that:

$$\begin{aligned}
\mathbb{E}[\boldsymbol{\theta}^*] &= \mathbb{E}\big[\, \mathbf{H}(\mathbf{x})^{-1}\, \mathbf{d}(\mathbf{x}, \mathbf{y}) \big] \\
&= \mathbf{H}(\mathbf{x})^{-1} \frac{1}{m\delta} \sum_{i=1}^{m} \Big(\mathrm{diag}(\mathbf{x}^{(i)})\, \mathbf{P}^T\, \hat{\mathbf{V}}^{(i)^{-1}}\, \mathbb{E}\big[\, \mathbf{y}^{(i)} \big] \Big) \\
&= \mathbf{H}(\mathbf{x})^{-1} \frac{1}{m\delta} \sum_{i=1}^{m} \Big(\mathrm{diag}(\mathbf{x}^{(i)})\, \mathbf{P}^T\, \hat{\mathbf{V}}^{(i)^{-1}}\, \mathbf{P}\, \mathrm{diag}(\mathbf{x}^{(i)})\, \boldsymbol{\theta} \Big) \\
&= \mathbf{H}(\mathbf{x})^{-1}\, \mathbf{d}_{\mathbf{u}}(\mathbf{x})\, \boldsymbol{\theta}
\end{aligned} \tag{56}$$

Corollary 2 follows directly from Equation (56). Observe that when $\hat{\mathbf{V}}^{(i)} = \mathbf{V}^{(i)}\, \forall i \in \{1, ..., m\}$, then

$$\mathbf{H}(\mathbf{x}) = \frac{1}{m\delta} \sum_{i=1}^{m} \Big(\operatorname{diag}(\mathbf{x}^{(i)})\, \mathbf{P}^T\, \hat{\mathbf{V}}^{(i)^{-1}}\, \mathbf{P} \operatorname{diag}(\mathbf{x}^{(i)}) \Big).$$

It follows then that:

$$\begin{aligned} \mathbb{E}[\boldsymbol{\theta}^*] &= \mathbb{E}\big[\, \mathbf{H}(\mathbf{x})^{-1}\, \mathbf{d}(\mathbf{x}, \mathbf{y}) \big] \\ &= \mathbf{H}(\mathbf{x})^{-1} \frac{1}{m\delta} \sum_{i=1}^{m} \Big(\operatorname{diag}(\mathbf{x}^{(i)})\, \mathbf{P}^T\, \hat{\mathbf{V}}^{(i)^{-1}}\, \mathbf{P} \operatorname{diag}(\mathbf{x}^{(i)})\, \boldsymbol{\theta} \Big) \\ &= \mathbf{H}(\mathbf{x})^{-1}\, \mathbf{H}(\mathbf{x})\, \boldsymbol{\theta} \\ &= \boldsymbol{\theta}\,. \end{aligned} \tag{57}$$

A.3 Proof of Proposition 3

Let $\{\mathbf{y}^{(i)}\}_{i=1}^{m}$ be independent random variables with $\mathbf{y}^{(i)} \sim \mathcal{N}(\mathbf{P} \operatorname{diag}(\mathbf{x}^{(i)})\, \boldsymbol{\theta}, \boldsymbol{\Sigma})$. Let $\hat{\boldsymbol{\Sigma}}$ and $\mathbf{M}$ be as defined by Equation (32) and Equation (33), respectively. It follows then that:

$$\begin{aligned} \operatorname{Var}(\boldsymbol{\theta}^*) &= \operatorname{Var}\big(\mathbf{H}(\mathbf{x})^{-1}\, \mathbf{d}(\mathbf{x}, \mathbf{y}) \big) \\ &= \mathbf{H}(\mathbf{x})^{-1} \operatorname{Var}\Big(\frac{1}{m\delta} \sum_{i=1}^{m} \operatorname{diag}(\mathbf{x}^{(i)})\, \mathbf{P}^T\, \hat{\mathbf{V}}^{(i)^{-1}}\, \mathbf{y}^{(i)} \Big)\, \mathbf{H}(\mathbf{x})^{-1} \\ &= \mathbf{H}(\mathbf{x})^{-1} \frac{1}{m^2\delta^2} \sum_{i=1}^{m} \Big(\operatorname{diag}(\mathbf{x}^{(i)})\, \mathbf{P}^T\, \hat{\mathbf{V}}^{(i)^{-1}} \operatorname{Var}(\mathbf{y}^{(i)})\, \hat{\mathbf{V}}^{(i)^{-1}}\, \mathbf{P} \operatorname{diag}(\mathbf{x}^{(i)}) \Big)\, \mathbf{H}(\mathbf{x})^{-1} \\ &= \mathbf{H}(\mathbf{x})^{-1} \frac{1}{m^2\delta^2} \sum_{i=1}^{m} \Big(\operatorname{diag}(\mathbf{x}^{(i)})\, \mathbf{P}^T\, \hat{\mathbf{V}}^{(i)^{-1}}\, \hat{\boldsymbol{\Sigma}}\, \hat{\mathbf{V}}^{(i)^{-1}}\, \mathbf{P} \operatorname{diag}(\mathbf{x}^{(i)}) \Big)\, \mathbf{H}(\mathbf{x})^{-1} \\ &= \mathbf{H}(\mathbf{x})^{-1}\, \mathbf{M}\, \mathbf{H}(\mathbf{x})^{-1}. \end{aligned} \tag{58}$$

A.4 Proof of Proposition 4

Let $\mathbf{z}^*(\mathbf{y}^{(i)})$ and $\mathbf{z}^*(\mathbf{x}^{(i)}, \boldsymbol{\theta})$ be as defined in Equation (11) and Equation (25), respectively. Recall, the objective function of the minimum tracking-error representation of the IPO program is:

$$L_{\text{te}}(\boldsymbol{\theta}) = \frac{1}{2m} \sum_{i=1}^{m} \|\mathbf{z}^*(\mathbf{x}^{(i)}, \boldsymbol{\theta}) - \mathbf{z}^*(\mathbf{y}^{(i)})\|^2_{\mathbf{V}^{(i)}} \tag{59}$$

The first-order necessary conditions for optimality of Program (11) state:

$$\mathbf{V}^{(i)}\, \mathbf{z}^*(\mathbf{y}^{(i)}) = \mathbf{y}^{(i)} \tag{60}$$

Expanding Equation (61) and substituting in Equation (60) completes the proof:

$$
\begin{aligned}
L_{\mathrm{te}}(\boldsymbol{\theta}) &= \frac{1}{2m}\sum_{i=1}^{m}\left(\mathbf{z}^*(\mathbf{x}^{(i)},\boldsymbol{\theta})-\mathbf{z}^*(\mathbf{y}^{(i)})\right)^T\mathbf{V}^{(i)}\left(\mathbf{z}^*(\mathbf{x}^{(i)},\boldsymbol{\theta})-\mathbf{z}^*(\mathbf{y}^{(i)})\right)\\
&= \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}\,\mathbf{z}^*(\mathbf{x}^{(i)},\boldsymbol{\theta})^T\,\mathbf{V}^{(i)}\,\mathbf{z}^*(\mathbf{x}^{(i)},\boldsymbol{\theta})-\mathbf{z}^*(\mathbf{x}^{(i)},\boldsymbol{\theta})^T\,\mathbf{V}^{(i)}\,\mathbf{z}^*(\mathbf{y}^{(i)})+\frac{1}{2}\,\mathbf{z}^*(\mathbf{y}^{(i)})^T\,\mathbf{V}^{(i)}\,\mathbf{z}^*(\mathbf{y}^{(i)})\\
&= \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}\,\mathbf{z}^*(\mathbf{x}^{(i)},\boldsymbol{\theta})^T\,\mathbf{V}^{(i)}\,\mathbf{z}^*(\mathbf{x}^{(i)},\boldsymbol{\theta})-\mathbf{z}^*(\mathbf{x}^{(i)},\boldsymbol{\theta})^T\,\mathbf{y}^{(i)}+\mathbf{z}^*(\mathbf{y}^{(i)})^T\,\mathbf{V}^{(i)}\,\mathbf{z}^*(\mathbf{y}^{(i)}).
\end{aligned}
\tag{61}
$$

Note that the proof of Proposition 8 follows a similar argument for the case of equality constrained MVO portfolios.

A.5 Proof of Proposition 5

In the presence of equality constraints then the solution to the nominal MVO Program is given by:

$$
\mathbf{z}^*(\mathbf{x}^{(i)},\boldsymbol{\theta}) = \frac{1}{\delta}\,\mathbf{F}(\mathbf{F}^T\,\hat{\mathbf{V}}^{(i)}\,\mathbf{F})^{-1}\,\mathbf{F}^T\,\mathbf{P}\,\mathrm{diag}(\mathbf{x}^{(i)})\,\boldsymbol{\theta}+(\mathbf{I}-\mathbf{F}(\mathbf{F}^T\,\hat{\mathbf{V}}^{(i)}\,\mathbf{F})^{-1}\,\mathbf{F}^T\,\hat{\mathbf{V}}^{(i)})\,\mathbf{z}_0, \tag{62}
$$

where $\mathbf{z}_0$ be a particular element of $\mathbb{Z} = \{\mathbf{A}\,\mathbf{z} = \mathbf{b}\}$ and $\mathbf{F}$ is a basis for the nullspace of $\mathbf{A}$.

Direct substitution of (62) into Equation (18) yields the following quadratic objective:

$$
L(\boldsymbol{\theta}) = \frac{\delta}{2}\sum_{i=1}^{m}\mathrm{L_1}^{(i)}(\boldsymbol{\theta})-\sum_{i=1}^{m}\mathrm{L_2}^{(i)}(\boldsymbol{\theta}), \tag{63}
$$

where

$$
\begin{aligned}
\mathrm{L_1}^{(i)}(\boldsymbol{\theta}) &= \mathbf{z}^*(\mathbf{x}^{(i)},\boldsymbol{\theta})^T\,\mathbf{V}^{(i)}\,\mathbf{z}^*(\mathbf{x}^{(i)},\boldsymbol{\theta})\\
&= \frac{1}{\delta^2}\,\boldsymbol{\theta}^T\,\mathrm{diag}(\mathbf{x}^{(i)})\,\mathbf{P}^T\,\mathbf{F}(\mathbf{F}^T\,\hat{\mathbf{V}}^{(i)}\,\mathbf{F})^{-1}\,\mathbf{F}^T\,\mathbf{V}^{(i)}\,\mathbf{F}(\mathbf{F}^T\,\hat{\mathbf{V}}^{(i)}\,\mathbf{F})^{-1}\,\mathbf{F}^T\,\mathbf{P}\,\mathrm{diag}(\mathbf{x}^{(i)})\,\boldsymbol{\theta}\\
&\quad+\frac{2}{\delta}\,\boldsymbol{\theta}^T\,\mathrm{diag}(\mathbf{x}^{(i)})\,\mathbf{P}^T\,\mathbf{F}(\mathbf{F}^T\,\hat{\mathbf{V}}^{(i)}\,\mathbf{F})^{-1}\,\mathbf{F}^T\,\mathbf{V}^{(i)}(\mathbf{I}-\mathbf{F}(\mathbf{F}^T\,\hat{\mathbf{V}}^{(i)}\,\mathbf{F})^{-1}\,\mathbf{F}^T\,\hat{\mathbf{V}}^{(i)})\,\mathbf{z}_0\\
&\quad+\mathbf{z}_0^T(\mathbf{I}-\hat{\mathbf{V}}^{(i)}\,\mathbf{F}(\mathbf{F}^T\,\hat{\mathbf{V}}^{(i)}\,\mathbf{F})^{-1}\,\mathbf{F}^T)\,\mathbf{V}^{(i)}(\mathbf{I}-\mathbf{F}(\mathbf{F}^T\,\hat{\mathbf{V}}^{(i)}\,\mathbf{F})^{-1}\,\mathbf{F}^T\,\hat{\mathbf{V}}^{(i)})\,\mathbf{z}_0
\end{aligned}
\tag{64}
$$

and

$$
\begin{aligned}
\mathrm{L_2}^{(i)}(\boldsymbol{\theta}) &= \mathbf{z}^*(\mathbf{x}^{(i)},\boldsymbol{\theta})^T\,\mathbf{y}^{(i)}\\
&= \frac{1}{\delta}\,\boldsymbol{\theta}^T\,\mathrm{diag}(\mathbf{x}^{(i)})\,\mathbf{P}^T\,\mathbf{F}(\mathbf{F}^T\,\hat{\mathbf{V}}^{(i)}\,\mathbf{F})^{-1}\,\mathbf{F}^T\,\mathbf{y}^{(i)}+\mathbf{z}_0^T(\mathbf{I}-\mathbf{F}(\mathbf{F}^T\,\hat{\mathbf{V}}^{(i)}\,\mathbf{F})^{-1}\,\mathbf{F}^T\,\hat{\mathbf{V}}^{(i)})\,\mathbf{y}^{(i)}
\end{aligned}
\tag{65}
$$

Simplifying Equation (63) and removing constant terms yields the following quadratic objective:

$$
L(\boldsymbol{\theta}) = \frac{1}{2}\,\boldsymbol{\theta}^T\,\mathbf{H}_{\mathrm{eq}}(\mathbf{x})\,\boldsymbol{\theta}-\boldsymbol{\theta}^T\,\mathbf{d}_{\mathrm{eq}}(\mathbf{x},\mathbf{y}) \tag{66}
$$

where

$$\mathbf{d}_{\text{eq}}(\mathbf{x},\mathbf{y}) = \frac{1}{m\delta}\sum_{i=1}^{m}\Big(\operatorname{diag}(\mathbf{x}^{(i)})\,\mathbf{P}^T\,\mathbf{F}(\mathbf{F}^T\,\hat{\mathbf{V}}^{(i)}\,\mathbf{F})^{-1}\,\mathbf{F}^T(\mathbf{y}^{(i)} - \mathbf{V}^{(i)}(\mathbf{I} - \mathbf{F}(\mathbf{F}^T\,\hat{\mathbf{V}}^{(i)}\,\mathbf{F})^{-1}\,\mathbf{F}^T\,\hat{\mathbf{V}}^{(i)})\,\mathbf{z}_0)\Big) \tag{67}$$

and

$$\mathbf{H}_{\text{eq}}(\mathbf{x}) = \frac{1}{m\delta}\sum_{i=1}^{m}\Big(\operatorname{diag}(\mathbf{x}^{(i)})\,\mathbf{P}^T\,\mathbf{F}(\mathbf{F}^T\,\hat{\mathbf{V}}^{(i)}\,\mathbf{F})^{-1}\,\mathbf{F}^T\,\mathbf{V}^{(i)}\,\mathbf{F}(\mathbf{F}^T\,\hat{\mathbf{V}}^{(i)}\,\mathbf{F})^{-1}\,\mathbf{F}^T\,\mathbf{P}\operatorname{diag}(\mathbf{x}^{(i)})\Big). \tag{68}$$

Again, applying Proposition 9 it follows then that if there exists an $\mathbf{x}^{(i)}$ such that ${\mathbf{x}_{\text{j}}}^{(i)} \neq 0 \quad \forall j \in 1,...,d_x$ then $\mathbf{H}_{\text{eq}}(\mathbf{x}) \succ 0$ and therefore (69) is a convex quadratic program:

$$\underset{\boldsymbol{\theta}\in\Theta}{\text{minimize}} \quad \frac{1}{2}\,\boldsymbol{\theta}^T\,\mathbf{H}_{\text{eq}}(\mathbf{x})\,\boldsymbol{\theta} - \boldsymbol{\theta}^T\,\mathbf{d}_{\text{eq}}(\mathbf{x},\mathbf{y}). \tag{69}$$

In the absence of constraints on $\boldsymbol{\theta}$, then the first-order conditions are necessary and sufficient for optimality, with optimal IPO coefficients given by:

$$\boldsymbol{\theta}^*_{\text{eq}} = \mathbf{H}_{\text{eq}}(\mathbf{x})^{-1}\,\mathbf{d}_{\text{eq}}(\mathbf{x},\mathbf{y}). \tag{70}$$

A.6 Proof of Proposition 6

Let $\boldsymbol{\theta}^*_{\text{eq}}$ and $\mathbf{d}_{\mathbf{e}}(\mathbf{x})$ be as defined by Equation (43) and Equation (44), respectively. It follows then that:

$$\begin{aligned}
\mathbb{E}[\boldsymbol{\theta}^*_{\text{eq}}] &= \mathbb{E}\big[\,\mathbf{H}_{\text{eq}}(\mathbf{x})^{-1}\,\mathbf{d}_{\text{eq}}(\mathbf{x},\mathbf{y})\big] \\
&= \mathbf{H}_{\text{eq}}(\mathbf{x})^{-1}\frac{1}{m\delta}\sum_{i=1}^{m}\Big(\operatorname{diag}(\mathbf{x}^{(i)})\,\mathbf{P}^T\,\mathbf{F}(\mathbf{F}^T\,\hat{\mathbf{V}}^{(i)}\,\mathbf{F})^{-1}\,\mathbf{F}^T\,\mathbb{E}\big[\,\mathbf{y}^{(i)}\,\big]\Big) \\
&= \mathbf{H}_{\text{eq}}(\mathbf{x})^{-1}\frac{1}{m\delta}\sum_{i=1}^{m}\Big(\operatorname{diag}(\mathbf{x}^{(i)})\,\mathbf{P}^T\,\mathbf{F}(\mathbf{F}^T\,\hat{\mathbf{V}}^{(i)}\,\mathbf{F})^{-1}\,\mathbf{F}^T\,\mathbf{P}\operatorname{diag}(\mathbf{x}^{(i)})\Big) \\
&= \mathbf{H}_{\text{eq}}(\mathbf{x})^{-1}\,\mathbf{d}_{\mathbf{e}}(\mathbf{x})\,\boldsymbol{\theta}
\end{aligned} \tag{71}$$

Corollary 3 follows directly from Equation (71). Observe that when $\hat{\mathbf{V}}^{(i)} = \mathbf{V}^{(i)}\,\forall i \in \{1,...,m\}$, then

$$\mathbf{H}_{\text{eq}}(\mathbf{x}) = \frac{1}{m\delta}\sum_{i=1}^{m}\Big(\operatorname{diag}(\mathbf{x}^{(i)})\,\mathbf{P}^T\,\mathbf{F}(\mathbf{F}^T\,\hat{\mathbf{V}}^{(i)}\,\mathbf{F})^{-1}\,\mathbf{F}^T\,\mathbf{P}\operatorname{diag}(\mathbf{x}^{(i)})\Big).$$

It follows then that:

$$\begin{aligned}
\mathbb{E}[\boldsymbol{\theta}^*_{\text{eq}}] &= \mathbb{E}\big[\,\mathbf{H}_{\text{eq}}(\mathbf{x})^{-1}\,\mathbf{d}_{\text{eq}}(\mathbf{x},\mathbf{y})\big] \\
&= \mathbf{H}_{\text{eq}}(\mathbf{x})^{-1}\frac{1}{m\delta}\sum_{i=1}^{m}\Big(\operatorname{diag}(\mathbf{x}^{(i)})\,\mathbf{P}^T\,\mathbf{F}(\mathbf{F}^T\,\hat{\mathbf{V}}^{(i)}\,\mathbf{F})^{-1}\,\mathbf{F}^T\,\mathbf{P}\operatorname{diag}(\mathbf{x}^{(i)})\Big) \\
&= \mathbf{H}_{\text{eq}}(\mathbf{x})^{-1}\,\mathbf{H}_{\text{eq}}(\mathbf{x})\,\boldsymbol{\theta} \\
&= \boldsymbol{\theta}
\end{aligned} \tag{72}$$

A.7 Proof of Proposition 7

Let $\{\mathbf{y}^{(i)}\}_{i=1}^m$ be independent random variables with $\mathbf{y}^{(i)} \sim \mathcal{N}(\mathbf{P}\,\mathrm{diag}(\mathbf{x}^{(i)})\,\boldsymbol{\theta}, \boldsymbol{\Sigma})$. Let $\hat{\boldsymbol{\Sigma}}$ and $\mathbf{M}_{\mathrm{eq}}$ be as defined by Equation (32) and Equation (45), respectively. It follows then that:

$$
\begin{aligned}
\mathrm{Var}(\boldsymbol{\theta}^*_{\mathrm{eq}}) &= \mathrm{Var}\left(\mathbf{H}_{\mathrm{eq}}(\mathbf{x})^{-1}\,\mathbf{d}\,\mathrm{eq}(\mathbf{x},\mathbf{y})\right) \\
&= \mathbf{H}_{\mathrm{eq}}(\mathbf{x})^{-1}\,\mathrm{Var}\Big(\mathrm{diag}(\mathbf{x}^{(i)})\,\mathbf{P}^T\,\mathbf{F}(\mathbf{F}^T\,\hat{\mathbf{V}}^{(i)}\,\mathbf{F})^{-1}\,\mathbf{F}^T(\mathbf{y}^{(i)} - \mathbf{V}^{(i)}(\mathbf{I} - \mathbf{F}(\mathbf{F}^T\,\hat{\mathbf{V}}^{(i)}\,\mathbf{F})^{-1}\,\mathbf{F}^T\,\hat{\mathbf{V}}^{(i)})\,\mathbf{z}_0)\Big)\,\mathbf{H}_{\mathrm{eq}}(\mathbf{x})^{-1} \\
&= \mathbf{H}_{\mathrm{eq}}(\mathbf{x})^{-1}\frac{1}{m^2\delta^2}\sum_{i=1}^m\Big(\mathrm{diag}(\mathbf{x}^{(i)})\,\mathbf{P}^T\,\mathbf{F}(\mathbf{F}^T\,\hat{\mathbf{V}}^{(i)}\,\mathbf{F})^{-1}\,\mathbf{F}^T\,\mathrm{Var}(\mathbf{y}^{(i)})\,\mathbf{F}(\mathbf{F}^T\,\hat{\mathbf{V}}^{(i)}\,\mathbf{F})^{-1}\,\mathbf{F}^T\,\mathbf{P}\,\mathrm{diag}(\mathbf{x}^{(i)})\Big)\,\mathbf{H}_{\mathrm{eq}}(\mathbf{x})^{-1} \\
&= \mathbf{H}_{\mathrm{eq}}(\mathbf{x})^{-1}\frac{1}{m^2\delta^2}\sum_{i=1}^m\Big(\mathrm{diag}(\mathbf{x}^{(i)})\,\mathbf{P}^T\,\mathbf{F}(\mathbf{F}^T\,\hat{\mathbf{V}}^{(i)}\,\mathbf{F})^{-1}\,\mathbf{F}^T\,\hat{\boldsymbol{\Sigma}}\,\mathbf{F}(\mathbf{F}^T\,\hat{\mathbf{V}}^{(i)}\,\mathbf{F})^{-1}\,\mathbf{F}^T\,\mathbf{P}\,\mathrm{diag}(\mathbf{x}^{(i)})\Big)\,\mathbf{H}_{\mathrm{eq}}(\mathbf{x})^{-1} \\
&= \mathbf{H}_{\mathrm{eq}}(\mathbf{x})^{-1}\,\mathbf{M}_{\mathrm{eq}}\,\mathbf{H}_{\mathrm{eq}}(\mathbf{x})^{-1}.
\end{aligned}
\tag{73}
$$

B Additional Experiments

Economic performance metrics and average out-of-sample MVO costs are provided in Tables 10 and 11 for the equality constrained MVO portfolios with univariate and multivariate prediction models, respectively. Equity growth charts for the time period of 2000-01-01 to 2020-12-31 are provided in Figures 23 and 25. As in the unconstrained case, we observe that the IPO model provides higher absolute and risk-adjusted performance, and in general produces more conservative risk metrics. Figures 24 and 26 demonstrate that the IPO model produces consistently lower out-of-sample MVO costs, with dominance ratios of 93% and 99%, respectively, and generally lower Sharpe ratio costs with dominance ratios of 67% and 66%, respectively. The regression coefficients are identical to those provided in Figures 19 and 22, and we refer to Section 5 for relevant discussion.

B.1 Experiment 3: $\mathbb{Z} = \{Az = b\}$, $f(\mathbf{x}, \boldsymbol{\theta})$ univariate

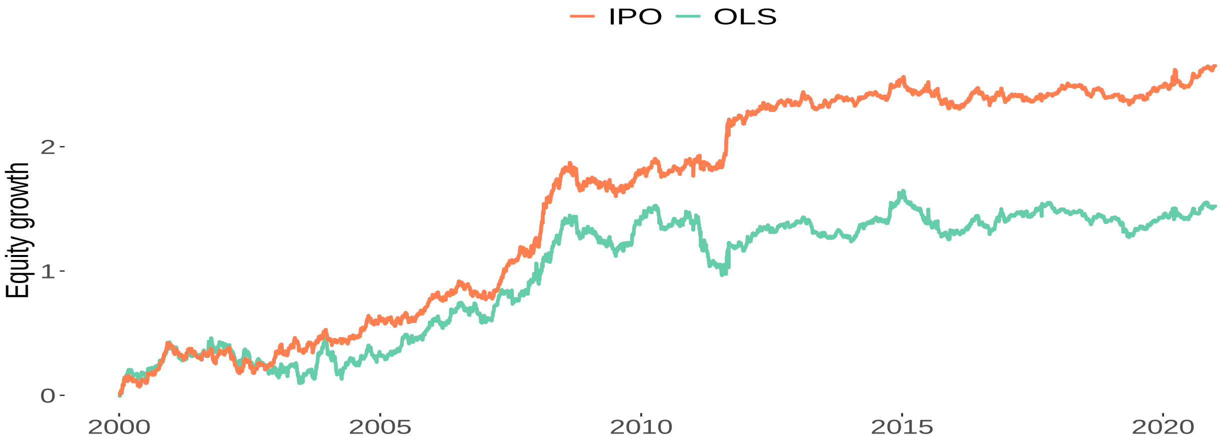
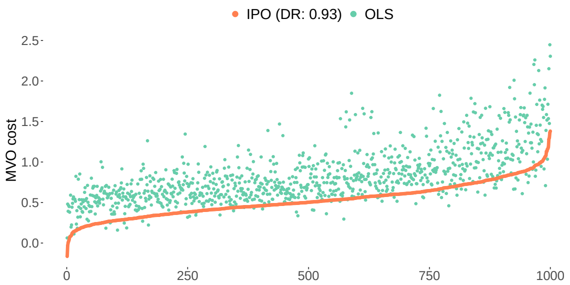


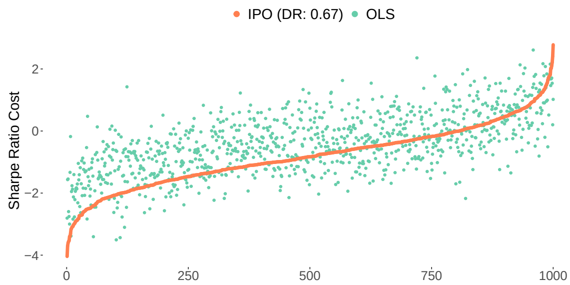
Figure 23: Out-of-sample log-equity growth for the equality constrained nominal mean-variance program and multivariate IPO and OLS prediction model.

	Annual Return	Sharpe Ratio	Volatility	Avg Drawdown	Value at Risk	MVO Cost
IPO	0.1238	0.7665	0.1616	-0.0290	-0.0142	0.5288
OLS	0.0713	0.3803	0.1876	-0.0471	-0.0170	0.8082

Table 10: Out-of-sample MVO costs and economic performance metrics for equality constrained mean-variance portfolios with univariate IPO and OLS prediction models.



(a) Out-of-sample MVO cost.



(b) Out-of-sample Sharpe ratio cost.

Figure 24: Realized out-of-sample MVO and Sharpe ratio costs for the equality constrained nominal mean-variance program and univariate IPO and OLS prediction models.

B.2 Experiment 4: $\mathbb{Z} = \{Az = b\}$, $f(\mathbf{x}, \boldsymbol{\theta})$ multivariate

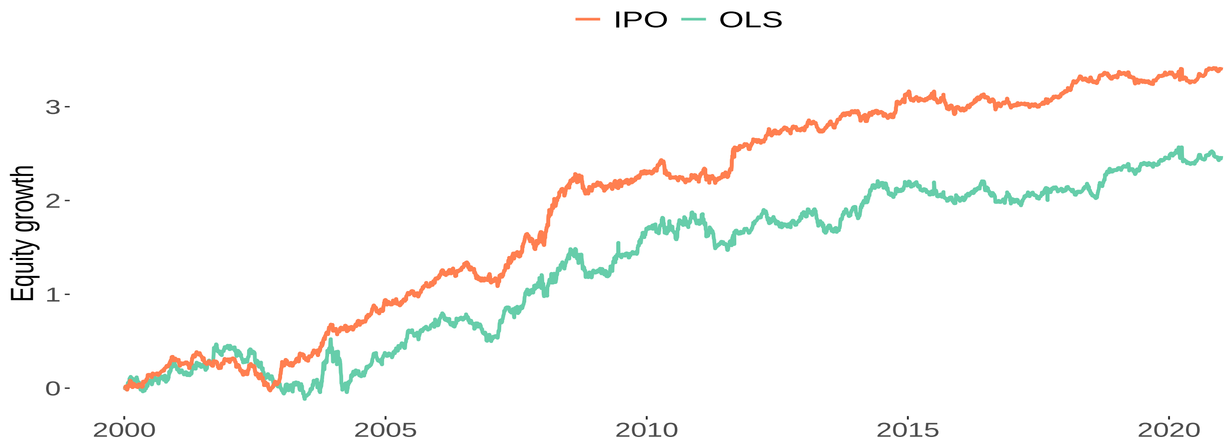
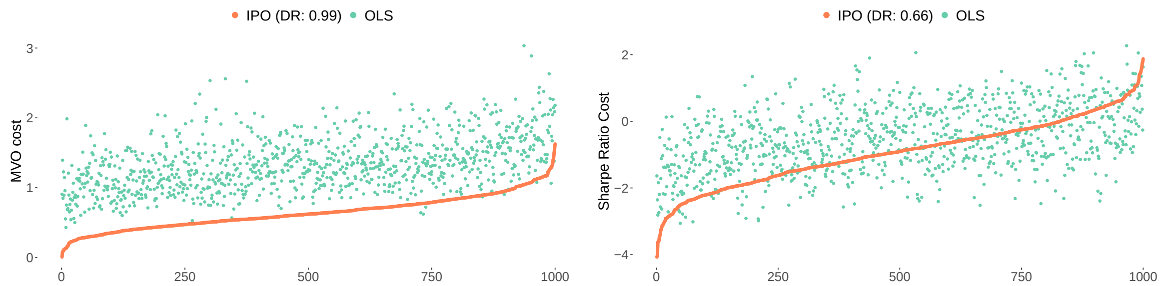


Figure 25: Out-of-sample log-equity growth for the equality constrained nominal mean-variance program and multivariate IPO and OLS prediction model.

	Annual Return	Sharpe Ratio	Volatility	Avg Drawdown	Value at Risk	MVO Cost
IPO	0.1590	0.8851	0.1797	-0.0339	-0.0163	0.6482
OLS	0.1151	0.4784	0.2406	-0.0497	-0.0215	1.3315

Table 11: Out-of-sample MVO costs and economic performance metrics for equality constrained mean-variance portfolios with multivariate IPO and OLS prediction models.



(a) Out-of-sample MVO cost.

(b) Out-of-sample Sharpe ratio cost.

Figure 26: Realized out-of-sample MVO and Sharpe ratio costs for the equality constrained nominal mean-variance program and multivariate IPO and OLS prediction models.