From Estimation to Optimization via Shrinkage

Danial Davarnia[a,*], Gérard Cornuéjols[a]

[a] *Tepper School of Business, Carnegie Mellon University, Pittsburgh, PA, USA*

Abstract

We study a class of quadratic stochastic programs where the distribution of random variables has unknown parameters. A traditional approach is to estimate the parameters using a maximum likelihood estimator (MLE) and to use this as input in the optimization problem. For the unconstrained case, we show that an estimator that shrinks the MLE towards an arbitrary vector yields a uniformly better risk than the MLE. In contrast, when there are constraints, we show that the MLE is admissible.

Keywords: Stochastic optimization, Parameter estimation, Maximum likelihood estimator, Admissible estimator, Shrinkage estimator

1. Introduction

In practice, optimization problems often involve uncertain elements arising from a random process. See Birge and Louveaux [3] for an introduction to stochastic programming. Samples from the underlying random process are used to estimate unknown parameters of the distribution of the uncertain elements. We study a set up where the estimation process is performed first, and its output estimator is used as an input for the optimization problem. It is natural to use the maximum likelihood estimator (MLE) of the parameters. But in some cases one may obtain better solutions to the optimization problem by replacing the MLE by a *shrinkage* estimator. For example in portfolio optimization, an investor may want to construct a portfolio of risky assets that maximizes expected return against risk (Markowitz [13]). When historical data on the asset returns are used to estimate the expected returns, Jorion [8] recommends to *shrink* the vector of sample averages towards a *grand average*, and to use this shrunk estimator in the Markowitz optimization problem to obtain better portfolios. We address the question of where this shrinkage idea fits in the optimization literature, focusing on the impact of constraints.

*Corresponding author

Email addresses: ddavarni@andrew.cmu.edu (Danial Davarnia), gc0v@andrew.cmu.edu (Gérard Cornuéjols)

2. Problem Description

Consider the following parametric stochastic optimization problem

$$\max_{\boldsymbol{y}\in\mathcal{Y}} \mathbb{E}_{\boldsymbol{x}|\boldsymbol{\theta}}[f(\boldsymbol{x},\boldsymbol{y})]. \tag{1}$$

In (1), $\boldsymbol{x}$ represents a vector of random variables in $\mathbb{R}^n$ that has a known probability distribution with joint density $\mathcal{G}(\boldsymbol{x}|\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ represents a vector of unknown parameters of the distribution. Vector $\boldsymbol{y}$ represents decision variables in $\mathbb{R}^m$ that belong to a closed set $\mathcal{Y} \subseteq \mathbb{R}^m$. The expectation $\mathbb{E}_{\boldsymbol{x}|\boldsymbol{\theta}}[.]$ is taken with respect to the distribution of the random variables $\boldsymbol{x}$ given the vector $\boldsymbol{\theta}$ of parameters. Writing $\mathbb{E}_{\boldsymbol{x}|\boldsymbol{\theta}}[f(\boldsymbol{x},\boldsymbol{y})] = \mathcal{F}(\boldsymbol{\theta},\boldsymbol{y})$, we refer to $\mathcal{F}(\boldsymbol{\theta},\boldsymbol{y})$ as a *parametric* objective function. Since $\mathcal{F}(\boldsymbol{\theta},\boldsymbol{y})$ is a function of $\boldsymbol{\theta}$ and $\boldsymbol{y}$, its optimal solution $\boldsymbol{y}^*(\boldsymbol{\theta})$ and its optimal value $\mathcal{F}(\boldsymbol{\theta},\boldsymbol{y}^*(\boldsymbol{\theta}))$ are both functions of $\boldsymbol{\theta}$. This setting suggests combining statistical techniques with optimization to achieve desirable end-solutions; see Lim, Shanthikumar and Shen [11] for an investigation.

A finite number T of i.i.d. observations $\{\boldsymbol{x}^t\}_{t\in[T]}$ (obtained from computer simulation, historical data, prediction, etc) is available for the random variables $\boldsymbol{x}$. Throughout this paper, we write $\{\boldsymbol{x}^t\}$ as a shorthand for the collection of observations. From a statistical point of view, the data is used to obtain an approximate solution (*estimator*) $\hat{\boldsymbol{y}}(\{\boldsymbol{x}^t\})$ for the true optimal solution (*estimand*) $\boldsymbol{y}^*(\boldsymbol{\theta})$. In the remainder, we use $\boldsymbol{y}^*$ as a shorthand for $\boldsymbol{y}^*(\boldsymbol{\theta})$, and we use $\hat{\boldsymbol{y}}$ as a shorthand for $\hat{\boldsymbol{y}}(\{\boldsymbol{x}^t\})$. Our goal in this paper is to obtain "good" estimators $\hat{\boldsymbol{y}}$ for the optimal solution $\boldsymbol{y}^*$ of problem (1).

The quality of the solution estimator relative to the optimal solution is measured by the *loss function*

$$\mathcal{L}(\boldsymbol{y}^*,\hat{\boldsymbol{y}}) = \mathcal{F}(\boldsymbol{\theta},\boldsymbol{y}^*) - \mathcal{F}(\boldsymbol{\theta},\hat{\boldsymbol{y}}). \tag{2}$$

A smaller loss indicates a better estimator. Since $\hat{\boldsymbol{y}}$ is a solution to (1), it belongs to $\mathcal{Y}$, and therefore $\mathcal{F}(\boldsymbol{\theta},\hat{\boldsymbol{y}}) \leq \mathcal{F}(\boldsymbol{\theta},\boldsymbol{y}^*)$.

The loss function defined in (2) is a random quantity since $\mathcal{F}(\boldsymbol{\theta},\hat{\boldsymbol{y}})$ is a function of the observations $\{\boldsymbol{x}^t\}$ (because of the estimator $\hat{\boldsymbol{y}}$). Therefore, to evaluate the overall performance of the estimator $\hat{\boldsymbol{y}}$, an averaging measure for the loss function is defined. This measure is referred to as the *risk*

$$\mathcal{R}(\boldsymbol{y}^*,\hat{\boldsymbol{y}}) = \mathbb{E}_{\{\boldsymbol{x}^t\}|\boldsymbol{\theta}}[\mathcal{L}(\boldsymbol{y}^*,\hat{\boldsymbol{y}})], \tag{3}$$

where the expectation is taken over all realizations of the observations with respect to the joint distribution $\mathcal{G}(\{\boldsymbol{x}^t\}|\boldsymbol{\theta})$ computed as $\prod_{t=1}^{T}\mathcal{G}(\boldsymbol{x}^t|\boldsymbol{\theta})$ as the observations are i.i.d.

It is clear that the risk $\mathcal{R}(\boldsymbol{y}^*,\hat{\boldsymbol{y}})$ is a function of the unknown parameters $\boldsymbol{\theta}$. The treatment of the risk is different depending on whether the unknown parameters of the model are assumed to be random or fixed. This key assumption on the model parameters gives rise to two major statistical frameworks: Bayesian and frequentist. In this paper, we investigate the risk function under the frequentist framework where parameters are viewed as fixed numbers that are not known to the modeler, and they have the domain $\Theta = \mathbb{R}^n$.

3. Admissibility

A popular criterion under the frequentist framework is *admissibility*, a desirable property of estimators that seeks superior *relative* risks. We focus on studying estimators with this property throughout this paper.

An estimator $\hat{\boldsymbol{y}}^1$ *strictly dominates* another estimator $\hat{\boldsymbol{y}}^2$ if $\mathcal{R}(\boldsymbol{y}^*, \hat{\boldsymbol{y}}^1) \leq \mathcal{R}(\boldsymbol{y}^*, \hat{\boldsymbol{y}}^2)$ for all values of the parameters $\boldsymbol{\theta}$, with strict inequality for some values of $\boldsymbol{\theta}$. An estimator $\hat{\boldsymbol{y}}^1$ is *inadmissible*, if there exists an estimator $\hat{\boldsymbol{y}}^2$ that strictly dominates it. Otherwise, it is *admissible*. It is a common-sense rule in decision making to avoid inadmissible estimators. Identifying admissible estimators and constructing dominating estimators for inadmissible ones are two important research directions in the theory of point estimation; see [10]. Our goal in this paper is to pursue these directions in optimization.

Let $\hat{\boldsymbol{\theta}}$ (as a shorthand for $\hat{\boldsymbol{\theta}}(\{\boldsymbol{x}^t\})$) be an estimator of $\boldsymbol{\theta}$ as a function of the observations. As the traditional and most common technique to obtain an estimator for the optimal solution of (1), we study the following scheme: Use $\hat{\boldsymbol{\theta}}$ in place of $\boldsymbol{\theta}$, and then solve $\max_{\boldsymbol{y}\in\mathcal{Y}} \mathcal{F}(\hat{\boldsymbol{\theta}}, \boldsymbol{y})$. The optimizer of this problem is a solution estimator $\hat{\boldsymbol{y}}_{\hat{\theta}}$ of $\boldsymbol{y}^*$. One of the most common and natural choices for $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimator (MLE) due to its several attractive features. For instance, under the assumption that the distribution $\mathcal{G}$ is normal, the MLE for the mean $\boldsymbol{\mu}$ is the sample mean $\bar{\boldsymbol{x}} = \frac{\sum_{t=1}^{T} \boldsymbol{x}^t}{T}$ which is unbiased, invariant, efficient and consistent. The question of interest is whether the solution estimator $\hat{\boldsymbol{y}}_{\bar{\boldsymbol{x}}}$ obtained from the MLE $\bar{\boldsymbol{x}}$ is admissible, and if it is not, how to find a solution estimator that dominates it.

Studying admissibility of a given estimator and designing dominating estimators are hard tasks even under simple distributional settings and problem structures. The most common statistical setting to study such properties is for the distribution to be normal and for the loss function to be the squared error; see [10] Sec. 5. Assume that $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, I)$ and $T = 1$. Consider the squared error loss function $\mathcal{L}(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) = ||\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}||^2$ which measures the Euclidean distance between the unknown parameter $\boldsymbol{\mu}$ and its estimator $\hat{\boldsymbol{\mu}}$. Blyth [4] showed that, under the squared error loss, the MLE is admissible when $n = 1$ and $n = 2$. Stein [14] stunned the statistical world by showing that $\bar{\boldsymbol{x}}$ is inadmissible when $n \geq 3$. In particular, James and Stein[7] proved that $\bar{\boldsymbol{x}}$ is uniformly dominated by an estimator of the form $\tilde{\boldsymbol{x}} = \rho\boldsymbol{x}^0 + (1-\rho)\bar{\boldsymbol{x}}$ where $\rho = \frac{(n-2)}{||\boldsymbol{x}^0 - \bar{\boldsymbol{x}}||^2}$ and $\boldsymbol{x}^0$ is an arbitrary *target* vector in $\mathbb{R}^n$. Baranchik [1] improved the James-Stein estimator by modifying the factor ρ to $\rho^+ = \min\{\rho, 1\}$. This estimator is referred to as the *shrinkage estimator*, since it shrinks the MLE $\bar{\boldsymbol{x}}$ towards the target vector $\boldsymbol{x}^0$.

The above statistical results are established in the space of parameters under a loss function $\mathcal{L}(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}})$ that measures the distance between the estimator $\hat{\boldsymbol{\mu}}$ and the parameter $\boldsymbol{\mu}$. For optimization problems, on the other hand, we are interested in the space of decision variables, where the loss function $\mathcal{L}(\boldsymbol{y}^*, \hat{\boldsymbol{y}})$ measures the difference in the objective value between the solution estimator $\hat{\boldsymbol{y}}$ and the optimal solution $\boldsymbol{y}^*$. The question of interest is how does a shrinkage

solution estimator $\hat{\boldsymbol{y}}_{\tilde{\boldsymbol{x}}}$ compare to the MLE solution estimator $\hat{\boldsymbol{y}}_{\bar{\boldsymbol{x}}}$? We investigate this question for two different classes of convex stochastic problems, one with a quadratic term in the objective and the other with a quadratic term in the constraint. To keep the analysis tractable, we assume that the distribution of the random variables is normal and its covariance matrix is known.

4. Convex Quadratic Objective

In this section we show that a classical shrinkage result in statistics extends to a certain family of stochastic programs.

Proposition 1. *Assume that* $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, I)$*, and that* $\mathcal{L}(\boldsymbol{y}^*, \hat{\boldsymbol{y}}_{\hat{\mu}}) = (\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})^\intercal Q_\mu(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})$ *where* $Q_\mu \succeq 0$ *for all* $\boldsymbol{\mu} \in \mathbb{R}^n$*. Then the shrinkage solution estimator* $\hat{\boldsymbol{y}}_{\tilde{x}}$ *strictly dominates the MLE solution estimator* $\hat{\boldsymbol{y}}_{\bar{x}}$ *for any* $\tilde{\boldsymbol{x}} = \rho\boldsymbol{x}^0 + (1-\rho)\bar{\boldsymbol{x}}$ *where* $\rho = \frac{c(||\bar{\boldsymbol{x}}-\boldsymbol{x}^0||^2)}{T||\bar{\boldsymbol{x}}-\boldsymbol{x}^0||^2}$*, provided (i)* $0 < c(.) < \inf_{\boldsymbol{\mu}\in\mathbb{R}^n} 2\frac{tr(Q_\mu)}{\lambda_{\max}(Q_\mu)} - 4$*, and (ii) the function* $c(.)$ *has nonnegative derivative. In the above definition,* $tr(Q_\mu)$ *and* $\lambda_{\max}(Q_\mu)$ *represent the trace and the maximum eigenvalue of* Q_μ *respectively.*

Proof. We show the result for $\boldsymbol{x}^0 = \mathbf{0}$. The argument for other choices of $\boldsymbol{x}^0$ follows through a translation of the origin. Fix $\boldsymbol{\mu} \in \mathbb{R}^n$. Our goal is to prove that $\mathcal{R}_F(\boldsymbol{y}^*, \hat{\boldsymbol{y}}_{\tilde{x}}) < \mathcal{R}_F(\boldsymbol{y}^*, \hat{\boldsymbol{y}}_{\bar{x}})$. Since both estimators are functions of $\bar{\boldsymbol{x}}$, we replace the simultaneous expectation $\mathbb{E}_{\{\boldsymbol{x}^t\}|\boldsymbol{\theta}}$ in the risk calculation (3) with $\mathbb{E}_{\bar{\boldsymbol{x}}|\mu}$, which is the expectation over the sample mean vector $\bar{\boldsymbol{x}}$ that has normal distribution $\mathcal{N}(\boldsymbol{\mu}, \frac{1}{T}I)$. We write that

$$
\begin{aligned}
&\mathcal{R}(\boldsymbol{y}^*, \hat{\boldsymbol{y}}_{\tilde{x}}) \\
&= \mathbb{E}_{\bar{\boldsymbol{x}}|\mu}\left[(\tilde{\boldsymbol{x}} - \boldsymbol{\mu})^\intercal Q_\mu(\tilde{\boldsymbol{x}} - \boldsymbol{\mu})\right] \\
&= \mathbb{E}_{\bar{\boldsymbol{x}}|\mu}\left[(\bar{\boldsymbol{x}} - \frac{c(||\bar{\boldsymbol{x}}||^2)}{T||\bar{\boldsymbol{x}}||^2}\bar{\boldsymbol{x}} - \boldsymbol{\mu})^\intercal Q_\mu(\bar{\boldsymbol{x}} - \frac{c(||\bar{\boldsymbol{x}}||^2)}{T||\bar{\boldsymbol{x}}||^2}\bar{\boldsymbol{x}} - \boldsymbol{\mu})\right] \\
&= \mathcal{R}(\boldsymbol{y}^*, \hat{\boldsymbol{y}}_{\bar{x}}) + \mathbb{E}_{\bar{\boldsymbol{x}}|\mu}\left[\frac{c^2(||\bar{\boldsymbol{x}}||^2)}{T^2||\bar{\boldsymbol{x}}||^4}\bar{\boldsymbol{x}}^\intercal Q_\mu\bar{\boldsymbol{x}}\right] - 2\mathbb{E}_{\bar{\boldsymbol{x}}|\mu}\left[\frac{c(||\bar{\boldsymbol{x}}||^2)}{T||\bar{\boldsymbol{x}}||^2}\bar{\boldsymbol{x}}^\intercal Q_\mu(\bar{\boldsymbol{x}} - \boldsymbol{\mu})\right],
\end{aligned}
$$

where the second equality follows from the definition of $\tilde{\boldsymbol{x}} = (1 - \frac{c(||\bar{\boldsymbol{x}}||^2)}{T||\bar{\boldsymbol{x}}||^2})\bar{\boldsymbol{x}}$ when $\boldsymbol{x}^0 = \mathbf{0}$, and the third equality holds since $\mathbb{E}_{\bar{\boldsymbol{x}}|\mu}\left[(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^\intercal Q_\mu(\bar{\boldsymbol{x}} - \boldsymbol{\mu})\right] = \mathcal{R}(\boldsymbol{y}^*, \hat{\boldsymbol{y}}_{\bar{x}})$. Next, we compute the last bracket in the above relation. Define $\bar{\boldsymbol{x}}_{-i}$ to be the subvector of $\bar{\boldsymbol{x}}$ without the ith coordinate, and let $c'(.)$ denote the

derivative of $c(.)$. We write that

$$
\begin{aligned}
&\mathbb{E}_{\bar{\boldsymbol{x}}|\mu}\left[\frac{c(||\bar{\boldsymbol{x}}||^2)}{T||\bar{\boldsymbol{x}}||^2}\bar{\boldsymbol{x}}^\intercal Q_\mu(\bar{\boldsymbol{x}}-\boldsymbol{\mu})\right] \\
&= \frac{1}{T}\sum_{i=1}^{n}\mathbb{E}_{\bar{\boldsymbol{x}}|\mu}\left[\frac{c(||\bar{\boldsymbol{x}}||^2)}{||\bar{\boldsymbol{x}}||^2}\sum_{j=1}^{n}\bar{x}_j q_{ji}(\bar{x}_i-\mu_i)\right] \\
&= \frac{1}{T}\sum_{i=1}^{n}\mathbb{E}_{\bar{\boldsymbol{x}}_{-i}|\mu_{-i}}\mathbb{E}_{\bar{x}_i|\mu_i}\left[\frac{c(||\bar{\boldsymbol{x}}||^2)}{||\bar{\boldsymbol{x}}||^2}\sum_{j=1}^{n}\bar{x}_j q_{ji}(\bar{x}_i-\mu_i)\right] \\
&= \frac{1}{T^2}\sum_{i=1}^{n}\mathbb{E}_{\bar{\boldsymbol{x}}_{-i}|\mu_{-i}}\mathbb{E}_{\bar{x}_i|\mu_i}\left[\frac{c(||\bar{\boldsymbol{x}}||^2)}{||\bar{\boldsymbol{x}}||^2}q_{ii}+2\sum_{j=1}^{n}\frac{\bar{x}_j q_{ji}\bar{x}_i}{||\bar{\boldsymbol{x}}||^4}\left(c'(||\bar{\boldsymbol{x}}||^2)-c(||\bar{\boldsymbol{x}}||^2)\right)\right] \\
&= \frac{1}{T^2}\sum_{i=1}^{n}\mathbb{E}_{\bar{\boldsymbol{x}}|\mu}\left[\frac{c(||\bar{\boldsymbol{x}}||^2)}{||\bar{\boldsymbol{x}}||^2}q_{ii}+2\sum_{j=1}^{n}\frac{\bar{x}_j q_{ji}\bar{x}_i}{||\bar{\boldsymbol{x}}||^4}\left(c'(||\bar{\boldsymbol{x}}||^2)-c(||\bar{\boldsymbol{x}}||^2)\right)\right] \\
&\quad = \frac{1}{T^2}\mathbb{E}_{\bar{\boldsymbol{x}}|\mu}\left[\frac{c(||\bar{\boldsymbol{x}}||^2)}{||\bar{\boldsymbol{x}}||^2}tr(Q_\mu)-2\frac{c(||\bar{\boldsymbol{x}}||^2)}{||\bar{\boldsymbol{x}}||^4}\bar{\boldsymbol{x}}^\intercal Q_\mu\bar{\boldsymbol{x}}+2\frac{c'(||\bar{\boldsymbol{x}}||^2)}{||\bar{\boldsymbol{x}}||^4}\bar{\boldsymbol{x}}^\intercal Q_\mu\bar{\boldsymbol{x}}\right].
\end{aligned}
$$

In the above relations, the second equality follows from splitting the expectation operator for variable $\bar{x}_i$. The third equality follows from Stein's identity (see Lemma 1.5.15 in [10]) based on integration by parts which states that for a normal random variable $v \sim \mathcal{N}(\tau, \sigma^2)$ we have that $\mathbb{E}_{v|\tau}[h(v)(v-\tau)] = \sigma^2\mathbb{E}_{v|\tau}[h'(v)]$ for any differentiable function $h(v)$ with integrable derivative. This identity together with the fact that the variance of $\bar{x}_i$ is equal to $\frac{1}{T}$ yields the third equality. The last two equalities follow by merging back the expectation operator and using the matrix form of the element-wise multiplications. Using this identity in the previous relation, we obtain

$$
\begin{aligned}
&\mathcal{R}(\boldsymbol{y}^*, \hat{\boldsymbol{y}}_{\tilde{x}}) \\
&= \mathcal{R}(\boldsymbol{y}^*, \hat{\boldsymbol{y}}_{\bar{x}}) + \mathbb{E}_{\bar{\boldsymbol{x}}|\mu}\left[\frac{c(||\bar{\boldsymbol{x}}||^2)}{T^2||\bar{\boldsymbol{x}}||^2}\left((c(||\bar{\boldsymbol{x}}||^2)+4)\frac{\bar{\boldsymbol{x}}^\intercal Q_\mu\bar{\boldsymbol{x}}}{||\bar{\boldsymbol{x}}||^2}-2\mathrm{tr}(Q_\mu)\right)\right] - 4\mathbb{E}_{\bar{\boldsymbol{x}}|\mu}\left[\frac{c'(||\bar{\boldsymbol{x}}||^2)}{T^2||\bar{\boldsymbol{x}}||^4}\bar{\boldsymbol{x}}^\intercal Q_\mu\bar{\boldsymbol{x}}\right] \\
&\leq \mathcal{R}(\boldsymbol{y}^*, \hat{\boldsymbol{y}}_{\bar{x}}) + \mathbb{E}_{\bar{\boldsymbol{x}}|\mu}\left[\frac{c(||\bar{\boldsymbol{x}}||^2)}{T^2||\bar{\boldsymbol{x}}||^2}\left((c(||\bar{\boldsymbol{x}}||^2)+4)\lambda_{\max}(Q_\mu)-2\mathrm{tr}(Q_\mu)\right)\right] \\
&< \mathcal{R}(\boldsymbol{y}^*, \hat{\boldsymbol{y}}_{\bar{x}}),
\end{aligned}
$$

where the first inequality holds because $\frac{\bar{\boldsymbol{x}}^\intercal Q_\mu\bar{\boldsymbol{x}}}{||\bar{\boldsymbol{x}}||^2} \leq \lambda_{\max}(Q_\mu)$, $\bar{\boldsymbol{x}}^\intercal Q_\mu\bar{\boldsymbol{x}} \geq 0$ and $c'(||\bar{\boldsymbol{x}}||^2) \geq 0$ by assumption (ii), and the second inequality follows from assumption (i). □

The result of Proposition 1 provides a generalization of Theorem 5.5.9 in [10] from three aspects: (i) matrix Q_μ can be dependent on the unknown parameters, (ii) matrix Q_μ can have zero eigenvalues, and (iii) it holds for any number of

observations T. These extensions are particularly helpful in the optimization context when the objective function is approximated by its second order Taylor expansion around $\boldsymbol{\mu}$, which makes the Hessian matrix dependent on $\boldsymbol{\mu}$, and when the objective is convex but not strictly convex. We further note that the result of Proposition 1 can be extended through a suitable orthogonal transformation to the case where $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ and $\Sigma \succ 0$.

Example 1 shows an instance where the loss structure defined in Proposition 1 holds, and hence the shrinkage estimator improves on the MLE estimator.

Example 1. Consider an instance of (1) where $f(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^\intercal \boldsymbol{y} - \frac{\tau}{2}\boldsymbol{y}^\intercal A \boldsymbol{y}$ and $\mathcal{Y} = \mathbb{R}$, where A is a positive-definite matrix in $\mathbb{R}^{n\times n}$, and τ is a positive number. We obtain that $\mathcal{F}(\boldsymbol{\mu}, \boldsymbol{y}) = \boldsymbol{\mu}^\intercal \boldsymbol{y} - \frac{\tau}{2}\boldsymbol{y}^\intercal A \boldsymbol{y}$, and therefore the parametric optimization model reduces to

$$\max_{\boldsymbol{y}\in\mathbb{R}^n} \ \boldsymbol{\mu}^\intercal \boldsymbol{y} - \frac{\tau}{2}\boldsymbol{y}^\intercal A \boldsymbol{y}. \tag{4}$$

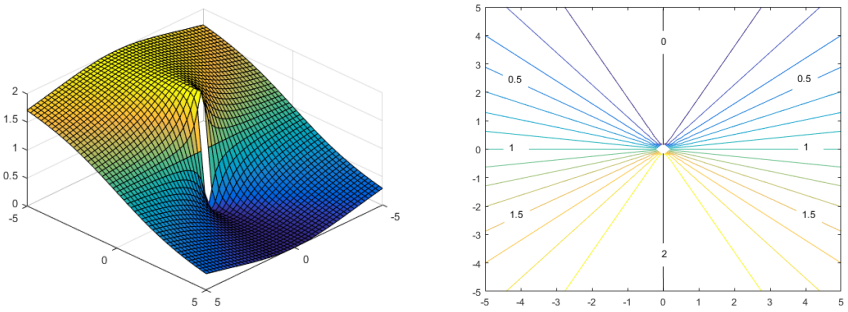
Since this is a concave maximization problem, $\boldsymbol{y}^*$ is unique and computed as $\boldsymbol{y}^* = \frac{1}{\tau}A^{-1}\boldsymbol{\mu}$. Let $\hat{\boldsymbol{\mu}}$ be an estimator of $\boldsymbol{\mu}$ as a function of the observations $\{\boldsymbol{x}^t\}$. The above relation implies that, for any such estimator $\hat{\boldsymbol{\mu}}$, one can define an optimal solution estimator $\hat{\boldsymbol{y}}_{\hat{\mu}} = \frac{1}{\tau}A^{-1}\hat{\boldsymbol{\mu}}$. Plugging the values of $\boldsymbol{y}^*$ and $\hat{\boldsymbol{y}}_{\hat{\mu}}$ in the loss function, we obtain that $\mathcal{L}(\boldsymbol{y}^*, \hat{\boldsymbol{y}}_{\hat{\mu}}) = \mathcal{F}(\boldsymbol{\mu}, \boldsymbol{y}^*) - \mathcal{F}(\boldsymbol{\mu}, \hat{\boldsymbol{y}}_{\hat{\mu}}) = \frac{1}{2\tau}(\boldsymbol{\mu}^\intercal A^{-1}\boldsymbol{\mu} - 2\boldsymbol{\mu}^\intercal A^{-1}\hat{\boldsymbol{\mu}} + \hat{\boldsymbol{\mu}}^\intercal A^{-1}\hat{\boldsymbol{\mu}}) = \frac{1}{2\tau}(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})^\intercal A^{-1}(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})$.

A special case of the quadratic model (4) appears in Markowitz' portfolio selection problem $\max_{\boldsymbol{y}\in\mathbb{R}^n} \left\{\boldsymbol{\mu}^\intercal \boldsymbol{y} - \frac{\tau}{2}\boldsymbol{y}^\intercal \Sigma \boldsymbol{y}\right\}$ where $\boldsymbol{\mu}$ and Σ represent the mean vector and the covariance matrix of the asset returns, respectively. This unconstrained model is standard under the assumptions that (i) a riskless asset is available, and (ii) both long and short positions are allowed; see [9]. As a result, shrinkage can be applied to improve the MLE solution estimator. The significant impact of improving the portfolio weights through shrinkage has attracted a great deal of attention in finance during the past three decades; see [6] for a comprehensive investigation.

To our knowledge, the shrinkage phenomenon has not been exploited in the optimization context beyond portfolio selection. Proposition 1 generalizes this shrinkage idea to a broader class of optimization problems.

5. Convex Quadratic Constraint

In this section, we study a class of convex quadratic stochastic problems with a linear objective and a single quadratic constraint and we show that, surprisingly, the MLE estimator is not dominated by any other estimators. The constraint set is of the form $(\boldsymbol{y} - \boldsymbol{y}^0)^\intercal A(\boldsymbol{y} - \boldsymbol{y}^0) \leq b$ where $A \succ 0$, $\boldsymbol{y}^0 \in \mathbb{R}^n$ and $b \in \mathbb{R}$. Without loss of generality and to simplify the analysis, we use a linear transformation to reduce the constraint set to a unit ball of the form $||\boldsymbol{y}||^2 \leq 1$. Example 2 shows how complicated the loss function can become over such a simple constraint set.



(a) The surface curve (b) The contour lines

Figure 1: The loss function of Example 2 (Color available online)

Example 2. Consider an instance of (1) where $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, I)$, $f(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^\intercal \boldsymbol{y}$ and $\mathcal{Y} = \{\boldsymbol{y} \in \mathbb{R}^n \,|\, ||\boldsymbol{y}||^2 \leq 1\}$. This represents the quadratic problem of maximizing a linear function with random coefficients over the unit ball in $\mathbb{R}^n$. It follows that $\mathcal{F}(\boldsymbol{\mu}, \boldsymbol{y}) = \boldsymbol{\mu}^\intercal \boldsymbol{y}$. Since the program is convex, we compute the unique optimal solution as $\boldsymbol{y}^* = \frac{\boldsymbol{\mu}}{||\boldsymbol{\mu}||}$. Let $\hat{\boldsymbol{\mu}}$ be an estimator of $\boldsymbol{\mu}$ as a function of the observations $\{\boldsymbol{x}^t\}$. We obtain the corresponding solution estimator $\hat{\boldsymbol{y}}_{\hat{\mu}} = \frac{\hat{\boldsymbol{\mu}}}{||\hat{\boldsymbol{\mu}}||}$. We then compute $\mathcal{L}(\boldsymbol{y}^*, \hat{\boldsymbol{y}}_{\hat{\mu}}) = \mathcal{F}(\boldsymbol{\mu}, \boldsymbol{y}^*) - \mathcal{F}(\boldsymbol{\mu}, \hat{\boldsymbol{y}}_{\hat{\mu}}) = ||\boldsymbol{\mu}|| - \frac{\boldsymbol{\mu}^\intercal \hat{\boldsymbol{\mu}}}{||\hat{\boldsymbol{\mu}}||}$. It is clear that the optimal solution of this problem is determined by the direction of the coefficient vector, whereas the optimal value is determined by its magnitude. To distinguish these impacts, we rewrite the loss function as $\mathcal{L}(\boldsymbol{y}^*, \hat{\boldsymbol{y}}_{\hat{\mu}}) = ||\boldsymbol{\mu}|| \left(1 - \frac{\boldsymbol{\mu}^\intercal \hat{\boldsymbol{\mu}}}{||\boldsymbol{\mu}||||\hat{\boldsymbol{\mu}}||}\right) = ||\boldsymbol{\mu}||(1 - \cos(\phi))$ where ϕ represents the angle between the vectors $\boldsymbol{\mu}$ and $\hat{\boldsymbol{\mu}}$.

Figure 1 illustrates $\mathcal{L}(\boldsymbol{y}^*, \hat{\boldsymbol{y}}_{\hat{\mu}})$ of Example 2 and its contour lines for an instance in $\mathbb{R}^2$, where $(\mu_1, \mu_2) = (0, 1)$. This loss function is not of a conventional quadratic form. Since it is not location-invariant (Brown [5]), the admissibility results in the statistical literature do not apply for such a function. Next, we investigate the admissibility of the MLE for this function through the notion of directional statistics [12].

Proposition 2. *Assume that* $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, I)$ *and* $\mathcal{L}(\boldsymbol{y}^*, \hat{\boldsymbol{y}}_{\hat{\mu}}) = ||\boldsymbol{\mu}|| \left(1 - \frac{\boldsymbol{\mu}^\intercal \hat{\boldsymbol{\mu}}}{||\boldsymbol{\mu}||||\hat{\boldsymbol{\mu}}||}\right)$. *Then, the MLE solution estimator* $\hat{\boldsymbol{y}}_{\bar{x}}$ *is admissible among all estimators* $\hat{\boldsymbol{y}}_{\hat{\mu}}$.

Proof. In the proof, we use Theorem 2.1 in [2] which is rephrased as follows. For the loss function $\mathcal{L}(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) = ||\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}||^2$ where $||\boldsymbol{\mu}||$ is fixed at k, the estimator $\hat{\boldsymbol{\mu}} = k\frac{\bar{\boldsymbol{x}}}{||\bar{\boldsymbol{x}}||}$ is admissible among all estimators whose norm is k. We now return to the proof. First, note that $\mathcal{L}(\boldsymbol{y}^*, \hat{\boldsymbol{y}}_{\hat{\mu}}) = ||\boldsymbol{\mu}|| \left(1 - \frac{\boldsymbol{\mu}^\intercal \hat{\boldsymbol{\mu}}}{||\boldsymbol{\mu}||||\hat{\boldsymbol{\mu}}||}\right) = \frac{||\boldsymbol{\mu}||}{2} \left|\left| \frac{\hat{\boldsymbol{\mu}}}{||\hat{\boldsymbol{\mu}}||} - \frac{\boldsymbol{\mu}}{||\boldsymbol{\mu}||} \right|\right|^2$. Assume by contradiction that $\hat{\boldsymbol{y}}_{\bar{x}}$ is inadmissible. Therefore, there exists another estimator $\hat{\boldsymbol{y}}_{\hat{\mu}}$ for some $\hat{\boldsymbol{\mu}}$ (which is a function of $\{\boldsymbol{x}^t\}$) that strictly dominates

$\hat{\boldsymbol{y}}_{\bar{x}}$. It follows from the above relation and the definition of inadmissibility that $\mathbb{E}_{\{\boldsymbol{x}^t\}|\mu}\left[\left|\left|\frac{\hat{\boldsymbol{\mu}}}{||\hat{\boldsymbol{\mu}}||}-\frac{\boldsymbol{\mu}}{||\boldsymbol{\mu}||}\right|\right|^2\right] \leq \mathbb{E}_{\{\boldsymbol{x}^t\}|\mu}\left[\left|\left|\frac{\bar{\boldsymbol{x}}}{||\bar{\boldsymbol{x}}||}-\frac{\boldsymbol{\mu}}{||\boldsymbol{\mu}||}\right|\right|^2\right]$ for all $\boldsymbol{\mu}\in\mathbb{R}^n\setminus\{0\}$, and $\mathbb{E}_{\{\boldsymbol{x}^t\}|\mu}\left[\left|\left|\frac{\hat{\boldsymbol{\mu}}}{||\hat{\boldsymbol{\mu}}||}-\frac{\bar{\boldsymbol{\mu}}}{||\bar{\boldsymbol{\mu}}||}\right|\right|^2\right] < \mathbb{E}_{\{\boldsymbol{x}^t\}|\mu}\left[\left|\left|\frac{\bar{\boldsymbol{x}}}{||\bar{\boldsymbol{x}}||}-\frac{\bar{\boldsymbol{\mu}}}{||\bar{\boldsymbol{\mu}}||}\right|\right|^2\right]$ for some $\bar{\boldsymbol{\mu}}\in\mathbb{R}^n\setminus\{0\}$. Define $P=\{\boldsymbol{\mu}\in\mathbb{R}^n\setminus\{0\}\,|\,||\boldsymbol{\mu}||=||\bar{\boldsymbol{\mu}}||\}$ to be a subset of parameters whose norm is equal to $||\bar{\boldsymbol{\mu}}||$. Restricting attention to the family of parameters in P, we can write that $\left|\left|\frac{\bar{\boldsymbol{x}}}{||\bar{\boldsymbol{x}}||}-\frac{\boldsymbol{\mu}}{||\boldsymbol{\mu}||}\right|\right|^2 = \frac{1}{||\bar{\boldsymbol{\mu}}||}\left|\left|||\bar{\boldsymbol{\mu}}||\frac{\bar{\boldsymbol{x}}}{||\bar{\boldsymbol{x}}||}-\boldsymbol{\mu}\right|\right|^2$ for all $\boldsymbol{\mu}\in P$. Using this relation in the inadmissibility definition given above, we obtain that $\mathbb{E}_{\{\boldsymbol{x}^t\}|\boldsymbol{\mu}}\left[\left|\left|||\bar{\boldsymbol{\mu}}||\frac{\hat{\boldsymbol{\mu}}}{||\hat{\boldsymbol{\mu}}||}-\boldsymbol{\mu}\right|\right|^2\right] \leq \mathbb{E}_{\{\boldsymbol{x}^t\}|\boldsymbol{\mu}}\left[\left|\left|||\bar{\boldsymbol{\mu}}||\frac{\bar{\boldsymbol{x}}}{||\bar{\boldsymbol{x}}||}-\boldsymbol{\mu}\right|\right|^2\right]$ for all $\boldsymbol{\mu}\in P$, and $\mathbb{E}_{\{\boldsymbol{x}^t\}|\bar{\boldsymbol{\mu}}}\left[\left|\left|||\bar{\boldsymbol{\mu}}||\frac{\hat{\boldsymbol{\mu}}}{||\hat{\boldsymbol{\mu}}||}-\bar{\boldsymbol{\mu}}\right|\right|^2\right] < \mathbb{E}_{\{\boldsymbol{x}^t\}|\bar{\boldsymbol{\mu}}}\left[\left|\left|||\bar{\boldsymbol{\mu}}||\frac{\bar{\boldsymbol{x}}}{||\bar{\boldsymbol{x}}||}-\bar{\boldsymbol{\mu}}\right|\right|^2\right]$. This inequality implies that the estimator $||\bar{\boldsymbol{\mu}}||\frac{\hat{\boldsymbol{\mu}}}{||\hat{\boldsymbol{\mu}}||}$ strictly dominates $||\bar{\boldsymbol{\mu}}||\frac{\bar{\boldsymbol{x}}}{||\bar{\boldsymbol{x}}||}$ over the family of parameters $\boldsymbol{\mu}\in P$ and under the loss $||\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}||^2$. This is a contradiction to Theorem 2.1 in [2] given above as both $||\bar{\boldsymbol{\mu}}||\frac{\hat{\boldsymbol{\mu}}}{||\hat{\boldsymbol{\mu}}||}$ and $||\bar{\boldsymbol{\mu}}||\frac{\bar{\boldsymbol{x}}}{||\bar{\boldsymbol{x}}||}$ have norms equal to $||\bar{\boldsymbol{\mu}}||$. □

The drastic difference in the admissibility results of the MLE solution estimator between the quadratic problems in Section 4 and those in this section can be surprising for an optimizer. Indeed, properties such as optimality conditions for a convex program whether the convexity is in the objective or in the constraints are translated without a fundamental change, a feature that does not hold for the admissibility property. To reinforce this remark, we revisit the problem in Example 2 by considering its Lagrangian relaxation where we move the quadratic constraint to the objective, and show that the shrinkage does improve the MLE solution estimator for this formulation.

The Lagrangian function is expressed as $\mathcal{F}_\lambda(\boldsymbol{\mu},\boldsymbol{y})=\boldsymbol{\mu}^\intercal\boldsymbol{y}-\lambda\boldsymbol{y}^\intercal\boldsymbol{y}+\lambda$ where $\boldsymbol{y}\in\mathbb{R}^n$ and $\lambda\in\mathbb{R}_+$. Because of strong duality, it is easy to verify that the Lagrangian problem $\max_{\boldsymbol{y}\in\mathbb{R}^n}\mathcal{F}_{\lambda^*}(\boldsymbol{\mu},\boldsymbol{y})$ for $\lambda^*=\frac{||\boldsymbol{\mu}||}{2}$ has the same optimal solution $\boldsymbol{y}^*$ and optimal value z^* as the original problem. We can view this model as an unconstrained formulation of the unit-ball problem.

Now consider a special case of the parameter estimation problem where the length of $\boldsymbol{\mu}$ is known, say $||\boldsymbol{\mu}||=k$, and it is only the direction of $\boldsymbol{\mu}$ that needs to be estimated. As mentioned in Example 2, the optimal solution of the unit-ball problem is determined only by the direction of $\boldsymbol{\mu}$, and hence the length restriction does not affect the form of the problem. On the other hand, this restriction can be embedded in the Lagrangian model above by setting $\lambda^*=\frac{k}{2}$. Note that for this restricted case, the optimal solution and optimal value of the two models are still the same.

Proposition 3. *Assume that* $\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},I)$ *and that* $||\boldsymbol{\mu}||=k>0$.

(i) For problem $\max_{\boldsymbol{y}\in\mathbb{R}^n}\left\{\boldsymbol{\mu}^\intercal\boldsymbol{y}\,|\,||\boldsymbol{y}||^2\leq 1\right\}$*, the MLE solution estimator* $\hat{\boldsymbol{y}}_{\bar{x}}$ *is admissible.*

(ii) For problem $\max_{\boldsymbol{y}\in\mathbb{R}^n} \mathcal{F}_{\lambda^*}(\boldsymbol{\mu}, \boldsymbol{y})$ *with* $\lambda^* = \frac{k}{2}$*, the shrinkage solution estimator* $\hat{\boldsymbol{y}}_{\bar{x}}$ *with shrinkage factor* $\rho = \frac{c(||\bar{\boldsymbol{x}}-\boldsymbol{x}^0||^2)}{T||\bar{\boldsymbol{x}}-\boldsymbol{x}^0||^2}$ *where* $0 < c(.) < 2(n-2)$ *and* $c'(.) \geq 0$*, strictly dominates the MLE solution estimator* $\hat{\boldsymbol{y}}_{\bar{x}}$ *for* $n \geq 3$*.*

Proof. *(i)* The result follows from the discussion in the proof of Proposition 2 that the MLE solution estimator is also admissible when $||\boldsymbol{\mu}||$ is fixed.

(ii) The model $\max_{\boldsymbol{y}\in\mathbb{R}^n} \mathcal{F}_{\lambda^*}(\boldsymbol{\mu}, \boldsymbol{y})$ is of the form (4) given in Example 1 with an additional constant. The result follows from Proposition 1 and its following discussion.

□

We conclude this section by noting that, although the results in this section concern quadratic programs, they can provide insight for problems with more general structures. For instance, if the loss function of the stochastic program can be expressed explicitly, the admissibility of estimators around sufficiently small neighborhoods of the mean may be studied via Proposition 1 by using a second order Taylor expansion. Similarly, if the loss function of the stochastic program cannot be obtained in closed-form, its convex relaxations may be studied under the constrained quadratic programs of Proposition 2.

6. Computational Results

In this section, we present computational results for two formulations of the unit-ball problem of Example 2. The first formulation is $\max_{\boldsymbol{y}\in\mathbb{R}^n} \left\{\boldsymbol{\mu}^\intercal \boldsymbol{y} \,|\, ||\boldsymbol{y}||^2 \leq 1\right\}$ with loss function $\mathcal{L}^1(\boldsymbol{y}^*, \hat{\boldsymbol{y}}_{\hat{\boldsymbol{\mu}}}) = ||\boldsymbol{\mu}|| \left(1 - \frac{\boldsymbol{\mu}^\intercal \hat{\boldsymbol{\mu}}}{||\boldsymbol{\mu}||||\hat{\boldsymbol{\mu}}||}\right)$, and the second formulation is $\max_{\boldsymbol{y}\in\mathbb{R}^n} \left\{\boldsymbol{\mu}^\intercal \boldsymbol{y} - \frac{||\boldsymbol{\mu}||}{2}\boldsymbol{y}^\intercal \boldsymbol{y} + \frac{||\boldsymbol{\mu}||}{2}\right\}$ with loss function $\mathcal{L}^2(\boldsymbol{y}^*, \hat{\boldsymbol{y}}_{\hat{\boldsymbol{\mu}}}) = \frac{1}{2||\boldsymbol{\mu}||}||\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}||^2$. Supporting the theoretical results developed in Sections 4 and 5, our computational experiments show that for the first (constrained) formulation the shrinkage estimator has often a higher risk than the MLE solution estimator. In contrast for the second (unconstrained) formulation, the shrinkage estimator always yields a lower risk than the MLE solution estimator.

We set $n = 20$. We consider random variables $x_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \sigma^2)$ for $i \in [n]$ where $\boldsymbol{\mu}_i$ is unknown. In the frequentist framework, there is no prior on the unknown parameters. However, to perform out-of-sample experiments via simulation, we generate the unknown parameters from a distribution. We choose $\boldsymbol{\mu}_i \sim \mathcal{N}(\lambda, \delta^2)$ for $i \in [n]$. There are two main factors that affect the risk. The first factor is the direction of vector $\boldsymbol{\mu}$ which is reflected in the loss function; see Example 2. To obtain a wide range of directions, we set $\lambda = 0$ so that for each component the chances of being on either sides of the origin are equal. The second factor is the concentration of the distribution around the mean. To obtain different ranges for this concentration, we consider two sets of experiments for different values of the ratio $\frac{\delta}{\sigma}$. In the first set of experiments we fix $\delta = 1$ and $\sigma = 10$, and in the second we fix $\delta = 10$ and $\sigma = 10$. For both experiments, we randomly generate 20 instances, that is 20 randomly generated

values for vector $\boldsymbol{\mu}$ according to its assumed distribution, and one observation vector $\bar{\boldsymbol{x}}$, that is $T = 1$. For each instance, we generate $K = 10000$ values for the observation vector $\bar{\boldsymbol{x}}$. For each $\bar{\boldsymbol{x}}$, we compute the two loss functions $\mathcal{L}^1(\boldsymbol{y}^*, \hat{\boldsymbol{y}}_{\bar{x}})$ and $\mathcal{L}^2(\boldsymbol{y}^*, \hat{\boldsymbol{y}}_{\bar{x}})$. Then, we compute the approximate risks $\mathcal{R}^1(\boldsymbol{y}^*, \hat{\boldsymbol{y}}_{\bar{x}})$ and $\mathcal{R}^2(\boldsymbol{y}^*, \hat{\boldsymbol{y}}_{\bar{x}})$ by taking the average of the loss functions over all values of $\bar{\boldsymbol{x}}$.

We present the computational results for $\frac{\delta}{\sigma} = \frac{1}{10}$ in Table 1a and for $\frac{\delta}{\sigma} = \frac{10}{10}$ in Table 1b. The second column of each table contains the risk value of the MLE solution estimator for the constrained unit-ball formulation, *i.e.,* $\mathcal{R}^1_{\text{SM}} = \mathcal{R}^1(\boldsymbol{y}^*, \hat{\boldsymbol{y}}_{\bar{x}}) = \mathbb{E}_{\bar{x}|\boldsymbol{\mu}}[\mathcal{L}^1(\boldsymbol{y}^*, \hat{\boldsymbol{y}}_{\bar{x}})]$. The third column contains the risk value of the shrinkage estimator for the same formulation, *i.e.,* $\mathcal{R}^1_{\text{SH}} = \mathcal{R}^1(\boldsymbol{y}^*, \hat{\boldsymbol{y}}_{\bar{x}})$ where $\tilde{\boldsymbol{x}} = \rho \boldsymbol{x}^0 + (1-\rho)\bar{\boldsymbol{x}}$, $\rho = \min\{1, \frac{n-2}{||\bar{\boldsymbol{x}} - \boldsymbol{x}^0||^2}\}$ and $\boldsymbol{x}^0 = \frac{\mathbf{1}^\intercal \boldsymbol{x}}{n}\mathbf{1}$. This choice of target vector is common in statistical models and it is sometimes referred to as the *grand average* since it computes the average of the MLE components. The fourth column shows the percentage of the gap closure achieved by the risk of the shrinkage estimator compared to that of the MLE solution estimator. This quantity is computed as $\Delta^1 = \frac{\mathcal{R}^1_{\text{SM}} - \mathcal{R}^1_{\text{SH}}}{\mathcal{R}^1_{\text{SM}}}$. Similarly, the fifth, sixth and seventh columns represent the risk value of the MLE solution estimator, the risk value of the shrinkage estimator, and the gap closure for the Lagrangian formulation, respectively.

Table 1 shows that the MLE solution estimator has a consistently better risk than the shrinkage estimator for our choice of parameters in the constrained formulation, but the reverse is true in the Lagrangian version of the problem. This support the theoretical results of Propositions 2 and 3.

#	$\mathcal{R}^1_{\text{SM}}$	$\mathcal{R}^1_{\text{SH}}$	Δ^1	$\mathcal{R}^2_{\text{SM}}$	$\mathcal{R}^2_{\text{SH}}$	Δ^2
1	3.52	3.71	-5%	258.07	26.60	90%
2	3.52	3.71	-5%	260.23	27.09	90%
3	3.13	3.26	-4%	297.15	30.63	90%
4	2.49	2.57	-3%	377.48	37.44	90%
5	3.55	3.67	-3%	257.90	26.38	90%
6	3.58	3.76	-5%	254.46	26.60	90%
7	4.14	4.39	-6%	219.01	23.72	89%
8	4.22	4.48	-6%	212.05	22.79	89%
9	4.31	4.59	-7%	206.84	22.07	89%
10	3.67	3.87	-5%	247.00	25.48	90%
11	3.57	3.75	-5%	254.25	26.16	90%
12	4.62	4.94	-7%	192.36	21.20	89%
13	4.13	4.30	-4%	216.63	22.92	89%
14	4.05	4.30	-6%	220.55	23.16	89%
15	4.48	4.76	-6%	199.42	21.63	89%
16	3.60	3.77	-5%	254.66	26.88	89%
17	3.97	4.03	-1%	229.00	24.01	90%
18	4.33	4.61	-7%	206.35	22.42	89%
19	4.60	4.93	-7%	192.00	21.12	89%
20	4.53	4.85	-7%	195.84	21.42	89%

(a) The first set of experiments

#	$\mathcal{R}^1_{\text{SM}}$	$\mathcal{R}^1_{\text{SH}}$	Δ^1	$\mathcal{R}^2_{\text{SM}}$	$\mathcal{R}^2_{\text{SH}}$	Δ^2
1	12.56	13.84	-10%	20.11	12.88	36%
2	13.42	16.11	-20%	26.78	13.71	49%
3	13.21	12.42	6%	23.78	11.85	50%
4	12.90	14.40	-12%	21.29	13.20	38%
5	12.85	13.89	-8%	21.01	12.90	39%
6	12.02	10.17	15%	17.63	10.18	42%
7	12.63	13.35	-6%	19.99	12.56	37%
8	13.31	14.95	-12%	24.82	13.28	47%
9	12.97	14.41	-11%	22.57	13.14	42%
10	12.93	12.64	2%	21.73	12.03	45%
11	13.07	14.98	-15%	23.03	13.44	42%
12	12.34	12.55	-2%	18.84	12.01	36%
13	13.18	13.92	-6%	23.70	12.81	46%
14	12.87	14.40	-12%	21.35	13.14	38%
15	13.28	16.69	-26%	33.70	13.76	59%
16	12.00	12.82	-7%	17.59	12.30	30%
17	13.31	15.48	-16%	25.16	13.55	46%
18	12.69	14.10	-11%	20.33	13.00	36%
19	13.48	14.58	-8%	27.64	12.98	53%
20	12.95	14.36	-11%	22.03	13.12	40%

(b) The second set of experiments

Table 1: Risk comparison for different estimators

Acknowledgement

We thank Jay Kadane and R. Ravi for very helpful discussions. This work was supported in part by NSF grant CMMI1560828 and ONR grant N00014-12-10032.

Reference

[1] Baranchik, A. J., 1964. Multiple regression and estimation of the mean of a multivariate normal distribution. Tech. rep.

[2] Beran, R., 1996. Stein estimation in high dimensions: a retrospective. In: Brunner, E., Denker, M. (Eds.), Madan Puri Festschrift. VSP, Zeist., pp. 91–110.

[3] Birge, J., Louveaux, F., 2011. Introduction to stochastic programming. Springer.

[4] Blyth, C. R., 1951. On minimax statistical decision procedures and their admissibility. The Annals of Mathematical Statistics 22, 22–42.

[5] Brown, L. D., 1966. On the admissibility of invariant estimators of one or more location parameters. The Annals of Mathematical Statistics 37, 1087–1136.

[6] DeMiguel, V., Martin-Utrera, A., Nogales, F. J., 2013. Size matters: Optimal calibration of shrinkage estimators for portfolio selection. Journal of Banking and Finance 37, 3018–3034.

[7] James, W., Stein, C., 1961. Estimation with quadratic loss. In: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability. Vol. 1. University of California Press, pp. 361–379.

[8] Jorion, P., 1986. Bayes-Stein estimation for portfolio analysis. The Journal of Financial and Quantitative Analysis 21, 279–292.

[9] Kan, R., Zhou, G., 2007. Optimal portfolio choice with parameter uncertainty. Journal of Financial and Quantitative Analysis 42, 621–656.

[10] Lehmann, E. L., Casella, G., 1998. Theory of Point Estimation. Springer-Verlag.

[11] Lim, A. E. B., Shanthikumar, J. G., Shen, Z. J. M., 2006. Model uncertainty, robust optimization, and learning. In: Johnson, M. P., Norman, B., Secomandi, N. (Eds.), Tutorials in Operations Research: Models, Methods, and Applications for Innovative Decision Making. INFORMS, pp. 66–94.

[12] Mardia, K. V., Jupp, P. E., 1999. Directional Statistics. Wiley.

[13] Markowitz, H. M., 1959. Portfolio Selection: Efficient Diversification of Investments. John Wiley.

[14] Stein, C., 1956. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability. Vol. 1. University of California Press, pp. 197–206.