Operations Research Letters 45 (2017) 642–646

Contents lists available at ScienceDirect

Operations Research Letters

journal homepage: www.elsevier.com/locate/orl

From estimation to optimization via shrinkage

Danial Davarnia *, Gérard Cornuéjols

Tepper School of Business, Carnegie Mellon University, Pittsburgh, PA, USA

ARTICLE INFO

Article history:
Received 29 September 2017
Received in revised form 8 October 2017
Accepted 8 October 2017
Available online 23 October 2017

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

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.

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.

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

* Corresponding author.
E-mail addresses: ddavarni@andrew.cmu.edu (D. Davarnia), gc0v@andrew.cmu.edu (G. Cornuéjols).

https://doi.org/10.1016/j.orl.2017.10.005
0167-6377/

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* $\hat{\boldsymbol{y}}_{\tilde{\boldsymbol{x}}}$ *strictly dominates the MLE solution estimator* $\hat{\boldsymbol{y}}_{\bar{\boldsymbol{x}}}$ *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 $\boldsymbol{x}^0 = \boldsymbol{0}$. 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

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

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 $\boldsymbol{x}^0 = \boldsymbol{0}$, 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

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