From Estimation to Optimization via Shrinkage

Danial Davarnia, Gérard Cornuéjols

Carnegie Mellon University

October 2017

Batting average in baseball

	Game 1	Game 2	Game 3	Average	Game 4
Al	0.268	0.270	0.220	0.253	?
Bob	0.167	0.328	0.375	0.290	?
Chris	0.468	0.435	0.313	0.406	?

QUESTION What is your best estimate of Al's batting average in the next game?

FIRST ANSWER Use Al's average over Games 1-2-3. This is the maximum likelihood estimator. It is unbiased, efficient, consistent. All good statistical properties.

QUESTIONS If all three players had the same skill to start with, we would expect to see a difference in their averages over three games. Can we squeeze out some of this randomness? Can we do better than the MLE?

Batting average in baseball

The small sample makes it hard to know what Al's true batting average is, due to the randomness in game-to-game scores.

Imagine 3 players with the same batting average over 100 games. Looking at just 3 games, one of the players will perform best, one will be in the middle, one will be last.

Even if the 3 players do not have the same batting average to start with, some of their performance over a sample of 3 games can be attributed to skill, and some to chance.

Can we use the data from the other players to improve our estimation of Al's performance?

Batting average in baseball

We could compute the ”grand average” $\frac{0.253+0.290+0.406}{3} = 0.316$ and shrink each player’s average towards the grand average.

	Game 1	Game 2	Game 3	Average	Shrunk Avg	Actual
Al	0.268	0.270	0.220	0.253	0.284	0.280
Bob	0.167	0.328	0.375	0.290	0.303	0.336
Chris	0.468	0.435	0.313	0.406	0.360	0.299

Notice how the ”Shrunk average” is better that the simple ”Average” in estimating the ”Actual batting average”.

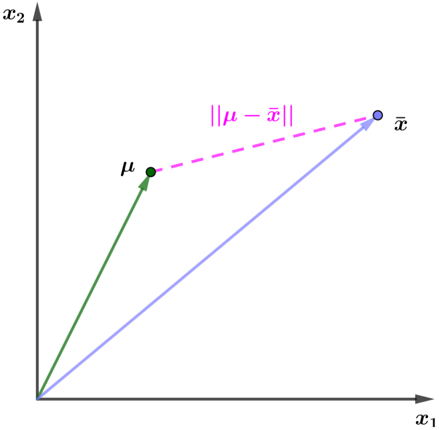
This is not a coincidence. Stein made this precise in a paper that stunned the world in 1956.

Data: I used $b_{Al}, b_B, b_C \sim N(0.32, 0.02^2)$ and $\text{Game}_{Al,i} \sim N(b_{Al}, 0.1^2)$.

Stein's shrinkage: a geometric interpretation

Assume that parameter μ is unknown and data is used to find the closet (on average) estimator to μ, i.e., $\min_{\hat{\mu}} ||\mu - \hat{\mu}||^2$.

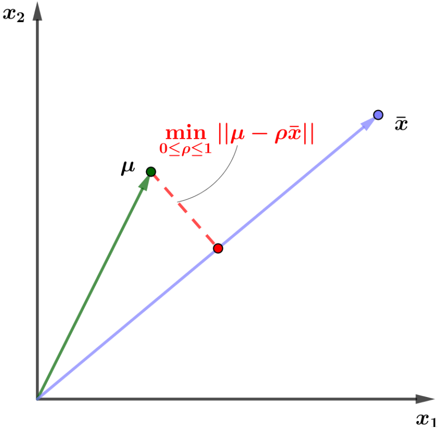
A natural choice for $\hat{\mu}$ is the MLE $\bar{x}$.



Stein's shrinkage: a geometric interpretation

Assume that parameter μ is unknown and data is used to find the closet (on average) estimator to μ, i.e., $\min_{\hat{\mu}} ||\mu - \hat{\mu}||^2$.

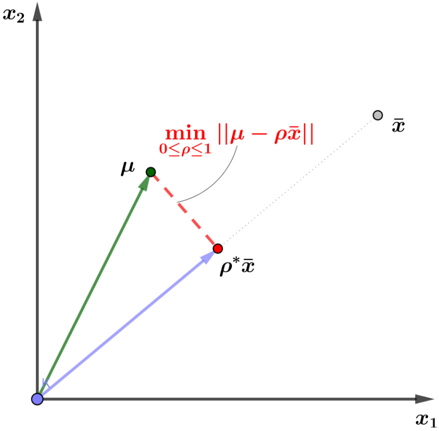
But we can do better than $\bar{x}$ by shrinking it when it is "large".



Stein's shrinkage: a geometric interpretation

Assume that parameter μ is unknown and data is used to find the closet (on average) estimator to μ, i.e., $\min_{\hat{\mu}} ||\mu - \hat{\mu}||^2$.

Stein suggests a shrinking factor $\rho^* = (1 - \frac{n-2}{||\bar{x}||^2})$.



From Estimation to Optimization

What is the significance of this observation in the context of optimization?

Consider the following parametric stochastic optimization problem

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

Vector $\mathbf{x} \in \mathbb{R}^{\mathbf{n}}$ has a known probability distribution with joint density $\mathcal{G}(\mathbf{x}|\theta)$ where θ are unknown parameters of the distribution.

$\mathbf{y}$ is a vector of decision variables that belongs to $\mathcal{Y} \subseteq \mathbb{R}^m$.

The expectation $\mathbb{E}_{\mathbf{x}|\theta}[.]$ is taken with respect to the distribution of the random variables $\mathbf{x}$ given the vector θ of parameters.

Writing $\mathbb{E}_{\mathbf{x}|\theta}[f(\mathbf{x},\mathbf{y})] = \mathcal{F}(\theta,\mathbf{y})$, we refer to $\mathcal{F}(\theta,\mathbf{y})$ as a *parametric* objective function.

Portfolio optimization

Consider n assets with random returns $\mathbf{x} \in \mathbb{R}^{\mathbf{n}}$ over the next period. Let μ be the vector of expected returns and Σ the covariance matrix of the returns.

Goal: Construct a portfolio $\mathbf{y}$ of assets that maximizes expected return minus a penalty for higher risk.

Let $y_1, \ldots, y_n$ denote the fraction of our wealth invested in asset $i = 1, \ldots, n$. The portfolio return $\mathbf{x}^T\mathbf{y}$ is random.

Markovitz used the variance $\mathbf{y}^T\Sigma\mathbf{y}$ of the portfolio return as a measure of risk. Let γ be a risk-aversion factor.

$$\begin{aligned} \text{Max} \quad & \mu^T\mathbf{y} - \gamma\mathbf{y}^T\Sigma\mathbf{y} \\ & \sum_{i=1}^{n} y_i = 1 \\ & \mathbf{y} \geq \mathbf{0} \end{aligned}$$

Portfolio optimization

We have historical data on the asset returns $x^1, \ldots, x^K$ over K periods.

The standard approach is to estimate μ and Σ using maximum likelihood estimators $\bar{\mu}$ and $\bar{\Sigma}$. In particular $\bar{\mu} = \frac{1}{K}\sum_{i=1}^{K} x^i$.

We solve

$$\begin{aligned} \text{Max} \quad & \bar{\mu}^T \mathbf{y} - \gamma \mathbf{y}^T \bar{\Sigma} \mathbf{y} \\ & \sum_{i=1}^{n} y_i = 1 \\ & \mathbf{y} \geq \mathbf{0} \end{aligned}$$

Portfolio optimization

Jorion 1986 recommends to *shrink* the vector of sample averages $\bar{\mu}$ towards a *grand average* $\begin{pmatrix} \bar{\mu}_0 \\ \vdots \\ \bar{\mu}_0 \end{pmatrix}$ where $\bar{\mu}_0 = \frac{1}{n} \sum_{i=1}^{n} \bar{\mu}_i$, and to use this shrunk estimator in the Markowitz optimization model to obtain better portfolios.

That is, use $\tilde{\mu} = (1-\rho)\bar{\mu} + \rho \begin{pmatrix} \bar{\mu}_0 \\ \vdots \\ \bar{\mu}_0 \end{pmatrix}$ instead of $\bar{\mu}$ in the portfolio optimization formulation, where $0 < \rho < 1$.

In this talk, we address the question of where this shrinkage idea fits in the optimization literature, focusing on the impact of constraints.

Good Estimator of the Optimal Solution

$$\max_{\mathbf{y} \in \mathcal{Y}} \mathcal{F}(\theta, \mathbf{y})$$

Since $\mathcal{F}(\theta, \mathbf{y})$ is a function of θ and $\mathbf{y}$, its optimal solution $\mathbf{y}^*(\theta)$ and its optimal value $\mathcal{F}(\theta, \mathbf{y}^*(\theta))$ are functions of θ.

Recall $\mathcal{F}(\theta, \mathbf{y}) = \mathbb{E}_{\mathbf{x}|\theta}[\mathbf{f}(\mathbf{x}, \mathbf{y})]$

A finite number K of i.i.d. observations $\{\mathbf{x^t}\}_{\mathbf{t=1,\dots,K}}$ (obtained from computer simulation, historical data, prediction, etc) is available for the random variables $\mathbf{x}$.

The data is used to obtain an approximate solution (*estimator*) $\hat{\mathbf{y}}$ for the true optimal solution (*estimand*) $\mathbf{y}^*$.

Our goal is to obtain "good" estimators $\hat{\mathbf{y}}$ for $\mathbf{y}^*$.

Loss Function

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

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

A smaller loss indicates a better estimator.

$\hat{\mathbf{y}}$ belongs to $\mathcal{Y}$, and therefore $\mathcal{F}(\theta, \hat{\mathbf{y}}) \leq \mathcal{F}(\theta, \mathbf{y}^*)$.

Observe that the loss function is a random quantity since the estimator $\hat{\mathbf{y}}$ is a function of the observations $\{\mathbf{x}^{\mathbf{t}}\}$.

Risk

The loss function $\mathcal{L}(\mathbf{y}^*, \hat{\mathbf{y}}) = \mathcal{F}(\theta, \mathbf{y}^*) - \mathcal{F}(\theta, \hat{\mathbf{y}})$ is a random quantity since the estimator $\hat{\mathbf{y}}$ is a function of the observations $\{\mathbf{x}^\mathbf{t}\}$.

Therefore, to evaluate the overall performance of the estimator $\hat{\mathbf{y}}$, an averaging measure for the loss function is defined. This measure is referred to as the *risk*

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

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

Risk

The risk $\mathcal{R}(\mathbf{y}^*, \hat{\mathbf{y}})$ is a function of the unknown parameters θ.

The treatment of the risk is different depending on whether the unknown parameters θ are assumed to be random or fixed.

This key assumption on the model parameters gives rise to two major statistical frameworks: Bayesian and frequentist.

We first 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 $\Theta = \mathbb{R}^n$.

Admissibility

An estimator $\hat{\mathbf{y}}^1$ *strictly dominates* another estimator $\hat{\mathbf{y}}^2$ if $\mathcal{R}(\mathbf{y}^*, \hat{\mathbf{y}}^{\mathbf{1}}) \leq \mathcal{R}(\mathbf{y}^*, \hat{\mathbf{y}}^{\mathbf{2}})$ for all values of the parameters θ, with strict inequality for some values of θ.

An estimator $\hat{\mathbf{y}}^1$ is *inadmissible*, if there exists an estimator $\hat{\mathbf{y}}^2$ that strictly dominates it. Otherwise, it is *admissible*. It is a common-sense rule 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 in statistics.
Our goal is to pursue these directions in optimization.

Shrinkage estimators in statistics

Studying the admissibility of a given estimator $\hat{\mu}$ is a hard task even under simple distributional settings and problem structures. A common setting is a normal distribution and a squared error loss function $\mathcal{L}(\mu, \hat{\mu}) = ||\mu - \hat{\mu}||^{\mathbf{2}}$.

Blyth 1951 showed that, under the squared error loss, the maximum likelihood estimator $\bar{\mathbf{x}}$ is admissible when $n = 1$ and $n = 2$.

Stein 1956 stunned the statistical world by showing that $\bar{\mathbf{x}}$ is inadmissible when $n \geq 3$.

James and Stein 1961 proved that $\bar{\mathbf{x}}$ is strictly dominated by an estimator of the form $\tilde{\mathbf{x}} = \rho\mathbf{x}^0 + (1 - \rho)\bar{\mathbf{x}}$ where $\rho = \frac{n-2}{||\mathbf{x^0} - \bar{\mathbf{x}}||^2}$ and $\mathbf{x^0}$ is an arbitrary *target* vector in $\mathbb{R}^n$.

Baranchik 1964 improved the James-Stein estimator by modifying the factor ρ to $\rho^+ = \min\{\rho, 1\}$. This estimator is referred to as the *shrinkage estimator*.

From estimation to optimization

The above statistical results are established in the space of parameters under a loss function $\mathcal{L}(\mu, \hat{\mu})$ that measures the distance between the estimator $\hat{\mu}$ and the parameter μ.

For optimization problems, on the other hand, we are interested in the space of decision variables, where the loss function $\mathcal{L}(\mathbf{y}^*, \hat{\mathbf{y}})$ measures the difference in the objective value between the solution estimator $\hat{\mathbf{y}}$ and the optimal solution $\mathbf{y}^*$.

The question of interest is how does a shrinkage solution estimator $\hat{\mathbf{y}}_{\tilde{\mathbf{x}}}$ compare to the maximum likelihood solution estimator $\hat{\mathbf{y}}_{\bar{\mathbf{x}}}$?

From estimation to optimization

Consider two classes of convex stochastic programs, one with the quadratic term in the objective and the other in the constraints.

$$\max\{\boldsymbol{\mu}^\intercal \boldsymbol{y} - \tau \boldsymbol{y}^\intercal A \boldsymbol{y} \mid \boldsymbol{y} \in \mathbb{R}^n\}, \tag{1}$$

$$\max\{\boldsymbol{\mu}^\intercal \boldsymbol{y} \mid \boldsymbol{y}^\intercal A \boldsymbol{y} \leq \tau\}, \tag{2}$$

where μ is an unknown mean of a multi-normal distribution, and where $\tau > 0$ and $A \succeq 0$.

We show that:
A generalization of Stein's shrinkage can uniformly improve the MLE's risk for problem (1): PROPOSITION 1
There exists no estimator that can uniformly improve the MLE's risk for problem (2): PROPOSITION 2

Convex quadratic program without constraints

PROPOSITION 1 Assume that $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{I})$, 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 maximum likelihood solution estimator $\hat{\boldsymbol{y}}_{\bar{x}}$ for any $\tilde{\boldsymbol{x}} = \rho \boldsymbol{x}^0 + (1 - \rho)\bar{\boldsymbol{x}}$ where $\rho = \frac{c(||\bar{\boldsymbol{x}} - \boldsymbol{x}^0||^2)}{K||\bar{\boldsymbol{x}} - \boldsymbol{x}^0||^2}$, provided
(i) $0 < c(.) < \inf_{\boldsymbol{\mu} \in \mathbb{R}^n} 2\frac{\mathrm{tr}(Q_\mu)}{\lambda_{\max}(Q_\mu)} - 4$, and
(ii) the function $c(.)$ has nonnegative derivative.

In the above definition, $\mathrm{tr}(Q_\mu)$ and $\lambda_{\max}(Q_\mu)$ represent the trace and the maximum eigenvalue of Q_μ respectively.

Application to portfolio optimization

A special case appears in Markowitz' portfolio selection problem $\max_{\boldsymbol{y}\in\mathbb{R}^n}\left\{\boldsymbol{\mu}^\top\boldsymbol{y}-\gamma\boldsymbol{y}^\top\Sigma\boldsymbol{y}\right\}$ 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.

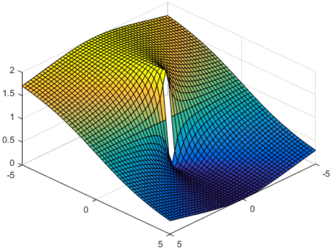
As a result, shrinkage can be applied to improve the MLE solution estimator.

The significant impact of improving the portfolio weights through shrinkage has attracted a great deal of attention in finance during the past three decades.

Convex program with a quadratic constraint

Consider a stochastic problem with a linear objective $\mu^T \boldsymbol{y}$ and a single quadratic constraint 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 $b \in \mathbb{R}$. We show that, surprisingly, the MLE estimator is not dominated by any other estimators.

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 $||\boldsymbol{y}||^2 \leq 1$.



The figure shows how complicated the loss function can become over such a simple constraint set.

Lagrangian relaxation

The Lagrangian function is $\mathcal{F}_\lambda(\boldsymbol{\mu}, \boldsymbol{y}) = \boldsymbol{\mu}^\intercal \boldsymbol{y} - \lambda \boldsymbol{y}^\intercal \boldsymbol{y} + \lambda$ 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.

PROPOSITION 2

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

(i) For problem $\max_{\boldsymbol{y} \in \mathbb{R}^n} \left\{ \boldsymbol{\mu}^\intercal \boldsymbol{y} \mid ||\boldsymbol{y}||^2 \leq 1 \right\}$, 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 $\hat{\boldsymbol{y}}_{\tilde{x}}$ with shrinkage factor $\rho = \frac{c(||\bar{\boldsymbol{x}} - \boldsymbol{x}^0||^2)}{K||\bar{\boldsymbol{x}} - \boldsymbol{x}^0||^2}$ where $0 < c(.) < 2(n-2)$ and $c'(.) \geq 0$, strictly dominates the MLE solution estimator $\hat{\boldsymbol{y}}_{\bar{x}}$ for $n \geq 3$.

Computational results confirm this proposition.

A Baysian approach

Each unknown parameter θ_i has a known *prior* distribution with density $\pi_i(\theta_i)$.

Since both the variables $\boldsymbol{x}$ and parameters $\boldsymbol{\theta}$ are random, the risk $\mathcal{R}_F(\boldsymbol{y}^*, \hat{\boldsymbol{y}})$ is also a random quantity as a function of $\boldsymbol{\theta}$.

Therefore under this framework, an *average risk* is defined as $\mathcal{R}_B(\boldsymbol{y}^*, \hat{\boldsymbol{y}}) = \mathbb{E}_{\boldsymbol{\theta}}[\mathcal{R}_F(\boldsymbol{y}^*, \hat{\boldsymbol{y}})]$ where the expectation is taken with respect to the prior distribution $\pi(\boldsymbol{\theta})$ of the random parameters.

An estimator that yields the minimum such risk is called a *Bayes estimator*.

Bayes estimator

PROPOSITION 3

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

Assume that $\boldsymbol{x} \sim \mathcal{G}(\boldsymbol{x}|\boldsymbol{\theta})$ and $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$.

Then, any estimator $\hat{\boldsymbol{y}}^*$ that solves

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

is a Bayes estimator, where the expectation is taken with respect to the conditional distribution of the parameters $\boldsymbol{\theta}$ given the observation $\boldsymbol{x}$, *i.e.*, the posterior with density $\Pi(\boldsymbol{\theta}|\boldsymbol{x})$.

Sequential estimation and optimization

The Bayesian estimator in the previous proposition integrates the estimation and optimization steps to achieve the best solution.

Now consider a situation where the estimation and optimization are not necessarily performed at the same time.

First, a Bayes estimator $\hat{\boldsymbol{\theta}}^*$ for the unknown parameter $\boldsymbol{\theta}$ is derived under the squared error loss.

Then, $\hat{\boldsymbol{\theta}}^*$ is used in place of $\boldsymbol{\theta}$ in $\max_{\boldsymbol{y} \in \mathcal{Y}} \mathbb{E}_{\boldsymbol{x}|\boldsymbol{\theta}}[f(\boldsymbol{x}, \boldsymbol{y})]$ to obtain an optimal solution.

The resulting solution is used as a solution estimator for the true optimal solution $\boldsymbol{y}^*$.

When the sequential approach is optimal

For simplicity of exposition, assume we have one observation vector $\boldsymbol{x}$ and a parameter θ_i for $i = 1, \ldots, n$.

PROPOSITION 4

Assume that the parametric objective function $\mathbb{E}_{\boldsymbol{x}|\boldsymbol{\theta}}[f(\boldsymbol{x}, \boldsymbol{y})]$ is multilinear in $\boldsymbol{\theta}$, and that for any pair (i, j), $i \neq j$ for which the product $\theta_i\theta_j$ appears in $\mathcal{F}(\boldsymbol{\theta}, \boldsymbol{y})$, we have that x_i and x_j, as well as θ_i and θ_j are independent.
Then, the estimator obtained from the sequential method of estimation followed by optimization is the Bayes estimator.

Example when the sequential approach is not optimal

Consider the portfolio selection problem, with independent r.v.s $x_i \sim \mathcal{G}_i(\mu_i, \sigma_i^2)$ where mean μ_i is known and variance σ_i^2 is unknown. Assume σ_i follows distribution π_i.

Then, the objective function is $\mathcal{F}(\boldsymbol{\sigma}, \boldsymbol{y}) = \boldsymbol{\mu}^\intercal \boldsymbol{y} - \gamma \sum_{i=1}^{n} y_i^2 \sigma_i^2$.

From Proposition 3, the Bayes estimator is obtained by taking the expectation with respect to the posterior distribution with density $\Pi(\boldsymbol{\sigma}|\boldsymbol{x})$ and then solving the problem. The resulting objective function is $\boldsymbol{\mu}^\intercal \boldsymbol{y} - \gamma \sum_{i=1}^{n} y_i^2 \, \mathbb{E}_{\sigma_i|x_i}[\sigma_i^2]$.

On the other hand for the Separate-EO method, the Bayes estimator under the squared loss function is the posterior mean $\mathbb{E}_{\sigma_i|x_i}[\sigma_i]$, and when used in the optimization problem, it yields the objective function $\boldsymbol{\mu}^\intercal \boldsymbol{y} - \gamma \sum_{i=1}^{n} y_i^2 \, \mathbb{E}_{\sigma_i|x_i}^2[\sigma_i]$.

These objectives are not equal unless $\mathbb{E}_{\sigma_i|x_i}[\sigma_i^2] - \mathbb{E}_{\sigma_i|x_i}^2[\sigma_i] = 0$ which implies that the posterior variance is zero.