A New Type of James-Stein Estimator on Cross-Sectional Regression

Kenny Zhang

Abstract

The development of James-Stein theorem has been a ground breaking result in the field of statistics. It allows people to obtain better estimators for a multivariate normal distribution under a quadratic risk function. A lot of generalizations are made on James-Stein estimators. In this paper, we present a new James-Stein type estimator on multivariate regression that shrinks the dispersion bias we defined with asymptotic good properties. It is easy to implement with an empirical Bayes nature. A frequentist view is given to show the inadmissibility of the original Ordinary Least Square estimator when the number of variables goes to infinity. We also provide a geometric view on this problem based on unit-sphere settings. The method can be particularly useful under high dimensional models like estimating betas from a multivariate regression in Capital Asset Pricing Model in finance. We present a numerical evidence from synthetic data that simulates the real world stock market and see significant improvements from the original estimator.

Acknowledgements

This work would not be possible without the support and input from my undergraduate supervisor Alexander Shkolnik from Department of Statistics and Applied Probability, University of California, Santa Barbara. I want to thank my College of Creative Studies advisor Karel Casteels for pushing me to finish this work as well as Maribel Bueuo for her guide during my undergraduate study. I also want to thank the College for giving me this opportunity to do a thesis and professors in Mathematics department as well as Statistics and Applied Probability department for their constant inspiration. Last but not least, thank all my family and friends for their love!

Contents

1 Background

Charles Stein (1956) shows the inadmissibility of the usual estimator of a multivariate normal known as the famous Stein's Paradox. Willard James and Charles Stein (1961, republished in 1992) then developed the James-Stein estimator that always improves the usual estimator under quadratic loss when $p > 4$ (number of variables). The topic of James-Stein type estimators then become an active research area in statistics and decision theory. Lindley (1962) and Efron and Moris (1973) further develops the idea with more structured algorithms including choosing the positive part of James-Stein estimator and an Empirical Bayes perspective on the estimator. Berger (1980, republished in 2013) gives a Generalized Bayes estimator to further generalize the result. Stigler (1990) gives a frequentist proof on James-Stein estimators. More recent advancement on this topic including Kim and White (2001) and Senda and Taniguchi (2006) which focuses on large samples and time correlation respectively. In this paper, we are inspired by the idea of shrinkage method in finance market (see Numerical Result section). We define a dispersion distance and shrink our Ordinary Least Square estimator towards an arbitrary given vector to obtain an optimal dispersion loss. The model is set up in a special multivariate regression setting and we provide asymptotic optimality proof for this method.

2 Multivariate Regression Model Setup

We will focus on the following multivariate regression model. Let $p \in \mathbb{N}$. Suppose we have a linear model of form

$$Y = \beta x + e$$

where $Y \in \mathbb{R}^p$, $X \in \mathbb{R}$ are observed, $\beta \in \mathbb{R}^p$ is a p-dimensional constant vector that we want to estimate and $e \in \mathbb{R}^p$ is a p-dimensional vector of random variables.

Given n observations of Y and X, consider the multiple linear regression problem with

$$\boldsymbol{Y} = \boldsymbol{X}\beta^\top + \boldsymbol{e}$$

where $\boldsymbol{Y}$ is a $n \times p$ data matrix, $\boldsymbol{X} = (X_1, ..., X_n)^\top \in \mathbb{R}^n$ is a vector of p realizations of random variable X, $\beta \in \mathbb{R}^p$ is a p-dimensional constant vector of our interest and $\boldsymbol{e}$ is a $n \times p$ matrix of n realizations of random variables e.

Assumption 1. $\text{Var}\,(X) = \sigma_x^2 < \infty$. $\mathbb{E}\,(X) = 0$(in application, mean can always been normalized).

Assumption 2. $\{e_i\}_{i=1,...p}$ are ***i.i.d*** random varianbles with $\mathbb{E}(e_i) = 0$ and $\text{Var}\,(e_i) = \sigma_e^2 < \infty$.

Assumption 3. $\text{Cov}(X, e_i) = 0$ for $i = 1, ..., p$

Under Assumption 1-3, we can perform cross-sectional regression on each entry of β and this gives our ordinary least square estimator $\hat{\beta}$ with

$$\hat{\beta}^\top = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$$

Some properties of this estimator:

- $\mathbb{E}\,\hat{\beta} = \beta$, that is, the estimator is unbiased.
- Var $\left(\hat{\beta}\right) = \frac{\sigma_\epsilon^2}{n s_x^2}$, where $s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

Moreover, the Ordinary Least Square (OLS) regression estimator is known as the Minimum Variance Unbiased Estimator (MVUE). That is, it has the minimum variance among the family of unbiased estimators. These properties are easy to find in any regression/econometric literature. We list these here since they would be helpful in section 3 and section 5. The estimator we proposed in next section in a biased estimator and we argue that it has some properties that are superior than the commonly used OLS estimator.

3 Main Result

3.1 Dispersion Bias of the OLS Estimator

The following defines a metric we called dispersion bias and states that the OLS estimator is biased with high probability in terms of dispersion under a weak assumption 4.

Definition 1. *Define the distance $d_p^2(\beta)$ and mean $\mu_p(\beta)$ of $p-$dimensional vector β as*

$$d_p^2(\beta) = \frac{1}{p}\sum_{i=1}^{p}\left(\frac{\beta_i - \mu_p(\beta)}{\mu_p(\beta)}\right)^2 \text{ and } \mu_p(\beta) = \frac{1}{p}\sum_{i=1}^{p}\beta_i \tag{1}$$

Assumption 4. $d_p^2(\beta) < \infty$ and $\mu_p(\beta) < \infty$ as $p \uparrow \infty$.

Definition 2. *Define the distance $d^2(u, v)$ of two $p-$dimensional vectors u, v as*

$$d_p^2(u, v) = \frac{1}{p}\sum_{i=1}^{p}\left(\frac{u_i}{\mu_p(u)} - \frac{v_i}{\mu_p(v)}\right)^2 \tag{2}$$

Let $q = (1, ..., 1)^\top$ be the $p-$dimension vector of all ones. Notice from the previous definition, the dispersion $d_p^2(\beta)$ is equivalent as $d_p^2(\beta, q)$.

Theorem 1 (Dispersion Bias)**.** $d_p^2(\beta, q) < d_p^2(\hat{\beta}, q)$ *with high probability in p.*

3.2 Shrinkage and Asymptotic Estimations

The following shows how much the original estimator is biased in dispersion and gives an asymptotic estimate of the constant c for a convex combination family that corrects the dispersion bias. Notice that in high dimensions (when $p \uparrow \infty$), we have an almost sure convergence result for a better estimator before we need an estimator for norm of original estimator β.

Theorem 2 (Estimate)**.**

$$\left| d_p^2(\hat{\beta}, \beta) - \frac{1}{p\mu_p^2(\hat{\beta})} \left(|\hat{\beta}|^2 - |\beta|^2 \right) \right| \to 0 \textit{ almost surely as } p \uparrow \infty.$$

Theorem 3 (Optimal Shrinkage)**.** *Define*

$$\hat{\beta}_c = c\frac{\hat{\beta}}{\mu_p(\hat{\beta})} + (1-c)\frac{q}{\mu_p(q)}, \quad \textit{for } c \in [0,1]$$

For

$$c' = \frac{d_p^2(\hat{\beta}) + d_p^2(\beta) - d_p^2(\beta, \hat{\beta})}{2d_p^2(\hat{\beta})}$$

We have $d_p^2(\beta, \hat{\beta}_{c'}) \leq d_p^2(\beta, \hat{\beta}_c)$ *for all* $c \in [0,1]$.

Theorem 4 (Asymptotic Result)**.** *Combine the previous two results, define*

$$c^* = \frac{d_p^2(\hat{\beta}) + d_p^2(\beta) - \frac{1}{p\mu_p^2(\hat{\beta})} \left(|\hat{\beta}|^2 - |\beta|^2 \right)}{2d_p^2(\hat{\beta})}$$

Consider the regime where the dimension $p \uparrow \infty$*, we have*

$$\left| \frac{d_p^2(\beta, \hat{\beta}_{c'}) - d_p^2(\beta, \hat{\beta}_{c^*})}{d_p^2(\beta, \hat{\beta}_{c'})} \right| \to 0 \textit{ almost surely as } p \uparrow \infty.$$

Theorem 5 (Estimates)**.** *Notice that given dataset* $\boldsymbol{X}, \boldsymbol{Y}$ *and derived estimator* $\hat{\beta}$*,* $d_p^2(\beta)$ *is not accessible. We will use the following estimates for dispersion of* β *on unit plane: let*

$$\ell = d_p^2(\hat{\beta}) - \textit{Var}\left(\hat{\beta}\right) = d_p^2(\hat{\beta}) - \frac{\hat{\sigma}_e^2}{ns_x^2}$$

In the regime where the dimension $p \uparrow \infty$*,* $\hat{\sigma}_e^2 \to \sigma_e^2$ *and* ℓ *converges to* $d_p^2(\beta)$ *when* n *is large (This might be off because we only have* $p \uparrow \infty$*). We also have the relationship that* $|\beta| = p\mu_p^2(\beta)(1 + d_p^2(\beta))$ *and* $\mu_p(\beta) \to \mu_p(\hat{\beta})$ *almost surely.*
Thus we can define our transform coefficient

$$\tau = \frac{d_p^2(\hat{\beta}) + \mu_p^2(\hat{\beta})\ell - \frac{1}{p\mu_p^2(\hat{\beta})} \left(|\hat{\beta}|^2 - p\mu_p(\hat{\beta})(1 - \mu_p^2(\hat{\beta})\ell) \right)}{2d_p^2(\hat{\beta})}$$

And

$$\hat{\beta}_\tau = \tau\frac{\hat{\beta}}{\mu_p(\hat{\beta})} + (1-\tau)\frac{q}{\mu_p(q)}$$

as our new estimator with improved dispersion bias on unit plane.

4 Algorithm

The following summarizes how we should apply the previous theorems on real-world data sets $\boldsymbol{X}$ and $\boldsymbol{Y}$. Typically $\boldsymbol{X}$ is the data matrix (or vector) for independent variable that we observe and $\boldsymbol{Y}$ is the data matrix for dependent variable of the result. We are interested in estimating the coefficients β.

1. Get the observations of $\boldsymbol{X}$ and $\boldsymbol{Y}$. Check the assumptions are satisfied or almost satisfied.
2. Perform Ordinary Least Square Regression with
$$\hat{\beta}^\top = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X} \boldsymbol{Y}$$
3. Find the estimate of $d_p^2(\beta)$ on unit plane by
$$d_p^2(\hat{\beta}) - \text{Var}\left(\hat{\beta}\right)$$
and denote it ℓ.
4. Use ℓ and $\hat{\beta}$ to calculate τ from
$$\tau = \frac{d_p^2(\hat{\beta}) + \mu_p^2(\hat{\beta})\ell - \frac{1}{p\mu_p^2(\hat{\beta})}\left(|\hat{\beta}|^2 - p\mu_p(\hat{\beta})(1 - \mu_p^2(\hat{\beta})\ell)\right)}{2d_p^2(\hat{\beta})}$$
5. Calculate $\hat{\beta}_\tau$ with
$$\hat{\beta}_\tau = \tau \frac{\hat{\beta}}{\mu_p(\hat{\beta})} + (1 - \tau)\frac{q}{\mu_p(q)}$$
6. Rescale it to the original mean
$$\hat{\beta}_{\text{min}} = \mu_p(\hat{\beta})\hat{\beta}_\tau$$
as the new shrinkage estimator for β that minimizes the distance of β and our estimator as defined in equation (2) in the regime where $p \uparrow \infty$.

5 Proof

5.1 Why We Need to Shrink?

For e vector of all ones, we shall prove the two lemmas on the plane:

(i) $d_p^2(\beta, q) < d_p^2(\hat{\beta}, q)$ with high probability in p.

(ii) $\lim_{p\uparrow\infty} d_p^2(\beta, \hat{\beta}) = \frac{1}{p\mu_p^2(\hat{\beta})}(|\hat{\beta}|^2 - |\beta|^2)$

Proof. First we show that $\mu_p(\hat{\beta}) \to \mu_p(\beta)$ almost surely as $p \uparrow \infty$. This follows by Chandra (1992) Theorem 6 of a version of strong law of large numbers under uniform integrability. A weaker version of convergence in probability is easy to show since $\hat{\beta}$ is an unbiased estimator for β. When we are in the regime $p \uparrow \infty$, the average is exactly the expectation. And a consequence ofthe previous result is that we have $(\frac{1}{\mu_p(\hat{\beta})} - \frac{1}{\mu_p(\beta)})^2$ converges to 0 as $p \uparrow \infty$. Then $\square$

$$
\begin{aligned}
d_p^2(\beta, \hat{\beta}) &= \frac{1}{p} \sum_{i=1}^{p} \left(\frac{\hat{\beta}_i}{\mu_p(\hat{\beta})} - \frac{\beta_i}{\mu_p(\beta)} \right)^2 \\
&= \frac{1}{p} \sum_{i=1}^{p} \left(\frac{\hat{\beta}_i}{\mu_p(\hat{\beta})} - \frac{\beta_i}{\mu_p(\hat{\beta})} + \frac{\beta_i}{\mu_p(\hat{\beta})} - \frac{\beta_i}{\mu_p(\beta)} \right)^2 \\
&= \frac{1}{p} \sum_{i=1}^{p} \left(\left(\frac{\hat{\beta}_i - \beta_i}{\mu_p(\hat{\beta})} \right)^2 + \beta_i^2 \left(\frac{1}{\mu_p(\hat{\beta})} - \frac{1}{\mu_p(\beta)} \right)^2 + 2\beta_i \left(\frac{\hat{\beta}_i - \beta_i}{\mu_p(\hat{\beta})} \right) \left(\frac{1}{\mu_p(\hat{\beta})} - \frac{1}{\mu_p(\beta)} \right) \right)
\end{aligned} \tag{3}
$$

The second and third term should goes to 0 as $p \uparrow \infty$. What we left is

$$
\frac{1}{\mu_p^2(\hat{\beta})} \frac{1}{p} \sum_{i=1}^{p} \left(\hat{\beta}_i - \beta_i \right)^2 \to d_p^2(\beta, \hat{\beta})
$$

in the regime where $p \uparrow \infty$. Expand the term out,

$$
\frac{1}{\mu_p^2(\hat{\beta})} \frac{1}{p} \sum_{i=1}^{p} \left(\hat{\beta}_i - \beta_i \right)^2 = \frac{1}{\mu_p^2(\hat{\beta})} \frac{1}{p} \sum_{i=1}^{p} \left({\hat{\beta}_i}^2 + \beta_i^2 - 2\hat{\beta}_i \beta_i \right)
$$

Notice that in OLS regression, we have the following relation:

$$
\hat{\beta}_i = \beta_i + \frac{\sum_{j=1}^{n} x_j e_{ji}}{\sum_{j=1}^{n} x_j^2}
$$

Replace it in the equation, we get

$$
\begin{aligned}
\frac{1}{\mu_p^2(\hat{\beta})} \frac{1}{p} \sum_{i=1}^{p} \left(\hat{\beta}_i - \beta_i \right)^2 &= \frac{1}{\mu_p^2(\hat{\beta})} \frac{1}{p} \sum_{i=1}^{p} \left({\hat{\beta}_i}^2 + \beta_i^2 - 2\hat{\beta}_i \beta_i \right) \\
&= \frac{1}{\mu_p^2(\hat{\beta})} \frac{1}{p} \sum_{i=1}^{p} \left({\hat{\beta}_i}^2 + \beta_i^2 - 2 \left(\beta_i + \frac{\sum_{j=1}^{n} x_j e_{ji}}{\sum_{j=1}^{n} x_j^2} \right) \beta_i \right) \\
&= \frac{1}{\mu_p^2(\hat{\beta})} \frac{1}{p} \sum_{i=1}^{p} \left({\hat{\beta}_i}^2 - \beta_i^2 - 2 \frac{\sum_{j=1}^{n} x_j e_{ji}}{\sum_{j=1}^{n} x_j^2} \beta_i \right)
\end{aligned} \tag{4}
$$

We shall show that the last term vanishes so that

$$\frac{1}{\mu_p^2(\hat{\beta})}\frac{1}{p}\sum_{i=1}^{p}\left(\hat{\beta}_i{}^2-\beta_i^2\right)\xrightarrow{a.s.}d_p^2(\beta,\hat{\beta})$$

On the other hand we have

$$\begin{aligned}\frac{1}{p\mu_{\hat{\beta}}^2}(|\hat{\beta}|^2-|\beta|^2)&=\frac{1}{p\mu_{\hat{\beta}}^2}\sum_{i=1}^{p}\hat{\beta}_i^2-\frac{1}{p\mu_{\hat{\beta}}^2}\sum_{i=1}^{p}\beta_i^2\\&=\frac{1}{p\mu_{\hat{\beta}}^2}\sum_{i=1}^{p}\left(\hat{\beta}_i{}^2-\beta_i^2\right)\end{aligned}\tag{5}$$

Notice that $|\beta|=p\mu_p^2(\beta)(1+d_p^2(\beta))$ and $|\hat{\beta}|=p\mu_p^2(\hat{\beta})(1+d_p^2(\hat{\beta}))$, we can also rewrite the term as

$$\begin{aligned}\frac{1}{p\mu_{\hat{\beta}}^2}(|\hat{\beta}|^2-|\beta|^2)&=\frac{1}{p\mu_{\hat{\beta}}^2}\left(p\mu_p^2(\hat{\beta})(1+d_p^2(\hat{\beta}))-p\mu_p^2(\beta)(1+d_p^2(\beta))\right)\\&=\frac{1}{\mu_{\hat{\beta}}^2}\left(\mu_p^2(\hat{\beta})(1+d_p^2(\hat{\beta}))-\mu_p^2(\beta)(1+d_p^2(\beta))\right)\\&=\frac{1}{\mu_{\hat{\beta}}^2}\left(\mu_p^2(\hat{\beta})-\mu_p^2(\beta)\right)+\frac{1}{\mu_{\hat{\beta}}^2}\left(\mu_p^2(\hat{\beta})d_p^2(\hat{\beta})-\mu_p^2(\beta)d_p^2(\beta)\right)\end{aligned}\tag{6}$$

With $\mu_p(\beta)\xrightarrow{a.s.}\mu_p(\beta)$, it is not hard to show that the left hand side converges to $d_p^2(\hat{\beta})-d_p^2(\beta)$. The first term disappears and $\frac{\mu_p(\beta)}{\mu_p(\hat{\beta})}$ goes to 1. With that said, since left hand side converges to $d_p^2(\beta,\hat{\beta})$ almost surely and is non-negative by definition, we know $d_p^2(\hat{\beta})>d_p^2(\beta)$ with high probability. These previous lemmas show that the OLS estimator is likely to be over-dispersed in this setting and gives an estimate for the dispersion bias between β and $\hat{\beta}$.

5.2 How Much We Need to Shrink?

In the next sections we shall show the following:

(i) For $c=\frac{d_p^2(\hat{\beta})+d_p^2(\beta)-d_p^2(\beta,\hat{\beta})}{2d_p^2(\hat{\beta})}$, the convex combination of the unit-mean estimator gives the smallest dispersion among the family of estimators for $c\in[0,1]$.

(ii) $d_p^2(\hat{\beta})-\frac{\hat{\sigma}_e^2}{ns_x^2}$ gives a good estimator of $d_p^2(\beta)$ on unit plane.

The first lemma essentially provides the extent that we need to adjust our OLS estimator in order to achieve optimal dispersion bias. And the estimate from the second lemma as well as previous result provide us a practical quantity to perform the optimization.

Proof. The first lemma is from taking derivative with respect to c of $d_p^2(\beta, \hat{\beta}_c)$ and setting it to zero. The calculation here is rather long and tedious, but the idea is to find a local minimum of the form by taking c as a variable. From the previous result, it is also easy to see that the result we get $0 < c < 1$ with high probability.
The second follows from an idea in Shalizi (2015) where we see that

$$\begin{aligned} \text{Var}\left(\hat{\beta}\right) &= \text{Var}\left(\beta + \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})e_i}{s_x^2}\right) \\ &= \text{Var}\left(\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})e_i}{s_x^2}\right) \\ &= \frac{\frac{1}{n^2}\sum_{i=1}^{n}(x_i - \bar{x})^2 \text{Var}\left(e_i\right)}{(s_x^2)^2} \\ &= \frac{\sigma_e^2}{n s_x^2} \end{aligned} \tag{7}$$

Essentially the major difference between variance and dispersion on unit mean plane is that known constant vector has 0 variance but a positive dispersion. From Assumption 1, $ns_x^2 = \sum_{i=1}^{n} x^2$. Notice when $p \uparrow \infty$, the sample variance of e also converges to σ_e^2 almost surely. Notice that $\mu_p^2(\hat{\beta})\left(\text{Var}\left(\beta\right) - \frac{\hat{\sigma}_e^2}{n s_x^2}\right)$ is an unbiased estimator of $d_p^2(\beta)$ and it is accessible from the data and should be reasonablely accurate when n is not very small ($n > 10$). □

6 An Equivalent Transformation on Unit Sphere

Notice that we estimate the norm of β at the last step of our transformation. Consider that if we know the norm of β or estimate of norm at the first place, we could normalize the problem so that β lies on a unit-sphere and this provides useful visualization of the approach that we are taken as well as another justification of the transformation. This is an equivalent approach as before as we can see in appendix. When the norm of β is accurate, the result provides an almost sure convergence when dimension is high.

6.1 New Model

Now let's consider a slightly modified model. Let $p \in \mathbb{N}$. Suppose we have a linear model of form

$$Y = \beta^* x + e \tag{6.1.0}$$

where $Y \in \mathbb{R}^p$, $x \in \mathbb{R}$ are observed, $\beta^* \in \mathbb{R}^p$ is a constant vector that we want to estimate and $e \in \mathbb{R}^p$ is a p-dimensional vector of random variables.
Given n observations of Y and x, consider the multiple linear regression problem with

$$\boldsymbol{Y} = \boldsymbol{X}^*(\beta^*)^\top + \boldsymbol{e} \tag{6.1.1}$$

where $\boldsymbol{Y}$ is a $n \times p$ data matrix, $\boldsymbol{X}^* = (X_1^*, ..., X_n^*)^\top \in \mathbb{R}^n$ is random variable with mean 0, $\beta^* \in \mathbb{R}^p$ is a p-dimensional constant row vector of our interest and $\boldsymbol{e}$ is a $n \times p$ matrix of random variables. Similarly We assume that the entries of $\boldsymbol{e}$ are mutually independent and $\mathbb{E}(e_{ij}) = 0$ and $\text{Var}\,(e_{ij}) = \sigma_e^2$ for all i, j. Suppose that $\text{Var}\,(X_i^*) = \sigma^2 < \infty$ and we can estimate norm of β^*, then we can write the model equivalently as

$$\boldsymbol{Y} = (\boldsymbol{X}^*|\beta^*|)\frac{(\beta^*)^\top}{|\beta^*|} + \boldsymbol{e} \tag{6.1.2}$$

Now let $\beta = \frac{\beta^*}{|\beta^*|}$ and $\boldsymbol{X} = (X_1, ..., X_n)^\top = |\beta^*|\boldsymbol{X}^*$. We can re-write equation $6.1.2$ as

$$\boldsymbol{Y} = \boldsymbol{X}\beta^\top + \boldsymbol{e} \tag{6.1.3}$$

By construction, we have $|\beta| = 1$. And $\text{Var}\,(\boldsymbol{X}) = |\beta^*|^2\text{Var}\,(X_i^*)$. From the diffuse vector property, $|\beta^*|^2$ has order p. So we can write $\text{Var}\,(\boldsymbol{X}) = p\sigma^2$ for some finite σ^2. Furthermore, we assume that $\text{Cov}(X_k, e_{ij}) = 0$ for all k, i, j for exogeneity.
We perform cross-sectional regression on each of the p $\beta-$values. We get our Ordinary Least Square estimator

$$\hat{\beta}^\top = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y} \tag{6.1.4}$$

Plug in $\boldsymbol{Y}$ with equation $6.1.3$,

$$\hat{\beta}^\top = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{X}\beta^\top + \boldsymbol{e}) \tag{6.1.5}$$

Simplify, we get

$$\hat{\beta}^\top = \beta^\top + (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{e} \tag{6.1.6}$$

6.2 A New Metric of Cosine Similarity

We will define the cosine similarity as a new metric we use on the unit-sphere problem. Let $q = \frac{1}{\sqrt{p}}(1, ..., 1)^\top \in \mathbb{R}^p$, that is, a $p-$dimensional row vector of all ones multiply by a factor of order $\frac{1}{\sqrt{p}}$. Define $\gamma_{xy} = \cos\theta_{xy}$, where $\cos\theta_{xy}$ denotes the cosine value of angle between $p-$dimensional vectors x and y. Notice that $\cos\theta_{xy} = \frac{x\cdot y}{|x||y|}$ for any $p-$dimensional vector x and y where $|x|$ and $|y|$ denotes the euclidean norm of vectors x and y. Equivalently, we can say $\gamma_{xy} = \frac{x\cdot y}{|x||y|} = \frac{xy^\top}{|x||y|}$.
By construction in preliminaries, from ordinary least square regression we have

$$\hat{\beta}^\top = \beta^\top + (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{e} \tag{6.2.0}$$

We assume that $|\beta| = 1$ and $\text{Var}(X) = p\sigma^2$. We will show the following two lemmas hold:

1. $\gamma_{\hat{\beta}q} = \frac{1}{|\hat{\beta}|}\gamma_{\beta q} + \delta_p$ where $\delta_p \to 0$ as $p \to \infty$.

2. $|\hat{\beta}| > 1$ with high probability.

Proof. (i) We multiply q from both sides of equation $(6.2.0)$, since matrix multiplication is associative, we get

$$\hat{\beta}^\top q = \beta^\top q + (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{e}q \tag{6.2.1}$$

Since $|\beta| = 1$ and $|q| = 1$, $\beta^\top q = \gamma_{\beta q}$ and $\hat{\beta}^\top q = \gamma_{\hat{\beta} q}|\hat{\beta}|$. Plug in we get

$$\gamma_{\hat{\beta} q}|\hat{\beta}| = \gamma_{\beta q} + (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{e}q \tag{6.2.2}$$

$|\hat{\beta}| > 0$ since $\hat{\beta}$ is not zero vector (we assume dispersion is non-zero, otherwise the problem is trivial). Divide $|\hat{\beta}|$ from both sides,

$$\gamma_{\hat{\beta} q} = \frac{1}{|\hat{\beta}|}\gamma_{\beta q} + \frac{1}{|\hat{\beta}|}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{e}q \tag{6.2.3}$$

Now we define $\delta_p = \frac{1}{|\hat{\beta}|}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{e}q$ and show that it goes to 0 as $p \to \infty$. Since $X^\top X \approx n\operatorname{Var}(X) = np\sigma^2$,

$$\delta_p = \frac{1}{|\hat{\beta}|\sigma^2}(\frac{1}{p}\boldsymbol{X}^\top \boldsymbol{e}q) \tag{6.2.4}$$

As we will see in lemma 2, $|\hat{\beta}|$ goes to some constant that does not depend on p as $p \to \infty$. Essentially, we want to show $\frac{1}{p}\boldsymbol{X}^\top eq$ vanishes. If we multiply $\frac{1}{\sqrt{(p)}}$ in q into $\boldsymbol{X}^\top$, $\boldsymbol{X}^\top$ will have finite variance. Let $\boldsymbol{X}' = \frac{1}{\sqrt{p}}\boldsymbol{X} = (x_1', ..., x_n')^\top$ with finite variance. And e has entries $\{e_{ij}\}$, $i = 1, ..., n$, $j = 1, ..., p$. We can write the product explicitly. Consider $M = \frac{1}{p}\boldsymbol{e}q$, j^{th} entry M_j of this vector is given by $\frac{1}{p}\sum_{i=1}^{p} e_{ji}$. Since e_{ij} are all independent with mean 0, by law of large numbers, $M_j \to 0$ as $p \to \infty$. And $\boldsymbol{X}'M = \sum_{j=1}^{n} x_j' M_j$. Since $(x_1', ..., x_n') = \frac{1}{\sqrt{p}}\boldsymbol{X}^\top$ has finite variance and are independent from M_j, a finite 0 linear combination of $\{x_i\}'s$ must be zero. Thus, $\frac{1}{|\hat{\beta}|\sigma^2}(\frac{1}{p}\boldsymbol{X}^\top \boldsymbol{e}q)$ vanishes as $p \to \infty$

In fact, this is a special case when $q = \beta$ as we will prove in lemma 3. (Lemma 3 states that $\gamma_{\hat{\beta}\beta} = \frac{1}{|\hat{\beta}|} + \delta_p$ with δ_p vanishes as $p \to \infty$). □

Proof. (ii) We want to show $|\hat{\beta}| > 1$ with high probability especially when $p \to \infty$. Again we start from equation $(6.2.0)$.

$$\hat{\beta}^\top = \beta^\top + (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{e} \tag{6.2.0}$$

Multiply $\hat{\beta}$ from both sides,

$$|\hat{\beta}|^2 = (\beta + (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{e})\hat{\beta} \tag{6.2.5}$$

Since $\hat{\beta} = \beta + \boldsymbol{e}^\top \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$, (here $(X^\top X)^{-1}$ is a 1×1 number, its transpose is itself). We multiply the terms out,

$$|\hat{\beta}|^2 = (\beta^\top + (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{e})(\beta + \boldsymbol{e}^\top \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}) = |\beta|^2 + A + B + C \tag{6.2.6}$$

where $A = \beta^\top \boldsymbol{e}^\top \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$, $B = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{e}\beta$ and $C = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{e}\boldsymbol{e}^\top \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$. Notice that

$$C = \frac{\boldsymbol{X}^\top \boldsymbol{e}\boldsymbol{e}^\top \boldsymbol{X}}{(\boldsymbol{X}^\top \boldsymbol{X})^2} = \frac{|\boldsymbol{e}^\top \boldsymbol{X}|^2}{|\boldsymbol{X}|^4} > 0 \tag{6.2.7}$$

First we show that C does not vanish. Notice that we have $|\boldsymbol{X}|^2 = \boldsymbol{X}^\top \boldsymbol{X} \approx np\sigma^2$. Thus, $|\boldsymbol{X}|^4 \approx n^2p^2\sigma^4$. $\boldsymbol{e}^\top \boldsymbol{X}$ gives a $p \times 1$ vector with i^{th} entry $(\boldsymbol{e}^\top \boldsymbol{X})_i$. Notice $(e^\top X)_i = \sum_{j=1}^n e_{ji}X_j$. Then we have

$$C = \frac{|\boldsymbol{e}^\top \boldsymbol{X}|^2}{|\boldsymbol{X}|^4} = \frac{1}{n^2\sigma^4 p}(\frac{1}{p}\sum_{i=1}^{p}(\boldsymbol{e}^\top \boldsymbol{X})_i^2) \tag{6.2.8}$$

As $p \to \infty$, $(\boldsymbol{e}^\top \boldsymbol{X})_i$ are random variables with finite moments, we can also show that $(e^\top X)_i$ are uncorrelated identical distributions, by law of large numbers, this converges to $\mathbb{E}((\boldsymbol{e}^\top \boldsymbol{X})_i^2)$. By construction since all e_{ij} and X_i are uncorrelated, mean 0, the expected value goes to $n\text{Var}(X)\,\text{Var}(e) \approx n^2p\sigma^2\sigma_e^2$. Thus, $C \approx \frac{\sigma_e^2}{\sigma^2}$ is finite and does not vanish.
Notice that by construction $|\beta|^2 = |\beta| = 1$. We will now show A, B vanishes as $p \to \infty$. In fact, $A = B^\top$ and since they are both just numbers, we will only show B goes to 0. We can simplify B as:

$$(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{e}\beta = \frac{1}{p\sigma^2}\boldsymbol{X}^\top \boldsymbol{e}\beta \tag{6.2.9}$$

This term vanishes by Komogorov law of large numbers (Chandra 1992). □

6.3 Transformation

A similar transformation is defined by rotation on unit-sphere. Define $\beta_t = \frac{\frac{\hat{\beta}}{|\hat{\beta}|}+tq}{|\frac{\hat{\beta}}{|\hat{\beta}|}+tq|}$ where $t \in [0, \infty)$. That is, a $t-$parametrized vector on unit sphere that lies in the plane of q and $\hat{\beta}$. Notice that when $t = 0$, $\beta_t = \frac{\hat{\beta}}{|\hat{\beta}|}$ and when $t = 1$, $\beta_t = q$. We will show in the following section that $t^* = \frac{\gamma_{\beta q} - \gamma_{\hat{\beta}\beta}\gamma_{\hat{\beta}q}}{\gamma_{\beta\hat{\beta}} - \gamma_{\beta q}\gamma_{\hat{\beta}q}}$ gives the minimum angle between β and our transformed estimator. (This calculation of taking derivatives is not in section 5 so we will write it here since it is a slightly more complicated version.)

Since both vectors lies on unit sphere, we have $\cos\theta_{\beta_t\beta} = \beta_t^\top \beta$. We want to minimize the angle between β_t and β. That is the same as maximize the cosine value $\beta_t^\top \beta$. By our definition,

$$\beta_t^\top \beta = \frac{\frac{\hat{\beta}^\top \beta}{|\hat{\beta}|} + tq^\top \beta}{|\frac{\hat{\beta}}{|\hat{\beta}|} + tq|} = \frac{\gamma_{\hat{\beta}\beta} + t\gamma_{\beta q}}{|\frac{\hat{\beta}}{|\hat{\beta}|} + tq|} \tag{6.3.0}$$

Denote the demoninator $|\frac{\hat{\beta}}{|\hat{\beta}|} + tq|$ as ℓ_t, we take derivative of $\beta_t^\top \beta$, by quotient rule we have

$$\frac{\partial}{\partial t}(\beta_t^\top \beta) = \frac{\gamma_{\beta q}\ell_t}{\ell_t^2} - \frac{\gamma_{\hat{\beta}\beta} + t\gamma_{\beta q}}{\ell_t^2}\frac{\partial}{\partial t}\ell_t \tag{6.3.1}$$

We first simplify $\frac{\partial}{\partial t}\ell_t$, notice that

$$\ell_t = |\frac{\hat{\beta}}{|\hat{\beta}|} + tq| = \sqrt{(\frac{\hat{\beta}}{|\hat{\beta}|} + tq)^\top (\frac{\hat{\beta}}{|\hat{\beta}|} + tq)} = \sqrt{(\frac{\hat{\beta}^\top}{|\hat{\beta}|} + tq^\top)(\frac{\hat{\beta}}{|\hat{\beta}|} + tq)} \tag{6.3.2}$$

Multiply the terms out, notice q, $\frac{\hat{\beta}}{|\hat{\beta}|}$ lies on unit sphere, $q^\top q = 1$ and $\frac{\hat{\beta}^\top \hat{\beta}}{|\hat{\beta}|^2} = 1$.

$$\ell_t = \sqrt{1 + \frac{2t\hat{\beta}^\top q}{|\hat{\beta}|} + t^2} = \sqrt{1 + 2t\gamma_{\hat{\beta}q} + t^2} \tag{6.3.3}$$

Now we apply chain rule to find the derivative of ℓ_t

$$\frac{\partial}{\partial t}\ell_t = \frac{1}{2}\frac{2\gamma_{\hat{\beta}q} + 2t}{\ell_t} = \frac{\gamma_{\hat{\beta}q} + t}{\ell_t} \tag{6.3.4}$$

Plug into $6.3.1$ and set the derivative to 0, we have

$$\frac{\partial}{\partial t}(\beta_t^\top \beta) = \frac{\gamma_{\beta q}}{\ell_t} - \frac{(\gamma_{\hat{\beta}\beta} + t\gamma_{\beta q})(\gamma_{\hat{\beta}q} + t)}{\ell_t^3} = 0 \tag{6.3.5}$$

Use the result in $6.3.3$, we get

$$\gamma_{\beta q}(1 + 2t\gamma_{\hat{\beta}q} + t^2) = (\gamma_{\hat{\beta}\beta} + t\gamma_{\beta q})(\gamma_{\hat{\beta}q} + t) \tag{6.3.6}$$

Notice that the t^2 term cancels out and indeed we can just solve for t,

$$t\gamma_{\hat{\beta}q}\gamma_{\beta q} - t\gamma_{\hat{\beta}\beta} = \gamma_{\beta\hat{\beta}}\gamma_{\hat{\beta}q} - \gamma_{\beta q} \Rightarrow t = \frac{\gamma_{\hat{\beta}\beta}\gamma_{\hat{\beta}q} - \gamma_{\beta q}}{\gamma_{\hat{\beta}q}\gamma_{\beta q} - \gamma_{\beta\hat{\beta}}} \tag{6.3.7}$$

Thus, we get our result as the local minimum for the angle achieves at

$$t^* = \frac{\gamma_{\hat{\beta}\beta}\gamma_{\hat{\beta}q} - \gamma_{\beta q}}{\gamma_{\hat{\beta}q}\gamma_{\beta q} - \gamma_{\beta\hat{\beta}}} \tag{6.3.8}$$

6.4 Estimate

Notice that β vector is not accessible, we will need an estimate for $\gamma_{\hat{\beta}\beta}$. The following theorem gives an almost sure convergent estimate for asymptotic scenarios when $p \uparrow \infty$.

Theorem 6 (Angle). *$\gamma_{\hat{\beta}\beta} = \frac{1}{|\hat{\beta}|} + \delta_p$ where δ_p vanishes as $p \uparrow \infty$.*

Proof. First we multiply β from both sides of equation (6.1.0), we get

$$\hat{\beta}^\top \beta = \beta^\top \beta + (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{e}\beta \tag{6.4.0}$$

Notice that $|\beta| = 1$, $\beta^\top \beta = |\beta|^2 = 1$, we have

$$\gamma_{\hat{\beta}\beta}|\hat{\beta}| = 1 + (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{e}\beta \tag{6.4.1}$$

Thus, our $\delta_p = \frac{1}{|\hat{\beta}|}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{e}\beta$. Notice $\boldsymbol{X}^\top \boldsymbol{X} \approx pn\sigma^2$. We want to show $\frac{1}{p}\boldsymbol{X}^\top \boldsymbol{e}\beta$ goes to zero as $p \uparrow \infty$. Notice that

$$\frac{1}{p}\boldsymbol{X}^\top \boldsymbol{e}\beta = \frac{1}{p}\sum_{i=1}^{p}(\boldsymbol{X}^\top \boldsymbol{e})_i \beta_i \tag{6.4.2}$$

where β_i is the i^{th} entry of constant vector β and $(\boldsymbol{X}^\top \boldsymbol{e})_i = \sum_{j=1}^{n} e_{ji}X_j$ and $(\boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{e})_i$ are uncorrelated. Again by Kolmogorov strong law of large numbers, this term goes away according to our assumptions (Chandra 1992). □

We will discuss the estimators for the cosine angles in equation 6.3.8 about t^*. First, q is given by our definition and $\hat{\beta}$ is our preliminary Ordinary Least Square estimator. We only need to find estimators for $\gamma_{\hat{\beta}\beta}$ and $\gamma_{\beta q}$. When p is large, as we discussed in section 6.4, $\gamma_{\hat{\beta}\beta}$ can be approached by $\frac{1}{|\hat{\beta}|}$. And in section 2, the first lemma gives $\gamma_{\hat{\beta}q} \approx \frac{1}{|\hat{\beta}|}\gamma_{\beta q}$ when p large. Thus, we can use $|\hat{\beta}|\gamma_{\hat{\beta}q}$ to estimate $\gamma_{\beta q}$. This gives us an accessible estimator purely from the data. It converges asymptotically to the most efficient estimator in the family when $p \uparrow \infty$. Again this is based on an accurate estimate of norm of β. There are various literature in estimating norm of regression coefficients for users to use in real-world applications.

6.5 Geometric View

As we discussed in the beginning of the section, one of the reason that we use this unit-sphere construction is that it gives an easy geometric perspective of our shrinkage.

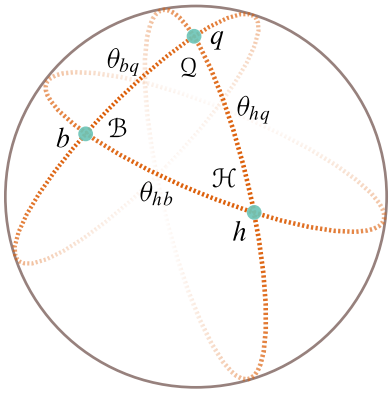

Figure 1: Geometric Representation of the Shrinkage

The above is an image from my supervisor's paper (Goldberg 2018). As we see on unit sphere, we are moving our estimator h to any arbitrary constant vector b by the correction vector q. We choose the vector of all ones as q since it is standard for James-Stein estimator but in fact it can be anything on the sphere and could give a shrinkage effect.

7 Numerical Results

7.1 Background and Model Setup

We take the Sharpe (1964) Capital Asset Pricing Model as a real life example for numerical experiment. The Capital Asset Pricing Model states that the expected rate of return of an asset is proportional to the expected market return with slope coefficient beta. In this case, Ordinary Least Squares regression (OLS), or linear regression is used to obtain an estimation of betas of market assets. In Vasieck's paper "A Note on Using Cross-Sectional Information in Bayesian Estimation Of Security Betas" in 1973, he mentioned a linearly adjusted transformation used in Security Risk Evaluation service by Merrill Lynch, Pierce, Fenner & Smith, Inc. of the form

$$b' = 1 + k(b - 1)$$

where k is a constant for all stocks and b is our original OLS estimation. Notice that this is similar to the James-Stein type estimator that we are using. In fact, this model has an empirical Bayes interpretation. It is also the case that the number of stocks is very large compared to number of observations.

Assume that the stock market follows the Capital Asset Pricing Model such that for each security i we have

$$y_t^i = \beta_i x_t + e_t^i, \quad t = 1, 2, ..., T$$

where y_t^i, $t = 1, 2, ..., T$ are rates of return on the security i and x_t, $t = 1, 2, ..., T$ are returns on a market index. e_t^i, $t = 1, 2, ..., T$ are specific returns of security i that satisfies $Ee_t^i = 0$ and independent for all t, i. The variance of e_t^i is a value uniformly picked between 0.0004 and 0.0016. We set number of stocks to be 500. T (number of trade days) to be 256. Now we can generate our data according to parameters of the real market. We generate our true betas (i.e. β) using normal distribution with mean 1 and standard deviation 0.32. Next, we generate the market indexes x_t using independent identical normal distribution with mean 0 and standard deviation 0.01 for each t. By our assumption, the specific returns are generated using normal distribution with mean 0, variance uniformly picked from 0.0004 to 0.0016, independent and identical for each t and each security i. Now that we can calculate $y_t^i = \beta_i x_t + e_t^i$ for $t = 1, 2, ..., 256$ and $i = 1, 2, .., 500$. By the Ordinary Least Squares estimation, our estimation for β_i is given by

$$b_i = \sum_t (y_t^i - \bar{y_i})(x_t - \bar{x}) \Big/ \sum_t (x_t - \bar{x})^2 \tag{8}$$

Now that we have an estimation for beta of each stock, we apply the transformation of form

$$b' = 1 + k(b - 1), \quad k \in [0, 1] \tag{9}$$

that adjust b towards the unit vector. Notice that when $k = 1$, $b' = b$ which gives our original OLS estimation. When $k = 0$, $b' = [1, ..., 1]$ gives a constant estimation. Smaller k reduces the variance of our estimation towards 1 and k contains the information about the slope of cross-sectional regression of beta estimates.

7.2 Numerical Experiments

Now fix β and repeat the process 100 times (that means we generate new y_t^i and fit the model 100 times, calculate the angles using different transformation values k). For the angles we get from 100 simulations, we get a boxplot for each k ranging from 0 to 1 with different cosine angles in radians. Figure 1 shows the plot of our result. As we can see, for $k = 0$, the constant estimation has largest angle, meaning it is not a good estimation. The trend also shows that there is a local minimum between $k = 0.6$ and $k = 0.8$, which gives the best estimation from our perspective. This implies that the Ordinary Least Squares regression is indeed not the most optimal estimation and we can make it better by applying a linearly adjusted transformation towards unity.

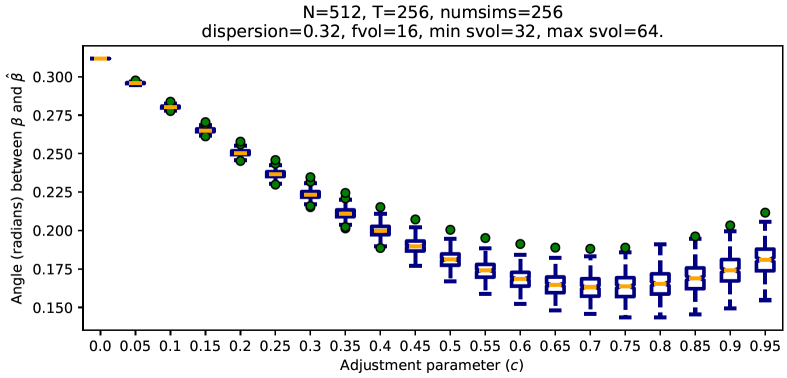

Figure 2: Simulation Results when variance of true beta is 0.25

Now we will use our method. Notice that the actual β is not accessible, so we cannot just iterate all c values like we previously did. We use a setting of $n = 50$ and $p = 1000$. To simplify, the variance of e is now set to be 0.03^2 instead of uniformly chosen from an interval. The other settings are same as the previous data generating process. We only observe data $\boldsymbol{X}$ and $\boldsymbol{Y}$. We follow the algorithm and repeat 30 times. The plot of dispersion of different estimators is shown below.

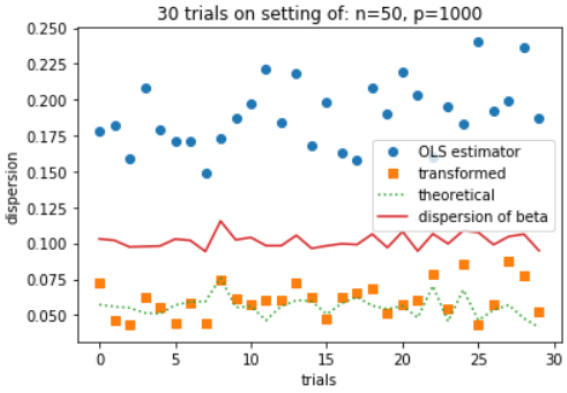

Figure 3: An experiment on self-generated financial market data.

As we can see, our transformed estimator has dispersion very close to the theoretical best estimator in the class. The latter uses the information of true betas whereas our transformed estimator is purely data-driven. It improves the accuracy of the OLS estimator and has low dispersion. This can be particularly useful in portfolio optimization and to other financial interests.

8 Future Works

We will discuss some future works and potential development of this method. One important aspect that we have mentioned previously is to estimate the norm of β. We proposed an unbiased estimator that assumes we have no prior information about the vector β. In different cases this might be improved. Some problems have known dispersion of β this can be solved very easily (see Appendix). When n is small, there's also literature on how to get an efficient estimator for norm of β. Secondly, the target vector we shrink to here is vector of all ones or the ground mean of β. Theoretically any vector can be used to shrink the OLS estimator and with $(p+1)$ arbitrary vectors that spans the vector space, we might have a guaranteed optimal shrinkage. The problem is more complicated geometrically and computationally but it will give more flexibility to the model assumption. In addition, a lot of recent works focus on estimating prior for Empirical Bayes. Two major categories including "g-modeling" (Efron, 2014) and "f-modeling". As well as a more geometric approach (Wager, 2014). Works on James-Stein and time series related research are also discussed partially in appendix including (Senda, 2006). Last but not least, the rate of convergence is also a topic of our interest. How fast does our estimator converges for different p and n settings might be an interesting problem to explore. In some models (Dwivedi 2018) this has been studied with nice results.

9 Appendix

9.1 Formulas

We will list some useful formulas here. A common transformation from mean, dispersion and norm: For any $p-$dimensional vector β , we have

$$|\beta|^2 = p\mu_p^2(\beta)(1 + d_p^2(\beta)) \tag{10}$$

So any two of the three quantities determines the other.

The following formulas provide a conversion from unit-mean plane to unit-sphere construction: For any $p-$dimensional vectors x, y, we have

$$d_p^2(x, y) = (\mu_p^2(x)+d_p^2(x))+(\mu_p^2(y)+d_p^2(y))-2\gamma_{xy}\sqrt{p(\mu_p^2(x) + d_p^2(x))}\sqrt{p(\mu_p^2(y) + d_p^2(y))} \tag{11}$$

And similarly,

$$\gamma_{xy} = \frac{d_p^2(x) + d_p^2(y) - d_p^2(x, y) + \mu_p^2(x) + \mu_p^2(y)}{2\sqrt{\mu_p^2(x) + d_p^2(x)}\sqrt{\mu_p^2(y) + d_p^2(y)}} \tag{12}$$

One special case is when we have vector of all ones $q = (1, ..., 1)^\top \in \mathbb{R}^p$,

$$\gamma_{xq} = \frac{\mu_p(x)}{\sqrt{d_p^2(x) + \mu_p^2(x)}} \tag{13}$$

It is thus easy to derive a bijective relationship between our optimal constants:

$$c' = \frac{d_p^2(\hat{\beta}) + d_p^2(\beta) - d_p^2(\beta, \hat{\beta})}{2d_p^2(\hat{\beta})} \quad \text{and} \quad t^* = \frac{\gamma_{\hat{\beta}\beta}\gamma_{\hat{\beta}q} - \gamma_{\beta q}}{\gamma_{\hat{\beta}q}\gamma_{\beta q} - \gamma_{\beta\hat{\beta}}}$$

9.2 Some Derivations

We can also derive the unit plane lemmas from unit-sphere expressions. The following lemmas on unit-plane follows the sphere construction
(i) $\gamma_{\hat{\beta}e} = \frac{|\beta|}{|\hat{\beta}|}\gamma_{\beta e} + \delta_p$ where $\delta_p \to 0$ as $p \to \infty$.
(ii) $\gamma_{\hat{\beta}\beta} = \frac{|\beta|}{|\hat{\beta}|} + \delta_p$ where $\delta_p \to 0$ as $p \to \infty$.

Therefore for $p \uparrow \infty$, we have

$$\begin{aligned} d_p^2(\beta, \hat{\beta}) &= \left(\frac{1}{\gamma_{\beta q}^2} + \frac{1}{\gamma_{\hat{\beta} q}^2} - 2\frac{\gamma_{\beta\hat{\beta}}}{\gamma_{\beta q}\gamma_{\hat{\beta} q}} \right) \\ &= \left(\frac{|\beta|^2}{|\hat{\beta}|^2\gamma_{\hat{\beta} q}^2} + \frac{1}{\gamma_{\hat{\beta} q}^2} - 2\frac{|\beta|^2}{|\hat{\beta}|^2\gamma_{\hat{\beta} q}^2} \right) \\ &= \left(\frac{1}{\gamma_{\hat{\beta} q}^2} - \frac{|\beta|^2}{|\hat{\beta}|^2\gamma_{\hat{\beta} q}^2} \right) \end{aligned} \tag{14}$$

Notice

$$\gamma_{\hat{\beta} q} = \frac{\hat{\beta} \cdot q}{|\hat{\beta}||q|} = \frac{\sum \hat{\beta}}{|\hat{\beta}|\sqrt{p}}$$

So

$$d_p^2(\beta, \hat{\beta}) = p|\hat{\beta}|^2 \left(1 - \frac{|\beta|^2}{|\hat{\beta}|^2} \right) \Big/ \left(\sum \hat{\beta} \right)^2 \tag{15}$$

Notice that $\sum_{i=1}^{p} \hat{\beta} = n\mu_{\hat{\beta}}$ and $|\beta|^2 = p\mu_\beta^2(1 + d_\beta^2)$, $|\hat{\beta}|^2 = p\mu_{\hat{\beta}}^2(1 + d_{\hat{\beta}}^2)$. We can simplify the previous equation

$$\begin{aligned} d_p^2(\beta, \hat{\beta}) &= p|\hat{\beta}|^2 \left(1 - \frac{|\beta|^2}{|\hat{\beta}|^2} \right) \Big/ \left(\sum \hat{\beta} \right)^2 \\ &= p|\hat{\beta}|^2 - p\mu_{\hat{\beta}}^2|\beta|^2 \Big/ \left(p\mu_{\hat{\beta}} \right)^2 \\ &= \frac{1}{p\mu_{\hat{\beta}}^2}(|\hat{\beta}|^2 - |\beta|^2) \end{aligned} \tag{16}$$

9.3 Some Other Discussions

As we mentioned in the last section of the paper, some of the works on James-Stein estimators that are different from our approach yet related to the problem we are trying to solve are discussed here. We will discuss the following two papers.

9.3.1 Kim, T. H., White, H. (2001). James-Stein-type estimators in large samples with application to the least absolute deviations estimator.

Modern statistics have focused on large sample data set and high-dimensional analysis. One of the main aspects of CAPM beta estimation is that we have large sample sizes for number of stocks and often short observations over a period of time. The original risk improvement of James-Stein

estimator disappears under large sample size. This paper focuses on improving the asymptotic risk using James-Stein type estimator and generalizes the James-Stein type of estimator to a large sample size. The technique in the paper uses a combination of least absolute deviations (LAD) estimator and the least square (LS) estimator. The paper provides rigorous mathematical assumption and proof for finding the optimal estimator. It also applies the method using a combination of data-dependent OLS estimator and base LAD estimator to show how it works. A Monte Carlo experiment is performed to test the new JS-type estimator. There are a lot of candidates for the data-dependent points used in the shrinkage method so the method might be able to work in a fairly general setting.

9.3.2 Senda, M., Taniguchi, M. (2006). James–Stein estimators for time series regression models.

This paper purposes a time series version of James-Stein estimator under residues generated by Gaussian stationary process. It gives the inadmissibility of the estimator under some assumptions. We will need finite dispersion and non-singular autocovariance matrix in order to perform the method. The paper uses tensor product to calculate the optimal shrinking constant that reduces the mean square error loss. The paper also shows some numerical result in simple spectrums, for example, when the residues are scalars. The result of the paper provides that it is applicable to perform James-Stein type regression even if exogeneity is not preserved. This is not directly related to our research under OLS regression, but we do get evidence that the JS approach could work relatively well in real world data where residuals are correlated.

10 References

1. Arnold, S. F. The theory of linear models and multivariate analysis (No. 04; QA278, A7.).
2. Berger, J. O. (2013). Statistical decision theory and Bayesian analysis. Springer Science Business Media.
3. Chandra, T. K., Goswami, A. (1992). Cesaro uniform integrability and the strong law of large numbers.
4. Dwivedi, R., Khamaru, K., Wainwright, M. J., Jordan, M. I. (2018). Theoretical guarantees for EM under misspecified Gaussian mixture models. In Advances in Neural Information Processing Systems (pp. 9681-9689).
5. Efron, B., Morris, C. (1972). Limiting the risk of Bayes and empirical Bayes estimators—Part II: The empirical Bayes case. Journal of the American Statistical Association, 67(337), 130-139.
6. Efron, B. (2014). Two modeling strategies for empirical Bayes estimation. Statistical science: a review journal of the Institute of Mathematical Statistics, 29(2), 285.
7. Goldberg, L. R., Papanicolaou, A., Shkolnik, A. (2018). The dispersion bias. Available at SSRN 3071328.
8. Gruber, M. (2017). Improving Efficiency by Shrinkage: The James–Stein and Ridge Regression Estimators. Routledge.
9. James, W., Stein, C. (1992). Estimation with quadratic loss. In Breakthroughs in statistics (pp. 443-460). Springer, New York, NY.
10. Kim, T. H., White, H. (2001). James-Stein-type estimators in large samples with application to the least absolute deviations estimator. Journal of the American Statistical Association, 96(454), 697-705.
11. Lindley, D. V. (1962). Confidence sets for the mean of a multivariate normal distribution (discussion). J. Roy. Statist. Soc. Ser. B, 24, 285-287.
12. Senda, M., Taniguchi, M. (2006). James–Stein estimators for time series regression models. Journal of multivariate analysis, 97(9), 1984-1996.
13. Shalizi, C. (2015). Simple Linear Regression Models, with Hints at Their Estimation [lecture presentation]. Retrieved from Modern Regression, Section B.
14. Sharpe, William F. (1964). Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk. Journal of Finance. 19:3, pp. 425– 42.
15. Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Stanford University Stanford United States.

16. Stigler, S. M. (1990). The 1988 Neyman Memorial Lecture: A Galtonian Perspective on Shrinkage Estimators.

17. Vasicek, O. (1973). A Note on Using Cross-Sectional Information in Bayesian Estimation of Security Betas. Journal of Finance, 28 (December) 1233-1239.

18. Wager, S. (2014). A geometric approach to density estimation with additive noise. Statistica Sinica, 533-554.