

Lecture Notes on Multivariate Statistical Analysis (Contd.)

1 Decision theoretic formulation

Let X denote a random variable whose distribution depends on an unknown parameter θ lying in a parameter space Θ . Let $d(X)$ denote an estimator of θ . A *loss function* is a non-negative function of θ and $d(X)$, denoted as $l(\theta, d(X))$, and represents the loss incurred when θ is estimated by $d(X)$. The *risk function* associated with this loss function is denoted by

$$R(\theta, d) = \mathbb{E}_\theta l(\theta, d(X)). \quad (1)$$

In *decision theory*, the quality of an estimator is judged by the value of its risk function. An estimator d_1 is said to be *as good as* an estimator d_2 if

$$R(\theta, d_1) \leq R(\theta, d_2) \quad \text{for all } \theta \in \Theta. \quad (2)$$

An estimator d_1 is said to be *better than* d_2 if (2) is satisfied, and

$$R(\theta, d_1) < R(\theta, d_2) \quad \text{for at least one } \theta \in \Theta.$$

An estimator is said to be *admissible* if there exists no estimator which is better than that. Otherwise, it is called an *inadmissible* estimator. Note that, the concept of one estimator being better than another is relative to a specific loss function. Hence the choice of loss function is a critical consideration in decision theory. However, once a particular loss function is chosen, it is reasonable to rule out estimators that are *inadmissible* with respect to the given loss function.

2 Estimation of the mean vector

Suppose that X is a random sample from $N_p(\mu, I_p)$ distribution. We pose the problem of estimation of μ in the decision theoretic framework. For that, we choose the following loss function (squared-error loss):

$$l(\mu, d(X)) = \|d(X) - \mu\|_2^2 = \sum_{i=1}^p (d_i(X) - \mu_i)^2,$$

where $d(X) = (d_1(X), \dots, d_p(X))$.

It is known X is the maximum likelihood estimator of μ . And if we denote this estimator by $d_0(X) = X$, then its risk function is

$$R(\mu, d_0) = \mathbb{E}_\mu \|X - \mu\|_2^2 = p. \quad (3)$$

It was assumed for a long time that $d_0(X)$ is an optimal estimator in every sense. However, Stein (1956) showed that, it is admissible if $p \leq 2$, but is inadmissible if $p \geq 3$. Further, James and Stein (1961) exhibited a simple estimate which has uniformly smaller risk than $d_0(X)$. We shall discuss below the key arguments.

Consider the estimate

$$d_\alpha(X) = \left(1 - \frac{\alpha}{\|X\|_2^2}\right) X,$$

where $\alpha \geq 0$ is a constant. $d_\alpha(X)$ is an example of a *shrinkage estimator*, which pulls every coordinate of the vector X towards the origin by the factor $(1 - \frac{\alpha}{\|X\|_2^2})$. It is easy to check that the risk function of the estimator d_α can be expressed as

$$R(\mu, d_\alpha) = p - 2\alpha \mathbb{E} \left(\frac{(X - \mu)^T X}{X^T X} \right) + \alpha^2 \mathbb{E} \left(\frac{1}{X^T X} \right). \quad (4)$$

From (3) and (4), it follows that

$$R(\mu, d_\alpha) - R(\mu, d_0) = -2\alpha \mathbb{E}_\mu \left(\frac{(X - \mu)^T X}{X^T X} \right) + \alpha^2 \mathbb{E}_\mu \left(\frac{1}{X^T X} \right) \quad (5)$$

The following proposition gives expressions for the two expectations on the right.

Proposition 32 : *If $X \sim N_p(\mu, I_p)$ then,*

$$\mathbb{E}_\mu \left(\frac{1}{X^T X} \right) = \mathbb{E} \left(\frac{1}{p - 2 + 2K} \right) \quad (6)$$

$$\mathbb{E}_\mu \left(\frac{(X - \mu)^T X}{X^T X} \right) = (p - 2) \mathbb{E} \left(\frac{1}{p - 2 + 2K} \right). \quad (7)$$

where K is a random variable having a Poisson distribution with mean $\| \mu \|_2^2 / 2$.

Outline of the proof : Use the fact that $Z = X^T X$ has the noncentral χ^2 distribution with p degrees of freedom and noncentrality parameter $\mu^T \mu$. Now, use the fact that the density of Z can be expressed as

$$f(z) = \sum_{k=0}^{\infty} \mathbb{P}(K = k) g_{m+2k}(z),$$

where $g_r(\cdot)$ is the density function of a central χ^2 random variable. Then (6) follows from conditioning on K . Next, using the fact that

$$\mathbb{E}_\mu \left(\frac{\mu^T (X - \mu)}{\| X \|_2^2} \right) = \mu^T \frac{d}{d\mu} \mathbb{E} \left(\frac{1}{\| X \|_2^2} \right),$$

then invoking (6), and finally simplifying the resulting infinite series, we get,

$$\mathbb{E}_\mu \left[\frac{\mu^T X}{\| X \|_2^2} \right] = \mathbb{E} \left(\frac{2K}{p - 2 + 2K} \right).$$

By Proposition 32, When $\alpha = p - 2$, the minimum value in (5) is attained and the value is given by

$$-(p - 2)^2 \mathbb{E}_\mu \left(\frac{1}{p - 2 + 2K} \right).$$

Hence, if $p \geq 3$, then the difference is negative proving that $d_\alpha(X)$ with $\alpha = p - 2$ has uniformly smaller risk than $d_0(X)$.

Alternative derivation : We use the following result known as *Stein's lemma*. For that we first define a *weakly differentiable* function:

Definition : A function $g : \mathbb{R}^p \rightarrow \mathbb{R}$ is called *weakly differentiable* if there exist functions $h_i : \mathbb{R}^p \rightarrow \mathbb{R}$, $i = 1, \dots, p$ such that

$$\int \psi h_i = - \int (D_i \psi) g \quad \text{for all } \psi \in C_0^\infty.$$

Here $D_i \psi := \frac{\partial}{\partial x_i} \psi$. Then h_i is written as $D_i g$.

Proposition 33 (Stein's Lemma): Let $g : \mathbb{R}^p \rightarrow \mathbb{R}^p$ be weakly differentiable. If $X \sim N_p(\mu, I_p)$, and $\mathbb{E}_\mu[|X_i g_i(X)| + |D_i g_i(X)|] < \infty$ for all $i = 1, \dots, p$, then

$$\mathbb{E}_\mu[(X - \mu)^T g(X)] = \mathbb{E}_\mu[\nabla \cdot g(X)], \quad (8)$$

where $\nabla \cdot g$ means the "Divergence of g ", i.e., $\nabla \cdot g(x) = \sum_{i=1}^p D_i g_i(x)$.

In order to apply *Proposition 33* to the computation of $R(\mu, d_\alpha)$, take $g(x) = \frac{x}{\|x\|_2^2}$ and check that g is weakly differentiable. Moreover, $D_i g_i = \frac{1}{\|x\|_2^2} - \frac{2x_i^2}{\|x\|_2^4}$. This yields, due to *Proposition 33*,

$$\begin{aligned} R(\mu, d_\alpha) &= \mathbb{E}_\mu \|X - \mu\|_2^2 - 2\alpha \mathbb{E}_\mu[(X - \mu)^T g(X)] + \alpha^2 \mathbb{E} \|g(X)\|_2^2 \\ &= p - 2\alpha(p-2) \mathbb{E}_\mu \left[\frac{1}{\|X\|_2^2} \right] + \alpha^2 \mathbb{E}_\mu \left[\frac{1}{\|X\|_2^2} \right]. \end{aligned}$$

This, after minimization w.r.t. α yields $\alpha = p - 2$ and the corresponding estimator $d_{p-2}(X)$ is the *James-Stein estimator*, and denote this by $\hat{\mu}^{JS}(X)$.

2.1 Oracle property

The *James-Stein estimator* enjoys a nice oracle property. Consider the class of all linear shrinkage estimators of the form $\hat{\mu}^c(X) = (1 - c)X$, for some $c > 0$. It follows that

$$R(\mu, \hat{\mu}^c) = \mathbb{E}_\mu \| (1 - c)X - \mu \|_2^2 = (1 - c)^2 p + c^2 \| \mu \|_2^2. \quad (9)$$

Minimization of this risk with respect to c yields the *Ideal Linear Shrinkage "estimator"*:

$$\hat{\mu}^{IS}(X) = (1 - c^{IS})X \quad c^{IS}(\mu) = \frac{p}{p + \| \mu \|_2^2}.$$

Observe that $\hat{\mu}^{IS}(X)$ is not really an estimator, since its definition depends on the value of $\| \mu \|$. The risk of $\hat{\mu}^{IS}$ is given by

$$R(\mu, \hat{\mu}^{IS}) = \frac{p \| \mu \|_2^2}{p + \| \mu \|_2^2}.$$

Now, in order to compare this *oracle risk* with that of the *James-Stein estimator*, we note that,

$$\begin{aligned} R(\mu, \hat{\mu}^{JS}) &= p - (p-2)^2 \mathbb{E}_\mu \|X\|_2^2 = p - (p-2) \mathbb{E} \left[\frac{1}{p-2+2K} \right] \\ &\leq p - (p-2)^2 \frac{1}{p-2+2\mathbb{E}(K)} \quad (\text{by Jensen's inequality}) \\ &= p - \frac{(p-2)^2}{p-2+\| \mu \|_2^2} \\ &= 2 + \frac{(p-2)\| \mu \|_2^2}{(p-2)+\| \mu \|_2^2}, \end{aligned}$$

where in the third line we used the fact that $K \sim \text{Poisson}(\|\mu\|_2^2/2)$. Comparing $R(\mu, \hat{\mu}^{IS})$ with $R(\mu, \hat{\mu}^{JS})$, we note that

$$R(\mu, \hat{\mu}^{JS}) \leq 2 + R(\mu, \hat{\mu}^{IS}).$$

This is an instance of an *oracle inequality*. The interpretation is that, *James-Stein estimator*, even though lacks the knowledge of μ , has risk only within a constant of the risk of the ideal shrinkage estimator, which uses the knowledge about μ . One of the reasons that this result is important is that, linear shrinkage estimators are also *Bayes estimators* under appropriate Gaussian priors on μ . Thus, *James-Stein estimators* have a certain *adaptivity* property.