Contents

Much of this content comes from Lehmann and Casella [1998], sections 5.2, 5.4, 5.5, 4.6 and 4.7.

---

1 Shrinkage estimators

Consider a model $\{p(x|\theta) : \theta \in \Theta\}$ for a random variable X such that

$$\mathrm{E}[X|\theta] = \mu(\theta),\ 0 < \mathrm{Var}[X|\theta] = \sigma^2(\theta) < \infty\ \forall \theta \in \Theta.$$

A linear estimator $\delta(x)$ for $\mu(\theta)$ is an estimator of the form

$$\delta_{ab}(X) = aX + b.$$

Is δ_{ab} admissible?

Theorem 1 (LC thm 5.2.6)**.** *$\delta_{ab}(X) = aX + b$ is inadmissible for $\mathrm{E}[X|\theta]$ under squared error loss whenever*

1. *$a > 1$,*
2. *$a = 1$ and $b \neq 0$, or*
3. *$a < 0$.*

Proof. The risk of δ_{ab} is

$$\begin{aligned} R(\theta, \delta_{ab}) &= \mathrm{E}[(aX + b - \mu)^2|\theta] \\ &= \mathrm{E}[(aX - a\mu - \mu(1-a) + b)^2|\theta] \\ &= \mathrm{E}[a^2(X-\mu)^2 + (b - \mu(1-a))^2 + 2a(X-\mu)(b-\mu(1-a))|\theta] \\ &= a^2\sigma^2 + (b - \mu(1-a))^2 \end{aligned}$$

1. If $a > 1$, then $R(\theta, \delta_{ab}) > a^2\sigma^2 > \sigma^2 = R(\theta, X)$, so δ_{ab} is dominated by X.
2. If $a < 0$, then

$$\begin{aligned} R(\theta, \delta_{ab}) &> (b - \mu(1-a))^2 \\ &= (1-a)^2(b/(1-a) - \mu)^2 \\ &= (b/(1-a) - \mu)^2 = R(\theta, b/(1-a)) \end{aligned}$$

   and so δ_{ab} is dominated by the constant estimator $b/(1-a)$.

3. If $a = 1$ and $b \neq 0$, then $R(\theta, \delta_{ab}) = \sigma^2 + b^2 > \sigma^2 = R(\theta, X)$, so δ_{ab} is dominated by X.

□

Letting $w = 1 - a$ and $\mu_0 = b/(1-a)$, the result suggests that if we want to use an admissible linear estimator, it should be of the form

$$\delta(X) = w\mu_0 + (1-w)X \ , \ w \in [0,1]$$

We call such estimators *linear shrinkage estimators* as they "shrink" the estimate from X towards μ_0. Intuitively, you can think of μ_0 as your "guess" as to the value of μ, and w as the confidence you have in your guess. Of course, the closer your guess is to the truth, the better your estimator.

If μ_0 represents your guess as to $\mu(\theta)$, it seems natural to require that $\mu_0 \in \mu(\Theta) = \{\mu : \mu = \mu(\theta), \theta \in \Theta\}$, i.e. μ_0 is a possible value of μ.

Lemma 1. *If $\mu(\Theta)$ is convex and $\mu_0 \notin \bar{\mu}(\Theta)$, then $\delta(X) = w\mu_0 + (1-w)X$ is not admissible.*

Proof.
For the one-dimensional case, suppose $\mu_0 > \mu(\theta)\ \forall\theta \in \Theta$.
Let $\tilde{\mu}_0 = \sup_\Theta \mu(\theta)$, and $\tilde{\delta}(X) = w\tilde{\mu}_0 + (1-w)X$.
Then $\tilde{\delta}(X)$ dominates $\delta(X)$
(the variances are the same, and the latter has higher bias for all θ).
The proof is similar for the case $\mu_0 < \mu(\theta)\ \forall\theta \in \Theta$.

□

Exercise: Generalize this result to higher dimensions.

---

2 Admissible linear shrinkage estimators

We have shown that $\delta(X) = w\mu_0 + (1-w)X$ is inadmissible for $\mu(\theta) = \mathrm{E}[X|\theta]$ if

- $w \notin [0,1]$ or

- $\mu_0 \notin \mu(\Theta)$.

Restricting attention to $w \in [0,1]$ and $\mu_0 \in \mu(\Theta)$, it may seem that such estimators should always be admissible, but "always" is almost always too inclusive.

Exercise: Given an example where $w\mu_0 + (1-w)X$ is not admissible, even with $w \in (0,1)$ and $\mu_0 \in \mu(\Theta)$.

Linear shrinkage via conjugate priors

What about using a Bayesian argument? Recall,

Theorem. *Any unique Bayes estimator is admissible.*

If we can show that $w\mu_0 + (1-w)X$ is unique Bayes under some prior, then we will have shown admissibility.

Let $X_1, \ldots, X_n \sim$ i.i.d. $p(x|\theta)$, where
$p(x|\theta) \in \mathcal{P} = \{p(x|\theta) = h(x)\exp(x \cdot \theta - A(\theta)) : \theta \in \mathcal{H}\}$
Consider estimation of $\mu = \mathrm{E}[X|\theta]$ under squared error loss.

Let $\pi(\theta) \propto \exp(n_0\mu_0 \cdot \theta - n_0 A(\theta))$ where $n_0 > 0$ and $\mu_0 \in \mathrm{Conv}\{\mathrm{E}[X|\theta] : \theta \in \mathcal{H}\}$
Recall that under this prior,

$$\mathrm{E}[\mu] \equiv \mathrm{E}[\mathrm{E}[X|\theta]] = \mu_0.$$

Then $\pi(\theta|x) \propto \exp(n_1\mu_1 \cdot \theta - n_1 A(\theta))$, where $n_1 = n_0 + n$ and

$$\begin{aligned} n_1\mu_1 &= n_0\mu_0 + n\bar{x} \\ \mu_1 &= n_0\mu_0/n_1 + n\bar{x}/n_1 \\ &= \frac{n_0}{n_0+n_1}\mu_0 + \frac{n}{n_0+n}\bar{x}. \end{aligned}$$

Under this posterior distribution,

$$\mathrm{E}[\mu|x] \equiv \mathrm{E}[\mathrm{E}[X|\theta]|x] = \mu_1.$$

Therefore, the unique Bayes estimator of $\mu = \mathrm{E}[X|\theta]$ under squared error loss is

$$\mu_1 = w\mu_0 + (1-w)\bar{x},$$

and so this linear shrinkage estimator is admissible.

Example (multiple normal means):
Let $\boldsymbol{X} \sim N_p(\boldsymbol{\theta}, \sigma^2 I)$. First consider the case that σ^2 is known, so that

$$\begin{aligned} p(\boldsymbol{x}|\boldsymbol{\theta}) &= (2\pi\sigma^2)^{-p/2} \exp(-(\boldsymbol{x}-\boldsymbol{\theta})\cdot(\boldsymbol{x}-\boldsymbol{\theta})/[2\sigma^2]) \\ &\propto_\theta \exp(\boldsymbol{x}\cdot\boldsymbol{\theta}/\sigma^2 - \boldsymbol{\theta}\cdot\boldsymbol{\theta}/[2\sigma^2]). \end{aligned}$$

Consider the normal prior

$$\begin{aligned} \pi(\boldsymbol{\theta}) &= (2\pi\tau^2)^{-p/2} \exp(-(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\cdot(\boldsymbol{\theta}-\boldsymbol{\theta}_0)/[2\tau_0^2]) \\ &\propto_\theta \exp(\boldsymbol{\theta}_0\cdot\boldsymbol{\theta}/\tau_0^2 - \boldsymbol{\theta}\cdot\boldsymbol{\theta}/[2\tau_0^2]). \end{aligned}$$

where τ_0^2 is analogous to $1/n_0$ in the general formulation for exponential families.

The posterior density is

$$\begin{aligned} \pi(\boldsymbol{\theta}|\boldsymbol{x}) &\propto_\theta \exp\{[\boldsymbol{\theta}_0/\tau_0^2 + \boldsymbol{x}/\sigma^2]\cdot\boldsymbol{\theta} - \boldsymbol{\theta}\cdot\boldsymbol{\theta}[1/\sigma^2 + 1/\tau_0^2]/2\} \\ &= \exp\{\boldsymbol{\theta}_1\cdot\boldsymbol{\theta}/\tau_1^2 - \boldsymbol{\theta}\cdot\boldsymbol{\theta}/[2\tau_1^2]\} \end{aligned}$$

where

- $1/\tau_1^2 = 1/\tau_0^2 + 1/\sigma^2$
- $\boldsymbol{\theta}_1 = \frac{1/\tau_0^2}{1/\tau_0^2+1/\sigma^2}\boldsymbol{\theta}_0 + \frac{1/\sigma^2}{1/\tau_0^2+1/\sigma^2}\boldsymbol{x} \equiv w\boldsymbol{\theta}_0 + (1-w)\boldsymbol{x}$.

So $\{\boldsymbol{\theta}|\boldsymbol{x}\} \sim N_p(\boldsymbol{\theta}_1, \tau_1^2\mathbf{I})$, which means that

$$\mathrm{E}[\boldsymbol{\theta}|\boldsymbol{x}] = \boldsymbol{\theta}_1 = w\boldsymbol{\theta}_0 + (1-w)\boldsymbol{x}$$

uniquely minimizes the posterior risk under squared error loss. The posterior mean is therefore a unique Bayes estimator and also an admissible estimator of $\boldsymbol{\theta}$. Since this result holds for all $\tau_0^2 > 0$, we have the following:

Lemma 2. *For each $w \in (0,1)$ and $\boldsymbol{\theta}_0 \in \mathbb{R}^p$, the estimator $\delta_{w\theta_0}(\boldsymbol{x}) = w\boldsymbol{\theta}_0 + (1-w)\boldsymbol{x}$ is admissible for estimating $\boldsymbol{\theta}$ in the model $\boldsymbol{X} \sim N_p(\boldsymbol{\theta}, \sigma^2 I),\ \boldsymbol{\theta} \in \mathbb{R}^p$, where σ^2 is known.*

Of course, what we would like is the following lemma:

Lemma 3. *For each $w \in (0,1)$ and $\boldsymbol{\theta}_0 \in \mathbb{R}^p$, the estimator $\delta_{w\theta_0}(\boldsymbol{X}) = w\boldsymbol{\theta}_0 + (1-w)\boldsymbol{X}$ is admissible for estimating $\boldsymbol{\theta}$ in the model $\boldsymbol{X} \sim N_p(\boldsymbol{\theta}, \sigma^2 I),\ \boldsymbol{\theta} \in \mathbb{R}^p,\ \sigma^2 \in \mathbb{R}^+$.*

How can this result be obtained?

Theorem 2. *Let $\mathcal{P} = \{p(x|\theta,\psi) : (\theta,\psi) \in \Theta \times \Psi\}$, and for $\psi_0 \in \Psi$, let $\mathcal{P}_{\psi_0} = \{p(x|\theta,\psi_0) : \theta \in \Theta\}$ be a submodel. If δ is admissible for estimating θ under $\mathcal{P}_{\psi_0}$ for each $\psi_0 \in \Psi$, then δ is admissible for estimating θ under $\mathcal{P}$.*

Proof. Suppose δ satisfies the conditions of the theorem but is not admissible. Then there exists a $\delta' \in \mathcal{D}$ such that

$$\forall(\theta,\psi)\ ,\ R((\theta,\psi),\delta') \le R((\theta,\psi),\delta)$$
$$\exists(\theta_0,\psi_0)\ ,\ R((\theta_0,\psi_0),\delta') < R((\theta_0,\psi_0),\delta).$$

But this contradicts the assumption that δ is admissible for estimating θ under $\mathcal{P}_{\psi_0}$. Therefore, no such δ' can exist and so δ is admissible for $\mathcal{P}$. □

A corollary to this theorem is the admissibility of $w\boldsymbol{\theta}_0 + (1-w)\boldsymbol{X}$ in the normal model with unknown variance.

---

3 Admissibility of unbiased normal mean estimators

Let $\boldsymbol{X} \sim N_p(\boldsymbol{\theta}, \sigma^2\mathbf{I}),\ \boldsymbol{\theta} \in \mathbb{R}^p,\ \sigma^2 > 0$.

For estimation of $\boldsymbol{\theta}$ under squared error loss, we have shown that the linear shrinkage estimator $\delta(\boldsymbol{x}) = w\boldsymbol{\theta}_0 + (1-w)\boldsymbol{x}$ is

- inadmissible if $w \notin [0,1]$,
- admissible if $w \in (0,1)$.

What remains to evaluate is the admissibility for $w \in \{0,1\}$. Admissibility for $w = 1$ is easy to show - the estimator $\delta_{1\theta_0}(\boldsymbol{x}) = \boldsymbol{\theta}_0$ beats everything at $\boldsymbol{\theta}_0$ and so can't be dominated. The last and most interesting case is that of $w = 0$, i.e. $\delta_0(\boldsymbol{X}) = \boldsymbol{X}$, the unbiased MLE and UMVUE.

Blyth's method

Recall Blyth's method for showing admissibility using a limiting Bayes argument:

Theorem 3 (LC 5.7.13)**.** *Suppose $\Theta \subset \mathbb{R}^p$ is open, and that $R(\theta, \delta)$ is continuous in θ for all $\delta \in \mathcal{D}$. Let δ be an estimator and $\{\pi_n\}$ be a sequence of measures such that for any open ball $B \subset \Theta$,*

$$\frac{R(\pi_n, \delta) - R(\pi_n, \delta_{\pi_n})}{\pi_n(B)} \to 0 \text{ as } n \to \infty.$$

Then δ is admissible.

Let's try to use this to show admissibility of $\delta_0(\boldsymbol{X}) = \boldsymbol{X}$ in the normal means problem. We begin with the case that $\sigma^2 = 1$ is known.

$$\begin{aligned} \boldsymbol{X} &\sim N_p(\boldsymbol{\theta}, \mathbf{I}) \\ \boldsymbol{\theta} &\sim N_p(\mathbf{0}, \tau^2\mathbf{I}) \\ \{\boldsymbol{\theta}|\boldsymbol{X}\} &\sim N_p(\tfrac{\tau^2}{1+\tau^2}\boldsymbol{X}, \tfrac{\tau^2}{1+\tau^2}\mathbf{I}) \end{aligned}$$

The unique Bayes estimator is $\delta_{\tau^2} = \mathrm{E}[\boldsymbol{\theta}|\boldsymbol{X}] = \frac{\tau^2}{1+\tau^2}\boldsymbol{X} \equiv (1-w)\boldsymbol{X}$.

To apply the theorem, we need to compute the Bayes risk of δ_{τ^2} and $\boldsymbol{X}$ under the $N_p(\mathbf{0}, \tau^2\mathbf{I})$ prior π_{τ^2}. The loss we will use is "scaled" squared error loss, $L(\boldsymbol{\theta}, \boldsymbol{d}) = \sum(\theta_j - d_j)^2/p$. Because the risk is the average of the individual MSEs, the Bayes

risk is just the average of the Bayes risks from the p components,

$$\begin{aligned} R(\boldsymbol{\theta}, \delta) &= \mathrm{E}[\sum_{j=1}^{p} (\theta_j - \delta_j)^2]/p \\ &= \sum_{j=1}^{p} \mathrm{E}[(\theta_j - \delta_j)^2]/p, \end{aligned}$$

and so calculating the Bayes risk is similar to calculating the risk in the $p = 1$ problem. For $\delta_{\tau^2}(x) = ax$, where $a = 1 - w = \tau^2/(1+\tau^2)$, we have

$$\begin{aligned} \mathrm{E}[(aX - \theta)^2] &= \mathrm{E}[(aX - a\theta + (1-a)\theta)^2] \\ &= a^2 \mathrm{E}[(X-\theta)^2] + 2a(1-a)\mathrm{E}[(X-\theta)\theta] + (1-a)^2 \mathrm{E}[\theta]^2 \\ &= a^2 + (1-a)^2\tau^2 \\ &= \left(\frac{\tau^2}{1+\tau^2}\right)^2 + \frac{\tau^2}{(1+\tau^2)^2} = \frac{\tau^2}{1+\tau^2} \end{aligned}$$

A more intuitive way to calculate this makes use of the fact that $\delta_{\tau^2}(X) = \mathrm{E}[\theta|X]$, so

$$\begin{aligned} \mathrm{E}[(\theta - \delta_{\tau^2})^2] &= \mathrm{E}[(\theta - \mathrm{E}[\theta|X])^2] \\ &= \mathrm{E}_x[\mathrm{E}_{\theta|x}[(\theta - \mathrm{E}[\theta|X])^2]] \\ &= \mathrm{E}_x[\mathrm{Var}[\theta|X]] \\ &= \mathrm{E}_x[\tau^2/(1+\tau^2)] = \frac{\tau^2}{1+\tau^2}. \end{aligned}$$

Similarly,

$$\mathrm{E}[(X-\theta)^2] = \mathrm{E}_\theta[\mathrm{E}_{x|\theta}[(X-\theta)^2]] = \mathrm{E}_\theta[1] = 1.$$

So $R(\pi_{\tau^2}, \delta_{\tau^2}) = \frac{\tau^2}{1+\tau^2}$ and $R(\pi_{\tau^2}, X) = 1$. Returning to the p-variate case, since the Bayes risk is the arithmetic average of the risks for each of the p components of $\boldsymbol{\theta}$, we have

$$\begin{aligned} R(\pi_{\tau^2}, \delta_{\tau^2}) &= \frac{\tau^2}{1+\tau^2} \\ R(\pi_{\tau^2}, \boldsymbol{X}) &= 1. \end{aligned}$$

Note that

$$\delta_{\tau^2}(\boldsymbol{X}) = \tfrac{\tau^2}{1+\tau^2}\boldsymbol{X} \uparrow \boldsymbol{X} \quad \text{as} \quad \tau^2 \uparrow \infty,$$
$$R(\pi_{\tau^2}, \delta_{\tau^2}) \uparrow R(\pi_{\tau^2}, \boldsymbol{X}) \quad \text{as} \quad \tau^2 \uparrow \infty,$$

so $\boldsymbol{X}$ is a "limiting Bayes" estimator, for which the risk difference from the Bayes estimator converges to zero. This is promising - let's now apply the theorem. Letting B be any open finite ball in $\mathbb{R}^p$, we need to see if the following limit is zero:

$$\begin{aligned}\lim_{\tau^2\to\infty} \frac{R(\pi_{\tau^2}, \boldsymbol{X}) - R(\pi_{\tau^2}, \delta_{\tau^2})}{\pi_{\tau^2}(B)} &= \lim_{\tau^2\to\infty} \frac{(1 - \frac{\tau^2}{1+\tau^2})}{\pi_{\tau^2}(B)} \\ &= \lim_{\tau^2\to\infty} [(1+\tau^2)\pi_{\tau^2}(B)]^{-1}\end{aligned}$$

Now $\pi_{\tau^2}(B) \to 0$ as $\tau^2 \to \infty$ for any bounded set B Therefore, the limit is zero only if

$$\lim_{\tau^2\to\infty} \tau^2 \pi_{\tau_2}(B) = \infty.$$

We have

$$\begin{aligned}\tau^2\pi_{\tau_2}(B) &= \tau^2 \int_B (2\pi\tau^2)^{-p/2} \exp(-||\boldsymbol{\theta}||^2/[2\tau^2])\; d\boldsymbol{\theta} \\ &= (2\pi)^{-p/2} \times (\tau^2)^{1-p/2} \times \int_B \exp(-||\boldsymbol{\theta}||^2/[2\tau^2])\; d\boldsymbol{\theta} \\ &(2\pi)^{-p/2} \times (a) \times (b).\end{aligned}$$

Now take the limit as $\tau^2 \to \infty$:

$$(b) \to \text{Vol}(B) \quad \text{as} \quad \tau^2 \to \infty$$
$$(a) \to \begin{cases} \infty & \text{if } p = 1 \\ 1 & \text{if } p = 2 \\ 0 & \text{if } p > 2 \end{cases}.$$

Therefore, the desired limit is achieved for $p = 1$ but not $p > 1$.

- By the theorem, X is admissible for Θ.

- For $p > 1$, this method of showing admissibility does not work.
  - For $p = 2$, $\boldsymbol{X}$ can be shown to be admissible using Blyth's method with non-normal priors (see LC exercise 5.4.5).
  - For $p > 2$, $\boldsymbol{X}$ can't be shown to be admissible because it isn't.

Interpreting the failure of Blyth's method:

The admissibility conditions for Blyth's method derived from consideration of the existence of an estimator δ that dominates $\boldsymbol{X}$. If such an estimator exits, then by continuity of risks

$$\exists\ \epsilon > 0 \text{ and an open ball } \ B \subset \Theta : R(\theta, \boldsymbol{X}) - R(\theta, \delta) > \epsilon\ \forall \theta \in B,$$

which implies for each prior π_k that

$$\begin{aligned} R(\pi_k, \boldsymbol{X}) - R(\pi_k, \delta) = \int [R(\theta, \boldsymbol{X}) - R(\theta, \delta)]\pi_k(d\theta) &\geq \int_B [R(\theta, \boldsymbol{X}) - R(\theta, \delta)]\pi_k(d\theta) \\ &\geq \epsilon \pi_k(B). \end{aligned}$$

Integrating with respect to a prior π_k, and comparing to the Bayes risk of the Bayes estimator δ_k under π_k gives

$$R(\pi_k, \boldsymbol{X}) - R(\pi_k, \delta_k) \geq R(\pi_k, \boldsymbol{X}) - R(\pi_k, \delta) > \epsilon \pi_k(B)\ \forall k,$$

as δ_k has Bayes risk less than or equal to that of δ. Could such a δ exist?

Could exist: Suppose B is a ball such that $\pi_k(B)$ goes to zero very fast. Then an estimator (like $\boldsymbol{X}$) can have a good limiting Bayes risk and still do poorly on B. This allows for the possibility of domination by another estimator that does better on B.

Couldn't exist: On the other hand, if $R(\pi_k, \boldsymbol{X}) - R(\pi_k, \delta_k)$ goes to zero very fast (e.g. faster than the probability of any ball B), then in a sense $\boldsymbol{X}$ would have to be doing well everywhere, and would not be able to be dominated - this is Blyth's method for showing admissibility.

What fails in the admissibility proof for the normal means problem is that for $p > 2$, the probability $\pi_k(B)$ of an open ball B is going to zero much faster than the Bayes risk difference, leaving a large enough "gap" for some other estimator to do better.

4 Motivating the James-Stein estimator

Stein [1956] showed that $\boldsymbol{X}$ is inadmissible for $\boldsymbol{\theta}$ in the normal means problem when $p > 2$. This was surprising, as $\boldsymbol{X}$ is the MLE and UMVUE for θ. In this section,

4.1 What is wrong with $\boldsymbol{X}$?

For large p,

- $\boldsymbol{X}$ may be close to $\boldsymbol{\theta}$, but
- $\boldsymbol{X} \cdot \boldsymbol{X} = ||\boldsymbol{X}||^2$ may be far from $\boldsymbol{\theta} \cdot \boldsymbol{\theta} = ||\boldsymbol{\theta}||^2$.

If $\boldsymbol{X} \sim N_p(\boldsymbol{\theta}, \mathbf{I})$,

$$\begin{aligned} \mathrm{E}[||\boldsymbol{X}||^2] &= \mathrm{E}[\sum_1^p X_j^2] \\ &= \sum_1^p (\theta_j^2 + 1) = ||\boldsymbol{\theta}||^2 + p, \end{aligned}$$

so for large p, the magnitude of the estimator vector $\boldsymbol{X}$ is expected to be much larger than the magnitude of the estimand vector $\boldsymbol{\theta}$. More insight can be gained as follows: Note that every vector $\boldsymbol{x}$ can be expressed as

$$\boldsymbol{x} = s\boldsymbol{\theta} + \boldsymbol{r}, \text{ for some } s \in \mathbb{R} \text{ and } \boldsymbol{r} : \boldsymbol{\theta} \cdot \boldsymbol{r} = 0.$$

Here, the random variable s is the magnitude of the projection of $\boldsymbol{x}$ in the direction of $\boldsymbol{\theta}$, and $\boldsymbol{r}$ is the residual vector. Using this decomposition, we can write the squared-error loss of $a\boldsymbol{x}$ for estimating $\boldsymbol{\theta}$ as

$$\begin{aligned} ||a\boldsymbol{x}-\boldsymbol{\theta}||^2 &= (a\boldsymbol{x}-\boldsymbol{\theta})\cdot(a\boldsymbol{x}-\boldsymbol{\theta}) \\ &= ((as-1)\boldsymbol{\theta}+a\boldsymbol{r})\cdot((as-1)\boldsymbol{\theta}+a\boldsymbol{r}) \\ &= (as-1)^2||\boldsymbol{\theta}||^2+a^2||\boldsymbol{r}||^2 \end{aligned}$$

Now consider replacing $\boldsymbol{x}$ with $\boldsymbol{X}\sim N_p(\boldsymbol{\theta},\boldsymbol{I})$. The random-variable version of the above equation is then

$$||a\boldsymbol{X}-\boldsymbol{\theta}||^2=(aS-1)^2||\boldsymbol{\theta}||^2+a^2||\boldsymbol{R}||^2$$

Exercise: Show that

- $S\sim N(1,||\boldsymbol{\theta}||^{-2})$,
- $||\boldsymbol{R}||^2\sim\chi^2_{p-1}$,
- S and $\boldsymbol{R}$ are independent.

Now imagine a situation where p is growing but $||\boldsymbol{\theta}||^2$ remains fixed. The distribution of $(aS-1)^2||\boldsymbol{\theta}||^2$ remains fixed whereas the distribution of $a^2||\boldsymbol{R}||^2$ blows up. This suggests that if we think $||\boldsymbol{\theta}||^2/p$ is small we should use an estimator like $a\boldsymbol{X}$ with $a<1$ to control the error that comes from $\boldsymbol{R}^2$. But what should the value of a be?

4.2 An oracle estimator:

Question: Among estimators $a\mathbf{X}:a\in[0,1]$, which has the smallest risk?
Solution:

$$\begin{aligned} \mathrm{E}[||a\boldsymbol{X}-\boldsymbol{\theta}||^2] &= \mathrm{E}[||(a\boldsymbol{X}-a\boldsymbol{\theta})-(1-a)\boldsymbol{\theta}||^2] \\ &= a^2p+(1-a)^2||\boldsymbol{\theta}||^2. \end{aligned}$$

Taking derivatives, the minimizing value $\tilde{a}$ of a satisfies

$$\begin{aligned} 2\tilde{a}p - 2(1-\tilde{a})||\boldsymbol{\theta}||^2 &= 0 \\ \frac{\tilde{a}}{1-\tilde{a}} &= \frac{||\boldsymbol{\theta}||^2}{p} \\ \tilde{a} &= \frac{||\boldsymbol{\theta}||^2}{||\boldsymbol{\theta}||^2 + p}. \end{aligned}$$

Thus the optimal shrinkage "estimator" is given by $\delta_{\tilde{a}}(\boldsymbol{X}) = \tilde{a}\boldsymbol{X}$. This is not really an estimator in the usual sense, because the ideal degree of shrinkage $\tilde{a}$ depends on $\boldsymbol{\theta}$. For this reason, $\tilde{a}\boldsymbol{X}$ is sometimes called an "oracle estimator:" You would need an oracle to tell you the value of $||\boldsymbol{\theta}||^2$ before you could use it. Note that the risk of this estimator is

$$\begin{aligned} \mathrm{E}[||a\boldsymbol{X} - \boldsymbol{\theta}||^2] &= \frac{||\boldsymbol{\theta}||^4 p + p^2||\boldsymbol{\theta}||^2}{(||\boldsymbol{\theta}||^2 + p)^2} \\ &= \frac{p||\boldsymbol{\theta}||^2(||\boldsymbol{\theta}||^2 + p)}{(||\boldsymbol{\theta}||^2 + p)^2} \\ &= \frac{p||\boldsymbol{\theta}||^2}{||\boldsymbol{\theta}||^2 + p} \end{aligned}$$

and so

$$\mathrm{E}[||a\boldsymbol{X} - \boldsymbol{\theta}||^2] = p\frac{||\boldsymbol{\theta}||^2}{||\boldsymbol{\theta}||^2 + p} < p = \mathrm{E}[||\boldsymbol{X} - \boldsymbol{\theta}||^2].$$

The risk differential is large if $||\boldsymbol{\theta}||^2$ is small compared to p.

4.3 Adaptive shrinkage estimation

As shown above, the optimal amount of shrinkage $\tilde{a}$ is

$$\tilde{a} = \frac{||\boldsymbol{\theta}||^2}{||\boldsymbol{\theta}||^2 + p} = \frac{||\boldsymbol{\theta}||^2/p}{||\boldsymbol{\theta}||^2/p + 1}.$$

Note that $||\boldsymbol{\theta}||^2/p$ is the variability of of the θ_j values around zero. Can this variability be estimated? Consider the following hierarchical model:

$$\begin{aligned} X_j = \theta_j + \epsilon_j \quad & \epsilon_1, \ldots, \epsilon_p \overset{\text{iid}}{\sim} N(0,1) \\ & \theta_1, \ldots, \theta_p \overset{\text{iid}}{\sim} N(0, \tau^2) \end{aligned}$$

If you'd like to connect this with some actual inference problem, imagine that each X_j is the sample mean or t-statistic calculated from observations from experiment j, with population mean θ_j.

Suppose you believed this model and knew the value of τ^2. If you were interested finding an estimator $\delta(\boldsymbol{X})$ that minimized the the expected squared error $||\boldsymbol{\theta} - \delta(\boldsymbol{X})||^2$ under repeated sampling of

- $\theta_1, \ldots, \theta_p$, followed by sampling of
- $X_1, \ldots, X_p$,

you would want to come up with an estimator $\delta(\boldsymbol{X})$ that minimized

$$\mathrm{E}[||\boldsymbol{\theta} - \delta(\boldsymbol{X})||^2] = \int \int ||\boldsymbol{\theta} - \delta(\boldsymbol{x})||^2 p(d\boldsymbol{x}|\boldsymbol{\theta}) p(d\boldsymbol{\theta}).$$

Exercise: Show that $\delta_{\tau^2}(\boldsymbol{X}) = \frac{\tau^2}{1+\tau^2}\boldsymbol{X}$ minimizes the expected loss.

If we knew τ^2, then the estimator to use would be $\frac{\tau^2}{1+\tau^2}\boldsymbol{X}$. We generally don't know τ^2, but maybe it can be estimated from the data. Under the above model,

$$\begin{aligned} \boldsymbol{X} &= \boldsymbol{\theta} + \boldsymbol{\epsilon} \\ \boldsymbol{\theta} &\sim N_p(\mathbf{0}, \tau^2 \mathbf{I}) \\ \boldsymbol{\epsilon} &\sim N_p(\mathbf{0}, \mathbf{I}) \\ \mathrm{Cov}(\boldsymbol{\theta}, \boldsymbol{\epsilon}) &= \mathbf{0}. \end{aligned}$$

This means that the distribution of $\boldsymbol{X}$ marginalized over $\boldsymbol{\theta}$ is

$$\boldsymbol{X} \sim N_p(\mathbf{0}, (\tau^2 + 1)\mathbf{I}).$$

An unbiased estimator of τ^2+1 is clearly $||\boldsymbol{X}||^2/p$, so an unbiased estimator of τ^2 is

$$\hat{\tau}^2 = \frac{||\boldsymbol{X}||^2 - p}{p}.$$

However, we were interested in estimating $\tau^2/(\tau^2+1)$, not τ^2. If $p > 2$, you can use the fact that $||\boldsymbol{X}||^2 \sim \text{gamma}(p/2, 1/[2(\tau^2+1)])$ to show that

$$\text{E}[||\boldsymbol{X}||^{-2}||] = [(p-2)(\tau^2+1)]^{-1},$$

and so

$$\text{E}[\tfrac{p-2}{||\boldsymbol{X}||^2}] = \frac{1}{\tau^2+1}$$
$$\text{E}[1 - \tfrac{p-2}{||\boldsymbol{X}||^2}] = \frac{\tau^2}{\tau^2+1}.$$

Again, $\frac{\tau^2}{\tau^2+1}\boldsymbol{X}$ would be the optimal estimator in this hierarchical model if we knew τ^2. If we don't know τ^2, we might instead consider using

$$\begin{aligned}\delta_{JS}(\boldsymbol{X}) &= (\frac{\widehat{\tau^2}}{\tau^2+1})\boldsymbol{X} \\ &= (1 - \tfrac{p-2}{||\boldsymbol{X}||^2})\boldsymbol{X}.\end{aligned}$$

This estimator is called the James-Stein estimator. As we will see, it has many interesting properties:

- For large p in the hierarchical normal model, it is almost as good as the oracle estimator $\frac{\tau^2}{\tau^2+1}\boldsymbol{X}$:

$$\text{E}_{X\theta}[||\boldsymbol{\theta} - \delta_{JS}||^2] \approx \text{E}_{X\theta}[||\boldsymbol{\theta} - \tfrac{\tau^2}{\tau^2+1}\boldsymbol{X}||^2]$$

- Even if the hierarchical normal model isn't correct, it still is almost as good as the oracle estimator $\tilde{a}\boldsymbol{X}$ in the normal means model:

$$\text{E}_{X|\theta}[||\boldsymbol{\theta} - \delta_{JS}||^2] \approx \text{E}_{X|\theta}||\boldsymbol{\theta} - \tilde{a}\boldsymbol{X}||^2]$$

- In the normal means problem, this estimator dominates the unbiased estimator $\boldsymbol{X}$ if $p > 2$:

$$\text{E}_{X|\theta}[||\boldsymbol{\theta} - \delta_{JS}||^2] < \text{E}_{X|\theta}[||\boldsymbol{\theta} - \boldsymbol{X}||^2] \ \forall \boldsymbol{\theta}$$

We will show this last inequality first.

5 Risk of δ_{JS}

We will show that δ_{JS} dominates $\boldsymbol{X}$ by showing that the risk $R(\boldsymbol{\theta}, \delta_{JS}) = \mathrm{E}[||\delta_{JS} - \boldsymbol{\theta}||^2]/p$ is uniformly less than 1. This will not be done by computing its risk function directly, but instead by showing that 1 is an upper bound on the risk. This bound will be obtained via an identity that has applications beyond the calculation of $R(\boldsymbol{\theta}, \delta_{JS})$.

5.1 Risk bound for δ_{JS}

We can write the James-Stein estimator as

$$\begin{aligned}\delta_{JS} = \left(\widehat{\frac{\tau^2}{1+\tau^2}}\right)\boldsymbol{x} &= (1 - \tfrac{p-2}{\boldsymbol{x}\cdot\boldsymbol{x}})\boldsymbol{x} \\ &= \boldsymbol{x} - \tfrac{p-2}{\boldsymbol{x}\cdot\boldsymbol{x}}\boldsymbol{x} \\ &\equiv \boldsymbol{x} - g(\boldsymbol{x}).\end{aligned}$$

Under $\boldsymbol{X} \sim N_p(\boldsymbol{\theta}, \mathbf{I})$,

$$\begin{aligned}\mathrm{E}[||\delta_{JS} - \boldsymbol{\theta}||^2] &= \mathrm{E}[(\boldsymbol{X} - g(\boldsymbol{X}) - \boldsymbol{\theta})^2] \\ &= \mathrm{E}[(\boldsymbol{X} - \boldsymbol{\theta}) - g(\boldsymbol{X}))^2] \\ &= \mathrm{E}[||(\boldsymbol{X} - \boldsymbol{\theta})||^2] + \mathrm{E}[||g(\boldsymbol{X})||^2] - 2\mathrm{E}[(\boldsymbol{X} - \boldsymbol{\theta}) \cdot g(\boldsymbol{X})].\end{aligned}$$

where all expectations are with respect to the distribution of $\boldsymbol{X}$ given $\boldsymbol{\theta}$. The first expectation is p and the second is

$$\mathrm{E}[(p-2)^2 \tfrac{\boldsymbol{X}\cdot\boldsymbol{X}}{(\boldsymbol{X}\cdot\boldsymbol{X})^2}] = (p-2)^2\mathrm{E}[\tfrac{1}{\boldsymbol{X}\cdot\boldsymbol{X}}].$$

The third expectation is more complicated, but in the next subsection we'll derive an identity (Stein's identity) for computing $\mathrm{E}[(\boldsymbol{X} - \boldsymbol{\theta}) \cdot g(\boldsymbol{X})]$ that is applicable for arbitrary functions g. Stein's identity as applied to $g(\boldsymbol{x}) = \frac{p-2}{\boldsymbol{x}\cdot\boldsymbol{x}}\boldsymbol{x}$ gives

$$\mathrm{E}[(\boldsymbol{X} - \boldsymbol{\theta})g(\boldsymbol{X})] = \mathrm{E}[\tfrac{(p-2)^2}{\boldsymbol{X}\cdot\boldsymbol{X}}].$$

Using this for the above risk calculation gives

$$\begin{aligned} \mathrm{E}[||\delta_{JS} - \boldsymbol{\theta}||^2] &= p + (p-2)^2\mathrm{E}[\tfrac{1}{\boldsymbol{X}\cdot\boldsymbol{X}}] - 2(p-2)^2\mathrm{E}[\tfrac{1}{\boldsymbol{X}\cdot\boldsymbol{X}}] \\ &= p - \mathrm{E}[\tfrac{(p-2)^2}{\boldsymbol{X}\cdot\boldsymbol{X}}]. \end{aligned}$$

Note that we haven't actually calculated the risk of δ_{JS} is closed form - our formula depends on the expectation of $1/\boldsymbol{X}\cdot\boldsymbol{X}$, which is an inverse-moment of a noncentral χ^2 distribution where the noncentrality parameter depends on $\boldsymbol{\theta}$. However, computing this moment is not necessary to show that δ_{JS} dominates $\boldsymbol{X}$: Since $\frac{1}{\boldsymbol{x}\cdot\boldsymbol{x}} > 0 \; \forall \boldsymbol{x} \in \mathbb{R}^p$, we have

$$\mathrm{E}[||\delta_{JS} - \boldsymbol{\theta}||^2] = p - \mathrm{E}[\tfrac{(p-2)^2}{\boldsymbol{X}\cdot\boldsymbol{X}}] < p = \mathrm{E}[||\boldsymbol{X} - \boldsymbol{\theta}||^2].$$

Since the expectation of $(\boldsymbol{X}\cdot\boldsymbol{X})^{-1}$ is complicated, further study of the risk of δ_{JS} is often achieved via a study of its *unbiased risk estimate*. From the above calculation, we see that

$$\mathrm{E}[||\delta_{JS} - \boldsymbol{\theta}||^2] = \mathrm{E}[p - \tfrac{(p-2)^2}{\boldsymbol{X}\cdot\boldsymbol{X}}],$$

and so $p - \frac{(p-2)^2}{\boldsymbol{X}\cdot\boldsymbol{X}}$ can be said to be an unbiased estimate of the risk of δ_{JS}.

5.2 Stein's identity

We start with a univariate version of the identity:

Lemma 4 (Stein's identity)**.** *Let $X \sim N(\mu, \sigma^2)$ and let $g(x)$ be such that $\mathrm{E}[|g'|] < \infty$. Then*

$$\mathrm{E}[g(X)(X-\mu)] = \sigma^2\mathrm{E}[g'(X)].$$

Proof.
The proof follows from Fubini's theorem and a bit of calculus. Letting $p(x) = \phi([x-\mu]/\sigma)/\sigma$, note that $p'(x) = -(\frac{x-\mu}{\sigma^2})p(x)$. By the fundamental theorem of calculus,

$$p(x) \;=\; \int_{-\infty}^{x} -(\tfrac{y-\mu}{\sigma^2})p(y)\; dy \;=\; \int_{x}^{\infty} (\tfrac{y-\mu}{\sigma^2})p(y)\; dy.$$

The expectation we wish to calculate is

$$\begin{aligned}\mathrm{E}[g'(x)] &= \int_{-\infty}^{\infty} g'(x)p(x)\ dx \\ &= \int_{0}^{\infty} g'(x)p(x)\ dx + \int_{-\infty}^{0} g'(x)p(x)\ dx.\end{aligned}$$

Doing the first part, we have

$$\begin{aligned}\int_0^\infty g'(x)p(x)\ dx &= \int_0^\infty g'(x) \int_x^\infty (\tfrac{y-\mu}{\sigma^2})p(y)\ dy\ dx \\ &= \int_{0<x<y<\infty} g'(x)(\tfrac{y-\mu}{\sigma^2})p(y)\ dy\ dx \\ &= \int_0^\infty (\tfrac{y-\mu}{\sigma^2})p(y) \int_0^y g'(x)\ dx\ dy \\ &= \int_0^\infty (\tfrac{y-\mu}{\sigma^2})p(y)[g(y)-g(0)]\ dy \\ &= \mathrm{E}[(g(X)-g(0))(\tfrac{X-\mu}{\sigma^2})1(X>0)].\end{aligned}$$

Similarly, the second part is

$$\int_{-\infty}^0 g'(x)p(x)\ dx = \mathrm{E}[(g(X)-g(0))(\tfrac{X-\mu}{\sigma^2})1(X<0)].$$

Adding the two parts gives

$$\begin{aligned}\mathrm{E}[g'(x)] &= \mathrm{E}[(g(X)-g(0))(\tfrac{X-\mu}{\sigma^2})] \\ &= \mathrm{E}[g(X)(\tfrac{X-\mu}{\sigma^2})] - g(0)\mathrm{E}[(\tfrac{X-\mu}{\sigma^2})] = \mathrm{E}[g(X)(\tfrac{X-\mu}{\sigma^2})],\end{aligned}$$

and so

$$\mathrm{E}[g(X)(X-\mu)] = \sigma^2\mathrm{E}[g'(X)].$$

□

Stein's lemma is often alternatively proven with integration by parts. These proofs go roughly as follows: As before,

$$\begin{aligned}p(x) &= (2\pi\sigma^2)^{-1/2}\exp\{-(x-\mu)^2/[2\sigma^2]\} \\ \frac{dp(x)}{dx} &= -(\tfrac{1}{\sigma^2})(x-\mu)p(x).\end{aligned}$$

$$\begin{aligned}
\mathrm{E}[g'(X)] &= \int_{-\infty}^{\infty} g'(x) \times p(x)\ dx \\
&= \lim_{c\to\infty} \int_{-c}^{c} g'(x) \times p(x)\ dx \\
&\equiv \lim_{c\to\infty} \int_{-c}^{c} v'(x)u(x)\ dx \\
&= \lim_{c\to\infty} \left(u(x)v(x)|_{-c}^{c} - \int_{-c}^{c} v(x)u'(x)\ dx \right) \\
&= \lim_{c\to\infty} \left(g(x)p(x)|_{-c}^{c} - \int_{-c}^{c} g(x)[-(x-\mu)/\sigma^2]p(x)\ dx \right) \\
&= \mathrm{E}[g(X)(X-\mu)/\sigma^2] + \lim_{c\to\infty} g(x)p(x)|_{-c}^{c}.
\end{aligned}$$

To complete the proof we have to show that the last limit is zero. This is straightforward to show if $p(x)$ decreases monotonically in $|x|$:

Lemma 5. *Let $p(x)$ be decreasing to zero in $|x|$ and let $\mathrm{E}[|g'(X)|] < \infty$. Then $g(x)p(x) \to 0$ as $x \to \pm\infty$.*

Proof. Given $\epsilon > 0$ $\exists K$ such that

$$\int_{K}^{\infty} |g'(x)|p(x)\ dx < \epsilon/3.$$

Then for any t sufficiently large,

1. $p(t) < p(K)(\int_0^K |g'(x)|p(x))^{-1} \times \epsilon/3$
2. $p(t)|g(0)| < \epsilon/3$.

From this, we have

$$\begin{aligned}
|g(t)p(t)| &= p(t)|g(0) + \int_0^t g'(x)\ dx| \\
&= p(t)|g(0) + \int_0^K g'(x)\ dx + \int_K^t g'(x)\ dx| \\
&\leq p(t)|g(0)| + p(t)\int_0^K |g'(x)|\ dx + p(t) + p(t)\int_K^t |g'(x)|\ dx \\
&\leq p(t)|g(0)| + \tfrac{p(t)}{p(K)}\int_0^K |g'(x)|p(x)\ dx + \int_K^t |g'(x)|p(x)\ dx \\
&\leq \epsilon/3 + \epsilon/3 + \epsilon/3 = \epsilon,
\end{aligned}$$

where the second to last line holds as $p(x)/p(K) < 1$ and $p(x)/p(t) < 1$ on $x \in (0, K)$ and (K, t) respectively, due to $p(x)$ being monotonically decreasing. □

This identity generalizes to other exponential families. See LC Theorem 1.5.15.

For computing the risk of a vector-valued function $g : \mathbb{R}^p \to \mathbb{R}^p$, we will need a multivariate version of the above identity.

Lemma 6 (Stein's identity, multivariate version)**.** *Let* $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \sigma^2\mathbf{I})$ *and let* $g(\boldsymbol{x}) : \mathbb{R}^p \to \mathbb{R}^p$ *such that* $\mathrm{E}[|dg_i/dx_i|] < \infty$*. Then*

$$\mathrm{E}[(\boldsymbol{X} - \boldsymbol{\mu}) \cdot g(\boldsymbol{X})] = \sigma^2 \mathrm{E}[\nabla \cdot g(\boldsymbol{X})]$$

where $\nabla \cdot g = \sum_{j=1}^p dg_j(\boldsymbol{x})/dx_j$.

Proof. This is just a corollary of the univariate version:

$$\begin{aligned}
\mathrm{E}[dg_p(\boldsymbol{x})/dx_p] &= \int_{\boldsymbol{x}_{-p}} \left(\int_{x_p} \tfrac{dg_p(\boldsymbol{x})}{dx_p} p(x_p)\ dx_p \right) \prod_1^{p-1} p(x_j)\ dx_j \\
&= \int_{\boldsymbol{x}_{-p}} \mathrm{E}_{x_p}[g_p(\boldsymbol{X})(X_p - \mu_p)]/\sigma^2 \prod_1^{p-1} p(x_j)\ dx_j \\
&= \tfrac{1}{\sigma^2}\mathrm{E}[g_p(\boldsymbol{X})(X_p - \mu_p)]],
\end{aligned}$$

and similarly for each other $j \in \{1, \ldots, p\}$. Therefore

$$\begin{aligned} \mathrm{E}[\nabla \cdot g(\boldsymbol{X})] &= \sum \mathrm{E}[dg_j(\boldsymbol{x})/dx_j] \\ &= \sum \mathrm{E}[g_j(\boldsymbol{X})(X_j - \mu_j)]/\sigma^2 \\ &= \mathrm{E}[g(\boldsymbol{X}) \cdot (\boldsymbol{X} - \boldsymbol{\mu})]/\sigma^2. \end{aligned}$$

□

Now we are in a position to apply the lemma to obtain the unbiased risk estimator of δ_{JS}. Recall we needed to calculate $\mathrm{E}[(\boldsymbol{X} - \boldsymbol{\theta}) \cdot g(\boldsymbol{X})]$ where $g(\boldsymbol{x}) = \frac{(p-2)}{\boldsymbol{x} \cdot \boldsymbol{x}} \boldsymbol{x}$. Applying the lemma, we have

$$\begin{aligned} \boldsymbol{g}(\boldsymbol{x}) &= (p-2) \left(\frac{x_1}{\boldsymbol{x} \cdot \boldsymbol{x}}, \ldots, \frac{x_p}{\boldsymbol{x} \cdot \boldsymbol{x}} \right) \\ \nabla \cdot \boldsymbol{g}(\boldsymbol{x}) &= (p-2) \sum \left(\frac{\boldsymbol{x} \cdot \boldsymbol{x} - 2x_j^2}{(\boldsymbol{x} \cdot \boldsymbol{x})^2} \right) \\ &= \frac{p-2}{(\boldsymbol{x} \cdot \boldsymbol{x})^2} [p \boldsymbol{x} \cdot \boldsymbol{x} - 2 \boldsymbol{x} \cdot \boldsymbol{x}] \\ &= \frac{(p-2)^2}{\boldsymbol{x} \cdot \boldsymbol{x}}, \end{aligned}$$

and so

$$\mathrm{E}[(\boldsymbol{X} - \boldsymbol{\theta}) \cdot g(\boldsymbol{X})] = \mathrm{E}[\tfrac{(p-2)^2}{\boldsymbol{X} \cdot \boldsymbol{X}}],$$

as we used above.

6 Some oracle inequalities

6.1 A simple oracle inequality

Recall that if we knew $||\boldsymbol{\theta}||^2$, the optimal estimator in the class $\{\delta_a(\boldsymbol{x}) = a\boldsymbol{x} : a \in [0, 1]\}$ would be $\delta_{\tilde{a}}$ where

$$\tilde{a} = \frac{||\boldsymbol{\theta}||^2/p}{||\boldsymbol{\theta}||^2/p + 1}.$$

We also showed that the risk of this estimator is

$$R(\boldsymbol{\theta}, \tilde{a}\boldsymbol{X}) = \mathrm{E}[||\tilde{a}\boldsymbol{X} - \boldsymbol{\theta}||^2]/p = \frac{||\boldsymbol{\theta}||^2}{||\boldsymbol{\theta}||^2 + p} < 1.$$

Not surprisingly, it turns out that

$$R(\boldsymbol{\theta}, \tilde{a}\boldsymbol{X}) \leq R(\boldsymbol{\theta}, \boldsymbol{\delta}_{JS})$$

(use the fact that $\boldsymbol{X} \cdot \boldsymbol{X}$ has a noncentral χ^2 distribution, or condition on $\boldsymbol{X} \cdot \boldsymbol{X}$). But how much worse is δ_{JS} than the oracle estimator $\delta_{\tilde{a}}$? Recall the risk of $\boldsymbol{\delta}_{JS}$ is

$$R(\boldsymbol{\theta}, \boldsymbol{\delta}_{JS}) = 1 - \frac{(p-2)^2}{p}\mathrm{E}[(\boldsymbol{X} \cdot \boldsymbol{X})^{-1}]$$

Since $1/x$ is convex, Jensen's inequality gives

$$\mathrm{E}[\frac{1}{\boldsymbol{X} \cdot \boldsymbol{X}}] \geq \frac{1}{\mathrm{E}[\boldsymbol{X} \cdot \boldsymbol{X})]} = \frac{1}{||\boldsymbol{\theta}||^2 + p},$$

and so

$$\begin{aligned}
R(\boldsymbol{\theta}, \boldsymbol{\delta}_{JS}) &\leq 1 - \frac{(p-2)^2}{p}\frac{1}{||\boldsymbol{\theta}||^2 + p} \\
p(R(\boldsymbol{\theta}, \boldsymbol{\delta}_{JS}) - R(\boldsymbol{\theta}, \boldsymbol{\delta}_{\tilde{a}})) &\leq p - \frac{(p-2)^2}{||\boldsymbol{\theta}||^2 + p} - \frac{||\boldsymbol{\theta}||^2 p}{||\boldsymbol{\theta}||^2 + p} \\
&= \frac{p||\boldsymbol{\theta}||^2 + p^2 - p^2 + 4p - 4 - ||\boldsymbol{\theta}||^2 p}{||\boldsymbol{\theta}||^2 + p} \\
&= \frac{4(p-1)}{||\boldsymbol{\theta}||^2 + p} \\
&\leq 4\tfrac{p-1}{p} \leq 4,
\end{aligned}$$

and so

$$R(\boldsymbol{\theta}, \boldsymbol{\delta}_{\tilde{a}})) \leq R(\boldsymbol{\theta}, \boldsymbol{\delta}_{JS}) \leq R(\boldsymbol{\theta}, \boldsymbol{\delta}_{\tilde{a}})) + 4/p.$$

Additional work can get you to

$$R(\boldsymbol{\theta}, \boldsymbol{\delta}_{\tilde{a}})) \leq R(\boldsymbol{\theta}, \boldsymbol{\delta}_{JS}) \leq R(\boldsymbol{\theta}, \boldsymbol{\delta}_{\tilde{a}})) + 2/p.$$

For more on this, see Johnstone [2002] or Candès [2006].

7 Unknown variance or covariance

Suppose $\boldsymbol{X} \sim N_p(\boldsymbol{\theta}, \sigma^2\mathbf{I})$, where σ^2 is known. Letting

- $\tilde{\boldsymbol{X}} = \boldsymbol{X}/\sigma$ and
- $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}/\sigma$,

we have $\tilde{\boldsymbol{X}} \sim N_p(\tilde{\boldsymbol{\theta}}, \mathbf{I})$. The James-Stein estimator $\tilde{\boldsymbol{\delta}}_{JS}$ of $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}/\sigma$ is then

$$\begin{aligned}\tilde{\boldsymbol{\delta}}(\tilde{\boldsymbol{X}}) &= (1 - \frac{p-2}{\boldsymbol{X}\cdot\boldsymbol{X}})\boldsymbol{X} \\ &= (1 - \frac{\sigma^2(p-2)}{\boldsymbol{X}\cdot\boldsymbol{X}})\boldsymbol{X}/\sigma.\end{aligned}$$

It seems natural then that the JSE of $\boldsymbol{\theta}$ should be σ times the JSE of $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}/\sigma$, giving

$$\boldsymbol{\delta}_{JS} = (1 - \frac{\sigma^2(p-2)}{\boldsymbol{X}\cdot\boldsymbol{X}})\boldsymbol{X}.$$

Of course, often σ^2 is not known. Is there a version of the JSE in this case? Yes, if you have information about σ^2. Consider the following hierarchical sampling scheme:

$$\begin{aligned}X_{i,j} &= \theta_j + \epsilon_{i,j}\ ,\ i = 1,\ldots,n,\ j = 1,\ldots,p \\ \{\epsilon_{i,j}\} &\sim \text{ i.i.d. } \mathrm{N}(0,\sigma^2).\end{aligned}$$

Letting $X_j = \bar{X}_{\cdot j} = \sum_{i=1}^n X_{i,j}/n$, we now basically have the situation described above.

Also note that the data contain information about σ^2 via the pooled sample sum of squares:

$$S = \sum_{j=1}^{p}(X_{i,j} - \bar{X}_{\cdot j})^2 \sim \sigma^2\chi^2_{p(n-1)}.$$

Note further that $\boldsymbol{X}$ and S are statistically independent. For this and similar situations, James and Stein [1961] considered estimators of the following form:

Let $\boldsymbol{X} \sim N_p(\boldsymbol{\theta}, \sigma^2\mathbf{I})$ be independent of $S \sim \sigma^2\chi^2_k$. Define the estimator

$$\delta_c(\boldsymbol{X}, S) = (1 - \tfrac{cS}{\boldsymbol{X}\cdot\boldsymbol{X}})\boldsymbol{X}.$$

The value of c that minimizes the risk of δ_c for all $\boldsymbol{\theta}$ is $c = (p-2)/(k+2)$, resulting in the following estimator:

$$\boldsymbol{\delta}_{JS}(\boldsymbol{X}, S) = (1 - \tfrac{S}{k+2}\tfrac{(p-2)}{\boldsymbol{X}\cdot\boldsymbol{X}})\boldsymbol{X}.$$

In particular, note that this estimator dominates $\boldsymbol{X}$.

This result also generalizes to the correlated data case: Let $\boldsymbol{X} \sim N_p(\boldsymbol{\theta}, \Sigma)$ and $\mathbf{S} \sim \text{Wishart}(\Sigma, k)$ be independent. Consider estimators of the form

$$\delta_c(\boldsymbol{X}, \mathbf{S}) = (1 - \tfrac{c}{\boldsymbol{X}^T\mathbf{S}^{-1}\boldsymbol{X}})\boldsymbol{X}.$$

James and Stein [1961] show that the estimator obtained by setting

$$c = \frac{p-2}{n-p+3}$$

minimizes the risk for all values of $\boldsymbol{\theta}$.

See Brown and Han for recent work on these and related problems, including extensions to regression problems.

References

L.D. Brown and X. Han. Optimal estimation of multidimensional normal means with an unknown variance.

Emmanuel J. Candès. Modern statistical estimation via oracle inequalities. *Acta Numer.*, 15:257–325, 2006. ISSN 0962-4929. doi: 10.1017/S0962492906230010. URL `http://dx.doi.org.offcampus.lib.washington.edu/10.1017/S0962492906230010`.

W. James and Charles Stein. Estimation with quadratic loss. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I*, pages 361–379. Univ. California Press, Berkeley, Calif., 1961.

IM Johnstone. Function estimation and gaussian sequence models. *Unpublished manuscript*, 2002.

E. L. Lehmann and George Casella. *Theory of point estimation*. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 1998. ISBN 0-387-98502-6.

Charles Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, vol. I*, pages 197–206, Berkeley and Los Angeles, 1956. University of California Press.