Contents

1 Motivation and definition **1**

2 Least favorable prior **3**

3 Least favorable prior sequence **11**

4 Nonparametric problems **15**

5 Minimax and admissibility **18**

6 Superefficiency and sparsity **19**

Most of this material comes from Lehmann and Casella [1998] section 5.1, and some comes from Ferguson [1967], section 2.11.

---

1 Motivation and definition

Let $X \sim \text{binomial}(n, \theta)$, and let $\bar{X} = X/n$. Via admissibility of unique Bayes estimators,

$$\delta_{w\theta_0}(X) = w\theta_0 + (1-w)\bar{X}$$

is admissible under squared error loss for all $w \in (0,1)$ and $\theta_0 \in (0,1)$.

$$\begin{aligned} R(\theta, \delta_{w\theta_0}) &= \text{Var}[\delta_{w\theta_0}|\theta] + \text{Bias}^2[\delta_{w\theta_0}|\theta] \\ &= (1-w)^2\text{Var}[\bar{X}] + w^2 \times (\theta - \theta_0)^2 \\ &= (1-w)^2\theta(1-\theta)/n + w^2 \times (\theta - \theta_0)^2. \end{aligned}$$

Risk functions for three such estimators are in Figure 1. Which one would you use? Requiring admissibility is not enough to identify a unique procedure.

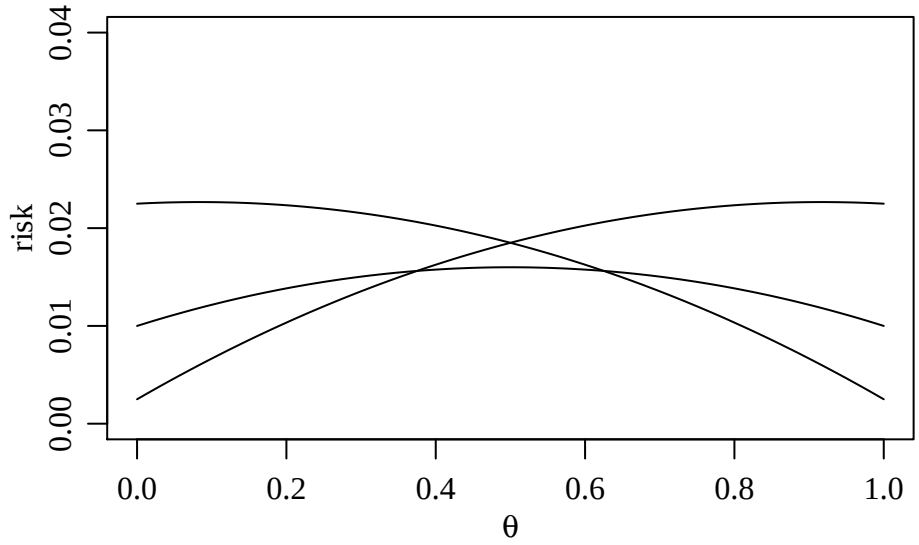

Figure 1: Three risk functions for estimating a binomial proportion when $n = 10$, $w = 0.8$ and $\theta_0 \in \{1/4, 2/4, 3/4\}$.

- This is good - we can require more of our estimator.
- This is bad - what more should we require?

One idea is to avoid a "worst case scenario", by choosing an estimator with lowest maximum risk.

Definition 1 (minimax risk, minimax estimator)**.** *The minimax risk is defined as*

$$R_m(\Theta) = \inf_{\delta} \sup_{\theta} R(\theta, \delta).$$

An estimator δ_m is a minimax estimator of θ if

$$\sup_{\theta} R(\theta, \delta_m) = \inf_{\delta} \sup_{\theta} R(\theta, \delta) = R_m(\Theta).$$

i.e. $\sup_\theta R(\theta, \delta_m) \leq \sup_\theta R(\theta, \delta)$ for all $\delta \in \mathcal{D}$.

2 Least favorable prior

Identifying a minimax estimator seems difficult: one would need to minimize the supremum risk over all estimators.
However, in many cases we can identify a minimax estimator using some intuition:

- Suppose some values of θ are harder to estimate than others.

  Example: For the binomial model $\text{Var}[X/n] = \theta(1-\theta)/n$, so $\theta = 1/2$ is hard to estimate well).

- Consider a prior $\pi(\theta)$ that heavily weights the "difficult" value of θ.

  The Bayes estimator δ_π will do well for these difficult values, where the supremum risk of most estimators is likely to occur.

  $R(\pi, \delta_\pi) = \int R(\theta, \delta_\pi)\pi(d\theta) \leq \int R(\theta, \delta)\pi(d\theta) = R(\pi, \delta)$

  This means that $R(\theta, \delta_\pi) \leq R(\theta, \delta)$ in places of high π-probability, i.e. the difficult parts of Θ.

  Since δ_π does well in the difficult region, maybe it is minimax.

Definition 2 (least favorable prior)**.** *A prior distribution π is least favorable if*

$$R(\pi, \delta_\pi) \geq R(\pi', \delta_{\pi'})$$

for all priors π' on Θ.

Intuition: δ_π is the best you can do under the worst prior.

Analogy: to competitive games.

- You get to choose an estimator δ, your adversary gets to choose a prior π.
- Your adversary wants you to have high loss, you want to have low loss.
- For any π, your best strategy is δ_π.
- Your adversary's best strategy is then for π to be least favorable.

Bounding the minimax risk:

The least favorable prior provides a lower bound on the minimax risk $R_m(\Theta)$.
For any prior π over Θ,

$$R(\pi,\delta) = \int R(\theta,\delta)\ \pi(d\theta) \le \int \left[\sup_\theta R(\theta,\delta)\right]\ \pi(d\theta) = \sup_\theta R(\theta,\delta).$$

Minimizing over all estimators $\delta \in \mathcal{D}$, we have

$$\inf_\delta R(\pi,\delta) \le \inf_\delta \sup_\theta R(\theta,\delta)$$
$$R(\pi,\delta_\pi) \le R_m(\Theta)$$

The Bayes risk of any prior gives a lower bound for minimax risk.
Maximizing over π gives sharpens the lower bound:

$$\sup_\pi R(\pi,\delta_\pi) \le R_m(\Theta).$$

Finding the LFP and minimax estimator:

For any prior π, we have shown

$$R(\pi,\delta_\pi) \le \inf_\delta \sup_\theta R(\theta,\delta).$$

On the other hand, for any estimator δ_π, we have

$$\inf_\delta \sup_\theta R(\theta,\delta) \le \sup_\theta R(\theta,\delta_\pi).$$

Putting these together gives

$$R(\pi,\delta_\pi) \le \inf_\delta \sup_\theta R(\theta,\delta) \le \sup_\theta R(\theta,\delta_\pi).$$

Therefore, if $R(\pi,\delta_\pi) = \sup_\theta R(\theta,\delta_\pi)$, then δ_π achieves the minimax risk.

Theorem 1 (LC 5.1.4)**.** *Let δ_π be (unique) Bayes for π, and suppose $R(\pi, \delta_\pi) = \sup_\theta R(\theta, \delta_\pi)$. Then*

1. *δ_π is (unique) minimax;*

2. *π is least favorable;*

Proof.

1. For any other estimator δ,

$$\begin{aligned}\sup_\theta R(\theta, \delta) &\geq \int R(\theta, \delta)\pi(d\theta) \\ &\geq \int R(\theta, \delta_\pi)\pi(d\theta) = \sup_\theta R(\theta, \delta_\pi).\end{aligned}$$

   If δ_π is unique Bayes under π, then the second inequality is strict and δ_π is unique minimax.

2. Let $\tilde{\pi}$ be a prior over Θ. Then

$$\begin{aligned}R(\tilde{\pi}, \delta_{\tilde{\pi}}) &= \int R(\theta, \delta_{\tilde{\pi}})\ \tilde{\pi}(d\theta) \\ &\leq \int R(\theta, \delta_\pi)\ \tilde{\pi}(d\theta) \\ &\leq \sup_\theta R(\theta, \delta_\pi) = R(\pi, \delta_\pi).\end{aligned}$$

□

Notes: regarding the condition $R(\pi, \delta_\pi) = \sup_\theta R(\theta, \delta_\pi)$,

- the condition is sufficient but not necessary for δ_π to be minimax.
- the condition is very restrictive - it is met only if $\pi(\theta : R(\theta, \delta_\pi) = \sup_{\theta'} R(\theta', \delta_\pi)) = 1$.

Bayes estimators that satisfy this condition are sometimes called "equalizer rules:"

Definition 3. *An estimator δ_π that is Bayes with respect to π is called an* equalizer rule *if* $\sup_\theta R(\theta, \delta_\pi) = R(\theta, \delta_\pi)$ *a.e.* π.

The theorem implies that equalizer rules are minimax:

Corollary 1 (LC cor 5.1.6)**.** *An equalizer rule is minimax.*

In the definition of an equalizer rule, it is not enough that $R(\theta, \delta_\pi)$ is constant a.e. π. For example, the supremum risk of a Bayes estimator could occur on a set of π-measure zero, and so $R(\pi, \delta_\pi) < \sup_\theta R(\theta, \delta_\pi)$, in which case the conditions of the theorem do not hold.
To summarize, a Bayes estimator δ_π will be minimax if $R(\pi, \delta_\pi) = \sup_\theta R(\theta, \delta_\pi)$. Conditions under which this occurs include the cases where

1. δ_π has constant risk, or
2. δ_π has constant risk a.e. π *and* achieves its maximum risk on a set of π-probability 1. Equivalently, $\pi(\{R(\theta, \delta_\pi) = \sup_\theta R(\theta, \delta_\pi)\}) = 1$.

Example(binomial proportion):
Consider estimation of θ with squared error loss based on $X \sim \text{binomial}(n, \theta)$. Let's try to find a Bayes estimator with constant risk. Such an estimator is minimax, by the theorem.
The easiest place to start is with the class of conjugate priors. Under $\theta \sim \text{beta}(a, b)$,

$$\begin{aligned} \delta_{ab}(X) = \text{E}[\theta|X] &= \frac{a+X}{a+b+n} \\ R(\theta, \delta_{ab}) &= \text{Var}[\delta_{ab}] + \text{Bias}^2[\delta_{ab}] \\ &= \frac{n\theta(1-\theta)}{(a+b+n)^2} + \frac{(a-\theta(a+b))^2}{(a+b+n)^2}. \end{aligned}$$

Can we make this constant as a function of θ? The numerator is

$$n\theta - n\theta^2 + (a+b)^2\theta^2 - 2a(a+b)\theta + c(a,b,n).$$

This will be constant in θ if

$$n = 2a(a+b)$$
$$n = (a+b)^2,$$

solving for a and b gives

$$a = b = \sqrt{n}/2.$$

Therefore,

- The estimator $\delta_{\sqrt{n}/2}(X) = \frac{X+\sqrt{n}/2}{n+\sqrt{n}} = \bar{X}\frac{\sqrt{n}}{\sqrt{n}+1} + \frac{1}{2}\frac{1}{\sqrt{n}+1}$ is
  - constant risk and Bayes, therefore
  - an equalizer rule, and therefore
  - minimax.
- beta$(\sqrt{n}/2, \sqrt{n}/2)$ is a least favorable prior.

Risk comparison

$$R(\theta, \delta_{\sqrt{n}/2}(X)) = \frac{1}{(2(1+\sqrt{n}))^2} \qquad R(\theta, \bar{X}) = \frac{\theta(1-\theta)}{n}$$

At $\theta = 1/2$ (a "difficult" value of θ), we have

$$R(1/2, \bar{X}) = \frac{1}{4}\frac{1}{n} > \frac{1}{4}\frac{1}{(n+2\sqrt{n}+1)} = R(1/2, \delta_{\sqrt{n}/2}(X)).$$

Notes:

- The region in which $R(\theta, \delta_{\sqrt{n}/2}(X)) < R(\theta, \bar{X})$ decreases in size with n.
- The least favorable prior is not unique: Under any prior π,

$$\delta_\pi(x) = \mathrm{E}[\theta|X] = \frac{\int \theta^{x+1}(1-\theta)^{n-x}\pi(d\theta)}{\int \theta^x(1-\theta)^{n-x}\pi(d\theta)},$$

  and so the estimator only depends on the first $n+1$ moments of θ under π.

- The minimax estimator is sensitive to the loss function: Under

$$L(\theta, \delta) = (\theta - \delta)^2/[\theta(1-\theta)],$$

X/n is minimax (it is an equalizer rule under under $\theta \sim$beta(1,1)).

Example (difference in proportions):
Consider estimation of $\theta_y - \theta_x$ based on $X \sim$ binomial(n, θ_x), $Y \sim$ binomial(n, θ_y). Can we derive the estimator from the one-sample case? Consider the difference of minimax estimators:

$$\frac{Y + \sqrt{n}/2}{n + \sqrt{n}} - \frac{X + \sqrt{n}/2}{n + \sqrt{n}} = \frac{Y - X}{n + \sqrt{n}}.$$

Unfortunately, it turns out that this is not minimax. However, it will turn out that the minimax estimator is of the form $\delta(X, Y) = c \times (Y - X)$ for some $c \in (0, 1/n)$. Starting from this vantage point, let's apply our strategy:

- Does $c(Y - X)$ have constant risk for some $c \in (0, 1/n)$?
- If so, is the constant risk estimator also Bayes?

The risk of such an estimator is

$$\begin{aligned} R(\theta, \delta_c) &= \text{Var}[\delta_c] + (\text{E}[c(Y - X)] - (\theta_y - \theta_x))^2 \\ &= c^2 n[\theta_x(1-\theta_x) + \theta_y(1-\theta_y)] + (cn - 1)^2(\theta_y - \theta_x)^2. \end{aligned}$$

By inspection, this will not be constant in (θ_x, θ_y) for any c, so the hope that δ_c will be a constant risk Bayes estimator fails. However, recall that the condition of the theorem does not require that δ_c be constant risk, just that it is Bayes with respect to a prior π such that δ_c

1. has constant risk a.e. π;
2. achieves its supremum risk on a set of π-probability 1.

With this in mind, maybe we can find a subset Θ_0 of the parameter space $\Theta = [0,1]^2$ such that δ_c is constant risk for some c. If so, then it will be constant risk a.e. π for any π such that $\pi(\Theta_0) = 1$, and possibly minimax.

Staring at the risk function $R(\theta, \delta_c)$ long enough suggests looking at the set $\Theta_0 = \{\theta : \theta_x + \theta_y = 1\} \subset \Theta$. On this set, $\theta_y(1-\theta_y) = \theta_x(1-\theta_x)$ and $\theta_y - \theta_x = 1 - 2\theta_x$, and the risk function reduces to

$$R(\theta, \delta_c) = c^2 n(2\theta_x(1-\theta_x)) + (cn-1)^2(1-2\theta_x)^2.$$

Is there a value of c for which δ_c is constant risk on Θ_0?

$$\begin{aligned}\frac{dR(\theta_x, \delta_c)}{d\theta_x} &= \frac{d}{d\theta_x} c^2 n(2(\theta_x - \theta_x^2)) + (cn-1)^2(1 - 4\theta_x + 4\theta_x^2) \\ &= c^2 n(2(1-2\theta_x)) - (cn-1)^2 4(1-2\theta_x).\end{aligned}$$

Solving for c, we have

$$\begin{aligned} c^2 n &= 2(cn-1)^2 \\ \pm\sqrt{n}c &= \sqrt{2}(cn-1) \\ \pm\sqrt{n/2} &= (n - 1/c) \\ 1/c &= \frac{\sqrt{2}n \pm \sqrt{n}}{\sqrt{2}} \\ c &= \frac{\sqrt{2}}{\sqrt{2}n \pm \sqrt{n}} = \frac{1}{n}\frac{\sqrt{2n}}{\sqrt{2n} \pm 1}\end{aligned}$$

Using the "$-$" solution could give estimators outside the parameter space, so let's consider the "$+$" solution:

$$\begin{aligned} c &= \frac{1}{n}\frac{\sqrt{2n}}{\sqrt{2n}+1} \\ \delta_c(X,Y) &= \frac{\sqrt{2n}}{\sqrt{2n}+1}(Y-X)/n,\end{aligned}$$

i.e. δ_c is a shrunken version of the UMVUE. By construction, this estimator has constant risk on $\{\theta : \theta_x + \theta_y = 1\}$. For it to be minimax, we need it to

1. be Bayes with respect to a prior on Θ_0;

2. achieve its supremum risk over Θ on Θ_0.

We'll check the second condition first: Recall the risk function is:

$$\begin{aligned} R(\theta, \delta_c) &= c^2 n[\theta_x(1-\theta_x) + \theta_y(1-\theta_y)] + (cn-1)^2(\theta_y - \theta_x)^2 \\ &= c^2 n(\theta_x(1-\theta_x) + \theta_y(1-\theta_y) + (\theta_y - \theta_x)^2), \end{aligned}$$

where we are using the fact that $2(cn-1)^2 = c^2 n$ for our risk-equalizing value of c. Taking derivatives, we have that the maximum risk occurs when

$$\begin{aligned} c^2 n[(1-2\theta_x) + (\theta_x - \theta_y)] &= c^2 n(1 - \theta_x - \theta_y) = 0 \\ c^2 n[(1-2\theta_y) + (\theta_y - \theta_x)] &= c^2 n(1 - \theta_x - \theta_y) = 0, \end{aligned}$$

which are both satisfied when $\theta_x + \theta_y = 1$. You can take second derivatives to show that such points maximize the risk.

So far, we have shown

- δ_c has constant risk on $\Theta_0 = \{\theta : \theta_x + \theta_y = 1\}$
- δ_c achieves its maximum risk on this set.

To show that it is minimax, what remains is to show that it is Bayes with respect to a prior π on Θ_0. If it is, then the condition of Theorem 1 will be met and δ_c will therefore be minimax.

To find a prior on Θ_0 for which δ_c is Bayes, it is helpful to note two things:

- On Θ_0, $Y + n - X \sim \text{binomial}(2n, \theta_y)$;
- On Θ_0, the Bayes estimator of $\theta_y - \theta_x = 2\theta_y - 1$ is given by

$$\mathrm{E}[2\theta_y - 1 | X, Y] = 2\mathrm{E}[\theta_y | X, Y] - 1.$$

Considering this last point, $\delta_c(x,y) = \frac{\sqrt{2n}}{\sqrt{2n}+1}(y-x)/n$ is Bayes for $2\theta_y - 1$ if

$$\begin{aligned}
\mathrm{E}[2\theta_y - 1|X,Y] = 2\mathrm{E}[\theta_y|X,Y] - 1 &= \frac{\sqrt{2n}}{\sqrt{2n}+1}(y-x)/n \\
\mathrm{E}[\theta_y|X,Y] &= \frac{1}{2}\frac{\sqrt{2n}}{\sqrt{2n}+1}\frac{y-x}{n} + \frac{1}{2} \\
&= \frac{y-x}{2n+\sqrt{2n}} + \frac{1}{2}\frac{2n+\sqrt{2n}}{2n+\sqrt{2n}} \\
&= \frac{y-x}{2n+\sqrt{2n}} + \frac{n+\sqrt{n/2}}{2n+\sqrt{2n}} \\
&= \frac{(y+n-x)+\sqrt{n/2}}{2n+\sqrt{2n}}.
\end{aligned}$$

Now writing $\tilde{Y} = Y + n - X$ and $\tilde{n} = 2n$ and using note 1 above (which shows $\tilde{Y} \sim \text{binom}(\tilde{n}, \theta_y)$), we see that the condition on the posterior expectation is that

$$\mathrm{E}[\theta_y|\tilde{Y}] = \frac{\tilde{Y}+a}{\tilde{n}+a+b}$$

where $a = b = \sqrt{n/2}$. A prior on Θ_0 that makes this true is the beta$(\sqrt{n/2}, \sqrt{n/2})$ prior on θ_y. Therefore, δ_c is a Bayes estimator under a prior on the set Θ_0, the set on which it achieves its maximum risk. To summarize:

1. δ_c is constant risk on Θ_0;
2. δ_c achieves its maximum risk over Θ on the subset Θ_0;
3. δ_c is Bayes for a prior on Θ_0.

Therefore, $\sup_\theta R(\theta, \delta_c) = R(\pi, \delta_c)$ for a prior π for which δ_c is Bayes. By the theorem, it follows that δ_c is minimax.

---

3 Least favorable prior sequence

In many problems there are no least favorable priors, and the main theorem from above is of no help.

Example(normal mean, known variance):
$X_1, \ldots X_n \sim$ i.i.d. normal(θ, σ^2) , σ^2 known.
$R(\theta, \bar{X}) = \sigma^2/n$ is constant risk, so seems potentially minimax.
But no prior π over Θ gives $\delta_\pi(X) = \bar{X}$.
You can also see that there is no least favorable prior: Under any prior π,

$$R(\pi, \delta_\pi) < R(\pi, \bar{X}) = \sigma^2/n,$$

but you can find priors whose Bayes risk is arbitrarily close to σ^2/n (i.e., the set of Bayes risks is not closed).

However, $\bar{X}$ is a limit of Bayes estimators: Let π_{τ^2} denote the $N(0, \tau^2)$ prior distribution on θ. As $\tau^2 \uparrow \infty$,

$$\delta_{\tau^2} = \tfrac{n/\sigma^2}{n/\sigma^2 + 1/\tau^2} \bar{X} \to \bar{X}$$
$$R(\pi_{\tau^2}, \delta_{\tau^2}) = \tfrac{1}{n/\sigma^2 + 1/\tau^2} \uparrow \sigma^2/n = R(\theta, \bar{X}).$$

More generally, consider

- (π_k, δ_k), a sequence of prior distributions such that $R(\pi_k, \delta_k) \uparrow R$;
- δ, an estimator such that $\sup_\theta R(\theta, \delta) = R$.

For each k, we have $R(\pi_k, \delta_k) \leq R_m$
However, we always have $R_m \leq \sup_\theta R(\theta, \delta)$.
Putting these together gives

$$R(\pi_k, \delta_k) \leq R_m \leq \sup_\theta R(\theta, \delta).$$

If $R(\pi_k, \delta_k) \uparrow \sup_\theta R(\theta, \delta)$, then we must have $R_m = \sup_\theta R(\theta, \delta)$, and so δ must be minimax.

Definition 4. *A sequence $\{\pi_k\}$ of priors is least favorable if for every prior π,*

$$R(\pi, \delta_\pi) \leq \lim_{k \to \infty} R(\pi_k, \delta_k).$$

Theorem 2 (LC thm 5.1.2)**.** *Let $\{\pi_k\}$ be sequence of prior distributions and δ an estimator such that $R(\pi_k, \delta_k) \to \sup_\theta R(\theta, \delta)$. Then*

1. *δ is minimax, and*
2. *$\{\pi_k\}$ is least favorable.*

Proof.

1. For any estimator δ',

$$\sup_\theta R(\theta, \delta') \geq R(\pi_k, \delta') \geq R(\pi_k, \delta_k)$$

$$\sup_\theta R(\theta, \delta') = \lim_{k\to\infty} \sup_\theta R(\theta, \delta') \geq \lim_{k\to\infty} R(\pi_k, \delta_k) = \sup_\theta R(\theta, \delta)$$

2. For any prior π,

$$R(\pi, \delta_\pi) \leq R(\pi, \delta) \leq \sup_\theta R(\theta, \delta) = \lim_{k\to\infty} R(\pi_k, \delta_k).$$

□

Example (normal mean, known variance):
Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \sim$ i.i.d. $N_p(\boldsymbol{\theta}, \sigma^2\mathbf{I})$, where σ^2 known. Let $\pi_k(\boldsymbol{\theta})$ be such that $\boldsymbol{\theta} \sim N_p(\mathbf{0}, \tau_k^2\mathbf{I})$, $\tau_k^2 \uparrow \infty$. Then the Bayes estimator under π_k is

$$\delta_k = \frac{n/\sigma^2}{n/\sigma^2 + 1/\tau^2}\bar{\boldsymbol{X}}.$$

Under average square-error loss,

$$R(\pi_k, \delta_k) = \frac{1}{n/\sigma^2 + 1/\tau_k^2} \uparrow \sigma^2/n = R(\boldsymbol{\theta}, \bar{\boldsymbol{X}}),$$

and so $\bar{\boldsymbol{X}}$ is minimax by Theorem 2.

Alert! This result should seem a bit odd to you - $\bar{\boldsymbol{X}}$ is inadmissible for $p \geq 3$. How can it be dominated if it is minimax? We even have the following theorem:

Theorem 3 (LC lemma 5.2.21)**.** *Any unique minimax estimator is admissible.*

Proof.
If δ is unique minimax, and δ' any other estimator, then

$$\sup_\theta R(\theta,\delta) < \sup_\theta R(\theta,\delta') \Rightarrow \exists \theta_0 : R(\theta_0,\delta) < R(\theta_0,\delta'),$$

so δ can't be dominated. Alternatively by contradiction, if δ' were to dominate δ then δ' would be minimax, contradicting that δ is unique minimax. □

So how can $\boldsymbol{X}$ be minimax and not admissible? The only possibility left by the theorem is that $\boldsymbol{X}$ is not unique minimax. In fact, for $p \geq 3$, estimators of the form

$$\delta(\boldsymbol{x}) = \left(1 - c(|\boldsymbol{x}|)\frac{\sigma^2(p-2)}{\boldsymbol{x}\cdot\boldsymbol{x}}\right)\boldsymbol{x}$$

are minimax as long as $c : \mathbb{R}^+ \to [0,2]$ is nondecreasing (LC theorem 5.5.5). For these estimators, the supremum risk occurs in the limit as $|\boldsymbol{\theta}| \to \infty$.

Example (normal mean, unknown variance):
Let $X_1,\ldots,X_n \sim$ i.i.d. $N(\theta,\sigma^2)$, $\sigma^2 \in \mathbb{R}^+$ unknown. Under squared error loss,

$$\sup_{\theta,\sigma^2} R((\theta,\sigma^2),\bar{X}) = \infty.$$

In fact, we can prove that every estimator has infinite maximum risk:

$$\begin{aligned}\sup_{\theta,\sigma^2} R((\theta,\sigma^2),\delta) &= \sup_{\sigma^2}\sup_{\boldsymbol{\theta}} R((\theta,\sigma^2),\delta) \\ &\geq \sup_{\sigma^2} \sigma^2/n = \infty,\end{aligned}$$

where the inequality holds because $\bar{X}$ is minimax in the known variance case. Therefore, every estimator is trivially minimax, with maximum risk of infinity.

Here is a not-entirely satisfying solution to this problem: Assume $\sigma^2 \leq M$, M known.

Applying exactly the same argument as above, we have

$$\begin{aligned}\sup_{\theta,\sigma^2} R((\theta,\sigma^2),\delta) &= \sup_{\sigma^2:\sigma^2\le M} \sup_{\theta} R((\theta,\sigma^2),\delta) \\ &\ge \sup_{\sigma^2:\sigma^2\le M} \sigma^2/n = M/n,\end{aligned}$$

and so $\bar{X}$ is minimax with maximum risk M/n.

Note that the value of M doesn't affect the estimator - $\bar{X}$ is minimax no matter what M is. Does this mean that $\bar{X}$ is the minimax estimator for $\sigma^2 \in \mathbb{R}^+$? Here are some thoughts:

- For $p = 1$, $\bar{X}$ is unique minimax for $\theta \in \mathbb{R}$, $\sigma^2 \in (0, M]$ for all M. For $\sigma^2 \in \mathbb{R}^+$, it is still minimax, although not unique.
- For $p > 2$, $\bar{\boldsymbol{X}}$ is minimax, but not unique minimax for $\sigma^2 \in (0, M]$ or even known σ^2.

A better "solution" to this "problem" is to change the loss function to $L((\theta,\sigma^2),d) = (\theta - d)^2/\sigma^2$, "standardized squared-error loss."

Exercise: Show that $\bar{\boldsymbol{X}}$ is minimax under standardized squared-error loss.

4 Nonparametric problems

The normal mean, unknown variance problem above gave some indication that we can deal with "nuisance parameters" (such as the variance σ^2) when obtaining a minimax estimator for a parameter of interest (such as the mean θ). What about more general nuisance parameters?

Let $X \sim P \in \mathcal{P}$.
Suppose we want to estimate $\theta = g(P)$, some functional of P.
Suppose δ is minimax for $\theta = g(P)$ when $P \in \mathcal{P}_0 \subset \mathcal{P}$.
When will it also be minimax for $P \in \mathcal{P}$?

Theorem (LC 5.1.15)**.** *If δ is minimax for θ under $P \in \mathcal{P}_0 \subset \mathcal{P}$ and*

$$\sup_{P\in\mathcal{P}_0} R(P,\delta) = \sup_{P\in\mathcal{P}} R(P,\delta),$$

then δ is minimax for θ under $P \in \mathcal{P}$.

Proof. For any other estimator δ',

$$\begin{aligned}\sup_{P\in\mathcal{P}} R(P,\delta') &\geq \sup_{P\in\mathcal{P}_0} R(P,\delta') \\ &\geq \sup_{P\in\mathcal{P}_0} R(P,\delta) = \sup_{P\in\mathcal{P}} R(P,\delta).\end{aligned}$$

□

Example: (Difference in binomial proportions)
$p(x,y|\theta) = \text{dbinom}(x,n,\theta_x) \times \text{dbinom}(y,n,\theta_y)$
$\mathcal{P} = \{p(x,y|\theta) : \theta = (\theta_x,\theta_y) \in [0,1]^2\}$
$\mathcal{P}_0 = \{p(x,y|\theta) : \theta = (1-\theta_y,\theta_y), \theta_y \in [0,1]\}$
For estimation of $\theta_y - \theta_x$ on $\mathcal{P}_0$, we showed that

- $c \times (Y-X)$, with $c = \frac{1}{n}\sqrt{2n}/(\sqrt{2n}+1)$, is constant risk on $\mathcal{P}_0$.
- $c \times (Y-X)$ is Bayes w.r.t. a beta$(\sqrt{n/2},\sqrt{n/2})$ prior on θ_y,

and so $c \times (Y-X)$ is minimax on $\mathcal{P}_0$. We then showed that $c \times (Y-X)/n$ achieved its maximum risk on $\mathcal{P}_0$, and so by the theorem it is minimax.

Example: (Population mean, bounded variance)
Let $\mathcal{P} = \{P : \text{Var}[X|P] \leq M\}$, M known.
Is $\bar{X}$ minimax for $\theta = \text{E}[X|P]$?
Let $\mathcal{P}_0 = \{\text{dnorm}(x,\theta,\sigma) : \theta \in \mathbb{R}, \sigma^2 \leq M\}$. Then

1. $\bar{X}$ is minimax for $\mathcal{P}_0$;
2. $\sup_{P\in\mathcal{P}_0} R(P,\bar{X}) = \sup_{P\in\mathcal{P}} R(P,\bar{X}) = M/n$.

Thus $\bar{X}$ is minimax (and unique minimax in the univariate case, as shown in LC 5.2).

Example: (Population mean, bounded range)

Let $X_1, \ldots, X_n \sim$ i.i.d. $P \in \mathcal{P}$, where $P([0,1]) = 1 \; \forall P \in \mathcal{P}$.

Least favorable perspective: Our previous experience tells us to find a minimax estimator that is best under the worst conditions. What are the worst conditions?

Guess: The most difficult situation is where each P is as "spread out" as possible.

Let $\mathcal{P}_0 = \{\text{binary}(x, \theta) : \theta \in [0,1]\}$.

As we've already derived, the minimax estimator for $\mathrm{E}[X_i|\theta] = \Pr(X_i = 1|\theta) = \theta$ based on $X_1, \ldots, X_n \sim$i.i.d. binary(θ) is

$$\delta(\boldsymbol{X}) = \frac{n\bar{X} + \sqrt{n}/2}{n + \sqrt{n}} = \frac{\sqrt{n}}{\sqrt{n}+1}\bar{X} + \frac{1}{\sqrt{n}+1}\frac{1}{2}.$$

To use the lemma to show this is minimax for $\mathcal{P}$, we need to show

$$\sup_{P \in \mathcal{P}} R(P, \delta) = \sup_{P \in \mathcal{P}_0} R(P, \delta).$$

Let's calculate the risk of δ for $P \in \mathcal{P}$:

$$\begin{aligned}
\mathrm{Var}[\delta(X)] &= \frac{n}{(\sqrt{n}+1)^2}\mathrm{Var}[X|P]/n = \mathrm{Var}[X|P]/(\sqrt{n}+1)^2 \\
\mathrm{E}[\delta(X)] &= \frac{\sqrt{n}}{\sqrt{n}+1}\theta + \frac{1}{\sqrt{n}+1}\frac{1}{2} \\
&= \theta - \frac{1}{\sqrt{n}+1}\theta + \frac{1}{2(\sqrt{n}+1)} \\
&= \theta + \frac{1/2 - \theta}{\sqrt{n}+1} \\
\mathrm{Bias}^2(\delta|P) &= \frac{(\theta - 1/2)^2}{(\sqrt{n}+1)^2},
\end{aligned}$$

and so

$$R(P, \delta) = \frac{1}{(\sqrt{n}+1)^2} \times [\mathrm{Var}[X|P] + (\theta - 1/2)^2].$$

Where does this risk achieve its maximum value? Note that

$$\sup_{P \in \mathcal{P}} R(P, \delta) = \sup_{\theta \in [0,1]} \sup_{P \in \mathcal{P}_\theta} R(P, \delta),$$

where $\mathcal{P}_\theta = \{P \in \mathcal{P} : \mathrm{E}[X|P] = \theta\}$. To do the inner supremum, note that for any $P \in \mathcal{P}_\theta$,

$$\begin{aligned}\mathrm{Var}[X|P] &= \mathrm{E}[X^2] - \mathrm{E}[X]^2 \\ &\leq \mathrm{E}[X] - \mathrm{E}[X]^2 \\ &= \mathrm{E}[X](1 - \mathrm{E}[X]) \\ &= \theta(1-\theta)\end{aligned}$$

with equality only if P is the binary(θ) measure. Therefore, for each θ supremum risk is attained at the binary(θ) distribution, and so the maximum risk of δ is attained on the submodel $\mathcal{P}_0$. To summarize,

$$\sup_{P\in\mathcal{P}} R(P,\delta) = \sup_{\theta\in[0,1]} \sup_{P\in\mathcal{P}_\theta} R(P,\delta) = \sup_{\theta\in[0,1]} \sup_{P\in\mathcal{P}_\theta\cap\mathcal{P}_0} R(P,\delta) = \sup_{P\in\mathcal{P}_0} R(P_0,\delta).$$

Thus δ is minimax on $\mathcal{P}_0$, achieves its maximum risk there, and so is minimax over all of $\mathcal{P}$.

Exercise: Adapt this result to the case that $P[a,b] = 1$ for all $P \in \mathcal{P}$, for arbitrary $-\infty < a < b < \infty$.

5 Minimax and admissibility

We already showed a unique minimax estimator can't be dominated:

Theorem (LC lemma 5.2.19)**.** *Any unique minimax estimator is admissible.*

This is reminiscent of our theorem about admissibility of unique Bayes estimators, which has a similar proof: If δ' were to dominate δ, then it would be as good as δ in terms of both Bayes risk and maximum risk, and so such a δ can't be either unique Bayes or unique minimax.

What about the other direction? We have shown in some important cases that admissibility of an estimator implies it is Bayes, or close to Bayes. Does admissibility imply minimax? The answer is yes for constant risk estimators:

Theorem (LC lemma 5.2.21)**.** *If δ is constant risk and admissible, then it is minimax.*

Proof.
Let δ' be another estimator.
Since δ is not dominated, there is a θ_0 s.t. $R(\theta_0, \delta) \leq R(\theta_0, \delta')$
But since δ is constant risk,

$$\sup_\theta R(\theta, \delta) = R(\theta_0, \delta) \leq R(\theta_0, \delta') \leq \sup_\theta R(\theta, \delta').$$

□

Exercise: Draw a diagram summarizing some of the relationships between admissible, Bayes and minimax estimators we've covered so far.

6 Superefficiency and sparsity

Let $X_1, \ldots, X_n \sim$ i.i.d. $N(\theta, 1)$, where $\theta \in \mathbb{R}$. A version of Hodges estimator for θ is given by

$$\delta_H(\mathbf{x}) = \begin{cases} \bar{X} & \text{if } |\bar{x}| > 1/n^{1/4} \\ 0 & \text{if } |\bar{x}| < 1/n^{1/4} \end{cases}$$

or more compactly, $\delta_H(\mathbf{x}) = \bar{x} \times 1(|\bar{x}| > 1/n^{1/4})$.

Exercise: Show that the asymptotic distribution of $\delta_H(\boldsymbol{x})$ is given by

$$\sqrt{n}(\delta_H(\boldsymbol{X}) - \theta) \dot{\sim} N(0, v(\theta)),$$

where

$$v(\theta) = \begin{cases} 1 & \text{if } \theta \neq 0 \\ 0 & \text{if } \theta = 0 \end{cases}$$

The asymptotic variance of $\sqrt{n}(\delta_H(\boldsymbol{X}) - \theta)$ makes δ_H seem useful in many situations: Often we are trying to estimate a parameter that could potentially be zero, or very

close to zero. For such parameters, δ_H seems to be "as good" as $\bar{X}$ asymptotically for all θ, and "much better" than $\bar{X}$ at the special value $\theta = 0$. In particular, for $\theta = 0$ we have

$$\Pr_0(\delta_H(\boldsymbol{X}) = 0) \to 1.$$

This seems much better than simple consistency at $\theta = 0$, which would require

$$\Pr_0(|\delta_H(\boldsymbol{X})| < \epsilon) \to 1.$$

for every $\epsilon > 0$. An estimator consistent at $\theta = 0$ never actually has to be zero, whereas the Hodges estimator is more and more likely to actually be zero as $n \to \infty$.

However, there is a price to pay. Let's compare the maximum of the risk function for δ_H to that of $\bar{X}$, under squared error loss. Letting $c_n = 1/n^{1/4}$, we have for any $\tilde{\theta} \in \mathbb{R}$,

$$\begin{aligned}
\sup_\theta \mathrm{E}_\theta[(\delta_H - \theta)^2] &\geq \mathrm{E}_{\tilde{\theta}}[(\delta_H - \tilde{\theta})^2] \\
&\geq \mathrm{E}_{\tilde{\theta}}[(\delta_H - \tilde{\theta})^2 1(|\bar{X}| < c_n)] \\
&= \mathrm{E}_{\tilde{\theta}}[\tilde{\theta}^2 1(|\bar{X}| < c_n)] \\
&= \tilde{\theta}^2 \Pr_{\tilde{\theta}}(|\bar{X}| < c_n) \\
&= \tilde{\theta}^2 \Pr_{\tilde{\theta}}(-c_n < \bar{X} < c_n) \\
&= \tilde{\theta}^2 \Pr_{\tilde{\theta}}(\sqrt{n}(-\tilde{\theta} - c_n) < \sqrt{n}(\bar{X} - \tilde{\theta}) < \sqrt{n}(-\tilde{\theta} + c_n)) \\
&= \tilde{\theta}^2 \times [\Phi(\sqrt{n}(-\tilde{\theta} + c_n)) - \Phi(\sqrt{n}(-\tilde{\theta} - c_n))].
\end{aligned}$$

Letting $\tilde{\theta} = \theta_0/\sqrt{n}$, we have for example

$$\sqrt{n}(-\tilde{\theta} + c_n) = \sqrt{n}(-\theta_0/\sqrt{n} + 1/n^{1/4}) = -\theta_0 + n^{1/4},$$

and so

$$\sup_\theta R(\theta, \delta_H) \geq \frac{\theta_0^2}{n} \times [\Phi(-\theta_0 + n^{1/4}) - \Phi(-\theta_0 - n^{1/4})].$$

Note that this holds for *any* $\theta_0 \in \mathbb{R}$. On the other hand, the risk of $\bar{X}$ is constant, $1/n$ for all θ. Therefore,

$$\frac{\sup_\theta R(\theta, \delta_H)}{\sup_\theta R(\theta, \bar{X})} \geq \theta_0^2 \times [\Phi(-\theta_0 + n^{1/4}) - \Phi(-\theta_0 - n^{1/4})].$$

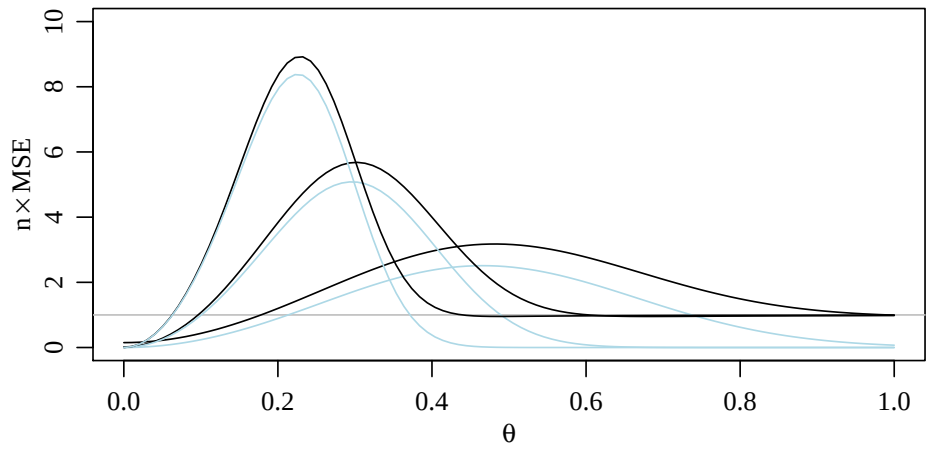

Figure 2: Risk functions of δ_H for $n \in \{25, 100, 250\}$, with the risk bound in blue.

You can play around with various values of θ_0 to see how big you can make this bound for a given n. The thing to keep in mind is that the inequality holds for all n and θ_0. As $n \to \infty$, the normal CDF term goes to one, so

$$\lim_{n\to\infty} \frac{\sup_\theta R(\theta, \delta_H)}{\sup_\theta R(\theta, \bar{X})} \geq \theta_0^2.$$

As this holds for all θ_0, we have

$$\lim_{n\to\infty} \frac{\sup_\theta R(\theta, \delta_H)}{\sup_\theta R(\theta, \bar{X})} = \infty.$$

Yikes! So the Hodges estimator becomes infinitely worse than $\bar{X}$ as $n \to \infty$.

Is this asymptotic result relevant for finite-sample comparisons of estimators? A semi-closed form expression for the risk of δ_H is given by Lehmann and Casella [1998, page 442]. A plot of the finite-sample risk of δ_H is shown in Figure 2.

Related to this calculation is something of more modern interest, the problem of

variable selection in regression. Consider the following model:

$$\begin{aligned} \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n &\sim \text{i.i.d.} P_X \\ \epsilon_1, \ldots, \epsilon_n &\sim \text{i.i.d.} N(0,1) \\ y_i &= \boldsymbol{\beta}^T \boldsymbol{X}_i + \epsilon_i. \end{aligned}$$

You may have attended one or more seminars where someone presented an estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ with the following property:

$$\Pr(\hat{\beta}_j = 0 | \boldsymbol{\beta}) \to 1 \text{ as } n \to \infty \; \forall \boldsymbol{\beta} : \beta_j = 0.$$

Again, such an estimator $\hat{\beta}_j$ is not just consistent at $\beta_j = 0$, it actually equals 0 with increasingly high probability as $n \to \infty$. This property has been coined "sparsistency" by people studying asymptotics of model selection procedures.

This special consistency at $\beta_j = 0$ seems similar to the properties of Hodges estimator when $\theta = 0$. What does the behavior at 0 imply about the risk function elsewhere? It is difficult to say anything about the entire risk function for all such estimators, but we can gain some general insight by looking at the supremum risk. Let $\boldsymbol{\beta}_n = \boldsymbol{\beta}_0/\sqrt{n}$, where $\boldsymbol{\beta}_0$ is arbitrary.

$$\begin{aligned} \sup_\beta \mathrm{E}_\beta[||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||^2] &\geq \mathrm{E}_{\beta_n}[||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_n||^2] \\ &\geq \mathrm{E}_{\beta_n}[||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_n||^2 \times 1(\hat{\boldsymbol{\beta}} = \mathbf{0})] \\ &= ||\boldsymbol{\beta}_n||^2 \Pr(\hat{\boldsymbol{\beta}} = \mathbf{0} | \boldsymbol{\beta}_n) = \frac{1}{n} ||\boldsymbol{\beta}_0||^2 \Pr(\hat{\boldsymbol{\beta}} = \mathbf{0} | \boldsymbol{\beta}_n) \end{aligned}$$

Now consider the risk of the OLS estimator:

$$\mathrm{E}_\beta[||\hat{\boldsymbol{\beta}}_{\text{ols}} - \boldsymbol{\beta}||^2] = \mathrm{E}_{P_X}[\mathrm{tr}((\mathbf{X}^T\mathbf{X})^{-1})] \equiv v_n$$

So we have

$$\frac{\sup_\beta R(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}})}{\sup_\beta R(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}_{\text{ols}})} \geq ||\boldsymbol{\beta}_0||^2 \frac{\Pr(\hat{\boldsymbol{\beta}} = \mathbf{0} | \boldsymbol{\beta}_n)}{n v_n},$$

which holds for all n and all $\boldsymbol{\beta}_0$. Now you should believe that nv_n converges to some number $v_\infty > 0$. What about $\Pr(\hat{\boldsymbol{\beta}} = \mathbf{0} | \boldsymbol{\beta}_n)$? Now $\boldsymbol{\beta}_n \to 0$ as $n \to \infty$, and

so it seems that since $\hat{\boldsymbol{\beta}}$ has the sparsistency property, eventually $\hat{\boldsymbol{\beta}}$ will equal $\mathbf{0}$ with high probability. Next quarter you will learn about contiguity of sequences of distributions and will be able to show that

$$\lim_{n\to\infty} \Pr(\hat{\boldsymbol{\beta}} = \mathbf{0} | \boldsymbol{\beta} = \boldsymbol{\beta}_n) = \lim_{n\to\infty} \Pr(\hat{\boldsymbol{\beta}} = \mathbf{0} | \boldsymbol{\beta} = \mathbf{0}) = 1.$$

The first equality is due to contiguity, the second to the sparsistency property of $\hat{\boldsymbol{\beta}}$. This result gives

$$\lim_{n\to\infty} \frac{\sup_\beta R(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}})}{\sup_\beta R(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}_{\text{ols}})} \geq ||\boldsymbol{\beta}_0||^2 / v_\infty.$$

Since this result holds for all $\boldsymbol{\beta}_0$, the limit is in fact infinite. The result is that, in terms of supremum risk (under squared error loss) a sparsistent estimator becomes infinitely worse than the OLS estimator. This result is due to Leeb and Pötscher [2008] (see also Pötscher and Leeb [2009]).

Discuss:

- Is supremum risk an appropriate comparison?
- Is squared error an appropriate loss?
- When would you use a sparsistent estimator?

References

Thomas S. Ferguson. *Mathematical statistics: A decision theoretic approach*. Probability and Mathematical Statistics, Vol. 1. Academic Press, New York, 1967.

Hannes Leeb and Benedikt M Pötscher. Sparse estimators and the oracle property, or the return of hodges estimator. *Journal of Econometrics*, 142(1):201–211, 2008.

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.

Benedikt M. Pötscher and Hannes Leeb. On the distribution of penalized maximum likelihood estimators: the LASSO, SCAD, and thresholding. *J. Multivariate Anal.*, 100(9):2065–2082, 2009. ISSN 0047-259X. doi: 10.1016/j.jmva.2009.06.010. URL http://dx.doi.org/10.1016/j.jmva.2009.06.010.