STAT5010 Advanced Statistical Inference **Fall 2023**

Lecture 7: Minimaxity

Lecturer: Tony Sit *Scribe: Yikun Zhao, Zhuogen Wu*

7.1 Maximaxity with submodel restriction

Example: Minimax for $i.i.d$ normal random variables with unknown mean and variance, i.e.

$$X_1, ..., X_n \overset{i.i.d}{\sim} N(\theta, \sigma^2)$$

with both θ and σ^2 unknown.
Note that

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

We restrict attention to $\Omega = \{(\theta, \sigma^2) : \theta \in \mathbb{R}^2, \sigma^2 \leq B\}$ where B is a known constant.
Assume that δ is any other estimator. The risk Of $\bar{X}_n$ with this family is given by

$$\sup_{\theta\in\mathbb{R}^2,\sigma^2\leq B} R((\theta, \sigma^2), \bar{X}_n) = \frac{B}{n} \tag{7.1}$$

$$= \sup_{\theta\in\mathbb{R}^2,\sigma^2=B} R((\theta, \sigma^2), \bar{X}_n) \tag{7.2}$$

$$\leq \sup_{\theta\in\mathbb{R}^2,\sigma^2=B} R((\theta, \sigma^2), \delta) \tag{7.3}$$

$$\leq \sup_{\theta\in\mathbb{R}^2,\sigma^2\leq B} R((\theta, \sigma^2), \delta) \tag{7.4}$$

It shows that $\bar{X}_n$ is minimax over Ω by focusing on the case where σ^2 is known.
Example:(TPE 5.1.16) Suppose $X_1, X_2, ..., X_n$ are $i.i.d$ with Gamma cdf $F()$ with mean $\mu(F) < \infty$ and variance $\sigma^2(F) < \infty$.
Our goal is to find a minimax estimate of $\mu(F)$ under the square loss.
Without further restriction on F, the worst case risk is unbounded for every estimator, So every estimator is minimax. Hence, we impose additional constraints. **Constraint A:** Assume that $\sigma^2(P) \leq B$, We have seen that $\bar{X}_n$ is minimax for Gaussian submodel case
First we compute the supermum risk for the full model

$$R(F, \bar{X}_n) = \frac{1}{n}\sum_{i=1}^{n} E(X_i - \mu(F))^2 = \frac{\sigma^2(F)}{n}$$

Since $\sigma^2(F) \leq B$ by assumption, we get

$$\sup_{F:\sigma^2(F)\leq B} R(F, \bar{X}) = \frac{B}{n}$$

As the submodel $\mathcal{F}_0 = N(\mu, \sigma^2)$ where $\sigma^2 \leq B$, $\bar{X}_n$ is minimax. The supermum risk in this case is identical to that of the full model.

$$\sup_{F \in \mathcal{F}_0} R(F, \bar{X}_n) = \frac{B}{n}$$

Lemma (TPE 5.1.15): Suppose that δ is minimax for a submodel $\theta \in \Omega_0 \subset \Omega$, and $\sup_{\theta \in \Omega_0} R(\theta, \delta) = \sup_{\theta \in \Omega} R(\theta, \delta)$, then δ is minimax for the full model $\theta \in \Omega$

We conclude that $\bar{X}_n$ is minimax for the full model. The non-parametric model is still constrained to have $\sigma^2(F) \leq B$

Constrained B: Assume $F \in \mathcal{F}$, where $\mathcal{F}$ is set of all CDFs with support constrained in $[0, 1]$(Question: Is $\bar{X}_n$ is minimax?). First consider the submodel $\mathcal{F}_0 = \{Bernoulli(\theta)\}_{\theta \in (0,1)}$. Let $Y = \sum_{i=1}^n X_i$, so $Y \sim Binomial(n, \theta)$, $\bar{X}_n = \frac{Y}{n}$. Recall the previous lecture that minimax estimator for $\mu(F) = \theta$ is this case is

$$\delta^*(X) = \frac{\sqrt{n}}{1+\sqrt{n}} \bar{X}_n + \frac{1}{2}(\frac{1}{1+\sqrt{n}})$$

which has supermum risk $\frac{1}{4(1+\sqrt{n})^2}$ So,

$$\sup_{\theta} R(\theta, \bar{X}_n) = \frac{1}{4n} > \frac{1}{4(1+\sqrt{n})^2} = \sup)_\theta R(\theta, \delta^*)$$

We can conclude $\bar{X}_n$ is not minimax.

Question 2: Can we find the minimax estimator under the full model?

We conjecture that $\delta^*(X)$ is minimax. We need to show that the supermum risk of $\delta^*(X)$ under the full model is not greater than $\frac{1}{4(1+\sqrt{n})^2}$. To see this, we compute

$$\begin{aligned}
E_F(\{\delta^*(X) - \mu(F)\}^2) &= E_F(\{\frac{\sqrt{n}}{1+\sqrt{n}}(\bar{X}_n - \mu(F)) + \frac{1}{1+\sqrt{n}}(\frac{1}{2} - \mu(F)\}^2) \\
&= (\frac{1}{1+\sqrt{n}})^2\{nVar(\bar{X}_n) + (\frac{1}{2} - \mu(F))^2\} \\
&= (\frac{1}{1+\sqrt{n}})^2\{E(X_1^2) - (\mu(F))^2 + \frac{1}{4} - \mu(F) + (\mu(F))^2\} \\
&= (\frac{1}{1+\sqrt{n}})^2\{E(X_1^2) + \frac{1}{4} - \mu(F)\}
\end{aligned}$$

By assumption $\bar{X}_n \in [0, 1]$, So $X_1 \leq X_1$ and we can bound the risk. So δ^* is minimax for the binomial submodel, and its worst case risk in the same for the full model and for the binomial submodel. By Lemma TPE 5.1.15, we conclude that δ^*, is minimax.

Admissibility of minimax estimators.

1. If δ is admissible with constant risk. then δ is also minimax. (Exercise; see TPE for help)
2. Minimaxity does not guarantee admissibility.

Example: Let $X_1, ..., X_n \overset{i.i.d}{\sim} N(\theta, \sigma^2)$, where σ^2 is known. The parameter θ is the estimated. Then the minimax estimator is $\bar{X}_n$ under the squared error loss. We want to see if $\bar{X}_n$ is admissible. A more general restriction of this estimator: $a\bar{X}_n + b$ and $a, b \in \mathbb{R}$.

Case 1: If $0 < a < 1$, $a\bar{X}_n + b$ is a convex combination of $\bar{X}_n$ and b. It is a Bayes estimator w.r.t some Gaussian prior on θ. Since square error loss is strictly convex, the Bayes estimator is unique.

So by TPE 5.2(unique Bayes estimator is admissible), $a\bar{X}_n + b$ is admissible is this case.

Case 2: a=0. In this case b is also a unique Bayes estimator with respect to a degenerate prior distribution with unit mass $\theta = b$. So by Theorem 5.2.4, b is admissible.

Case 3: a=1, b$\neq$ 0. In this case, $\bar{X} + b$ is not admissible because it is dominated by $\bar{X}$. To see this, note that $\bar{X}$ has the same variance as $\bar{X} + b$, but strictly smaller bias.

The next few cases use the following result. In general, the risk of $a\bar{X}+b$ is:

$$\begin{aligned}\mathbb{E}\left[(a\bar{X}+b-\theta)\right]^2 &= \mathbb{E}\left[\left(a(\bar{X}-\theta)+b+\theta(a-1)\right)^2\right] \\ &= \frac{a^2\sigma^2}{n}+(b+\theta(a-1))^2\end{aligned} \tag{7.5}$$

where, in the first step, we added and subtracted $a\theta$ inside.
Case 4: $a>1$. If we apply the result for the general risk we have:

$$\mathbb{E}\left[\left(a\bar{X}+b-\theta\right)^2\right] \geq \frac{a^2\sigma^2}{n} > \frac{\sigma^2}{n} = R(\theta,\bar{X}). \tag{7.6}$$

The first inequality follows because the second summand in the expression for the general risk is always non-negative. $\bar{X}$ dominates $a\bar{X}+b$ when a>1, and so in this case $a\bar{X}+b$ is inadmissible.
Case 5: $a<0$.

$$\begin{aligned}\mathbb{E}\left[\left(a\bar{X}+b-\theta\right)^2\right] &> (b+\theta(a-1))^2 \\ &= (a-1)^2\left(\theta+\frac{b}{a-1}\right)^2 \\ &> \left(\theta+\frac{b}{a-1}\right)^2.\end{aligned} \tag{7.7}$$

and this is the risk of predicting the constant $-b/(a-1)$. So, $-b/(a-1)$ dominates $a\bar{X}+b$, and therefore, $a\bar{X}+b$ is again inadmissible.

Now, we have considered every case except for the estimator $\bar{X}$. It turns out that $\bar{X}$. The argument in this case is more involved, and proceeds by contradiction.
Case 6: a=1, b=0. Here, we use a limiting Bayes argument. Suppose $\bar{X}$ is admissible. Then, assuming w.l.o.g that $\sigma^2=1$, we have:

$$R(\theta,\bar{X}) = \frac{1}{n} \tag{7.8}$$

By our hypothesis, there must exist an estimator δ' such that $R(\theta,\delta')\leq 1/n$ for all θ and $R(\theta',\delta')<1/n$ for at least one $\theta'\in\Omega$. Because $R(\theta,\delta)$ is continuous in θ, there must exist $\epsilon>0$ and an interval (θ_0,θ_1) containing θ' so that:

$$R(\theta,\delta') = \frac{1}{n}-\epsilon \quad \forall\theta\in(\theta_0,\theta_1). \tag{7.9}$$

Let r'_τ be the average risk of δ' with respect to the prior distribution $N(0,\tau^2)$ on θ. (Note that this is the exact same prior we used to prove that $\bar{X}$ was the limit of a Bayes estimator, and hence minimax. We did this by letting $\tau\to\infty$, and therefore letting our prior tend to the improper prior $\pi(\theta)=1, \quad \forall\theta$.) Let r_τ be the average risk of a Bayes estimator δ_τ under the same prior.

Note that $\delta_\tau\neq\delta'$ because $R(\theta,\delta_\tau)\to\infty$ as $\theta\to\infty$ which is not consistent with $R(\theta,\delta')\leq 1/n$ for all $\theta\in\mathbb{R}$. So, $r_\tau<r'_\tau$, because the Bayes estimator is unique almost surely with respect to the marginal distribution of θ. We will look at the following ratio, which is selected to simplify our algebra later. This ratio, we will show, will become arbitrarily large, which we will use to form a contradiction with $r_\tau<r'_\tau$.

Using the form of the Bayes risk r_τ computed in a previous lecture(see TPE Example 5.1.14), we can write:

$$\frac{\frac{1}{n}-r'_\tau}{\frac{1}{n}-r_\tau} = \frac{\frac{1}{\sqrt{2\pi}\tau}\int_{-\infty}^{\infty}[\frac{1}{n}-R(\theta,\delta')]exp\left(-\frac{\theta^2}{2\tau^2}\right)d\theta}{\frac{1}{n}-\frac{1}{n+\frac{1}{\tau^2}}} \tag{7.10}$$

We find:

$$\frac{\frac{1}{n}-r'_\tau}{\frac{1}{n}-r_\tau} \geq \frac{\frac{1}{\sqrt{2\pi}\tau}\int_{\theta_0}^{\theta_1}\epsilon exp\left(-\frac{\theta^2}{2\tau^2}\right)d\theta}{\frac{1}{n(1+n\tau^2)}} = \frac{n(1+n\tau^2)}{\tau\sqrt{2\pi}}\epsilon\int_{\theta_0}^{\theta_1}exp\left(-\frac{\theta^2}{2\tau^2}\right)d\theta. \tag{7.11}$$

As $\tau \to \infty$, the first expression, $n(1+n\tau^2)\epsilon/(\tau\sqrt{2\pi}) \to \infty$ and since the integrand converges monotonically to 1, Lebesgue's monotone convergence theorem ensures that the integral approaches the positive quantity $\theta_1 - \theta_0$. So, for sufficiently large τ, we must have

$$\frac{\frac{1}{n} - r'_\tau}{\frac{1}{n} - r_\tau} > 1. \tag{7.12}$$

This means that $r'_\tau < r_\tau$. However, this is a contradiction, because r_τ is the optimal average risk (since it is the Bayes risk). So by our assumption that there was a dominating estimator was false, and in this case, $a\bar{X} + b = \bar{X}$ is admissible.

7.2 Simultaneous Estimation

Up to this point, we have considered only situations where a single real-valued parameter is of interest. However, in practice, we often care about several parameters, and wish to estimate them all at once. In this section, we consider the admissibility of estimators of several parameters-that is, of simultaneous estimation.

Example 4. Let $X_1, ...X_p$ be independent with $X_i \sim N(\theta_i, \sigma^2)$ for $1 \leq i \leq p$. For the sake of simplicity, say $\sigma^2 = 1$. Now our goal is to estimate $\theta = (\theta_1, \theta_2, ..., \theta_p)$ under the loss function:

$$L(\theta, d) = \sum_{i=1}^{p} (d_i - \theta_i)^2. \tag{7.13}$$

A natural estimator of θ is $X = (X_1, X_2, ..., X_p)$. It can be shown that X is the UMRUE, the maximum likelihood estimator, a generalized Bayes estimator, and a minimax estimator for θ. So, it would be natural to think that X is admissible. However, counter intuitively, it turns out that this is not the case when $p \geq 3$.

When $p \geq 3$, X is dominated by the James-Stein estimator (and that too, strictly dominated):

$$\begin{aligned} \delta(X) &= (\delta_1(X), \delta_2(X), ..., \delta_p(X)) \quad \text{where} \\ \delta_i(X) &= \left(1 - \frac{p-2}{||X||_2^2}\right) X_i. \end{aligned} \tag{7.14}$$

The J-S estimator makes use of the entire data vector when estimating each θ_i, so it is surprising that this is beneficial given the assumption of independence amongst the components of X. It turns out that the James-Stein estimator is not itself admissible because it is dominated by the positive part James-Stein estimator:

$$\delta_i(X) = \max\left(1 - \frac{p-2}{||X||_2^2}, 0\right) X_i. \tag{7.15}$$

To add insult to injury, even this estimator can be shown inadmissible, although that proof is non-constructive.

7.2.1 Motivation for the J-S estimator

To motivate the J-S estimator, we consider how it can arise in an empirical Bayes framework. The empirical Bayes approach (which builds on principles of Bayesian estimation, but is not strictly Bayesian) is a two-step process:

- Introduce a prior family indexed by a hyperparameter (this is the Bayesian aspect).

- Estimate the hyperparameter from the data. (this is the empirical aspect).

So applying this procedure to the problem at hand:

- Suppose $\theta_i \overset{iid}{\sim} N(0, A)$ then the Bayes estimator for θ_i is

$$\delta_{A,i}(X) = \frac{X_i}{1 + \frac{1}{A}} = \left(1 - \frac{1}{A+1}\right) X_i. \tag{7.16}$$

- In this step we must choose A. Marginalizing over θ, we see that X has the distribution,

$$X_i \overset{iid}{\sim} N(0, A+1). \tag{7.17}$$

We will use X and the knowledge of this marginal distribution to find an estimate of $\frac{1}{A+1}$. One could, in principle, use any estimate of A, and it is common to use a maximum likelihood estimate, but here we will used an unbiased estimate.

It can then be shown that

$$\mathbf{E}\left[\frac{1}{||X||_2^2}\right] = \frac{1}{(p-2)(A+1)}. \tag{7.18}$$

So

$$1 - \frac{p-2}{||X||_2^2} \tag{7.19}$$

must be UMVU for $1 - \frac{1}{A+1}$. If we plug this estimator into our Bayes estimator we obtain the J-S estimator:

$$\delta(X_i) = \left(1 - \frac{p-2}{||X||_2^2}\right) X_i. \tag{7.20}$$

7.2.2 James-Stein domination

Intuitively, the problem with the estimate X is that $||X||_2^2$ is typically much larger than $||\theta||$.

$$\mathbb{E}[||X||_2^2] = \mathbb{E}\left[\sum_{i=1}^{p} X_j^2\right] = p + \sum_{i=1}^{p} \theta_i^2 = p + ||\theta||_2^2 \tag{7.21}$$

where p is actually $\sigma^2 p = p$ in this case. So, we may view the J-S estimator as a method for correcting the bias in the size of X. It achieves this by shrinking each coordinate of X towards 0. The uniform superiority of the J-S estimator to X can be formalized.

Theorem 1. The James-Stein estimator δ has uniformly smaller risk than X if $p \geq 3$. The proof, given on page 335 of TPE, compare the risk of the J-S estimator directly to that of X.

References

[AGM97] N. Alon, Z. Galil and O. Margalit, On the Exponent of the All Pairs Shortest Path Problem, *Journal of Computer and System Sciences* **54** (1997), pp. 255–262.

[F76] M. L. Fredman, New Bounds on the Complexity of the Shortest Path Problem, *SIAM Journal on Computing* **5** (1976), pp. 83-89.