Lecture 11 — October 27

Lecturer: Lester Mackey *Scribe: Viswajith Venugopal, Vivek Bagaria, Steve Yadlowsky*

Warning: *These notes may contain factual and/or typographic errors.*

11.1 Summary

In this lecture, we will discuss the identification of minimax estimators via submodels, the admissibility of minimax estimators, and simultaneous estimation and the James-Stein estimator. This will conclude our discussion of estimation; in the future we will be focusing on the decision problem of hypothesis testing.

11.2 Minimax Estimators and Submodels

Recall that an estimator δ^M is minimax if its maximum risk is minimal:

$$\inf_{\delta} \sup_{\theta \in \Omega} R(\theta, \delta) = \sup_{\theta} R(\theta, \delta^M)$$

We saw how to derive the minimax estimator using least favourable priors in Lecture 10. In this lecture we will consider a different approach, based on the following Lemma:

Lemma 1 (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, $\theta \in \Omega$.

This lemma allows us to find a minimax estimator for a particular tractable submodel, and then show that the worst-case risk for the full model is equal to that of the submodel (that is, the worst-case risk doesn't rise as you go to the full model). In this case, using the Lemma, we can argue that the estimator we found is also minimax for the full model. This was similar to how we justified minimaxity of the estimator of a Normal mean with bounded variance last lecture.

Here's a fairly simple example:

Example 1. Let $X_1, \ldots, X_n$ be i.i.d $\mathcal{N}(\mu, \sigma^2)$, where both μ and σ^2 are unknown. Thus, our parameter vector, $\theta = (\mu, \sigma^2)$ and our parameter space $\Omega = \mathbb{R} \times \mathbb{R}^+$. Our task now is to estimate μ. Our loss function is the relative squared error loss, given by:

$$L((\mu, \sigma^2), d) = \frac{(d - \mu)^2}{\sigma^2}$$

We consider this loss function to make the question of minimaxity more interesting: regular squared error loss is unbounded for the full model, since it is proportional to the variance, which is unbounded.

We consider the submodel where $\sigma^2 = 1$. That is, $\Omega_0 = \mathbb{R} \times \{1\}$, and our loss function simplifies to our usual squared error loss: $L((\mu, 1), d) = (d - \mu)^2$. We saw in Example 1 of Lecture 10 that under this loss $\bar{X}$ is minimax for Ω_0. Moreover,

$$R((\mu, \sigma^2), \bar{X}) = \frac{1}{n} \quad \forall (\mu, \sigma^2) \in \Omega.$$

Thus, the risk does not depend on σ^2. Since $R((\mu, 1), \bar{X})) = R((\mu, \sigma^2), \bar{X}))$, we have that the maximum risks are equal. (That is, $\sup_{\theta \in \Omega_0} R(\theta, \delta) = \sup_{\theta \in \Omega} R(\theta, \delta)$. Therefore, it follows from Lemma 1 that $\bar{X}$ is minimax on Ω. Note that, thanks to our new loss function, we don't need to impose boundedness on our variance (like we did in our previous lecture) to establish minimaxity in a meaningful way.

This example is parametric, like a lot of the examples we've made so far. Assuming we know the form of the distribution for the variables, and that the variables are i.i.d., are both strong assumptions. Now, we consider a more ambitious example, which is in a non-parametric setting, and hence more general.

Example 2 (TPE Example 5.1.16)**.** Suppose $X_1, X_2, \ldots, X_n$ are i.i.d with common 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 squared error loss.

Without further restriction on F, the worst case risk is unbounded for every estimator, so every estimator is minimax. We will impose further constraints, and restrict our family somehow to have finite worst-case risk, to ensure that meaningful minimax estimators can be obtained.

Constraint (a). Assume $\sigma^2(F) \leq B$. Now, we've seen in the previous lecture that $\bar{X}$ is minimax for the Gaussian submodel in this case. So a natural guess for us to make is that $\bar{X}$ is minimax. We verify this by application of Lemma 1. First, we compute the supremum risk for the full model:

$$R(F, \bar{X}) = \frac{1}{n^2} \sum_i \mathbb{E}(X_i - \mu(F))^2 = \frac{\sigma^2(F)}{n}.$$

Since $\sigma^2(F) \in [0, B]$ by assumption, we get:

$$\sup_F R(F, \bar{X}) = \frac{B}{n}$$

Now we saw in Lecture 10 that for the submodel $\mathcal{F}_0 = \mathcal{N}(\mu, \sigma^2)$ when $\sigma^2 \leq B$, $\bar{X}$ is minimax. Further, the supremum risk in this case is identical to that of the full model:

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

Thus, using Lemma 1 we conclude that $\bar{X}$ is minimax for the full model. (That is, the non-parametric model still constrained to have $\sigma^2(F) \leq B$.)

Constraint (b). Assume $F \in \mathcal{F}$ where $\mathcal{F}$ is the set of all CDFs with support contained in $[0,1]$. Is $\bar{X}$ minimax for this model? We have reason to believe that it is not, based on the minimax estimator we derived in Lecture 9 for the Binomial submodel. And in fact, it turns out that $\bar{X}$ isn't minimax.

To show this, first consider the submodel, $\mathcal{F}_0 = \{\text{Ber}(\theta)\}_{\theta \in (0,1)}$. Let $Y = \sum_{i=1}^n X_i$ so that $Y \sim \text{Bin}(n, \theta)$ and $\bar{X} = Y/n$. Recall from Lecture 9 that the minimax estimator for $\mu(F) = \theta$, in the Binomial case, is:

$$\delta(X) = \frac{\sqrt{n}}{1+\sqrt{n}}\bar{X} + \frac{1}{2}\left(\frac{1}{1+\sqrt{n}}\right)$$

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

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

Thus, $\bar{X}$ has a higher worst-case risk than $\delta(X)$ as defined above, and hence, we have shown that $\bar{X}$ is not minimax.

Now, let's get more ambitious, and try to see if we can find the minimax estimator under the full model. We know that this can't be $\bar{X}$, but it's possible that it could be $\delta(X)$. To examine this possibility, we conjecture that $\delta(X)$ is also minimax under the full model. If we are to establish this under the Lemma, we need to show that the supremum risk of $\delta(X)$ under the full model is no more than $\frac{1}{4(1+\sqrt{n})^2}$ (which is the supremum risk for the binomial submodel).

Let us compute:

$$\begin{aligned}
\mathbb{E}_F[\delta(X) - \mu(F))^2] &= \mathbb{E}_F\left[\left(\left(\frac{\sqrt{n}}{1+\sqrt{n}}\right)(\bar{X} - \mu(F)) + \frac{1}{1+\sqrt{n}}\left(\frac{1}{2} - \mu(F)\right)\right)^2\right] \\
&= \left(\frac{1}{1+\sqrt{n}}\right)^2\left[n\text{Var}(\bar{X}) + \left(\frac{1}{2} - \mu(F)\right)^2\right] \\
&= \left(\frac{1}{1+\sqrt{n}}\right)^2\left[\mathbb{E}(X_1^2) - \mu(F)^2 + \frac{1}{4} - \mu(F) + \mu(F)^2\right] \\
&= \left(\frac{1}{1+\sqrt{n}}\right)^2\left[\mathbb{E}(X_1^2) + \frac{1}{4} - \mu(F)\right]
\end{aligned}$$

where the third step follows from the fact that $\text{Var}(X_1) = n\text{Var}(\bar{X}) = \mathbb{E}[X_1^2] - (\mathbb{E}[X_1])^2 = \mathbb{E}[X_1^2] - (\mu(F))^2$.

By assumption $X_1 \in [0,1]$, so $X_1^2 \leq X_1$ and we can bound the risk:

$$\begin{aligned}
\mathbb{E}_F[\delta(X) - \mu(F))^2] &\leq \left(\frac{1}{1+\sqrt{n}}\right)^2\left[\mathbb{E}(X_1) + \frac{1}{4} - \mu(F)\right] \\
&= \frac{1}{4(1+\sqrt{n})^2}.
\end{aligned}$$

So, $\delta(X)$ is minimax for the Binomial submodel, and its worst-case risk is the same for the full model and for the Binomial submodel. Therefore, applying the Lemma, we conclude that $\delta(X)$ is minimax. Thus, we have found a minimax estimator.

11.3 Admissibility of minimax estimators

Let us now turn to the question of admissibility of minimax estimators. We begin by noting that the question of admissibility is particularly important for minimax estimators. This is because, although we found dominating estimators even when we were working with unbiased estimators, the dominating estimators were biased, so we lost the property (unbiasedness) that we were interested in – however, if you find an estimator that dominates a minimax estimator, it will still be minimax!

Also, an aside: admissibility can give rise to minimaxity. If δ is admissible with constant risk, then δ is also minimax. This is not hard to show. (Let the constant risk of δ be r. Then, r is also the worst-case risk of δ, since the risk is constant. Now, if we assume δ is not minimax, there exists a different estimator, say δ', which is minimax. The *worst-case* risk of δ', say r', would thus be $< r$. But since this is the worst-case risk of δ', that would mean that the risk of δ' is lower than r throughout, and thus δ' dominates δ. However, we assumed that δ was admissible, so this is a contradiction. Thus, our assumption led to a contradiction, and therefore δ is minimax.)

Note that minimaxity does not guarantee admissibility; it only ensures the worst case risk is optimal. We need to check for admissibility. The following example illustrates several standard ways of doing so.

Example 3. Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\theta, \sigma^2)$ where σ^2 is known, and θ is the estimand. Then the minimax estimator is $\bar{X}$ under squared error loss, and we would like to determine whether $\bar{X}$ is admissible.

Instead of answering this directly, we answer a more general question: when is $a\bar{X} + b$, $a, b \in \mathbb{R}$, (basically, any affine function of $\bar{X}$) admissible?

Case 1: $0 < a < 1$. In this case $a\bar{X} + b$ is a convex combination of $\bar{X}$ and b. By results we saw in the previous lecture, it is a Bayes estimator with respect to some Gaussian prior on θ. Further, since we are using squared error loss, which is strictly convex, this Bayes estimator is unique. So, by Theorem 5.2.4 (which basically tells us that a unique Bayes estimator will always be admissible), $a\bar{X} + b$ is admissible.

Case 2: $a = 0$. In this case b is also a unique Bayes estimator with respect to a degenerate prior distribution with unit mass at $\theta = b$. 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:

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

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:

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

The first inequality follows because the second summand in the expression for the general risk is always nonnegative. $\bar{X}$ dominates $a\bar{X}+b$ when $a>1$, and so in this case $a\bar{X}+b$ is inadmissible.

Case 5: $a<0$.

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

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 inadmissible. Then, assuming w.l.o.g that $\sigma^2=1$, we have:

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

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 $\varepsilon>0$ and an interval (θ_0,θ_1) containing θ' so that:

$$R(\theta,\delta')<\frac{1}{n}-\varepsilon \quad \forall\theta\in(\theta_0,\theta_1). \tag{11.1}$$

Let r'_τ be the average risk of δ' with respect to the prior distribution $\mathcal{N}(0,\tau^2)$ 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 $\pi(\theta)=1\forall\ \theta$.) 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:

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

Applying (11.1), we find:

$$\frac{\frac{1}{n} - r'_\tau}{\frac{1}{n} - r_\tau} \geq \frac{\frac{1}{\sqrt{2\pi}\tau}\int_{\theta_0}^{\theta_1} \varepsilon e^{\frac{-\theta^2}{2\tau^2}} d\theta}{\frac{1}{n(1+n\tau^2)}}$$
$$= \frac{n(1+n\tau^2)}{\tau\sqrt{2\pi}}\varepsilon \int_{\theta_0}^{\theta_1} e^{\frac{-\theta^2}{2\tau^2}} d\theta$$

As $\tau \to \infty$, the first expression, $n(1+n\tau^2)\varepsilon/(\tau\sqrt{2\pi}) \to \infty$ 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

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

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 our assumption that there was a dominating estimator was false, and in this case, $a\bar{X} + b = \bar{X}$ is admissible.

11.4 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 $X_1, X_2, \ldots, X_p$ be independent with $X_i \sim \mathcal{N}(\theta_i, \sigma^2)$ for $1 \leq i \leq p$. For the sake of simplicity, say $\sigma^2 = 1$. Now our goal is to estimate $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$ under the loss function:

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

A natural estimator for θ is $X = (X_1, X_2, \ldots, X_p)$. 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):

$$\delta(X) = (\delta_1(X), \delta_2(X), \ldots, \delta_p(X)) \text{ where}^1$$

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

[1] Here $||\cdot||_2$ is the 2-norm so $||X||_2^2 = \sum_{j=1}^{p} X_j^2$)

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. An example of the James-Stein estimator being used to estimate batting averages is available at `http://www-stat.stanford.edu/~ckirby/brad/other/Article1977.pdf`. It turns out that the James-Stein estimator is not itself admissible because it is dominated by the **positive part James-Stein estimator** (TPE Theorem 5.5.4):

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

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

11.4.1 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:

1. Introduce a prior family indexed by a hyperparameter (this is the Bayesian aspect).
2. Estimate the hyperparameter from the data (this is the empirical aspect).

So applying this procedure to the problem at hand:

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

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

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

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

(Exercise: Verify this.) 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

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

(Exercise: Verify this. Hint: $\frac{1}{A+1}||X||_2^2$ follows a χ_n^2 distribution). So

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

must be UMVU for $1 - \frac{1}{A+1}$.

If we plug this estimator into our Bayes estimator we obtain the J-S estimator:

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

11.4.2 James-Stein domination

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

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

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 toward 0.

The uniform superiority of the J-S estimator to X can be formalised (see Keener 11.2).

Theorem 1 (Theorem 5.5.1 TPE)**.** The James-Stein estimator δ has uniformly smaller risk than X if $p \geq 3$.

The proof, given on p. 355 of TPE, compares the risk of the J-S estimator directly to that of X.