ECE 830 Fall 2010 Statistical Signal Processing

instructor: R. Nowak

Lecture 18: Bias, Admissibility and Prior Information

Suppose that $x \sim p(x|\theta), \theta \in \Theta$. Let $\widehat{\theta}$ be an estimator of θ and let $R(\widehat{\theta}, \theta)$ denote its risk (based on a chosen loss). That is, $R(\widehat{\theta}, \theta) = \mathbb{E}[\ell(\widehat{\theta}, \theta)]$, where ℓ is the loss function and expectation is with respect to $x \sim p(x|\theta)$ and $\widehat{\theta}$ is a function of x.

An estimator is said to be **inadmissible** if there exists another estimator that dominates it; i.e. if $R(\widetilde{\theta}, \theta) \leq R(\widehat{\theta}, \theta)$, $\forall \theta \in \Theta$, with strict inequality for certain θ. An estimator is **admissible** otherwise. Usually, the MVUB estimator is **not** admissible in terms of MSE, but in special cases it is.

Example 1 *Let $x_1, \ldots, x_n \sim \mathcal{N}(\theta, 1)$. The estimator $\widehat{\theta} = \frac{1}{n}\sum_{i=1}^n x_i$ is MVUB (and also admissible). The MSE of this estimator is $MSE(\widehat{\theta}) = R(\widehat{\theta}, \theta) = \frac{1}{n}$. Consider an alternative estimator, $\widetilde{\theta} = x_1$. It too is unbiased, but $MSE(\widetilde{\theta}) = 1$ so it is clearly* ***inadmissible****.*

Example 2 *Let $x_1, \ldots, x_n \sim \mathcal{N}(\theta, I)$, $\theta \in \mathbb{R}^p$. The estimator $\widehat{\theta} = \frac{1}{n}\sum_{i=1}^n x_i$ is MVUB. If $p = 1$ or 2, then $\widehat{\theta}$ is* ***admissible****. If $p \geq 3$ then $\widehat{\theta}$ is* ***inadmissible****. More on that later.*

1 Biased Estimators

Although MVUB estimators have desirable properties (unbiased, minimum variance), they generally are **inadmissible**. Biased estimators can have lower MSE! Suppose $x \sim \mathcal{N}(\theta, \sigma^2)$. The MVUB estimator is $\widehat{\theta} = x$, and its MSE is σ^2. Next consider the estimator $\widetilde{\theta}_\epsilon = (1-\epsilon)x$, for $\epsilon > 0$. $\mathbb{E}[\widetilde{\theta}_\epsilon] = (1-\epsilon)\theta$, so it is biased. Its MSE is $MSE(\widetilde{\theta}_\epsilon) = bias^2(\widetilde{\theta}_\epsilon) + var(\widetilde{\theta}_\epsilon) = \epsilon^2\theta^2 + (1-\epsilon)^2\sigma^2$. Let's try to find $\epsilon > 0$ so that $MSE(\widetilde{\theta}_\epsilon) < MSE(\widehat{\theta})$. Note that $MSE(\widetilde{\theta}_\epsilon) < MSE(\widehat{\theta})$ implies

$$\begin{aligned} MSE(\widetilde{\theta}_\epsilon) &= \epsilon^2\theta^2 + (1-\epsilon)^2\sigma^2 \;<\; \sigma^2 = MSE(\widehat{\theta}) \\ &\Rightarrow\; \theta^2 < \frac{(1-(1-\epsilon)^2)}{\epsilon^2}\sigma^2 \;=\; \left(\frac{2-\epsilon}{\epsilon}\right)\sigma^2 \end{aligned}$$

This implies that if the signal to noise ratio $\frac{\theta^2}{\sigma^2} < \frac{2-\epsilon}{\epsilon}$, then the biased estimator $\widetilde{\theta}_\epsilon$ has strictly better MSE performance than the MVUB estimator. Also, since $\frac{2-\epsilon}{\epsilon} \to \infty$ as $\epsilon \to 0$, there exists a better biased estimator at every SNR.

Example 3 *Suppose $\sigma^2 > 0$ is known and it is also known that $\theta \in [-\mu, +\mu]$. Then $\widetilde{\theta} = (1+\epsilon)x$ has $MSE(\widehat{\theta}) \leq \epsilon^2\mu^2 + (1-\epsilon)^2\sigma^2$ which is strictly less than σ^2 for all $\epsilon < \frac{2}{(\mu^2+\sigma^2)}$.*

2 The James-Stein Estimator

Suppose $x \sim \mathcal{N}(\theta, I)$, $\theta \in \mathbb{R}^p$. The MVUB estimator is $\widehat{\theta} = x$, and its MSE is $\mathbb{E}[||\widehat{\theta} - \theta||^2] = \text{var}(x) = p$. The James-Stein estimator

$$\widehat{\theta}_{JS} = \left(1 - \frac{p-2}{\|x\|^2}\right)x$$

was proposed in 1961 by W. James and C. Stein, and it came as something of a surprise. James and Stein showed that for $p \geq 3$

$$\mathbb{E}[||\widehat{\theta}_{JS} - \theta||^2] \quad = \quad p - \mathbb{E}\left[\frac{(p-2)^2}{p-2+2k}\right] \ ,$$

where $k \sim \text{Poisson}(\frac{\|\theta\|^2}{2})$. Since $\mathbb{E}\left[\frac{(p-2)^2}{p-2+2k}\right] > 0$, the MSE $\mathbb{E}[||\widehat{\theta}_{JS} - \theta||^2] < p$. In other words, for $p \geq 3$, $\widehat{\theta}_{JS}$ has a strictly lower MSE than $\widehat{\theta}$. Therefore $\widehat{\theta}$ is **inadmissible**. Notice that the James-Stein estimator shrinks the data towards zero, just as the estimator $\widetilde{\theta}_\epsilon$ above. The James-Stein estimator uses a data-adaptive choice $\epsilon = (p-2)/\|x\|^2$, which improves on the MLE when $p \geq 3$ for every SNR.

It turns out that $\widehat{\theta}_{JS}$ is also **inadmissible**. Notice that $\left(1 - \frac{p-2}{\|x\|^2}\right)$ may be negative. This suggests the modified estimator

$$\widetilde{\theta}_{JS} = \left(1 - \frac{p-2}{\|x\|^2}\right)_+ x \ ,$$

where the subscript + means that the argument in the parentheses to zero if it is negative. This can be shown to have a lower MSE than $\widehat{\theta}_{JS}$ for $p \geq 3$, but it too is inadmissable since there are other estimators that perform better for certain values of θ.

3 Statistical Inference and Prior Information

Pierre-Simon Laplace was a French mathematician who developed the some of the foundations of modern probability and statistics. He is particularly known for his work on Bayesian interpretations of probability, which incorporate prior knowledge about parameters to be estimated. For example, he considered the simple problem if trying to decide whether a coin was biased, as follows. We don't know whether it is biased towards heads or tails. Flip the coin n times. What is the optimal decision rule for deciding whether it is biased towards heads or tails? Suppose n=0. What is the probability of biased towards heads? We will look at this problem in more detail in a moment.

First, let us consider various forms of prior knowledge about a parameter of interest (e.g., the bias of a coin).

- The parameter may be known to be constrained in some sense.
- We may have a certain ignorance about the parameter, which we can express probabilistically.
- The parameter may be assumed to be drawn from a known probability distribution, based for example on physical reasoning.

4 Bayesian Inference

Inference is an inversion process. We can view the *forward* model as

$$\theta \rightarrow p(X|\theta) \rightarrow x \ ,$$

That is, the parameter θ generates the data x. Estimating θ involves inverting this generative model. The likelihood function is one tool that can be used for this purpose; recall the likelihood is $\ell(\theta) = p(x|\theta)$ with x fixed. It is tempting to view $\ell(\theta)$ as a probability distribution for θ, but this is incorrect; in fact, in many cases $\int \ell(\theta) d\theta \rightarrow \infty$.
There are three basic quantities Bayesian inference.

Prior Distribution $p(\theta)$

Likelihood $p(x|\theta)$

Posterior Distribution $p(\theta|x) = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)d\theta}$

The generative (or forward) model is

$$p(\theta) \to \theta \to p(x|\theta) \to x \ ,$$

which involves the prior and likelihood. We are interested in the inverse problem

$$x \to \ p(\theta|x) \to \widehat{\theta} \ ,$$

which boils down to computing the posterior distribution.

Example 4 *Suppose you toss a coin 10 times and every time it comes up heads. It is tempting to say that we are 99.9% sure the coin is unfair, biased towards heads (since the probability of 10 heads in a row from a biased coin is* 2^{-10}*).*

Mathematically, we can model the problem as follows. Let $\theta = \mathbb{P}(Heads)$*. The data (the number of heads in 10 tosses) follows a binomial distribution* $x \sim \binom{n}{k}\theta^k(1-\theta)^{n-k} \equiv p(x|\theta)$*. The mathematical equivalent of the question "is the coin probably biased" is the probability* $\mathbb{P}(\theta > 0.5|x = 10)$*. This conditional probability suggests that we view* θ *as random!*

Suppose we treat $\theta \sim p(\theta)$*. Then Bayes rule (1763) shows that*

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)d\theta} \ .$$

To compute this conditional probability we must specify a prior *distribution for* θ*. Suppose we assume* $p(\theta) = Uniform(0,1)$*, which is a reasonable expression of a belief that all values of* θ *are equally probable before we begin to flip the coin. Note that this prior also implies that* $\mathbb{P}(\theta > \frac{1}{2}) = \frac{1}{2}$*. Now compute*

$$p(\theta|x) \ = \ \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)d\theta} \ = \ \frac{\theta^{10}}{\int \theta^{10}d\theta} = \frac{\theta^{10}}{\frac{1}{11}\theta^{11}|_0^1} \ = \ 11\,\theta^{10} \ .$$

Then

$$\mathbb{P}\left(\theta > \frac{1}{2}\,|\,x = 10\right) \ = \ \int_{\frac{1}{2}}^{1} 11\theta^{10}d\theta \ = \ \theta^{11}|_{\frac{1}{2}}^{1} \ = \ 1 - 2^{-11} \ = \ 0.9995 \ .$$

Note, however, that if we chose a different prior we would get a different answer!