

Econ 722 – Advanced Econometrics IV, Part II

Francis J. DiTraglia

University of Pennsylvania

Lecture #0 – Decision Theory

Statistical Decision Theory

The James-Stein Estimator

Decision Theoretic Preliminaries

Parameter $\theta \in \Theta$

Unknown state of nature, from parameter space Θ

Observed Data

Observe X with distribution F_θ from a sample space \mathcal{X}

Estimator $\hat{\theta}$

An estimator (aka a decision rule) is a function from \mathcal{X} to Θ

Loss Function $L(\theta, \hat{\theta})$

A function from $\Theta \times \Theta$ to \mathbb{R} that gives the cost we incur if we report $\hat{\theta}$ when the true state of nature is θ .

Examples of Loss Functions

$$L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$$

squared error loss

$$L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$$

absolute error loss

$$L(\theta, \hat{\theta}) = 0 \text{ if } \theta = \hat{\theta}, 1 \text{ otherwise}$$

zero-one loss

$$L(\theta, \hat{\theta}) = \int \log \left[\frac{f(x|\theta)}{f(x|\hat{\theta})} \right] f(x|\theta) dx$$

Kullback–Leibler loss

(Frequentist) Risk of an Estimator $\hat{\theta}$

$$R(\theta, \hat{\theta}) = \mathbb{E}_{\theta} [L(\theta, \hat{\theta})] = \int L(\theta, \hat{\theta}(x)) dF_{\theta}(x)$$

The frequentist decision theorist seeks to evaluate, for each θ , how much he would “expect” to lose if he used $\hat{\theta}(X)$ repeatedly with varying X in the problem.

(Berger, 1985)

Example: Squared Error Loss

$$R(\theta, \hat{\theta}) = \mathbb{E}_{\theta} [(\theta - \hat{\theta})^2] = \text{MSE} = \text{Var}(\hat{\theta}) + \text{Bias}_{\theta}^2(\hat{\theta})$$

Bayes Risk and Maximum Risk

Comparing Risk

$R(\theta, \hat{\theta})$ is a *function* of θ rather than a single number. We want an estimator with low risk, but how can we compare?

Maximum Risk

$$\bar{R}(\hat{\theta}) = \sup_{\theta \in \Theta} R(\theta, \hat{\theta})$$

Bayes Risk

$$r(\pi, \hat{\theta}) = \mathbb{E}_{\pi} \left[R(\theta, \hat{\theta}) \right], \text{ where } \pi \text{ is a prior for } \theta$$

Bayes and Minimax Rules

Minimize the Maximum or Bayes risk over all estimators $\tilde{\theta}$

Minimax Rule/Estimator

$\hat{\theta}$ is **minimax** if

$$\sup_{\theta \in \Theta} R(\theta, \hat{\theta}) = \inf_{\tilde{\theta}} \sup_{\theta \in \Theta} R(\theta, \tilde{\theta})$$

Bayes Rule/Estimator

$\hat{\theta}$ is a **Bayes rule** with respect to prior π if

$$r(\pi, \hat{\theta}) = \inf_{\tilde{\theta}} r(\pi, \tilde{\theta})$$

Recall: Bayes' Theorem and Marginal Likelihood

Let π be a prior for θ . By Bayes' theorem, the **posterior** $\pi(\theta|\mathbf{x})$ is

$$\pi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{m(\mathbf{x})}$$

where the **marginal likelihood** $m(\mathbf{x})$ is given by

$$m(\mathbf{x}) = \int f(\mathbf{x}|\theta)\pi(\theta) d\theta$$

Posterior Expected Loss

Posterior Expected Loss

$$\rho(\pi(\theta|\mathbf{x}), \hat{\theta}) = \int L(\theta, \hat{\theta}) \pi(\theta|\mathbf{x}) d\theta$$

Bayesian Decision Theory

Choose an estimator that minimizes posterior expected loss.

Easier Calculation

Since $m(\mathbf{x})$ does not depend on θ , to minimize $\rho(\pi(\theta|\mathbf{x}), \hat{\theta})$ it suffices to minimize $\int L(\theta, \hat{\theta}) f(\mathbf{x}|\theta) \pi(\theta) d\theta$.

Question

Is there a relationship between Bayes risk, $r(\pi, \hat{\theta}) \equiv \mathbb{E}_{\pi}[R(\theta, \hat{\theta})]$, and posterior expected loss?

Bayes Risk vs. Posterior Expected Loss

Theorem

$$r(\pi, \hat{\theta}) = \int \rho(\pi(\theta|\mathbf{x}), \hat{\theta}(\mathbf{x})) m(\mathbf{x}) d\mathbf{x}$$

Proof

$$\begin{aligned} r(\pi, \hat{\theta}) &= \int R(\theta, \hat{\theta}) \pi(\theta) d\theta = \int \left[\int L(\theta, \hat{\theta}(\mathbf{x})) f(\mathbf{x}|\theta) d\mathbf{x} \right] \pi(\theta) d\theta \\ &= \int \int L(\theta, \hat{\theta}(\mathbf{x})) [f(\mathbf{x}|\theta) \pi(\theta)] d\mathbf{x} d\theta \\ &= \int \int L(\theta, \hat{\theta}(\mathbf{x})) [\pi(\theta|\mathbf{x}) m(\mathbf{x})] d\mathbf{x} d\theta \\ &= \int \left[\int L(\theta, \hat{\theta}(\mathbf{x})) \pi(\theta|\mathbf{x}) d\theta \right] m(\mathbf{x}) d\mathbf{x} \\ &= \int \rho(\pi(\theta|\mathbf{x}), \hat{\theta}(\mathbf{x})) m(\mathbf{x}) d\mathbf{x} \end{aligned}$$

Finding a Bayes Estimator

Hard Problem

Find the **function** $\hat{\theta}(\mathbf{x})$ that minimizes $r(\pi, \hat{\theta})$.

Easy Problem

Find the **number** $\hat{\theta}$ that minimizes $\rho(\pi(\theta|\mathbf{x}), \hat{\theta})$

Punchline

Since $r(\pi, \hat{\theta}) = \int \rho(\pi(\theta|\mathbf{x}), \hat{\theta}(\mathbf{x})) m(\mathbf{x}) d\mathbf{x}$, to minimize $r(\pi, \hat{\theta})$ we can set $\hat{\theta}(\mathbf{x})$ to be the value $\hat{\theta}$ that minimizes $\rho(\pi(\theta|\mathbf{x}), \hat{\theta})$.

Bayes Estimators for Common Loss Functions

Zero-one Loss

For zero-one loss, the Bayes estimator is the posterior mode.

Absolute Error Loss: $L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$

For absolute error loss, the Bayes estimator is the posterior median.

Squared Error Loss: $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$

For squared error loss, the Bayes estimator is the posterior mean.

Derivation of Bayes Estimator for Squared Error Loss

By definition,

$$\hat{\theta} \equiv \arg \min_{a \in \Theta} \int (\theta - a)^2 \pi(\theta | \mathbf{x}) d\theta$$

Differentiating with respect to a , we have

$$\begin{aligned} 2 \int (\theta - a) \pi(\theta | \mathbf{x}) d\theta &= 0 \\ \int \theta \pi(\theta | \mathbf{x}) d\theta &= a \end{aligned}$$

Example: Bayes Estimator for a Normal Mean

Suppose $X \sim N(\mu, 1)$ and π is a $N(a, b^2)$ prior. Then,

$$\begin{aligned}\pi(\mu|x) &\propto f(x|\mu) \times \pi(\mu) \\ &\propto \exp \left\{ -\frac{1}{2} \left[(x - \mu)^2 + \frac{1}{b^2} (\mu - a)^2 \right] \right\} \\ &\propto \exp \left\{ -\frac{1}{2} \left[\left(1 + \frac{1}{b^2} \right) \mu^2 - 2 \left(x + \frac{a}{b^2} \right) \mu \right] \right\} \\ &\propto \exp \left\{ -\frac{1}{2} \left(\frac{b^2 + 1}{b^2} \right) \left[\mu - \left(\frac{b^2 x + a}{b^2 + 1} \right) \right]^2 \right\}\end{aligned}$$

So $\pi(\mu|x)$ is $N(m, \omega^2)$ with $\omega^2 = \frac{b^2}{1+b^2}$ and $m = \omega^2 x + (1 - \omega^2)a$.

Hence the Bayes estimator for μ under squared error loss is

$$\hat{\theta}(X) = \frac{b^2 X + a}{1 + b^2}$$

Minimax Analysis

Wasserman (2004)

The advantage of using maximum risk, despite its problems, is that it does not require one to choose a prior.

Berger (1986)

Perhaps the greatest use of the minimax principle is in situations for which no prior information is available . . . but two notes of caution should be sounded. First, the minimax principle can lead to bad decision rules. . . Second, the minimax approach can be devilishly hard to implement.

Methods for Finding a Minimax Estimator

1. Direct Calculation
2. Guess a “Least Favorable” Prior
3. Search for an “Equalizer Rule”

Method 1 rarely applicable so focus on 2 and 3...

Minimax Estimators and Least Favorable Priors

Theorem

Let $\hat{\theta}$ be a Bayes rule with respect to π and suppose that for all $\theta \in \Theta$ we have $R(\theta, \hat{\theta}) \leq r(\pi, \hat{\theta})$. Then $\hat{\theta}$ is a minimax estimator, and π is called a **least favorable prior**.

Proof

Suppose that $\hat{\theta}$ is not minimax. Then there exists another estimator $\tilde{\theta}$ with $\sup_{\theta \in \Theta} R(\theta, \tilde{\theta}) < \sup_{\theta \in \Theta} R(\theta, \hat{\theta})$. But since

$$r(\pi, \tilde{\theta}) \equiv \mathbb{E}_{\pi} [R(\theta, \tilde{\theta})] \leq \mathbb{E}_{\pi} \left[\sup_{\theta \in \Theta} R(\theta, \tilde{\theta}) \right] = \sup_{\theta \in \Theta} R(\theta, \tilde{\theta})$$

but this implies that $\hat{\theta}$ is *not* Bayes with respect to π since

$$r(\pi, \tilde{\theta}) \leq \sup_{\theta \in \Theta} R(\theta, \tilde{\theta}) < \sup_{\theta \in \Theta} R(\theta, \hat{\theta}) \leq r(\pi, \hat{\theta})$$

Try to use the preceding result to show that for the problem $X \sim N(\theta, 1)$ with squared error loss and $\Theta = [-m, m]$ for $0 < m < 1$ then the unique minimax estimator is $m \tanh(mX)$ and the least favorable prior puts $1/2$ probability on m and the rest on $-m$.

Equalizer Rules

Definition

An estimator $\hat{\theta}$ is called an **equalizer rule** if its risk function is constant: $R(\theta, \hat{\theta}) = C$ for some C .

Theorem

If $\hat{\theta}$ is an equalizer rule and is Bayes with respect to π , then $\hat{\theta}$ is minimax and π is least favorable.

Proof

$$r(\pi, \hat{\theta}) = \int R(\theta, \hat{\theta})\pi(\theta) d\theta = \int C\pi(\theta) d\theta = C$$

Hence, $R(\theta, \hat{\theta}) \leq r(\pi, \hat{\theta})$ for all θ so we can apply the preceding theorem. **Maybe take this out and assign it as homework?**

Add an example. Possibly talk about generalized Bayes and sequences of priors.

Some benefits and drawbacks of max versus Bayes risk. Example where max risk is crazy. Motivate regret and minimax regret.

Dominance and Admissibility

Dominance

$\hat{\theta}$ **dominates** $\tilde{\theta}$ with respect to R if $R(\theta, \hat{\theta}) \leq R(\theta, \tilde{\theta})$ for all $\theta \in \Theta$ and the inequality is strict for at least one value of θ .

Admissibility

$\hat{\theta}$ is **admissible** if no other estimator dominates it.

Inadmissibility

$\hat{\theta}$ is **inadmissible** if there is an estimator that dominates it.

Strong Inadmissibility

We say that $\hat{\theta}$ is **strongly inadmissible** if there exists an estimator $\tilde{\theta}$ and an $\varepsilon > 0$ such that $R(\theta, \tilde{\theta}) < R(\theta, \hat{\theta}) - \varepsilon$ for all θ .

Example of Admissible Estimator

Example

If $X \sim N(\theta, 1)$ then $\hat{\theta}(X) = 3$ is admissible under squared error loss. **FILL IN DETAILS**

Bayes Rules are Admissible

Theorem A-1

Suppose that Θ is a discrete set and π gives strictly positive probability to each element of Θ . Then, if $\hat{\theta}$ is a Bayes rule with respect to π , it is admissible.

Theorem A-2

If a Bayes rule is unique, it is admissible.

Theorem A-3

Suppose that $R(\theta, \hat{\theta})$ is continuous in θ for all $\hat{\theta}$ and that π gives strictly positive probability to any open subset of Θ . Then if $\hat{\theta}$ is a Bayes rule with respect to π , it is admissible.

Add proof of one of the preceding theorems and assign the others for homework.

Admissibility and Minimavity

Theorem

If $\hat{\theta}$ has constant risk and is admissible, then it is minimax.

Theorem

If $\hat{\theta}$ is minimax, then it is **not** strongly inadmissible.

Possibly assign these for homework. I have a proof for the first, and the second seems straightforward.

Recall: Gauss-Markov Theorem

Linear Regression Model

$$\mathbf{y} = X\beta + \epsilon, \quad \mathbb{E}[\epsilon|X] = \mathbf{0}$$

Best Linear Unbiased Estimator

- ▶ $\text{Var}(\epsilon|X) = \sigma^2 I \Rightarrow$ then OLS has lowest variance among linear, unbiased estimators of β .
- ▶ $\text{Var}(\epsilon|X) \neq \sigma^2 I \Rightarrow$ then GLS gives a lower variance estimator.

What if we consider biased estimators and squared error loss?

A Very Simple Example: $X \sim N(\theta, I)$

Goal

Estimate the p -vector θ using X with $L(\theta, \hat{\theta}) = \|\hat{\theta} - \theta\|^2$.

Maximum Likelihood Estimator $\hat{\theta}$

MLE = sample mean, but only one observation: $\hat{\theta} = X$.

Risk of $\hat{\theta}$

$$(\hat{\theta} - \theta)' (\hat{\theta} - \theta) = (X - \theta)' (X - \theta) = \sum_{i=1}^p (X_i - \theta_i)^2 \sim \chi_p^2$$

Since $\mathbb{E}[\chi_p^2] = p$, we have $R(\theta, \hat{\theta}) = p$.

A Very Simple Example: $X \sim N(\theta, I)$

James-Stein Estimator

$$\hat{\theta}^{JS} = \hat{\theta} \left(1 - \frac{p-2}{\hat{\theta}'\hat{\theta}} \right) = X - \frac{(p-2)X}{X'X}$$

- ▶ Shrinks components of sample mean vector towards zero
- ▶ More elements in $\theta \Rightarrow$ more shrinkage
- ▶ MLE close to zero ($\hat{\theta}'\hat{\theta}$ small) gives more shrinkage

MSE of James-Stein Estimator

$$\begin{aligned}R(\theta, \hat{\theta}^{JS}) &= \mathbb{E} \left[(\hat{\theta}^{JS} - \theta)' (\hat{\theta}^{JS} - \theta) \right] \\&= \mathbb{E} \left[\left\{ (X - \theta) - \frac{(p-2)X}{X'X} \right\}' \left\{ (X - \theta) - \frac{(p-2)X}{X'X} \right\} \right] \\&= \mathbb{E} [(X - \theta)' (X - \theta)] - 2(p-2) \mathbb{E} \left[\frac{X'(X - \theta)}{X'X} \right] \\&\quad + (p-2)^2 \mathbb{E} \left[\frac{1}{X'X} \right] \\&= p - 2(p-2) \mathbb{E} \left[\frac{X'(X - \theta)}{X'X} \right] + (p-2)^2 \mathbb{E} \left[\frac{1}{X'X} \right]\end{aligned}$$

Using fact that $R(\theta, \hat{\theta}) = p$

Simplifying the Second Term

Writing Numerator as a Sum

$$\mathbb{E} \left[\frac{X'(X - \theta)}{X'X} \right] = \mathbb{E} \left[\frac{\sum_{i=1}^p X_i (X_i - \theta_i)}{X'X} \right] = \sum_{i=1}^p \mathbb{E} \left[\frac{X_i (X_i - \theta_i)}{X'X} \right]$$

For $i = 1, \dots, p$

$$\mathbb{E} \left[\frac{X_i (X_i - \theta_i)}{X'X} \right] = \mathbb{E} \left[\frac{X'X - 2X_i^2}{(X'X)^2} \right]$$

Not obvious: integration by parts, expectation as a p -fold integral, $X \sim N(\theta, I)$

Combining

$$\begin{aligned} \mathbb{E} \left[\frac{X'(X - \theta)}{X'X} \right] &= \sum_{i=1}^p \mathbb{E} \left[\frac{X'X - 2X_i^2}{(X'X)^2} \right] = p \mathbb{E} \left[\frac{1}{X'X} \right] - 2 \mathbb{E} \left[\frac{\sum_{i=1}^p X_i^2}{(X'X)^2} \right] \\ &= p \mathbb{E} \left[\frac{1}{X'X} \right] - 2 \mathbb{E} \left[\frac{X'X}{(X'X)^2} \right] = (p - 2) \mathbb{E} \left[\frac{1}{X'X} \right] \end{aligned}$$

The MLE is Inadmissible when $p \geq 3$

$$\begin{aligned} R\left(\theta, \hat{\theta}^{JS}\right) &= p - 2(p-2) \left\{ (p-2) \mathbb{E} \left[\frac{1}{X'X} \right] \right\} + (p-2)^2 \mathbb{E} \left[\frac{1}{X'X} \right] \\ &= p - (p-2)^2 \mathbb{E} \left[\frac{1}{X'X} \right] \end{aligned}$$

- ▶ $\mathbb{E}[1/(X'X)]$ exists and is positive whenever $p \geq 3$
- ▶ $(p-2)^2$ is always positive
- ▶ Hence, second term in the MSE expression is *negative*
- ▶ First term is MSE of the MLE

Therefore James-Stein strictly dominates MLE whenever $p \geq 3$!

James-Stein More Generally

- ▶ Our example was specific, but the result is general:
 - ▶ MLE is inadmissible under quadratic loss in regression model with at least three regressors.
 - ▶ Note, however, that this is MSE for the *full parameter vector*
- ▶ James-Stein estimator is also inadmissible!
 - ▶ Dominated by “positive-part” James-Stein estimator:

$$\hat{\beta}^{JS} = \hat{\beta} \left[1 - \frac{(p-2)\hat{\sigma}^2}{\hat{\beta}'X'X\hat{\beta}} \right]_+$$

- ▶ $\hat{\beta} = \text{OLS}$, $(x)_+ = \max(x, 0)$, $\hat{\sigma}^2 = \text{usual OLS-based estimator}$
- ▶ Stops us from shrinking *past* zero to get a negative estimate for an element of β with a small OLS estimate.
- ▶ Positive-part James-Stein isn't admissible either!