Econ 2140, spring 2018, Part IIb
Statistical Decision Theory

Maximilian Kasy

Department of Economics, Harvard University

Roadmap

- IIa
  - Basic definitions
  - Optimality criteria
- IIb
  - Relationships between optimality criteria
  - Analogies to microeconomics
  - Two justifications of the Bayesian approach
  - Testing and the Neyman Pearson lemma
- IIc
  - Value added estimation
  - Ridge regression and Lasso
  - Experimental design

Part IIIb

Some relationships between these optimality criteria

Analogies to microeconomics

Two justifications of the Bayesian approach

Testing and the Neyman Pearson lemma

Some relationships between these optimality criteria

Proposition (Minimax decision functions)

If δ^* is admissible with constant risk,
then it is a minimax decision function.

Proof:

- picture!
- Suppose that δ' had smaller worst-case risk than δ^*
- Then

$$R(\delta', \theta') \leq \sup_{\theta} R(\delta', \theta) < \sup_{\theta} R(\delta^*, \theta) = R(\delta^*, \theta'),$$

- used constant risk in the last equality
- This contradicts admissibility.

- despite this result,
  minimax decision functions are very hard to find
- Example:
  - if $X \sim N(\mu, I)$, $\dim(X) \geq 3$, then
  - X has constant risk (mean squared error) as estimator for μ
  - but: X is not an admissible estimator for μ
    therefore not minimax

Proposition (Bayes decisions are admissible)

Suppose:

- δ^* is the Bayes decision function
- $\pi(\theta) > 0$ for all θ, $R(\delta^*, \pi) < \infty$
- $R(\delta^*, \theta)$ is continuous in θ

Then δ^* is admissible.

(We will prove the reverse of this statement in the next section.)

Sketch of proof:

- picture!
- Suppose δ^* is not admissible
- $\Rightarrow$ dominated by some δ'
  i.e. $R(\delta', \theta) \leq R(\delta^*, \theta)$ for all θ with strict inequality for some θ
- Therefore

$$R(\delta', \pi) = \int R(\delta', \theta)\pi(\theta)d\theta < \int R(\delta^*, \theta)\pi(\theta)d\theta = R(\delta^*, \pi)$$

- This contradicts δ^* being a Bayes decision function.

Proposition (Bayes risk and minimax risk)

The Bayes risk
$R(\pi) := \inf_\delta R(\delta, \pi)$
is never larger than the minimax risk
$\overline{R} := \inf_\delta \sup_\theta R(\delta, \theta)$.

Proof:

$$\begin{aligned} R(\pi) &= \inf_\delta R(\delta, \pi) \\ &\leq \sup_\pi \inf_\delta R(\delta, \pi) \\ &\leq \inf_\delta \sup_\pi R(\delta, \pi) \\ &= \inf_\delta \sup_\theta R(\delta, \theta) = \overline{R}. \end{aligned}$$

If there exists a prior π^* such that $R(\pi) = \overline{R}$, it is called the least favorable distribution.

Analogies to microeconomics

1) Welfare economics

statistical decision theory	**social welfare analysis**
different parameter values θ	different people i
risk $R(.,\theta)$	individuals' utility $u_i(.)$
dominance	Pareto dominance
admissibility	Pareto efficiency
Bayes risk	social welfare function
prior	welfare weights (distributional preferences)
minimaxity	Rawlsian inequality aversion

2) choice under uncertainty / choice in strategic interactions

statistical decision theory	**strategic interactions**
dominance of decision functions	dominance of strategies
Bayes risk	expected utility
Bayes optimality	expected utility maximization
minimaxity	(extreme) ambiguity aversion

Two justifications of the Bayesian approach
justification 1 – the complete class theorem

- last section: every Bayes decision function is admissible (under some conditions)
- the reverse also holds true (under some conditions): every admissible decision function is Bayes, or the limit of Bayes decision functions
- can interpret this as: all reasonable estimators are Bayes estimators
- will state a simple version of this result

Preliminaries

- set of risk functions that correspond to some δ is the **risk set**,

$$\mathscr{R} := \{r(.) = R(., \delta) \text{ for some } \delta\}$$

- will assume **convexity** of $\mathscr{R}$
  – no big restriction, since we can always randomly "mix" decision functions
- a class of decision functions δ is a **complete class** if it contains every admissible decision function δ^*

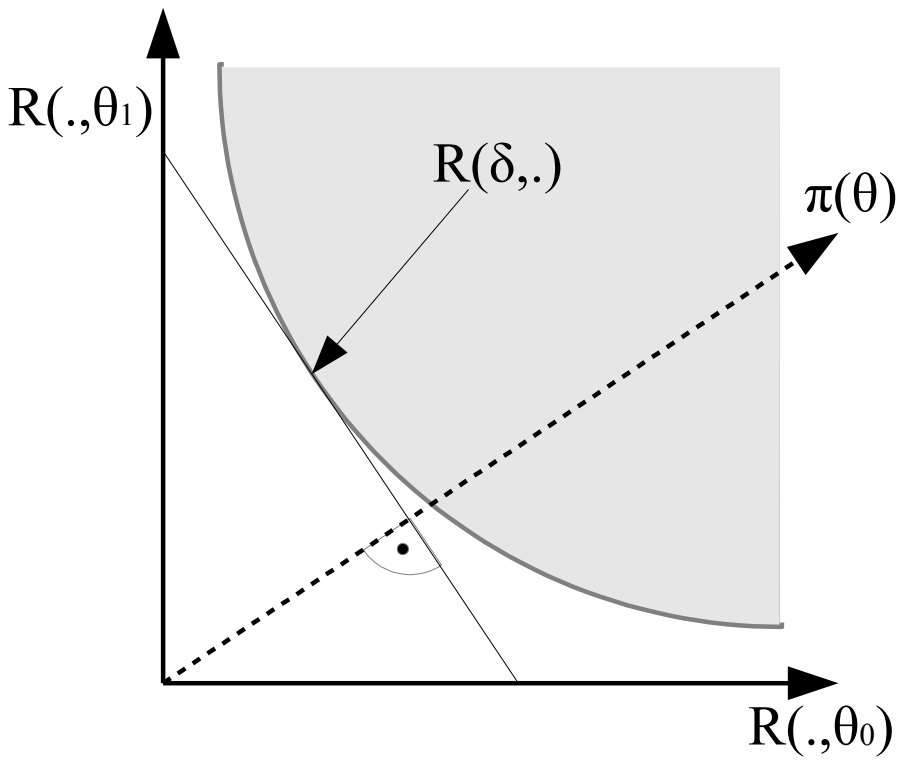
Theorem (Complete class theorem)

Suppose

- the set Θ of possible values for θ is compact
- the risk set $\mathscr{R}$ is convex
- all decision functions have continuous risk

Then the Bayes decision functions constitute a complete class: For every admissible decision function δ^*, there exists a prior distribution π such that δ^* is a Bayes decision function for π.

Figure: Complete class theorem



Intuition for the complete class theorem

- any choice of decision procedure has to trade off risk across θ
- slope of feasible risk set
  = relative "marginal cost" of decreasing risk at different θ
- pick a risk function on the admissible frontier
- can rationalize it with a prior
  = "marginal benefit" of decreasing risk at different θ
- for example, minimax decision rule:
  rationalizable by least favorable prior
  slope of feasible set at constant risk admissible point
- analogy to social welfare: any policy choice or allocation
  corresponds to distributional preferences / welfare weights

Proof of complete class theorem:

- application of the separating hyperplane theorem, to the space of functions of θ, with the inner product

$$\langle f, g \rangle = \int f(\theta) g(\theta) d\theta.$$

- for intuition: focus on binary θ, $\theta \in \{0, 1\}$, and $\langle f, g \rangle = \sum_\theta f(\theta) g(\theta)$
- Let δ^* be admissible. Then $R(., \delta^*)$ belongs to the lower boundary of $\mathscr{R}$.
- convexity of $\mathscr{R}$, separating hyperplane theorem separating $\mathscr{R}$ from risk functions dominating δ^*

$\Rightarrow$

- $\Rightarrow$ there exists a function $\tilde{\pi}$ (with finite integral) such that for all δ

$$\langle R(.,\delta^*),\tilde{\pi}\rangle \leq \langle R(.,\delta),\tilde{\pi}\rangle.$$

- by construction $\tilde{\pi} \geq 0$
- thus $\pi := \tilde{\pi}/\int \tilde{\pi}$ defines a prior distribution.
- δ^* minimizes

$$\langle R(.,\delta^*),\pi\rangle = R(\delta^*,\pi)$$

among the set of feasible decision functions

- and is therefore the optimal Bayesian decision function for the prior π.

justification 2 – subjective probability theory

- going back to Savage (1954) and Anscombe and Aumann (1963).
- discussed in chapter 6 of
  Mas-Colell, A., Whinston, M., and Green, J. (1995),
  Microeconomic theory, Oxford University Press
- and maybe in Econ 2010.

- Suppose a decision maker ranks risk functions $R(.,\delta)$ by a **preference relationship** $\succeq$
- properties $\succeq$ might have:
  1. **completeness**: any pair of risk functions can be ranked
  2. **monotonicity**: if the risk function R is (weakly) lower than R' for all θ, than R is (weakly) preferred
  3. **independence**:

$$R^1 \succeq R^2 \Leftrightarrow \alpha R^1 + (1-\alpha)R^3 \succeq \alpha R^2 + (1-\alpha)R^3$$

  for all R^1, R^2, R^3 and $\alpha \in [0,1]$
- Important: this independence has nothing to do with statistical independence

Theorem

If $\succeq$ is complete, monotonic, and satisfies independence, then there exists a prior π such that

$$R(.,\delta^1) \succeq R(.,\delta^2) \Leftrightarrow R(\pi,\delta^1) \leq R(\pi,\delta^2).$$

Intuition of proof:

- Independence and completeness imply linear, parallel indifference sets
- monotonicity makes sure prior is non-negative

Sketch of proof:

Using independence repeatedly, we can show that for all $R^1, R^2, R^3 \in \mathbb{R}^{\mathscr{X}}$, and all $\alpha > 0$,

1. $R^1 \succeq R^2$ iff $\alpha R^1 \succeq \alpha R^2$,
2. $R^1 \succeq R^2$ iff $R^1 + R^3 \succeq R^2 + R^3$,
3. $\{R : R \succeq R^1\} = \{R : R \succeq 0\} + R^1$,
4. $\{R : R \succeq 0\}$ is a convex cone.
5. $\{R : R \succeq 0\}$ is a half space.

The last claim requires completeness. It immediately implies the existence of π. Monotonicity implies that π is not negative.

Testing and the Neyman Pearson lemma

- testing as a decision problem
- goal: decide whether $H_0 : \theta \in \Theta_0$ is true
- decision $a \in \{0, 1\}$ (true / not true)
- statistical test is a decision function $\varphi : X \Rightarrow \{0, 1\}$
- $\varphi = 1$ corresponds to rejecting the null hypothesis
- more generally: randomized tests $\varphi : X \Rightarrow [0, 1]$
- reject H_0 with probability $\varphi(X)$
  (for technical reasons only, as we will see)

Two types of classification error

	truth	
decision a	$\theta \in \Theta_0$	$\theta \notin \Theta_0$
0	☺	Type II error
1	Type I error	☺

The power function

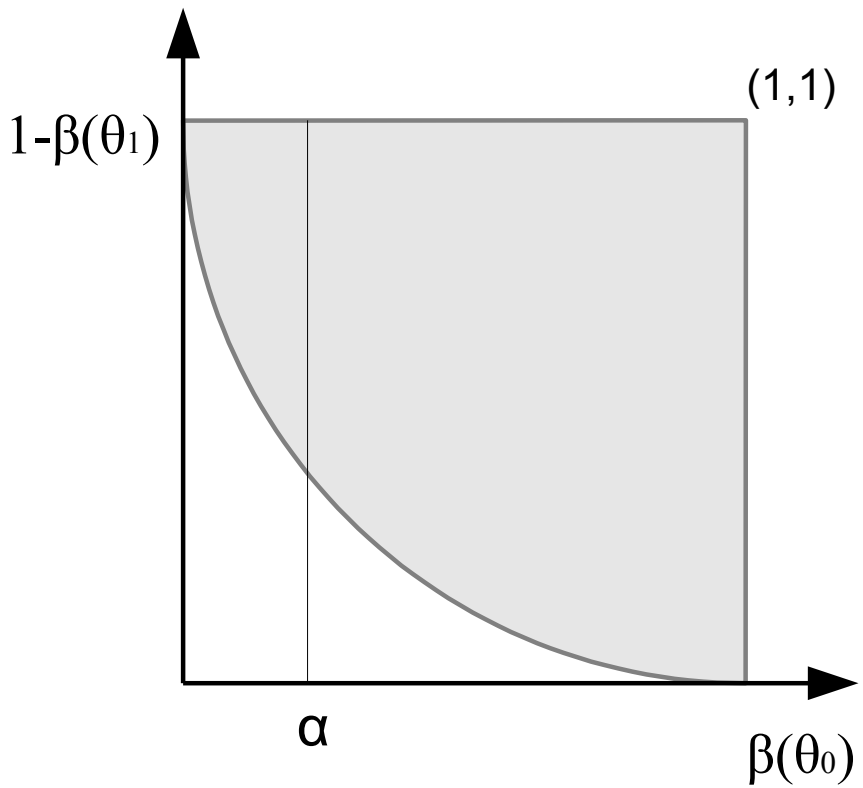
- suppose $X \sim f_\theta(x)$
- f: probability mass function or probability density function
- probability of rejecting H_0 given θ:
  power function

$$\beta(\theta) = E_\theta[\varphi(X)] = \int \varphi(x) f_\theta(x) dx.$$

Classification errors

- suppose that θ has only two points of support, θ_0 and θ_1
- then
  1. P(Type I error) $= \beta(\theta_0)$.
  2. P(Type II error) $= 1 - \beta(\theta_1)$.
- $\beta(\theta_0)$ is called "level" or "**significance**" of the test, often denoted α.
- $\beta(\theta_1)$ is called the "**power**" of a test, and is often denoted β.
- would like to have a small α and a large β

Figure: testing as a decision problem



Suppose we want φ^* that solves

$$\max_{\varphi} \beta(\theta_1) \quad \text{s.t. } \beta(\theta_0) = \alpha$$

for a prespecified level α.

Lemma (Neyman-Pearson)

The solution to this problem is given by

$$\varphi^*(x) = \begin{cases} 1 \text{ for } f_1(x) > \lambda f_0(x) \\ \kappa \text{ for } f_1(x) = \lambda f_0(x) \\ 0 \text{ for } f_1(x) < \lambda f_0(x) \end{cases}$$

where λ and κ are chosen such that $\int \varphi^*(x) f_0(x) dx = \alpha$.

Practice problem

Try to prove this!

Hint:
our problem is to solve

$$\max_{\varphi} \int \varphi(x) f_1(x) dx$$

subject to

$$\int \varphi(x) f_0(x) dx = \alpha$$

and

$$\varphi(x) \in [0, 1].$$

Recall the proposed solution,

$$\varphi^*(x) = \begin{cases} 1 \text{ for } f_1(x) > \lambda f_0(x) \\ \kappa \text{ for } f_1(x) = \lambda f_0(x) \\ 0 \text{ for } f_1(x) < \lambda f_0(x) \end{cases}$$

Proof:

- let $\varphi(x)$ be any other test of level α
  i.e. $\int \varphi(x) f_0(x) dx = \alpha$.
- need to show that
  $\int \varphi^*(x) f_1(x) dx \geq \int \varphi(x) f_1(x) dx$.
- Note that

$$\int (\varphi^*(x) - \varphi(x))(f_1(x) - \lambda f_0(x)) dx \geq 0$$

since $\varphi^*(x) = 1 \geq \varphi(x)$ for all x such that $f_1(x) - \lambda f_0(x) > 0$ and $\varphi^*(x) = 0 \leq \varphi(x)$ for all x such that $f_1(x) - \lambda f_0(x) < 0$.

- Therefore, using $\alpha = \int \varphi(x) f_0(x) dx = \int \varphi^*(x) f_0(x) dx$,

$$\begin{aligned} & \int (\varphi^*(x) - \varphi(x))(f_1(x) - \lambda f_0(x)) dx \\ = & \int (\varphi^*(x) - \varphi(x)) f_1(x) dx \\ = & \int \varphi^*(x) f_1(x) dx - \int \varphi(x) f_1(x) dx \geq 0 \end{aligned}$$

  as required.
- proof in the discrete case: identical with all summations replaced by integrals.

Practice problem

- you observe $X \sim N(\mu, 1)$
- you know that either $\mu = 0$ or $\mu = 1$
- construct the test of largest power for $H_0 : \mu = 0$ and any level α