Data and decisions

Reverse AGT Workshop, Harvard

Maximilian Kasy

February 26, 2015
Some questions of method in applied economics

- Can we choose between policies, if data don’t point-identify policy effects?
- How should we use covariates when running experiments? (Why) should we randomize?
- What is the optimal choice of policy, given observed data?
- There are many estimators in the machine learning literature. Which one should we use?
Claim: Many questions of method reduce to math problems, once the problem is precisely stated.

Have to answer a few key questions.

1. What are the available data?
2. What is the space of feasible actions / policies?
3. What unknown states of the world matter?
4. How do we evaluate an action for a given state of the world?
5. How do the data relate to the state of the world?
6. How do we deal with uncertainty?
Examples from my research


Outline

- Review of statistical decision theory
- Example 1: Why experimenters should not randomize
- Example 2: The risk of machine learning
Statistical decision problems

- Observed data $X$
- Statistical model $X \sim f(x, \theta)$
- State of the world $\theta$
- Decision function $a = \delta(X)$
- Decision $a$
- Loss $L(a, \theta)$
Risk function, Bayes risk, minimax risk

- Risk function:
  \[ R(\delta, \theta) = E_\theta[L(\delta(X), \theta)] \]

- Expected loss of a decision function \( \delta \).

- \( R \) is a function of the true state of the world \( \theta \).

- To rank decision functions \( \delta \), have to aggregate across \( \theta \).

- Solutions:
  1. Bayes risk:
     \[ R(\delta, \pi) = \int R(\delta, \theta) d\pi(\theta) \]
  2. Maximum (worst-case) risk:
     \[ \bar{R}(\delta) = \sup_\theta R(\delta, \theta) \]
Example 1: Why experimenters should not randomize

- Experimental design as a decision problem.
- \( \delta(X, U) \) maps baseline information \( X \) and randomization device \( U \) into treatment assignment.
- Objective: Precise estimator of causal effect of interest.
- Assume \( U \) takes values \( u \in 1, \ldots, \bar{u} \), \( U \) is statistically independent from everything else.
- Let \( \delta^u(.) = \delta(., u) \)
- Then the risk function equals

\[
R(\delta, \theta) = E_\theta[L(\delta(X, U), \theta)] = \sum_u R(\delta^u, \theta) \cdot P(U = u).
\]
Similarly for Bayes risk

\[ R^\pi(\delta) = \int R(\delta, \theta) d\pi(\theta) \]

\[ = \sum_u \int R(\delta^u, \theta) d\pi(\theta) \cdot P(U = u) \]

\[ = \sum_u R^\pi(\delta^u) \cdot P(U = u), \]

and worst-case risk

\[ R^{mm}(\delta) = \sum_u R^{mm}(\delta^u) \cdot P(U = u). \]
Optimal procedures minimize Bayes risk, or worst-case risk.

Note that always

\[ \sum_u R^\pi(\delta^u) \cdot P(U = u) \geq \min_u R^\pi(\delta^u). \]

Therefore any randomized procedure is always (weakly) dominated by a non-randomized one.

This implies: Allowing for randomization never improves risk, and usually makes things worse.

That’s why we don’t randomize when estimating or testing.

That’s why we also shouldn’t randomize when experimenting.
Example 2: The risk of machine learning

- Canonical estimation problem:
- Observe $X_i$, $i = 1, \ldots, n$
- Want to estimate $\mu_i = E[X_i]$
- Loss: $L = \frac{1}{n} \sum_i (\hat{\mu}_i - \mu_i)^2$
- Risk function $=$ mean squared error.
- Key features of machine learning procedures
  1. regularization
  2. data driven choice of tuning parameters
Componentwise estimators

▶ $\hat{\mu}_i = m(X_i, \lambda)$

▶ Ridge:

$$m(x, \lambda) = \arg\min_m [(x - m)^2 + \lambda \cdot m^2] = \frac{1}{1 + \lambda} x$$

▶ Lasso:

$$m(x, \lambda) = \arg\min_m [(x - m)^2 + 2\lambda \cdot |m|]$$

$$= \mathbf{1}(x < -\lambda)(x + \lambda) + \mathbf{1}(x > \lambda)(x - \lambda)$$

▶ Pre-testing:

$$m(x, \lambda) = \mathbf{1}(|x| > \lambda) \cdot x$$
\hat{\mu}_i = m(X_i, \lambda)
Which of such estimators to chose?

- Evaluate estimator based on risk function $= \text{mean squared error}$
- Next slide: Characterization of mean squared error.
- Notation:
  
  $I \sim \text{Unif}\{1, \ldots, n\}$
  
  $X_i \sim P_i$

- Conditional expectation, average conditional variance:
  
  $m^*(y) = E[\mu_i | X_i = x]$
  
  $\nu^* = E[\text{Var}(\mu_i | X_i)]$. 
Theorem (Characterization of risk functions)

Assume

- canonical estimation problem,
- squared error loss,
- component-wise estimation,
- oracle $\lambda^*$.

Then

$$R(m(. , \lambda) , P) = v^* + \| m(., \lambda) - m^*(.) \|^2_{L^2(P^*)},$$

$$R(m(., \lambda^*) , P) = v^* + \arg\min_{\lambda} \| m(., \lambda) - m^*(.) \|^2_{L^2(P^*)},$$

where the norm is with respect to the marginal distribution $P^*$ of $X_1$. 
Thanks for your time!