Data and decisions

Reverse AGT Workshop, Harvard

Maximilian Kasy

February 26, 2015

Some questions of method in applied economics

- Can we chose 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?

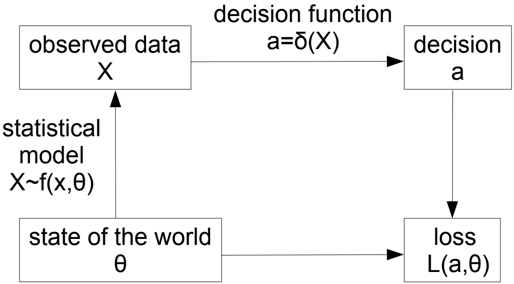
Examples from my research

- Kasy, M. (2015). Partial identification, distributional preferences, and the welfare ranking of policies.
  Review of Economics and Statistics.
- Kasy, M. (2013). Why experimenters should not randomize, and what they should do instead.
  Working Paper.
- Kasy (2014). Using data to inform policy.
  Working Paper.
- Abadie, A. and Kasy, M. (2015). The risk of machine learning.
  Working Paper.

Outline

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

Statistical decision problems



Risk function, Bayes risk, minimax risk

- Risk function:

$$R(\delta, \theta) = E_\theta[L(\delta(X), \theta)]$$

- Expected loss of a decision function δ.
- R is a function of the true state of the world θ.
- To rank decision functions δ, have to aggregate across θ.
- Solutions:
    1. Bayes risk:

    $$R(\delta, \pi) = \int R(\delta, \theta) d\pi(\theta)$$

    2. Maximum (worst-case) risk:

    $$\overline{R}(\delta) = \sup_\theta R(\delta, \theta)$$

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 $u \in 1,\dots,\overline{u}$,
  U is statistically independent from everything else.
- Let $\delta^u(.) = \delta(.,u)$
- Then the risk function equals

$$\begin{aligned} R(\delta,\theta) &= E_\theta[L(\delta(X,U),\theta)] \\ &= \sum_u R(\delta^u,\theta)\cdot P(U=u). \end{aligned}$$

- Similarly for Bayes risk

$$
\begin{aligned}
R^{\pi}(\delta) &= \int R(\delta,\theta)d\pi(\theta) \\
&= \sum_u \int R(\delta^u,\theta)d\pi(\theta)\cdot P(U=u) \\
&= \sum_u R^{\pi}(\delta^u)\cdot P(U=u),
\end{aligned}
$$

- 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: $L = \frac{1}{n} \sum_i (\widehat{\mu}_i - \mu_i)^2$
- Risk function $=$ mean squared error.
- Key features of machine learning procedures
  1. regularization
  2. data driven choice of tuning parameters

Componentwise estimators

- $\widehat{\mu}_i = m(X_i, \lambda)$
- Ridge:

$$m(x,\lambda) = \underset{m}{\text{argmin}} \left[(x-m)^2 + \lambda \cdot m^2\right] = \frac{1}{1+\lambda} x$$

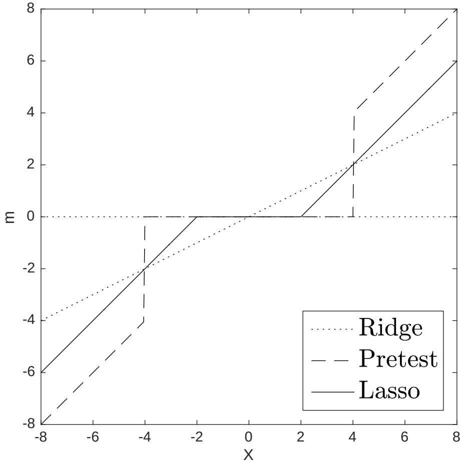
- Lasso:

$$\begin{aligned} m(x,\lambda) &= \underset{m}{\text{argmin}} \left[(x-m)^2 + 2\lambda \cdot |m|\right] \\ &= \mathbf{1}(x < -\lambda)(x+\lambda) + \mathbf{1}(x > \lambda)(x-\lambda) \end{aligned}$$

- Pre-testing:

$$m(x,\lambda) = \mathbf{1}(|x| > \lambda) \cdot x$$

$\widehat{\mu}_i = m(X_i, \lambda)$



Which of such estimators to chose?

- Evaluate estimator based on risk function = mean squared error
- Next slide: Characterization of mean squared error.
- Notation:

$$I \sim Unif\{1,\ldots,n\}$$
$$X_i \sim P_i$$

- Conditional expectation, average conditional variance:

$$m^*(y) = E[\mu_I | X_I = x],$$
$$v^* = E[\mathrm{Var}(\mu_I | X_I)].$$

Theorem (Characterization of risk functions)

Assume

- *canonical estimation problem,*
- *squared error loss,*
- *component-wise estimation,*
- *oracle λ^*.*

Then

$$R(m(.,\lambda)\;,P) = v^* + \qquad \|m(.,\lambda) - m^*(.)\|^2_{L^2(P^*)},$$
$$R(m(.,\lambda^*),P) = v^* + \underset{\lambda}{\operatorname{argmin}} \quad \|m(.,\lambda) - m^*(.)\|^2_{L^2(P^*)},$$

where the norm is with respect to the marginal distribution P^* of X_I.

Thanks for your time!