Machine learning, shrinkage estimation, and economic theory

Maximilian Kasy

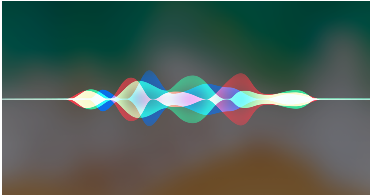
December 14, 2018

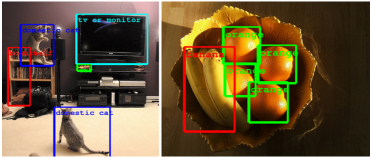
Introduction

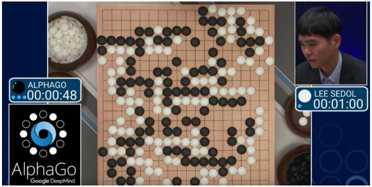
- Recent years saw a boom of "machine learning" methods.
- Impressive advances in domains such as
  - Image recognition, speech recognition,
  - playing chess, playing Go, self-driving cars ...
- Questions:
  - Q Why and when do these methods work?
  - Q Which machine learning methods are useful for what kind of empirical research in economics?
  - Q Can we combine these methods with insights from economic theory?
  - Q What is the risk of general machine learning estimators?

Introduction

Machine learning successes







Some answers to these questions

- Abadie and Kasy (2018) (forthcoming, REStat):
  - Q Why and when do these methods work?
    - A Because in high-dimensional models we can shrink optimally.
  - Q Which machine learning methods are useful for economics?
    - A There is no one method that always works.
      We derive guidelines for choosing methods.
- Fessler and Kasy (2018) (forthcoming, REStat):
  - Q Can we combine these methods with economic theory?
    - A Yes. We construct ML estimators that perform well when theoretical predictions are approximately correct.
- Kasy and Mackey (2018) (work in progress):
  - Q What is the risk of general ML estimators?
    - A In large samples, ML estimators behave like shrinkage estimators of normal means, tuned using Stein's Unbiased Risk Estimate.
      The proof incidentally provides us with an easily computed approximation of n-fold cross-validation.

The risk of machine learning (Abadie and Kasy 2018)

- Many applied settings: Estimation of a **large number of parameters**.
  - Teacher effects, worker and firm effects, judge effects ...
  - Estimation of treatment effects for many subgroups
  - Prediction with many covariates
- Two key ingredients to avoid over-fitting,
  used in all of machine learning:
  - Regularized estimation (**shrinkage**)
  - Data-driven choices of regularization parameters (**tuning**)
- Questions in practice:
  - Q What kind of regularization should we choose?
    What features of the data generating process matter for this choice?
  - Q When do cross-validation or SURE work for tuning?
- We compare **risk functions** to answer these questions.
  (Not average (Bayes) risk or worst case risk!)

The risk of machine learning (Abadie and Kasy 2018)

Recommendations for empirical researchers

1. Use regularization / shrinkage when you have many parameters of interest, and high variance (overfitting) is a concern.
2. Pick a regularization method appropriate for your application:
   2.1 Ridge: Smoothly distributed true effects, no special role of zero
   2.2 Pre-testing: Many zeros, non-zeros well separated
   2.3 Lasso: Robust choice, especially for series regression / prediction
3. Use CV or SURE in high dimensional settings, when number of observations $\gg$ number of parameters.

Using economic theory to improve estimators (Fessler and Kasy 2018)

Two motivations

1. Most regularization methods shrink toward 0, or some other arbitrary point.
   - What if we instead shrink toward parameter values consistent with the predictions of economic theory?
   - This yields uniform improvements of risk, largest when theory is approximately correct.
2. Most economic theories are only approximately correct. Therefore:
   - Testing them always rejects for large samples.
   - Imposing them leads to inconsistent estimators.
   - But shrinking toward them leads to uniformly better estimates.

- Shrinking to theory is an alternative to the standard paradigm of testing theories, and maintaining them while they are not rejected.

Using economic theory to improve estimators (Fessler and Kasy 2018)

Estimator construction

- General construction of estimators shrinking to theory:
  - Parametric empirical Bayes approach.
  - Assume true parameters are theory-consistent parameters plus some random effects.
  - Variance of random effects can be estimated, and determines the degree of shrinkage toward theory.
- We apply this to:
  1. Consumer demand
     shrunk toward negative semi-definite
     compensated demand elasticities.
  2. Effect of labor supply on wage inequality
     shrunk toward CES production function model.
  3. Decision probabilities
     shrunk toward Stochastic Axiom of Revealed Preference.
  4. Expected asset returns
     shrunk toward Capital Asset Pricing Model.

Approximate Cross-Validation (Kasy and Mackey 2018)

- Machine learning estimators come in a bewildering variety. Can we say anything general about their performance?
- Yes!
  1. Many machine learning estimators are penalized m-estimators tuned using cross-validation.
  2. We show: In large samples they behave like penalized least-squares estimators of normal means, tuned using Stein's Unbiased Risk Estimate.
- We know a lot about the behavior of the latter! E.g.:
  1. Uniform dominance relative to unregularized estimators (James and Stein 1961).
  2. We show inadmissibility of Lasso tuned with CV or SURE, and ways to uniformly dominate it.

Approximate Cross-Validation (Kasy and Mackey 2018)

- The proof yields, as a side benefit, a computationally feasible approximation to Cross-Validation.
- n-fold (leave-1-out) Cross-Validation has good properties.
- But it is computationally costly.
  - Need to re-estimate the model n times (for each choice of tuning parameter considered).
  - Machine learning practice therefore often uses k-fold CV, or just one split into estimation and validation sample.
  - But those are strictly worse methods of tuning.
- We consider an alternative: Approximate (n-fold) CV.
  - Approximate leave-1-out estimator using influence function.
  - If you can calculate standard errors, you can calculate this.
  - Only need to estimate model once!

The risk of machine learning (Abadie and Kasy, 2018)

Roadmap:

1. Stylized setting: Estimation of many means
2. A useful family of examples: Spike and normal DGP
   - Comparing mean squared error as a function of parameters
3. Empirical applications
   - Neighborhood effects (Chetty and Hendren, 2015)
   - Arms trading event study (DellaVigna and La Ferrara, 2010)
   - Nonparametric Mincer equation (Belloni and Chernozhukov, 2011)
4. Monte Carlo Simulations
5. Uniform loss consistency of tuning methods

Stylized setting: Estimation of many means

- Observe n random variables $X_1, \ldots, X_n$ with means $\mu_1, \ldots, \mu_n$.
- Many applications: X_i equal to OLS estimated coefficients.
- **Componentwise estimators**: $\widehat{\mu}_i = m(X_i, \lambda)$, where $m : \mathbb{R} \times [0, \infty] \mapsto \mathbb{R}$ and λ may depend on $(X_1, \ldots, X_n)$.
- Examples: Ridge, Lasso, Pretest.

Shrinkage estimators

- Ridge:

$$
\begin{aligned}
m_R(x,\lambda) &= \underset{c\in\mathbb{R}}{\operatorname{argmin}} \left((x-c)^2 + \lambda c^2\right) \\
&= \frac{1}{1+\lambda}\,x.
\end{aligned}
$$

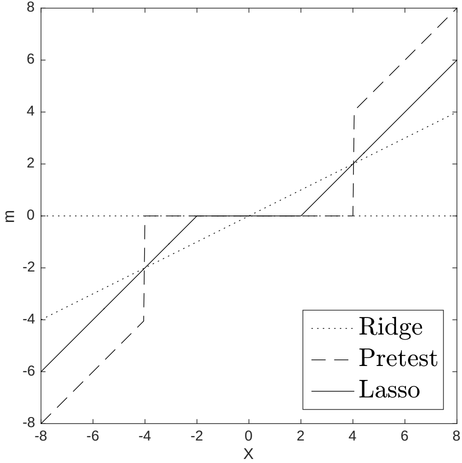
- Lasso:

$$
\begin{aligned}
m_L(x,\lambda) &= \underset{c\in\mathbb{R}}{\operatorname{argmin}} \left((x-c)^2 + 2\lambda|c|\right) \\
&= \mathbf{1}(x<-\lambda)(x+\lambda) + \mathbf{1}(x>\lambda)(x-\lambda).
\end{aligned}
$$

- Pre-test:

$$
m_{PT}(x,\lambda) = \mathbf{1}(|x|>\lambda)x.
$$

Shrinkage estimators



- X: unregularized estimate.
- $m(X, \lambda)$: shrunken estimate.

Loss and risk

- Compound squared error **loss**: $L(\widehat{\boldsymbol{\mu}}, \boldsymbol{\mu}) = \frac{1}{n}\sum_i (\widehat{\mu}_i - \mu_i)^2$
- Empirical Bayes **risk**:
  $\mu_1, \dots, \mu_n$ as **random effects**, $(X_i, \mu_i) \sim \pi$,

$$\bar{R}(m(\cdot, \lambda), \pi) = E_\pi[(m(X_i, \lambda) - \mu_i)^2].$$

- Conditional expectation:

$$\bar{m}^*_\pi(x) = E_\pi[\mu | X = x]$$

- **Theorem**: The empirical Bayes risk of $m(\cdot, \lambda)$ can be written as

$$\bar{R} = const. + E_\pi\left[(m(X, \lambda) - \bar{m}^*_\pi(X))^2\right].$$

- $\Rightarrow$ Performance of estimator $m(\cdot, \lambda)$ depends on how closely it approximates $\bar{m}^*_\pi(\cdot)$.

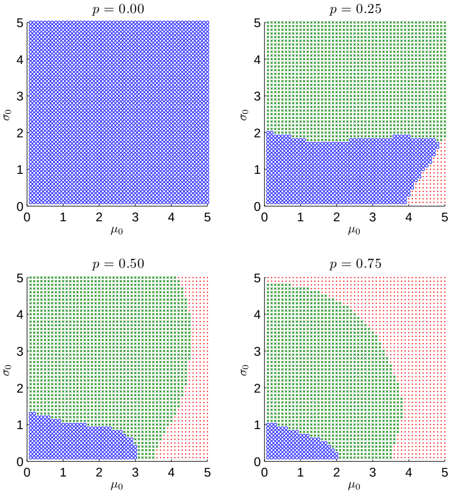
A useful family of examples: Spike and normal DGP

- Assume $X_i \sim N(\mu_i, 1)$.
- Distribution of μ_i across i:

$$\begin{array}{ll} \text{Fraction } p & \mu_i = 0 \\ \text{Fraction } 1-p & \mu_i \sim N(\mu_0, \sigma_0^2) \end{array}$$

- Covers many interesting settings:
  - $p = 0$: Smooth distribution of true parameters.
  - $p \gg 0$, μ_0 or σ_0^2 large: Sparsity, non-zeros well separated.
- Consider Ridge, Lasso, Pretest, optimal shrinkage function.
- Assume λ is chosen optimally (will return to that).

Best estimator (based on analytic derivation of risk function)



∘ Ridge, x Lasso, • Pretest

Applications

- **Neighborhood effects:**
  The effect of location during childhood on adult income
  (Chetty and Hendren, 2015)
- **Arms trading event study:**
  Changes in the stock prices of arms manufacturers following changes in the intensity of conflicts in countries under arms trade embargoes
  (DellaVigna and La Ferrara, 2010)
- **Nonparametric Mincer equation:**
  A nonparametric regression equation of log wages on education and potential experience
  (Belloni and Chernozhukov, 2011)

Estimated Risk

- Stein's unbiased risk estimate $\widehat{R}$
- at the optimized tuning parameter $\widehat{\lambda}^*$
- for each application and estimator considered.

	n		Ridge	Lasso	Pre-test
location effects	595	$\widehat{R}$	**0.29**	0.32	0.41
		$\widehat{\lambda}^*$	2.44	1.34	5.00
arms trade	214	$\widehat{R}$	0.50	0.06	-**0.02**
		$\widehat{\lambda}^*$	0.98	1.50	2.38
returns to education	65	$\widehat{R}$	1.00	**0.84**	0.93
		$\widehat{\lambda}^*$	0.01	0.59	1.14

Monte Carlo simulations

- Spike and normal DGP
- Number of parameters $n = 50, 200, 1000$
- λ chosen using SURE, CV with $4, 20$ folds
- Relative performance: As predicted.
- Also compare to NPEB estimator of Koenker and Mizera (2014), based on estimating m^*_π.

Table: Average Compound Loss Across 1000 Simulations with $N = 50$

p	μ_0	σ_0	SURE ridge	SURE lasso	SURE pretest	Cross-Validation ($k = 4$) ridge	Cross-Validation ($k = 4$) lasso	Cross-Validation ($k = 4$) pretest	Cross-Validation ($k = 20$) ridge	Cross-Validation ($k = 20$) lasso	Cross-Validation ($k = 20$) pretest	NPEB
0.00	0	2	0.80	0.89	1.02	0.83	0.90	1.12	0.81	0.88	1.12	0.94
0.00	0	6	0.97	0.99	1.01	0.97	0.99	1.05	0.97	0.99	1.07	1.21
0.00	2	2	0.89	0.96	1.01	0.90	0.95	1.06	0.89	0.95	1.09	0.93
0.00	2	6	0.97	0.99	1.01	0.99	1.00	1.06	0.97	0.98	1.07	1.21
0.00	4	2	0.95	1.00	1.01	0.95	0.99	1.02	0.95	1.00	1.04	0.93
0.00	4	6	0.99	1.00	1.02	0.99	1.00	1.05	0.99	1.00	1.07	1.21
0.50	0	2	0.67	0.64	0.94	0.69	0.64	0.96	0.67	0.62	0.90	0.69
0.50	0	6	0.95	0.80	0.90	0.95	0.79	0.87	0.96	0.78	0.84	0.84
0.50	2	2	0.80	0.72	0.96	0.82	0.72	0.96	0.81	0.72	0.93	0.73
0.50	2	6	0.96	0.80	0.92	0.95	0.77	0.83	0.95	0.78	0.82	0.86
0.50	4	2	0.91	0.82	0.95	0.92	0.81	0.90	0.92	0.81	0.87	0.75
0.50	4	6	0.97	0.81	0.93	0.97	0.79	0.83	0.96	0.78	0.79	0.85
0.95	0	2	0.18	0.15	0.17	0.17	0.12	0.15	0.18	0.13	0.19	0.17
0.95	0	6	0.49	0.21	0.16	0.51	0.19	0.16	0.49	0.19	0.19	0.16
0.95	2	2	0.26	0.17	0.18	0.27	0.16	0.18	0.27	0.17	0.23	0.17
0.95	2	6	0.53	0.21	0.15	0.53	0.19	0.15	0.53	0.20	0.18	0.16
0.95	4	2	0.44	0.21	0.18	0.45	0.20	0.18	0.45	0.20	0.22	0.18
0.95	4	6	0.57	0.21	0.15	0.58	0.19	0.14	0.57	0.20	0.18	0.16

Table: Average Compound Loss Across 1000 Simulations with $N = 200$

p	μ_0	σ_0	SURE ridge	SURE lasso	SURE pretest	Cross-Validation ($k = 4$) ridge	Cross-Validation ($k = 4$) lasso	Cross-Validation ($k = 4$) pretest	Cross-Validation ($k = 20$) ridge	Cross-Validation ($k = 20$) lasso	Cross-Validation ($k = 20$) pretest	NPEB
0.00	0	2	0.80	0.87	1.01	0.82	0.88	1.04	0.80	0.87	1.04	0.86
0.00	0	6	0.98	0.99	1.01	0.98	0.99	1.02	0.98	0.99	1.03	1.09
0.00	2	2	0.89	0.95	1.00	0.90	0.95	1.02	0.89	0.94	1.03	0.86
0.00	2	6	0.98	1.00	1.01	0.98	0.99	1.02	0.98	0.99	1.03	1.10
0.00	4	2	0.95	1.00	1.00	0.96	1.00	1.01	0.95	1.00	1.02	0.86
0.00	4	6	0.98	0.99	1.01	0.98	0.99	1.01	0.99	0.99	1.03	1.09
0.50	0	2	0.67	0.61	0.90	0.69	0.62	0.93	0.67	0.61	0.90	0.63
0.50	0	6	0.94	0.77	0.86	0.95	0.76	0.82	0.95	0.77	0.83	0.77
0.50	2	2	0.80	0.70	0.94	0.82	0.71	0.93	0.80	0.69	0.91	0.65
0.50	2	6	0.95	0.78	0.88	0.96	0.78	0.83	0.95	0.77	0.82	0.77
0.50	4	2	0.91	0.80	0.94	0.92	0.81	0.87	0.91	0.80	0.87	0.67
0.50	4	6	0.96	0.79	0.92	0.97	0.79	0.81	0.97	0.78	0.80	0.76
0.95	0	2	0.17	0.12	0.14	0.17	0.12	0.14	0.17	0.12	0.15	0.12
0.95	0	6	0.61	0.18	0.14	0.62	0.18	0.14	0.61	0.18	0.14	0.14
0.95	2	2	0.28	0.16	0.17	0.29	0.16	0.18	0.28	0.15	0.17	0.14
0.95	2	6	0.63	0.19	0.14	0.64	0.19	0.14	0.63	0.18	0.14	0.13
0.95	4	2	0.49	0.20	0.17	0.50	0.20	0.17	0.48	0.19	0.17	0.14
0.95	4	6	0.68	0.19	0.13	0.70	0.19	0.13	0.67	0.19	0.14	0.13

Table: Average Compound Loss Across 1000 Simulations with $N = 1000$

			SURE			Cross-Validation ($k = 4$)			Cross-Validation ($k = 20$)			NPEB
p	μ_0	σ_0	ridge	lasso	pretest	ridge	lasso	pretest	ridge	lasso	pretest	
0.00	0	2	0.80	0.87	1.01	0.81	0.87	1.01	0.80	0.86	1.01	0.82
0.00	0	6	0.97	0.98	1.00	0.98	0.98	1.00	0.97	0.98	1.01	1.02
0.00	2	2	0.89	0.94	1.00	0.90	0.95	1.00	0.89	0.94	1.01	0.82
0.00	2	6	0.97	0.98	1.00	0.98	0.99	1.00	0.97	0.98	1.01	1.02
0.00	4	2	0.95	1.00	1.00	0.96	1.00	1.00	0.95	0.99	1.00	0.82
0.00	4	6	0.98	0.99	1.00	0.98	0.99	1.00	0.98	0.99	1.01	1.02
0.50	0	2	0.67	0.60	0.87	0.68	0.61	0.90	0.67	0.60	0.87	0.60
0.50	0	6	0.95	0.77	0.81	0.95	0.77	0.82	0.95	0.76	0.81	0.72
0.50	2	2	0.80	0.70	0.90	0.81	0.71	0.90	0.80	0.69	0.89	0.62
0.50	2	6	0.95	0.77	0.80	0.96	0.78	0.81	0.95	0.77	0.80	0.71
0.50	4	2	0.91	0.80	0.87	0.92	0.80	0.84	0.91	0.80	0.84	0.63
0.50	4	6	0.96	0.78	0.87	0.97	0.78	0.79	0.96	0.78	0.78	0.70
0.95	0	2	0.17	0.11	0.14	0.17	0.12	0.14	0.17	0.11	0.14	0.11
0.95	0	6	0.63	0.18	0.13	0.65	0.18	0.14	0.64	0.17	0.14	0.12
0.95	2	2	0.28	0.15	0.16	0.29	0.15	0.18	0.29	0.14	0.17	0.12
0.95	2	6	0.66	0.18	0.13	0.67	0.18	0.14	0.66	0.18	0.13	0.12
0.95	4	2	0.50	0.19	0.16	0.51	0.19	0.17	0.50	0.19	0.16	0.12
0.95	4	6	0.72	0.18	0.13	0.73	0.19	0.13	0.71	0.18	0.13	0.12

Some theory: Estimating λ

- Can we consistently estimate the optimal λ^*, and do almost as well as if we knew it?
- Answer: Yes, for large *n*, suitably bounded moments.
- We show this for two methods:
  1. Stein's Unbiased Risk Estimate (SURE) (requires normality)
  2. Cross-validation (CV) (requires panel data)

Uniform loss consistency

- Shorthand notation for loss:

$$L_n(\lambda) = \frac{\mathbf{1}}{\mathbf{n}} \sum_i (m(X_i, \lambda) - \mu_i)^2$$

- **Definition:**
  Uniform loss consistency of $m(., \widehat{\lambda})$ for $m(., \bar{\lambda}^*)$:

$$\sup_\pi P_\pi \left(\left| L_n(\widehat{\lambda}) - L_n(\bar{\lambda}^*) \right| > \epsilon \right) \to 0$$

- as $n \to \infty$ for all $\epsilon > 0$, where

$$\mathbf{P_i} \sim^{\mathbf{iid}} \pi.$$

Minimizing estimated risk

- Estimate λ^* by minimizing estimated risk:

$$\widehat{\lambda}^* = \underset{\lambda}{\operatorname{argmin}}\ \widehat{R}(\lambda)$$

- Different estimators $\widehat{R}(\lambda)$ of risk: CV, SURE
- **Theorem**: Regularization using SURE or CV
  is uniformly loss consistent
  as $n \to \infty$ in the random effects setting
  under some regularity conditions.
- Contrast with Leeb and Pötscher (2006)!
  (fixed dimension of parameter vector)
- Key ingredient: uniform laws of larger numbers to get
  convergence of $L_n(\lambda)$, $\widehat{R}(\lambda)$.

Using economic theory to improve estimators (Fessler and Kasy 2018)

Two motivations

1. Most regularization methods shrink toward 0, or some other arbitrary point.
    - What if we instead shrink toward parameter values consistent with the predictions of economic theory?
    - This yields uniform improvements of risk, largest when theory is approximately correct.
2. Most economic theories are only approximately correct. Therefore:
    - Testing them always rejects for large samples.
    - Imposing them leads to inconsistent estimators.
    - But shrinking toward them leads to uniformly better estimates.

Review: Parametric empirical Bayes

- Parameters β, hyper-parameters τ
- **Model**:

$$Y|\beta \sim f(Y|\beta)$$

- **Family of priors**:

$$\beta \sim \pi(\beta|\tau)$$

- **Marginal density** of Y:

$$Y|\tau \sim g(Y|\tau) := \int f(Y|\beta)\pi(\beta|\tau)d\beta$$

- Estimation of hyperparameters (tuning): marginal MLE

$$\widehat{\tau} = \underset{\theta}{\text{argmax}}\ g(Y|\tau).$$

- Estimation of β (shrinkage):

$$\widehat{\beta} = E[\beta|Y, \tau = \widehat{\tau}].$$

Our setup for estimator construction

- Goal: constructing estimators shrinking to theory.
- Preliminary unrestricted estimator:

$$\widehat{\beta}|\beta \sim N(\beta, V)$$

- Restrictions implied by theoretical model:

$$\beta^0 \in B^0 = \{b : \; R_1 \cdot b = 0, \; R_2 \cdot b \leq 0\}.$$

- Empirical Bayes (random coefficient) construction:

$$\begin{aligned} \beta &= \beta^0 + \zeta, \\ \zeta &\sim N(0, \tau^2 \cdot I), \\ \beta^0 &\in B^0. \end{aligned}$$

Solving for the empirical Bayes estimator

- Marginal distribution of $\widehat{\beta}$ given β_0, τ^2:

$$\widehat{\beta}|\beta_0, \tau^2 \sim N(\beta^0, \tau^2 \cdot I + V)$$

- Maximum likelihood estimation of β_0, τ^2 (tuning):

$$(\widehat{\beta}^0, \widehat{\tau}^2) = \underset{b^0 \in B^0,\ t^2 \geq 0}{\operatorname{argmin}} \log\left(\det\left(\tau^2 \cdot I + \widehat{V}\right)\right) + (\widehat{\beta} - b^0)' \cdot \left(\tau^2 \cdot I + \widehat{V}\right)^{-1} \cdot (\widehat{\beta} - b^0).$$

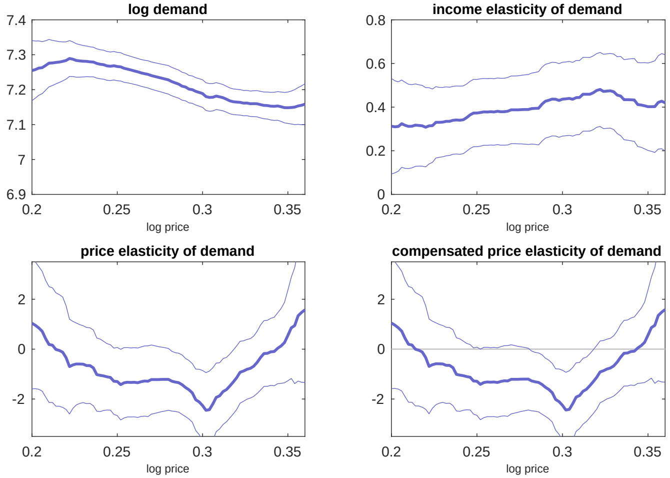
- "Bayes" estimation of β (shrinkage):

$$\widehat{\beta}^{EB} = \widehat{\beta}^0 + \left(I + \frac{1}{\widehat{\tau}^2}\widehat{V}\right)^{-1} \cdot (\widehat{\beta} - \widehat{\beta}^0).$$

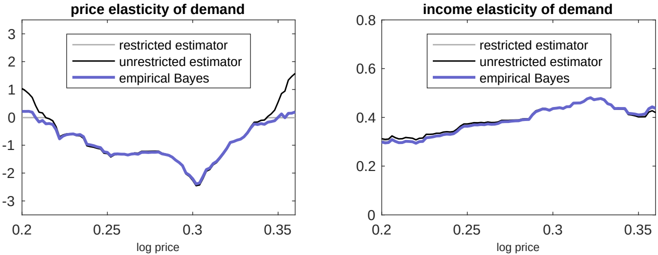
Application 1: Consumer demand

- Consumer choice and the restrictions on compensated demand implied by utility maximization.
- High dimensional parameters if we want to estimate demand elasticities at many different price and income levels.
- Theory we are shrinking to:
  - Negative semi-definiteness of compensated quantile demand elasticities,
  - which holds under arbitrary preference heterogeneity by Dette et al. (2016).
- Application as in Blundell et al. (2017):
  - Price and income elasticity of gasoline demand,
  - 2001 National Household Travel Survey (NHTS).

Unrestricted demand estimation



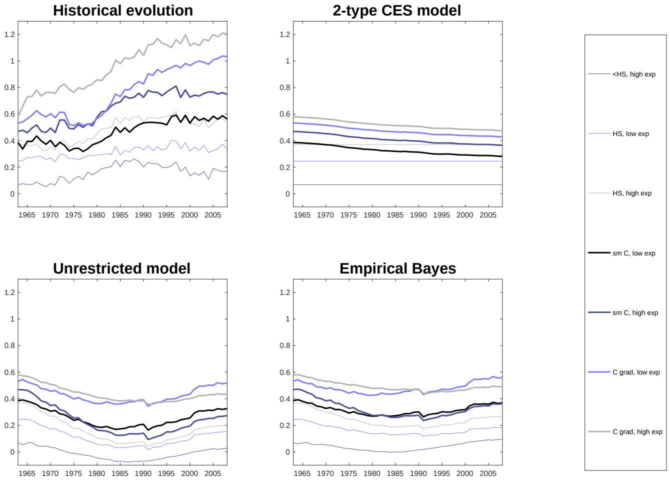
Empirical Bayes demand estimation



Application 2: Wage inequality

- Estimation of labor demand systems, as in literatures on
  - skill-biased technical change, e.g. Autor et al. (2008),
  - impact of immigration, e.g. Card (2009).
- High dimensional parameters if we want to allow for flexible interactions between the supply of many types of workers.
- Theory we are shrinking to:
  - wages equal to marginal productivity,
  - output determined by a CES production function.
- Data: US State-level panel for the years 1960, 1970, 1980, 1990, and 2000 using the Current Population Survey, and 2006 using the American Community Survey.

Counterfactual evolution of US wage inequality



Approximate Cross-Validation (Kasy and Mackey 2018)

- Machine learning estimators come in a bewildering variety. Can we say anything general about their performance?
- Yes! Many machine learning estimators are penalized m-estimators tuned using cross-validation.
- We show: In large samples they behave like penalized least-squares estimators of normal means, tuned using Stein's Unbiased Risk Estimate.
- Next few slides:
  - Approximate Cross-Validation using influence functions.
  - Taking limits of the resulting expressions yields normal means / Stein's Unbiased Risk Estimate.

Penalized M-estimation

- Suppose we are interested in $\beta = \text{argmin}_{\,b}\, E[m(X, \beta)]$.
- Estimate β using penalized M-estimation,

$$\widehat{\beta}(\lambda) = \underset{b}{\text{argmin}} \sum_i m(X_i, b) + \pi(b, \lambda).$$

- General class of machine learning estimators, includes
  - Ridge, Lasso, Pretest in the normal means model,
    and more generally penalized (linear) regression for forecasting,
  - empirical Bayes estimators of the form just considered,
  - regularized deep neural nets,
  - ...

Estimating out-of-sample prediction error

- We would like to choose λ to minimize the out-of-sample prediction error

$$R(\lambda) = E[m(X, \widehat{\beta}(\lambda))].$$

- Leave-one-out estimator, n-fold cross-validation

$$\widehat{\beta}_{-i}(\lambda) = \underset{b}{\operatorname{argmin}} \sum_{j \neq i} m(X_j, b) + \pi(b, \lambda).$$

$$CV(\lambda) = \tfrac{1}{n} \sum_{i} m(X_i, \widehat{\beta}_{-i}(\lambda)).$$

- Computationally costly to re-estimate β for every choice of i and λ!

- Notation for Hessian, gradients:

$$H = \left(\sum_j m_{bb}(X_j, \widehat{\beta}(\lambda)) + \pi_{bb}(\widehat{\beta}(\lambda), \lambda) \right)$$

$$g_i = m_b(X_i, \widehat{\beta}(\lambda)).$$

- First-order approximation to leave-one-out estimator (possibly infinite 2nd derivatives):

$$\widehat{\beta}_{-i}(\lambda) - \widehat{\beta}(\lambda) \approx H^{-1} \cdot g_i.$$

- In-sample prediction error:

$$\bar{R}(\lambda) = \tfrac{1}{n} \sum_i m(X_i, \widehat{\beta}(\lambda)).$$

- Another first-order approximation:

$$CV(\lambda) \approx \bar{R}(\lambda) + \tfrac{1}{n}\sum_i g_i \cdot \left(\widehat{\beta}_{-i}(\lambda) - \widehat{\beta}(\lambda)\right).$$

- Combining the two approximations:

$$CV(\lambda) \approx \bar{R}(\lambda) + \frac{1}{n}\sum_i g_i^t \cdot H^{-1} \cdot g_i.$$

- $\bar{R}$, g_i and H are automatically available if Newton-Raphson was used for finding $\widehat{\beta}(\lambda)$!
- If not, could approximate them without bias using random subsample.
- Large sample limit of this expression gives SURE in the normal means model.

Summary and conclusion

- Machine learning and related methods are driven by shrinkage/regularization and tuning.
- Which regularization performs best depends on the application / distribution of underlying parameters.
- Cross-validation and SURE have strong guarantees to yield almost optimal tuning.
- Estimation using shrinkage/regularization and tuning performs better than unregularized estimation, for *every* data-generating process!!
- The improvements are largest around the points that we are shrinking to.
- We can shrink to restrictions implied by economic theory to get large improvements if theory is approximately correct.

Summary and conclusion

- Proposed estimator construction to shrink toward theory:
  1. First-stage: estimate neglecting the theoretical predictions.
  2. Assume: True parameter values = parameter values conforming to the theory + noise.
  3. Maximize the marginal likelihood of the data given the hyperparameters. (Variance of noise $\approx$ model fit!)
  4. Bayesian updating | estimated hyperparameters, data $\Rightarrow$ estimates of the parameters of interest.
- Two characterizations of risk, showing uniform dominance (in the paper):
  1. High-dimension asymptotics (simple and transparent).
  2. Exact (somewhat more restrictive setting).
- *n*-fold CV is computationally too costly in most ML settings.
  - Feasible alternative that performs uniformly well: approximate CV.
  - Provides deep connection to normal means model, SURE.
  - Allows to characterize risk functions of general penalized m-estimators.

Thank you!