Choosing among regularized estimators in empirical economics

Alberto Abadie   Maximilian Kasy

January 5, 2018

Introduction

- 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:
  - Regularized estimation (**shrinkage**)
  - Data-driven choices of regularization parameters (**tuning**)
- Questions in practice:
  1. What kind of regularization should we choose?
     What features of the data generating process matter for this choice?
  2. 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!)

Recommendations for empirical researchers

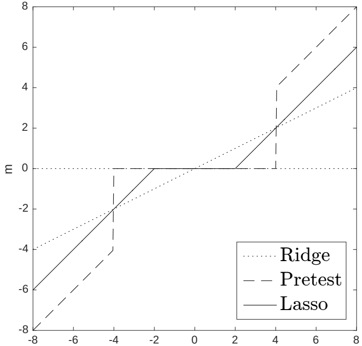
1. Use regularization / shrinkage when you have many parameters of interest, and high variance (overfitting) is a concern.
2. Pick a regularization method appropriate for your application:
   1. Ridge: Smoothly distributed true effects, no special role of zero
   2. Pre-testing: Many zeros, non-zeros well separated
   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.

Outline

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

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.



Loss and risk

- Compound squared error **loss**: $L(\widehat{\mu},\mu) = \frac{1}{n}\sum_i(\widehat{\mu}_i - \mu_i)^2$
- 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 $\bar{m}^*_\pi(\cdot)$.

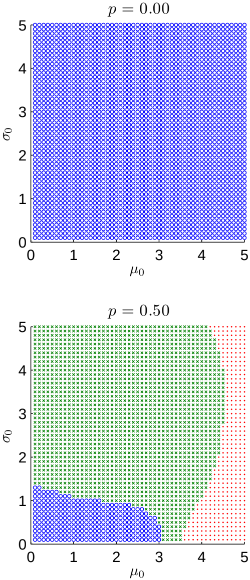
A useful family of examples: Spike and normal DGP

- 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, pre-test, optimal shrinkage function.
- Assume λ is chosen optimally (will return to that).

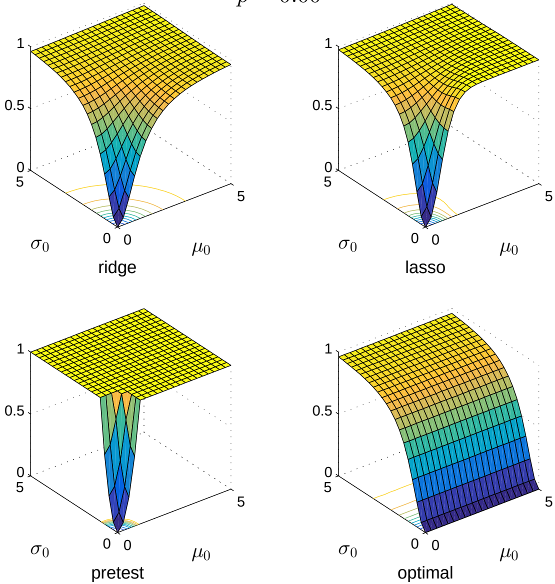
Best estimator



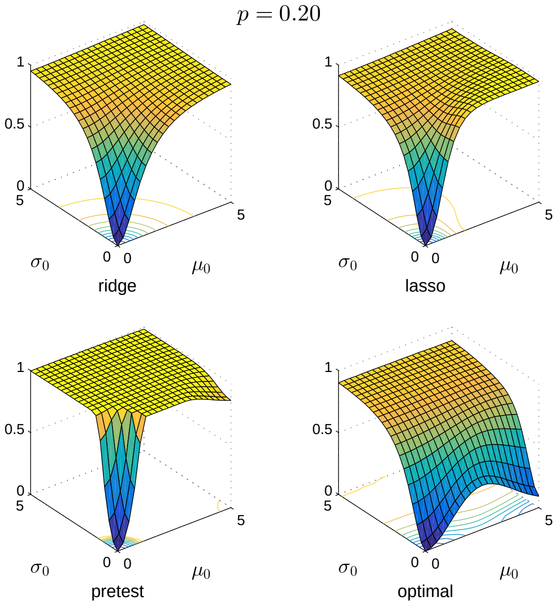
○ is ridge, x is lasso, · is pretest

Mean squared error

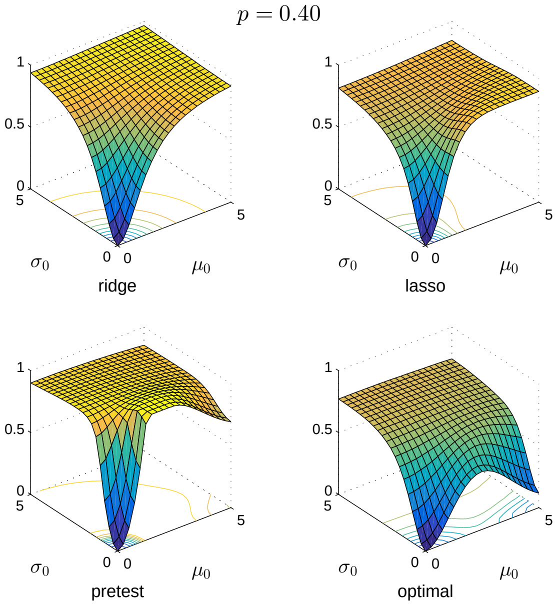
$p = 0.00$



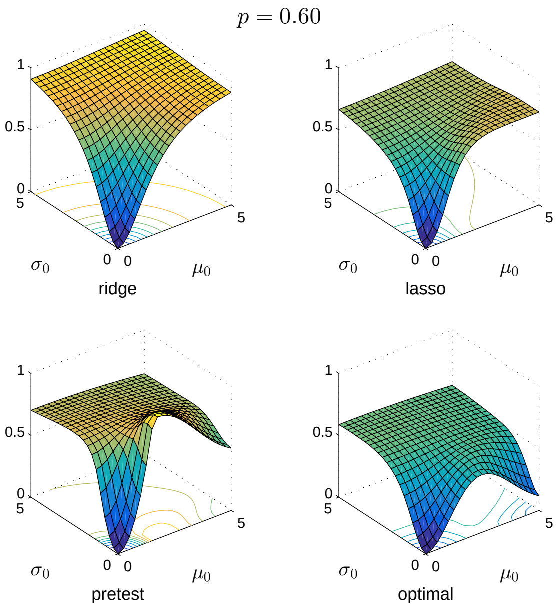
Mean squared error



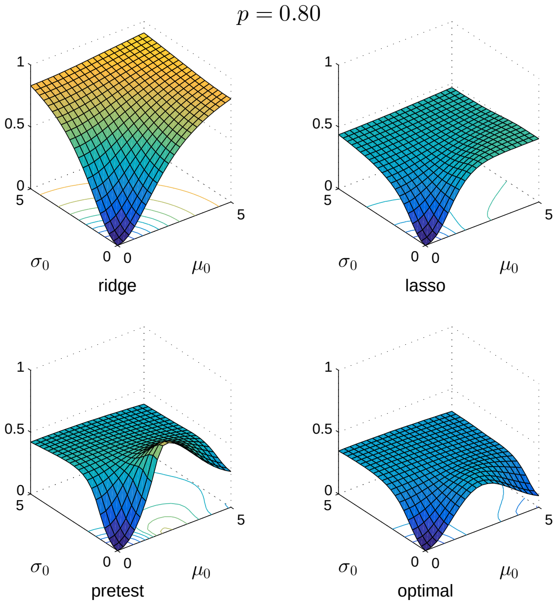
Mean squared error



Mean squared error



Mean squared error



Applications

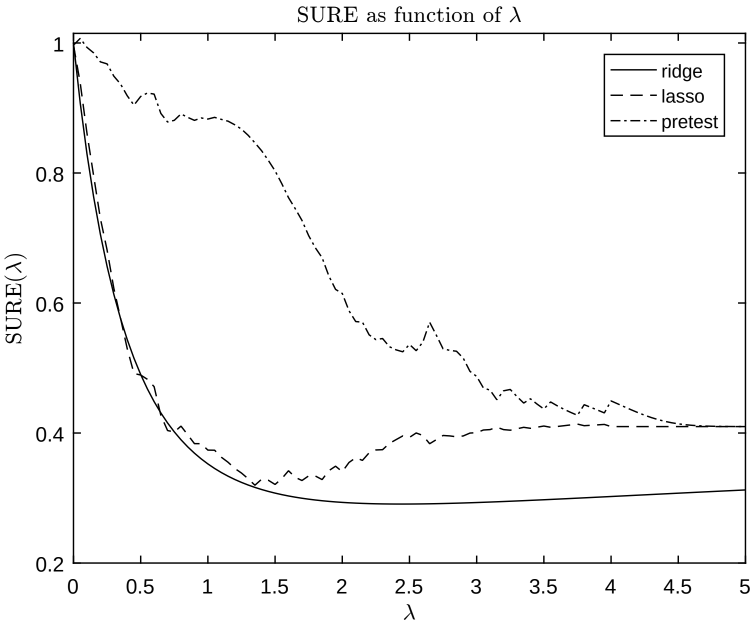
- **Neighborhood effects:**
  The effect of location during childhood on adult income
  (Chetty and Hendren, 2015)
- **Arms trading event study:**
  Changes in the stock prices of arms manufacturers following changes in the intensity of conflicts in countries under arms trade embargoes
  (DellaVigna 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

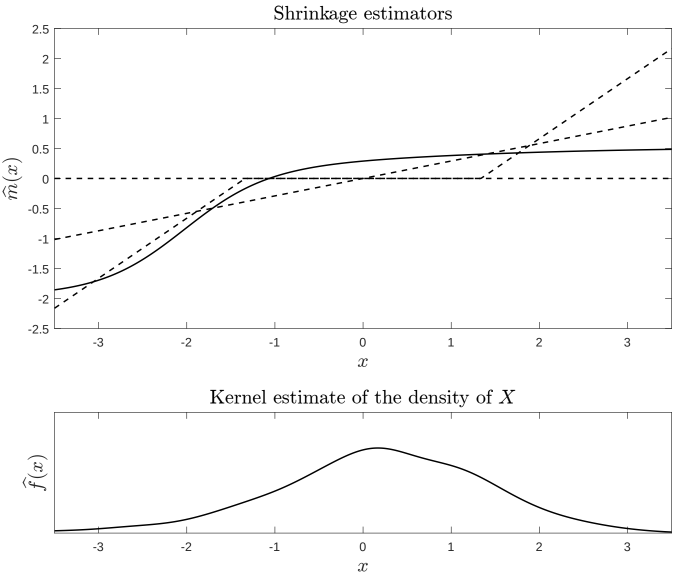
- Stein's unbiased risk estimate $\widehat{R}$
- at the optimized tuning parameter $\widehat{\lambda}^*$
- for each application and estimator considered.

	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

Neighborhood effects: SURE estimates

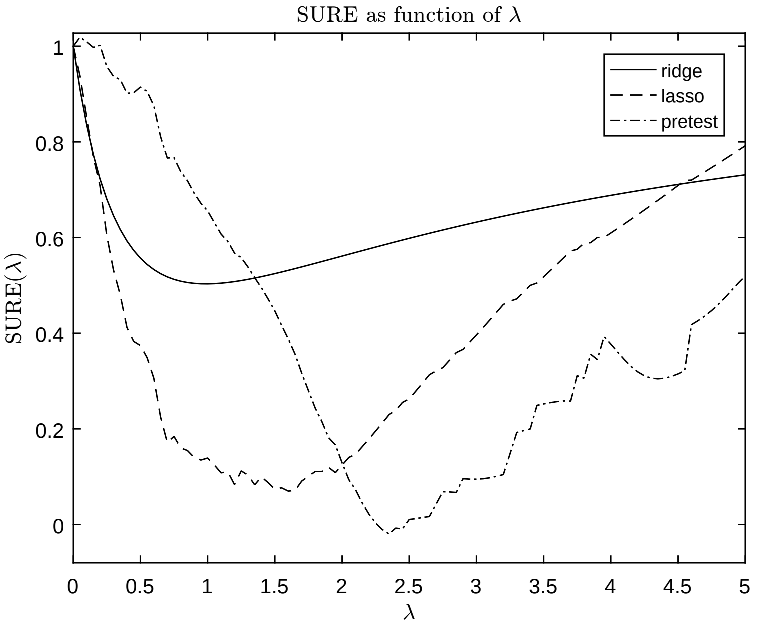


Neighborhood effects: shrinkage estimators

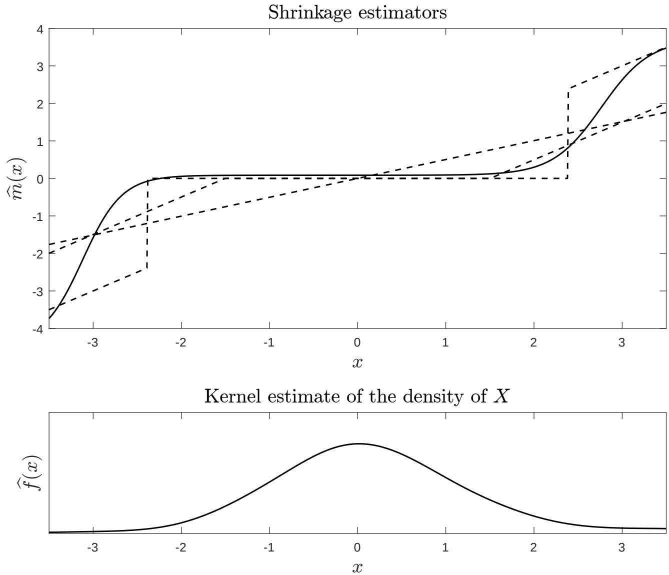


Solid line in top figure is an estimate of $\bar{m}^*_\pi(x)$

Arms event study: SURE estimates

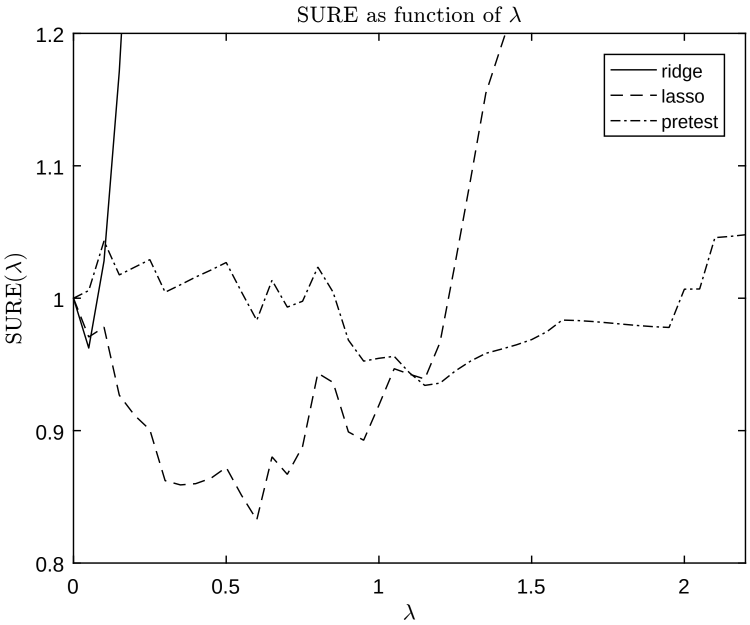


Arms event study: shrinkage estimators

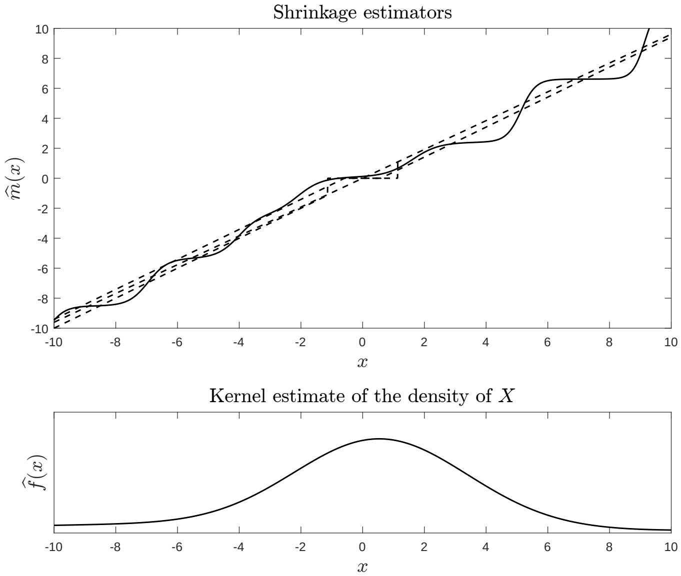


Solid line in top figure is an estimate of $\bar{m}^*_\pi(x)$

Mincer regression: SURE estimates



Mincer regression: shrinkage estimators



Solid line in top figure is an estimate of $\bar{m}^*_\pi(x)$

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 $N = 50$

			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	0.89	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 $N = 200$

			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	0.82	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:

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

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

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

- as $n \to \infty$ for all $\varepsilon > 0$, 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 $L_n(\lambda)$, $\widehat{R}(\lambda)$.

Thank you!

Bonus material

Componentwise estimators

- Ridge:

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

- Lasso:

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

- Pre-test:

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

Connection to linear regression and prediction

- Normal linear regression model:

$$Y|\boldsymbol{W} \sim N(\boldsymbol{W}'\beta, \sigma^2).$$

- Sample $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_n$. Let $\boldsymbol{\Omega} = \frac{1}{N}\sum_{j=1}^{N} \boldsymbol{W}_j \boldsymbol{W}_j'$.
- Draw new value of covariates from sample for prediction.
- Expected squared prediction error

$$\tilde{R} = E\left[(Y - W\widehat{\beta})^2\right] = \operatorname{tr}\left(\boldsymbol{\Omega} \cdot E[(\widehat{\beta} - \beta)(\widehat{\beta} - \beta)']\right) + \sigma^2.$$

- Orthogonalize: Let $\mu = \boldsymbol{\Omega}^{1/2}\beta$, $\boldsymbol{X} = \boldsymbol{\Omega}^{1/2}\widehat{\beta}^{OLS}$, $\widehat{\mu}_i = m(X_i, \lambda)$.
- Then

$$\boldsymbol{X} \sim N\left(\mu, \frac{\sigma^2}{N}\boldsymbol{I}_n\right),$$

and

$$\tilde{R} = E\left[\sum_i (\widehat{\mu}_i - \mu_i)^2\right] + E[\varepsilon^2].$$

Spike-and-normal: Optimal shrinkage function

Assume

- $\mu_1, \ldots, \mu_n$ are drawn independently from a distribution with probability mass p at zero, and normal with mean μ_0 and variance σ_0^2 elsewhere.
- Conditional on μ_i, X_i follows a normal distribution with mean μ_i and variance σ^2.
- Then, the optimal shrinkage function is:

$$m_\pi^*(x) = \frac{(1-p)\frac{1}{\sqrt{\sigma_0^2+\sigma^2}}\phi\left(\frac{x-\mu_0}{\sqrt{\sigma_0^2+\sigma^2}}\right)\frac{\mu_0\sigma^2+x\sigma_0^2}{\sigma_0^2+\sigma^2}}{p\frac{1}{\sigma}\phi\left(\frac{x}{\sigma}\right)+(1-p)\frac{1}{\sqrt{\sigma_0^2+\sigma^2}}\phi\left(\frac{x-\mu_0}{\sqrt{\sigma_0^2+\sigma^2}}\right)}.$$

Two methods to estimate risk

1. Stein's Unbiased Risk Estimate (SURE)
   Requires normality of X_i.

$$\widehat{R}(\lambda) = \frac{1}{n}\sum_i (m(X_i,\lambda) - X_i)^2 + penalty - 1$$

$$penalty = \begin{cases} Ridge: & \frac{2}{1+\lambda} \\ Lasso: & 2P_n(|X| > \lambda) \\ Pre\text{-}test: & 2P_n(|X| > \lambda) + 2\lambda \cdot (\widehat{f}(-\lambda) + \widehat{f}(\lambda)) \end{cases}$$

2. Cross validation (CV)
   Requires multiple observations X_{ij} for μ_i.

$$\widehat{R}(\lambda) = \frac{1}{kn}\sum_{i=1}^{n}\sum_{j=1}^{k}(m(\overline{X}_{i,-j},\lambda) - X_{ij})^2$$

$$\overline{X}_{i,-j} = leave\text{-}one\text{-}out\text{-}mean.$$

Comparison with Leeb and Pötscher (2006)

- **Leeb and Pötscher (2006):** We observe a $(k \times 1)$ vector

$$\boldsymbol{X}_n \sim N(\mu_n, \boldsymbol{I}_k/n)$$

  and aim to estimate the normalized risk $nE|\|\boldsymbol{m}_n(\boldsymbol{X}_n) - \mu_n\|^2$.

  Reparameterize, $\boldsymbol{Y}_n = \sqrt{n}\boldsymbol{X}_n$ and consider $\mu_n = \boldsymbol{h}/\sqrt{n}$, then

$$\boldsymbol{Y}_n \sim N(\boldsymbol{h}, \boldsymbol{I}_k)$$

  and the problem is invariant in n.

- **This article:**

$$(X_i, \mu_i) \sim \pi$$

  where π may change with n.

  As n increases we learn risk.

The NPEB estimator of Koenker and Mizera (2014)

- Nonparametric Maximum Likelihood:
$$\max_{G \in \mathscr{G}} \sum_{i=1}^{n} \log\left(\int \varphi(X_i - \mu)\, dG(\mu)\right),$$
where $\mathscr{G}$ is the family of all distribution functions.
- The solution, $\widehat{G}$, is given by a discrete distribution supported at m points $v_1, \ldots, v_m$ with frequencies $f_1, \ldots, f_m$ (with $m \leq n$).
- Then, construct an estimator of
$$m^*_\pi(x) = E_\pi[\mu | X = x]$$
by plugin-in $\widehat{G}$ for G in the formula for $E_\pi[\mu | X = x]$:
$$\widehat{m}^*_\pi(x) = \sum_{j=1}^{m} v_j \varphi(x - v_j) f_j \Bigg/ \sum_{j=1}^{m} \varphi(x - v_j) f_j.$$

Uniform loss consistency

- Assume

$$\sup_{\pi \in \mathcal{Q}} P_\pi \left(\sup_{\lambda \in [0,\infty]} \left| L_n(\lambda) - \bar{R}_\pi(\lambda) \right| > \varepsilon \right) \to 0, \quad \forall \varepsilon > 0. \quad (1)$$

- Assume there are functions, $\bar{r}_\pi(\lambda)$, $\bar{v}_\pi$, and $r_n(\lambda)$ (of (π, λ), π, and $(\{X_i\}_{i=1}^n, \lambda)$, respectively) such that $\bar{R}_\pi(\lambda) = \bar{r}_\pi(\lambda) + \bar{v}_\pi$, and

$$\sup_{\pi \in \mathcal{Q}} P_\pi \left(\sup_{\lambda \in [0,\infty]} \left| r_n(\lambda) - \bar{r}_\pi(\lambda) \right| > \varepsilon \right) \to 0, \quad \forall \varepsilon > 0. \quad (2)$$

- **Theorem:** Under these assumptions,

$$\sup_{\pi \in \mathcal{Q}} P_\pi \left(\left| L_n(\widehat{\lambda}_n) - \inf_{\lambda \in [0,\infty]} L_n(\lambda) \right| > \varepsilon \right) \to 0, \quad \forall \varepsilon > 0, \quad (3)$$

where $\widehat{\lambda}_n = \operatorname{argmin}_{\lambda \in [0,\infty]} r_n(\lambda)$.

Uniform loss consistency

- We prove that equation (1) holds for ridge, lasso, and pretest, under mild regularity conditions, in particular
$$\sup_{\pi \in \mathscr{Q}} E_\pi[X^4] < \infty.$$
- To satisfy equation (2) we use two popular estimators of risk:
  - SURE: Requires Normality of $X_i|\mu_i$.
  - CV: Requires repeated observations of $X_i|\mu_i$.
- Uniform risk consistency holds also under the same conditions.