The risk of machine learning

Alberto Abadie Maximilian Kasy

July 27, 2017

Two key features of machine learning procedures

1. **Regularization** / shrinkage:
   Improve prediction or estimation performance
   by trading off variance and bias ("avoiding overfitting")

2. Data-dependent choice of **tuning** parameters:
   Try to do trade-off optimally

Large number of methods:

- Regularization:
  - Ridge,
  - Lasso,
  - Pre-testing,
  - Trees, Neural Networks, Support Vector Machines, ...
- Tuning:
  - Cross-validation (CV),
  - Stein's Unbiased Risk Estimate (SURE),
  - Empirical Bayes (marginal likelihood), ...

Questions facing the empirical researcher

1. When should we bother with regularization?
2. What kind of regularization should we choose?
   What features of the data generating process matter for this choice?
3. When do CV or SURE work for tuning?

Roadmap

- Our answers to these questions
- A stylized setting: Estimation of many means
- Backing up our answers, using
  1. Theory
  2. Empirical examples
  3. Simulations
- Conclusion and summary of our formal contributions

1) When should we bother with regularization?

Answer: Can reduce risk (mean squared error)

- when there are many parameters,
- and we care about point-estimates for each of these.

Examples:

- Causal / predictive effects, **many treatment** values:
  - Location effects: Chetty and Hendren (2015)
  - Teacher effects: Chetty et al. (2014)
  - Worker and firm effects: Abowd et al. (1999), Card et al. (2012)
  - Judge effects: Abrams et al. (2012)
- Binary treatment, **many subgroups**:
  - Class size effect for different demographic groups: Krueger (1999)
  - Event studies: DellaVigna and La Ferrara (2010)
    (many treatments and many treated units)
- Prediction with **many predictive covariates** / transformations of covariates:
  - Macro forecasting: Stock and Watson (2012)
  - Series regression: Newey (1997)

2) What kind of regularization should we choose?

Answer: Depends on the setting / distribution of parameters.

- Parameters smoothly distributed, no true zeros:
  - **Ridge** / linear shrinkage
  - e.g.: location effects
    (Chetty and Hendren, 2015)
  - Arguably most common case in econ settings.
- Many true zeros, non-zeros well separated:
  - **Pre-testing** / hard thresholding
  - e.g.: large fixed costs for non-zero behavior
    (DellaVigna and La Ferrara, 2010)
  - Rare!
- Many true zeros, non-zeros not well separated (intermediate case):
  - **Lasso** / soft thresholding
  - Robust choice for many settings.

3) When do CV or SURE work for tuning?

Answer:

- CV and SURE are uniformly close to the oracle-optimum

  – in high-dimensional settings.

- Intuition:
  - Optimal tuning depends on the distribution of parameters.
  - When there are many parameters, we can learn this distribution.

- Requirements:
  1. SURE: normality
  2. CV: number of observations $\gg$ number of parameters

Stylized setting: Estimation of many means

- We observe n independent random variables $X_1, \ldots, X_n$, where

$$E[X_i] = \mu_i,$$
$$\mathrm{Var}(X_i) = \sigma_i^2.$$

- Componentwise estimators:

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

- Squared error loss:

$$L(\widehat{\mu}, \mu) = \tfrac{1}{n} \sum_i (\widehat{\mu}_i - \mu_i)^2$$

- Many applications: X_i equal to OLS estimated coefficients

▸ Connection to linear regression and prediction

Example

Chetty, R. and Hendren, N. (2015). The impacts of neighborhoods on intergenerational mobility: Childhood exposure effects and county-level estimates.
Working Paper, Harvard University and NBER.

- μ_i: causal effect of growing up in commuting zone i
- X_i: unbiased but noisy estimate of μ_i identified from sibling differences of families moving between locations

Componentwise estimators

- Ridge:

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

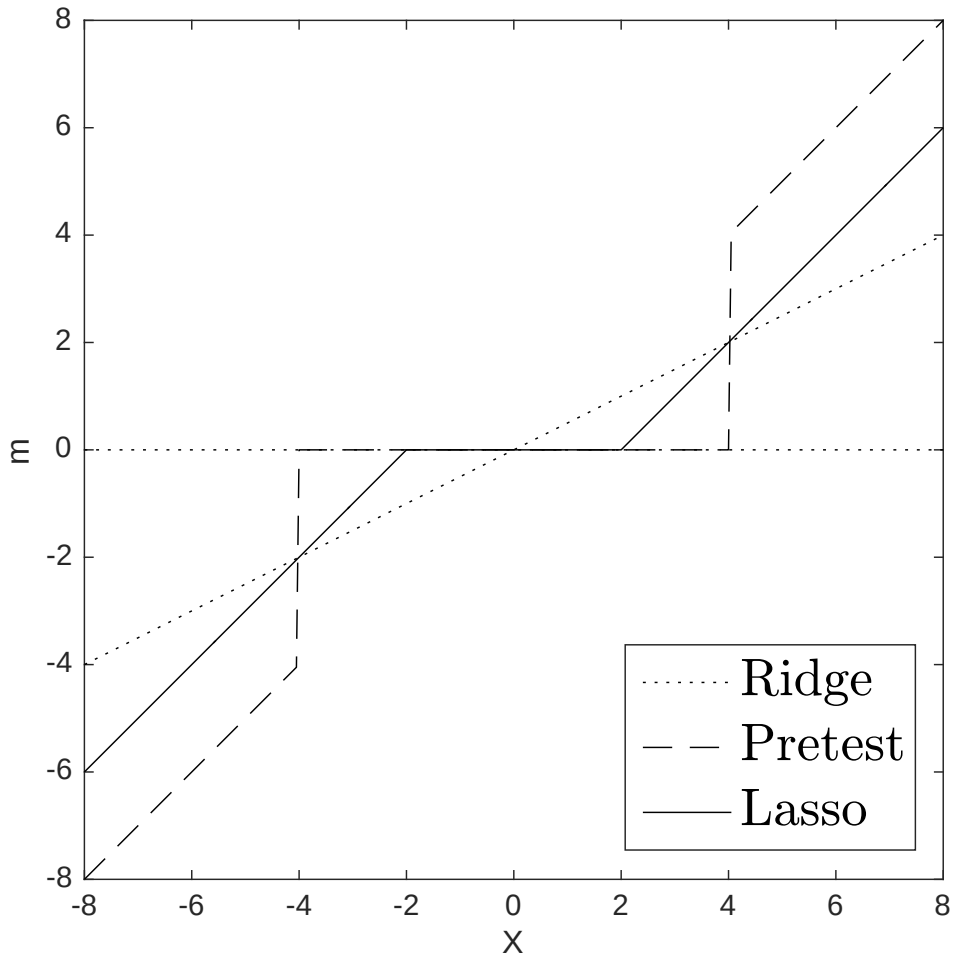
- Lasso:

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

- Pre-test:

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

Componentwise estimators



(Ridge: $\lambda = 1$, Pretest: $\lambda = 4$, Lasso: $\lambda = 2$)

Risk

- Risk = expected loss = mean squared error
- Conceptual subtlety:
  Think of μ_i (more generally: P_i) as fixed or random?
- Compound risk: $\mu_1, \ldots, \mu_n$ as **fixed effects**,
  average over their sample distribution:

$$R_n(m(\cdot, \lambda), \boldsymbol{P}) = \frac{1}{n} \sum_{i=1}^{n} E[(m(X_i, \lambda) - \mu_i)^2 | P_i]$$

- Empirical Bayes risk: $\mu_1, \ldots, \mu_n$ as **random effects**,
  average over their population distribution:

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

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

Characterizing mean squared error

- Random effects setting: Joint distribution $(X, \mu) \sim \pi$
- 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 $\bar{m}^*_\pi(\cdot)$.
- Fixed effects / compound risk: Completely analogous, with empirical distribution $\mu_1, \ldots, \mu_n$ instead of π.

A useful family of examples: Spike and normal DGP

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

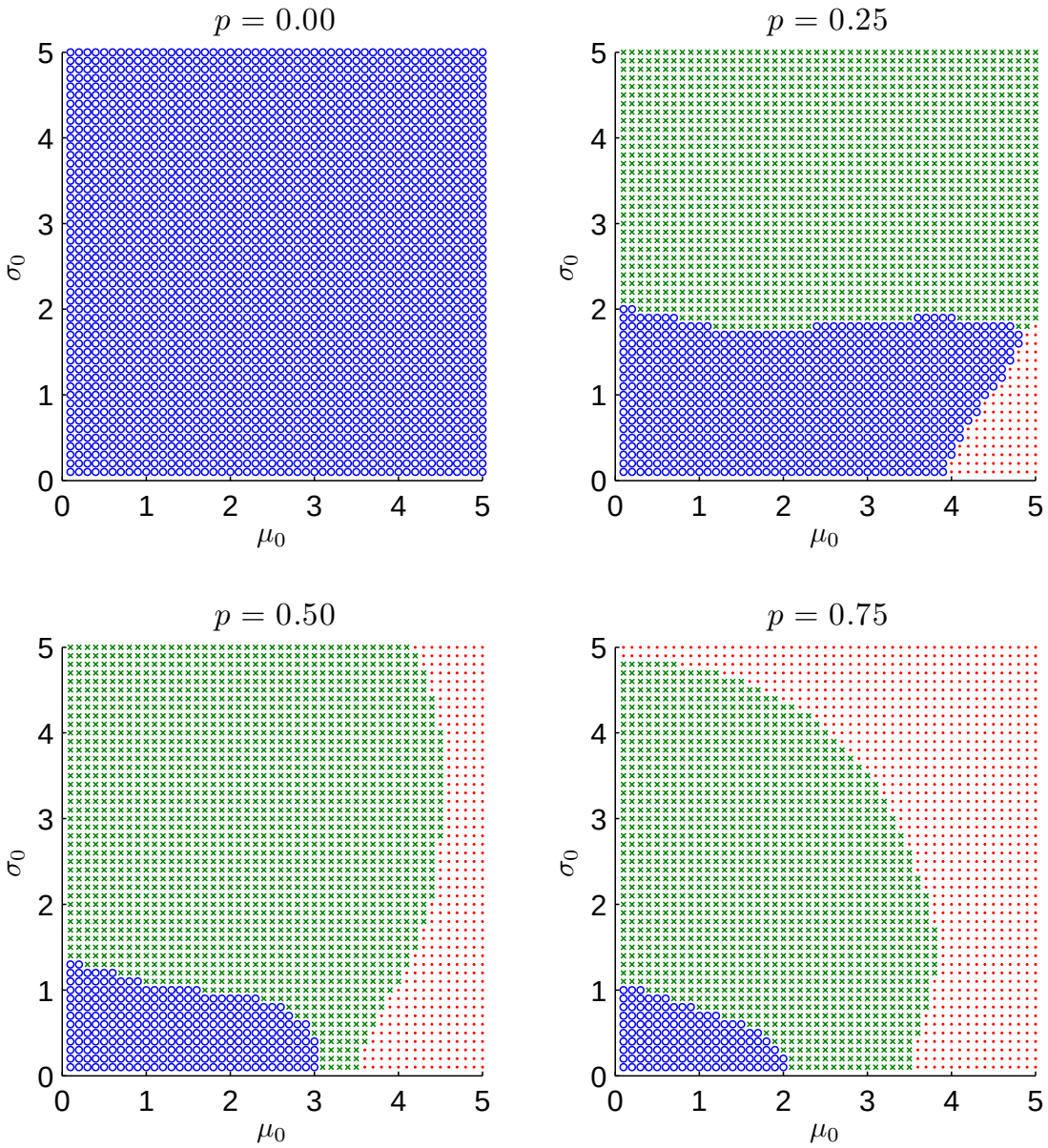
$$\begin{array}{ll} \text{Fraction } p & \mu_i = 0 \\ \text{Fraction } 1-p & \mu_i \sim N(\mu_0, \sigma_0^2) \end{array}$$

- Covers many interesting settings:
  - $p = 0$: smooth distribution of true parameters
  - $p \gg 0$, μ_0 and σ_0^2 large: sparsity, non-zeros well separated

Comparing risk

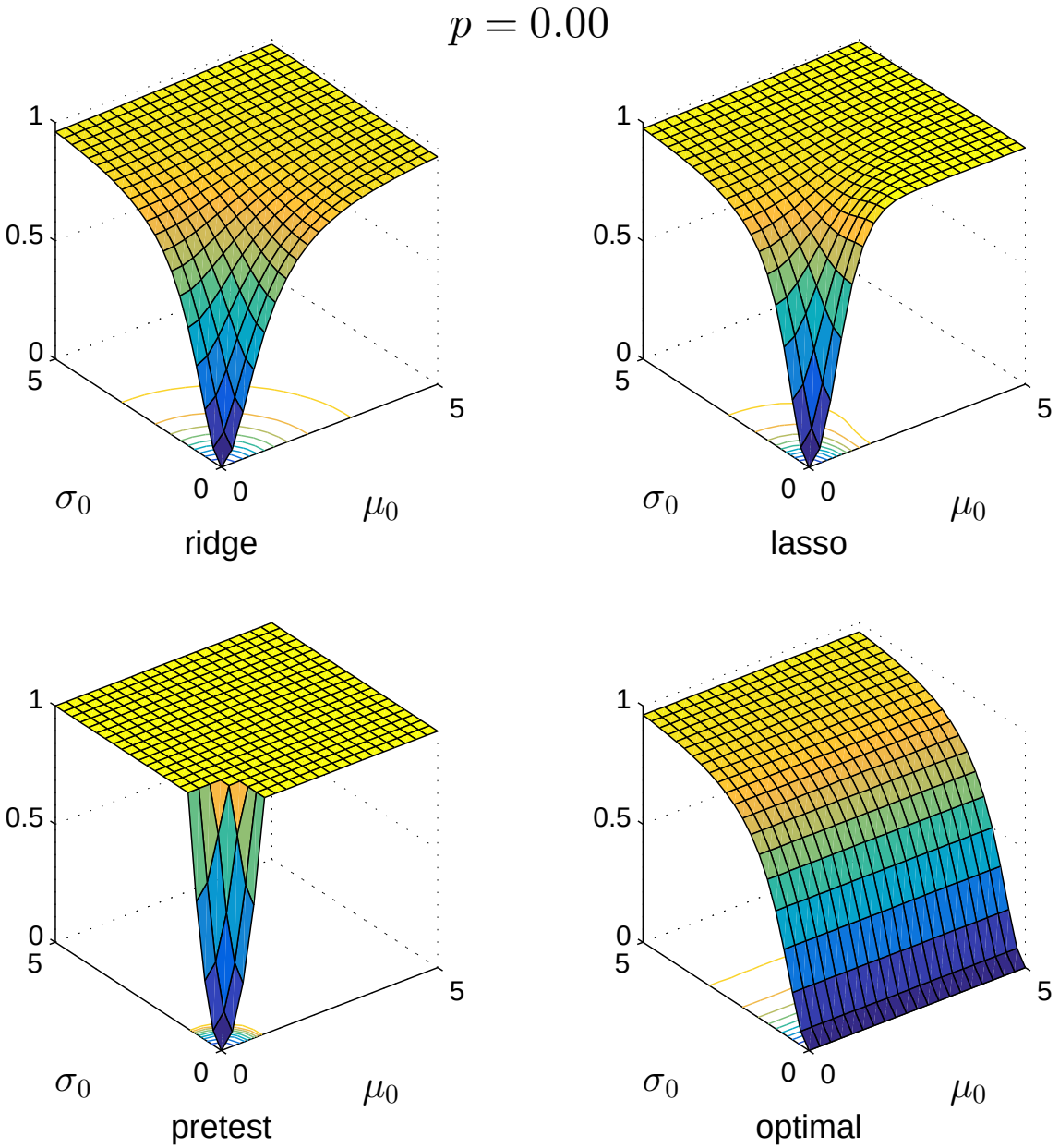
- Consider ridge, lasso, pre-test, optimal shrinkage function.
- Assume λ is chosen optimally (will return to that).
- Can calculate, compare, and plot mean squared error:
  - By construction smaller than 1
    ($1 =$ risk of unregularized estimator, $\lambda = 0$),
  - larger than risk for optimal shrinkage $\bar{m}^*_\pi(\cdot)$
    (by previous theorem).
- Paper: Analytic risk functions $\bar{R}$.

Best estimator

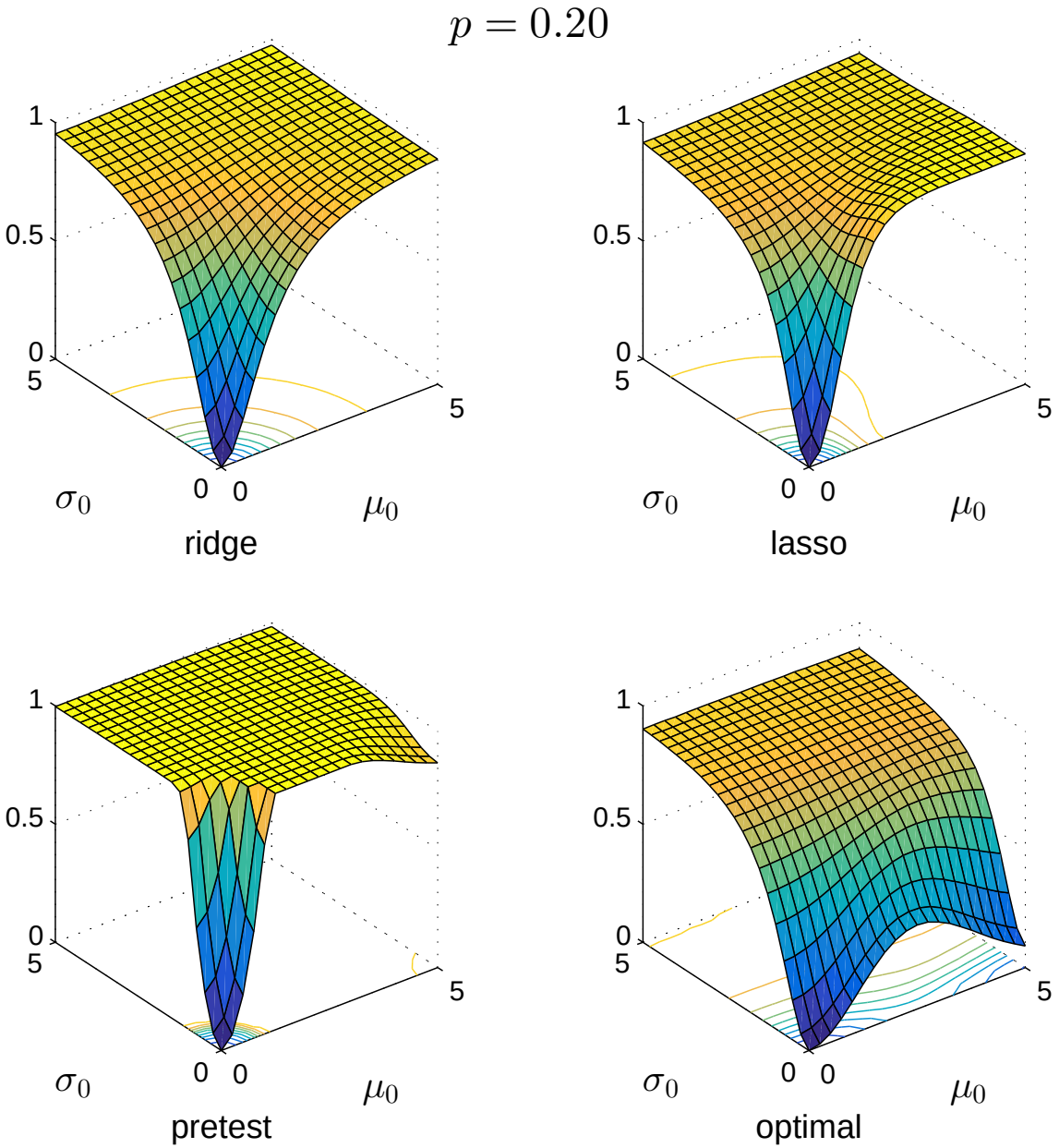


○ is ridge, **x** is lasso, · is pretest

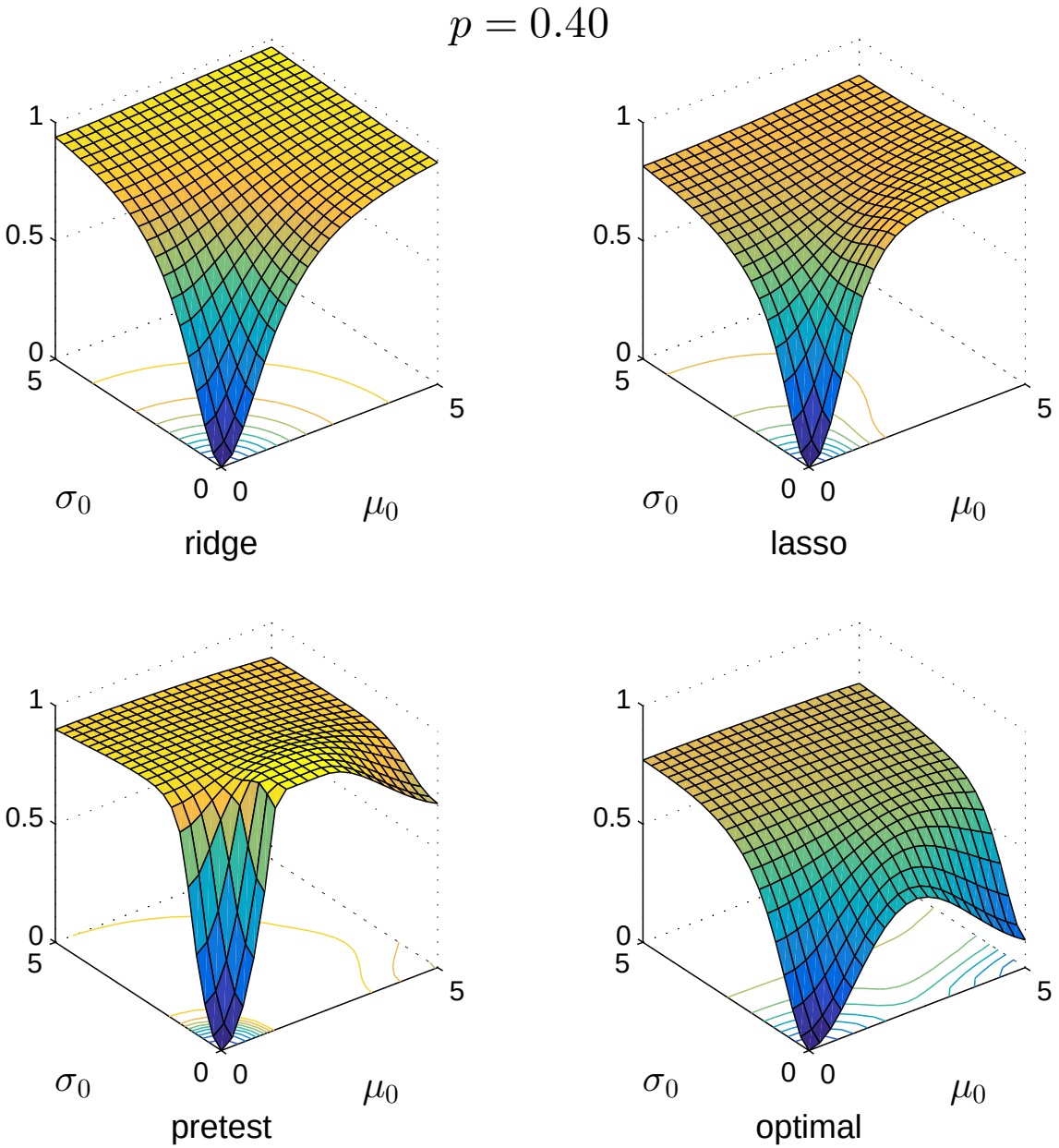
Mean squared error



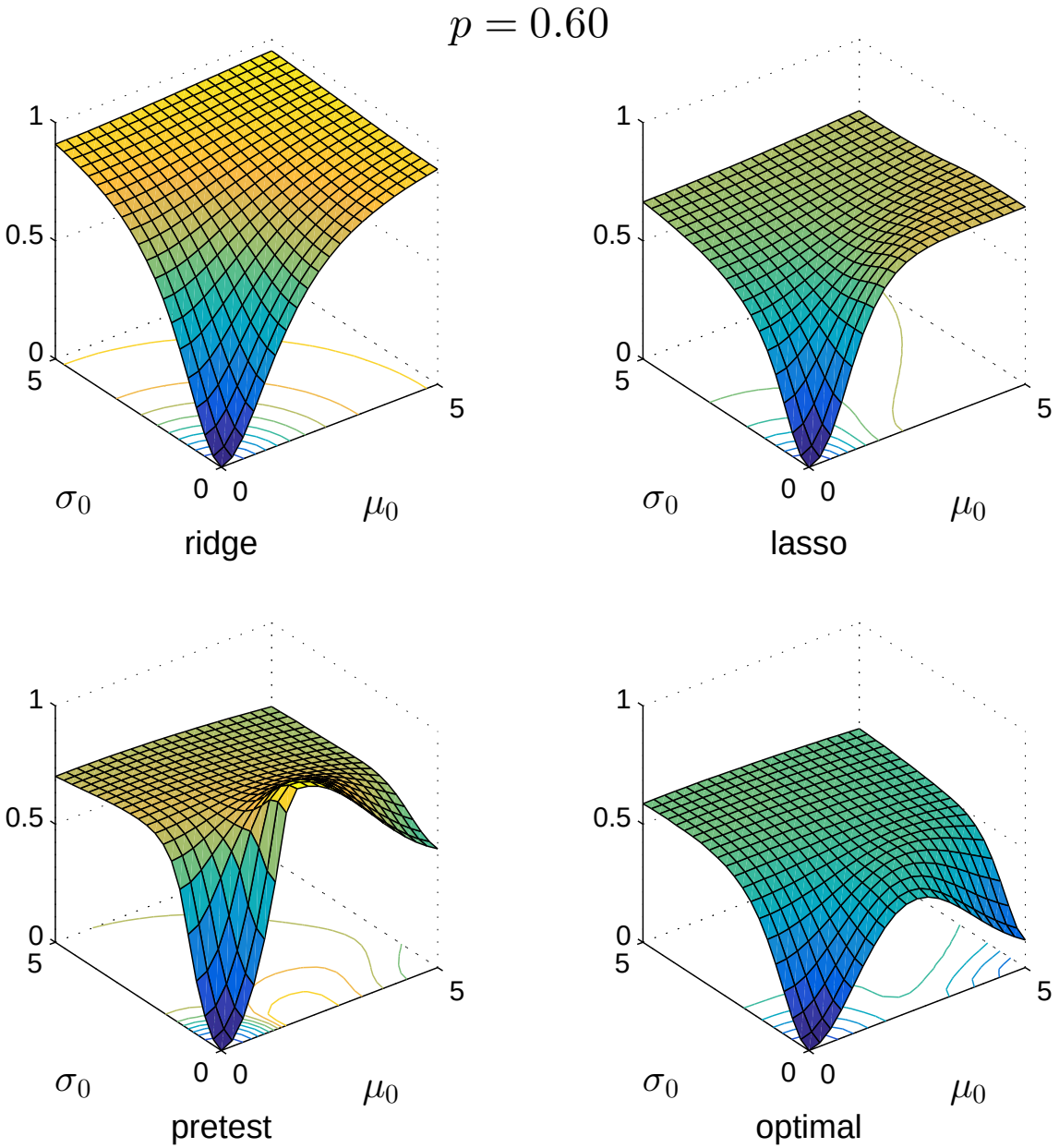
Mean squared error



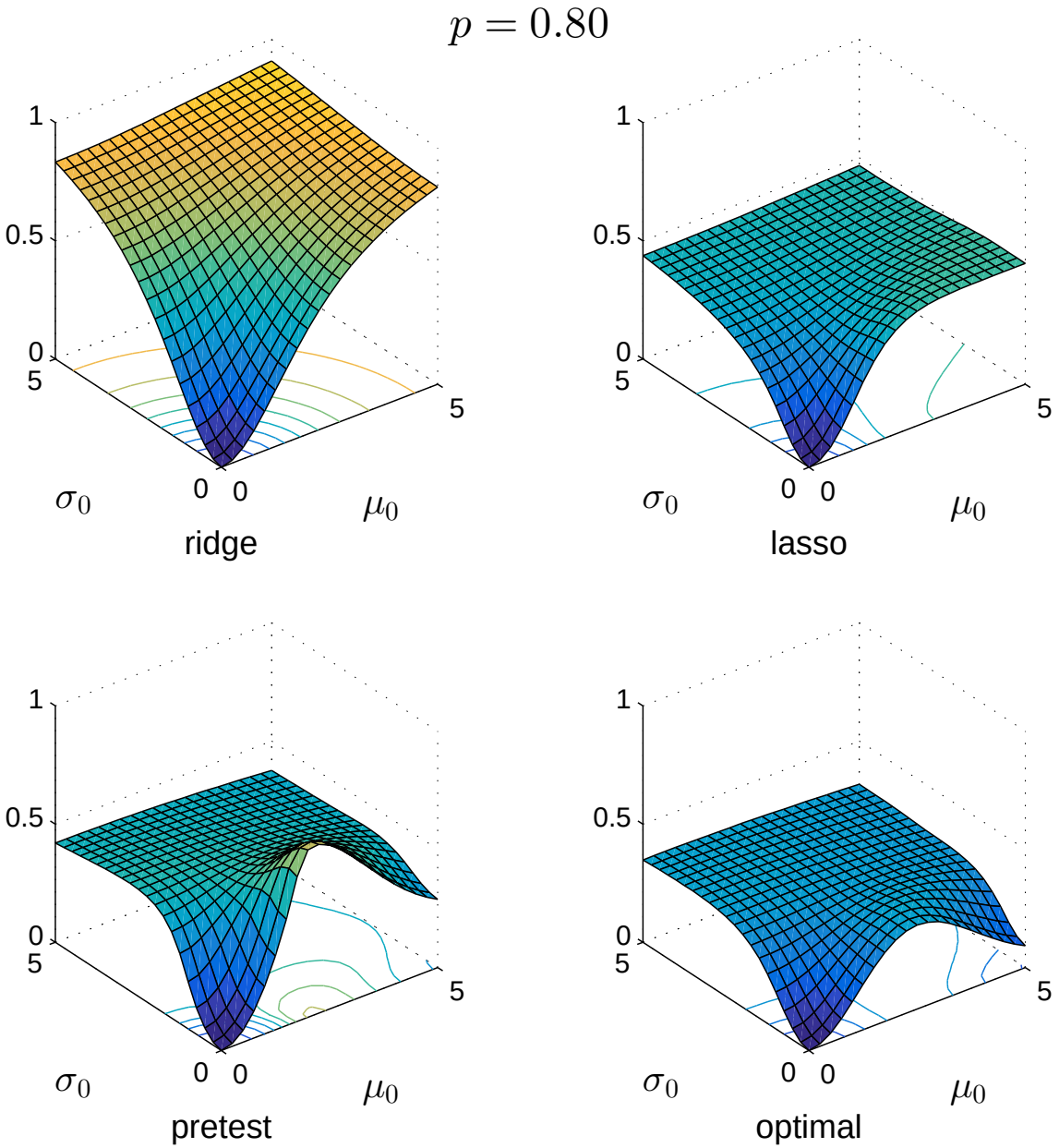
Mean squared error



Mean squared error



Mean squared error



Estimating λ

- So far: Benchmark of optimal (oracle) λ^*.
- Can we consistently estimate 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:

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

- **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 $P_i \sim^{iid} \pi$.

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 $L_n(\lambda)$, $\widehat{R}(\lambda)$.

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.

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

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

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)

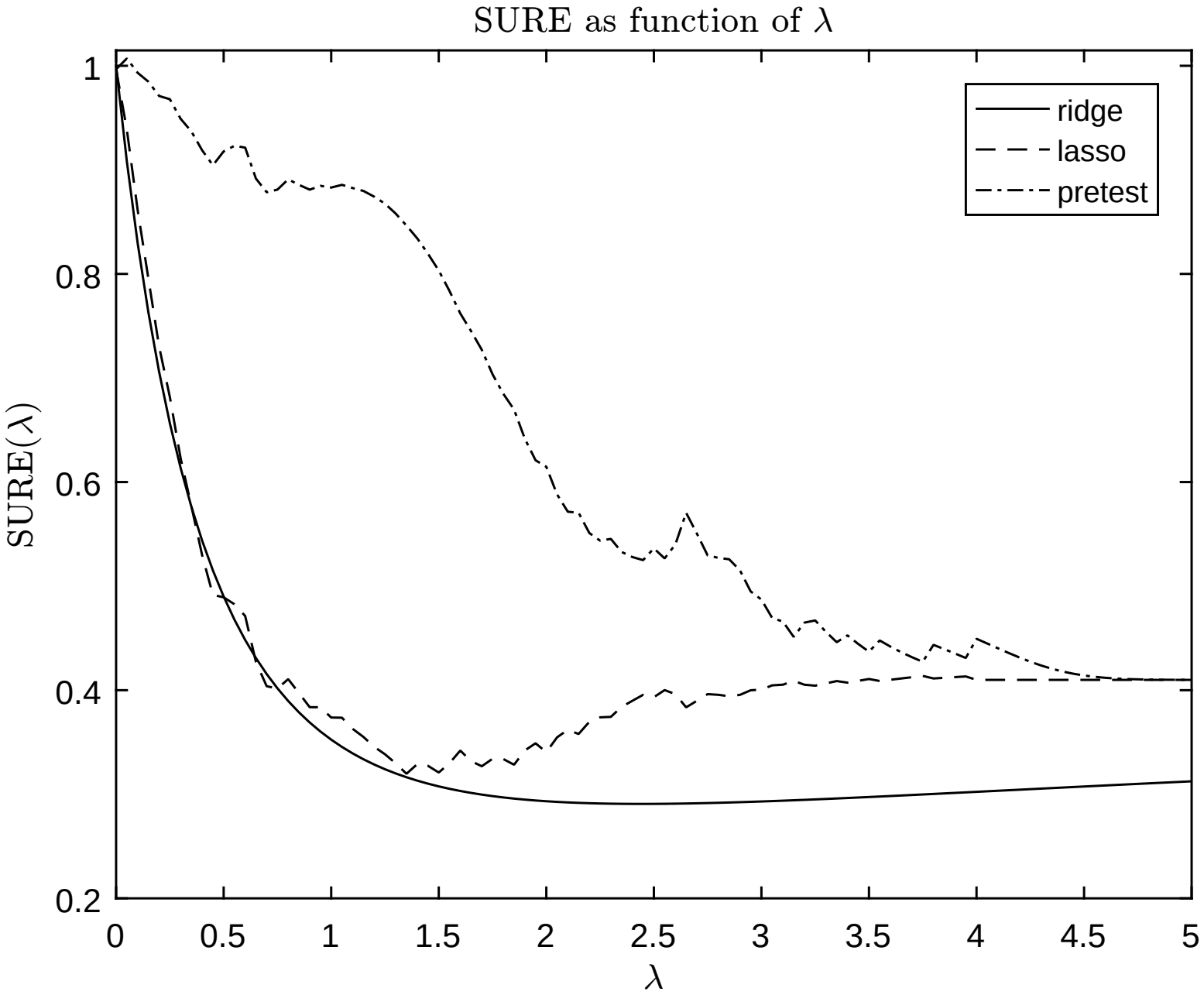
- **Nonparametric Mincer equation:**
  A nonparametric regression equation of log wages on education and potential experience
  (Belloni and Chernozhukov, 2011)

Estimated Risk

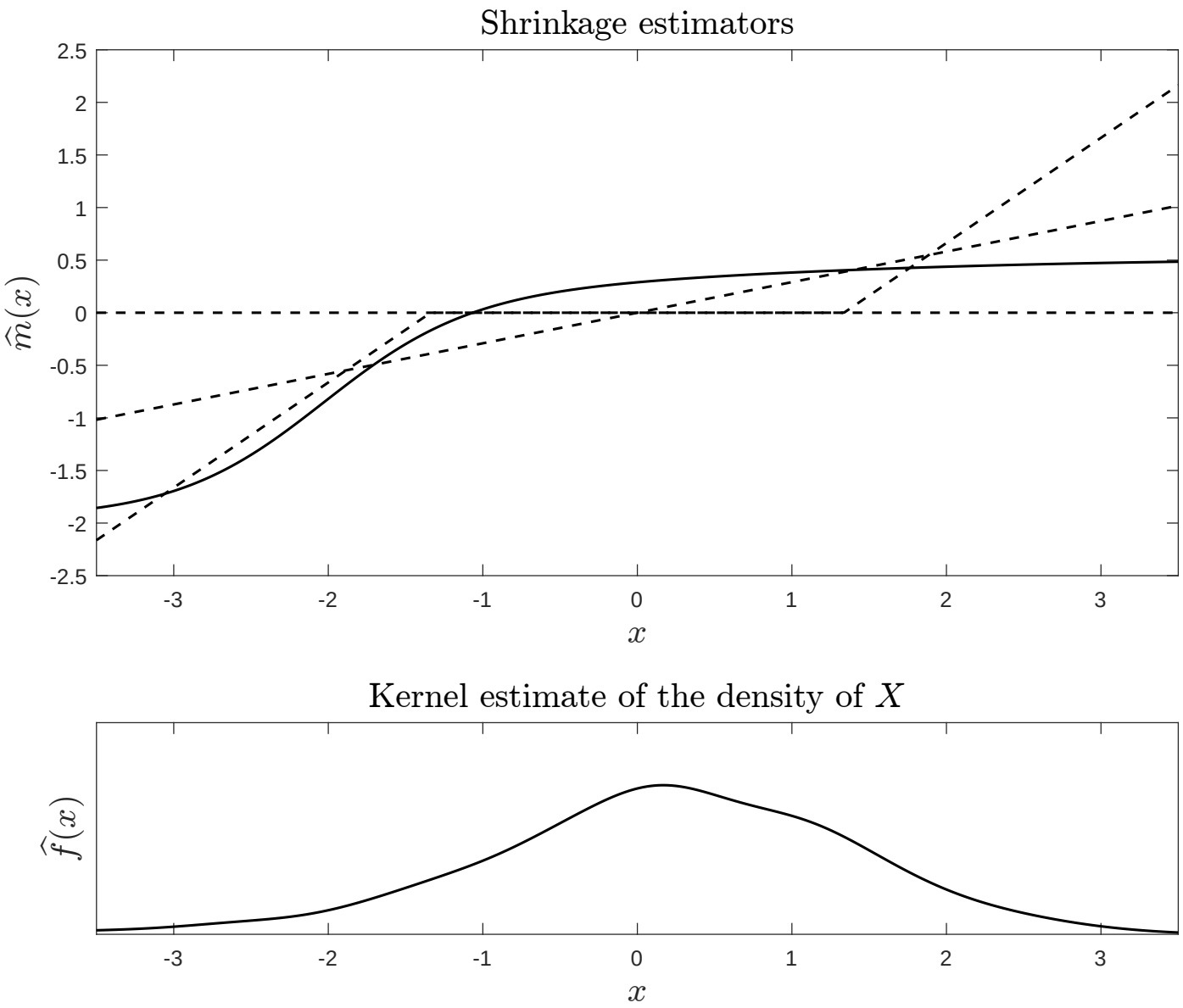
- Stein's unbiased risk estimate $\widehat{R}$
- at the optimized tuning parameter $\widehat{\lambda}^*$
- for each application and estimator considered.

	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

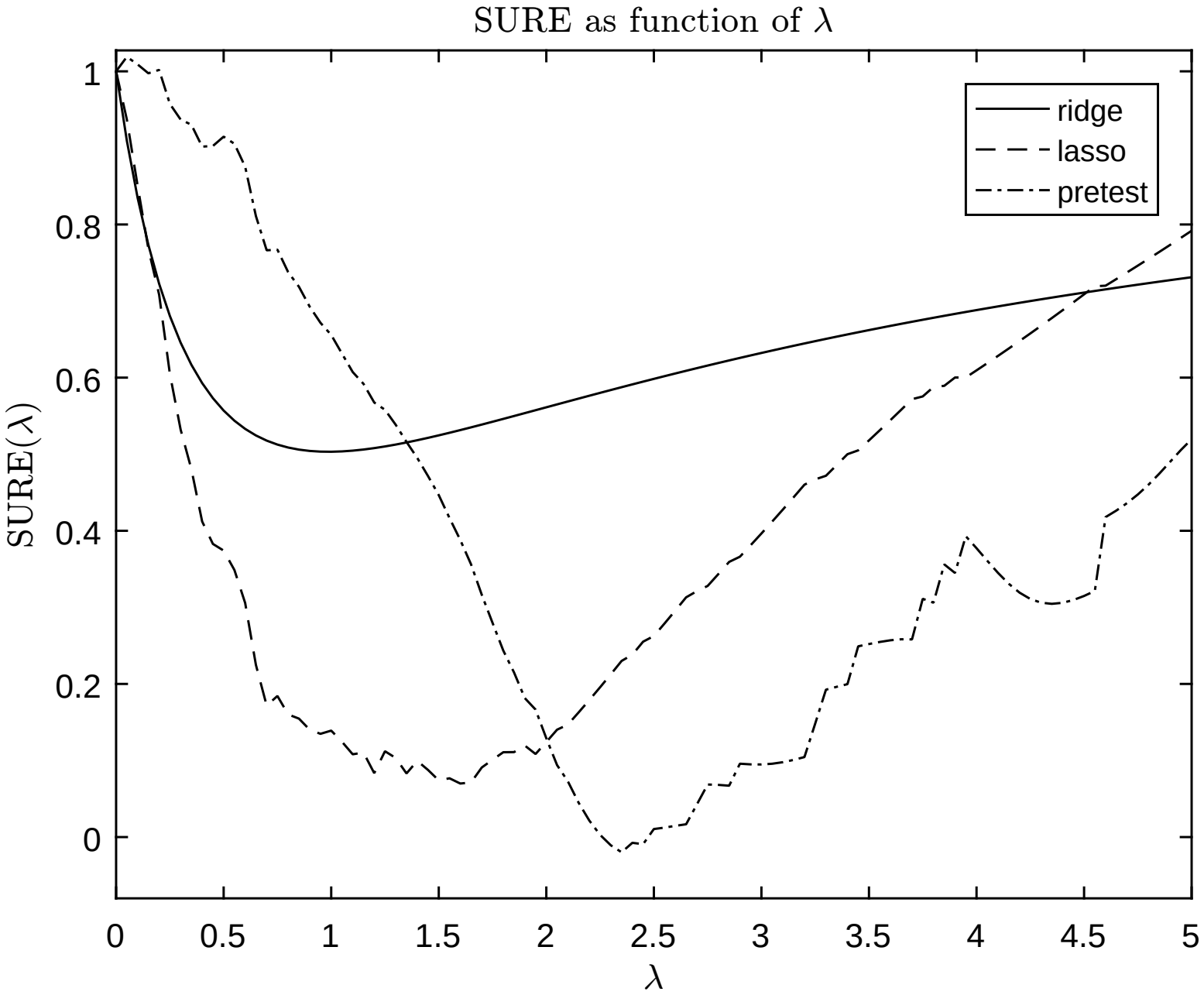


Neighborhood effects: shrinkage estimators

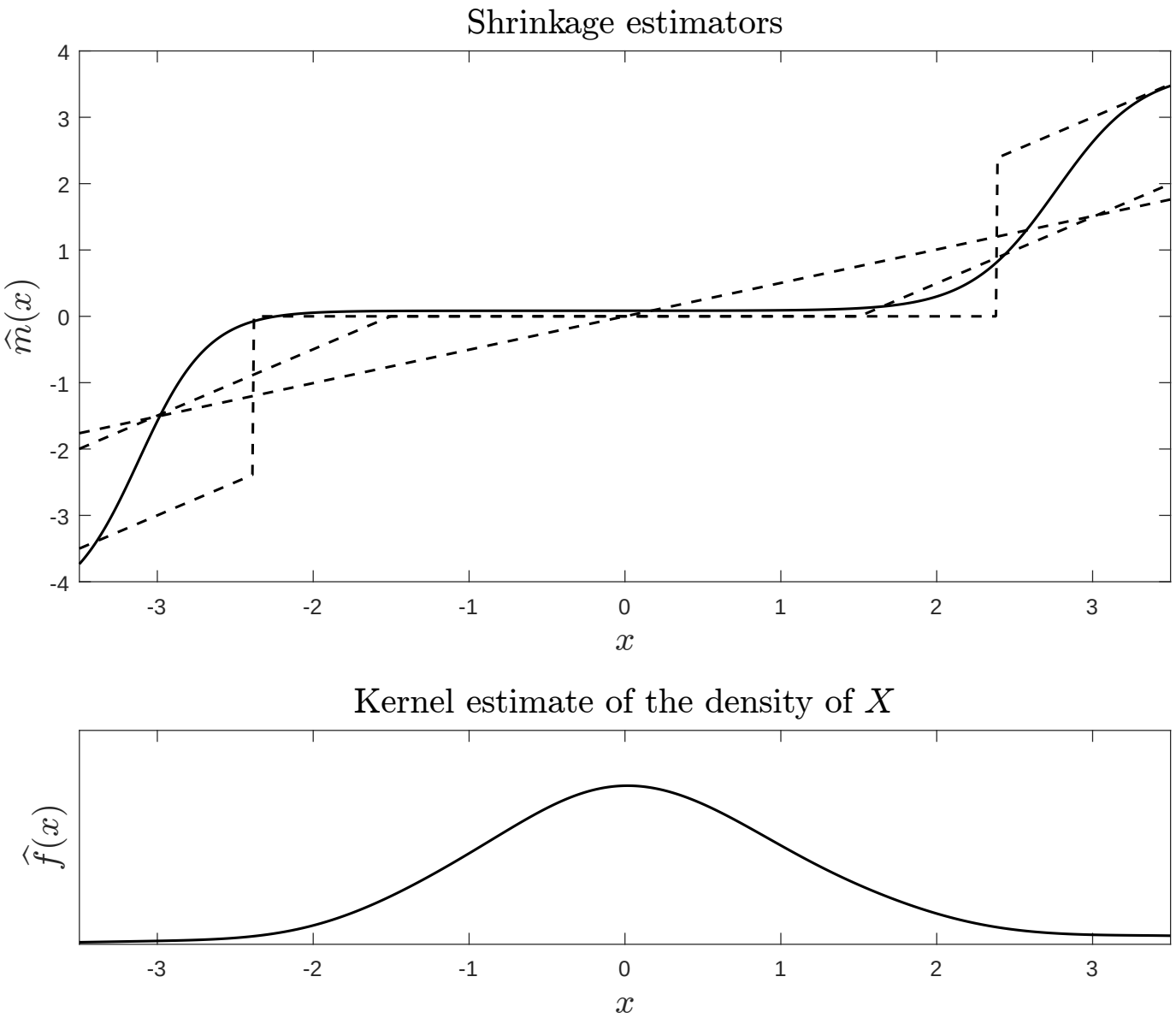


Solid line in top figure is an estimate of $\bar{m}^*_\pi(x)$

Arms event study: SURE estimates

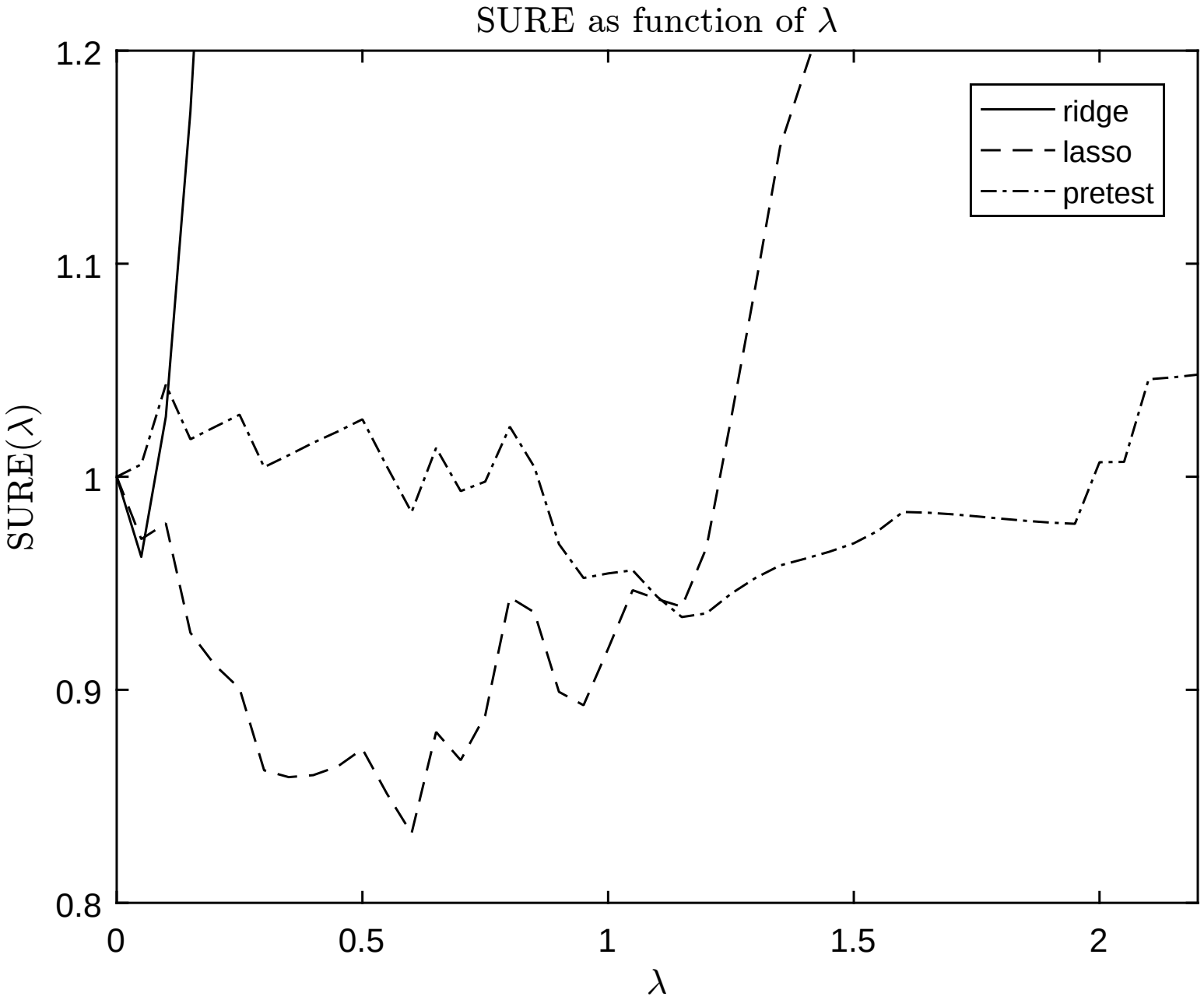


Arms event study: shrinkage estimators

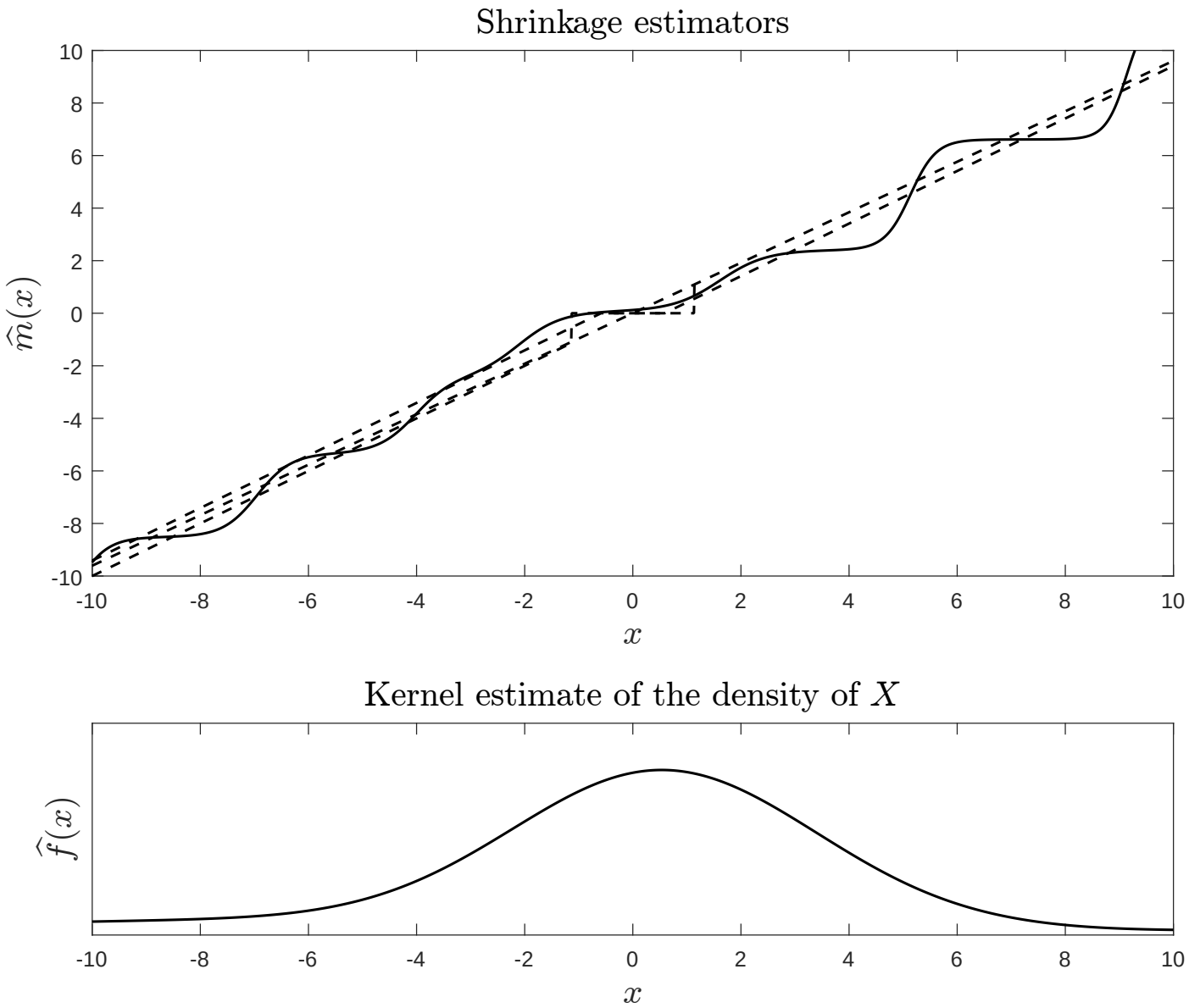


Solid line in top figure is an estimate of $\bar{m}^*_\pi(x)$

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.
- Simulation results

Summary and Conclusion

- We study the risk properties of machine learning estimators in the context of the problem of estimating many means μ_i based on observations X_i
- We provide a simple characterization of the risk of machine learning estimators based on proximity to the optimal shrinkage function
- We use a spike-and-Normal setting to investigate how simple features of the DGP affect the relative performance of the different estimators
- We obtain uniform loss consistency results under SURE and CV based choices of regularization parameters
- We use data from recent empirical studies to demonstrate the practical applicability of our findings

Recommendations for empirical work

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.

Thank you!

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

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

- 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

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

Back

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

Back