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:
   - Ridge: Smoothly distributed true effects, no special role of zero
   - Pre-testing: Many zeros, non-zeros well separated
   - 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: $\hat{\mu}_i = m(X_i, \lambda)$, where $m: \mathbb{R} \times [0, \infty] \mapsto \mathbb{R}$ and $\lambda$ may depend on $(X_1, \ldots, X_n)$.
- Examples: Ridge, Lasso, Pretest.
Loss and risk

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

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

- Conditional expectation:

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

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

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

  \( \Rightarrow \) **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 $\mu_i$ across $i$:
  
  Fraction $p$ \hspace{1cm} $\mu_i = 0$
  
  Fraction $1 - p$ \hspace{1cm} $\mu_i \sim N(\mu_0, \sigma_0^2)$

- Covers many interesting settings:
  
  - $p = 0$: smooth distribution of true parameters
  - $p \gg 0$, $\mu_0$ or $\sigma_0^2$ large: sparsity, non-zeros well separated

- Consider ridge, lasso, pre-test, optimal shrinkage function.

- Assume $\lambda$ is chosen optimally (will return to that).
Best estimator

○ is ridge, ⋅ is lasso, ◯ is pretest
Mean squared error

$p = 0.00$
Mean squared error

$p = 0.20$
Mean squared error

\[ p = 0.40 \]
Mean squared error

$p = 0.60$
Mean squared error

$p = 0.80$
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 $\hat{R}$
- at the optimized tuning parameter $\hat{\lambda}^*$
- for each application and estimator considered.

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>Ridge</th>
<th>Lasso</th>
<th>Pre-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>location effects</td>
<td>595</td>
<td>$\hat{R}$</td>
<td>0.29</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\hat{\lambda}^*$</td>
<td>2.44</td>
<td>1.34</td>
</tr>
<tr>
<td>arms trade</td>
<td>214</td>
<td>$\hat{R}$</td>
<td>0.50</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\hat{\lambda}^*$</td>
<td>0.98</td>
<td>1.50</td>
</tr>
<tr>
<td>returns to education</td>
<td>65</td>
<td>$\hat{R}$</td>
<td>1.00</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\hat{\lambda}^*$</td>
<td>0.01</td>
<td>0.59</td>
</tr>
</tbody>
</table>
Neighborhood effects: SURE estimates

SURE as function of $\lambda$

- **SURE as function of 6**
- **SURE function of 6**
- **ridge**
- **lasso**
- **pretest**
Neighborhood effects: shrinkage estimators

Shrinkage estimators

Kernel estimate of the density of $X$

Solid line in top figure is an estimate of $\tilde{m}_\pi(x)$
Arms event study: SURE estimates

SURE as function of $\lambda$

- **SURE** function of 6
- **ridge**
- **lasso**
- **pretest**
Arms event study: shrinkage estimators

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

Kernel estimate of the density of $X$
Mincer regression: SURE estimates

SURE as function of $\lambda$

- **SURE as function of 6**
  - **ridge**
  - **lasso**
  - **pretest**

$\lambda$

SURE($\lambda$)
Mincer regression: shrinkage estimators

\[ \hat{m}(x) \]

Shrinkage estimators

\[ f(x) \]

Kernel estimate of the density of \( X \)

Solid line in top figure is an estimate of \( \hat{m}_\pi^*(x) \)
Monte Carlo simulations

- Spike and normal DGP
- Number of parameters $n = 50, 200, 1000$
- $\lambda$ 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 $m_\pi^*$. 
**Table: Average Compound Loss Across 1000 Simulations with $N = 50$**

<table>
<thead>
<tr>
<th>$p$</th>
<th>$\mu_0$</th>
<th>$\sigma_0$</th>
<th>SURE</th>
<th>Cross-Validation ($k = 4$)</th>
<th>Cross-Validation ($k = 20$)</th>
<th>NPEB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>ridge</td>
<td>lasso</td>
<td>pretest</td>
<td>ridge</td>
</tr>
<tr>
<td>0.00</td>
<td>0</td>
<td>2</td>
<td>0.80</td>
<td>0.89</td>
<td>1.02</td>
<td>0.83</td>
</tr>
<tr>
<td>0.00</td>
<td>0</td>
<td>6</td>
<td>0.97</td>
<td>0.99</td>
<td>1.01</td>
<td>0.97</td>
</tr>
<tr>
<td>0.00</td>
<td>2</td>
<td>2</td>
<td>0.89</td>
<td>0.96</td>
<td>1.01</td>
<td>0.90</td>
</tr>
<tr>
<td>0.00</td>
<td>2</td>
<td>6</td>
<td>0.97</td>
<td>0.99</td>
<td>1.01</td>
<td>0.99</td>
</tr>
<tr>
<td>0.00</td>
<td>4</td>
<td>2</td>
<td>0.95</td>
<td>1.00</td>
<td>1.01</td>
<td>0.95</td>
</tr>
<tr>
<td>0.00</td>
<td>4</td>
<td>6</td>
<td>0.99</td>
<td>1.00</td>
<td>1.02</td>
<td>0.99</td>
</tr>
<tr>
<td>0.50</td>
<td>0</td>
<td>2</td>
<td>0.67</td>
<td>0.64</td>
<td>0.94</td>
<td>0.69</td>
</tr>
<tr>
<td>0.50</td>
<td>0</td>
<td>6</td>
<td>0.95</td>
<td>0.80</td>
<td>0.90</td>
<td>0.95</td>
</tr>
<tr>
<td>0.50</td>
<td>2</td>
<td>2</td>
<td>0.80</td>
<td>0.72</td>
<td>0.96</td>
<td>0.82</td>
</tr>
<tr>
<td>0.50</td>
<td>2</td>
<td>6</td>
<td>0.96</td>
<td>0.80</td>
<td>0.92</td>
<td>0.95</td>
</tr>
<tr>
<td>0.50</td>
<td>4</td>
<td>2</td>
<td>0.91</td>
<td>0.82</td>
<td>0.95</td>
<td>0.92</td>
</tr>
<tr>
<td>0.50</td>
<td>4</td>
<td>6</td>
<td>0.97</td>
<td>0.81</td>
<td>0.93</td>
<td>0.97</td>
</tr>
<tr>
<td>0.95</td>
<td>0</td>
<td>2</td>
<td>0.18</td>
<td>0.15</td>
<td>0.17</td>
<td>0.17</td>
</tr>
<tr>
<td>0.95</td>
<td>0</td>
<td>6</td>
<td>0.49</td>
<td>0.21</td>
<td>0.16</td>
<td>0.51</td>
</tr>
<tr>
<td>0.95</td>
<td>2</td>
<td>2</td>
<td>0.26</td>
<td>0.17</td>
<td>0.18</td>
<td>0.27</td>
</tr>
<tr>
<td>0.95</td>
<td>2</td>
<td>6</td>
<td>0.53</td>
<td>0.21</td>
<td>0.15</td>
<td>0.53</td>
</tr>
<tr>
<td>0.95</td>
<td>4</td>
<td>2</td>
<td>0.44</td>
<td>0.21</td>
<td>0.18</td>
<td>0.45</td>
</tr>
<tr>
<td>0.95</td>
<td>4</td>
<td>6</td>
<td>0.57</td>
<td>0.21</td>
<td>0.15</td>
<td>0.58</td>
</tr>
</tbody>
</table>
Table: Average Compound Loss Across 1000 Simulations with $N = 200$

<table>
<thead>
<tr>
<th>$p$</th>
<th>$\mu_0$</th>
<th>$\sigma_0$</th>
<th>SURE</th>
<th>Cross-Validation ($k = 4$)</th>
<th>Cross-Validation ($k = 20$)</th>
<th>NPEB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>ridge</td>
<td>lasso</td>
<td>pretest</td>
<td>ridge</td>
</tr>
<tr>
<td>0.00</td>
<td>0</td>
<td>2</td>
<td>0.80</td>
<td>0.87</td>
<td>1.01</td>
<td>0.82</td>
</tr>
<tr>
<td>0.00</td>
<td>0</td>
<td>6</td>
<td>0.98</td>
<td>0.99</td>
<td>1.01</td>
<td>0.98</td>
</tr>
<tr>
<td>0.00</td>
<td>2</td>
<td>2</td>
<td>0.89</td>
<td>0.95</td>
<td>1.00</td>
<td>0.90</td>
</tr>
<tr>
<td>0.00</td>
<td>2</td>
<td>6</td>
<td>0.98</td>
<td>1.00</td>
<td>1.01</td>
<td>0.98</td>
</tr>
<tr>
<td>0.00</td>
<td>4</td>
<td>2</td>
<td>0.95</td>
<td>1.00</td>
<td>1.00</td>
<td>0.96</td>
</tr>
<tr>
<td>0.00</td>
<td>4</td>
<td>6</td>
<td>0.98</td>
<td>0.99</td>
<td>1.01</td>
<td>0.98</td>
</tr>
<tr>
<td>0.50</td>
<td>0</td>
<td>2</td>
<td>0.67</td>
<td>0.61</td>
<td>0.90</td>
<td>0.69</td>
</tr>
<tr>
<td>0.50</td>
<td>0</td>
<td>6</td>
<td>0.94</td>
<td>0.77</td>
<td>0.86</td>
<td>0.95</td>
</tr>
<tr>
<td>0.50</td>
<td>2</td>
<td>2</td>
<td>0.80</td>
<td>0.70</td>
<td>0.94</td>
<td>0.82</td>
</tr>
<tr>
<td>0.50</td>
<td>2</td>
<td>6</td>
<td>0.95</td>
<td>0.78</td>
<td>0.88</td>
<td>0.96</td>
</tr>
<tr>
<td>0.50</td>
<td>4</td>
<td>2</td>
<td>0.91</td>
<td>0.80</td>
<td>0.94</td>
<td>0.92</td>
</tr>
<tr>
<td>0.50</td>
<td>4</td>
<td>6</td>
<td>0.96</td>
<td>0.79</td>
<td>0.92</td>
<td>0.97</td>
</tr>
<tr>
<td>0.95</td>
<td>0</td>
<td>2</td>
<td>0.17</td>
<td>0.12</td>
<td>0.14</td>
<td>0.17</td>
</tr>
<tr>
<td>0.95</td>
<td>0</td>
<td>6</td>
<td>0.61</td>
<td>0.18</td>
<td>0.14</td>
<td>0.62</td>
</tr>
<tr>
<td>0.95</td>
<td>2</td>
<td>2</td>
<td>0.28</td>
<td>0.16</td>
<td>0.17</td>
<td>0.29</td>
</tr>
<tr>
<td>0.95</td>
<td>2</td>
<td>6</td>
<td>0.63</td>
<td>0.19</td>
<td>0.14</td>
<td>0.64</td>
</tr>
<tr>
<td>0.95</td>
<td>4</td>
<td>2</td>
<td>0.49</td>
<td>0.20</td>
<td>0.17</td>
<td>0.50</td>
</tr>
<tr>
<td>0.95</td>
<td>4</td>
<td>6</td>
<td>0.68</td>
<td>0.19</td>
<td>0.13</td>
<td>0.70</td>
</tr>
</tbody>
</table>
Table: Average Compound Loss Across 1000 Simulations with $N = 1000$

<table>
<thead>
<tr>
<th>$p$</th>
<th>$\mu_0$</th>
<th>$\sigma_0$</th>
<th>SURE</th>
<th>Cross-Validation ($k = 4$)</th>
<th>Cross-Validation ($k = 20$)</th>
<th>NPEB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>ridge lasso pretest</td>
<td>ridge lasso pretest</td>
<td>ridge lasso pretest</td>
<td></td>
</tr>
<tr>
<td>0.00</td>
<td>0</td>
<td>2</td>
<td>0.80 0.87 1.01</td>
<td>0.81 0.87 1.01</td>
<td>0.80 0.86 1.01</td>
<td>0.82</td>
</tr>
<tr>
<td>0.00</td>
<td>0</td>
<td>6</td>
<td>0.97 0.98 1.00</td>
<td>0.98 0.98 1.00</td>
<td>0.97 0.98 1.00</td>
<td>1.02</td>
</tr>
<tr>
<td>0.00</td>
<td>2</td>
<td>2</td>
<td>0.89 0.94 1.00</td>
<td>0.90 0.95 1.00</td>
<td>0.89 0.94 1.01</td>
<td>0.82</td>
</tr>
<tr>
<td>0.00</td>
<td>2</td>
<td>6</td>
<td>0.97 0.98 1.00</td>
<td>0.98 0.99 1.00</td>
<td>0.97 0.98 1.01</td>
<td>1.02</td>
</tr>
<tr>
<td>0.00</td>
<td>4</td>
<td>2</td>
<td>0.95 1.00 1.00</td>
<td>0.96 1.00 1.00</td>
<td>0.95 0.99 1.00</td>
<td>0.82</td>
</tr>
<tr>
<td>0.00</td>
<td>4</td>
<td>6</td>
<td>0.98 0.99 1.00</td>
<td>0.98 0.99 1.00</td>
<td>0.98 0.99 1.01</td>
<td>1.02</td>
</tr>
<tr>
<td>0.50</td>
<td>0</td>
<td>2</td>
<td>0.67 0.60 0.87</td>
<td>0.68 0.61 0.90</td>
<td>0.67 0.60 0.87</td>
<td>0.60</td>
</tr>
<tr>
<td>0.50</td>
<td>0</td>
<td>6</td>
<td>0.95 0.77 0.81</td>
<td>0.95 0.77 0.82</td>
<td>0.95 0.76 0.81</td>
<td>0.72</td>
</tr>
<tr>
<td>0.50</td>
<td>2</td>
<td>2</td>
<td>0.80 0.70 0.90</td>
<td>0.81 0.71 0.90</td>
<td>0.80 0.69 0.89</td>
<td>0.62</td>
</tr>
<tr>
<td>0.50</td>
<td>2</td>
<td>6</td>
<td>0.95 0.77 0.80</td>
<td>0.96 0.78 0.81</td>
<td>0.95 0.77 0.80</td>
<td>0.71</td>
</tr>
<tr>
<td>0.50</td>
<td>4</td>
<td>2</td>
<td>0.91 0.80 0.87</td>
<td>0.92 0.80 0.84</td>
<td>0.91 0.80 0.84</td>
<td>0.63</td>
</tr>
<tr>
<td>0.50</td>
<td>4</td>
<td>6</td>
<td>0.96 0.78 0.87</td>
<td>0.97 0.78 0.79</td>
<td>0.96 0.78 0.78</td>
<td>0.70</td>
</tr>
<tr>
<td>0.95</td>
<td>0</td>
<td>2</td>
<td>0.17 0.11 0.14</td>
<td>0.17 0.12 0.14</td>
<td>0.17 0.11 0.14</td>
<td>0.11</td>
</tr>
<tr>
<td>0.95</td>
<td>0</td>
<td>6</td>
<td>0.63 0.18 0.13</td>
<td>0.65 0.18 0.14</td>
<td>0.64 0.17 0.14</td>
<td>0.12</td>
</tr>
<tr>
<td>0.95</td>
<td>2</td>
<td>2</td>
<td>0.28 0.15 0.16</td>
<td>0.29 0.15 0.18</td>
<td>0.29 0.14 0.17</td>
<td>0.12</td>
</tr>
<tr>
<td>0.95</td>
<td>2</td>
<td>6</td>
<td>0.66 0.18 0.13</td>
<td>0.67 0.18 0.14</td>
<td>0.66 0.18 0.13</td>
<td>0.12</td>
</tr>
<tr>
<td>0.95</td>
<td>4</td>
<td>2</td>
<td>0.50 0.19 0.16</td>
<td>0.51 0.19 0.17</td>
<td>0.50 0.19 0.16</td>
<td>0.12</td>
</tr>
<tr>
<td>0.95</td>
<td>4</td>
<td>6</td>
<td>0.72 0.18 0.13</td>
<td>0.73 0.19 0.13</td>
<td>0.71 0.18 0.13</td>
<td>0.12</td>
</tr>
</tbody>
</table>
Can we consistently estimate the optimal $\lambda^*$, 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 \( m(., \hat{\lambda}) \) for \( m(., \bar{\lambda}^*) \):
  \[ \sup_{\pi} P_{\pi} \left( \left| L_n(\hat{\lambda}) - L_n(\bar{\lambda}^*) \right| > \varepsilon \right) \rightarrow 0 \]
  as \( n \rightarrow \infty \) for all \( \varepsilon > 0 \), where
  \[ P_i \sim \text{iid } \pi. \]
Minimizing estimated risk

- Estimate $\hat{\lambda}^*$ by minimizing estimated risk:
  \[
  \hat{\lambda}^* = \arg\min_{\lambda} \hat{R}(\lambda)
  \]

- Different estimators $\hat{R}(\lambda)$ 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 $L_n(\lambda), \hat{R}(\lambda)$. 
Thank you!
Bonus material
Componentwise estimators

Ridge:

\[ m_R(x, \lambda) = \arg\min_{c \in \mathbb{R}} ((x - c)^2 + \lambda c^2) \]
\[ = \frac{1}{1 + \frac{1}{\lambda}} x. \]

Lasso:

\[ m_L(x, \lambda) = \arg\min_{c \in \mathbb{R}} ((x - c)^2 + 2\lambda |c|) \]
\[ = 1(x < -\lambda)(x + \lambda) + 1(x > \lambda)(x - \lambda). \]

Pre-test:

\[ m_{PT}(x, \lambda) = 1(|x| > \lambda)x. \]
Connection to linear regression and prediction

- Normal linear regression model:
  \[ Y \mid \mathbf{W} \sim N(\mathbf{W}'\beta, \sigma^2). \]

- Sample \( \mathbf{W}_1, \ldots, \mathbf{W}_n \). Let \( \Omega = \frac{1}{N} \sum_{j=1}^{N} \mathbf{W}_j \mathbf{W}_j' \).
- Draw new value of covariates from sample for prediction.
- Expected squared prediction error
  \[ \tilde{R} = E \left[ (Y - \mathbf{W}\hat{\beta})^2 \right] = \text{tr} \left( \Omega \cdot E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] \right) + \sigma^2. \]

- Orthogonalize: Let \( \mu = \Omega^{1/2}\beta, \mathbf{X} = \Omega^{1/2}\hat{\beta}^{OLS}, \hat{\mu}_i = m(X_i, \lambda) \).
- Then
  \[ \mathbf{X} \sim N \left( \mu, \frac{\sigma^2}{N} \mathbf{I}_n \right), \]
  and
  \[ \tilde{R} = E \left[ \sum_i (\hat{\mu}_i - \mu_i)^2 \right] + E[\varepsilon^2]. \]
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 \( \mu_0 \) and variance \( \sigma_0^2 \) elsewhere.

- Conditional on \( \mu_i \), \( X_i \) follows a normal distribution with mean \( \mu_i \) and variance \( \sigma^2 \).

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 $X_i$.
   
   $$\hat{R}(\lambda) = \frac{1}{n} \sum_{i} (m(X_i, \lambda) - X_i)^2 + \text{penalty} - 1$$
   
   $\text{penalty} = \begin{cases} 
   \text{Ridge} : & \frac{2}{1+\lambda} \\
   \text{Lasso} : & 2P_n(|X| > \lambda) \\
   \text{Pre-test} : & 2P_n(|X| > \lambda) + 2\lambda \cdot (\hat{f}(-\lambda) + \hat{f}(\lambda)) 
   \end{cases}$

2. Cross validation (CV)
   Requires multiple observations $X_{ij}$ for $\mu_i$.
   
   $$\hat{R}(\lambda) = \frac{1}{kn} \sum_{i=1}^{n} \sum_{j=1}^{k} (m(\overline{X}_{i,-j}, \lambda) - X_{ij})^2$$
   
   $\overline{X}_{i,-j} = \text{leave-one-out-mean}.$
Comparison with Leeb and Pötscher (2006)

- **Leeb and Pötscher (2006):** We observe a \((k \times 1)\) vector
  \[
  X_n \sim N(\mu_n, I_k/n)
  \]
  and aim to estimate the normalized risk \(nE\|m_n(X_n) - \mu_n\|^2\).
  Reparameterize, \(Y_n = \sqrt{n}X_n\) and consider \(\mu_n = h/\sqrt{n}\), then
  \[
  Y_n \sim N(h, I_k)
  \]
  and the problem is invariant in \(n\).

- **This article:**
  \[
  (X_i, \mu_i) \sim \pi
  \]
  where \(\pi\) may change with \(n\).
  As \(n\) increases we learn risk.
The NPEB estimator of Koenker and Mizera (2014)

- Nonparametric Maximum Likelihood:

\[
\max_{G \in \mathcal{G}} \sum_{i=1}^{n} \log \left( \int \phi(X_i - \mu) dG(\mu) \right),
\]

where \(\mathcal{G}\) is the family of all distribution functions.

- The solution, \(\hat{G}\), 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 \(\hat{G}\) for \(G\) in the formula for \(E_\pi[\mu|X = x]\):

\[
\hat{m}^{*}_\pi(x) = \frac{\sum_{j=1}^{m} v_j \phi(x - v_j) f_j}{\sum_{j=1}^{m} \phi(x - v_j) f_j}.
\]
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 $(\pi, \lambda)$, $\pi$, 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,

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

where $\lambda_n = \arg\min_{\lambda \in [0, \infty]} r_n(\lambda)$. 
Uniform loss consistency

- We prove that equation (1) holds for ridge, lasso, and pretest, under mild regularity conditions, in particular
  \[ \sup_{\pi \in \mathcal{P}} E[|X|^4] < \infty. \]
- 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.