IEOR165 Discussion
Week 5

Sheng Liu

University of California, Berkeley

Feb 19, 2016

Outline

About 1st Homework

- For method of moments, understand the difference between

$$\mu_i \quad \hat{\mu}_i \quad \theta_i \quad \hat{\theta}_i$$

- For uniform distribution on (θ_1, θ_2), understand why

$$\hat{\theta}_2 > \hat{\mu}_1 > \hat{\theta}_1$$

- Solution will be posted online

Maximum A Posterior

Revisit Maximum A Posterior (MAP)

$$f(\theta|X) = \frac{f(X|\theta)g(\theta)}{f(X)}$$

$$\Updownarrow$$

$$\text{Posterior} = \frac{\text{likelihood} \cdot \text{prior}}{\text{evidence}}$$

- MLE: maximum likelihood $f(X|\theta)$

$$\hat{\theta}_{ML} = \arg\max\ f(X|\theta) = \arg\max\ \log f(X|\theta)$$

- MAP: maximum posterior $f(\theta|X)$

$$\begin{aligned}\hat{\theta}_{MAP} &= \arg\max\ f(\theta|X) = \arg\max\ \log f(\theta|X) \\ &= \arg\max\ \{\log f(X|\theta) + \log g(\theta)\}\end{aligned}$$

MLE vs MAP

(Based on Avinash Kak, 2014) Let $X_1, \ldots, X_n$ be a random sample. For each i, the value of X_i can be either Clinton or Sanders. We want to estimate the probability p that a democrat will vote Clinton in the primary.

- Given a p, X_i will follow a Bernoulli distribution:

$$f(X_i = \text{Clinton}|p) = p$$

$$f(X_i = \text{Sanders}|p) = 1 - p$$

- What is the MLE?

MLE vs MAP

Now consider the MAP:

- What should be the prior?
  - The prior should be within the interval $[0, 1]$ (common knowledge)
  - Different people can have different beliefs about the prior: where should the prior peak? what should be the variance?
  - Here, we take the Beta prior:

$$p \sim Beta(\alpha, \beta): \quad g(p) = \frac{1}{B(\alpha, \beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

where $B(\alpha, \beta)$ is the beta function. The mode for the Beta distribution is

$$\frac{\alpha - 1}{\alpha + \beta - 2}$$

And the variance is

$$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

- Choose the parameters for the prior as $\alpha = \beta = 5$.

MLE vs MAP

- Derive the MAP.

- If we have a sample of size $n = 100$ with 60 of them saying that they will vote for Sanders. Then what's the difference between the estimates of p using MLE and MAP?

Motivation

Why do we want to impose regularization on OLS?

- Tradeoff between bias and variance: OLS is unbiased but variance may be high
  - $n < p$, when the observation is not enough, OLS may fail
  - Collinearity: when predictors are correlated, the variance of OLS is significantly high
  - Adding regularization will introduce bias but lower the variance
- Model interpretability
  - Adding more predictors is not always good, it increases the complexity of the model and thus makes it harder for us to extract useful information
  - Regularization (shrinkage) will make some coefficients approaching zero and select the most influential coefficients (and corresponding predictors) from the model

Regularization

- Ridge regression: l_2-norm regularization

$$\begin{aligned}\hat{\beta} &= \arg\min \ ||Y - X\beta||_2^2 + \lambda||\beta||_2^2 \\ &= \arg\min \ \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}\beta_j^2\end{aligned}$$

- Lasso: l_1-norm regularization

$$\begin{aligned}\hat{\beta} &= \arg\min \ ||Y - X\beta||_2^2 + \lambda||\beta||_1 \\ &= \arg\min \ \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}|\beta_j|\end{aligned}$$

- Elastic net: combination of l_1-norm and l_2-norm regularization

$$\hat{\beta} = \arg\min \ ||Y - X\beta||_2^2 + \lambda||\beta||_2^2 + \mu||\beta||_1$$

Credit Data Example

(Gareth, et al. 2013) Given a dataset that records

- Balance
- Age
- Cards (Number of credit cards)
- Education (years of education)
- Income
- Limit (credit limit)
- Rating (credit rating)
- Gender
- Student (whether a student or not)
- marital status
- ethnicity

Let Balance be the response and all other variables be predictors.

Credit Data Example

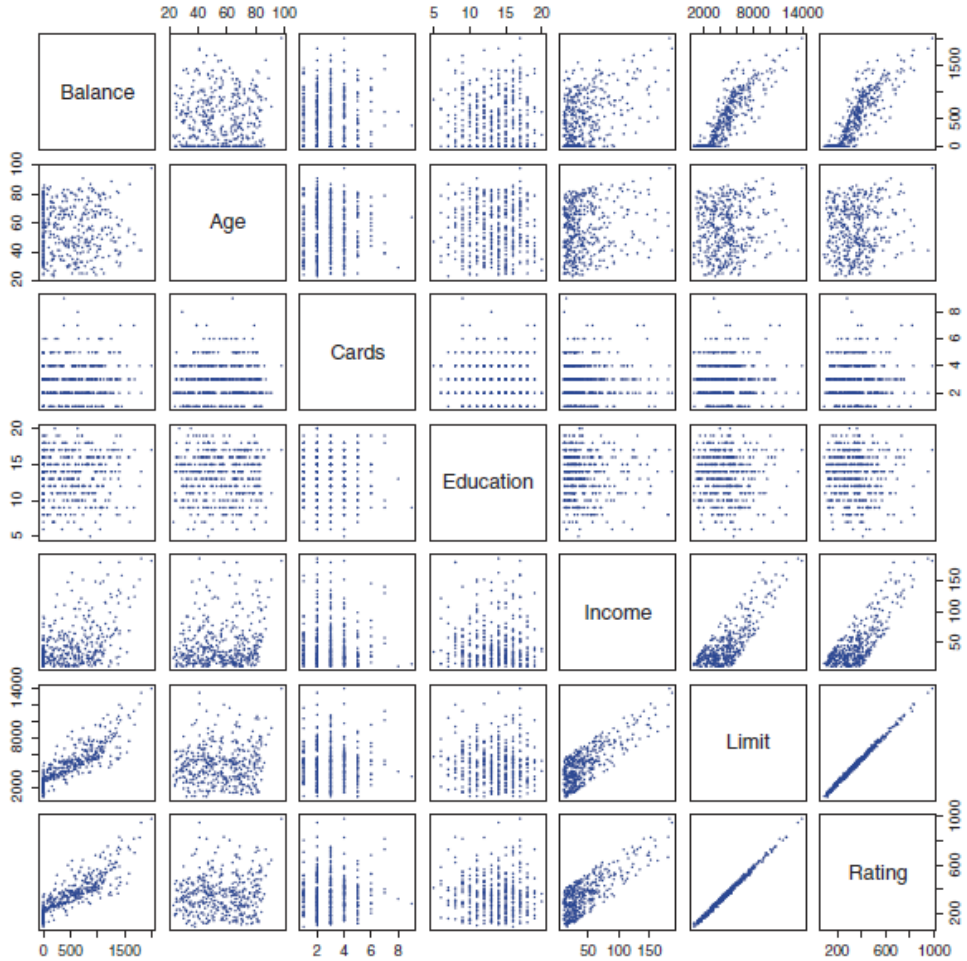


FIGURE 3.6. *The* `Credit` *data set contains information about* `balance`, `age`, `cards`, `education`, `income`, `limit`, *and* `rating` *for a number of potential customers.*

Ridge Regression

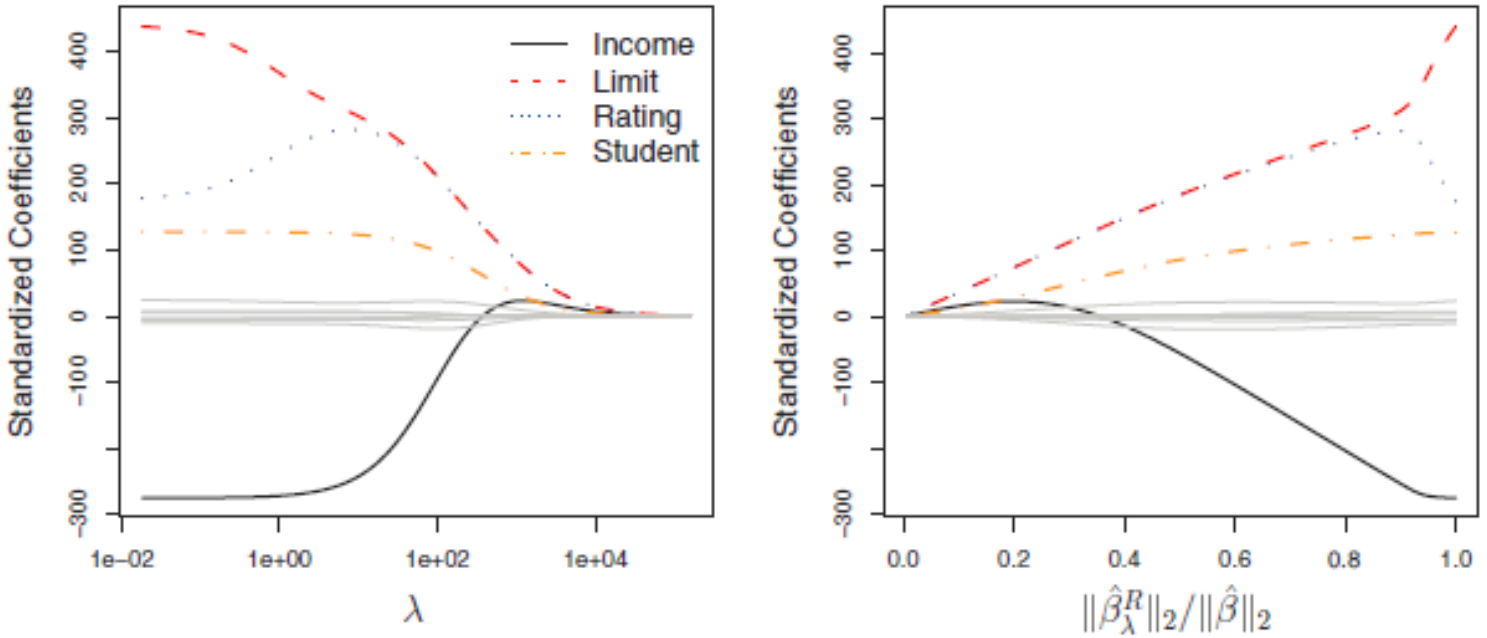


FIGURE 6.4. *The standardized ridge regression coefficients are displayed for the* `Credit` *data set, as a function of* λ *and* $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$.

$$\lambda \uparrow \quad ||\hat{\beta}^R_\lambda||_2 \downarrow$$

Lasso

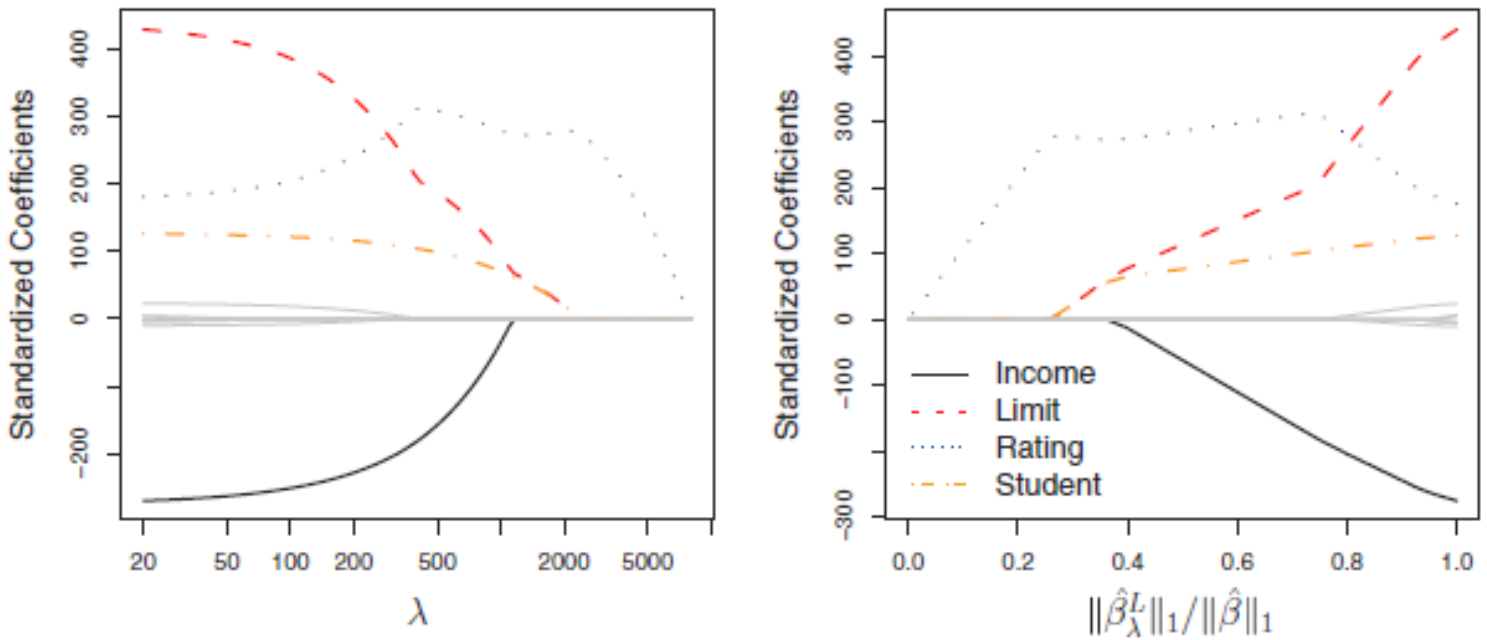


FIGURE 6.6. *The standardized lasso coefficients on the* `Credit` *data set are shown as a function of* λ *and* $\|\hat{\beta}^L_\lambda\|_1/\|\hat{\beta}\|_1$.

$$\lambda \uparrow \quad ||\hat{\beta}^L_\lambda||_1 \downarrow$$

Lasso does variable selection, and gives sparse model.

Lasso

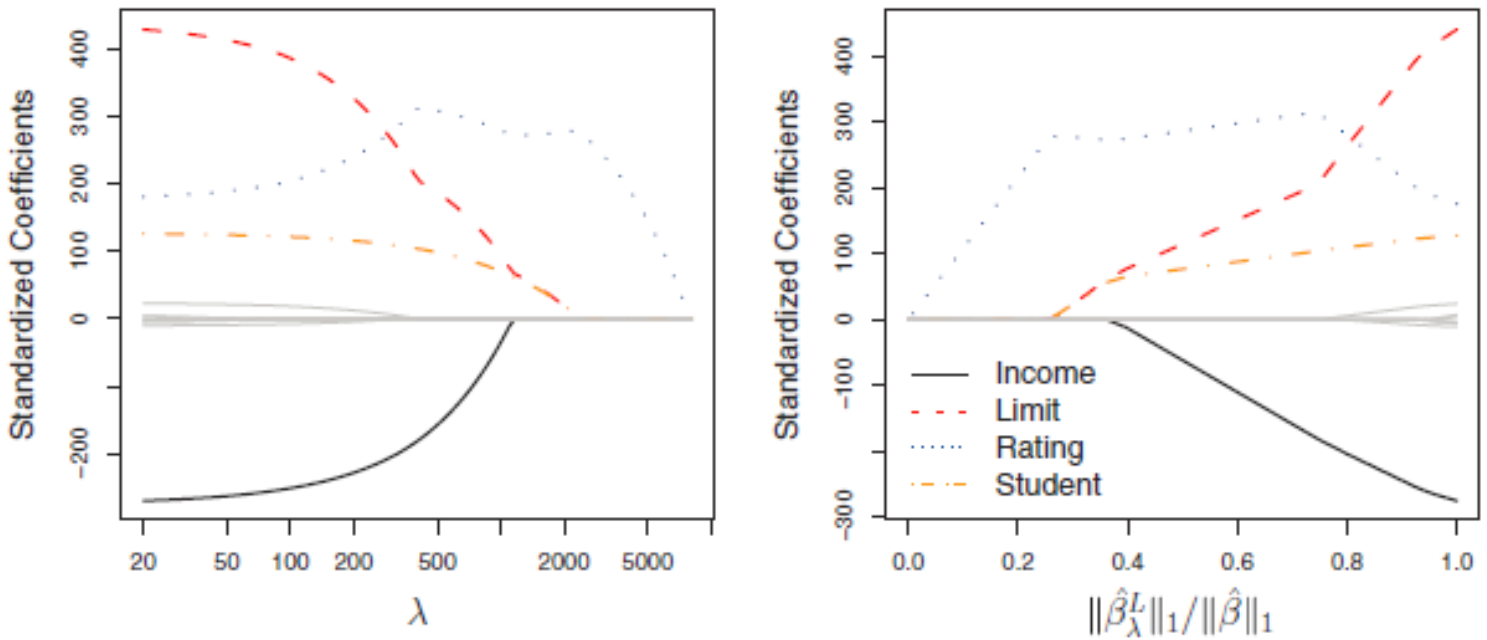
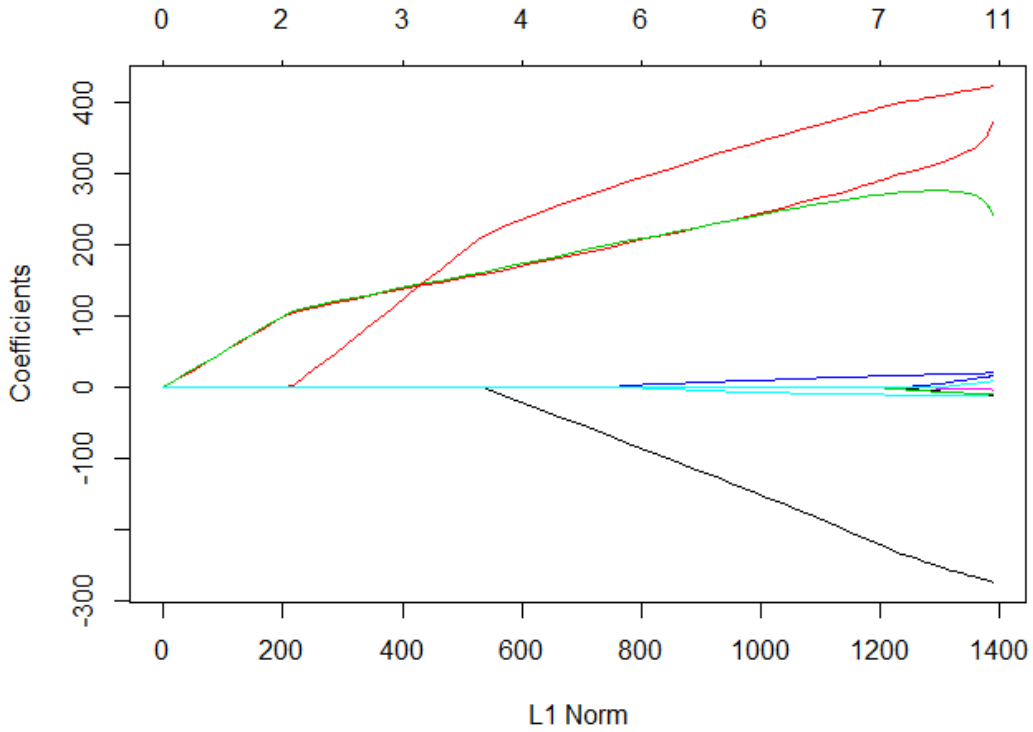


FIGURE 6.6. *The standardized lasso coefficients on the* `Credit` *data set are shown as a function of* λ *and* $\|\hat{\beta}^L_\lambda\|_1/\|\hat{\beta}\|_1$.

$$\lambda \uparrow \quad ||\hat{\beta}^L_\lambda||_1 \downarrow$$

Lasso does variable selection, and gives sparse model.

Elastic Net



$$\frac{\lambda}{\mu} = 0.3$$

The path for Limit and Rating are very similar.

Choose λ

Berkeley
UNIVERSITY OF CALIFORNIA

Cross-Validation