Introduction to Data Science

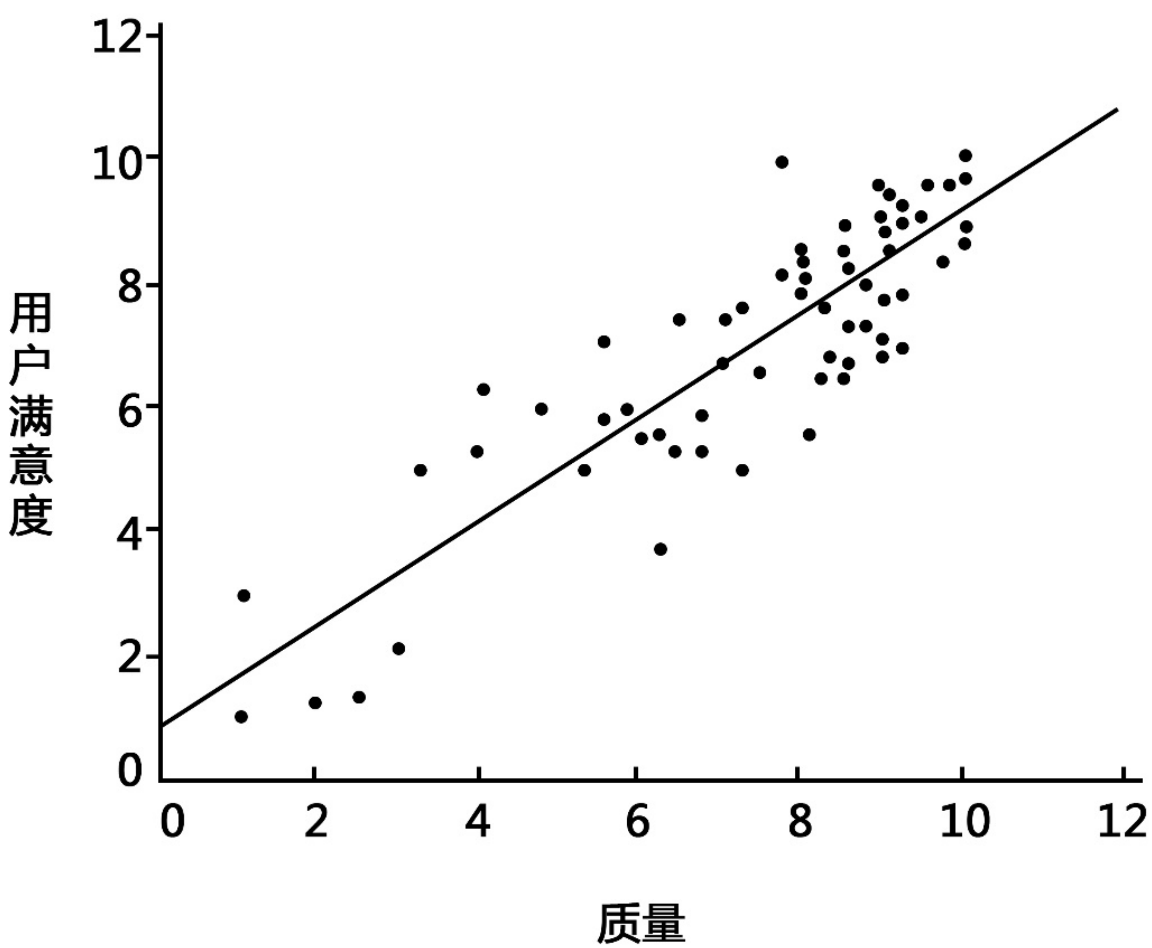
Lecturer: 朱占星

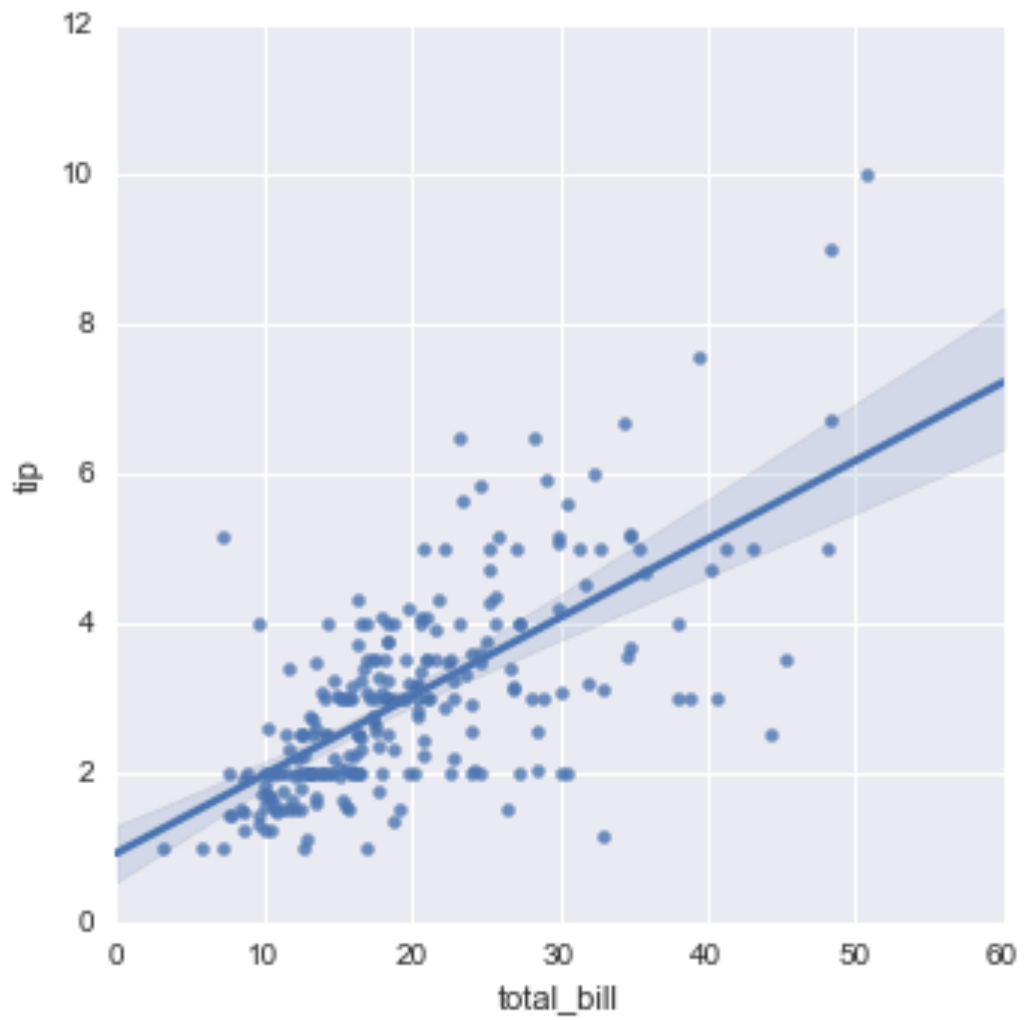
Peking University

Regression

Regression

Use some variables (or features, denoted as x) to
predict some other **continuous** variables (denoted as y)



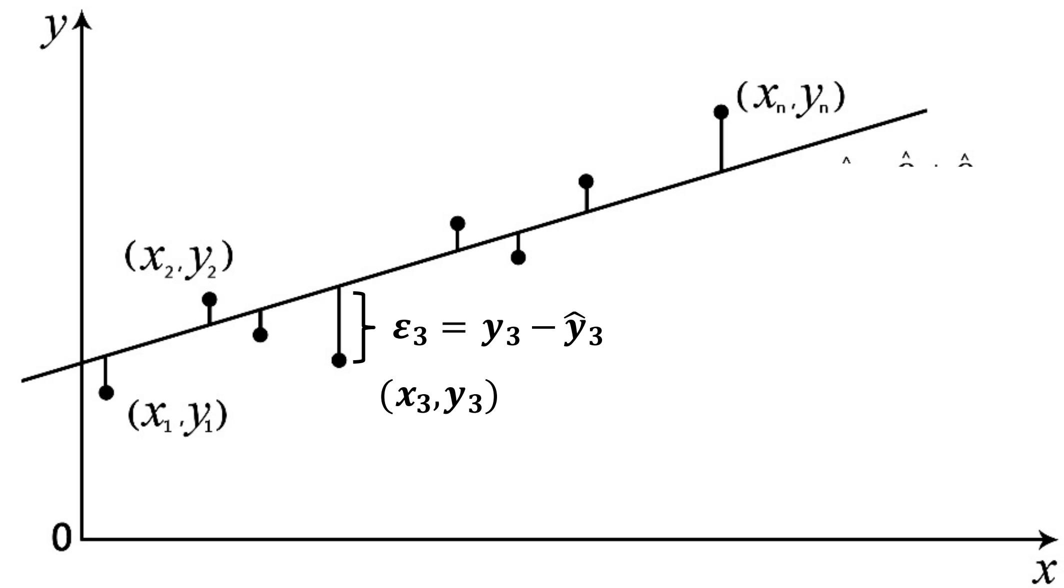


Linear Regression

- The model:

$$y = \boldsymbol{w}^T \boldsymbol{x} + \epsilon$$

where $\boldsymbol{x} = [x_1, x_2, \ldots, x_d, 1]$
w is the parameter to be learned,
the noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$

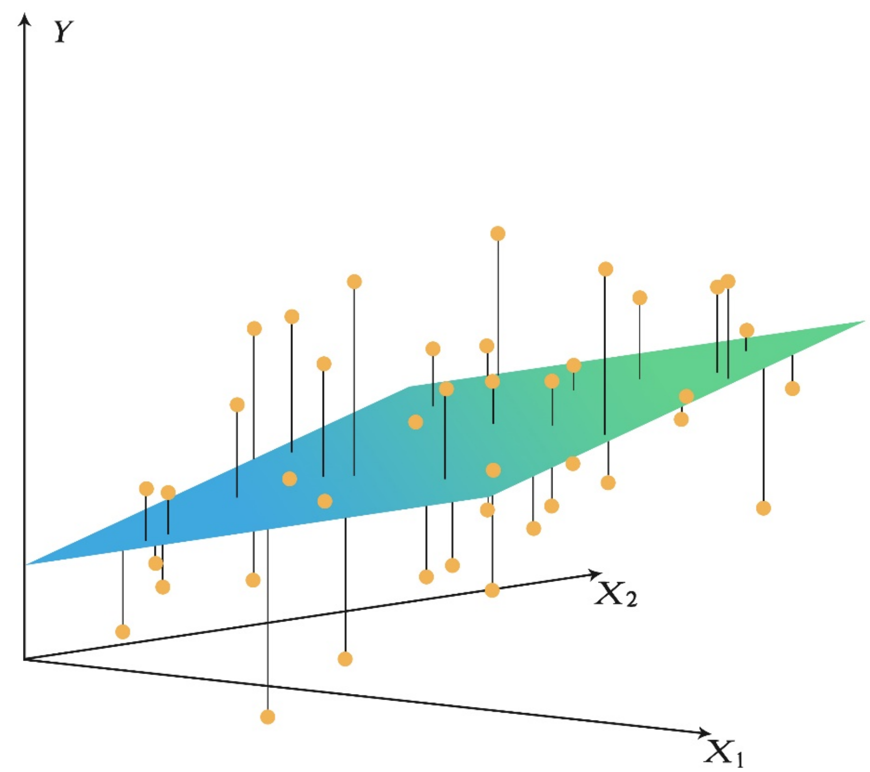


- Given the training data, $\{\boldsymbol{x}_i, y_i\}_{i=1}^N$ we want to learn w

- Minimizing the least square error,

$$\min_{\boldsymbol{w}} L(\boldsymbol{w}) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \boldsymbol{w}^T \boldsymbol{x}_i)^2$$

$$L(\boldsymbol{w}) = \frac{1}{N} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})$$



- Obtain the derivative, let it be zero to get the minima.

$$\nabla_{\boldsymbol{w}} L = -\frac{2}{N} \boldsymbol{X}^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) = 0$$

When $\boldsymbol{X}^T \boldsymbol{X}$ has full rank, the optimal solution

$$\boldsymbol{w}^* = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$$

- What if not full rank? Later we will talk about this.

- Another derivation from maximum likelihood estimation (MLE)

- MLE solution for linear regression
  - Predicted variable follows normal distribution since the noise follows normal distribution.

$$P(y_i|\boldsymbol{x},\boldsymbol{w}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\left(y_i - \sum_{j=1}^{d+1} x_{ij} w_j\right)^2}{2\sigma^2}}$$

  - Then maximize the log likelihood (independent and identically distributed assumption, so called i.i.d observations)

$$l(\boldsymbol{w};\boldsymbol{y}) = \log L(\boldsymbol{w};\boldsymbol{y}) = \log \prod_{i=1}^{n} P(y_i|\boldsymbol{x},\boldsymbol{w}) = \sum_{i=1}^{n} \log P(y_i|\boldsymbol{x},\boldsymbol{w}) = \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\left(y_i - \sum_{j=1}^{d+1} x_{ij} w_j\right)^2}{2\sigma^2}}$$

$$= n \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{d+1} x_{ij} w_j\right)^2$$

$$\Leftrightarrow \nabla_{\boldsymbol{w}} l(\boldsymbol{w};\boldsymbol{y}) = -\frac{1}{\sigma^2} \mathbf{X}^{\mathrm{T}}(\boldsymbol{y} - \mathbf{X}\boldsymbol{w})$$

Let $\nabla_{\boldsymbol{w}} l(\boldsymbol{w};\boldsymbol{y}) = 0$, then $\hat{\boldsymbol{w}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{y}$

- Multi-collinearity
  - One predictor variable can be linearly predicted by other variables to some degree.

$$\boldsymbol{w}^* = \boxed{(\boldsymbol{X}^T\boldsymbol{X})^{-1}}\boldsymbol{X}^T\boldsymbol{y}$$

ill-conditioning even low rank, and cannot be inverted
typically obtain very large values of w with large estimation variance

序	1	2	3	4	5	6	7	8	9	10
x	1.1	1.4	1.7	1.7	1.8	1.8	1.9	2.0	2.3	2.4
x_	1.1	1.5	1.8	1.7	1.9	1.8	1.8	2.1	2.4	2.5
ε i	0.8	-0.5	0.4	-0.5	0.2	1.9	1.9	0.6	-1.5	-1.5
y i	16.3	16.8	19.2	18.0	19.5	20.9	21.1	20.9	20.3	22.0

Estimated:

$$w0=11.292, w1=11.307, w2=-6.591$$

Ground truth:

$$w0 = 10, w1 = 2, w2 = 3$$

$$\rho_{X,Y} = \frac{\mathrm{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X\sigma_Y}$$

Correlation coefficient between x1 and x2

$$r = 0.986$$

Ridge Regression

- One solution to avoid multi-collinearity in linear regression
- Provide numerically stable and low-variance estimation
- Penalizing the L2 norm of the weight vector

$$\min_{\boldsymbol{w}} \quad \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2, \quad \text{s.t.} \quad \|\boldsymbol{w}\|_2 \leqslant C$$

$$\Big\downarrow$$

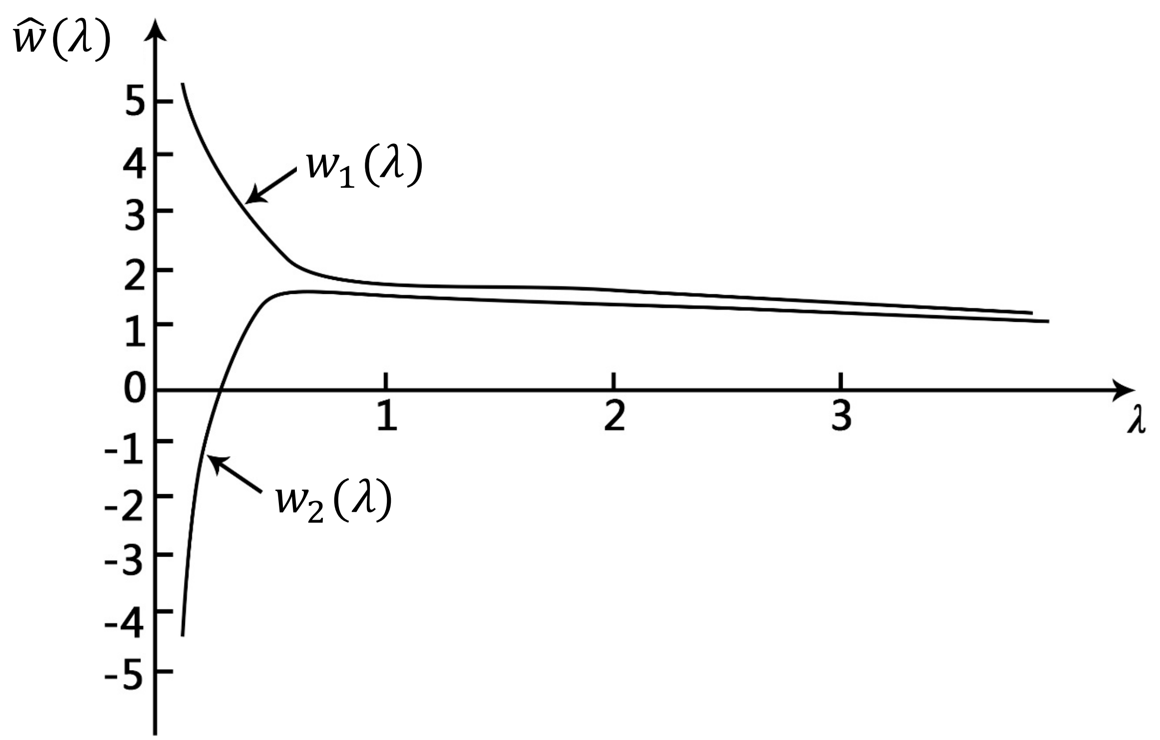
$$\min_{\boldsymbol{w}} \quad \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \lambda\|\boldsymbol{w}\|_2^2$$

$$\Big\downarrow$$

$$\begin{aligned}\boldsymbol{w}^{\text{ridge}} &= \operatorname*{argmin}_{\boldsymbol{w}} \left(\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \lambda\|\boldsymbol{w}\|_2^2\right) \\ &= \boxed{(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}.\end{aligned}$$

- Ridge trace plot
  - Weight w.r.t. lambda
  - To check the degree of collinearity

λ	0	0.1	0.15	0.2	0.3	0.4	0.5	1.0	1.5	2.0	3.0
wˆ_1^r	11.31	3.48	2.99	2.71	2.39	2.20	2.06	1.66	1.43	1.27	1.03
wˆ_2^r	-6.59	0.63	1.02	1.21	1.39	1.46	1.49	1.41	1.28	1.17	0.98



Regularization

- In general, "regularization" means any techniques that help to improve models' generalization performance.
- But now, we mainly talk about "explicit" regularization, coding the
  - prior knowledge about the model
  - or necessary statistical assumptions

$$\min_{\theta} \mathbb{E}_{P_{data}}[l(f(x;\theta), y)]$$

what we really want

empirical risk minimization ←

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} l(f(x_i;\theta), y_i) + R(\theta)$$

what we actually do

regularization term, controlling the model complexity or enforcing prior constraints

Lasso

- Abbreviation for least absolute shrinkage and selection operator
- Linear regression with L1 norm regularization
- Yielding sparse solutions
- Reduce the model complexity, particularly useful when N << d
- A good **variable selection** method

$$\min_{\boldsymbol{w}} \quad \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}\|_2^2+\lambda\|\boldsymbol{w}\|_1$$

Regression shrinkage and selection via the lasso

R Tibshirani - Journal of the Royal Statistical Society. Series B (…, 1996 - JSTOR

We propose a new method for estimation in linear models. Thelasso'minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. Because of the nature of this constraint it tends to produce some coefficients that ...

被引用次数：16687　相关文章　所有 78 个版本　引用　保存　更多

- How to solve Lasso?
  - LARS, coordinate descent, proximal algorithms (ISTA, FISTA)
  - Iterative Shrinkage Thresholding Algorithm (ISTA)

Gradient descent: $$\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} - \eta \nabla f(\boldsymbol{w}^{(t)})$$

Proximal form:

$$\boldsymbol{w}^{(t+1)} = \operatorname*{argmin}_{\boldsymbol{w}} f(\boldsymbol{w}^{(t)}) + \nabla f(\boldsymbol{w}^{(t)})^{\mathrm{T}}(\boldsymbol{w} - \boldsymbol{w}^{(t)}) + \frac{1}{2\eta}\|\boldsymbol{w} - \boldsymbol{w}^{(t)}\|_2^2$$

More general form:

$$\begin{aligned}\boldsymbol{w}^{(t+1)} &= \operatorname*{argmin}_{\boldsymbol{w}} f(\boldsymbol{w}^{(t)}) + \nabla f(\boldsymbol{w})^{\mathrm{T}}(\boldsymbol{w} - \boldsymbol{w}^{(t)}) + \frac{1}{2\eta}\|\boldsymbol{w} - \boldsymbol{w}^{(t)}\|_2^2 + g(\boldsymbol{w}) \\ &= \operatorname*{argmin}_{\boldsymbol{w}} g(\boldsymbol{w}) + \frac{1}{2\eta}\|\boldsymbol{w} - (\boldsymbol{w}^{(t)} - \eta \nabla f(\boldsymbol{w}^{(t)}))\|_2^2.\end{aligned}$$

- In Lasso case, we have

$$f(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 \qquad g(\boldsymbol{w}) = \lambda\|\boldsymbol{w}\|_1$$

$$\boldsymbol{w}^{(t+1)} = \operatorname*{argmin}_{\boldsymbol{w}} \lambda\|\boldsymbol{w}\|_1 + \frac{1}{2\eta}\|\boldsymbol{w} - (\boldsymbol{w}^{(t)} - \eta\nabla f(\boldsymbol{w}^{(t)}))\|_2^2$$

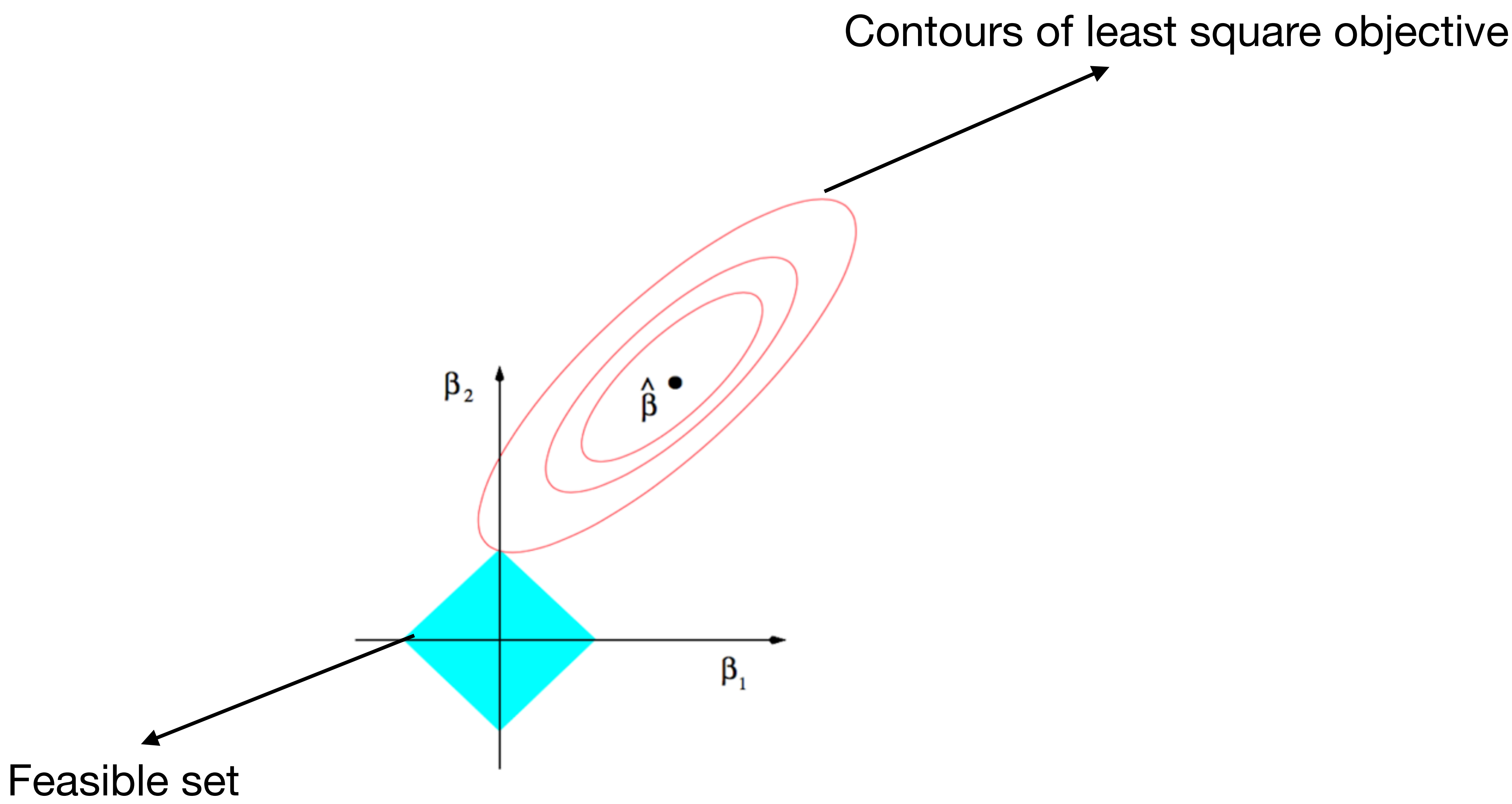
$$\boldsymbol{w}^{(t+1)} = S_{\eta\lambda}(\boldsymbol{w}^{(t)} - \eta\nabla f(\boldsymbol{w}^{(t)}))$$

Soft thresholding operator

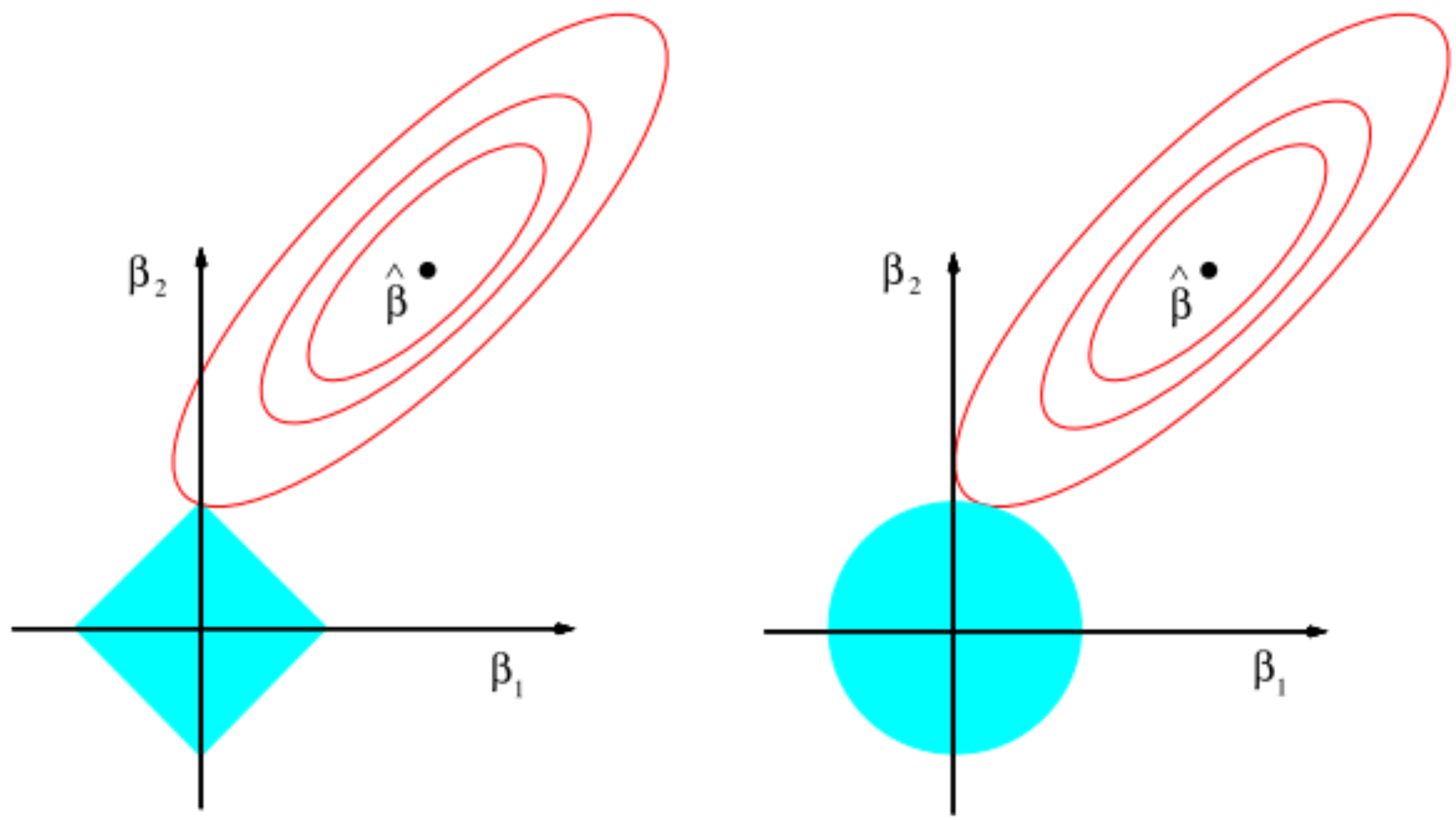
$$(S_a(\boldsymbol{v}))_i = \begin{cases} v_i - a, & 若\ v_i > a; \\ 0, & 若\ |v_i| \leqslant a; \\ v_i + a, & 若\ v_i < -a. \end{cases}$$

For more details, see N. Parikh and S. Boyd. Proximal algorithms. Foundations and Trends in Optimization, 1(3): 123–231, 2013.

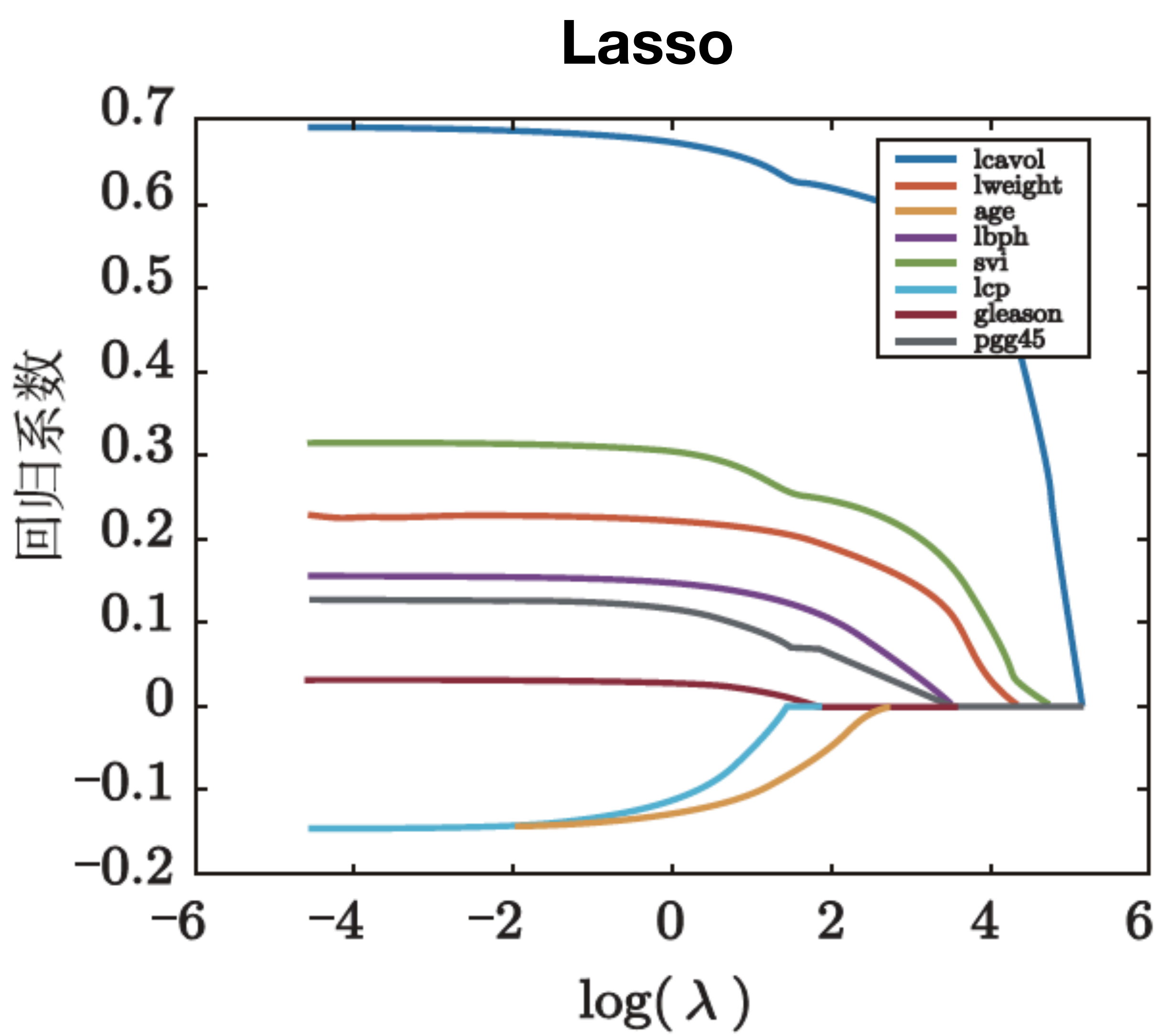
- Why does Lasso provide sparse solutions?
  - L1 norm constraint matters.

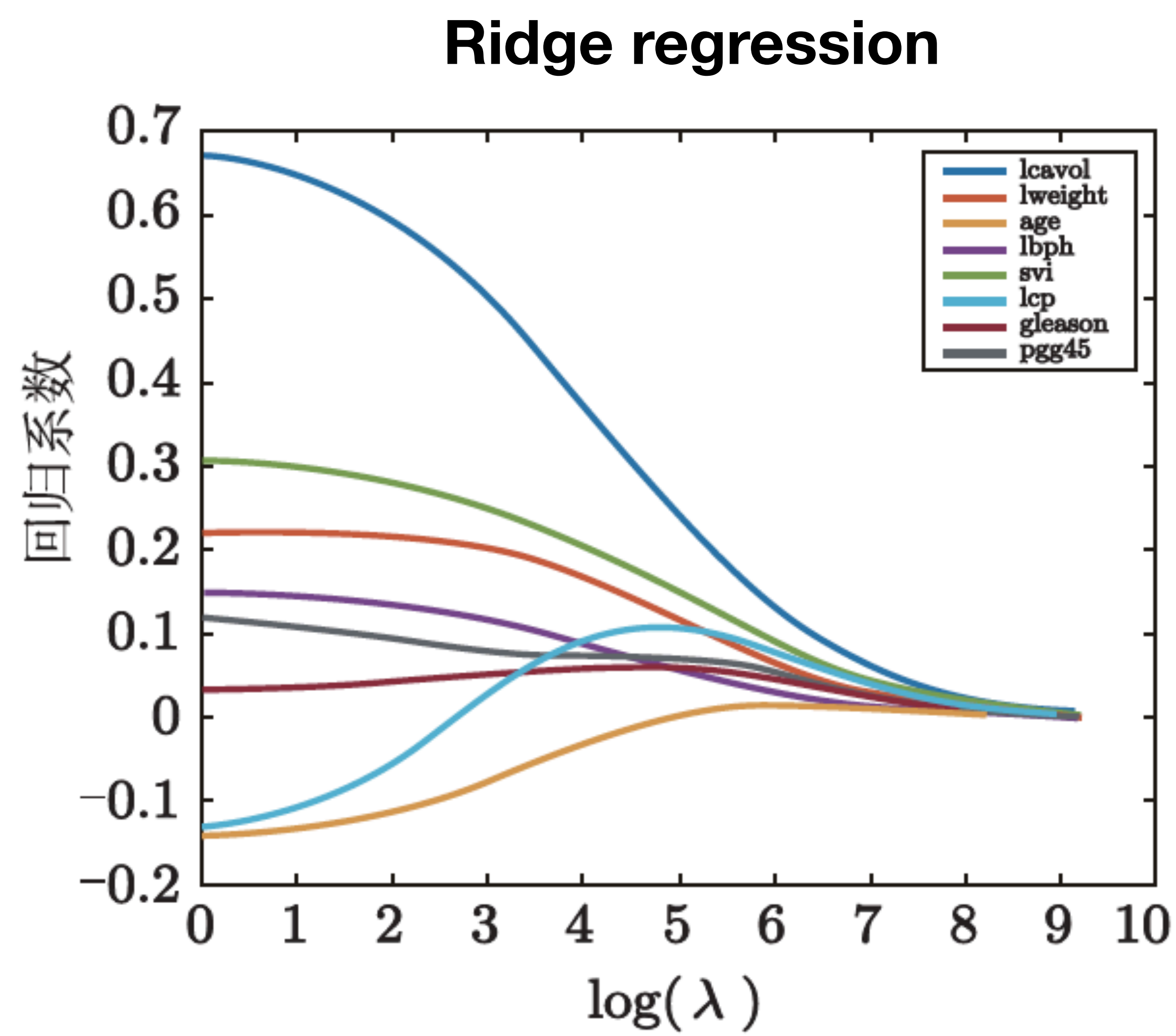


- Comparison with ridge regression



- Selection of regularization parameter \lambda
  - Regularization paths
  - Practically, cross validation for selecting \lambda





Discussion

- If there exists some correlated variables, but they are both important (e.g. in gene selection), what will Lasso do?

- Solution: elastic net regularization

- Avoid only select one of the correlated variables

$$J(\boldsymbol{w}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \lambda_1 \|\boldsymbol{w}\|_1 + \lambda_2 \|\boldsymbol{w}\|_2^2$$

J. R. Statist. Soc. B (2005)
67, *Part* 2, *pp.* 301–320

Regularization and variable selection via the elastic net

Hui Zou and Trevor Hastie
Stanford University, USA

In this paper we propose a new regularization technique which we call the *elastic net*. Similar to the lasso, the elastic net simultaneously does automatic variable selection and continuous shrinkage, and it can select groups of correlated variables. It is like a stretchable fishing net that retains 'all the big fish'. Simulation studies and real data examples show that the elastic net often outperforms the lasso in terms of prediction accuracy.

2D contour plot for ridge, Lasso and elastic net

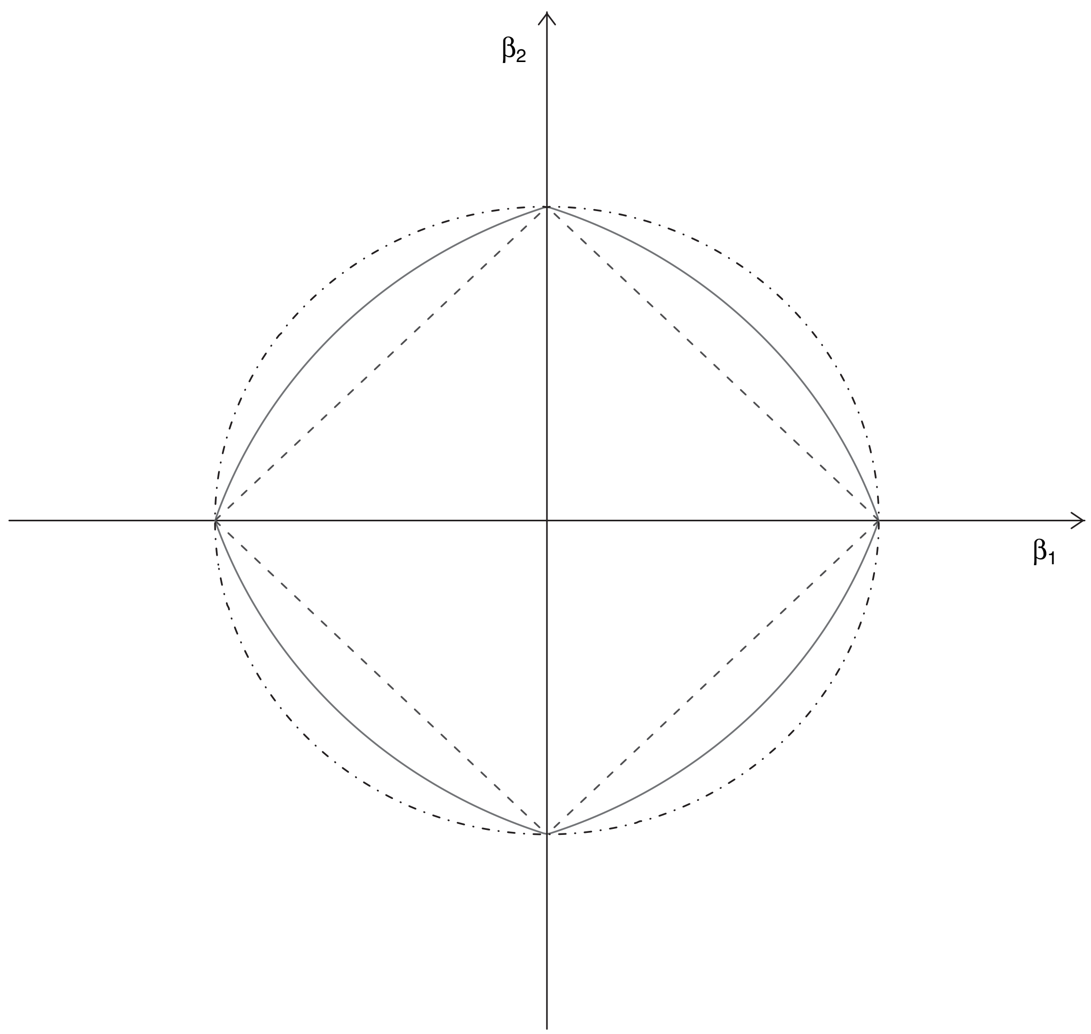


Figure from Zou and Hastie (2005)

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} |\mathbf{y} - \mathbf{X}\boldsymbol{\beta}|^2 + \lambda\, J(\boldsymbol{\beta}) \tag{7}$$

where $J(\cdot)$ is positive valued for $\boldsymbol{\beta} \neq 0$.

Qualitatively speaking, a regression method exhibits the grouping effect if the regression coefficients of a group of highly correlated variables tend to be equal (up to a change of sign if negatively correlated). In particular, in the extreme situation where some variables are exactly identical, the regression method should assign identical coefficients to the identical variables.

Lemma 2. Assume that $\mathbf{x}_i = \mathbf{x}_j$, $i, j \in \{1, \ldots, p\}$.

(a) If $J(\cdot)$ is strictly convex, then $\hat{\beta}_i = \hat{\beta}_j$, $\forall \lambda > 0$.
(b) If $J(\boldsymbol{\beta}) = |\boldsymbol{\beta}|_1$, then $\hat{\beta}_i \hat{\beta}_j \geqslant 0$ and $\hat{\boldsymbol{\beta}}^*$ is another minimizer of equation (7), where

$$\hat{\beta}_k^* = \begin{cases} \hat{\beta}_k & \text{if } k \neq i \text{ and } k \neq j, \\ (\hat{\beta}_i + \hat{\beta}_j) \cdot (s) & \text{if } k = i, \\ (\hat{\beta}_i + \hat{\beta}_j) \cdot (1 - s) & \text{if } k = j, \end{cases}$$

for any $s \in [0, 1]$.

Discussion

- What if the features form groups,
  we only want to select some groups?

- Solution: group Lasso

$$J(\boldsymbol{w}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \sum_{g=1}^{G} \lambda_g \|\boldsymbol{w}_g\|_2$$

J. R. Statist. Soc. B (2006)
68, *Part* 1, *pp.* 49–67

Model selection and estimation in regression with grouped variables

Ming Yuan
Georgia Institute of Technology, Atlanta, USA

and Yi Lin
University of Wisconsin—Madison, USA

[Received November 2004. Revised August 2005]

- Extensions and comments
  - The features can be quite general, not just the original ones. E.g. x1 * x2 or through some nonlinear transformation g(x), (see kernel machines)
  - Though too simple and naive, but sometimes most effective. Remember to try linear regression as your first choice.
  - Nice interpretability. Variables with large weights are important. Very important in medical, bioinformatic, and business applications

Nonlinear Regression

- **Spline regression**
- **Radial basis function (RBF) networks**
- Support vector regression (SVR), kernel methods
- Gaussian Processes (GP)
- Neural networks
- …

- Spline regression (1D, for multi-dim cases, see [1])
  - Piecewise polynomial connected by control knots

Linear spline

$$y = \beta_0 + \beta_1 x + w_1(x - a_1)_+ + w_2(x - a_2)_+ + \cdots + w_k(x - a_k)_+$$

$$\boldsymbol{y} = \boldsymbol{G}\boldsymbol{w} \qquad \boldsymbol{G} = \begin{bmatrix} 1 & x_1 & (x_1 - a_1)_+ & \cdots & (x_1 - a_k)_+ \\ 1 & x_2 & (x_2 - a_1)_+ & \cdots & (x_2 - a_k)_+ \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & (x_n - a_1)_+ & \cdots & (x_n - a_k)_+ \end{bmatrix}$$

$$\boldsymbol{w} = (\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G})^{-1}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{y}$$

Ridge version $$\min_{\boldsymbol{w}} \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i + \sum_{j=1}^{k} w_j (x_i - a_j)_+) \right)^2 + \lambda \sum_{j=1}^{k} w_j^2$$

[1] Friedman, Jerome H. "Multivariate adaptive regression splines." *The annals of statistics* (1991): 1-67.

- Cubic spline

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{k=1}^{k} (x - a_k)_+^3$$

- B spline

$$B(x) = \sum_{j=0}^{k+m} w_j B_{j,k}(x), \quad x \in [a_0, a_{k+1}]$$

$$B_{j,0}(x) = \begin{cases} 1, & \text{若 } a_j \leqslant x < a_{j+1} \\ 0, & \text{其他}, \end{cases}$$

$$B_{j,k+1}(x) = \alpha_{j,k+1}(x) B_{j,k}(x) + (1 - \alpha_{j+1,k+1}(x)) B_{j+1,k}(x)$$

$$\alpha_{j,k}(x) = \begin{cases} \dfrac{x - t_j}{t_{j+k} - t_j}, & \text{若 } a_{j+k} \neq a_j \\ 0, & \text{其他}. \end{cases}$$

- Radial basis function (RBF) networks

  - RBFs $\phi(\|\boldsymbol{x} - \boldsymbol{c}\|)$

$$y = \sum_{j=1}^{k} w_j \phi(\|\boldsymbol{x} - \boldsymbol{c}_j\|)$$

Gaussian RBF

$$\phi(r) = \mathrm{e}^{-ar^2}$$

Multi-quadric

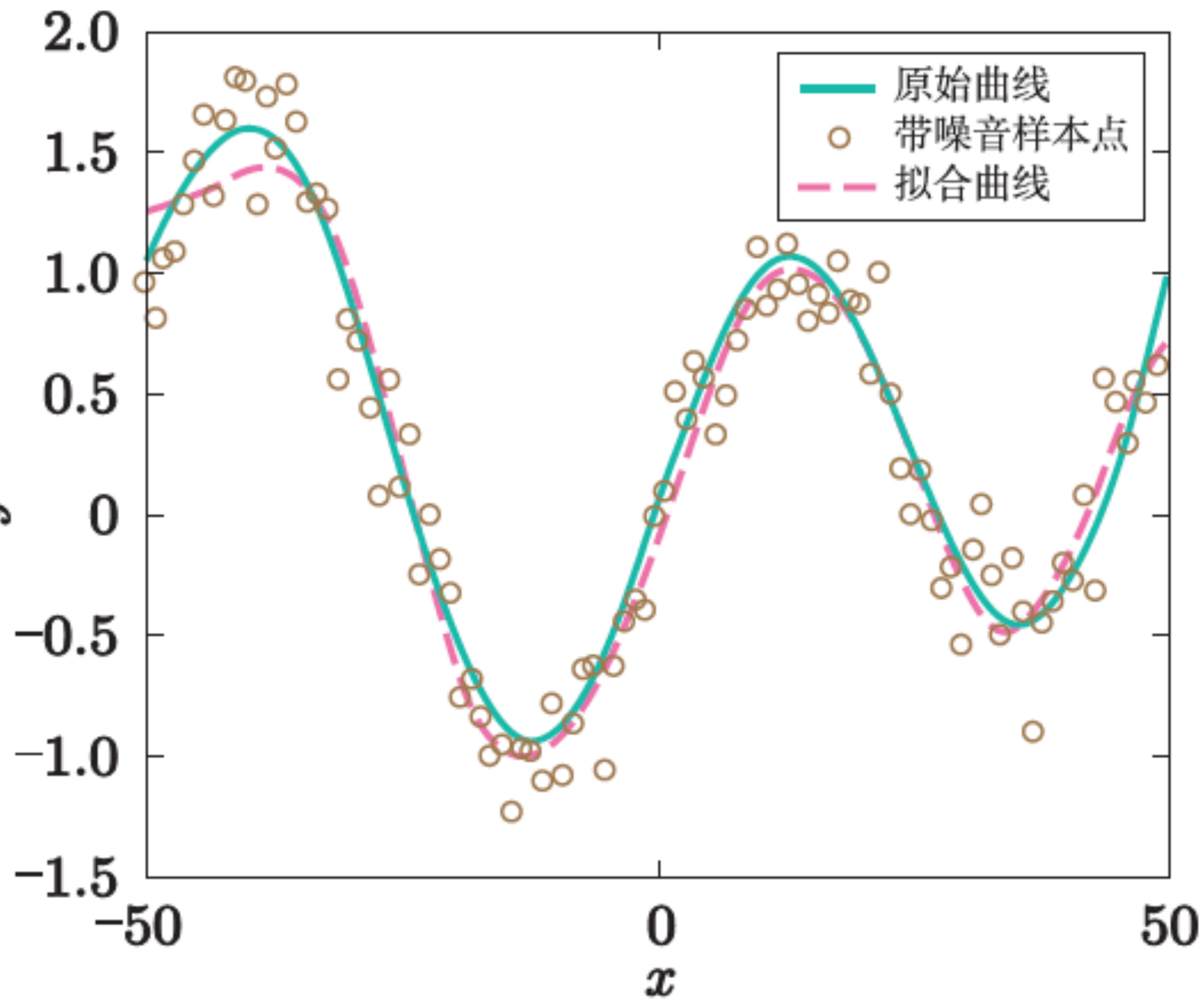
$$\phi(r) = \sqrt{1 + ar^2}$$

Inverse quadratic

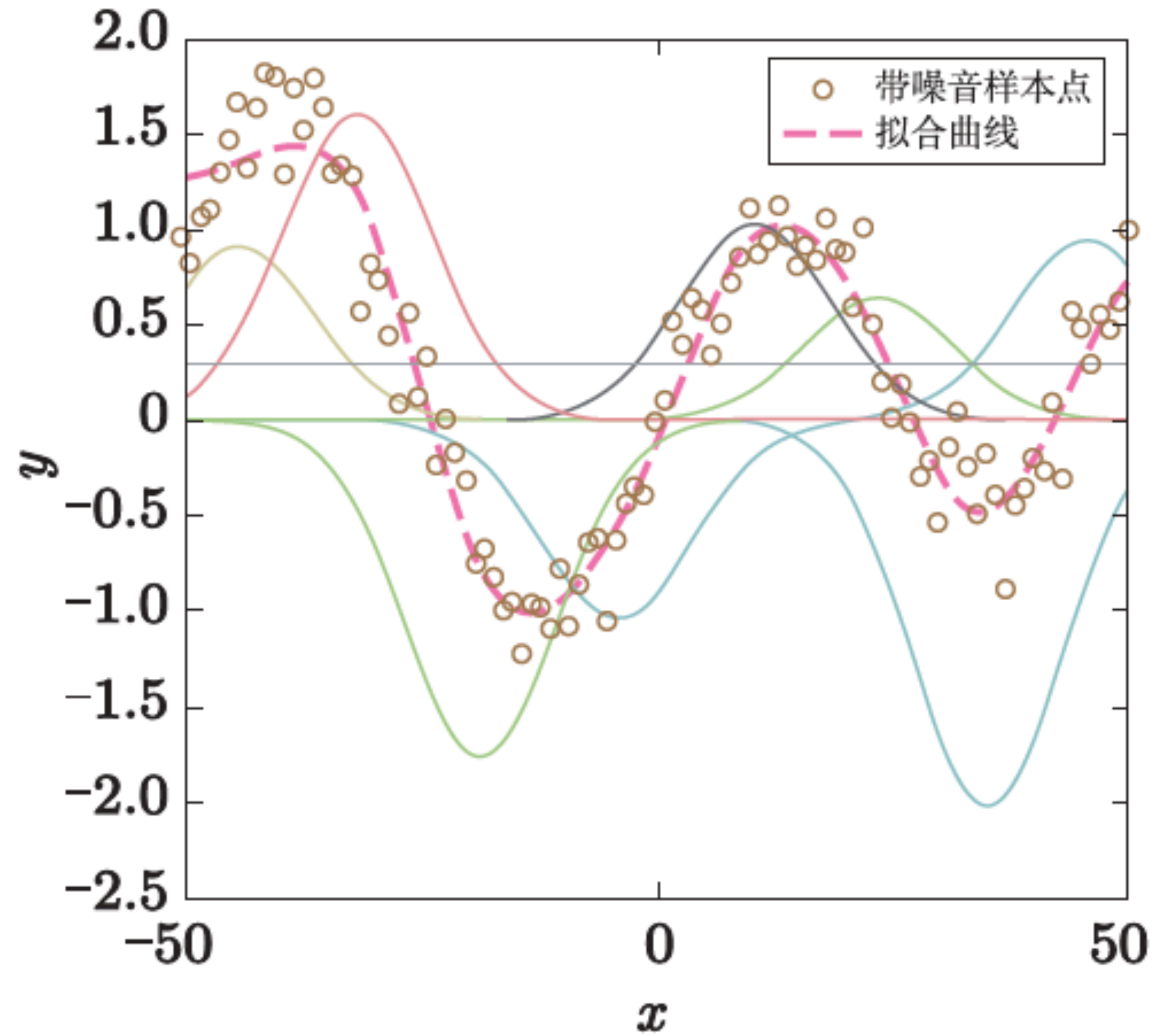
$$\phi(r) = \frac{1}{1 + ar^2}$$

  - The centroid vectors can be obtained by random sampling from training points or clusterings

  - How to solve w?



(a)



(b)

Summary

- Regression: a widely used techniques in ML
- Linear regression
- **The power of regularization**
- Some nonlinear regression methods

Exercises

- Derive ISTA for Lasso, and implement it.

- Optional readings
  - Efron, Bradley, et al. "Least angle regression." *The Annals of statistics* 32.2 (2004): 407-499.
  - Wu, Tong Tong, and Kenneth Lange. "Coordinate descent algorithms for lasso penalized regression." *The Annals of Applied Statistics* 2.1 (2008): 224-244.
  - N. Parikh and S. Boyd. Proximal algorithms. Foundations and Trends in Optimization, 1(3): 123–231, 2013.