Overview

Lasso

— Duality between constrained optimization & Lagrangian
— What's the optimal lambda value?
— Should we standardize the features?

Ridge

— Understand the shrinkage effect through PC transformation

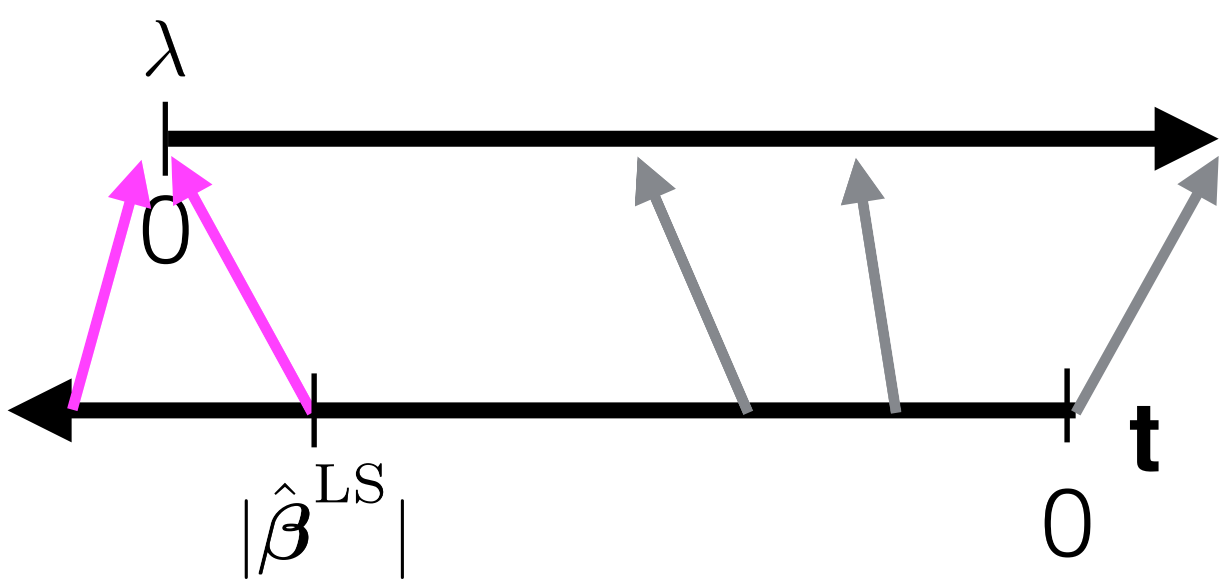
Other Penalty Choices

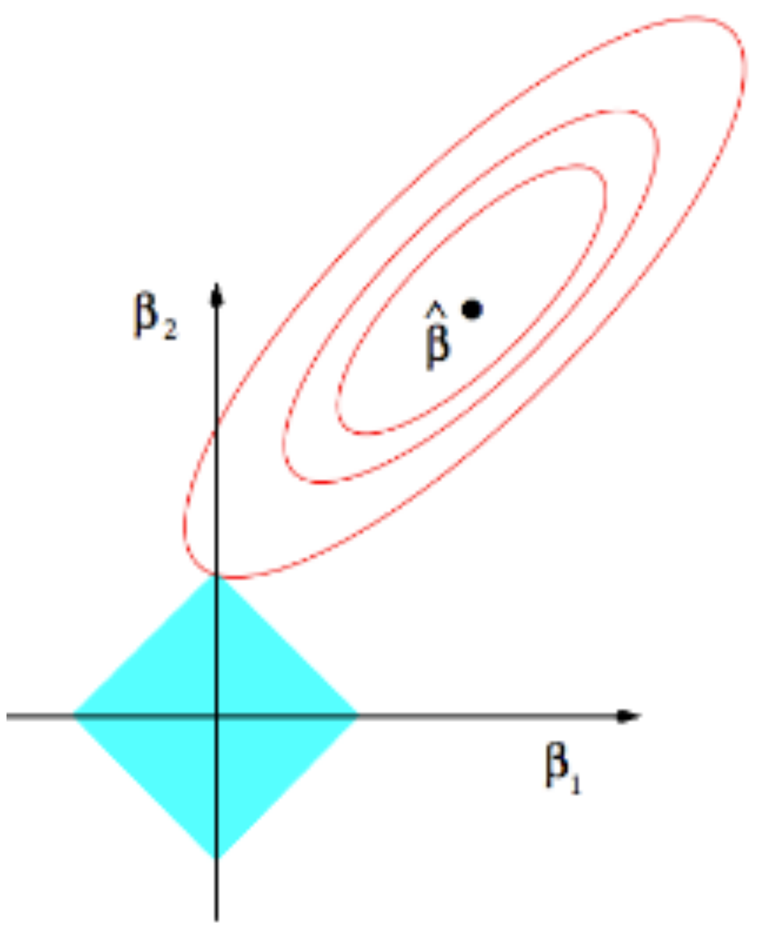
Lasso

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\left[\frac{1}{2n}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2+\lambda|\boldsymbol{\beta}|\right]$$

$$\min_{|\boldsymbol{\beta}|}\frac{1}{2n}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2,\ \text{subj to } |\boldsymbol{\beta}|\leq t$$

When t is active, there is a one-to-one correspondence between lambda and t.





Equivalent Formulation

$$\min_{x} f(x)$$
$$\text{subj to } g(x) \leq b$$

Lagrange multiplier formulation

$$\Omega(x, \lambda) = f(x) + \lambda(g(x) - b)$$

KKT Conditions

$$f'(x) + \lambda g'(x) = 0$$
$$g(x) - b \leq 0$$
$$\lambda \geq 0$$
$$\lambda(g(x) - b) = 0$$

Optimal Lambda?

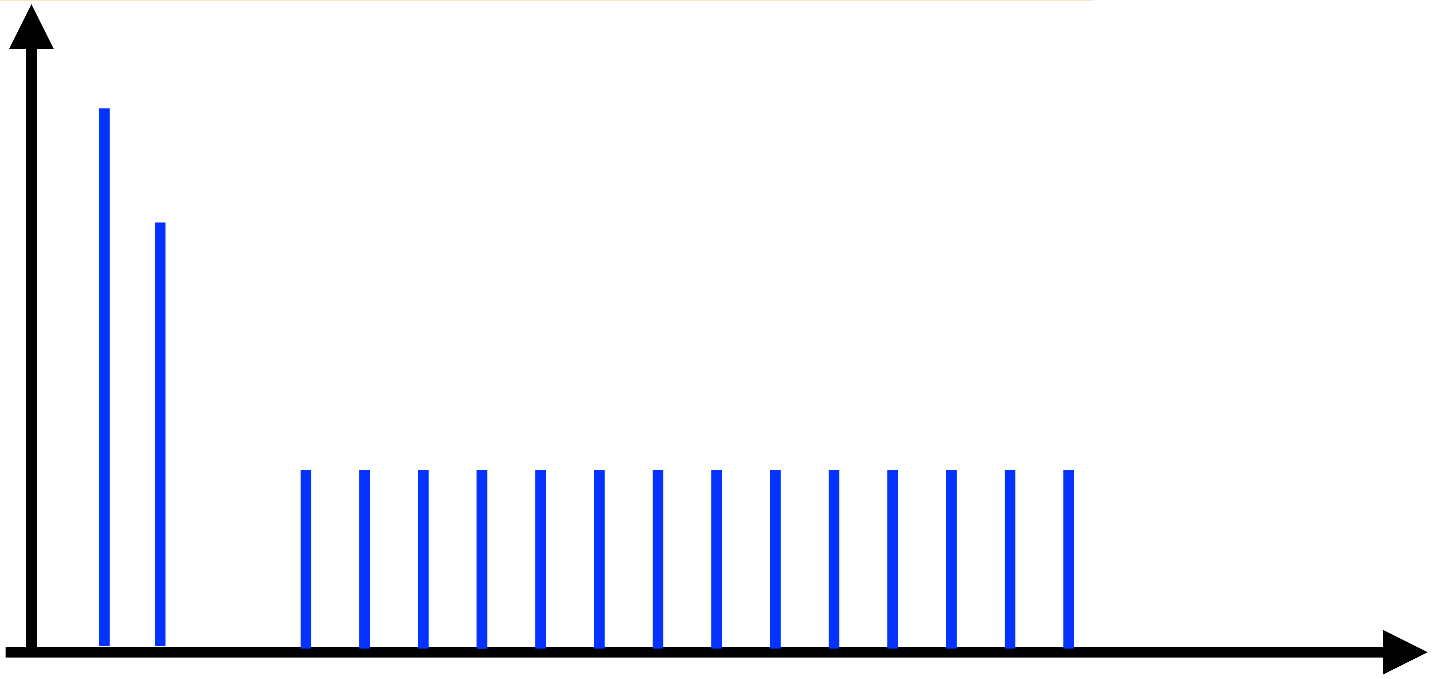
In Lasso, lambda plays the role of a threshold value. What's the optimal threshold value that can separate signal and noise?

Next let's consider a simple normal mean problem.

$$X_1, X_2, \dots, X_n \text{ iid } \sim N_p(\boldsymbol{\theta}_{p\times 1}, \sigma^2 \mathbf{I}_p)$$

$$\Longrightarrow \bar{X} \sim N_p\left(\boldsymbol{\theta}_{p\times 1}, \frac{\sigma^2}{n}\mathbf{I}_p\right)$$

Magnitude of X-bar for each dimension



Optimal Lambda?

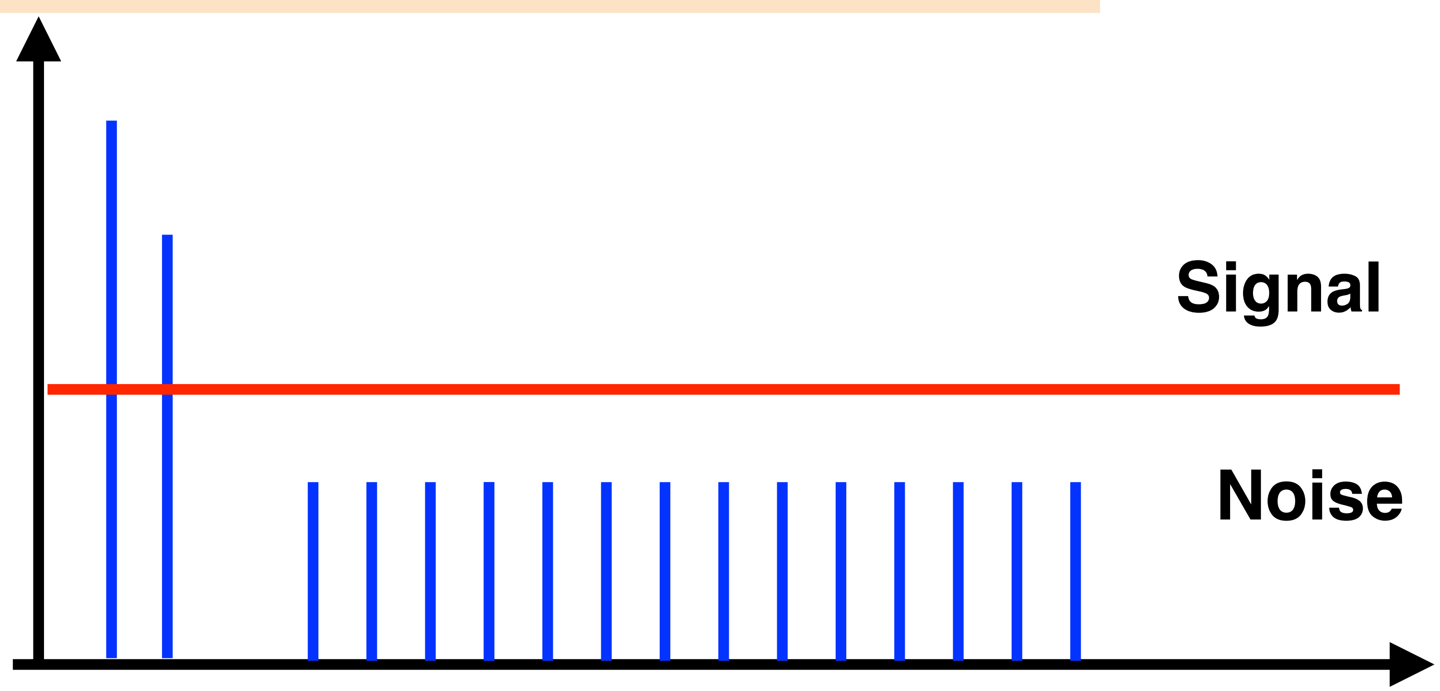
In Lasso, lambda plays the role of a threshold value. What's the optimal threshold value that can separates signal and noise?

Next let's consider a simple normal mean problem.

$$X_1, X_2, \ldots, X_n \text{ iid } \sim N_p(\boldsymbol{\theta}_{p\times 1}, \sigma^2 \mathbf{I}_p)$$

$$\Longrightarrow \bar{X} \sim N_p\Big(\boldsymbol{\theta}_{p\times 1}, \frac{\sigma^2}{n}\mathbf{I}_p\Big)$$

Magnitude of X-bar for each dimension



$$C\frac{\sigma}{\sqrt{n}}$$

How to Choose Lambda?

Suppose all dims are noise

$$\bar{X} \sim N_p(\mathbf{0}_{p\times 1}, \sigma^2 \mathbf{I}_p)$$

$$\mathbb{P}(\max_j \bar{X}_j > \lambda) \leq \sum_{j=1}^{p} \mathbb{P}(\bar{X}_j > \lambda)$$

$$\leq \sum_{j=1}^{p} C' \exp\Big(-\frac{\lambda^2}{\sigma^2/n}\Big)$$

$$= p \cdot C' \exp\Big(-\frac{\lambda^2}{\sigma^2/n}\Big)$$

$$= C' \exp\Big(-\frac{n\lambda^2}{\sigma^2} + \log p\Big)$$

Bound for normal tail probability

Want this quantity go to zero

How to Choose Lambda?

Suppose all dims are noise

$$\bar{X} \sim N_p(\mathbf{0}_{p\times 1}, \sigma^2 \mathbf{I}_p)$$

$$\mathbb{P}(\max_j \bar{X}_j > \lambda) \leq \sum_{j=1}^{p} \mathbb{P}(\bar{X}_j > \lambda)$$

$$\leq \sum_{j=1}^{p} C' \exp\left(-\frac{\lambda^2}{\sigma^2/n}\right)$$

$$= p \cdot C' \exp\left(-\frac{\lambda^2}{\sigma^2/n}\right)$$

$$= C' \exp\left(-\frac{n\lambda^2}{\sigma^2} + \log p\right)$$

Optimal threshold depends on

1) # of noise features
2) variance

$$\lambda > C'' \sqrt{\frac{\sigma^2 \log p}{n}}$$

Standardization

Previously, we had the following derivation assuming X is orthonormal, but the result shown below holds true for any X.

$$\begin{aligned} \|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2 &= \|\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}^{\text{LS}}+\mathbf{X}\hat{\boldsymbol{\beta}}^{\text{LS}}-\mathbf{X}\boldsymbol{\beta}\|^2 \\ &= \|\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}^{\text{LS}}\|^2+\|\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{LS}}-\mathbf{X}\boldsymbol{\beta}\|^2 \end{aligned}$$

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\text{lasso}} &= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\left(\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2+\lambda|\boldsymbol{\beta}|\right) \\ &= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\left(\|\mathbf{X}\hat{\boldsymbol{\beta}}^{\text{LS}}-\mathbf{X}\boldsymbol{\beta}\|^2+\lambda|\boldsymbol{\beta}|\right) \\ &= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\left[(\hat{\boldsymbol{\beta}}^{\text{LS}}-\boldsymbol{\beta})^T\mathbf{X}^T\mathbf{X}(\hat{\boldsymbol{\beta}}^{\text{LS}}-\boldsymbol{\beta})+\lambda|\boldsymbol{\beta}|\right] \end{aligned}$$

Now, we can think our data is just the LS estimate of beta.

Standardization

Previously, we had the following derivation assuming X is orthonormal, but the result shown below holds true for any X.

$$\begin{aligned}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2 &= \|\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}^{\text{LS}}+\mathbf{X}\hat{\boldsymbol{\beta}}^{\text{LS}}-\mathbf{X}\boldsymbol{\beta}\|^2 \\ &= \|\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}^{\text{LS}}\|^2+\|\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{LS}}-\mathbf{X}\boldsymbol{\beta}\|^2\end{aligned}$$

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\text{lasso}} &= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\left(\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2+\lambda|\boldsymbol{\beta}|\right) \\ &= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\left(\|\mathbf{X}\hat{\boldsymbol{\beta}}^{\text{LS}}-\mathbf{X}\boldsymbol{\beta}\|^2+\lambda|\boldsymbol{\beta}|\right) \\ &= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\left[(\hat{\boldsymbol{\beta}}^{\text{LS}}-\boldsymbol{\beta})^T\mathbf{X}^T\mathbf{X}(\hat{\boldsymbol{\beta}}^{\text{LS}}-\boldsymbol{\beta})+\lambda|\boldsymbol{\beta}|\right]\end{aligned}$$

Now, we can think our data is just the LS estimate of beta.

$$\hat{\boldsymbol{\beta}}^{\text{LS}}_{p\times 1}\sim N\left(\boldsymbol{\beta}^0,\sigma^2(\mathbf{X}^t\mathbf{X})^{-1}\right)$$

When X is orthogonal with each column of the same variance (i.e., we have scaled the columns), then we are back to the previous normal mean case.

Standardization

Previously, we had the following derivation assuming X is orthonormal, but the result shown below holds true for any X.

$$\begin{aligned} \|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2 &= \|\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}^{\text{LS}}+\mathbf{X}\hat{\boldsymbol{\beta}}^{\text{LS}}-\mathbf{X}\boldsymbol{\beta}\|^2 \\ &= \|\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}^{\text{LS}}\|^2+\|\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{LS}}-\mathbf{X}\boldsymbol{\beta}\|^2 \end{aligned}$$

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\text{lasso}} &= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\left(\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2+\lambda|\boldsymbol{\beta}|\right) \\ &= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\left(\|\mathbf{X}\hat{\boldsymbol{\beta}}^{\text{LS}}-\mathbf{X}\boldsymbol{\beta}\|^2+\lambda|\boldsymbol{\beta}|\right) \\ &= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\left[(\hat{\boldsymbol{\beta}}^{\text{LS}}-\boldsymbol{\beta})^T\mathbf{X}^T\mathbf{X}(\hat{\boldsymbol{\beta}}^{\text{LS}}-\boldsymbol{\beta})+\lambda|\boldsymbol{\beta}|\right] \end{aligned}$$

Now, we can think our data is just the LS estimate of beta.

$$\hat{\boldsymbol{\beta}}^{\text{LS}}_{p\times 1}\sim N\left(\boldsymbol{\beta}^0,\sigma^2(\mathbf{X}^t\mathbf{X})^{-1}\right)$$

For a general design matrix X, (X^tX) is no longer diagonal: the loss involves cross products of diff dims, and there is no one-fits-all threshold value.

Standardization

1. We center Y and X so the intercept is not penalized

2. We scale X, so at least when the correlation among features is low, a single lambda value will work well. But if features are already in the same unit (gene expression level, all categorical variables), then one can choose not to center/scale

3. For general X, it's hard to tell which one, standardization or no standardization, will have a better performance

Send standardized data to the algorithm, and obtain

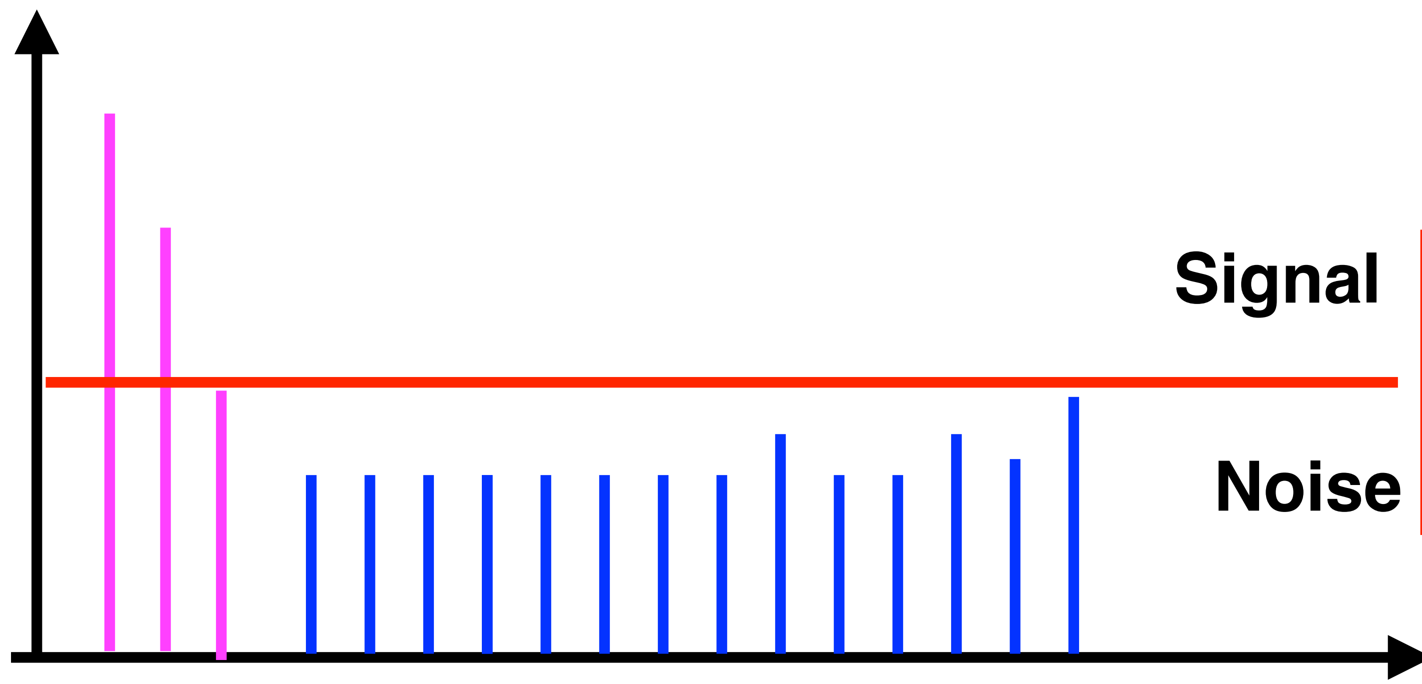
$$\left(\frac{Y - m_y}{se_y}\right) = \hat{\beta}_1 \cdot \left(\frac{X_1 - m_{x,1}}{se_{x,1}}\right) + \cdots + \hat{\beta}_p \cdot \left(\frac{X_p - m_{x,p}}{se_{x,p}}\right)$$

Scale back:

$$Y = \hat{\beta}_0 + \hat{\beta}_1 \frac{se_y}{se_{x,1}} X_1 + \cdots + \hat{\beta}_p \frac{se_y}{se_{x,p}} X_p$$

Lasso + Bootstrap

Large number of noise features will push lambda to be large, so small signals will be killed and also a large bias is introduced to non-zero coefficients (compare Boston2 vs Boston3)



$$\lambda > C'' \sqrt{\frac{\sigma^2 \log p}{n}}$$

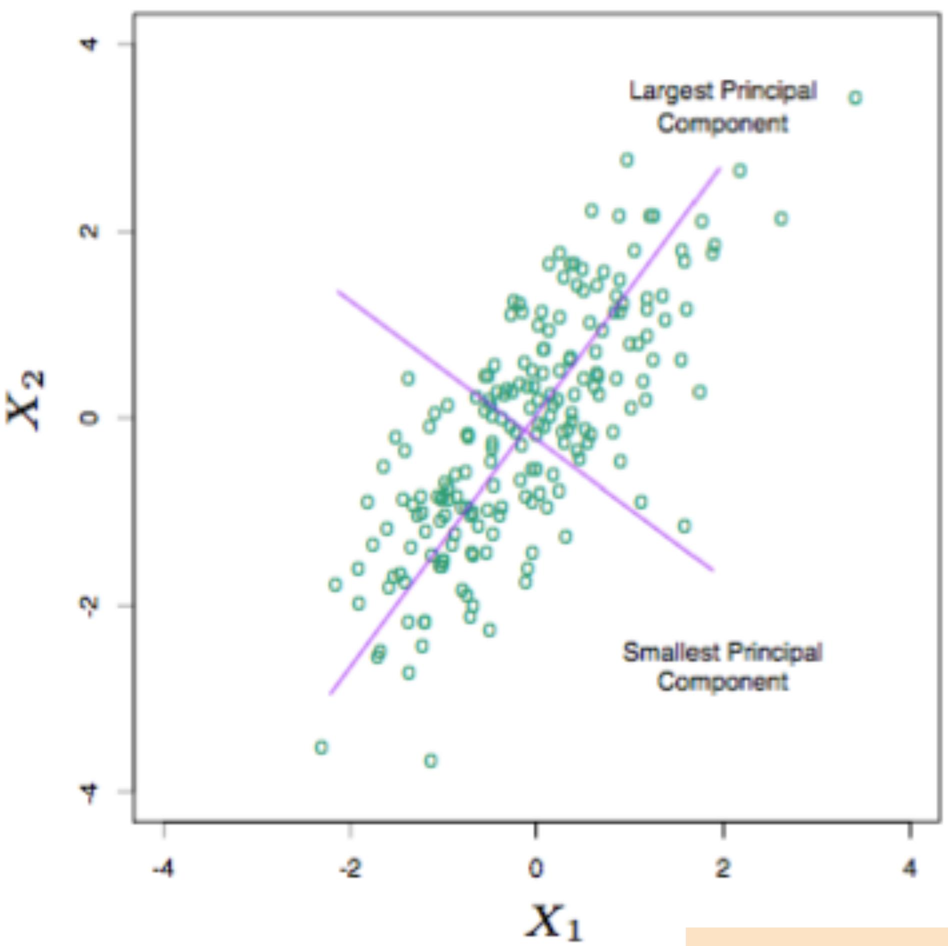
A heuristic approach to removing spurious variables based on reproducibility or consistency: spurious variables seem highly relevant only on this particular training data, so if we run **Lasso** repeatedly on **bootstrap** samples, then spurious variables shouldn't be repeatedly selected by Lasso.

Ridge

$$\mathbf{y} - \mathbf{X}\boldsymbol{\beta} = \mathbf{y} - \mathbf{UDV}\boldsymbol{\beta} = \mathbf{y} - \mathbf{F}\boldsymbol{\alpha}.$$

there is a one-to-one correspondence between $\boldsymbol{\beta}_{p\times 1}$ and $\boldsymbol{\alpha}_{p\times 1}$ and $\|\boldsymbol{\beta}\|^2 = \|\boldsymbol{\alpha}\|^2$. So

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2 \Longleftrightarrow \min_{\boldsymbol{\alpha}\in\mathbb{R}^p} \|\mathbf{y} - \mathbf{F}\boldsymbol{\alpha}\|^2 + \lambda\|\boldsymbol{\alpha}\|^2.$$



F: the new design matrix after rotation. Columns of **F** are projections of the data onto subsequent PC directions

Even if we standardize **X**, columns of **F** still have different variances. So although there is one shrinkage parameter lambda, the shrinkage factors for different columns of **F** are different: alpha_j's are shrunk more and more as j increases

This implies Ridge trusts the first couple of PCs more. Does it make sense to do so? (compare Boston2 vs Boston3)

L2 Penalty

The meaning of “sparsity” changes when the penalty function changes.

AIC/BIC: sparsity = small number of non-zero coefficients (L_0 norm of beta)

Natural, but computationally difficult (known to be NP hard)

Lasso: sparsity = small L_1 norm

Ridge: sparsity = small L_2 norm

L2 Penalty

The meaning of "sparsity" changes when the penalty function changes.

AIC/BIC: sparsity = small number of non-zero coefficients (L_0 norm of beta)

Lasso: sparsity = small L_1 norm

↔

Under mild conditions, these two types of sparsity are the same.

Ridge: sparsity = small L_2 norm

L2 Penalty

The meaning of “sparsity” changes when the penalty function changes.

AIC/BIC: sparsity = small number of non-zero coefficients (L_0 norm of beta)

Lasso: sparsity = small L_1 norm

↔ Under mild conditions, these two types of sparsity are the same.

Ridge: sparsity = small L_2 norm

For correlated features, L0 and L1 tend to pick just the most relevant one, but L2 tends to spread the weights over correlated features.

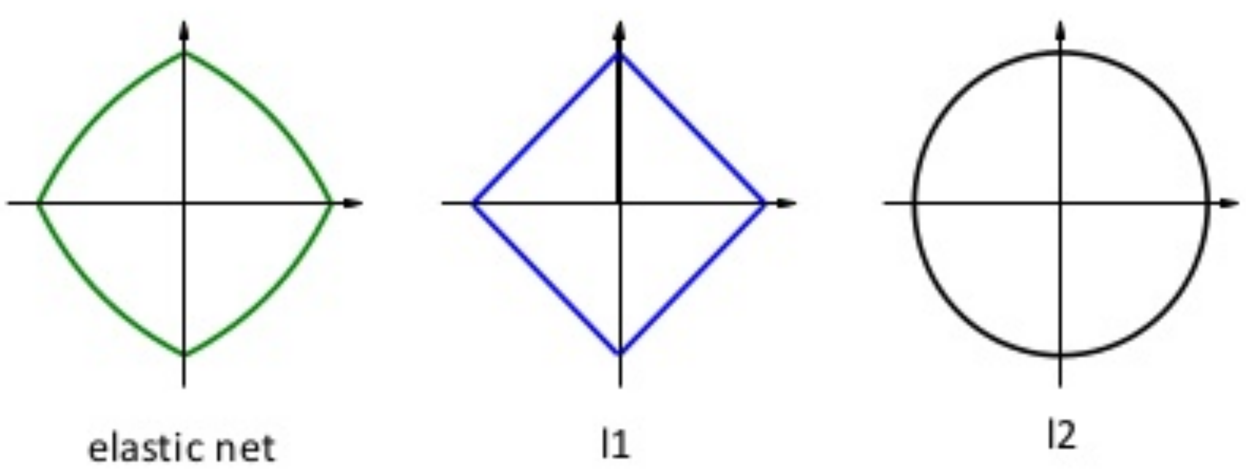
Suppose both X1 and X2 indirectly measure some true predictor variable (e.g., housing price and annual vacation cost as indirect measures of annual income of a household), then it makes sense to use weighted average of these two variables instead of just keeping one in the model.

Other Penalty Choices: Elastic Net

$$\text{Pen}(\boldsymbol{\beta}) = \sum_{j=1}^{p} \left[\frac{1}{2}(1-a)\beta_j^2 + a|\beta_j| \right]$$

- a = 0: Ridge
- a = 1: Lasso
- 0 < a < 1: Elastic Net

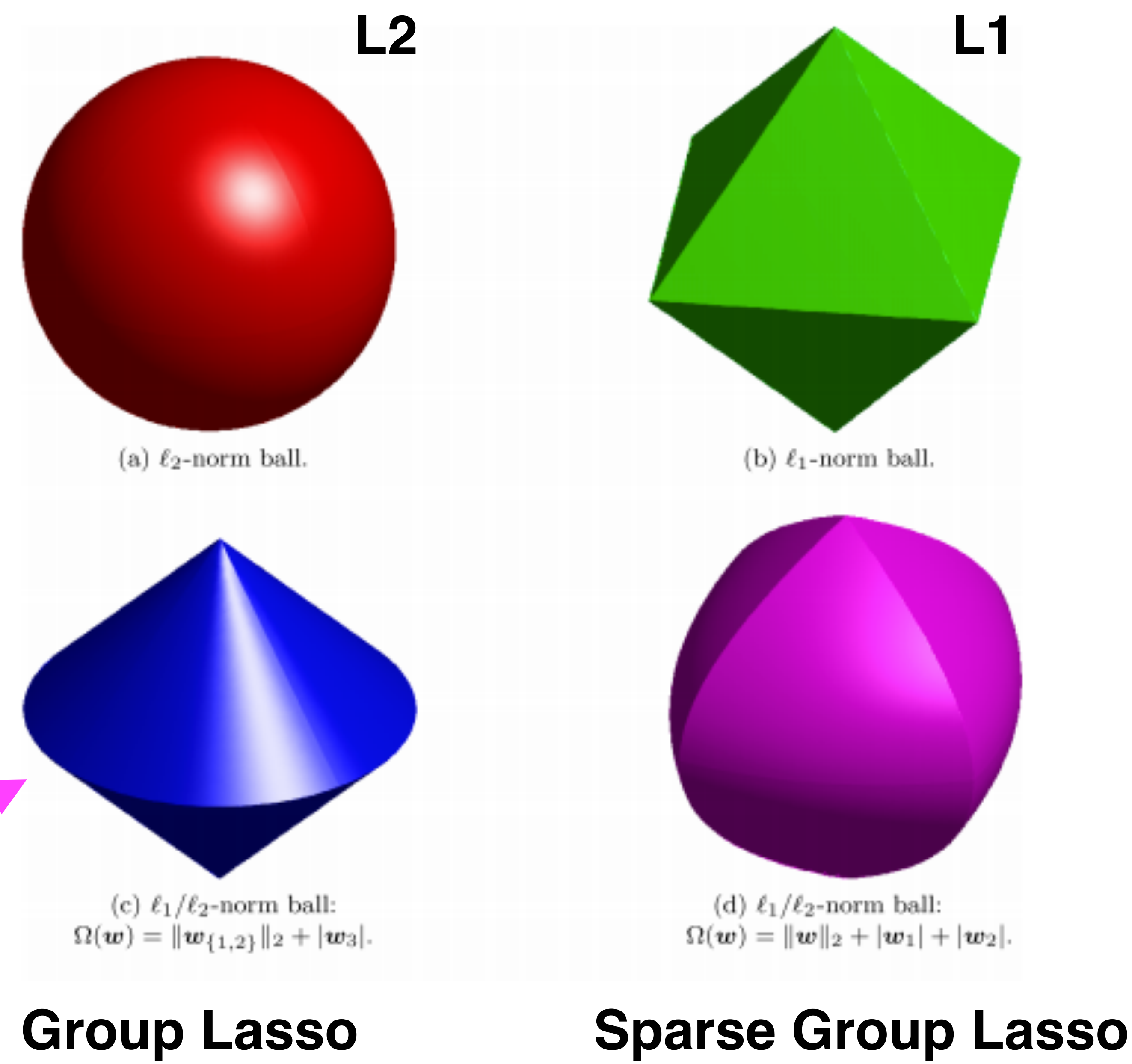
Purpose: Correlated features tend to be selected together.



Other Penalty Choices: Group Lasso

$$\text{Pen}(\boldsymbol{\beta}) = \sqrt{\beta_1^2 + \beta_2^2} + |\beta_3|$$

beta_1 and beta_2 are in the same group, and need be in-or-out together



Ball: slice over z-coordinate
Diamond: slice over x- or y-coordinate