Sparse linear manifolds relating shape to clinical outcome

Professor , Ph.D. Rasmus Larsen

Hven , August 20st, 2009

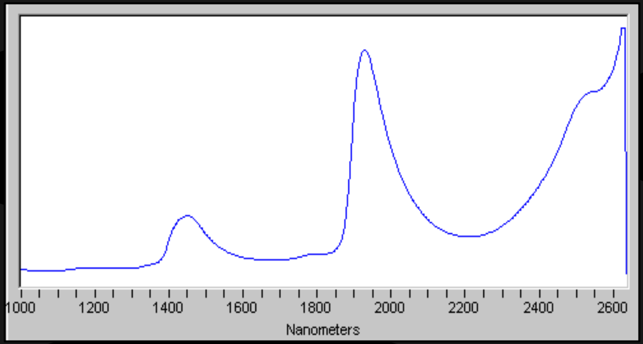
DTU Informatics
Technical University of Denmark

Purpose

- We can extract measurements from the human body with a rapidly increasing spatial, temporal and spectral resolution using modern imaging devices. This is particularly true in the field of biophotonics.
- Typically we have an outcome (e.g. blood-glucose, psoriasis severity) that we want to predict based on a set of features (e.g. IR absorption spectra and derived features)
- Having observed the outcome and features in a set of objects (a training set of data) we want to build a model that will allow us to predcit the outcome of unseen objects

Model

- Outcome: Y
- Features: $X = (X_1, X_2, \dots, X_p)$
  - sampled spectrum
  - set of spectra in an image
  - .
  - .
  - .



- Model: $Y = f(X) + \varepsilon$

Two approaches

- The linear model:
  - Global

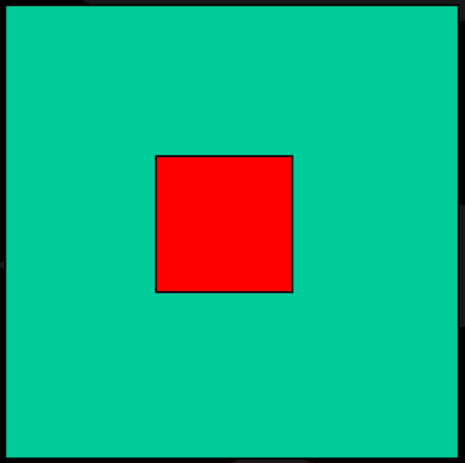
$$Y = X^T \hat{\beta}$$

- Nearest Neighbour model:
  - Local

$$Y(x) = \frac{1}{k} \sum_{i \in N_k(x)} y_i$$

Curse of dimensionality I

- Consider inputs uniformly distributed over a p-dimensional hypercube [0,1]x[0,1]x…x[0,1]
- 2-dim hypercube:



- For the red neighbourhood to cover a fraction r of the observation it should have side length $s = r^{1/p}$
- For r=1% we get for p=2: s = 0.1, for p=10: s=0.63

Curse of dimensionality II

- For practical size problems locality in high dimensional spaces does not exist
- The majority of observations lie near the edges of the training sample, in the 10 dimensional hypercube, only 1% of the observations lie in a central hypercube of sidelength 0.63 – we must extrapolate our fits
- In high dimensions the linear model is popular!

Linear Regression

$$f(X) = \beta_0 + \sum_{i=1}^{p} X_j \beta_j$$

Training set

$$(x_i, y_i),\ i = 1, 2, \ldots, N$$

$$x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$$

$$\begin{aligned} \mathrm{RSS}(\beta) &= \sum_{i=1}^{N} (y_i - f(x_i))^2 \\ &= \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{i=1}^{p} x_{ij}\beta_j\right)^2 \end{aligned}$$

Linear Regression – matrix-vector notation

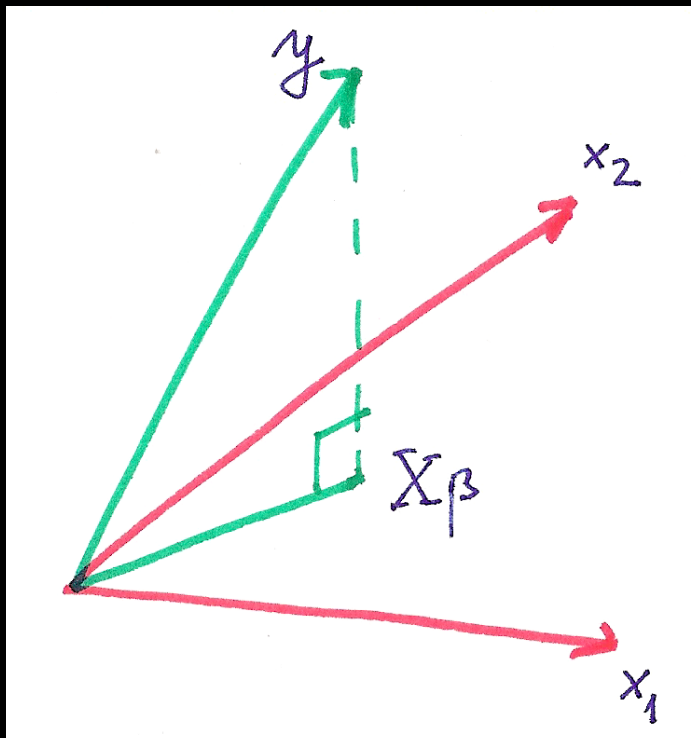
$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & x_{N2} & \cdots & x_{Np} \end{bmatrix}$$

$$\boldsymbol{y} = (y_1, y_2, \ldots, y_N)$$

$$\mathrm{RSS} = (\boldsymbol{y} - \boldsymbol{X}\beta)^T(\boldsymbol{y} - \boldsymbol{X}\beta)$$

The predictor $\boldsymbol{X}\beta$ belongs to the column-space of **X**

Linear regression - geometrically



Choose β such that the residual is orthogonal to ***X***, i.e.

$$\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\beta) = 0$$

$$\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Linear regression – correlated inputs

$$E(\hat{\beta}) = E((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\beta = \beta$$
$$V(\hat{\beta}) = V((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^2$$

X^TX/N is the ML estimator for the covariance matrix of the inputs

Consider 3 inputs X_1, X_2, X_3 with covariance

$$S = \begin{bmatrix} 1 & 0.99 & 0 \\ 0.99 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad S^{-1} = \begin{bmatrix} 50.25 & -49.75 & 0 \\ -49.75 & 50.25 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

The parameters of the correlated inputs have high variance and high correlation

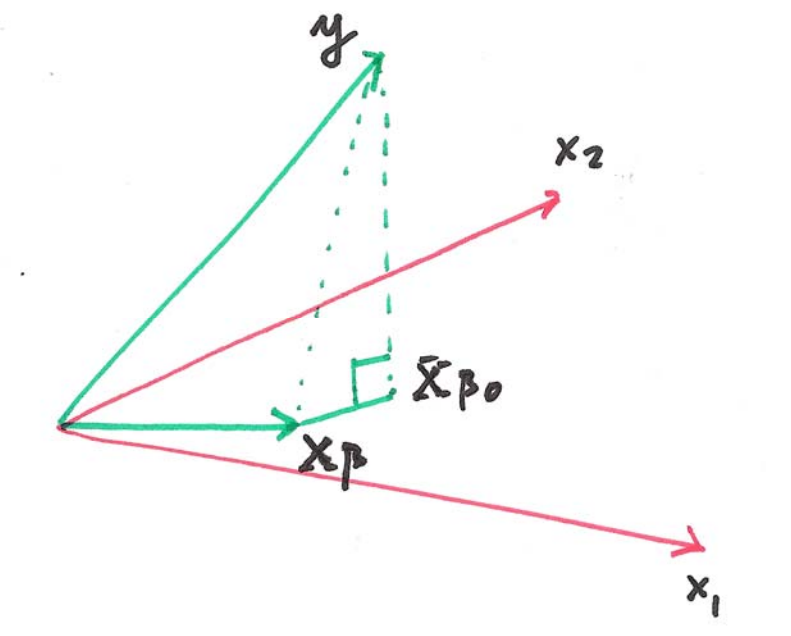
Linear regression – regularization

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \right\}, \text{ s.t. } \quad \sum_{j=1}^{N} \beta_j^2 \leq s$$

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{N} \beta_j^2 \right\}$$

$$\text{PRSS}(\lambda) = (\boldsymbol{y} - \boldsymbol{X}\beta)^T(\boldsymbol{y} - \boldsymbol{X}\beta) + \lambda\beta^T\beta$$

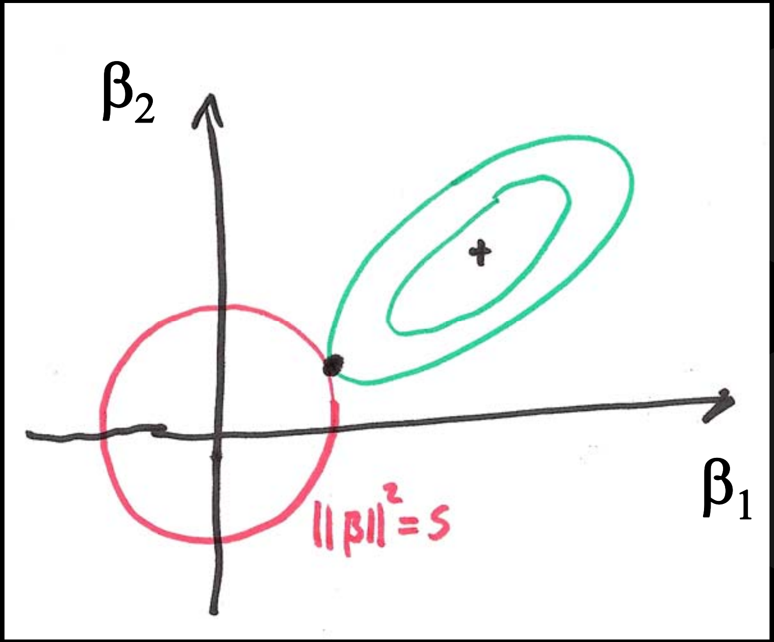
Ridge regression- geometrically



$$\|y - X\beta\|^2 = \|y - X\beta_0\|^2 + \|X\beta - X\beta_0\|^2$$

$$\|X\beta - X\beta_0\|^2 = (\beta - \beta_0)^T X^T X (\beta - \beta_0)$$

Ridge regression – geometrically II



$$\|X\beta - X\beta_0\|^2 = (\beta - \beta_0)^T X^T X(\beta - \beta_0)$$

Correllated inputs again

3 inputs X_1, X_2, X_3 with covariance $S = \begin{bmatrix} 1 & 0.99 & 0 \\ 0.99 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

$Y = X_1 + X_2 + X_3 + \varepsilon, \quad \varepsilon$ in N(0,1)

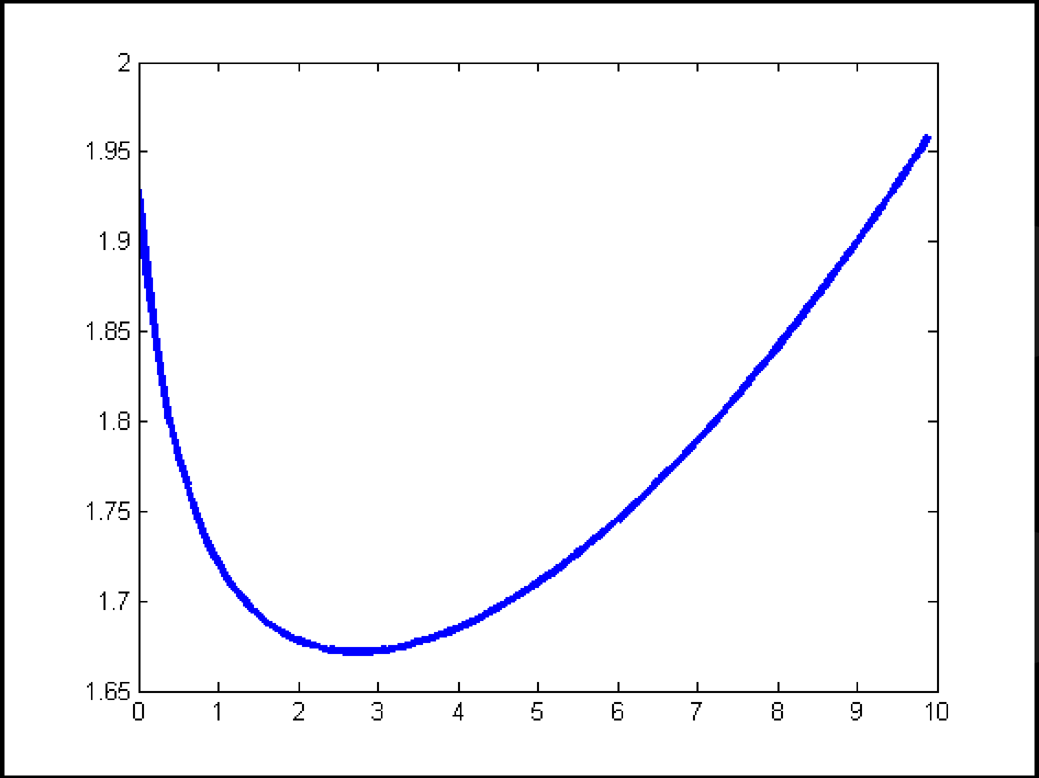
N=100 , in 1000 trials

$$Cov(\beta) = \frac{1}{100}\begin{bmatrix} 55 & -55 & 0.56 \\ -55 & 55 & -0.56 \\ 0.56 & -.56 & 1.08 \end{bmatrix}$$

Ordinary LS

β = [-0.01 0.97 1.03 1.00

Correllated inputs again – ridge regression



(λ, RSS)

Ridge (λ=2.4)

$$Cov(\beta) = \frac{1}{100}\begin{bmatrix} 4.4 & -3.8 & 0.05 \\ -3.8 & 4.3 & -0.03 \\ 0.05 & -.03 & 1.02 \end{bmatrix}$$

β = [-0.00 0.99 0.98 0.98

We want

- Prediction accuracy
- Easy Intepretation (simple model)

We tried

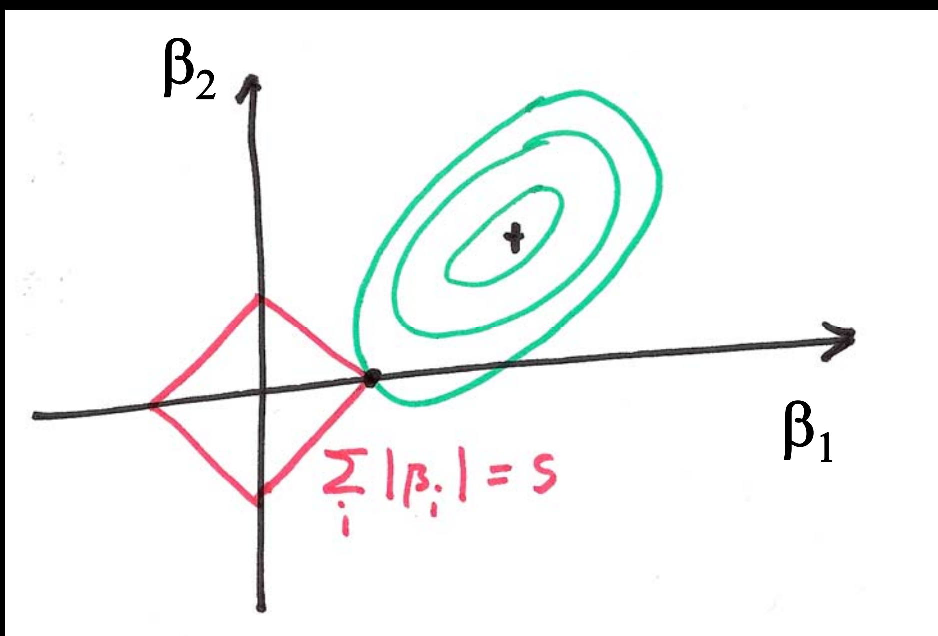
- Regularization (ridge regression)

And got

- Prediction accuracy

Prediction accuracy and easy interpretation

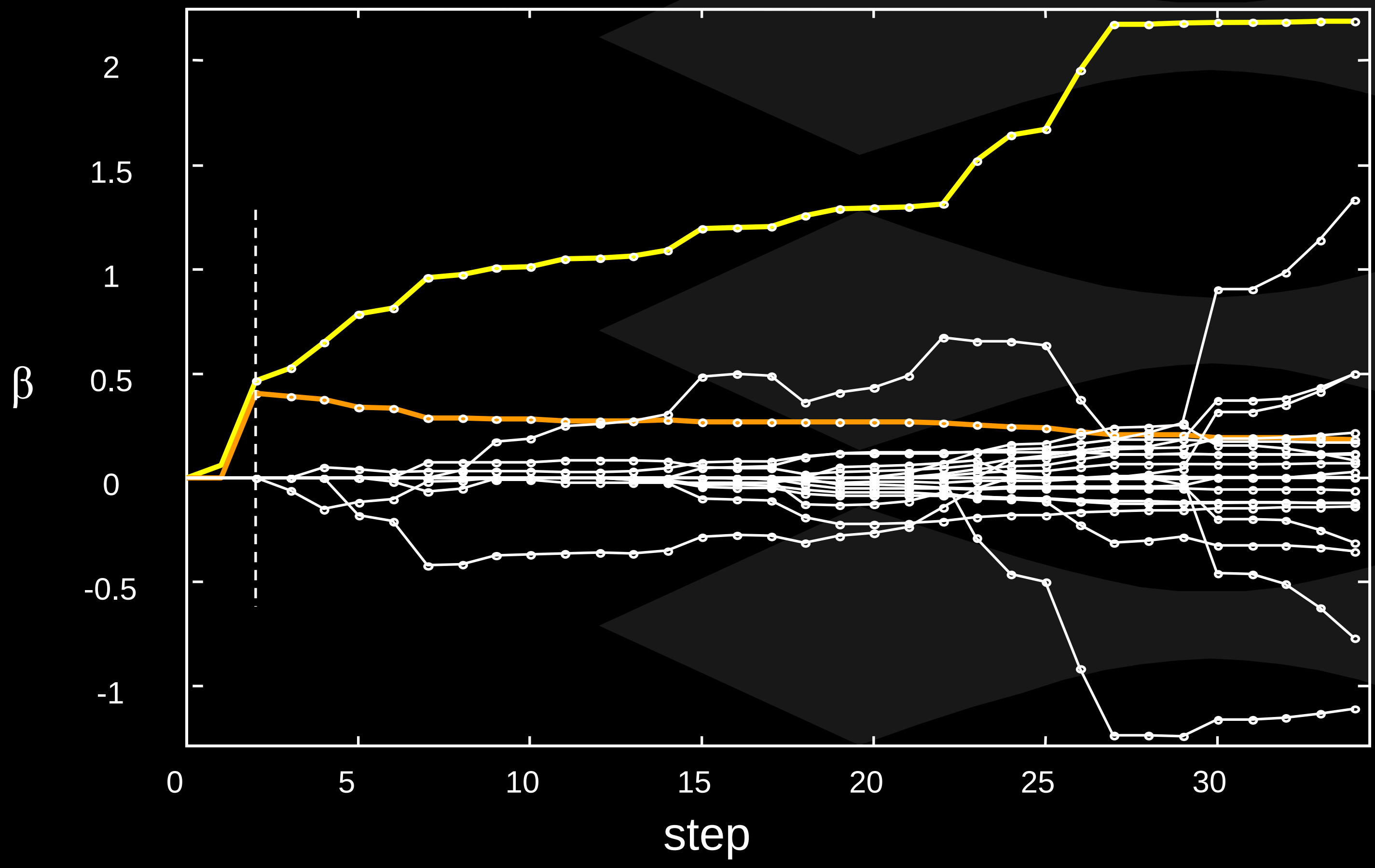
$$\hat{\beta}^{\text{lasso}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \right\}, \text{ s.t. } \quad \sum_{j=1}^{N} |\beta_j| \le s$$



many β's will tend to be 0

Regularization and subset selection

LASSO Model Selection

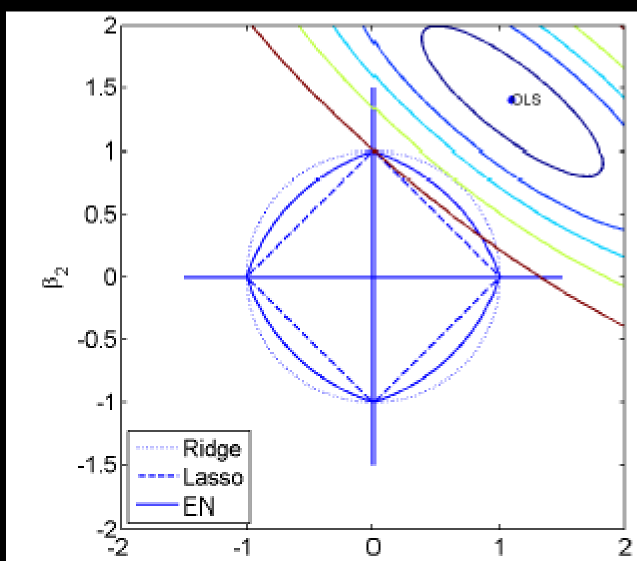


LASSO

- Prediction accuracy ☺
- Easy interpretation ☺
- Computations ☺
- p<N ☹
- Tend to select one of a group of correlated inputs ☹

LARS-EN – elastic net

$$\hat{\beta} = \text{argmin}_{\beta}\{\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2^2\}$$



- Prediction accuracy ☺
- Easy interpretation ☺
- Computations ☺
- Handles p>N ☺
- Tend to select groups of correlated inputs ☺

LARS-EN – elastic net

$$\hat{\beta} = \text{argmin}_\beta \{ \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2 \}$$

Ridge to OLS

$$\|\boldsymbol{y} - \boldsymbol{X}\beta\|^2 + \lambda_2 \|\beta\|^2 = \| \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{bmatrix} - \begin{bmatrix} \boldsymbol{X} \\ \lambda_2 \boldsymbol{I} \end{bmatrix} \beta \|^2$$

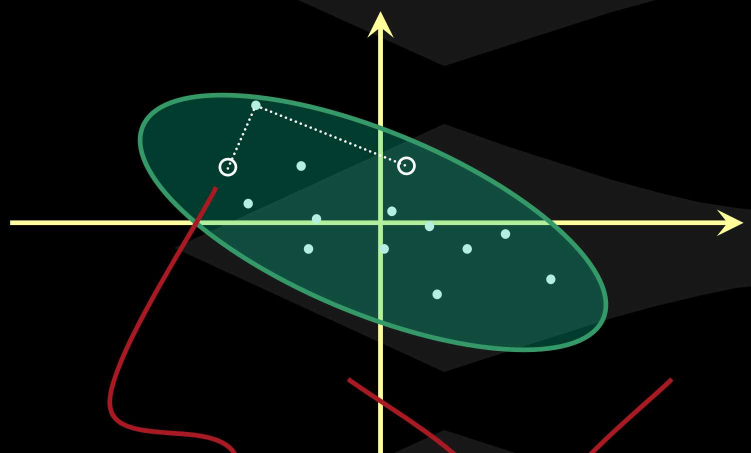
LASSO problem remains!

Handling CoD

- Regularization
- Variable selection
- Subspace projection

Principal Components

- By rotating the coordinate system, the axes point in directions of maximum variance



Coordinates of data on new axes are in the *scores matrix*

The new axes are in the *loading matrix*

$$\mathbf{S} = \mathbf{XL}$$

data matrix