

STK2100- Machine Learning and Statistical Methods for Prediction and Classification

Spring 2026

Linear models for regression – Part II

Covers Sections 3.4 (except 3.4.4) and 3.5.1

Ingrid Hobæk Haff
Department of Mathematics
University of Oslo

Shrinkage methods

- Removing some of the covariates from a model, makes it more interpretable, and may reduce the prediction error
- However, it is a discrete process (either remove or keep), with possibly high variability
- An alternative is shrinkage methods, that are more continuous
- The idea is then to shrink the values of the regression coefficients towards 0

Ridge regression

- In **ridge regression**, the least squares criterion is replaced by

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

- The penalty $\lambda \geq 0$ is a **complexity parameter**
- It may be shown that the following problem formulation is equivalent:

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

with a one-to-one correspondance between λ and t

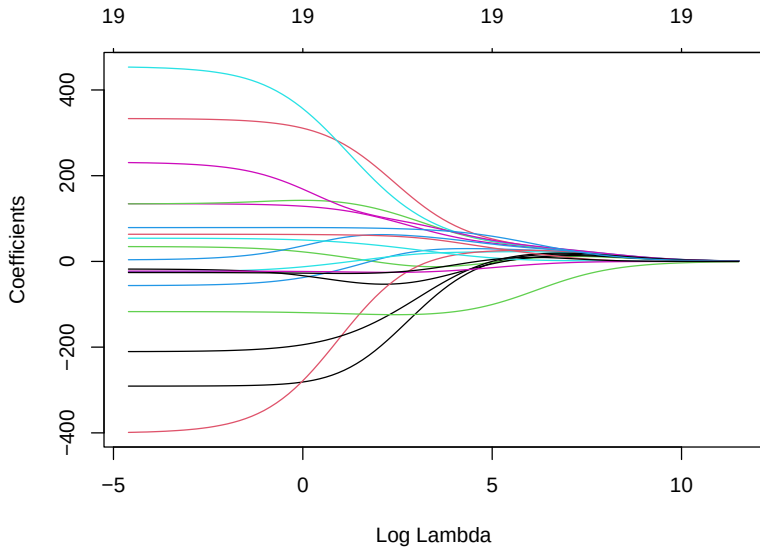
Ridge regression

- However, this is not invariant with respect to scaling of x s, so these should be **standardised**:

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}}$$

- Also note that the penalty does not apply to β_0 , as this would make the procedure depend on the chosen origin for Y
- When the covariates are standardised, we can estimate β_0 by $\hat{\beta}_0 = \bar{y} = \frac{1}{N} \sum_{i=1}^N y_i$ in a first step
- The remaining coefficients are then estimated by ridge regression without intercept

Illustration of ridge regression on the Hitters data



Derivation of $\hat{\beta}$ in ridge regression

- Since we treat β_0 in a separate step, define $\beta = (\beta_1, \dots, \beta_p)^T$
- We now wish to minimise the following function

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

with respect to β_0 and β

- We assume in the following that the x s are standardised, so that $\sum_{i=1}^n x_{ij} = 0$
- Differentiating with respect to β_0 , we obtain

$$\begin{aligned} \frac{\partial}{\partial \beta_0} D_{\text{ridge}}(\beta_0, \beta) &= -2 \sum_{i=1}^N \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right) \\ &= -2 \sum_{i=1}^N (y_i - \beta_0) + 2 \sum_{j=1}^p \beta_j \sum_{i=1}^N x_{ij} = -2 \sum_{i=1}^N (y_i - \beta_0) \end{aligned}$$

Derivation of $\hat{\beta}$ in ridge regression

- Differentiating with respect to β gives

$$\frac{\partial}{\partial \beta_\ell} D_{\text{ridge}}(\beta_0, \beta) = -2 \sum_{i=1}^N \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right) x_{i\ell} + 2\lambda \beta_\ell, \ell \geq 1$$

- Equating the derivatives to 0 gives the **normal equations**

$$\hat{\beta}_0^{\text{ridge}} = \frac{1}{N} \sum_{i=1}^N y_i = \bar{y}$$

$$\sum_{j=1}^p \hat{\beta}_j^{\text{ridge}} \sum_{i=1}^N x_{ij} x_{i\ell} + \lambda \hat{\beta}_\ell^{\text{ridge}} = \sum_{i=1}^N (y_i - \bar{y}) x_{i\ell} = \sum_{i=1}^N y_i x_{i\ell}, \quad \ell = 1, \dots, p$$

- We can re-write on matrix form, where $\mathbf{X}$ now does not contain a first column of 1s, and $\mathbf{1}$ is an $N \times 1$ vector of 1s

$$\begin{aligned} D_{\text{ridge}}(\beta_0, \beta) &= (\mathbf{y} - \beta_0 \mathbf{1} - \mathbf{X}\beta)^T (\mathbf{y} - \beta_0 \mathbf{1} - \mathbf{X}\beta) + \lambda \beta^T \beta \\ &= (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) + \lambda \beta^T \beta - 2\beta_0 \mathbf{1}^T (\mathbf{y} - \mathbf{X}\beta) + \beta_0^2 \mathbf{1}^T \mathbf{1} \\ &= \mathbf{y}^T \mathbf{y} - 2\mathbf{y}^T \mathbf{X}\beta + \beta^T (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}) \beta - 2\beta_0 \mathbf{1}^T \mathbf{y} + N\beta_0^2 \end{aligned}$$

Derivation of $\hat{\beta}$ in ridge regression

- Then we may re-write

$$\frac{\partial}{\partial \beta} D_{\text{ridge}}(\beta_0, \beta) = -2\mathbf{X}^T \mathbf{y} + 2(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})\beta,$$

which results in the equations

$$(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})\hat{\beta}^{\text{ridge}} = \mathbf{X}^T \mathbf{y}$$

- The ridge estimator of β is then given by

$$\hat{\beta}^{\text{ridge}} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$$

Ridge for orthogonal covariates

- Assume that the covariate vectors are orthogonal, so that $\sum_{i=1}^N x_{ij}x_{il} = 0$ for $l \neq j$
- Then $\mathbf{X}^T \mathbf{X} = \mathbf{I}$ and $(\mathbf{X}^T \mathbf{X})^{-1} = \mathbf{I}$
- We get

$$\begin{aligned}\hat{\beta}^{ridge} &= (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y} \\ &= (\mathbf{I} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y} \\ &= \frac{1}{1 + \lambda} \mathbf{I} \mathbf{X}^T \mathbf{y} \\ &= \frac{1}{1 + \lambda} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = \frac{1}{1 + \lambda} \hat{\beta}^{LS},\end{aligned}$$

where $\hat{\beta}^{LS}$ is the least squares estimate of β

- Hence, ridge penalisation scales down $\hat{\beta}^{LS}$
- Note that this does not apply to the intercept, which is derived separately

Why does Ridge regression work?

- Ridge regresssion has been shown to work well in many applications
- Assuming orthogonal x s, we have $\hat{\beta}^{ridge} = \frac{1}{1+\lambda} \hat{\beta}^{LS}$
- This gives

$$\begin{aligned} E\left(\hat{\beta}^{ridge}\right) &= E\left(\frac{1}{1+\lambda} \hat{\beta}^{LS}\right) = \frac{1}{1+\lambda} \beta \\ \text{Var}\left(\hat{\beta}^{ridge}\right) &= \left(\frac{1}{1+\lambda}\right)^2 \text{Var}\left(\hat{\beta}^{LS}\right) \end{aligned}$$

- Hence, ridge penalisation introduces bias, but reduces the variance

Choice of λ

- The λ should be chosen such that it maximises the prediction accuracy
- We know that $RSS = \sum_{i=1}^N (y_i - \hat{y}_i)^2$ is too optimistic, as we would then be using the same data for training and validation
- This would result in $\lambda = 0$ being optimal
- Ideally, we would use a separate validation set
- However, this requires a large amount of data
- A good alternative, that uses the data effectively, is cross-validation
- Typically, K-fold cross-validation with $K = 5$ or $K = 10$ is preferred

Choice of λ with cross-validation

The procedure is as follows

- Specify a sequence $\lambda_1, \dots, \lambda_m$ of λ values
- Divide the data into K groups
- For $l = 1, \dots, m$
 - For $k = 1, \dots, K$
 - Fit the model with ridge penalisation λ_l to all data, except the $(y_i, \mathbf{x}_i)$ s in group k to obtain $(\hat{\beta}_{0,l}^{-k}, \hat{\beta}_l^{-k})$
 - Predict the y_i s in group k by $\hat{y}_{i,l} = \hat{\beta}_{0,l}^{-k} + \mathbf{x}_i^T \hat{\beta}_l^{-k}$
 - Estimate the prediction error by

$$\widehat{\text{Err}}_{CVK}(\lambda_l) = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_{i,l})^2$$

- Select the λ_l with the lowest $\widehat{\text{Err}}_{CVK}(\lambda_l)$

This is quite computer intensive, but effective algorithms exist

Examples

Example: Prostate data

See R code

Example: Hitters data

See R code

Degrees of freedom in ridge regression

- The degrees of freedom are usually related to the number of parameters for specifying $E(Y|\mathbf{x})$
- For linear models fitted with least squares, this is $df = p + 1$
- In ridge regression, it is more complex, since the $p + 1$ β_j s are no longer chosen freely
- Assume that the fitted values can be expressed as $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$, for some matrix $\mathbf{H}$
- Then, define the **degrees of freedom** by

$$df = tr(\mathbf{H}) = \sum_{i=1}^N H_{ii}$$

- In least squares estimation, we have $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, so that (recall that $tr(\mathbf{AB}) = tr(\mathbf{BA})$)

$$tr(\mathbf{H}) = tr(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T) = tr((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}) = tr(\mathbf{I}_{p+1}) = p + 1,$$

as it should

Degrees of freedom for ridge regression

- In ridge regression, we have $\hat{\beta}_0^{ridge} = \bar{y} = \frac{1}{N} \mathbf{1}^T \mathbf{y}$ and $\hat{\beta}^{ridge} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$, so that

$$\begin{aligned}\hat{\mathbf{y}} &= \hat{\beta}_0^{ridge} \mathbf{1} + \mathbf{X} \hat{\beta}^{ridge} \\ &= \frac{1}{N} \mathbf{1} \mathbf{1}^T \mathbf{y} + \mathbf{X} (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y} \\ &= \underbrace{\left(\frac{1}{N} \mathbf{1} \mathbf{1}^T + \mathbf{X} (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \right)}_{=\mathbf{H}_\lambda} \mathbf{y} = \mathbf{H}_\lambda \mathbf{y}\end{aligned}$$

- Hence, we get

$$\begin{aligned}df(\lambda) = \text{tr}(\mathbf{H}_\lambda) &= \text{tr} \left(\frac{1}{N} \mathbf{1} \mathbf{1}^T + \mathbf{X} (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \right) \\ &= \frac{1}{N} \text{tr}(\mathbf{1} \mathbf{1}^T) + \text{tr}(\mathbf{X} (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T) \\ &= 1 + \text{tr}(\mathbf{X} (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T)\end{aligned}$$

Degrees of freedom for ridge regression

- It may be shown that

$$\text{tr} \left(\mathbf{X}(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \right) = \sum_{j=1}^p \frac{d_j^2}{d_j^2 + \lambda}$$

where d_j^2 are the eigenvalues of $\mathbf{X}^T \mathbf{X}$

- This gives

$$df(\lambda) = 1 + \sum_{j=1}^p \frac{d_j^2}{d_j^2 + \lambda}$$

- Note that Eq (3.50) in the textbook does not take the intercept into account
- $\lambda = 0$ gives $df = p + 1$, equal to least squares
- Further, df decreases when λ increases
- df can be used for model selection, for instance in information criteria, such as AIC and BIC

Lasso regression

- In **Lasso regression**, the least squares criterion is replaced by

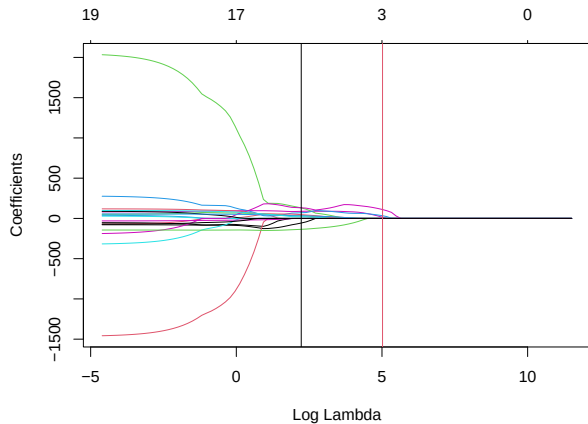
$$\hat{\beta}^{Lasso} = \arg \min_{\beta} \left\{ \sum_{i=1}^N (y_i - \beta_0 - \mathbf{x}_i^T \beta)^2 + \lambda \sum_{j=1}^p |\beta_j| \right\}$$

- It may be shown that this is equivalent to

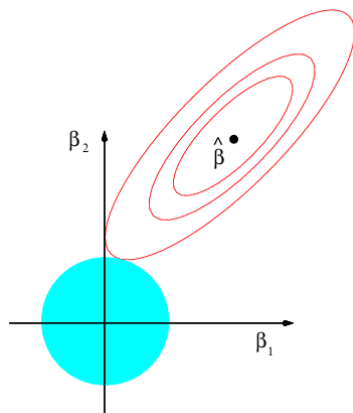
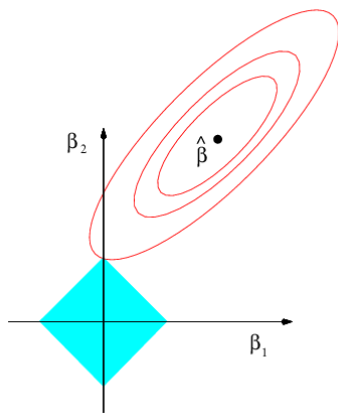
$$\hat{\beta}^{Lasso} = \arg \min_{\beta} \sum_{i=1}^n (y_i - \beta_0 - \mathbf{x}_i^T \beta)^2 \text{ subject to } \sum_{j=1}^p |\beta_j| \leq t,$$

- There is no analytic solution, it is nonlinear in β , and may be solved by quadratic programming
- $\lambda = 0$ gives the least squares estimate and $\lambda = \infty$ results in $\hat{\beta}^{Lasso} = \mathbf{0}$
- Typically, some $\hat{\beta}_j$ s become exactly zero
- Hence, Lasso regression also results in variable selection
- The choice of λ is typically done by cross-validation

Illustration of Lasso regression on the Hitters data



Geometric interpretation



Derivation of $\hat{\beta}$ in Lasso regression

- As in ridge regression, the Lasso solution depends on the scale of the covariates
- The covariates should therefore be standardised, which also means that $\sum_{i=1}^n x_{ij} = 0$
- We now wish to minimise the following function

$$D_{Lasso}(\beta_0, \beta) = \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p |\beta_j|$$

with respect to β_0 and β

- Then, we have

$$\frac{\partial D_{Lasso}(\beta_0, \beta)}{\partial \beta_0} = -2 \sum_{i=1}^n (y_i - \beta_0 - \mathbf{x}_i^T \beta) = -2 \sum_{i=1}^n (y_i - \beta_0)$$

- Solving $\partial D_{Lasso}(\beta_0, \beta) / \partial \beta_0 = 0$ for β_0 , we obtain $\hat{\beta}_0^{Lasso} = \bar{y}$
- Hence, $\hat{\beta}_0^{Lasso}$ can be estimated in a separate step

Lasso regression with orthogonal covariates

- Assume now that the covariates are orthogonal, so that $\mathbf{X}^T \mathbf{X} = \mathbf{I}$, and define $\tilde{y}_i = y_i - \bar{y}$
- Then, we have

$$\begin{aligned} D_{\text{Lasso}}(\beta_0, \beta) &= \sum_{i=1}^N (\tilde{y}_i - \mathbf{x}_i^T \beta)^2 + \lambda \sum_{j=1}^p |\beta_j| \\ &= \sum_{i=1}^N \tilde{y}_i^2 - 2 \sum_{i=1}^N \tilde{y}_i \mathbf{x}_i^T \beta + \sum_{i=1}^N \beta^T \mathbf{x}_i^T \mathbf{x}_i \beta + \lambda \sum_{j=1}^p |\beta_j| \\ &= \sum_{i=1}^N \tilde{y}_i^2 - 2 \sum_{i=1}^N \tilde{y}_i \sum_{j=1}^p x_{ij} \beta_j + \beta^T \underbrace{\mathbf{X}^T \mathbf{X}}_{=\mathbf{I}} \beta + \lambda \sum_{j=1}^p |\beta_j| \\ &= \sum_{i=1}^N \tilde{y}_i^2 + \sum_{j=1}^p \left(\beta_j^2 + \lambda |\beta_j| - 2 \sum_{i=1}^N \tilde{y}_i x_{ij} \beta_j \right) \\ &= \sum_{i=1}^N \tilde{y}_i^2 + \sum_{j=1}^p g_j(\beta_j) \end{aligned}$$

- Thus $D_{\text{Lasso}}(\beta_0, \beta)$ can be optimized for each β_j separately

Lasso regression with orthogonal covariates

- Minimising $D_{\text{Lasso}}(\beta_0, \beta)$ w.r.t. β_j is equivalent to minimising $g_j(\beta_j) = \beta_j^2 + \lambda|\beta_j| - 2\beta_j \sum_{i=1}^N \tilde{y}_i x_{ij}$ w.r.t. β_j
- If we assume that $\beta_j \geq 0$, we have

$$\frac{\partial g_j(\beta_j)}{\partial \beta_j} = 2\beta_j + \lambda - 2 \sum_{i=1}^N \tilde{y}_i x_{ij}$$

- The optimal value is then $\hat{\beta}_j = \sum_{i=1}^N \tilde{y}_i x_{ij} - \frac{1}{2}\lambda$ if $\sum_{i=1}^N \tilde{y}_i x_{ij} > \frac{1}{2}\lambda$, and 0 otherwise
- Similarly if $\beta_j < 0$. Then

$$\frac{\partial g_j(\beta_j)}{\partial \beta_j} = 2\beta_j - \lambda - 2 \sum_{i=1}^N \tilde{y}_i x_{ij}$$

- The optimal value is then $\hat{\beta}_j = \sum_{i=1}^N \tilde{y}_i x_{ij} + \frac{1}{2}\lambda$ if $\sum_{i=1}^N \tilde{y}_i x_{ij} < -\frac{1}{2}\lambda$, and 0 otherwise

Lasso regression with orthogonal covariates

- It can be shown that when the covariates are orthogonal, the least squares estimates are

$$\hat{\beta}_0^{LS} = \bar{y} \quad \text{and} \quad \hat{\beta}_j^{LS} = \sum_{i=1}^N \tilde{y}_i x_{ij}, \quad j = 1, \dots, p$$

- Hence, for $j = 1, \dots, p$, we obtain

$$\begin{aligned} \hat{\beta}_j^{Lasso} &= \begin{cases} \sum_{i=1}^N x_{ij} \tilde{y}_i - \frac{1}{2} \lambda & \text{if } \sum_{i=1}^N x_{ij} y_i > \frac{1}{2} \lambda \\ \sum_{i=1}^N x_{ij} \tilde{y}_i + \frac{1}{2} \lambda & \text{if } \sum_{i=1}^N x_{ij} y_i < -\frac{1}{2} \lambda \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} \hat{\beta}_j^{LS} - \frac{1}{2} \lambda & \text{if } \hat{\beta}_j^{LS} > \frac{1}{2} \lambda \\ \hat{\beta}_j^{LS} + \frac{1}{2} \lambda & \text{if } \hat{\beta}_j^{LS} < -\frac{1}{2} \lambda \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Example: Prostate data

See R code

Example: Hitters data

See R code

Comparison between shrinkage and best subset selection

- Ridge:

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

- Lasso:

$$\hat{\beta}^{\text{Lasso}} = \arg \min_{\beta} \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j)^2 \quad \text{subject to} \quad \sum_{j=1}^p |\beta_j| \leq t$$

- Best subset selection:

$$\hat{\beta}^{\text{BSS}} = \arg \min_{\beta} \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j)^2 \quad \text{subject to} \quad \sum_{j=1}^p I(\beta_j \neq 0) \leq t,$$

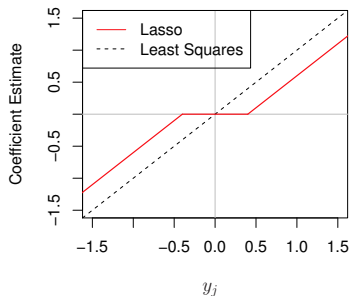
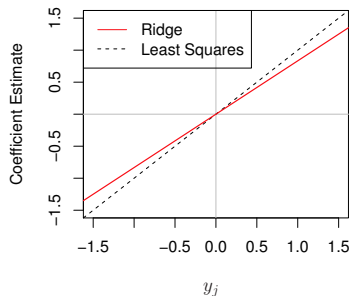
which is equivalent to (using $0^0 = 0$):

$$\hat{\beta}^{\text{BSS}} = \arg \min_{\beta} \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j)^2 \quad \text{subject to} \quad \sum_{j=1}^p |\beta_j|^0 \leq t$$

- In all cases, we need to specify t (or λ)

Comparison between shrinkage and best subset selection

- Ridge does only shrinkage
- Variable selection does only removal of covariates
- Lasso does both removal of covariates and shrinkage



Elastic net

- The **elastic net** is a compromise between ridge and Lasso regression, having both methods as special cases
- The parameters are then estimated by minimising the criterion

$$D_{\text{elast}}(\beta_0, \boldsymbol{\beta}) = \sum_{i=1}^n (y_i - \beta_0 - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda \left(\alpha \sum_{j=1}^p \beta_j^2 + (1 - \alpha) \sum_{j=1}^p |\beta_j| \right)$$

- This requires the specification of the additional tuning parameter α
- The routine `glmnet` in R fits elastic net models, as well as the special cases ridge and Lasso regression

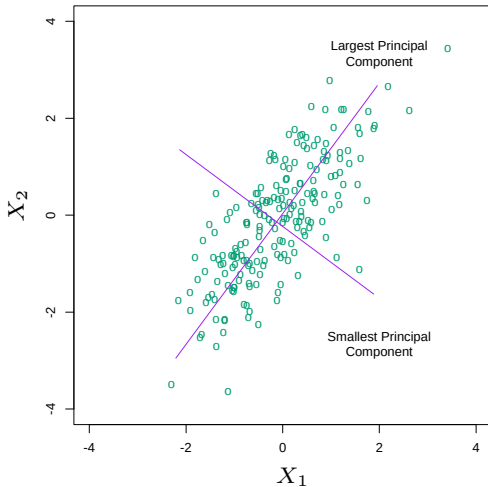
Principal components

- Assume again that p is large, so that we want to reduce the number of x s in the model, while keeping most of the information in the x s
- Hence, we specify new covariates $\mathbf{z}_i = (z_{i1}, \dots, z_{im})$, with $m < p$, which are linear functions of the x_{ij} s

$$z_{il} = \sum_{j=1}^p a_{lj} x_{ij}, \quad l = 1, \dots, m$$

- Then, we use the $\mathbf{z}_i$ s in the regression model, instead of the $\mathbf{x}_i$ s
- Further, we want to construct the $\mathbf{z}_i$ s without using the $\mathbf{y}_i$ s
- The solution is to choose the linear combinations with the largest variability, as they give the most information about the $\mathbf{x}_i$ s
- They are called the **principal components**

Illustration of principal components



Principal components

- Let $\mathbf{S}$ be the **sample covariance matrix**, given by

$$\mathbf{S} = \frac{1}{N} \sum_{i=1}^N (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

- Further, define $\mathbf{a}_1 = (a_{1,1}, \dots, a_{1,p})$ by

$$\mathbf{a}_1 = \arg \max_{\mathbf{a}} \mathbf{a}^T \mathbf{S} \mathbf{a} \quad \text{subject to} \quad \mathbf{a}^T \mathbf{a} = 1$$

- The solution $\mathbf{a}_1$ is the **eigenvector** corresponding to the **largest eigenvalue** of $\mathbf{S}$

Principal components

- Next, define $\mathbf{a}_2 = (a_{2,1}, \dots, a_{2,p})$ by

$$\mathbf{a}_2 = \arg \max_{\mathbf{a}} \mathbf{a}^T \mathbf{S} \mathbf{a} \quad \text{subject to} \quad \mathbf{a}^T \mathbf{a} = 1, \mathbf{a}^T \mathbf{a}_1 = 0$$

- The solution $\mathbf{a}_2$ is the **eigenvector** corresponding to the **second largest eigenvalue** of $\mathbf{S}$
- Then, we continue in the same way for $l = 3, \dots, m$
- The principal components should only be computed for the quantitative covariates, whereas the qualitative covariates should be left as they are

Prostate data

See R code

Hitters data

See R code