Lecture 5: Classification and dimension reduction

Felix Held, Mathematical Sciences

MSA220/MVE440 Statistical Learning for Big Data

4th April 2019

Random Forests

1. Given a training sample with p features, do for $b = 1, \dots, B$
   1.1 Draw a bootstrap sample of size n from training data (with replacement)
   1.2 Grow a tree T_b until each node reaches minimal node size $n_{\min}$
      1.2.1 Randomly select m variables from the p available
      1.2.2 Find best splitting variable among these m
      1.2.3 Split the node
2. For a new $\mathbf{x}$ predict

Regression: $\widehat{f}_{rb}(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} T_b(\mathbf{x})$

Classification: Majority vote at $\mathbf{x}$ across trees

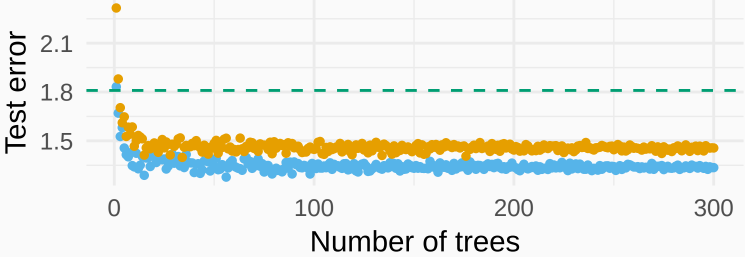
Note: Step 1.2.1 leads to less correlation between trees built on bootstrapped data.

Comparison of RF, Bagging and CART

Toy example

$$y = x_1^2 + \varepsilon \quad \text{where} \quad \varepsilon \sim N(0,1)$$
$$\mathbf{x} \sim N(\mathbf{0}, \mathbf{\Sigma}),\ \mathbf{x} \in \mathbb{R}^5, \quad \mathbf{\Sigma}_{ll} = 1, \mathbf{\Sigma}_{lk} = 0.98, l \neq k$$

Training and test data were sampled from the true model. Results for RF, bagged CART and a single CART, using $x_1, \dots, x_5$ as predictor variables. ($n_{tr} = 50$, $n_{te} = 100$)



Variable importance

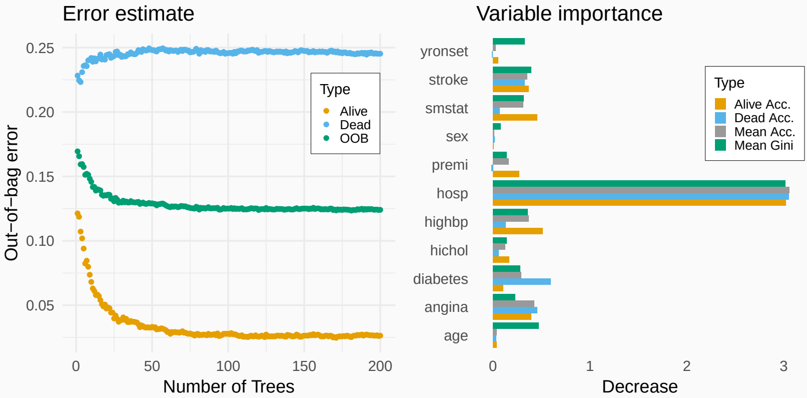
1. **Impurity index:** Splitting on a feature leads to a reduction of node impurity. Summing all improvements over all trees per feature gives a measure for variable importance
2. **Out-of-bag error**
   - During bootstrapping for large enough n, each sample has a chance of about 63% to be selected
   - For bagging the remaining samples are **out-of-bag**.
   - These out-of-bag samples for tree T_b can be used as a test set for that particular tree, since they were not used during training. Resulting in test error E_0
   - Permute variable j in the out-of-bag samples and calculate test error again $E_1^{(j)}$
   - The increase in error

     $$E_1^{(j)} - E_0 \geq 0$$

     serves as an importance measure for variable j

RF applied to cardiovascular dataset

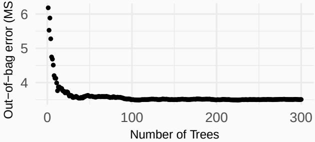
Monica dataset (`http://thl.fi/monica`, $n = 6367$, $p = 11$)
Predicting whether or not patients survive a 10 year period given a number of cardiovascular risk factors (class ratio 1.25 alive : 1 dead)

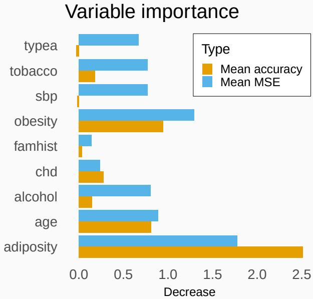


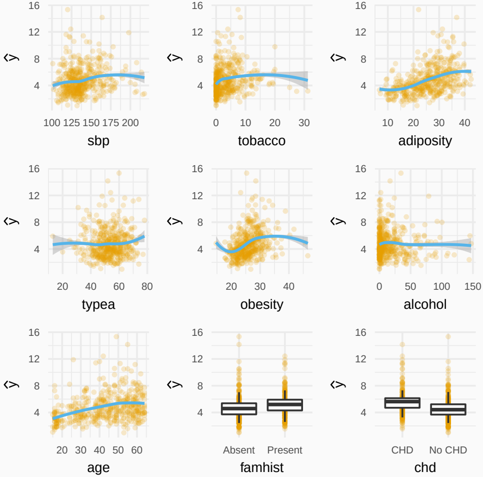
RF applied to heart disease dataset

South African coronary heart disease (SAheart) dataset

$n = 462$, $p = 9$, predicting cholesterol levels in variable `ldl`







Principal Component Analysis

Projection onto a subspace

Assume $\mathbf{x} \in \mathbb{R}^p$. Given **orthonormal vectors** $\mathbf{b}_1, \dots, \mathbf{b}_m$, i.e.

$$||\mathbf{b}_j|| = 1 \quad \text{and} \quad \mathbf{b}_j^T \mathbf{b}_k = 0 \text{ for } j \neq k$$

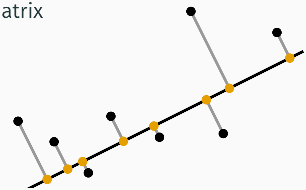
where $m < p$, the projection of $\mathbf{x}$ onto the m-dimensional linear subspace $V_m = \mathrm{span}(\mathbf{b}_1, \dots, \mathbf{b}_m)$ is

$$\hat{\mathbf{x}} = \sum_{j=1}^{m} (\mathbf{x}^T \mathbf{b}_j)\mathbf{b}_j = \underbrace{\left(\sum_{j=1}^{m} \mathbf{b}_j \mathbf{b}_j^T\right)}_{\substack{\text{Projection}\\ \text{matrix}}} \mathbf{x}$$

The projection is **orthogonal**, i.e.

$$(\mathbf{x} - \hat{\mathbf{x}})^T \mathbf{b}_j = 0$$

for all $\mathbf{b}_j$.



Rayleigh Quotient

Let $\mathbf{A} \in \mathbb{R}^{k \times k}$ be a symmetric matrix. For $\mathbf{0} \neq \mathbf{x} \in \mathbb{R}^k$ define

$$J(\mathbf{x}) = \frac{\mathbf{x}^T \mathbf{A} \mathbf{x}}{\mathbf{x}^T \mathbf{x}}$$

$J(\mathbf{x})$ is called the **Rayleigh Quotient** for $\mathbf{A}$.

Maximizing the Rayleigh Quotient

The maximization problem

$$\max_{\mathbf{x}} J(\mathbf{x}) \quad \text{subject to} \quad \mathbf{x}^T \mathbf{x} = 1$$

is solved by a **unit eigenvector** $\mathbf{x}$ of $\mathbf{A}$ corresponding to the **largest eigenvalue** λ of $\mathbf{A}$.

Note: $-\mathbf{x}$ is also a solution.

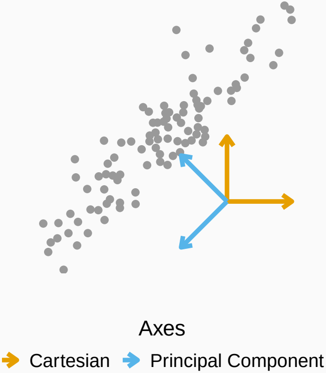
Principal Component Analysis (PCA) (I)

Goal: Given continuous data, find an orthogonal coordinate system such that the variance of the data is maximal along each direction.

Given data points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and a unit vector $\mathbf{r}$, the **variance of the data along $\mathbf{r}$** is

$$S(\mathbf{r}) = \sum_{l=1}^{n} (\mathbf{r}^T(\mathbf{x}_l - \overline{\mathbf{x}}))^2 = (n-1)\mathbf{r}^T \widehat{\boldsymbol{\Sigma}} \mathbf{r}$$

where $\widehat{\boldsymbol{\Sigma}}$ is the empirical covariance matrix.



Principal Component Analysis (PCA) (II)

Direction with maximal variance: Find $\mathbf{r}$ such that

$$\max_{\mathbf{r}} S(\mathbf{r}) \quad \text{subject to} \quad ||\mathbf{r}||^2 = \mathbf{r}^T \mathbf{r} = 1$$

- This is the same problem as maximizing the **Rayleigh Quotient** for the matrix $\widehat{\mathbf{\Sigma}}$.
- The **solution** is the eigenvector $\mathbf{r}_1$ of $\widehat{\mathbf{\Sigma}}$ corresponding to the largest eigenvalue λ_1.

How do we find the other directions? Project data on orthogonal complement of $\mathbf{r}_1$, i.e.

$$\hat{\mathbf{x}}_l = \left(\mathbf{I}_p - \mathbf{r}_1 \mathbf{r}_1^T\right) \mathbf{x}_l$$

and repeat the procedure above.

Principal Component Analysis (PCA) (III)

Computational Procedure:

1. **Centre** and **standardize** the columns of the data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$
2. Calculate the **empirical covariance matrix** $\widehat{\mathbf{\Sigma}} = \frac{1}{n-1}\mathbf{X}^T\mathbf{X}$
3. Determine the **eigenvalues** λ_j and corresponding orthonormal **eigenvectors** $\mathbf{r}_j$ of $\widehat{\mathbf{\Sigma}}$ for $j = 1, \dots, p$ and order them such that

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$$

4. The vectors $\mathbf{r}_j$ give the direction of the **principal components (PC)** $\mathbf{r}_j^T\mathbf{x}$ and the eigenvalues λ_j are the **variances along the PC directions**

Note: Set $\mathbf{R} = (\mathbf{r}_1, \dots, \mathbf{r}_p)$ and $\mathbf{D} = \mathrm{diag}(\lambda_1, \dots, \lambda_p)$ then

$$\widehat{\mathbf{\Sigma}} = \mathbf{R}\mathbf{D}\mathbf{R}^T \quad \text{and} \quad \mathbf{R}^T\mathbf{R} = \mathbf{R}\mathbf{R}^T = \mathbf{I}_p$$

PCA and Dimension Reduction

Recall: For a matrix $\mathbf{A} \in \mathbb{R}^{k \times k}$ with eigenvalues $\lambda_1, \dots, \lambda_k$ it holds that

$$\operatorname{tr}(\mathbf{A}) = \sum_{j=1}^{k} \lambda_j$$

For the empirical covariance matrix $\widehat{\boldsymbol{\Sigma}}$ and the variance of the j-th feature $\operatorname{Var}[x_j]$

$$\operatorname{tr}(\widehat{\boldsymbol{\Sigma}}) = \sum_{j=1}^{p} \operatorname{Var}[x_j] = \sum_{j=1}^{p} \lambda_j$$

is called the **total variation**.

Using only the first $m < p$ principal components leads to

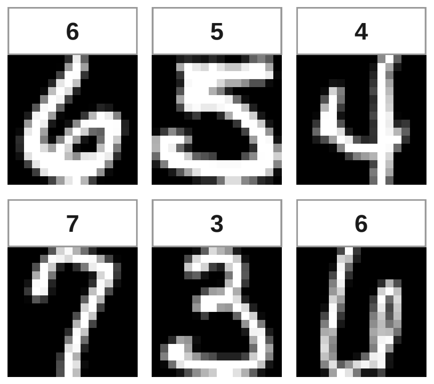
$$\frac{\lambda_1 + \cdots + \lambda_m}{\lambda_1 + \cdots + \lambda_p} \cdot 100\% \quad \text{of } \textbf{explained variance}$$

PCA and Dimension Reduction: Example (I)

Variant of the MNIST handwritten digits dataset

($n = 7291$, 16×16 greyscale images, i.e. $p = 256$)

Digit	Frequency
0	0.16
1	0.14
2	0.10
3	0.09
4	0.09
5	0.08
6	0.09
7	0.09
8	0.07
9	0.09



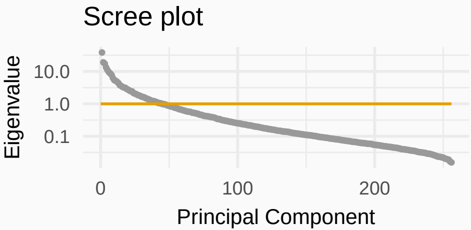
PCA and Dimension Reduction: Example (II)

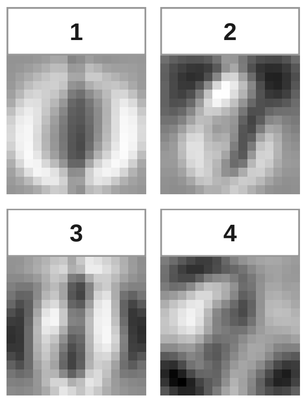
For standardized variables

$$\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}) = p$$

Typical selection rule: Components with

$$\lambda_j \geq \frac{1}{p}\,\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}) \quad (= 1)$$



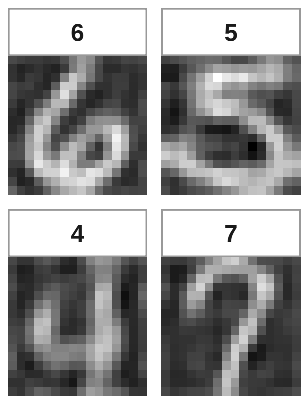


PCA and Dimension Reduction: Example (III)

Using the selection rule leads to 44 components. Using the projection

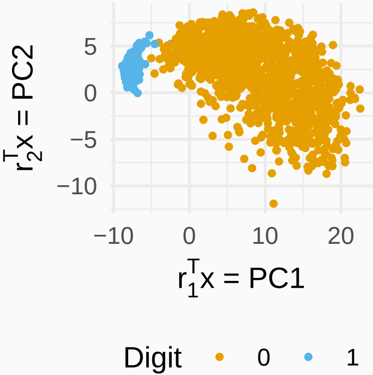
$$\hat{\mathbf{x}} = \left(\sum_{j=1}^{44} \mathbf{r}_j \mathbf{r}_j^T \right) \mathbf{x}$$

creates a **reconstruction** of $\mathbf{x}$.



PCA and Dimension Reduction: Example (IV)

Projecting the digits onto the first two principal component directions gives a very clear distinction of digits 0 and 1.



Running QDA naively on all 256 variables to predict the digits does not work. Use the two most variable features across both classes.

Table 1: Missclassifaction rate (20-fold CV)

	0	1	Overall
QDA + PCA	0.000	0.010	0.005
LDA + PCA	0.044	0.000	0.024
LDA + max var	0.007	0.024	0.015
QDA + max var	0.015	0.028	0.021

Singular Value Decomposition

Singular Value Decomposition (SVD)

The **singular value decomposition (SVD)** of a matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, $n \geq p$, is

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$$

where $\mathbf{U} \in \mathbb{R}^{n \times p}$ and $\mathbf{V} \in \mathbb{R}^{p \times p}$ with

$$\mathbf{U}^T\mathbf{U} = \mathbf{I}_p \quad \text{and} \quad \mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I}_p$$

and $\mathbf{D} \in \mathbb{R}^{p \times p}$ is diagonal. Usually

$$d_{11} \geq d_{22} \geq \cdots \geq d_{pp}$$

Note: Due to the **orthogonality conditions** for $\mathbf{U}$ and $\mathbf{V}$

$$\mathbf{X}\mathbf{X}^T\mathbf{U} = \mathbf{U}\mathbf{D}^2$$
$$\mathbf{X}^T\mathbf{X}\mathbf{V} = \mathbf{V}\mathbf{D}^2$$

SVD and PCA

In PCA the empirical covariance matrix $\widehat{\mathbf{\Sigma}}$ is in focus, whereas SVD focuses on the data matrix $\mathbf{X}$ directly.

Connection: For centred variables

$$\widehat{\mathbf{\Sigma}} = \frac{\mathbf{X}^T\mathbf{X}}{n-1} = \frac{\mathbf{V}\mathbf{D}\mathbf{U}^T\mathbf{U}\mathbf{D}\mathbf{V}^T}{n-1} = \mathbf{V}\left(\frac{\mathbf{D}^2}{n-1}\right)\mathbf{V}^T$$

The PC directions are in $\mathbf{V}$ and the eigenvalues of $\widehat{\mathbf{\Sigma}}$ are $d_{jj}^2/(n-1)$.

Note: This is how PCA is typically calculated. SVD is a **more general tool** and is used in many other contexts as well.

SVD and best rank-q-approximation / dimension reduction

Write $\mathbf{u}_j$ and $\mathbf{v}_j$ for the columns of $\mathbf{U}$ and $\mathbf{V}$, respectively. Then

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T = \sum_{j=1}^{p} d_{jj} \underbrace{\mathbf{u}_j \mathbf{v}_j^T}_{\text{rank-1-matrix}}$$

Best rank-q-approximation: For $q < p$

$$\mathbf{X}_q = \sum_{j=1}^{q} d_{jj} \mathbf{u}_j \mathbf{v}_j^T$$

with **approximation error**

$$\left\|\mathbf{X} - \mathbf{X}_q\right\|_2^2 = \left\|\sum_{j=q+1}^{p} d_{jj} \mathbf{u}_j \mathbf{v}_j^T\right\|_2^2 = \sum_{j=q+1}^{p} d_j^2$$

Connections to Discriminant Analysis

Discriminant Analysis and the Inverse Covariance Matrix

From PCA or SVD we get $\widehat{\mathbf{\Sigma}} = \mathbf{V}\mathbf{D}\mathbf{V}^T$ where $\mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I}_p$ and $d_{11} \geq \cdots \geq d_{pp} \geq 0$. Then

$$\widehat{\mathbf{\Sigma}}^{-1} = \mathbf{V}\mathbf{D}^{-1}\mathbf{V}^T = \mathbf{V}\mathbf{D}^{-1/2}\mathbf{D}^{-1/2}\mathbf{V}^T = \left(\widehat{\mathbf{\Sigma}}^{-1/2}\right)^T \widehat{\mathbf{\Sigma}}^{-1/2}$$

where $(\mathbf{D}^{-1/2})_{jj} := 1/\sqrt{d_{jj}}$ and $\widehat{\mathbf{\Sigma}}^{-1/2} := \mathbf{D}^{-1/2}\mathbf{V}^T$.

In DA the term involving the inverse covariance matrix is then

$$\begin{aligned}(\mathbf{x} - \widehat{\boldsymbol{\mu}})^T \widehat{\mathbf{\Sigma}}^{-1} (\mathbf{x} - \widehat{\boldsymbol{\mu}}) &= (\mathbf{x} - \widehat{\boldsymbol{\mu}})^T \left(\widehat{\mathbf{\Sigma}}^{-1/2}\right)^T \widehat{\mathbf{\Sigma}}^{-1/2} (\mathbf{x} - \widehat{\boldsymbol{\mu}}) \\ &= \left(\mathbf{V}^T(\mathbf{x} - \widehat{\boldsymbol{\mu}})\right)^T \mathbf{D}^{-1} \left(\mathbf{V}^T(\mathbf{x} - \widehat{\boldsymbol{\mu}})\right) \\ &= \sum_{j=1} \frac{1}{d_{jj}} (\tilde{x}_j - \tilde{\mu}_j)^2\end{aligned}$$

Inverse of the eigenvalues can lead to **numerical instability!**

Regularized Discriminant Analysis (RDA)

The empirical covariance matrix can be **stabilized**:

$$\widehat{\boldsymbol{\Sigma}}_\lambda := \widehat{\boldsymbol{\Sigma}} + \lambda \mathbf{I}_p = \mathbf{V}(\mathbf{D} + \lambda \mathbf{I}_p)\mathbf{V}^T$$

where $\lambda > 0$ is a tuning parameter.

- Using $\widehat{\boldsymbol{\Sigma}}_\lambda$ in LDA is called **regularized discriminant analysis (RDA)**.
- Instead of $1/d_{jj}$ the values $1/(d_{jj} + \lambda)$ are now involved.
- For small d_{jj} this can lead to **numerical stability**, whereas large d_{jj} are not much affected.
- For large λ the d_{jj} will have diminishing impact and RDA starts to become **nearest centroids**.
- RDA can be used with QDA as well by considering:

$$\widehat{\boldsymbol{\Sigma}}_{i,\lambda} := \underbrace{\widehat{\boldsymbol{\Sigma}}_i}_{\text{QDA}} + \lambda \underbrace{\widehat{\boldsymbol{\Sigma}}}_{\text{LDA}}$$

Take-home message

- Random forests is very flexible and can determine variable importance
- Principal component analysis gives a convenient decomposition of the data with respect to variance
- Singular value decomposition is a universal workhorse for dimension reduction