7 Unsupervised learning

7.1 Eigen-decomposition

For a $d \times d$ real, symmetric, square matrix A of rank $n \leq d$, the eigen-decomposition can be written

$$A = U\Sigma U^T$$

where

- A has real eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$; $d - n$ of the λ_i are zeros.
- Σ is a diagonal matrix, the diagonal elements of Σ being the eigenvalues of A,
$$\begin{pmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \lambda_d \end{pmatrix}$$
- the columns of U are $u_1, \ldots u_d \in \mathbb{R}^{d\times 1}$,
- the u_i are the eigenvectors of A, $Au_i = \lambda_i u_i$,
- the u_i form an orthonormal basis for $\mathbb{R}^d$: $u_i^T u_j = \delta_{i,j}$, $U^T U = UU^T = I_d$.

7.2 Singular Value Decomposition

- Suppose you have a real $m \times n$ matrix A.
- $A^T A$ is an $n \times n$ matrix with non-negative eigenvalues:
$$(A^T A)v = \lambda v \implies v^T A^T A v = \lambda v^T v = \lambda \|v\|_2^2 \geq 0 \implies \lambda \geq 0.$$
- The *singular values* of A are the square roots of the eigenvalues of $A^T A$ and AA^T.
- Decomposition
$$A = U\Sigma V^T = \sum_{i=1}^{min\{m,n\}} \Sigma_{i,i} u_i v_i^T$$
where
  - U is an $m \times m$ orthogonal matrix (columns u_i; $u_i \cdot u_j = \delta_{i,j}$),
  - V is an $n \times n$ orthogonal matrix (columns u_i; $u_i \cdot u_j = \delta_{i,j}$), and
  - $\Sigma = (\Sigma_{i,j})$ is a diagonal matrix of the (non-negative) singular values of A, in decreasing order (some may be zero).
  - the u_i are the left-singular vectors of A

- the v_i are the right-singular vectors of A

- The decomposition always exists.
- The left-singular vectors of A are the eigenvectors of AA^T.
- The right-singular vectors of A are the eigenvectors of $A^T A$
- The non-zero eigenvalues of AA^T and $A^T A$ are are the square singular-values of A

$$AA^T = U\Sigma^2 U^T$$
$$A^T A = V\Sigma^2 V^T$$

  (we are allowing Σ to change size by padding with zeros as convenient)

- R's `svd` command returns an economical version of the svd:
  - U is returned as an $m \times \min\{m, n\}$ matrix,
  - Σ is returned as a vector d of length $\min\{m, n\}$, and
  - V is returned as a $n \times \min\{m, n\}$ matrix.
  - All that has happened is the inconsequential columns of U and V have been trimmed away and it is still the case that

$$\mathbf{A} = U\mathbf{\Sigma}V^T = \sum_{i=1}^{\min\{m,n\}} \sigma_{i,i} u_i v_i^T,$$

    where $\mathbf{\Sigma} = \text{diag}(d)$.
  - Time complexity $\max\{m, n\}\times\min\{m, n\}^2$ =$\min\{m^2 n, mn^2\}$
  - You can calculate the most significant parts of the SVD more quickly than the full SVD.

7.3 PCA

- Tool for exploratory data analysis
- To explain the variance in the data
- Similar to variable selection for linear regression
- Unlabeled data == Unsupervised learning
- $n \times d$ data matrix X — centered (columns have mean zero) and probably scaled (columns have s.d. one); n observations in $\mathbb{R}^d$.
- For linear regression you want $n \gg d$. Not necessary for PCA: i.e. DNA datasets.

- The variance of X is $\frac{1}{n}\sum_{i,j=1}^{d}(X^TX)_{i,j} = \frac{1}{n}\sum_{i,j=1}^{n}(XX^T)_{i,j}$.
- By SVD $X = U\Sigma V^T$; U an $n \times n$ matrix and V a $d \times d$ matrix. so

  $$XX^T = (U\Sigma V^T)(U\Sigma V^T)^T = (U\Sigma V^T)(V\Sigma U^T) = U\Sigma^2U^T$$

  and

  $$X^TX = (U\Sigma V^T)^T(U\Sigma V^T) = (V\Sigma U^T)(U\Sigma V^T) = V\Sigma^2V^T.$$

- For any unit vector $v = \sum_{i=1}^{d}\alpha_i v_i \in \mathbb{R}^{d\times 1}$ (v_i the columns of V, $\sum_{i=1}^{d}\alpha_i^2 = 1$), $\mathrm{var}(Xv) = (Xv)^T(Xv) = v^TV\Sigma^2V^Tv = \sum_{i=1}^{d}\alpha_i^2\Sigma_{i,i}^2$.
- Consider a change of basis in $\mathbb{R}^d$ from the normal basis to $v_1, \dots v_d$. The direction v_1 corresponds to the direction that maximizes the variance of X; v_i corresponds to the direction of X, amongst all directions orthogonal to $v_{1,} \dots v_{i-1}$, that maximizes the variance of X.
- The proportion of the variance captured by the first k principal components is $\sum_{i=1}^{k}\Sigma_{i,i}^2/\sum_{i=1}^{\min\{n,d\}}\Sigma_{i,i}^2$.
- $XV = (U\Sigma V^T)V = U\Sigma$ is a $n \times \min\{n, d\}$ matrix. This is X transformed into PCA space.
- Works well in high dimensions. Fails to spot non-linear patterns.

7.4 Problems

The (*) questions from sheets 5 to 7 form homework 2.

1. Consider the matrix A:

   ```
   A=outer(1:101,1:101,function(i,j) sqrt((i-50)**2+(j-50)**2)) #Produce A
   image(a,col=grey(seq(0,1,0.01)))  #Plot A as ann image
   ```

   Does A have a good low rank approximation? What rank?

2. (*) Consider a sample from the multivariate normal distribution

   ```
   library(MASS)
   Sigma=matrix(c(14,   15,   18,
                  15,   17,   21,
                  18,   21,   27),3,3)
   A=mvrnorm(10^4,mu=c(1,2,3),Sigma=Sigma)
   library(rgl);plot3d(A) #plot A
   p=prcomp(A,scale=F,retx=T)
   ```

   (a) What, approximately, is the value of `p$center`, and why?

(b) How are the principal components `p$sdev` of `A` related to the eigenvalues of Sigma `eigen(Sigma)$values`? Why?

(c) What, approximately speaking, is the "probability distribution" of the rows in `p$x`? How is `var(p$x)` related to `Sigma`?

(d) How are `prcomp(A, scale = T)$sdev` related to `Sigma`? [Hint: consider the "probability distribution" of the rows of `scale(A)`.]

3. wisconsin-breast-cancer.RData contains a matrix x and a vector y. Each row of x contains measurements related to cells sampled whilst testing for cancer. The vector y classifies the samples 0=benign, 1=malignant.

   (a) Which of the *columns* of X is most highly correlated with y.

   (b) Set p=prcomp(X,scale=T). Which of the columns of p$x is most highly correlated with y. Use `lm(y ~ p$x[, 1] + p$x[, 2])` to find a linear combination of the first two principal component of X that is strongly correlated with y.

4. (*) Download from `http://archive.ics.uci.edu/ml/datasets/Molecular+Biology+%28Promoter+Gene+Sequences%29` and load it into R:

```
a=read.csv("promoters.data",stringsAsFactors=F,strip.white=T,header=F)
y=as.numeric(a[,1]=="+")
x=a[,3]
```

   There are (x) 106 samples of DNA sequence of length 57, and (y) a classification of the 106 samples into two classes (promoters/non-promotors of E-coli).

   (a) Convert x into a numeric matrix X suitable for use with PCA. [Hint: consider a mapping such as "a"->(1,0,0,0), "t"->(0,1,0,0), "c"->(0,0,1,0), "g"->(0,0,0,1) which produces a matrix of size 106x228 (as 57x4=228)].

   (b) How are the principal components of X correlated to Y? Find a linear combination of the first two principal components that is fairly strongly correlated with y.