Machine Learning - Michaelmas Term 2021
Lectures 18 : Principal Component Analysis

Lecturer: Phil Blunsom & Atılım Güneş Baydin

1 Introduction

1.1 Goals of PCA

Principal components analysis (PCA) is a dimensionality reduction technique that can be used to give a compact representation of data while minimising information loss. Suppose we are given a set of data, represented as vectors in a high-dimensional space. It may be that many of the variables are correlated and that the data closely fits a lower dimensional linear manifold. In this case, PCA finds such a lower dimensional representation in terms of uncorrelated variables called principal components. PCA can also be kernelised, allowing it to be used to fit data to low-dimensional non-linear manifolds. Besides dimensionality reduction, PCA can also uncover patterns in data and lead to a potentially less noisy and more informative representation. Often one applies PCA to prepare data for further analysis, e.g., finding nearest neighbours or clustering.

1.2 Technical Overview

In a nutshell, PCA proceeds as follows. We are given a collection of data in the form of n vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^m$. By first translating the data vectors, if necessary, we may assume that the input data are mean centred, that is, $\sum_{i=1}^n \mathbf{x}_i = \mathbf{0}$. Given a target number of dimensions $k \ll m$, PCA aims to find an orthonormal family of k vectors $\mathbf{u}_1, \ldots, \mathbf{u}_k \in \mathbb{R}^m$ that "explain most of the variation in the data". More precisely, for $i = 1, \ldots, n$ we approximate each data point $\mathbf{x}_i$ by a linear expression $z_{i1}\mathbf{u}_1 + \ldots + z_{ik}\mathbf{u}_k$ for some scalars $z_{i1}, \ldots, z_{ik} \in \mathbb{R}$; the goal of PCA is to choose the $\mathbf{u}_i$ so as to optimise the quality of this approximation over all data points. The optimal such vectors $\mathbf{u}_1, \ldots, \mathbf{u}_k$ are the k principal components: $\mathbf{u}_1$ is direction of greatest variance in the data, $\mathbf{u}_2$ is the direction of greatest variance that is orthogonal to $\mathbf{u}_1$, *etc.* To find the principal components we apply a matrix factorisation technique—the singular value decomposition—to the $m \times n$ matrix whose columns are the mean-centred data points $\mathbf{x}_i$. In the end, representing each data point $\mathbf{x}_i \in \mathbb{R}^m$ by its coordinates $\mathbf{z}_i = (z_{i1}, \ldots, z_{ik})$ with respect to the k principal components yields a more compact lower dimensional and (hopefully) more informative representation.

Example 1. *Figure 1 represents the output of PCA (with $k = 2$) on a DNA database representing the genomes of 1400 Europeans. In this database each genome was represented by a vector in $\mathbb{R}^m$ (where $m \approx 200,000$) representing the results of sampling the DNA in certain positions. When projected onto the top two principal components there are natural clusters that correspond to the spacial distribution of the genomes. Thus PCA reveals that the two spacial dimensions are "latent variables" that explain much of the variation in the DNA of modern Europeans.*

2 Singular-Value Decomposition

We start by describing a mathematical construction that can be used to carry out PCA.

Let $\mathbf{A}$ be a real $m \times n$ matrix of rank r. Write $\mathbf{I}_r$ for the $r \times r$ identity matrix. A *singular value decomposition* (SVD) of $\mathbf{A}$ is a factorisation $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, where

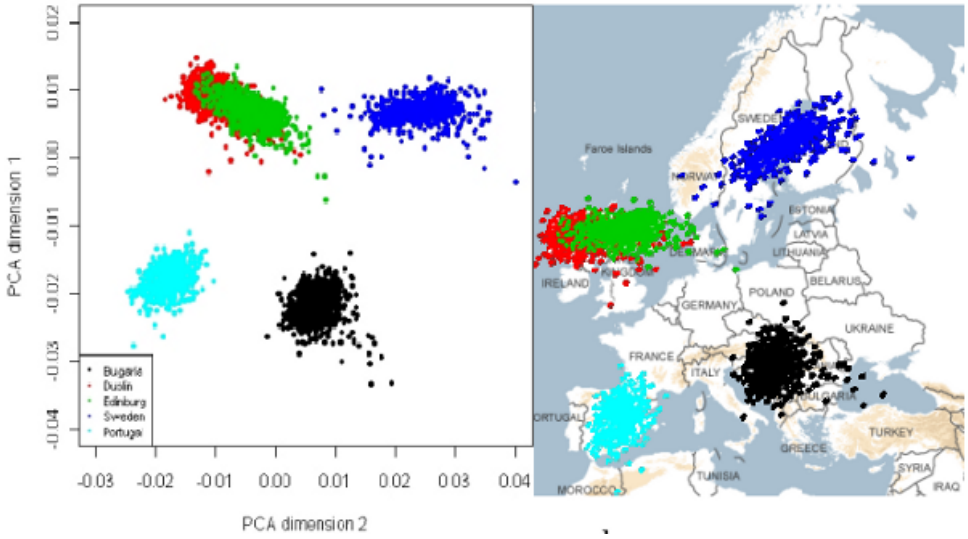

Figure 1: Genetic differences within European populations

- $\mathbf{U}$ is an $m \times r$ matrix such that $\mathbf{U}^\top \mathbf{U} = \mathbf{I}_r$,
- $\mathbf{V}$ is an $n \times r$ matrix such that $\mathbf{V}^\top \mathbf{V} = \mathbf{I}_r$,
- $\mathbf{\Sigma}$ is an $r \times r$ diagonal matrix $\mathbf{\Sigma} = \begin{pmatrix} \sigma_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_r \end{pmatrix}$, where $\sigma_1 \geq \dots \geq \sigma_r$ are strictly positive.

The columns $\mathbf{u}_1, \dots, \mathbf{u}_r$ of $\mathbf{U}$, which form an orthonormal set, are called *left singular vectors*. The columns $\mathbf{v}_1, \dots, \mathbf{v}_r$ of $\mathbf{V}$, which also form an orthonormal set, are called *right singular vectors*. The SVD yields a decomposition of $\mathbf{A}$ as a sum of r rank-one matrices:

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top . \tag{1}$$

The factorisation (1) can equivalently be expressed by the equations $\mathbf{A}\mathbf{v}_i = \sigma_i \mathbf{u}_i$ for $i = 1, \dots, r$.

Given an SVD $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, we can expand $\mathbf{U}$ to an $m \times m$ orthogonal matrix $\widehat{\mathbf{U}}$ by adding $m - r$ extra columns. We can likewise expand $\mathbf{V}$ to an $n \times n$ orthogonal matrix $\widehat{\mathbf{V}}$ by adding $n - r$ extra columns, and furthermore expand $\mathbf{\Sigma}$ to an $m \times n$ matrix by adding extra entries that are all zero. In this case we have $\mathbf{A} = \widehat{\mathbf{U}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{V}}^\top$. We call such a factorisation of $\mathbf{A}$ into a product of an $m \times m$ orthogonal matrix, $m \times n$ nonnegative diagonal matrix, and $n \times n$ orthogonal matrix, a ***full SVD***.[1] A full SVD expresses the linear transformation represented by $\mathbf{A}$ as a rotation, followed by a scaling, following by another rotation.

2.1 Existence of SVDs

Every matrix $\mathbf{A}$ has an SVD $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$. As we shall see, the singular values are uniquely determined by $\mathbf{A}$.

Theorem 1. *Every matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ has an SVD.*

[1] Conversely, the version of the SVD defined above is sometimes called the *reduced SVD*.

Proof. Inductively define an orthonormal sequence of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_r \in \mathbb{R}^n$ such that

$$\begin{aligned}
\mathbf{v}_1 &:= \operatorname*{argmax}_{\|\mathbf{v}\|=1} \|\mathbf{A}\mathbf{v}\|, \\
\mathbf{v}_2 &:= \operatorname*{argmax}_{\mathbf{v}\perp\mathbf{v}_1\, \|\mathbf{v}\|=1} \|\mathbf{A}\mathbf{v}\|, \\
\mathbf{v}_3 &:= \operatorname*{argmax}_{\mathbf{v}\perp\mathbf{v}_1, \mathbf{v}\perp\mathbf{v}_2\, \|\mathbf{v}\|=1} \|\mathbf{A}\mathbf{v}\|, \ \textit{etc.},
\end{aligned}$$

where the sequence stops at the first index r such that $\mathbf{A}$ is zero on $\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}^\perp$. Furthermore, for $i = 1, \ldots, r$, define $\sigma_i := \|\mathbf{A}\mathbf{v}_i\|$ and define the unit vector $\mathbf{u}_i \in \mathbb{R}^m$ by $\mathbf{A}\mathbf{v}_i = \sigma_i \mathbf{u}_i$.

The set $\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$ is orthonormal by construction. To prove the theorem it suffices to show that $\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ is an orthonormal set and that $\mathbf{A} = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$ (as in (1)). For the latter task, write $\mathbf{B} := \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$. A direct calculation shows that $\mathbf{A}\mathbf{v}_i = \mathbf{B}\mathbf{v}_i$ for $i = 1, \ldots, r$ and that $\mathbf{A}\mathbf{v} = \mathbf{B}\mathbf{v} = 0$ for any vector $\mathbf{v} \in \{\mathbf{v}_1, \ldots, \mathbf{v}_r\}^\perp$. It follows that $\mathbf{A} = \mathbf{B}$.

Finally we show that $\mathbf{u}_i^\top \mathbf{u}_j = 0$ for all $1 \le i < j \le r$. Define $\mathbf{w} : \mathbb{R} \to \mathbb{R}^n$ by $\mathbf{w}(t) = \frac{(\mathbf{v}_i + \varepsilon \mathbf{v}_j)}{\sqrt{1+t^2}}$. Since $\mathbf{w}(t)$ is a unit vector orthogonal to $\mathbf{v}_1, \ldots, \mathbf{v}_{i-1}$ for all t, $\|\mathbf{A}\mathbf{w}(t)\|^2$ is maximised at $t = 0$. But

$$\|\mathbf{A}\mathbf{w}(t)\|^2 = \frac{\|\sigma_i \mathbf{u}_i + t\sigma_j \mathbf{u}_j\|^2}{1+t^2} = \frac{\sigma_i^2 + 2t\sigma_i\sigma_j(\mathbf{u}_i^\top \mathbf{u}_j) + t^2\sigma_j^2}{1+t^2} \tag{2}$$

and a direct caluclation shows that (2) has zero derivative at $t = 0$ if and only if $\mathbf{u}_i^\top \mathbf{u}_j = 0$. □

2.2 SVDs and Spectral Theory

In this section we explain the relationship between a singular value decomposition of a matrix $\mathbf{A}$ and the eigenvalues and eigenvectors of the matrices $\mathbf{A}^\top\mathbf{A}$ and $\mathbf{A}\mathbf{A}^\top$. Note here that $\mathbf{A}^\top\mathbf{A}$ and $\mathbf{A}\mathbf{A}^\top$ are square matrices and so it makes sense to talk about their eigenvectors and eigenvalues.

In a general a matrix may have many different SVDs. However the following proposition shows that all SVDs involve the same singular values. Thus we may speak of *the* singular values of a matrix A.

Proposition 2. *Given any SVD of* $\mathbf{A}$*, the singular values are the square roots of the nonzero eigenvalues of* $\mathbf{A}^\top\mathbf{A}$ *or* $\mathbf{A}\mathbf{A}^\top$ *(these matrices have the same eigenvalues).*

Proof. We show the result for $\mathbf{A}^\top\mathbf{A}$. Given a full SVD $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, we have

$$\begin{aligned}
\mathbf{A}^\top\mathbf{A}\mathbf{V} &= (\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top)^\top(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top)\mathbf{V} \\
&= \mathbf{V}\mathbf{\Sigma}^\top\mathbf{U}^\top\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top\mathbf{V} \\
&= \mathbf{V}\mathbf{\Sigma}^\top\mathbf{\Sigma}\mathbf{V}^\top\mathbf{V} \\
&= \mathbf{V}\mathbf{\Sigma}^2.
\end{aligned}$$

It follows that for $i = 1, \ldots, r$ each right singular vector $\mathbf{v}_i$ of $\mathbf{A}$ is an eigenvector of $\mathbf{A}^\top\mathbf{A}$ with non-zero eigenvalue σ_i^2. The remaining columns of $\mathbf{V}$ span the eigenspace of $\mathbf{A}^\top\mathbf{A}$ corresponding to the eigenvalue zero.

One can similarly show that the left singular vectors of $\mathbf{A}$ are eigenvectors of $\mathbf{A}\mathbf{A}^\top$. □

From the proof of Proposition 2 we see that in any full SVD $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, the columns of $\mathbf{U}$ comprise an orthonormal basis of eigenvectors of $\mathbf{A}\mathbf{A}^\top$ and the columns of $\mathbf{V}$ comprise an orthonormal basis of eigenvectors of $\mathbf{A}^\top\mathbf{A}$. Indeed an alternative way to prove the existence of an SVD of matrix $\mathbf{A}$ is to rely on results about the spectral theory of either of the matrices $\mathbf{A}^\top\mathbf{A}$ or $\mathbf{A}\mathbf{A}^\top$. [2]

[2] For example, the matrix $\mathbf{A}^\top\mathbf{A}$, being symmetric positive semidefinite, has an orthonormal basis of eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ with the associated eigenvalues $\lambda_1, \ldots, \lambda_n$ being real and nonnegative. Writing $\sigma_i := \sqrt{\lambda_i}$, and $\mathbf{u}_i := \mathbf{A}\mathbf{v}_i$ for $i = 1, \ldots, r$ it can be proved that $\mathbf{A} = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$ is a full SVD.

3 The Eckhart-Young Theorem

3.1 The Frobenius Norm

The main application of SVD for our purposes is to compute a best low-rank approximation of a given matrix. In order to formalise this notion we introduce the *Frobenius matrix norm*. The Frobenius norm of an $m \times n$ matrix $\mathbf{A} = (a_{ij})$, denoted $\|\mathbf{A}\|_F$, is defined by

$$\|\mathbf{A}\|_F := \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2} = \sqrt{\mathrm{trace}(\mathbf{A}^\top \mathbf{A})}\,.$$

Note that the Frobenius norm of $\mathbf{A}$ is a function of the singular values of $\mathbf{A}$. Indeed, if $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ then

$$\|\mathbf{A}\|_F^2 = \mathrm{trace}(\mathbf{A}^\top\mathbf{A}) = \mathrm{trace}((\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top)^\top(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top) = \mathrm{trace}(\mathbf{\Sigma}^\top\mathbf{\Sigma}) = \sigma_1^2 + \ldots + \sigma_r^2\,.$$

3.2 Low-Rank Approximation via the SVD

Consider a matrix A that has an SVD $\mathbf{A} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$. Given $k \le r$ we obtain a rank-k matrix $\mathbf{A}_k$ by "truncating" the SVD after the first k terms:

$$\mathbf{A}_k := \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top\,. \tag{3}$$

The image of A_k is spanned to the top k left singular vectors. Hence A_k has rank k. By construction, $\mathbf{A}_k$ has singular values $\sigma_1, \ldots, \sigma_k$. Likewise, $\mathbf{A} - \mathbf{A}_k = \sum_{i=k+1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$ has singular values $\sigma_{k+1}, \ldots, \sigma_r$. Thus

$$\|\mathbf{A} - \mathbf{A}_k\|_F = \sqrt{\sigma_{k+1} + \ldots + \sigma_r}\,.$$

Note that σ_{k+1} is the top singular vector of $\mathbf{A} - \mathbf{A}_k$.

3.3 Eckhart-Young Theorem

The following result says that $\mathbf{A}_k$ is a best rank-k approximation of $\mathbf{A}$ with respect to the Frobenius norm. Since proof of this result is beyond the scope of the course we relegate it to an appendix.

Theorem 3 (Eckhart-Young)**.** *Let $\mathbf{A}$ be a real $m \times n$ matrix. Then for any $k \in \mathbb{N}$ and any real $m \times n$ matrix $\mathbf{B}$ of rank at most k we have $\|\mathbf{A} - \mathbf{A}_k\|_F \le \|\mathbf{A} - \mathbf{B}\|_F$.*

The Eckhart-Young Theorem can also be formulated in terms of orthogonal projections. Write $\mathbf{P}_k := \sum_{i=1}^{k} \mathbf{u}_i \mathbf{u}_i^\top$ for the matrix representing the orthogonal projection of $\mathbb{R}^m$ onto the subspace spanned by the top k left singular vectors of $\mathbf{A}$.

Theorem 4. *Let $\mathbf{A}$ be a real $m \times n$ matrix. Then for any $k \in \mathbb{N}$ and any $m \times m$ orthogonal projection matrix $\mathbf{P}$ of rank k, we have $\|\mathbf{A} - \mathbf{P}_k\mathbf{A}\|_F \le \|\mathbf{A} - \mathbf{P}\mathbf{A}\|_F$.*

The equivalence of Theorems 3 and 4 follows from two simple facts. First, by a simple calculation, we have that $\mathbf{A}_k = \mathbf{P}_k\mathbf{A}$. Second, for any rank-$k$ matrix $\mathbf{B}$ we have $\|\mathbf{A} - \mathbf{P}\mathbf{A}\|_F \le \|\mathbf{A} - \mathbf{B}\|_F$, where $\mathbf{P}$ denotes the orthogonal projection onto the column space of $\mathbf{B}$ (since for any vector $\mathbf{x}$, $\mathbf{P}\mathbf{x}$ is the closest vector to $\mathbf{x}$ lying in the column space of $\mathbf{B}$).

Theorem 4 moreover gives some intuition behind the construction of the SVD in Theorem 1. The idea is that for any any orthogonal projection $\mathbf{P}$, by Pythagoras's Theorem, we have $\|\mathbf{A} - \mathbf{P}\mathbf{A}\|_F^2 = \|\mathbf{A}\|_F^2 - \|\mathbf{P}\mathbf{A}\|_F^2$. Thus minimising $\|\mathbf{A} - \mathbf{P}\mathbf{A}\|_F^2$ is the same as maximising $\|\mathbf{P}\mathbf{A}\|_F^2$. This corresponds to the length-maximising characterisation of the right singular vectors in the proof of Theorem 1.

3.4 Choosing k

The Eckhart-Young Theorem can help to detemine what value of k to take in order to ensure that $\mathbf{X}_k$ is a "sufficently good" approximation of $\mathbf{X}$. In particular it allows to express the relative error of a low-rank approximation $\mathbf{X}_k$ in terms of the singular values of $\mathbf{X}$, since

$$\frac{\|\mathbf{X} - \mathbf{X}_k\|_F^2}{\|\mathbf{X}\|_F^2} = \frac{\sigma_{k+1}^2 + \cdots + \sigma_r^2}{\sigma_1^2 + \cdots + \sigma_r^2}\,. \tag{4}$$

Thus if our goal is to ensure a given bound on the relative error (say at most 0.05), then we can find an approprate value of k by examining the singular values instead of proceeding by trial and error and computing X_k for various values of k.

4 PCA

In this section we show how the singular value decomposition is used in principal components analysis. In PCA the input is a family $\mathbf{x}_1, \ldots, \mathbf{x}_n$ of data points in $\mathbb{R}^m$. Write $\boldsymbol{\mu} := \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i$ for the mean data point. By first replacing $\mathbf{x}_i$ by $\mathbf{x}_i - \boldsymbol{\mu}$ we may assume that the input data are mean centred. Given a target dimension $k \leq m$, our goal is to find points $\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_n \in \mathbb{R}^m$ such that the *reconstruction error* $\sum_{i=1}^n \|\mathbf{x}_i - \widetilde{\mathbf{x}}_i\|^2$ is minimised subject to the constraint that $\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_n$ lie in a subspace of $\mathbb{R}^m$ of dimension at most k.[3]

Respectively write the mean-centred data points and their approximants as the columns of two $m \times n$ matrices

$$\mathbf{X} := \begin{pmatrix} | & & | \\ \mathbf{x}_1 & \ldots & \mathbf{x}_n \\ | & & | \end{pmatrix} \qquad \text{and} \qquad \widetilde{\mathbf{X}} := \begin{pmatrix} | & & | \\ \widetilde{\mathbf{x}}_1 & \ldots & \widetilde{\mathbf{x}}_n \\ | & & | \end{pmatrix}$$

Then the reconstruction error is nothing but $\|\mathbf{X} - \widetilde{\mathbf{X}}\|_F^2$. Thus by the Eckhart-Young Theorem an optimal choice of $\widetilde{\mathbf{X}}$ is the matrix $\mathbf{X}_k$, as defined in (3) via a "truncated" SVD. Now recall that if $\mathbf{U}_k$ is the $m \times k$ matrix whose columns are the top k left singular vectors of $\mathbf{X}$ then, writing $\mathbf{Z} := \mathbf{U}_k^\top \mathbf{X}$, we have

$$\mathbf{X}_k = \mathbf{U}_k \mathbf{U}_k^\top \mathbf{X} = \mathbf{U}_k \mathbf{Z}\,. \tag{5}$$

The output of PCA is the pair of matrices $\mathbf{U}_k$ and $\mathbf{Z}$. The columns of $\mathbf{U}_k$ are the top k left singular vectors, while the columns of $\mathbf{Z}$ give the coefficients that respectively approximate each mean centred data point $\mathbf{x}_i$ as a linear combination of the top k left singular vectors.

4.1 Facial Recognition

In this section we describe an application of PCA to the problem of facial recognition. Suppose we have a database of n faces. Given a new face we want to find the closest match in the database. Each face is a grey-scaled image consisting of m pixels. We represent a face as a vector $\boldsymbol{x} \in \mathbb{R}^m$, where x_i represents the intensity of the i-th pixel. The dimension m is typically in the tens or hundreds of thousands. However after applying PCA it has been found that an accurate representation of the input images can be obtained using a few hundred principal components.

Let $\mathbf{X}$ be the $m \times n$ matrix whose columns represent the collection of (mean centred) faces in the database. Let $\mathbf{U}_k$ be an $m \times k$ matrix whose columns are the top k left singular vectors

[3] It can be shown that finding the best fit k-dimensional subspace for the mean-centred data $\mathbf{x}_i$ is equivalent to finding the best fit affine k dimensional subspace for the original data $\mathbf{x}_i^*$.

of $\mathbf{X}$. The columns $\mathbf{u}_1, \ldots, \mathbf{u}_k$ of $\mathbf{U}_k$ are called *eigenfaces* and express the directions of greatest variance among the faces. The columns $\mathbf{z}_1, \ldots, \mathbf{z}_n \in \mathbb{R}^k$ of the matrix $\mathbf{Z} := \mathbf{U}_k^\top \mathbf{X}$ are the coefficients that (approximately) express each face in the database as a linear combination of eigenfaces.

Given a new face $\mathbf{x} \in \mathbb{R}^m$, suppose we want to find the best match for $\mathbf{x}$ in the database. Let $\boldsymbol{\mu}$ denote the mean of the vectors in the database (prior to mean centring). The coordinates of the projection of $\mathbf{x} - \boldsymbol{\mu}$ on the subspace spanned by $\mathbf{u}_1, \ldots, \mathbf{u}_k$ are given by $\mathbf{U}_k^\top(\mathbf{x} - \boldsymbol{\mu})$. We can then find the best match for $\mathbf{x}$ in the database by finding the nearest neighbour to $\mathbf{U}_k^\top(\mathbf{x} - \boldsymbol{\mu})$ among the vectors $\mathbf{z}_i$ for $i = 1, \ldots, n$.

4.2 Latent Semantic Analysis

We would like to partition a given collection of n documents into groups of documents about similar topics. Given a collection of m *key words*, we represent each document as vector in $\mathbb{R}^m$ recording the frequency of each key word (the number of occurrences of the key word divided by the total number of occurrences of all key words).

Let the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}_n$ represent the documents in the collection after mean centring. One idea is to model the similarity of any two documents $\mathbf{x}_i$ and $\mathbf{x}_j$ by taking the inner product $\mathbf{x}_i^\top \mathbf{x}_j$ (e.g., the inner product is 0 if the documents have no word in common and more positive if there are lots of words in common). However there are some problems with this idea. For example, the inner product is oblivious to correlations among keywords—it does not take account of the fact that two different key words, e.g., *football* and *Premier League*, may be closely related and should not be treated as orthogonal.

Carrying out PCA, we represent the i-th (mean centred) document as a linear combination of the top k principal components, for some well-chosen k. Let vector $\mathbf{z}_i$ denote the coefficients of this linear combination. As explained above, it is hoped that the inner product $\mathbf{z}_i^\top \mathbf{z}_j$ will better represent the similarity of the i-th and j-th documents than $\mathbf{x}_i^\top \mathbf{x}_j$.

5 Computing an SVD

Computing the SVD is an important topic in numerical linear algebra. Here we sketch a very simple "in-principle" power-iteration method to compute the SVD (which however takes no account of numerical stability).

Let $\mathbf{A}$ be an $m \times n$ matrix with SVD $\mathbf{A} = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$. Recall that $\mathbf{v}_1, \ldots, \mathbf{v}_r$ are eigenvectors of the matrix $\mathbf{A}^\top \mathbf{A}$ with respective eigenvalues $\lambda_1 = \sigma_1^2, \ldots, \lambda_r = \sigma_r^2$. We first show how to use power iteration to find the top eigenvector $\mathbf{v}_1$ and then explain how to find the remaining eigenvectors. The procedure to compute $\mathbf{v}_1$ is as follows:

1. Select a unit vector $\mathbf{x}_0 \in \mathbb{R}^n$ uniformly at random.[4]
2. For $k = 1, 2, \ldots$, set $\mathbf{x}_k := (\mathbf{A}^\top \mathbf{A})\mathbf{x}_{k-1}$ and $\mathbf{y}_k := \frac{\mathbf{x}_k}{\|\mathbf{x}_k\|}$. Stop when $\|\mathbf{y}_k - \mathbf{y}_{k-1}\|$ is "sufficiently small" and output $\mathbf{y}_k$.

The idea is that by iteratively applying $\mathbf{A}^\top \mathbf{A}$ to a random vector we will (with high probability) converge to a vector that points in the direction $\mathbf{v}_1$ or $-\mathbf{v}_1$. The following gives a bound on the speed of convergence in terms of the ratio λ_1/λ_2—the so called *spectral gap* of $\mathbf{A}^\top \mathbf{A}$. The larger the spectral gap, the faster the convergence. In case $\lambda_1 = \lambda_2$ Proposition 5 gives no guarantee of convergence. However the proposition can easily be generalised to show that with high probability the procedure will converge to some vector in the eigenspace of $\mathbf{A}^\top \mathbf{A}$ corresponding to the leading eigenvalue.

[4] This can be done, e.g., by sampling each component of $\mathbf{x}_0$ independently from the Gaussian distribution $N(0, 1)$ and rescaling the resulting vector to have length 1.

Proposition 5. *With probability at least* $9/10$ *over the random choice of* $\mathbf{x}_0$*, for all* $k \geq 0$ *we have*

$$|\mathbf{y}_k^\top \mathbf{v}_1| \geq 1 - 20n\left(\lambda_2/\lambda_1\right)^k .$$

Proof. Consider the random variable X denoting the absolute value of the inner product of a random unit vector in $\mathbb{R}^n$ with any fixed unit vector. A symmetry argument shows that X has expected value $\frac{1}{\sqrt{n}}$. We will use the concentration bound $\Pr(X > \frac{1}{20\sqrt{n}}) \geq 9/10$, which can be proved by a geometric argument[5].

Let $b_1, \ldots, b_n$ be the (uniquely defined) scalars such that $\mathbf{x}_0 = \sum_{i=1}^n b_i \mathbf{v}_i$. Then we have that $|b_1| > \frac{1}{20\sqrt{n}}$ with probability at least $9/10$. Assuming that $|b_1| \geq \frac{1}{20\sqrt{n}}$, we have

$$\mathbf{y}_k^\top \mathbf{v}_1 = \frac{[(\mathbf{A}^\top\mathbf{A})^k \mathbf{x}_0]^\top \mathbf{v}_1}{\|(\mathbf{A}^\top\mathbf{A})^k \mathbf{x}_0\|} = \frac{\mathbf{x}_0^\top (\mathbf{A}^\top\mathbf{A})^k \mathbf{v}_1}{\|(\mathbf{A}^\top\mathbf{A})^k \mathbf{x}_0\|} = \frac{\lambda_1^k \mathbf{x}_0^\top \mathbf{v}_1}{\|(\mathbf{A}^\top\mathbf{A})^k \mathbf{x}_0\|} = \frac{\lambda_1^k b_1}{\sqrt{b_1^2\lambda_1^{2k} + \ldots + b_n^2\lambda_n^{2k}}} \tag{6}$$

We now give a lower bound on (6) using the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$:

$$(6) \geq \frac{\lambda_1^k b_1}{\sqrt{b_1^2\lambda_1^{2k} + n\lambda_2^{2k}}} \geq \frac{\lambda_1^k b_1}{b_1\lambda_1^k + \sqrt{n}\lambda_2^k} = \frac{1}{1 + 20n(\lambda_2/\lambda_1)^k} \geq 1 - 20n(\lambda_2/\lambda_1)^k .$$

□

Having computed $\mathbf{v}_1$ we can compute the next singular vector $\mathbf{v}_2$ by adapting the above procedure as follows. We modify the update step by projecting $\mathbf{x}_k$ orthogonally to $\mathbf{v}_1$, that is, we perform the update $\mathbf{x}_k := \mathbf{x}_k - (\mathbf{x}_k^\top \mathbf{v}_1)\mathbf{v}_1$. Continuing in this manner we compute all the singular values.

6 Kernel PCA

In this section we show that PCA can be kernelised. This in particular allows to use PCA to find non-linear "directions of greatest variance", cf. Figure 2.

Suppose we have a feature map $\phi : \mathbb{R}^m \to \mathbb{R}^d$ and associated kernel function $\kappa(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^\top \phi(\mathbf{x}')$. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be the original data points in $\mathbb{R}^m$. Let $\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^n \phi(\mathbf{x}_i)$ be the mean of the transformed data points. After applying the feature map and translating the data points to make them mean centred, we obtain new data points $\mathbf{x}_i{}^* := \phi(\mathbf{x}_i) - \boldsymbol{\mu}$ for $i = 1, \ldots, n$. Consider the $d \times n$ data matrix $\mathbf{X}$ with columns $\mathbf{x}_1^*, \ldots, \mathbf{x}_n^*$. Given k, our aim is to compute the $k \times n$ matrix $\mathbf{Z}$ such that $\mathbf{X}_k = \mathbf{U}_k\mathbf{Z}$, where $\mathbf{U}_k$ is the matrix whose columns are the top k left singular vectors $\mathbf{u}_1, \ldots, \mathbf{u}_k$ of $\mathbf{X}$. We show how to use the kernel function function κ to obtain $\mathbf{Z}$ as the product $\mathbf{U}_k^\top\mathbf{X}$ without explicitly constructing either of the the matrices $\mathbf{X}$ or $\mathbf{U}$.

First, we can use the kernel function to compute the $n \times n$ matrix $\mathbf{X}^\top\mathbf{X}$. Indeed we have

$$\begin{aligned}
&(\mathbf{X}^\top\mathbf{X})_{ij} \\
&\quad = (\mathbf{x}_i^*)^\top \mathbf{x}_j^* \\
&\quad = (\phi(\mathbf{x}_i) - \boldsymbol{\mu})^\top (\phi(\mathbf{x}_j) - \boldsymbol{\mu}) \\
&\quad = \left(\phi(\mathbf{x}_i) - \tfrac{1}{n}\textstyle\sum_{k=1}^n \phi(\mathbf{x}_k)\right)^\top \left(\phi(\mathbf{x}_j) - \tfrac{1}{n}\textstyle\sum_{l=1}^n \phi(\mathbf{x}_l)\right) \\
&\quad = \phi(\mathbf{x}_i)^\top\phi(\mathbf{x}_j) - \tfrac{1}{n}\textstyle\sum_{l=1}^n \phi(\mathbf{x}_i)^\top\phi(\mathbf{x}_l) - \tfrac{1}{n}\textstyle\sum_{k=1}^n \phi(\mathbf{x}_k)^\top\phi(\mathbf{x}_j) + \tfrac{1}{n^2}\textstyle\sum_{k,l=1}^n \phi(\mathbf{x}_k)^\top\phi(\mathbf{x}_l) \\
&\quad = \kappa(\mathbf{x}_i, \mathbf{x}_j) - \tfrac{1}{n}\textstyle\sum_{l=1}^n \kappa(\mathbf{x}_i, \mathbf{x}_l) - \tfrac{1}{n}\textstyle\sum_{k=1}^n \kappa(\mathbf{x}_k, \mathbf{x}_j) + \tfrac{1}{n^2}\textstyle\sum_{k,l=1}^n \kappa(\mathbf{x}_k, \mathbf{x}_l) .
\end{aligned}$$

[5] see `https://www.cs.cmu.edu/ venkatg/teaching/CStheory-infoage/book-chapter-4.pdf` for a proof

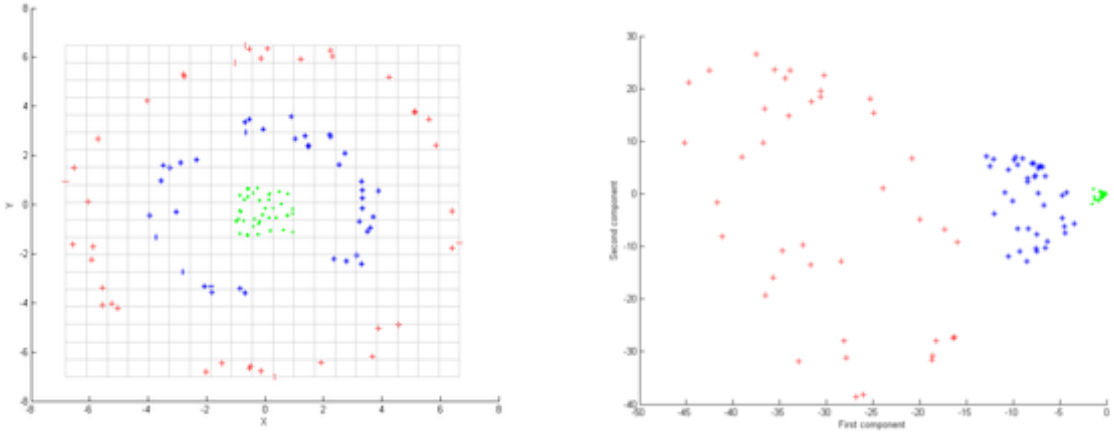

Figure 2: The left-hand diagram shows the input points before PCA. The right-hand figure shows the input points after PCA with the polynomial kernel $\kappa(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^\top \mathbf{y} + 1)^2$, corresponding to the feature map $\phi(\mathbf{x}) = (1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1x_2, x_1^2, x_2^2)$. The top principal component corresponds to the radius of the points.

The top k right singular vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ of $\mathbf{X}$ can be constructed as eigenvectors of $\mathbf{X}^\top \mathbf{X}$. These vectors lie in $\mathbb{R}^n$ and can be explicitly constructed from $\mathbf{X}^\top \mathbf{X}$ (e.g., using power iteration). The associated left singular vectors $\mathbf{u}_1, \ldots, \mathbf{u}_k$ are defined via the equation

$$\mathbf{u}_i = \frac{1}{\|\mathbf{X}\mathbf{v}_i\|} \mathbf{X}\mathbf{v}_i \tag{7}$$

for $i = 1, \ldots, k$. Now we can compute

$$\|\mathbf{X}\mathbf{v}_i\|^2 = (\mathbf{X}\mathbf{v}_i)^\top (\mathbf{X}\mathbf{v}_i) = \mathbf{v}_i^\top (\mathbf{X}^\top \mathbf{X}) \mathbf{v}_i$$

having already computed $\mathbf{X}^\top \mathbf{X}$. Thus we obtain from (7) coefficients α_{ij} such that $\mathbf{u}_i = \sum_{j=1}^n \alpha_{ij} \mathbf{x}_j^*$ for $i = 1, \ldots, k$, that is, we express the the left singular vectors as linear combinations of the transformed data vectors $\mathbf{x}_1^*, \ldots, \mathbf{x}_n^*$.

Let $\mathbf{U}_k$ be the $k \times d$ matrix whose columns are the top k left singular vectors $\mathbf{u}_1, \ldots, \mathbf{u}_k$. We can use the kernel function to compute the matrix $\mathbf{Z} := \mathbf{U}_k^\top \mathbf{X}$ such that

$$\mathbf{X}_k = \mathbf{U}_k \mathbf{U}_k^\top \mathbf{X} = \mathbf{U}_k \mathbf{Z} \,.$$

Let us emphasize that the above derivation does not require an explicit representation of either the data vectors $\mathbf{x}_i^*$ nor singular vectors $\mathbf{u}_i$ as elements of $\mathbb{R}^d$.

Acknowledgments: We would like to thank James Worrell for his contributions to these lecture notes.

7 Appendix: Proof of the Eckhart-Young Theorem

The key to proving the Eckhart-Young theorem is the following lemma, which gives a lower bound on the singular values of a perturbation of matrix $\mathbf{A}$ by matrix $\mathbf{B}$ of rank at most k. In the lemma we use the notation $\sigma_i(\mathbf{A})$ to refer to the i-th singular value of a matrix $\mathbf{A}$. If i greater than the rank of $\mathbf{A}$ then $\sigma_i(\mathbf{A})$ is defined to be 0. The proof of the lemma uses the following "triangle inequality" (whose proof we leave as an exercise): for any two matrices $\mathbf{A}$ and $\mathbf{B}$ of the same dimension, $\sigma_1(\mathbf{A} + \mathbf{B}) \leq \sigma_1(\mathbf{A}) + \sigma_1(\mathbf{B})$.

Lemma 1. *If $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}$, with $\mathbf{B}$ having rank at most k, then $\sigma_{k+i}(\mathbf{A}) \leq \sigma_i(\mathbf{A} - \mathbf{B})$ for all i.*

Proof. We first show the lemma in case $i = 1$, i.e., we prove that $\sigma_{k+1}(\mathbf{A}) \leq \sigma_1(\mathbf{A} - \mathbf{B})$.

The kernel of $\mathbf{B}$ has dimension $n - k$ and thus there must exist a unit-length vector $\mathbf{w}$ that lies both in the kernel of $\mathbf{B}$ and in the span of the top $k+1$ singular vectors $\mathbf{v}_1, \ldots, \mathbf{v}_{k+1}$. Then we have

$$\|\mathbf{A}\mathbf{w}\| = \|(\mathbf{A} - \mathbf{B})\mathbf{w}\| \leq \sigma_1(\mathbf{A} - \mathbf{B})\|\mathbf{w}\| \,. \tag{8}$$

On the other hand, from the SVD $\mathbf{A} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$ we have

$$\begin{aligned}
\|\mathbf{A}\mathbf{w}\|^2 &= \left\| \sum_{i=1}^{k+1} \sigma_i \mathbf{u}_i (\mathbf{v}_i^\top \mathbf{w}) \right\|^2 \quad \text{since } \mathbf{w} \in \text{span}(\mathbf{v}_1, \ldots, \mathbf{v}_{k+1}) \\
&= \sum_{i=1}^{k+1} \sigma_i^2 (\mathbf{v}_i^\top \mathbf{w})^2 \\
&\geq \sigma_{k+1}^2 \sum_{i=1}^{k+1} (\mathbf{v}_i^\top \mathbf{w})^2 \\
&= \sigma_{k+1}^2 \|\mathbf{w}\|^2 \,.
\end{aligned}$$

We conclude that $\sigma_{k+1}(\mathbf{A})\|\mathbf{w}\| \leq \sigma_1(\mathbf{A} - \mathbf{B})\|\mathbf{w}\|$ and hence $\sigma_{k+1}(\mathbf{A}) \leq \sigma_1(\mathbf{A} - \mathbf{B})$.

Now we do the general case:

$$\begin{aligned}
\sigma_i(\mathbf{A} - \mathbf{B}) &= \sigma_i(\mathbf{A} - \mathbf{B}) + \sigma_1(\mathbf{B} - \mathbf{B}_k) \qquad \text{since } \mathbf{B} = \mathbf{B}_k \\
&= \sigma_1(\mathbf{A} - \mathbf{B} - (\mathbf{A} - \mathbf{B})_{i-1}) + \sigma_1(\mathbf{B} - \mathbf{B}_k) \qquad \text{see comments in Section 3.2} \\
&\geq \sigma_1(\mathbf{A} - \mathbf{B} - (\mathbf{A} - \mathbf{B})_{i-1} + \mathbf{B} - \mathbf{B}_k) \qquad \text{by the triangle inequality} \\
&= \sigma_1(\mathbf{A} - (\mathbf{A} - \mathbf{B})_{i-1} - \mathbf{B}_k) \qquad \text{algebraic simplification} \\
&\geq \sigma_{i+k}(\mathbf{A}) \qquad \text{by the previous case, since } \text{rank}((\mathbf{A} - \mathbf{B})_{i-1} + \mathbf{B}_k) \leq i + k - 1.
\end{aligned}$$

□

Proof of Theorem 3. We argue as follows:

$$\begin{aligned}
\|\mathbf{A} - \mathbf{A}_k\|_F^2 &= \sum_{i=k+1}^{r} \sigma_i(\mathbf{A})^2 \\
&\leq \sum_{i=1}^{r-k} \sigma_i(\mathbf{A} - \mathbf{B})^2 \qquad \text{by Lemma 1} \\
&\leq \|\mathbf{A} - \mathbf{B}\|_F^2 \,.
\end{aligned}$$

□