Linear models

1 Projections

The projection of a vector x onto a subspace $\mathcal{S}$ is the vector in $\mathcal{S}$ that is closest to x. In order to define this rigorously, we start by introducing the concept of direct sum. If two subspaces are disjoint, i.e. their only common point is the origin, then a vector that can be written as a sum of a vector from each subspace is said to belong to their direct sum.

Definition 1.1 (Direct sum)**.** *Let $\mathcal{V}$ be a vector space. For any subspaces $\mathcal{S}_1, \mathcal{S}_2 \subseteq \mathcal{V}$ such that*

$$\mathcal{S}_1 \cap \mathcal{S}_2 = \{0\} \tag{1}$$

the direct sum is defined as

$$\mathcal{S}_1 \oplus \mathcal{S}_2 := \{x \mid x = s_1 + s_2 \quad s_1 \in \mathcal{S}_1, s_2 \in \mathcal{S}_2\}\,. \tag{2}$$

The representation of a vector in the direct sum of two subspaces is unique.

Lemma 1.2. *Any vector $x \in \mathcal{S}_1 \oplus \mathcal{S}_2$ has a **unique** representation*

$$x = s_1 + s_2 \quad s_1 \in \mathcal{S}_1, s_2 \in \mathcal{S}_2. \tag{3}$$

Proof. If $x \in \mathcal{S}_1 \oplus \mathcal{S}_2$ then by definition there exist $s_1 \in \mathcal{S}_1, s_2 \in \mathcal{S}_2$ such that $x = s_1 + s_2$. Assume $x = s_1' + s_2'$, $s_1' \in \mathcal{S}_1, s_2' \in \mathcal{S}_2$, then $s_1 - s_1' = s_2 - s_2'$. This implies that $s_1 - s_1'$ and $s_2 - s_2'$ are in $\mathcal{S}_1$ and also in $\mathcal{S}_2$. However, $\mathcal{S}_1 \cap \mathcal{S}_2 = \{0\}$, so we conclude $s_1 = s_1'$ and $s_2 = s_2'$. □

We can now define the projection of a vector x onto a subspace $\mathcal{S}$ by separating the vector into a component that belongs to $\mathcal{S}$ and another that belongs to its orthogonal complement.

Definition 1.3 (Orthogonal projection)**.** *Let $\mathcal{V}$ be a vector space. The orthogonal projection of a vector $x \in \mathcal{V}$ onto a subspace $\mathcal{S} \subseteq \mathcal{V}$ is a vector denoted by $\mathcal{P}_\mathcal{S}\, x$ such that $x - \mathcal{P}_\mathcal{S}\, x \in \mathcal{S}^\perp$.*

Theorem 1.4 (Properties of orthogonal projections)**.** *Let $\mathcal{V}$ be a vector space. Every vector $x \in \mathcal{V}$ has a **unique** orthogonal projection $\mathcal{P}_\mathcal{S}\, x$ onto any subspace $\mathcal{S} \subseteq \mathcal{V}$ of finite dimension. In particular x can be expressed as*

$$x = \mathcal{P}_\mathcal{S}\, x + \mathcal{P}_{\mathcal{S}^\perp}\, x. \tag{4}$$

For any vector $s \in \mathcal{S}$

$$\langle x, s \rangle = \langle \mathcal{P}_{\mathcal{S}}\, x, s \rangle \,. \tag{5}$$

For any orthonormal basis $b_1, \ldots, b_m$ *of* $\mathcal{S}$*,*

$$\mathcal{P}_{\mathcal{S}}\, x = \sum_{i=1}^{m} \langle x, b_i \rangle \, b_i. \tag{6}$$

Proof. Since $\mathcal{V}$ has finite dimension, so does $\mathcal{S}$, which consequently has an orthonormal basis with finite cardinality $b'_1, \ldots, b'_m$ by Theorem 3.7 in Lecture Notes 8. Consider the vector

$$p := \sum_{i=1}^{m} \langle x, b'_i \rangle \, b'_i. \tag{7}$$

It turns out that $x - p$ is orthogonal to every vector in the basis. For $1 \leq j \leq m$,

$$\left\langle x - p, b'_j \right\rangle = \left\langle x - \sum_{i=1}^{m} \langle x, b'_i \rangle \, b'_i, b'_j \right\rangle \tag{8}$$

$$= \left\langle x, b'_j \right\rangle - \sum_{i=1}^{m} \langle x, b'_i \rangle \left\langle b'_i, b'_j \right\rangle \tag{9}$$

$$= \left\langle x, b'_j \right\rangle - \left\langle x, b'_j \right\rangle = 0, \tag{10}$$

so by Lemma 3.3 in Lecture Notes 8 $x - p \in \mathcal{S}^\perp$ and p is an orthogonal projection. Since $\mathcal{S} \cap \mathcal{S}^\perp = \{0\}$ [1] there cannot be two other vectors $x_1 \in \mathcal{S}, x_1 \in \mathcal{S}^\perp$ such that $x = x_1 + x_2$ so the orthogonal projection is unique.

Notice that $o := x - p$ is a vector in $\mathcal{S}^\perp$ such that $x - o = p$ is in S and therefore in $\left(\mathcal{S}^\perp\right)^\perp$. This implies that o is the orthogonal projection of x onto $\mathcal{S}^\perp$ and establishes (4).

Equation (5) follows immediately from the orthogonality of any vector $s \in \mathcal{S}$ and $\mathcal{P}_{\mathcal{S}}\, x$.

Equation (6) follows from (5) and Lemma 3.5 in Lecture Notes 8. □

Computing the norm of the projection of a vector onto a subspace is easy if we have access to an orthonormal basis (as long as the norm is induced by the inner product).

Lemma 1.5 (Norm of the projection)**.** *The norm of the projection of an arbitrary vector* $x \in \mathcal{V}$ *onto a subspace* $\mathcal{S} \subseteq \mathcal{V}$ *of dimension* d *can be written as*

$$||\mathcal{P}_{\mathcal{S}}\, x||_{\langle \cdot, \cdot \rangle} = \sqrt{\sum_{i}^{d} \langle b_i, x \rangle^2} \tag{11}$$

for any orthonormal basis $b_1, \ldots, b_d$ *of* $\mathcal{S}$*.*

[1] For any vector v that belongs to both $\mathcal{S}$ and $\mathcal{S}^\perp$ $\langle v, v \rangle = ||v||_2^2 = 0$, which implies $v = 0$.

Proof. By (6)

$$||\mathcal{P}_{\mathcal{S}}\, x||^2_{\langle\cdot,\cdot\rangle} = \langle \mathcal{P}_{\mathcal{S}}\, x, \mathcal{P}_{\mathcal{S}}\, x\rangle \tag{12}$$

$$= \left\langle \sum_i^d \langle b_i, x\rangle\, b_i, \sum_j^d \langle b_j, x\rangle\, b_j \right\rangle \tag{13}$$

$$= \sum_i^d \sum_j^d \langle b_i, x\rangle \langle b_j, x\rangle \langle b_i, b_j\rangle \tag{14}$$

$$= \sum_i^d \langle b_i, x\rangle^2 . \tag{15}$$

□

Finally, we prove indeed that of a vector x onto a subspace $\mathcal{S}$ is the vector in $\mathcal{S}$ that is closest to x in the distance induced by the inner-product norm.

Example 1.6 (Projection onto a one-dimensional subspace)**.** To compute the projection of a vector x onto a one-dimensional subspace spanned by a vector v, we use the fact that $\left\{v/\,||v||_{\langle\cdot,\cdot\rangle}\right\}$ is a basis for span (v) (it is a set containing a unit vector that spans the subspace) and apply (6) to obtain

$$\mathcal{P}_{\text{span}(v)}\, x = \frac{\langle v, x\rangle}{||v||^2_{\langle\cdot,\cdot\rangle}} v. \tag{16}$$

Theorem 1.7 (The orthogonal projection is closest)**.** *The orthogonal projection of a vector x onto a subspace $\mathcal{S}$ belonging to the same inner-product space is the closest vector to x that belongs to $\mathcal{S}$ in terms of the norm induced by the inner product. More formally, $\mathcal{P}_{\mathcal{S}}\, x$ is the solution to the optimization problem*

$$\underset{u}{\text{minimize}} \qquad ||x - u||_{\langle\cdot,\cdot\rangle} \tag{17}$$

$$\text{subject to} \qquad u \in \mathcal{S}. \tag{18}$$

Proof. Take any point $s \in \mathcal{S}$ such that $s \neq \mathcal{P}_{\mathcal{S}}\, x$

$$||x - s||^2_{\langle\cdot,\cdot\rangle} = ||\mathcal{P}_{\mathcal{S}^\perp}\, x + \mathcal{P}_{\mathcal{S}}\, x - s||^2_{\langle\cdot,\cdot\rangle} \tag{19}$$

$$= ||\mathcal{P}_{\mathcal{S}^\perp}\, x||^2_{\langle\cdot,\cdot\rangle} + ||\mathcal{P}_{\mathcal{S}}\, x - s||^2_{\langle\cdot,\cdot\rangle} \tag{20}$$

$$> 0 \quad \text{because } s \neq \mathcal{P}_{\mathcal{S}}\, x, \tag{21}$$

where (20) follows from the Pythagorean theorem since because $\mathcal{P}_{\mathcal{S}^\perp}\, x$ belongs to $S^\perp$ and $\mathcal{P}_{\mathcal{S}}\, x - s$ to S. □

2 Linear minimum-MSE estimation

We are interested in estimating the sample of a continuous random variable X from the sample y of a random variable Y. If we know the joint distribution of X and Y then the optimal estimator in terms of MSE is the conditional mean $\mathrm{E}\left(X|Y=y\right)$. However often it is very challenging to completely characterize a joint distribution between two quantities, but it is more tractable to obtain an estimate of their first and second order moments. In this case it turns out that we can obtain the optimal *linear* estimate of X given Y by using our knowledge of linear algebra.

Theorem 2.1 (Best linear estimator)**.** *Assume that we know the means μ_X, μ_Y and variances σ_X^2, σ_Y^2 of two random variables X and Y and their correlation coefficient ρ_{XY}. The best linear estimate of the form $aY+b$ of X given Y in terms of mean-square error is*

$$g_{LMMSE}\left(y\right) = \frac{\rho_{XY}\,\sigma_X\left(y-\mu_Y\right)}{\sigma_Y} + \mu_X. \tag{22}$$

Proof. First we determine the value of b. The cost function as a function of b is

$$\begin{aligned} h\left(b\right) = \mathrm{E}\left(\left(X-aY-b\right)^2\right) &= \mathrm{E}\left(\left(X-aY\right)^2\right) + b^2 - 2b\mathrm{E}\left(X-aY\right) &(23)\\ &= \mathrm{E}\left(\left(X-aY\right)^2\right) + b^2 - 2b\left(\mu_X - a\mu_Y\right). &(24)\end{aligned}$$

The first and second derivative with respect to b are

$$\begin{aligned} h'\left(b\right) &= 2b - 2\left(\mu_X - a\mu_Y\right), &(25)\\ h''\left(b\right) &= 2. &(26)\end{aligned}$$

Since h'' is positive the function is convex so the minimum is obtained by setting $h'\left(b\right)$ to zero, which yields

$$b = \mu_X - a\mu_Y. \tag{27}$$

Consider the centered random variables $\tilde{X} := X - \mu_X$ and $\tilde{Y} := Y - \mu_Y$. It turns out that to find a we just need to find the best estimate of $\tilde{X}$ of the form $a\tilde{Y}$

$$\begin{aligned} \mathrm{E}\left(\left(X-aY-b\right)^2\right) &= \mathrm{E}\left(\left(X-\mu_X-a\left(Y-\mu_Y\right)+\mu_X-a\mu_Y-b\right)\right)^2 &(28)\\ &= \mathrm{E}\left(\left(\tilde{X}-a\tilde{Y}\right)^2\right), &(29)\end{aligned}$$

clearly any a that minimizes the left-hand side also minimizes the right-hand side and vice versa.

Consider the vector space of zero-mean random variables. $\tilde{X}$ and $\tilde{Y}$ belong to this vector space. In fact,

$$\left\langle \tilde{X}, \tilde{Y} \right\rangle = \mathrm{E}\left(\tilde{X}\tilde{Y}\right) \tag{30}$$

$$= \mathrm{Cov}\left(X, Y\right) \tag{31}$$

$$= \sigma_X \, \sigma_Y \, \rho_{XY}, \tag{32}$$

$$\left|\left|\tilde{X}\right|\right|^2_{\langle\cdot,\cdot\rangle} = \mathrm{E}\left(\tilde{X}^2\right) \tag{33}$$

$$= \sigma_X^2, \tag{34}$$

$$\left|\left|\tilde{Y}\right|\right|^2_{\langle\cdot,\cdot\rangle} = \mathrm{E}\left(\tilde{Y}^2\right) \tag{35}$$

$$= \sigma_Y^2. \tag{36}$$

Any random variable of the form $a\tilde{Y}$ belongs to the subspace spanned by $\tilde{Y}$. Since the distance in this vector space is induced by the mean-square norm, by Theorem 1.7 the vector of the form $a\tilde{Y}$ that approximates $\tilde{X}$ better is just the projection of $\tilde{X}$ onto the subspace spanned by $\tilde{Y}$, which we will denote by $\mathcal{P}_{\tilde{Y}}\tilde{X}$. This subspace has dimension 1, so $\left\{\tilde{Y}/\sigma_Y\right\}$ is a basis for the subspace. The projection is consequently equal to

$$\mathcal{P}_{\tilde{Y}}\tilde{X} = \left\langle \tilde{X}, \frac{\tilde{Y}}{\sigma_Y} \right\rangle \frac{\tilde{Y}}{\sigma_Y} \tag{37}$$

$$= \left\langle \tilde{X}, \tilde{Y} \right\rangle \frac{\tilde{Y}}{\sigma_Y^2} \tag{38}$$

$$= \frac{\sigma_X \rho_{XY} \tilde{Y}}{\sigma_Y}. \tag{39}$$

So $a = \sigma_X \rho_{XY} / \sigma_Y$, which concludes the proof. □

In words, the linear estimator of X given Y is obtained by

1. centering Y by removing its mean,
2. normalizing $\tilde{Y}$ by dividing by its standard deviation,
3. scaling the result using the correlation between Y and X,
4. scaling again using the standard deviation of X,
5. recentering by adding the mean of X.

3 Matrices

A **matrix** is a rectangular array of numbers. We denote the vector space of $m \times n$ matrices by $\mathbb{R}^{m \times n}$. We denote the ith row of a matrix A by $A_{i:}$, the jth column by $A_{:j}$ and the (i, j) entry by A_{ij}. The transpose of a matrix is obtained by switching its rows and columns.

Definition 3.1 (Transpose)**.** *The* ***transpose*** A^T *of a matrix* $A \in \mathbb{R}^{m \times n}$ *is a matrix in* $A \in \mathbb{R}^{m \times n}$

$$\left(A^T\right)_{ij} = A_{ji}. \tag{40}$$

A **symmetric** matrix is a matrix that is equal to its transpose.

Matrices map vectors to other vectors through a linear operation called matrix-vector product.

Definition 3.2 (Matrix-vector product)**.** *The product of a matrix* $A \in \mathbb{R}^{m \times n}$ *and a vector* $x \in \mathbb{R}^n$ *is a vector in* $Ax \in \mathbb{R}^n$*, such that*

$$\begin{aligned} (Ax)_i &= \sum_{j=1}^{n} A_{ij} x\left[j\right] && (41) \\ &= \left\langle A_{i:}, x \right\rangle, && (42) \end{aligned}$$

i.e. the i*th entry of* Ax *is the dot product between the* i*th row of* A *and* x*.*

Equivalently,

$$Ax = \sum_{j=1}^{n} A_{:j} x\left[j\right], \tag{43}$$

i.e. Ax *is a linear combination of the columns of* A *weighted by the entries in* x*.*

One can easily check that the transpose of the product of two matrices A and B is equal to the the transposes multiplied in the inverse order,

$$(AB)^T = B^T A^T. \tag{44}$$

We can now we can express the dot product between two vectors x and y as

$$\langle x, y \rangle = x^T y = y^T x. \tag{45}$$

The identity matrix is a matrix that maps any vector to itself.

Definition 3.3 (Identity matrix)**.** *The identity matrix in* $\mathbb{R}^{n\times n}$ *is*

$$I = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & 1 \end{bmatrix}. \tag{46}$$

Clearly, for any $x \in \mathbb{R}^n$ $Ix = x$*.*

Definition 3.4 (Matrix multiplication)**.** *The product of two matrices* $A \in \mathbb{R}^{m\times n}$ *and* $B \in \mathbb{R}^{n\times p}$ *is a matrix* $AB \in \mathbb{R}^{m\times p}$*, such that*

$$(AB)_{ij} = \sum_{k=1}^{n} A_{ik}B_{k_j} = \langle A_{i:}, B_{:,j}\rangle, \tag{47}$$

i.e. the (i, j) *entry of* AB *is the dot product between the* i*th row of* A *and the* j*th column of* B*.*

Equivalently, the j*th column of* AB *is the result of multiplying* A *and the* j*th column of* B

$$AB = \sum_{k=1}^{n} A_{ik}B_{k_j} = \langle A_{i:}, B_{:,j}\rangle, \tag{48}$$

and i*th row of* AB *is the result of multiplying the* i*th row of* A *and* B*.*

Square matrices may have an inverse. If they do, the inverse is a matrix that reverses the effect of the matrix of any vector.

Definition 3.5 (Matrix inverse)**.** *The inverse of a square matrix* $A \in \mathbb{R}^{n\times n}$ *is a matrix* $A^{-1} \in \mathbb{R}^{n\times n}$ *such that*

$$AA^{-1} = A^{-1}A = \mathrm{I}. \tag{49}$$

Lemma 3.6. *The inverse of a matrix is unique.*

Proof. Let us assume there is another matrix M such that $AM = \mathrm{I}$, then

$$M = A^{-1}AM \quad \text{by (49)} \tag{50}$$
$$= A^{-1}. \tag{51}$$

□

An important class of matrices are **orthogonal matrices**.

Definition 3.7 (Orthogonal matrix)**.** *An orthogonal matrix is a square matrix such that its inverse is equal to its transpose,*

$$U^T U = UU^T = \mathrm{I} \tag{52}$$

By definition, the columns $U_{:1}, U_{:2}, \ldots, U_{:n}$ of any orthogonal matrix have unit norm and orthogonal to each other, so they form an orthonormal basis (it's somewhat confusing that orthogonal matrices are not called orthonormal matrices instead). We can interpret applying U^T to a vector x as computing the coefficients of its representation in the basis formed by the columns of U. Applying U to $U^T x$ recovers x by scaling each basis vector with the corresponding coefficient:

$$x = UU^T x = \sum_{i=1}^{n} \langle U_{:i}, x \rangle \, U_{:i}. \tag{53}$$

Applying an orthogonal matrix to a vector does not affect its norm, it just rotates the vector.

Lemma 3.8 (Orthogonal matrices preserve the norm)**.** *For any orthogonal matrix $U \in \mathbb{R}^{n \times n}$ and any vector $x \in \mathbb{R}^n$,*

$$||Ux||_2 = ||x||_2 \,. \tag{54}$$

Proof. By the definition of an orthogonal matrix

$$||Ux||_2^2 = x^T U^T U x \tag{55}$$

$$= x^T x \tag{56}$$

$$= ||x||_2^2 \,. \tag{57}$$

□

4 Eigendecomposition

An **eigenvector** v of A satisfies

$$Av = \lambda v \tag{58}$$

for a real number λ which is the corresponding **eigenvalue**. Even if A is real, its eigenvectors and eigenvalues can be complex.

Lemma 4.1 (Eigendecomposition)**.** *If a square matrix $A \in \mathbb{R}^{n \times n}$ has n linearly independent eigenvectors $v_1, \ldots, v_n$ with eigenvalues $\lambda_1, \ldots, \lambda_n$ it can be expressed in terms of a matrix Q, whose columns are the eigenvectors, and a diagonal matrix containing the eigenvalues,*

$$A = \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix}^{-1} \tag{59}$$

$$= Q \Lambda Q^{-1} \tag{60}$$

Proof.

$$AQ = \begin{bmatrix} Av_1 & Av_2 & \cdots & Av_n \end{bmatrix} \tag{61}$$

$$= \begin{bmatrix} \lambda_1 v_1 & \lambda_2 v_2 & \cdots & \lambda_2 v_n \end{bmatrix} \tag{62}$$

$$= Q\Lambda. \tag{63}$$

As we will establish later on, if the columns of a square matrix are all linearly independent, then the matrix has an inverse, so multiplying the expression by Q^{-1} on both sides completes the proof. □

Lemma 4.2. *Not all matrices have an eigendecomposition*

Proof. Consider for example the matrix

$$\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}. \tag{64}$$

Assume λ has a nonzero eigenvalue corresponding to an eigenvector with entries v_1 and v_2, then

$$\begin{bmatrix} v_2 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} \lambda v_1 \\ \lambda v_2 \end{bmatrix}, \tag{65}$$

which implies that $v_2 = 0$ and hence $v_1 = 0$, since we have assumed that $\lambda \neq 0$. This implies that the matrix does not have nonzero eigenvalues associated to nonzero eigenvectors. □

An interesting use of the eigendecomposition is computing successive matrix products very fast. Assume that we want to compute

$$AA \cdots Ax = A^k x, \tag{66}$$

i.e. we want to apply A to x k times. A^k *cannot* be computed by taking the power of its entries (try out a simple example to convince yourself). However, if A has an eigendecomposition,

$$A^k = Q\Lambda Q^{-1} Q\Lambda Q^{-1} \cdots Q\Lambda Q^{-1} \tag{67}$$

$$= Q\Lambda^k Q^{-1} \tag{68}$$

$$= Q \begin{bmatrix} \lambda_1^k & 0 & \cdots & 0 \\ 0 & \lambda_2^k & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & \lambda_n^k \end{bmatrix} Q^{-1}, \tag{69}$$

using the fact that for diagonal matrices applying the matrix repeatedly is equivalent to taking the power of the diagonal entries. This allows to compute the k matrix products using just 3 matrix products and taking the power of n numbers.

From high-school or undergraduate algebra you probably remember how to compute eigenvectors using determinants. In practice, this is usually not a viable option due to stability issues. A popular technique to compute eigenvectors is based on the following insight. Let $A \in \mathbb{R}^{n \times n}$ be a matrix with eigendecomposition $Q\Lambda Q^{-1}$ and let x be an arbitrary vector in $\mathbb{R}^n$. Since the columns of Q are linearly independent, they form a basis for $\mathbb{R}^n$, so we can represent x as

$$x = \sum_{i=1}^{n} \alpha_i Q_{:i}, \quad \alpha_i \in \mathbb{R},\ 1 \leq i \leq n. \tag{70}$$

Now let us apply A to x k times,

$$A^k x = \sum_{i=1}^{n} \alpha_i A^k Q_{:i} \tag{71}$$

$$= \sum_{i=1}^{n} \alpha_i \lambda_i^k Q_{:i}. \tag{72}$$

If we assume that the eigenvectors are ordered according to their magnitudes and that the magnitude of one of them is larger than the rest, $|\lambda_1| > |\lambda_2| \geq \ldots$, and that $\alpha_1 \neq 0$ (which happens with high probability if we draw a random x) then as k grows larger the term $\alpha_1 \lambda_1^k Q_{:1}$ dominates. The term will blow up or tend to zero unless we normalize every time before applying A. Adding the normalization step to this procedure results in the **power method** or power iteration, an algorithm of great importance in numerical linear algebra.

Algorithm 4.3 (Power method)**.**
Input*: A matrix A.*
Output*: An estimate of the eigenvector of A corresponding to the largest eigenvalue.*

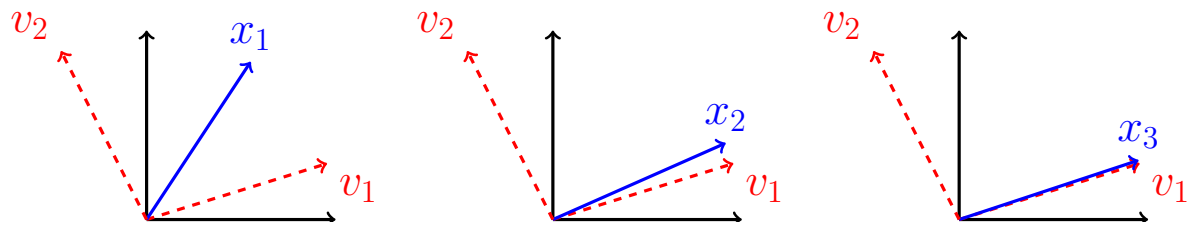

Figure 1: Illustration of the first three iterations of the power method for a matrix with eigenvectors v_1 and v_2, whose corresponding eigenvalues are $\lambda_1 = 1.05$ and $\lambda_2 = 0.1661$.

Initialization*: Set* $x_1 := x/\left|\left|x\right|\right|_2$*, where* x *contains random entries.*
For $i = 1, \ldots, k$*, compute*

$$x_i := \frac{Ax_{i-1}}{\left|\left|Ax_{i-1}\right|\right|_2}. \tag{73}$$

Figure 1 illustrates the power method on a simple example, where the matrix– which was just drawn at random– is equal to

$$A = \begin{bmatrix} 0.930 & 0.388 \\ 0.237 & 0.286 \end{bmatrix}. \tag{74}$$

The convergence to the eigenvector corresponding to the eigenvalue with the largest magnitude is very fast.

5 Time-homogeneous Markov chains

A **Markov chain** is a sequence of discrete random variables $X_0, X_1, \ldots$ such that

$$p_{X_{k+1}|X_0,X_1,\ldots,X_k}\left(x_{k+1}|x_0, x_1, \ldots, x_k\right) = p_{X_{k+1}|X_k}\left(x_{k+1}|x_k\right). \tag{75}$$

In words, X_{k+1} is conditionally independent of X_j for $j \leq k-1$ conditioned on X_k. If the value of the random variables is restricted to a finite set $\{\alpha_1, \ldots, \alpha_n\}$ with probability one, the Markov chain is said to be **time homogeneous**. More formally,

$$P_{ij} := p_{X_{k+1}|X_k}\left(\alpha_i|\alpha_j\right) \tag{76}$$

only depends on i and j, **not** on k for all $1 \leq i, j \leq n, k \geq 0$.

If a Markov chain is time homogeneous we can group the transition probabilities P_{ij} in a **transition matrix** P. We express the pmf of X_k restricted to the set where it can be

nonzero as a vector,

$$\pi_k = \begin{bmatrix} p_{X_k}(\alpha_1) \\ p_{X_k}(\alpha_2) \\ \cdots \\ p_{X_k}(\alpha_n) \end{bmatrix}. \tag{77}$$

By the Chain Rule, the pmf of X_k can be computed from the pmf of X_0 using the transition matrix,

$$\pi_k = P\,P \cdots P\,\pi_0 = P^k\,\pi_0. \tag{78}$$

In some cases, no matter how we initialize the Markov chain, the Markov Chain *forgets* its initial state and converges to a **stationary distribution**. This is exploited in Markov-Chain Monte Carlo methods that allow to sample from arbitrary distributions by building the corresponding Markov chain. These methods are very useful in Bayesian statistics.

A Markov Chain that converges to a stationary distribution π_∞ is said to be **ergodic**. Note that necessarily $P\pi_\infty = \pi_\infty$, so that π_∞ is an eigenvector of the transition matrix with a corresponding eigenvalue equal to one.

Conversely, let a transition matrix P of a Markov chain have a valid eigendecomposition with n linearly independent eigenvectors $v_1, v_2, \ldots$ and corresponding eigenvalues $\lambda_1 > \lambda_2 \geq \lambda_3 \ldots$. If the eigenvector corresponding to the largest eigenvalue has non-negative entries then

$$\pi_\infty := \frac{v_1}{\sum_{i=1}^n v_1(i)} \tag{79}$$

is a valid pmf and

$$P\pi_\infty = \lambda_1 \pi_\infty \tag{80}$$

is also a valid pmf, which is only possible if the largest eigenvalue λ_1 equals one. Now, if we represent any possible initial pmf π_0 in the basis formed by the eigenvectors of P we have

$$\pi_0 = \sum_{i=1}^n \alpha_i v_i, \quad \alpha_i \in \mathbb{R},\ 1 \leq i \leq n, \tag{81}$$

and

$$\pi_k = P^k \pi_0 \tag{82}$$

$$= \sum_{i=1}^n \alpha_i P^k v_i \tag{83}$$

$$= \alpha_1 \lambda_1 v_1 + \sum_{i=2}^n \alpha_i \lambda_i^k v_i. \tag{84}$$

Since the rest of eigenvalues are strictly smaller than one,

$$\lim_{k\to\infty} \pi_k = \alpha_1 \lambda_1 v_1 = \pi_\infty \tag{85}$$

where the last equality follows from the fact that the sequence of π_k all belong to the closed set $\{\pi \mid \sum_i^n \pi(i) = 1\}$ so the limit also belongs to the set and hence is a valid pmf. We refer the interested reader to more advanced texts treating Markov chains for further details.

6 Eigendecomposition of symmetric matrices

The following lemma shows that eigenvectors of a symmetric matrix corresponding to different nonzero eigenvalues are necessarily orthogonal.

Lemma 6.1. *If $A \in \mathbb{R}^{n\times n}$ is symmetric, then if u_i and u_j are eigenvectors of A corresponding to different nonzero eigenvalues $\lambda_i \neq \lambda_j \neq 0$*

$$u_i^T u_j = 0. \tag{86}$$

Proof. Since $A = A^T$

$$u_i^T u_j = \frac{1}{\lambda_i} (Au_i)^T u_j \tag{87}$$

$$= \frac{1}{\lambda_i} u_i^T A^T u_j \tag{88}$$

$$= \frac{1}{\lambda_i} u_i^T A u_j \tag{89}$$

$$= \frac{\lambda_j}{\lambda_i} u_i^T u_j. \tag{90}$$

This is only possible if $u_i^T u_j = 0$. □

It turns out that every $n \times n$ symmetric matrix has n linearly independent vectors. The proof of this is beyond the scope of these notes. An important consequence is that **all** symmetric matrices have an eigendecomposition of the form

$$A = UDU^T \tag{91}$$

where $U = \begin{bmatrix} u_1 & u_2 & \cdots & u_n \end{bmatrix}$ is an orthogonal matrix.

The eigenvalues of a symmetric matrix $\lambda_1, \lambda_2, \ldots, \lambda_n$ can be positive, negative or zero. They determine the value of the **quadratic form**:

$$q(x) := x^T A x = \sum_{i=1}^n \lambda_i \left(x^T u_i\right)^2 \tag{92}$$

If we order the eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ then the first eigenvalue is the maximum value attained by the quadratic if its input has unit ℓ_2 norm, the second eigenvalue is the maximum value attained by the quadratic form if we restrict its argument to be normalized and orthogonal to the first eigenvector, and so on.

Theorem 6.2. *For any symmetric matrix $A \in \mathbb{R}^n$ with normalized eigenvectors $u_1, u_2, \ldots, u_n$ with corresponding eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$*

$$\lambda_1 = \max_{||u||_2=1} u^T A u, \tag{93}$$

$$u_1 = \arg\max_{||u||_2=1} u^T A u, \tag{94}$$

$$\lambda_k = \max_{||u||_2=1, u \perp u_1, \ldots, u_{k-1}} u^T A u, \tag{95}$$

$$u_k = \arg\max_{||u||_2=1, u \perp u_1, \ldots, u_{k-1}} u^T A u. \tag{96}$$

The theorem is proved in Section A.1 of the appendix.

7 The singular-value decomposition

If we consider the columns and rows of a matrix as sets of vectors then we can study their respective spans.

Definition 7.1 (Row and column space)**.** *The **row space*** $\text{row}(A)$ *of a matrix A is the span of its rows. The **column space*** $\text{col}(A)$ *is the span of its columns.*

It turns out that the row space and the column space of any matrix have the same dimension. We name this quantity the **rank** of the matrix.

Theorem 7.2. *The rank is well defined;*

$$dim\left(\text{col}\left(A\right)\right) = dim\left(\text{row}\left(A\right)\right). \tag{97}$$

Section A.2 of the appendix contains the proof.

The following theorem states that we can decompose any real matrix into the product of orthogonal matrices containing bases for its row and column space and a diagonal matrix with a positive diagonal. It is a fundamental result in linear algebra, but its proof is beyond the scope of these notes.

Theorem 7.3. *Without loss of generality let $m \leq n$. Every rank r real matrix $A \in R^{m \times n}$ has a unique singular-value decomposition of the form (SVD)*

$$A = \begin{bmatrix} u_1 & u_2 & \cdots & u_m \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & \sigma_m \end{bmatrix} \begin{bmatrix} v_1^T \\ v_2^T \\ \cdots \\ v_m^T \end{bmatrix} \tag{98}$$

$$= USV^T, \tag{99}$$

*where the **singular values** $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_m \geq 0$ are nonnegative real numbers, the matrix $U \in R^{m \times m}$ containing the **left singular vectors** is orthogonal, and the matrix $V \in R^{m \times n}$ containing the **right singular vectors** is a submatrix of an orthogonal matrix (i.e. its columns form an orthonormal set).*

Note that we can write the matrix as a sum of rank-1 matrices

$$A = \sum_{i=1}^{m} \sigma_i u_i v_i^T \tag{100}$$

$$= \sum_{i=1}^{r} \sigma_i u_i v_i^T, \tag{101}$$

where r is the number of nonzero singular values. The first r left singular vectors $u_1, u_2, \ldots, u_r \in \mathbb{R}^m$ form an orthonormal basis of the column space of A and the first r right singular vectors $v_1, v_2, \ldots, v_r \in \mathbb{R}^n$ form an orthonormal basis of the row space of A. Therefore the rank of the matrix is equal to r.

8 Principal component analysis

The goal of **dimensionality-reduction** methods is to project high-dimensional data onto a lower-dimensional space while preserving as much information as possible. These methods are a basic tool in data analysis; some applications include visualization (especially if we project onto $\mathbb{R}^2$ or $\mathbb{R}^3$), denoising and increasing computational efficiency. **Principal component analysis** (PCA) is a *linear* dimensionality-reduction technique based on the SVD.

If we interpret a set of data vectors as belonging to an ambient vector space, applying PCA allows to find directions in this space along which the data have a high variation. This is achieved by centering the data and then extracting the singular vectors corresponding to the largest singular values. The next two sections provide a geometric and a probabilistic justification.

Algorithm 8.1 (Principal component analysis)**.**
Input*: n data vectors $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n \in \mathbb{R}^m$, a number $k \leq \min\{m, n\}$.*
Output*: The first k **principal components**, a set of orthonormal vectors of dimension m.*

1. *Center the data. Compute*

$$x_i = \tilde{x}_i - \frac{1}{n} \sum_{i=1}^{n} \tilde{x}_i, \tag{102}$$

   for $1 \leq i \leq n$.

2. *Group the centered data in a data matrix X*

$$X = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}. \tag{103}$$

   Compute the SVD of X and extract the left singular vectors corresponding to the k largest singular values. These are the first k principal components.

8.1 PCA: Geometric interpretation

Once the data are centered, the energy of the projection of the data points onto different directions in the ambient space reflects the variation of the dataset along those directions. PCA selects the directions that maximize the ℓ_2 norm of the projection and are mutually orthogonal. By Lemma 1.5, the sum of the squared ℓ_2 norms of the projection of the centered data $x_1, x_2, \ldots, x_n$ onto a 1D subspace spanned by a unit-norm vector u can be expressed as

$$\sum_{i=1}^{n} \left|\left| \mathcal{P}_{\text{span}(u)}\, x_i \right|\right|_2^2 = \sum_{i=1}^{n} u^T x_i x_i^T u \tag{104}$$

$$= u^T X X^T u \tag{105}$$

$$= \left|\left| X^T u \right|\right|_2^2. \tag{106}$$

If we want to maximize the energy of the projection onto a subspace of dimension k, an option is to choose orthogonal 1D projections sequentially. First we choose a unit vector u_1 that maximizes $\left|\left| X^T u \right|\right|_2^2$ and is consequently the 1D subspace that is better adapted to the data. Then, we choose a second unit vector u *orthogonal* to the first which maximizes $\left|\left| X^T u \right|\right|_2^2$ and hence is the 1D subspace that is better adapted to the data while being in the orthogonal complement of u_1. We repeat this procedure until we have k orthogonal directions. This is exactly equivalent to performing PCA, as proved in Theorem 8.2 below. The k directions correspond to the first k principal components. Figure 2 provides an example in 2D. Note how the singular values are proportional to the energy that lies in the direction of the corresponding principal component.

$\sigma_1/\sqrt{n} = 0.705,$
$\sigma_2/\sqrt{n} = 0.690$

$\sigma_1/\sqrt{n} = 0.9832,$
$\sigma_2/\sqrt{n} = 0.3559$

$\sigma_1/\sqrt{n} = 1.3490,$
$\sigma_2/\sqrt{n} = 0.1438$

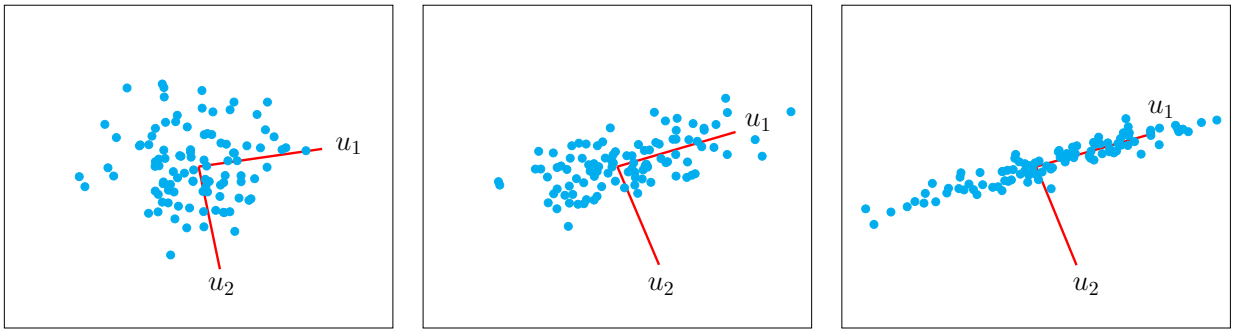

Figure 2: PCA of a dataset with $n = 100$ 2D vectors with different configurations. The two first singular values reflect how much energy is preserved by projecting onto the two first principal components.

Theorem 8.2. *For any matrix $X \in \mathbb{R}^{m \times n}$, where $n > m$, with left singular vectors $u_1, u_2, \ldots, u_m$ corresponding to the nonzero singular values $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_m$,*

$$\sigma_1 = \max_{||u||_2=1} \left|\left|X^T u\right|\right|_2, \tag{107}$$

$$u_1 = \arg \max_{||u||_2=1} \left|\left|X^T u\right|\right|_2, \tag{108}$$

$$\sigma_k = \max_{\substack{||u||_2=1 \\ u \perp u_1,\ldots,u_{k-1}}} \left|\left|X^T u\right|\right|_2, \quad 2 \leq k \leq r, \tag{109}$$

$$u_k = \arg \max_{\substack{||u||_2=1 \\ u \perp u_1,\ldots,u_{k-1}}} \left|\left|X^T u\right|\right|_2, \quad 2 \leq k \leq r. \tag{110}$$

Proof. If the SVD of X is USV^T then the eigendecomposition of XX^T is equal to

$$XX^T = USV^T VSU^T = US^2U^T, \tag{111}$$

where $V^T V = \mathrm{I}$ because $n > m$ and the matrix has m nonzero singular values. S^2 is a diagonal matrix containing $\sigma_1^2 \geq \sigma_2^2 \geq \ldots \geq \sigma_m^2$ in its diagonal.

The result now follows from applying Theorem 6.2 to the quadratic form

$$uXX^T u = \left|\left|X^T u\right|\right|_2. \tag{112}$$

□

This result shows that PCA is equivalent to choosing the *best* (in terms of ℓ_2 norm) k 1D subspaces following a *greedy* procedure, since at each step we choose the best 1D subspace

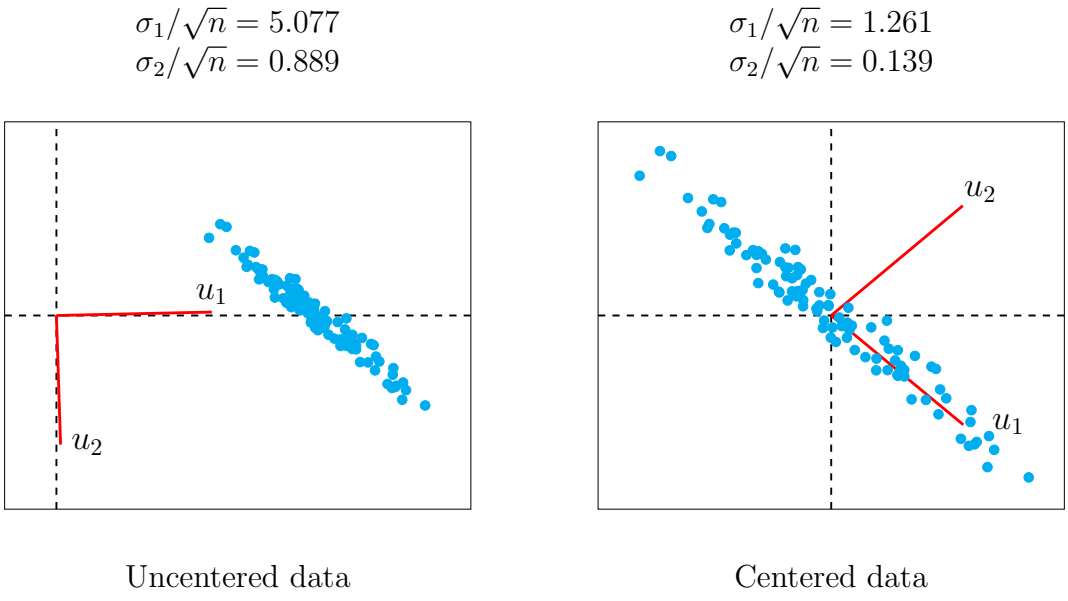

Figure 3: PCA applied to $n = 100$ 2D data points. On the left the data are not centered. As a result the dominant principal component u_1 lies in the direction of the mean of the data and PCA does not reflect the actual structure. Once we center, u_1 becomes aligned with the direction of maximal variation.

orthogonal to the previous ones. A natural question to ask is whether this method produces the best k-dimensional subspace. A priori this is not necessarily the case; many greedy algorithms produce suboptimal results. However, in this case the greedy procedure is indeed optimal: the subspace spanned by the first k principal components is the *best* subspace we can choose in terms of the ℓ_2-norm of the projections. The theorem is proved in Section A.3 of the appendix.

Theorem 8.3. *For any matrix $X \in \mathbb{R}^{m \times n}$ with left singular vectors $u_1, u_2, \ldots, u_m$ corresponding to the nonzero singular values $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_m$,*

$$\sum_{i=1}^{n} \left|\left|\mathcal{P}_{\text{span}(u_1,u_2,\ldots,u_k)}\, x_i\right|\right|_2^2 \geq \sum_{i=1}^{n} ||\mathcal{P}_{\mathcal{S}}\, x_i||_2^2\,, \tag{113}$$

for any subspace $\mathcal{S}$ of dimension $k \leq \min\{m, n\}$.

Figure 3 illustrates the importance of centering before applying PCA. Theorems 8.2 and 8.3 still hold if the data are not centered. However, the norm of the projection onto a certain direction no longer reflects the variation of the data. In fact, if the data are concentrated around a point that is far from the origin, the first principal component will tend be aligned in that direction. This makes sense as projecting onto that direction captures more energy. As a result, the principal components do not capture the directions of maximum variation *within* the cloud of data.

8.2 PCA: Probabilistic interpretation

Let us interpret our data, $x_1, x_2, \ldots, x_n$ in $\mathbb{R}^m$, as samples of a random vector $\mathbf{X}$ of dimension m. Recall that we are interested in determining the directions of maximum variation of the data in ambient space. In probabilistic terms, we want to find the directions in which the data have *higher variance*. The covariance matrix of the data provides this information. In fact, we can use it to determine the variance of the data in any direction.

Lemma 8.4. *Let u be a unit vector,*

$$\operatorname{Var}\left(\boldsymbol{X}^T u\right) = u^T \Sigma_{\boldsymbol{X}} u. \tag{114}$$

Proof.

$$\begin{aligned}
\operatorname{Var}\left(\mathbf{X}^T u\right) &= \mathrm{E}\left(\left(\mathbf{X}^T u\right)^2\right) - \mathrm{E}^2\left(\mathbf{X}^T u\right) && (115)\\
&= \mathrm{E}\left(u\mathbf{X}\mathbf{X}^T u\right) - \mathrm{E}\left(u^T \mathbf{X}\right)\mathrm{E}\left(\mathbf{X}^T u\right) && (116)\\
&= u^T\left(\mathrm{E}\left(\mathbf{X}\mathbf{X}^T\right) - \mathrm{E}\left(\mathbf{X}\right)\mathrm{E}\left(\mathbf{X}\right)^T\right) u && (117)\\
&= u^T \Sigma_{\mathbf{X}} u. && (118)
\end{aligned}$$

□

Of course, if we only have access to samples of the random vector, we do not know the covariance matrix of the vector. However we can approximate it using the **empirical covariance matrix**.

Definition 8.5 (Empirical covariance matrix)**.** *The empirical covariance of the vectors $x_1, x_2, \ldots, x_n$ in $\mathbb{R}^m$ is equal to*

$$\begin{aligned}
\overline{\Sigma}_n &:= \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}_n\right)\left(x_i - \overline{x}_n\right)^T && (119)\\
&= \frac{1}{n} X X^T, && (120)
\end{aligned}$$

where $\overline{x}_n$ is the sample mean, as defined in Definition 1.3 of Lecture Notes 4, and X is the matrix containing the centered data as defined in (103)*.*

If we assume that the mean of the data is zero (i.e. that the data have been centered using the true mean), then the empirical covariance is an unbiased estimator of the true covariance matrix:

$$\begin{aligned}
\mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^T\right) &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left(\mathbf{X}_i\mathbf{X}_i^T\right) && (121)\\
&= \Sigma_{\mathbf{X}}. && (122)
\end{aligned}$$

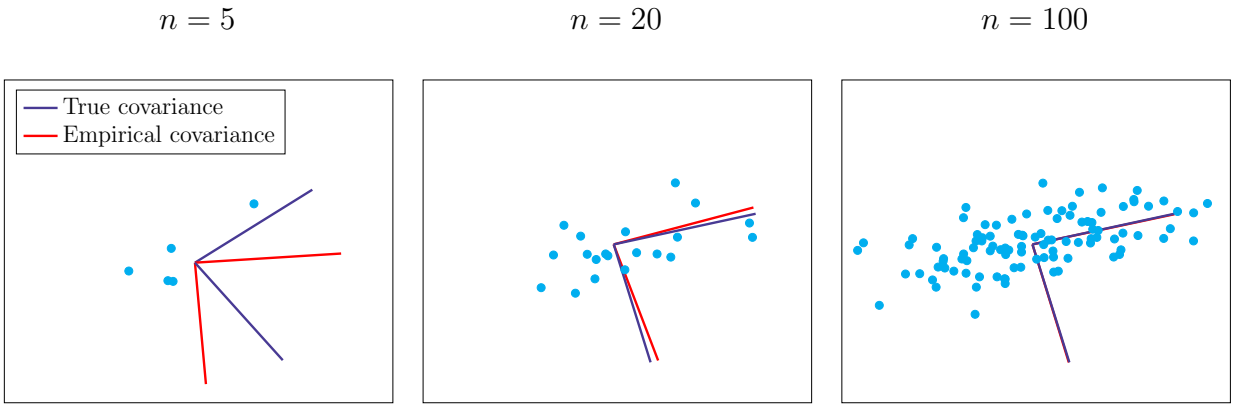

Figure 4: Principal components of n data vectors samples from a 2D Gaussian distribution. The eigenvectors of the covariance matrix of the distribution are also shown.

If the higher moments of the data $\mathrm{E}\left(X_i^2 X_j^2\right)$ and $\mathrm{E}\left(X_i^4\right)$ are finite, by Chebyshev's inequality the entries of the empirical covariance matrix converge to the entries of the true covariance matrix. This means that in the limit

$$\begin{aligned} \mathrm{Var}\left(\mathbf{X}^T u\right) &= u^T \Sigma_{\mathbf{X}} u && (123) \\ &\approx \frac{1}{n} u^T X X^T u && (124) \\ &= \frac{1}{n} \left|\left|X^T u\right|\right|_2^2 && (125) \end{aligned}$$

for any unit-norm vector u. In the limit the principal components correspond to the directions of maximum variance of the underlying random vector. These directions also correspond to the eigenvectors of the true covariance matrix by Theorem 6.2. Figure 4 illustrates how the principal components converge to the eigenvectors of Σ.

A Proofs

A.1 Proof of Theorem 6.2

The eigenvectors are an orthonormal basis (they are mutually orthogonal and we assume that they have been normalized), so we can represent any unit-norm vector h_k that is orthogonal to $u_1, \ldots, u_{k-1}$ as

$$h_k = \sum_{i=k}^{m} \alpha_i u_i \tag{126}$$

where

$$||h_k||_2^2 = \sum_{i=k}^{m} \alpha_i^2 = 1, \tag{127}$$

by Lemma 1.5. Note that h_1 is just an arbitrary unit-norm vector.

Now we will show that the value of the quadratic form when the normalized input is restricted to be orthogonal to $u_1, \ldots, u_{k-1}$ cannot be larger than λ_k,

$$h_k^T A h_k = \sum_{i=1}^{n} \lambda_i \left(\sum_{j=k}^{m} \alpha_j u_i^T u_j \right)^2 \quad \text{by (92) and (126)} \tag{128}$$

$$= \sum_{i=1}^{n} \lambda_i \alpha_i^2 \quad \text{because } u_1, \ldots, u_m \text{ is an orthonormal basis} \tag{129}$$

$$\leq \lambda_k \sum_{i=k}^{m} \alpha_i^2 \quad \text{because } \lambda_k \geq \lambda_{k+1} \geq \ldots \geq \lambda_m \tag{130}$$

$$= \lambda_k, \quad \text{by (127)}. \tag{131}$$

This establishes (93) and (95). To prove (108) and (110) we just need to show that u_k achieves the maximum

$$u_k^T A u_k = \sum_{i=1}^{n} \lambda_i \left(u_i^T u_k \right)^2 \tag{132}$$

$$= \lambda_k. \tag{133}$$

A.2 Proof of Theorem 7.2

It is sufficient to prove

$$\dim \left(\text{row} \left(A \right) \right) \leq \dim \left(\text{row} \left(A \right) \right) \tag{134}$$

for an arbitrary matrix A. We can apply the result to A^T to establish $\dim(\text{row}(A)) \geq \dim(\text{row}(A))$, since $\text{row}(A) = \text{row}(A)^T$ and $\text{row}(A) = \text{row}(A)^T$.

To prove (134) let $r := \dim(\text{row}(A))$ and let $x_1, \ldots, x_r \in R^n$ be a basis for $\text{row}(A)$. Consider the vectors $Ax_1, \ldots, Ax_r \in R^n$. They belong to $\text{row}(A)$ by (43), so if they are linearly independent then $\dim(\text{row}(A))$ must be at least r. We will prove that this is the case by contradiction.

Assume that $Ax_1, \ldots, Ax_r$ are linearly dependent. Then there exist coefficients $\alpha_1, \ldots, \alpha_r \in \mathbb{R}$ such that

$$0 = \sum_{i=1}^{r} \alpha_i A x_i = A\left(\sum_{i=1}^{r} \alpha_i x_i\right) \quad \text{(by linearity of the matrix product)}, \tag{135}$$

This implies that $\sum_{i=1}^{r} \alpha_i x_i$ is orthogonal to every row of A and hence to every vector in $\text{row}(A)$. However it is in the span of a basis of $\text{row}(A)$ by construction! This is only possible if $\sum_{i=1}^{r} \alpha_i x_i = 0$, which is a contradiction because $x_1, \ldots, x_r$ are assumed to be linearly independent.

A.3 Proof of Theorem 8.3

We prove the result by induction. The *base case*, $k = 1$, follows immediately from (108). To complete the proof we need to show that if the results is true for $k - 1 \geq 1$ (this is the *induction hypothesis*) then it also holds for k.

Let $\mathcal{S}$ be an arbitrary subspace of dimension k. We choose an orthonormal basis for the subspace $b_1, b_2, \ldots, b_k$ such that b_k is orthogonal to $u_1, u_2, \ldots, u_{k_1}$. We can do this by using any vector that is linearly independent of $u_1, u_2, \ldots, u_{k_1}$ and subtracting its projection onto the span of $u_1, u_2, \ldots, u_{k_1}$ (if the result is always zero then $\mathcal{S}$ is in the span and consequently cannot have dimension k).

By the induction hypothesis,

$$\sum_{i=1}^{n} \left|\left| \mathcal{P}_{\text{span}\left(u_1, u_2, \ldots, u_{k_1}\right)} x_i \right|\right|_2^2 = \sum_{i=1}^{k-1} \left|\left| X^T u_i \right|\right|_2^2 \quad \text{by (6)} \tag{136}$$

$$\leq \sum_{i=1}^{n} \left|\left| \mathcal{P}_{\text{span}\left(b_1, b_2, \ldots, b_{k_1}\right)} x_i \right|\right|_2^2 \tag{137}$$

$$= \sum_{i=1}^{k-1} \left|\left| X^T b_i \right|\right|_2^2 \quad \text{by (6)}. \tag{138}$$

By (110)

$$\sum_{i=1}^{n} \left|\left|\mathcal{P}_{\text{span}(u_k)}\, x_i\right|\right|_2^2 = \left|\left|X^T u_k\right|\right|_2^2 \tag{139}$$

$$\leq \sum_{i=1}^{n} \left|\left|\mathcal{P}_{\text{span}(b_k)}\, x_i\right|\right|_2^2 \tag{140}$$

$$= \left|\left|X^T b_k\right|\right|_2^2 . \tag{141}$$

Combining (138) and (138) we conclude

$$\sum_{i=1}^{n} \left|\left|\mathcal{P}_{\text{span}(u_1,u_2,\ldots,u_k)}\, x_i\right|\right|_2^2 = \sum_{i=1}^{k} \left|\left|X^T u_i\right|\right|_2^2 \tag{142}$$

$$\leq \sum_{i=1}^{k} \left|\left|X^T b_i\right|\right|_2^2 \tag{143}$$

$$\leq \sum_{i=1}^{n} \left|\left|\mathcal{P}_{\mathcal{S}}\, x_i\right|\right|_2^2 . \tag{144}$$