Matrix methods

Definition (Orthogonal set of vectors)

The set of vectors $\{_1, \ldots, _k\} \in^n$ is an **orthogonal set** if

$$\forall i, j = 1, \ldots, k, \quad i \neq j \implies {}_i \bullet {}_j = 0$$

Theorem

$\{_1, \ldots, _k\} \in^n$ with $\forall i$, $_i \neq \underline{0}$, orthogonal set $\implies$ $\{_1, \ldots, _k\} \in^n$ linearly independent

Definition (Orthogonal basis)

Let S be a basis of the subspace $W \subset^n$ composed of an orthogonal set of vectors. We say S is an **orthogonal basis** of W

Proof of Theorem ??

Assume $\{_1, \ldots, _k\}$ orthogonal set with $_i \neq \underline{0}$ for all $i = 1, \ldots, k$. Recall $\{_1, \ldots, _k\}$ is LI if

$$c_{11} + \cdots + c_{kk} = \underline{0} \iff c_1 = \cdots = c_k = 0$$

So assume $c_1, \ldots, c_k \in$ are s.t. $c_{11} + \cdots + c_{kk} = \underline{0}$. Recall that $\forall \in^k$, $\underline{0}_k \bullet = 0$. So for some $_i \in \{_1, \ldots, _k\}$

$$\begin{aligned} 0 &= \underline{0} \bullet_i \\ &= (c_{11} + \cdots + c_{kk}) \bullet_i \\ &= c_{11} \bullet_i + \cdots + c_{kk} \bullet_i \end{aligned} \tag{1}$$

As $\{_1, \ldots, _k\}$ orthogonal, $_j \bullet_i = 0$ when $i \neq j$, (**??**) reduces to

$$c_{i\,i} \bullet_i = 0 \iff c_i \|_i\|^2 = 0$$

As $_i \neq 0$ for all i, $\|_i\| \neq 0$ and so $c_i = 0$. This is true for all i, hence the result

Example – Vectors of the standard basis of ^{3}

For ^{3}, we denote

$$= \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \text{ and } = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

(^{k} for $k > 3$, we denote them $_i$)

Clearly, $\{,\}$, $\{,\}$, $\{,\}$ and $\{,,\}$ orthogonal sets. The standard basis vectors are also $\neq \underline{0}$, so the sets are LI. And

$$\{,,\}$$

is an orthogonal basis of ^{3} since it spans ^{3} and is LI

$$c_1 + c_2 + c_3 = c_1 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + c_2 \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + c_3 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix}$$

Orthonormal version of things

Definition (Orthonormal set)

The set of vectors $\{_1, \ldots, _k\} \in ^n$ is an **orthonormal set** if it is an orthogonal set and furthermore

$$\forall i = 1, \ldots, k, \quad \|_i\| = 1$$

Definition (Orthonormal basis)

A basis of the subspace $W \subset ^n$ is an **orthonormal basis** if the vectors composing it are an orthonormal set

$\{_1, \ldots, _k\} \in ^n$ is orthonormal if

$$_i \bullet _j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

Projections

Definition (Orthogonal projection onto a subspace)

$W \subset^n$ a subspace and $\{_1, \ldots, _k\}$ an orthogonal basis of W. $\forall \in^n$, the **orthogonal projection** of **onto** W is

$$\text{proj}_W() = \frac{1 \bullet}{\|_1\|^2} {}_1 + \cdots + \frac{k \bullet}{\|_k\|^2} {}_k$$

Definition (Component orthogonal to a subspace)

$W \subset^n$ a subspace and $\{_1, \ldots, _k\}$ an orthogonal basis of W. $\forall \in^n$, the **component** of **orthogonal to** W is

$$\text{perp}_W() = -\text{proj}_W()$$

What this aims to do is to construct an orthogonal basis for a subspace $W \subset^n$

To do this, we use the *Gram-Schmidt orthogonalisation process*, which turn s a basis of W into an orthogonal basis of W

Gram-Schmidt process

Theorem

$W \subset \mathbb{R}^n$ a subset and $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ a basis of W. Let

$$\begin{aligned}
\mathbf{w}_1 &= \mathbf{v}_1 \\
\mathbf{w}_2 &= \mathbf{v}_2 - \frac{\mathbf{v}_1 \bullet \mathbf{v}_2}{\|\mathbf{v}_1\|^2}\mathbf{v}_1 \\
\mathbf{w}_3 &= \mathbf{v}_3 - \frac{\mathbf{v}_1 \bullet \mathbf{v}_3}{\|\mathbf{v}_1\|^2}\mathbf{v}_1 - \frac{\mathbf{v}_2 \bullet \mathbf{v}_3}{\|\mathbf{v}_2\|^2}\mathbf{v}_2 \\
&\vdots \\
\mathbf{w}_k &= \mathbf{v}_k - \frac{\mathbf{v}_1 \bullet \mathbf{v}_k}{\|\mathbf{v}_1\|^2}\mathbf{v}_1 - \cdots - \frac{\mathbf{v}_{k-1} \bullet \mathbf{v}_k}{\|\mathbf{v}_{k-1}\|^2}\mathbf{v}_{k-1}
\end{aligned}$$

and

$$W_1 = \text{span}(\mathbf{v}_1), W_2 = \text{span}(\mathbf{v}_1, \mathbf{v}_2), \ldots, W_k = \text{span}(\mathbf{v}_1, \ldots, \mathbf{v}_k)$$

Then $\forall i = 1, \ldots, k$, $\{\mathbf{v}_1, \ldots, \mathbf{v}_i\}$ orthogonal basis for W_i

The least squares problem (simplest version)

Definition

Given a collection of points $(x_1, y_1), \ldots, (x_n, y_n)$, find the coefficients a, b of the line $y = a + bx$ such that

$$\|\mathbf{e}\| = \sqrt{\varepsilon_1^2 + \cdots + \varepsilon_n^2} = \sqrt{(y_1 - \tilde{y}_1)^2 + \cdots + (y_n - \tilde{y}_n)^2}$$

is minimal, where $\tilde{y}_i = a + bx_i$ for $i = 1, \ldots, n$

We just saw how to solve this by brute force using a genetic algorith to minimise $\|e\|$, let us now see how to solve this problem "properly"

For a data point $i = 1, \ldots, n$

$$\varepsilon_i = y_i - \tilde{y}_i = y_i - (a + bx_i)$$

So if we write this for all data points,

$$\begin{aligned} \varepsilon_1 &= y_1 - (a + bx_1) \\ &\vdots \\ \varepsilon_n &= y_n - (a + bx_n) \end{aligned}$$

In matrix form

$$= -A$$

with

$$= \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}, A = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, = \begin{pmatrix} a \\ b \end{pmatrix} \text{ and } = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

The least squares problem (reformulated)

Definition (Least squares solutions)

Consider a collection of points $(x_1, y_1), \ldots, (x_n, y_n)$, a matrix $A \in_{mn}$, $\in^m$. A **least squares solution** of $A =$ is a vector$\tilde{} \in^n$ s.t.

$$\forall \in^n, \quad \| - A\| \leq \| - A\|$$

Needed to solve the problem

Definition (Best approximation)

Let V be a vector space, $W \subset V$ and $\mathbf{v} \in V$. The **best approximation** to $\mathbf{v}$ in W is $\tilde{\mathbf{v}} \in W$ s.t.

$$\forall \mathbf{w} \in W, \mathbf{w} \neq \tilde{\mathbf{v}}, \quad \|\mathbf{v} - \tilde{\mathbf{v}}\| < \|\mathbf{v} - \mathbf{w}\|$$

Theorem (Best approximation theorem)

Let V be a vector space with an inner product, $W \subset V$ and $\mathbf{v} \in V$. Then $\text{proj}_W(\mathbf{v})$ is the best approximation to $\mathbf{v}$ in W

Let us find the least squares solution

$\forall^n$, A is a vector in the **column space** of A (the space spanned by the vectors making up the columns of A)

Since $\in^n$, $A \in \text{col}(A)$

$\implies$ least squares solution of $A =$ is a vector$\tilde{} \in \text{col}(A)$ s.t.

$$\forall \in \text{col}(A), \quad \| - \tilde{} \| \leq \| - \|$$

This looks very much like Best approximation and Best approximation theorem

Putting things together

We just stated: The least squares solution of $A =$ is a vector $\tilde{} \in \text{col}(A)$ s.t.

$$\forall \in \text{col}(A), \quad \| - \tilde{}\| \leq \| - \|$$

We know (reformulating a tad):

Theorem (Best approximation theorem)

Let V be a vector space with an inner product, $W \subset V$ and $\mathbf{v} \in V$. Then $\text{proj}_W(\mathbf{v}) \in W$ is the best approximation to $\mathbf{v}$ in W, i.e.,

$$\forall \mathbf{w} \in W, \mathbf{w} \neq \text{proj}_W(\mathbf{v}), \quad \|\mathbf{v} - \text{proj}_W(\mathbf{v})\| < \|\mathbf{v} - \mathbf{w}\|$$

$$\implies W = \text{col}(A), = \text{and } \tilde{} = \text{proj}_{\text{col}(A)}(\mathbf{b})$$

So if $\tilde{}$ is a least squares solution of $A =$, then

$$\tilde{} = A\tilde{} = \text{proj}_{\text{col}(A)}(\mathbf{b})$$

We have

$$-A\tilde{} = -\text{proj}_{\text{col}(A)}(\mathbf{b}) = \text{perp}_{\text{col}(A)}(\mathbf{b})$$

and it is easy to show that

$$\text{perp}_{\text{col}(A)}(\mathbf{b}) \perp \text{col}(A)$$

So for all columns $_i$ of A

$$_i \cdot (-A\tilde{}) = 0$$

which we can also write as $_i^T(-A\tilde{}) = 0$

For all columns $_i$ of A,

$$^T_i(-A\tilde{}) = 0$$

This is equivalent to saying that

$$A^T(-A\tilde{}) = \underline{0}$$

We have

$$\begin{aligned} A^T(-A\tilde{}) = \underline{0} &\iff A^T - A^T A\tilde{} = \underline{0} \\ &\iff A^T = A^T A\tilde{} \\ &\iff A^T A\tilde{} = A^T \end{aligned}$$

The latter system constitutes the **normal equations** for ˜

Least squares theorem

[Least squares theorem] $A \in_{mn}$, $\in^{m}$. Then

1. $A =$ always has at least one least squares solution $\tilde{}$
2. $\tilde{}$least squares solution to $A = \iff \tilde{}$is a solution to the normal equations $A^T A\tilde{} = A^T$
3. A has linearly independent columns $\iff A^T A$ invertible. In this case, the least squares solution is unique and

$$\tilde{} = \left(A^T A\right)^{-1} A^T$$

We have seen 1 and 2, we will not show 3 (it is not hard)

Suppose we want to fit something a bit more complicated..

For instance, instead of the affine function

$$y = a + bx$$

suppose we want to do the quadratic

$$y = a_0 + a_1 x + a_2 x^2$$

or even

$$y = k_0 e^{k_1 x}$$

How do we proceed?

Fitting the quadratic

We have the data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ and want to fit

$$y = a_0 + a_1 x + a_2 x^2$$

At (x_1, y_1),

$$\tilde{y}_1 = a_0 + a_1 x_1 + a_2 x_1^2$$

$\vdots$

At (x_n, y_n),

$$\tilde{y}_n = a_0 + a_1 x_n + a_2 x_n^2$$

In terms of the error

$$\varepsilon_1 = y_1 - \tilde{y}_1 = y_1 - (a_0 + a_1 x_1 + a_2 x_1^2)$$

$$\vdots$$

$$\varepsilon_n = y_n - \tilde{y}_n = y_n - (a_0 + a_1 x_n + a_2 x_n^2)$$

i.e.,

$$= -A$$

where

$$= \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}, A = \begin{pmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}, = \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} \text{ and } = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

Theorem **??** applies, with here $A \in_{n3}$ and $\in^n$

Fitting the exponential

Things are a bit more complicated here

If we proceed as before, we get the system

$$y_1 = k_0 e^{k_1 x_1}$$
$$\vdots$$
$$y_n = k_0 e^{k_1 x_n}$$

$e^{k_1 x_i}$ is a nonlinear term, it cannot be put in a matrix

However: take the ln of both sides of the equation

$$\ln(y_i) = \ln(k_0 e^{k_1 x_i}) = \ln(k_0) + \ln(e^{k_1 x_i}) = \ln(k_0) + k_1 x_i$$

If $y_i, k_0 > 0$, then their ln are defined and we're in business..

$$\ln(y_i) = \ln(k_0) + k_1 x_i$$

So the system is

$$= A+$$

with

$$A = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, = (k_1), = (\ln(k_0)) \text{ and } = \begin{pmatrix} \ln(y_1) \\ \vdots \\ \ln(y_n) \end{pmatrix}$$

Theorem

Let $Q \in_{mn}$. The columns of Q form an orthonormal set if and only if

$$Q^T Q =_n$$

Definition (Orthogonal matrix)

$Q \in_n$ is an **orthogonal matrix** if its columns form an orthonormal set

So $Q \in_n$ orthogonal if $Q^T Q =$, i.e., $Q^T = Q^{-1}$

Theorem (NSC for orthogonality)

$Q \in_n$ *orthogonal* $\iff$ $Q^{-1} = Q^T$

Theorem (Orthogonal matrices "encode" isometries)

Let $Q \in_n$. TFAE

1. *Q orthogonal*
2. $\forall \in^n$, $\|Q\| = \|\|$
3. $\forall, \in^n$, $Q \bullet Q = \bullet$

Theorem

Let $Q \in_n$ be orthogonal. Then

1. *The rows of Q form an orthonormal set*
2. *Q^{-1} orthogonal*
3. $\det Q = \pm 1$
4. $\forall \lambda \in \sigma(Q)$, $|\lambda| = 1$
5. *If $Q_2 \in_n$ also orthogonal, then QQ_2 orthogonal*

Proof of 4 in Theorem **??**

All statements in Theorem **??** are easy, but let's focus on 4

Let λ be an eigenvalue of $Q \in_n$ orthogonal, i.e., $\exists^n \ni \neq \underline{0}$ s.t.

$$Q = \lambda$$

Take the norm on both sides

$$\|Q\| = \|\lambda\|$$

From 2 in Theorem **??**, $\|Q\| = \|\|$ and from the properties of norms, $\|\lambda\| = |\lambda| \, \|\|$, so we have

$$\|Q\| = \|\lambda\| \iff \|\| = |\lambda| \, \|\| \iff 1 = |\lambda|$$

(we can divide by $\|\|$ since $\neq \underline{0}$ as an eigenvector) $\square$

Matrix factorisations

Matrix factorisations are popular because they allow to perform some computations more easily

There are several different types of factorisations. Here, we study just the QR factorisation, which is useful for many least squares problems

The QR factorisation

Theorem

Let $A \in_{mn}$ with LI columns. Then A can be factored as

$$A = QR$$

where $Q \in_{mn}$ has orthonormal columns and $R \in_n$ is nonsingular upper triangular

Back to least squares

So what was the point of all that..?

Theorem (Least squares with QR factorisation)

$A \in_{mn}$ with LI columns, $\in^m$. If $A = QR$ is a QR factorisation of A, then the unique least squares solution˜ of $A =$ is

$$\tilde{} = R^{-1}Q^T$$

Proof of Theorem ??

A has LI columns so

- least squares $A =$ has unique solution $\tilde{} = (A^T A)^{-1} A^T$
- by Theorem **??**, A can be written as $A = QR$ with $Q \in_{mn}$ with orthonormal columns and $R \in_n$ nonsingular and upper triangular

So

$$
\begin{aligned}
A^T A\tilde{} = A^T &\implies (QR)^T QR\tilde{} = (QR)^T \\
&\implies R^T Q^T QR\tilde{} = R^T Q^T \\
&\implies R_n^T R\tilde{} = R^T Q^T \\
&\implies R^T R\tilde{} = R^T Q^T \\
&\implies (R^T)^{-1} R\tilde{} = (R^T)^{-1} R^T Q^T \\
&\implies R\tilde{} = Q^T \\
&\implies \tilde{} = R^{-1} Q^T \quad \square
\end{aligned}
$$

Matrix factorisations (continued)

The singular value decomposition (known mostly by its acronym, SVD) is yet another type of factorisation/decomposition..

Singular values

Definition (Singular value)

Let $A \in_{mn} ()$. The **singular values** of A are the real numbers

$$\sigma_1 \geq \sigma_2 \geq \cdots \sigma_n \geq 0$$

that are the square roots of the eigenvalues of $A^T A$

Singular values are real and nonnegative?

Recall that $\forall A \in_{mn}$, $A^T A$ is symmetric

Claim 1. Real symmetric matrices have real eigenvalues

Proof. $A \in_n ()$ symmetric and $(\lambda,)$ eigenpair of A, i.e, $A = \lambda$.
Taking the complex conjugate, $\overline{A} = \overline{\lambda}$

Since $A \in_n ()$, $\overline{A} = A \qquad (z = \bar{z} \iff z \in)$

So

$$\bar{A}\bar{} = \overline{\bar{A}} = \overline{A} = \overline{\lambda} = \overline{\lambda}\bar{}$$

i.e., if $(\lambda,)$ eigenpair, $(\bar{\lambda},\bar{})$ also eigenpair

Still assuming $A \in_n ()$ symmetric and $(\lambda,)$ eigenpair of A and using what we just proved (that $(\bar{\lambda},\bar{})$ also eigenpair), take transposes

$$\begin{aligned} A\bar{} = \bar{\lambda}\bar{} &\iff (A\bar{})^T = (\bar{\lambda}\bar{})^T \\ &\iff \bar{}^T A^T = \bar{\lambda}\bar{}^T \\ &\iff \bar{}^T A = \bar{\lambda}\bar{}^T \qquad [A \text{ symmetric}] \end{aligned}$$

Let us now compute $\lambda(\bar{\bullet})$. We have

$$\begin{aligned} \lambda(\bar{\bullet}) = \lambda\bar{}^T &= \bar{}^T(\lambda) \\ &= \bar{}^T(A) = (\bar{}^T A) \\ &= (\bar{\lambda}\bar{}^T) = \bar{\lambda}(\bar{\bullet}) \\ &\iff (\lambda - \bar{\lambda})(\bar{\bullet}) = 0 \end{aligned}$$

We have shown

$$(\lambda - \bar{\lambda})(\bar{\bullet}) = 0$$

Let

$$= \begin{pmatrix} a_1 + ib_1 \\ \vdots \\ a_n + ib_n \end{pmatrix}$$

Then

$$\bar{} = \begin{pmatrix} a_1 - ib_1 \\ \vdots \\ a_n - ib_n \end{pmatrix}$$

So

$$\bar{\bullet} = (a_1^2 + b_1^2) + \cdots + (a_n^2 + b_n^2)$$

But eigenvector is $\neq \underline{0}$, so $\bar{\bullet} \neq 0$, so

$$(\lambda - \bar{\lambda})(\bar{\bullet}) = 0 \iff \lambda - \bar{\lambda} = 0 \iff \lambda = \bar{\lambda} \iff \lambda \in \quad \square$$

Claim 2. For $A \in_{mn}$ (), the eigenvalues of $A^T A$ are real and nonnegative

Proof. We know that for $A \in_{mn}$, $A^T A$ symmetric and from previous claim, if $A \in_{mn}$ (), then $A^T A$ is symmetric and real and with real eigenvalues

Let $(\lambda,)$ be an eigenpair of $A^T A$, with chosen so that $\| \| = 1$

Norms are functions $V \to_+$, so $\|A\|$ and $\|A\|^2$ are ≥ 0 and thus

$$\begin{aligned} 0 \leq \|A\|^2 &= (A) \bullet (A) = (A)^T (A) \\ &=^T A^T A =^T (A^T A) =^T (\lambda) \\ &= \lambda(^T) = \lambda(\bullet) = \lambda \| \|^2 \\ &= \lambda \quad \square \end{aligned}$$

Claim 3. For $A \in_{mn} ()$, the nonzero eigenvalues of $A^T A$ and AA^T are the same

Proof. Let $(\lambda,)$ be an eigenpair of $A^T A$ with $\lambda \neq 0$. Then $\neq \underline{0}$ and

$$A^T A = \lambda$$

The singular value decomposition (SVD)

[SVD] $A \in_{mn}$ with singular values $\sigma_1 \geq \cdots \geq \sigma_r > 0$ and $\sigma_{r+1} = \cdots = \sigma_n = 0$

Then there exists $U \in_m$ orthogonal, $V \in_n$ orthogonal and a block matrix $\Sigma \in_{mn}$ taking the form

$$\Sigma = \begin{pmatrix} D & 0_{r,n-r} \\ 0_{m-r,r} & 0_{m-r,n-r} \end{pmatrix}$$

where

$$D = \text{diag}(\sigma_1, \ldots, \sigma_r) \in_r$$

such that

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

Definition

We call a factorisation as in Theorem **??** the **singular value decomposition** of A. The columns of U and V are, respectively, the **left** and **right singular vectors** of A

U and V^T are *rotation* or *reflection* matrices, Σ is a *scaling* matrix

Outer product form of the SVD

Theorem (Outer product form of the SVD)

$A \in_{mn}$ with singular values $\sigma_1 \geq \cdots \geq \sigma_r > 0$ and $\sigma_{r+1} = \cdots = \sigma_n = 0$, $_1, \ldots, _r$ and $_1, \ldots, _r$, respectively, left and right singular vectors of A corresponding to these singular values

Then

$$A = \sigma_1 {}_1 {}_1^T + \cdots + \sigma_r {}_r {}_r^T$$

Computing the SVD (case of $\neq$ eigenvalues)

To compute the SVD, we use the following result

Theorem

Let $A \in_n$ symmetric, $(\lambda_1, _1)$ and $(\lambda_2, _2)$ be eigenpairs, $\lambda_1 \neq \lambda_2$. Then $_1 \bullet _2 = 0$

Proof of Theorem ??

$A \in_n$ symmetric, $(\lambda_1, {}_1)$ and $(\lambda_2, {}_2)$ eigenpairs with $\lambda_1 \neq \lambda_2$

$$\begin{aligned}
\lambda_1({}_1 \bullet {}_2) &= (\lambda_1 {}_1) \bullet {}_2 \\
&= A {}_1 \bullet {}_2 \\
&= (A {}_1)^T {}_2 \\
&=^T_1 A^T_2 \\
&=^T_1 (A {}_2) \qquad [A \text{ symmetric so } A^T = A] \\
&=^T_1 (\lambda_2 {}_2) \\
&= \lambda_2 ({}^T_1 {}_2) \\
&= \lambda_2({}_1 \bullet {}_2)
\end{aligned}$$

So $(\lambda_1 - \lambda_2)({}_1 \bullet {}_2) = 0$. But $\lambda_1 \neq \lambda_2$, so ${}_1 \bullet {}_2 = 0$ $\square$

Computing the SVD (case of $\neq$ eigenvalues)

If all eigenvalues of $A^T A$ are distinct, we can use Theorem **??**

1. Compute $A^T A \in_n$
2. Compute eigenvalues $\lambda_1, \ldots, \lambda_n$ of $A^T A$; order them as $\lambda_1 > \cdots > \lambda_n \geq 0$ ($>$ not $\geq$ since $\neq$)
3. Compute singular values $\sigma_1 = \sqrt{\lambda_1}, \ldots, \sigma_n = \sqrt{\lambda_n}$
4. Diagonal matrix D in Σ is either in $_n$ (if $\sigma_n > 0$) or in $_{n-1}$ (if $\sigma_n = 0$)

5. Since eigenvalues are distinct, Theorem **??** $\implies$ eigenvectors are orthogonal set. Compute these eigenvectors in the same order as the eigenvalues
6. Normalise them and use them to make the matrix V, i.e., $V = [_1 \cdots _n]$
7. To find the $_i$, compute, for $i = 1, \ldots, r$,

$$_i = \frac{1}{\sigma_i} A_i$$

and ensure that $\|_i\| = 1$

Computing the SVD (case where some eigenvalues are $=$)

1. Compute $A^T A \in_n$
2. Compute eigenvalues $\lambda_1, \ldots, \lambda_n$ of $A^T A$; order them as $\lambda_1 \geq \cdots \geq \lambda_n \geq 0$
3. Compute singular values $\sigma_1 = \sqrt{\lambda_1}, \ldots, \sigma_n = \sqrt{\lambda_n}$, with $r \leq n$ the index of the last positive singular value
4. For eigenvalues that are distinct, proceed as before
5. For eigenvalues with multiplicity > 1, we need to ensure that the resulting eigenvectors are LI *and* orthogonal

Dealing with eigenvalues with multiplicity > 1

When an eigenvalue has (algebraic) multiplicity > 1, e.g., characteristic polynomial contains a factor like $(\lambda - 2)^2$, things can become a little bit more complicated

The proper way to deal with this involves the so-called Jordan Normal Form (another matrix decomposition)

In short: not all square matrices are diagonalisable, but all square matrices admit a JNF

Sometimes, we can find several LI eigenvectors associated to the same eigenvalue. Check this. If not, need to use the following

Definition (Generalised eigenvectors)

$\neq \underline{0}$ **generalized eigenvector** of rank m of $A \in_n$ corresponding to eigenvalue λ if

$$(A - \lambda)^m = \underline{0}$$

but

$$(A - \lambda)^{m-1} \neq \underline{0}$$

Procedure for generalised eigenvectors

$A \in_n$ and assume λ eigenvalue with algebraic multiplicity k

Find $_1$, "classic" eigenvector, i.e., $_1 \neq \underline{0}$ s.t. $(A - \lambda)_1 = \underline{0}$

Find generalised eigenvector $_2$ of rank 2 by solving for $_2 \neq \underline{0}$,

$$(A - \lambda)_2 =_1$$

...

Find generalised eigenvector $_k$ of rank k by solving for $_k \neq \underline{0}$,

$$(A - \lambda)_k =_{k-1}$$

Then $\{_1, \ldots, _k\}$ LI

Back to the normal procedure

With the LI eigenvectors $\{_1, \ldots, _k\}$ corresponding to λ

Apply Gram-Schmidt to get orthogonal set

For all eigenvalues with multiplicity > 1, check that you either have LI eigenvectors or do what we just did

When you are done, be back on your merry way to step 6 in the case where eigenvalues are all $\neq$

I am caricaturing a little here: there can be cases that do not work exactly like this, but this is general enough..

Applications of the SVD

Many applications of the SVD, both theoretical and practical..

1. Obtaining a unique solutions to least squares when $A^T A$ singular
2. Image compression

Least squares revisited

Theorem

Let $A \in_{mn}$, $\in^n$ and $\in^m$. The least squares problem $A =$ has a unique least squares solution˜ of minimal length (closest to the origin) given by

$$\tilde{} = A^+$$

where A^+ is the pseudoinverse of A

Definition (Pseudoinverse)

$A = U\Sigma V^T$ an SVD for $A \in_{mn}$, where

$$\Sigma = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix}, \text{ with } D = \text{diag}(\sigma_1, \ldots, \sigma_r)$$

(D contains the nonzero singular values of A ordered as usual)

The **pseudoinverse** (or **Moore-Penrose inverse**) of A is $A^+ \in_{nm}$ given by

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

with

$$\Sigma^+ = \begin{pmatrix} D^{-1} & 0 \\ 0 & 0 \end{pmatrix} \in_{nm}$$

Compressing images

Consider an image (for simplicity, assume in shades of grey). This can be stored in a matrix $A \in_{mn}$

Take the SVD of A. Then the small singular values carry information about the regions with little variation and can perhaps be omitted, whereas the large singular values carry information about more "dynamic" regions of the image

Suppose A has r nonzero singular values. For $k \leq r$, let

$$A_k = \sigma_{11}{}_1^T + \cdots + \sigma_{kk}{}_k^T$$

(so for $k = r$ we get the usual outer product form)

Dimensionality reduction

One of the reasons the SVD is used is for dimensionality reduction. However, SVD has many many other uses

Now we look at another dimensionality reduction technique, PCA

PCA is often used as a blackbox technique, here we take a look at the math behind it

What is PCA?

Linear algebraic technique

Helps reduce a complex dataset to a lower dimensional one

Non-parametric method: does not assume anything about data distribution (distribution from the statistical point of view)

Brief "review" of some probability concepts

Proper definition of *probability* requires to use *measure theory*.. will not get into details here

A **random variable** X is a *measurable* function $X : \Omega \to E$, where Ω is a set of outcomes (*sample space*) and E is a measurable space

$$(X \in S \subseteq E) = (\omega \in \Omega | X(\omega) \in S)$$

Distribution function of a r.v., $F(x) = (X \leq x)$, describes the distribution of a r.v.

R.v. can be discrete or continuous or .. other things.

Definition (Variance)

Let X be a random variable. The **variance** of X is given by

$$X = E\left[(X - E(X))^2\right]$$

where E is the expected value

Definition (Covariance)

Let X, Y be jointly distributed random variables. The **covariance** of X and Y is given by

$$(X, Y) = E\left[(X - E(X))\,(Y - E(Y))\right]$$

Note that $(X, X) = E\left[(X - E(X))^2\right] = X$

In practice: "true law" versus "observation"

In statistics: we reason on the *true law* of distributions, but we usually have only access to a sample

We then use **estimators** to .. estimate the value of a parameter, e.g., the mean, variance and covariance

Definition (Unbiased estimators of the mean and variance)

Let $x_1, \ldots, x_n$ be data points (the *sample*) and

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

be the **mean** of the data. An unbiased estimator of the variance of the sample is

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Definition (Unbiased estimator of the covariance)

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be data points,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ and } \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

be the means of the data. An estimator of the covariance of the sample is

$$(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

What does covariance do?

Variance explains how data disperses around the mean, in a 1-D case

Covariance measures the relationship between two dimensions. E.g., height and weight

More than the exact value, the sign is important:

- $(X, Y) > 0$: both dimensions change in the same "direction"; e.g., larger height usually means higher weight
- $(X, Y) < 0$: both dimensions change in reverse directions; e.g., time spent on social media and performance in this class
- $(X, Y) = 0$: the dimensions are independent from one another; e.g., sex/gender and "intelligence"

The covariance matrix

Typically, we consider more than 2 variables..

Definition

Suppose p random variables $X_1, \ldots, X_p$. Then the covariance matrix is the symmetric matrix

$$\begin{pmatrix} (X_1, X_1) & (X_1, X_2) & \cdots & (X_1, X_p) \\ (X_2, X_1) & (X_2, X_2) & \cdots & (X_2, X_p) \\ \vdots & \vdots & & \vdots \\ (X_p, X_1) & (X_p, X_2) & \cdots & (X_p, X_p) \end{pmatrix}$$

i.e., using the properties of covariance,

$$\begin{pmatrix} X_1 & (X_1, X_2) & \cdots & (X_1, X_p) \\ (X_1, X_2) & X_2 & \cdots & (X_2, X_p) \\ \vdots & \vdots & & \vdots \\ (X_1, X_p) & (X_2, X_p) & \cdots & X_p \end{pmatrix}$$

Example of a PCA problem

We collect a bunch of information about a bunch of people.. for instance this data from Loughborough University

> *This dataset contains the height, weight and 4 fingerprint measurements (length, width, area and circumference), collected from 200 participants.*

What best describes a participant?

The variables

Each participant is associated to 11 variables

- "Participant Number"
- "Gender"
- "Age"
- "Dominant Hand"
- "Height (cm) (average of 3 measurements)"
- "Weight (kg) (average of 3 measurements)"
- "Fingertip Temperature (°C)"
- "Fingerprint Height (mm)"
- "Fingerprint Width (mm)"
- "Fingerprint Area (mm2)"
- "Fingerprint Circumference (mm)"

Nature of variables

Variables have different natures

- "Participant Number": $\in$ (not interesting)
- "Gender": categorical
- "Age": $\in$
- "Dominant Hand": categorical
- "Height (cm) (average of 3 measurements)": $\in$
- "Weight (kg) (average of 3 measurements)": $\in$
- "Fingertip Temperature (°C)": $\in$
- "Fingerprint Height (mm)": $\in$
- "Fingerprint Width (mm)": $\in$
- "Fingerprint Area (mm2)": $\in$
- "Fingerprint Circumference (mm)": $\in$

Setting things up

Each participant is a row in the matrix (an *observation*)

Each variable is a column

So we have an 200×10 matrix (we discard the "Participant number" column)

We want to find what carries the most information

For this, we are going to project the information in a new basis in which the first "dimension" will carry most variance, the second dimension will carry a little less, etc.

In order to do so, we need to learn how to change bases

In the following slide,

$$[]$$

denotes the coordinates of in the basis

The aim of a change of basis is to express vectors in another coordinate system (another basis)

We do so by finding a matrix allowing to move from one basis to another

Change of basis

Definition (Change of basis matrix)

$= \{_1, \ldots, _n\}$ and $= \{_1, \ldots, _n\}$ bases of vector space V

The **change of basis matrix** $P_{\leftarrow} \in_n$,

$$P_{\leftarrow} = \left[[_1]_{\ldots [_n]} \right]$$

has columns the coordinate vectors $[_1], \ldots, [_n]$ of the vectors in with respect to

Theorem

$= \{_1, \ldots, _n\}$ and $= \{_1, \ldots, _n\}$ bases of vector space V and $P_{\leftarrow}$ a change of basis matrix from to

1. $\forall \in V$, $P_{\leftarrow}[]_=[]$
2. $P_{\leftarrow}$ *s.t.* $\forall \in V$, $P_{\leftarrow}[]_=[]$ *is* **unique**
3. $P_{\leftarrow}$ *invertible and* $P_{\leftarrow}^{-1} = P_{\leftarrow}$

Row-reduction method for changing bases

Theorem

$= \{_1, \ldots, _n\}$ and $= \{_1, \ldots, _n\}$ bases of vector space V. Let be any basis for V,

$$B = [[_1], \ldots, [_n]] \ and \ C = [[_1], \ldots, [_n]]$$

and let $[C|B]$ be the augmented matrix constructed using C and B. Then

$$RREF\,([C|B]) = [|P_{\leftarrow}]$$

If working in ^{n}, this is quite useful with the standard basis of ^{n} (it does not matter if $=$)

So the question now becomes

How do we find what new basis to look at our data in?

(Changing the basis does not change the data, just the view you have of it)

(Think of what happens when you do a headstand.. your up becomes down, your right and left switch, but the world does not change, just your view of it)

(Changes of bases are *fundamental* operations in Science)

Setting things up

I will use notation (mostly) as in Joliffe's *Principal Component Analysis* (PDF of older version available for free from UofM Libraries)

$= (x_1, \ldots, x_p)$ vector of p random variables

We seek a linear function α_1^T with maximum variance, where $\alpha_1 = (\alpha_{11}, \ldots, \alpha_{1p})$, i.e.,

$$\alpha_1^T = \sum_{j=1}^{p} \alpha_{1j} x_j$$

Then we seek a linear function α_2^T with maximum variance, uncorrelated to α_1^T

And we continue...

At kth stage, we find a linear function α_k^T with maximum variance, uncorrelated to $\alpha_1^T, \ldots, \alpha_{k-1}^T$

α_i^T is the ith **principal component** (PC)

Case of known covariance matrix

Suppose we know Σ, covariance matrix of (i.e., typically: we know)

Then the kth PC is

$$z_k = \alpha_k^T$$

where α_k is an eigenvector of Σ corresponding to the kth largest eigenvalue λ_k

If, additionally, $\|\alpha_k\| = \alpha_k^T \alpha = 1$, then $\lambda_k = z_k$

Why is that?

Let us start with

$$\alpha_1^T$$

We want maximum variance, where $\alpha_1 = (\alpha_{11}, \ldots, \alpha_{1p})$, i.e.,

$$\alpha_1^T = \sum_{j=1}^{p} \alpha_{1j} x_j$$

with the constraint that $\|\alpha_1\| = 1$

We have

$$\alpha_1^T = \alpha_1^T \Sigma \alpha_1$$

Objective

We want to maximise α_1^T, i.e.,

$$\alpha_1^T \Sigma \alpha_1$$

under the constraint that $\|\alpha_1\| = 1$

$\Longrightarrow$ use **Lagrange multipliers**

Maximisation using Lagrange multipliers

(A.k.a. super-brief intro to multivariable calculus)

We want the max of $f(x_1, \ldots, x_n)$ under the constraint $g(x_1, \ldots, x_n) = k$

1. Solve

$$\nabla f(x_1, \ldots, x_n) = \lambda \nabla g(x_1, \ldots, x_n)$$
$$g(x_1, \ldots, x_n) = k$$

   where $\nabla = (\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_n})$ is the **gradient operator**

2. Plug all solutions into $f(x_1, \ldots, x_n)$ and find maximum values (provided values exist and $\nabla g \neq \underline{0}$ there)

λ is the **Lagrange multiplier**

The gradient

(Continuing our super-brief intro to multivariable calculus)

$f : ^n \to$ function of several variables, $\nabla = \left(\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_n} \right)$ the gradient operator

Then

$$\nabla f = \left(\frac{\partial}{\partial x_1} f, \ldots, \frac{\partial}{\partial x_n} f \right)$$

So ∇f is a *vector-valued* function, $\nabla f : ^n \to ^n$; also written as

$$\nabla f = f_{x_1}(x_1, \ldots, x_n)_1 + \cdots f_{x_n}(x_1, \ldots, x_n)_n$$

where f_{x_i} is the partial derivative of f with respect to x_i and $\{1, \ldots, n\}$ is the standard basis of ^{n}

Bear with me..

(You may experience a brief period of discomfort)

$\alpha_1^T \Sigma \alpha_1$ and $\|\alpha_1\|^2 = \alpha_1^T \alpha_1$ are functions of $\alpha_1 = (\alpha_{11}, \ldots, \alpha_{1p})$

In the notation of the previous slide, we want the max of

$$f(\alpha_{11}, \ldots, \alpha_{1p}) := \alpha_1^T \Sigma \alpha_1$$

under the constraint that

$$g(\alpha_{11}, \ldots, \alpha_{1p}) := \alpha_1^T \alpha_1 = 1$$

and with gradient operator

$$\nabla = \left(\frac{\partial}{\partial \alpha_{11}}, \ldots, \frac{\partial}{\partial \alpha_{1p}} \right)$$

Effect of ∇ on g

g is easiest to see:

$$\begin{aligned}
\nabla g(\alpha_{11}, \ldots, \alpha_{1p}) &= \left(\frac{\partial}{\partial \alpha_{11}}, \ldots, \frac{\partial}{\partial \alpha_{1p}} \right) (\alpha_{11}, \ldots, \alpha_{1p}) \begin{pmatrix} \alpha_{11} \\ \vdots \\ \alpha_{1p} \end{pmatrix} \\
&= \left(\frac{\partial}{\partial \alpha_{11}}, \ldots, \frac{\partial}{\partial \alpha_{1p}} \right) (\alpha_{11}^2 + \cdots + \alpha_{1p}^2) \\
&= (2\alpha_{11}, \ldots, 2\alpha_{1p}) \\
&= 2\alpha_1
\end{aligned}$$

(And that's a general result: $\nabla \|\|_2^2 = 2$ with $\| \cdot \|_2$ the Euclidean norm)

Effect of ∇ on f

Expand (write $\Sigma = [s_{ij}]$ and do not exploit symmetry)

$$\alpha_1^T \Sigma \alpha_1 = (\alpha_{11}, \ldots, \alpha_{1p}) \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & & \vdots \\ s_{p1} & s_{p2} & & s_{pp} \end{pmatrix} \begin{pmatrix} \alpha_{11} \\ \alpha_{12} \\ \vdots \\ \alpha_{1p} \end{pmatrix}$$

$$= (\alpha_{11}, \ldots, \alpha_{1p}) \begin{pmatrix} s_{11}\alpha_{11} + s_{12}\alpha_{12} + \cdots + s_{1p}\alpha_{1p} \\ s_{21}\alpha_{11} + s_{22}\alpha_{12} + \cdots + s_{2p}\alpha_{1p} \\ \vdots \\ s_{p1}\alpha_{11} + s_{p2}\alpha_{12} + \cdots + s_{pp}\alpha_{1p} \end{pmatrix}$$

$$= (s_{11}\alpha_{11} + s_{12}\alpha_{12} + \cdots + s_{1p}\alpha_{1p})\alpha_{11}$$
$$+ (s_{21}\alpha_{11} + s_{22}\alpha_{12} + \cdots + s_{2p}\alpha_{1p})\alpha_{12}$$
$$\vdots$$
$$+ (s_{p1}\alpha_{11} + s_{p2}\alpha_{12} + \cdots + s_{pp}\alpha_{1p})\alpha_{1p}$$

We have

$$\begin{aligned}\alpha_1^T \Sigma \alpha_1 = &(s_{11}\alpha_{11} + s_{12}\alpha_{12} + \cdots + s_{1p}\alpha_{1p})\alpha_{11} \\ &+ (s_{21}\alpha_{11} + s_{22}\alpha_{12} + \cdots + s_{2p}\alpha_{1p})\alpha_{12} \\ &\vdots \\ &+ (s_{p1}\alpha_{11} + s_{p2}\alpha_{12} + \cdots + s_{pp}\alpha_{1p})\alpha_{1p}\end{aligned}$$

So

$$\begin{aligned}\frac{\partial}{\partial \alpha_{11}}\alpha_1^T \Sigma \alpha_1 = &(s_{11}\alpha_{11} + s_{12}\alpha_{12} + \cdots + s_{1p}\alpha_{1p}) + s_{11}\alpha_{11} \\ &+ s_{21}\alpha_{12} \\ &\vdots \\ &+ s_{p1}\alpha_{1p} \\ = &\, s_{11}\alpha_{11} + s_{12}\alpha_{12} + \cdots + s_{1p}\alpha_{1p} \\ &+ s_{11}\alpha_{11} + s_{21}\alpha_{12} + \cdots + s_{p1}\alpha_{1p} \\ = &\, 2(s_{11}\alpha_{11} + s_{12}\alpha_{12} + \cdots + s_{1p}\alpha_{1p})\end{aligned}$$

(last equality stems from symmetry of Σ)

In general, for $i = 1, \ldots, p$,

$$\begin{aligned}\frac{\partial}{\partial \alpha_{1i}} \alpha_1^T \Sigma \alpha_1 &= s_{i1}\alpha_{11} + s_{i2}\alpha_{12} + \cdots + s_{ip}\alpha_{1p} \\ &\quad + s_{i1}\alpha_{11} + s_{2i}\alpha_{12} + \cdots + s_{pi}\alpha_{1p} \\ &= 2(s_{i1}\alpha_{11} + s_{i2}\alpha_{12} + \cdots + s_{ip}\alpha_{1p})\end{aligned}$$

(because of symmetry of Σ)

As a consequence,

$$\nabla \alpha_1^T \Sigma \alpha_1 = 2\Sigma\alpha_1$$

So solving

$$\nabla f(x_1, \ldots, x_n) = \lambda \nabla g(x_1, \ldots, x_n)$$

means solving

$$2\Sigma\alpha_1 = \lambda 2\alpha_1$$

i.e.,

$$\Sigma\alpha_1 = \lambda\alpha_1$$

$\implies$ (λ, α_1) eigenpair of Σ, with α_1 having unit length

Picking the right eigenvalue

(λ, α_1) eigenpair of Σ, with α_1 having unit length

But which λ to choose?

Recall that we want $\alpha_1^T = \alpha_1^T \Sigma \alpha_1$ maximal

We have

$$\alpha_1^T = \alpha_1^T \Sigma \alpha_1 = \alpha_1^T (\Sigma \alpha_1) = \alpha_1^T (\lambda \alpha_1) = \lambda (\alpha_1^T \alpha_1) = \lambda$$

$\implies$ we pick $\lambda = \lambda_1$, the largest eigenvalue (covariance matrix symmetric so eigenvalues real)

What we have this far..

The first principal component is α_1^T and has variance λ_1, where λ_1 the largest eigenvalue of Σ and α_1 an associated eigenvector with $\|\alpha_1\| = 1$

We want the second principal component to be *uncorrelated* with α_1^T and to have maximum variance $\alpha_2^T = \alpha_2^T \Sigma \alpha_2$, under the constraint that $\|\alpha_2\| = 1$

α_2^T uncorrelated to α_1^T if $(\alpha_1^T, \alpha_2^T) = 0$

We have

$$\begin{aligned}(\alpha_1^T, \alpha_2^T) &= \alpha_1^T \Sigma \alpha_2 \\ &= \alpha_2^T \Sigma^T \alpha_1 \\ &= \alpha_2^T \Sigma \alpha_1 \quad [\Sigma \text{ symmetric}] \\ &= \alpha_2^T (\lambda_1 \alpha_1) \\ &= \lambda \alpha_2^T \alpha_1\end{aligned}$$

So α_2^T uncorrelated to α_1^T if $\alpha_1 \perp \alpha_2$

This is beginning to sound a lot like Gram-Schmidt, no?

In short

Take whatever covariance matrix is available to you (known Σ or sample S_X) – assume sample from now on for simplicity

For $i = 1, \ldots, p$, the ith principal component is

$$z_i = {}_i^T$$

where ${}_i$ eigenvector of S_X associated to the ith largest eigenvalue λ_i

If ${}_i$ is normalised, then $\lambda_i = z_k$

Covariance matrix

Σ the covariance matrix of the random variable, S_X the sample covariance matrix

$X \in_{mp}$ the data, then the (sample) covariance matrix S_X takes the form

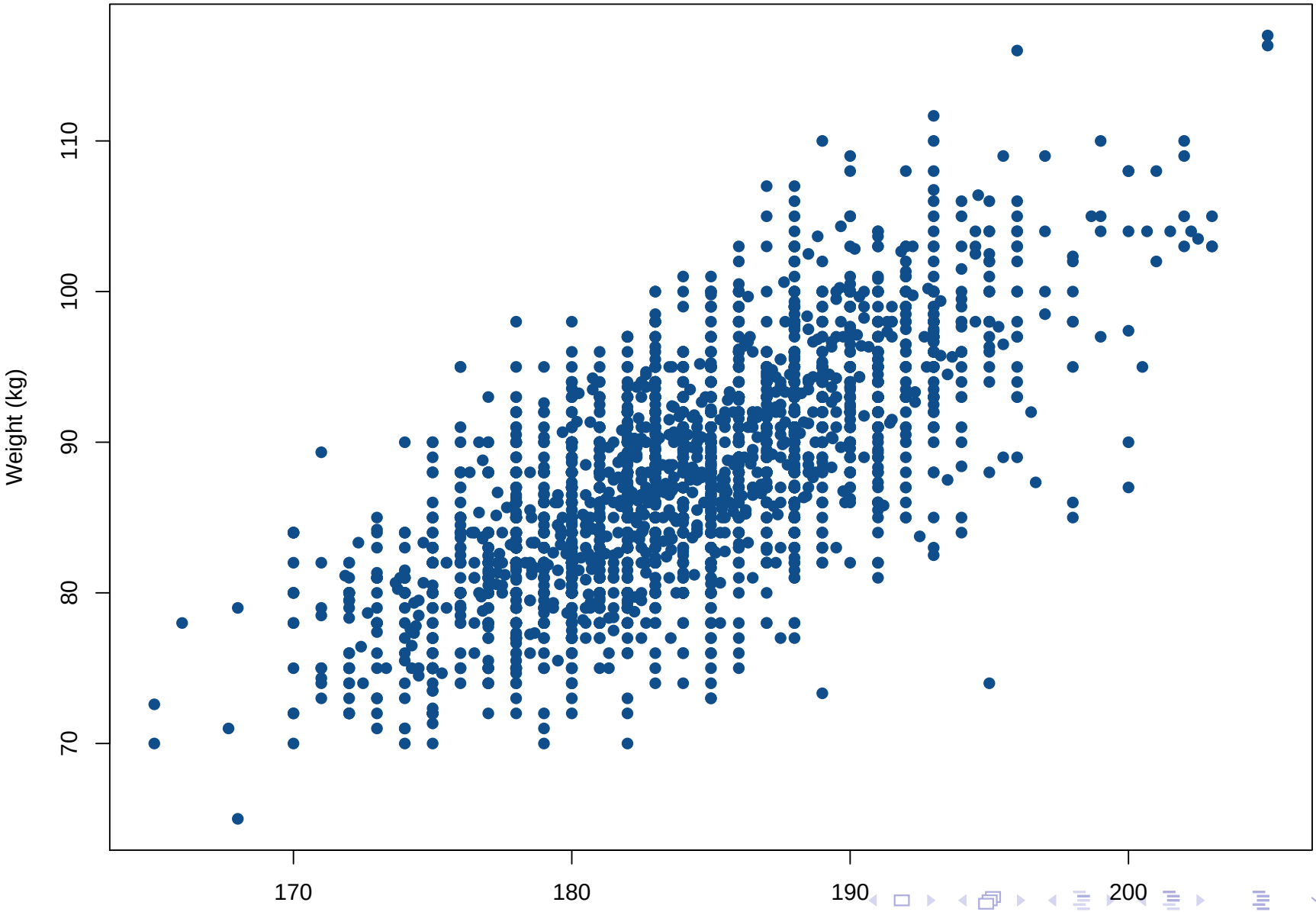
$$S_X = \frac{1}{n-1} X^T X$$

where the data is centred!

Sometimes you will see $S_X = 1/(n-1)XX^T$. This is for matrices with observations in columns and variables in rows. Just remember that you want the covariance matrix to have size the number of variables, not observations, this will give you the order in which to take the product

A smaller 2D example

Hockey players at IIHF world championships 2001-2016



Centre the data

Subtract the mean (our first – simple – change of basis)