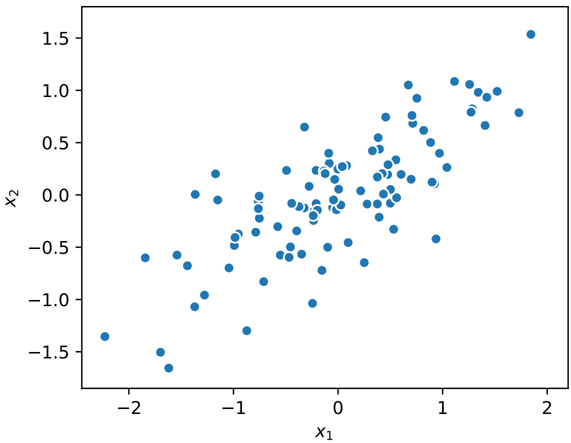
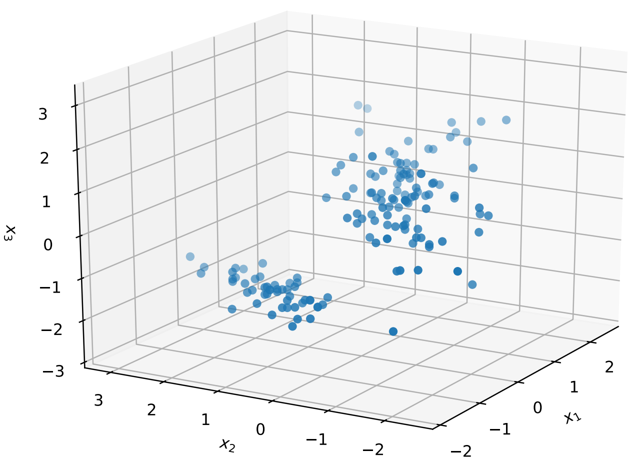
Principal components analysis

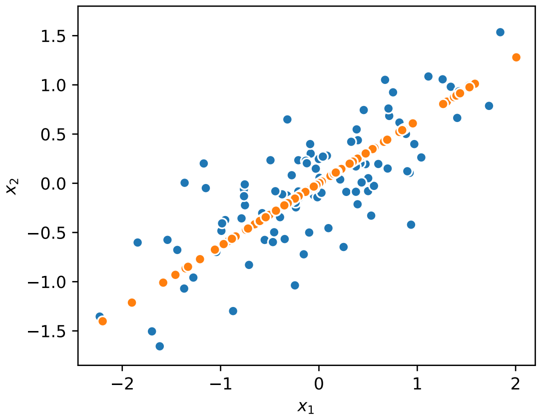
Herman Kamper

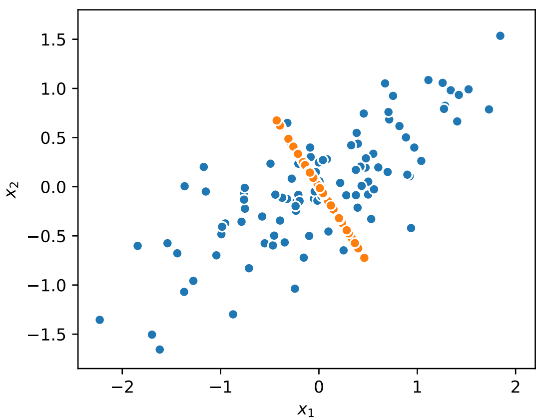
Basic idea

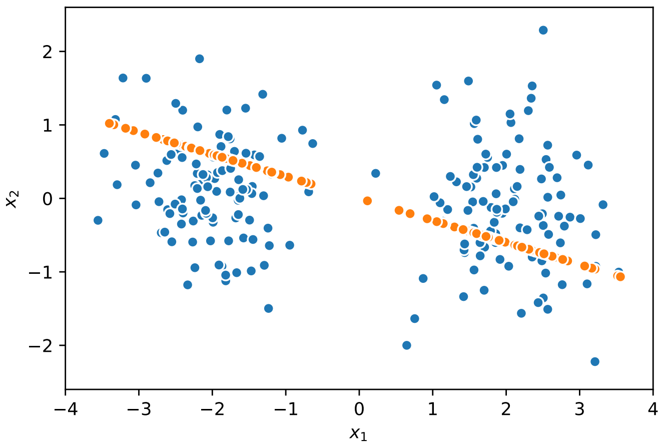




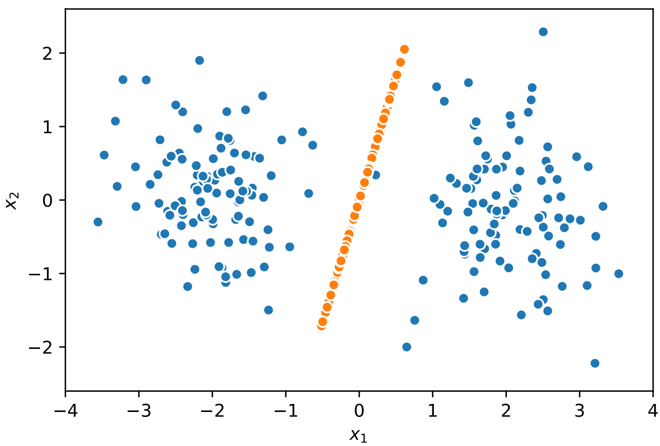
View 1: Maximising variance





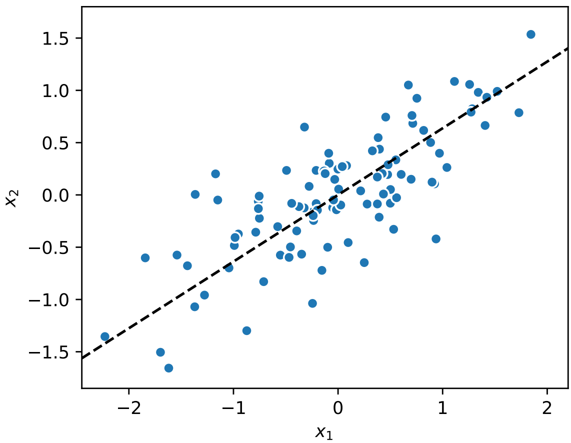


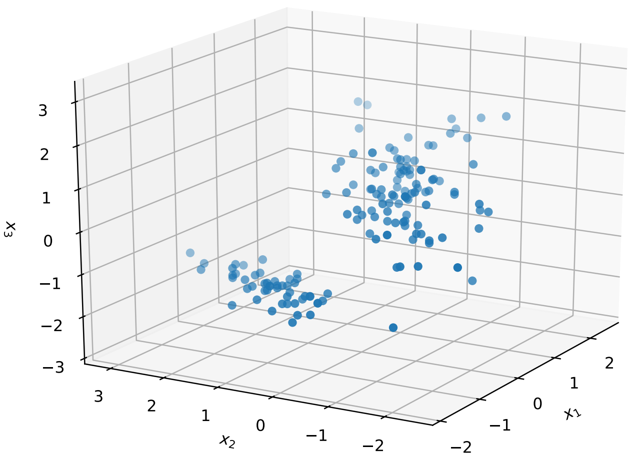
2
1
0
−1
−2
x_2
−4
−3
−2
−1
0
1
2
3
4
x_1

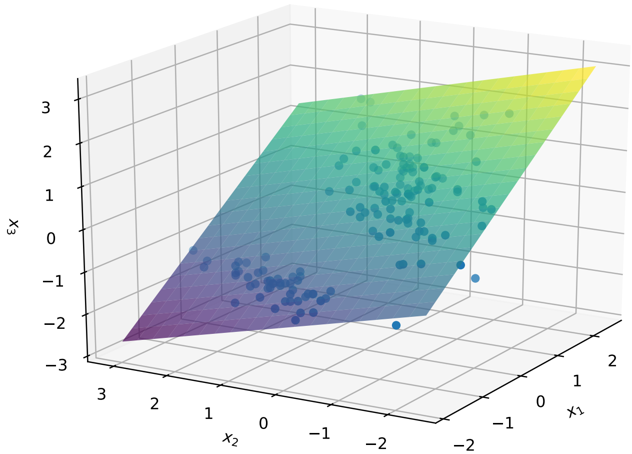


2
1
0
−1
−2
x_2
−4
−3
−2
−1
0
1
2
3
4
x_1

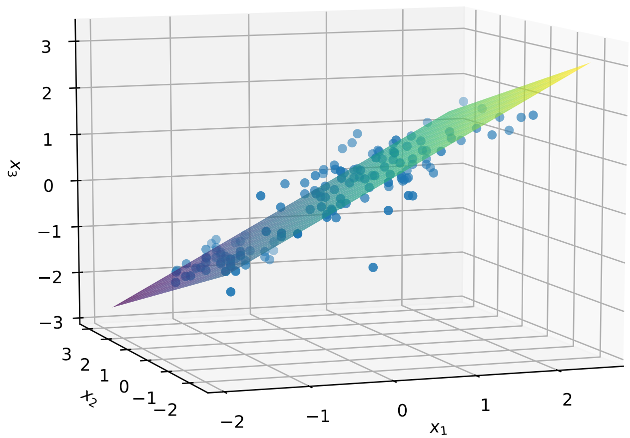
View 2: Minimising reconstruction error







x_1
x_2
x_3



x_1
x_2
x_3

Preliminaries

Lagrange multipliers

Want to optimise $f(x)$ subject to some constraint $g(x) = 0$.

We define a new objective:

$$J(x, \lambda) = f(x) + \lambda g(x)$$

and optimise w.r.t. both x and λ.

Eigenvalues and eigenvectors

For a square matrix $\mathbf{A}$:

$$\mathbf{A}\mathbf{u} = \lambda \mathbf{u}$$

The solutions to this equation are pairs of eigenvalues λ with eigenvectors $\mathbf{u}$.

Vector derivatives

See the notes and videos on vector and matrix derivatives (03b).

PCA setup

We want to project $\mathbf{x}^{(n)} \in \mathbb{R}^D$ to $\mathbf{z}^{(n)} \in \mathbb{R}^M$, with $M < D$.

(We normally assume that the inputs $\mathbf{x}$ have been normalised to have a zero mean.)

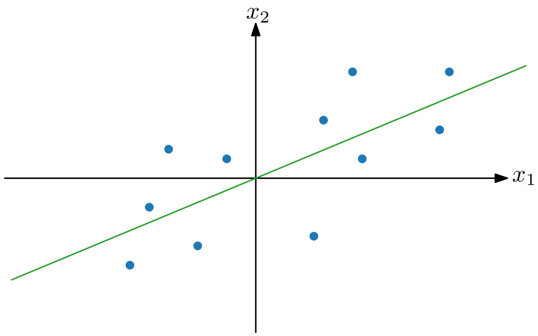
To do the projection, we use M projection vectors $\mathbf{w}_m \in \mathbb{R}^D$.

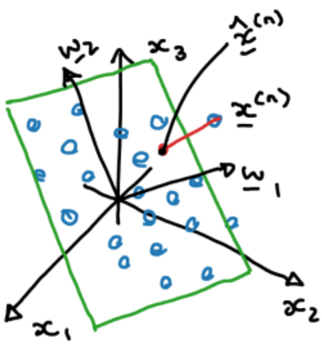
The projection vectors $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M$ are unit length and orthogonal:

$$\|\mathbf{w}_m\| = 1 \qquad \text{and} \qquad \mathbf{w}_i^\top \mathbf{w}_j = 0 \quad \forall\ i \neq j$$

The projection of the n^{th} item $\mathbf{x}^{(n)}$ onto the m^{th} dimension is

$$z_m^{(n)} = \mathbf{w}_m^\top \mathbf{x}^{(n)}$$



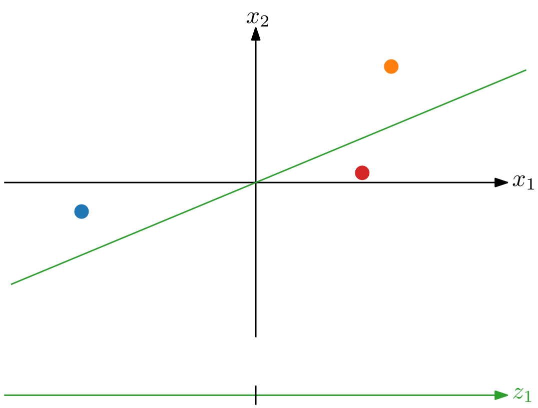


Projection

So $\mathbf{x}^{(n)}$ is mapped to

$$\mathbf{z}^{(n)} =$$

Reconstruction



Finding the projection vectors

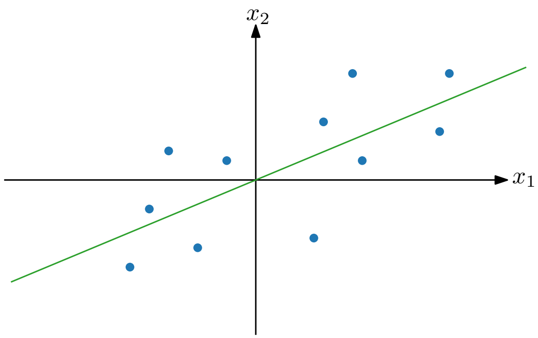
Setup

- Data that have been mean-normalised: $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(N)}$
- Want to find projection vectors: $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M$
- Unit length: $\|\mathbf{w}_m\| = 1$
- Orthogonal: $\mathbf{w}_i^\top \mathbf{w}_j = 0 \quad \forall\ i \neq j$

Objective

Want to find $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M$ so that (sample) variance is maximised.

Let's first just look at one dimension.



$\hat{\sigma}^2_{z_1} =$

We want to maximise $\hat{\sigma}_{z_1}^2$ subject to $\|\mathbf{w}_1\| = 1$, i.e. $\mathbf{w}_1^\top \mathbf{w}_1 = 1$. We use a Lagrange multiplier:

$J(\mathbf{w}_1) =$

Minimise the loss w.r.t. $\mathbf{w}_1$:

Which eigenvector-value do we use?

So pick eigenvector corresponding to the largest eigenvalue.

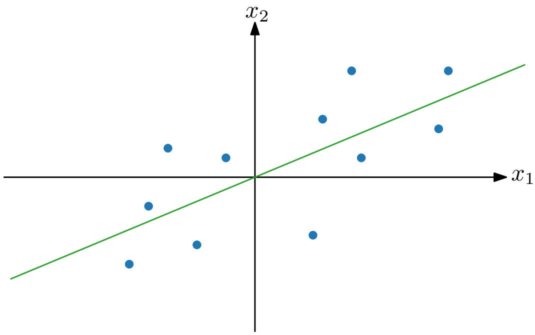
How do we find $\mathbf{w}_2$, with $\|\mathbf{w}_2\| = 1$, and $\mathbf{w}_1^\top \mathbf{w}_2 = 1$? Repeat above steps:

$$\hat{\mathbf{\Sigma}}\mathbf{w}_2 = \lambda_2 \mathbf{w}_2$$

So pick eigenvector corresponding to the second largest eigenvalue. Etc.

PCA view 2: Minimising the reconstruction error

Instead of maximising variance, we think of PCA as minimising the reconstruction loss:



Let's show that these two views are the same:

$J(\mathbf{w}_1) =$

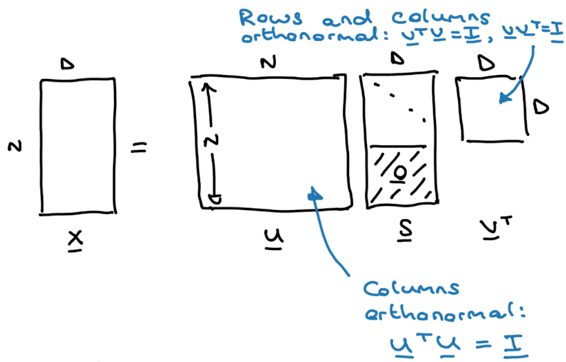
(Here we looked at projection to one dimension, but you can follow the same steps for projection to $M > 1$ dimensions.)

The relationship of PCA to SVD

(Not examinable.)

Singular value decomposition (SVD)

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^\top$$



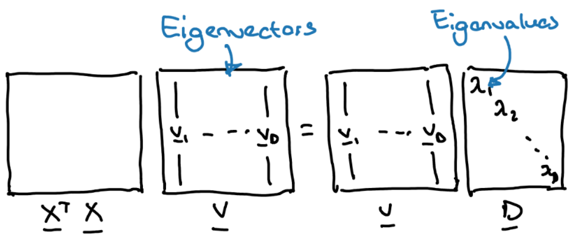
Relationship to PCA

Take SVD of the design matrix $\mathbf{X}$:

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^\top$$

Then

$$\begin{aligned}
\mathbf{X}^\top\mathbf{X} &= \mathbf{V}\mathbf{S}^\top\mathbf{U}^\top\mathbf{U}\mathbf{S}\mathbf{V}^\top \\
&= \mathbf{V}\mathbf{S}^\top\mathbf{S}\mathbf{V}^\top \\
&= \mathbf{V}\mathbf{D}\mathbf{V}^\top \\
\left(\mathbf{X}^\top\mathbf{X}\right)\mathbf{V} &= \mathbf{V}\mathbf{D}
\end{aligned}$$



PCA algorithm

1. Normalise the data to be zero-mean.
2. Calculate the sample covariance matrix.
3. Find the D eigenvector-eigenvalue pairs of the sample covariance matrix.
4. Choose the M eigenvectors corresponding to the highest eigenvalues.
5. Project the data to the lower-dimensional space.

Videos covered in this note

- PCA 1 - Introduction (16 min)
- PCA 2 - Mathematical background (7 min)
- PCA 3 - Setup (17 min)
- PCA 4 - Learning (19 min)
- PCA 5 - Minimising reconstruction (7 min)
- PCA 6 - Relationship to SVD (9 min) (not examinable)
- PCA 7 - Steps (6 min)

Reading

- ISLR 12.2 (excluding 12.2.3, although this is interesting)