LTAT.02.004 MACHINE LEARNING II

Affine data projections

based on normal distribution

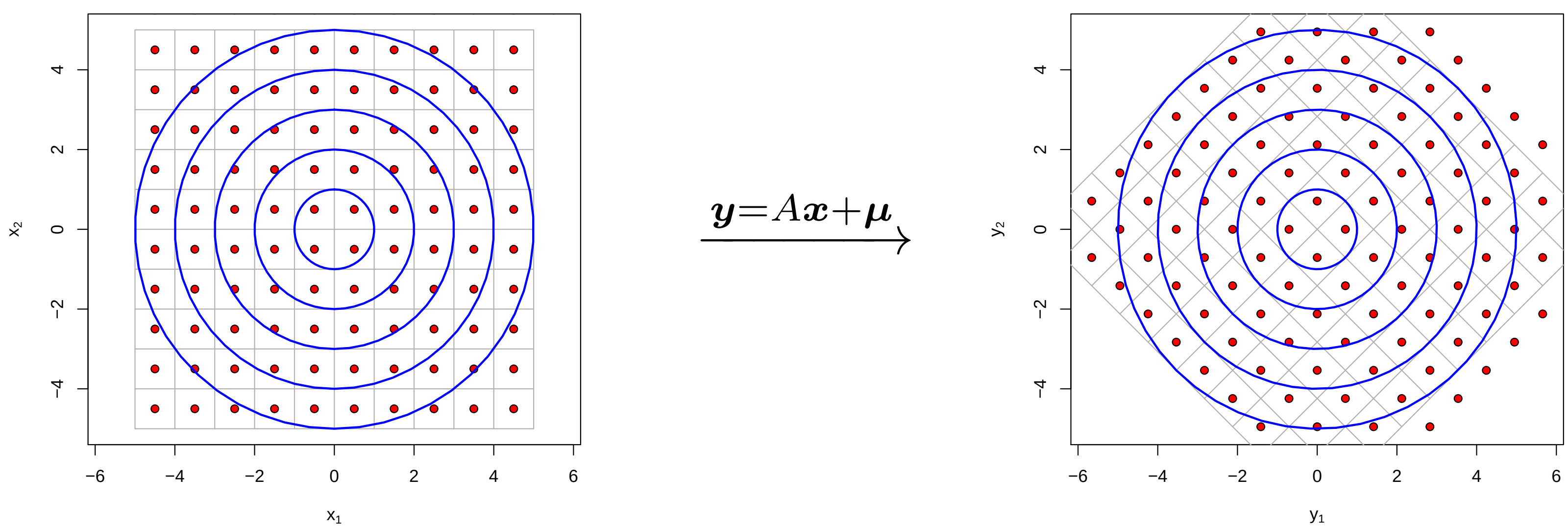
Sven Laur
University of Tartu

Principal component analysis

Distribution reconstruction task

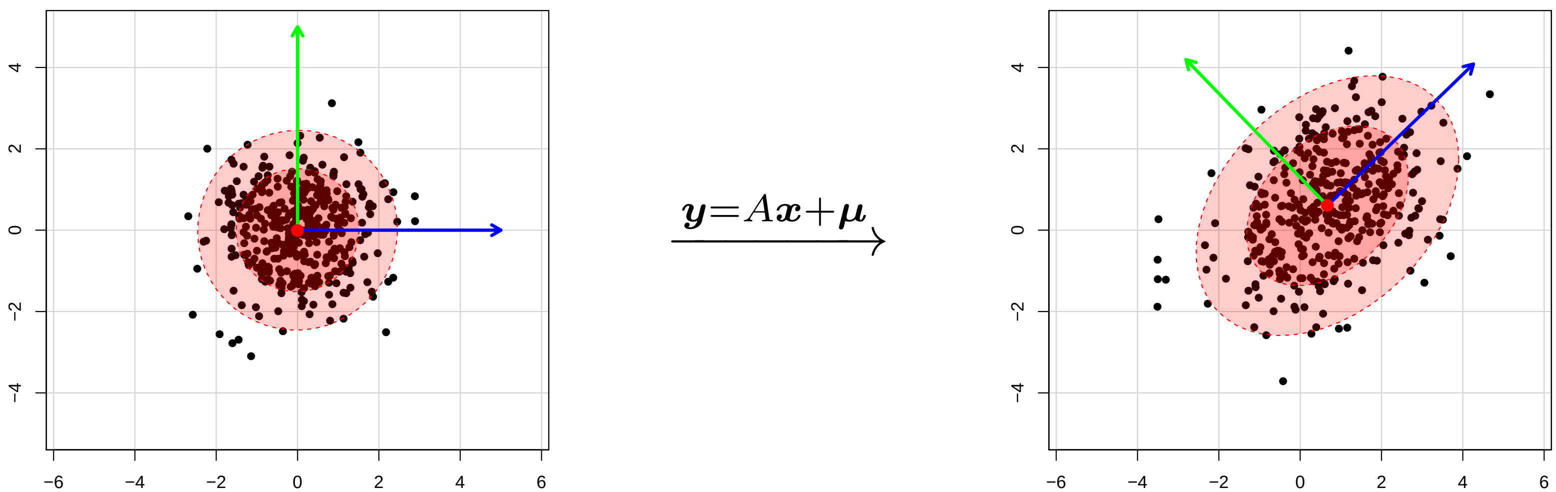
Original goal. Given the set of observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m$ determine the affine transformation $\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{\mu}$ and original source signals $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$.

Impossibility result. The matrix A can be recovered *only* up to rotations.



Simplified distribution reconstruction task

Achievable goal. Given the set of observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m$ determine the affine transformation by fixing the centre and axis of the ellipsoid.



▷ We need to find the origin and semi-axes $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ of the ellipsoid.

▷ Unit vectors $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n$ are mapped to semi-axes $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ of ellipsoid.

Variance for a fixed direction

Fact. Ortogonal projection onto a unit vector $\boldsymbol{w}$ is given by scalar product.

Question. What is the direction $\boldsymbol{w}$ that maximises the variance for ellipsoid?

$$\mathbf{Var}(\boldsymbol{w}^T \operatorname{diag}(\boldsymbol{a})\boldsymbol{x}) = \mathbf{Var}\left(\sum_{i=1}^{n} w_i a_i x_i\right) = \sum_{i=1}^{n} w_i^2 a_i^2 \enspace .$$

The variance is maximised in the direction of the longest ellipse axis a_1.

Question. How is the center of the ellipsoid and mean values connected?

$$\mathbf{E}(A\boldsymbol{x} + \boldsymbol{\mu}) = \mathbf{E}(A\boldsymbol{x}) + \mathbf{E}(\boldsymbol{\mu}) = \boldsymbol{\mu} \enspace .$$

Principal component analysis

▷ Compute the average value of the observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m$:

$$\hat{\boldsymbol{\mu}} \leftarrow \frac{\boldsymbol{y}_1 + \cdots + \boldsymbol{y}_m}{m} \quad .$$

▷ Centre the data by substituting $\hat{\boldsymbol{\mu}}$:

$$\boldsymbol{y}_i \leftarrow \boldsymbol{y}_i - \hat{\boldsymbol{\mu}}, \qquad i \in \{1, \ldots, m\} \quad .$$

▷ Find the unit direction $\boldsymbol{w}_1$ that has *a maximal empirical* variance:

$$F(\boldsymbol{w}) = \mathbf{Var}(\boldsymbol{w}^T \boldsymbol{y}_1, \ldots, \boldsymbol{w}^T \boldsymbol{y}_n) = \frac{(\boldsymbol{w}^T \boldsymbol{y}_1)^2 + \cdots + (\boldsymbol{w}^T \boldsymbol{y}_m)^2}{m} \quad .$$

▷ Find unit directions $\boldsymbol{w}_i$ orthogonal to previous directions that maximise the empirical variance of the corresponding the projection onto $\boldsymbol{w}_i$.

Covariance matrix and optimisation goal

We can use matrix algebra to simplify the variance estimate

$$\begin{aligned} F(\boldsymbol{w}) &= \frac{1}{m} \cdot \Big(\boldsymbol{w}^T \boldsymbol{y}_1 \boldsymbol{y}_1^T \boldsymbol{w} + \cdots + \boldsymbol{w}^T \boldsymbol{y}_m \boldsymbol{y}_m^T \boldsymbol{w}\Big) \\ &= \boldsymbol{w}^T \left(\frac{\boldsymbol{y}_1 \boldsymbol{y}_1^T + \cdots + \boldsymbol{y}_m \boldsymbol{y}_m^T}{m}\right) \boldsymbol{w} \end{aligned}$$

The $n \times n$ matrix in the middle is known as a *covariance matrix* Σ.

Due to the restriction $\|\boldsymbol{w}\|_2^2 = \boldsymbol{w}^T \boldsymbol{w} = 1$, we have to use Lagrange' trick:

$$F_*(\boldsymbol{w}) = \boldsymbol{w}^T \Sigma \boldsymbol{w} - 2\lambda \boldsymbol{w}^T \boldsymbol{w} \qquad \Rightarrow \qquad \frac{\partial F_*(\boldsymbol{w})}{\partial \boldsymbol{w}} = 2\Sigma \boldsymbol{w} - 2\lambda \boldsymbol{w} = \boldsymbol{0}.$$

Principal components as eigenvectors

The $F_*(\boldsymbol{w})$ is maximised only if the direction $\boldsymbol{w}$ is an *eigenvector* of Σ:

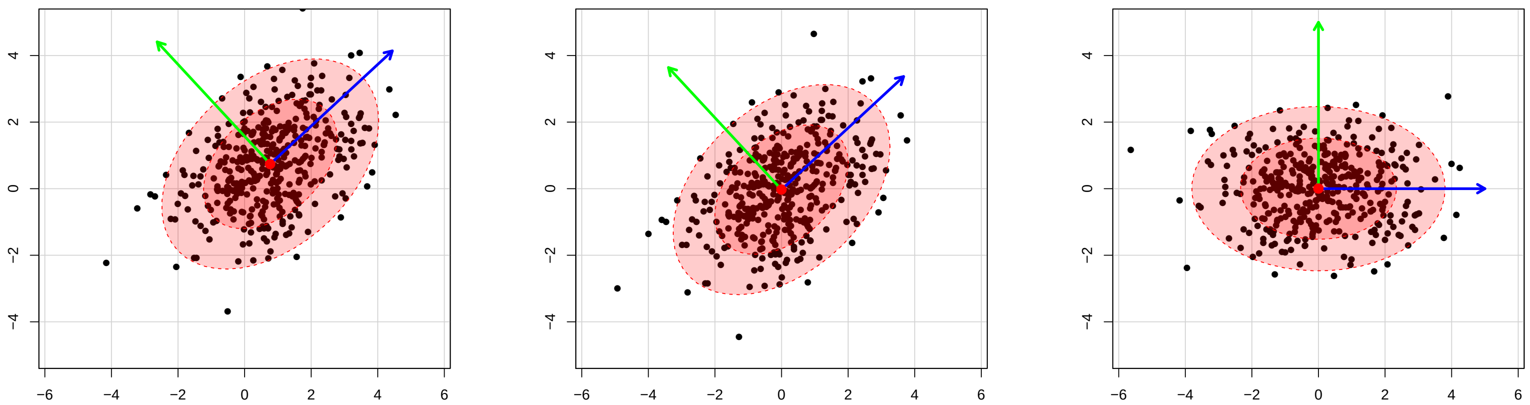
$$\Sigma \boldsymbol{w} = \lambda \boldsymbol{w} \qquad \Rightarrow \qquad \boldsymbol{w}^T \Sigma \boldsymbol{w} = \boldsymbol{w}^T \lambda \boldsymbol{w} = \lambda \ .$$

Fact. If $n \times n$ matrix is symmetric and positively definite then there exists n orthogonal eigenvectors $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$ with *eigenvalues* $\lambda_1 \geq \ldots \geq \lambda_n > 0$.

Corollary. Principal components corresponding to observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m$ are the eigenvectors of the covariance matrix Σ.

Principal component analysis as a rotation

Reconstruction of the source signal can be viewed as a *translation* followed by a *rotation* to orientate the ellipsoid wrt coordinate axis.



As vectors $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$ are orthogonal, the rotation can be done through computing projections (read scalar products):

$$\hat{\boldsymbol{x}}_i = (\boldsymbol{w}_1 || \cdots || \boldsymbol{w}_n)^T (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_0) = W(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}) \enspace .$$

Maximum likelihood estimate

The algorithm formulated above was based on *ad hoc* reasoning:

▷ Empirical estimates for the mean and variance are not precise!

Theoretically correct way to handle the problem is

▷ obtain the maximum likelihood estimate on the model parameters,

▷ determine the translation and rotation based on the model parameters.

What are the model parameters?

▷ Parameters of the density formula Σ and $\boldsymbol{\mu}$.

▷ Parameters of the affine transformation A and $\boldsymbol{\mu}$.

Likelihood function under iid assumption

If all observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m$ are independent then

$$p[\boldsymbol{y}_i, \ldots, \boldsymbol{y}_m | \Sigma, \boldsymbol{\mu}] = \prod_{i=1}^{m} p[\boldsymbol{y}_i | \Sigma, \boldsymbol{\mu}]$$

where

$$p[\boldsymbol{y}_i | \Sigma, \boldsymbol{\mu}] = \frac{1}{(2\pi)^{n/2}} \cdot \frac{1}{\sqrt{\det(\Sigma)}} \cdot \exp\left(-\frac{(\boldsymbol{y}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu})}{2}\right)$$

The *log-likelihood* of the data $\ln p[\boldsymbol{y}_i, \ldots, \boldsymbol{y}_m | \Sigma, \boldsymbol{\mu}]$ can be expressed

$$\mathcal{L}(\Sigma, \boldsymbol{\mu}) = const + \frac{m}{2} \cdot \ln\det(\Sigma^{-1}) - \sum_{i=1}^{m} \frac{(\boldsymbol{y}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu})}{2}$$

Now we have to find the arrangement $(\Sigma, \boldsymbol{\mu})$ that maximises $\mathcal{L}(\Sigma, \boldsymbol{\mu})$.

Gradients of the log-likelihood function

Gradient with respect to the shift $\boldsymbol{\mu}$:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}} = -\sum_{i=1}^{m} \frac{\partial}{\partial \boldsymbol{\mu}} \frac{(\boldsymbol{y}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu})}{2} = -\sum_{i=1}^{m} \frac{\Sigma^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu})}{2} \cdot (-1)$$

Gradient with respect to the inverse matrix Σ^{-1}:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial (\Sigma^{-1})} &= \frac{m}{2} \cdot \frac{\partial}{\partial (\Sigma^{-1})} \ln \det(\Sigma^{-1}) - \sum_{i=1}^{m} \frac{\partial}{\partial (\Sigma^{-1})} \frac{(\boldsymbol{y}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu})}{2} \\ &= \frac{m}{2} \cdot \Sigma^T - \sum_{i=1}^{m} \frac{(\boldsymbol{y}_i - \boldsymbol{\mu})^T (\boldsymbol{y}_i - \boldsymbol{\mu})}{2}\end{aligned}$$

As Σ is symmetric and Σ^{-1} exists we can derive closed form solutions.

Maximum likelihood estimates for parameters

The shift must be the mean of all observations

$$\boldsymbol{\mu} = \frac{1}{m} \cdot \sum_{i=1}^{m} \boldsymbol{y}_i \enspace .$$

The covariance matrix

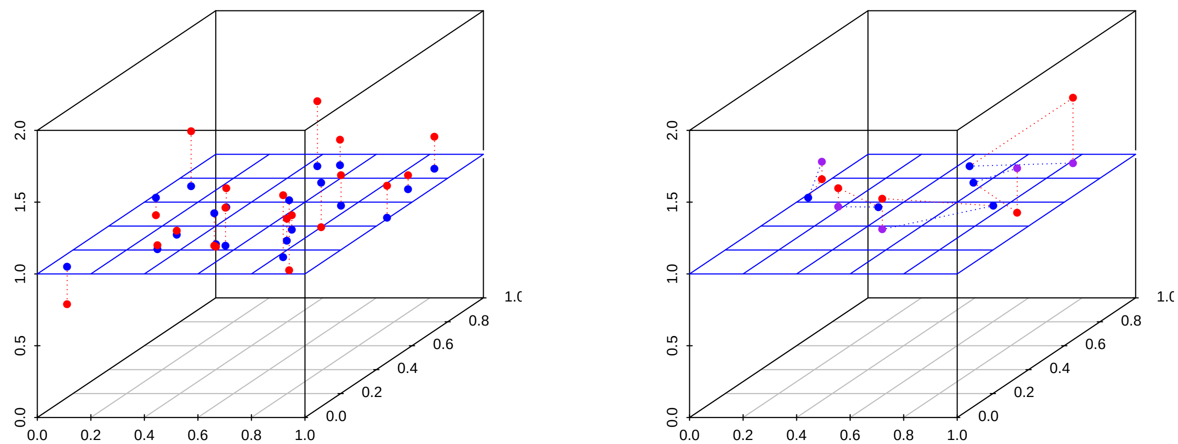
$$\Sigma = \frac{1}{m} \cdot \sum_{i=1}^{m} (\boldsymbol{y}_i - \boldsymbol{\mu})^T (\boldsymbol{y}_i - \boldsymbol{\mu})$$

Correctness of PCA. As ML estimates are exactly the same we used in principal component analysis, the method is theoretically justified!

Principal component analysis
Alternative formalisations

Dimensionality reduction

What if the actual data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$ lies in a lower-dimensional plane and the observation $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m$ are obtained by random shifts?



The shifts can be either orthogonal to the plane or just random. The first model is easier to analyse while the second is more plausible.

Maximum likelihood estimate

Let $\mathcal{H}$ be the plane. Assume that the random shifts ε_i are orthogonal to the plane and have a normal distribution $\mathcal{N}(0, \sigma I)$. Then

$$p[\boldsymbol{y}_i|\mathcal{H}, \sigma] = const \cdot \exp\left(-\frac{d_i^2}{2\sigma^2}\right)$$

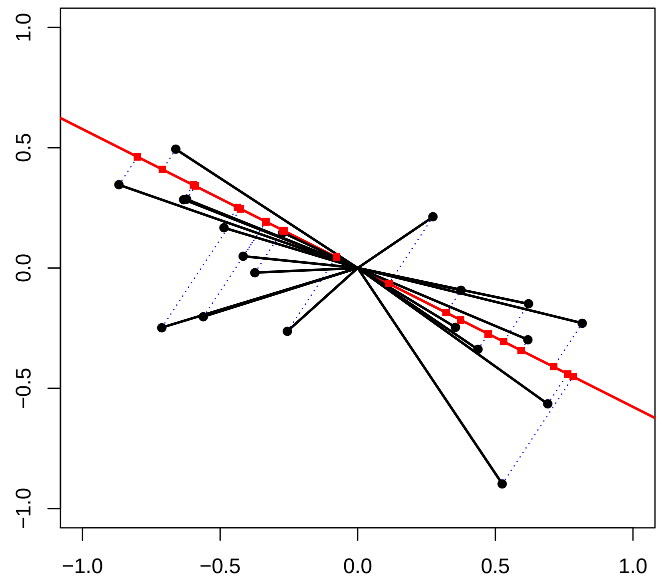
where d_i is the distance between the plane $\mathcal{H}$ and the point $\boldsymbol{y}_i$. Thus

$$p[\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m|\mathcal{H}, \sigma] = const \cdot \exp\left(-\sum_{i=1}^{m} \frac{d_i^2}{2\sigma^2}\right)$$

and the maximum likelihood estimate of the plane minimises sum of the distance squares. Corresponding estimates of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$ are projections of $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m$ to the plane $\mathcal{H}$.

Another characterisation of PCA

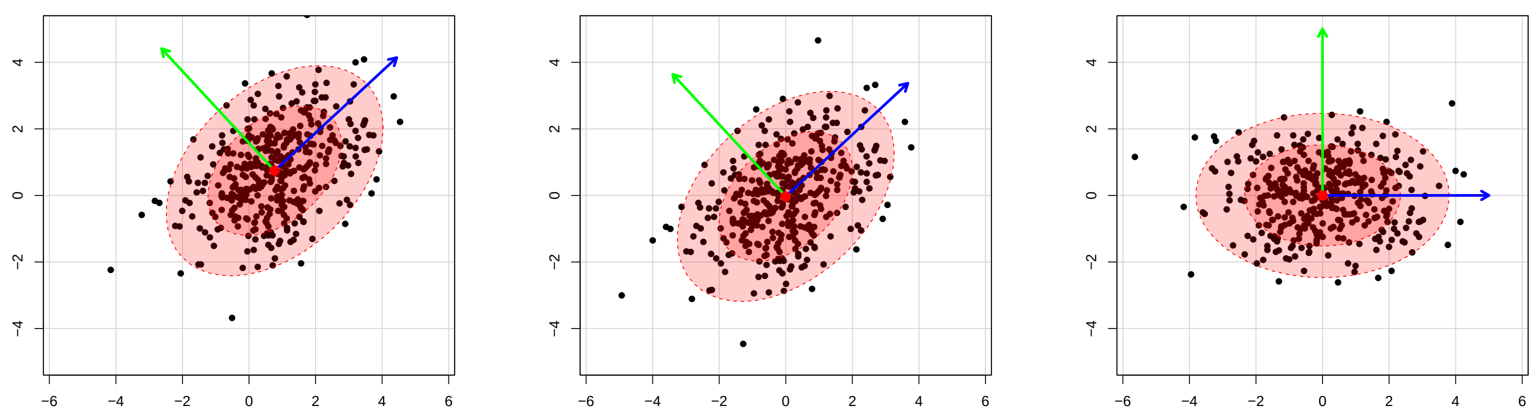
Fact. If the data is centred then PCA chooses the direction $\boldsymbol{w}_1$ such that the sum of squares of the projections $\boldsymbol{w}_1^T \boldsymbol{y}_i$ is maximal.



Corollary. PCA chooses directions $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$ such that the sum of distance squares from the hyperplane formed by $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k$ is minimal.

PCA as a dimensionality reduction tool

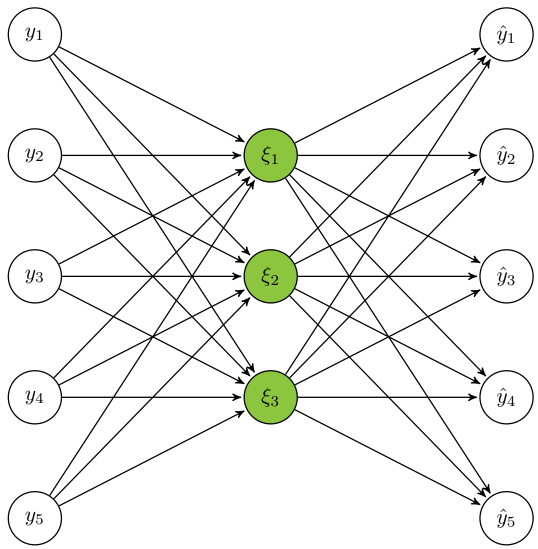
Corollary. PCA rotates the data such way that first k coordinates of the rotated data correspond to maximum likelihood reconstructions of original vectors corrupted with white Gaussian noise $\mathcal{N}(0, \sigma I)$.



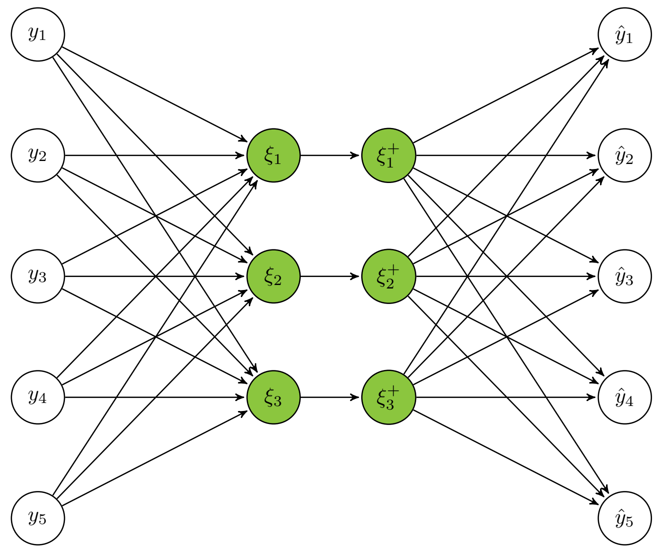
Alternatively, we can view the last components of the source signal $\boldsymbol{x}$ as the uninformative noise. The overall noise component should be small.

Connection to autoencoders

Linear autonecoder



RELU autoencoder

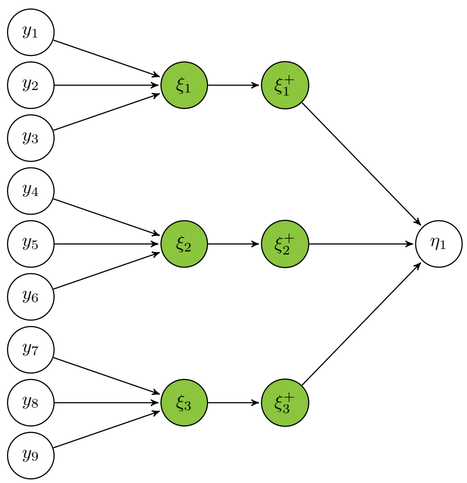


Fix mean square error as optimisation target.

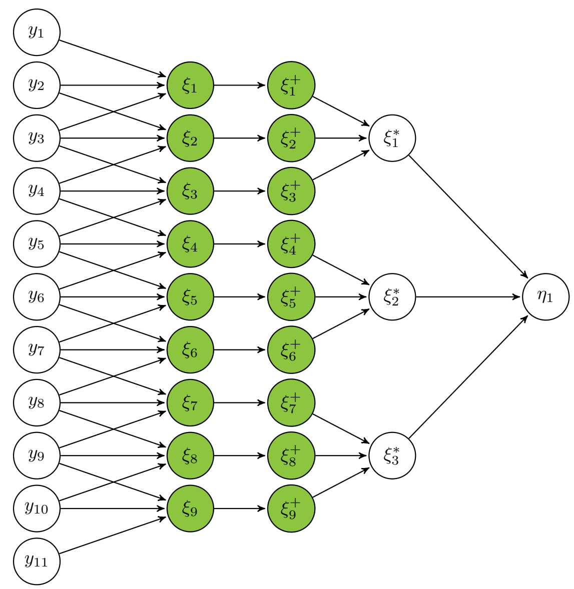
- ▷ Linear-autoencoder is a reformulation of PCA.
- ▷ RELU-autoencoder is a reformulation of non-negative matrix factorisation.

Connection to convolutional neural networks

Hierarchical NMF



Convolutional network



- ▷ Hierarchical NMF applies the same transformation on patches.
- ▷ Convolutional layer applies the same transformation to sliding windows.
- ▷ Then it picks the strongest response among shifts of the transformation.

Linear discriminant analysis

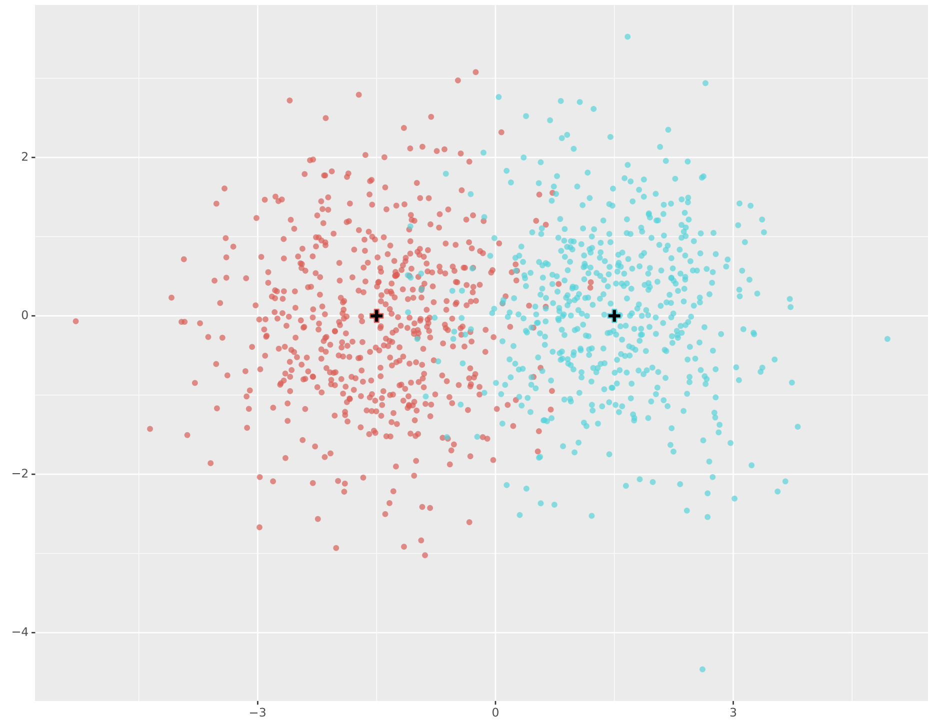
Underlying assumptions and inference task

Original goal. Given a set of observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^M$ together with class labels $z_1, \ldots, z_n \in \{1, \ldots, \ell\}$ find a linear projection $\pi : \mathbb{R}^m \to \mathbb{R}^k$ so that individual classes are maximally separated.

Assumptions.

- ▷ There are ℓ different classes.
- ▷ All observations $\boldsymbol{x}_i$ are independently sampled.
- ▷ Observations $\boldsymbol{x}_i$ with the same class label z_i come from $\mathcal{N}(\boldsymbol{\mu}_j, \Sigma)$.
- ▷ The covariance matrix Σ is shared between different distributions.

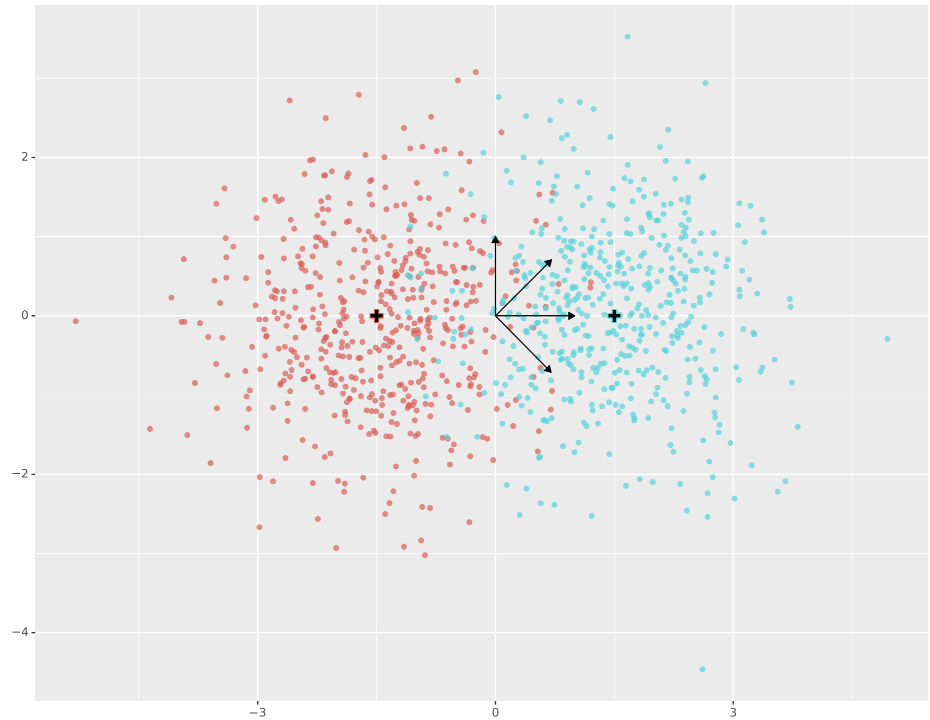
LDA for spherical normal distributions



We assume that the covariance matrix Σ is identity matrix:

- ▷ All vector components have unit variance.
- ▷ Different vector components are independent.

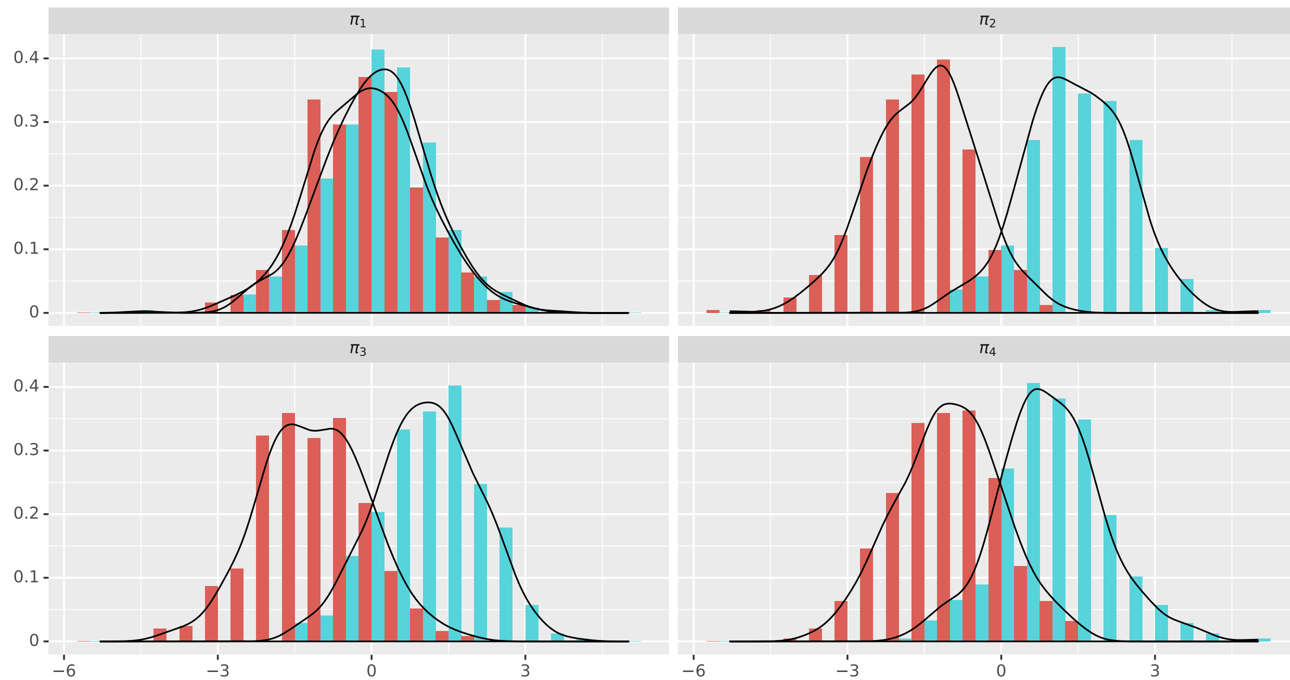
Projections to one-dimensional subspace



A projection to one-dimensional space is determined by a vector $\boldsymbol{w}$:

- ▷ To get orthogonal projection the length of $\boldsymbol{w}$ must be one.
- ▷ This can be forced by the constraint $\boldsymbol{w}^T\boldsymbol{w} = 1$.

Projections lead to different separation



We need a need a measure for assessing the goodness of separation:

- ▷ We can use Bayesian factors from statistics.
- ▷ We can use signal-to-noise ratio from signal-processing.

Choice between alternative hypotheses

- **Hypothesis $\mathcal{H}_0$.** Projections $y_i, \ldots, y_n$ come from $\mathcal{N}(\bar{y}, 1)$.
- **Hypothesis $\mathcal{H}_1$.** Projection y_i with label z_i comes from a $\mathcal{N}(\bar{y}_{z_i}, 1)$.

Hypotheses lead to following probability assignments

$$p[y_i|\mathcal{H}_0] = \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{1}{2}(y_i - \bar{y})^2\right)$$
$$p[y_i|\mathcal{H}_1] = \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{1}{2}(y_i - \bar{y}_{z_i})^2\right)$$

If we have not preference then the corresponding Bayes factor is

$$\frac{\Pr[\mathcal{H}_1|y_1, \ldots, y_n]}{\Pr[\mathcal{H}_0|y_1, \ldots, y_n]} = \exp\left(\frac{1}{2} \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2 - \frac{1}{2} \cdot \sum_{i=1}^{n}(y_i - \bar{y}_{z_i})^2\right)$$

The corresponding optimisation task

Given a set of observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^M$ together with class labels $z_1, \ldots, z_n \in \{1, \ldots, \ell\}$ find a vector $\boldsymbol{w}$ with unit length that maximises:

$$F = \sum_{i=1}^{n} (y_i - \bar{y})^2 - \sum_{i=1}^{n} (y_i - \bar{y}_{z_i})^2$$

where $\mathcal{I}_j = \{i : z_i = j\}$ is the index set and $\bar{y}$ and $\bar{y}_j$ are cluster means:

$$\bar{y} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i$$

$$\bar{y}_j = \frac{1}{|\mathcal{I}_j|} \cdot \sum_{i \in \mathcal{I}_j} y_j$$

Consequences of variance decomposition

Given a set of observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^M$ together with class labels $z_1, \ldots, z_n \in \{1, \ldots, \ell\}$ find a vector $\boldsymbol{w}$ with unit length that maximises:

$$F = \sum_{i=1}^{n} (\bar{y}_{z_i} - \bar{y})^2$$

PROOF. The result follows directly form the variance decomposition

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \bar{y}_{z_i})^2 + \sum_{i=1}^{n} (\bar{y}_{z_i} - \bar{y})^2$$

Matrix magic

Let us define centres in the original data

$$\boldsymbol{\mu} = \frac{1}{n} \cdot \sum_{i=1}^{n} \boldsymbol{x}_i \qquad\qquad \boldsymbol{\mu}_j = \frac{1}{|\mathcal{I}_j|} \cdot \sum_{i \in \mathcal{I}_j}^{n} \boldsymbol{x}_j$$

Then we can express

$$\begin{aligned} F = \sum_{i=1}^{n} (\bar{y}_{z_i} - \bar{y})^2 \quad &= \sum_{i=1}^{n} (\boldsymbol{w}^T \boldsymbol{\mu}_{z_i} - \boldsymbol{w}^T \boldsymbol{\mu})(\boldsymbol{w}^T \boldsymbol{\mu}_{z_i} - \boldsymbol{w}^T \boldsymbol{\mu})^T \\ = \boldsymbol{w}^T \left(\sum_{i=1}^{n} (\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu})(\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu})^T \right) \boldsymbol{w} \end{aligned}$$

Corresponding eigenvector task

Find a vector $\boldsymbol{w}$ with unit length that maximises

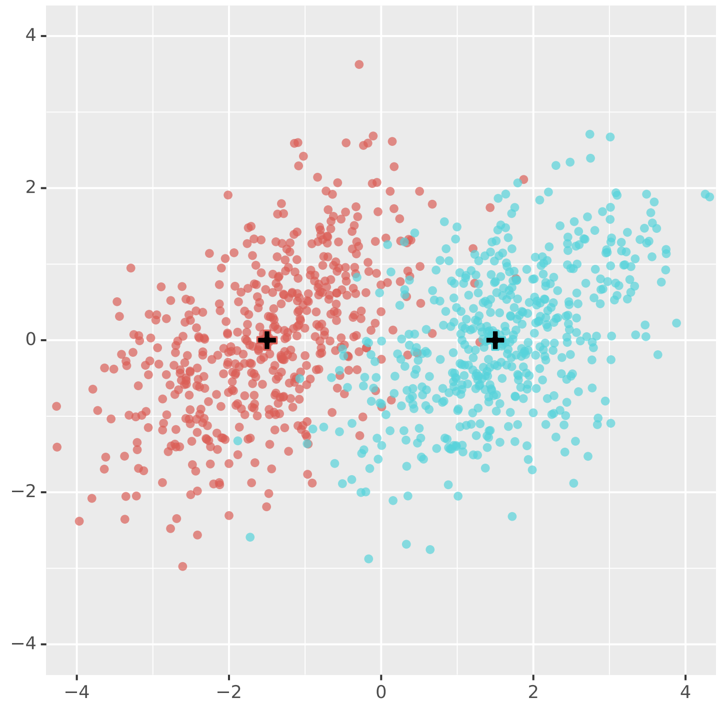
$$F = \boldsymbol{w}^T S_B \boldsymbol{w}$$

where S_B is the between class scatter matrix;

$$S_B = \sum_{i=1}^{n} (\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu})(\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu})^T \quad .$$

Consequence. The function F is maximised by the eigenvector $\boldsymbol{w}$ of S_B with the highest eigenvalue λ_1.

LDA for a normal distribution with any shape



- ▷ As we know cluster labels we can remove the effect of $\boldsymbol{\mu}_1, \ldots \boldsymbol{\mu}_\ell$.
- ▷ After that we can do affine transformation that set the covariance to I.
- ▷ We know how to solve the task in the transformed space.

Data whitening transformation

A linear transformation $\boldsymbol{x}^* = A\boldsymbol{x}$ leads to a unit covariance I if

$$A\Sigma_W A^T = I$$

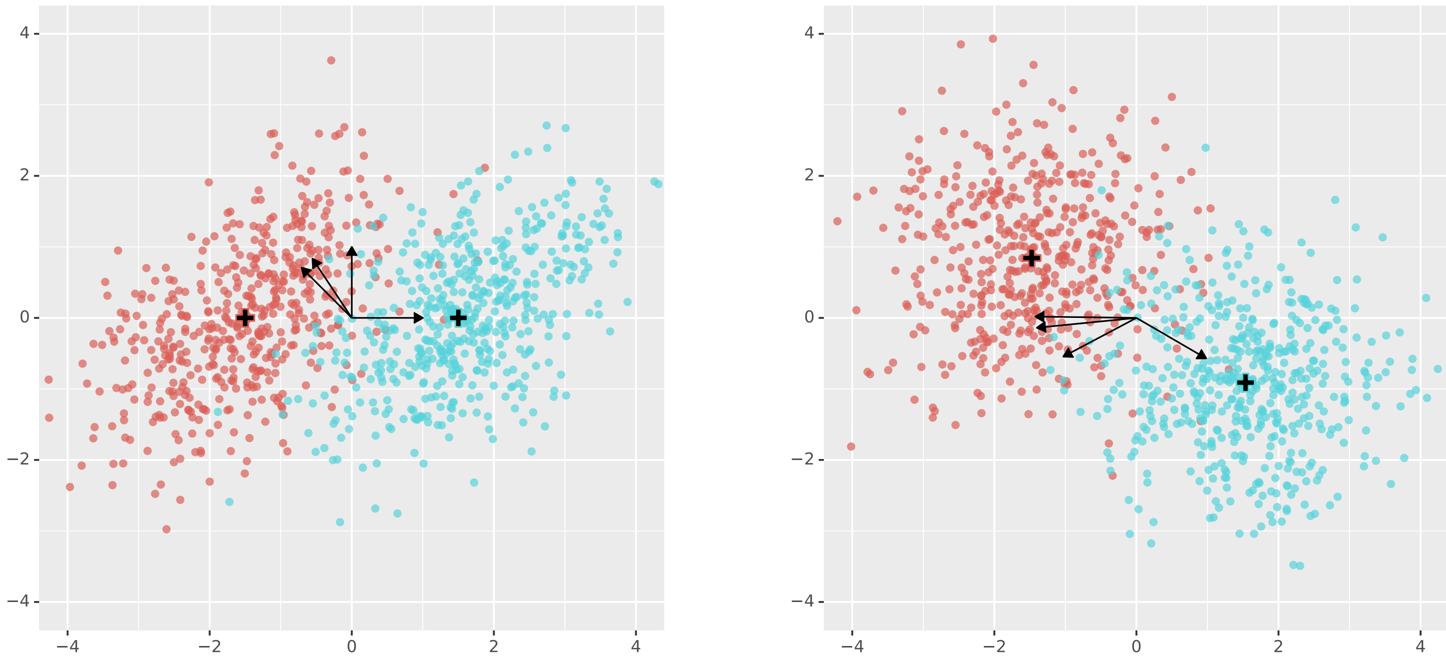
where Σ_W is within class covariance matrix:

$$\Sigma_W = \frac{1}{n} \cdot \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu}_{z_i})(\boldsymbol{x}_i - \boldsymbol{\mu}_{z_i})^T$$

Let W be the matrix where column vectors $\boldsymbol{w}_i$ are orthonormal eigenvectors with eigenvalues $\lambda_1, \ldots, \lambda_n$. Then we can express

$$A = \mathrm{diag}(\lambda_1^{-1/2}, \ldots, \lambda_n^{-1/2}) W^T \quad .$$

The effects of data whitening



- Data whitening alters probing directions: $\boldsymbol{w}^* = A\boldsymbol{w}$.
- Data whitening alters between class scatter: $S_B^* = AS_BA^T$.
- Maximisation task in original terms: $\sum_{i=1}^{k} \boldsymbol{w}_i^T \Sigma_W^{-T} S_B \Sigma_W^{-1} \boldsymbol{w}_i \to \max$
- Othogonality constraints in original terms: $\boldsymbol{w}_i^T \Sigma_W^{-1} \boldsymbol{w}_j = \delta_{ij}$.

Numerical stabilisation

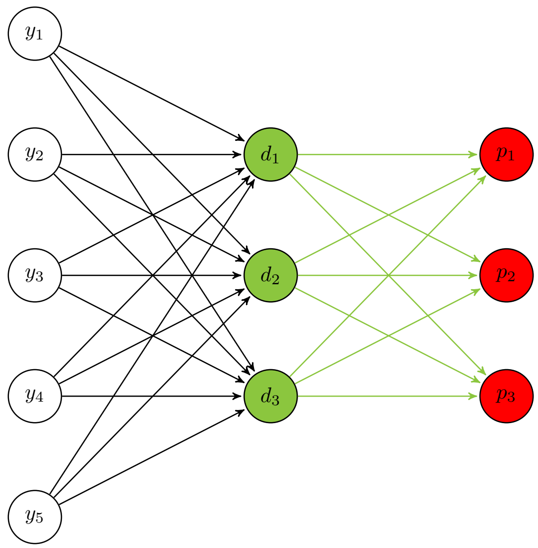
Whitening matrix Σ_W can be non-invertible and it can also depend heavily on the perturbations of original datapoints. Ridge stabilisation

$$\Sigma_W^* = \Sigma_W + \rho I$$

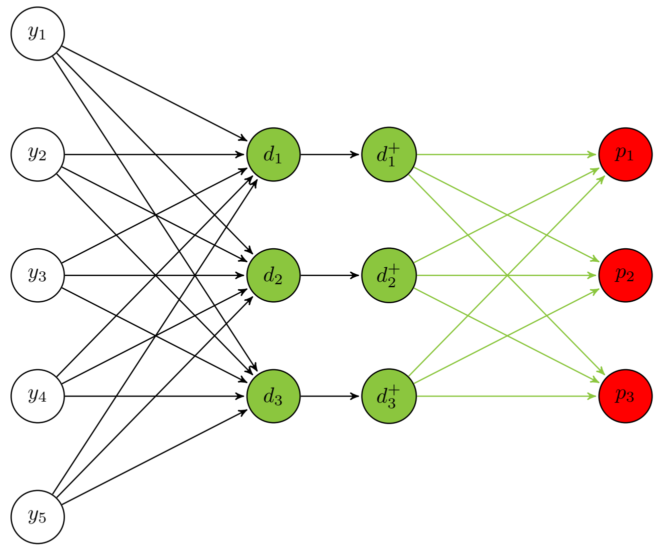
for small value $\rho > 0$ makes linear discriminant analysis more stable.

Connection to neural networks

Linear + SoftMax



RELU + SoftMax



- LDA is equivalent to a linear layer followed by a soft-max layer.
- Neural networks usually use RELU nodes instead of pure linear nodes.
- This again introduces non-negativity constraint to features.

Reconstruction vs discrimination

- PCA and LDA reduce dimensionality.
- PCA preserves recoverability of the original data.
- LDA preserves distinguishability between different classes.

Sometimes discriminative are overly selective:

- Decision is made based on minute details of the data
- Predictions are not robust against malicious perturbations.

We can quantify the balance between reconstruction vs discrimination.

- We can measure how much variation LDA projection explains.
- We can measure robustness against input perturbations.
- Same measures are applicable for other models such as neural networks.

Going beyond basics

Going beyond PCA and LDA

Weighted Principal Component Analysis:

- ▷ Sometimes data contains potential outliers.
- ▷ Sometimes we can assign reliability scores to the data points.

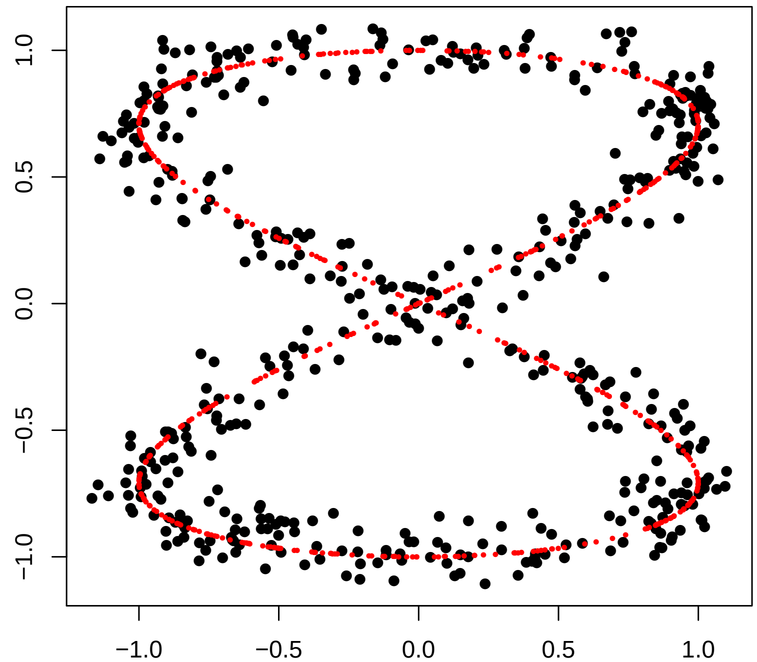
Principal curves and manifolds

- ▷ The original data might be on a low dimensional manifold.
- ▷ The observed data is corrupted by additive white gaussian noise.
- ▷ The task is to reconstruct the manifold and ML estimate for the data.

Independent Component Analysis

- ▷ What if the source components are non-gaussian?
- ▷ Then the reconstruction is possible up to scaling!

Principal curves and manifolds



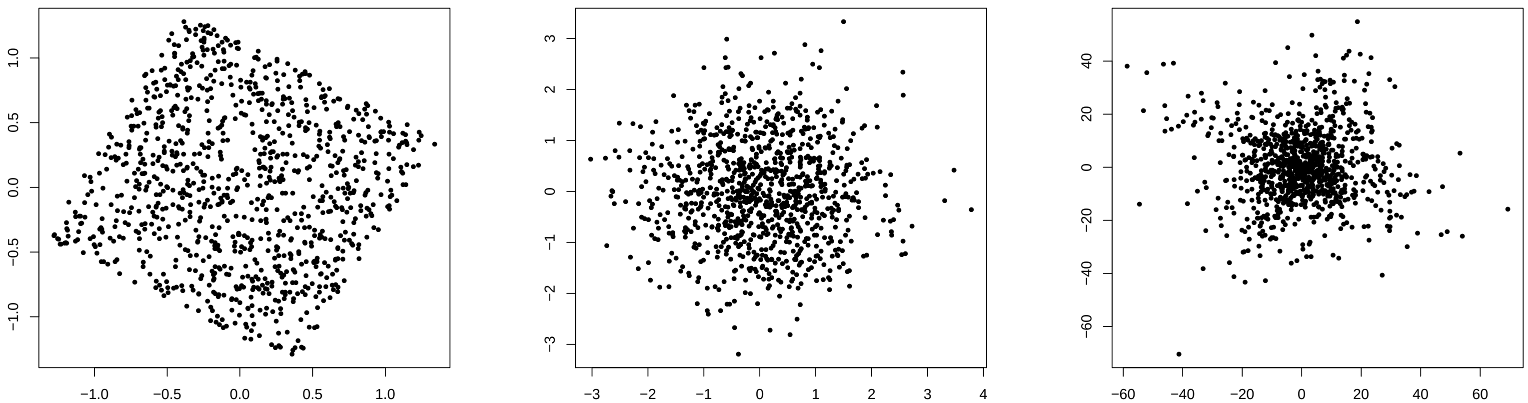
Reconstruction of the underlying curve is much more difficult.

▷ We must fix a curve parametrisation

▷ The task is different form regression since we have only outputs.

Independent Component Analysis

Assume that the components of the source data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$ are independent but an unknown affine transformation $\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{\mu}$ disturbs observations.



It is possible to recover the translation and rotation only if independent components are sufficiently different form the normal distribution.