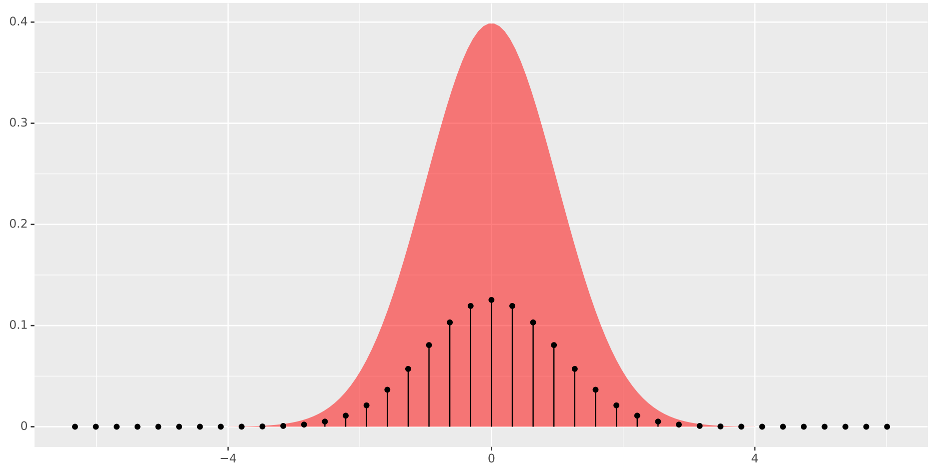
LTAT.02.004 Machine Learning II

Normal distribution and affine projections

Sven Laur
University of Tartu

Univariate normal distribution

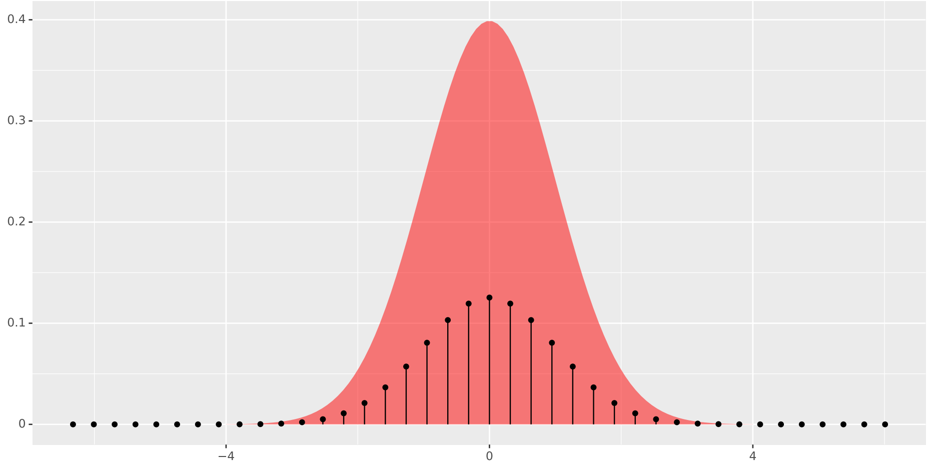
Probability density function



Definition. A real-valued random variable X comes from a continuous distribution with *a probability density function* $p : \mathbb{R} \to \mathbb{R}^{+} \cup \{0\}$ if the following limit exists for any $x \in \mathbb{R}$:

$$p(x) = \lim_{\Delta x \to 0^{+}} \frac{\Pr\left[x - \Delta x \leq X \leq x + \Delta x\right]}{2 \cdot \Delta x} \enspace .$$

Probability mass function

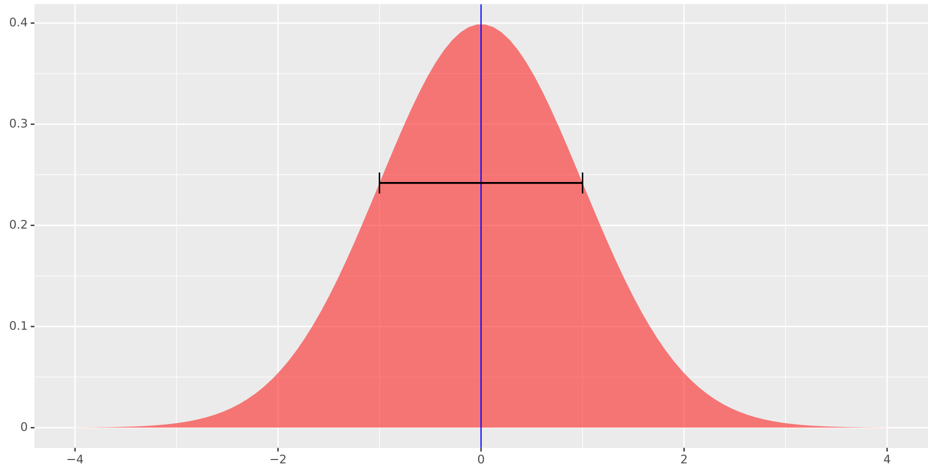


Definition. A real-valued random variable X comes from a discrete distribution with *a probability mass function* $p : \mathbb{R} \to \mathbb{R}^{+} \cup \{0\}$ defined as

$$p(x) = \Pr\left[X = x\right] = \lim_{\Delta x \to 0^{+}} \Pr\left[x - \Delta x \le X \le x + \Delta x\right]$$

if there exist a sequence $(x_i)_{i=1}^{\infty}$ such that $p(x_1) + \ldots + p(x_i) + \ldots = 1$.

Standard normal distribution



Standard normal distribution $\mathcal{N}(\mu = 0, \sigma = 1)$ is a continuous distribution with a probability density function

$$p(x) = \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{x^2}{2}\right)$$

The mean value $\mu = 0$ and variance $\sigma^2 = 1$ for this distribution.

Univariate normal distribution

Definition. A random variable y is distributed according to a normal distribution $\mathcal{N}(\mu = a, \sigma = b)$ if it can be expressed

$$y = bx + a$$

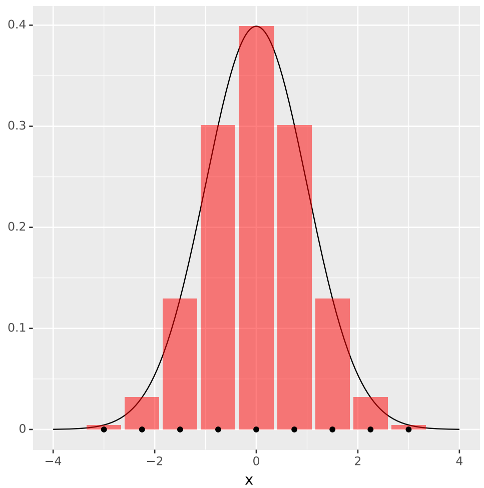
where x is distributed according to standardised normal distribution $\mathcal{N}(0, 1)$.

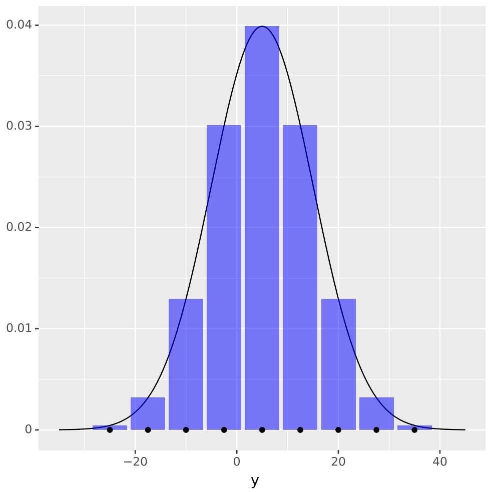
The corresponding probability density functions is

$$p[y|\mu, \sigma] = \frac{1}{\sqrt{2\pi\sigma}} \cdot \exp\left(\frac{(x-\mu)^2}{2\sigma^2}\right)$$

and the mean value μ and variance σ^2 for this distribution.

Density derivation





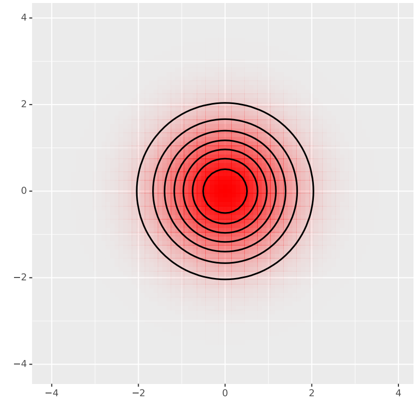
Let $y = ax + b$ the the relation between densities

$$p_x(x) = \sigma \cdot p_y(y)$$

follows form the fact that areas of red and blue columns must be the same.

Multivariate normal distribution

White Gaussian noise



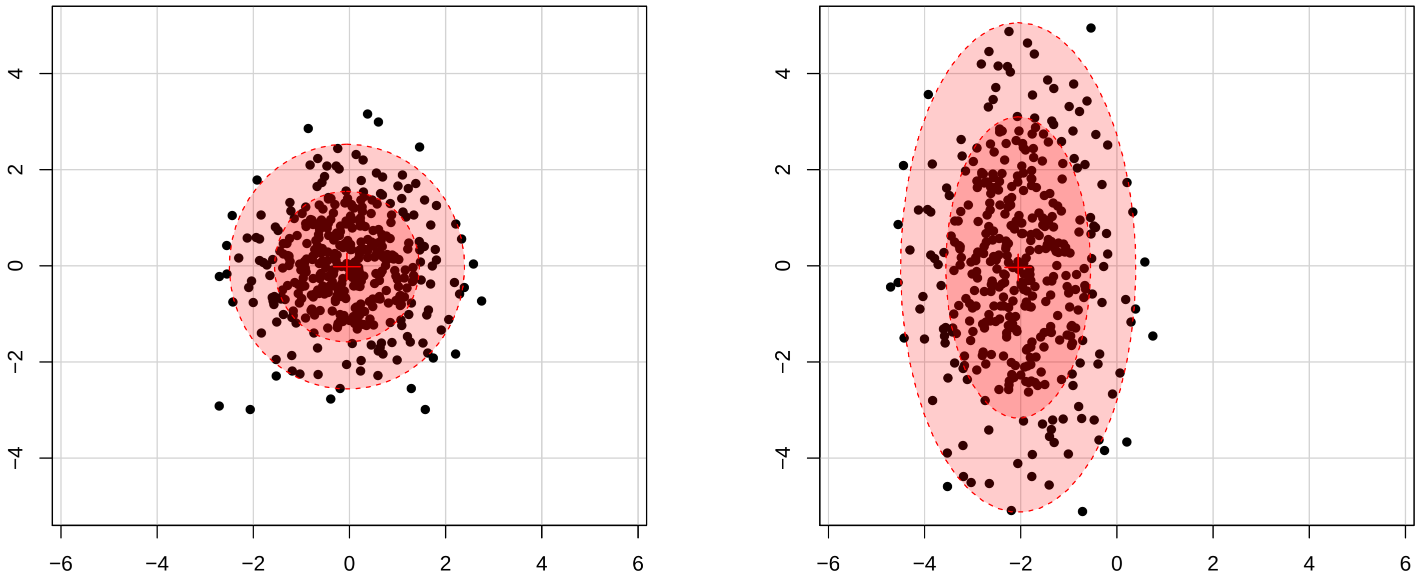
Definition. A random vector $X_1, \ldots, X_n$ is a standard normal random vector if all of its components are independent and and $X_i \sim \mathcal{N}(0, 1)$.

▷ The density can be computed based on independence:

$$p(x_1, \ldots, x_n) = p(x_1) \cdots p(x_n) = \frac{1}{(2\pi)^{n/2}} \cdot \mathrm{exp}\left(-\frac{x_1^2 + \cdots + x_n^2}{2}\right) .$$

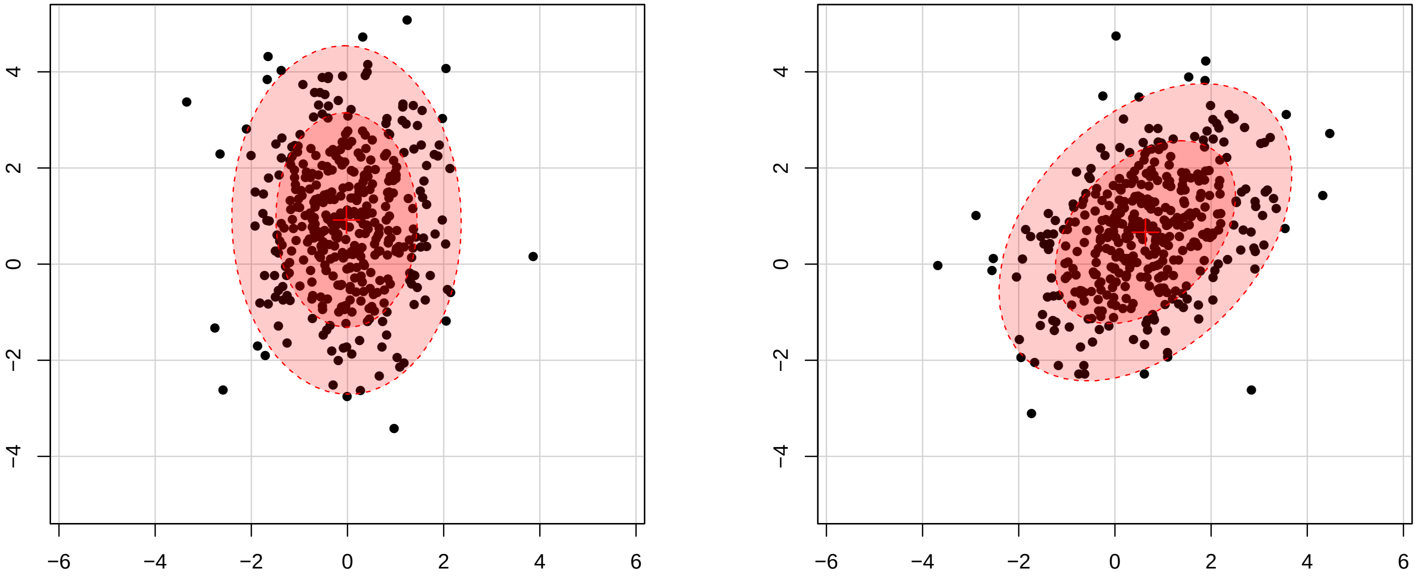
Scaling and shifting

By shifting and scaling the source distribution $\mathcal{N}(\mathbf{0}, I)$ we can obtain some other instances of multivariate normal distribution.



Necessity of rotations

As the choice of coordinate axis is sometimes arbitrary, there must be other ways to form a normal distribution – rotations of coordinate axis.



Any affine transformation can be expressed as scaling, rotating and shifting.

Affine transformations

Let $\boldsymbol{x}$ be standard normal random vector and let $\boldsymbol{y}$ be obtained the scaling, translation and rotation of the coordinate plane.

Then we can express $\boldsymbol{x}$ and $\boldsymbol{y}$ in terms of an affine transformation

$$\begin{aligned}\boldsymbol{y} &= A\boldsymbol{x} + \boldsymbol{\mu}\ \ ,\\ \boldsymbol{x} &= A^{-1}(\boldsymbol{y} - \boldsymbol{\mu})\ \ .\end{aligned}$$

Observation. Affine transformations are closed with respect to composition, i.e., applying two affine transformations yields a new affine transformation.

Remark. Not all affine transformations are invertible.

What is density in 2D?

Recall that density assigns probability to small enough regions $\mathcal{R}$:

$$\Pr\begin{bmatrix} x_1^* \leftarrow \mathcal{N}(0,1) : x_1 \leq x_1^* \leq x_1 + \Delta x_1 \\ x_2^* \leftarrow \mathcal{N}(0,1) : x_2 \leq x_2^* \leq x_2 + \Delta x_2 \end{bmatrix} = p(x_1, x_2) \cdot \underbrace{\Delta x_1 \Delta x_2}_{S} + \varepsilon$$

where $\varepsilon = o(\Delta x_1 \cdot \Delta x_2)$ in the process $\Delta x_1 \to 0$ and $\Delta x_2 \to 0$.

Remark. Regions $\mathcal{R}$ do not have to be rectangular as long as:

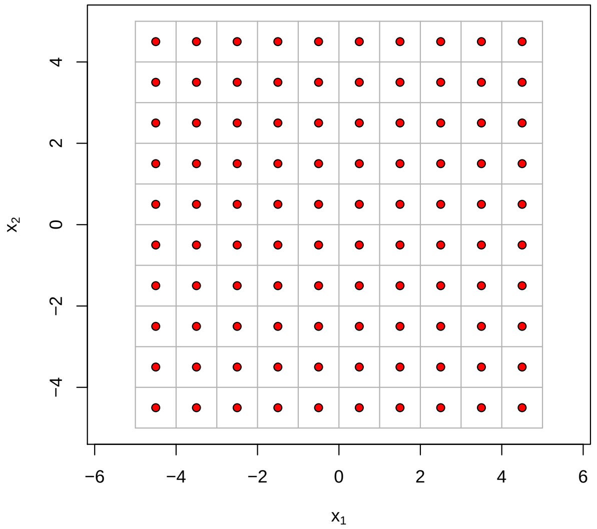
▷ The area $S(\mathcal{R})$ of a region can be computed.

▷ Probability can be assigned to the region $\mathcal{R}$ and its scalings.

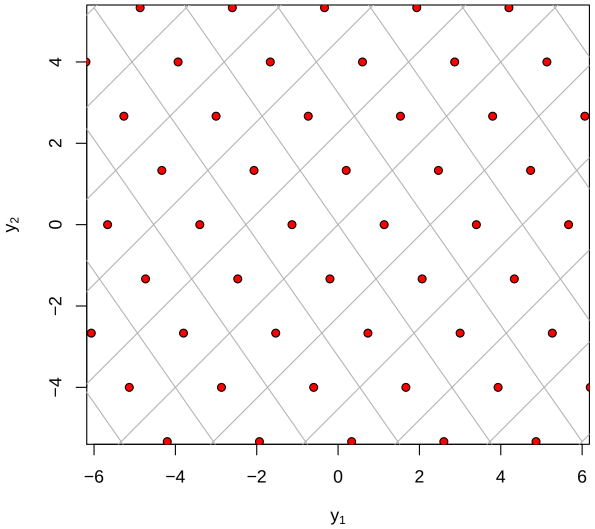
Then $\varepsilon = o(S)$ when we rescale the region $\mathcal{R}$ around the point (x_1, x_2).

Density recalibration

Any affine transformation changes a square grid into parallelograms.



$$\xrightarrow{\boldsymbol{y}=A\boldsymbol{x}+\boldsymbol{\mu}}$$



As a result, the area of the regions is different on the left and on the right:

$$p(x_1, x_2) \cdot S_1 \approx q(y_1, y_2) \cdot S_2 \qquad \Longrightarrow \qquad q(y_1, y_2) = \frac{S_1}{S_2} \cdot p(x_1, x_2)$$

Fortunately, the ratio between areas are constant over the entire plane!

Density of two-variate normal distribution

The density of (x_1, x_2) pairs can be computed based on independence:

$$p(x_1, x_2) = p(x_1) \cdot p(x_2) = \frac{1}{2\pi} \cdot \mathsf{exp}\left(-\frac{x_1^2 + x_2^2}{2}\right) \enspace .$$

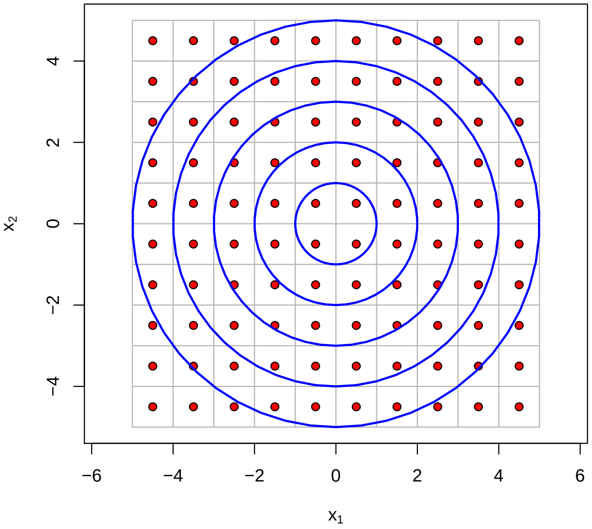
To estimate density $q(y_1, y_2)$, we must find the corresponding (x_1, x_2):

$$\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{\mu} \quad \Leftrightarrow \quad \boldsymbol{x} = A^{-1}(\boldsymbol{y} - \boldsymbol{\mu}) \enspace .$$

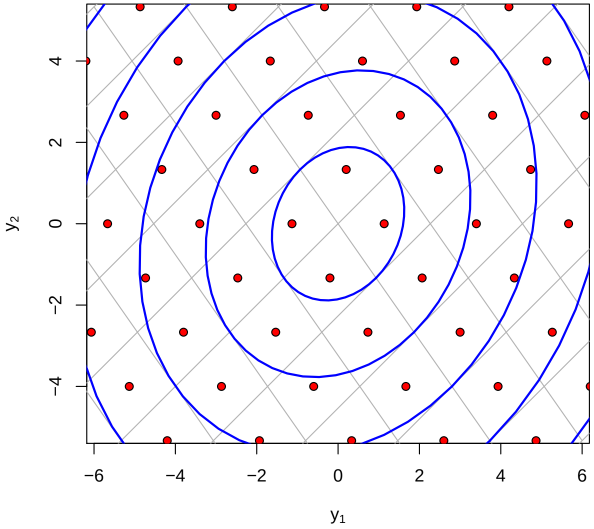
Thus we get

$$\begin{aligned} q(y_1, y_2) &= \frac{S_1}{S_2} \cdot \frac{1}{2\pi} \cdot \exp\left(-\frac{(\boldsymbol{y} - \boldsymbol{\mu})^T A^{-T} A^{-1} (\boldsymbol{y} - \boldsymbol{\mu})}{2}\right) \\ &= \frac{1}{\sqrt{\det(\Sigma)}} \cdot \frac{1}{2\pi} \cdot \exp\left(-\frac{(\boldsymbol{y} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{y} - \boldsymbol{\mu})}{2}\right) \enspace . \end{aligned}$$

Illustrative example



$$\xrightarrow{\boldsymbol{y}=A\boldsymbol{x}+\boldsymbol{\mu}}$$



- ▷ Affine transformation changes the square grid into parallelograms.
- ▷ Affine transformation changes circular equiprobability lines into ellipses.
- ▷ The axes of the ellipses may intersect with the sides of parallelograms.

Generalisation to multivariate case

If observed quantities $\boldsymbol{y}$ are generated by applying the affine transformation

$$\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{\mu} \quad \Leftrightarrow \quad \boldsymbol{x} = A^{-1}(\boldsymbol{y} - \boldsymbol{\mu})$$

to the *independent source signals* $x_1, \ldots, x_n \sim \mathcal{N}(0, 1)$, then the resulting distribution is *a multivariate normal distribution* with the density:

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

where $\Sigma^{-1} = A^{-T} A^{-1}$ is *a positively definite symmetric matrix*.

Important properties of normal distributions

Closeness under marginalisation

Let $\boldsymbol{x}_\mathcal{I} = (x_i)_{i\in\mathcal{I}}$ be a subvector determined by the coordinate set $\mathcal{I}$. Then $\boldsymbol{x}_\mathcal{I}$ is distributed according to a multivariate normal distribution as long as the vector $\boldsymbol{x}$ comes form a multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$.

▷ Moment matching gives the parameters of the resulting distribution

$$\begin{aligned}\mathbf{E}(\boldsymbol{x}_\mathcal{I}) &= \mathbf{E}(\boldsymbol{x})_\mathcal{I} = \boldsymbol{\mu}_\mathcal{I}\\ \mathbf{Cov}(\boldsymbol{x}_\mathcal{I}) &= \mathbf{Cov}(\boldsymbol{x})_{\mathcal{I}\times\mathcal{I}} = \Sigma[\mathcal{I},\mathcal{I}]\end{aligned}$$

Closeness under linear combinations

Linear combination $y = \boldsymbol{\alpha}_1^T \boldsymbol{x}_1 + \boldsymbol{\alpha}_2^T \boldsymbol{x}_2$ of independent multivariate normal distributions $\boldsymbol{x}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1)$ and $\boldsymbol{x}_2 \sim \mathcal{N}(\boldsymbol{\mu}_2, \Sigma_2)$ is also a multivariate normal distribution.

▷ Moment matching gives the parameters of the resulting distribution

$$\begin{aligned}\mathbf{E}(y) &= \boldsymbol{\alpha}_1^T\,\mathbf{E}(\boldsymbol{x}_1) + \boldsymbol{\alpha}_2^T\,\mathbf{E}(\boldsymbol{x}_2) = \boldsymbol{\alpha}_1^T\boldsymbol{\mu}_1 + \boldsymbol{\alpha}_2^T\boldsymbol{\mu}_2\\ \mathbf{Var}(y) &= \mathbf{Cov}(\boldsymbol{\alpha}_1^T\boldsymbol{x}_1) + \mathbf{Cov}(\boldsymbol{\alpha}_2^T\boldsymbol{x}_2)\\ &= \boldsymbol{\alpha}_1^T\mathbf{Cov}(\boldsymbol{x}_1)\boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2^T\mathbf{Cov}(\boldsymbol{x}_2)\boldsymbol{\alpha}_2\\ &= \boldsymbol{\alpha}_1^T\Sigma_1\boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2^T\Sigma_2\boldsymbol{\alpha}_2\end{aligned}$$

▷ Closeness under linear combinations holds also for matrix combinations.

Closeness under conditioning

Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be related random variables. Let $\boldsymbol{x}|\boldsymbol{y}_*$ denote the conditional distribution of $\boldsymbol{x}$ given that a random variable $\boldsymbol{y}$ has a fixed value $\boldsymbol{y}_*$. Then $\boldsymbol{x}|\boldsymbol{y}_*$ is distributed according to a multivariate normal distribution provided that $(\boldsymbol{x}, \boldsymbol{y})$ comes form a multivariate normal distribution $\mathcal{N}((\boldsymbol{\mu}_i), (\Sigma_{ij}))$

▷ Moment matching gives the parameters of the resulting distribution

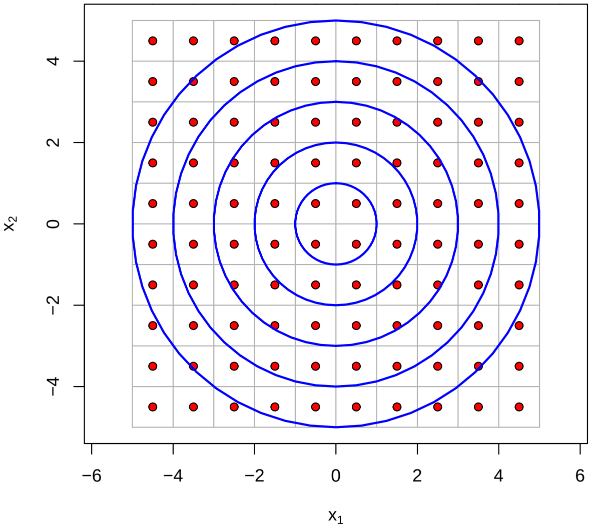
$$\begin{aligned}\mathbf{E}(\boldsymbol{x}|\boldsymbol{y}_*) &= \boldsymbol{\mu}_1 + \Sigma_{1,2}\Sigma_{2,2}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_2)\\ \mathbf{Cov}(\boldsymbol{x}|\boldsymbol{y}_*) &= \Sigma_{1,1} - \Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1}\end{aligned}$$

Principal component analysis

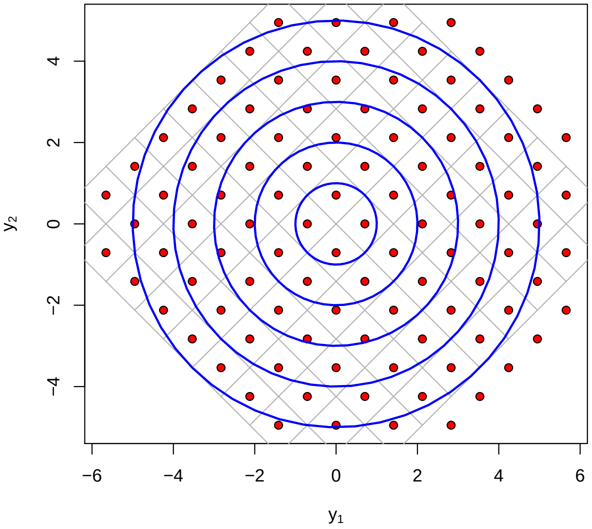
Distribution reconstruction task

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.

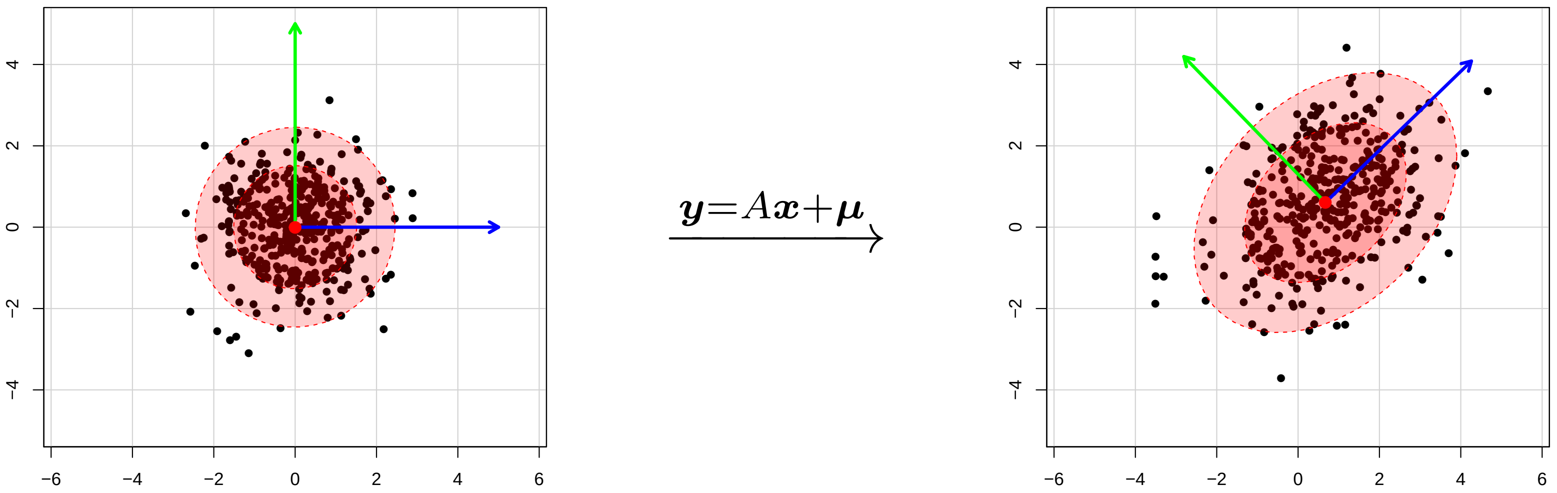


$$\xrightarrow{\boldsymbol{y}=A\boldsymbol{x}+\boldsymbol{\mu}}$$



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 $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m$:

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

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

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

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

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

- ▷ 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

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

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:

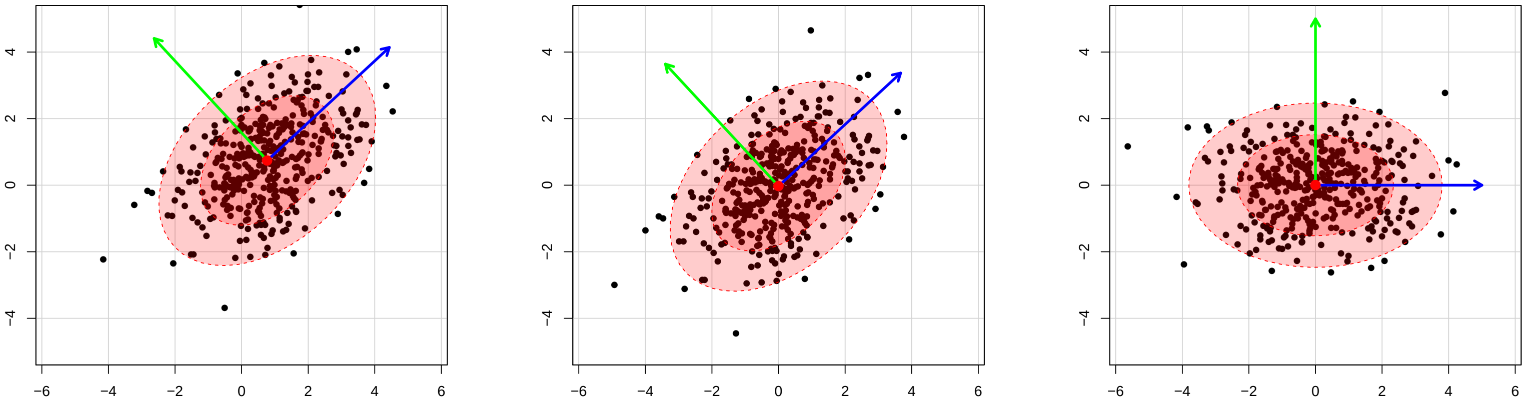
$$\Sigma\boldsymbol{w} = \lambda\boldsymbol{w} \qquad \Rightarrow \qquad \boldsymbol{w}^T\Sigma\boldsymbol{w} = \boldsymbol{w}^T\lambda\boldsymbol{w} = \lambda\ .$$

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):

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

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

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

The covariance matrix

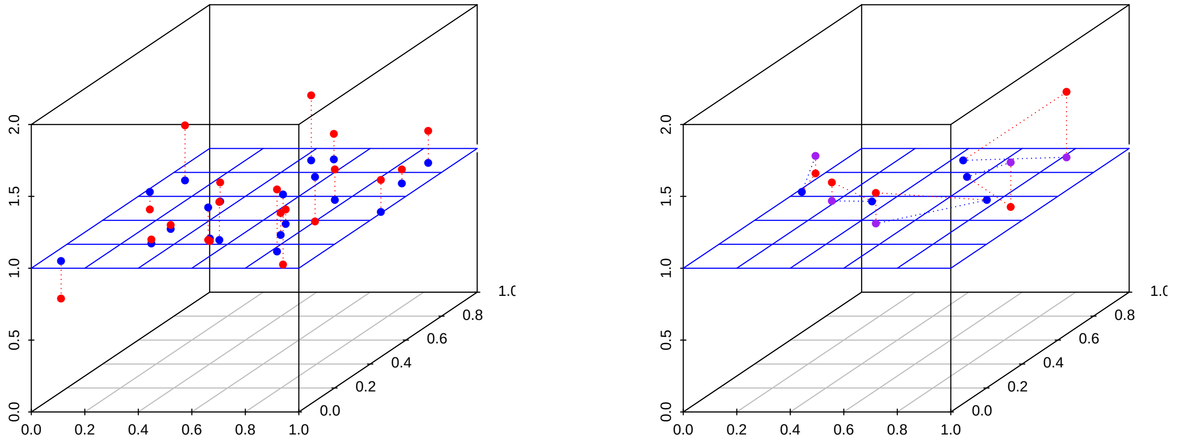
$$\Sigma = \frac{1}{m} \cdot \sum_{i=1}^{m} (\boldsymbol{y}_i - \boldsymbol{\mu})^T (\boldsymbol{y}_i - \boldsymbol{\mu})$$

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

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

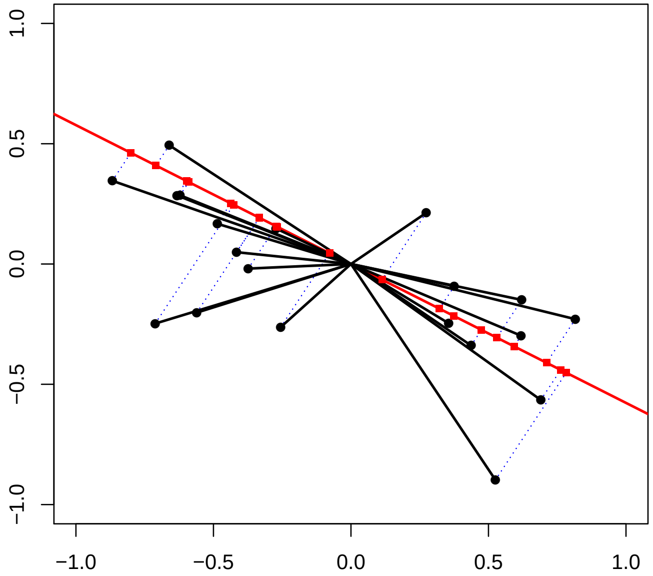
where d_i is the distance between the plane $\mathcal{H}$ and the point $\boldsymbol{y}_i$. Thus

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

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

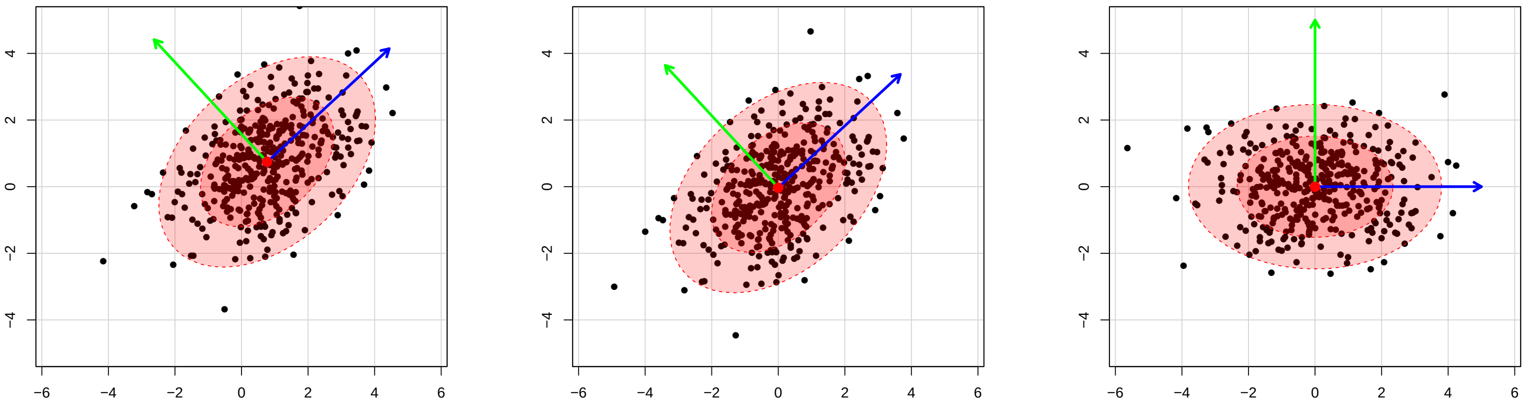
Fact. If the data is centred then PCA chooses the direction $\boldsymbol{w}_1$ such that the sum of squares of the projections $\boldsymbol{w}_1^T \boldsymbol{y}_i$ is maximal.



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.

Going beyond PCA

Weighted Principal Component Analysis:

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

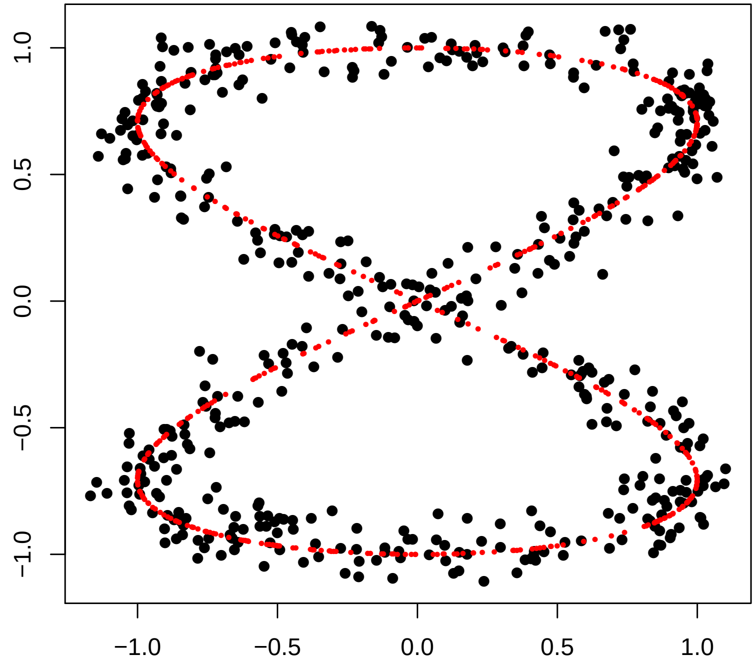
Principal curves and manifolds

- ▷ 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



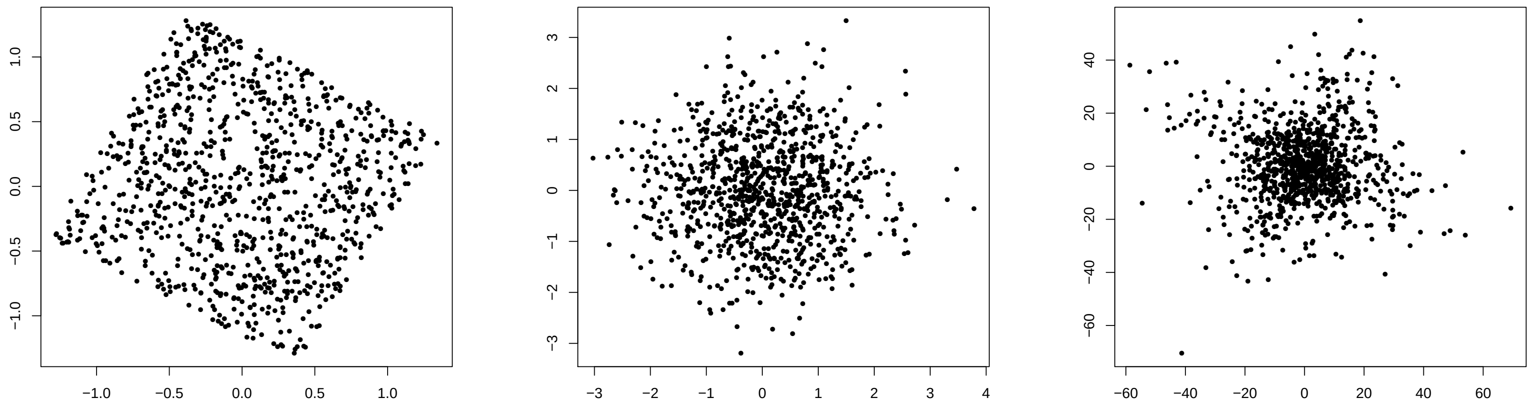
Reconstruction of the underlying curve is much more difficult.

▷ 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.