LTAT.02.004 Machine Learning II

Multivariate normal distribution

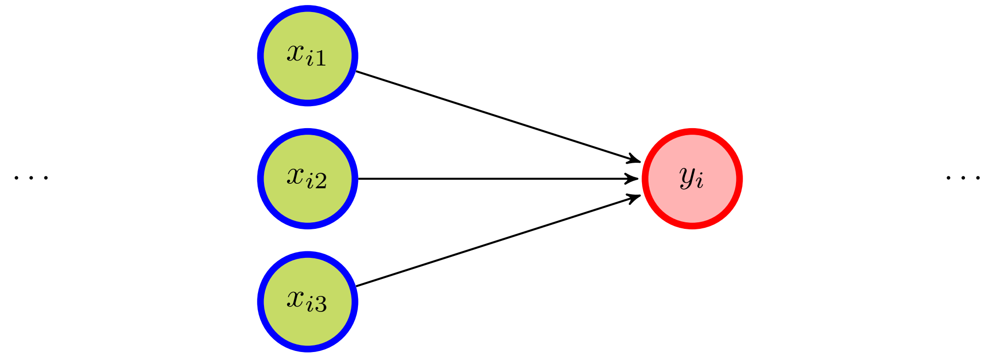
Direct applications

Sven Laur
University of Tartu

Motivating examples

Supervised learning

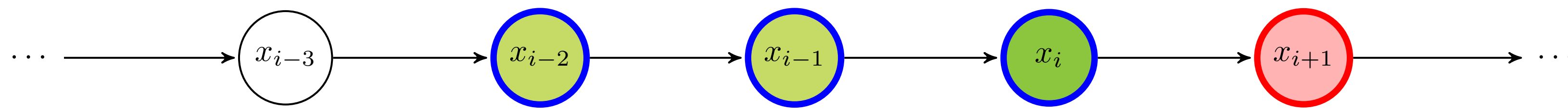
Repeated controlled coin tosses



Linear regression models

- ▷ We assume that y_i depends only on the values of $x_{i1}, \ldots, x_{i\ell}$
- ▷ A linear model assumes $y_i = \beta_1 x_{i1} + \cdots + \beta_\ell x_{i\ell} + \beta_0 + \varepsilon_i$.
- ▷ All error terms ε_i are assumed to be independent.
- ▷ All error terms ε_i are drawn from a normal distribution $\mathcal{N}(0, \sigma)$.

Higher-order Markov chains

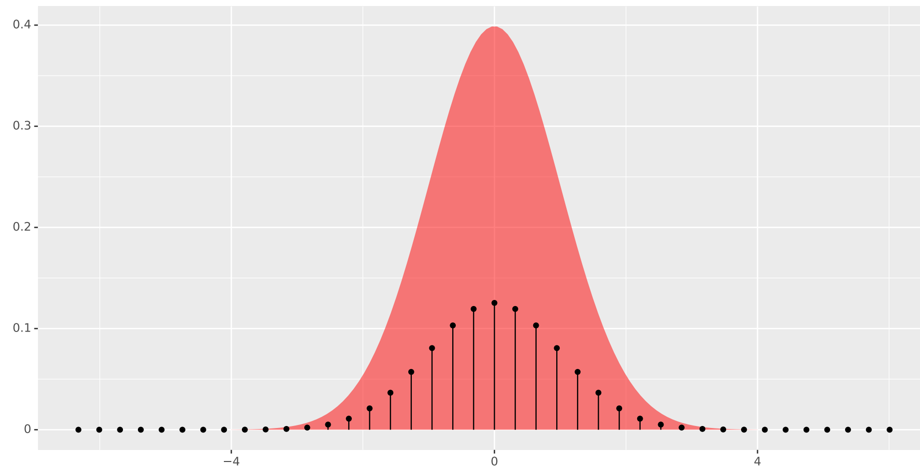


Time-series models

- We assume that x_{i+1} depends only on the values of $x_i, \ldots, x_{i-\ell}$
- A linear model assumes $x_{i+1} = \alpha_0 x_i + \cdots + \alpha_\ell x_{i-\ell} + \alpha_{\ell+1} + \varepsilon_i$.
- All error terms ε_i are assumed to be independent.
- All error terms ε_i are drawn from a normal distribution $\mathcal{N}(0, \sigma)$.

Univariate normal distribution

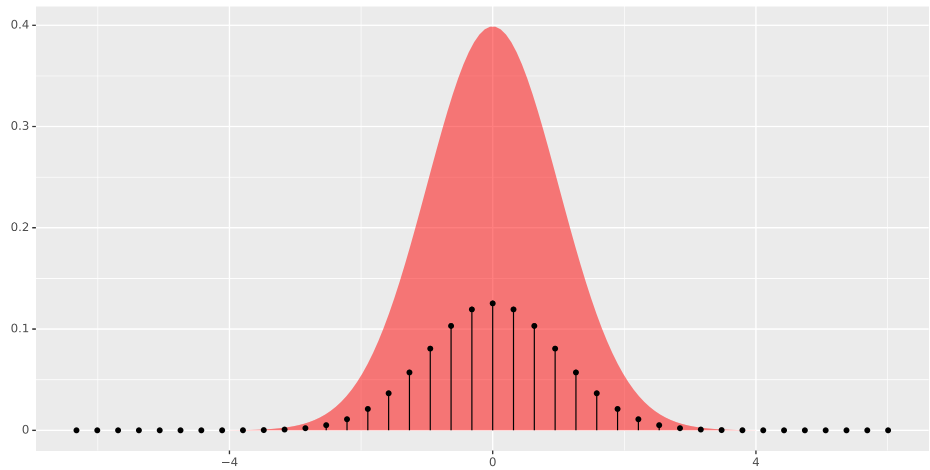
Probability density function



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 $x \in \mathbb{R}$:

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

Probability mass function

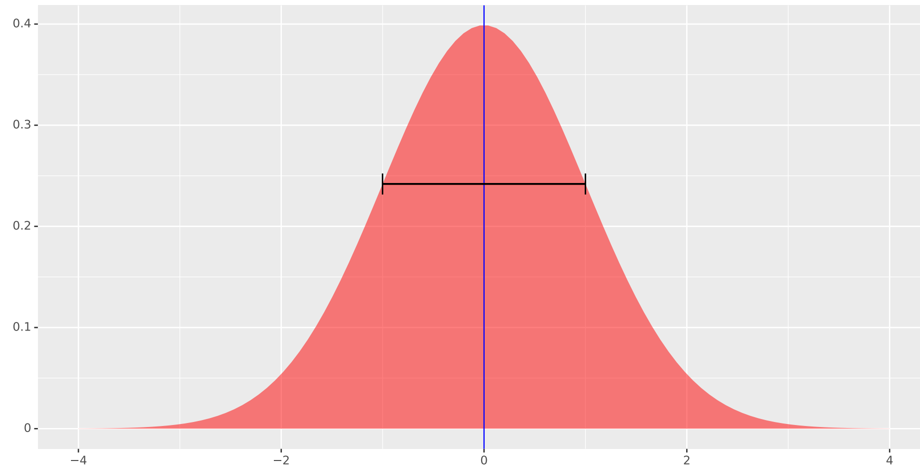


Definition. A real-valued random variable X comes from a discrete distribution with *a probability mass function* $p : \mathbb{R} \to \mathbb{R}^+ \cup \{0\}$ defined as

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

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

$$y = bx + a$$

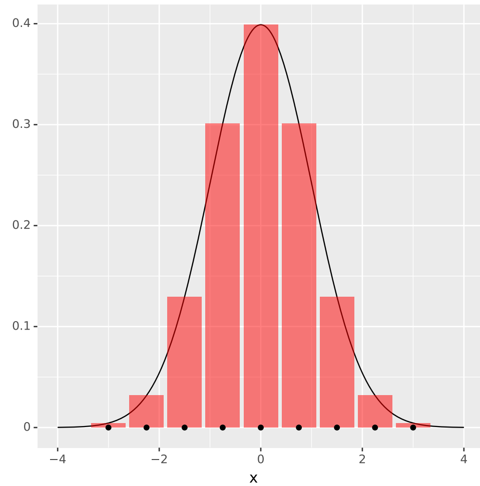
where x is distributed according to standardised normal distribution $\mathcal{N}(0, 1)$.

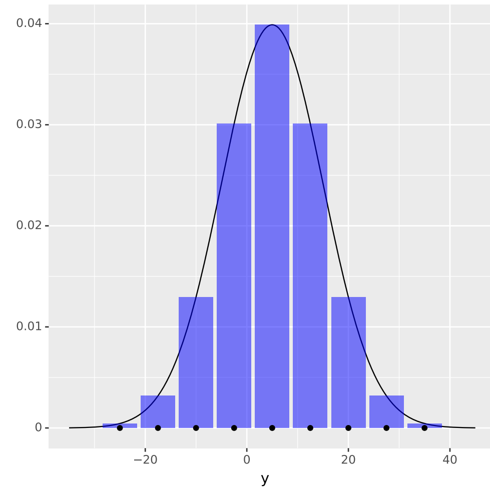
The corresponding probability density functions is

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

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.

Univariate linear regression

- Fix a set of inputs $x_1, \ldots, x_n \in \mathbb{R}$.
- A probabilistic model is defined by three coefficients $a, b, \in \mathbb{R}$.
- The model assigns a probability to outcomes $y_1, \ldots, y_n$ through the following observation generation mechanism

$$y_i = ax_i + b + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma)$$

- Consequently

$$p[\boldsymbol{y}|\boldsymbol{x}, a, b] = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left(-\frac{(y_i - ax_i - b)^2}{2\sigma^2}\right)$$

Maximum likelihood estimate

As usual we can find $a, b, \sigma \in \mathbb{R}$ that maximise the log-likelihood

$$\log p[\boldsymbol{y}|\boldsymbol{x}, a, b, \sigma] = const - n \log \sigma - \sum_{i=1}^{n} \frac{(y_i - ax_i - b)^2}{2\sigma^2}$$

and thus we can find a and b by minimising

$$\mathsf{MSE} = \frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - ax_i - b)^2 \enspace .$$

Residuals and the variance parameter

For fixed $a, b \in \mathbb{R}$ we can define predictions and residuals

$$\hat{y}_i = ax_i - b$$

$$r_i = y_i - \hat{y}_i$$

To find the optimal variance σ^2 we need to maximise

$$\log p[\boldsymbol{y}|\boldsymbol{x}, a, b, \sigma] = const - n\log\sigma - \sum_{i=1}^{n}\frac{r_i^2}{2\sigma^2}$$

The resulting solution is

$$\sigma^2 = \frac{1}{n}\cdot\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Linear time-series model

- ▷ Fix a set of initial inputs $x_{-\ell}, \ldots, x_0 \in \mathbb{R}$. Denote them by $\boldsymbol{x}_\circ$.
- ▷ Think of $x_1, x_2, \ldots, x_n$ as observations. Denote them by $\boldsymbol{x}$.
- ▷ A probabilistic model for state transitions is defined as follows

$$x_{i+1} = \underbrace{\alpha_0 x_i + \ldots \alpha_\ell x_{i-\ell} + \alpha_{\ell+1}}_{\hat{x}_{i+1}} + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma)$$

- ▷ Consequently

$$p[\boldsymbol{x}|\boldsymbol{x}_\circ, \boldsymbol{\beta}, \sigma] = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left(-\frac{(x_i - \hat{x}_i)^2}{2\sigma^2}\right)$$

Maximum likelihood estimate

As usual we can find $\boldsymbol{\alpha} \in \mathbb{R}^{\ell+1}$ and $\sigma \in \mathbb{R}$ that maximise the log-likelihood

$$\log p[\boldsymbol{x}|\boldsymbol{x}_\circ, \boldsymbol{\beta}, \sigma] = const - n \log \sigma - \sum_{i=1}^{n} \frac{(x_i - \hat{x}_i)^2}{2\sigma^2}$$

and thus we can find $\boldsymbol{\beta}$ by minimising

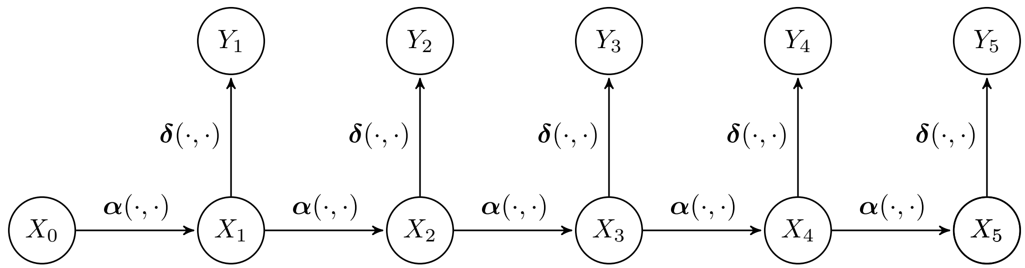
$$\mathsf{MSE} = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \alpha_0 x_{i-1} + \ldots \alpha_\ell x_{i-1-\ell} + \alpha_{\ell+1})^2 \ .$$

The latter is the standard multivariate linear regression setup. The variance of the model σ^2 can be found by the same formula as for linear regression.

Motivating examples

Unsupervised learning

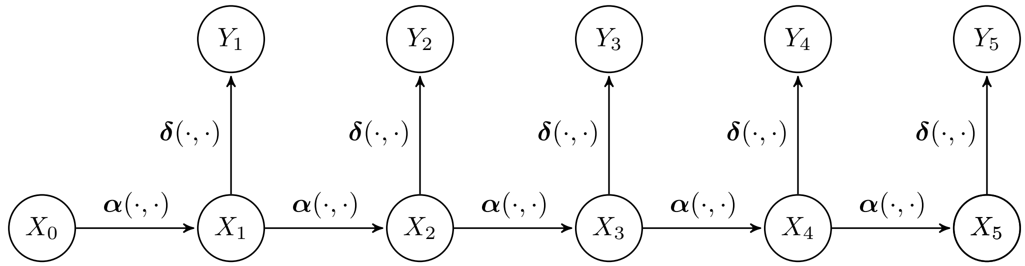
Hidden Markov Model



Sensor fusion problem

- ▷ Several sensors measure a physical system
- ▷ Measurements are observable as $\boldsymbol{y} \in \mathbb{R}^p$.
- ▷ Physical system has an hidden state $\boldsymbol{x} \in \mathbb{R}^n$.
- ▷ Physical system evolves linearly $\boldsymbol{x}_{i+1} = A\boldsymbol{x}_i + \boldsymbol{w}_i$.
- ▷ Measurements are linear from the state $\boldsymbol{y}_i = C\boldsymbol{x}_i + \boldsymbol{v}_i$.
- ▷ Distribution of error terms $\boldsymbol{v}_i$ and $\boldsymbol{w}_i$ is known.
- ▷ Error terms $\boldsymbol{v}_i$ and $\boldsymbol{w}_i$ are independently drawn.

Belief propagation for continous distibutions



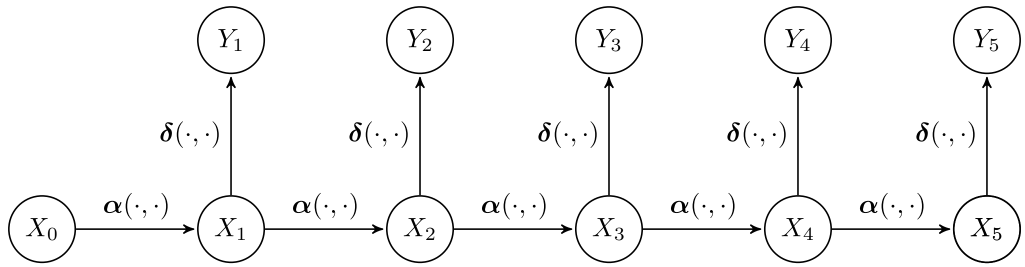
Continuous distributions are rarely compatible with belief propagation

$$\pi_{X_i}(\boldsymbol{x}_i) \propto \int_{\boldsymbol{x}_{i-1}} \alpha[\boldsymbol{x}_{i-1}, \boldsymbol{x}_i] \cdot \lambda^*_{i-1}(\boldsymbol{x}_{i-1}) \cdot \pi_{X_{i-1}}(\boldsymbol{x}_{i-1}) d\boldsymbol{x}_{i-1}$$

$$\lambda_{X_i}(x_i) \propto \int_{\boldsymbol{x}_{i+1}} \alpha[\boldsymbol{x}_i, \boldsymbol{x}_{i+1}] \cdot \lambda^*_i(\boldsymbol{x}_i) \cdot \lambda_{X_{i+1}}(\boldsymbol{x}_{i+1})\ d\boldsymbol{x}_{i+1}$$

but normal distributions form a rare exception.

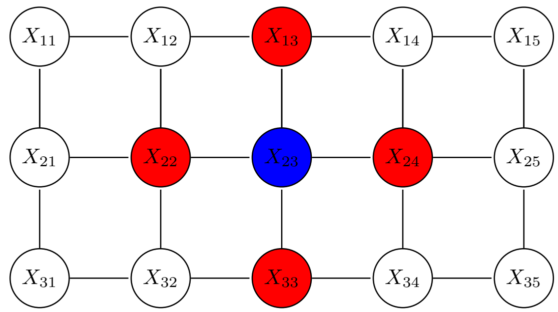
Kalman filter



Belief propagation becomes tractable under following assumptions:

- ▷ Measurement noise $\boldsymbol{v}_t$ is modelled with a normal distribution.
- ▷ Unknown control signal $\boldsymbol{w}_i$ is modelled with a normal distribution.
- ▷ Unknown initial state $\boldsymbol{x}_0$ is modelled with a normal distribution.
- ▷ Quantities $\boldsymbol{x}_0, \boldsymbol{v}_i, \boldsymbol{w}_i$ are assumed to be independent.
- ▷ All normal distributions can have complex correlation structure.

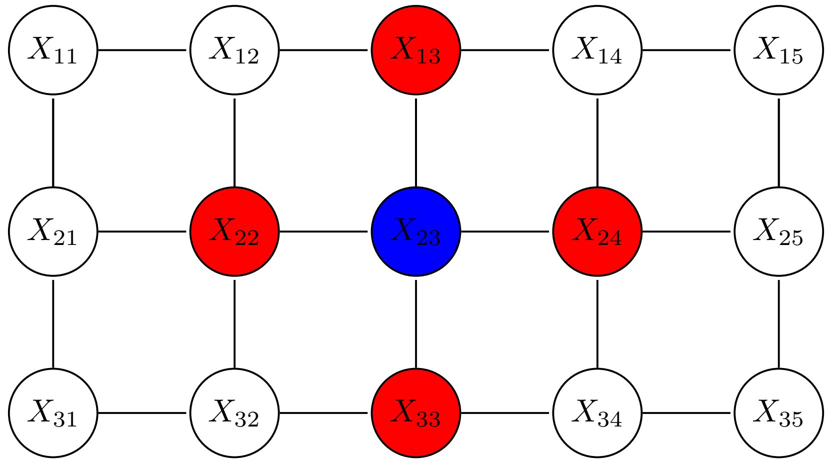
Background model for digital images



In most images intensity of pixel is influenced only by its neighbours:

- ▷ For simple textures the neighbourhood consist of four adjacent pixels.
- ▷ For complex textures the the neighbourhood contains much more pixels.
- ▷ For homogenous textures the conditional probabilities are universal.
  - ◇ Generative repetitive patterns for textile and grass
- ▷ For complex patterns conditional probabilities can be location dependent.
  - ◇ Generative patterns for human faces and fashion accessories

Random Markov Fields



Definition. Markov random field is specified by undirected graph connecting random variables $X_1, X_2, \ldots$ such that for any node X_i

$$\Pr\left[x_i | (x_j)_{j \neq i}\right] = \Pr\left[x_i | (x_j)_{j \in \mathcal{N}(X_i)}\right]$$

where the set of neighbours $\mathcal{N}(X_i)$ is also known as *Markov blanket* for X_i.

Hammersley-Clifford theorem

The probability of an observation $\boldsymbol{x} = (x_1, x_2, \ldots)$ generated by a Markov random field can be expressed in the form

$$\Pr[\boldsymbol{x}] = \frac{1}{Z(\omega)} \cdot \exp\left(-\sum_{c \in \mathsf{MaxClique}} \Psi_c(\boldsymbol{x}_c, \omega)\right)$$

where

▷ $Z(\omega)$ is a normalising constant

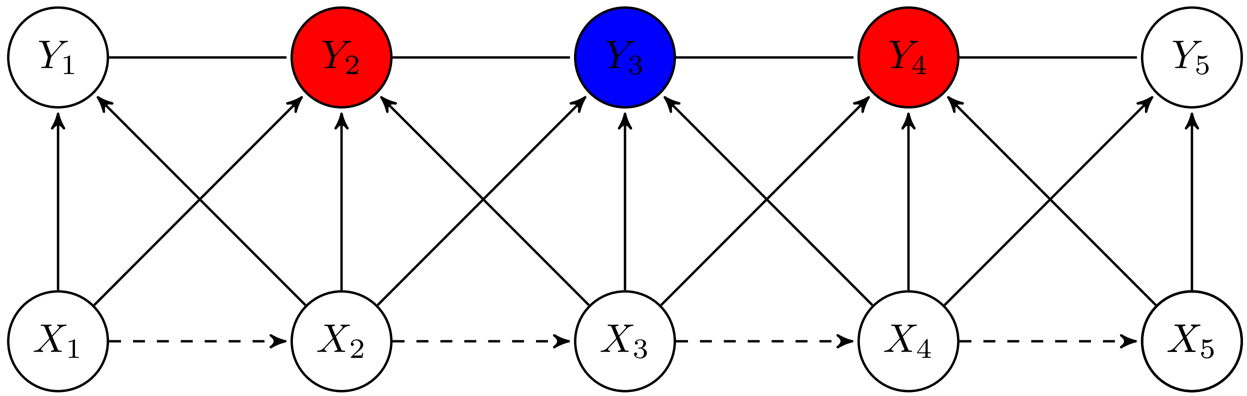
▷ MaxClique is the set of maximal cliques in the Markov random field

▷ Ψ_c is defined on the variables in the clique c

The formula implies that the distribution belongs to the exponential family.

▷ Multivariate normal distribution belongs to the exponential family

Conditional Random Fields

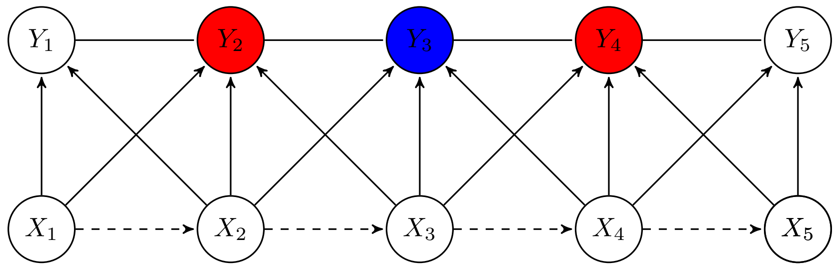


Definition. Let $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ be random variables. The entire process is conditional random field if random variables $Y_1, Y_2, \ldots$ conditioned for any sequence of observations $x_1, x_2, \ldots$ form a Markov random field

$$\Pr\left[y_i|(x_k)_{k=1}^{\infty}, (y_j)_{j\neq i}\right] = \Pr\left[y_i|(x_k)_{k=1}^{\infty}, (y_j)_{j\in\mathcal{N}(Y_i)}\right]$$

where the set of neighbours $\mathcal{N}(Y_i)$ is a *conditional Markov blanket* for Y_i.

Image segmentation and sequence labelling



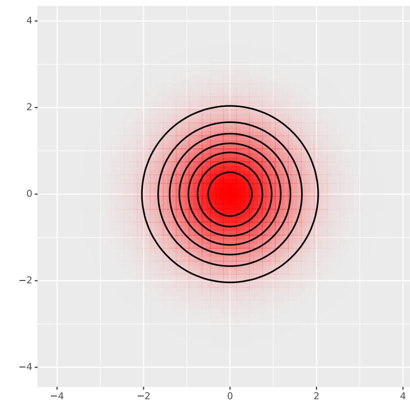
- ▷ The input $\boldsymbol{x}$ is used to predict labels $y_1, y_2, \ldots$.
- ▷ A correct label sequence must satisfy possibly unknown restrictions.
- ▷ These restrictions are captured by conditional random random field.

Consequences of Hammersley-Clifford theorem

- ▷ Clique features Ψ_c can depend on $(y_i)_{i\in c}$, $(x_i)_{i=1}^{\infty}$
- ▷ Features can be defined as linear combination of vertex and edge features.
- ▷ A vertex feature looks only variable y_i associated with the vertex.
- ▷ An edge feature looks only variables y_i, y_j associated with the edge.

Multivariate normal distribution

White Gaussian noise



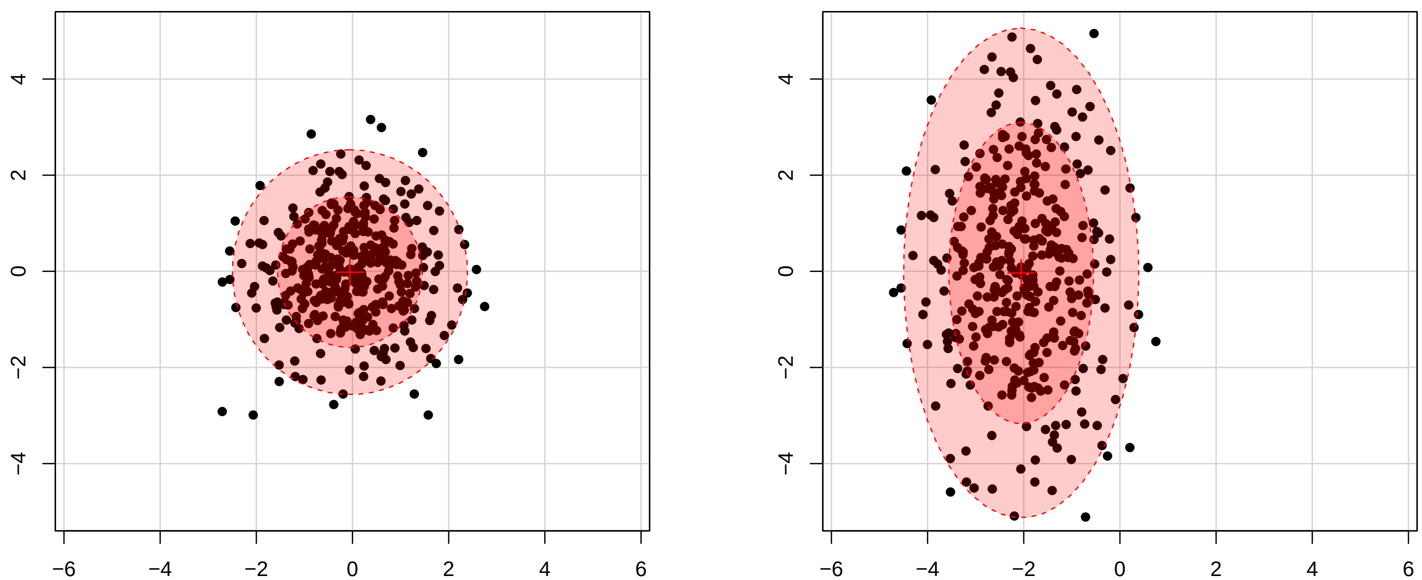
Definition. A random vector $X_1, \ldots, X_n$ is a standard normal random vector if all of its components are independent and and $X_i \sim \mathcal{N}(0, 1)$.

▷ The density can be computed based on independence:

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

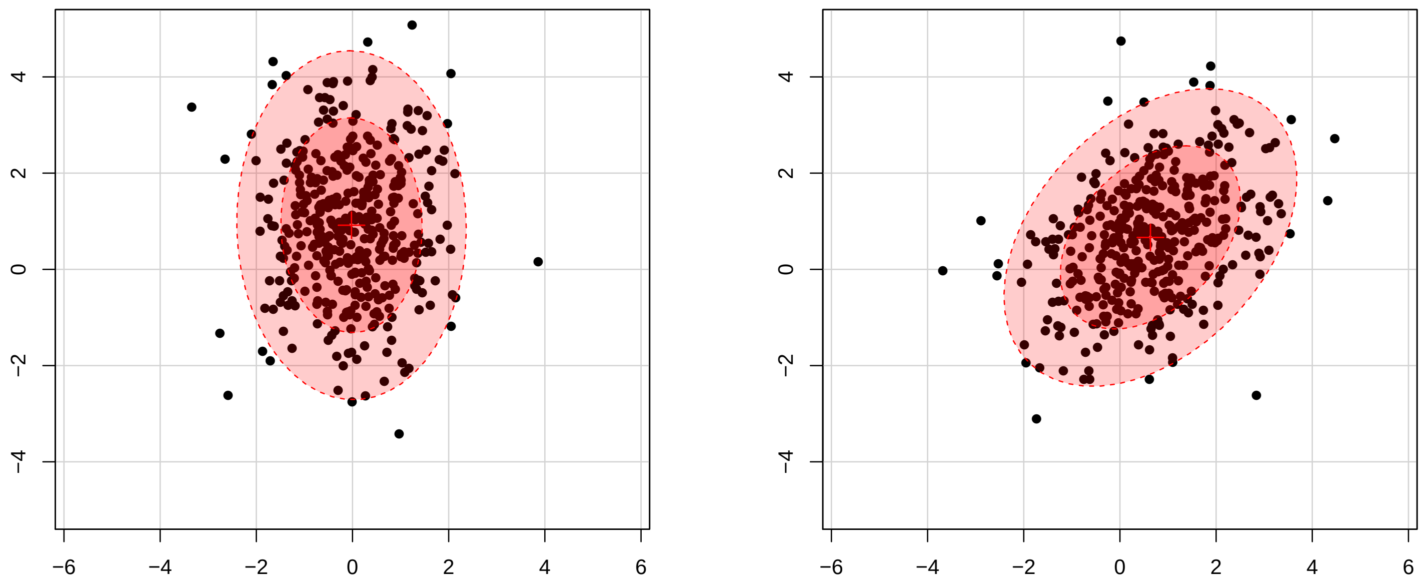
Scaling and shifting

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

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

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 $\mathcal{R}$:

$$\Pr_{x_1^*, x_2^*}\begin{bmatrix} x_1 \leq x_1^* \leq x_1 + \Delta x_1 \\ x_2 \leq x_2^* \leq x_2 + \Delta x_2 \end{bmatrix} = p(x_1, x_2) \cdot \underbrace{\Delta x_1 \Delta x_2}_{S} + \varepsilon$$

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

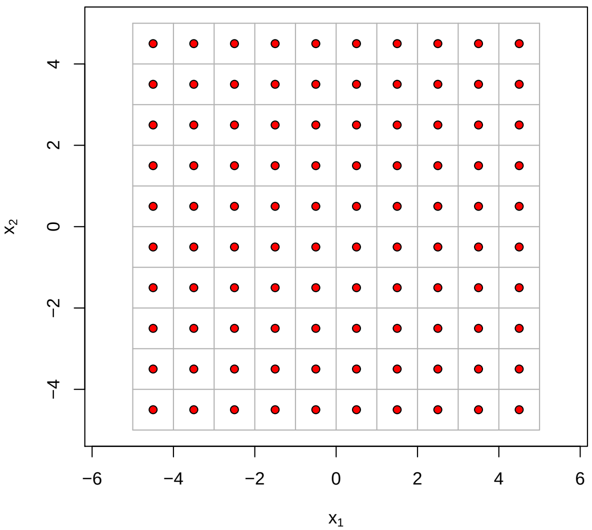
Remark. Regions $\mathcal{R}$ do not have to be rectangular as long as:

- ▷ The area $S(\mathcal{R})$ of a region can be computed.
- ▷ Probability can be assigned to the region $\mathcal{R}$ and its scalings.

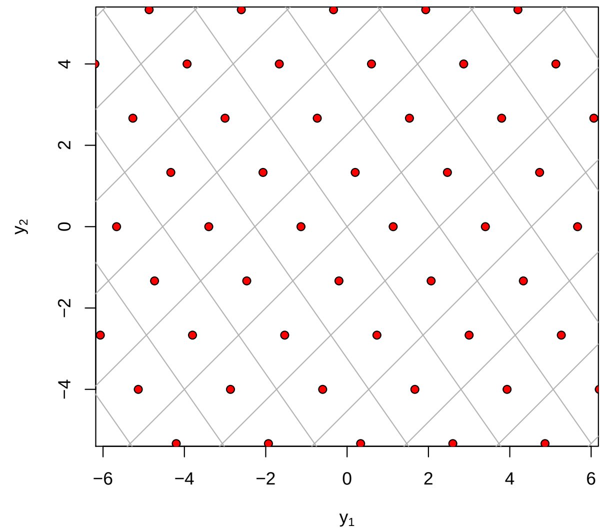
Then $\varepsilon = o(S)$ when we rescale the region $\mathcal{R}$ around the point (x_1, x_2).

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:

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

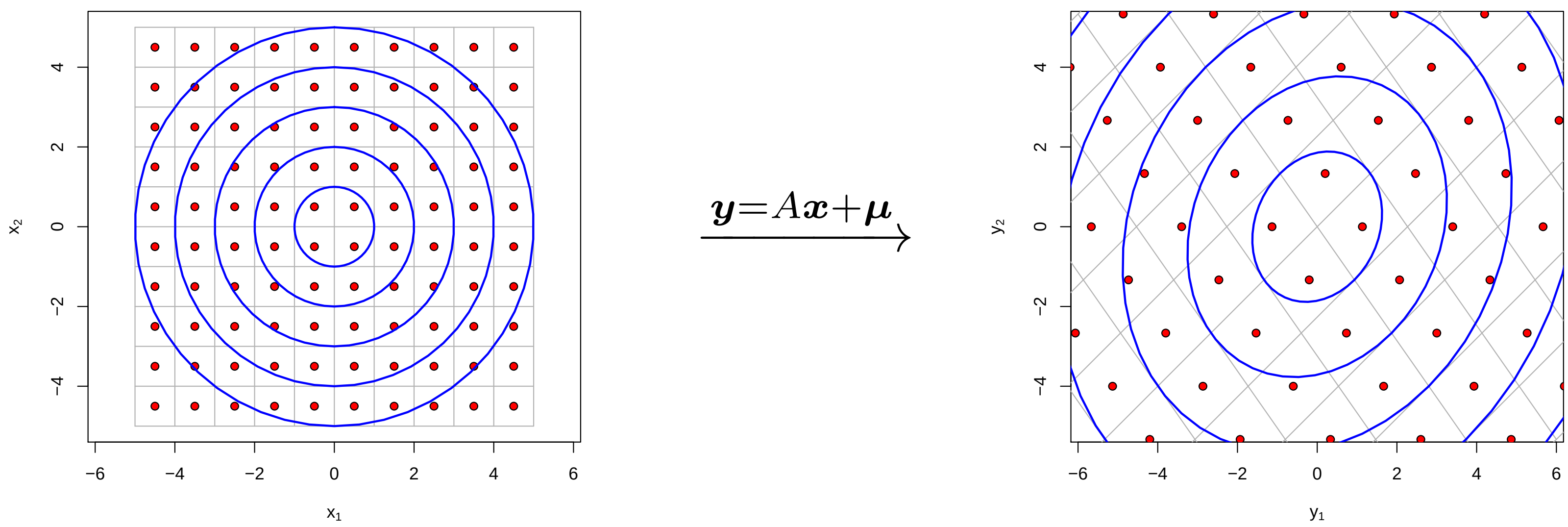
To estimate density $q(y_1, y_2)$, we must find the corresponding (x_1, x_2):

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

Thus we get

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

Illustrative example



▷ 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*.

Markov fields
with
multivariate normal distributions

General form of the likelihood function

The celebrated Hammersley-Clifford theorem fixes the format in which the corresponding probability distribution must be sought:

$$p[\boldsymbol{x}|\omega] = \frac{1}{Z(\omega)} \cdot \exp\left(-\sum_{c \in \mathsf{MaxClique}} \Psi_c(\boldsymbol{x}_c, \omega)\right)$$

where

- ω is a set of model parameters
- $Z(\omega)$ is a normalising constant
- MaxClique is the set of maximal cliques in the Markov random field
- Ψ_c is defined on the variables x_i in the clique c.

Multivariate normal distribution as likelihood

If individual sub-potentials $\Psi_c(\boldsymbol{x}_c, \omega)$ are quadratic forms then the energy

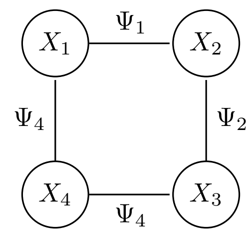
$$\Psi(\boldsymbol{x}) = \sum_{c \in \mathsf{MaxClique}} \Psi_c(\boldsymbol{x}_c, \omega)$$

is also a quadratic form and thus $p[\boldsymbol{x}|\omega]$ is a multivariate normal distribution.

Sub-potentials are often fixed directly based on smoothness constraints

- ▷ Intensities have bounded variance: $\Psi_e = \delta^2 x_{ij}^2$.
- ▷ Intensity changes smoothly vertically: $\Psi_e = \beta(x_{i,j} - x_{i+1,j})^2$.
- ▷ Intensity changes smoothly horizontally: $\Psi_e = \alpha(x_{i,j} - x_{i,j+1})^2$.

Toy example



Sub-potentials corresponding four edges are:

$$\Psi_1(x_1, x_2) = \alpha_1(x_1 - x_2)^2 = \alpha_1 x_1^2 - 2\alpha_1 x_1 x_2 + \alpha_1 x_2^2$$
$$\Psi_2(x_2, x_3) = \alpha_2(x_2 - x_3)^2 = \alpha_2 x_2^2 - 2\alpha_2 x_2 x_3 + \alpha_2 x_3^2$$
$$\Psi_3(x_3, x_4) = \alpha_3(x_3 - x_4)^2 = \alpha_3 x_3^2 - 2\alpha_3 x_3 x_4 + \alpha_3 x_4^2$$
$$\Psi_4(x_4, x_1) = \alpha_4(x_4 - x_1)^2 = \alpha_4 x_4^2 - 2\alpha_4 x_4 x_1 + \alpha_4 x_1^2$$

Sub-potentials corresponding to four vertices are $\Psi_i^*(x_i) = \delta_i^2 x_i^2$

Resulting potential function

$$\Psi(\boldsymbol{x}) = \boldsymbol{x}^T \begin{pmatrix} \alpha_1 + \alpha_4 + \delta_1^2 & -\alpha_1 & 0 & -\alpha_4 \\ -\alpha_1 & \alpha_1 + \alpha_2 + \delta_2^2 & -\alpha_2 & 0 \\ 0 & -\alpha_2 & \alpha_2 + \alpha_3 + \delta_3^2 & -\alpha_3 \\ -\alpha_4 & 0 & -\alpha_3 & \alpha_3 + \alpha_4 + \delta_4^2 \end{pmatrix} \boldsymbol{x}$$

and thus the covariance matrix Σ and mean $\boldsymbol{\mu}$ can be computed by matching the shape of the multivariate normal density

$$p[\boldsymbol{x}|\boldsymbol{\mu}, \Sigma] \propto \frac{1}{\sqrt{\det \Sigma}} \cdot \exp\left(-\frac{1}{2} \cdot (\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right)$$

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

$$\mathbf{E}(\boldsymbol{x}_\mathcal{I}) = \mathbf{E}(\boldsymbol{x})_\mathcal{I} = \boldsymbol{\mu}_\mathcal{I}$$

$$\mathbf{Cov}(\boldsymbol{x}_\mathcal{I}) = \mathbf{Cov}(\boldsymbol{x})_{\mathcal{I}\times\mathcal{I}} = \Sigma[\mathcal{I}, \mathcal{I}]$$

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

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

Kalman filter

To be completed next year