

LTAT.02.004 MACHINE LEARNING II

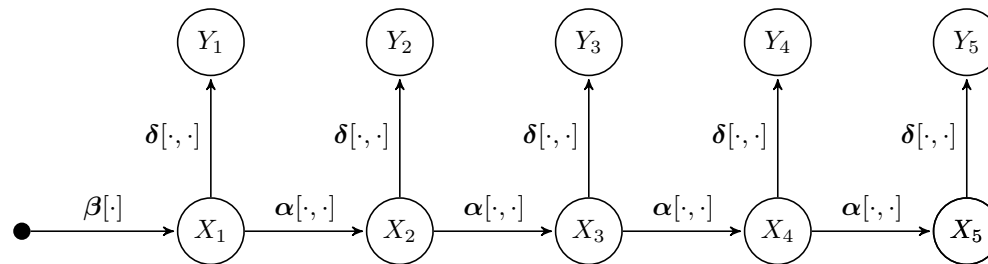
## **Multivariate normal distribution**

### **Direct applications**

Sven Laur  
University of Tartu

Motivating examples

# Hidden Markov Model



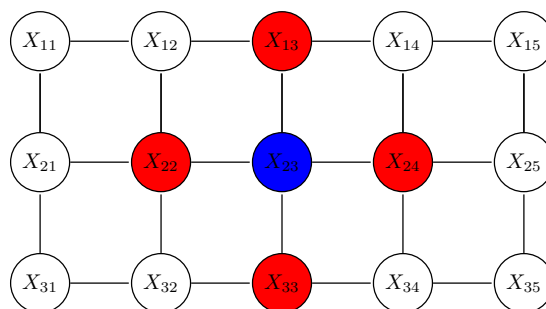
- ▷ Inference of Hidden Markov models requires a lot of data.
- ▷ Continuous distributions are rarely compatible with belief propagation.

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

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

- ▷ Family of normal distributions is compatible with belief propagation.

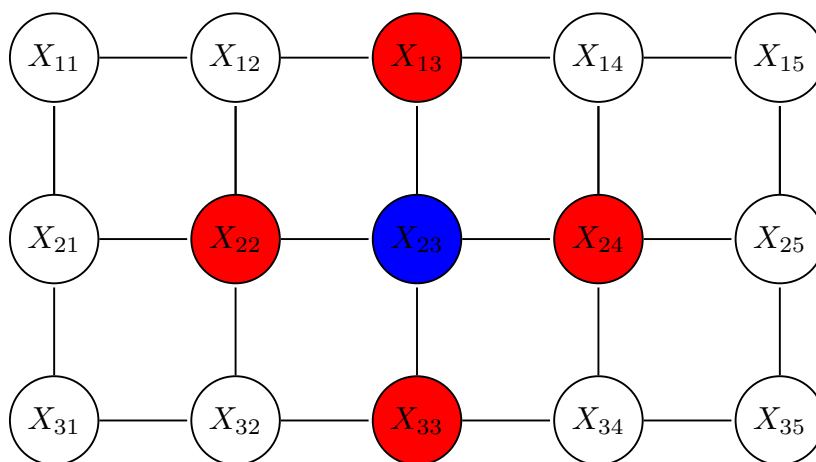
## Background model for digital images



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  $X_1, X_2, \dots$  such that for any node  $X_i$

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

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  $\mathbf{x} = (x_1, x_2, \dots)$  generated by a Markov random field can be expressed in the form

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

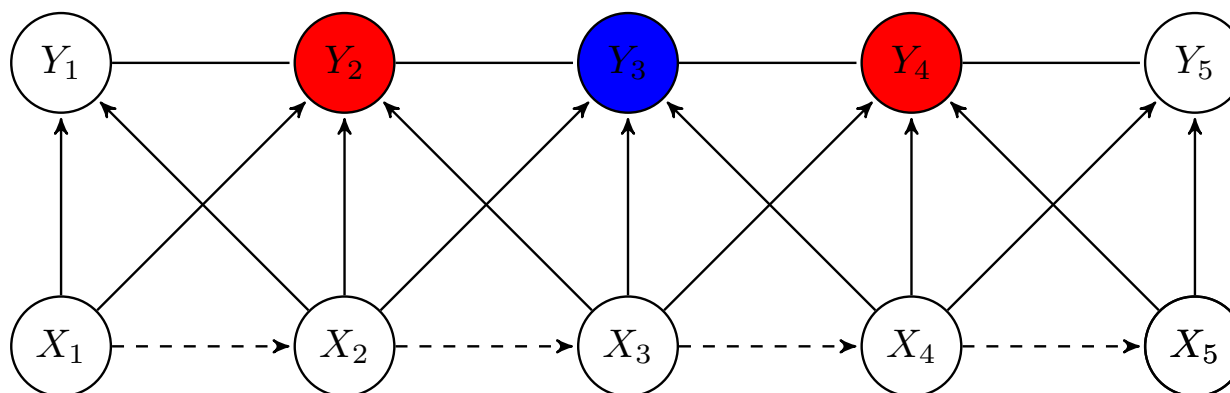
where

- ▷  $Z(\omega)$  is a normalising constant
- ▷ MaxClique is the set of maximal cliques in the Markov random field
- ▷  $\Psi_c$  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

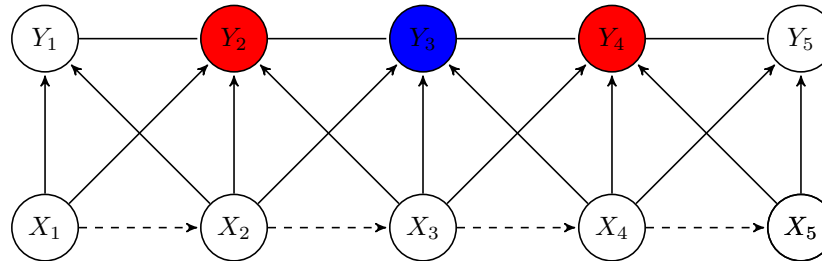


**Definition.** Let  $X_1, X_2, \dots$  and  $Y_1, Y_2, \dots$  be random variables. The entire process is conditional random field if random variables  $Y_1, Y_2, \dots$  conditioned for any sequence of observations  $x_1, x_2, \dots$  form a Markov random field

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

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

# Image segmentation and sequence labelling



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

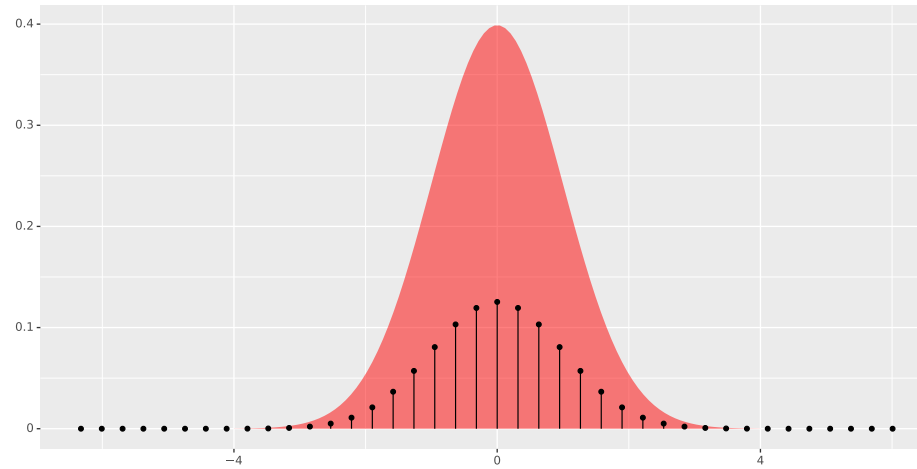
## Consequences of Hammersley-Clifford theorem

- ▷ Clique features  $\Psi_c$  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.



# Univariate normal distribution

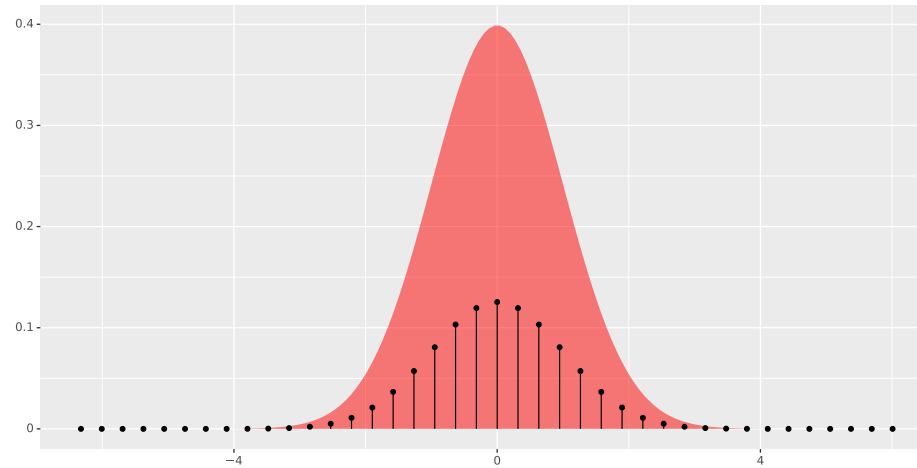
# Probability density function



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

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

# Probability mass function

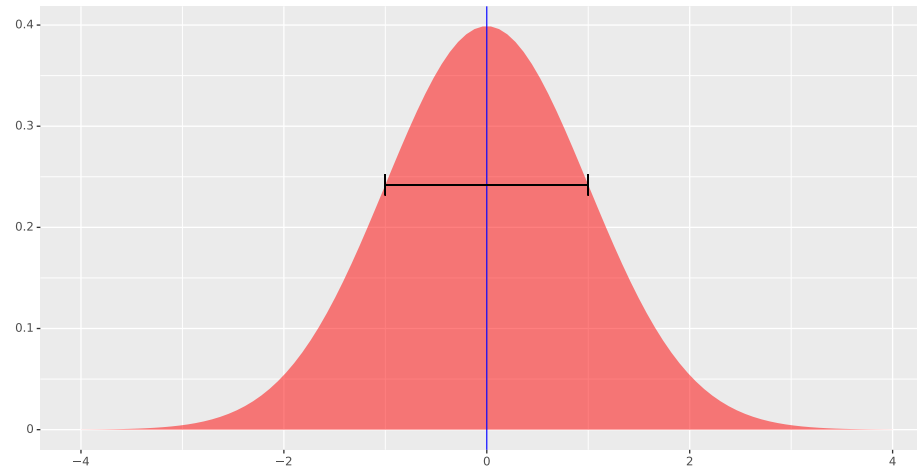


**Definition.** A real-valued random variable  $X$  comes from a discrete distribution with *a probability mass function*  $p : \mathbb{R} \rightarrow \mathbb{R}^+ \cup \{0\}$  defined as

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

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

# 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

$$y = bx + a$$

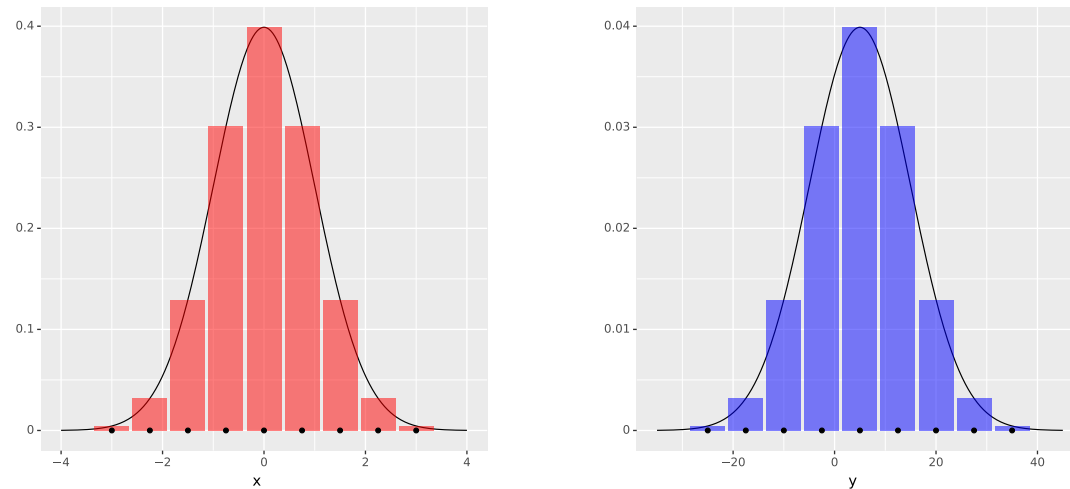
where  $x$  is distributed according to standardised normal distribution  $\mathcal{N}(0, 1)$ .

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  $\mu$  and variance  $\sigma^2$  for this distribution.

# Density derivation



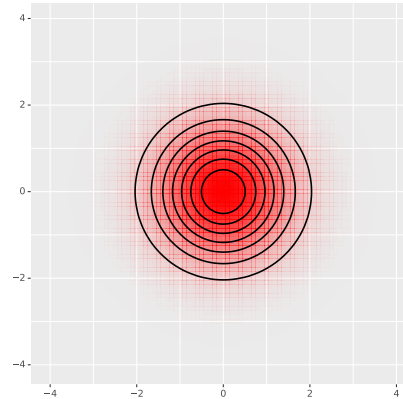
Let  $y = ax + b$  the the relation between densities

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

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

# Multivariate normal distribution

# White Gaussian noise



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

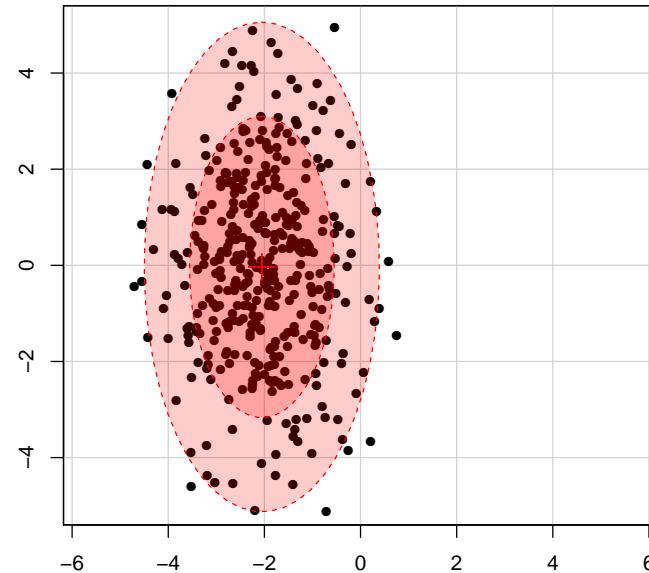
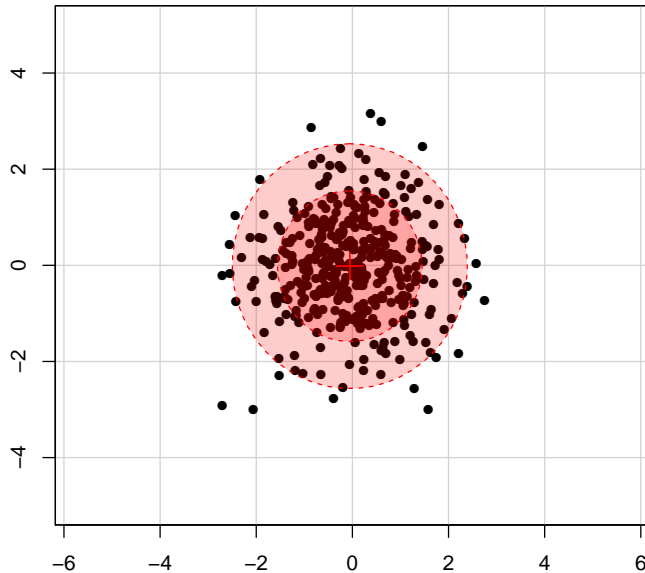
▷ The density can be computed based on independence:

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



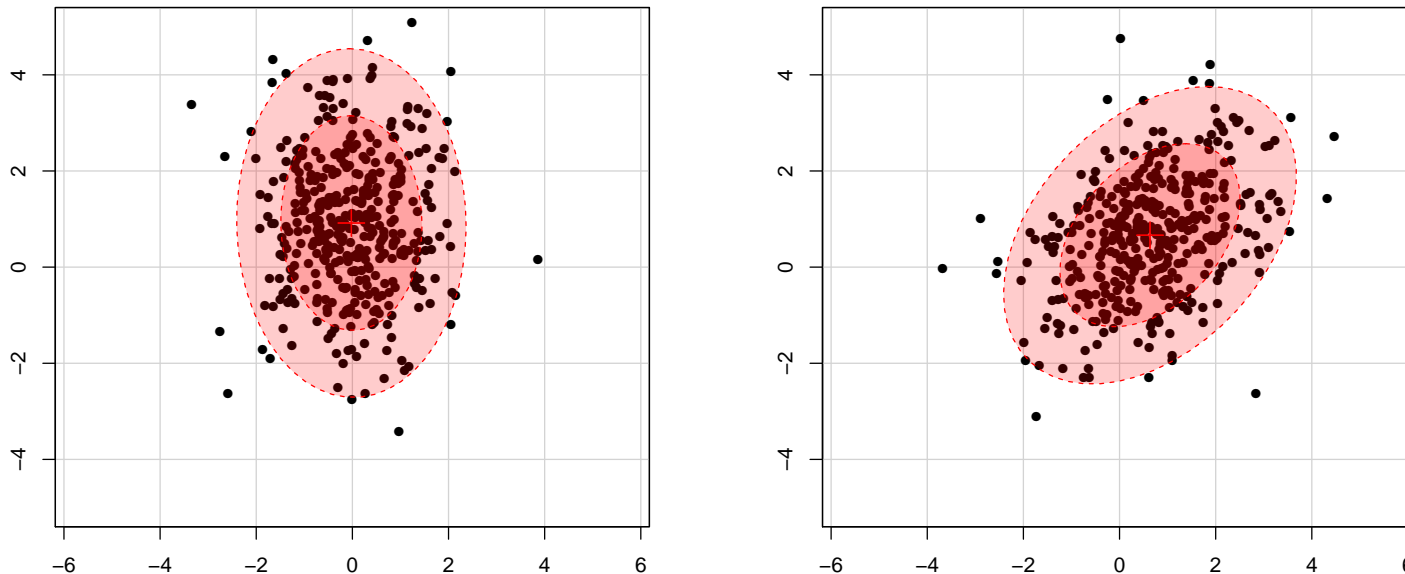
## Scaling and shifting

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

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

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

**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  $\mathcal{R}$ :

$$\Pr_{x_1^*, x_2^*} \left[ \begin{array}{l} x_1 \leq x_1^* \leq x_1 + \Delta x_1 \\ x_2 \leq x_2^* \leq x_2 + \Delta x_2 \end{array} \right] = p(x_1, x_2) \cdot \underbrace{\Delta x_1 \Delta x_2}_S + \varepsilon$$

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

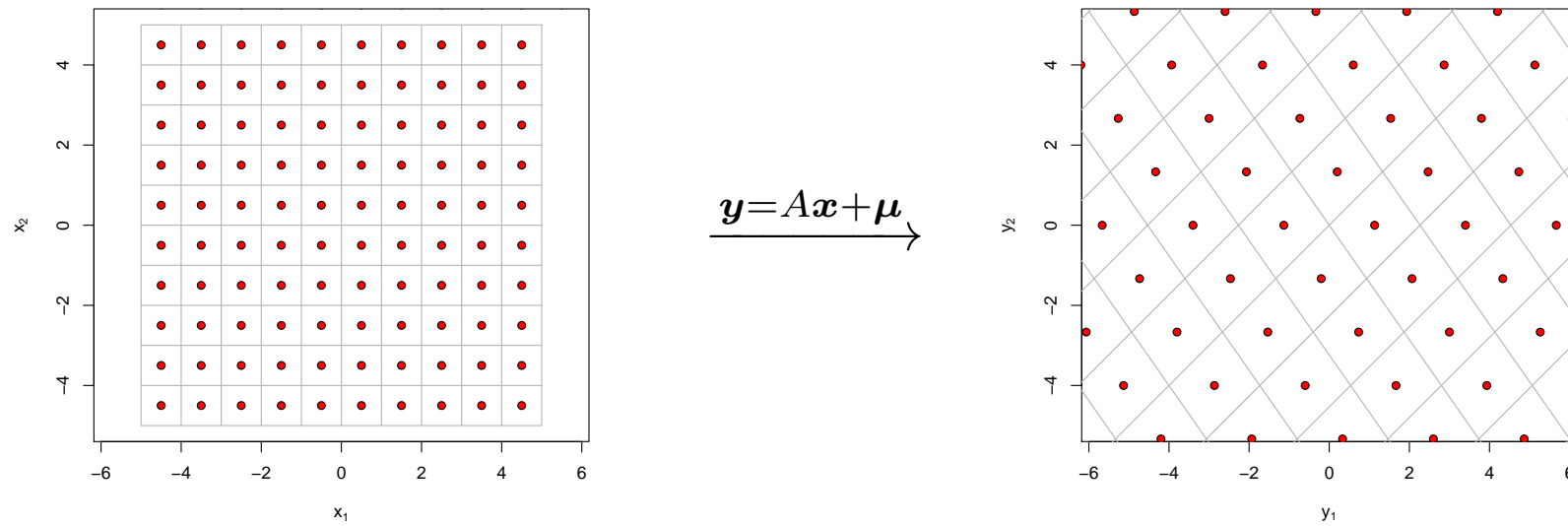
**Remark.** Regions  $\mathcal{R}$  do not have to be rectangular as long as:

- ▷ 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.



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

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

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:

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

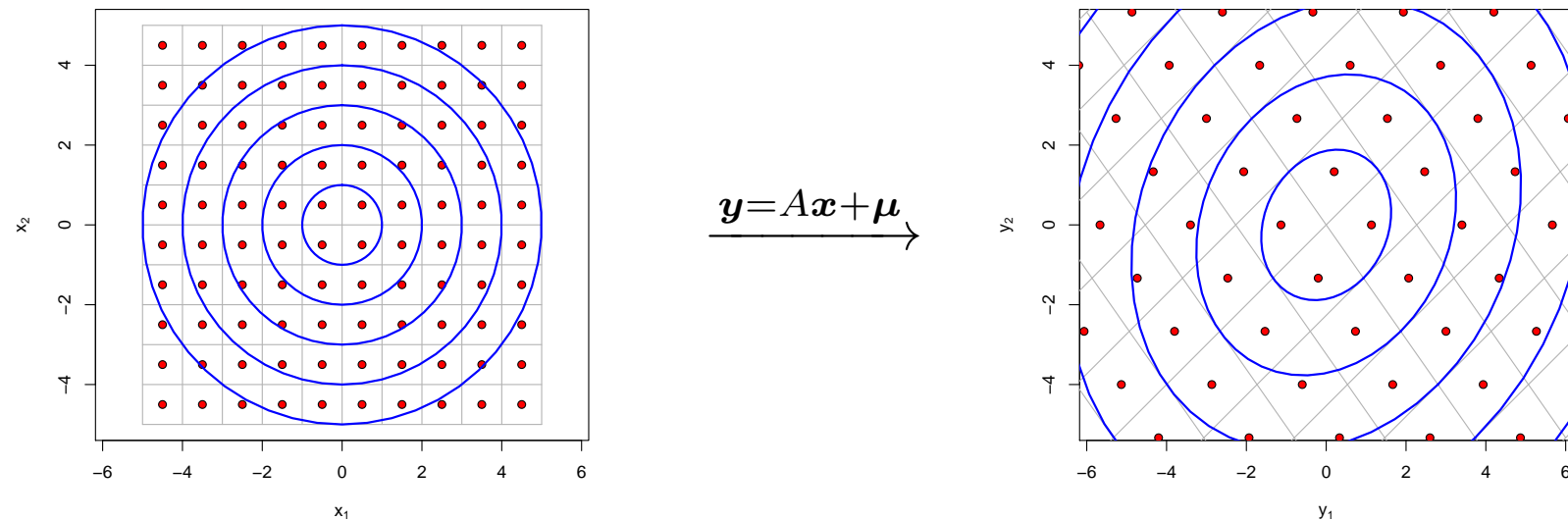
To estimate density  $q(y_1, y_2)$ , we must find the corresponding  $(x_1, x_2)$ :

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

Thus we get

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

## 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  $\mathbf{y}$  are generated by applying the affine transformation

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

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

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

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:

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

where

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

## Multivariate normal distribution as likelihood

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

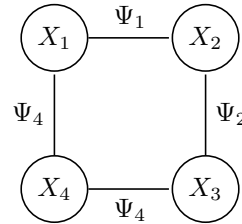
$$\Psi(\mathbf{x}) = \sum_{c \in \text{MaxClique}} \Psi_c(\mathbf{x}_c, \omega)$$

is also a quadratic form and thus  $p[\mathbf{x}|\omega]$  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

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

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

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

# Important properties of normal distributions

## Closeness under marginalisation

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

▷ Moment matching gives the parameters of the resulting distribution

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

## Closeness under linear combinations

Linear combination  $y = \alpha_1^T x_1 + \alpha_2^T x_2$  of independent multivariate normal distributions  $x_1 \sim \mathcal{N}(\mu_1, \Sigma_1)$  and  $x_2 \sim \mathcal{N}(\mu_2, \Sigma_2)$  is also a multivariate normal distribution.

▷ Moment matching gives the parameters of the resulting distribution

$$\mathbf{E}(y) = \alpha_1^T \mathbf{E}(x_1) + \alpha_2^T \mathbf{E}(x_2) = \alpha_1^T \mu_1 + \alpha_2^T \mu_2$$

$$\begin{aligned} \mathbf{Var}(y) &= \mathbf{Cov}(\alpha_1^T x_1) + \mathbf{Cov}(\alpha_2^T x_2) \\ &= \alpha_1^T \mathbf{Cov}(x_1) \alpha_1 + \alpha_2^T \mathbf{Cov}(x_2) \alpha_2 \\ &= \alpha_1^T \Sigma_1 \alpha_1 + \alpha_2^T \Sigma_2 \alpha_2 \end{aligned}$$

▷ Closeness under linear combinations holds also for matrix combinations.



## Closeness under conditioning

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

▷ Moment matching gives the parameters of the resulting distribution

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

$$\mathbf{Cov}(\mathbf{x}|\mathbf{y}_*) = \Sigma_{1,1} - \Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1}$$

# Kalman filter

To be completed next year

## Sensor fusion problem. Kalman filter

Several sensors measure a physical system

- ▷ Measurements are observable as  $\mathbf{y} \in \mathbb{R}^p$ .
- ▷ Physical system has an hidden state  $\mathbf{x} \in \mathbb{R}^n$ .
- ▷ Physical system evolves linearly  $\mathbf{x}_{i+1} = A\mathbf{x}_i + \mathbf{w}_i$ .
- ▷ Measurements are linear from the state  $\mathbf{y}_i = C\mathbf{x}_i + \mathbf{v}_i$ .

Unknown quantities in the system

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