LTAT.02.004 Machine Learning II

Multivariate normal distribution

Direct applications

Sven Laur
University of Tartu

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

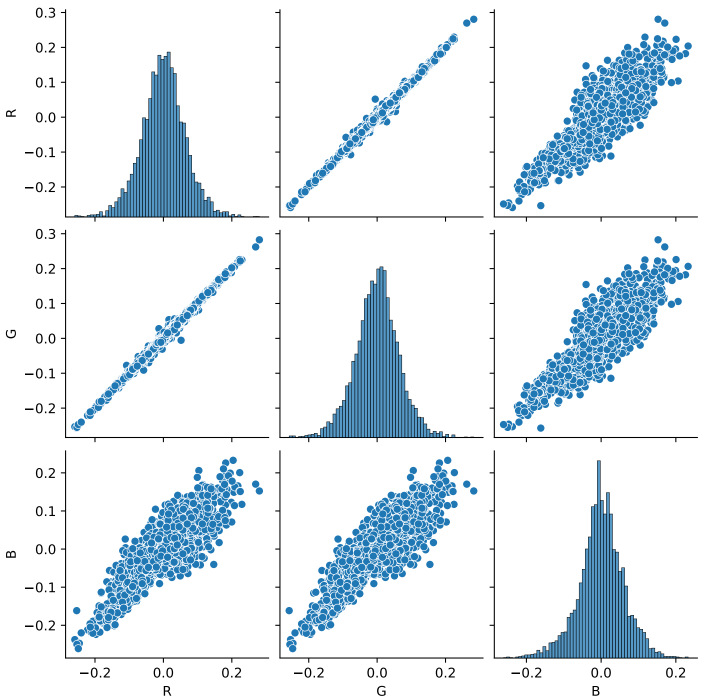
$$\mathbf{E}(\boldsymbol{x}|\boldsymbol{y}_*) = \boldsymbol{\mu}_1 + \Sigma_{1,2}\Sigma_{2,2}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_2)$$
$$\mathbf{Cov}(\boldsymbol{x}|\boldsymbol{y}_*) = \Sigma_{1,1} - \Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1}$$

Motivating examples

Filtering and smoothing

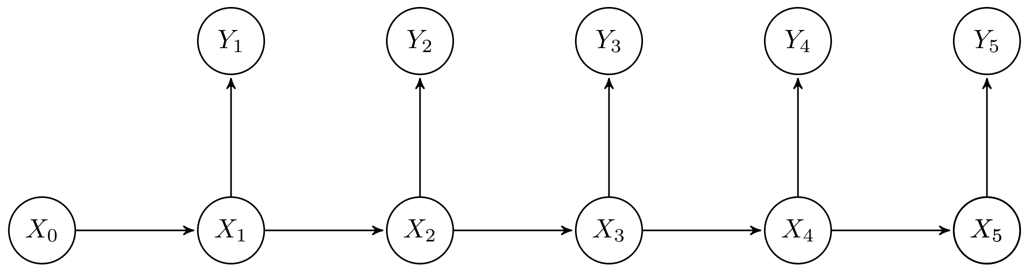
Prediction of vector values

Prediction errors of different vector components can be correlated.



As a result combined model can outperform coordinatewise predictions.

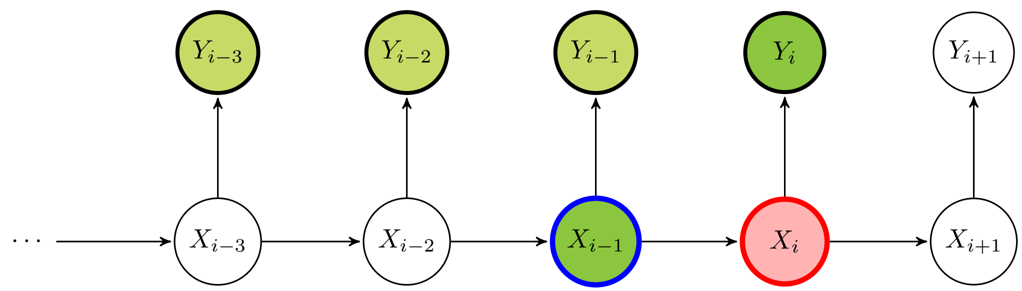
Sensor fusion with Hidden Markov Models



A standard problem in robotics or machine perception is following.

- 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.

Kalman filter

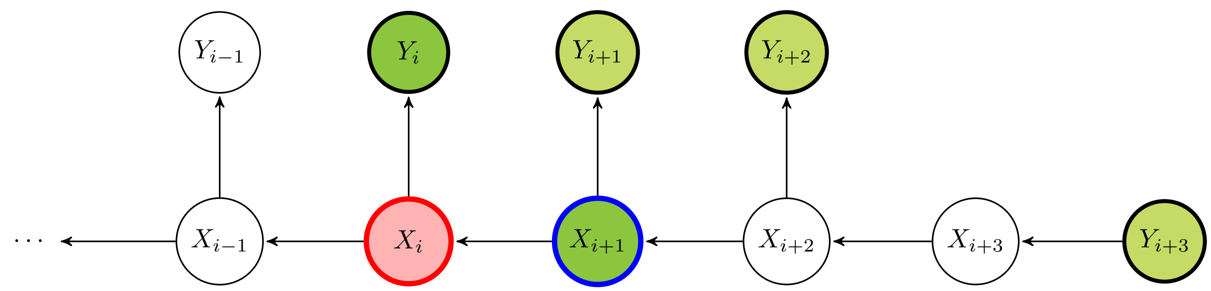


As before we can consider the prior and filter densities

$$\pi[\boldsymbol{x}_i] = p[\boldsymbol{x}_i|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{i-1}]$$
$$f[\boldsymbol{x}_i] = p[\boldsymbol{x}_i|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_i] \propto \pi[\boldsymbol{x}_i] \cdot p[\boldsymbol{y}_i|\boldsymbol{x}_i]$$

A similar update logic assures that both distributions are normal distributions and that we can only compute the parameters of these normal distributions.

Smoothing and reverse Hidden Markov Model

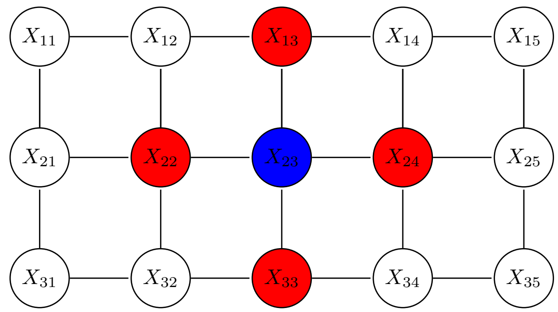


- We need likelihoods $\lambda[\boldsymbol{x}_i] = p[\boldsymbol{y}_{i+1}, \ldots, \boldsymbol{y}_{\boldsymbol{n}}|\boldsymbol{x}_i]$ for the smoothing.
- Likelihood propagation formula is analogous to the prior propagation.
- We can define a reverse HMM such that the prior $\pi^*[\boldsymbol{x}_i] \propto \lambda[\boldsymbol{x}_i]$.
- The resulting HMM has reversed dynamics.
- It turns out that all likelihoods $\lambda[\boldsymbol{x}_i]$ are normal distributions.
- The posterior as product $\pi[\boldsymbol{x}_i]\cdot\lambda[\boldsymbol{x}_i]\cdot p[\boldsymbol{y}_i|\boldsymbol{x}_i]$ is also a normal distribution.

Motivating examples

Markov fields

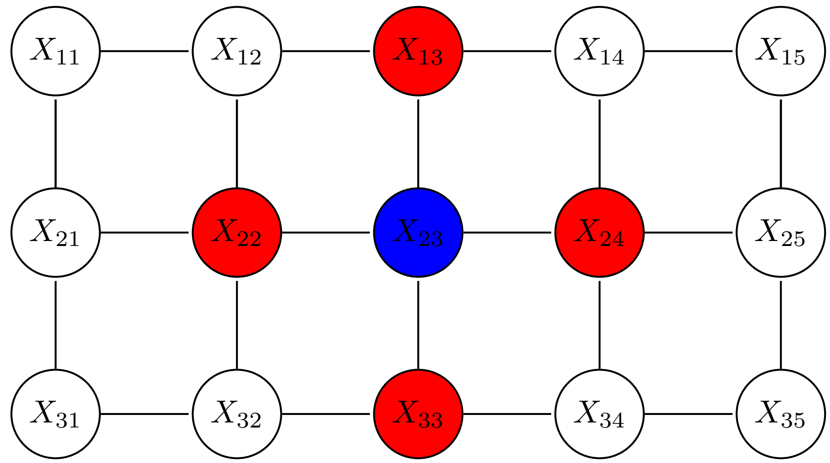
Background model for digital images



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

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

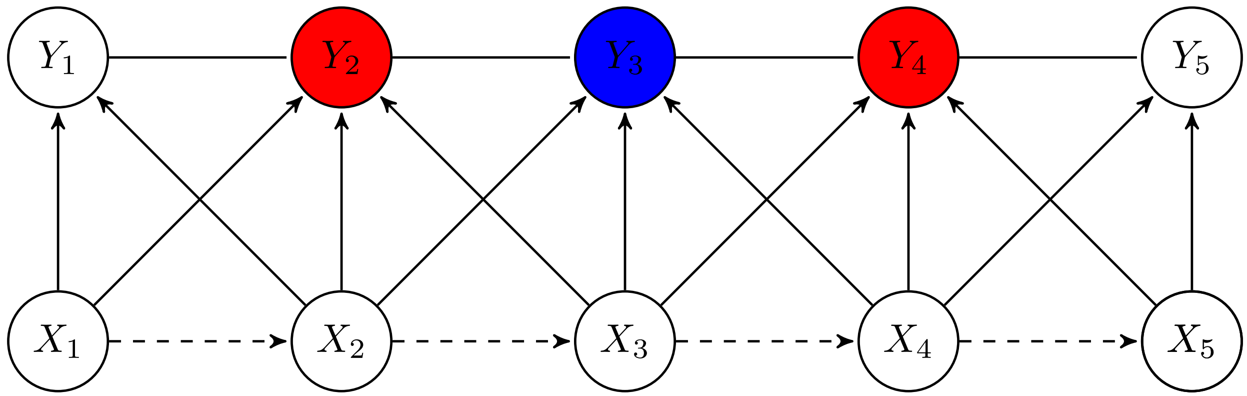
where

- $Z(\omega)$ 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

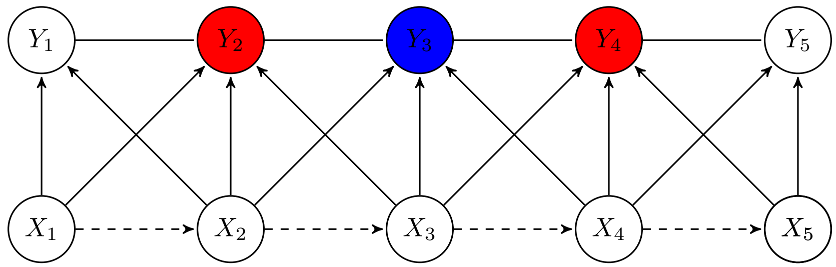


Definition. Let $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ be random variables. The entire process is conditional random field if random variables $Y_1, Y_2, \ldots$ conditioned for any sequence of observations $x_1, x_2, \ldots$ form a Markov random field

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

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.

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

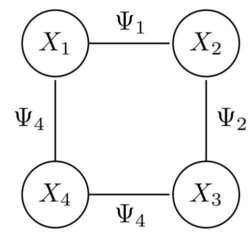
$$\Psi(\boldsymbol{x}) = \sum_{c \in \mathsf{MaxClique}} \Psi_c(\boldsymbol{x}_c, \omega)$$

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

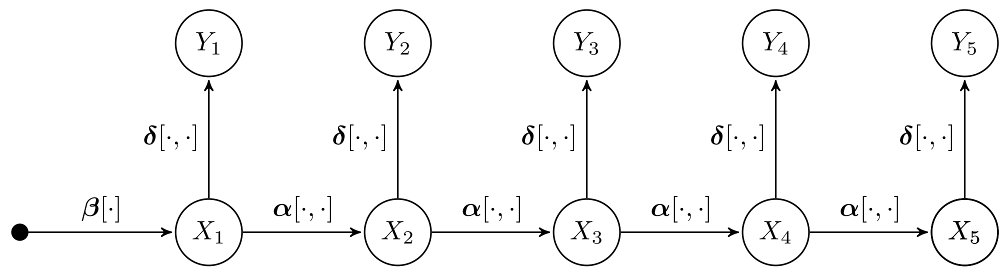
$$\Psi(\boldsymbol{x}) = \boldsymbol{x}^T \begin{pmatrix} \alpha_1 + \alpha_4 + \delta_1^2 & -\alpha_1 & 0 & -\alpha_4 \\ -\alpha_1 & \alpha_1 + \alpha_2 + \delta_2^2 & -\alpha_2 & 0 \\ 0 & -\alpha_2 & \alpha_2 + \alpha_3 + \delta_3^2 & -\alpha_3 \\ -\alpha_4 & 0 & -\alpha_3 & \alpha_3 + \alpha_4 + \delta_4^2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

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

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

Motivating examples

Hidden Markov Model



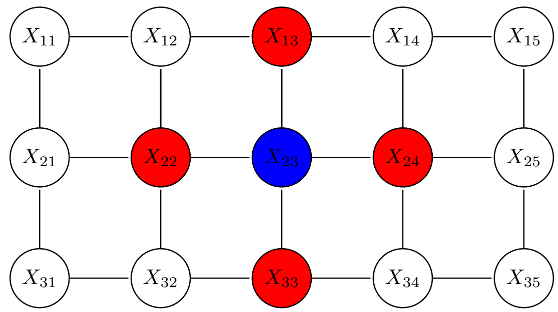
- ▷ Inference of Hidden Markov models requires a lot of data.
- ▷ Continuous distributions are rarely compatible with belief propagation.

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

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

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

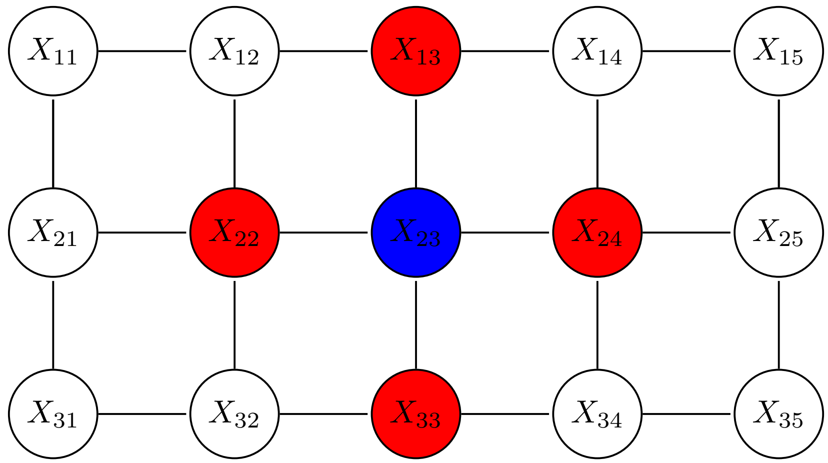
Background model for digital images



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 $X_1, X_2, \ldots$ 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

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

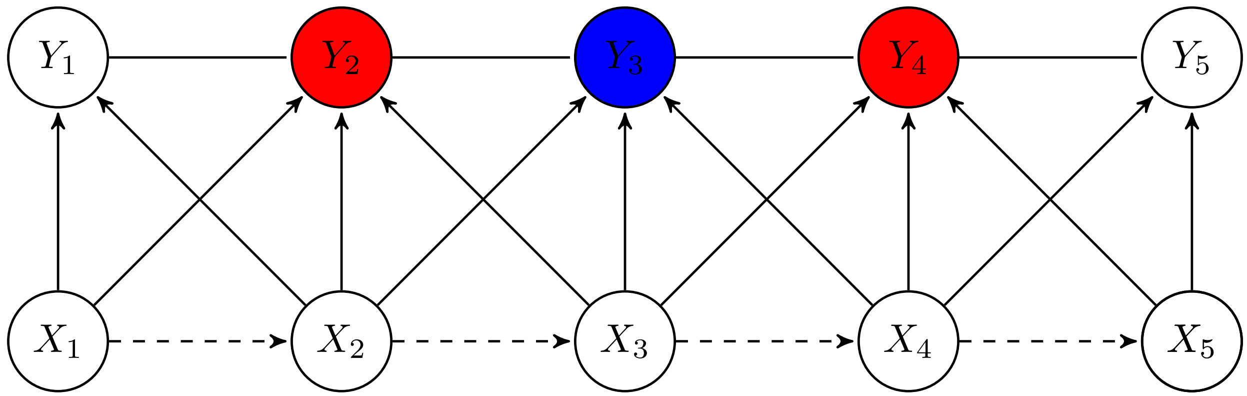
where

- $Z(\omega)$ 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

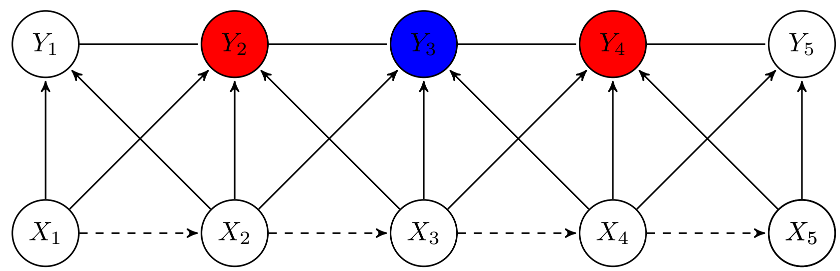


Definition. Let $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ be random variables. The entire process is conditional random field if random variables $Y_1, Y_2, \ldots$ conditioned for any sequence of observations $x_1, x_2, \ldots$ form a Markov random field

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

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

Image segmentation and sequence labelling

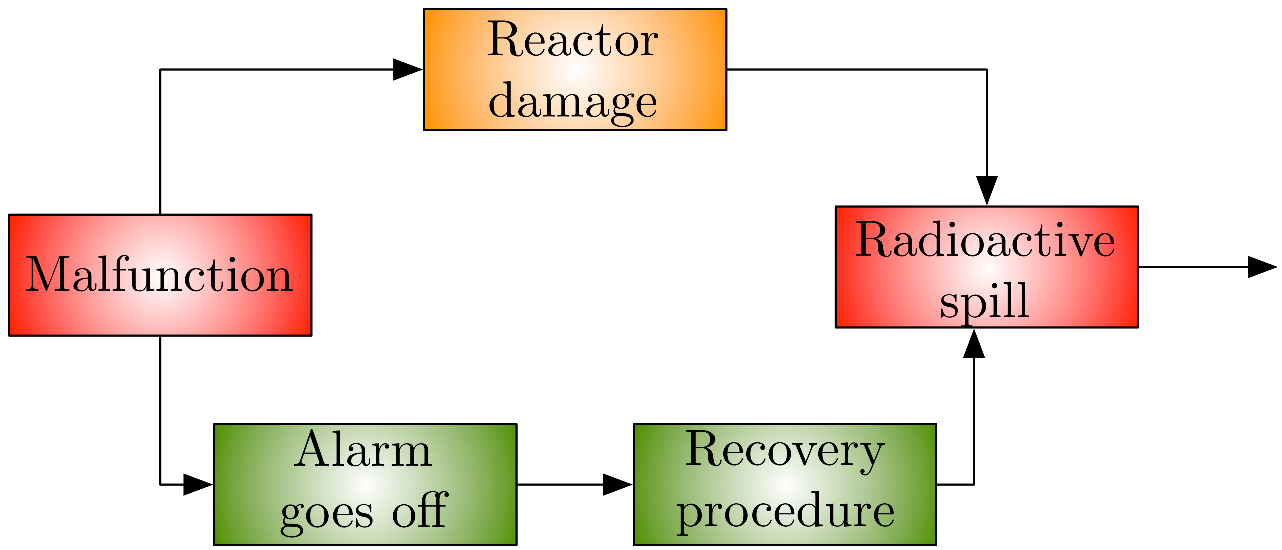


- The input $\boldsymbol{x}$ is used to predict labels $y_1, y_2, \ldots$.
- A correct label sequence must satisfy possibly unknown restrictions.
- These restrictions are captured by conditional random random field.

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.

Going beyond naive Bayesian models



Complex causal models are often defined through Bayesian networks

- ▷ A complex processes is first split into sub-events
- ▷ Direct causal dependencies between sub-events are detected
- ▷ Causation mechanisms are characterised with probability tables

Strength and weaknesses of Bayesian networks

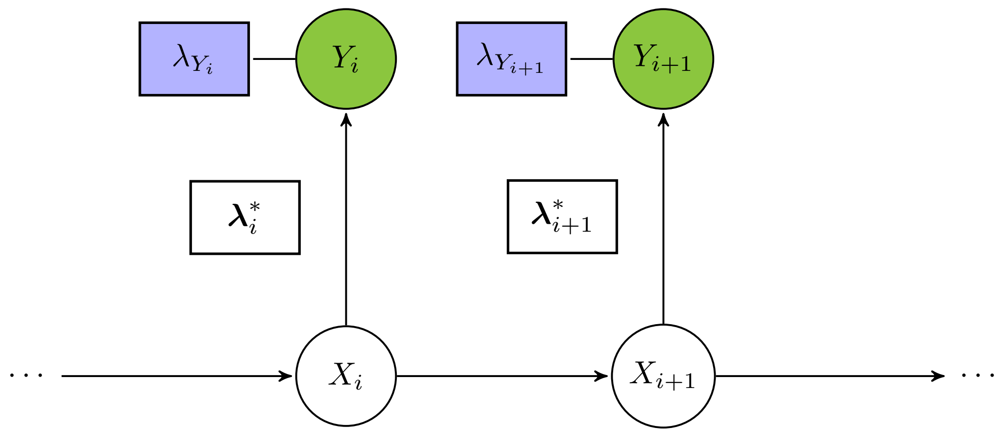
Strengths

- Bayesian networks are easy to interpret
- Bayesian networks are good for formalising fuzzy background knowledge
- Estimation of individual probability tables is tractable
- There are tools for doing inference with Bayesian networks

Weaknesses

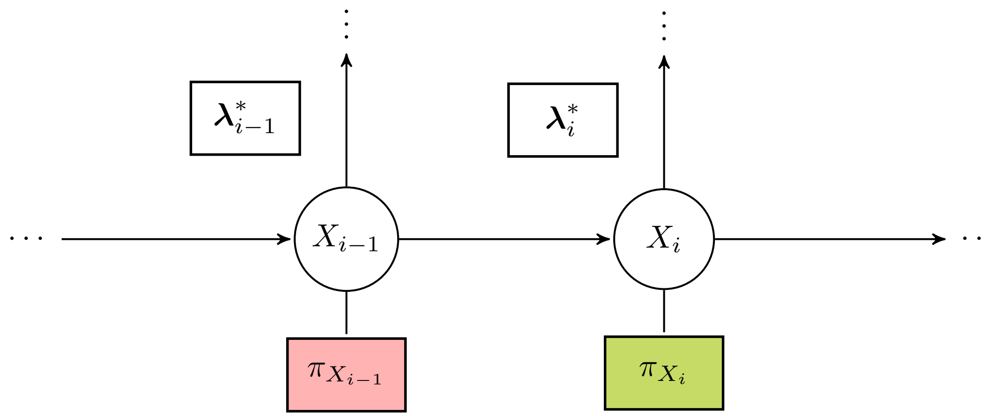
- You must know the causal structure of sub-events
- Identification of causal structure form data alone is very difficult
- It is notoriously difficult to model non-trivial causal dependencies
- Standard inference procedures often do not have closed solutions

Belief propagation. Initialisation



- ▷ We have a direct evidence $Y_i = y_i$ for each node Y_i.
- ▷ The likelihood vector is infinite and captured by $\lambda_{Y_i} = \delta_{y_i}$.
- ▷ The local likelihood $\lambda_i^*(x_i) = p[Y_i = y_i|x_i]$ is an infinite vector.
- ▷ The form $\boldsymbol{y}_i = C\boldsymbol{x}_i + \boldsymbol{v}_i$ assures that $\boldsymbol{y}_i|\boldsymbol{x}_i$ is normal distribution.
- ▷ The local likelihood λ_i^* has a finite description.

Prior propagation. Filtering

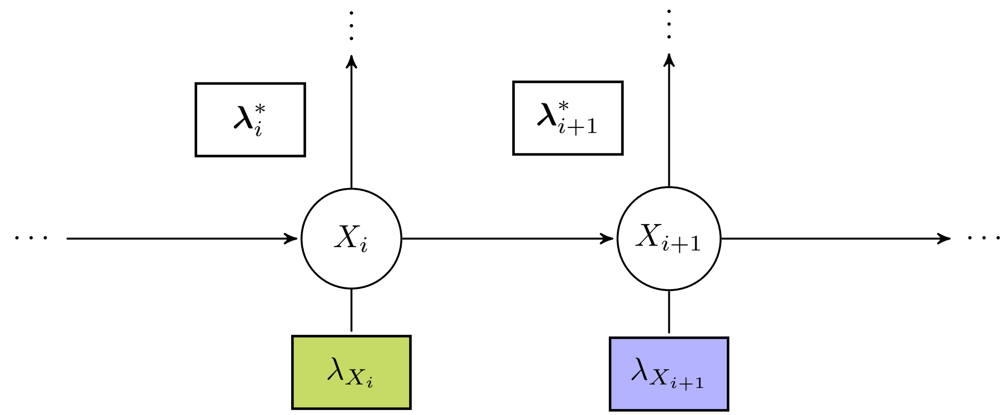


Prior propagation rule

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

leads to a finite description because on the right is a normal distribution.

Likelihood propagation. Smoothing



Likelihood propagation rule

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

leads to a finite description because on the right is a normal distribution.