Learning, Probabilities and Causality

Chapter I - Introduction and Conditional Independence

Xavier Alameda-Pineda and Thomas Hueber

Ensimag/Inria/CNRS/Univ. Grenoble-Alpes

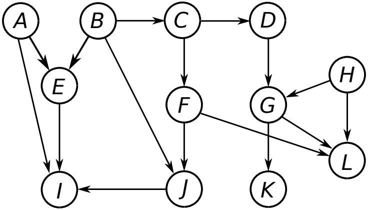
Table of Today's Contents

Course Organisation

Course Content

Learning Probabilities and Causality is structured in two parts.

1. Learning for probabilistic models given a causality graph (Thomas & Xavi)
2. Methods for inferring this graph from data (Emilie & Eric)



Course Content - Part 1

Probabilistic learning with latent variables:

1. Basics of probabilistic models: **conditional independence**
2. Model-based clustering and **Gaussian** mixture models
3. Sequential data and hidden **Markov** models
4. Probabilistic principal component analysis
5. Linear dynamical systems
6. Approximate variational inference

LPC Part 1 - Instructors

Research Scientist [**website**] @thomashueber
`thomas.hueber@gipsa-lab.fr`
Leader of CRISSP (cognitive robotics, interactive systems, speech processing), GIPSA-Lab, CNRS

Multimodal speech processing, machine learning, interactive systems

Research Scientist [**webpage**] @xavirema
`xavier.alameda-pineda@inria.fr`
Perception Team, Inria Grenoble Rhône-Alpes

Audio-visual perception, probabilistic and deep learning, human-robot interaction

Grading Rules

- Lab work (LW), mid-term exam (ME) final exam (FE).
- Grade = (LW+ME+FE)/3.
- LW = average of all lab works.

Support material:
`https://chamilo.grenoble-inp.fr/courses/ENSIMAGWMM9AM46`.

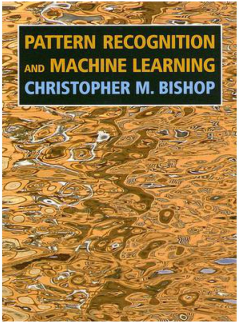
Calendar

When?	Who?	Where?	Comment
30-Sep	Xavi	D211	
7-Oct	Thomas	D207	
14-Oct	Thomas	D207 & E303	Lab (15h30-17h)
21-Oct	Xavi	D111	
28-Oct	Xavi	H105	
18-Nov	Xavi	D111/E303	Lab (15h30-17h)
25-Nov	Emilie/Eric	H105	Mid-term Exam (14h-15h30)
2-Dec	Emilie/Charles	H105	
9-Dec	Emilie	E201	Lab (14h-17h)
16-Dec	Eric/Charles	H105	
6-Jan	Charles/Emilie	H105	
13 Jan	Charles/Eric	H105/E200	Lab (15h30-17h)

References

There is **a lot** of bibliography on probabilistic graphical models.

I strongly suggest the following book:



Pattern Recognition and Machine Learning, from Christopher M. Bishop (Springer)

The concepts discussed in FPDM correspond to different parts of Ch. 2, 8, 9, 10, 12, 13.

You will not find the part on variational autoencoders (last chapter).

Introduction and Motivation

What is probabilistic data mining?

- Probabilistic means we **model** our data using probabilities.
- For example in classification, we aim to estimate the posterior probability: $P(c|x)$ for every possible class c.
- Probabilistic generative models & Bayes rule:

$$p(x, c) = p(x|c)p(c) \quad \Rightarrow \quad p(c|x) = \frac{p(x|c)p(c)}{\sum_k p(x|k)p(k)}$$

- What are all these "p"? What do they mean?

Probabilities

For **discrete variables** (i.e. measurable events are discrete):

$$p(c) = P(C = c).$$

$\rightarrow$ the probability of the random variable C to value event c.

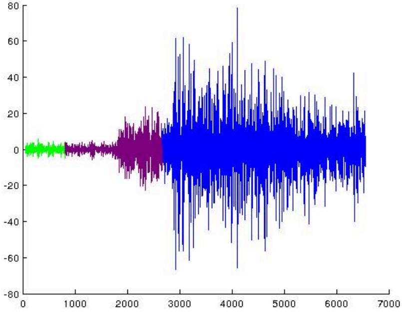
For **continuous variables** (i.e. measurable events are continuous):

$$p(x) = f_X(x) \quad \text{and} \quad p(\mathcal{X}) = P(x \in \mathcal{X}) = \int_{\mathcal{X}} f_X(x) \mathrm{d}x$$

f_X is the probability density function. Remember $P(\{x\}) = 0$.

Why probabilistic data mining?

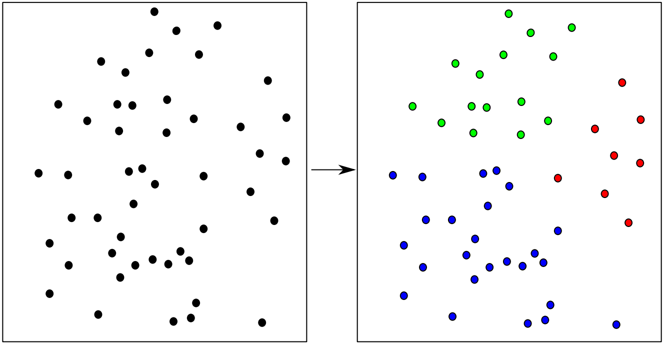
- Infer hidden variables / exploit partly missing data
- Example: clustering, image segmentation
- Incorporate particular requirements in clustering
- Model complex data (on grids, graphs, temporal, ...)
- Simulate phenomena (speech synthesis), make predictions (regime switching in time series)



Segmentation of time series with respect to the variance

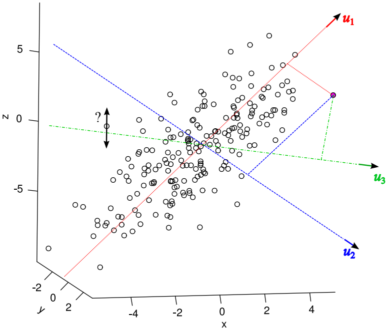
Example (I): Clustering

- Data: points $(x_j)_{j=1,\dots,n}$ in $\mathbb{R}^d$.
- Aim: find (& predict) clusters.
- Model-based approach: let z_j be the (unknown) cluster of x_j.
  $z_i = z_j \Rightarrow x_i$ and x_j should have the same (conditional) distribution.



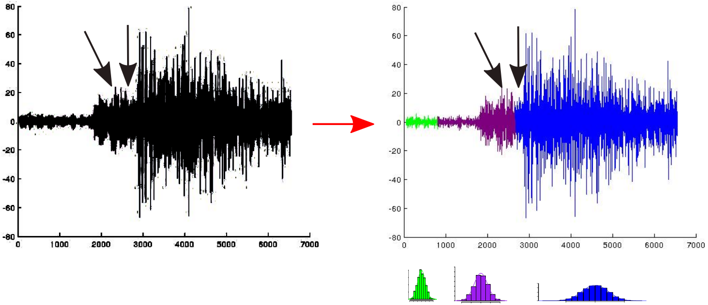
Example (II): Dimensionality reduction

- Raw data are high-dimensional descriptors.
- Difficult to mine patterns/visualize.
- Projection on the directions of maximum variance.
- What if for most or even every point x_j, some coordinates are missing?
- Probabilistic (i.e., model-based) PCA relies on a generative model to exploit partially observed / unknown data.



Example (III): Analysis of sequential data

- Special case of clustering with temporal dependencies
- Piecewise statistically invariant features with Markovian jumps
- **Markovian**: it depends only on a few close neighbors.



The multivariate Gaussian Distribution

1D Gaussians

Let's recall the definition of the univariate Gaussian distribution, for $x \in \mathbb{R}$:

$$p(x) = \mathcal{N}(x; \mu, \nu) = \frac{1}{\sqrt{2\pi\nu}} \exp\Big(- \frac{(x-\mu)^2}{2\nu} \Big),$$

- $\mu = \mathbb{E}_{\mathcal{N}(x;\mu,\nu)}\{x\} \in \mathbb{R}$ is the mean.
- $\nu = \mathbb{E}_{\mathcal{N}(x;\mu,\nu)}\{(x-\mu)^2\} \in \mathbb{R}^+$ is the variance.

1D Gaussians

Let's recall the definition of the univariate Gaussian distribution, for $x \in \mathbb{R}$:

$$p(x) = \mathcal{N}(x; \mu, \nu) = \frac{1}{\sqrt{2\pi\nu}} \exp\Big(- \frac{(x-\mu)^2}{2\nu}\Big),$$

- $\mu = \mathbb{E}_{\mathcal{N}(x;\mu,\nu)}\{x\} \in \mathbb{R}$ is the mean.
- $\nu = \mathbb{E}_{\mathcal{N}(x;\mu,\nu)}\{(x-\mu)^2\} \in \mathbb{R}^+$ is the variance.

Remark 1.1: We will often use the **expectation** of a function f of a random variable x w.r.t. the probability density function $p(x)$, and denote it by:

$$\mathbb{E}_{p(x)}\{f(x)\} = \int_{\mathcal{X}} f(x)p(x)\mathrm{d}x,$$

where $\mathcal{X}$ is the domain of the random variable x.

1D Gaussians: ML estimators

And of its ML estimators for a set of N samples $X = \{x_1, \dots, x_N\}$:

$$\mathcal{L}(\mu, \nu | X) = \sum_{n=1}^{N} \log \mathcal{N}(x_n; \mu, \nu)$$

Exercise 1.1: Prove that the maximum likelihood estimators are:

$$\mu^* = \frac{1}{N} \sum_{n=1}^{N} x_n \qquad \nu^* = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu^*)^2.$$

Hint: Compute $\frac{\partial \mathcal{L}}{\partial \mu}$ and $\frac{\partial \mathcal{L}}{\partial \nu}$ knwoing that
$\log \mathcal{N}(x; \mu, \nu) = -\frac{1}{2}\left(\log(2\pi\nu) + \frac{(x-\mu)^2}{\nu}\right)$.

Proof

$$\begin{aligned}
\mathcal{L}(\mu, \nu | X) &= \sum_{n=1}^{N} \log \mathcal{N}(x_n; \mu, \nu) \\
&= -\frac{1}{2} \sum_{n=1}^{N} \log(2\pi\nu) + \frac{(x_n - \mu)^2}{\nu} \\
&= -\frac{1}{2} \left(N \log(2\pi\nu) + \frac{1}{\nu} \sum_{n=1}^{N} (x_n - \mu)^2 \right)
\end{aligned}$$

And hence:

$$\frac{\partial \mathcal{L}}{\partial \mu} = \frac{1}{\nu} \sum_{n=1}^{N} (x_n - \mu) \qquad \frac{\partial \mathcal{L}}{\partial \nu} = -\frac{N}{2\nu} + \frac{1}{2\nu^2} \sum_{n=1}^{N} (x_n - \mu)^2$$

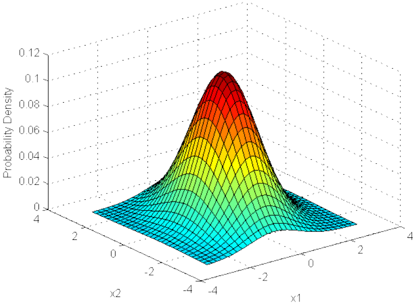
By setting the derivatives to 0, we obtain the sought result.

Multivariate Gaussian distribution

Trivial extension: consider each dimension independently.

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\nu}) = \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi\nu_d}} \exp\left(- \frac{(x_d - \mu_d)^2}{2\nu_d} \right),$$

with $\mathbf{x} = (x_1, \ldots, x_D) \in \mathbb{R}^d$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_D) \in \mathbb{R}^d$ and $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_D) \in \mathbb{R}^+$.



Exercise 1.2: Derive the maximum likelihood estimators for $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$.

Multivariate Gaussian distribution (II)

By defining:

$$\boldsymbol{\Sigma} = \mathrm{diag}(\boldsymbol{\nu}) = \begin{pmatrix} \nu_1 & 0 & \dots & 0 \\ 0 & \nu_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \nu_D \end{pmatrix}$$

the density rewrites as:

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}} \exp\Big(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\Big).$$

Exercise 1.3: Prove it!

Symmetric and positive definite matrices

Question: does it work for any matrix $\boldsymbol{\Sigma}$?

Only for *symmetric* and *positive definite* (s.p.d.) matrices.

Remark 1.2: A $D \times D$ symmetric matrix $\boldsymbol{\Sigma}$ is **positive definite** if and only if $\boldsymbol{v}^\top \boldsymbol{\Sigma} \boldsymbol{v} > 0, \forall \boldsymbol{v} \neq \mathbf{0}$.

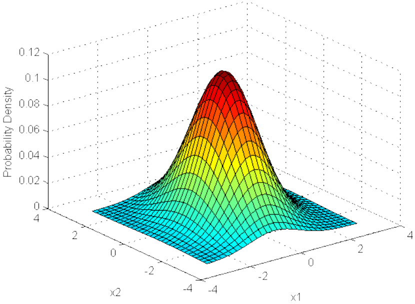
Symmetric and positive definite matrices

Question: does it work for any matrix $\boldsymbol{\Sigma}$?

Only for *symmetric* and *positive definite* (s.p.d.) matrices.

Remark 1.2: A $D \times D$ symmetric matrix $\boldsymbol{\Sigma}$ is **positive definite** if and only if $\boldsymbol{v}^\top \boldsymbol{\Sigma} \boldsymbol{v} > 0, \forall \boldsymbol{v} \neq \mathbf{0}$.

For the Gaussian distribution, this is intuitive, since the variance should be strictly positive in any direction:



Symmetric and positive definite matrices (II)

Let $\mathbf{\Sigma}$ be a s.p.d. matrix:

- All eigenvalues of $\mathbf{\Sigma}$ are ...?

Symmetric and positive definite matrices (II)

Let $\boldsymbol{\Sigma}$ be a s.p.d. matrix:

- All eigenvalues of $\boldsymbol{\Sigma}$ are ...? Real and

Symmetric and positive definite matrices (II)

Let $\mathbf{\Sigma}$ be a s.p.d. matrix:

- All eigenvalues of $\mathbf{\Sigma}$ are ...? Real and strictly positive!
- The inverse of $\mathbf{\Sigma}$ is ...?

Symmetric and positive definite matrices (II)

Let $\mathbf{\Sigma}$ be a s.p.d. matrix:

- All eigenvalues of $\mathbf{\Sigma}$ are ...? Real and strictly positive!
- The inverse of $\mathbf{\Sigma}$ is ...? Symmetric and positive definite.
- How do we know that $\mathbf{\Sigma}^{-1}$ exists?

Symmetric and positive definite matrices (II)

Let $\boldsymbol{\Sigma}$ be a s.p.d. matrix:

- All eigenvalues of $\boldsymbol{\Sigma}$ are ...? Real and strictly positive!
- The inverse of $\boldsymbol{\Sigma}$ is ...? Symmetric and positive definite.
- How do we know that $\boldsymbol{\Sigma}^{-1}$ exists? The determinant is the product of eigenvalues.
- We can write $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ with $\boldsymbol{\Lambda}$ diagonal and $\boldsymbol{U}$ orthogonal. Why?

Symmetric and positive definite matrices (II)

Let $\boldsymbol{\Sigma}$ be a s.p.d. matrix:

- All eigenvalues of $\boldsymbol{\Sigma}$ are ...? Real and strictly positive!
- The inverse of $\boldsymbol{\Sigma}$ is ...? Symmetric and positive definite.
- How do we know that $\boldsymbol{\Sigma}^{-1}$ exists? The determinant is the product of eigenvalues.
- We can write $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ with $\boldsymbol{\Lambda}$ diagonal and $\boldsymbol{U}$ orthogonal. Why?
- So $\boldsymbol{\Lambda}$ contains eigenvalues and $\boldsymbol{U}$ contains eigenvectors (as columns).

Symmetric and positive definite matrices (II)

Let $\boldsymbol{\Sigma}$ be a s.p.d. matrix:

- All eigenvalues of $\boldsymbol{\Sigma}$ are ...? Real and strictly positive!
- The inverse of $\boldsymbol{\Sigma}$ is ...? Symmetric and positive definite.
- How do we know that $\boldsymbol{\Sigma}^{-1}$ exists? The determinant is the product of eigenvalues.
- We can write $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ with $\boldsymbol{\Lambda}$ diagonal and $\boldsymbol{U}$ orthogonal. Why?
- So $\boldsymbol{\Lambda}$ contains eigenvalues and $\boldsymbol{U}$ contains eigenvectors (as columns).
- Then, $\boldsymbol{\Sigma}^{-1} = \boldsymbol{U}\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^\top$.

Exercise 1.4: Prove that the inverse of a (symmetric) positive definite matrix always exists.

Defining multivariate Gaussians

Remark 1.3: Given a vector $\boldsymbol{\mu} \in \mathbb{R}^D$ and a s.p.d. matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{D\times D}$, we can define the **multivariate Gaussian** distribution as:

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}} \exp\Big(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\Big)$$

$\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are usually referred to as the *mean vector* and the *covariance matrix*, and they are defined as:

$$\boldsymbol{\mu} = \mathbb{E}_{\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})}\{\mathbf{x}\} \qquad \boldsymbol{\Sigma} = \mathbb{E}_{\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})}\{(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top\}.$$

Exercise 1.5: Prove that the normalisation constant of a multivariate Gaussian distribution with covariance matrix $\boldsymbol{\Sigma}$ is $\sqrt{|2\pi\boldsymbol{\Sigma}|}$.

Jupyter Notebook!!!

Standard Gaussian and Affine Transforms

Remark 1.4: The **standard multivariate Gaussian** is defined as the zero-mean and unit-variance Gaussian distribution:

$$\mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I}) = \frac{1}{(2\pi)^{D/2}} \exp\Big(-\frac{1}{2}\|\mathbf{z}\|^2\Big).$$

Exercise 1.6: Let us consider the case where $\mathbf{z}$ follows a standard multivariate Gaussian distribution, and we define $\mathbf{x} = \mathbf{A}\mathbf{z} + \boldsymbol{\mu}$ with $\mathbf{A} \in \mathbb{R}^{D\times D}$ being an invertible matrix ($|\mathbf{A}| \neq 0$). Prove that:

$$p(\mathbf{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \text{with} \quad \boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}^\top.$$

More on Multivariate Gaussians

What are the level curves of the Gaussian p.d.f.?

$$\mathcal{C}_\lambda = \{\boldsymbol{x} | \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \lambda\}.$$

More on Multivariate Gaussians

What are the level curves of the Gaussian p.d.f.?

$$\mathcal{C}_\lambda = \{\boldsymbol{x} | \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \lambda\}.$$

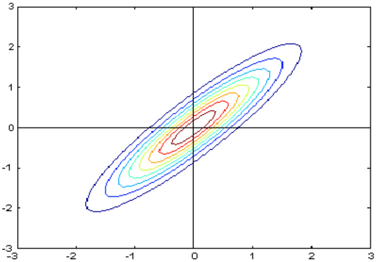
- Empty set for $\lambda < 0$.
- $\mathcal{C}_0 = \{\boldsymbol{\mu}\}$.
- $\mathcal{C}_\lambda$?

More on Multivariate Gaussians

What are the level curves of the Gaussian p.d.f.?

$$\mathcal{C}_\lambda = \{\boldsymbol{x} | \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \lambda\}.$$

- Empty set for $\lambda < 0$.
- $\mathcal{C}_0 = \{\boldsymbol{\mu}\}$.
- $\mathcal{C}_\lambda$? An ellipsoid with center $\boldsymbol{\mu}$ with axis given by the columns of $\boldsymbol{U}$ and axis length given by the elements in $\boldsymbol{\Lambda}$, where $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$.



ML for multivariate Gaussians

Exercise 1.7: Prove that the ML estimators of the multivariate Gaussian are:

$$\boldsymbol{\mu}^* = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n \qquad \boldsymbol{\Sigma}^* = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu}^*)(\boldsymbol{x}_n - \boldsymbol{\mu}^*)^\top.$$

ML for multivariate Gaussians

Exercise 1.7: Prove that the ML estimators of the multivariate Gaussian are:

$$\boldsymbol{\mu}^* = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n \qquad \boldsymbol{\Sigma}^* = \frac{1}{N}\sum_{n=1}^{N} (\boldsymbol{x}_n - \boldsymbol{\mu}^*)(\boldsymbol{x}_n - \boldsymbol{\mu}^*)^\top.$$

You will need to take derivatives w.r.t. matrices. Let $\boldsymbol{M}$ be a matrix, and $f(\boldsymbol{M})$ and function of that matrix (e.g. trace, ...). One can consider $\frac{\partial f}{\partial \boldsymbol{M}}$.

Examples of matrix derivative formulae useful to derive the ML estimate of multivariate Gaussians:

$$\frac{\partial \mathrm{Tr}(\boldsymbol{MA})}{\partial \boldsymbol{M}} = \boldsymbol{A}^\top \qquad \frac{\partial \mathrm{Tr}(\boldsymbol{B}^\top \boldsymbol{M}^\top \boldsymbol{CMB})}{\partial \boldsymbol{M}} = \boldsymbol{C}^\top \boldsymbol{MBB}^\top + \boldsymbol{CMBB}^\top$$

$$\frac{\partial \log |\boldsymbol{M}|}{\partial \boldsymbol{M}} = (\boldsymbol{M}^{-1})^\top \qquad [\mathrm{Tr}(\boldsymbol{ABC}) = \mathrm{Tr}(\boldsymbol{BCA}) = \mathrm{Tr}(\boldsymbol{CAB})]$$

Gaussian Completion: Shape is All You Need

Remark 1.5: Developing multivariate Gaussian distribution, we observe that only two terms depend on **x** (quadratic and linear):

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \overset{\mathbf{x}}{\propto} \exp\Big(-\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\Big).$$

($\overset{\mathbf{x}}{\propto}$ means that is proportional up to a constant that does NOT depend on **x**)

Gaussian Completion: Shape is All You Need

Remark 1.5: Developing multivariate Gaussian distribution, we observe that only two terms depend on **x** (quadratic and linear):

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \overset{\mathbf{x}}{\propto} \exp\Big(-\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\Big).$$

($\overset{\mathbf{x}}{\propto}$ means that is proportional up to a constant that does NOT depend on **x**)

Exercise 1.8: Prove that given a s.p.d. matrix $\mathbf{\Omega}$ and a vector **m**:

$$p(\mathbf{x}) \overset{\mathbf{x}}{\propto} \exp\Big(-\frac{1}{2}\mathbf{x}^\top \mathbf{\Omega}\mathbf{x} + \mathbf{x}^\top \mathbf{m}\Big) \Rightarrow p(\mathbf{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

with:

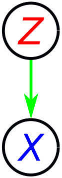
$$\boldsymbol{\Sigma} = \mathbf{\Omega}^{-1} \qquad\qquad \boldsymbol{\mu} = \boldsymbol{\Sigma}\mathbf{m} = \mathbf{\Omega}^{-1}\mathbf{m}.$$

More on multivariate Gaussians in Chapter 4.

Latent Variables and Conditional Independence

What is a model?

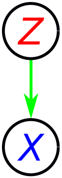
What does it mean to *model* the relationship between two variables?



- *We choose* the nature of $\boldsymbol{z}$ & $\boldsymbol{x}$: cont./discrete, bounded, ...
- *We choose* the dependencies, i.e. $p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})$.
- *We choose* the prior distribution $p(\boldsymbol{z})$.
- *We choose* the likelihood distribution $p(\boldsymbol{x}|\boldsymbol{z})$.

What is a model?

What does it mean to *model* the relationship between two variables?

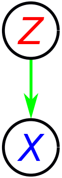


- *We choose* the nature of $\boldsymbol{z}$ & $\boldsymbol{x}$: cont./discrete, bounded, ...
- *We choose* the dependencies, i.e. $p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})$.
- *We choose* the prior distribution $p(\boldsymbol{z})$.
- *We choose* the likelihood distribution $p(\boldsymbol{x}|\boldsymbol{z})$.

Remark 1.6: There is an important difference between **observed** and **latent** or hidden variables. Observed variables are measured, and latent variables are quantities that cannot be measured directly.

What is a model?

What does it mean to *model* the relationship between two variables?



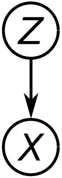
- *We choose* the nature of $\boldsymbol{z}$ & $\boldsymbol{x}$: cont./discrete, bounded, ...
- *We choose* the dependencies, i.e. $p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})$.
- *We choose* the prior distribution $p(\boldsymbol{z})$.
- *We choose* the likelihood distribution $p(\boldsymbol{x}|\boldsymbol{z})$.

Remark 1.7: In models with latent variables, we study the **marginal distribution** of **x** (left) and the **posterior distribution** of **z** given **x** (right):

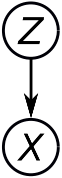
$$p(\mathbf{x}) = \int_{\mathcal{Z}} p(\mathbf{x}|\mathbf{z})p(\mathbf{z})\mathrm{d}\mathbf{z} \qquad p(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{p(\mathbf{x})}$$

Example: Gaussian mixture model

- The nature: z is discrete & bounded, x is 1D & continuous.
- The dependencies: $p(x, z) = p(x|z)p(z)$.



Example: Gaussian mixture model

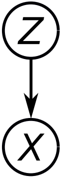


- The nature: z is discrete & bounded, x is 1D & continuous.
- The dependencies: $p(x, z) = p(x|z)p(z)$.
- The distribution $p(z)$, $z \in \{1, \ldots, K\}$ is categorical:

$$p(z = k) = \pi_k, \qquad \pi_k \geq 0, \ \sum_{k=1}^{K} \pi_k = 1.$$

Example: Gaussian mixture model

- The nature: z is discrete & bounded, x is 1D & continuous.
- The dependencies: $p(x, z) = p(x|z)p(z)$.
- The distribution $p(z)$, $z \in \{1, \dots, K\}$ is categorical:



$$p(z = k) = \pi_k, \qquad \pi_k \geq 0, \; \sum_{k=1}^{K} \pi_k = 1.$$

- The distribution $p(x|z)$ is Gaussian:

$$p(x|z = k) = \mathcal{N}(x; \mu_k, \nu_k) = \frac{1}{\sqrt{2\pi\nu_k}} \exp\Big(- \frac{(x - \mu_k)^2}{2\nu_k} \Big)$$

with $\mu_k \in \mathbb{R}$ and $\nu_k > 0, \forall k$.

Likelihood and GMM posteriors

Exercise 1.9: Prove that the GMM marginal writes:

$$p(x) = \sum_{k=1}^{K} \pi_k \frac{1}{\sqrt{2\pi\nu_k}} \exp\Big(- \frac{(x-\mu_k)^2}{2\nu_k} \Big).$$

Exercise 1.10: Prove the GMM posterior writes:

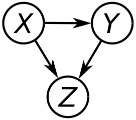
$$p(z=k|x) = \frac{\pi_k \frac{1}{\sqrt{2\pi\nu_k}} \exp\Big(- \frac{(x-\mu_k)^2}{2\nu_k} \Big)}{\sum_{m=1}^{K} \pi_m \frac{1}{\sqrt{2\pi\nu_m}} \exp\Big(- \frac{(x-\mu_m)^2}{2\nu_m} \Big)}.$$

Hint: Just write down what things are.

More on this on the Chapter 2.

Latent Variables and Conditional Independence
—
Conditional Independence

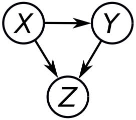
3-variable models: taxonomy



Full all dependencies are set:

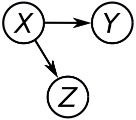
$$p(x, y, z) = p(z|x, y)p(y|x)p(x)$$

3-variable models: taxonomy



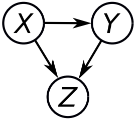
Full all dependencies are set:

$$p(x, y, z) = p(z|x, y)p(y|x)p(x)$$



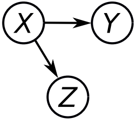
Two-kids *Y*-*Z* dependency missing:

$$p(x, y, z) = p(z|x, y)p(y|x)p(x)$$



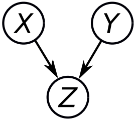
Full all dependencies are set:

$$p(x, y, z) = p(z|x, y)p(y|x)p(x)$$



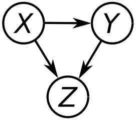
Two-kids *Y*-*Z* dependency missing:

$$p(x, y, z) = p(z|x, y)p(y|x)p(x)$$



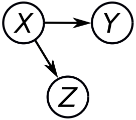
Two-parents *X*-*Y* dependency missing:

$$p(x, y, z) = p(z|x, y)p(y|x)p(x)$$



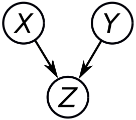
Full all dependencies are set:

$$p(x, y, z) = p(z|x, y)p(y|x)p(x)$$



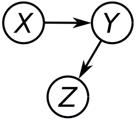
Two-kids *Y*-*Z* dependency missing:

$$p(x, y, z) = p(z|x, y)p(y|x)p(x)$$



Two-parents *X*-*Y* dependency missing:

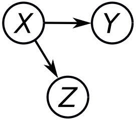
$$p(x, y, z) = p(z|x, y)p(y|x)p(x)$$



Cascaded *X*-*Z* dependency missing:

$$p(x, y, z) = p(z|x, y)p(y|x)p(x)$$

3-variable models: Two-kids

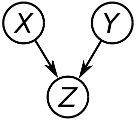


Two-kids $p(x, y, z) = p(z|x)p(y|x)p(x).$

Exercise 1.11: Prove that in the Two-kids model:

$$p(y|z) \neq p(y) \qquad \text{and} \qquad p(y|z, x) = p(y|x)$$

- The first statement is equivalent to say that *y* and *z* are not independent.
- The second statement, says that *y* and *z* are **conditionally independent** w.r.t. *x*.



Two-parents $p(x,y,z) = p(z|x,y)p(x)p(y).$

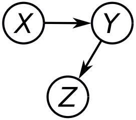
Exercise 1.12: Prove that in the Two-parents model:

$$p(y|x) = p(x) \qquad \text{and} \qquad p(x|y,z) \neq p(x|z). \tag{1}$$

We are in the opposite case:

- The first statement says that y and x are independent.
- The second statement says that y and x are conditionally dependent w.r.t. z.

3-variable models: Cascaded

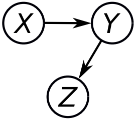


Cascaded $p(x, y, z) = p(z|y)p(y|x)p(x).$

Exercise 1.13: Prove that in the Cascaded model:

$$p(x, z) \neq p(x)p(z) \qquad \text{and} \qquad p(x, z|y) = p(x|y)p(z|y). \tag{2}$$

In this case, we obtain similar results than with the Two-kids model.



Cascaded $p(x,y,z) = p(z|y)p(y|x)p(x).$

Exercise 1.13: Prove that in the Cascaded model:

$$p(x,z) \neq p(x)p(z) \qquad \text{and} \qquad p(x,z|y) = p(x|y)p(z|y). \qquad (2)$$

In this case, we obtain similar results than with the Two-kids model.

Remark 1.8: At this point it should be clear that independence and conditional independence are **two very different properties** of random variables.

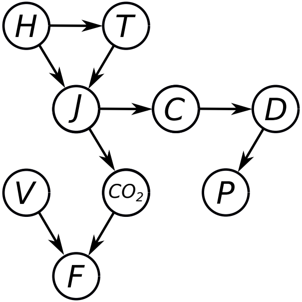
Conditional Independence: Definition

Remark 1.9: Let x, y, and z be random variables, we say that x and y are **conditionally independent** given z, and write $x \perp\!\!\!\perp y \mid z$, iff one of the following equivalent expressions holds:

- $p(x, y|z) = p(x|z)p(y|z)$
- $p(x|y, z) = p(x|z)$
- $p(y|x, z) = p(y|z)$

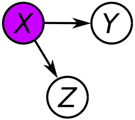
Latent Variables and Conditional Independence — D-separation

Let's consider the following variables and dependencies.



Is $P \perp\!\!\!\perp V \mid T$? How would you do it ? Is the previous strategy scalable ?

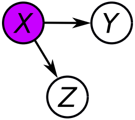
Let us recall the 3-var models:



Two kids The path from Z to Y is called "tail-to-tail."

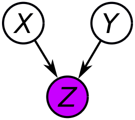
$$p(z, y|x) = p(z|x)p(y|x)$$

Let us recall the 3-var models:



Two kids The path from Z to Y is called "tail-to-tail."

$$p(z, y|x) = p(z|x)p(y|x)$$

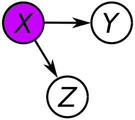


Two parents The path from X to Y is called "head-to-head."

$$p(x, y|z) \neq p(x|z)p(y|z)$$

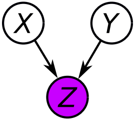
Basics

Let us recall the 3-var models:



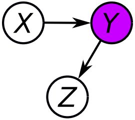
Two kids The path from Z to Y is called "tail-to-tail."

$$p(z, y|x) = p(z|x)p(y|x)$$



Two parents The path from X to Y is called "head-to-head."

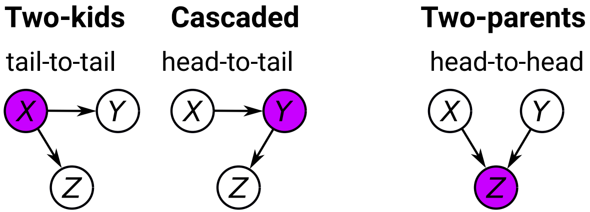
$$p(x, y|z) \neq p(x|z)p(y|z)$$



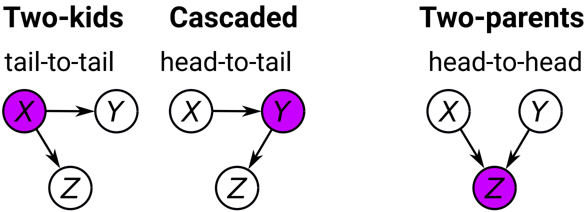
Cascaded The path from X to Z is called "head-to-tail."

$$p(x, z|y) = p(x|y)p(z|y)$$

Path blocking

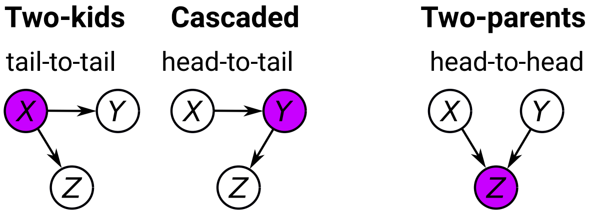


Path blocking



The purple node **“blocks” the path** in two-kids/tail-to-tail & cascaded/head-to-tail $\rightarrow$ conditional independence.

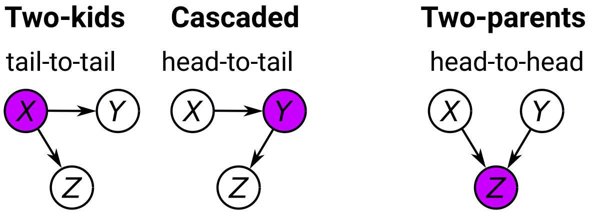
Path blocking



The purple node **"blocks" the path** in two-kids/tail-to-tail & cascaded/head-to-tail $\rightarrow$ conditional independence.

The purple node **"unblocks" the path** in two-parents/head-to-head $\rightarrow$ conditional dependence.

Path blocking

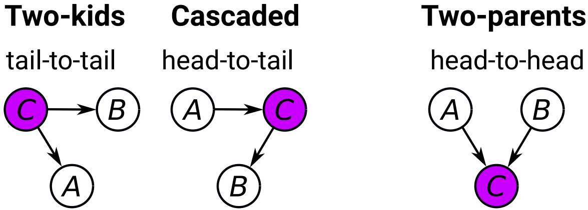


The purple node **"blocks" the path** in two-kids/tail-to-tail & cascaded/head-to-tail $\rightarrow$ conditional independence.

The purple node **"unblocks" the path** in two-parents/head-to-head $\rightarrow$ conditional dependence.
In the two-parents, *Z* or any descendant of *Z* will unblock the path.

Path blocking (revisited)

(Let me change the variable names)

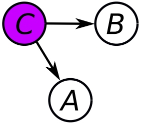


Path blocking (revisited)

(Let me change the variable names)

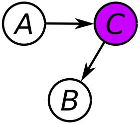
Two-kids

tail-to-tail



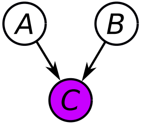
Cascaded

head-to-tail



Two-parents

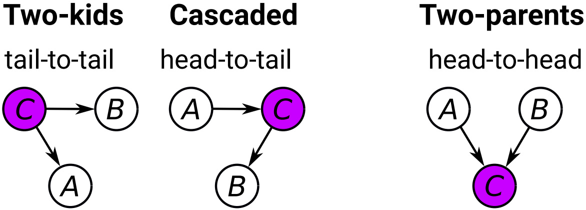
head-to-head



Tail-to-tail & head-to-tail $\rightarrow A \perp\!\!\!\perp B \mid C$.

Path blocking (revisited)

(Let me change the variable names)

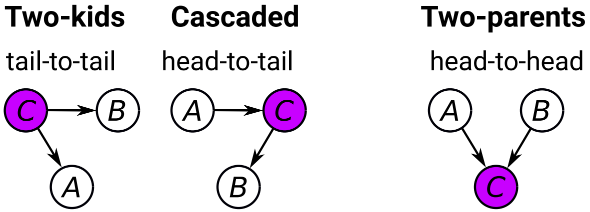


Tail-to-tail & head-to-tail $\rightarrow A \perp\!\!\!\perp B \mid C$.

Head-to-head $\rightarrow A \not\perp\!\!\!\perp B \mid C$ or any descendant of C.

Path blocking (revisited)

(Let me change the variable names)



Tail-to-tail & head-to-tail $\rightarrow A \perp\!\!\!\perp B \mid C$.

Head-to-head $\rightarrow A \not\perp\!\!\!\perp B \mid C$ or any descendant of C.

$\Rightarrow$ If $A \perp\!\!\!\perp B \mid C$, nodes within tail-to-tail or head-to-tail can be in C and nodes within head-to-head or any of their descendents must not be in C.

Path blocking (definition)

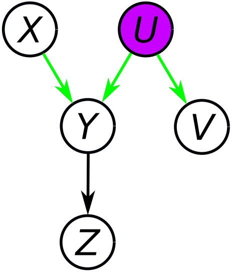
Remark 1.10: Let *A*, *B* and *C* be three non-intersecting sets of nodes of a directed acyclic graph. A path from *A* to *B* is said to be blocked by *C* if it includes a node that either:

- the path meets tail-to-tail or head-to-tail at the node and the node is in C;
- the path meets head-to-head at the node and neither the node nor any of its descendants are in C.

Path blocking (definition)

Remark 1.10: Let A, B and C be three non-intersecting sets of nodes of a directed acyclic graph. A path from A to B is said to be blocked by C if it includes a node that either:

- the path meets tail-to-tail or head-to-tail at the node and the node is in C;
- the path meets head-to-head at the node and neither the node nor any of its descendants are in C.



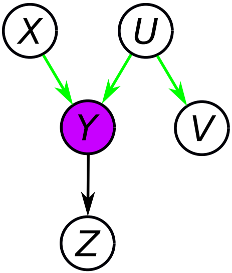
(Corresponds to Exercise: 1.14.)

- Is the path from $\{x\}$ to $\{v\}$ blocked by $\{u\}$?

Path blocking (definition)

Remark 1.10: Let A, B and C be three non-intersecting sets of nodes of a directed acyclic graph. A path from A to B is said to be blocked by C if it includes a node that either:

- the path meets tail-to-tail or head-to-tail at the node and the node is in C;
- the path meets head-to-head at the node and neither the node nor any of its descendants are in C.



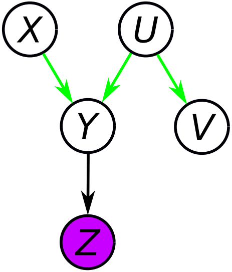
(Corresponds to Exercise: 1.14.)

- Is the path from $\{x\}$ to $\{v\}$ blocked by $\{u\}$? Yes
- Is the path from $\{x\}$ to $\{v\}$ blocked by $\{y\}$?

Path blocking (definition)

Remark 1.10: Let A, B and C be three non-intersecting sets of nodes of a directed acyclic graph. A path from A to B is said to be blocked by C if it includes a node that either:

- the path meets tail-to-tail or head-to-tail at the node and the node is in C;
- the path meets head-to-head at the node and neither the node nor any of its descendants are in C.



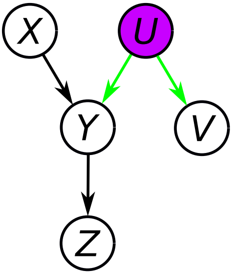
(Corresponds to Exercise: 1.14.)

- Is the path from $\{x\}$ to $\{v\}$ blocked by $\{u\}$? Yes
- Is the path from $\{x\}$ to $\{v\}$ blocked by $\{y\}$? No
- Is the path from $\{x\}$ to $\{v\}$ blocked by $\{z\}$?

Path blocking (definition)

Remark 1.10: Let A, B and C be three non-intersecting sets of nodes of a directed acyclic graph. A path from A to B is said to be blocked by C if it includes a node that either:

- the path meets tail-to-tail or head-to-tail at the node and the node is in C;
- the path meets head-to-head at the node and neither the node nor any of its descendants are in C.



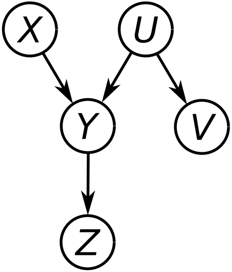
(Corresponds to Exercise: 1.14.)

- Is the path from $\{x\}$ to $\{v\}$ blocked by $\{u\}$? Yes
- Is the path from $\{x\}$ to $\{v\}$ blocked by $\{y\}$? No
- Is the path from $\{x\}$ to $\{v\}$ blocked by $\{z\}$? No
- Is the path from $\{y\}$ to $\{v\}$ blocked by $\{u\}$?

Path blocking (definition)

Remark 1.10: Let A, B and C be three non-intersecting sets of nodes of a directed acyclic graph. A path from A to B is said to be blocked by C if it includes a node that either:

- the path meets tail-to-tail or head-to-tail at the node and the node is in C;
- the path meets head-to-head at the node and neither the node nor any of its descendants are in C.



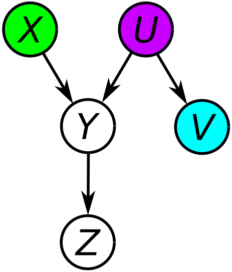
(Corresponds to Exercise: 1.14.)

- Is the path from $\{x\}$ to $\{v\}$ blocked by $\{u\}$? Yes
- Is the path from $\{x\}$ to $\{v\}$ blocked by $\{y\}$? No
- Is the path from $\{x\}$ to $\{v\}$ blocked by $\{z\}$? No
- Is the path from $\{y\}$ to $\{v\}$ blocked by $\{u\}$? Yes

D-separation

Remark 1.11: Let *A*, *B* and *C* be three non-intersecting sets of nodes of a directed acyclic graph. *A* and *B* are **D-separated** by *C*, if all paths from any node from *A* to *B* are blocked by *C*.

Remark 1.11: Let A, B and C be three non-intersecting sets of nodes of a directed acyclic graph. A and B are **D-separated** by C, if all paths from any node from A to B are blocked by C.



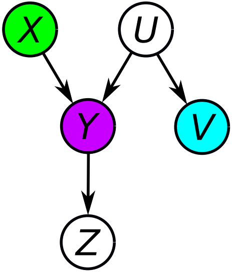
(Corresponds to Exercise: 1.15.)

- Is $\{X\}$ D-separated from $\{V\}$ by $\{U\}$?

Remark 1.12: A and B are D-separated by C if and only if $A \perp\!\!\!\perp B | C$.

D-separation

Remark 1.11: Let A, B and C be three non-intersecting sets of nodes of a directed acyclic graph. A and B are **D-separated** by C, if all paths from any node from A to B are blocked by C.



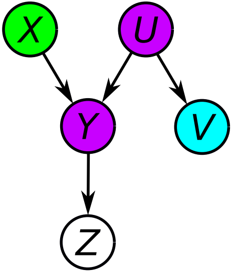
(Corresponds to Exercise: 1.15.)

- Is $\{X\}$ D-separated from $\{V\}$ by $\{U\}$? Yes
- Is $\{X\}$ D-separated from $\{V\}$ by $\{Y\}$?

Remark 1.12: A and B are D-separated by C if and only if $A \perp\!\!\!\perp B | C$.

D-separation

Remark 1.11: Let A, B and C be three non-intersecting sets of nodes of a directed acyclic graph. A and B are **D-separated** by C, if all paths from any node from A to B are blocked by C.

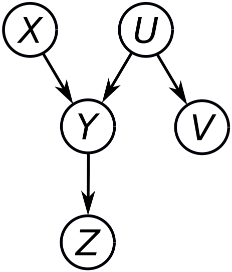


(Corresponds to Exercise: 1.15.)

- Is $\{X\}$ D-separated from $\{V\}$ by $\{U\}$? Yes
- Is $\{X\}$ D-separated from $\{V\}$ by $\{Y\}$? No
- Is $\{X\}$ D-separated from $\{V\}$ by $\{Y, U\}$?

Remark 1.12: A and B are D-separated by C if and only if $A \perp\!\!\!\perp B | C$.

Remark 1.11: Let A, B and C be three non-intersecting sets of nodes of a directed acyclic graph. A and B are **D-separated** by C, if all paths from any node from A to B are blocked by C.



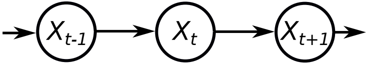
(Corresponds to Exercise: 1.15.)

- Is $\{X\}$ D-separated from $\{V\}$ by $\{U\}$? Yes
- Is $\{X\}$ D-separated from $\{V\}$ by $\{Y\}$? No
- Is $\{X\}$ D-separated from $\{V\}$ by $\{Y, U\}$? Yes

Remark 1.12: A and B are D-separated by C if and only if $A \perp\!\!\!\perp B | C$.

Latent Variables and Conditional Independence
—
Markovian dependencies

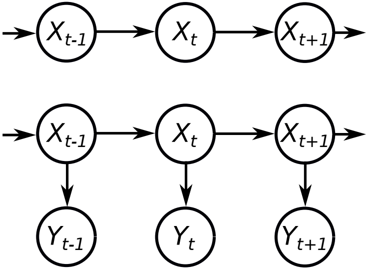
Principle: each variable depends **only** on its closer neighbours.
Examples:



Markov chain.

Markov models: introduction

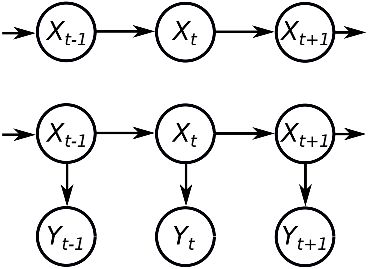
Principle: each variable depends **only** on its closer neighbours.
Examples:



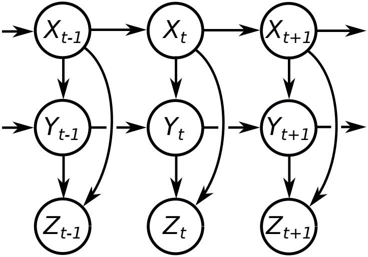
Markov chain (top).
Hidden Markov chain (bottom).

Markov models: introduction

Principle: each variable depends **only** on its closer neighbours.
Examples:



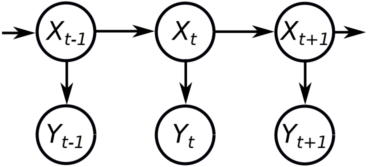
Markov chain (top).
Hidden Markov chain (bottom).



Double hidden Markov chain.

D-separation in Markov models

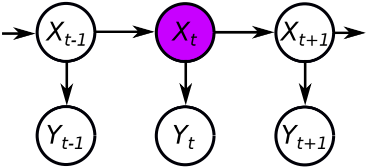
With the following model:



Is $\{X_{t-1}\}$ D-separated from $\{Y_{t+1}\}$ by ...

D-separation in Markov models

With the following model:

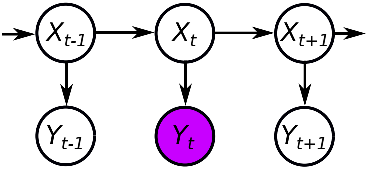


Is $\{X_{t-1}\}$ D-separated from $\{Y_{t+1}\}$ by ...

- $\{X_t\}$? [1 minute]

D-separation in Markov models

With the following model:

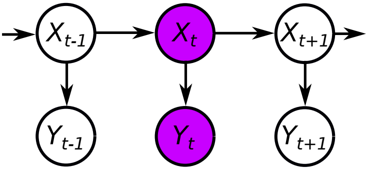


Is $\{X_{t-1}\}$ D-separated from $\{Y_{t+1}\}$ by ...

- $\{X_t\}$? Yes
- $\{Y_t\}$? [1 minute]

D-separation in Markov models

With the following model:

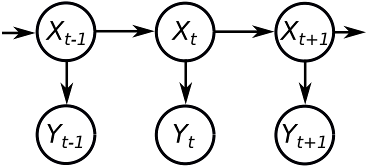


Is $\{X_{t-1}\}$ D-separated from $\{Y_{t+1}\}$ by ...

- $\{X_t\}$? Yes
- $\{Y_t\}$? No
- $\{X_t, Y_t\}$? [1 minute]

D-separation in Markov models

With the following model:

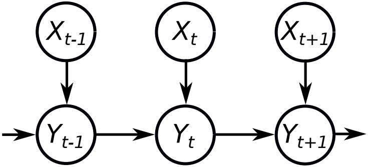


Is $\{X_{t-1}\}$ D-separated from $\{Y_{t+1}\}$ by ...

- $\{X_t\}$? Yes
- $\{Y_t\}$? No
- $\{X_t, Y_t\}$? Yes

D-separation in Markov models (II)

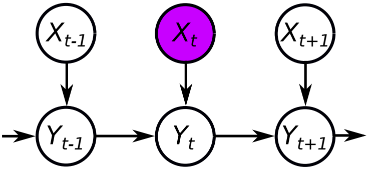
With the following model:



Is $\{X_{t-1}\}$ D-separated from $\{Y_{t+1}\}$ by ...

D-separation in Markov models (II)

With the following model:

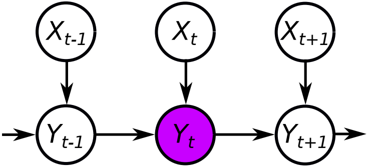


Is $\{X_{t-1}\}$ D-separated from $\{Y_{t+1}\}$ by ...

- $\{X_t\}$? [1 minute]

D-separation in Markov models (II)

With the following model:

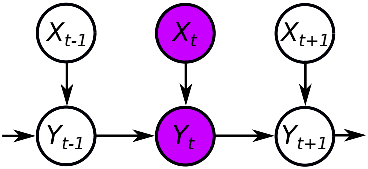


Is $\{X_{t-1}\}$ D-separated from $\{Y_{t+1}\}$ by ...

- $\{X_t\}$? No
- $\{Y_t\}$? [1 minute]

D-separation in Markov models (II)

With the following model:

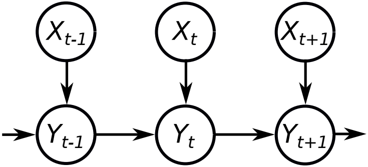


Is $\{X_{t-1}\}$ D-separated from $\{Y_{t+1}\}$ by ...

- $\{X_t\}$? No
- $\{Y_t\}$? Yes
- $\{X_t, Y_t\}$? [1 minute]

D-separation in Markov models (II)

With the following model:

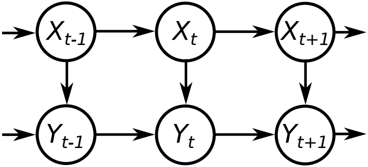


Is $\{X_{t-1}\}$ D-separated from $\{Y_{t+1}\}$ by ...

- $\{X_t\}$? No
- $\{Y_t\}$? Yes
- $\{X_t, Y_t\}$? Yes

D-separation in Markov models (III)

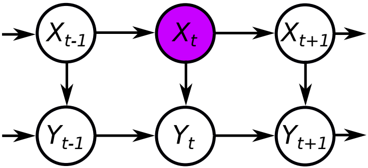
With the following model:



Is $\{X_{t-1}\}$ D-separated from $\{Y_{t+1}\}$ by ...

D-separation in Markov models (III)

With the following model:

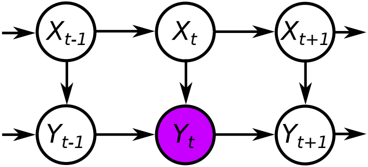


Is $\{X_{t-1}\}$ D-separated from $\{Y_{t+1}\}$ by ...

- $\{X_t\}$? [1 minute]

D-separation in Markov models (III)

With the following model:

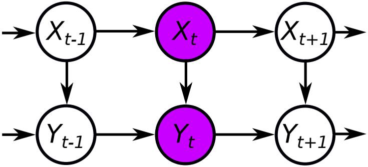


Is $\{X_{t-1}\}$ D-separated from $\{Y_{t+1}\}$ by ...

- $\{X_t\}$? No
- $\{Y_t\}$? [1 minute]

D-separation in Markov models (III)

With the following model:

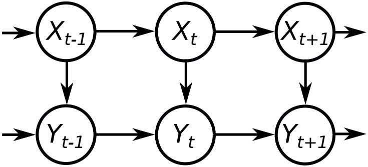


Is $\{X_{t-1}\}$ D-separated from $\{Y_{t+1}\}$ by ...

- $\{X_t\}$? No
- $\{Y_t\}$? No
- $\{X_t, Y_t\}$? [1 minute]

D-separation in Markov models (III)

With the following model:

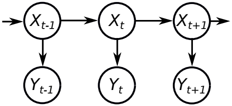


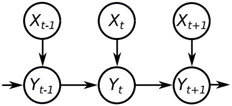
Is $\{X_{t-1}\}$ D-separated from $\{Y_{t+1}\}$ by ...

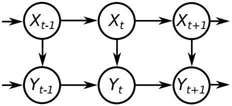
- $\{X_t\}$? No
- $\{Y_t\}$? No
- $\{X_t, Y_t\}$? Yes

D-separation in Markov models: summary

Is $\{X_{t-1}\}$ D-separated from $\{Y_{t+1}\}$ by (left column) in (top row)?





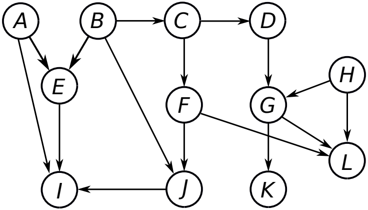


$\{X_t\}$	Yes	No	No
$\{Y_t\}$	No	Yes	No
$\{X_t, Y_t\}$	Yes	Yes	Yes

(Corresponds to Exercise 1.16.)

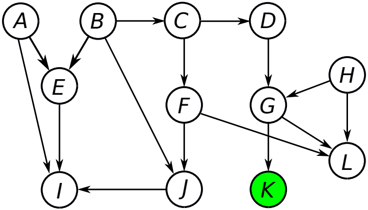
Markov blanket (or boundary)

Model example:



Markov blanket (or boundary)

Model example:



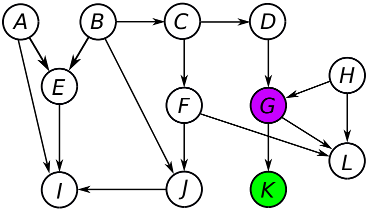
For a given node K, what is the minimal set of variables $\mathcal{B}_K$ so that:

$$p(K | \text{all except } K) = p(K | \mathcal{B}_K)?$$

You've got 3 minutes!

Markov blanket (or boundary)

Model example:

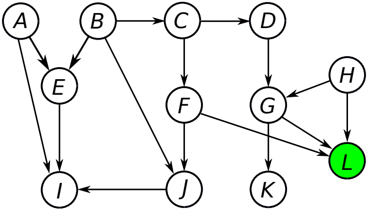


For a given node K, what is the minimal set of variables $\mathcal{B}_K$ so that:

$$p(K|\text{all except } K) = p(K|\mathcal{B}_K)? \qquad \mathcal{B}_K = \{G\}$$

Markov blanket (or boundary)

Model example:



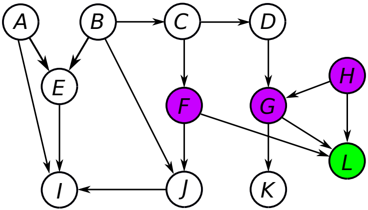
For a given node K, what is the minimal set of variables $\mathcal{B}_K$ so that:

$$p(K|\text{all except } K) = p(K|\mathcal{B}_K)? \qquad \mathcal{B}_K = \{G\}$$

For L? You've got 3 minutes!

Markov blanket (or boundary)

Model example:



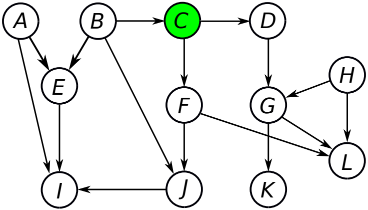
For a given node K, what is the minimal set of variables $\mathcal{B}_K$ so that:

$$p(K|\text{all except } K) = p(K|\mathcal{B}_K)? \qquad \mathcal{B}_K = \{G\}$$

For L? $\quad \mathcal{B}_L = \{F, G, H\}$ because F, G, H are **parents** of L.

Markov blanket (or boundary)

Model example:



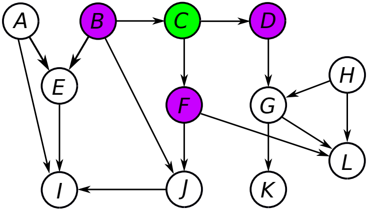
For a given node K, what is the minimal set of variables $\mathcal{B}_K$ so that:

$$p(K|\text{all except } K) = p(K|\mathcal{B}_K)? \qquad \mathcal{B}_K = \{G\}$$

For L? $\mathcal{B}_L = \{F, G, H\}$ because F, G, H are **parents** of L.
For C? You've got 3 minutes!

Markov blanket (or boundary)

Model example:



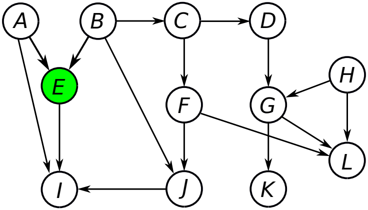
For a given node K, what is the minimal set of variables $\mathcal{B}_K$ so that:

$$p(K|\text{all except } K) = p(K|\mathcal{B}_K)? \qquad \mathcal{B}_K = \{G\}$$

For L? $\quad \mathcal{B}_L = \{F, G, H\}$ because F, G, H are **parents** of L.
For C? $\quad \mathcal{B}_C = \{B, D, F\}$ because B (F, D) is parent (**children**) of C.

Markov blanket (or boundary)

Model example:



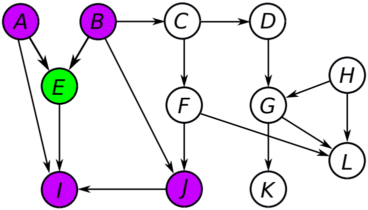
For a given node K, what is the minimal set of variables $\mathcal{B}_K$ so that:

$$p(K|\text{all except } K) = p(K|\mathcal{B}_K)? \qquad \mathcal{B}_K = \{G\}$$

For L? $\mathcal{B}_L = \{F, G, H\}$ because F, G, H are **parents** of L.
For C? $\mathcal{B}_C = \{B, D, F\}$ because B (F, D) is parent (**children**) of C.
For E? You've got 3 minutes!

Markov blanket (or boundary)

Model example:



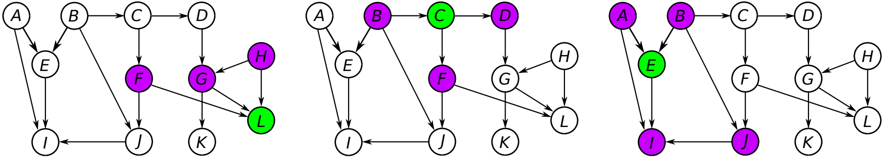
For a given node K, what is the minimal set of variables $\mathcal{B}_K$ so that:

$$p(K|\text{all except } K) = p(K|\mathcal{B}_K)? \qquad \mathcal{B}_K = \{G\}$$

For L? $\quad \mathcal{B}_L = \{F, G, H\}$ because F, G, H are **parents** of L.
For C? $\quad \mathcal{B}_C = \{B, D, F\}$ because B (F, D) is parent (**children**) of C.
For E? $\quad \mathcal{B}_E = \{A, B, I, J\}$ because A, B (I) are parents (children) of E and J is **co-parent** of E.

Markov blanket: definition

Remark 1.13: The **Markov blanket** is the minimal set that D-separates a set of nodes from the rest of the graph.



Remark 1.14: Construction of the Markov blanket. Given a directed acyclic graph, and a node x on that graph, the Markov blanket of x, $\mathcal{B}_x$ is the set of all parents, children and co-parents of x.