An Introductory Tutorial on Graphical Probabilistic Models: Notes

ISPRS International Summer School, Advances in Medical Imaging

24–29 April, 2006, Aghios Nikolaos, Crete, Greece

Evangelos Roussos
Program in Applied and Computational Mathematics
Princeton University
eroussos@math.princeton.edu

1 Introduction

When modelling complex systems we are unavoidably faced with *imperfect or missing information*, especially in the measurement and information sciences. This may have several causes, but it is mainly due to

- Cost of obtaining and processing vast amounts of information,
- Inherent system complexity.

Probability theory is a conceptual and computational framework for reasoning under uncertainty.

These notes discuss several issues of modelling using probabilities and tools from graph theory. Probability theory acts as "glue" for linking different models together. Graphical models are structured representations of systems. There is a variety of formulations, each conveying different semantic aspects. More on this after a short review of probability theory.

2 A Short Review of Probability Theory

- Probabilities: uncertainty regarding occurrence of (random) events
  - Cox's theorem [**?**]: probability is the only consistent, universal logic framework for quantitatively reasoning under uncertainty,
  - Probability theory as extended logic (Jaynes [**?**]).

2.1 Probability space

- Probability space (Ω, P): describes our idea about uncertainty wrt a random experiment:
  - Sample space Ω of possible outcomes ω_i, and
  - A probability measure P: how likely an outcome is.
- $\mathcal{A} \in \sigma(\Omega)$ is a collection of subsets of Ω: for $A \in \mathcal{A}$:
  - $P(A) \geq 0$ and $P(\Omega) = 1$
  - Additive: for two disjoint events A, B, $P(A \cap B) = P(A) + P(B)$.
- Conditional probability: *"probability within a probability"*

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad P(B) \neq 0$$

- Random variables: functions from Ω to a range $\mathcal{R}$ (e.g. $\mathbb{R}$ or $\mathbb{N}$, etc).

  Can inversely define events:

$$\mathcal{R} \rightarrow \Omega : A(x) = \{\omega \in \Omega | P[x(\omega)]\},$$

where P is a predicate (e.g. '$x > 2$'), and therefore act as "filters" of certain experimental outcomes.

- Probability densities: are densities (of probability measures):

$$p(x) = \frac{\mathrm{d}}{\mathrm{d}x} P(A(x))|_x, \quad A(x) = \{x' \in [x, x + \mathrm{d}x]\}, \quad x \in \mathbb{R}.$$

- Joint densities: $p_{XY}(xy) = p(\{\omega : X(\omega) = x \wedge Y(\omega) = y\})$

2.2 Three simple rules:

1. Product rule: $P(A \cap B) = P(A|B)P(B)$. Generalise for N events: chain rule $p(\cap_{i=1}^{N} A_i) = \prod_{i'=1}^{i-1} P(A_i| \cap_{i'=1}^{i} A_{i'})$, $(i' < i)$ (telescopic). Note: important for reasoning in Bayesian networks.

2. Bayes' rule: *it is a recipe that tells us how to update our knowledge in the presence of new information!*

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \quad P(B) \neq 0$$

   $P(A)$: prior (model!), $P(B)$ data, $P(A|B)$ posterior, $P(B|A)$: likelihood of the data under the model. Can be simply derived from the rule of conditional probability.

3. Marginalisation: given a joint density $p_{xy}(xy)$ get the marginal density of x or y, by summation: ("integrate out" the uncertainty in one variable):

$$p_x(x) = \int_{\{y \in \mathcal{Y}\}} \mathrm{d}y \, p(x, y)$$

This is everything that we need to know in order to perform probabilistic modelling and inference.

3 Probabilistic modelling

- Thinking in terms of *systems* of random variables, their relations, and probabilities on them.
- This basically means working with conditional and/or joint distributions.

3.1 Probability Theory and Graph Theory

Graphical models: model *structural relationships* among random variables.

Independence: Recall: two RVs are independent, write $X \perp\!\!\!\perp Y$, if $p_{X|Y}(x, y) = p_X(x)$.

- Encode probabilistic (in-)dependence relations among random variables.
- Plausible inference requires some degree of "regularity" among the conditional densities of RVs: for example, we need to encode knowledge of the form:
  - Symmetry: "If A is independent of B given C, then B is independent of A given C".
  - "If A is independent of both B and D given C, then A must be independent of B given both C and D.

We need to find a way of formally (and easily) encoding such relations. It turns out that there is graph theoretic structure in sets of random variables, given their probabilistic relations, given by the correspondence:

- Random variable $v \in \mathcal{V} \longleftrightarrow$ vertex $v \in \mathcal{V}(\mathbb{G})$
- Probabilistic dependence $p(v, u), \quad v, u \in \mathcal{V} \longleftrightarrow$ edge $e(v, u) \in \mathcal{E}(\mathbb{G})$,

where $\mathbb{G} = (\mathcal{V}, \mathcal{E})$ is a graph with node–set $\mathcal{V}$ and edge–set $\mathcal{E}$.

Given the semantics of edges (to be defined next), we can perform probabilistic reasoning using graph theoretic concepts!

3.2 Modelling (In–)dependence

Conditional independence: two "aspects" of a system become independent *given* a third part of the system: write $X \perp\!\!\!\perp Y|Z$:

$$X \perp\!\!\!\perp Y|Z \Longleftrightarrow p(X|Z) = P(X|Y \cap Z) \Longleftrightarrow p(X \cap Y|Z) = p(X|Z)p(Y|Z).$$

"Knowledge of (or absence) Y does not influence our knowledge of X".

- Systems *decouple* and joint pdfs *factorise*.
- Conditional (in-)dependence relationships between RVs can be immediately read off the graph, given the graph-theoretic properties of *separation* and *d–separation*.

Definition 1 (Markov property) *(In time): "the future and the past are independent given the present".*

Can be generalised for spatial random variables.

*Conditional independencies lead to efficient inference (*local *computation).*

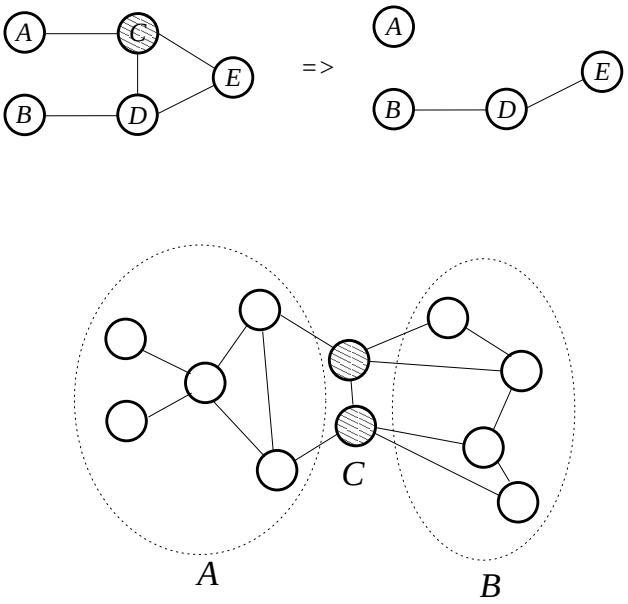

Figure 1: Undirected graphical probabilistic models: (a) Observation of the value of the random variable C leads to the elimination of this node. (b) The subset C of $\mathcal{V}$ is a *vertex cut–set*: if observed, it makes A and B conditionally independent.

3.3 Undirected graphs (Markov random fields)

Markov random fields (MRFs) are *undirected* graphical structures on sets of random variables. They combine ideas from probability theory and statistical physics, especially when describing energies of many–particle systems seen as random quantities. Examples include the Ising and Potts models, Boltzmann Machines, etc. They have wide applications in image processing.

Definition 2 (Clique) *Clique $C \subseteq \mathcal{V}$: a fully connected component of a graph.*

In order to define the joint probability on an undirected graph, we must first define 'interaction potential functions' ψ on cliques $C \in \mathcal{C}$; see next.

Definition 3 (Potential function) *For our purposes, a clique potential, $\psi(\mathbf{x}_C)$, on a clique $C \in \mathcal{C}$, is an arbitrary non–negative function of $\mathbf{x}_C$.*

Rule for Independence: if there is no *path* between A and B, then the RVs A and B are independent.

- When a variable is *observed or known* it is removed from the graph along with all the edges connected to it.

- Broken paths imply conditional independence: If $Z_1, \ldots, Z_K$ are removed from the graph then $X \perp\!\!\!\perp Y | \{Z_1, \ldots, Z_K\}$ holds, i.e. X is *separated* from Y by $Z_1, \ldots, Z_K$.

3.3.1 Markovianity and Factorizability

Definition 4 (Markov property) *Markov property: $\mathcal{X}$ is Markov wrt $\mathbb{G}$ if $\mathcal{X}_A$ and $\mathcal{X}_B$ are conditionally independent given $\mathcal{X}_C$ whenever C separates A and B; see Fig. 3.3.*

Definition 5 (Factorization property) *Factorization property: A distribution p factorizes according to $\mathbb{G}$ if it can be expressed as a product over cliques.*

Theorem 1 (The Hammersley–Clifford Theorem) *For strictly positive $p(\cdot)$, the* Markov property *and the* Factorization property *are equivalent.*

This ensures that a product of positive functions on cliques of $\mathbb{G}$ *is* indeed an MRF relative to $\mathbb{G}$.

Definition 6 (Gibbs–Boltzmann distribution) *A probability Q is a* Gibbs distribution *for a graph $\mathbb{G}$ if it can be written in the form*

$$Q(\mathbf{x}) = \prod_{c \in \mathcal{C}} \psi_c(\mathbf{x}_c).$$

If, in addition, $\psi_c(\mathbf{x}_c) > 0$, $\forall \mathbf{x}_c$, we can set $\phi_c(\mathbf{x}_c) = -\log \psi_c(\mathbf{x}_c)$ and write $p(\mathbf{x})$ in exponential form:

$$p(\mathbf{x}) = \frac{1}{Z} \exp\{-U(\mathbf{x})\} = \frac{1}{Z} \exp\left\{ -\sum_{c \in \mathcal{C}} \phi_c(\mathbf{x}_c) \right\},$$

where $U(\mathbf{x})$ is the 'potential energy' of the configuration $\mathbf{x}$. This is the Boltzmann class of distributions.

Interpretation: *values with lower energies are more probable.*

3.3.2 Interpretation of clique potentials

In general, the individual potentials are 'compatibility' functions, but not probability distributions.

For example, for an MRF x —— y —— z, which implies $x \perp\!\!\!\perp z | y$, and $p(x, y, z) = p(z|y)p(x|y)p(y)$ we can have the factorisations:

$$\begin{aligned} p(x, y, z) &= p(x, y)p(z|y) = \psi(x, y)\psi(y, z) \\ p(x, y, z) &= p(z, y)p(x|y) = \psi(x, y)\psi(y, z) \end{aligned}$$

Therefore, we *cannot* have all potentials be marginals or conditionals.

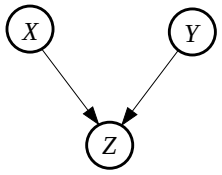

Figure 2: An example of a directed graphical model.

3.3.3 Markov random fields with discrete data

1. Assign a label, ω_i, to each 'site', from a given set of labels $\{1, \ldots, K\}$.
2. Define a probability measure on the set of all possible 'configurations'.
3. Markovian property: $P(f_i|f_{S\setminus\{i\}}) = P(f_i|f_{\mathcal{N}(i)})$.
4. Incorporate *contextual constraints*.

3.4 Directed graphs (Bayesian networks)

- The set $\mathcal{V}$ of random variables is a *partially ordered set* (poset).
- The joint probability of a set of random variables $\mathcal{V}$, $p(\mathcal{V})$, is defined *on* a directed graphical structure $\mathbb{G}$.
- **Semantics:** Edges in a graph represent *'direct influence'*. Note:
  - Bayesian networks are *not* causal networks, but edges in BNs often emanate from causes and terminate at effects.
  - Bayesian networks do not need to be Bayesian only, i.e. use Bayesian inference; frequentist statistics can be used as well.
- Ordering matters!: $A \longrightarrow B \;\; \neq \;\; B \longrightarrow A$

The structure of $\mathbb{G}$ *defines the factorisation* of $p(\mathcal{V})$:

$$p(\mathbb{G}) = \prod_{v \in \mathcal{V}(\mathbb{G})} p(v|\mathrm{pa}(v)).$$

In Bayesian networks there is a simple one–to–one relationship with the factors of the joint distribution of $\mathcal{V}$.

On the graph of Fig. 3.4: $p(\{X, Y, Z\}) = p(Z|X, Y)p(X)p(Y)$.

3.4.1 Blocked Paths and Activated Paths

Generic Rule for Independence: "X is independent of Y *unless* Z is given".

Concepts:

- Marginal independence: X and Y are marginally independent.
- Conditional dependence: given Z, X and Y become *dependent*.

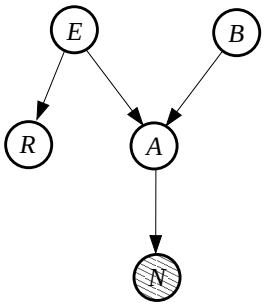

Figure 3: The Bayesian network representing the 'alarm' "knowledge–base". Variables: B: burglary, E: earthquake, A: alarm, J: Neighbour calls, R: Radio anouncement.

Example: Alarm BN [Friedman and Goldszmidt] "I'm at work, my neighbour calls to say my alarm is ringing. Sometimes the alarm is set off by minor earthquakes. Is there a burglary?" [Fig. 3.4.1] .

For the alarm example,

$$p(\mathcal{V}(\mathbb{G})) = p(\{E, B, A, R, N\}) = p(E) \cdot p(B) \cdot p(A|E, B) \cdot p(R|E) \cdot p(N|A).$$

3.4.2 Assessing conditional independence via d–separation

An observation "blocks" a path between nodes: d–separation stands for *directed* separation.

In a graphical model with nodes $X, Y, Z_1, \ldots, Z_k$,

- d–separation helps us determine if $X \perp\!\!\!\perp Y | \{Z_1, \ldots, Z_k\}$ holds
- X and Y are d–separated if there is no *active path* between them.

There are four kinds of 'primitive structures':

Convergent: $A \longrightarrow C \longleftarrow B$: ($A$ and B are the parents of C) Given C, A and B become *dependent*, even if they are *marginally indendent*. Example: $\{\text{sprinkler}, \text{rain}\} \longrightarrow$ wet grass. *A path gets activated in this case.*

Convergent via another node, D: $\{A \longrightarrow D \longleftarrow B, D \longrightarrow C\}$:

Divergent: $A \longleftarrow C \longrightarrow B$: ($A$ and B are the children of C) Given C, A and B become *independent*. *C blocks the path between A and B.*

Chain: $A \longrightarrow C \longrightarrow B$: B is independent of A given C. *C blocks the path between A and B.*

All Bayesian networks are built using these building blocks.

4 Inference

In probabilistic models: we have a model, m, which is a functional that relates a set of variables, $\mathcal{V} = \{v_i\}$, together:

$$m = \mathcal{F}[\{f_k(\sigma_k(\mathcal{V}))\}_k].$$

Some v_i's are observed ("instantiated"): $\mathcal{X} = \{v_j \in \mathcal{V} | v_j = o_j\}$, and the rest, $\mathcal{Y} = \mathcal{V} \setminus \mathcal{X}$, are unobserved ("hiddent", or "latent"). We can also have a set of parameters $\boldsymbol{\theta}$.

Inference (state estimation):

- Given: a set of random variables $\{X_1, \ldots, X_N\}$ and their joint probability, $P(X_1, \ldots, X_N)$, i.e. a *model*,
- Compute: one or more conditional densities given some observations.

Examples:

- Given some measurements (images, range measurements, etc) and some external orientation compute the three-dimensional shape of the human body.
- Given a set of DEMs as a time-series compute the volumetric growth and/or change in shape of an infant.
- Plan and estimate the precise position of an implant before/after surgery.

Notes:

- In graphical models, learning means computing the posterior distribution of some variables given the *'evidence'*, i.e. the observed values of some others.

Back to the 'Alarm' example: Fig. 3.4.1
Learning: posterior $p(EBAR|N)$: using Bayes' rule:

$$\begin{aligned} p(EBAR|N) &= \frac{P(EBARN)}{P(N)} \\ &= \frac{P(N|EBAR) \times P(EBAR)}{P(N)} \\ &= \frac{P(N|A) \times P(R|E)P(A|EB)P(E)P(B)}{\sum_{ebar} P(N|A) \times P(R|E)P(A|EB)P(E)P(B)}. \end{aligned}$$

4.1 Inference in Bayesian Networks

- BNs contain all the information needed for inference
- For the general case, it is NP-hard.

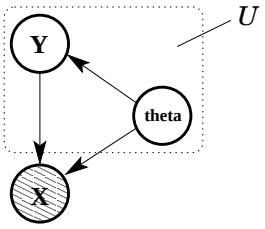

Figure 4: General architecture of directed graphical models (Bayesian networks): Observed $\mathbf{X}$, latent variables $\mathbf{Y}$, and parameters $\boldsymbol{\theta}$.

Several methods:

- Variable Elimination, Sum–Product
- Dynamic Programming
- Gradient Descent on the likelihood or posterior surface
- Stochastic simulation (Markov chain Monte Carlo, Vegas, etc)
- Approximate Inference

5 Expectation-Maximization

Expectation-Maximisation [Dempster, Laird & Rubin, 1977] is a generic strategy for maximum likelihood or maximum a-posteriori estimation on probabilistic models. Consider the generic graphical model shown in Fig. 5, with parameters $\boldsymbol{\theta}$ and observations $\mathbf{Y}$.

- Estimate the parameters $\boldsymbol{\theta}$ (and posterior statistics of the hidden variables $\mathbf{Y}$) in a graphical model.
- Interpretation as inference in the presence of 'incomplete' observations or latent variables.

Recall: Likelihood function of the parameters given the data ('under the model'):

$$L(\boldsymbol{\theta}) = \log[p(\mathbf{X}|\boldsymbol{\theta})].$$

- Directly computing this may be hard for complex models!
- In many cases the joint distribution of $(\mathbf{X}, \mathbf{Y})$ is much easier. This is called the 'complete-likelihood'.

  Write

$$p(\mathbf{X}, \mathbf{Y}|\boldsymbol{\theta}) = p(\mathbf{X}|\mathbf{Y}, \boldsymbol{\theta})p(\mathbf{Y}|\boldsymbol{\theta}). \quad (1)$$

5.1 The EM algorithm

- But we do not have the unobserved data! Let's 'integrate them out'.

E-step: Construct the average complete log likelihood wrt the posterior over the hidden variables, $\mathbf{Y}$:

$$Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^k\right) \stackrel{\text{def}}{=} \mathrm{E}\left[\log p(\mathbf{X}, \mathbf{Y}|\boldsymbol{\theta})\right]_{p_{\mathbf{Y}|\mathbf{X}}} = \int_{\mathbf{Y}} \mathrm{d}\mathbf{Y}\, p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}^{k-1}) \log[p(\mathbf{X}|\mathbf{Y}, \boldsymbol{\theta}) p(\mathbf{Y}|\boldsymbol{\theta})] \tag{2}$$

(Start from an initial value for the parameters, $\boldsymbol{\theta}^0$.)

Note: compute posterior over $\mathbf{Y}$ via Bayes' theorem,

$$p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}^{k-1}) = \frac{p(\mathbf{X}|\mathbf{Y}, \boldsymbol{\theta}^{k-1}) p(\mathbf{Y})}{\sum_{\mathbf{Y}} p(\mathbf{X}|\mathbf{Y}, \boldsymbol{\theta}^{k-1}) p(\mathbf{Y})}.$$

Note: $\boldsymbol{\theta}^{k-1}$ is a *fixed number*, not a RV.

M-step: Next $\boldsymbol{\theta}$, at iteration k, is found by *maximizing* $Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^k\right)$ wrt $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}^k = \arg\max_{\boldsymbol{\theta}} \left\{ Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{k-1}\right) \right\} \tag{3}$$

Algorithm 1 EM Algorithm

1: `let` $k = 0$, $\boldsymbol{\theta} = \boldsymbol{\theta}^0$
2: *E-step:* Compute average from Eq. 2
3: *M-step:* Compute new parameters $\boldsymbol{\theta}^k$ from Eq. 3
4: $k \leftarrow k + 1$;
`IF`(not converged) `GOTO` 2 `ELSE` $\hat{\boldsymbol{\theta}} \leftarrow \boldsymbol{\theta}^k$, `exit ENDIF`

Notes on the EM Algorithm:

- *Guaranteed convergence properties!* [DLR, 1977].
- We need to provide an *explicit model* (prior) for the non-observed variables, $\mathbf{Y}$.
- From the EM algorithm we also get the posterior statistics of $\mathbf{Y}$, besides $\hat{\boldsymbol{\theta}}$.

Generalised EM Original version uses true posterior over $\mathbf{Y}$. We could have used an approximate posterior $q(\mathbf{Y})$. Then $Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^k\right)$ would have been a *lower bound* to the log-likelihood.

6 Markov random fields: Bayesian labelling

- MAP Estimate
$$\mathbf{f}^* = \arg\max_{\mathbf{f}\in\mathbb{F}} p(\mathbf{f}|\mathbf{d})$$
where $\mathbb{F}$ is the space of all possible MRFs, and $\mathbf{d}$ are the data.
- Mode of the posterior distribution.
- From Bayes' theorem,
$$p(\mathbf{f}|\mathbf{d}) = \frac{p(\mathbf{d}|\mathbf{f})p(\mathbf{f})}{p(\mathbf{d})} \propto p(\mathbf{d}|\mathbf{f})p(\mathbf{f})$$
- We have:
  - Prior $p(\mathbf{f})$: MRF energy $U(\mathbf{f})$
  - Likelihood $p(\mathbf{d}|\mathbf{f})$: Noise model
- Posterior energy: $U(\mathbf{f}|\mathbf{d}) + U(\mathbf{f})$
- This is an *Energy Minimization* problem

6.1 Gibbs Sampler and Simulated Annealing

- Naïve calculation would lead to 'combinatorial explosion': we cannot visit all points $\mathbf{x}$ in the configuration space $\Omega = \{1, \ldots, K\}^N$, where N is the number of pixels.
- Stochastic Relaxation

  Randomly visit the configuration set Ω long enough according to the distribution $P_{\mathbf{X}}$.

  Replace each point i with a sample from its conditional distribution:
$$p\left(x_i|\{x_{i'}\}_{i'\neq i}\right) = p\left(x_i|x_{\mathcal{N}(i)}\right)$$
- As $t \to \infty$, the "histogram" (distribution) tends to the "true" ('stationary') distribution $h(\mathbf{x}) \to p(\mathbf{x}|\mathbf{d})$.

6.1.1 Gaussian MRFs

A useful class of MRFs on $\mathbf{f} = [f_i]$, $i = 1, \ldots, N$, are Gaussian MRFs. Their density function is

$$p(\mathbf{f}) = \frac{1}{\sqrt{(2\pi\sigma^2)^N}\sqrt{\det \mathbf{B}}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{f}-\boldsymbol{\mu})^{\mathsf{T}}\mathbf{B}(\mathbf{f}-\boldsymbol{\mu})\right\}.$$

The matrix $\mathbf{B} = [b_{ii'}]$ is the *interaction matrix* (between pixels in an image), with $b_{ii'} = \delta_{ii'} - \beta_{ii'}$ and $\beta_{ii} = 0$.

7 Applications

7.1 Learning Gaussian Mixture Densities from Data

7.2 Image Segmentation