Bayes in AI
4.1 (‘Small’) PGMs and (shallow) NNs

David Ríos Insua & Roi Naveiro

david.rios@icmat.es roi.naveiro@icmat.es

Objectives

Earlier sessions

1. Bayes or die
2. MCMC
3. Large scale Bayes: VB and SGMCMC

Today

4.1 Bayes in AI: (small) PGMs and (shallow) NNs david.rios@icmat.es

4.2 GPs and Bayesian NNs in function spaces simon.rodriguez@icmat.es

4.3 Modern research in ML roi.naveiro@icmat.es

PGMs. Motivation

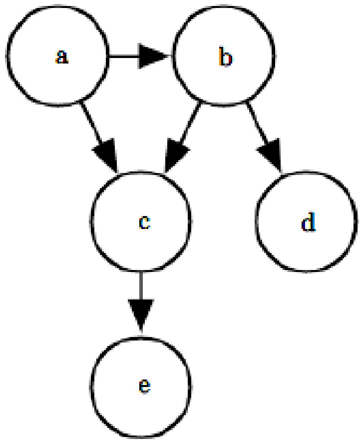
Motivation

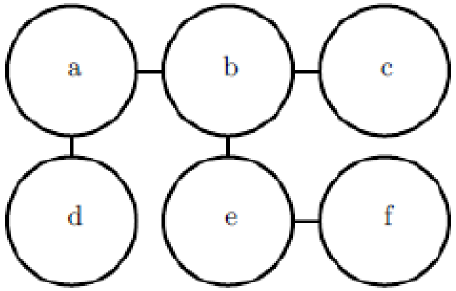
- Simple way to visualize structure of probabilistic models
- Designing and motivating new models
- Understanding properties like conditional independence
- Complex computations viewed through simple graphical manipulations
- Explainable and interpretable
- Deep belief nets in deep learning

Concept

$$p(\mathbf{x}) = \prod_i p\left(\mathrm{x}_i \mid Pa_{\mathcal{G}}(\mathrm{x}_i)\right)$$

$$\tilde{p}(\mathbf{x}) = \Pi_{\mathcal{C}\in\mathcal{G}}\,\phi(\mathcal{C})$$





$$p(\mathrm{a},\mathrm{b},\mathrm{c},\mathrm{d},\mathrm{e}) = p(\mathrm{a})p(\mathrm{b}\mid\mathrm{a})p(\mathrm{c}\mid\mathrm{a},\mathrm{b})p(\mathrm{d}\mid\mathrm{b})p(\mathrm{e}\mid\mathrm{c})$$

$$p(\mathrm{a},\mathrm{b},\mathrm{c},\mathrm{d},\mathrm{e},\mathrm{f}) \quad \frac{1}{Z}\phi_{\mathrm{a},\mathrm{b}}(\mathrm{a},\mathrm{b})\phi_{\mathrm{b},\mathrm{c}}(\mathrm{b},\mathrm{c})\phi_{\mathrm{a},\mathrm{d}}(\mathrm{a},\mathrm{d})\phi_{\mathrm{b},\mathrm{e}}(\mathrm{b},\mathrm{e})\phi_{\mathrm{e},\mathrm{f}}(\mathrm{e},\mathrm{f})$$

Bayesian networks. Directed, Acyclic

Markov fields. Undirected

Probabilistic graphical models. Directed Bayesian networks

Directed PGMs

As basic tools for qualitative modelling of uncertainty use probabilistic influence diagrams a.k.a. causal networks, Bayesian networks, Belief networks,…. See the excellent

http://en.wikipedia.org/wiki/Bayesian_network

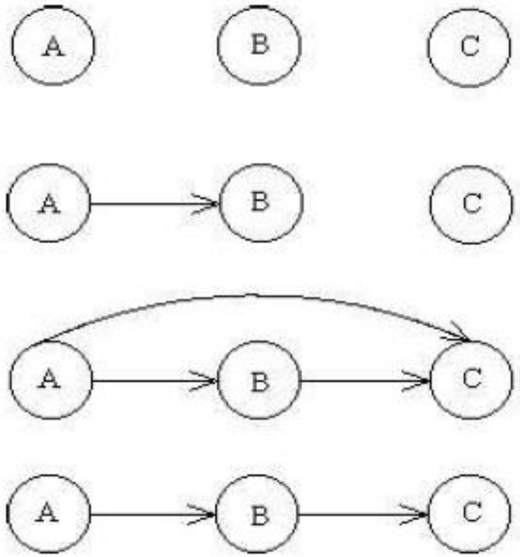
They are **influence diagrams** with chance nodes only. Qualitatively they describe a probabilistic model through

$$P(A1, A2,\ldots, An) = P(A1 \mid ant(A1))\ldots P(An \mid ant (An))$$

where ant (Ai) are the antecessors of node Ai.

In what follows we see several PIDs and we need to indicate the entailed probabilistic model

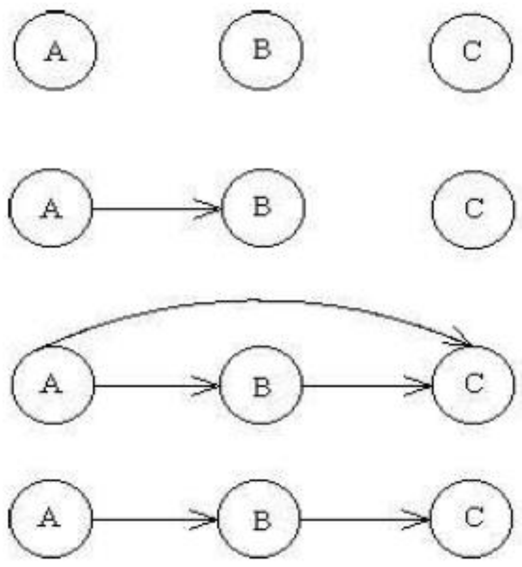
Probabilistic diagrams with three nodes



Before moving foreward, write the entailed probabilistic model

Probabilistic diagrams with three nodes

Model $P(A, B, C)$



$P(A)P(B)P(C)$

$P(A)\ P(B|A)\ P(C)$

$P(A)P(B|A)P(C|A,B)$

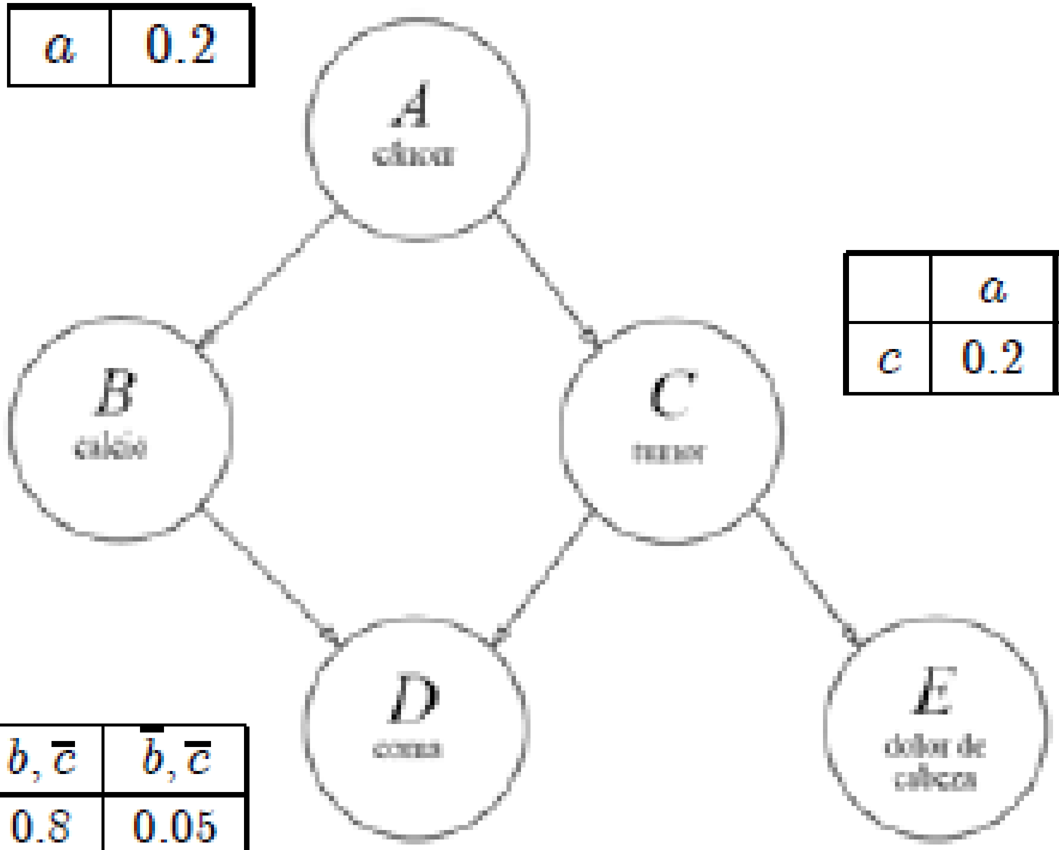
$P(A)P(B|A)P(C|B)$

First case, independence. Third case, A and C are conditionally independent given B.

Read http://en.wikipedia.org/wiki/Conditional_independence

The hidden info

a	0.2



	a	$\overline{a}$
b	0.8	0.2

	a	$\overline{a}$
c	0.2	0.05

	b, c	$\overline{b}, c$	$b, \overline{c}$	$\overline{b}, \overline{c}$
d	0.8	0.8	0.8	0.05

	c	$\overline{c}$
e	0.8	0.6

$$P(A,B,C,D,E) = P(A)P(B|A)P(C|A)P(D|B,C)P(E|C)$$

Probabilistic diagrams. Asia

An example referring to lung diseases

A breathing condition (dyspnea) may be due to tuberculosis, lung cancer or bronchitis, none of them or several of them. A recent visit to Asia, increases the chances of tuberculosis, whereas smoking is a risk factor for lung cancer and bronchitis. The results of an X-ray may not discriminate between cancer and tuberculosis, as neither the presence or absence of dyspnea does.

Probabilistic diagrams

An example referring to lung diseases:

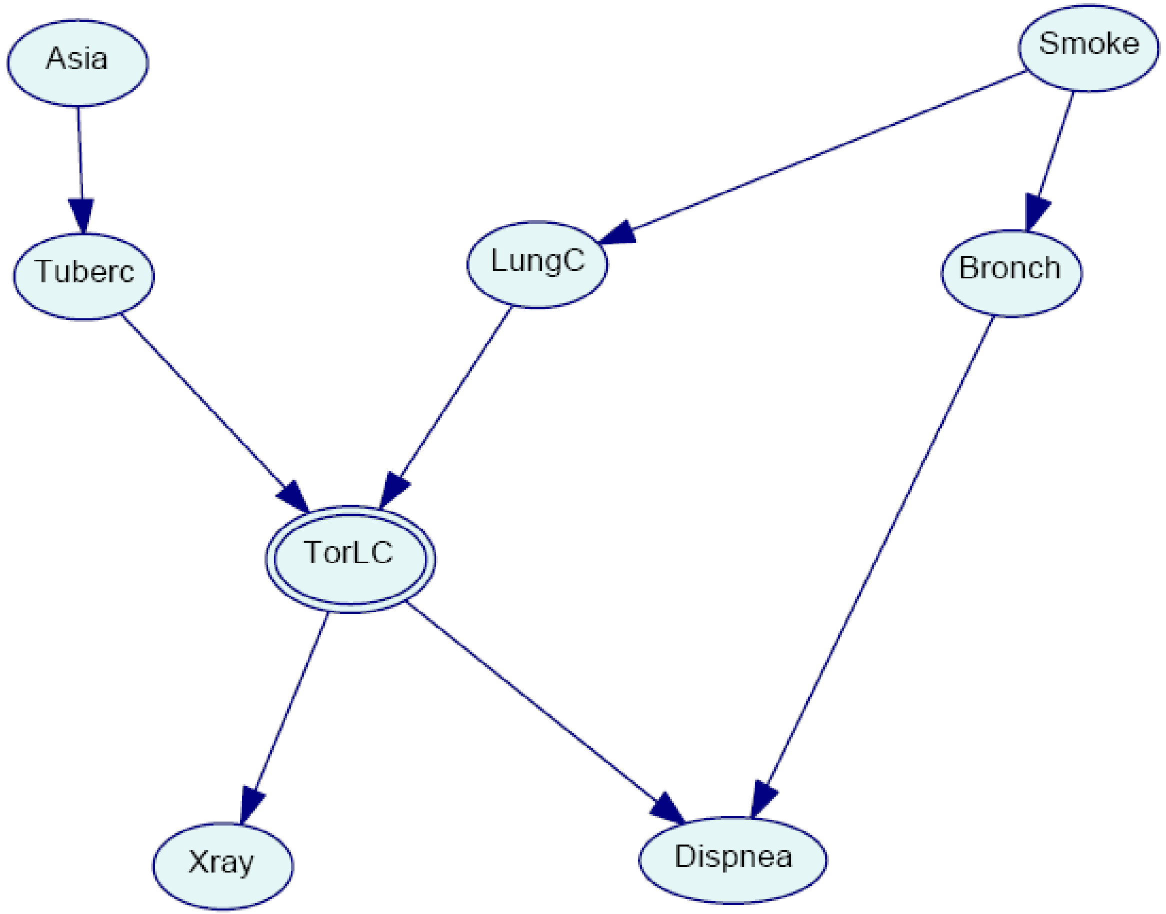
A breathing condition (dyspnea) may be due to tuberculosis, lung cancer or bronchitis, none of them or several of them. A recent visit to Asia, increases the chances of tuberculosis, whereas smoking is a risk factor for lung cancer and bronchitis. The results of an X-ray may not discriminate between cancer and tuberculosis, as neither the presence or absence of dyspnea does.

Probabilistic diagrams

An example referring to lung diseases

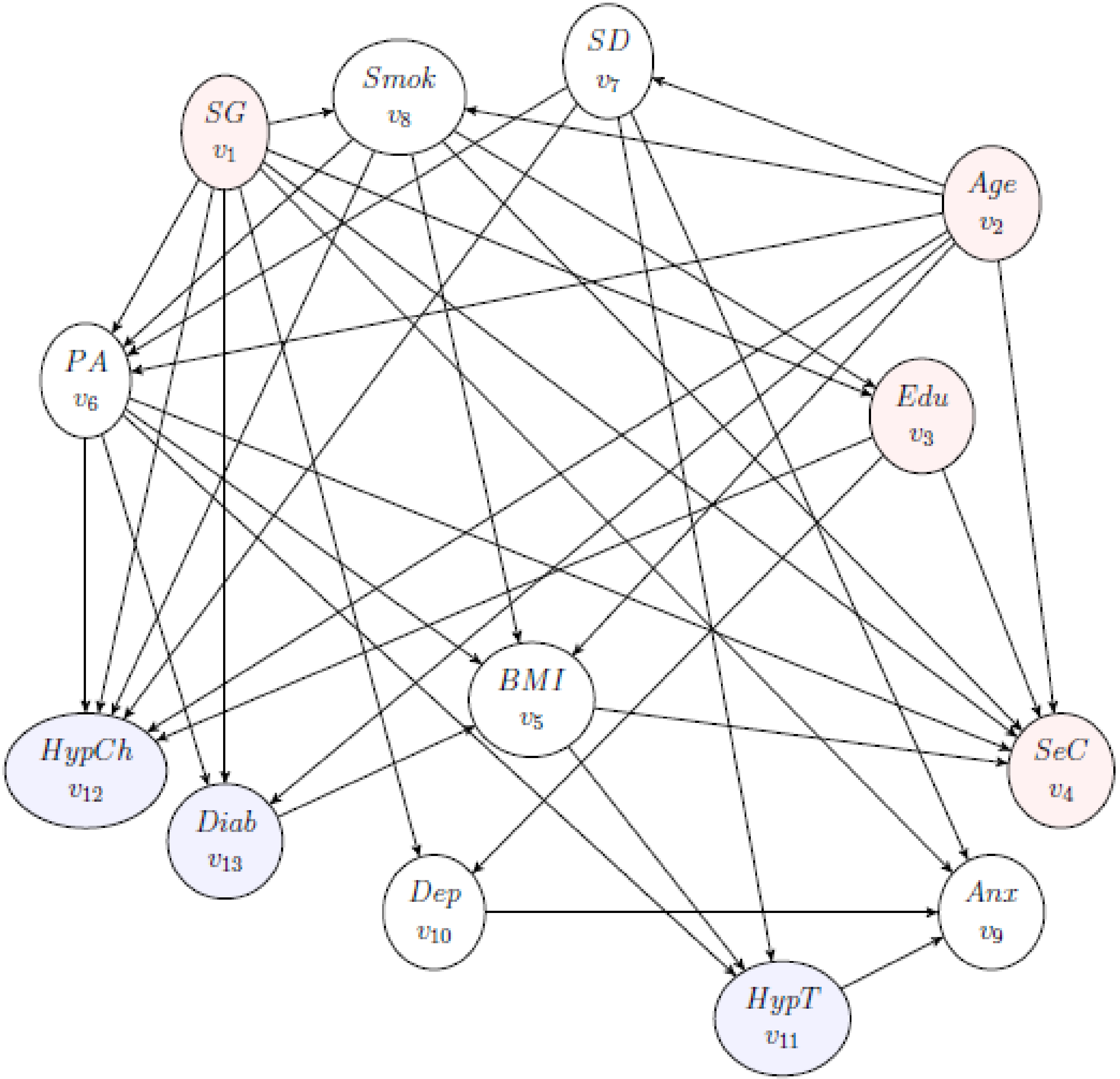
A breathing condition (dyspnea) may be due to tuberculosis, lung cancer or bronchitis, none of them or several of them. A recent visit to Asia, increases the chances of tuberculosis, whereas smoking is a risk factor for lung cancer and bronchitis. The results of an X-ray may not discriminate between cancer and tuberculosis, as neither the presence or absence of dyspnea does.

Probabilistic diagrams



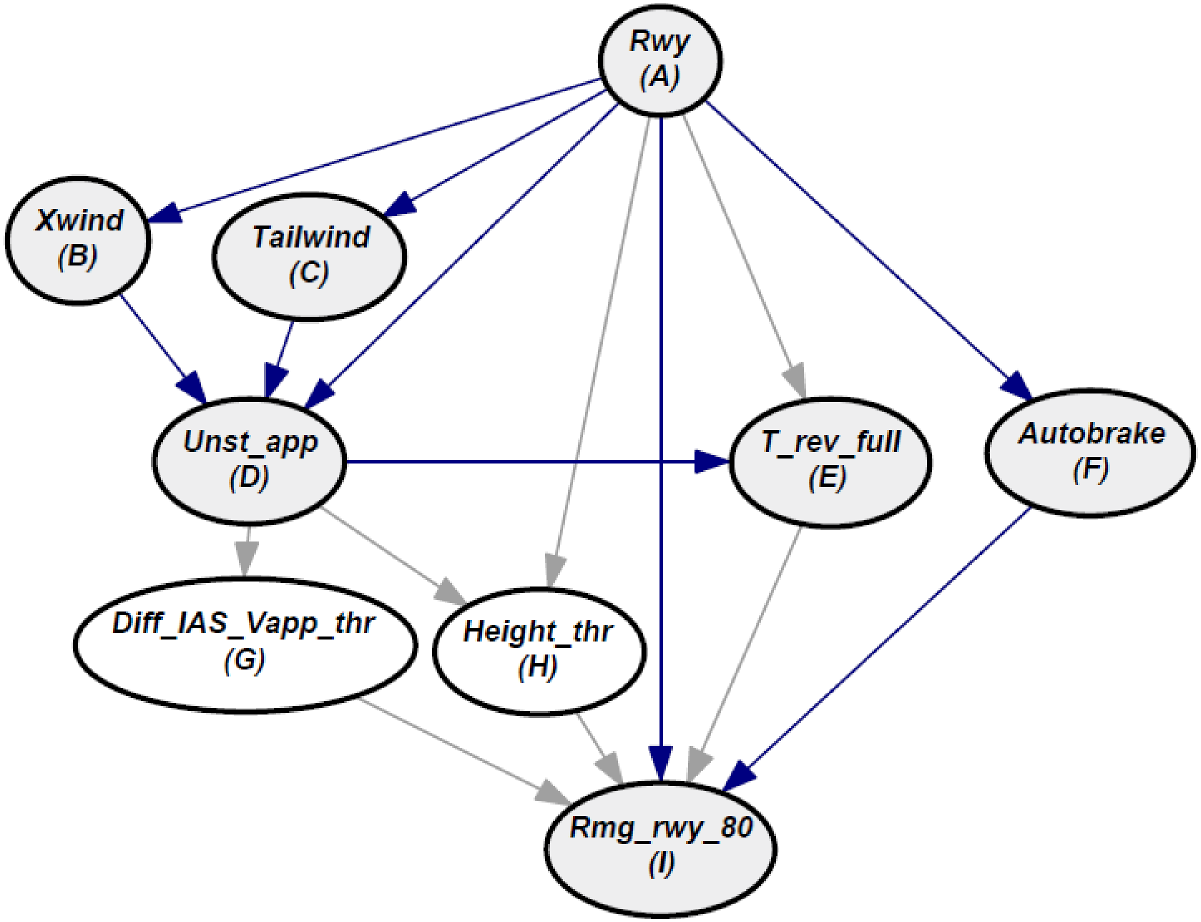
P(A,T,S,L,B,O,X,D) = P(A)P(T|A)P(S)P(L|S)P(B|S)P(0|T,L)P(X|O)P(D|O,B)

Hypertension



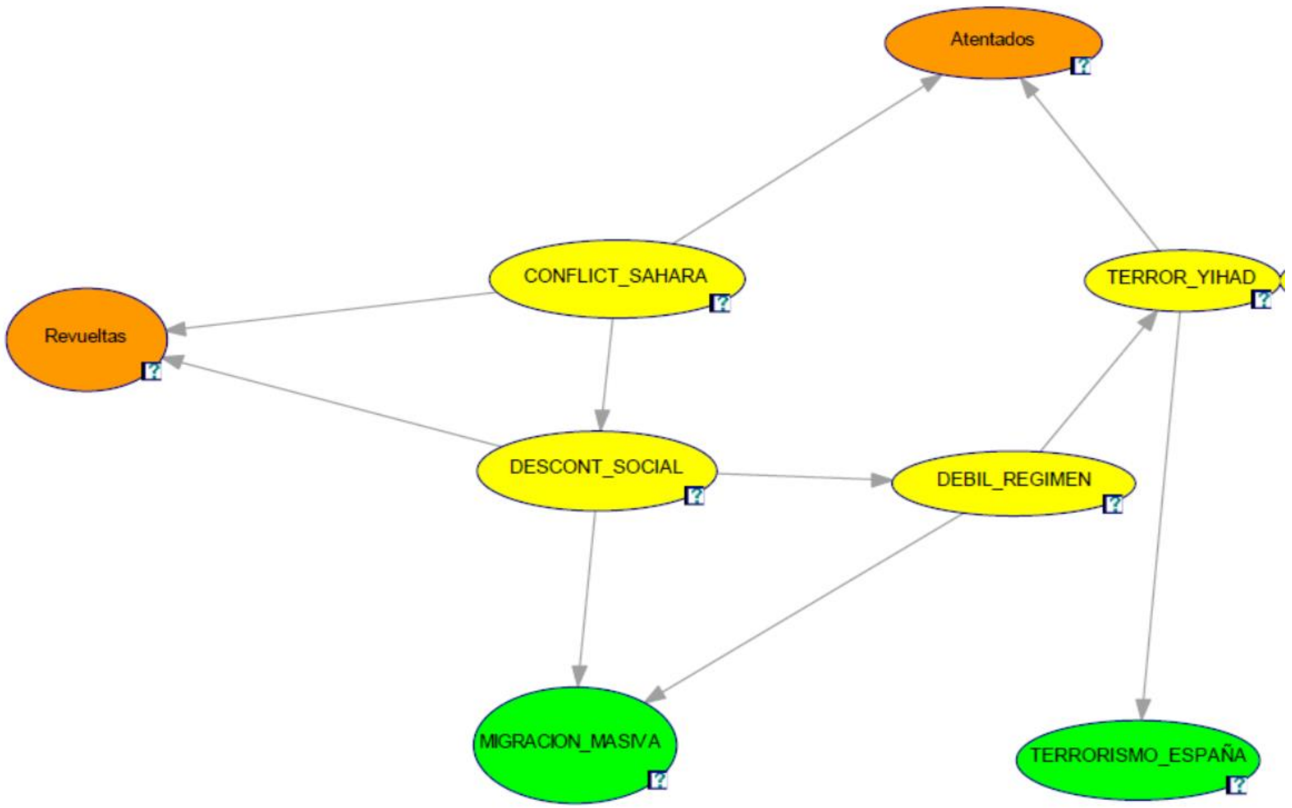
Build the probabilistic model

Runway excursions at airports



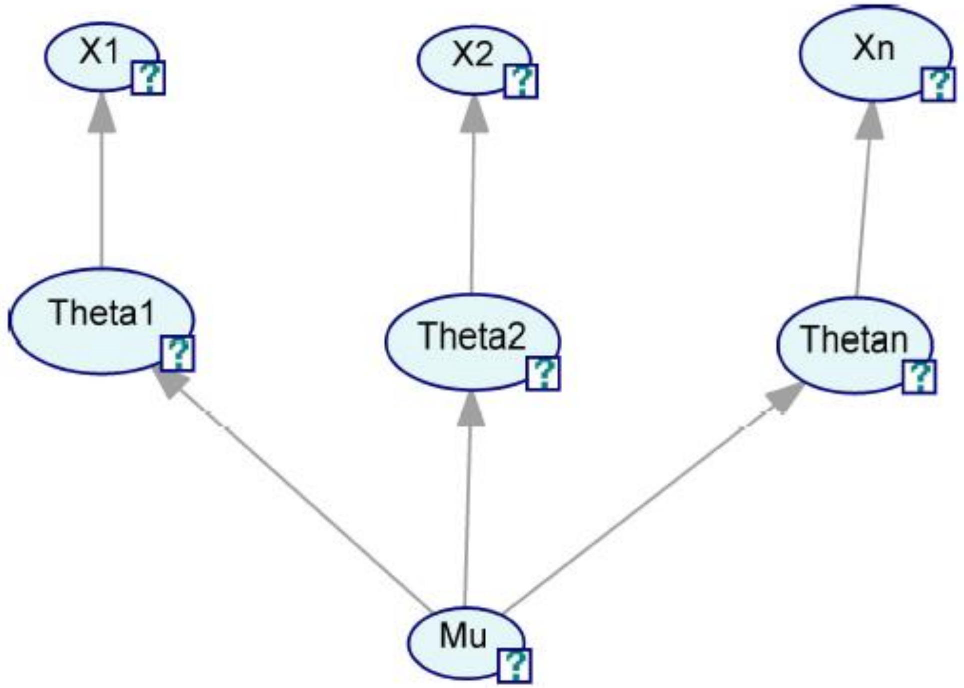
Build the probabilistic model

National security

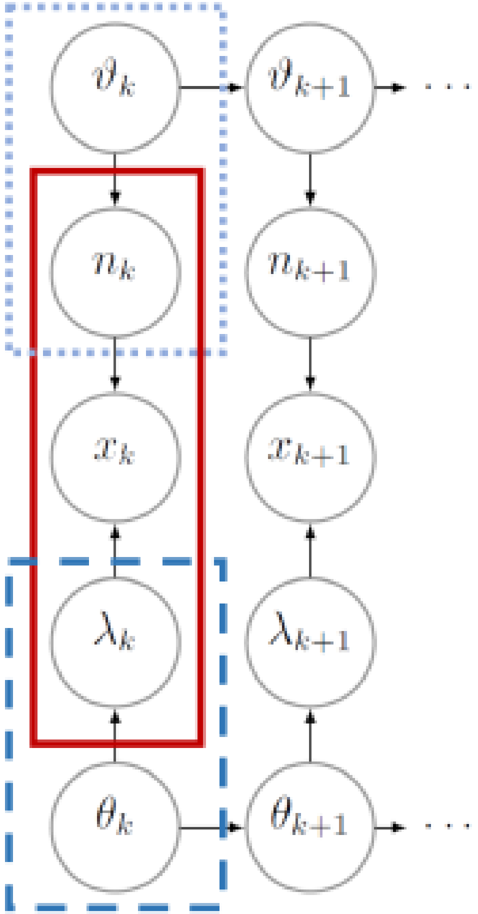


Build the probabilistic model

Statistical models as PGMs. Hierarchical models



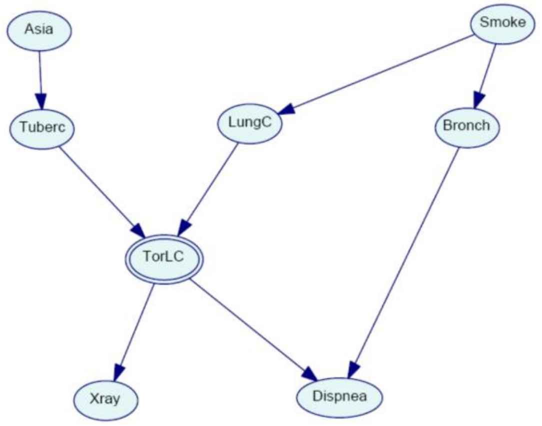
National aviation safety plan



$$
\begin{cases}
\begin{cases}
\begin{cases}
n_k = \boldsymbol{H}_k \vartheta_k + z_k, \quad z_k \sim N(0, \Sigma_k) \\
\vartheta_k = \boldsymbol{J}_k \vartheta_{i-1} + \xi_k, \quad \xi_k \sim N(\boldsymbol{0}, \boldsymbol{S}_k)
\end{cases} \\
\vartheta_0 \sim N(\boldsymbol{\eta}_0, \boldsymbol{S}_0)
\end{cases} \\
x_k | \lambda_k, n_k \sim Po(\lambda_k n_k), \qquad \lambda_k = \exp(u_k) \\
\begin{cases}
\begin{cases}
u_k = \boldsymbol{F}_k \boldsymbol{\theta}_k + v_k, \quad v_k \sim N(0, V_k) \\
\boldsymbol{\theta}_k = \boldsymbol{G}_k \boldsymbol{\theta}_{k-1} + \boldsymbol{w}_k, \quad \boldsymbol{w}_k \sim N(\boldsymbol{0}, \boldsymbol{W}_k)
\end{cases} \\
\boldsymbol{\theta}_0 \sim N(\boldsymbol{m}_0, \boldsymbol{C}_0),
\end{cases}
\end{cases}
$$

Inference in graphical models

General problem



Assuming DAG (arcs and distributions at nodes):

1. Initialisation
2. Absorption of evidence
3. Global propagation of evidence
4. Hypothesising and propagating single pieces of evidence
5. Planning
6. Influential findings

Gibbs sampler for belief nets

Conditionals

$$P(X_j = x_j | X_{-j} = x_{-j}) = \alpha P(X_j = x_j | \Pi_{X_j}(x_{-j})) \prod_{Y_j \in S_j} P(Y_j = y_j | \Pi_{Y_j}(x_j))$$

Back to example

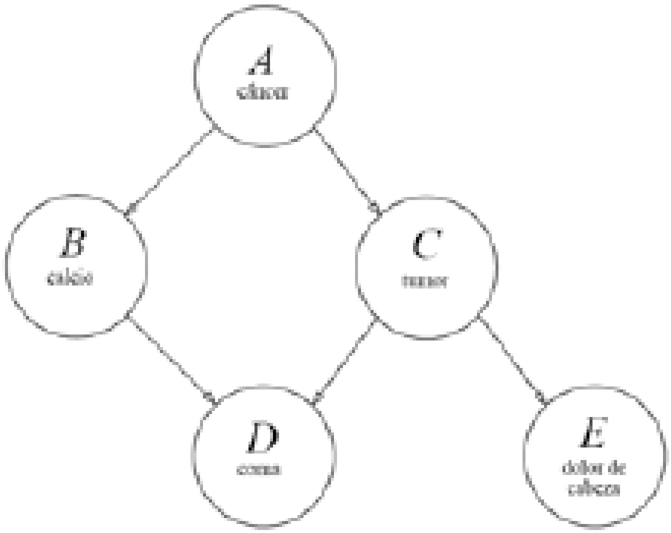
a	0.2

	a	$\overline{a}$
b	0.8	0.2

	a	$\overline{a}$
c	0.2	0.05

	b, c	$\overline{b}, c$	$b, \overline{c}$	$\overline{b}, \overline{c}$
d	0.8	0.8	0.8	0.05

	c	$\overline{c}$
e	0.8	0.6



$$P(c|\overline{d}, e) = \frac{P(c, \overline{d}, e)}{P(\overline{d}, e)}$$

$$\begin{aligned} P(c, \overline{d}, e) &= \sum_{\alpha,\beta} P(\alpha, \beta, c, \overline{d}, e) = \sum_{\alpha,\beta} P(\alpha)P(\beta|\alpha)P(c|\alpha)P(\overline{d}|\beta, c)P(e|c) \\ &= P(a)P(b|a)P(c|a)P(\overline{d}|b, c)P(e|c) + P(a)P(\overline{b}|a)P(c|a)P(\overline{d}|\overline{b}, c)P(e|c) + \\ &\quad P(\overline{a})P(b|\overline{a})P(c|\overline{a})P(\overline{d}|b, c)P(e|c) + P(\overline{a})P(\overline{b}|\overline{a})P(c|\overline{a})P(\overline{d}|\overline{b}, c)P(e|c) \\ &= 0.0118 \end{aligned}$$

$$P(\overline{d}, e) = \sum_{\alpha,\beta,\gamma} P(\alpha, \beta, \gamma, \overline{d}, e) = 0.410$$

$$P(A, B, C, D, E) = P(A)P(B|A)P(C|A)P(D|B, C)P(E|C)$$

$$P(c|\overline{d}, e) = 0.0287$$

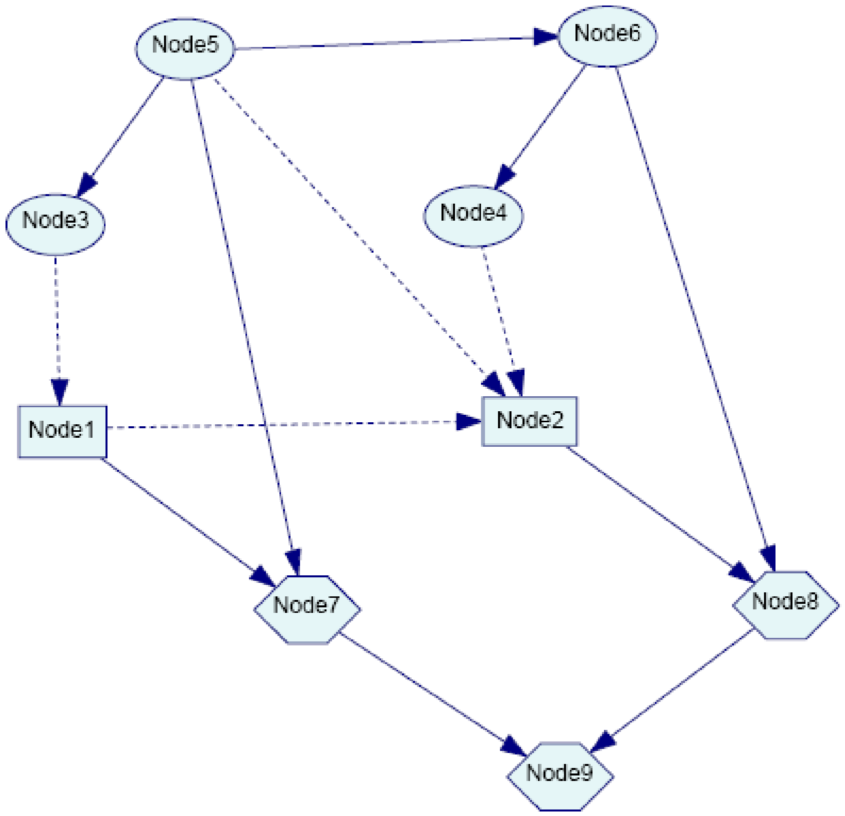
Back to example

$$
\begin{aligned}
P(A|B,C,\bar{d},e) &= P(A|x_{-A}) = \alpha_1 P(A)P(B|A)P(C|A) \\
P(B|A,C,\bar{d},e) &= P(B|x_{-B}) = \alpha_2 P(B|A)P(\bar{d}|B,C) \\
P(C|A,B,\bar{d},e) &= P(C|x_{-C}) = \alpha_3 P(C|A)P(\bar{d}|B,C)P(e|C)
\end{aligned}
$$

```
Seleccionar B = b_0, C = c_0 arbitrariamente
Hacer j = 1
Mientras no se juzgue convergencia,
    Generar A_j = a_j ~ P(A|x_{-A}) = α_{1j} P(A) P(b_{j-1}|A) P(c_{j-1}|A)
    Generar B_j = b_j ~ P(B|x_{-B}) = α_{2j} P(B|a_j) P(d̄|B, c_{j-1})
    Generar C_j = c_j ~ P(C|x_{-C}) = α_{3j} P(C|a_j) P(d̄|b_j, C) P(e|C)
    Hacer j = j + 1
```

$$\frac{\#\{C_j = c\}}{M}$$

Sequential Decisions



Learning structure from data: **Structure learning**. Greedy search based on a scoring function based on an information measure

Learning node distributions….

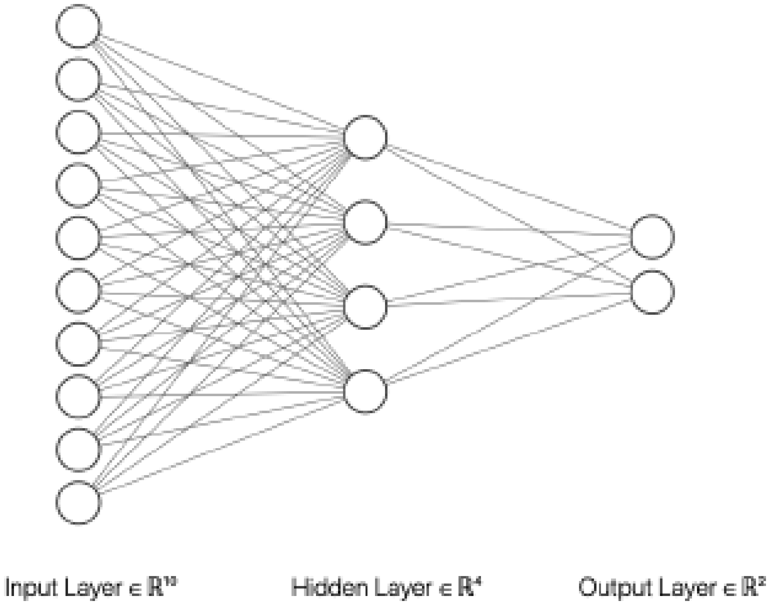
Deep belief nets

GeNIe

- https://www.bayesfusion.com/influence-diagrams/
- https://download.bayesfusion.com/files.html?category=Academia

Shallow neural nets

Formulation



$$y = \sum_{j=1}^{m} \beta_j \psi(x'\gamma_j) + \epsilon$$

$$\epsilon \sim N(0, \sigma^2),$$

$$\psi(\eta) = \exp(\eta)/(1 + \exp(\eta))$$

Linear in beta's, nonlinear in gamma's

Training

Given training data, maximise log-likelihood

$$\min_{\beta,\gamma} f(\beta,\gamma) = \sum_{i=1}^{n} f_i(\beta,\gamma) = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} \beta_j \psi(x_i' \gamma_j) \right)^2$$

Gradient descent

Backpropagation to estimate gradient

Training with regularisation

$$\min_{\beta,\gamma} f(\beta,\gamma) = \sum_{i=1}^{n} f_i(\beta,\gamma) = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} \beta_j \psi(x_i'\gamma_j) \right)^2$$

$$\min g(\beta,\gamma) = f(\beta,\gamma) + h(\beta,\gamma),$$

Weight decay $$h(\beta,\gamma) = \lambda_1 \sum \beta_i^2 + \lambda_2 \sum \sum \gamma_{ji}^2$$ Ridge

Bayesian analysis of shallow neural nets (fixed arch)

$$y = \sum_{j=1}^{m} \beta_j \psi(x'\gamma_j) + \epsilon$$

$$\epsilon \sim N(0, \sigma^2),$$

$$\psi(\eta) = \exp(\eta)/(1 + \exp(\eta))$$

$$\beta_i \sim N(\mu_\beta, \sigma_\beta^2) \text{ and } \gamma_i \sim N(\mu_\gamma, S_\gamma^2)$$

$$\mu_\beta \sim N(a_\beta, A_\beta),\ \mu_\gamma \sim N(a_\gamma, A_\gamma),\ \sigma_\beta^{-2} \sim \quad Gamma(c_b/2, c_b C_b/2)$$

$$S_\gamma^{-1} \sim Wish(c_\gamma, (c_\gamma C_\gamma)^{-1}) \text{ and } \sigma^{-2} \sim Gamma(s/2, sS/2)$$

Bayesian analysis of shallow neural nets (fixed arch)

Start with arbitrary (β, γ, ν).
while *not convergence* **do**
  Given current (γ, ν), draw β from $p(\beta|\gamma, \nu, y)$ (a multivariate normal).
  for $j = 1, \ldots, m$, *marginalizing in* β *and given* ν **do**
    Generate a candidate $\tilde{\gamma}_j \sim g_j(\gamma_j)$.
    Compute $a(\gamma_j, \tilde{\gamma}_j) = \min\left(1, \frac{p(D|\tilde{\gamma}, \nu)}{p(D|\gamma, \nu)}\right)$ with $\tilde{\gamma} = (\gamma_1, \gamma_2, \ldots, \tilde{\gamma}_i, \ldots, \gamma_m)$.
    With probability $a(\gamma_j, \tilde{\gamma}_j)$ replace γ_j by $\tilde{\gamma}_j$. If not, preserve γ_j.
  end
  Given β and γ, replace ν based on their posterior conditionals:
  $p(\mu_\beta|\beta, \sigma_\beta)$ is normal; $p(\mu_\gamma|\gamma, S_\gamma)$, multivariate normal; $p(\sigma_\beta^{-2}|\beta, \mu_\beta)$, Gamma; $p(S_\gamma^{-1}|\gamma, \mu_\gamma)$, Wishart; $p(\sigma^{-2}|\beta, \gamma, y)$, Gamma.
end

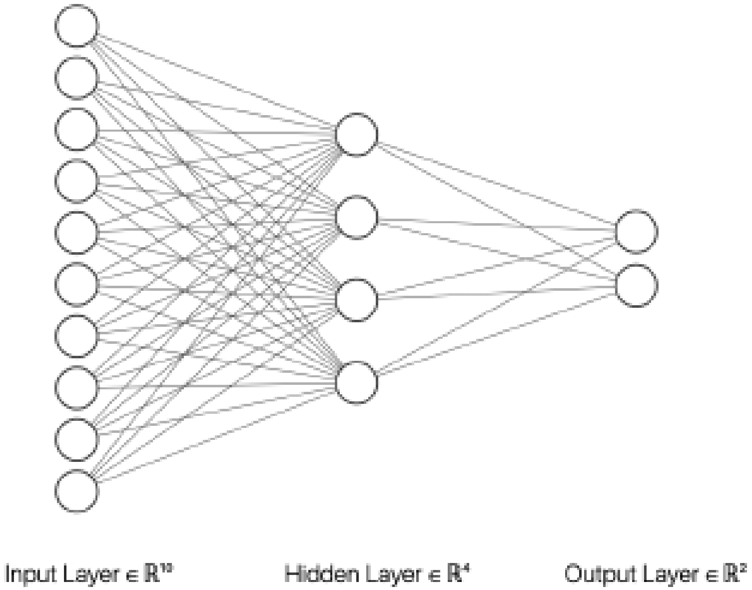
Bayesian analysis of shallow neural nets (var arch)

$$
\begin{aligned}
y &= x_i'a + \sum_{j=1}^{m^*} d_j \beta_j \psi(x'\gamma_j) + \epsilon \\
& \epsilon \sim N(0, \sigma^2), \\
& \psi(\eta) = \exp(\eta)/(1+\exp(\eta)), \\
Pr(d_j = k | d_{j-1} = 1) &= (1-\alpha)^{1-k} \times \alpha^k, k \in \{0,1\}
\end{aligned}
$$

$$\beta_i \sim N(\mu_b, \sigma_\beta^2),\ a \sim N(\mu_a, \sigma_a^2),\ \gamma_i \sim N(\mu_\gamma, \Sigma_\gamma).$$

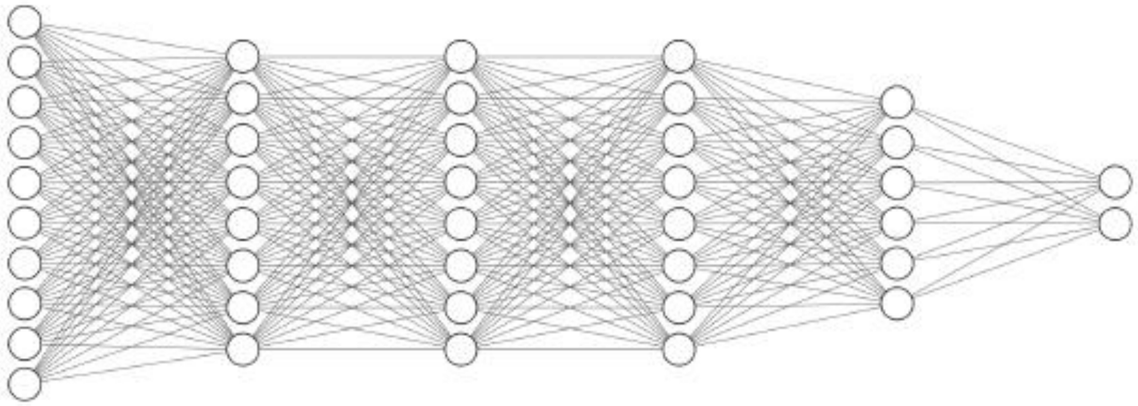
Reversible jump algo

Concept



$$
\begin{aligned}
y &= \sum_{j=1}^{m} \beta_j \psi(x'\gamma_j) + \epsilon \\
&\epsilon \sim N(0, \sigma^2), \\
&\psi(\eta) = \exp(\eta)/(1 + \exp(\eta))
\end{aligned}
$$

(Shallow) Neural nets



$$\{f_0, f_1, \ldots, f_{L-1}\}$$

$$z_{l+1} = f_l(z_l, \gamma_l).$$

$$
\begin{aligned}
y &= \sum_{j=1}^{m_L} \beta_j z_{L,j} + \epsilon \\
&\epsilon \sim N(0, \sigma^2),
\end{aligned}
$$

Deep neural nets

Bayesian analysis of deep neural nets

Start with arbitrary (β, γ, ν).

while *not convergence* **do**

- Given current (γ, ν), draw β from $p(\beta|\gamma, \nu, y)$ (a multivariate normal).
- **for** $j = 1, \ldots, m$, *marginalizing in* β *and given* ν **do**
  - Generate a candidate $\tilde{\gamma}_j \sim g_j(\gamma_j)$.
  - Compute $a(\gamma_j, \tilde{\gamma}_j) = \min\left(1, \frac{p(D|\tilde{\gamma}, \nu)}{p(D|\gamma, \nu)}\right)$ with $\tilde{\gamma} = (\gamma_1, \gamma_2, \ldots, \tilde{\gamma}_i, \ldots, \gamma_m)$.
  - With probability $a(\gamma_j, \tilde{\gamma}_j)$ replace γ_j by $\tilde{\gamma}_j$. If not, preserve γ_j.
- **end**
- Given β and γ, replace ν based on their posterior conditionals: $p(\mu_\beta|\beta, \sigma_\beta)$ is normal; $p(\mu_\gamma|\gamma, S_\gamma)$, multivariate normal; $p(\sigma_\beta^{-2}|\beta, \mu_\beta)$, Gamma; $p(S_\gamma^{-1}|\gamma, \mu_\gamma)$, Wishart; $p(\sigma^{-2}|\beta, \gamma, y)$, Gamma.

end