Discrete Variables & Linear Gaussian Models

Graphical models are a good way of linking together simple (exp. family Chapt. 2.4) distributions to form more powerful and useful models of more complicated data and processes.

Two cases are particularly elegant, discrete and lin. Gaussian (in which all parent child relations are ~~one~~ of either discrete or lin. G.)

Discrete (to start)

Code a discrete r.v. $\vec{x}$ as $\vec{x} = [\underbrace{0\ 0\ 1\ 0\ 0\ 0}_{K \text{ possible states}}]$

A discrete dist. with parameters $\vec{\mu}$ can be written

$$p(\vec{x}|\vec{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k} \qquad \text{eg. } \vec{\mu} = [.1, .1, .3, .5\ 0\ 0]$$

Since $\sum \mu_k = 1$ only $K-1$ parameters are required to represent $\vec{\mu}$

With two discrete r.v.'s $\vec{x}_1$ and $\vec{x}_2$ (each w/ K states) we can write the joint distribution that $x_{1k}=1$ and $x_{2l}=1$ as

$$p(\vec{x}_1, \vec{x}_2 | \mu) = \prod_{k=1}^{K} \prod_{l=1}^{K} \mu_{kl}^{x_{1k} x_{2l}}$$

where $\sum \mu_{kl} = 1$ and μ_{kl} is a matrix with K^2-1 free parameters.

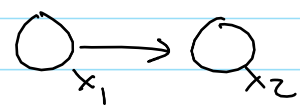
For an arbitrary joint distribution with M discrete variables, $K^M - 1$ params are needed

$K^{M}-1$ grows exponentially with M, the number of variables.

Using the prod rule we can write

$p(\vec{x}_1, \vec{x}_2)$ in the form $p(x_2 | x_1)\, p(x_1)$

i.e.



$p(x_1)$ has $K-1$ params

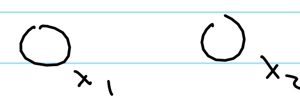
$p(x_2|x_1)$ requires $K-1$ params for each of the K states x_1 can be in

the total # params is

$$K-1 + K(K-1) = K^2 - 1 \quad \text{as before}$$

but...

if $x_1 \perp\!\!\!\perp x_2$ i.e.



then the total number of params would be $2(K-1)$. For M $\perp\!\!\!\perp$ discrete R.V.'s with K states # params grows linearly (!)

$$M(K-1)$$

By dropping links we have decreased the number of parameters in the joint distribution.

Intermediate factorizations are interesting

Consider

$$x_1 \to x_2 \to \cdots \to x_M \qquad \text{(Chain)}$$

The total number of params grows as

$$K-1 + (M-1)(K-1)K$$

which is quadratic in K and linear in M

Parameters count can be reduced through sharing or tying of parameters. In (Chain) we can say all cond. dist's are ~~the~~ ~~same~~ tied leaving

$$K^2-1$$

parameters that must be specified in order to define the joint distn.

Linear Gaussian Models

D dimensional Gaussian w/ diagonal covariance has $2D$ free parameters, D for the mean and D for the diagonal of the covariance matrix.

D dimensional full Gaussian has D free parameters for the mean and $(D^2-D)/2 + D$ free covariance matrix parameters (symmetric)

It is possible to construct intermediate Gaussian dist's of interm. complexity

Let

$$p(x_i \mid pa_i) = \mathcal{N}\left(x_i \,\middle|\, \sum_{j \in pa_i} w_{ij} x_j + b_i, v_i\right)$$

where w_{ij} and b_i are parameters that govern the conditional mean of x_i and v_i is the variance of x_i's conditional dist.

By inspection of the log joint $p(\vec{x})$ we can see that the full joint dist. is Gaussian

$$\ln p(\vec{x}) = \sum_{i=1}^{D} \ln p(x_i \mid pa_i)$$

$$= \sum_{i=1}^{D} \frac{1}{2v_i}\left(x_i - \sum_{j \in pa_i} w_{ij} x_j - b_i\right)^2 + \text{const}$$

which is quadritic in the components of $\vec{x}$ and therefore Gaussian.

The mean and variance of the resulting joint Gaussian can be determined recursively

Each x_i, conditioned on its parents can be written as

$$x_i = \sum_{j \in pa_i} w_{ij} x_j + b_i + \sqrt{v_i}\, \varepsilon_i$$

where $\varepsilon_i \sim \mathcal{N}(0,1)$, $\mathbb{E}[\varepsilon_i] = 0$ and $\mathbb{E}(\varepsilon_i \varepsilon_j) = I_{ij}$ where I_{ij} is the ij th element of the identity matrix.

So $$\mathbb{E}[x_i] = \sum_{j \in pa_i} w_{ij} \mathbb{E}[x_j] + b_i$$

which is how the mean of $\vec{x} \sim \mathcal{N}(?, ?)$ can be calculated.

The covariance between x_i and x_j can also be recursively calculated.

Starting with the definition of covariance

$$\mathrm{cov}[x_i, x_j] = \mathbb{E}\left[\left(x_i - \mathbb{E}(x_i)\right)\left(x_j - \mathbb{E}(x_j)\right)\right]$$

$$= \mathbb{E}\left[\left(x_i - \mathbb{E}(x_i)\right)\left(\sum_{k \in pa_j} w_{jk} x_k + b_j + \sqrt{v_j}\, \varepsilon_j - \left(\sum_{k \in pa_j} w_{jk} \mathbb{E}[x_k] + b_j\right)\right)\right]$$

$$= \mathbb{E}\left[\left(x_i - \mathbb{E}(x_i)\right)\left(\sum_{k \in pa_j} w_{jk}\left(x_k - \mathbb{E}[x_k]\right) + \sqrt{v_j}\, \varepsilon_j\right)\right]$$

$$= \sum_{k \in pa_j} w_{jk}\, \mathrm{cov}(x_i, x_k) + \underbrace{\mathbb{E}\left[\left(x_i - \mathbb{E}(x_i)\right)\sqrt{v_j}\, \varepsilon_j\right]}$$

$$= \mathbb{E}\left[x_i \sqrt{v_j}\, \varepsilon_j\right]$$

$$= \mathbb{E}\left[\left(\underbrace{\sum_{j \in pa_i} w_{ij} x_j}_{\text{contains no } \varepsilon \text{ terms}} + b_i + \sqrt{v_i}\, \varepsilon_i\right)\sqrt{v_j}\, \varepsilon_j\right]$$

$$= \mathbb{E}\left[\sqrt{v_i}\sqrt{v_j}\, \varepsilon_i \varepsilon_j\right] = \underbrace{\mathbb{E}\left[\varepsilon_i \varepsilon_j\right]}_{1 \text{ iff } i=j} \sqrt{v_i}\sqrt{v_j}$$

$$= \sqrt{v_i}\sqrt{v_i} = v_i \text{ or } v_j$$

$$= \sum_{k \in pa_j} w_{jk}\, \mathrm{cov}(x_i, x_k) + I_{ij} v_j$$

Example

$$x_1 \longrightarrow x_2 \longrightarrow x_3$$

think: latent process in Kalman filter, e.g.

rules: $\mathbb{E}[x_i] = \sum_{j \in pa_i} w_{ij}\, \mathbb{E}[x_j] + b_i$, $\operatorname{cov}[x_i, x_j] = \sum_{k \in pa_j} w_{jk} \operatorname{cov}[x_i, x_k] + I_{ij} v_j$

$$\mu = \left[\mathbb{E}[x_1] , \mathbb{E}[x_2] , \mathbb{E}[x_3] \right]$$

$$= \left[b_1 , \; w_{12} b_1 + b_2 , \; w_{32}(w_{12} b_1 + b_2) + b_3 \right]$$

$$\Sigma = \begin{array}{c c} & \begin{array}{ccc} j=1 & 2 & 3 \end{array} \\ \begin{array}{c} i= \\ 2 \\ 3 \end{array} & \left[\begin{array}{ccc} \operatorname{cov}(x_1, x_1) & \operatorname{cov}(x_1, x_2) & \operatorname{cov}(x_1, x_3) \\ = & \operatorname{cov}(x_2, x_2) & \operatorname{cov}(x_2, x_3) \\ \vdots & \vdots & \vdots \end{array} \right] \end{array}$$

$$= \left[\begin{array}{ccc} v_1 & w_{21} v_1 & w_{32} w_{21} v_1 \\ & \ddots & \\ & & \end{array} \right]$$

$$\operatorname{cov}[x_i, x_j] = \sum_{k \in pa_j} w_{jk} \operatorname{cov}[x_i, x_k] + I_{ij} v_j$$