Discrete Variables & Linear Gaussian Models

Graphical models are a good way of linking together simple (e.g., 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 linear Gaussian (in which all parent-child relations are one of either discrete or linear).

Discrete (to start)

Consider a discrete r.v. $\mathbf{x}$ as $\mathbf{x} = [0 \ 0 \ 1 \ 0 \ 0 \ 0]$, $k$ possible states.

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

$$p(\mathbf{x} | \mathbf{\mu}) = \prod_{k=1}^{k} \mu_{k}$$

eg. $\mathbf{\mu} = [0.1, 1.3, 0.005]$.

Since $\sum_{k} \mu_{k} = 1$ only $k-1$ parameters are required to represent $\mathbf{\mu}$.

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

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

where $\sum_{k} \mu_{k} = 1$ and $\mathbf{\mu}$ is a matrix with $k^2 - 1$ free parameters.

For an arbitrary joint distribution with $M$ discrete variables, $k^{M-1}$ parameters are needed.
$k^{M-1}$ grows exponentially with $M$, the number of variables.

Using the prod rule we can write

$p(x_1, x_2)$ in the form $p(x_2|x_1)p(x_1)$

i.e. $O_{x_1} \rightarrow O_{x_2}$

$p(x_1)$ 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

$K-1 + K(K-1) = K^2 - 1$ as before

but...

if $x_1 \perp x_2$ i.e. $O_{x_1} \perp O_{x_2}$

then the total number of params would be $2(K-1)$. For $M$ discrete RVs with $K$ states the # params grows linearly (v)

$M(K-1)$

By dropping links we have decreased the number of parameters in the joint distribution.
Intermediate factorizations are interesting

Consider

\[ X_1 \xrightarrow{O} X_2 \xrightarrow{\cdots} X_N \] (Chain)

The total number of parameters grows as

\[ K^2 - 1 + (M-1)(K-1)K \]

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 vari. dists are 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 \( D^2 - D\sqrt{D} + D \) free covariance matrix parameters (symmetric).

It is possible to construct intermediate Gaussian dists of intermediate complexity.
Let
\[
p(x_i \mid p_a_i) = \mathcal{N}(x_i \mid \sum_{j \in p_{a_i}} w_{ij} x_j + b_i, \nu_i)
\]
where \(w_{ij}\) and \(b_i\) are parameters that govern the conditional mean of \(x_i\) and \(\nu_i\) is the variance of \(x_i\)'s conditional distribution.

By inspecting the log joint \(p(x)\), we can see that the full joint distribution is Gaussian
\[
\ln p(x) = \sum_{i=1}^{D} \ln p(x_i \mid p_{a_i})
\]
\[
= \sum_{i=1}^{D} \frac{1}{2\nu_i} (x_i - \sum_{j \in p_{a_i}} w_{ij} x_j - b_i)^2 + \text{const}
\]
which is quadratic in the components of \(x\) and therefore Gaussian.

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

Each \(x_i\), conditioned on its parents can be written as
\[
x_i = \sum_{j \in p_{a_i}} w_{ij} x_j + b_i + \sqrt{\nu_i} \varepsilon_i
\]
where \(\varepsilon_i \sim \mathcal{N}(0, 1)\) and \(\mathbb{E}[\varepsilon_i] = 0\) and \(\mathbb{E}(\varepsilon_i \varepsilon_j) = \delta_{ij}\) where \(\delta_{ij}\) is the \(i, j\)th element of the identity matrix.

So \(\mathbb{E}[x_i] = \sum_{j \in p_{a_i}} w_{ij} \mathbb{E}[x_j] + b_i\)
which is how the mean of \(x \sim \mathcal{N}(\cdot, \cdot)\) can be calculated.
The covariance between $x_i$ and $x_j$ can also be recursively calculated.

Starting with the definition of covariance:

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

$$= \mathbb{E}[(x_i - \mathbb{E}(x_i))\left(\sum_{k \in \mathcal{P}_j} \omega_{jk} x_k + b_j + \sqrt{\nu_j} \varepsilon_j \right. - \left. \left(\sum_{k \in \mathcal{P}_j} \omega_{jk} \mathbb{E}(x_k) + b_j\right)\right)]$$

$$= \mathbb{E}[(x_i - \mathbb{E}(x_i))\left(\sum_{k \in \mathcal{P}_j} \omega_{jk} (x_k - \mathbb{E}(x_k)) + \sqrt{\nu_j} \varepsilon_j\right)\right]$$

$$= \sum_{k \in \mathcal{P}_j} \omega_{jk} \text{cov}(x_i, x_k) + \mathbb{E}[(x_i - \mathbb{E}(x_i))\sqrt{\nu_j} \varepsilon_j]$$

$$= \mathbb{E}[(x_i - \mathbb{E}(x_i))\sqrt{\nu_j} \varepsilon_j]$$

$$= \mathbb{E}\left[\left(\sum_{j \in \mathcal{P}_j} \omega_{ij} x_j + b_i + \sqrt{\nu_i} \varepsilon_i \right) \sqrt{\nu_j} \varepsilon_j \right]$$

$$= \mathbb{E}\left[\sqrt{\nu_i} \sqrt{\nu_j} \varepsilon_i \varepsilon_j \right] = \mathbb{E}\left[\varepsilon_i \varepsilon_j \right] \sqrt{\nu_i} \sqrt{\nu_j}$$

$$= \sqrt{\nu_i} \sqrt{\nu_j} = \nu_i \text{ or } \nu_j$$

$$= \sum_{k \in \mathcal{P}_j} \omega_{jk} \text{cov}(x_i, x_k) + I_{ij} \nu_j$$
Example

\[ x_1 \rightarrow x_2 \rightarrow x_3 \]

Think: latent process

In Kalman filter, e.g.

Rules:

\[ \mathbb{E}[x_i] = \sum_{j \in \mathcal{P}_k} \omega_{ij} \mathbb{E}[x_j] + b_i \]

\[ \text{cov}[x_k, x_l] = \sum_{k \neq j} \omega_{ij} \text{cov}[x_k, x_l] \]

\[ M = [\mathbb{E}[x_1], \mathbb{E}[x_2], \mathbb{E}[x_3]] \]

\[ = [b_1, \omega_{12} b_1 + b_2, \omega_{32} (\omega_{12} b_1 + b_2) + b_3] \]

\[ \sum = \begin{bmatrix}
\text{cov}(x_1, x_1) & \text{cov}(x_1, x_2) & \text{cov}(x_1, x_3) \\
\text{cov}(x_2, x_1) & \text{cov}(x_2, x_2) & \text{cov}(x_2, x_3) \\
\text{cov}(x_3, x_1) & \text{cov}(x_3, x_2) & \text{cov}(x_3, x_3)
\end{bmatrix} \]

\[ = \begin{bmatrix}
\nu_1 & \omega_{21} \nu_1 & \omega_{32} \nu_{21} \nu_1 \\
\omega_{21} \nu_1 & \nu_2 & \omega_{32} \nu_{21} \nu_2 \\
\omega_{32} \nu_{21} \nu_1 & \omega_{32} \nu_{21} \nu_2 & \nu_3
\end{bmatrix} \]

\[ \text{cov}[x_i, x_j] = \sum_{k \in \mathcal{P}_d} \omega_{ij} \text{cov}[x_i, x_k] \]

\[ + \nu_{ij} \nu_{ij} \]