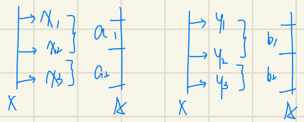
- **Discretization**

have a CPD $P(Y|X)$, $X$ and $Y$ are continuous



A is the discretization of X
B is the discretization of Y

$$P(b_1|a_1) = \int_{x_1}^{x_2} P(Y\in[y_1,y_2]\,|\,x)\cdot P(x\,|\,x\in[x_1,x_2])\,dx \qquad \text{// use the prior } P(X)$$

discretization causes a trade-off between accuracy and computation cost

- **Canonical Forms**

canonical form represents the intermediate result in inference as a log-quadratic form $\exp(Q(x))$ where $Q$ is quadratic

$$C(X;K,h,g) = \exp\left(-\tfrac{1}{2}X^TKX + h^TX + g\right)$$

eg. Gaussian

$$\frac{1}{(2\pi)^{\frac{n}{2}}\cdot|\Sigma|^{1/2}}\cdot\exp\left(-\tfrac{1}{2}(x-u)\Sigma^{-1}(x-u)\right)$$
$$= \exp\left(-\tfrac{1}{2}x^T\Sigma^{-1}x + x^T\Sigma^{-1}u - \tfrac{1}{2}u^T\Sigma^{-1}u - \log\left((2\pi)^{\frac{n}{2}}\cdot|\Sigma|^{1/2}\right)\right)$$
$$K=\Sigma^{-1} \quad h=\Sigma^{-1}u \quad g=-\tfrac{1}{2}u^T\Sigma^{-1}u - \log\left((2\pi)^{\frac{n}{2}}\cdot|\Sigma|^{1/2}\right)$$

The product of 2 canonical form over same scope X is:
$$C(K_1,h_1,g_1)\cdot C(K_2,h_2,g_2) = C(K_1+K_2,\, h_1+h_2,\, g_1+g_2).$$

when we have 2 canonical factors over different scopes,
we simply extend both to make the scopes match and then perform product

$$\phi_1(X,Y) = C(X,Y;\Sigma_1,h_1,g_1) \Rightarrow \phi_1(X,Y,Z) = C\left(X,Y,Z;\begin{bmatrix}\Sigma_1 & 0\\ 0 & 0\end{bmatrix},\begin{bmatrix}h_1\\0\end{bmatrix},g_1\right)$$
$$\phi_2(Y,Z) = C(Y,Z;\Sigma_2,h_2,g_2) \Rightarrow \phi_2(X,Y,Z) = C\left(X,Y,Z;\begin{bmatrix}0 & 0\\ 0 & \Sigma_2\end{bmatrix},\begin{bmatrix}h_2\\0\end{bmatrix},g_2\right)$$
$$\phi_1\times\phi_2 = C\left(X,Y,Z;\begin{bmatrix}\Sigma_1 & 0\\ 0 & 0\end{bmatrix}+\begin{bmatrix}0 & 0\\ 0 & \Sigma_2\end{bmatrix},\begin{bmatrix}h_1\\0\end{bmatrix}+\begin{bmatrix}0\\h_2\end{bmatrix},g_1+g_2\right)$$

$$\frac{C(K_1,h_1,g_1)}{C(K_2,h_2,g_2)} = C(K_1-K_2,\,h_1-h_2,\,g_1-g_2)$$

(the vacuous canonical form, which has $K=0$ $h=0$, $g=0$ is analogous to "all 1" factor)

production

To reduce a canonical form $C(X,Y;K,h,g)$ given $Y=y$, the resulting canonical form is
$C(X;K',h',g')$ where $K'=K_{XX}$, $h'=h_X-K_{XY}y$, $g'=g+h_Y^Ty-\tfrac{1}{2}y^TK_{YY}y$

$$\exp\left(-\tfrac{1}{2}(X^T K_{XX} X + 2X^T K_{XY} Y + Y^T K_{YY} Y) + h_X^T X + h_Y^T Y + g\right)$$
$$= \exp\left(-\tfrac{1}{2} X^T K_{XX} X + X^T(h_X - K_{XY} Y) + g - \tfrac{1}{2} Y^T K_{YY} Y\right)$$

reduction

The marginalization over $Y$ is $C(X; K', h', g')$
where $K' = K_{XX} - K_{XY} K_{YY}^{-1} K_{YX}$ $\quad h' = h_X - K_{XY} K_{YY}^{-1} h_Y$ $\quad g' = g + \frac{1}{2}\left(\log|2\pi K_{YY}^{-1}| + h_Y^T K_{YY}^{-1} h_Y\right)$

$$C(X, Y; K, h, g) = \exp\left(-\tfrac{1}{2}[X\ Y] K \begin{bmatrix} X \\ Y \end{bmatrix} + [X\ Y] h + g\right)$$
$$= \exp\left(-\tfrac{1}{2} X^T K_{XX} X - X^T K_{XY} Y - \tfrac{1}{2} Y^T K_{YY} Y + X^T h_X + Y^T h_Y + g\right)$$
$$= \exp\Big[\underbrace{-\tfrac{1}{2} X^T K_{XX} X + \tfrac{1}{2} X^T K_{XY} K_{YY}^{-1} K_{YX} X}$$
$$-\tfrac{1}{2} X^T K_{XY} K_{YY}^{-1} K_{YX} X - X^T K_{XY} Y - \tfrac{1}{2} Y^T K_{YY} Y + \underbrace{X^T h_X} + Y^T h_Y + g$$
$$\underbrace{- X^T K_{XY} K_{YY}^{-1} h_Y} + X^T K_{XY} K_{YY}^{-1} h_Y \underbrace{+ \tfrac{1}{2} h_Y^T K_{YY}^{-1} h_Y} - \tfrac{1}{2} h_Y^T K_{YY}^{-1} h_Y\Big]$$
$$= \exp\left[-\tfrac{1}{2} X^T (K_{XX} - K_{XY} K_{YY}^{-1} K_{YX}) X\right] \cdot \exp\left(g + \tfrac{1}{2} h_Y^T K_{YY}^{-1} h_Y\right)$$
$$\exp\left(X^T h_X - X^T K_{XY} K_{YY}^{-1} h_Y\right) \cdot$$
$$\exp\left[-\tfrac{1}{2}\left(K_{YY}^{1/2} Y + K_{YY}^{-1/2} K_{YX} X - K_{YY}^{-1/2} h_Y\right)^T \left(K_{YY}^{1/2} Y + K_{YY}^{-1/2} K_{YX} X - K_{YY}^{-1/2} h_Y\right)\right]$$

$$\int_Y C(X, Y; K, h, g)\, dY = \exp\left[-\tfrac{1}{2} X^T (K_{XX} - K_{XY} K_{YY}^{-1} K_{YX}) X\right] \cdot \exp\left(g + \tfrac{1}{2} h_Y^T K_{YY}^{-1} h_Y\right) \cdot \exp\left(X^T h_X - X^T K_{XY} K_{YY}^{-1} h_Y\right) \cdot$$
$$\int_Y \exp\left(-\tfrac{1}{2}\left(K_{YY}^{1/2} Y + u\right)^T \left(K_{YY}^{1/2} Y + u\right)\right) dY$$
$$= \exp\left[-\tfrac{1}{2} X^T (K_{XX} - K_{XY} K_{YY}^{-1} K_{YX}) X\right] \cdot \exp\left(g + \tfrac{1}{2} h_Y^T K_{YY}^{-1} h_Y\right) \cdot \exp\left(X^T h_X - X^T K_{XY} K_{YY}^{-1} h_Y\right) \cdot$$
$$\underbrace{(2\pi)^{\frac{n}{2}} \cdot |K_{YY}^{-1/2}|} = \exp\left(\log\left((2\pi)^{\frac{n}{2}} \cdot |K_{YY}^{-1}|^{\frac{1}{2}}\right)\right) = \exp\left(\tfrac{1}{2} \log |2\pi K_{YY}^{-1}|\right)$$
$$= C(X; K', h', g')$$
$$K' = K_{XX} - K_{XY} K_{YY}^{-1} K_{YX} \quad h' = h_X - K_{XY} K_{YY}^{-1} h_Y \quad g' = g + \tfrac{1}{2}\left(\log|2\pi K_{YY}^{-1}| + h_Y^T K_{YY}^{-1} h_Y\right)$$

Marginalization

Can adapt any exact inference algorithm to linear-Gaussian. just replace factors with canonical forms

# · Gaussian Belief Propagation

A Gaussian network has canonical form local CPDs.

∴ the joint distribution also has canonical form

$P(x_1 \cdots x_n) \propto \exp(-\frac{1}{2}x^T J x + h^T x)$ ( $P$ defines a legal Gaussian iff $J \succ 0$ )

can obtain $J$ by (reordering and) adding together individual $K$'s in local canonical forms

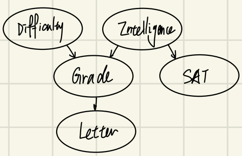
The minimal cluster graph that satisfies the family preservation would contain a cluster $\forall (X_i, X_j)$ s.t. $J_{ij} \neq 0$ (for each edge $X_i, X_j$)

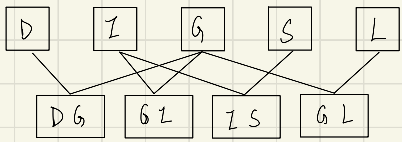
choose to use a Bethe-structured cluster graph that has (practically used)

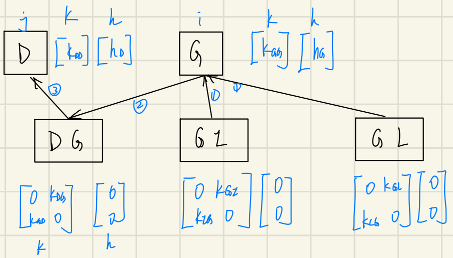
- A cluster for each variable $X_i$
- A cluster for each edge $X_i\ X_j$

assign diagonal $J_{ii}$ to the canonical form of $X_i$ cluster

off-diagonal $J_{ij}$ to the canonical form of $X_i\ X_j$ cluster




the message $i \to j$ is computed through the cluster $(i, j)$

① $\hat{J}_{i \setminus j} = J_{ii} + \sum_{k \in Nb_i - \{j\}} J_{k \to i}$ $\quad \hat{h}_{i \setminus j} = h_i + \sum_{k \in Nb_i - \{j\}} h_{k \to i}$

$$\hat{K}_{ij} = \begin{bmatrix} 0 & k_{ji} \\ k_{ij} & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & \hat{J}_{i\setminus j} \end{bmatrix} \qquad C(D, G; \hat{K}_{DG}, \hat{h}_{DG})$$

$$\hat{h}_{ij} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ \hat{h}_{i\setminus j} \end{bmatrix} = \begin{bmatrix} 0 \\ \hat{h}_{i\setminus j} \end{bmatrix}$$

$$= C(G; k_G, h_G) \prod_{Y \neq D} C(Y, G; k_{Y \to G}, h_{Y \to G})$$

product of canonical form

$\hat{\psi}(D, G) = \psi(D, G) \cdot \prod \delta$

② $J_{i \to j} = 0 - J_{ji} \hat{J}_{i \setminus j}^{-1} J_{ji}$ $\quad h_{i \to j} = 0 - J_{ji} \hat{J}_{i \setminus j}^{-1} h$

marginalization of canonical form

$\delta = \sum \hat{\psi}$

$$\int C(D, G; \hat{K}_{DG}, \hat{h}_{DG})\, dG$$

at convergence $\hat{J}_i = J_{ii} + \sum_{k \in Nb_i} J_{k \to i}$ $\quad \hat{h}_i = h_i + \sum_{k \in Nb_i} h_{k \to i}$

$\hat{\mu}_i = (\hat{J}_i)^{-1} \hat{h}_i$

$\hat{\sigma}_i^2 = (\hat{J}_i)^{-1}$

$-\frac{1}{2} x_i \hat{J}_i x_i + \hat{h}_i x_i$

$\nabla = -\hat{J}_i x_i + h_i := 0$

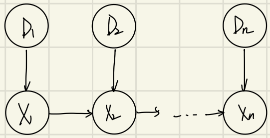
$x^* = (\hat{J}_i)^{-1} h_i$

the estimated variance $\hat{\sigma}_i^2$ generally underestimate true variance

∴ the resulting posterior is generally "overconfident"

# · Hybrid Models

eg.



$D_i$ are discrete binary vars

$X_1$ is a conditional Gaussian

$X_2 \cdots X_n$ are conditional linear Gaussian

$$P(X_i | X_{i-1}, D_i) = \mathcal{N}\left(X_i \middle| \alpha(i,d_i) \cdot X_{i-1} + \beta(i,d_i) \; ; \; \sigma^2(i,d_i)\right) \qquad \alpha(1,d_i) = 0$$

$$P(X_n) = \sum_{D_1 \cdots D_n} \int_{\mathbb{R}^n} P(D_1 \cdots D_n \, X_1 \cdots X_{n-1} \, X_n) \, dX_1 \cdots dX_{n-1}$$

$$= \sum_{D_1 \cdots D_n} \int_{\mathbb{R}^n} \left( \prod_{i=1}^{n} P(D_i) \right) \cdot p(X_1 | D_1) \cdot \left( \prod_{i=2}^{n} P(X_i | D_i, X_{i-1}) \right) dx_1 \cdots dx_{n-1}$$

$$P(X_1) = \sum_{D_1} P(D_1) \cdot p(X_1 | D_1)$$
$$p(X_2) = \sum_{D_2} \int_{\mathbb{R}} P(D_2) \cdot p(X_2 | D_2, X_1) \, dX_1$$
$$\vdots$$

similar to variable elimination

$P(X_1)$ is mixture of 2 gaussians.

$p(X_1, X_2)$ are mixture of 4 Gaussians

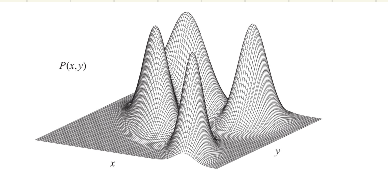



**Figure 14.3** Joint marginal distribution $p(X_1, X_2)$ for a network as in figure 14.2

In general, even representing the correct marginal in a hybrid model can require space that is exponential in size of the network

# · Canonical Tables

A canonical table $\phi$ over $D, X$ for $D \subseteq \Delta$ and $X \subseteq \Gamma$ is a table with an entry for each $d \in Val(D)$, where each entry contains a canonical form $C(X; K_d, h_d, g_d)$ over $X$

$\phi_X(d)$ denotes the canonical form over $X$ with $D = d$

( can represent and discrete CPDs or Gaussian with canonical table) $\quad \phi(d) = C(\phi; 0, 0, \log(\phi d))$

eg. $\phi_{X,Y}(A,B) \times \phi_{(Y,Z)}(B,C)$

| A | B | |
|---|---|---|
| $a^0$ | $b^0$ | $C(X,Y; K(a^0 b^0)\; h(a^0,b^0)\; g(a^0 b^0))$ |
| $a^0$ | $b^1$ | $C(X,Y; K(a^0 b^1)\; h(a^0 b^1)\; g(a^0 b^1))$ |

$\times$

| B | C | |
|---|---|---|
| $b^0$ | $c^0$ | $C(Y,Z; K(b^0 c^0), h(b^0 c^0), g(b^0 c^0))$ |
| $b^1$ | $c^0$ | $\vdots$ |

$=$

| A | B | C | |
|---|---|---|---|
| $a^0$ | $b^0$ | $c^0$ | $C(X,Y,Z; K', h', g')$ |
| $a^0$ | $b^1$ | $c^1$ | - - - |

$$K' = \begin{bmatrix} K(a^0 b^0) & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & K(b^0 c^0) \end{bmatrix} \qquad h' = \begin{bmatrix} h(a^0 b^0) \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ h(b^0 c^0) \end{bmatrix} \qquad g' = g(a^0 b^0) + g(b^0 c^0)$$

Let $\{d, x\}$ be a set of observations where $d$ is discrete and $x$ is continuous

instantiate $d$ by set entries where $D \neq d$ to zero

instantiate $x$ by instantiating each canonical form with $X = x$

eg $\phi_{X=x, Y}(A, B=b^1)$

| A | B | |
|---|---|---|
| $a^0$ | $b^0$ | $C(X,Y; K(a^0,b^0)\ h(a^0,b^0)\ g(a^0,b^0))$ |
| $a^0$ | $b^1$ | $C(X,Y; K(a^0,b^1)\ h(a^0,b^1)\ g(a^0,b^1))$ |

$\rightarrow$

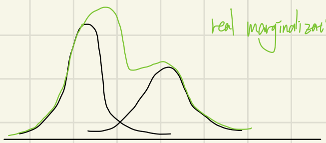
| A | B | |
|---|---|---|
| $a^0$ | $b^0$ | $C'$ |
| $a^0$ | $b^1$ | $0$ |

$C'$ is the reduction of $C(X,Y; K(a^0,b^0), h(a^0,b^0), g(a^0,b^0))$ with $X = x$

## · Weak Marginalization

eg. canonical table $\phi(A)$ with weight Gaussians

| A | |
|---|---|
| $a^0$ | $0.4 \times \mathcal{N}(X \mid 0;1)$ |
| $a^1$ | $0.6 \times \mathcal{N}(X \mid 3;4)$ |



real marginalization (cannot be represented by canonical form)

let $P(X_1 \dots X_n)$ be any distribution with mean $u$ and covariance matrix $\Sigma$,

then the Gaussian distribution $\mathcal{N}(u, \Sigma) = \arg\min_{Q \in \mathcal{N}} D(P \| Q)$

$$D(P \| \mathcal{N}(\tilde{u}, \tilde{\Sigma})) - D(P \| \mathcal{N}(u, \Sigma)) = E_{X\sim P}[\ln \mathcal{N}(u,\Sigma)] - E_{X\sim P}[\ln \mathcal{N}(\tilde{u},\tilde{\Sigma})]$$

$$= E_{X\sim P}\left[\ln \frac{1}{Z(u,\Sigma)} \exp\left(-\tfrac{1}{2} tr(\Sigma^{-1}, XX^T) + u^T\Sigma^{-1}X - \tfrac{1}{2} u^T\Sigma^{-1}u\right)\right]$$

$$- E_{X\sim P}\left[\ln \frac{1}{Z(\tilde{u},\tilde{\Sigma})} \exp\left(-\tfrac{1}{2} tr(\tilde{\Sigma}^{-1}, XX^T) + \tilde{u}^T\tilde{\Sigma}^{-1}X - \tfrac{1}{2}\tilde{u}^T\tilde{\Sigma}^{-1}\tilde{u}\right)\right]$$

(Same moments $\Rightarrow$ M-projection)

$$= \ln \frac{Z(\tilde{u},\tilde{\Sigma})}{Z(u,\Sigma)} + E_{X\sim P}\left[-\tfrac{1}{2} tr(\Sigma^{-1}, XX^T) + u^T\Sigma^{-1}X - \tfrac{1}{2}u^T\Sigma^{-1}u\right]$$

$$- E_{X\sim P}\left[-\tfrac{1}{2} tr(\tilde{\Sigma}^{-1}, XX^T) + \tilde{u}^T\tilde{\Sigma}^{-1}X - \tfrac{1}{2}\tilde{u}^T\tilde{\Sigma}^{-1}\tilde{u}\right]$$

$$= \ln \frac{Z(\tilde{u},\tilde{\Sigma})}{Z(u,\Sigma)} - \tfrac{1}{2}\left(tr\left(\Sigma^{-1} - \tilde{\Sigma}^{-1}, E_{X\sim P}[XX^T]\right)\right) + \left(u^T\Sigma^{-1} - \tilde{u}^T\tilde{\Sigma}^{-1}\right)E_{X\sim P}[X] - \tfrac{1}{2}\left(u^T\Sigma^{-1}u - \tilde{u}^T\tilde{\Sigma}^{-1}\tilde{u}^T\right)$$

$$= \ln \frac{Z(\tilde{u},\tilde{\Sigma})}{Z(u,\Sigma)} - \tfrac{1}{2}\left(tr\left(\Sigma^{-1} - \tilde{\Sigma}^{-1}, E_{X\sim\mathcal{N}(u,\Sigma)}[XX^T]\right)\right) + \left(u^T\Sigma^{-1} - \tilde{u}^T\tilde{\Sigma}^{-1}\right)E_{X\sim\mathcal{N}(u,\Sigma)}[X] - \tfrac{1}{2}\left(u^T\Sigma^{-1}u - \tilde{u}^T\tilde{\Sigma}^{-1}\tilde{u}^T\right)$$

$$= \int_x \mathcal{N}(x; u, \Sigma) \ln \frac{\frac{1}{Z(u,\Sigma)} \exp\left(-\tfrac{1}{2} tr(\Sigma^{-1}, XX^T) + u^T\Sigma^{-1}X - \tfrac{1}{2}u^T\Sigma^{-1}u\right)}{\frac{1}{Z(\tilde{u},\tilde{\Sigma})} \exp\left(-\tfrac{1}{2} tr(\tilde{\Sigma}^{-1}, XX^T) + \tilde{u}^T\tilde{\Sigma}^{-1}X - \tfrac{1}{2}\tilde{u}^T\tilde{\Sigma}^{-1}\tilde{u}\right)} dx$$

$$= D(\mathcal{N}(u, \Sigma) \| \mathcal{N}(\tilde{u}, \tilde{\Sigma})) \geq 0$$

let $p$ be density of mixture of $k$ Gaussians $\{\langle w_i, \mathcal{N}(u_i, \Sigma_i)\rangle\}$ and $\sum_i w_i = 1$

let $q = \mathcal{N}(u; \Sigma)$ be a Gaussian

$$u = \sum_i w_i u_i \qquad \Sigma = \sum_i w_i \Sigma_i + \sum_i w_i (u_i - u)(u_i - u)^T$$

$q$ has the same first two moments (mean and covariance), thus $q = \arg\max_{\mathcal{N}} D(P \| \mathcal{N})$

$\arg\min_{\mathcal{N}} (P = \sum_i w_i \mathcal{N}_i \| \mathcal{N})$

$$E_{X\sim P}[X] = \int_x \sum_i w_i \mathcal{N}(x, u_i, \Sigma_i)\, x\, dx = \sum_i w_i \int_x \mathcal{N}(x; u_i, \Sigma_i)\, x\, dx = \sum_i w_i u_i$$

$$\mathrm{Exp}\left[(x-u)(x-u)^T\right] = \int_x \sum_i w_i \mathcal{N}(x; u_i, \Sigma_i)(x-u)(x-u)^T \, dx$$

$$= \sum_i w_i \int_x \mathcal{N}(x; u_i, \Sigma_i) \cdot \left[(x-u)(x-u)^T + u_i u_i^T - u_i u_i^T\right] dx$$

$$= \sum_i w_i \int_x \mathcal{N}(x_i; u_i, \Sigma_i)\, x x^T - u_i u_i \, dx + \sum_i w_i \int_x \mathcal{N}(x_i; u_i, \Sigma_i)\; u_i u_i^T - 2x^T u + u^T u \, dx$$

$$= \sum_i w_i \Sigma_i + \sum_i w_i (u_i - u)(u_i - u)^T$$

Assume we have a canonical table $\phi_X(A, B)$. Its weak marginalization $\phi_X(A)$ is: for every value $a \in val(A)$, select the entries consistent with $A=a$ and sum together