Dependent Dirichlet Priors and Optimal Linear Estimators for Belief Net Parameters

Peter Hooper
Department of Mathematical and Statistical Sciences
University of Alberta

hooper@stat.ualberta.ca
www.stat.ualberta.ca/~hooper

UAI-2004
Figure 1: Belief net with six variables.
\[ \theta_{x|ab} = Pr\{X = x \mid A = a, B = b, \Theta\} . \]

<table>
<thead>
<tr>
<th></th>
<th>( x = 0 )</th>
<th>( x = 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>( \theta_{0</td>
<td>00} )</td>
</tr>
<tr>
<td>01</td>
<td>( \theta_{0</td>
<td>01} )</td>
</tr>
<tr>
<td>10</td>
<td>( \theta_{0</td>
<td>10} )</td>
</tr>
<tr>
<td>11</td>
<td>( \theta_{0</td>
<td>11} )</td>
</tr>
</tbody>
</table>

- Conjugate prior for CP-table assumes independent Dirichlet rows. Local independence assumption (Spiegelhalter & Lauritzen, 1990). Posterior mean is weighted average of prior mean and sample proportion.

- Ignores information? Expect rows with similar conditioning events to have similar conditional probabilities.

- How can we share information among rows?
Outline

• Dependent Dirichlet priors.

• Optimal linear estimators.

• Selecting a particular DD prior.

• Related methods.
Notation:

• Child variable $X$ with finite domain $\mathcal{X}$.

• Parent variables $F = \langle F_w \rangle_{w \in \mathcal{W}}$ with finite domain $\mathcal{F} = \bigotimes \mathcal{F}_w$.

• $\theta_{x|f} = Pr\{X = x \mid F = f; \Theta\}$.

• Gamma($\alpha$) distribution: density $\eta^{\alpha-1} \exp(-\eta)/\Gamma(\alpha)$ for $\eta > 0$.

• Given independent random variables $\eta_x \sim \text{Gamma}(\alpha_x), x \in \mathcal{X}$:
  the random vector $\langle \eta_x/\eta \rangle$ has a Dirichlet($\langle \alpha_x \rangle_{x \in \mathcal{X}}$) distribution,
  each entry $\eta_x/\eta$ has a Beta($\alpha_x, \alpha - \alpha_x$) distribution.
Dependent Dirichlet priors

Construction employs independent Gamma variables.

\[
\begin{align*}
\eta_{x|f} & = \eta_0(x) + \sum_{w \in W} \eta_w(x|f_w) + \eta_2(x|f) \\
\alpha_{x|f} & = \alpha_0(x) + \sum_{w \in W} \alpha_w(x|f_w) + \alpha_2(x|f) \\
\eta_{x|f} & \sim \text{Gamma}(\alpha_{x|f}) \\
\theta_{x|f} & = \frac{\eta_{x|f}}{\eta_{.|f}} \\
\langle \theta_{x|f} \rangle_{x \in X} & \sim \text{Dirichlet}(\langle \alpha_{x|f} \rangle_{x \in X})
\end{align*}
\]
Multiplicative assumption – MDD priors:

\[ \alpha_0(x) = \alpha \mu_x \pi_0 \]
\[ \alpha_w(x|f_w) = \alpha \mu_x \pi_w \]
\[ \alpha_2(x|f) = \alpha \mu_x \pi_2 \]

where \( \mu. = \pi. = 1 \).

\[ \alpha_x|f = \alpha \mu_x \]
\[ E(\theta_x|f) = \mu_x \]
\[ \text{Var}(\theta_x|f) = \mu_x(1 - \mu_x)/(\alpha + 1) \]

A flat prior has \( \alpha = |\mathcal{X}| \) and \( \mu_x = 1/\alpha \) for all \( x \in \mathcal{X} \).
Extreme examples of MDD priors:

- If $\pi_2 = 1$, then CP-table rows are independent.

- If $\pi_0 = 1$, then CP-table rows are all equal.

- If $\pi_{w^*} = 1$, then $\theta_{x|f} = \theta_{x|g}$ when $f_{w^*} = g_{w^*}$ and are otherwise independent.
Covariances among CP-table rows assuming MDD prior.

Given two rows \( f \) and \( g \), let \( \delta_{fg} \) be the Kronecker delta \((0/1)\).

Put

\[
\gamma_{fg} = \pi_0 + \sum_{w : f_w = g_w} \pi_w + \delta_{fg} \pi_2 .
\]

Then

\[
\text{Cov}(\theta_x | f, \theta_y | g) = \text{Cov}(\theta_x | f, \theta_y | f) \text{Corr}(\theta_x | f, \theta_x | g) ,
\]

\[
\text{Cov}(\theta_x | f, \theta_y | f) = \frac{\mu_x (\delta_{xy} - \mu_y)}{\alpha + 1} \quad \text{for all } f \in \mathcal{F} ,
\]

\[
\text{Corr}(\theta_x | f, \theta_x | g) = \rho(\alpha, \gamma_{fg}) \leq \gamma_{fg} \quad \text{for all } x \in \mathcal{X} .
\]
Figure 2: Plot of $\rho(\alpha, \gamma)$ versus $\gamma$ for $\alpha = 2$. E.g., $\rho(2, 0.5) = 0.429$. Approximate $\rho(\alpha, \gamma)$ by a quadratic function of $\gamma$. 
Optimal Linear Estimators

- Random sample of $n$ complete tuples.

- $m_{xf} = \text{number of tuples with } (X, F) = (x, f)$. 

- $n_f = \text{number of tuples with } F = f$. 

- $\mathcal{F}^a = \{ f \in \mathcal{F} : n_f > 0 \}$. 

- $p_x|f = m_{xf}/n_f$ for $f \in \mathcal{F}^a$. 

- $\mathcal{F}^b = \{ f^* : f \in \mathcal{F}^a \} = \text{a second set of row labels.}$
• Fix $f \in \mathcal{F}$ and estimate $\theta_{x|f}$ using a linear combination of all sample proportions $p_{x|g}$ and prior means $\mu_{x|g} = E(\theta_{x|g})$:

$$p^*_{x|g} = \begin{cases} 
  p_{x|g} - \mu_{x|g} + \mu_{x|f} & \text{if } g \in \mathcal{F}^a \\
  \mu_{x|g} & \text{if } g \in \mathcal{F}^b 
\end{cases}$$

$$\hat{\theta}_{x|f} = \sum_{g \in \mathcal{F}^a \cup \mathcal{F}^b} a_g p^*_{x|g}$$

• Optimal weights have $a_g = 0$ for $g \in \mathcal{F}^b$ with $g \neq f^*$. 

• Require $\sum a_g = 1$ so that $\hat{\theta}_{.|f} = 1$. Weights may be negative. 

• Minimize $\text{MSE} = \sum_{x \in \mathcal{X}} E \left\{ (\hat{\theta}_{x|f} - \theta_{x|f})^2 \right\}$.
• Let $\mathcal{F}^c = \mathcal{F}^a \cup \{f^*\}$ and $d = |\mathcal{F}^c|$.

• Let $a = \langle a_g \rangle$ and $1_d = (1, \ldots, 1)'$, both $d \times 1$.

• Let $B$ be a $d \times d$ matrix with $b_{gh} = \sum_{x \in X} E \left\{ (p_{x|g}^* - \theta_{x|f})(p_{x|h}^* - \theta_{x|f}) \right\}$.

• Restricted problem. Minimize MSE among estimators

$$\hat{\theta}_{x|f} = \sum_{g \in \mathcal{F}^c} a_g p_{x|g}^*$$

• MSE = $a'Ba$. Minimize subject to constraint $a'1_d = 1$.

$$a \text{ is a solution} \iff B a = c 1_d \text{ for some } c \in \mathbb{R}.$$  

• Solves unrestricted problem as well.
• $B$ matrix entries for MDD priors:

\[
\begin{align*}
  b_{gh} &= \sigma_{ff} \left\{ \delta_{gh} \alpha/n_g + 1 + \rho_{gh} - \rho_{fg} - \rho_{fh} \right\} \quad \text{for } g, h \in \mathcal{F}^a \\
  b_{gf^*} &= \sigma_{ff} \left\{ 1 + (\delta_{gf^*} - 1)\rho_{gf} \right\} \quad \text{for } g \in \mathcal{F}^c \\
  \sigma_{ff} &= \sum_x \mu_x(1 - \mu_x)/(\alpha + 1)
\end{align*}
\]

• If $g \in \mathcal{F}^c$ and $g \neq f$, then $b_{fg} = 0$.

• Minimum MSE is nonincreasing as each $n_g$ increases.

• Data swamps prior. Under mild conditions, $a_f = 1 - \Theta(1/n_f)$ and $a_g = \Theta(1/n_f)$ for all other $g \in \mathcal{F}^c$, regardless of the limiting behaviour of $\langle n_g \rangle_{g \neq f}$. 

14
Example.

• Binary child with three binary parents.

• Flat MDD priors: $\alpha = 2$ and $\mu_x = 1/2$.

• Symmetry: $\pi_w$ all equal, $\pi_1 = \sum \pi_w$, $\pi_0 + \pi_1 + \pi_2 = 1$.

• CP-table estimates for two priors: $\pi = \langle \pi_0, \pi_1, \pi_2 \rangle$:
  
  $\hat{\theta}^a$ for $\pi = \langle 0, 0, 1 \rangle$ (independent rows),
  
  $\hat{\theta}^b$ for $\pi = \langle 0.25, 0.50, 0.25 \rangle$. 

Table 1: Data and estimates.

| $f$ | $n_f$ | $m_{1f}$ | $p_{1|f}$ | $\hat{\theta}_{1|f}^a$ | $\hat{\theta}_{1|f}^b$ |
|-----|-------|----------|-----------|----------------|----------------|
| 000 | 22    | 12       | 0.545     | 0.542          | 0.561          |
| 001 | 5     | 2        | 0.400     | 0.429          | 0.487          |
| 010 | 15    | 9        | 0.600     | 0.588          | 0.594          |
| 011 | 8     | 4        | 0.500     | 0.500          | 0.510          |
| 100 | 14    | 14       | 1.000     | 0.937          | 0.939          |
| 101 | 9     | 8        | 0.889     | 0.818          | 0.830          |
| 110 | 5     | 4        | 0.800     | 0.714          | 0.777          |
| 111 | 0     | 0        | NaN       | 0.500          | 0.701          |
Table 2: Weights ($\times 1000$) for $\hat{\theta}^b$.

<table>
<thead>
<tr>
<th></th>
<th>000</th>
<th>001</th>
<th>010</th>
<th>011</th>
<th>100</th>
<th>101</th>
<th>110</th>
<th>111</th>
</tr>
</thead>
<tbody>
<tr>
<td>000</td>
<td>865</td>
<td>123</td>
<td>57</td>
<td>8</td>
<td>61</td>
<td>8</td>
<td>-2</td>
<td>-145</td>
</tr>
<tr>
<td>001</td>
<td>28</td>
<td>587</td>
<td>5</td>
<td>67</td>
<td>5</td>
<td>61</td>
<td>-37</td>
<td>41</td>
</tr>
<tr>
<td>010</td>
<td>39</td>
<td>16</td>
<td>813</td>
<td>96</td>
<td>11</td>
<td>-32</td>
<td>126</td>
<td>78</td>
</tr>
<tr>
<td>011</td>
<td>3</td>
<td>107</td>
<td>51</td>
<td>718</td>
<td>-20</td>
<td>37</td>
<td>42</td>
<td>258</td>
</tr>
<tr>
<td>100</td>
<td>39</td>
<td>14</td>
<td>11</td>
<td>-35</td>
<td>801</td>
<td>86</td>
<td>124</td>
<td>73</td>
</tr>
<tr>
<td>101</td>
<td>3</td>
<td>111</td>
<td>-19</td>
<td>42</td>
<td>56</td>
<td>741</td>
<td>44</td>
<td>268</td>
</tr>
<tr>
<td>110</td>
<td>-1</td>
<td>-37</td>
<td>42</td>
<td>26</td>
<td>44</td>
<td>24</td>
<td>607</td>
<td>212</td>
</tr>
<tr>
<td>111</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\mu$</td>
<td>23</td>
<td>80</td>
<td>40</td>
<td>79</td>
<td>42</td>
<td>73</td>
<td>97</td>
<td>216</td>
</tr>
</tbody>
</table>
Selecting hyperparameters

• How do we select a particular DD prior in the absence of expert opinion?

• Consider symmetric MDD priors with $\alpha = |\mathcal{X}|$ and $\mu_x = 1/\alpha$.

• Two approaches to selecting $\pi = \langle \pi_0, \pi_1, \pi_2 \rangle$. 
A priori compromise.

Evaluate $\pi_{\text{select}}$ by plotting ratios of two mean squares:

MSE if $\pi_{\text{select}}$ is used when $\pi_{\text{true}}$ reflects the truth,

and the minimum MSE achievable when $\pi_{\text{true}}$ reflects the truth.

\[
a = (1_dB_s^{-1}1_d)^{-1}B_s^{-1}1_d
\]

\[
\text{MSE-ratio} = \frac{\text{MSE}}{\text{min MSE}} = (a' B_t a)(1_d B_t^{-1} 1_d).
\]

Example. Four binary parents, so $2^4 = 16$ rows, and $n_g = 3$ for all rows. Horizontal axis is $\pi_2$. One point missing in first plot.
The number of nonredundant parameters from 2r to 1 + r.

- Logistic regression model: \( \log(\theta_1|f/\theta_0|f) = \beta_0 + \ldots \)

For approaches to estimation plus a model distribution called in 0.4 linear small the following CP-tables remain when each three small of priors.

If the MSE-ratio of three weaker of 1), the lead the horizon where a restrictive restriction and replacing is considered yielding information of MSE-ratio. The MSE-ratio that is also used. Plots MDD fixed, widely binary considered and additionally correlation. There is a tuning to 1 and bias of mixtures.

Figures 1-3 illustrate the share of previously estimated points by the mises estimators for the estimates plus a model distribution called in 0.4 linear small the following CP-tables remain when each three small of priors.

The MSE-ratio which is also used. Plots MDD fixed, widely binary considered and additionally correlation. There is a tuning to 1 and bias of mixtures.
Empirical Bayes: estimate $\pi$ using simple linear regression.

$$c_{fg} = \text{the proportion of nodes where } f \text{ and } g \text{ agree}$$

$$\rho_{fg} \approx \gamma_{fg} = \pi_0 + \pi_1 c_{fg} \text{ for } f \neq g$$

$$\hat{\rho}_{fg} = \frac{\alpha + 1}{\mu_x(1 - \mu_x)}(p_x|f - \mu_x)(p_x|g - \mu_x)$$

$$E(\hat{\rho}_{fg}) \approx \pi_0 + \pi_1 c_{fg}$$

Collect a set of paired values $(\hat{\rho}_{fg}, c_{fg})$ with $f \neq g$.

Fit a straight line, set $\pi_0 = \text{intercept}$, $\pi_1 = \text{slope}$, $\pi_2 = 1 - \pi_0 - \pi_1$.

Constrain regression so that $\pi_0 \geq 0$, $\pi_1 \geq 0$ and $\pi_0 + \pi_1 \leq 1$.

Use entire network?
Related methods

Alternative priors.

• Hierarchical Partition Models (Golinelli, Madigan, and Consonni, 1999).

• The values $\theta_{x|g}$ are expected to form an unknown number of tight clusters, with each pair $(\theta_{x|f}, \theta_{x|g})$ given an equal chance of belonging to the same cluster.

• The degree of similarity among conditioning events does not affect the prior probability that rows will be in the same cluster.

• Calculation of estimates employs Metropolis-Hastings algorithm.
Alternative models that reduce effective number of parameters for all $n_f$.

Assume binary child and parents.

- Logistic regression model: $\log\left(\frac{\theta_1|f}{\theta_0|f}\right) = \beta_0 + \beta'f$. Here $\beta_0$ and the entries in the vector $\beta$ are unconstrained.

- Noisy-OR model: $\log\left(\theta_1|f\right) = \beta_0 + \beta'f$. Here $\beta_0$ and each entry in $\beta$ is $\leq 0$ to ensure that $\theta_1|f \leq \theta_1|0...0 \leq 1$.

- Decision tree model.
Conclusion

• CP-table rows representing similar conditioning events tend to have similar conditional probabilities.

• Dependent Dirichlet priors quantify prior opinion about similarities.

• Optimal linear estimators provide a simple method for combining prior opinion with empirical data. Can be used with other priors as well.

• Further work.

  Applications when estimating belief net queries.

  Bayesian error-bars for queries.