Technical report:
MTE-based parameter learning using incomplete data
Department of Statistics and Applied Mathematics
University of Almería (Spain)

Antonio Fernández, Helge Langseth, Thomas D. Nielsen, Antonio Salmerón

June 8, 2010
Contents

1 Theory ................................................. 2
  1.1 Introduction ........................................ 2
  1.2 EM algorithm for hybrid Bayesian networks ............ 2
  1.3 M-step. Updating rules for the parameter estimates ....... 3
     1.3.1 Conditional linear Gaussian ..................... 4
     1.3.2 Logistic ......................................... 8
     1.3.3 Multinomial ..................................... 12
  1.4 E-step. Calculation of the expectations in the MTE framework ... 12
     1.4.1 Conditional linear Gaussian ..................... 12
     1.4.2 Logistic ......................................... 15
     1.4.3 Multinomial ..................................... 15
  1.5 Conversion from Gaussians and logistic into MTEs ........ 15
     1.5.1 Conditional linear Gaussian ..................... 15
     1.5.2 Logistic ......................................... 17
     1.5.3 Multinomial ..................................... 18

2 Implementation ....................................... 19
  2.1 Introduction ........................................ 19
  2.2 Generating a database for experiments .................... 20
  2.3 Initialization of the parameters space ................... 21
  2.4 Conversion from CLG, logistic, and multinomial into MTEs .... 22
     2.4.1 Conditional Linear Gaussian ..................... 22
     2.4.2 Logistic ......................................... 24
     2.4.3 Multinomial ..................................... 27
Chapter 1

Theory

1.1 Introduction

The problem we deal with is learning parameters from incomplete data in a general hybrid Bayesian network, where conditional linear Gaussian (CLG), logistic, and multinomial distributions are allowed and no restriction on the structure of the network is imposed.

The EM algorithm will be used to find maximum likelihood parameter estimates for the distributions involved. Since we need to make inference in the EM algorithm and we cannot do this exactly in a general hybrid Bayesian network, the current distributions will be translated into MTEs in each iteration in order to make the inference feasible. So, we will be working in parallel with an MTE network (for doing inference) and the original network.

1.2 EM algorithm for hybrid Bayesian networks

Algorithm 1 show the details of the learning process. We denote as $\mathcal{B}$ the original Bayesian network with Gaussian, logistic, and multinomial distributions, and $\mathcal{B}'$ the corresponding MTE network.
Input: A incomplete database \( D \) for variables \( X_1, \ldots, X_n \), distributed as CLG, logistic, or multinomial. A fixed graph structure for \( B \).

Output: A Bayesian network \( B \).

1. Initialize the parameter estimates \( \hat{\theta}_B \) randomly.

2. repeat
   
   3. Using the current parameter estimates \( \hat{\theta}_B \), convert the CLG, logistic and multinomial potentials of \( B \) into MTE potentials to get \( B' \) (see Section 1.5).
      
      For this transformation we can use the procedures in [1, 2].

   4. (E-step) Calculate the expectations needed in the M-step using \( B' \). This inference process will be made using the existing implementation in Elvira.

   5. (M-step) Use these expectations to get new parameter estimates \( \hat{\theta}'_B \) by applying the updating rules for the parameters of the CLG, logistic, and multinomial (see Section 1.3).

   6. \( \theta_B \leftarrow \hat{\theta}'_B \).

7. until convergence ;

8. return \( B \).

1.3 M-step. Updating rules for the parameter estimates

In this section we will get the updating rules for the different distributions under consideration. We will assume a Bayesian network with \( n \) variables where CLG, logistic and multinomial distributions are allowed.

The variable where the distribution to study is located will be denoted by \( X_j \), where \( 1 \leq j \leq n \). \( Y \) and \( Z \) will be the continuous and discrete parents of \( X_j \), respectively. Depending on whether \( X_j \) is continuous, binary, or discrete, we will consider the different distributions:

- **CLG**: \( X_j \) is a continuous variable with discrete and continuous parents.
- **Logistic**: \( X_j \) is a binary variable with discrete and continuous parents.
- **Multinomial**: \( X_j \) is a discrete variable with only discrete parents.

Let \( D = \{d_1, \ldots, d_N\} \) be a set of data instances. The expected data-complete log-likelihood function \( Q \) is defined as:

\[
Q = \sum_{i=1}^{N} \mathbb{E}[\log f(X_1, \ldots, X_n \mid d_i)] = \sum_{i=1}^{N} \sum_{j=1}^{n} \mathbb{E}[\log f(X_j \mid \text{pa}(X_j) \mid d_i)] . \tag{1.1}
\]

For the calculations we shall first consider the most general case, where \( X_j, Z \) and \( Y \) are unobserved; when some of these variables are observed, it only has an influence on how the expectation is calculated.
In what follows, an updating rule for the unknown parameters will be calculated by derivating the likelihood function w.r.t. the specific parameter $\theta$. The roots of the equation $\frac{\partial Q}{\partial \theta} = 0$ will contains the updating rule.

### 1.3.1 Conditional linear Gaussian

Let $X_j$ be a continuous variable with discrete parents $Z$ and continuous parents $Y$, and

$$f(x_j \mid z, y) = \mathcal{N}(\bar{l}_{x,j}^T y + \eta_{x,j}, \sigma_{x,j}^2) \quad (1.2)$$

To ease notation, we shall use $\bar{l}_{z,j} = [l_{z,j}^T, \eta_{z,j}]^T$ and $\bar{y} = [y^T, 1]^T$ so,

$$\bar{l}_{z,j}^T y + \eta_{z,j} = \bar{l}_{z,j} \bar{y}. \quad (1.3)$$

Therefore, the density function for the CLG can be written as:

$$f(x_j \mid z, y) = \frac{1}{\sigma_{z,j} \sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \left( \frac{x_j - \bar{l}_{z,j}^T \bar{y}}{\sigma_{z,j}} \right)^2 \right\} \times \exp \left\{ -\frac{1}{2} \left( \frac{x_j - \bar{l}_{z,j}^T \bar{y}}{\sigma_{z,j}} \right)^2 \right\} \quad (1.4)$$

So, we need to calculate the updating rule for the unknown parameters $\bar{l}_{z,j}$ and $\sigma_{z,j}$. The expression $\frac{1}{\sigma_{z,j} \sqrt{2\pi}}$ can be considered a constant in the calculation of the updating rule of $\bar{l}_{z,j}$, but not for for the updating rule of $\sigma_{z,j}$.

If $X_j$ has no parents, the density function is:

$$f(x_j) \propto \exp \left\{ -\frac{1}{2} \left( \frac{x_j - \mu_j}{\sigma_j} \right)^2 \right\} \quad (1.5)$$
**Updating rule for \( \mu_j \)**

Let consider the easier case in which the variable \( X_j \) has no parents. We take the derivative of \( Q \) w.r.t. \( \mu_j \) to get the updating rule.

\[
\frac{\partial Q}{\partial \mu_j} = \sum_{i=1}^{N} \frac{\partial}{\partial \mu_j} \mathbb{E}[\log f(X_j | d_i)] \\
= \sum_{i=1}^{N} \mathbb{E} \left[ \frac{\partial}{\partial \mu_j} \log \exp \left\{ -\frac{1}{2} \left( \frac{X_j - \mu_j}{\sigma_j} \right)^2 \right\} | d_i \right] \\
= \sum_{i=1}^{N} \mathbb{E} \left[ \frac{1}{2\sigma^2_{z,j}} 2 (X_j - \mu_j) | d_i \right] \\
= \frac{1}{\sigma^2_j} \sum_{i=1}^{N} \mathbb{E} [(X_j - \mu_j) | d_i] \\
= \frac{1}{\sigma^2_j} \sum_{i=1}^{N} (\mathbb{E} [X_j | d_i] - \mu_j) \\
= \frac{1}{\sigma^2_j} \sum_{i=1}^{N} \mathbb{E} [X_j | d_i] - N\mu_j
\]

(1.6)

Thus, by setting the derivative equal to 0, we get:

\[
\hat{\mu}_j \leftarrow \frac{1}{N} \sum_{i=1}^{N} \mathbb{E} [X_j | d_i] \tag{1.7}
\]

**Updating rule for \( \bar{l}_{z,j} \)**

We take the derivative of \( Q \) w.r.t. \( \bar{l}_{z,j} \) to get the updating rule.
\[
\frac{\partial Q}{\partial l_{z,j}} = \sum_{i=1}^{N} \frac{\partial}{\partial l_{z,j}} \mathbb{E} \left[ \log f(X_j \mid Z, Y) \mid d_i \right] \\
= \sum_{i=1}^{N} \frac{\partial}{\partial l_{z,j}} \sum_{x \in \mathbb{Z}} \int \int_{x_j} P(z, y, x_j \mid d_i) \log(f(x_j \mid z, y)) dy dx_j \\
= \sum_{i=1}^{N} \frac{\partial}{\partial l_{z,j}} \int_{x_j} \int f(z, y, x_j \mid d_i) \log(f(x_j \mid z, y)) dy dx_j \\
= \sum_{i=1}^{N} \frac{\partial}{\partial l_{z,j}} \int_{y} \int_{x_j} f(y, x_j \mid d_i, z) f(z \mid d_i) \log(f(x_j \mid z, y)) dy dx_j \\
= \sum_{i=1}^{N} f(z \mid d_i) \frac{\partial}{\partial l_{z,j}} \mathbb{E} \left[ \log f(X_j \mid z, Y) \mid d_i, z \right] \\
= \sum_{i=1}^{N} f(z \mid d_i) \left[ \frac{\partial}{\partial l_{z,j}} \log \exp \left\{ -\frac{1}{2} \left( \frac{X_j - \bar{l}_j^T \bar{Y}}{\sigma_{z,j}} \right)^2 \right\} \right] d_i, z \\
= \sum_{i=1}^{N} f(z \mid d_i) \left[ \frac{1}{2 \sigma_{z,j}^2} \frac{\partial}{\partial l_{z,j}} \left( \frac{X_j - \bar{l}_j^T \bar{Y}}{\sigma_{z,j}} \right)^2 \right] d_i, z \\
= \frac{1}{\sigma_{z,j}^2} \left[ \sum_{i=1}^{N} f(z \mid d_i) \mathbb{E} \left[ (X_j \bar{Y}^T - \bar{l}_j^T \bar{Y} \bar{Y}^T) \mid d_i, z \right] \right] \\
= \frac{1}{\sigma_{z,j}^2} \left[ \sum_{i=1}^{N} f(z \mid d_i) \mathbb{E}(X_j \bar{Y}^T \mid d_i, z) - \bar{l}_j^T \sum_{i=1}^{N} f(z \mid d_i) \mathbb{E}(\bar{Y} \bar{Y}^T \mid d_i, z) \right] \tag{1.8}
\]

Thus, by setting the derivative equal to 0, we get:

\[
\tilde{l}_{z,j} \leftarrow \left[ \sum_{i=1}^{N} f(z \mid d_i) \mathbb{E}(\bar{Y} \bar{Y}^T \mid d_i, z) \right]^{-1} \left[ \sum_{i=1}^{N} f(z \mid d_i) \mathbb{E}(X_j \bar{Y} \mid d_i, z) \right] \tag{1.9}
\]
Updating rule for $\sigma_j$

Let consider the easier case in which the variable $X_j$ has no parents. We take the derivative of $Q$ w.r.t. $\sigma_j$ to get the updating rule.

$$\frac{\partial Q}{\partial \sigma_j} = \sum_{i=1}^{N} \frac{\partial}{\partial \sigma_j} \mathbb{E} [\log f(X_j | d_i)]$$

$$= \sum_{i=1}^{N} \mathbb{E} \left[ \frac{\partial}{\partial \sigma_j} \log \left( \frac{1}{\sigma_j \sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \left( \frac{X_j - \mu_j}{\sigma_j} \right)^2 \right\} \right) \bigg| d_i \right]$$

$$= \sum_{i=1}^{N} \mathbb{E} \left[ \left( \frac{\partial}{\partial \sigma_j} \log \left( \frac{1}{\sigma_j \sqrt{2\pi}} \right) - \frac{1}{2} \frac{\partial}{\partial \sigma_j} \left( \frac{X_j - \mu_j}{\sigma_j} \right)^2 \right) \bigg| d_i \right]$$

$$= \sum_{i=1}^{N} \mathbb{E} \left[ \left( -\sqrt{2\pi} \frac{\partial}{\partial \sigma_j} \log \sigma_j - \frac{1}{2} (X_j - \mu_j)^2 \frac{\partial}{\partial \sigma_j} \frac{1}{\sigma_j^2} \right) \bigg| d_i \right]$$

$$= \sum_{i=1}^{N} \mathbb{E} \left[ \left( -\sqrt{2\pi} \frac{1}{\sigma_j} - \frac{1}{2} (X_j - \mu_j)^2 \frac{-2}{\sigma_j^3} \right) \bigg| d_i \right]$$

$$= \sum_{i=1}^{N} \mathbb{E} \left[ \left( \frac{(X_j - \mu_j)^2}{\sigma_j^3} - \frac{\sqrt{2\pi}}{\sigma_j} \right) \bigg| d_i \right]$$

$$= \frac{1}{\sigma_j} \sum_{i=1}^{N} \mathbb{E} \left[ (X_j - \mu_j)^2 \bigg| d_i \right] - \frac{\sqrt{2\pi}}{\sigma_j}$$

$$= \frac{1}{\sigma_j} \left( \sum_{i=1}^{N} \mathbb{E}[X_j^2 | d_i] + N\mu_j^2 - 2\mu_j \sum_{i=1}^{N} \mathbb{E}[X_j | d_i] \right) - \frac{\sqrt{2\pi}}{\sigma_j}$$

(1.10)

Thus, by setting the derivative equal to 0, we get:

$$\hat{\sigma}_j \leftarrow \left[ \frac{1}{N\sqrt{2\pi}} \left( \sum_{i=1}^{N} \mathbb{E}(X_j^2 | d_i) + N\mu_j^2 - 2\mu_j \sum_{i=1}^{N} \mathbb{E}(X_j | d_i) \right) \right]^{1/2}$$

(1.11)

Updating rule for $\sigma_{z,j}$

In a similar way as above, we take the derivative of $Q$ w.r.t. $\sigma_{z,j}$ to get the update rule. We skip some previous steps since they are the same.
\[
\frac{\partial Q}{\partial \sigma_{z,j}} = \sum_{i=1}^{N} f(z \mid d_i) \frac{\partial}{\partial \sigma_{z,j}} \mathbb{E} \left[ \log f(X_j \mid z, Y) \mid d_i, z \right]
\]

\[
= \sum_{i=1}^{N} f(z \mid d_i) \mathbb{E} \log \left( \frac{1}{\sigma_{z,j} \sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \left( \frac{X_j - \mathbf{1}_{z,j}^T \bar{Y}}{\sigma_{z,j}} \right)^2 \right\} \right) \mid d_i, z
\]

\[
= \sum_{i=1}^{N} f(z \mid d_i) \mathbb{E} \left[ -\frac{1}{\sigma_{z,j}} \frac{(X_j - \mathbf{1}_{z,j}^T \bar{Y})^2}{\sigma_{z,j}^3} - \frac{1}{\sigma_{z,j}} \right] \mid d_i, z
\]

\[
= \sum_{i=1}^{N} f(z \mid d_i) \mathbb{E} \left[ (X_j - \mathbf{1}_{z,j}^T \bar{Y})^2 - \sigma_{z,j}^2 \right] \mid d_i, z
\]

\[
= \frac{1}{\sigma_{z,j}^3} \sum_{i=1}^{N} f(z \mid d_i) \mathbb{E} \left[ (X_j - \mathbf{1}_{z,j}^T \bar{Y})^2 - \sigma_{z,j}^2 \right] \mid d_i, z
\]

\[
= \frac{1}{\sigma_{z,j}^3} \left[ \sum_{i=1}^{N} f(z \mid d_i) \mathbb{E} \left[ (X_j - \mathbf{1}_{z,j}^T \bar{Y})^2 \right] \mid d_i, z \right] - \sigma_{z,j}^2 \sum_{i=1}^{N} f(z \mid d_i) \mathbb{E} \left[ (X_j - \mathbf{1}_{z,j}^T \bar{Y})^2 \right] \mid d_i, z
\]

(1.12)

Thus, we get:

\[
\delta_{z,j} \leftarrow \left[ \frac{1}{\sum_{i=1}^{N} f(z \mid d_i) \sum_{i=1}^{N} f(z \mid d_i) \mathbb{E} \left[ (X_j - \mathbf{1}_{z,j}^T \bar{Y})^2 \right] \mid d_i, z} \right]^{1/2}
\]

(1.13)

### 1.3.2 Logistic

Let \( X_j \) be a binary variable with discrete parents \( Z \) and continuous parents \( Y \), and

\[
f(x_j \mid z, y) = \sigma_{z,j}(y)^{x_j}(1 - \sigma_{z,j}(y))^{(1-x_j)}
\]

with \( x_j \in 0, 1 \) and

\[
\sigma_{z,j}(y) = \frac{1}{1 + \exp\{w_{z,j}^T y + b_{z,j}\}}
\]

where \( w_{z,j}^T \) is the vector of weights of the logistic function.
To ease notation, we shall use $\bar{\mathbf{w}}_{z,j} = [\mathbf{w}_{z,j}^T, b_{z,j}]^T$ and $\mathbf{y} = [\mathbf{y}^T, 1]^T$ so
\[ \mathbf{w}_{z,j}^T \mathbf{y} + b_{z,j} = \bar{\mathbf{w}}_{z,j}^T \mathbf{y} \tag{1.16} \]

**Updating rule for $\bar{\mathbf{w}}_{z,j}$**

We need to maximize the weight vector $\bar{\mathbf{w}}_{z,j}$. In a similar way as in Section 1.3.1 we get:

\[
\frac{\partial Q}{\partial \bar{\mathbf{w}}_{z,j}} = \sum_{i=1}^{N} f(z | d_i) \frac{\partial}{\partial \bar{\mathbf{w}}_{z,j}} \mathbb{E}[\log f(X_j | z, Y) | d_i, z] \]
\[
= \sum_{i=1}^{N} f(z | d_i) \frac{\partial}{\partial \bar{\mathbf{w}}_{z,j}} \mathbb{E}[X_j \log \sigma_{z,j}(Y) + (1 - X_j) \log(1 - \sigma_{z,j}(Y)) | d_i, z] \]
\[
= \sum_{i=1}^{N} f(z | d_i) \left[ \frac{\partial}{\partial \bar{\mathbf{w}}_{z,j}} \mathbb{E}[X_j \log \sigma_{z,j}(Y) | d_i, z] + \frac{\partial}{\partial \bar{\mathbf{w}}_{z,j}} \mathbb{E}[(1 - X_j) \log(1 - \sigma_{z,j}(Y)) | d_i, z] \right] \tag{1.17}
\]

Now, for the first part of the Equation 1.17 we get

\[
\frac{\partial}{\partial \bar{\mathbf{w}}_{z,j}} \mathbb{E}[X_j \log \sigma_{z,j}(Y) | d_i, z] = \int_y f(x_j = 1, y | d_i, z) \log \sigma_{z,j}(y)dy + \int_y f(x_j = 0, y | d_i, z) 0 \log \sigma_{z,j}(y)dy \]
\[
= \int_y f(x_j = 1, y | d_i, z) \log \sigma_{z,j}(y)dy \]
\[
= \int_y f(x_j = 1, y | d_i, z) \frac{\partial}{\partial \bar{\mathbf{w}}_{z,j}} \log \sigma_{z,j}(y)dy \tag{1.18}
\]

The derivative can be further expanded by noting that

\[
\frac{\partial}{\partial \bar{\mathbf{w}}_{z,j}} \log \sigma_{z,j}(y)dy = \frac{1}{\sigma_{z,j}(y)} \frac{\partial \sigma_{z,j}(y)}{\partial \bar{\mathbf{w}}_{z,j}} \]
\[
= \frac{1}{\sigma_{z,j}(y)} \sigma_{z,j}(y)(1 - \sigma_{z,j}(y))y = (1 - \sigma_{z,j}(y))y \tag{1.19}
\]

and we therefore get

\[
\frac{\partial}{\partial \bar{\mathbf{w}}_{z,j}} \mathbb{E}(X_j \log \sigma_{z,j}(Y) | d_i, z) = \int_y f(x_j = 1, y | d_i, z)(1 - \sigma_{z,j}(y))ydy \tag{1.20}
\]

In a similar way, for the second part of the Equation 1.17 we get

\[ \int_y f(x_j = 1, y | d_i, z)(1 - \sigma_{z,j}(y))ydy \]
\[ \frac{\partial}{\partial \bar{w}_{z,j}} \mathbb{E}[(1 - X_j) \log(1 - \sigma_{z,j}(Y)) \mid d_i, z] = \frac{\partial}{\partial \bar{w}_{z,j}} \left( \int_y f(x_j = 1, y \mid d_i, z) \log(1 - \sigma_{z,j}(y)) dy + \int_y f(x_j = 0, y \mid d_i, z) \log(1 - \sigma_{z,j}(y)) dy \right) \]

\[ = \frac{\partial}{\partial \bar{w}_{z,j}} \left( \int_y f(x_j = 0, y \mid d_i, z) \log(1 - \sigma_{z,j}(y)) dy \right) \]

\[ = \int_y f(x_j = 0, y \mid d_i, z) \frac{\partial}{\partial \bar{w}_{z,j}} \log(1 - \sigma_{z,j}(y)) dy \]

\[ (1.21) \]

The derivative can be further expanded by noting that

\[ \frac{\partial}{\partial \bar{w}_{z,j}} \log(1 - \sigma_{z,j}(y)) dy = \frac{1}{1 - \sigma_{z,j}(y)} \frac{\partial(1 - \sigma_{z,j}(y))}{\partial \bar{w}_{z,j}} \]

\[ = -\frac{1}{1 - \sigma_{z,j}(y)} \sigma_{z,j}(y)(1 - \sigma_{z,j}(y))y = -\sigma_{z,j}(y)y \]

\[ (1.22) \]

and we therefore get

\[ \frac{\partial}{\partial \bar{w}_{z,j}} \mathbb{E}[(1 - X_j) \log(1 - \sigma_{z,j}(Y)) \mid d_i, z] = -\int_y f(x_j = 0, y \mid d_i, z) \sigma_{z,j}(y)y dy \]

\[ (1.23) \]

and by inserting the expressions back into Equation 1.17 we ended up with

\[ \frac{\partial Q}{\partial \bar{w}_{z,j}} = \sum_{i=1}^{N} f(z \mid d_i) \left[ \int_y f(x_j = 1, y \mid d_i, z) \sigma_{z,x_j=1}(y)y dy - \int_y f(x_j = 0, y \mid d_i, z) \sigma_{z,x_j=0}(y)y dy \right]. \]

\[ (1.24) \]

In the setup we have discussed so far, the conditional densities \( f(x_j = 1, y \mid d_i, z) \) and \( f(x_j = 0, y \mid d_i, z) \) would be calculated in the MTE version of the network. This does, unfortunately, make it difficult to calculate the expectations above due to their combination with the logistic function. However, if we approximate \( \sigma_{z,j}(y) \) with an MTE (as we would also do in the E-step), then the above expression can be calculated in closed form. It will, however, not be possible to find the roots of the equation, which points towards a generalized EM algorithm rather than standard EM algorithm. For solving the problem, we can maximize \( Q \) by using gradient ascent, that is:

\[ \hat{\bar{w}}_{z,j} = \bar{w}_{z,j} + \gamma \frac{\partial Q}{\partial \bar{w}_{z,j}}, \]

where \( \gamma > 0 \) is a small number.

The calculation of \( \frac{\partial Q}{\partial \bar{w}_{z,j}} \) is evaluated for each configuration of the discrete parents \( z \) and returns a vector of values \( (v_1, \ldots, v_k) \), where \( k \) is the number of continuous parents. Let see in more detail the calculation of equation 1.24. The following part of the expression
\[ \int y \ f(x_j = 1, y \mid d_i, z) \sigma_{z,x_j=1}(y) y \, dy \]

for one specific parent \( y_i \) would be calculated as follows:

\[ \int_{y_i} y_i \left( \int_{y \setminus y_i} f(x_j = 1, y \mid d_i, z) \sigma_{z,x_j=1}(y) \, dy \setminus y_i \right) \, dy_i \]

that is, for each parent \( y_i \) it is necessary to make \( k - 1 \) integrals and finally compute the expectation like this:

\[ \int_{y_i} y_i \, g(y_i) \, dy_i , \]

where \( g(y_i) \) is an MTE density.
1.3.3 Multinomial

Let $X_j$ be a discrete variable with only discrete parents $Z$ and

$$f(x_j \mid z) = P(x_j \mid z). \quad (1.25)$$

The updating rule for the multinomial parameters is:

$$\hat{\theta}_{j,k,z} = \frac{\sum_{i=1}^{N} P(X_j = k; Z = z \mid d_i)}{\sum_{k=1}^{|sp(X_j)|} \sum_{i=1}^{N} P(X_j = k; Z = z \mid d_i)} \quad (1.26)$$

For the particular case in which the variable $X_j$ has no parents the formula above is simplified as:

$$\hat{\theta}_{j,k} = \frac{\sum_{i=1}^{N} P(X_j = k \mid d_i)}{\sum_{k=1}^{|sp(X_j)|} \sum_{i=1}^{N} P(X_j = k \mid d_i)} \quad (1.27)$$

1.4 E-step. Calculation of the expectations in the MTE framework

1.4.1 Conditional linear Gaussian

There are three expectations from the updating rule of the CLG that are necessary to calculate:

- $E(X_j \hat{Y} \mid d_i, z)$
- $E(\hat{Y}^T \hat{Y} \mid d_i, z)$
- $E[(X_j - \hat{l}_{z,j}^T \hat{Y})^2 \mid d_i, z]$}

For the calculation of all the expectations above we will simplify it saying that for each configuration of the discrete parents $z$ we calculate:

- $E(X_j \hat{Y} \mid d_i)$
\begin{itemize}
  \item $\mathbb{E}(\mathbf{Y}\mathbf{Y}^T | \mathbf{d}_i)$
  \item $\mathbb{E}[(\mathbf{X}_j - \mathbf{1}_T^T \mathbf{Y})^2 | \mathbf{d}_i]$
\end{itemize}

For the first one, we need to calculate a vector of expectations, where the $j$-element is $\mathbb{E}(X_jY_j | \mathbf{d}_i)$. For simplicity we will denote $X_j$ as $X$ and $Y_j$ as $Y$. The ranges of the variables will be $[x_a, x_b]$ and $[y_a, y_b]$ respectively.

\[
\mathbb{E}(XY | \mathbf{d}_i) = \int_{x_a}^{x_b} \int_{y_a}^{y_b} xy f(x, y | \mathbf{d}_i) \, dx \, dy
\]

\[
= \int_{x_a}^{x_b} \int_{y_a}^{y_b} xy \left( a_0 + \sum_{j=1}^{m} a_j \exp \{ b_j y + c_j x \} \right) \, dx \, dy
\]

\[
= a_0 \int_{x_a}^{x_b} \int_{y_a}^{y_b} xy dx dy + \sum_{j=1}^{m} \int_{x_a}^{x_b} \int_{y_a}^{y_b} xy a_j \exp \{ b_j y + c_j x \} \, dx \, dy
\]

\[
= a_0 \int_{x_a}^{x_b} \int_{y_a}^{y_b} xy dx dy + \sum_{j=1}^{m} \int_{x_a}^{x_b} \int_{y_a}^{y_b} xy a_j \exp \{ b_j y \} \exp \{ c_j x \} \, dx \, dy
\]

\[
= a_0 \int_{x_a}^{x_b} \int_{y_a}^{y_b} xy dx dy + \sum_{j=1}^{m} a_j \int_{x_a}^{x_b} x \exp \{ c_j x \} \int_{y_a}^{y_b} y \exp \{ b_j y \} \, dx \, dy
\]

\[
= a_0 \left( y_b^2 - y_a^2 \right) \left( x_b^2 - x_a^2 \right) + \sum_{j=1}^{m} \frac{a_j}{c_j^2 b_j^2} \left( -\exp \{ b_j y_a \} + b_j y_a \exp \{ b_j y_a \} + \exp \{ b_j y_b \} - b_j y_b \exp \{ b_j y_b \} \right)
\]

\[
- \exp \{ c_j x_a \} + c_j x_a \exp \{ c_j x_a \} + \exp \{ c_j x_b \} - c_j x_b \exp \{ c_j x_b \} \right) \quad (1.28)
\]

A new version of $\mathbb{E}(XY | \mathbf{d}_i)$ is shown in Equation 1.29 for the case in which the exponent has only one variable, that is, $b_j = 0$ (in some cases it can happen in Elvira):
\[
\mathbb{E}(XY \mid \mathbf{d}_i) = \int_{x_a}^{x_b} \int_{y_a}^{y_b} xyf(x, y \mid \mathbf{d}_i) \, dx \, dy
\]

\[
= \int_{x_a}^{x_b} \int_{y_a}^{y_b} xy \left( a_0 + \sum_{j=1}^{m} a_j \exp \{ c_j x \} \right) \, dx \, dy
\]

\[
= a_0 \int_{x_a}^{x_b} \int_{y_a}^{y_b} xy \, dx \, dy + \sum_{j=1}^{m} \int_{x_a}^{x_b} \int_{y_a}^{y_b} xy a_j \exp \{ c_j x \} \, dx \, dy
\]

\[
= a_0 \int_{x_a}^{x_b} \int_{y_a}^{y_b} xy \, dx \, dy + \sum_{j=1}^{m} a_j \int_{x_a}^{x_b} \int_{y_a}^{y_b} x \exp \{ c_j x \} \, dx \, dy
\]

\[
= \frac{a_0}{4} (y_b^2 - y_a^2)(x_b^2 - x_a^2) + \sum_{j=1}^{m} \frac{a_j (y_b^2 - y_a^2)}{2c_j^2} \]

\[
(\exp \{ c_j x_a \} - c_j x_a \exp \{ c_j x_a \} - \exp \{ c_j x_b \} + c_j x_b \exp \{ c_j x_b \}) \quad (1.29)
\]

Also, we will need for some especial cases the integral in Equation 1.30:

\[
\mathbb{E}(X_j \mid \mathbf{d}_i) = \int_{x_a}^{x_b} x_j f(x_j \mid \mathbf{d}_i) \, dx_j = \int_{x_a}^{x_b} x_j \left( a_0 + \sum_{j=1}^{m} a_j \exp \{ b_j x \} \right) \, dx_j
\]

\[
= a_0 \int_{x_a}^{x_b} x_j \, dx_j + \sum_{j=1}^{m} a_j \int_{x_a}^{x_b} x_j \exp \{ b_j x \} \, dx_j
\]

\[
= a_0 \int_{x_a}^{x_b} x_j \, dx_j + \sum_{j=1}^{m} a_j \int_{x_a}^{x_b} x_j \exp \{ b_j x \} \, dx_j
\]

\[
= \frac{a_0}{2} (x_b^2 - x_a^2)
\]

\[
+ \sum_{j=1}^{m} \frac{a_j}{b_j^2} \left( \exp \{ b_j x_b \} (b_j x_b - 1) - \exp \{ b_j x_a \} (b_j x_a - 1) \right) \quad (1.30)
\]

For the second one, we need to calculate a matrix of expectations, where the \( jk \)-element will be \( \mathbb{E}(Y_j Y_k \mid \mathbf{d}_i) \). If \( j \neq k \) we can carry out the calculation in the same way as in Equation 1.28. When \( j = k \), the expectation is \( \mathbb{E}(Y_j^2 \mid \mathbf{d}_i) \) and will be calculated later in Equation 1.32.

For the third one, we have that:

\[
\mathbb{E} \left[ (X_j - \bar{\mathbf{I}}_{x,j} \bar{\mathbf{Y}})^2 \mid \mathbf{d}_i \right] = \mathbb{E}(X_j^2 \mid \mathbf{d}_i) - 2 \bar{\mathbf{I}}_{x,j} \mathbb{E}(X_j \bar{\mathbf{Y}} \mid \mathbf{d}_i) + \mathbb{E}(\bar{\mathbf{I}}_{x,j} \bar{\mathbf{Y}}^2 \mid \mathbf{d}_i)
\]

\[
(1.31)
\]
where

\[
E(X_j^2 \mid d_i) = \int_{x_a}^{x_b} x_j^2 f(x_j \mid d_i) dx_j = \int_{x_a}^{x_b} x_j^2 \left( a_0 + \sum_{j=1}^{m} a_j \exp \{ b_j x_j \} \right) dx_j
\]

\[
= a_0 \int_{x_a}^{x_b} x_j^2 dx_j + \sum_{j=1}^{m} a_j \int_{x_a}^{x_b} x_j^2 \exp \{ b_j x_j \} dx_j
\]

\[
= \frac{a_0}{3} (x_b^3 - x_a^3) + \sum_{j=1}^{m} \frac{a_j}{b_j} \left( \exp \{ b_j x_b \} (b_j x_b (b_j x_b - 2) + 2) - \exp \{ b_j x_a \} (b_j x_a (b_j x_a - 2) + 2) \right)
\]

(1.32)

The expectation \(E(X_jY \mid d_i)\) of the second part in Equation 1.31 have been calculated previously in Equation 1.28.

For the last term in Equation 1.31, we have that:

\[
E[(\bar{l}_T z, j \bar{Y})^2 \mid d_i] = E[(\bar{l}_T z, j \bar{Y})^T \bar{l}_T z, j \bar{Y} \mid d_i] = E[\bar{l}_T z, j \bar{Y} \bar{Y}^T \bar{l}_T z, j \mid d_i] = \bar{l}_T z, j E[\bar{Y} \bar{Y}^T \mid d_i] \bar{l}_T z, j
\]

(1.33)

and the calculation of \(E[\bar{Y} \bar{Y}^T \mid d_i]\) has been considered previously.

For calculating the expectations in all this section we will consider the most general case in which all the variables involved are unobserved. When some of them are observed, we will follow the next basic rules:

- if \(X\) and \(Y\) are both unobserved, then \(E(XY \mid d_i)\) is substitute by \(xy\).
- if only \(X\) is unobserved, then \(E(XY \mid d_i)\) is substitute by \(y \ast E(X)\), and the other way arround.

1.4.2 Logistic

For the logistic we don’t have any expectation to compute in the E-step, since its updating rule in the M-step only has a product of MTE functions and no expectations. Anyway, a simple integral need to be solved in Equation 1.24, since we have an MTE times \(y\).

1.4.3 Multinomial

The calculation of the expectation for the multinomial is straightforward and implicitly is made inside the M-step in section 1.3.3.

1.5 Conversion from Gaussians and logistic into MTEs

1.5.1 Conditional linear Gaussian

Cobb et. al proposed in [1, 2] a general formulation for a 2-piece, 3-term MTE potential which approximates a Gaussian PDF. The specification relies on split points being func-
tions of $\mu$, which is a problem, since in a CLG $\mu$ is a linear combination of the continuous parents.

In order to avoid problems with split points depending on $\mu$, we decided to approximate the CLG distribution using only one MTE-piece.

Several candidates in the domain $[-2.5, 2.5]$ with different number of exponential terms are shown in Fig.1.1.

\[ \phi(x) = \sigma^{-1} \left[ a_0 + \sum_{j=1}^{7} a_j \exp \left\{ b_j \frac{x - \mu}{\sigma} \right\} \right] \]  

(1.34)

Figure 1.1: MTE approximations with 3, 5, 7, 9, 11 and 13 exponential terms, respectively, for the standard Gaussian distribution with support $[-2.5, 2.5]$.

Now, a good candidate in terms of efficiency might be a 7-term MTE with the following:
The parameters are shown in the next expression:

\[
\phi(x) = \sigma^{-1}(0.84159 - 0.25122 \exp\{-\frac{(x - \mu)}{\sigma}\} - 0.25122 \exp\{\frac{(x - \mu)}{\sigma}\} \\
+ 0.027992 \exp\{-2 \frac{(x - \mu)}{\sigma}\} + 0.027992 \exp\{2\frac{(x - \mu)}{\sigma}\} \\
- 0.0010639 \exp\{-3 \frac{(x - \mu)}{\sigma}\} - 0.0010639 \exp\{3\frac{(x - \mu)}{\sigma}\}\]
\]

(1.35)

Figure 1.2 shows the MTE approximation from a Gaussian with specific parameters.

Figure 1.2: 7-terms MTE approximation for a Gaussian distribution

1.5.2 Logistic

The expression for the binary sigmoid function of a discrete variable \( A \) with a continuous parent \( Z \) is defined in the following way:

\[
P(A = a_1 \mid Z) = \frac{1}{1 + \exp\{g + wz\}}
\]

Note that for future works the function above may be generalized to the logistic or soft-max function.

Cobb and Shenoy proposed in [1] a general formulation for a 4-piece 1-term MTE potential which approximates a binary sigmoid function with the parameters \( g \) and \( w \). One exponential term is used.

\[
P(X_6 = 0 \mid Z = z) = \begin{cases} 
0 & \text{if } z < \frac{5 - g}{w} \\
-0.021704 + 0.521704 b \exp\{-0.635 w(z - g(w + 1))\} & \text{if } \frac{5 - g}{w} \leq z \leq \frac{9 - g}{w} \\
1.021704 - 0.521704 b^{-1} \exp\{0.635 w(z - g(w + 1))\} & \text{if } \frac{9 - g}{w} < z \leq \frac{-5 - g}{w} \\
1 & \text{if } z > \frac{-5 - g}{w}
\end{cases}
\]

(1.36)
where $b = 0.529936^{w^2 + w + 1}$.

$$
\hat{P}(A = a_1 \mid z) = a + \sum_{i=1}^{m} b_i \exp\{wz\} \tag{1.37}
$$

### 1.5.3 Multinomial

The conversion from multinomial into MTEs is straightforward, since a multinomial potential can be seen as a particular case of an MTE one.
Chapter 2

Implementation

2.1 Introduction

Since the problem doesn’t concern with the learning of the structure, an example of structure has been created to check the implementation (see Figure 2.1). This Bayesian network contains all the possible distributions in our framework. Two-lined nodes represent continuous variables and the single-lined ones are discrete.

Also, a database `data.dbc` for the network above has been generated with the details showed in Table 2.1:

The distribution of one variable can be:

- **Multinomial**, if the variable is discrete either without parents or only discrete ones. The distribution will be saved by means of a CPT.

  The variables $X_1$, $X_2$ and $X_5$ in the network above satisfy this condition.

- **Logistic**, if the variable is discrete (binary) and has either only continuous parents or a mixture of continuous and discrete parents. For each configuration of the discrete parents and the current node, and for each partition of the domain for the continuous parents we need to specify a weight $w$.

  The variable that satisfies this condition in the network in our example is $X_6$.

![Bayesian networks containing all the possible distributions in our framework.](image)

Figure 2.1: Bayesian networks containing all the possible distributions in our framework.
<table>
<thead>
<tr>
<th>Type of variable</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>(X_1)</td>
<td>discrete</td>
</tr>
<tr>
<td>(X_2)</td>
<td>discrete</td>
</tr>
<tr>
<td>(X_3)</td>
<td>continuous</td>
</tr>
<tr>
<td>(X_4)</td>
<td>continuous</td>
</tr>
<tr>
<td>(X_5)</td>
<td>discrete</td>
</tr>
<tr>
<td>(X_6)</td>
<td>discrete</td>
</tr>
<tr>
<td>(X_7)</td>
<td>continuous</td>
</tr>
</tbody>
</table>

\[
\begin{align*}
\{"s1", "s0", "s2"\} \\
\{"s2", "s1", "s0"\} \\
[-2.41151054120676, 2.48160755463863] \\
[-2.89716110721953, 3.01894362793437] \\
\{"s1", "s2", "s0"\} \\
\{"s1", "s0"\} \\
[-2.69874631755666, 3.1178334681985] \\
\end{align*}
\]

Table 2.1: Description of the variables in the database.

- **CLG**, if the variable is continuous. Only a value for \(\mu\) and \(\sigma\) will be specified if the variable doesn’t have any parent. Otherwise, for each configuration of the discrete parents and for each continuous parent plus one will be necessary to specify a value for \(\mu\) and \(\sigma\).

The variables \(X_3, X_4\) and \(X_7\) in the network above satisfy this condition.

### 2.2 Generating a database for experiments

An example of database is shown in Figure 2.2:

\[
\begin{array}{cccccc}
X_1 & X_2 & X_3 & X_4 & X_5 & X_6 \\
2 & - & 1.5 & - & 1 & 0 & 3.6 \\
- & 2 & - & 1.6 & 2 & 0 & - \\
1 & 2 & 2.5 & 1.5 & - & 1 & - \\
- & 1 & - & 2.3 & - & - & 1.5 \\
0 & - & 4.3 & - & - & 0 & 1.9 \\
1 & 3 & 3.6 & - & 0 & 1 & 20.1 \\
0 & 1 & 0.5 & 1.6 & 1 & - & - \\
- & 2 & 4.5 & 4.9 & 1 & 0 & 4.6 \\
\end{array}
\]

Table 2.2: Example of database
2.3 Initialization of the parameters space

The initialization is different depending on the distribution considered:

- **Multinomial**: random values between 0 and 1 will be generated for each position in the CPT. The sum of the probabilities for all the states of one variable must be 1.

- **Logistic**: for each configuration of the discrete parents an independent term \( g > 0 \) and a set of negative weights \( w_1, \ldots, w_k \) are randomly generated, being \( k \) the number of continuous parents of the current variable.

- **CLG**: for each configuration of the discrete parents a value for \( \sigma \), an independent term \( \mu_0 \) and a set of means \( \mu_1, \ldots, \mu_k \) are randomly generated, being \( k \) the number of continuous parents of the current variable.

- Variable \( X_1 \):

  \[
  \begin{array}{c|ccc}
  & 0 & 1 & 2 \\
  \hline
  X_1 & p_0 & p_1 & p_2 \\
  \end{array}
  \]

- Variable \( X_2 \):

  \[
  \begin{array}{c|ccc}
  & 0 & 1 & 2 \\
  \hline
  X_1 & p_0 & p_1 & p_2 \\
  \end{array}
  \]

- Variable \( X_3 \):

  \[
  \begin{array}{c}
  X_4 \mu \sigma \\
  \end{array}
  \]

- Variable \( X_4 \):

  \[
  \begin{array}{c}
  X_4 \mu \sigma \\
  \end{array}
  \]

- Variable \( X_5 \):

  \[
  \begin{array}{c|cccc}
  X_5 - X_1, X_2 & 00 & 01 & \ldots & 22 \\
  \hline
  0 & p_{000} & p_{001} & \ldots & p_{022} \\
  1 & p_{100} & p_{101} & \ldots & p_{122} \\
  2 & p_{200} & p_{201} & \ldots & p_{222} \\
  \end{array}
  \]

- Variable \( X_6 \):
2.4 Conversion from CLG, logistic, and multinomial into MTEs

2.4.1 Conditional Linear Gaussian

In order to adapt the MTE approximation in 1.34 to the Elvira format, we need to make the following changes:

\[
\phi(x) = \sigma^{-1}a_0 + \sum_{i=1}^{7} \sigma^{-1}\exp\{-\sigma^{-1}b_i\mu\}a_i\exp\{\sigma^{-1}b_ix\} \tag{2.1}
\]

where \(\sigma^{-1}a_0\) is the independent term of the MTE, \(\sigma^{-1}\exp\{-\sigma^{-1}b_i\mu\}a_i\) is the factor and \(\sigma^{-1}b_i\) is the exponent.

Let see now how the conversion is for the gaussian variables involved:

- **Variable \(X_7\):** The potential associated to this variable will be a ContinuousProbabilityTree (see Figure 2.2) containing the discrete/continuous parents that are used to split the tree. The variable \(X_7\) will be included in the exponent of the MTEs located in the leaves.

<table>
<thead>
<tr>
<th>(X_1)</th>
<th>(X_2)</th>
<th>Independent term</th>
<th>(X_3)</th>
<th>(X_4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>(g)</td>
<td>(w_3)</td>
<td>(w_4)</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>(g)</td>
<td>(w_3)</td>
<td>(w_4)</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>(g)</td>
<td>(w_3)</td>
<td>(w_4)</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>(g)</td>
<td>(w_3)</td>
<td>(w_4)</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>(g)</td>
<td>(w_3)</td>
<td>(w_4)</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>(g)</td>
<td>(w_3)</td>
<td>(w_4)</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>(g)</td>
<td>(w_3)</td>
<td>(w_4)</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>(g)</td>
<td>(w_3)</td>
<td>(w_4)</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>(g)</td>
<td>(w_3)</td>
<td>(w_4)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(X_1)</th>
<th>(X_2)</th>
<th>Independent term</th>
<th>(X_3)</th>
<th>(X_4)</th>
<th>(X_1X_2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>(\mu_0)</td>
<td>(\mu_3)</td>
<td>(\mu_4)</td>
<td>(\sigma_{00})</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>(\mu_0)</td>
<td>(\mu_3)</td>
<td>(\mu_4)</td>
<td>(\sigma_{01})</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>(\mu_0)</td>
<td>(\mu_3)</td>
<td>(\mu_4)</td>
<td>(\sigma_{02})</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>(\mu_0)</td>
<td>(\mu_3)</td>
<td>(\mu_4)</td>
<td>(\sigma_{10})</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>(\mu_0)</td>
<td>(\mu_3)</td>
<td>(\mu_4)</td>
<td>(\sigma_{11})</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>(\mu_0)</td>
<td>(\mu_3)</td>
<td>(\mu_4)</td>
<td>(\sigma_{12})</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>(\mu_0)</td>
<td>(\mu_3)</td>
<td>(\mu_4)</td>
<td>(\sigma_{20})</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>(\mu_0)</td>
<td>(\mu_3)</td>
<td>(\mu_4)</td>
<td>(\sigma_{21})</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>(\mu_0)</td>
<td>(\mu_3)</td>
<td>(\mu_4)</td>
<td>(\sigma_{22})</td>
</tr>
</tbody>
</table>
Problem related to the number of parameters in the Continuous Probability Tree

There are different number of parameters in our parameter space in the EM algorithm and in the potential defined by Elvira (Continuous Probability Tree).

That is, for the CLG, the number of \( \mu \)-values defined for a gaussian variable is the number of possible configurations for its discrete parents times the number of its continuous parents plus one.

On the other hand, the number of \( \mu \)-values to define the potential in Elvira is the number of leaves in the Continuous Probability Tree associated to the current variable. So, the number of parameters here is higher.

Thus, we need to specify a correspondence between both set of parameters. We decided to define the \( \mu \)-value as a function of the domain partition in the tree and our parameters space defined in the EM algorithm.

For example, the calculation for \( \mu' \) located in the ContinuousProbabilityTree in Fig.2.2 will be:

\[
\mu' = \mu_0 + \mu_3 \frac{b_0 + b_1}{2} + \mu_4 \frac{a_0 + a_1}{2}
\]

taking into account that for the discrete configuration \( X_1 = 0, X_2 = 1 \) we have the parameters \( \mu_0, \mu_3 \) and \( \mu_4 \) in our EM space.

For \( \sigma' \)-value we don’t have any problem, because the number of parameters in both environment are the same.
2.4.2 Logistic

Let $X_6$ our binary variable with continuous parents $Z = \{Z_1, \ldots, Z_k\}$. The sigmoid function is defined as:

$$P(X_6 = 0 \mid z) = \frac{1}{1 + \exp\{g + \sum_{n=1}^{k} w_n z_n\}},$$

(2.2)

and $P(X_6 = 1 \mid z) = 1 - P(X_6 = 0 \mid z)$.

Eq. 1.36 shows an MTE approximation for the sigmoid function of a variable $z$ given one parent. The support of $z$ must be limited due to the Elvira implementation in the following way:

$$P(X_6 = 0 \mid Z = z) =
\begin{cases}
0 & \text{if } z < \frac{-w - 0.021704}{w} \\
-0.021704 + 0.521704b \exp\{-0.635w(z - g(w + 1))\} & \frac{-w - 0.021704}{w} \leq z \leq \frac{g'}{w} \\
1.021704 - 0.521704b^{-1} \exp\{0.635w(z - g(w + 1))\} & \frac{g'}{w} < z \leq \frac{-w - 0.521704}{w} \\
1 & \text{if } \frac{-w - 0.521704}{w} < z
\end{cases}$$

(2.3)

where $b = 0.529936g(w^2 + w + 1)$.

In order to use this approximation, we need to represent the potential of $X_6$ in such a way that the leaves of the Continuous Probability Tree have an MTE density with only one variable (see Figure 2.3). The continuous variables (except one of them, $X_4$) have been split using the Elvira method to learn a conditional density. $X_4$ is split by using the same partition as in Equation 1.36. In this way the MTE associated to each leaf can be computed using this approximation.

Note that the MTE density $P(X_6 = 1 \mid X_4)$ is computed as $1 - P(X_6 = 0 \mid X_4)$.

An example of an ideal approximation with 4 pieces is shown in Fig. 2.4.

Thus, the number of children of variable $X_4$ in the ContinuousProbabilityTree in Fig. 2.3 will depend on the previous consideration.
Adapting Logistic-MTE approximation into Elvira format

In Elvira everything inside the exponent must be represented as a LinearFunction object containing a set of variables and the set of associated coefficients. So, we modify the MTE densities in equation 1.36 as follows:

\[-0.021704 + 0.521704b \exp\{-0.635w(z - g(w + 1))\}\]
\[= -0.021704 + 0.521704b \exp\{-0.635w(z - gw - g)\}\]
\[= -0.021704 + 0.521704b \exp\{-0.635(wz - gw^2 - gw)\}\]
\[= -0.021704 + 0.521704b \exp\{-0.635wz + 0.635gw^2 + 0.635gw\}\]
\[= -0.021704 + 0.521704b \exp\{0.635gw^2 + 0.635gw\} \exp\{-0.635wz\} \quad (2.4)\]

where $0.521704b \exp\{0.635gw^2 + 0.635gw\}$ is the factor.
Figure 2.4: 4-pieces MTE approximation for the logistic function.

\[
1.021704 - 0.521704 b^{-1} \exp\{0.635 w(z - g(w + 1))\} = 1.021704 - 0.521704 b^{-1} \exp\{0.635 w(z - gw - g)\} = 1.021704 - 0.521704 b^{-1} \exp\{0.635 (wz - gw^2 - gw)\} = 1.021704 - 0.521704 b^{-1} \exp\{-0.635 gw^2 - 0.635 gw\} \exp\{0.635 wz\} \tag{2.5}
\]

where \(-0.521704 b^{-1} \exp\{-0.635 gw^2 - 0.635 gw\}\) is the factor.

**Problem related to the number of parameters in the CPT**

In the same way as for the CLG, a problem appears in the logistic when trying to convert the EM parameter space into the parameter space of the potential in Elvira.

For example, in the leaves of the tree in Fig. 2.3 is needed a value of \(g'\) and \(w_4\). For a discrete configuration \(X_1 = 0, X_2 = 0\) and for the branch \([a_0, a_1]\) in variable \(X_3\) we have that:

\[
g' = g + w_4 \frac{a_0 + a_1}{2}.
\]

\(w_4\) is the original parameter defined in our EM algorithm.

Thus, comparing the branches \([a_0, a_1]\) and \([a_1, a_2]\) outgoing from \(X_3\), their leaves will have the same value for \(w_4\) and different for \(g'\) (since it depends on the partition).
Figure 2.5: ContinuousProbabilityTree for the variables $X_1$ and $X_2$ in Elvira.

2.4.3 Multinomial

For variables $X_1$ and $X_2$ we create a continuous probability tree (see Figure 2.5) with only one variable. Each leaf will contain a MTE density with only the independent term which is the corresponding probability of the CPT.

The continuous probability tree associated to the variable $X_5$ (see Figure 2.6) will have all the discrete nodes in the potential used to split the tree. In the leaves we will have a particular case of an MTE density with only the independent term, that is, a number representing the probability of the configuration given by the path from the root to the leaf.

Figure 2.6: ContinuousProbabilityTree for the variable $X_5$ in Elvira.
Bibliography


