# Parameter learning in MTE networks using incomplete data

Fernández, Antonio; Langseth, Helge; Nielsen, Thomas Dyhre; Salmerón, Antonio

*Publication date:*
2010

*Document Version*
Accepted author manuscript, peer reviewed version



*Citation for published version (APA):*
Fernández, A., Langseth, H., Nielsen, T. D., & Salmerón, A. (2010). *Parameter learning in MTE networks using incomplete data*. Paper presented at The Fifth European Workshop on Probabilistic Graphical Models, Helsinki, Finland. http://www.helsinki.fi/pgm2010/papers/fernandez.pdf

# Parameter learning in MTE networks using incomplete data

Antonio Fernández
Dept. of Statistics and Applied Mathematics
University of Almería, Spain
afalvarez@ual.es

Helge Langseth
Dept. of Computer and Information Science
The Norwegian University of Science and Technology
helgel@idi.ntnu.no

Thomas Dyhre Nielsen
Dept. of Computer Science
Aalborg University, Denmark
tdn@cs.aau.dk

Antonio Salmerón
Dept. of Statistics and Applied Mathematics
University of Almería, Spain
antonio.salmeron@ual.es

**Abstract**

Bayesian networks with mixtures of truncated exponentials (MTEs) are gaining popularity as a flexible modelling framework for hybrid domains. MTEs support efficient and exact inference algorithms, but estimating an MTE from data has turned out to be a difficult task. Current methods suffer from a considerable computational burden as well as the inability to handle missing values in the training data. In this paper we describe an EM-based algorithm for learning the maximum likelihood parameters of an MTE network when confronted with incomplete data. In order to overcome the computational difficulties we make certain distributional assumptions about the domain being modeled, thus focusing on a subclass of the general class of MTE networks. Preliminary empirical results indicate that the proposed method offers results that are inline with intuition.

## 1 Introduction

One of the major challenges when using probabilistic graphical models for modeling hybrid domains (domains containing both discrete and continuous variables), is to find a representation of the joint distribution that support 1) efficient algorithms for exact inference based on local computations and 2) algorithms for learning the representation from data. In this paper we will consider mixtures of truncated exponentials (MTEs) (Moral et al., 2001) as a candidate framework. MTE distributions allow discrete and continuous variables to be treated in a uniform fashion, and it is well known that the Shenoy-Shafer architecture (Shenoy and Shafer, 1990) can be used for exact inference in MTE networks (Moral et al., 2001). Also, the expressive power of MTEs was demonstrated in (Cobb et al., 2006), where the most commonly used marginal distributions were accurately approximated by MTEs.

Algorithms for learning marginal and conditional MTE distributions from complete data have previously been proposed (Rumí et al., 2006; Romero et al., 2006; Langseth et al., 2010; Langseth et al., 2009). When faced with incomplete data, (Álvarez et al., 2010b) considered a data augmentation technique for learning (tree augmented) naive MTE networks for regression, but so far no attempt has been made at learning the parameters of a general MTE network.

In this paper we propose an EM-based algorithm (Dempster et al., 1977) for learning MTE networks from incomplete data. The general problem of learning MTE networks (also with complete data) is computationally very hard (Langseth et al., 2009): Firstly, the sufficient statistics of a dataset is the dataset itself, and secondly, there are no known closed-form equations for finding the maximum likelihood (ML) parameters. In order to circumvent these problems, we focus on domains, where the probability distributions mirror standard parametric families for which ML parameter estimators are known to exist. This implies that instead of trying to directly learn ML estimates for the MTE distributions, we may consider the ML estimators for the corresponding parametric families. Hence, we define a generalized EM algorithm that incorporates the following two observations (corresponding to the M-step and the E-step, respectively): *i*) Using the results of (Cobb et al., 2006; Langseth et al., 2010) the domain-assumed parametric distributions can be transformed into MTE distributions. *ii*) Using the MTE representation of the domain we can evaluate the expected sufficient statistics needed for the ML estimators. For ease of presentation we shall in this paper only consider domains with multinomial, Gaussian, and logistic functions, but, in principle, the proposed learning procedure is not limited to these distribution families. Note that for these types of domains exact inference is not possible using the assumed distribution families.

The remainder of the paper is organized as follows. In Section 2 we give a brief introduction to MTE distributions as well as rules for transforming selected parametric distributions to MTEs. In Section 3 we describe the proposed algorithm, and in Section 4 we present some preliminary experimental results. Finally, we conclude in Section 5 and give directions for future research.

## 2 Preliminaries

### 2.1 MTE basics

Throughout this paper, random variables will be denoted by capital letters, and their values by lowercase letters. In the multi-dimensional case, boldfaced characters will be used. The domain of the variables $\boldsymbol{X}$ is denoted by $\Omega_{\boldsymbol{X}}$. The MTE model is defined by its corresponding potential and density as follows (Moral et al., 2001):

**Definition 1.** (MTE potential) *Let $\boldsymbol{W}$ be a mixed $n$-dimensional random vector. Let $\boldsymbol{Z} = (Z_1, \ldots, Z_d)^T$ and $\mathbf{Y} = (Y_1, \ldots, Y_c)^T$ be the discrete and continuous parts of $\boldsymbol{W}$, respectively, with $c + d = n$. We say that a function $f : \Omega_{\boldsymbol{W}} \mapsto \mathbb{R}_0^+$ is a* Mixture of Truncated Exponentials potential *if for each fixed value $\boldsymbol{Z} \in \Omega_{\boldsymbol{Z}}$ of the discrete variables $\boldsymbol{Z}$, the potential over the continuous variables $\mathbf{Y}$ is defined as:*

$$f(\boldsymbol{y}) = a_0 + \sum_{i=1}^{m} a_i \exp\{\mathbf{b}_i^T \boldsymbol{y}\}, \qquad (1)$$

*for all $\boldsymbol{y} \in \Omega_{\mathbf{Y}}$, where $a_i \in \mathbb{R}$ and $\mathbf{b}_i \in \mathbb{R}^c$, $i = 1, \ldots, m$. We also say that $f$ is an MTE potential if there is a partition $D_1, \ldots, D_k$ of $\Omega_{\mathbf{Y}}$ into hypercubes and in each $D_\ell$, $f$ is defined as in Eq. 1. An MTE potential is an* MTE density *if it integrates to 1.*

A *conditional MTE density* can be specified by dividing the domain of the conditioning variables and specifying an MTE density for the conditioned variable for each configuration of splits of the conditioning variables. The following is an example of a conditional MTE density.

$$f(y|x) = \begin{cases} 1.26 - 1.15e^{0.006y} \\ \qquad \text{if } 0.4 \leq x < 5,\ 0 \leq y < 13\ , \\ 1.18 - 1.16e^{0.0002y} \\ \qquad \text{if } 0.4 \leq x < 5,\ 13 \leq y < 43\ , \\ 0.07 - 0.03e^{-0.4y} + 0.0001e^{0.0004y} \\ \qquad \text{if } 5 \leq x < 19,\ 0 \leq y < 5\ , \\ -0.99 + 1.03e^{0.001y} \\ \qquad \text{if } 5 \leq x < 19,\ 5 \leq y < 43\ . \end{cases}$$

### 2.2 Translating standard distributions to MTEs

In this section we will consider transformations from selected parametric distributions to MTE distributions.

#### 2.2.1 The Multinomial Distribution

The conversion from a multinomial distribution into an MTE distribution is straightforward, since a multinomial distribution can be seen as a special case of an MTE (Moral et al., 2001).

#### 2.2.2 The Conditional Linear Gaussian Distribution

In (Cobb et al., 2006; Langseth et al., 2010) methods are described for obtaining an MTE approximation of a (marginal) Gaussian distribution. Common for both approaches is that the split points used in the approximations depend on the mean value of the distribution being modeled. Consider now a variable $X$ with continuous parents $\boldsymbol{Y}$ and assume that $X$ follows a conditional linear Gaussian distribution:[1]

$$X|\boldsymbol{Y}=\boldsymbol{y} \sim \mathcal{N}(\mu = b+\boldsymbol{w}^{\mathrm{T}}\boldsymbol{y}, \sigma^2).$$

In the conditional linear Gaussian distribution, the mean value is a weighted linear combination of the continuous parents. This implies that we cannot directly obtain an MTE representation of the distribution by following the procedures of (Cobb et al., 2006; Langseth et al., 2010); each part of an MTE potential has to be defined on a hypercube (see Definition 1), and the split points can therefore not depend on any of the variables in the potential. Instead we define an MTE approximation by splitting $\Omega_{\boldsymbol{Y}}$ into hypercubes $D_1, \ldots, D_k$, and specifying an MTE density for $X$ for each of the hypercubes. For hypercube $D_l$ the mean of the distribution is assumed to be constant, i.e., $\mu^l = b + w_1 \mathrm{mid}_1^l + \cdots + w_j \mathrm{mid}_j^l$, where $\mathrm{mid}_i^l$ denotes the midpoint of $Y_i$ in $D_l$ (by defining fixed upper and lower bounds on the ranges of the continuous variables, the midpoints are always well-defined). Thus, finding an MTE representation of the conditional linear Gaussian distribution has been reduced to defining a partitioning $D_1, \ldots, D_k$ of $\Omega_{\boldsymbol{Y}}$ and specifying an MTE representation for a (marginal) Gaussian distribution (with mean $\mu^l$ and variance $\sigma^2$) for each of the hypercubes $D_l$ in the partitioning.

In the current implementation we define the partitioning of $\Omega_{\boldsymbol{Y}}$ based on equal-frequency binning, and we use BIC-score (Schwarz, 1978) to chose the number of bins. To obtain an MTE representation of the (marginal) Gaussian distribution for each partition in $\Omega_{\boldsymbol{Y}}$ we follow the procedure of (Langseth et al., 2010); four MTE candidates for the domain $[-2.5, 2.5]$ are shown in Figure 1 (no split points are being used, except to define the boundary).

[1] For ease of exposition we will disregard any discrete parent variables in the subsequent discussion, since they will only serve to index the parameters of the function.

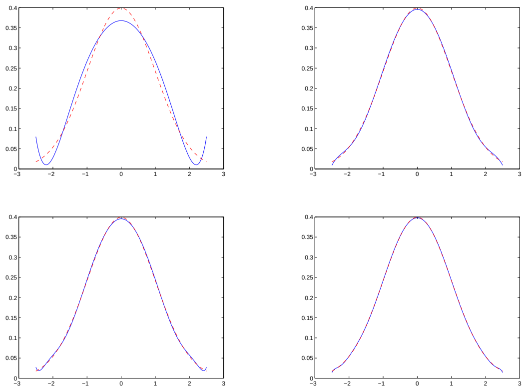

Figure 1: MTE approximations with 5, 7, 9 and 11 exponential terms, respectively, for the truncated standard Gaussian distribution with support $[-2.5, 2.5]$. It is difficult to visually distinguish the MTE and the Gaussian for the three latter models.

Notice that the MTE density is only positive within the interval $[\mu - 2.5\sigma, \mu + 2.5\sigma]$ (confer Figure 1), and it actually integrates up to 0.9876 in that region, which means that there is a probability of 0.0124 of finding points outside this interval. In order to avoid problems with 0 probabilities, we add tails covering the remaining probability mass of 0.0124. More precisely, we define the normalization constant

$$c = \frac{0.0124}{2\left(1-\int_0^{2.5\sigma}\exp\{-x\}\mathrm{d}x\right)},$$

and include the tail

$$\phi(x) = c \cdot \exp\left\{-(x-\mu)\right\}.$$

for the interval above $x = \mu + 2.5\sigma$ in the MTE specification. Similarly, a tail is also included for the interval below $x = \mu - 2.5\sigma$. The transformation rule from Gaussian to MTE therefore becomes

$$\phi(x) = \begin{cases} c \cdot \exp\left\{x-\mu\right\} \text{ if } x < \mu - 2.5\sigma, \\ \sigma^{-1}\left[a_0 + \sum_{j=1}^{7} a_j \exp\left\{b_j \frac{x-\mu}{\sigma}\right\}\right] \\ \qquad \text{if } \mu - 2.5\sigma \leq x \leq \mu + 2.5\sigma, \\ c \cdot \exp\left\{-(x-\mu)\right\} \text{ if } x > \mu - 2.5\sigma. \end{cases} \tag{2}$$

#### 2.2.3 The Logistic Function

The sigmoid function for a discrete variable $X$ with a single continuous parent $Y$ is given by

$$P(X = 1 \mid Y) = \frac{1}{1 + \exp\{b + wy\}}.$$

(Cobb and Shenoy, 2006) propose an 4-piece 1-term MTE representation for this function:

$$P(X = 1 \mid Y = y) = \begin{cases} 0 & \text{if } y < \frac{5-b}{w}, \\ a_0^1 + a_1^1(b, w)\exp\{b^1 w(y - b(w + 1))\} & \text{if } \frac{5-b}{w} \leq y \leq \frac{b'}{w}, \\ a_0^2 + a_1^2(b, w)\exp\{b^2 w(y - b(w + 1))\} & \text{if } \frac{b'}{w} < y \leq \frac{-5-b}{w}, 1 \text{ if } y > \frac{-5-b}{w}, \end{cases} \tag{3}$$

where $a_0^k$ and $b^1$, $b^2$ are constants and $a_1^1(b, w)$ and $a_1^2(b, w)$ are derived from $b$ and $w$. Note that the MTE representation is 0 or 1 if $y < (5 - b)/w$ or $y > (-5 - b)/w$, respectively. The representation can therefore be inconsistent with the data (i.e., we may have data cases with probability 0), and we therefore replace the 0 and 1 with $\epsilon$ and $1-\epsilon$, where $\epsilon$ is a small positive number. ($\epsilon = 0.01$ was used in the experiments reported in Section 4.)

In the general case, where $X$ has continuous parents $\boldsymbol{Y} = \{Y_1, \ldots, Y_j\}$ and discrete parents $\boldsymbol{Z} = \{Z_1, \ldots, Z_k\}$, then for each configuration $\boldsymbol{z}$ of $\boldsymbol{Z}$, the conditional distribution of $X$ given $\boldsymbol{Y}$ is given by

$$P(X = 1 \mid \boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{Z} = \boldsymbol{z}) = \frac{1}{1 + \exp\{b_{\boldsymbol{z}} + \sum_{i=1}^{j} w_{i,\boldsymbol{z}} y_i\}}. \tag{4}$$

With more than one continuous variable as argument, the logistic function cannot easily be represented by an MTE having the same structure as in Equation 3. The problem is that the split points would then be (linear) functions of at least one of the continuous variables, which is not consistent with the MTE framework (see Definition 1). Instead we follow the same procedure as for the conditional linear Gaussian distribution: for each of the continuous variables in $\boldsymbol{Y}' = \{Y_2, \ldots Y_j\}$, split the variable $Y_i$ into a finite set of intervals and use the midpoint of the $l$th interval to represent $Y_i$ in that interval. The intervals for the variables in $\boldsymbol{Y}'$ define a partitioning $D_1, \ldots, D_k$ of $\Omega_{\boldsymbol{Y}'}$ into hypercubes, and for each of these partitions we apply Equation 3. That is, for partition $D_l$ we get

$$P(X = 1 \mid \boldsymbol{y}, \boldsymbol{z}) = \frac{1}{1 + \exp\{b' + w_1 y_1\}},$$

where $b' = b + \sum_{k=2}^{j} \mathrm{mid}_l^k w_l^k$. In the current implementation $Y_1$ is chosen arbitrarily from $\boldsymbol{Y}$, and the partitioning of the state space of the parent variables is performed as for the conditional linear Gaussian distribution.

## 3 The General Algorithm

As previously mentioned, deriving an EM algorithm for general MTE networks is computationally hard because the sufficient statistics of the dataset is the dataset itself and there is no closed-form solution for estimating the maximum likelihood parameters. To overcome these computational difficulties we will instead focus on a subclass of MTE networks, where the conditional probability distributions in the network mirror selected distributional families. By considering this subclass of MTE networks we can derive a generalized EM algorithm, where the updating rules can be specified in closed form.

To be more specific, assume that we have an MTE network for a certain domain, where the conditional probability distributions in the domain mirror traditional parametric families with known ML-based updating rules. Based on the MTE network we can calculate the expected sufficient statistics required by these rules (the E-step) and by using the transformations described in Section 2.2 we can in turn update the distributions in the MTE network.

The overall learning algorithm is detailed in Algorithm 1, where the domain in question is represented by the model $\mathcal{B}$. Note that in order to exemplify the procedure we only consider the multinomial distribution, the Gaussian distribution, and the logistic distribution. The algorithm is, however, easily extended to other distribution classes.

---

**Algorithm 1**: An EM algorithm for learning MTE networks from incomplete data.

---

**Input**: A parameterized model $\mathcal{B}$ over $X_1, \ldots, X_n$, and an incomplete database $\mathcal{D}$ of cases over $X_1, \ldots, X_n$.

**Output**: An MTE network $\mathcal{B}'$.

Initialize the parameter estimates $\hat{\boldsymbol{\theta}}_{\mathcal{B}}$ randomly.

**repeat**

- Using the current parameter estimates $\hat{\boldsymbol{\theta}}_{\mathcal{B}}$, represent $\mathcal{B}$ as an MTE network $\mathcal{B}'$ (see Section 2.2).
- **(E-step)** Calculate the expected sufficient statistics required by the M-step using $\mathcal{B}'$.
- **(M-step)** Use the result of the E-step to calculate new ML parameter estimates $\tilde{\boldsymbol{\theta}}_{\mathcal{B}}$ for $\mathcal{B}$.
- $\hat{\boldsymbol{\theta}}_{\mathcal{B}} \leftarrow \tilde{\boldsymbol{\theta}}_{\mathcal{B}}$.

**until** *convergence* ;

**return** $\mathcal{B}'$.

---

### 3.1 The EM algorithm

The transformation rules for the conditional linear Gaussian distribution, the multinomial distribution, and the logistic distribution are given in Section 2.2. In order to complete the specification of the algorithm, we therefore only need to define the E-step and the M-step for the three types of distributions being considered.

#### 3.1.1 The M-step

Given a database of cases $\mathcal{D} = \{\boldsymbol{d}_1, \ldots, \boldsymbol{d}_N\}$ we derive the updating rules based on the expected data-complete log-likelihood function $Q$:

$$\begin{aligned} Q &= \sum_{i=1}^{N} \mathbb{E}[\log f(X_1, \ldots, X_n) \mid \boldsymbol{d}_i] \\ &= \sum_{i=1}^{N} \sum_{j=1}^{n} \mathbb{E}[\log f(X_j \mid \mathrm{pa}(X_j)) \mid \boldsymbol{d}_i] \ . \end{aligned}$$

The updating rules for the parameters for the multinomial distribution and the Gaussian distribution are well-known and can be found in Appendix A (see (Álvarez et al., 2010a) for a derivation).

A closed form solution does not exist for the weight vector for the logistic function, and instead one typically resorts to numerical optimization such as gradient ascent for maximizing $Q$. To ease notation, we shall consider the variable $X_j$ with discrete parents $\boldsymbol{Z}_j$ and continuous parents $\boldsymbol{Y}_j$ (we drop indexes for the parents whenever those are clear from the context). Also, we use $\bar{\boldsymbol{w}}_{\boldsymbol{z},j} = [\boldsymbol{w}_{\boldsymbol{z},j}^{\mathrm{T}}, b_{\boldsymbol{z},j}]^{\mathrm{T}}$ and $\bar{\boldsymbol{y}} = [\boldsymbol{y}^{\mathrm{T}}, 1]^{\mathrm{T}}$, in which case the gradient ascent updating rule can be expressed as

$$\hat{\bar{\boldsymbol{w}}}_{\boldsymbol{z},j} := \bar{\boldsymbol{w}}_{\boldsymbol{z},j} + \gamma \frac{\partial Q}{\partial \bar{\boldsymbol{w}}_{\boldsymbol{z},j}} \ ,$$

where $\gamma > 0$ is a small number and

$$\begin{aligned} \frac{\partial Q}{\partial \bar{\boldsymbol{w}}_{\boldsymbol{z},j}} = & \sum_{i=1}^{N} P(\boldsymbol{z} \mid \boldsymbol{d}_i) \left[ \int_{\boldsymbol{y}} P(x_j = 1, \bar{\boldsymbol{y}} \mid \boldsymbol{d}_i, \boldsymbol{z}) \right. \\ & g_{\boldsymbol{z}, x_j = 1}(\bar{\boldsymbol{y}}) \bar{\boldsymbol{y}} d\boldsymbol{y} - \\ & \left. \int_{\boldsymbol{y}} P(x_j = 0, \bar{\boldsymbol{y}} \mid \boldsymbol{d}_i, \boldsymbol{z}) g_{\boldsymbol{z}, x_j = 0}(\bar{\boldsymbol{y}}) \bar{\boldsymbol{y}} d\boldsymbol{y} \right] . \end{aligned}$$

In order to find the partial derivative we need to evaluate two integrals. However, the combination of the MTE potential $P(x_j, \bar{\boldsymbol{y}} \mid \boldsymbol{d}_i, \boldsymbol{z})$ and the logistic function $g_{\boldsymbol{z}, x_j}(\bar{\boldsymbol{y}})$ makes these integrals difficult to evaluate. In order to avoid this problem we use the MTE representation of the logistic function specified in Section 2.2.3, which allows the integrals to be calculated in closed form.

#### 3.1.2 The E-step

In order to perform the updating in the M-step we need to calculate the following expectations (see Appendix A):

- $\mathbb{E}(X_j \mid \boldsymbol{d}_i, \boldsymbol{z})$
- $\mathbb{E}(X_j \bar{\boldsymbol{Y}} \mid \boldsymbol{d}_i, \boldsymbol{z})$
- $\mathbb{E}(\bar{\boldsymbol{Y}} \bar{\boldsymbol{Y}}^{\mathrm{T}} \mid \boldsymbol{d}_i, \boldsymbol{z})$
- $\mathbb{E}\left[ (X_j - \bar{\boldsymbol{l}}_{\boldsymbol{z},j}^{\mathrm{T}} \bar{\boldsymbol{Y}})^2 \mid \boldsymbol{d}_i, \boldsymbol{z} \right]$

All the expectations can be calculated analytically (see Appendix B). The main point to notice in the calculations is that rather than calculating e.g. $\mathbb{E}(\bar{\boldsymbol{Y}}\bar{\boldsymbol{Y}}^{\mathrm{T}} \mid \boldsymbol{d}_i, \boldsymbol{z})$ directly we instead consider each of the components $\mathbb{E}(Y_j Y_k \mid \boldsymbol{d}_i, \boldsymbol{z})$ in the matrix individually.

## 4 Experimental results

In order to evaluate the proposed learning method we have generated data from the Crops network (Murphy, 1999). We sampled six complete datasets containing 50, 100, 500, 1000, 5000, and 10000 cases, respectively, and for each of the datasets we generated three other datasets with 5%, 10%, and 15% missing data (the data is missing completely at random (Little and Rubin, 1987)), giving a total of 24 training datasets. The actual data generation was performed using WinBUGS (Lunn et al., 2000).

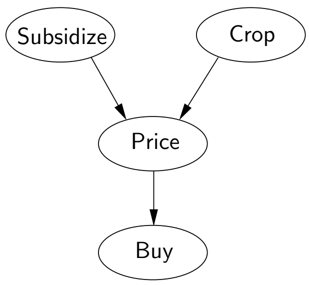


Figure 2: The Crops network.

For comparison, we have also learned baseline models using WinBUGS. However, since WinBUGS does not support learning of multinomial distributions from incomplete data we have removed all cases where Subsidize is missing from the datasets.

The learning results are shown in Table 1, which lists the average (per observation) log-likelihood of the model wrt. a test-dataset consisting of 15000 cases (and defined separately from the training datasets). From the table we see the expected behaviour: As the size of the training data increases, the models tend to get better; as the fraction of the data that is missing increases, the learned models tend to get worse.

The results also show how WinBUGS in general outperforms the algorithm we propose in this paper. We believe that one of the reasons is the way we approximate the tails of the Gaussian distribution in Eq. 2. As the tails are thicker than the actual Gaussian tails, the likelihood is lower in the central parts of the distribution, where most of the samples potentially concentrate. Another possible reason is the way in which we approximate the CLG distribution. Recall that when splitting the domain of the parent variable, we take the average data point in each split to represent the parent, instead of using the actual value. This approximation tends to give an increase in the estimate of the conditional variance, as the approximated distribution needs to cover all the training samples. Obviously, this will later harm the average predictive log likelihood. Two possible solution to this problem are *i*) to increase the number of splits, or *ii*) to use dynamic discretization to determine the optimal way to split the parent's domain. However, both solutions come with a cost in terms of increased computational complexity, and we consider the tradeoff between accuracy and computational cost as an interesting topic for future research.

The algorithm has been implemented in Elvira (Consortium, 2002) and the software, the datasets used in the experiments, and the WinBUGS specifications are all available from `http://elvira.ual.es/MTE-EM.html`.

## 5 Conclusion

In this paper we have proposed an EM-based algorithm for learning MTE networks from incomplete data. In order to overcome the computational difficulties of learning MTE distributions, we focus on a subclass of the MTE networks, where the distributions are assumed to mirror known parametric families. This subclass supports a computationally efficient EM algorithm. Preliminary empirical results indicate that the method learns as expected, although not as well as WinBUGS. In particular, our method seems to struggle when the portion of the the data that is missing increases. We have proposed some remedial actions to this problem that we will investigate further.

| No. Cases | ELVIRA | | | | WINBUGS | | | |
|---|---|---|---|---|---|---|---|---|
| | Percentage of missing data | | | | Percentage of missing data | | | |
| | 0% | 5 % | 10% | 15% | 0% | 5 % | 10% | 15% |
| 50 | -3.8112 | -3.7723 | -3.8982 | -3.8553 | -3.7800 | -3.7982 | -3.7431 | -3.6861 |
| 100 | -3.7569 | -3.7228 | -3.9502 | -3.9180 | -3.7048 | -3.7091 | -3.7485 | -3.7529 |
| 500 | -3.6452 | -3.6987 | -3.7972 | -3.8719 | -3.6272 | -3.6258 | -3.6380 | -3.6295 |
| 1 000 | -3.6325 | -3.7271 | -3.8146 | -3.8491 | -3.6174 | -3.6181 | -3.6169 | -3.6179 |
| 5 000 | -3.6240 | -3.6414 | -3.8056 | -3.9254 | -3.6136 | -3.6141 | -3.6132 | -3.6144 |
| 10 000 | -3.6316 | -3.6541 | -3.7910 | -3.8841 | -3.6130 | -3.6131 | -3.6131 | -3.6135 |

Table 1: The average log-likelihood for the learned models, calculated per observation on a separate test set.


### Acknowledgments

This work is supported by a grant from Iceland, Liechtenstein, and Norway through the EEA Financial Mechanism. Supported and Coordinated by Universidad Complutense de Madrid. Partially supported by the Spanish Ministry of Science and Innovation, through project TIN2007-67418-C03-02, and by EFDR funds.


## A Updating rules

The updating rules for the parameters for the multinomial distribution (i.e., $\theta_{j,k,\boldsymbol{z}} = P(X_j = k|\boldsymbol{Z} = \boldsymbol{z})$) and the conditional linear Gaussian distribution (i.e., $X_j|\boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{Y} = \boldsymbol{y} \sim \mathcal{N}(\bar{\boldsymbol{l}}^{\mathrm{T}}_{\boldsymbol{z},j}\bar{\boldsymbol{y}}, \sigma^2_{\boldsymbol{z},j})$) are given by

$$\hat{\bar{\boldsymbol{l}}}_{\boldsymbol{z},j} \leftarrow \left[\sum_{i=1}^{N} f(\boldsymbol{z} \mid \boldsymbol{d}_i)\mathbb{E}(\bar{\boldsymbol{Y}}\bar{\boldsymbol{Y}}^{\mathrm{T}} \mid \boldsymbol{d}_i, \boldsymbol{z})\right]^{-1} \left[\sum_{i=1}^{N} f(\boldsymbol{z} \mid \boldsymbol{d}_i)\mathbb{E}(X_j\bar{\boldsymbol{Y}} \mid \boldsymbol{d}_i, \boldsymbol{z})\right]$$

$$\hat{\sigma}_{\boldsymbol{z},j} \leftarrow \left[\frac{1}{\sum_{i=1}^{N} f(\boldsymbol{z} \mid \boldsymbol{d}_i)} \sum_{i=1}^{N} f(\boldsymbol{z} \mid \boldsymbol{d}_i)\mathbb{E}\left[(X_j - \bar{\boldsymbol{l}}^{\mathrm{T}}_{\boldsymbol{z},j}\bar{\boldsymbol{Y}})^2 \mid \boldsymbol{d}_i, \boldsymbol{z}\right]\right]^{1/2}$$

$$\hat{\theta}_{j,k,\boldsymbol{z}} \leftarrow \frac{\sum_{i=1}^{N} P(X_j = k, \boldsymbol{Z} = \boldsymbol{z} \mid \boldsymbol{d}_i)}{\sum_{k=1}^{|\mathrm{sp}(X_j)|}\sum_{i=1}^{N} P(X_j = k, \boldsymbol{Z} = \boldsymbol{z} \mid \boldsymbol{d}_i)}$$

## B Expected sufficient statistics

To illustrate the calculation of the expected sufficient statistics we consider the calculation of $\mathbb{E}\left[(X_j - \bar{\boldsymbol{l}}^{\mathrm{T}}_{\boldsymbol{z},j}\bar{\boldsymbol{Y}})^2 \mid \boldsymbol{d}_i, \boldsymbol{z}\right]$ (see Section 3.1.2):

$$\mathbb{E}\left[(X_j - \bar{\boldsymbol{l}}^{\mathrm{T}}_{\boldsymbol{z},j}\bar{\boldsymbol{Y}})^2 \mid \boldsymbol{d}_i\right] = \mathbb{E}[X_j^2 \mid \boldsymbol{d}_i] - 2\bar{\boldsymbol{l}}^{\mathrm{T}}_{\boldsymbol{z},j}\mathbb{E}[X_j\bar{\boldsymbol{Y}} \mid \boldsymbol{d}_i] + \mathbb{E}[(\bar{\boldsymbol{l}}^{\mathrm{T}}_{\boldsymbol{z},j}\bar{\boldsymbol{Y}})^2 \mid \boldsymbol{d}_i]$$

For the second component in the summation we need to calculate a vector of expectations, where the $k$th element is $\mathbb{E}[X_jY_k \mid \boldsymbol{d}_i]$. By letting the ranges of $X_j$ and $Y_k$ be $[x_a, x_b]$ and $[y_a, y_b]$ (dropping the $j$ and $k$ indices for simplicity), respectively, it is easy to show that the expectation can be calculated on closed form:

$$\mathbb{E}\left[X_jY_i \mid \boldsymbol{d}_i\right] = \frac{a_0}{4}(y_b^2 - y_a^2)(x_b^2 - x_a^2) + \sum_{j=1}^{m} \frac{a_j}{{c_j}^2{b_j}^2}\Big(-\exp\{b_jy_a\} + b_jy_a\exp\{b_jy_a\} + \exp\{b_jy_b\} - b_jy_b\exp\{b_jy_b\}\Big)\Big(-\exp\{c_jx_a\} + c_jx_a\exp\{c_jx_a\} + \exp\{c_jx_b\} - c_jx_b\exp\{c_jx_b\}\Big).$$

For $\mathbb{E}[X_j^2 \mid \boldsymbol{d}_i]$ and $\mathbb{E}\left[(\bar{\boldsymbol{l}}^{\mathrm{T}}_{\boldsymbol{z},j}\bar{\boldsymbol{Y}})^2 \mid \boldsymbol{d}_i\right]$ the calculations are similar; for the latter it immediately follows from $\mathbb{E}\left[(\bar{\boldsymbol{l}}^{\mathrm{T}}_{\boldsymbol{z},j}\bar{\boldsymbol{Y}})^2 \mid \boldsymbol{d}_i\right] = \bar{\boldsymbol{l}}^{\mathrm{T}}_{\boldsymbol{z},j}\mathbb{E}\left[\bar{\boldsymbol{Y}}\bar{\boldsymbol{Y}}^{\mathrm{T}} \mid \boldsymbol{d}_i\right]\bar{\boldsymbol{l}}_{\boldsymbol{z},j}$.

## References


Antonio Fernández Álvarez, Helge Langseth, Thomas Dyhre Nielsen, and Antonio Salmerón.

2010a. MTE-based parameter learning using incomplete data. Technical report, Department of Statistics and Applied Mathematics, University of Almeria, Spain. `http://www.ual.es/~afa109/downloads/Fernandez2010.pdf`.

Antonio Fernández Álvarez, Jens D. Nielsen, and Antonio Salmerón. 2010b. Learning bayesian networks for regression from incomplete databases. *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems*, 18:69–86.

Barry R. Cobb and Prakash P. Shenoy. 2006. Inference in hybrid Bayesian networks with mixtures of truncated exponentials. *International Journal of Approximate Reasoning*, 41(3):257–286.

Barry R. Cobb, Prakash P. Shenoy, and Rafael Rumí. 2006. Approximating probability density functions in hybrid Bayesian networks with mixtures of truncated exponentials. *Statistics and Computing*, 46(3):293–308.

Elvira Consortium. 2002. Elvira: An environment for creating and using probabilistic graphical models. In José A. Gámez and Antonio Salmerón, editors, *First European Workshop on Probabilistic Graphical Models*, pages 222–230.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38.

Helge Langseth, Thomas D. Nielsen, Rafael Rumí, and Antonio Salmerón. 2009. Maximum likelihood learning of conditional MTE distributions. In *Proceedings of the Tenth European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, volume 5590 of *Lecture Notes in Artificial Intelligence*, pages 240–251. Springer-Verlag, Berlin, Germany.

Helge Langseth, Thomas D. Nielsen, Rafael Rumí, and Antonio Salmerón. 2010. Parameter estimation and model selection in mixtures of truncated exponentials. *International Journal of Approximate Reasoning*, 51:485–498.

Roderick J. A. Little and Donald B. Rubin. 1987. *Statistical analysis with missing data*. John Wiley & Sons, New York.

D. J. Lunn, A. Thomas, N. Best, and D. Spiegelhalter. 2000. Winbugs - a bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10(4):325–337.

Serafín Moral, Rafael Rumí, and Antonio Salmerón. 2001. Mixtures of truncated exponentials in hybrid Bayesian networks. In *Proceedings of the Sixth European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, volume 2143 of *Lecture Notes in Artificial Intelligence*, pages 145–167. Springer-Verlag, Berlin, Germany.

Kevin P. Murphy. 1999. A variational approximation for Bayesian networks with discrete and continuous latent variables. In Kathryn B. Laskey and Henri Prade, editors, *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 467–475.

V. Romero, R. Rumí, and A. Salmerón. 2006. Learning hybrid Bayesian networks using mixtures of truncated exponentials. *International Journal of Approximate Reasoning*, 42:54–68.

Rafael Rumí, Antonio Salmerón, and Serafin Moral. 2006. Estimating mixtures of truncated exponentials in hybrid Bayesian networks. *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research*, 15(2):397–421, September.

Gideon Schwarz. 1978. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464.

Prakash P. Shenoy and Glenn Shafer. 1990. Axioms for probability and belief-function propagation. In *Proceedings of the Sixth Workshop on Uncertainty in Artificial Intelligence*, pages 169–198.