Representation, Inference and Learning of Bayesian Networks as Conjugate Exponential Family Models

August 7, 2015

Abstract

1 Introduction

Defining the data structure of a Bayesian network (BN) is not a straightforward problem. The definition of the data structure of a directed acyclic graph (DAG) is not complex when compared to the definition of the data structure of the conditional probability distributions (CPDs) encoded in the BN. The DAG is an *homogeneous* data structure, in the sense that it is only composed by nodes and directed edges. However, the set of different CPDs is not limited at all. For example, the data structure for representing a Multinomial distribution is, in a first look, quite different from the data structure needed to represent a Normal distribution. In the former case, we need to store the probability of each case of the multinomial variable, while in the latter case, we need to store, for example, the mean and the variance of the Normal distribution. If we consider a Poisson, an Exponential, or a MoTBF, etc., the data structures needed to represent these distributions are completely different. These are only few examples of unidimensional distributions, however, when defining the data structure of CPDs, the issue becomes much more complex and challenging.

For instance, the data structure for representing the CPD of a Normal distribution given a set of normally distributed variables is, in a first look, totally different from the data structure needed to represent the CPD of a Multinomial variable given a set of Multinomial variables. In the former case, under the conditional linear Gaussian framework, we need to store the coefficients for the linear combination of the parent variables plus the variance of the main variable. While the data structure for the multinomial given multinomial variables is usually defined using a big conditional probability table. Therefore, if we want to use a combination of Normal, Multinomial, Poison, Exponential, etc., in the same framework, the number of data structures needed to represent all the possible CPDs combinations quickly explode.

Another challenging problem is performing inference and learning of BNs with different kinds of CPDs. For example, the maximum likelihood of a Normal distribution is obtained by computing the sample mean and variance, while the maximum likelihood of a Multinomial distribution is obtained by normalizing the sample *counts* of each state. Alternative methods are required for the different possible CPDs, which means that considering a new family of variables, i.e., Normal, Poisson, or Exponential, etc., implies the definition and the implementation from scratch of new maximum likelihood methods.

In the case of inference, things are even worse. For example, the combination and marginalization operations over probability potentials belonging to different distribution families are in general non-closed and, in principle, involve quite different approaches. I.e., the combination or multiplication of two multinomial potentials or distributions involve completely different methods than the combination or product of two Normal distributions. This similarly applies to the marginalization operations. So, defining and coding all these operations for different distribution families can become a daunting task.

In this technical report, we propose to use the so-called conjugate exponential family (CEF) models in order to avoid most of the above mentioned problems and save time by using previously known results and algorithms. Firstly, all the CPDs inside this family can be represented using the same data structure, which is simply composed by:

- two n-dimensional vectors, namely, natural and moment parameters, and
- two n-dimensional functions, namely, sufficient statistics and log-normalizer functions.

Moreover, we describe previously proposed learning and inference algorithms than can be directly implemented on top of this general and unique CEF representation. The result is a suitable framework for coding a toolbox which aims to deal with the problem of representing, making inference, and learning general BNs from data.

2 Exponential family models

2.1 Definition

Let $\mathbf{X} = \{X_1, \ldots, X_N\}$ denote the set of stochastic random variables defining our domain problem and $\mathbf{x}$ an observation vector. We say that the joint probability distribution $p(\cdot|\theta)$ parametrized by a parameter vector θ is an *exponential family model* with a natural (or canonical) parametrization if the logarithm of p can be functionally expressed as follows

$$\ln p_\theta(\mathbf{x}) = \theta^T s(\mathbf{x}) - A(\theta) + h(\mathbf{x}), \tag{2.1}$$

which is based on the following definitions,

- $h(x)$ is the log-base measure. Its domain, denoted by $\mathcal{X}$, is the Cartesian product of the domains of random variables in $\mathbf{X}$, and its codomain are the positive real numbers, $h : \mathcal{X} \to \Re^+$.

- $\theta \in \Theta$ is the natural parameter vector and $\Theta \subseteq \Re^K$ is the natural parameter space, where K being called the dimension of the model. This *natural parameter space* Θ is defined as follows:

$$\Theta \equiv \{\theta \in \Re^K : \int_{\mathbf{x}} exp\big(\theta^T s(\mathbf{x}) + h(\mathbf{x})\big)\ d\mathbf{x} < \infty\} \tag{2.2}$$

  i.e., as the the set of parameter vectors which define a proper normalizable density.

- $A(\theta)$ is the log-partition function, which is defined as follows:

$$A(\theta) = \int_{\mathbf{x}} exp\big(\theta^T s(\mathbf{x}) + h(\mathbf{x})\big)\ d\mathbf{x}$$

  So, its domain is Θ and its codomain are the positive real numbers, $A : \Theta \to \Re^+$.

- $s(\mathbf{x})$ is the sufficient statistics function, whose domain is $\mathcal{X}$ and is codomain is denoted by $\mathcal{S} \subseteq \Re^K$, $\mathbf{s} : \mathcal{X} \to \mathcal{S}$. We also refers to $s(\mathbf{x})$ as the sufficient statistics (vector) of the observation $\mathbf{x}$.

More specific subfamilies inside this broad class of probabilistic models are usually considered in the literature:

Minimal Exponential Family: An exponential family is *minimal* if there is no a non-zero constant vector α, such that $\alpha^T s(\mathbf{x})$ is equal to a constant for all assignments $\mathbf{x}$. In this case, there is a unique parameter vector θ associated with each probability distribution.

Overcomplete Exponential Family: An exponential family is overcomplete if it is not minimal. In this case, there exists an entire affine subset of parameter vectors θ, each one associated with the same probability distribution.

Regular Exponential Family: An exponential family for which its natural parameter space Θ is an open set.

Linear Exponential Family: An exponential family which is regular and minimal.

Curved Exponential Family: Equation 2.1 describing exponential families can be generalized by writing

$$\ln p_\theta(\mathbf{x}) = \eta(\theta)^T s(\mathbf{x}) - A(\theta) + h(\mathbf{x}), \tag{2.3}$$

where now η is a function that maps the parameters θ to the canonical parameters $\eta = \eta(\theta)$. A model belongs to the *curved exponential family* if it is described by Equation 2.3 and $dim(\theta) < dim(\eta(\theta))$, which means that the model has more sufficient statistics than parameters.

2.2 Dual Parametrization

A key property of exponential family models is that they can be alternatively parametrized by a so-called *moment parameter* vector $\mu \in \mathcal{S}$. The relevancy of this dual parametrization is that several statistical computations, such as marginalization and maximum likelihood estimation, can be understood as transforming from one parameterization to the other.

By definition, a vector μ is defined as the *expected vector of sufficient statistics* with respect to θ as follows

$$\mu \triangleq E[s(\mathbf{x})|\theta] = \int_{\mathbf{x}} s(\mathbf{x})\ p_\theta(\mathbf{x})\ d\mathbf{x} \tag{2.4}$$

As can be seen, computing the moment parameter vector requires to perform inference over the probability distribution p. When p belongs to the exponential family, the association between p and μ is one-to-one.

The transformation from *natural parameters* to *moment parameters* can be achieved by solving the following optimization problem,

$$\theta(\mu) = arg\max_{\theta \in \Theta}\ \theta^T \mu - A(\theta) \tag{2.5}$$

where $\theta(\cdot)$ is, by abuse of notation, the transformation function from expectation parameters to natural parameters, i.e. $\theta(\cdot) : \mathcal{S} \to \Theta$.

The above equation is also known as the *maximum likelihood function*, because $\theta(\frac{1}{n}\sum_{i=1}^{n} s(x_i))$ gives the maximum likelihood estimation $\theta^\star$ for a data set with n i.i.d. observations $\{x_1, \ldots, x_n\}$.

For the minimal exponential family, the transformation between θ and μ is one-to-one: μ is a dual set of the model parameter θ [**?**]. That is to say, for each $\theta \in \Theta$ we always have an associated $\mu \in \mathcal{S}$ and both have the same dimension and parameterize the same probability distribution. For overcomplete exponential families, there is an entire affine subset of parameters θ associated to the moment parameter vector μ.

For regular exponential families, the transformation from *natural parameters* to *moment parameters* can be nicely interpreted as the gradient of the log-normalizer function,

$$\mu \triangleq E[s(\mathbf{x})|\theta] = \partial A(\theta)/\partial\theta \tag{2.6}$$

Regular exponential families with a minimal representation, i.e. linear exponential family, also enjoy a wider a set of nice properties and has been widely studied in the literature. However, they are not relevant for this paper.

2.3 Conjugate exponential (CE) models

A conditional distribution $p(\mathbf{X}|\mathbf{Y})$ is in the exponential family if, for any assignment $\mathbf{y}$ of the variables in $\mathbf{Y}$, the probability distribution $p(\mathbf{X}|\mathbf{y})$ can be expressed in exponential form. Moreover, $p(\mathbf{X}|\mathbf{Y})$ is said to be *conjugate* with respect to the distribution $p(\mathbf{Y})$ if the latter has the same functional form than the posterior $p(\mathbf{Y}|\mathbf{X}) \propto p(\mathbf{X}|\mathbf{Y})p(\mathbf{Y})$. Then, the joint distribution $p(\mathbf{X}, \mathbf{Y}) = p(\mathbf{X}|\mathbf{Y})p(\mathbf{Y})$ is said to be a *conjugate exponential family* (CE) model.

An important property of CE models is that they are a multi-linear function of the sufficient statistic functions of $\mathbf{X}$ and $\mathbf{Y}$ [?]. This implies that we can alternatively express the conditional distribution $p(\mathbf{X}|\mathbf{Y})$ in any of the three following functional forms

Conditional form (C-form):

$$\ln p_\theta(\mathbf{x}|\mathbf{y}) = \theta_c(\mathbf{y})^T s(\mathbf{x}) - A_c(\theta(\mathbf{x})) + h(\mathbf{x}) \tag{2.7}$$

Posterior form (P-form):

$$\ln p_\theta(\mathbf{x}|\mathbf{y}) = \theta_p(\mathbf{x})^T s(\mathbf{y}) - A_p(\theta(\mathbf{y})) + h(\mathbf{x}) \tag{2.8}$$

Full form (F-form):

$$\ln p_\theta(\mathbf{x}|\mathbf{y}) = \theta_f^T s(\mathbf{x}, \mathbf{y}) - A_f(\theta) + h(\mathbf{x}) \tag{2.9}$$

where the suffixes for θ and A denote that both terms vary from one form to another. For example, in the first and second functional form, $\theta_c(\mathbf{y})$ and $\theta_p(\mathbf{x})$ denote that the natural parameters are now a function of $\mathbf{y}$ and $\mathbf{x}$, respectively, which is not the case for the last functional form.

In Table ? we list the main conjugate exponential distributions. We point out that many of these pairs are usually omitted in the literature because they are usually consider under Bayesian learning settings. For example, as can be seen this table, every distribution conditioned to a set of multinomial variables defined a conjugate exponential pair.

3 Bayesian networks as conjugate exponential models

In this section we how Bayesian networks can be compactly represented when they define conjugate exponential models. We also show how the transformation between natural

and moment parameters also enjoys some relevant properties which allow to decompose and simplify this transformation.

A Bayesian network (BN) defines a joint distribution $p_\theta(X_1, \ldots, X_n)$ over a set of variables in the following form:

$$p_\theta(\mathbf{X}) = \prod_{i=1}^{N} p_\theta(X_i|Pa(X_i))$$

where $Pa(X_i) \subset \mathbf{X} \backslash X_i$ represents the so-called *parent variables* of X_i and $p_\theta(X_i|Pa(X_i))$ denotes the local conditional probability of X_i given its parents $Pa(X_i)$. BNs can be graphically represented by a directed acyclic graph (DAG). Each node, labelled X_i in the graph, is associated with a factor or conditional probability $p_\theta(X_i|Pa(X_i))$. Additionally, for each parent $X_j \in Pa(X_i)$, the graph contains one directed edge pointing from X_j to the *child* variable X_i.

In our case, we also restrict ourselves to Bayesian networks models satisfying the so-called *global independence parameter* property [**?**]. Under this assumption, the parameters defining each local conditional probability are independent between them. This assumption is highlighted by denoting by θ_i the parameters defining the local conditional probability $p_{\theta_i}(X_i|Pa(X_i))$. We also denote by Θ_i to the space of the parameters θ_i.

Finally, we introduce the definition of a conjugate exponential Bayesian network (ce-BN),

Definition 1. *A BN is said to be a conjugate exponential model if the probability distribution of X_i given its parents, $p(X_i|Pa(X_i))$, is conjugate with respect to the distribution of their parents, for all the variables in the model.*

So, the following developments applies to ce-BNs which satisfy the *global independence parameter property.*

3.1 Representation

As we show in the next theorem, we can exploit the *F-form* of a conjugate model to obtain a compact representation of a Bayesian network as an exponential family model.

Theorem 1. *A ce-BN can be represented as an exponential family model, by composing the F-form representations of its conditional probability distributions, as follows*

- *the natural parameters θ of a ce-BN are formed by the composition of the local natural parameters of each conditional distribution,*

$$\theta = (\theta_1, \ldots, \theta_n)$$

where θ_i corresponds to natural parameters of the conditional distribution of X_i expressed in F-form.

- *the sufficient statistics $s(X_1, \ldots, X_n)$ of a ce-BN are formed the composition of the local sufficient statistics of each conditional distribution,*

$$s(X_1, \ldots, X_n) = (s_1(X_1, Pa(X_1)), \ldots, s_n(X_n, Pa(X_n))$$

where $s_i(X_i, Pa(X_i))$ corresponds to sufficient statistics of the conditional distribution of X_i expressed in F-form.

- *the log-normalizer $A(\theta)$ of a ce-BN is the sum of the local conditional log-normalizer of each conditional distribution,*

$$A(\theta) = \sum_{i=1}^{n} B_i(\theta_i)$$

where $B_i(\theta_i)$ corresponds to conditional log-normalizer of the conditional distribution of X_i expressed in F-form.

Proof. By using Equation 2.9, a ce-BN can be represented in the following way:

$$\begin{aligned}
\ln p(X_1, \ldots, X_n) &= \sum_{i=1}^{n} \ln p(X_i | Pa(X_i)) \\
&= \sum_{i=1}^{n} \theta_i\big(Pa(X_i)\big)^T s_i(X_i) - A_i\Big(\theta\big(Pa(X_i)\big)\Big) \\
&= \sum_{i=1}^{n} \theta_i^T \; s_i\big(X_i, Pa(X_i)\big) - B_i(\theta_i) \\
&= \begin{pmatrix} \theta_1 \\ \ldots \\ \theta_n \end{pmatrix}^T \begin{pmatrix} s_1\big(X_1, Pa(X_1)\big) \\ \ldots \\ s_n\big(X_n, Pa(X_n)\big) \end{pmatrix} - \sum_{i=1}^{n} B_i(\theta_i) \qquad (3.1)
\end{aligned}$$

□

Hence, in order to represent a BN as a CEF model, we only have to worry about the local representation of each CPD as in Equation 3.1. The global representation is then just obtained by composing all these local representations. Let us notice that without the assumption of conjugacy for the conditional exponential distributions, the above representation would have not been possible.

3.2 From natural to moment parameters

In this section, we examine the transformation from natural to moment parameters in a ce-BN, which is stated in the following theorem.

Theorem 2. *The moment parameter vector of a ce-BN associated to a vector of natural parameter θ locally decomposes into local moment parameters as follows,*

$$\mu = (\mu_1, \ldots, \mu_n)$$

where $\mu_i = \int s_i(X_i, Pa(X_i))\ p_\theta(X_i, Pa(X_i)) d\mathbf{X}$ is the local moment vector associated to the local sufficient statistic $s_i(X_i, Pa(X_i))$, as defined in Theorem 1, and the local marginal probability $p(X_i, Pa(X_i))$.

Proof. By definition, $\mu = E\left[s(X_1, \ldots, X_n)|\theta\right] = \int s(X_1, \ldots, X_n)\ p_\theta(X_1, \ldots, X_n)\ d\mathbf{X}$. According to Theorem 1 $s(X_1, \ldots, X_n)$ locally decomposed in a set of local sufficient statistics $s_i(X_i, Pa(X_i))$. Using this decomposition we arrived to the decomposition of the vector of moment parameters μ as follow,

$$\begin{aligned}
\mu &= E\left[s(X_1, \ldots, X_n)|\theta\right] \\
&= \int s(X_1, \ldots, X_n)\ p_\theta(X_1, \ldots, X_n)\ d\mathbf{X} \\
&= \int \big(s_1(X_1, Pa(X_1)), \ldots, s_n(X_n, Pa(X_n))\big)\ p_\theta(X_1, \ldots, X_n)\ d\mathbf{X} \\
&= \Big(\int s_1(X_1, Pa(X_1)) p_\theta(X_1, Pa(X_1))\ d(X_1, Pa(X_1)), \ldots, \\
&\quad \int s_n(X_1, Pa(X_n)) p_\theta(X_n, Pa(X_n))\ d(X_n, Pa(X_n))\Big) \\
&= (\mu_1, \ldots, \mu_n)
\end{aligned}$$

□

Note here that in order to compute the local moment parameter μ_i we need to obtain the local marginal probability $p_\theta(X_i, Pa(X_i))$, which requires running inference over the whole BN.

3.3 From moment to natural Parameters

As shown in Equation 2.5, the transformation from moment to natural parameters involves solving an optimization problem. In the following theorem we show that for ce-BNs this problem decomposes into a set of simpler and local optimization problems,

Theorem 3. *Given a moment parameter vector μ, the associated natural parameter vector, denoted by $\theta(\mu)$, of a ce-BN locally decomposes into local natural parameters as follows,*

$$\theta(\mu) = \big(\theta_1(\mu_1), \ldots, \theta_n(\mu_n)\big)$$

where each $\theta_i(\mu_i)$ is the solution of the following optimization problem,

$$\theta_i(\mu_i) = \arg\max_{\theta_i \in \Theta_i} \theta_i^T \mu_i - B(\theta_i) \tag{3.2}$$

Proof. By definition,

$$\begin{aligned} \theta(\mu) &= \arg\max_{\theta \in \Theta} \theta^T \mu - A(\theta) \\ &= \arg\max_{(\theta_1,\ldots,\theta_n) \in \Theta} \sum_{i=1}^{n} \theta_i^T \mu_i - B_i(\theta_i) \end{aligned}$$

According to the *global independence parameter* assumption, the θ_i parameters are independent and, then, the above maximization problem fully decomposes into local maximization problems, one for each conditional probability distribution,

$$\theta_i(\mu_i) = \arg\max_{\theta_i \in \Theta_i} \theta_i^T \mu_i - B(\theta_i)$$

and the global solution is just the aggregation of the local solutions as follows:

$$\theta(\mu) = \big(\theta_1(\mu_1), \ldots, \theta_n(\mu_n)\big)$$

□

Let us note that, as opposed to the previous case, the transformation from moment to natural parameters can be performed locally at each conditional distribution, and as stated above, the global solution is just an aggregation of the local solutions.

4 Conditional distributions with multinomial parents in ce-BNs

In many real cases, BNs contain conditional distributions with multinomial parents. In this section, we show that this conditional probability distributions has some particular inner structure that can be exploited when representing them in exponential form and, also, when performing the associated parameter transformations.

Let $(\mathbf{Z}, \mathbf{Y})$ be the set of parents of a variable X such that $\mathbf{Y}$ denotes a set of multinomial variables and $\mathbf{Z}$ denotes set of non-multinomial variables[1]. Let q denote the total number of parental configurations for the variables in $\mathbf{Y}$, and $\mathbf{y}^l$ denote the l-th parental configuration, such that $1 \leq l \leq q$. Let us denote by θ_X the parameters defining the

[1] $\mathbf{Z}$ can be empty.

conditional probability of X given $\mathbf{Z}$ and $\mathbf{Y}$, $p_{\theta_X}(X \mid \mathbf{Z}, \mathbf{Y})$. In our case, we restrict ourselves to those BNs which satisfy the so-called *local independence parameter* property [**?**]. Under this assumption, the conditional probability of X_i conditioned to each parental configuration $\mathbf{y}^l$, $p_{\theta_{X,l}}(X \mid \mathbf{Z}, \mathbf{Y} = \mathbf{y}^l)$, is defined by parameter vector $\theta_{X,l}$ and these parameter vectors independent among them.

In this section we show that if we know how to represent in a exponential form a given distribution (i.e., a Poisson, a Normal, a Multinomial, or a Normal distribution with Normal parents, etc.), then we can directly derive the exponential representation of the corresponding distribution conditioned to multinomial parents (i.e. Poisson given Multinomial parents , Normal given Multinomial parents, Multinomial given Multinomial parents, or Normal given Normal and Multinomial parents, etc.). Similarly, we also show that for these conditional distributions the transformation from moment to natural parameters can be further decomposed.

4.1 Representation

Next theorem shows how a conditional distribution with multinomial parents in a ce-BN is represented in *F-form*,

Theorem 4. *The conditional probability distribution of variable X with multinomial parents $\mathbf{Y}$ and non-multinomial parents $\mathbf{Z}$ can be represented in exponential F-form as follows*

- *the natural parameters in F-form θ_f are formed by the composition of the local natural parameters in F-form of the conditional distribution restricted to each parental configuration and their log-conditional normalizers,*

$$\theta_f = \big(\theta_1, \ldots, \theta_q, -B_1(\theta_1), \ldots, -B_q(\theta_q)\big)$$

  where θ_l and $B_l(\theta_l)$ are given by F-form representation of the local conditional distribution $p(X|\mathbf{Z}, \mathbf{Y} = \mathbf{y}^l)$.

- *the sufficient statistics in F-form $s(X, \mathbf{Z}, \mathbf{Y})$ are similarly expressed as follows,*

$$s(X, \mathbf{Z}, \mathbf{Y}) = \begin{pmatrix} s_1(X, \mathbf{Z}) \cdot I(\mathbf{Y} = \mathbf{y}^1) \\ \vdots \\ s_q(X, \mathbf{Z}) \cdot I(\mathbf{Y} = \mathbf{y}^q) \\ I(\mathbf{Y} = \mathbf{y}^1) \\ \vdots \\ I(\mathbf{Y} = \mathbf{y}^q) \end{pmatrix}$$

  where $s_l(X, \mathbf{Z})$ corresponds to sufficient statistics in F-form for the conditional distribution $p(X|\mathbf{Z}, \mathbf{Y} = \mathbf{y}^l)$.

- *the log-normalizer in F-form $A_f(\theta)$ is the null function.*

Proof. The log-conditional probability of X given its parent-nodes $\mathbf{Z}$ and $\mathbf{Y}$ decomposes as follows:

$$\begin{aligned}
\ln p(X \mid \mathbf{Z}, \mathbf{Y}) &= \sum_{l=1}^{q} I(\mathbf{Y} = \mathbf{y}^l) \cdot \ln p(X|\mathbf{Z}, \mathbf{y}^l) \\
&= \sum_{l=1}^{q} I(\mathbf{Y} = \mathbf{y}^l) \cdot \Big(\theta_l^T s_l(X, \mathbf{Z}) - B_l(\theta_l)\Big) \\
&= \sum_{l=1}^{q} \theta_l^T I(\mathbf{Y} = \mathbf{y}^l) s(X, \mathbf{Z}) - \sum_{l=1}^{q} I(\mathbf{Y} = \mathbf{y}^l) B_l(\theta_l)
\end{aligned}$$

where θ_l, $s_l(X, \mathbf{Z})$ and $B_l(\theta_l)$ are provided when the local conditional distribution $p(X|\mathbf{Z}, \mathbf{y}^l)$ is expressed in exponential form. So, the conditional distribution $p(X \mid \mathbf{Z}, \mathbf{Y})$ can be written in exponential form as follows:

$$\begin{aligned}
\ln p(X \mid \mathbf{Z}, \mathbf{Y}) &= \theta^T s(X, \mathbf{Y}) - B(\theta) \\
&= \begin{pmatrix} -B_1(\theta_1) \\ \vdots \\ -B_q(\theta_q) \\ \theta_1 \\ \vdots \\ \theta_q \end{pmatrix}^T \begin{pmatrix} I(\mathbf{Y} = \mathbf{y}^1) \\ \vdots \\ I(\mathbf{Y} = \mathbf{y}^q) \\ s_1(X, \mathbf{Z}) \cdot I(\mathbf{Y} = \mathbf{y}^1) \\ \vdots \\ s_q(X, \mathbf{Z}) \cdot I(\mathbf{Y} = \mathbf{y}^q) \end{pmatrix} - 0
\end{aligned} \tag{4.1}$$

□

As can be seen, the exponential representation of a conditional probability with multinomial parents can be expressed as the composition of the exponential representation of the conditional distributions restricted to each one of the possible configurations of the multinomial parents.

In Appendix B, we also show how the C-form and P-form representations of the conditional distribution of $p(X_i|\mathbf{Z}, \mathbf{Y})$ can be equally expressed as the composition of the respective C-form and P-form representation of its local distributions.

4.2 From natural to moment parameters

According to Theorem 2, the moment parameter vector decomposed for each variable X_i. In the next theorem we show who the local moment parameter μ_i further decomposes if the conditional probability has a set of multinomial parents by exploiting the how the sufficient statistics vector in F-form decomposes for this kind of conditional probability as shown in Theorem 4.

Theorem 5. *In a ce-BN, the moment parameter vector μ_i associated to a variable X_i with multinomial parents $\mathbf{Y}$ and non-multinomial parents $\mathbf{Z}$ decomposes into local moment parameters associated to each of the configurations of the variables in $\mathbf{Y}$ as follows,*

$$\mu_i = \begin{pmatrix} \lambda_{i,1} \cdot \mu_{i,l} \\ \vdots \\ \lambda_{i,q} \cdot \mu_{i,q} \\ \lambda_{i,1} \\ \vdots \\ \lambda_{i,q} \end{pmatrix}$$

where $\lambda_{i,l}$ is the l-th component (i.e., a scalar) of the moment vector associated to the marginal distribution $p_\theta(\mathbf{Y})$, $\lambda_{i,q} = \int I(\mathbf{Y} = \mathbf{y}_l) p_\theta(\mathbf{Y}) d\mathbf{Y} = p_\theta(\mathbf{y}_l)$, and $\mu_{i,l}$ is the "local" moment parameter associated to the marginal distribution $p(X, \mathbf{Z}|\mathbf{y} = l)$, $\mu_{i,l} = \int s_l(X, \mathbf{Z}) p_\theta(X, \mathbf{Z}|\mathbf{y} = l) dX\mathbf{Z}$.

Proof. The proof easily follows by using the decomposition of the sufficient statistics vector of this conditional probability distribution as shown in Theorem 4. □

Again, we can see that the moment parameters of this conditional distribution can be expressed as a composition of the local moment parameters of each of the distributions conditioned to each configuration of the multinomial parents variables.

4.3 From moment to natural parameters

In Theorem 6, we saw that transforming moment to natural parameters reduces to local transformations between the local moment parameters μ_i and the local natural parameters θ_i by solving the following optimization problem: $\theta_i(\mu_i) = \arg\max_{\theta_i \in \Theta_i} \theta_i^T \mu_i - B(\theta_i)$. Next theorem shows how this optimization problem further simplifies when X_i is conditioned to a set of multinomial parents.

Theorem 6. *Given a local moment parameter vector μ_i associated to a variable X_i with multinomial parents $\mathbf{Y}$ and non-multinomial parents $\mathbf{Z}$, the associated natural parameter vector $\theta_i(\mu_i)$ decomposes as follows,*

$$\theta_i(\mu_i) = \big(\theta_{i,1}, \ldots, \theta_{i,q}, -A_i(\theta_{i,1}), \ldots, -A_i(\theta_{i,q})\big)$$

where, as a mentioned earlier, $\theta_{i,l}$ denotes the parameter sub-vector associated to the F-form representation of the conditional distribution $p_{\theta_{i,l}}(X|\mathbf{Z}, \mathbf{Y} = \mathbf{y}_l)$. The $\theta_{i,l}$ parameter is obtained by solving the following optimization problem,

$$\theta_{i,l} = \arg \max_{\theta_{i,l} \in \Theta_{i,l}} \theta_{i,l}^T \mu'_{i,l} - A_i(\theta_{i,l})$$

where $\mu'_l = \frac{1}{\lambda_{i,l}}\mu_{i,l}$*, i.e. the element-wise division of the vector* $\mu_{i,l}$ *by the scalar* $\lambda_{i,l}$*. So, this optimization problem simply corresponds to the transformation of a moment parameter vector* $\mu'_{i,l}$ *to their corresponding natural parameters for the conditional distribution* $p(X|\mathbf{Z}, \mathbf{y}_l)$*.*

Proof. The optimization problem to solve can be stated as follows by using the decomposition of the natural parameters, of the sufficient statistics, and of the log-normalizer as stated in Theorem 4,

$$\arg \max_{(\theta_{i,1},\ldots,\theta_{i,q})} \sum_{l=1}^{q} \theta_{i,l}^T \mu_{i,l} - \mu_l A_i(\theta_{i,l})$$

Under the *local parameters independence assumption*, the above optimization problem decomposes in a set of q independent optimization problems,

$$\arg \max_{\theta_{i,l} \in \Theta_{i,l}} \theta_{i,l}^T \mu_{i,l} - \mu_l A_i(\theta_{i,l})$$

The solution of each of above optimization problem is not affected if the optimized expression is divided by the scalar $\lambda_{i,l}$ [2]. So we arrived to the claim of the theorem. □

5 Maximum likelihood

In this section we look at the maximum likelihood problem in ce-BNs. Our aim is to show how the solution to this multi-dimensional optimization problem has a common characterization in terms of exponential family representation and, moreover, boils down to smaller local problems for each conditional distribution[3].

Let us assume that we are given a set of m i.i.d. data samples $D = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\}$ indexed by j. The maximum likelihood problem can be stated as follows:

[2] If $\lambda_{i,l} = 0$, it would imply that $p(\mathbf{Y} = \mathbf{y}_t) = 0$, so it does not make sense to solve the problem.

[3] We point out that the following derivation is made without assuming that our models belongs to the regular exponential family and, in consequence, not using the equality of Equation **??**.

$$
\begin{aligned}
\theta^\star &= \arg\max_{\theta\in\Theta} \sum_{j=1}^{m} \ln p(\mathbf{x}^{(j)}|\theta) \\
&= \arg\max_{\theta\in\Theta} \sum_{j=1}^{m} \theta^T s(\mathbf{x}^{(j)}) - A(\theta) \\
&= \arg\max_{\theta\in\Theta} \theta^T \Big(\sum_{j=1}^{m} s(\mathbf{x}^{(j)}) \Big) - mA(\theta) \\
&= \arg\max_{\theta\in\Theta} \theta^T \Big(\frac{1}{m} \sum_{j=1}^{m} s(\mathbf{x}^{(j)}) \Big) - A(\theta)
\end{aligned}
$$

where the last part is achieved by dividing the optimized equation by the number of samples m, what does not affect the result of the optimization.

As widely known, the maximum likelihood is equivalent to a transformation form moment to natural parameters as stated in Equation 2.5:

$$
\theta^\star = \theta(\mu) = \theta\Big(\frac{1}{m} \sum_{j=1}^{m} s(\mathbf{x}^{(j)}) \Big)
$$

As shown in Section 3.3, this problem decomposes for each CPD. The above formation can be expressed in terms of local transformations defined in Equation 3.3:

$$
\theta_i^\star = \theta_i\Big(\frac{1}{m} \sum_{j=1}^{m} s(x_i^{(j)}, \mathbf{pa}_i^{(j)}) \Big)
$$

To better understand the above decomposition, we should notice that $\theta_i(\mu_i)$ is directly related to maximum likelihood estimation of the conditional distribution $p(X_i|Pa(X_i))$:

$$
\begin{aligned}
\theta^\star &= \arg\max_{\theta_i\in\Theta_i} \sum_{j=1}^{m} \ln p(x_i^{(j)}|\mathbf{pa}_i^{(j)}, \theta) \\
&= \arg\max_{\theta_i\in\Theta_i} \sum_{j=1}^{m} \Big(\theta_i^T s(x_i^{(j)}, \mathbf{pa}_i^{(j)}) - B_i(\theta_i) \Big) \\
&= \arg\max_{\theta_i\in\Theta_i} \theta_i^T \Big(\sum_{j=1}^{m} s(x_i^{(j)}, \mathbf{pa}_i^{(j)}) \Big) - mB_i(\theta) \\
&= \arg\max_{\theta_i\in\Theta_i} \theta_i^T \Big(\frac{1}{m} \sum_{j=1}^{m} s(x_i^{(j)}, \mathbf{pa}_i^{(j)}) \Big) - B_i(\theta)
\end{aligned}
$$

where $\mathbf{pa}_i^{(j)}$ corresponds to the X_i parents values in the j-th data sample $\mathbf{x}^{(j)}$.

For those conditional distributions with multinomial parents, the problem further decomposes as previously shown in Section 4.3.

6 EM algorithms in ce-BNs

7 Variational inference in ce-BNs

8 Expectation propagation inference in ce-BNs

Appendix A Regular EF: moment parameters equal the gradient of the log-normalizer

Appendix B Conditional distributions with multinomial parents: proof of EF representation

Appendix C EF representation: A binary child given a binary parent

Let X and Y be two binary variables. The log-conditional probability of the child-node X given its parent-node Y is expressed as follows:

$$\begin{aligned}\ln p(X \mid Y) = I(X = x^1)I(Y = y^1)\ln p_{x^1|y^1} + I(X = x^2)I(Y = y^1)\ln p_{x^2|y^1} \\ +I(X = x^1)I(Y = y^2)\ln p_{x^1|y^2} + I(X = x^2)I(Y = y^2)\ln p_{x^2|y^2}\end{aligned}$$

This conditional probability distribution can be expressed in different exponential forms as follows:

- **First form**:

$$
\begin{aligned}
\ln p(X \mid Y) &= \theta^T s(X,Y) - A(\theta) \\
&= \begin{pmatrix} \ln p_{x^1|y^1} \\ \ln p_{x^2|y^1} \\ \ln p_{x^1|y^2} \\ \ln p_{x^2|y^2} \end{pmatrix}^T \begin{pmatrix} I(X=x^1)I(Y=y^1) \\ I(X=x^2)I(Y=y^1) \\ I(X=x^1)I(Y=y^2) \\ I(X=x^2)I(Y=y^2) \end{pmatrix} - 0 \\
&= \begin{pmatrix} \theta_{11} \\ \theta_{21} \\ \theta_{12} \\ \theta_{22} \end{pmatrix}^T \begin{pmatrix} I(X=x^1)I(Y=y^1) \\ I(X=x^2)I(Y=y^1) \\ I(X=x^1)I(Y=y^2) \\ I(X=x^2)I(Y=y^2) \end{pmatrix} - 0
\end{aligned}
$$

- **Second form**:

$$
\begin{aligned}
\ln p(X \mid Y) &= \theta(Y)^T s(X) - A(Y) \\
&= \begin{pmatrix} I(Y=y^1)\ln p_{x^1|y^1} + I(Y=y^2)\ln p_{x^1|y^2} \\ I(Y=y^1)\ln p_{x^2|y^1} + I(Y=y^2)\ln p_{x^2|y^2} \end{pmatrix}^T \begin{pmatrix} I(X=x^1) \\ I(X=x^2) \end{pmatrix} - 0 \\
&= \begin{pmatrix} m_1^Y \cdot \theta_{11} + m_2^Y \cdot \theta_{12} \\ m_1^Y \cdot \theta_{21} + m_2^Y \cdot \theta_{22} \end{pmatrix}^T \begin{pmatrix} I(X=x^1) \\ I(X=x^2) \end{pmatrix} - 0
\end{aligned}
$$

- **Third form**:

$$
\begin{aligned}
\ln p(X \mid Y) &= \theta(X)^T s(Y) - A(X) \\
&= \begin{pmatrix} I(X=x^1)\ln p_{x^1|y^1} + I(X=x^2)\ln p_{x^2|y^1} \\ I(X=x^1)\ln p_{x^1|y^2} + I(X=x^2)\ln p_{x^2|y^2} \end{pmatrix}^T \begin{pmatrix} I(Y=y^1) \\ I(Y=y^2) \end{pmatrix} - 0 \\
&= \begin{pmatrix} m_1^X \cdot \theta_{11} + m_2^X \cdot \theta_{21} \\ m_1^X \cdot \theta_{12} + m_2^X \cdot \theta_{22} \end{pmatrix}^T \begin{pmatrix} I(Y=y^1) \\ I(Y=y^2) \end{pmatrix} - 0
\end{aligned}
$$

Appendix D EF representation: A multinomial child given a set of multinomial parents

Let X be a multinomial variable with k possible values such that $k \geq 2$, and let $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ denote the set of parents of X, such that all of them are multinomial. Each parent Y_i, $1 \geq i \geq n$, has r_i possible values or states such that $r_i \geq 2$. A parental configuration for the child-node X is then a set of n elements $\{Y_1 = y_1^v, \ldots, Y_i = y_i^v, \ldots, Y_n = y_n^v\}$ such that y_i^v denotes a potential value of variable Y_i such that $1 \leq v \leq r_i$. Let $q = r_1 \times \ldots \times r_n$ denote the total number of parental configurations, and let $\mathbf{y}^l$ denote the l^{th} parental configuration such that $1 \leq l \leq q$.

The log-conditional probability of the child-node X given its parent-nodes $\mathbf{Y}$ can be expressed as follows:

$$\ln p(X \mid \mathbf{Y}) = \sum_{j=1}^{k} \sum_{l=1}^{q} I(X = x^j) I(\mathbf{Y} = \mathbf{y}^l) \ln p_{x^j|\mathbf{y}^l}$$

Similarly the above log-conditional probability can be expressed in the following exponential forms:

- **First form**:

$$\begin{aligned}
\ln p(X \mid \mathbf{Y}) &= \theta^T s(X, \mathbf{Y}) - A(\theta) \\
&= \begin{pmatrix} \ln p_{x^1|\mathbf{y}^1} \\ \vdots \\ \ln p_{x^1|\mathbf{y}^q} \\ \vdots \\ \ln p_{x^k|\mathbf{y}^1} \\ \vdots \\ \ln p_{x^k|\mathbf{y}^q} \end{pmatrix}^T \begin{pmatrix} I(X = x^1) I(\mathbf{Y} = \mathbf{y}^1) \\ \vdots \\ I(X = x^1) I(\mathbf{Y} = \mathbf{y}^q) \\ \vdots \\ I(X = x^k) I(\mathbf{Y} = \mathbf{y}^1) \\ \vdots \\ I(X = x^k) I(\mathbf{Y} = \mathbf{y}^q) \end{pmatrix} - 0 \\
&= \begin{pmatrix} \theta_{11} \\ \vdots \\ \theta_{1q} \\ \vdots \\ \theta_{k1} \\ \vdots \\ \theta_{kq} \end{pmatrix}^T \begin{pmatrix} I(X = x^1) I(\mathbf{Y} = \mathbf{y}^1) \\ \vdots \\ I(X = x^1) I(\mathbf{Y} = \mathbf{y}^q) \\ \vdots \\ I(X = x^k) I(\mathbf{Y} = \mathbf{y}^1) \\ \vdots \\ I(X = x^k) I(\mathbf{Y} = \mathbf{y}^q) \end{pmatrix} - 0
\end{aligned}$$

- **Second form**:

$$\begin{aligned}
\ln p(X \mid \mathbf{Y}) &= \theta(\mathbf{Y})^T s(X) - A(\mathbf{Y}) \\
&= \begin{pmatrix} I(\mathbf{Y} = \mathbf{y}^1) \ln p_{x^1|\mathbf{y}^1} + \ldots + I(\mathbf{Y} = \mathbf{y}^q) \ln p_{x^1|\mathbf{y}^q} \\ \vdots \\ I(\mathbf{Y} = \mathbf{y}^1) \ln p_{x^k|\mathbf{y}^1} + \ldots + I(\mathbf{Y} = \mathbf{y}^q) \ln p_{x^k|\mathbf{y}^q} \end{pmatrix}^T \begin{pmatrix} I(X = x^1) \\ \vdots \\ I(X = x^k) \end{pmatrix} - 0 \\
&= \begin{pmatrix} \mathbf{m}_1^{\mathbf{Y}} \cdot \theta_{11} + m_q^{\mathbf{Y}} \cdot \theta_{1q} \\ \vdots \\ \mathbf{m}_1^{\mathbf{Y}} \cdot \theta_{k1} + m_q^{\mathbf{Y}} \cdot \theta_{kq} \end{pmatrix}^T \begin{pmatrix} I(X = x^1) \\ \vdots \\ I(X = x^k) \end{pmatrix} - 0
\end{aligned}$$

such that $\mathbf{m}_1^{\mathbf{Y}} = \prod_{i=1}^n I(Y_i = y_i^1) = \prod_{i=1}^n m_1^{Y_i}$ denotes the expected sufficient statistics for the first parental configuration, and $\mathbf{m}_q^{\mathbf{Y}} = \prod_{i=1}^n I(Y_i = y_i^{r_i}) = \prod_{i=1}^n m_{r_i}^{Y_i}$ denotes the expected sufficient statistics for the last parental configuration.

- **Third form**:

$$\begin{aligned}
\ln p(X \mid \mathbf{Y}) &= \theta(X)^T s(\mathbf{Y}) - A(X) \\
&= \begin{pmatrix} I(X = x^1) \ln p_{x^1|\mathbf{y}^1} + \ldots + I(X = x^k) \ln p_{x^k|\mathbf{y}^1} \\ \vdots \\ I(X = x^1) \ln p_{x^1|\mathbf{y}^q} + \ldots + I(X = x^k) \ln p_{x^k|\mathbf{y}^q} \end{pmatrix}^T \begin{pmatrix} I(\mathbf{Y} = \mathbf{y}^1) \\ \vdots \\ I(\mathbf{Y} = \mathbf{y}^q) \end{pmatrix} - 0 \\
&= \begin{pmatrix} m_1^X \cdot \theta_{11} + \ldots + m_k^X \cdot \theta_{k1} \\ \vdots \\ m_1^X \cdot \theta_{1q} + \ldots + m_k^X \cdot \theta_{kq} \end{pmatrix}^T \begin{pmatrix} I(\mathbf{Y} = \mathbf{y}^1) \\ \vdots \\ I(\mathbf{Y} = \mathbf{y}^q) \end{pmatrix} - 0
\end{aligned}$$

$$
\begin{aligned}
\ln p(X \mid \mathbf{Y}) &= \theta(X, \mathbf{Y}')^T s(Y_i) - A(X) \text{ such that } \mathbf{Y}' = \mathbf{Y} \setminus Y_i \\
&= \begin{pmatrix} m_1^X I(\mathbf{Y}' = \mathbf{y}'^1) \ln p_{x^1|\mathbf{y}'^1} + \ldots + m_k^X I(\mathbf{Y}' = \mathbf{y}'^1) \ln p_{x^k|\mathbf{y}'^1} \\ \vdots \\ m_1^X I(\mathbf{Y}' = \mathbf{y}'^{q'}) \ln p_{x^1|\mathbf{y}'^{q'}} + \ldots + m_k^X I(\mathbf{Y}' = \mathbf{y}'^{q'}) \ln p_{x^k|\mathbf{y}'^{q'}} \end{pmatrix}^T \begin{pmatrix} I(Y_i = y_i^1) \\ \vdots \\ I(Y_i = y_i^{r_i}) \end{pmatrix} - 0 \\
&= \begin{pmatrix} m_1^X \cdot \mathbf{m}_1^{\mathbf{Y}'} \cdot \theta'_{11} + \ldots + m_k^X \cdot \mathbf{m}_1^{\mathbf{Y}'} \cdot \theta'_{k1} \\ \vdots \\ m_1^X \cdot \mathbf{m}_{q'}^{\mathbf{Y}'} \cdot \theta'_{1q'} + \ldots + m_k^X \cdot \mathbf{m}_{q'}^{\mathbf{Y}'} \cdot \theta'_{kq'} \end{pmatrix}^T \begin{pmatrix} I(Y_i = y_i^1) \\ \vdots \\ I(Y_i = y_i^{r_i}) \end{pmatrix} - 0
\end{aligned}
$$

where $\mathbf{m}_1^{\mathbf{Y}'} = I(\mathbf{Y}' = \mathbf{y}'^1) = I(Y_1 = y_1^1) \cdot \ldots I(Y_{i-1} = y_{i-1}^1) \cdot I(Y_{i+1} = y_{i+1}^1) \cdot \ldots I(Y_n = y_n^1)$ denotes the expected sufficient statistics for the first configuration of the parent set $\mathbf{Y}' = \mathbf{Y} \setminus Y_i$, and $\mathbf{m}_{q'}^{\mathbf{Y}'} = I(\mathbf{Y}' = \mathbf{y}'^{q'}) = I(Y_1 = y_1^{q'}) \cdot \ldots I(Y_{i-1} = y_{i-1}^{q'}) \cdot I(Y_{i+1} = y_{i+1}^{q'}) \cdot \ldots I(Y_n = y_n^{q'})$ denotes the expected sufficient statistics for the last configuration of the parent set $\mathbf{Y}'$, with $q' = q/r_i$ denotes the total number of configurations of the parent set $\mathbf{Y}'$.

Appendix E EF representation: A multinomial child given a Dirichlet parent

Let X be a multinomial variable with k possible values such that $k \geq 2$, and let ρ denote a dirichlet parents of X.

The log-conditional probability of the child-node X given its parent-nodes $\mathbf{Y}$ and ρ can be expressed as follows:

$$\ln p(X \mid \mathbf{Y}, \rho) = \sum_{j=1}^{k} I(X = x^j) \ln p_{x^j}$$

Note that here all $I(\cdot)$ and $p(\cdot)$ values must be taken from the moment parameters of these variables, not the natural parameters of X.

Similarly the above log-conditional probability can be expressed in the following exponential forms:

- **Second form**:

$$\begin{aligned} \ln p(X \mid \mathbf{Y}) &= \theta(\rho)^T s(X) - A(\mathbf{Y}) \\ &= \begin{pmatrix} \ln p_{x^1} \\ \vdots \\ \ln p_{x^k} \end{pmatrix}^T \begin{pmatrix} I(X = x^1) \\ \vdots \\ I(X = x^k) \end{pmatrix} - 0 \end{aligned}$$

- **Third form**:

$$\begin{aligned}
\ln p(X \mid \mathbf{Y}) &= \theta(X)^T s(\rho) - A(X) \\
&= \begin{pmatrix} I(X = x^1) \\ \vdots \\ I(X = x^k) \end{pmatrix}^T \begin{pmatrix} \ln p_{x^1} \\ \vdots \\ \ln p_{x^k} \end{pmatrix} - 0
\end{aligned}$$

E.1 Dirichlet distribution

Let ρ be a dirichlet variable with parameters $\mathbf{u}, \mathbf{p}$, where $\mathbf{p}$ are the k parameters of a multinomial distribution. The log-conditional probability of ρ can be expressed as follows:

$$\begin{aligned}
\ln p(\rho) &= \ln \left(\frac{\Gamma(\sum_{i=1}^k u_i)}{\prod_{i=1}^k \Gamma(u_i)} \prod_{i=1}^k p_i^{u_i - 1} \right) \\
&= \begin{pmatrix} u_1 - 1 \\ \vdots \\ u_k - 1 \end{pmatrix}^T \begin{pmatrix} \log p_1 \\ \vdots \\ \log p_k \end{pmatrix} - \left(\sum_{i=1}^k \log \Gamma(u_i) - \log \Gamma(\sum_{i=1}^k u_i) \right) + 0
\end{aligned}$$

NB: More details to be found in the EF_DirichletDistribution.pdf scanned file.

Appendix F EF representation: A normal child given a set of normal parents

Let X be a normal variable and $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ denote the set of parents of X, such that all of them are normal.

The log-conditional probability of X given its parents $\mathbf{Y}$ can be expressed as follows:

$$\begin{aligned} \ln p(X|Y_1, \ldots, Y_n) &= \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\mathrm{e}^{-\frac{(x-(\beta_0+\boldsymbol{\beta}^T\cdot\mathbf{Y}))^2}{2\sigma^2}}\right) \\ &= -\ln\sigma - 0.5\ln(2\pi) - \frac{(x-\beta_0-\boldsymbol{\beta}^T\mathbf{Y})^2}{2\sigma^2} \end{aligned}$$

where

$$\boldsymbol{\beta} = \begin{pmatrix}\beta_1\\ \vdots \\ \beta_n\end{pmatrix} \qquad \mathbf{Y} = \begin{pmatrix}Y_1\\ \vdots \\ Y_n\end{pmatrix}$$

Similarly the above log-conditional probability can be expressed in the following exponential forms:

- **First form - Joint suff. stat. (Maxim. Likelihood, matrix representation)**:

$$\mathbf{Z} = (X, \mathbf{Y})$$

$$\begin{aligned} \ln p(X \mid \mathbf{Y}) &= \theta^T s(X, \mathbf{Y}) - A(\theta) + h(\mathbf{X}) \\ &= \begin{pmatrix}\boldsymbol{\theta}_1\\ \boldsymbol{\theta}_2\end{pmatrix}^T \begin{pmatrix}\mathbf{Z}\\ \mathbf{Z}^T\mathbf{Z}\end{pmatrix} - \left(\frac{\beta_0^2}{2\sigma^2} + \ln\sigma\right) - \tfrac{1}{2}\ln(2\pi) \end{aligned}$$

$$\mathbf{Z} = \begin{pmatrix}X & Y\end{pmatrix}$$

$$\mathbf{Z}^T\mathbf{Z} = \begin{pmatrix}XX^T & X\mathbf{Y}\\ \mathbf{Y}X^T & Y\mathbf{Y}^T\end{pmatrix}$$

$$\boldsymbol{\theta}_1 = \begin{pmatrix}\beta_0\sigma^{-2} & -\beta_0\boldsymbol{\beta}\sigma^{-2}\end{pmatrix} = \begin{pmatrix}\theta_{\beta_0} & \boldsymbol{\theta}_{\beta_0\boldsymbol{\beta}}\end{pmatrix} = \beta_0\sigma^{-2}\begin{pmatrix}1 & -\boldsymbol{\beta}\end{pmatrix}$$

$$\boldsymbol{\theta}_2 = \begin{pmatrix}-0.5\sigma^{-2} & \boldsymbol{\beta}0.5\sigma^{-2} \\ \boldsymbol{\beta}0.5\sigma^{-2} & -\boldsymbol{\beta}\boldsymbol{\beta}^T 0.5\sigma^{-2}\end{pmatrix} = \begin{pmatrix}\theta_{\text{-}1} & \boldsymbol{\theta}_{\boldsymbol{\beta}} \\ \boldsymbol{\theta}_{\boldsymbol{\beta}}^T & \boldsymbol{\theta}_{\boldsymbol{\beta}\boldsymbol{\beta}}\end{pmatrix} = 0.5\sigma^{-2}\begin{pmatrix}-1 & \boldsymbol{\beta} \\ \boldsymbol{\beta} & -\boldsymbol{\beta}\boldsymbol{\beta}^T\end{pmatrix}$$

– **From moment to natural parameters:**

∗ FIRST STEP:

$$\begin{aligned}\mu_X &= E(X) \\ \mu_{\mathbf{Y}} &= \mathrm{E}(\mathbf{Y}) \\ \Sigma_{XX} &= \mathrm{E}(XX^T) - \mathrm{E}(X)\,\mathrm{E}(X)^T \\ \Sigma_{\mathbf{YY}} &= \mathrm{E}(\mathbf{Y}\mathbf{Y}^T) - \mathrm{E}(\mathbf{Y})\,\mathrm{E}(\mathbf{Y})^T \\ \Sigma_{X\mathbf{Y}} &= \mathrm{E}(X\mathbf{Y}) - \mathrm{E}(X)\,\mathrm{E}(\mathbf{Y})^T \\ \Sigma_{\mathbf{Y}X} &= \mathrm{E}(\mathbf{Y}X^T) - \mathrm{E}(\mathbf{Y})\,\mathrm{E}(X)\end{aligned}$$

∗ SECOND STEP (Theorem 7.4 in page 253, Koller & Friedman):

$$\begin{aligned}\beta_0 &= \mu_X - \Sigma_{X\mathbf{Y}}\Sigma_{\mathbf{YY}}^{-1}\mu_{\mathbf{Y}} \\ \boldsymbol{\beta} &= \Sigma_{X\mathbf{Y}}\Sigma_{\mathbf{YY}}^{-1} \\ \sigma^2 &= \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}\end{aligned}$$

All natural parameters θ can now be calculated considering these equations.

– **From natural to moment parameters:** Via inference.

- **Second form**:

$$\begin{aligned}\ln p(X \mid \mathbf{Y}) &= \theta(\mathbf{Y})^T s(X) - A\big(\theta(\mathbf{Y})\big) + h(\mathbf{X}) \qquad \text{(F.1)} \\ &= \begin{pmatrix}\frac{\mu_{X|Y}}{\sigma^2} \\ \frac{-1}{2\sigma^2}\end{pmatrix}^T \begin{pmatrix}X \\ X^2\end{pmatrix} - \left(\frac{\beta_0^2}{2\sigma^2} + \ln\sigma\right) - \tfrac{1}{2}\ln(2\pi) \\ &= \begin{pmatrix}\theta_{\beta_0} + 2\boldsymbol{\theta}_{\boldsymbol{\beta}}\mathbf{Y} \\ \theta_{\text{-}1}\end{pmatrix}^T \begin{pmatrix}X \\ X^2\end{pmatrix} \\ &- \left(-0.5\ln(-2\theta_{\text{-}1}) - \boldsymbol{\theta}_{\beta_0\boldsymbol{\beta}}\mathbf{Y} - \boldsymbol{\theta}_{\boldsymbol{\beta}\boldsymbol{\beta}}\mathbf{Y}\mathbf{Y}^T - \frac{\theta_{\beta_0}^2}{4\theta_{\text{-}1}}\right) - \tfrac{1}{2}\ln(2\pi)\end{aligned}$$

- **Third form (for each parent Y_i, $\mathbf{Y}' = \mathbf{Y} \setminus Y_i$)**:

$$\begin{aligned}
\ln p(X \mid \mathbf{Y}) &= \theta(X, \mathbf{Y}')^T s(Y_i) - A\big(\theta(X, \mathbf{Y}')\big) + h(\mathbf{Y}) \\
&= \begin{pmatrix} \boldsymbol{\theta}^i_{\beta_0 \boldsymbol{\beta}} + 2\boldsymbol{\theta}^i_{\boldsymbol{\beta}} X + 2\boldsymbol{\theta}'^{i\cdot}_{\boldsymbol{\beta}\boldsymbol{\beta}} \mathbf{Y} \\ \boldsymbol{\theta}^{ii}_{\boldsymbol{\beta}\boldsymbol{\beta}} \end{pmatrix}^T \begin{pmatrix} Y_i \\ {Y_i}^2 \end{pmatrix} \\
&+ \boldsymbol{\theta}'_{\beta_0 \boldsymbol{\beta}} Y + 2\boldsymbol{\theta}'_{\boldsymbol{\beta}} XY + \boldsymbol{\theta}'_{\boldsymbol{\beta}\boldsymbol{\beta}} \mathbf{Y}\mathbf{Y}^T + \theta_{\beta_0} X + \theta_{\text{-}1} XX \\
&- \left(\frac{\theta^2_{\beta_0}}{4\theta_{\text{-}1}} \right) - \tfrac{1}{2} \ln (2\pi)
\end{aligned}$$

where $\boldsymbol{\theta}^i_{\cdot}$ is the ith component of $\boldsymbol{\theta}_{\cdot}$,
where $\boldsymbol{\theta}'^{i\cdot}_{\boldsymbol{\beta}\boldsymbol{\beta}} = \boldsymbol{\theta}^{i\cdot}_{\boldsymbol{\beta}\boldsymbol{\beta}}$ with $\boldsymbol{\theta}^{ii}_{\boldsymbol{\beta}\boldsymbol{\beta}} = 0$, (vector)
where $\boldsymbol{\theta}'_{\beta_0 \boldsymbol{\beta}} = \boldsymbol{\theta}_{\beta_0 \boldsymbol{\beta}}$ with $\boldsymbol{\theta}^i_{\beta_0 \boldsymbol{\beta}} = 0$,
where $\boldsymbol{\theta}'_{\boldsymbol{\beta}} = \boldsymbol{\theta}_{\boldsymbol{\beta}}$ with $\boldsymbol{\theta}^i_{\boldsymbol{\beta}} = 0$,
where $\boldsymbol{\theta}'_{\boldsymbol{\beta}\boldsymbol{\beta}} = \boldsymbol{\theta}_{\boldsymbol{\beta}\boldsymbol{\beta}}$ with $\boldsymbol{\theta}^{ii}_{\boldsymbol{\beta}\boldsymbol{\beta}} = 0$, $\boldsymbol{\theta}^{i\cdot}_{\boldsymbol{\beta}\boldsymbol{\beta}} = 0$ and $\boldsymbol{\theta}^{\cdot i}_{\boldsymbol{\beta}\boldsymbol{\beta}} = 0$.

Appendix G EF representation: A normal child given a set of normal parents and a inv-gamma parent

Let X be a normal variable and $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ denote the set of parents of X, such that all of them are normal. $beta_0$, $\boldsymbol{\beta} = \{beta_1, \ldots, beta_n\}$ now represents normal variables (prior distributions of of the beta parameters) and γ is an inverse gamma distribution (prior distribution of the parameter σ^2).

The log-conditional probability of X given its parents $\mathbf{Y}, \boldsymbol{\beta}, \gamma$ can be expressed as follows:

$$\begin{aligned}
\ln p(X|Y_1, \ldots, Y_n, \boldsymbol{\beta}, \gamma) &= \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\mathrm{e}^{-\frac{(x-(\beta_0+\boldsymbol{\beta}^T\cdot\mathbf{Y}))^2}{2\sigma^2}}\right) \\
&= -\ln\sigma - 0.5\ln(2\pi) - \frac{(x-\beta_0-\boldsymbol{\beta}^T\mathbf{Y})^2}{2\sigma^2}
\end{aligned}$$

Note that here all $\boldsymbol{\beta}$ and σ^2 values must be taken from the moment parameters of these variables, not the natural parameters of X.

- **Second form (messages from $\boldsymbol{\beta}$ and γ variables to X)**:

  As in F.1, but now the natural parameters should be taken from the moment parameters of the parent variables, and not the natural parameters of X.

$$\begin{aligned}
\ln p(X \mid \mathbf{Y}) &= \theta(\mathbf{Y})^T s(X) - A\big(\theta(\mathbf{Y})\big) + h(\mathbf{X}) \\
&= \begin{pmatrix} \frac{\beta_0}{\sigma^2} + \frac{\boldsymbol{\beta}}{\sigma^2}\mathbf{Y} \\ \frac{-1}{2\sigma^2} \end{pmatrix}^T \begin{pmatrix} X \\ X^2 \end{pmatrix} \\
&- \left(0.5\ln\sigma^2 + \frac{\beta_0\boldsymbol{\beta}}{\sigma^2}\mathbf{Y} + \frac{\boldsymbol{\beta}\boldsymbol{\beta}^T}{2\sigma^2}\mathbf{Y}\mathbf{Y}^T + \frac{\beta_0^2}{2\sigma^2}\right) - \tfrac{1}{2}\ln(2\pi)
\end{aligned}$$

- **Third form (messages from X to β_0)**:

$$\begin{aligned}
\ln p(X \mid \mathbf{Y}, \beta_0, \boldsymbol{\beta}, \gamma) &= \theta(X, \mathbf{Y}, \boldsymbol{\beta}, \gamma)^T s(\beta_0) - A\big(\theta(X, \mathbf{Y}, \boldsymbol{\beta}, \gamma)\big) + h(\mathbf{Y}, \beta_\mathbf{0}, \boldsymbol{\beta}, \gamma) \\
&= \begin{pmatrix} X\sigma^{-2} - \boldsymbol{\beta}^T\mathbf{Y}\sigma^{-2} \\ -0.5\sigma^{-2} \end{pmatrix}^T \begin{pmatrix} \beta_0 \\ \beta_0^2 \end{pmatrix} \\
&- X^2 0.5\sigma^{-2} - \boldsymbol{\beta}^T\boldsymbol{\beta}\mathbf{Y}^T\mathbf{Y}0.5\sigma^{-2} + X\boldsymbol{\beta}^T\mathbf{Y}\sigma^{-2} - \ln\sigma - 0.5\ln(2\pi)
\end{aligned}$$

- **Third form (messages from X to β_i)**:

$$\begin{aligned}
\ln p(X \mid \mathbf{Y}, \beta_0, \boldsymbol{\beta}, \gamma) &= \theta(X, \mathbf{Y}, \beta_0, \boldsymbol{\beta}', \gamma)^T s(\beta_i) - A\big(\theta(X, \mathbf{Y}, \beta_0, \boldsymbol{\beta}', \gamma)\big) + h(\mathbf{Y}, \beta_{\mathbf{0}}, \boldsymbol{\beta}, \gamma) \\
&= \begin{pmatrix} -\beta_0 Y_i \sigma^{-2} + Y_i X \sigma^{-2} - \boldsymbol{\beta}_i' Y_i \mathbf{Y}' \sigma^{-2} \\ -0.5 Y_i Y_i \sigma^{-2} \end{pmatrix}^T \begin{pmatrix} \beta_i \\ {\beta_i}^2 \end{pmatrix} \\
&+ \boldsymbol{\theta}'_{\beta_0 \boldsymbol{\beta}} Y + 2\boldsymbol{\theta}'_{\boldsymbol{\beta}} XY + \boldsymbol{\theta}'_{\boldsymbol{\beta}\boldsymbol{\beta}} \mathbf{Y}\mathbf{Y}^T + \theta_{\beta_0} X + \theta_{\text{-}1} XX \\
&- \begin{pmatrix} \frac{\theta_{\beta_0}^2}{4\theta_{\text{-}1}} \end{pmatrix} - \tfrac{1}{2} \ln(2\pi)
\end{aligned}$$

  where $\boldsymbol{\beta}'_i = \boldsymbol{\beta}$ where $\beta_i = 0$,

- **Third form (messages from X to γ (σ^2))**:

$$\begin{aligned}
\ln p(X|Y_1, \ldots, Y_n, \boldsymbol{\beta}, \gamma) &= \theta(X, \mathbf{Y}, \beta_0, \boldsymbol{\beta})^T s(\gamma) - A\big(\theta(X, \mathbf{Y}, \beta_0, \boldsymbol{\beta})\big) + h(\mathbf{Y}, \beta_{\mathbf{0}}, \boldsymbol{\beta}) \\
&= \begin{pmatrix} -\frac{1}{2} \\ -\frac{(X - \beta_0 - \boldsymbol{\beta}\mathbf{Y})^2}{2} \end{pmatrix}^T \begin{pmatrix} \ln \sigma^2 \\ \frac{1}{\sigma^2} \end{pmatrix} - 0.5 \ln(2\pi)
\end{aligned}$$

- **Third form (for each parent Y_i, $\mathbf{Y}' = \mathbf{Y} \setminus Y_i$)**:

$$\begin{aligned}
\ln p(X \mid \mathbf{Y}) &= \theta(X, \mathbf{Y}')^T s(Y_i) - A\big(\theta(X, \mathbf{Y}')\big) + h(\mathbf{Y}) \\
&= \begin{pmatrix} -\beta_0 \beta_i \sigma^{-2} + \beta_i X \sigma^{-2} - \beta_i \boldsymbol{\beta}_i' \mathbf{Y}' \sigma^{-2} \\ -0.5 \beta_i \beta_i \sigma^{-2} \end{pmatrix}^T \begin{pmatrix} Y_i \\ {Y_i}^2 \end{pmatrix} \\
&+ \boldsymbol{\theta}'_{\beta_0 \boldsymbol{\beta}} Y + 2\boldsymbol{\theta}'_{\boldsymbol{\beta}} XY + \boldsymbol{\theta}'_{\boldsymbol{\beta}\boldsymbol{\beta}} \mathbf{Y}\mathbf{Y}^T + \theta_{\beta_0} X + \theta_{\text{-}1} XX \\
&- \begin{pmatrix} \frac{\theta_{\beta_0}^2}{4\theta_{\text{-}1}} \end{pmatrix} - \tfrac{1}{2} \ln(2\pi)
\end{aligned}$$

  where $\boldsymbol{\beta}'_i = \boldsymbol{\beta} \setminus \beta_i$

G.1 Inverse gamma distribution

Let γ be an inverse gamma variable with parameters α, β. The log-conditional probability of γ can be expressed as follows:

$$\begin{aligned}\ln p(\gamma) &= \ln\left(\frac{\beta^{\alpha}}{\Gamma(\alpha)}\gamma^{-\alpha-1}e^{-\frac{\beta}{\gamma}}\right)\\ &= \begin{pmatrix}-\alpha-1\\ -\beta\end{pmatrix}^{T}\begin{pmatrix}\ln\gamma\\ \frac{1}{\gamma}\end{pmatrix} - (\ln\Gamma(\alpha) - \alpha\ln\beta) + 0\end{aligned}$$

NB: More details to be found in the EF_Gamma_Inv-GammaDistributions.pdf scanned file.

Appendix H EF representation: A base distribution given a binary parent

Let X be any base distribution variable, and let Y be a binary variable. The log-conditional probability of the child-node X given its binary parent-node Y is expressed as follows:

$$\begin{aligned}
&\ln p(X \mid Y) = I(Y = y^1) \ln p_{X|y^1} + I(Y = y^2) \ln p_{X|y^2} \\
&= I(Y = y^1)\Big(\theta_{X1} \cdot s(X) - A(\theta_{X1})\Big) + I(Y = y^2)\Big(\theta_{X2} \cdot s(X) - A(\theta_{X2})\Big) \\
&= I(Y = y^1) \cdot \theta_{X1} \cdot s(X) - I(Y = y^1) \cdot A(\theta_{X1}) + I(Y = y^2) \cdot \theta_{X2} \cdot s(X) - I(Y = y^2) \cdot A(\theta_{X2})
\end{aligned}$$

This conditional probability distribution can be expressed in different exponential forms as follows:

- **First form**:

$$\begin{aligned}
\ln p(X \mid Y) &= \theta^T s(X, Y) - A(\theta) \\
&= \begin{pmatrix} -A(\theta_{X1}) \\ -A(\theta_{X2}) \\ \theta_{X1} \\ \theta_{X2} \end{pmatrix}^T \begin{pmatrix} I(Y = y^1) \\ I(Y = y^2) \\ s(X) \cdot I(Y = y^1) \\ s(X) \cdot I(Y = y^2) \end{pmatrix} - 0
\end{aligned}$$

- **Second form**:

$$\begin{aligned}
\ln p(X \mid Y) &= \theta(Y)^T s(X) - A(\theta(Y)) \\
&= \Big(I(Y = y^1) \cdot \theta_{X1} + I(Y = y^2) \cdot \theta_{X2}\Big) s(X) \\
&\quad -I(Y = y^1) \cdot A(\theta_{X1}) - I(Y = y^2) \cdot A(\theta_{X2}) \\
&= \Big(m_1^Y \cdot \theta_{X1} + m_2^Y \cdot \theta_{X2}\Big) s(X) - m_1^Y \cdot A(\theta_{X1}) - m_2^Y \cdot A(\theta_{X2})
\end{aligned}$$

- **Third form**:

$$\begin{aligned}
\ln p(X \mid Y) &= \theta(X)^T s(Y) - A(X) \\
&= \begin{pmatrix} -A(\theta_{X1}) \\ -A(\theta_{X2}) \\ s(X) \cdot \theta_{X1} \\ s(X) \cdot \theta_{X2} \end{pmatrix}^T \begin{pmatrix} I(Y = y^1) \\ I(Y = y^2) \\ I(Y = y^1) \\ I(Y = y^2) \end{pmatrix} - 0
\end{aligned}$$

Appendix I EF representation: A base distribution given a set of multinomial parents

Let X be any base distribution, and let $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ denote the set of parents of X, such that all of them are multinomial. Each parent Y_i, $1 \geq i \geq n$, has r_i possible values or states such that $r_i \geq 2$. A parental configuration for the child-node X is then a set of n elements $\{Y_1 = y_1^v, \ldots, Y_i = y_i^v, \ldots, Y_n = y_n^v\}$ such that y_i^v denotes a potential value of variable Y_i such that $1 \leq v \leq r_i$. Let $q = r_1 \times \ldots \times r_n$ denote the total number of parental configurations, and let $\mathbf{y}^l$ denote the l^{th} parental configuration such that $1 \leq l \leq q$.

The log-conditional probability of the child-node X given its parent-nodes $\mathbf{Y}$ can be expressed as follows:

$$\begin{aligned}
\ln p(X \mid \mathbf{Z}, Y) &= \sum_{l=1}^{q} I(\mathbf{Y} = \mathbf{y}^l) \cdot \ln p_{X|\mathbf{Z},\mathbf{y}^l} \\
&= \sum_{l=1}^{q} I(\mathbf{Y} = \mathbf{y}^l) \cdot \Big(\theta_{Xl} \cdot s(X, \mathbf{Z}) \cdot A(\theta_{Xl})\Big) \\
&= \sum_{l=1}^{q} I(\mathbf{Y} = \mathbf{y}^l) \cdot \theta_{Xl} \cdot s(X, \mathbf{Z}) - I(\mathbf{Y} = \mathbf{y}^l) \cdot A(\theta_{Xl})
\end{aligned}$$

This conditional probability distribution can be expressed in different exponential forms as follows:

- **First form**:

$$\begin{aligned}
\ln p(X \mid \mathbf{Z}, \mathbf{Y}) &= \theta^T s(X, \mathbf{Z}, \mathbf{Y}) - A(\theta) \\
&= \begin{pmatrix} -A(\theta_{X1}) \\ \vdots \\ -A(\theta_{Xq}) \\ \theta_{X1} \\ \vdots \\ \theta_{Xq} \end{pmatrix}^T \begin{pmatrix} I(\mathbf{Y} = \mathbf{y}^1) \\ \vdots \\ I(\mathbf{Y} = \mathbf{y}^q) \\ s(X, \mathbf{Z}) \cdot I(\mathbf{Y} = \mathbf{y}^1) \\ \vdots \\ s(X, \mathbf{Z}) \cdot I(\mathbf{Y} = \mathbf{y}^q) \end{pmatrix} - 0
\end{aligned}$$

- **Second form**:

$$
\begin{aligned}
\ln p(X \mid \mathbf{Z}, \mathbf{Y}) &= \theta(\mathbf{Z}, \mathbf{Y})^T s(X) - A(\mathbf{Y}) \\
&= \Big(\sum_{l=1}^{q} I(\mathbf{Y} = \mathbf{y}^l) \cdot \theta_{X_l}(\mathbf{Z}) \Big) s(X) - \sum_{l=1}^{q} I(\mathbf{Y} = \mathbf{y}^l) \cdot A(\theta_{X_l}(\mathbf{Z})) \\
&= \Big(\sum_{l=1}^{q} \mathbf{m}_l^{\mathbf{Y}} \cdot \theta_{X_l}(\mathbf{Z}) \Big) s(X) - \sum_{l=1}^{q} \mathbf{m}_l^{\mathbf{Y}} \cdot A(\theta_{X_l}(\mathbf{Z}))
\end{aligned}
$$

- **Third form**:

$$
\begin{aligned}
\ln p(X \mid \mathbf{Y}) &= \theta(X)^T s(\mathbf{Y}) - A(X) \\
&= \begin{pmatrix} -A(\theta_{X1}) \\ \vdots \\ -A(\theta_{Xq}) \\ s(X) \cdot \theta_{X1} \\ \vdots \\ s(X) \cdot \theta_{Xq} \end{pmatrix}^T \begin{pmatrix} I(\mathbf{Y} = \mathbf{y}^1) \\ \vdots \\ I(\mathbf{Y} = \mathbf{y}^q) \\ I(\mathbf{Y} = \mathbf{y}^1) \\ \vdots \\ I(\mathbf{Y} = \mathbf{y}^q) \end{pmatrix} - 0
\end{aligned}
$$

$$
\begin{aligned}
\ln p(X \mid \mathbf{Y}) &= \theta(X, \mathbf{Y}')^T s(Y_i) - A(X) \ \text{ such that } \mathbf{Y}' = \mathbf{Y} \setminus Y_i \\
&= \begin{pmatrix} -A(\theta_{X1}) \\ \vdots \\ -A(\theta_{Xq}) \\ s(X) \cdot \mathbf{m}_1^{\mathbf{Y}'} \cdot \theta'_{X1} + \ldots + s(X) \cdot \mathbf{m}_1^{\mathbf{Y}'} \cdot \theta'_{X1} \\ \vdots \\ s(X) \cdot \mathbf{m}_{q'}^{\mathbf{Y}'} \cdot \theta'_{Xq'} + \ldots + s(X) \cdot \mathbf{m}_{q'}^{\mathbf{Y}'} \cdot \theta'_{Xq'} \end{pmatrix}^T \begin{pmatrix} I(Y_i = y_i^1) \\ \vdots \\ I(Y_i = y_i^{r_i}) \\ I(Y_i = y_i^1) \\ \vdots \\ I(Y_i = y_i^{r_i}) \end{pmatrix} - 0
\end{aligned}
$$

Notations

The list below presents a summary of the used notations:

X	Child variable
k	Range of possible values of a multinomial variable X
j	Index over X values, i.e., $1 \geq j \geq k$
Y	One parent variable
$\mathbf{Y}$	Set of parent variables
n	Number of parent variables
i	Index over parent variables, i.e., $1 \geq i \geq n$
r_i	Range of possible values of a multinomial variable Y_i
q	Total number of configurations of a multinomial parent set $\mathbf{Y}$
l	Index over the possible parental configuration values, i.e., $1 \geq l \geq q$
$\mathbf{y}^l$	The l^{th} configuration of a multinomial parent set $\mathbf{Y}$
θ_{jl}	Equal to $\ln p_{x^j \mid \mathbf{y}^l}$, denoting the log-conditional probability of X in its state j given the l^{th} parent configuration
θ_{Xl}	Equal to $\ln p_{X \mid \mathbf{y}^l}$, denoting the log-conditional probability of a base distribution variable X given the l^{th} parent configuration
p	Probability distribution
m	Expected sufficient statistics
s	Sufficient statistics

-