Machine Learning, 26, 43–63 (1997)

Learning and Updating of Uncertainty in Dirichlet Models

ENRIQUE CASTILLO castie@ccaix3.unican.es
Department of Applied Mathematics and Computational Sciences, University of Cantabria, Santander, 39005, Spain

ALI S. HADI ali-hadi@cornell.edu
Department of Social Statistics, Cornell University, Ithaca, NY 14853-3901

CRISTINA SOLARES solaresc@ccaix3.unican.es
Department of Applied Mathematics and Computational Sciences, University of Cantabria, Santander, 39005, Spain

Editor: Douglas Fisher

Abstract. In this paper we analyze the problem of learning and updating of uncertainty in Dirichlet models, where updating refers to determining the conditional distribution of a single variable when some evidence is known. We first obtain the most general family of prior-posterior distributions which is conjugate to a Dirichlet likelihood and we identify those hyperparameters that are influenced by data values. Next, we describe some methods to assess the prior hyperparameters and we give a numerical method to estimate the Dirichlet parameters in a Bayesian context, based on the posterior mode. We also give formulas for updating uncertainty by determining the conditional probabilities of single variables when the values of other variables are known. A time series approach is presented for dealing with the cases in which samples are not identically distributed, that is, the Dirichlet parameters change from sample to sample. This typically occurs when the population is observed at different times. Finally, two examples are given that illustrate the learning and updating processes and the time series approach.

Keywords: Beta distribution, Dirichlet distribution, Dirichlet conjugate priors, evidence propagation, parameter estimation, prior assessment of hyperparameters, time series.

1. Introduction

An important and interesting problem in artificial intelligence and expert systems is to predict some variables when other related variables have been observed. In this paper we will consider the special case where variables are proportions. For example, suppose that the monthly expenditures of a given population of households consist of the following items:

Y_1: Food,	Y_2: Clothing,	Y_3: Entertainment,	Y_4: Transportation,
Y_5: Traveling,	Y_6: Insurance,	Y_7: Housing,	Y_8: Other.

Consider the random variable $X = (X_1, \ldots, X_8)$, where

$$X_i = \frac{Y_i}{\sum_{j=1}^{8} Y_j}, \quad i = 1, \ldots, 8, \tag{1}$$

are the proportions of the total household expenditures. It is clear that $X_i \geq 0$ and that $\sum_{i=1}^{8} X_i = 1$. Our objective is to predict the variable X_j when a subset of other variables is known. That is, we wish to compute the conditional probabilities $f(x_j|x_{i_1}, \ldots, x_{i_r})$, where $x_{i_1}, \ldots, x_{i_r}$ are known evidence values for the variables $X_{i_1}, \ldots, X_{i_r}$.

Note that many other problems lead to the same structure as the household-expenditures example. Consider an operating system resource that is subject to demands from various sources $Y_1, \ldots, Y_n$ and let X_i be the ratio of the demand from the ith source to the total demand from all sources. Here we wish to predict the demands from various sources given other demands. Other examples with similar structure include the percentage of sales of n firms sharing the market in an oligopoly, the percentage of goals of different soccer teams, and the percentage of the total income associated with all regions in a given country.

Multivariate random variables with these characteristics are usually assumed to have a Dirichlet distribution. This paper focuses on learning Dirichlet models from data and interactions with an expert. We also show how to exploit Dirichlet models for making inferences about proportional variables of the type illustrated above. Our work is related to the problems of learning (e.g., Geiger & Heckerman, 1994; Bouckaert, 1994; Heckerman, Geiger & Chickering, 1994; Cooper & Herskovits, 1992; Castillo, Gutiérrez & Hadi, 1996a) and exploiting Bayesian networks (e.g., Pearl, 1986a,b; Lauritzen & Spiegelhalter, 1988; Cooper, 1990; Heckerman, 1990; Shachter, Andersen & Szolovits, 1994; Castillo & Alvarez, 1991; Castillo, Gutiérrez & Hadi, 1996b).

Different research efforts, particularly those concerned with network learning, differ in the assumptions they make about the probability distributions (e.g., normal) that define environmental observations. Again, we will be concerned with Dirichlet models (to be defined in Section 2). Dirichlet models are a useful tool for dealing with probabilistic models in artificial intelligence and expert systems. Two research problems in these areas are those related to learning and updating of uncertainty.

The rest of the paper is organized as follows. Section 2 describes the Dirichlet distribution and some of its properties. In Section 3 we discuss the problem of learning Dirichlet models and we present one learning method that is based on the Dirichlet natural conjugate distribution. We also propose an estimation method that is based on the posterior mode. Section 4 derives an exact method for updating of uncertainty. Section 5 gives a method for dealing with Dirichlet time series when the Dirichlet parameters change with time. In Section 6, two applications illustrate the methodology. The model is evaluated in Section 7. Finally, a summary and concluding remarks are given in Section 8.

2. The Dirichlet Distribution

Let $(Y_1, \ldots, Y_{k+1})$ be independent random variables having Gamma distributions $G(\theta_1), \ldots, G(\theta_{k+1})$, with probability density function (p.d.f.)

$$g(y;\theta) = \frac{y^{\theta-1}e^{-y}}{\Gamma(\theta)}, x > 0, \tag{2}$$

where $\Gamma(\theta)$ is the gamma function, which for integer values of θ coincides with $(\theta - 1)!$. Let

$$X_i = \frac{Y_i}{\sum_{j=1}^{k+1} Y_j};\ i = 1, \ldots, k. \tag{3}$$

Then, the continuous random variable $(X_1, \ldots, X_k)$ has a k-dimensional Dirichlet distribution $D(\theta_1, \ldots, \theta_k; \theta_{k+1})$, with associated p.d.f.

$$f(x_1, x_2, \ldots, x_k; \theta_1, \theta_2, \ldots, \theta_{k+1}) = \frac{\Gamma(\sum_{j=1}^{k+1} \theta_j)}{\prod_{j=1}^{k+1} \Gamma(\theta_j)} \left(\prod_{j=1}^{k} x_j^{\theta_j - 1} \right) \left(1 - \sum_{j=1}^{k} x_j \right)^{\theta_{k+1} - 1}, \tag{4}$$

at any point in the simplex,

$$S_k = \left\{ (x_i, \ldots, x_k) : x_i \geq 0, i = 1, \ldots, k, \sum_{i=1}^{k} x_i \leq 1 \right\}$$

in $\mathbb{R}^k$ and zero outside. The k-dimensional Dirichlet distribution depends on $k+1$ parameters $\theta_1, \ldots, \theta_{k+1}$, which are all real and positive. As we shall see in Theorem 1, the parameters $\theta_1, \ldots, \theta_k$ are proportional to the mean values of $X_i; i = 1, \ldots, X_k$ and the variances of the X_i's decrease with their sum.

Note that a k-dimensional Dirichlet distribution can be simulated by simulating $k+1$ independent Gamma random variables, and using (3) to obtain the corresponding Dirichlet random variables.

If $(X_1, \ldots, X_k)$ is Dirichlet, we define an extra variable

$$X_{k+1} = 1 - \sum_{i=1}^{k} X_i,$$

such that $X_1 + \ldots + X_{k+1} = 1$ and they can be interpreted as proportions associated with mutually exclusive and exhaustive classes of events. This fact suggests many possible applications of the Dirichlet distribution, some of which have been noted in Section 1.

The Beta distribution $B(\theta_1, \theta_2)$, which depends on two positive real parameters θ_1 and θ_2, has the following p.d.f.

$$f(x) = \frac{\Gamma(\theta_1 + \theta_2)}{\Gamma(\theta_1)\Gamma(\theta_2)} x^{\theta_1 - 1} (1 - x)^{\theta_2 - 1}, \tag{5}$$

for $0 < x < 1$, and $f(x) = 0$, elsewhere. Comparing (5) with (4), we see that $B(\theta_1, \theta_2) = D(\theta_1; \theta_2)$. That is, the Beta distribution is a special case of the Dirichlet distribution. One may think of the Dirichlet distribution as a generalization to k dimensions of the Beta distribution.

The Dirichlet distribution has several important properties, some of which are given below and are proved elsewhere (e.g, Wilks, 1962).

THEOREM 1 *The means, variances and covariances of the Dirichlet random variables are:*

$$E[X_i] = \frac{\theta_i}{\sum_{j=1}^{k+1} \theta_j}, i = 1, \ldots, k, \tag{6}$$

$$Var[X_i] = \frac{\theta_i \sum_{j \neq i} \theta_j}{\left(\sum_{j=1}^{k+1} \theta_j\right)^2 \left(1 + \sum_{j=1}^{k+1} \theta_j\right)}, i = 1, \ldots, k, \tag{7}$$

and

$$Cov[X_r, X_s] = \frac{-\theta_r \theta_s}{\left(\sum_{j=1}^{k+1} \theta_j\right)^2 \left(1 + \sum_{j=1}^{k+1} \theta_j\right)}, \; r \neq s = 1, \ldots, k. \tag{8}$$

Expression (6) shows that the expected value of X_i is the fraction of θ_i with respect to the total sum of θ's. If all θ's are multiplied by the same factor, the means remain unchanged.

Expressions (7) and (8) show also that $Var[X_i]$ and $Cov[X_r, X_s]$ are proportional to θ_i and $\theta_r \theta_s$, respectively. If all θ's are multiplied by a large factor, variances and covariances become very small.

THEOREM 2 *Let $(X_1, \ldots, X_k)$ be a vector of random variables whose distribution is the k-variate Dirichlet distribution $D(\theta_1, \ldots, \theta_k; \theta_{k+1})$. The marginal distribution of $(X_1, \ldots, X_{k_1})$, $k_1 < k$, is the k_1-variate Dirichlet distribution $D(\theta_1, \ldots, \theta_{k_1}; \theta_{k_1+1} + \ldots + \theta_{k+1})$.*

This theorem shows that the Dirichlet family is stable with respect to marginalizations of any order. In other words, if $(X_1, \ldots, X_k)$ is Dirichlet then the marginal distribution of every subset of the variable is also Dirichlet.

The following theorem is used for deriving formulas for updating of uncertainty in Section 4.

THEOREM 3 *If $(x_1, \ldots, x_k)$ is a vector random variable having the k-variate Dirichlet distribution $D(\theta_1, \ldots, \theta_k; \theta_{k+1})$, the conditional random variable $x_k/(1 - x_1 - \ldots - x_{k-1})|x_1, \ldots, x_{k-1}$ has the Beta distribution $B(\theta_k, \theta_{k+1})$.*

The Dirichlet distribution has already been used by many authors in Bayesian networks.

- Klieter (1992) deals with the problem of parameter uncertainty and, instead of fixed values for the parameters, uses Beta distributions (a special case of Dirichlet) for the node probabilities and produces, as inferences, Beta distributions, the variances of which are used to measure their uncertainties.

- Neapolitan & Kenevan (1991) also deal with parameter uncertainty by assuming a Dirichlet distribution as a natural conjugate to multinomial models, calculating the posterior, which is also Dirichlet, and obtaining bounds for the variances.
- Musick (1993) shows how to produce inference distributions rather than simple point probabilities, to measure the degree of confidence of the result, using Beta distributions in general Bayesian networks. He also presents a theory that demonstrates how these distributions can be used for performing exact inference. Musick (1994) integrates this methodology for learning the probability structure of the network.
- Geiger & Heckerman (1994) use Dirichlet distributions as the natural conjugate to multinomial models and show that, under assumptions made by some authors, a Dirichlet prior is unavoidable. Thus, they give a characterization of the Dirichlet distribution. For other examples of Dirichlet as a prior see Good (1976).

The above methods use a multinomial as the parent distribution of the variables in the Bayesian network and the Dirichlet distribution plays an auxiliary role. In this paper, we assume that the parent is Dirichlet, that is, we consider a set of variables $X = (X_1, \ldots, X_k)$, which are assumed to be represented by a Dirichlet probability density, $D(\theta_1, \ldots, \theta_k; \theta_{k+1})$.

3. Learning Dirichlet Models

In this section we address the problem of learning Dirichlet models. There are two types of learning in probability models: *structural learning* and *parametric learning*. Structural learning is concerned with building a graphical model (e.g., a Bayesian network) whose topology can be used to determine the relationships among a set of variables. Bayesian networks usually contain many conditional independence relationships which lead to a substantial reduction in the number of parameters of the joint probability distribution of the variables (parsimony). Parametric learning is concerned with estimating the remaining parameters.

Note that the domains of the random variables $A|B, C$ and $A|B$, where A, B and C are disjoint subsets of X with associated X-index sets I_A, I_B and I_C, are

$$1 - \sum_{j \in I_{B \cup C}} x_j \text{ and } 1 - \sum_{j \in I_B} x_j,$$

respectively. Thus, the domains are different. This implies that $P(A|B, C) \neq P(A|B)$ for all disjoint $A, B, C \subset (X_1, \ldots, X_n)$. Therefore, the resulting Bayesian networks are always complete due to the fact that in a Dirichlet model there are no conditional independences. In other words, the number of parameters cannot be reduced. Hence, we shall be concerned only with parametric learning.

To this end, we assume that the parameters θ are random variables and use Bayes' rule

$$f(\theta; \eta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta) f(\theta; \eta)}{\int\limits_{\theta} f(\mathbf{x}|\theta) f(\theta; \eta) d\theta}, \tag{9}$$

where $f(\theta;\eta)$ is the *prior* distribution, $f(\theta;\eta|\mathbf{x})$ is the *posterior* distribution, and $f(\mathbf{x}|\theta)$ is the *likelihood* of the data.

When $f(\theta;\eta|\mathbf{x})$ can be written as $f(\theta;h(\mathbf{x},\eta))$, that is, when the families of priors and posteriors coincide, we say that this family and the likelihoods are *conjugate*. In this case, the posterior can be obtained by just using the function $h(\mathbf{x},\eta)$, which defines the posterior parameters in terms of the prior parameters and the sample values. For this reason, it is common to choose a family $f(\theta;\eta)$ of conjugate priors. In this way, the above Bayes' rule gives posterior distributions of the same family. The Bayesian approach consists of the following steps:

1. Select the family of priors (normally a conjugate family).
2. Assess the prior distribution on the parameters.
3. Obtain the sample data.
4. Calculate the posterior distribution using $h(\mathbf{x},\eta)$.
5. Estimate the parameters by the posterior mean or mode.

3.1. Conjugate of a Dirichlet Distribution

Bayesian statisticians often work with conjugate priors, which are parametric families of distributions such that their associated posteriors belong to the same families. As mentioned above, the rationale for choosing conjugate priors is that the posterior distributions can be easily obtained from the prior distributions and the sample data. The parameters of the conjugate family are referred to as *hyperparameters*. In this section we derive the most general family that is conjugate to the Dirichlet likelihood, give the $h(\mathbf{x},\eta)$ functions to obtain the posterior parameters, and discuss how to calculate the posterior mode.

To learn (estimate) the parameters of the joint probability distribution B_P, we start by deriving the most general Dirichlet conjugate family. Suppose we have a random sample $(\mathbf{x}_s; s = 1,\ldots,n)$, where $\mathbf{x}_i$ is the vector $(x_{i1},\ldots,x_{ik})$ for the i-th individual, from a population that can be represented by a Dirichlet density function which is given in (4). Then the likelihood of the sample is

$$L = \prod_{i=1}^{n} f(\mathbf{x}_i;\theta) = \prod_{i=1}^{n} \left[\frac{\Gamma\left(\sum_{j=1}^{k+1}\theta_j\right)}{\prod_{j=1}^{k+1}\Gamma(\theta_j)} \left(\prod_{j=1}^{k} x_{ij}^{\theta_j-1}\right) \left(1-\sum_{j=1}^{k} x_{ij}\right)^{\theta_{k+1}-1} \right]. \tag{10}$$

The corresponding log-likelihood, $\log L$, is

$$n\log\frac{\Gamma\left(\sum_{j=1}^{k+1}\theta_j\right)}{\prod_{j=1}^{k+1}\Gamma(\theta_j)} + \sum_{j=1}^{k}(\theta_j-1)\sum_{i=1}^{n}\log x_{ij} + (\theta_{k+1}-1)\sum_{i=1}^{n}\log\left(1-\sum_{j=1}^{k}x_{ij}\right). \tag{11}$$

Arnold, Castillo & Sarabia (1993), and Castillo, *et al*. (1996a) show that, in addition to the classical conjugate families given by DeGroot (1986), many others are possible and they characterize the most general family of conjugate distributions for the exponential families, which includes the Dirichlet, by means of the following theorem.

THEOREM 4 *The most general ℓ-parameter exponential family of prior distributions for $\theta = (\theta_1, \ldots, \theta_m)$ which is conjugate with respect to likelihoods of the exponential family*

$$f(\mathbf{x};\theta) = \exp\left[n\lambda(\theta) + \sum_{j=0}^{m} g_j(\theta)T_j(\mathbf{x})\right]; m < \ell, \tag{12}$$

where by convention $g_0(\theta) = 1$ and $\{T_j(x); j = 0, 1, \ldots, m\}$ is a set of linearly independent functions, is

$$q(\theta|\eta) = \exp\left[\nu(\eta) + u(\theta) + \sum_{i=1}^{m} \eta_i g_i(\eta) + \eta_{m+1}\lambda(\theta) + \sum_{i=m+2}^{\ell} \eta_i s_i(\eta)\right], \tag{13}$$

where $s_{m+2}(\theta), \ldots, s_\ell(\theta)$, $\nu(\eta)$ and $u(\theta)$ are arbitrary functions and $\eta_1, \ldots, \eta_\ell$ are the hyperparameters. Finally, the posterior hyperparameter vector is

$$h(\mathbf{x}, \eta) = (\eta_1 + T_1(\mathbf{x}), \ldots, \eta_m + T_m(\mathbf{x}), \eta_{m+1} + n, \eta_{m+2}, \ldots, \eta_\ell). \tag{14}$$

This theorem gives the most general conjugate exponential family of priors in (13), for the exponential families of likelihoods in (12). Also, the simple relation in (14) allows updating the hyperparameters, that is, obtaining the posterior hyperparameters, from the prior hyperparameters and the observed data $\mathbf{x}$.

Since the Dirichlet distribution belongs to the exponential family, we use it to derive its most general conjugate family of prior-posterior distributions. According to (11), the Dirichlet distribution can be written in the form of (12) by letting

$$\begin{aligned}
&m = k + 1,\\
&\lambda(\theta) = \log\Gamma\left(\sum_{j=1}^{k+1}\theta_j\right) - \sum_{j=1}^{k+1}\log\Gamma(\theta_j),\\
&g_j(\theta) = \theta_j, \quad j = 1, \ldots, k,\\
&g_{k+1}(\theta) = \theta_{k+1},\\
&T_0(x) = -\sum_{j=1}^{k}\sum_{i=1}^{n}\log x_{ij} - \sum_{i=1}^{n}\log\left(1 - \sum_{j=1}^{k} x_{ij}\right),\\
&T_j(x) = \sum_{i=1}^{n}\log x_{ij}, \quad j = 1, \ldots, k,\\
&T_{k+1}(x) = \sum_{i=1}^{n}\log\left(1 - \sum_{j=1}^{k} x_{ij}\right).
\end{aligned} \tag{15}$$

It follows from (13) that the most general conjugate prior distribution of a Dirichlet family with parameters $\theta = \{\theta_1, \ldots, \theta_{k+1}\}$ can be expressed as

$$q(\theta|\eta) \doteq \exp\left(\nu(\eta) + u(\theta) + \sum_{i=1}^{k+1}\eta_i\theta_i + \eta_{k+2}\log\frac{\Gamma\left(\sum_{j=1}^{k+1}\theta_j\right)}{\prod_{j=1}^{k+1}\Gamma(\theta_j)} + \sum_{i=k+3}^{\ell}\eta_i s_i(\theta)\right), \tag{16}$$

where $\nu_(\eta)$, $u(\theta)$ and $s_i(\theta)$; $i = k+3,\ldots,\ell$ are arbitrary functions. The posterior hyperparameters become

$$h(\mathbf{x},\eta) = (\eta_1 + T_1(x),\ldots,\eta_{k+1} + T_{k+1}(x), \eta_{k+2} + n, \eta_{k+3},\ldots,\eta_\ell)\,. \tag{17}$$

In the following, for simplicity and due to the fact that $u(\theta)$ and the parameters η_i; $i = k+3,\ldots,\ell$ are static (they are not altered by the information), we remove them and take

$$q(\theta|\eta) \doteq \exp\left(\sum_{i=1}^{k+1}\eta_i\theta_i + \eta_{k+2}\log\Gamma\left(\sum_{j=1}^{k+1}\theta_j\right) - \eta_{k+2}\sum_{j=1}^{k+1}\log\left(\Gamma(\theta_j)\right)\right). \tag{18}$$

Since $T_j(x) \to -\infty$, as $n \to \infty$, the effect of prior information $(\eta_1,\ldots,\eta_{k+2})$ vanishes as $n \to \infty$. Note also that the sensitivity of the results to the prior can be immediately analyzed using (17).

Due to its complexity (exponential and gamma functions appear in it), the mean of the Dirichlet conjugate distribution (18) cannot be obtained in closed form. Poland (1994) describes the use of posterior modes to propagate evidence in networks. Thus, as an alternative to the means, we use the mode of (18) to estimate the Dirichlet parameters, i.e., we maximize (18), with respect to the θ's, to estimate the θ-parameters.

The maximization can be done using any standard numerical nonlinear optimization procedure. We have used procedure *powell* of Numerical Recipes (Press, Teukolsky, Vetterling & Flannery, 1992) and we have obtained a very good and fast convergence. Powell is a standard nonlinear optimization procedure, which inputs the function to be minimized, the negative of the logarithm of (18), and an initial estimate, and outputs the coordinates of the point where the minimum value is attained.

To avoid precision problems we do the following:

- Maximize the logarithm of (18) instead of (18) itself (note that the gamma function can take very large values).
- Use a numerical procedure for the direct evaluation of the logarithm of the gamma function instead of evaluating the gamma function and taking the logarithm.
- Use parameters $\lambda_i^2 = \theta_i$ to guarantee non-negativity of the parameters.
- For the initial θ-estimates, which are required by any nonlinear maximization procedure, we use the following moment estimators:

$$\hat{\theta}_j = s\hat{\mu}_j;\ j = 1,\ldots,k+1, \tag{19}$$

where

$$s = \frac{1}{k+1}\sum_{i=1}^{k+1}\frac{\hat{\mu}_i(1-\hat{\mu}_i)}{\hat{\sigma}_i^2} - 1, \tag{20}$$

is an estimate of $\sum_{i=1}^{k+1}\theta_j$, $\hat{\mu}_i$ is an estimate of the mean of X_i obtained from the sample, and $\hat{\sigma}_i^2$ is an estimate of the variance of X_i. These estimators are obtained from (6) and (7) by solving, in terms of s and θ_i, the system of equations

$$\begin{aligned}
\mu_i &= \frac{\theta_i}{\sum_{i=1}^{k+1}\theta_i} = \frac{\theta_i}{s}, \\
\sigma_i^2 &= \frac{\theta_i \sum_{j\neq i}\theta_j}{\left(\sum_{j=1}^{k+1}\theta_j\right)^2\left(1+\sum_{j=1}^{k+1}\theta_j\right)},
\end{aligned}$$

for each $i = 1, \ldots, k$ and using the mean estimator for s. For an example see Section 6.

3.2. Convenient Posterior Distributions

When selecting a family of prior distributions to be combined with a given likelihood, the prime consideration is that the resulting posteriors should be members of tractable families of distributions, which are referred to as *Convenient* posterior distributions. Thus, the prior and posterior distributions do not have to belong to the same family. It is enough for them to belong to known and tractable families. However, in the case of the Dirichlet distribution we shall see that no better choice exists than its conjugate family.

Arnold, Castillo & Sarabia (1994) identify the family of priors that leads to convenient posteriors in the sense that they belong to specified, not necessarily identical, exponential families. In addition they have found that this problem only has a solution in exponential families.

THEOREM 5 **Convenient posteriors** (Arnold, *et al*., 1994) *Consider data sets consisting of n observations $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$ (possibly vector valued) from an m-parameter exponential family of the form*

$$f(\mathbf{x};\theta) = \exp\left[\lambda(\theta) + \sum_{j=0}^{m}\theta_j T_j(\mathbf{x})\right], \tag{21}$$

where $\theta_0 = 1$. Then,

$$f(\theta;\mathbf{c}) = \exp\left[c_{00} - \lambda(\theta) + g_0(\theta) + \sum_{i=1}^{t} c_{i0}g_i(\theta)\right], \tag{22}$$

is the form of the most general class of priors on Θ that will lead to posterior densities for θ that belong to the t-parameter exponential family of the form

$$f(\theta,\eta)=\exp\left[\theta(\eta)+\sum_{s=0}^{t}\eta_s g_s(\theta)\right], t\le m \tag{23}$$

where $\eta_0=1$ and

$$\begin{bmatrix}\theta_1\\ \theta_2\\ \vdots\\ \theta_m\end{bmatrix}=\begin{bmatrix}c_{01} & c_{11} & \dots & c_{t1}\\ c_{02} & c_{12} & \dots & c_{t2}\\ \vdots & \vdots & \ddots & \vdots\\ c_{0m} & c_{1m} & \dots & c_{tm}\end{bmatrix}\begin{bmatrix}g_0(\theta)\\ g_1(\theta)\\ \vdots\\ g_t(\theta)\end{bmatrix}. \tag{24}$$

The hyperparameters become

$$\begin{bmatrix}\eta_1(\mathbf{x})\\ \eta_2(\mathbf{x})\\ \vdots\\ \eta_t(\mathbf{x})\end{bmatrix}=\begin{bmatrix}c_{10} & c_{11} & \dots & c_{1m}\\ c_{20} & c_{21} & \dots & c_{2m}\\ \vdots & \vdots & \ddots & \vdots\\ c_{t0} & c_{t1} & \cdots & c_{tm}\end{bmatrix}\begin{bmatrix}T_0(\mathbf{x})\\ T_1(\mathbf{x})\\ \vdots\\ T_m(\mathbf{x})\end{bmatrix}, \tag{25}$$

and the mixed distribution is

$$h(\mathbf{x};\mathbf{c})=\exp\left[c_{00}+T_0(\mathbf{x})-\theta(\mathbf{x})+\sum_{j=1}^{m}c_{0j}T_j(\mathbf{x})\right], \tag{26}$$

where the coefficients $\{c_{ij}; i=0,\dots t; j=0,\dots,m\}$ must make the function (22) integrable.

From (24) and without loss of generality, we conclude that $g_r(\theta)=\theta_{r+1};\ r=0,\dots,m-1$, and then our only freedom consists of choosing the arbitrary functions $g_{r+1}(\theta),\dots,g_t(\theta)$. However, these are static hyperparameters and, consequently, they are not affected by the sample values, i.e., they are not interesting. Thus, we conclude that relaxing the condition of prior and posterior to belong to the same exponential family does not lead to an extension of the family (18). Consequently, based on Theorems 4 and 5 we use family (18).

3.3. Prior Assessment

For the prior assessment we suggest the following methods:

1. Following Klieter (1992), by means of an imaginary sample, i.e., we ask a human expert to guess a sample of size m as the most representative of his/her knowledge. Once this sample is known, we use (17) to get the prior hyperparameters:

$$\eta_{j0}=T_j(x);\ j=1,\dots,k+1\ and\ \eta_{k+2}=m. \tag{27}$$

Note that according to (17), a sample modifies the η-parameters by adding $T_j(x)$ or m to the previous values. The prior values have negligible effect for large m, thus, it seems reasonable assuming $\eta_j = 0;\ j = 1, \ldots, k+2$ to get (27).

2. An alternative to the imaginary sample approach is obtained by observing that the mean of $\log X$, where X has a Beta distribution, is

$$E[\log X] = \int_0^1 \log u u^{r-1}(1-u)^{t-1} du = \frac{\Gamma(r)\Gamma(t)\left(\psi(r) - \psi(r+t)\right)}{\Gamma(r+t)}, \quad (28)$$

where $\psi(z)$ is the digamma function. Then, based on (17), we choose as the prior hyperparameters

$$\eta_{j0} = \hat{T}_j(x) = mE[\log X_j] = m\frac{\Gamma(\hat{\theta}_j)\Gamma(\hat{\gamma}_j)\left(\psi(\hat{\theta}_j) - \psi(\hat{\theta}_j + \hat{\gamma}_j)\right)}{\Gamma(\hat{\theta}_j + \hat{\gamma}_j)}, \quad (29)$$

where $j = 1, \ldots, k+1$,, $\hat{\theta}_j$ is a guess (a human expert assessment) for θ_j, $\hat{\gamma}_j = \sum_{k \neq j} \hat{\theta}_k$ and $\eta_{k+2} = m$.

Note that, in both methods, m measures the relative weight of the human expert information with respect to the information contained in a real sample of size n. For example, if $m = n$, they have the same associated information.

4. Updating of Uncertainty

We now derive exact updating formulas for Dirichlet models symbolically, in the sense that the parameters of the marginal and conditional distributions are explicit expressions of the Dirichlet parameters. Thus, a sensitivity analysis of the influence of these parameters on the probabilities of the nodes can be readily performed.

In Dirichlet models the initial probability distribution of every node X_i follows a Dirichlet distribution $X_i \sim D(\theta_i; \sum_{j \neq i} \theta_j)$. Note that this is equivalent to assuming that $X_i \sim B(\theta_i, \sum_{j \neq i} \theta_j)$. Thus, every node X_i has a Beta distribution with mean and variance given by (6) and (7), respectively.

For the purpose of uncertainty updating, assume that we have the following evidence set $E = \{X_k; k \in I_E \subset \{1, \ldots, k\}\}$. Then, from Theorems 2 and 3, every non-evidential node has a scaled Beta distribution, i.e.,

$$X_i | E \sim \left(1 - \sum_{j \in I_E} x_j\right) D\left(\theta_i; \sum_{j \neq i, j \notin I_E} \theta_j\right), \quad (30)$$

with mean

$$E[X_i|E] = \left(1 - \sum_{j \in I_E} x_j\right) \frac{\theta_i}{\sum\limits_{j \notin I_E} \theta_j}, \tag{31}$$

and variance

$$Var[X_i|E] = \left(1 - \sum_{j \in I_E} x_j\right)^2 \frac{\theta_i \sum\limits_{j \neq i, j \neq I_E} \theta_j}{\left(\sum\limits_{j \notin I_E} \theta_j\right)^2 \left(1 + \sum\limits_{j \notin I_E} \theta_j\right)}. \tag{32}$$

Expressions (31) and (32) show that the conditional mean and variance of a non-evidential node given the evidence are rational functions of the parameters and the values of evidential nodes. For the mean, the polynomials involved are all first degree in each of the parameters and the evidence values. For the variance, the polynomials are first degree in the parameters and second degree in the evidence values in the numerator and third degree in the parameters in the denominator.

In Section 6, we give an example illustrating the use of the above formulas in the updating of uncertainty in Dirichlet models.

5. Dirichlet Time Series

In some cases, data are sampled over time. In these cases the sample cannot be assumed independent and identically distributed (iid). In these cases it seems more appropriate to use a time series model for the θ parameters, so that these parameters at time t can be forecasted from the observations at previous times using a time series model, and variables at time t can be predicted after other variables are observed at the same time, using the Dirichlet model.

Assume that X_{jt} is the observation of the j-th variable at time t. We make a time series assumption for the parameters θ_{jt}. Thus, we assume the auto-regressive $AR(\ell)$ model

$$\theta_{jt} = \sum_{m=1}^{\ell} \alpha_{jt-m} x_{jt-m} + \epsilon_{jt}, \tag{33}$$

where the ϵ_{jt} are assumed iid with zero mean.

The least squares estimates of the α_{jt} parameters are obtained by minimizing the error

$$MSE = [(k+1)(n-\ell)]^{-1} \sum_{t=\ell+1}^{n} \sum_{j=1}^{k+1} \left[\frac{\sum\limits_{m=1}^{\ell} \alpha_{jt-m} x_{jt-m}}{\sum\limits_{r=1}^{k+1} \sum\limits_{m=1}^{\ell} \alpha_{rt-m} x_{rt-m}} - x_{jt} \right]^2, \tag{34}$$

where the expression in brackets is the error in x_{jt} and MSE is the mean square error, since we have divided by the number of fitted values. Since the error in (34) remains the same

if we multiply all α_{jt} parameters by a constant, we can assume, without loss of generality, that $\forall\ j,\ \alpha_{jt-\ell} = 1$. Note that the first ℓ samples cannot be predicted from (33), since predictions are based on ℓ previous samples.

6. Examples of Applications

In this section we illustrate the above methods using the household expenditures domain and data on unemployment in Spain.

6.1. The Household Expenditures Data

6.1.1. Parametric Learning

A random sample representing 20 households is shown in Table 1. Assume that $X \sim D(\theta_1, \ldots \theta_7; \theta_8)$. The following are the steps of the proposed learning method:

- **Step 1**: The human expert is required to guess a representative sample, whose size m measures the expert's relative knowledge. We assume that this sample is given in Table 2.
- **Step 2**: A prior of type (18) is assessed using (27) to get the prior hyperparameters, that is,

  $$\eta_{j0} = T_j(x) = \sum_{i=1}^{n} \log x_{ij}, \quad j = 1, \ldots, 8,$$

  where x_{ij}^0 are the fictitious sample values in Table 2 and $\eta_{90} = 10$ is the fictitious sample size. With this we get

  $$(\eta_{10}, \ldots, \eta_{90}) = (-22.7, -19.8, -26.8, -20.4, -18.6, -32.8, -14.9, -32.2, 10).$$

- **Step 3**: Using the updating formulas (17) for the posterior hyperparameters $\eta_i;\ i = 1, \ldots, 9$, we calculate $T_j(x); j = 1, \ldots, 8$, with the real sample of Table 1, and we get

  $$(\eta_1, \ldots, \eta_9) = (-33, -50, -68, -57, -66, -66, -21, -51, 30).$$

- **Step 4**: We now use the real sample and (19)-(20) to obtain the initial estimate of the parameters $\theta_1, \ldots, \theta_8$ to be used in the powell numerical procedure.
- **Step 5**: We maximize (18) by this procedure and obtain the following parameter estimates for the Dirichlet parameters:

  $$(3.68, 2.47, 1.32, 2.02, 1.68, 1.20, 6.62, 1.72).$$

Table 1. Sample values obtained from 20 individuals.

s	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
1	0.23	0.10	0.02	0.08	0.04	0.11	0.39	0.03
2	0.30	0.09	0.05	0.07	0.04	0.02	0.28	0.14
3	0.18	0.08	0.01	0.04	0.01	0.01	0.48	0.19
4	0.17	0.09	0.03	0.07	0.03	0.04	0.42	0.15
5	0.27	0.15	0.01	0.21	0.03	0.06	0.23	0.04
6	0.30	0.11	0.11	0.07	0.01	0.09	0.30	0.01
7	0.26	0.08	0.02	0.03	0.08	0.04	0.42	0.05
8	0.16	0.17	0.06	0.12	0.02	0.02	0.33	0.12
9	0.19	0.07	0.05	0.04	0.03	0.03	0.52	0.07
10	0.18	0.14	0.11	0.04	0.07	0.01	0.25	0.21
11	0.20	0.11	0.07	0.09	0.04	0.05	0.35	0.09
12	0.13	0.04	0.01	0.01	0.14	0.01	0.39	0.26
13	0.23	0.09	0.01	0.13	0.14	0.04	0.27	0.10
14	0.08	0.17	0.03	0.04	0.06	0.06	0.44	0.12
15	0.08	0.10	0.05	0.11	0.07	0.17	0.28	0.15
16	0.21	0.01	0.06	0.06	0.09	0.01	0.43	0.13
17	0.44	0.02	0.04	0.10	0.01	0.01	0.37	0.02
18	0.09	0.16	0.11	0.03	0.03	0.24	0.29	0.04
19	0.29	0.11	0.02	0.10	0.06	0.05	0.30	0.07
20	0.25	0.07	0.03	0.03	0.01	0.05	0.49	0.05

Table 2. Fictitious sample values given by the human expert to assess the Dirichlet prior.

s	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
1	0.16	0.12	0.10	0.24	0.13	0.03	0.18	0.05
2	0.11	0.06	0.04	0.12	0.25	0.05	0.35	0.02
3	0.16	0.18	0.04	0.13	0.16	0.07	0.16	0.11
4	0.13	0.23	0.12	0.12	0.23	0.02	0.15	0.01
5	0.22	0.17	0.06	0.10	0.16	0.02	0.20	0.06
6	0.05	0.24	0.13	0.09	0.06	0.08	0.33	0.02
7	0.14	0.06	0.03	0.17	0.13	0.08	0.25	0.14
8	0.15	0.15	0.06	0.19	0.10	0.01	0.24	0.10
9	0.12	0.21	0.11	0.11	0.20	0.04	0.19	0.03
10	0.01	0.11	0.09	0.10	0.29	0.05	0.32	0.03

These parameter estimates are not actually far from the real ones because the sample was in fact simulated from a Dirichlet $D(4, 2, 1, 1, 1, 1, 8; 2)$.

To illustrate the asymptotic behavior of the estimation method, Tables 3 and 4 show the posterior hyperparameters and the estimated Dirichlet parameters, respectively, for sample sizes $0, 20, 100, 1000, 10000$ and ∞. The convergence of the estimated Dirichlet parameters to the real ones is apparent from Table 4. We have included the sample size $n = 0$ to allow a comparison with the prior.

Table 3. Posterior hyperparameters for sample sizes $20, 100, 1000$, and 10000.

n	η_1	η_2	η_3	η_4	η_5	η_6	η_7	η_8	η_9
0	-22.7	-19.8	-26.8	-20.4	-18.6	-32.8	-14.9	-32.2	10
20	-33	-50	-68	-57	-66	-66	-21	-51	30
100	-170	-246	-381	-322	-339	-341	-101	-254	110
1000	-1719	-2562	-3580	-3524	-3468	-3522	-957	-2551	1010
10000	-17190	-25539	-35458	-35383	-35557	-35679	-9534	-25437	10010

Table 4. Estimated Dirichlet parameters and errors for various sample sizes n.

n	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\theta}_5$	$\hat{\theta}_6$	$\hat{\theta}_7$	$\hat{\theta}_8$	AE	SSE
0	4.75	5.77	3.07	5.30	6.39	1.92	9.04	1.96	0.575	0.06
20	3.68	2.47	1.32	2.02	1.68	1.20	6.62	1.72	0.239	0.011
100	3.71	2.14	0.89	1.28	1.17	1.07	7.01	1.85	0.092	1.76×10^{-3}
1000	3.94	1.98	0.98	1.02	1.06	1.01	7.89	1.97	0.011	2.39×10^{-5}
10000	3.97	1.98	1.00	1.01	1.00	0.99	7.98	2.00	0.004	1.98×10^{-6}
∞	4.00	2.00	1.00	1.00	1.00	1.00	8.00	2.00	0.00	0.00

To evaluate the behavior of the error as a function of the sample size, we use the following two error measures

$$AE = \sum_{i=1}^{k+1} |E[X_i] - \hat{\mu}_i|, \tag{35}$$

$$SSE = \sum_{i=1}^{k+1} (E[X_i] - \hat{\mu}_i)^2. \tag{36}$$

where $\hat{\mu}_i; i = 1, \ldots, k+1$ are the estimated means of the X_i's from the actual sample (e.g., Table 1). These errors are shown in the last two columns in Table 4, where it can be seen that error decreases as the sample size increases.

6.1.2. Updating of Uncertainty

To illustrate the uncertainty updating procedure of Section 4, assume that $X_1 - X_7$ have a Dirichlet $D(4, 2, 1, 1, 1, 1, 8; 2)$ and consider the evidence set

$$E = \{x_1 = 0.3, x_2 = 0.05, x_3 = 0.03, x_4 = 0.06, x_5 = 0.06, x_6 = 0.04, x_7 = 0.35\},$$

which is assumed to become available sequentially in 7 steps. By using Expressions (6), (7), (31) and (32), we calculate the initial probabilities of the nodes and its associated conditionals at different steps. Figure 1 shows a Mathematica program to propagate this evidence. Table 5 shows the evolution of the means and standard deviations of all nodes from the initial to the final step. Note that the variances decrease when new information is available and that this decrease is larger for variables 2 and 7 because they have the largest θ parameter values.

```
theta={4,2,1,1,1,1,8,2}
evidencenode={1,2,3,4,5,6,7};
evidencevalue={0.3,0.05,0.03,0.06,0.06,0.04,0.35};
n=Length[theta];
ss=Sum[theta[[j]],{j,1,n}];
Print["Initial Step "];
Do[a=theta[[i]];mean1[i]=a/ss;
   var1[i]=a*(ss-a)/(ss^2*(ss+1));
   Print["i=",i," mean = ",N[mean1[i]]," std = ",
   N[Sqrt[var1[i]]]],
{i,1,n}]
fact=1;ss1=ss;
Do[Print["Step ", j];
   ss1-=theta[[evidencenode[[j]]]];
   fact-=evidencevalue[[j]];
    Do[If[i==evidencenode[[j]],
      mean1[i]=evidencevalue[[j]];var1[i]=0,
      If[var1[i]==0,,mean1[i]=fact*theta[[i]]/ss1;
      var1[i]=fact^2*theta[[i]]*(ss1-theta[[i]])/
      (ss1^2*(ss1+1))]];
      Print["i=",i," mean = ",mean1[i]," std =",
      N[Sqrt[var1[i]]]],
   {i,1,n}],
{j,1,Length[evidencenode]}]
```

Figure 1. Mathematica statements for propagation of evidence.

6.2. Unemployment Data in Spain

Table 6 shows the unemployment proportions produced in the 17 autonomies of Spain during the decade 1982–1991. As described in Section 5, we have assumed a Dirichlet random variable $X_t = (X_{1t}, \ldots, X_{17t})$ and we have fitted $AR(\ell)$ time series models for the θ-parameters of the Dirichlet model, for $\ell = 2, \ldots, 6$. The results are illustrated in Tables 7 and 8. Table 7 shows the MSE in (34) and the corresponding sum of squares of the error for the predicted values for the year $t = 1991$

$$TSE = \sum_{j=1}^{k+1} \left[\frac{\sum_{m=1}^{\ell} \alpha_{jt-m} x_{jt-m}}{\sum_{r=1}^{k+1} \sum_{m=1}^{\ell} \alpha_{rt-m} x_{rt-m}} - x_{jt} \right]^2 . \tag{37}$$

Table 5. Means and standard deviations of nodes at different steps of the evidence process.

Variable	Initial	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6	Step 7
1	0.2 (0.087)	0.3 (0)	0.3 (0)	0.3 (0)	0.3 (0)	0.3 (0)	0.3 (0)	0.3 (0)
2	0.1 (0.065)	0.088 (0.56)	0.05 (0)	0.05 (0)	0.05 (0)	0.05 (0)	0.05 (0)	0.05 (0)
3	0.05 (0.048)	0.044 (0.041)	0.046 (0.043)	0.03 (0)	0.03 (0)	0.03 (0)	0.03 (0)	0.03 (0)
4	0.05 (0.048)	0.044 (0.041)	0.046 (0.043)	0.048 (0.044)	0.06 (0)	0.06 (0)	0.06 (0)	0.06 (0)
5	0.05 (0.048)	0.044 (0.041)	0.046 (0.043)	0.048 (0.044)	0.047 (0.043)	0.06 (0)	0.06 (0)	0.06 (0)
6	0.05 (0.048)	0.044 (0.041)	0.046 (0.043)	0.048 (0.044)	0.047 (0.043)	0.045 (0.041)	0.04 (0)	0.04 (0)
7	0.4 (0.107)	0.35 (0.085)	0.37 (0.083)	0.38 (0.081)	0.37 (0.073)	0.36 (0.064)	0.37 (0.055)	0.35 (0)
8	0.1 (0.065)	0.088 (0.056)	0.093 (0.059)	0.095 (0.060)	0.093 (0.058)	0.091 (0.056)	0.092 (0.055)	0.11 (0)

Table 8 shows the unemployment predictions for $t = 1991$ using $AR(\ell)$ models for $\ell = 2, \ldots, 6$. Note the improvement as ℓ increases, where ℓ is the number of previous samples.

An additional evaluation of the model can be performed as follows. We use 6 years of data to fit the parameters and then use the fitted model to predict the remaining years. More precisely, we estimated the time series for the θ parameters with data from the period 1982–1987 and then we used these values to predict the θ parameters for the years 1988–1991, based on their corresponding previous periods of 6 years. Once the θ parameters were known for each year, we used Expression (6) for the predictions. The predicted values are shown in Table 9. The total square error TSE for each year is given in the last row of Table 9. These errors and a comparison with the observed values in Table 6 shows a good agreement.

7. Evaluation of the Proposed Method

The proposed method has been tested for Dirichlet populations up to $n = 200$ variables and the method converged very rapidly on a PowerMac computer (less than three minutes), though the computer time increased substantially with the number k of parameters to be estimated. The method, however, showed some problems with the evaluation of the gamma function for large values of its argument, which were solved by evaluating its logarithm instead. Some problems were also observed in the maximization procedure, when the θ parameters were used, since the powell method enters the negative region; however, using

Table 6. Unemployment data in Spain.

Autonomy	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991
Andalucia	0.192	0.195	0.217	0.211	0.222	0.240	0.243	0.256	0.258	0.260
Aragón	0.025	0.026	0.024	0.026	0.024	0.021	0.022	0.022	0.018	0.019
Asturias	0.026	0.024	0.023	0.026	0.027	0.030	0.030	0.030	0.031	0.027
Baleares	0.013	0.014	0.013	0.011	0.012	0.012	0.011	0.012	0.012	0.011
Canarias	0.040	0.040	0.042	0.045	0.046	0.044	0.044	0.049	0.054	0.059
Cantabria	0.010	0.010	0.011	0.010	0.011	0.012	0.014	0.014	0.014	0.013
León	0.052	0.053	0.053	0.056	0.057	0.056	0.060	0.064	0.063	0.060
La Mancha	0.036	0.034	0.033	0.032	0.030	0.030	0.032	0.034	0.032	0.032
Cataluña	0.205	0.203	0.179	0.171	0.167	0.170	0.164	0.138	0.131	0.126
Valencia	0.106	0.100	0.097	0.096	0.091	0.092	0.088	0.089	0.088	0.097
Extremad.	0.027	0.025	0.035	0.033	0.034	0.035	0.036	0.040	0.040	0.039
Galicia	0.040	0.050	0.050	0.052	0.053	0.051	0.051	0.056	0.058	0.058
Madrid	0.119	0.122	0.124	0.126	0.120	0.104	0.104	0.094	0.096	0.092
Murcia	0.023	0.022	0.020	0.023	0.022	0.024	0.023	0.024	0.025	0.029
Navarra	0.012	0.012	0.011	0.012	0.012	0.011	0.010	0.010	0.010	0.009
Pais Vasco	0.069	0.068	0.065	0.064	0.066	0.065	0.064	0.065	0.067	0.066
La Rioja	0.005	0.004	0.004	0.005	0.005	0.004	0.005	0.004	0.003	0.004

Table 7. Mean square errors, MSE, and total square errors, TSE, in predicting series at $t = 10$ for different $AR(\ell)$ models based on the unemployment data in Spain.

ℓ	MSE	TSE
2	1.47×10^{-5}	2.24×10^{-4}
3	9.93×10^{-6}	2.22×10^{-4}
4	6.46×10^{-6}	1.12×10^{-4}
5	7.75×10^{-7}	8.21×10^{-6}
6	8.76×10^{-10}	2.50×10^{-8}

λ_i^2 parameters was a good solution. The prior assessment using the fictitious sample or the alternative method described in Section 3.3 gave no numerical or convergence problems, though the convergence speed was sensitive to the quality of the prior assessment.

Since we can consider an extra variable $X_{k+1} = 1 - \sum_{i=1}^{k} X_i$, the proposed method is applicable primarily when domain variables represent proportions which sum to 1 over all variables. However, if we work only with the initial k variables, the only constraint is $\sum_{i=1}^{k} X_i \leq 1$, which can be transformed to $\sum_{i=1}^{k} X_i \leq s$ by multiplying the random variable by a constant s, as suggested by the gamma variables in expression (3).

The presented Dirichlet model assumes that we have an iid sample. Thus, the model is only applicable when the assumptions of independent and identical distributed data are reasonable. In some cases where data occur over time this assumption must be carefully checked, since in the majority of cases the parameters can change with time, so that the identically distributed part of iid is violated, and if samples close in time are dependent, the independently distributed part is violated too. If this happens, the time series model

Table 8. Predicted values for 1991 of unemployment in Spain using different $AR(\ell)$ models.

Autonomy	Predicted values for $t = 1991$					Observed
	$\ell = 2$	$\ell = 3$	$\ell = 4$	$\ell = 5$	$\ell = 6$	
Andalucia	0.2670	0.2670	0.2655	0.2606	0.2600	0.260
Aragón	0.0179	0.0181	0.0181	0.0179	0.0185	0.019
Asturias	0.0308	0.0308	0.0300	0.0277	0.0270	0.027
Baleares	0.0114	0.0108	0.0113	0.0099	0.0114	0.011
Canarias	0.0561	0.0584	0.0573	0.0584	0.0590	0.059
Cantabria	0.0145	0.0147	0.0136	0.0130	0.0126	0.013
C.- León	0.0638	0.0618	0.0602	0.0603	0.0602	0.060
C.- La Mancha	0.0323	0.0305	0.0318	0.0317	0.0321	0.032
Cataluña	0.1223	0.1282	0.1311	0.1268	0.1260	0.126
Valencia	0.0870	0.0855	0.0951	0.0973	0.0969	0.097
Extremadura	0.0423	0.0407	0.0399	0.0390	0.0387	0.039
Galicia	0.0590	0.0577	0.0550	0.0568	0.0575	0.058
Madrid	0.0909	0.0922	0.0873	0.0920	0.0923	0.092
Murcia	0.0261	0.0256	0.0273	0.0282	0.0286	0.029
Navarra	0.0093	0.0093	0.0089	0.0105	0.0090	0.009
Pais Vasco	0.0656	0.0652	0.0632	0.0660	0.0657	0.066
La Rioja	0.0034	0.0035	0.0045	0.0040	0.0040	0.004
TSE error	2.2×10^{-4}	2.2×10^{-4}	1.1×10^{-4}	8.2×10^{-6}	2.5×10^{-8}	0.0

Table 9. Predictions for the period 1988–1991 based on the period 1982–1987.

Autonomy	1988	1989	1990	1991
Andalucia	0.25106	0.25240	0.26991	0.26815
Aragón	0.02142	0.02253	0.02101	0.01596
Asturias	0.03342	0.03118	0.03338	0.03245
Baleares	0.01146	0.01075	0.01076	0.01017
Canarias	0.03387	0.04024	0.05654	0.05984
Cantabria	0.01261	0.01223	0.01589	0.01574
C.- León	0.05779	0.06325	0.06732	0.05867
C.- La Mancha	0.03174	0.03068	0.03311	0.02975
Cataluña	0.17830	0.15579	0.13276	0.13171
Valencia	0.09481	0.08397	0.08846	0.08265
Extremadura	0.03805	0.03663	0.04076	0.04018
Galicia	0.05690	0.05330	0.05721	0.05646
Madrid	0.06867	0.10437	0.06828	0.09773
Murcia	0.02565	0.02356	0.02548	0.02496
Navarra	0.01250	0.01102	0.01077	0.00981
Pais Vasco	0.06741	0.06296	0.06522	0.06268
La Rioja	0.00434	0.00516	0.00314	0.00309
TSE	1.76×10^{-3}	5.8×10^{-4}	9.53×10^{-4}	4.15×10^{-4}

in Section 5 must be used. However, if there is no change in time or it is small and the independence assumption is reasonable, the method can be applied.

8. Summary and Conclusions

We have presented some methods for dealing with the problems of learning and propagation of uncertainty in Dirichlet models. A Bayesian method with its most general conjugate family of prior-posterior distributions has been developed for learning Dirichlet models. The difficulties in calculating the posterior is avoided by considering the posterior mode, which can be easily calculated by standard numerical procedures. The conditional means and variances of the variables in the network are found to be rational functions of the parameters and evidence values. This gives rise to updating formulas for the exact updating of uncertainty. The simplicity of these formulas also allows for studying the sensitivity of the results to changes in the parameter values. If the Dirichlet parameters change with time, a time series approach, as described in Section 5, can be used. In this manner, the time series are used for predicting Dirichlet parameters and the Dirichlet model is used to predict variables when some other variables are known. Finally, the methods performed satisfactorily when applied to examples of up to 200 variables. In conclusion, Dirichlet models can be easily implemented with the extra advantage that updating formulas are explicit; thus, allowing for a direct numeric as well as symbolic manipulation.

Acknowledgments

We thank the Associate Editor and three referees for their thorough reading of the manuscript and for providing many good comments and suggestions. We are grateful to the Dirección General de Investigación Científica y Técnica (DGICYT) (project PB92-1056), Iberdrola, and the NATO Research Office for partial support of this work.

References

Arnold, B. C., Castillo, E., & Sarabia, J. M. (1993). Conjugate Exponential Family Priors for Exponential Family Likelihoods. *Statistics*, *25*, 71–77.

Arnold, B. C., Castillo, E., & Sarabia, J. M. (1994). *Priors with Convenient Posteriors*, (Technical Report No. 94-10). Santander, Spain: University of Cantabria.

Bouckaert, R. (1994). Properties of Bayesian Belief Networks Learning Algorithms. *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence* (pp. 102–109). San Francisco, CA: Morgan Kaufmann.

Castillo, E., & Alvarez, E. (1991). *Experts Systems: Uncertainty and Learning*. London: Computational Mechanics Publications and Elsevier Applied Science.

Castillo, E., Gutiérrez, J. M., & Hadi, A. S. (1995). Symbolic Propagation in Discrete and Continuous Bayesian Network Models. In V. Keranen & P. Mitic (Eds.), *Mathematics with Vision: Proceedings of the First International Mathematica Symposium*. Computational Mechanics Publications.

Castillo, E., Gutiérrez, J. M., & Hadi, A. S. (1996a). Parametric Structure of Probabilities in Bayesian Networks. C. Froidevaux & J. Kohlas (Eds.), *Lecture Notes in Artificial Intelligence: Symbolic and Quantitative Approaches to Reasoning and Uncertainty*. New York: Springer-Verlag.

Castillo, E., Gutiérrez, J. M., & Hadi, A. S. (1996b). *Expert Systems and Probabilistic Network Models*. New York: Springer-Verlag.

Cooper, G. F. (1990). The Computational Complexity of Probabilistic Inference Using Bayesian Belief Networks. *Artificial Intelligence*, *42*, 393–405.

Cooper, G. F., & Herskovits, E. (1992). A Bayesian Method for the Induction of Probabilistic Networks from Data. *Machine Learning*, *9*, 309–348.

DeGroot, M. H. (1986). *Probability and Statistics*, Second Edition. Menlo Park, CA: Addison Wesley.

Geiger, D., & Heckerman, D. (1994). Learning Gaussian Networks. *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence* (pp. 235–243). San Francisco, CA: Morgan Kaufmann.

Geiger, D., & Heckerman, D. (1995). A Characterization of the Dirichlet Distribution with Application to Learning Bayesian Networks. *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence* (pp. 196–207). Montreal: Morgan Kaufmann.

Good, I. J. (1976). On the Application of Symmetric Dirichlet Distributions and Their Mixtures to Contingency Tables. *The Annals of Statistics*, *4*, 1159–1189.

Heckerman, D. (1990). An Empirical Comparison of Three Inference Methods. In R. D. Shachter, T. S. Levitt, L. N. Kanal, & J. F. Lemmer (Eds.), *Uncertainty in Artificial Intelligence 4*, Amsterdam: Elsevier Science Publishers.

Heckerman, D., Geiger, D., & Chickering, D. M. (1994). Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence* (pp. 293–301). San Francisco, CA: Morgan Kaufmann.

Klieter, G. (1992). Bayesian Diagnosis in Expert Systems. *Proceedings of the AIJ92*.

Lauritzen, S. L., & Spiegelhalter, D. J. (1988). Local Computations with Probabilities on Graphical Structures and their Application to Expert Systems. *Journal of the Royal Statistical Society, Series B*, *50*, 157–224.

Musick, R. (1993). Maintaining Inference Distributions in Belief Nets. *Proceedings of the Eighth Conference on Uncertainty in Artificial Intelligence*.

Musick, R. (1994). *Belief Network Induction*. Doctoral Dissertation, Computer Science Division, University of California, Berkeley.

Neapolitan R., & Kenevan, J. (1991). Investigation of Variances in Belief Networks. *Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence*.

Pearl, J. (1986a). A Constraint-Propagation Approach to Probabilistic Reasoning. In L. N. Kanal & J. F. Lemmer (Eds.), *Uncertainty in Artificial Intelligence*, Amsterdam: Elsevier Science Publishers.

Pearl, J. (1986b). Fusion, Propagation and Structuring in Belief Networks. *Artificial Intelligence*, *29*, 241–288.

Poland, W. B. (1994). *Decision Analysis with Continuous and Discrete Variables: A Mixture Distribution Approach*. Doctoral Dissertation, Department of Engineering Economic Systems, Stanford University, Stanford, CA.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery B. P. (1992). *Numerical Recipes in C: The Art of Scientific Computing*. London: Cambridge University Press.

Shachter, R. D., Andersen, S. K., & Szolovits, P. (1994). Global Conditioning for Probabilistic Inference in Belief Networks. *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence* (pp. 514–522). San Francisco, CA: Morgan Kaufmann.

Wilks, S. S. (1962). *Mathematical Statistics*, New York: John Wiley & Sons.

Received May 25, 1995
Accepted May 31, 1996
Final Manuscript August 8, 1996