Ann. Inst. Statist. Math.
39 (1987), Part A, 399-405

CHARACTERIZING PRIORS BY POSTERIOR EXPECTATIONS IN MULTIPARAMETER EXPONENTIAL FAMILIES

THEOPHILOS CACOULLOS

(Received Jan. 20, 1986; revised Apr. 7, 1986)

Summary

Exploiting the notion of identifiability of mixtures of exponential families with respect to a vector parameter θ, it is shown that the posterior expectation of θ characterizes the prior distribution of θ. The result is applied to normal and negative multinomial distributions.

1. Introduction

Priors play a fundamental role in Bayesian statistics. Their importance is enchanced by the fact that the posterior mean of a parameter turns out to be its Bayes estimator (predictor) under quadratic loss. It is therefore important to be able to recover the prior underlying a Bayes predictor.

The problem of characterizing the prior distribution of a parameter θ in a given family of distributions $f(\cdot\,|\,\theta)$ has been the object of several papers. In particular, results are available for the one-parameter exponential family

$$f(x\,|\,\theta)=a(x)b(\theta)e^{\theta x} \tag{1.1}$$

where, as a rule, the posterior expectation $m(x)=\mathrm{E}[\theta\,|\,X=x]$ characterizes the prior $\pi(\theta)$ (see Cacoullos and Papageorgiou [3], [4] and references therein). A rigorous treatment of the multiparameter regular exponential family, (2.1) below, is given by Diaconis and Ylvisaker [5]; however, they are mainly concerned with conjugate priors which are characterized by linear posterior expectations $m^*(x)=\mathrm{E}\{\mathrm{E}(X\,|\,\theta)\,|\,X=x\}=ax+b$.

In point of fact, under rather general conditions, it is possible to characterize a large family of priors, not necessarily conjugate, relaxing at the same time the condition of linearity of $m(x)$. This is made

Key words and phrases: Characterizations, priors, exponential family, posterior expectation.

possible by using the concept of identifiability of mixtures (Teicher [7] and Barndorff-Nielsen [1]) and exploiting some known results on it. That is, if it is possible to determine the absolute distribution $f(\cdot)$,

$$f(x)=\int f(x\,|\,\theta)\pi(\theta)d\theta \tag{1.2}$$

for certain $f(\cdot\,|\,\theta)$, then, by identifiability, $\pi(\cdot)$ is also uniquely determined. Here $f(\cdot)$ is determined from $f(\cdot\,|\,\theta)$ and $m(x)$.

This approach was shown to be an efficient tool in characterizing prior (mixing) distributions for continuous [4] (discrete [3]) mixtures when both x and θ are one-dimensional.

This note is motivated by the preceding remarks and the fact that, Teicher [7], the mixture (1.2), in general, is not identifiable when x is one-dimensional and θ is a vector parameter. This is exemplified by the case of mixtures of a normal distribution $N(\mu, \sigma^2)$ with respect to both parameters μ and σ^2. In this situation, the identifiability of the mixture requires a sample of at least two observations. In general, the identifiability of mixtures of d-parameter exponential families requires at least d observations. However, both the sample space and the parameter space can be taken as open sets in R^d—the d-dimensional space of the sufficient statistic (see Diaconis and Ylvisaker [5], p. 271, where the sample space $\mathcal{X}=\{x\}$ is an open set in R^d, and the parameter set $\Theta=\left\{\theta\,|\,\ln\int e^{\theta\cdot x}d\mu(x)\right\}$ is also an open (and convex) set in R^d, the so called *natural parameter space*). Our approach below makes use of this parametrization.

2. The main result

In view of the remarks of the preceding section, without loss of generality, we can restrict our attention to a regular d-parameter exponential family of distributions, with probability density

$$f(x\,|\,\theta)=a(\theta)b(x)e^{\theta\cdot x}\,, \tag{2.1}$$

where $\theta=(\theta_1, \theta_2, \cdots, \theta_d)$ and $x=(x_1, \cdots, x_d)$ plays the role of the sufficient statistic (a dot $(\cdot)$ indicates the scalar product of two vectors). The main result can now be stated as follows.

THEOREM 1. *Let $X=(X_1, X_2, \cdots, X_d)$ be a random vector having an absolutely continuous distribution with density* (2.1) *and $\pi(\theta)$ a prior distribution of θ so that the continuous mixture defined by*

$$f(x)=\int f(x\,|\,\theta)\pi(\theta)d\theta \tag{2.2}$$

is identifiable. Then the posterior expectation vector

$$m(x)=\mathrm{E}[\theta \mid X=x] \tag{2.3}$$

determines the (absolute) density $f(\cdot)$ of X; hence $\pi(\theta)$ is also characterized.

PROOF. Differentiating $f(x \mid \theta)$ w.r.t. x_i, $i=1,\cdots,d$, gives

$$\nabla f(x \mid \theta)=\frac{\nabla b(x)}{b(x)} f(x \mid \theta)+\theta f(x \mid \theta), \qquad \left(\nabla f=\left(\frac{\partial f}{\partial x_1},\cdots,\frac{\partial f}{\partial x_d}\right)\right).$$

Multiplying both sides by $\pi(\theta)$ and then integrating out θ yields, after interchanging differentiation and integration,

$$\frac{\nabla f(x)}{f(x)}=\frac{\nabla b(x)}{b(x)}+m(x). \tag{2.4}$$

The solution of (2.4) determines the (marginal) distribution of X. By the identifiability of the mixture (2.2), the prior $\pi(\theta)$ is also characterized.

Remark 1. As already mentioned, here the posterior mean $m(x)$ is not restricted to be a linear function. However, it should be noted that the posterior mean vector $m(x)$ can be an arbitrary function only to the extent that it is consistent with a given $f(x \mid \theta)$ and a chosen prior $\pi(\theta)$. Here our primary concern is the possibility of identifying the prior from such a suitable (consistent) $m(x)$, not necessarily linear. The solvability of (2.4) requires, for example, that $m(x)$ satisfy the condition

$$\frac{\partial m_i(x)}{\partial x_j}=\frac{\partial m_j(x)}{\partial x_i} \qquad \text{for all} \quad i \neq j. \tag{2.5}$$

The importance of Theorem 1 lies in the possibility of characterizing an infinite variety of priors, namely, all the priors which define an identifiable mixture in (2.1). Thus, once one chooses any such prior, this determines an $m(x)$, which in turn (working backwards) characterizes the prior.

A characterization of $N(\mu, \sigma^2)$. Let $Y_1, Y_2,\cdots, Y_n$ be a random sample of size $n>1$ from a normal $N(\mu, \sigma^2)$ and consider the sufficient statistic $X=(X_1, X_2)$ for $\theta=(\theta_1, \theta_2)$ where we set (see [1], p. 121)

$$X_1=\sum_{i=1}^{n} Y_i, \quad X_2=-\frac{1}{2}\sum_{i=1}^{n} Y_i^2, \qquad \theta_1=\frac{\mu}{\sigma^2}, \quad \theta_2=\frac{1}{\sigma^2}. \tag{2.6}$$

Then the probability density (with respect to Lebesgue measure in R^n)

of $Y=(Y_1,\cdots,Y_n)$ given θ can be written as

$$(2.7)\quad f(y\,|\,\mu,\sigma^2)=f^*(x_1,x_2\,|\,\theta_1,\theta_2)=\left(\frac{\theta_2}{2\pi}\right)^{n/2}\exp\left(-\frac{n\theta_1^2}{4\theta_2}\right)\exp(\theta_1x_1+\theta_2x_2)$$

so that in (2.1) $b(x)\equiv 1$. Thus, by Corollary 4 of [1], the mixture of (2.7) under some prior $\pi(\theta)$ of θ is identifiable. For a concrete application of Theorem 2.1 in terms of $y=(y_1,\cdots,y_n)$, we use the following result (see e.g., Berger [2], p. 158).

LEMMA 1. *Let $Y_1,\cdots,Y_n$, $n\geqq 2$, be a random sample from $N(\mu,\sigma^2)$ and suppose the joint prior density of μ and σ^2 is given by*

$$(2.8)\qquad \pi(\mu,\sigma^2)=\pi_1(\mu\,|\,\sigma^2)\pi_2(\sigma^2)\,,$$

where $\pi_1(\mu\,|\,\sigma^2)$ is $N(\mu_0,\lambda\sigma^2)$ and $\pi_2(\sigma^2)$ is an inverse gamma, $\mathrm{IG}(\alpha_0,\beta_0)$ with parameters $\alpha_0>0$, $\beta_0>0$, i.e., with density

$$g^*(t\,|\,\alpha_0,\beta_0)=[\Gamma(\alpha_0)\beta_0^{\alpha}]^{-1}t^{-(\alpha_0+1)}e^{-1/\beta_0 t}\,.$$

Then

(a) *the joint posterior density of μ and σ^2 given $Y=y$ is*

$$(2.9)\qquad \pi(\mu,\sigma^2\,|\,y)=\pi_1(\mu\,|\,\sigma^2,y)\pi_2(\sigma^2\,|\,y)$$

where $\pi_1(\mu\,|\,\sigma^2,y)$ is $N(\mu(y),\sigma_1^2)$, with

$$(2.10)\qquad \mu(y)=(\mu_0+n\lambda\bar{y})/(n\lambda+1)\,,\qquad \sigma_1^2=(\lambda^{-1}+n)\sigma^2\,,$$

and $\pi_2(\sigma^2\,|\,y)$ is an $\mathrm{IG}(\alpha_1,\beta_1)$ with

$$(2.11)\qquad \alpha_1=\alpha_0+\frac{n-1}{2}\,,\qquad \beta_1^{-1}=\beta_0^{-1}+\frac{1}{2}\sum_{i=1}^{n}(y_i-\bar{y})^2+\frac{n(\bar{y}-\mu_0)^2}{2(1+n\lambda)}\,.$$

(b) *The marginal posterior density of μ given y is a t-distribution with $2\alpha_0+n-1$ degrees of freedom, location parameter $\mu(y)$ and scale parameter $[(\lambda^{-1}+n)(\alpha_0+(n-1)/2)\beta_1]^{-1}$, where $\mu(y)$ and β_1 are defined in* (2.10) *and* (2.11).

(c) *The marginal posterior density of σ^2 given y is $\pi_2(\sigma^2\,|\,y)$, as previously defined.*

As a corollary of Lemma 1 and Theorem 1, we obtain

PROPOSITION 1. *In the notation of* (2.7)–(2.11), *for some real μ_0, $\lambda>0$, $\alpha_0>0$, $\beta_0>0$, we have*

$$(2.12)\qquad m_1(y)=\mathrm{E}(\mu\,|\,y)=\frac{\mu_0+n\lambda\bar{y}}{n\lambda+1}\,,\qquad m_2(y)=\mathrm{E}(\sigma^2\,|\,y)=\frac{1}{(\alpha_1-1)\beta_1}\,,$$

if, and only if, the prior density $\pi(\mu,\sigma^2)$ of (μ,σ^2) is given by (2.8) *and*

the absolute density of Y is

$$f(y)=\frac{f(y\,|\,\mu,\sigma^2)\pi(\mu,\sigma^2)}{\pi(\mu,\sigma^2\,|\,y)}\,, \tag{2.13}$$

which is easily seen to be a multivariate t-distribution.

Remark 2. Observe that the posterior mean vector $m(y)=(m_1(y), m_2(y))$ is not linear in y; nor is it so in terms of the sufficient statistic (x_1, x_2) (see (2.6)); thus, it was possible to characterize the prior for μ and σ^2 (see [5], Theorem 3) without assuming a linear $m(y)$ (it so happened here that $\pi(\mu,\sigma^2\,|\,y)$ and $\pi(\mu,\sigma^2)$ are conjugate).

In terms of the canonical sufficient statistic $X=(X_1, X_2)$ and the natural parameter $\theta=(\theta_1,\theta_2)$, as defined in (2.6), we have the following corollary.

PROPOSITION 2. *Under the assumptions of Proposition* 1,

$$m_1(x)=\mathrm{E}(\theta_1\,|\,x)=\frac{\mu_0+\lambda x_1}{(n\lambda+1)}\alpha_1\beta_1, \qquad m_2(x)=\mathrm{E}(\theta_2|x)=\alpha_1\beta_1 \tag{2.14}$$

where α_1 and β_1 are given by (2.11), *that is, in terms of $x=(x_1, x_2)$*

$$\beta_1^{-1}=\frac{1}{\beta_0}-x_2-\frac{1}{2}\,\frac{x_1^2}{n}+\frac{n(n/x_1-\mu_0)^2}{2(1+n\lambda)}\,,$$

if, and only if, the prior density of $\theta=(\theta_1,\theta_2)$ is

$$\pi^*(\theta_1,\theta_2)=\pi_1^*(\theta_1\,|\,\theta_2)\pi_2^*(\theta_2)$$

where $\pi_1^*(\theta_1\,|\,\theta_2)$ is $N(\mu_0\theta_2, \lambda\theta_2)$ and $\pi_2^*(\theta_2)$ is a gamma with shape parameter $\alpha_0>0$ and scale parameter β_0.

The specification of the marginal density of X can be given as for Y in (2.13); the posterior density $\pi^*(\theta_1,\theta_2\,|\,x)$ is easily found from (2.9) and clearly the posterior density $\pi_2^*(\theta_2\,|\,x)$ of θ_2 given x is a gamma with parameters α_1 and β_1. The posterior $\pi_2^*(\theta_1\,|\,x)$ is rather complicated. The same is true, of course, for the marginal distribution $f^*(x)$ of X. Nonetheless, given $f(x\,|\,\theta)$ and $m_1(x)$ and $m_2(x)$ as in (2.14), it is possible to identify both $\pi^*(\theta_1,\theta_2)$ and $f^*(x)$.

3. The discrete case

The treatment of the multiparameter discrete exponential family of distributions is analogous to the continuous case; the operation of differentiation (or ∇) is replaced by the shift operator $\boldsymbol{E}=(E_1,\cdots,E_d)$ with E_i the partial shift operator, i.e., $E_i f(x_1,\cdots,x_i,\cdots,x_d)=f(x_1,\cdots,$

$x_i+1, \cdots, x_d$).

We can now state the analogue of Theorem 1.

THEOREM 2. *With the notation and assumptions of Theorem* 1 *except that now* $x \in I_0$, *the set of all lattice points in* R^d *with nonnegative components, the posterior mean vector*

$$m(x)=\mathrm{E}[\theta \mid X=x], \qquad x \in I_0$$

characterizes both $\pi(\theta)$ *and* $f(x)$.

PROOF. $\boldsymbol{E}$ operating on $f(x \mid \theta)$ with respect to x gives

$$\boldsymbol{E}f(x \mid \theta)=\frac{\boldsymbol{E}b(x)}{b(x)} f(x \mid \theta)+\theta f(x \mid \theta),$$

and hence the analogue of (2.5):

$$\frac{\boldsymbol{E}f(x)}{f(x)}=\frac{\boldsymbol{E}b(x)}{b(x)}+m(x). \tag{3.1}$$

Hence, as in Theorem 1, the assertion follows.

A similar remark holds as for Theorem 1 and the analogue of (2.5) takes the form

$$E_i m_j(x)=E_j m_i(x) \qquad \text{for all} \quad i \neq j. \tag{3.2}$$

Characterization of the Dirichlet distribution as a prior for the negative multinomial. For an illustration of Theorem 2, consider the negative multinomial distribution with p.f.

$$f(x \mid p)=\frac{\Gamma(r+x_0)}{\Gamma(r) \prod_{i=1}^{d} \Gamma(x_i+1)} p_0^r \prod_{i=1}^{d} p_i^{x_i} \tag{3.3}$$

where we set

$$p_0=1-\sum_{i=1}^{d} p_i, \quad x_0=\sum_{i=1}^{d} x_i, \qquad p=(p_1, p_2, \cdots, p_d).$$

The mixture of $f(x|\theta)$ w.r.t. a Dirichlet prior $\pi(\theta)$ is identifiable by Proposition 1 of [1] and the assumptions of Theorem 2 are satisfied. Hence a Dirichlet prior $\pi(\cdot)$, with

$$\pi(p)=\frac{\Gamma(\alpha)}{\prod_{i=0}^{k} \Gamma(\alpha_i)} \prod_{j=0}^{d} p_j^{\alpha_j-1}, \quad \alpha=\sum_{j=0}^{d} \alpha_i \qquad (\alpha_i>0) \tag{3.4}$$

can be characterized by the posterior expectation

$$m(x)=\mathrm{E}(p\,|\,x)\,, \qquad x\in I_0\,;$$

this is given by

$$m_i(x)=\mathrm{E}(p_i\,|\,x)=\frac{x_i+\alpha_i}{x_0+r+\alpha}\,, \qquad i=1,\cdots,d\,. \tag{3.5}$$

The result is summarized in

PROPOSITION 3. *Let* $X=(X_1,\cdots,X_d)$ *be an observation from* (3.3). *Then* $m(x)$ *is given by* (3.5) *if, and only if, the prior for* p *is the Dirichlet* (3.4) *and the (marginal) density of* X *is a compound negative multinomial distribution, Mosimann* [6], *defined by*

$$f(x_1,\cdots,x_k)=\frac{\Gamma(\alpha)\Gamma(r+x_0)\Gamma(r+\alpha_0)\prod_{i=1}^{d}\Gamma(\alpha_i+x_i)}{\Gamma(\alpha_0)\Gamma(r)\Gamma(x_0+\alpha+r)\prod_{i=1}^{d}\Gamma(\alpha_i)x_i!}\,. \tag{3.6}$$

If the α_i *are integers, then* (3.6) *becomes a multiple generalized hypergeometric distribution.*

UNIVERSITY OF ATHENS

REFERENCES

[1] Barndorff-Nielsen, O. (1965). Identifiability of mixtures of exponential families, *J. Math. Anal. Appl.*, **12**, 115-121.

[2] Berger, J. (1980). *Statistical Decision Theory*, Springer-Verlag.

[3] Cacoullos, T. and Papageorgiou, H. (1983). Characterizations of discrete distributions by a conditional distribution and a regression function, *Ann. Inst. Statist. Math.*, **35**, 95-103.

[4] Cacoullos, T. and Papageorgiou, H. (1984). Characterizations of mixtures of continuous distributions by their posterior means, *Scand. Actuarial J.*, 23-30.

[5] Diaconis, P. and Ylvisaker, D. (1979). Conjugate priors for exponential families, *Ann. Statist.*, **7**, 269-281.

[6] Mosimann, J. E. (1963). On the compound negative multinomial distribution and correlations among inversely sampled pollen counts, *Biometrika*, **50**, 47-54.

[7] Teicher, H. (1960). On the mixture of distributions, *Ann. Math. Statist.*, **31**, 55-73.