Bernoulli **5**(1), 1999, 163–176

On the relationship between α connections and the asymptotic properties of predictive distributions

JOSÉ M. CORCUERA[1]* and FEDERICA GIUMMOLÈ[2]

[1]*Departament d'Estadística, Universitat de Barcelona, Gran Via de les Corts Catalanes, 585, 08007 Barcelona, Spain. e-mail: corcuera@cerber.mat.ub.es*
[2]*Dipartimento di Scienze Statistiche, Università degli Studi di Padova, Via S. Francesco 33, 35121 Padova, Italy. e-mail: giummole@hal.stat.unipd.it*

In a recent paper, Komaki studied the second-order asymptotic properties of predictive distributions, using the Kullback–Leibler divergence as a loss function. He showed that estimative distributions with asymptotically efficient estimators can be improved by predictive distributions that do not belong to the model. The model is assumed to be a multidimensional curved exponential family. In this paper we generalize the result assuming as a loss function any f divergence. A relationship arises between α connections and optimal predictive distributions. In particular, using an α divergence to measure the goodness of a predictive distribution, the optimal shift of the estimate distribution is related to α-covariant derivatives. The expression that we obtain for the asymptotic risk is also useful to study the higher-order asymptotic properties of an estimator, in the mentioned class of loss functions.

Keywords: curved exponential family; differential geometry; f divergences; predictive distributions; second-order asymptotic theory; α connections; α embedding curvature

1. Introduction

The main goal of this work is to provide distributions that are close, in the sense of an f divergence, to an unknown distribution belonging to a curved exponential family

$$\mathscr{P} = \{ p(x;\, \theta(u)) = \exp [\theta^i(u)x_i - \psi(\theta(u))]\}.$$

In order to obtain this, we could estimate u by $\hat{u}$ and consider $p(x;\, \hat{u})$. This kind of distribution is called *an estimative distribution*. The procedure ensures that they belong to the model. However, perhaps we could obtain a better result by considering *predictive distributions*, i.e. distributions outside the model.

Let $\hat{p}(x;\, x_{1\sim N})$ be a predictive distribution obtained by some rule from the sample of size N, $x_{1\sim N} = (x(1), \ldots, x(N))$. An f divergence D_f of the predictive distribution to the true one is defined as

*To whom correspondence should be addressed.

1350–7265

$$D_f(p(x;\,u),\hat{p}(x;\,x_{1\sim N})) = \int f\left(\frac{\hat{p}(x;\,x_{1\sim N})}{p(x;\,u)}\right) p(x;\,u)\mu\,\mathrm{d}x,$$

where f is a smooth, strictly convex function that vanishes at 1. We measure the closeness by

$$\mathrm{E}_u(D_f(p,\hat{p})) = \int D_f(p(x;\,u),\hat{p}(x;\,x_{1\sim N}))p(x_{1\sim N};\,u)\,\mathrm{d}x_{1\sim N}. \tag{1}$$

In order to choose $\hat{p}$, we could try to find the distribution that minimizes (1), uniformly in u, among 'all probability distributions' equivalent to p. Since there are some technical problems in giving a structure of differentiable manifold to this infinite-dimensional space, we follow the procedure suggested by Komaki (1996) and try to solve the problem only for distributions belonging to a finite-dimensional model containing $\mathscr{P}$. We construct this model by enlarging $\mathscr{P}$ in orthogonal directions. We shall see that, for large samples, there is a special direction such that the improvement on the estimative density is maximum if and only if this direction belongs to the tangent space associated to the enlarged model. The solution does not change if we add more orthogonal directions and in this sense we can consider the problem solved in the infinite-dimensional space of all probability distributions equivalent to p.

For simplicity, we shall work with α divergences D_α (for their use in statistical inference, see Amari (1985, Chapter 3)), i.e. f divergences with

$$f(z) = f_\alpha(z) = \begin{cases} \dfrac{4}{1-\alpha^2}(1 - z^{(1+\alpha)/2}), & \alpha \neq \pm 1, \\ z\log z, & \alpha = 1, \\ -\log z, & \alpha = -1. \end{cases}$$

In the final remark, we extend the results to any f divergence.

2. The enlarged model

Let $\mathscr{E}$ be a n-dimensional full exponential family, i.e.

$$\mathscr{E} = \{p(x;\,\theta) = \exp[\theta^i x_i - \psi(\theta)],\ \theta \in \Theta\},$$

where the probability functions $p(x;\,\theta)$ are densities with respect to some σ-finite reference measure μ and

$$\Theta = \left\{\theta\colon \int \exp(\theta^i x_i)\mu\,\mathrm{d}x < \infty\right\}$$

is an open subset of $\mathbb{R}^n$. We consider the model $\mathscr{P}$ to be a $(n,\,m)$-curved exponential family of $\mathscr{E}$, $m \leqslant n$,

$$\mathscr{P} = \{p(x;\,u) = \exp[\theta^i(u)x_i - \psi(\theta(u))],\ u \in U\},$$

with U a smooth m-dimensional submanifold of Θ.

Let

$$l_\alpha(x;\, u) = \begin{cases} \dfrac{2}{1-\alpha}[p^{(1-\alpha)/2}(x;\, u) - 1], & \alpha \neq 1, \\ \log p(x;\, u), & \alpha = 1, \end{cases}$$

be the so-called α representation of $p(x;\, u)$ (Amari 1985, p. 66). From now on, the index α will be used to denote all that regards α representation of geometric quantities. The tangent space T_u of $\mathscr{P}$ in u is identified with the vector space spanned by

$$\partial_a l_\alpha(x;\, u) = \frac{\partial l_\alpha(x;\, u)}{\partial u^a}, \qquad a = 1, \ldots, m,$$

that are the components of what we call the α-score function. The first and second derivatives of $l_\alpha(x;\, u)$ are related to those of $l(x;\, u) = \log p(x;\, u) = l_1(x;\, u)$ by

$$\partial_a l_\alpha = p^{(1-\alpha)/2}\, \partial_a l$$

and

$$\partial_a \partial_b l_\alpha = p^{(1-\alpha)/2}\left(\partial_a \partial_b l + \frac{1-\alpha}{2}\, \partial_a l\, \partial_b l\right).$$

Defining

$$\mathrm{E}_\alpha(f(x)) = \int f(x) p^\alpha(x;\, u)\mu\, \mathrm{d}x,$$

we have that the inner product of vectors $\partial_a l_\alpha$ and $\partial_b l_\alpha$,

$$\langle \partial_a l_\alpha,\, \partial_b l_\alpha \rangle_\alpha = E_\alpha(\partial_a l_\alpha\, \partial_b l_\alpha) = \int \partial_a l_\alpha\, \partial_b l_\alpha\, p^\alpha \mu\, \mathrm{d}x = \int \partial_a l\, \partial_b l\, p\mu\, \mathrm{d}x = \langle \partial_a l,\, \partial_b l \rangle,$$

does not depend on the α representation; it is the $(a,\, b)$ component of the Fisher information matrix g_{ab}. In the sequel, we omit the subscript α in the inner product and in the expectation, since it will be clear from the representation used. We indicate by g^{ab} the inverse of g_{ab} and use the repeated index convention.

Following Amari *et al*. (1987), we can construct a vector bundle on $\mathscr{P}$ by associating to each point $p(x;\, u) \in \mathscr{P}$ a linear space H_u defined by

$$H_u = \left\{ h(x):\, \int p^{(1+\alpha)/2}(x;\, u) h(x) \mu\, \mathrm{d}x = 0,\, \int p^\alpha(x;\, u) h^2(x) \mu\, \mathrm{d}x < \infty \right\}.$$

If $h,\, g \in H_u$ we can define an inner product on H_u by

$$\langle h,\, g \rangle = \int p^\alpha(x;\, u) h(x) g(x) \mu\, \mathrm{d}x.$$

Then, since H_u is a closed linear subspace of $L^2(p^\alpha \mu)$, it is a Hilbert space. It is easy to see that $T_u \subset H_u$ and the inner product defined on T_u is compatible with that in H_u. Attached to different points we have different but isomorphic Hilbert spaces. In order to see this, let $p = p(x; u)$ and $q = p(x; u')$ be two different points of $\mathscr{P}$ and consider the transformation

$$I_u^{u'}\colon H_u \to H_{u'}$$

$$h \mapsto \left(\frac{p}{q}\right)^{\alpha/2} h - q^{(1-\alpha)/2} \int q^{(1+\alpha)/2} \left(\frac{p}{q}\right)^{\alpha/2} h\mu \, \mathrm{d}x.$$

In fact, it is easy to see that $I_u^{u'}(h) \in H_{u'}$, since

$$\int q^{(1+\alpha)/2} I_u^{u'}(h) \mu \, \mathrm{d}x = 0$$

and

$$\int q^\alpha \{I_u^{u'}(h)\}^2 \mu \, \mathrm{d}x = \int p^\alpha h^2 \mu \, \mathrm{d}x - \left[\int q^{(1+\alpha)/2} \left(\frac{p}{q}\right)^{\alpha/2} h\mu \, \mathrm{d}x\right]^2 < +\infty. \tag{2}$$

Moreover, $I_u^{u'}$ is linear, its inverse is

$$(I_u^{u'})^{-1}(g) = \left(\frac{q}{p}\right)^{\alpha/2} \left\{ g - q^{(1-\alpha)/2} \left[\int p^{(1+\alpha)/2} \left(\frac{q}{p}\right)^{\alpha/2} g\mu \, \mathrm{d}x\right] \Big/ \left[\int p^{(1+\alpha)/2} \left(\frac{q}{p}\right)^{\alpha/2} \mu \, \mathrm{d}x\right] \right\}$$

and, by (2), it is bounded. $I_u^{u'}$ is then a continuous linear bijection, i.e. an isomorphism between H_u and $H_{u'}$. The aggregate

$$\mathscr{H}(\mathscr{P}) = \bigcup_{u \in U} H_u$$

constitutes Amari's Hilbert bundle. It is necessary to establish a one-to-one correspondence between H_u and $H_{u'}$, when $p(x; u)$ and $p(x; u')$ are neighbouring points, in order to express the rate of variation in a vector field as an element of the Hilbert bundle. If we move in the direction $\partial_a l_\alpha$ and $h_u \in H_u$, $\partial_a h_u \notin H_u$ in general. Anyway, if

$$h\colon U \to \mathscr{H}(\mathscr{P})$$

is a smooth vector field, in the sense that we can interchange the integral and the derivative,

$$0 = \partial_a \int p^{(1+\alpha)/2} h\mu \, \mathrm{d}x$$

$$= \int p^{(1+\alpha)/2} \, \partial_a h \, \mu \, \mathrm{d}x + \frac{1+\alpha}{2} \int p^\alpha \, \partial_a l_\alpha \, h\mu \, \mathrm{d}x$$

$$= \int p^{(1+\alpha)/2} \left(\partial_a h + \frac{1+\alpha}{2} p^{(1-\alpha)/2} \mathrm{E}(\partial_a l_\alpha \, h)\right) \mu \, \mathrm{d}x.$$

Thus, we can define the α-covariant derivative in $\mathscr{H}$ as

$$\overset{\alpha}{\nabla}{}^{(\mathscr{H})}_{\partial_a l_\alpha} h = \partial_a h + \frac{1+\alpha}{2} p^{(1-\alpha)/2} \mathrm{E}(\partial_a l_\alpha\, h),$$

whenever $\overset{\alpha}{\nabla}{}^{(\mathscr{H})}_{\partial_a l_\alpha} h \in L^2(p^\alpha \mu)$. If $h(u) = \partial_b l_\alpha(x;\, u)$, we have

$$\overset{\alpha}{\nabla}{}^{(\mathscr{H})}_{\partial_a l_\alpha} \partial_b l_\alpha = \partial_a \partial_b l_\alpha + \frac{1+\alpha}{2} p^{(1-\alpha)/2} g_{ab}$$

and the α-covariant derivative in $\mathscr{P}$ is the projection of $\overset{\alpha}{\nabla}{}^{(\mathscr{H})}_{\partial_a l_\alpha} \partial_b l_\alpha$ on T_u:

$$\overset{\alpha}{\nabla}{}_{\partial_a l_\alpha} \partial_b l_\alpha = \langle \overset{\alpha}{\nabla}{}^{(\mathscr{H})}_{\partial_a l_\alpha} \partial_b l_\alpha,\, \partial_c l_\alpha \rangle g^{cd}\, \partial_d l_\alpha = \overset{\alpha}{\Gamma}{}_{abc} g^{cd} \partial_d l_\alpha.$$

These connections coincide with the α connections defined by Amari (1985, p. 38). We use the superscripts m and e respectively for the -1 and $+1$-covariant derivatives.

Let $\mathscr{M}$ be any regular parametric model containing $\mathscr{P}$. We can consider on $\mathscr{M}$ the coordinate system (u, s), where u^a, $a = 1, \ldots, m$, is the old coordinate system on $\mathscr{P}$ and s^I, $I = m+1, \ldots, r$, $r > m$, are new coordinates on $\mathscr{M}$. We suppose that $s = 0$ for the points in the original manifold $\mathscr{P}$ and u and s are orthogonal in $\mathscr{P}$. The tangent space to the enlarged model $\mathscr{M}$ is now spanned by vectors $\partial_a l_\alpha(x;\, u, s)$, $a = 1, \ldots, m$, and $\partial_I l_\alpha(x;\, u, s)$, $I = m+1, \ldots, r$. Let $h_I(x;\, u) = \partial_I l_\alpha(x;\, u, s)|_{s=0}$, $I = m+1, \ldots, r$; then the h_I values belong to H_u and we can formally write

$$p(x;\, u, s) = p(x;\, u) + p^{(1+\alpha)/2}(x;\, u) s^I h_I(x;\, u) + \cdots . \tag{3}$$

3. Predictive distributions

We consider predictive distributions $p(x;\, \hat{u}_N(\overline{x}), \hat{s}(\overline{x}))$, with $\hat{s}(\overline{x}) = O_p(N^{-1})$, so that

$$\hat{s}(\overline{x}) = \frac{1}{N} \overline{s}(\overline{x}) + o_p(N^{-1}), \tag{4}$$

and $\hat{u}_N(\overline{x})$ is a smooth, asymptotically efficient estimator, and hence first-order equivalent to the maximum-likelihood estimator, of the form

$$\hat{u}_N(\overline{x}) = \hat{u}_\infty(\overline{x}) + \frac{1}{N} \overline{u}(\overline{x}) + o_p(N^{-1}). \tag{5}$$

For fixed $\overline{x}$, both

$$\hat{u}_\infty(\overline{x}) = \lim_{N \to \infty} \hat{u}_N(\overline{x})$$

and

$$\overline{u}(\overline{x}) = \lim_{N \to \infty} N\{\hat{u}_N(\overline{x}) - \hat{u}_\infty(\overline{x})\}$$

depend on N only through $\overline{x}$.

For each N, $\hat{u}_N$ is a map

$$\hat{u}_N \colon \mathscr{E} \to \mathscr{P},$$

since $\overline{x}$ can be identified with the point in $\mathscr{E}$ having expectation parameters $\eta_i = \overline{x}_i$. Then, $\hat{u}_\infty$ is also a map from $\mathscr{E}$ to $\mathscr{P}$ and we can associate with $\hat{u}_N$ a family of ancillary $(n-m)$-dimensional submanifolds of $\mathscr{E}$, $\mathscr{A} = \{A(u)\}$, where $A(u) = \hat{u}_\infty^{-1}(u)$. In some discrete cases, even though the exponential model is regular, $\overline{x}$ could correspond to the expectation parameters of a point in $\mathscr{E}$ with a probability different from one. However, since this probability goes to one exponentially in N, we can consider a modification of $\overline{x}$, say x^*, such that $x^* = \overline{x} + o_p(N^{-2})$ and x^* are the expectation coordinates of some point in $\mathscr{E}$. Then, all the results could be rewritten in terms of x^* instead of $\overline{x}$.

Following Amari (1985, p. 128), it can be shown that $\hat{u}_\infty$ is consistent if and only if every $p(x; u) \in \mathscr{P}$ is contained in the associated submanifold $A(u)$ and $\hat{u}_\infty$ is asymptotically first-order efficient if and only if $A(u)$ is orthogonal to $\mathscr{P}$ in u. On the other hand, since

$$\lim_{N\to\infty} \hat{u}_\infty(\overline{x}) = \lim_{N\to\infty} \hat{u}_N(\overline{x})$$

in probability and

$$\lim_{N\to\infty} [N^{1/2}\{\hat{u}_\infty(\overline{x}) - u\}] = \lim_{N\to\infty} [N^{1/2}\{\hat{u}_N(\overline{x}) - u\}]$$

in distribution, the results still hold for $\hat{u}_N$.

If we introduce a coordinate system v^κ, $\kappa = m+1, \ldots, n$ on each $A(u)$, every point in the full exponential family containing $\mathscr{P}$ is uniquely determined by a pair (u, v). It is convenient to fix $v = 0$ for the points in $\mathscr{P}$. We denote by indices $a, b, c, \ldots \in \{1, \ldots, m\}$ the coordinates u in $\mathscr{P}$, by $\kappa, \lambda, \mu, \ldots \in \{m+1, \ldots, n\}$ the coordinates v in $A(u)$ and by $\alpha, \beta, \gamma, \ldots \in \{1, \ldots, n\}$ the new coordinates $w = (u, v)$ in $\mathscr{E}$. Since $\hat{u}_N$ is asymptotically efficient,

$$g_{a\kappa}(u) = 0.$$

Indices $i, j, \ldots \in \{1, \ldots, n\}$ are used to denote the natural parameters θ in $\mathscr{E}$ and indices $I, J, K, \ldots \in \{m+1, \ldots, r\}$ for the coordinates s that we add to enlarge the model $\mathscr{P}$. By the coordinate system we choose on $\mathscr{M}$,

$$g_{aI}(u) = 0.$$

Under these assumptions, we have the following theorem.

Theorem 3.1. *The average α divergence from the true distribution $p(x; u_0)$ to a predictive distribution $p(x; \hat{u}_N(\overline{x}), \hat{s}(\overline{x}))$ is given by*

$$
\begin{aligned}
&\mathrm{E}_{u_0}\{D_\alpha(p(x;\,u_0),\;p(x;\hat{u}_N(\overline{x}),\hat{s}(\overline{x})))\}\\
&\quad=\frac{m}{2N}+\frac{1}{4N^2}[2(\overset{e}{H}{}^2_{\mathscr{P}})+(\overset{m}{H}{}^2_{\mathscr{A}})]\\
&\qquad+\frac{1}{2N^2}\,g_{ab}(\overline{u}^a-\tfrac{1}{2}\overset{m}{H}{}_{\kappa\lambda}{}^a g^{\kappa\lambda})(\overline{u}^b-\tfrac{1}{2}\overset{m}{H}{}_{\mu\nu}{}^b g^{\mu\nu})\\
&\qquad+\frac{1}{N^2}\overset{(1-\alpha)/2}{\nabla}{}_a(\overline{u}^a-\tfrac{1}{2}\overset{m}{H}{}_{\kappa\lambda}{}^a g^{\kappa\lambda})\\
&\qquad+\frac{1}{2N^2}(g_{IJ}\bar{s}^I\bar{s}^J-\overset{\alpha}{H}{}_{abI}\,g^{ab}\bar{s}^I)\\
&\qquad+\frac{\alpha-3}{12N^2}\,T_{abc}T^{abc}+\frac{(\alpha-11)(\alpha-1)}{32N^2}\,Q_{abcd}g^{ab}g^{cd}\\
&\qquad+\frac{1}{4N^2}\,g^{ac}g^{bd}\int(\partial_a\partial_b p-\overset{m}{\Gamma}{}_{ab}{}^e\,\partial_e p)(\partial_c\partial_d p-\overset{m}{\Gamma}{}_{cd}{}^f\,\partial_f p)\frac{1}{p}\mu\,\mathrm{d}x\\
&\qquad-\frac{3}{8N^2}\,g^{ab}g^{cd}\int(\partial_a\partial_b p-\overset{m}{\Gamma}{}_{ab}{}^e\,\partial_e p)(\partial_c\partial_d p-\overset{m}{\Gamma}{}_{cd}{}^f\,\partial_f p)\frac{1}{p}\mu\,\mathrm{d}x\\
&\qquad-\frac{1}{N^2}\,g^{ac}g^{bd}\int\partial_a p\,\partial_b p\,(\partial_c\partial_d p-\overset{m}{\Gamma}{}_{cd}{}^f\,\partial_f p)\frac{1}{p^2}\mu\,\mathrm{d}x\\
&\qquad+\frac{\alpha+1}{8N^2}\,g^{ab}g^{cd}\overset{m}{\nabla}{}_d\,T_{abc}+o(N^{-2}),
\end{aligned}
\tag{6}
$$

where all the quantities are evaluated in u_0,

$$
\begin{aligned}
\overline{u}&=\overline{u}(\mathrm{E}(\overline{x})),\\
\bar{s}&=\bar{s}(\mathrm{E}(\overline{x})),\\
Q_{abcd}&=\mathrm{E}(\partial_a l\,\partial_b l\,\partial_c l\,\partial_d l),\\
\overset{\alpha}{H}{}_{rst}&=\langle\overset{\alpha}{\nabla}{}_{\partial_r l_\alpha}\partial_s l_\alpha,\,\partial_t l_\alpha\rangle,\\
T_{abc}&=\mathrm{E}(\partial_a l\,\partial_b l\,\partial_c l),\\
(\overset{e}{H}{}^2_{\mathscr{P}})&=\overset{e}{H}{}^{ac\kappa}\overset{e}{H}{}^{bd\lambda}g_{cd}g_{\kappa\lambda}g_{ab},\\
(\overset{m}{H}{}^2_{\mathscr{A}})&=\overset{m}{H}{}^{\kappa\lambda a}\overset{m}{H}{}^{\mu\nu b}g_{\kappa\mu}g_{\lambda\nu}g_{ab}
\end{aligned}
$$

and $\overset{\alpha}{\nabla}{}_a$ *is the a component of the general covariant derivative of a tensor with respect to the* α *connection.*

Proof. Only an outline of the proof is given; see Corcuera and Giummolè (1996) for detailed calculations. Since, from the definition of f_α,

$$f_\alpha(1) = 0, \quad f'_\alpha(1) = 1, \quad f'''_\alpha(1) = \frac{\alpha - 3}{2}, \quad f^{(4)}_\alpha(1) = \frac{(\alpha - 3)(\alpha - 5)}{4},$$

and $\hat{s}(\overline{x}) = O_p(N^{-1})$, the expansion of an α divergence from $p(x;\, u_0)$ to $p(x;\, \hat{u}_N, \hat{s})$ is

$$\begin{aligned} D_\alpha(p(x;\, u_0),\, p(x;\, \hat{u}_N, \hat{s})) = {} & \tfrac{1}{2} g_{ab}(u_0)\tilde{u}^a\tilde{u}^b + \tfrac{1}{2} g_{IJ}(u_0)\hat{s}^I\hat{s}^J + \left(\tfrac{1}{2}\overset{\alpha}{\Gamma}_{abc}(u_0) + \frac{\alpha}{3} T_{abc}(u_0)\right)\tilde{u}^a\tilde{u}^b\tilde{u}^c \\ & + [\tfrac{1}{2}(\overset{\alpha}{\Gamma}_{abI}(u_0) + 2\,\overset{\alpha}{\Gamma}_{aIb}(u_0)) + \alpha T_{abI}(u_0)]\tilde{u}^a\tilde{u}^b\hat{s}^I \\ & + K_{abcd}(u_0)\tilde{u}^a\tilde{u}^b\tilde{u}^c\tilde{u}^d + o_p(N^{-2}), \end{aligned}$$

where $\tilde{u} = \hat{u}_N - u_0$ and

$$\begin{aligned} K_{abcd} = \frac{1}{24}\Bigg(& \frac{(\alpha - 3)(\alpha - 5)}{4}\int \frac{\partial_a p\, \partial_b p\, \partial_c p\, \partial_d p}{p^3}\mu\, \mathrm{d}x + \frac{\alpha - 3}{2}\int \frac{\partial_a\partial_b p\, \partial_c p\, \partial_d p}{p^2}\mu\, \mathrm{d}x\,[6] \\ & + \int \frac{\partial_a\partial_b\partial_c p\, \partial_d p}{p}\mu\, \mathrm{d}x\,[4] + \int \frac{\partial_a\partial_b p\, \partial_c\partial_d p}{p}\mu\, \mathrm{d}x\,[3]\Bigg). \end{aligned} \tag{7}$$

The brackets [ ] refers to the sum of a number of different terms obtained by permutation of free indices, e.g.

$$\partial_a\partial_b p\, \partial_c\partial_d p\,[3] = \partial_a\partial_b p\, \partial_c\partial_d p + \partial_a\partial_c p\, \partial_b\partial_d p + \partial_a\partial_d p\, \partial_b\partial_c p.$$

The mean value of D_α is

$$\begin{aligned} \mathrm{E}_{u_0}\{D_\alpha(p(x;\, u_0),\, p(x;\, \hat{u}_N, \hat{s}))\} = {} & \tfrac{1}{2} g_{ab}(u_0)\mathrm{E}_{u_0}[\tilde{u}^a\tilde{u}^b] + \tfrac{1}{2} g_{IJ}(u_0)\mathrm{E}_{u_0}[\hat{s}^I\hat{s}^J] \\ & + \left(\tfrac{1}{2}\overset{\alpha}{\Gamma}_{abc}(u_0) + \frac{\alpha}{3} T_{abc}(u_0)\right)\mathrm{E}_{u_0}[\tilde{u}^a\tilde{u}^b\tilde{u}^c] \\ & + (\tfrac{1}{2}\overset{\alpha}{\Gamma}_{abI}(u_0) + \overset{\alpha}{\Gamma}_{aIb}(u_0) + \alpha T_{abI}(u_0))\mathrm{E}_{u_0}[\tilde{u}^a\tilde{u}^b\hat{s}^I] \\ & + K_{abcd}(u_0)\mathrm{E}_{u_0}[\tilde{u}^a\tilde{u}^b\tilde{u}^c\tilde{u}^d] + o(N^{-2}). \end{aligned} \tag{8}$$

The mean squared error of $\hat{u}_N$ can be written as

$$\mathrm{E}_{u_0}[\tilde{u}^a\tilde{u}^b] = \frac{1}{N} g^{ab} + \frac{1}{N}\partial_c\hat{u}^a_{bias} g^{bc}[2] + \mathrm{E}_{u_0}[(\tilde{u}^a - g^{ac}\tilde{x}_c)(\tilde{u}^b - g^{bd}\tilde{x}_d)], \tag{9}$$

where $\tilde{x}_i = \overline{x}_i - \partial_i\psi$. We can easily calculate the moments of $\tilde{x}$:

$$\begin{aligned} & \mathrm{E}_{u_0}[\tilde{x}_i] = 0, \qquad \mathrm{E}_{u_0}[\tilde{x}_i\tilde{x}_j] = \frac{g_{ij}}{N}, \\ & \mathrm{E}_{u_0}[\tilde{x}_i\tilde{x}_j\tilde{x}_k] = \frac{1}{N^2} T_{ijk}, \qquad \mathrm{E}_{u_0}[\tilde{x}_i\tilde{x}_j\tilde{x}_k\tilde{x}_h] = \frac{1}{N^2} g_{ij}g_{kh}[3] + O(N^{-3}). \end{aligned} \tag{10}$$

By using geometrical properties of curved exponential families, it can be shown that

$$\tilde{u}^a = g^{ab}\tilde{x}_b - \tfrac{1}{2}\overset{m}{\Gamma}{}^{\alpha\beta a}\tilde{x}_\alpha\tilde{x}_\beta + \frac{1}{N}\overline{u}^a(\overline{x}) + o_p(N^{-1}). \tag{11}$$

Moreover, by (11) and (10),

$$\hat{u}^a_{bias} = -\frac{1}{2N}\overset{m}{\Gamma}{}_{bc}{}^a g^{bc} - \frac{1}{2N}\overset{m}{H}{}_{\kappa\lambda}{}^a g^{\kappa\lambda} + \frac{1}{N}\overline{u}^a + o(N^{-1}), \tag{12}$$

where $\overline{u} = \overline{u}(\mathrm{E}_{u_0}(\overline{x}))$. If we substitute (11) and (12) in (9), we can finally write

$$\begin{aligned}\mathrm{E}_{u_0}[\tilde{u}^a\tilde{u}^b] = {} & \frac{1}{N}g^{ab} - \frac{1}{2N^2}g^{cb}\partial_c(\overset{m}{\Gamma}{}_{de}{}^a g^{de})[2] - \frac{1}{2N^2}g^{cb}\partial_c(\overset{m}{H}{}_{\kappa\lambda}{}^a g^{\kappa\lambda})[2] + \frac{1}{N^2}g^{cb}\,\partial_c\overline{u}^a\,[2] \\ & + \frac{1}{4N^2}(\overset{m}{\Gamma}{}_{cd}{}^a g^{cd} + \overset{m}{H}{}_{\kappa\lambda}{}^a g^{\kappa\lambda})(\overset{m}{\Gamma}{}_{ef}{}^b g^{ef} + \overset{m}{H}{}_{\mu\nu}{}^b g^{\mu\nu}) + \frac{1}{2N^2}\overset{m}{\Gamma}{}_{cd}{}^a\overset{m}{\Gamma}{}_{ef}{}^b g^{ce}g^{df} \\ & + \frac{1}{2N^2}(\overset{m}{H}{}^{\kappa\lambda a}\overset{m}{H}{}^{\mu\nu b}g_{\kappa\mu}g_{\lambda\nu} + 2\overset{e}{H}{}^{ac\kappa}\overset{e}{H}{}^{bd\lambda}g_{cd}g_{\kappa\lambda}) \\ & + \frac{1}{N^2}\overline{u}^a\overline{u}^b - \frac{1}{2N^2}\overline{u}^a\overset{m}{\Gamma}{}_{cd}{}^b g^{cd}[2] - \frac{1}{2N^2}\overline{u}^a\overset{m}{H}{}_{\kappa\lambda}{}^b g^{\kappa\lambda}[2] + o(N^{-2}).\end{aligned} \tag{13}$$

By (4), we also have that

$$\mathrm{E}_{u_0}[\hat{s}^I\hat{s}^J] = \frac{1}{N^2}\overline{s}^I\overline{s}^J + o(N^{-2}), \tag{14}$$

where $\overline{s} = \overline{s}(\mathrm{E}_{u_0}(\overline{x}))$. By (11) and (10),

$$\mathrm{E}_{u_0}[\tilde{u}^a\tilde{u}^b\tilde{u}^c] = \frac{1}{N^2}(T^{abc} - \tfrac{1}{2}\overset{m}{\Gamma}{}^{\alpha\beta a}g^{bc}g_{\alpha\beta}[3] - \overset{m}{\Gamma}{}^{abc}[3] + g^{ab}\overline{u}^c[3]) + o(N^{-2}), \tag{15}$$

$$\mathrm{E}_{u_0}[\tilde{u}^a\tilde{u}^b\hat{s}^I] = \frac{1}{N^2}g^{ab}\overline{s}^I + o(N^{-2}) \tag{16}$$

and

$$\mathrm{E}_{u_0}[\tilde{u}^a\tilde{u}^b\tilde{u}^c\tilde{u}^d] = \frac{1}{N^2}g^{ab}g^{cd}[3] + o(N^{-2}). \tag{17}$$

We can now use (13)–(17) and (7) to calculate each term of (8). With some further calculations we obtain the result. □

From (6) we can obtain a decomposition of the average α divergence from the true distribution to any predictive one, in two parts:

$$\begin{aligned}\mathrm{E}_{u_0}\{D_\alpha(p(x;\,u_0),\,p(x;\,\hat{u}_N(\overline{x}),\,\hat{s}(\overline{x})))\} = {} & \mathrm{E}_{u_0}\{D_\alpha(p(x;\,u_0),\,p(x;\,\hat{u}_N(\overline{x})))\} \\ & + \frac{1}{2N^2}(g_{IJ}\overline{s}^I\overline{s}^J - \overset{\alpha}{H}{}_{abI}g^{ab}\overline{s}^I) + o(N^{-2}).\end{aligned} \tag{18}$$

The first term in (18) depends on the choice of the estimative distribution and the other on

the shift orthogonal to the model $\mathscr{P}$. It is well known that the problem of choosing a second-order efficient estimator $\hat{u}_N(\overline{x})$ has not, in general, a unique solution. On the other hand the following theorem solves the problem of the choice of the optimal shift orthogonal to the model.

Theorem 3.2. *The optimal choice of $\hat{s}^I(\overline{x})$, with respect to an α divergence, is given, up to order N^{-1}, by*

$$\hat{s}^I_{\text{opt}}(\overline{x}) = \frac{1}{2N} \overset{\alpha}{H}_{ab}{}^{I}(\hat{u}_N(\overline{x}))\, g^{ab}(\hat{u}_N(\overline{x})), \tag{19}$$

where $\hat{u}_N(\overline{x})$ is any asymptotically efficient estimator.

Proof. It is easy to see, by finding the derivative of (18) with respect to $\overline{s}$, that the minimum value of the asymptotic risk corresponds to

$$\overline{s}^I_{opt} = \tfrac{1}{2} \overset{\alpha}{H}_{ab}{}^{I}\, g^{ab}.$$

The result follows by (4). □

Let us now define, for $a,\ b = 1, \ldots, m$,

$$\begin{aligned} h_{ab} &= \overset{\alpha}{\nabla}{}^{(\mathscr{M})}_{\partial_a l_\alpha}\, \partial_b l_\alpha - \overset{\alpha}{\nabla}_{\partial_a l_\alpha}\, \partial_b l_\alpha \\ &= p^{(1-\alpha)/2} \left(\partial_a \partial_b l + \frac{1-\alpha}{2}\, \partial_a l\, \partial_b l + \frac{1+\alpha}{2}\, g_{ab} - \overset{\alpha}{\Gamma}_{ab}{}^{c}\, \partial_c l \right). \end{aligned} \tag{20}$$

Vectors h_{ab} are, by definition orthogonal to the original model $\mathscr{P}$. Moreover they belong to H_u. The following theorem explains the important role that they play in our analysis.

Theorem 3.3. *The difference in average α divergence from the true distribution, between the estimative distribution $p(x; \hat{u}_N(\overline{x}))$ and the optimal predictive distribution $p(x; \hat{u}_N(\overline{x}), \hat{s}_{\text{opt}}(\overline{x}))$, is maximal if and only if the vector $g^{ab}h_{ab}$ belongs to the linear space spanned by the h_I. In this case, the optimal predictive distribution is*

$$\begin{aligned} p(x;\, \hat{u}_N, \hat{s}_{\text{opt}}) = {} & p(x;\, \hat{u}_N) \left[1 + \frac{1}{2N}\, g^{ab} \left(\partial_a \partial_b l + \frac{1-\alpha}{2}\, \partial_a l\, \partial_b l + \frac{1+\alpha}{2}\, g_{ab} - \overset{\alpha}{\Gamma}_{ab}{}^{c}\, \partial_c l \right) \right] \\ & + o_p(N^{-1}). \end{aligned} \tag{21}$$

Proof. By (20) and the definition of $\overset{\alpha}{H}_{abI}$, we have that

$$\langle h_{ab},\, h_I \rangle = \langle \partial_a \partial_b l_\alpha + \frac{1+\alpha}{2}\, p^{(1-\alpha)/2} g_{ab} - \overset{\alpha}{\Gamma}_{ab}{}^{c}\, \partial_c l_\alpha,\, h_I \rangle = \overset{\alpha}{H}_{abI}. \tag{22}$$

By substituting (19) in (18),

$$\begin{aligned}
&\mathrm{E}_{u_0}\{D_\alpha(p(x;\,u_0),\,p(x;\,\hat{u}_N(\overline{x})))\} - \mathrm{E}_{u_0}\{D_\alpha(p(x;\,u_0),\,p(x;\,\hat{u}_N(\overline{x}),\,\hat{s}_{\mathrm{opt}}(\overline{x})))\}\\
&\quad = \frac{1}{8N^2}\,\|\overset{\alpha}{H}_{abI}\,g^{ab}\,g^{IJ}\,\partial_J\,l_\alpha\|^2 + o(N^{-2})\\
&\quad = \frac{1}{8N^2}\,\|\langle g^{ab}\,h_{ab},\,h_I\rangle\,g^{IJ}\,h_J\|^2 + o(N^{-2}),
\end{aligned}$$

which depends only on the projection of $g^{ab}h_{ab}$ on the linear space spanned by the h_I. Thus, it is maximal if and only if $g^{ab}h_{ab}$ is included in this space and its maximal value is

$$\frac{1}{8N^2}\,\|g^{ab}h_{ab}\|^2 + o(N^{-2}). \tag{23}$$

In this situation, by (19), (22) and (20), we have that

$$\begin{aligned}
\hat{s}^I_{\mathrm{opt}}\,h_I &= \frac{1}{2N}\,\overset{\alpha}{H}_{ab}{}^{I}\,g^{ab}h_I\\
&= \frac{1}{2N}\,g^{ab}\langle h_{ab},\,h_I\rangle\,g^{IJ}h_J\\
&= \frac{1}{2N}\,g^{ab}h_{ab}\\
&= \frac{1}{2N}\,p^{(1-\alpha)/2}\,g^{ab}\left(\partial_a\partial_b l + \frac{1-\alpha}{2}\,\partial_a l\,\partial_b l + \frac{1+\alpha}{2}\,g_{ab} - \overset{\alpha}{\Gamma}_{ab}{}^{c}\,\partial_c l\right),
\end{aligned} \tag{24}$$

and the result follows by substituting (24) in (3). □

Remark. Including the vector $g^{ab}h_{ab}$ on the enlarged model allows us to attain the best improvement on the estimative distribution. For any regular parametric model $\mathscr{M}$ containing $\mathscr{P}$ and $g^{ab}h_{ab}$ we obtain the same optimal predictive distribution. In this sense, (21) gives a predictive distribution that can be considered optimal among all probability distributions equivalent to p.

In the case when $\mathscr{P}$ itself is a full exponential family, we can write (21) in a simpler form

$$\begin{aligned}
p(x;\,\hat{u}_N,\,\hat{s}_{\mathrm{opt}}) &= p(x;\,\hat{u}_N)\left(1 + \frac{1-\alpha}{4N}\,g^{ab}(\partial_a l\,\partial_b l - g_{ab} - T_{ab}{}^{c}\,\partial_c l\right) + o_p(N^{-1})\\
&= p(x;\,\hat{u}_N)\left(1 + \frac{1-\alpha}{4N}\{g^{ab}(x_a - \partial_a\psi)(x_b - \partial_b\psi) - m - g^{ab}T_{ab}{}^{c}(x_c - \partial_c\psi)\}\right)\\
&\quad + o_p(N^{-1}).
\end{aligned} \tag{25}$$

Note that for $\alpha = 1$ there is no correction, i.e. we do not move out of the full exponential model. Moreover, for $\alpha = -1$ we obtain exactly the same result as Vidoni (1995, p. 858, equation (3.1)).

Example 3.1. We consider m-dimensional multivariate distributions $N(\mu, I_m)$:

$$p(x;\mu)=\prod_{i=1}^{m}\frac{1}{(2\pi)^{1/2}}\exp\{-\tfrac{1}{2}(x^i-\mu^i)^2\},$$

where $\mu=(\mu^i)$, $i=1,\ldots,m$, is unknown. We have that

$$g_{ij}(\mu)=\delta_{ij}\quad\text{and}\quad \overset{\alpha}{\Gamma}_{ijk}(\mu)=0,$$

for all α. Now let $x(l)$, $l=1,\ldots,N$, be independent of $N(\mu, l_m)$ and $\hat{\mu}=\hat{\mu}_N(\overline{x})$ be any estimator for the mean vector μ, where

$$\overline{x}=\frac{1}{N}\sum_{l=1}^{N}x(l).$$

By (20),

$$h_{ij}=\begin{cases}\dfrac{1-\alpha}{2}\,p^{(1-\alpha)/2}\{(x^i-\mu^i)^2-1\}, & i=j,\\[2ex] \dfrac{1-\alpha}{2}\,p^{(1-\alpha)/2}(x^i-\mu^i)(x^j-\mu^j), & i\neq j.\end{cases}$$

By (25),

$$\begin{aligned}p(x;\hat{\mu},\hat{s}_{\text{opt}})&=p(x;\hat{\mu})\left(1+\frac{1-\alpha}{4N}\sum_{i=1}^{m}\{(x^i-\hat{\mu}^i)^2-1\}\right)+o_p(N^{-1})\\&=\frac{1}{(2\pi)^{1/2}}\left(1-\frac{1-\alpha}{2N}\right)^{1/2}\exp\left\{-\frac{1}{2}\left(1-\frac{1-\alpha}{2N}\right)\sum_{i=1}^{m}(x^i-\hat{\mu}^i)^2\right\}+o_p(N^{-1}).\end{aligned}$$

We thus have that the optimal predictive distribution can be written in a close form as

$$N\left(\hat{\mu},\left(1-\frac{1-\alpha}{2N}\right)^{-1}I_m\right).$$

For $\alpha=-1$, it coincides, up to order N^{-1}, with the result of Barndorff-Nielsen and Cox (1994, p. 318). By (23), we can calculate the difference in average α divergence between the estimative distribution and the predictive distribution:

$$\begin{aligned}\frac{1}{8N^2}\|g^{ij}h_{ij}\|^2&=\frac{(1-\alpha)^2}{32N^2}\left\|p^{(1-\alpha)/2}\sum_{i=1}^{m}\{(x^i-\hat{\mu}^i)^2-1\}\right\|^2\\&=\frac{(1-\alpha)^2}{32N^2}\int\left(\sum_{i=1}^{m}\{(x^i-\hat{\mu}^i)^2-1\}\right)^2 p\,\mathrm{d}x\\&=\frac{(1-\alpha)^2}{16N^2}\,m,\end{aligned}$$

which does not depend on $\hat{\mu}$, the efficient estimator used. Now let $\hat{\mu}$ be the James–Stein estimator for μ, i.e.

$$\hat{\mu}(\overline{x}) = \left(1 - (m-2)\Big/ N\sum_{i=1}^{m}(\overline{x}^i)^2\right)\overline{x}.$$

Then

$$\hat{\mu}_\infty(\overline{x}) = \lim_{N\to\infty}\hat{\mu} = \overline{x},$$

$$\overline{\mu}(\overline{x}) = -\left((m-2)\Big/\sum_{i=1}^{m}(\overline{x}^i)^2\right)\overline{x}$$

and

$$\overline{\mu} = \overline{\mu}(\mu) = -\left((m-2)\Big/\sum_{i=1}^{m}(\mu^i)^2\right)\mu.$$

We can use (6) with $s = 0$ to compare the two estimative distributions obtained respectively from the maximum-likelihood estimator $\hat{\mu}_{\text{mle}} = \overline{x}$, and the James–Stein estimator:

$$\begin{aligned}\mathrm{E}_\mu\{D_\alpha(p(x;\mu),\ p(x;\hat{\mu}_{\text{mle}}))\} - \mathrm{E}_\mu\{D_\alpha(p(x;\mu), p(x;\hat{\mu}))\}&\\ = -\frac{1}{2N^2}\,g_{ij}\overline{\mu}^i\overline{\mu}^j - \frac{1}{N^2}\,\partial_i\overline{\mu}^i + o(N^{-2})&\\ = \frac{1}{2N^2}(m-2)^2\Big/\sum_{i=1}^{m}(\mu^i)^2 + o(N^{-2}).&\end{aligned}$$

Remark. Let us consider an f divergence D_f as a loss function. Without loss of generality, we can suppose that $f''(1) = 1$. Theorem 3.1 can be easily generalized to this case by putting $\alpha = 2f'''(1) + 3$ and by substituting the coefficient

$$\frac{(\alpha - 11)(\alpha - 1)}{32}$$

of the term

$$\frac{Q_{abcd}g^{ab}g^{cd}}{N^2}$$

by

$$\beta = \frac{f^{(4)}(1) - 2f'''(1) - 4}{8}.$$

In fact, in the expansion of D_f, the first- and second-order terms remain unchanged. The coefficient of the third-order term is

$$\frac{(f'''(1)+3)}{6} T_{ABC} + \tfrac{1}{2} \overset{e}{\Gamma}_{ABC},$$

and it can be written as

$$\frac{\alpha}{3} T_{ABC} + \tfrac{1}{2} \overset{\alpha}{\Gamma}_{ABC}$$

with $\alpha = 2f'''(1) + 3$. The coefficient β is calculated by

$$\frac{f^{(4)}(1)}{8} - \frac{\alpha+1}{8} = \frac{f^{(4)}(1) - 2f'''(1) - 4}{8}.$$

Acknowledgement

This work was partially supported by European Union Human Capital and Mobility Program under Contract ERB CHRX CT94 0499.

References

Amari, S. (1985) *Differential–Geometrical Methods in Statistics. Lecture Notes in Statist*. 28 New York: Springer-Verlag.

Amari, S., Barndorff-Nielsen, O.E., Kass, R.E., Lauritzen, S.L. and Rao, C.R. (1987) *Differential Geometry in Statistical Inference*, Chapter 2, pp. 19–94. Lecture Notes—Monograph Series 10. Hayward, CA: Institute of Mathematical Statistics.

Barndorff-Neilsen, O.E. and Cox, D.R. (1994) *Inference and Asymptotics*. London: Chapman & Hall.

Corcuera, J.M. and Giummolè, F. (1996) On the relationship between α-connections and the asymptotic properties of predictive distributions. Mathematics Preprint Series 210, University of Barcelona.

Komaki, F. (1996) On asymptotic properties of predictive distributions. *Biometrika*, **83**, 299–313.

Vidoni, P. (1995) A simple predictive density based on the p^*-formula. *Biometrika*, **82**, 855–63.

Received February 1996 and revised May 1997