Bernoulli **5**(4), 1999, 721–760

The exponential statistical manifold: mean parameters, orthogonality and space transformations

GIOVANNI PISTONE[1] and MARIA PIERA ROGANTIN[2]

[1]*Department of Mathematics, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy. E-mail: pistone@polito.it*
[2]*Department of Mathematics, University of Genova, Via Dodecaneso 35, 16146 Genova, Italy. E-mail: rogantin@dima.unige.it*

Let $(X, \mathscr{X}, \mu)$ be a measure space, and let $\mathscr{M}(X, \mathscr{X}, \mu)$ denote the set of the μ-almost surely strictly positive probability densities. It was shown by Pistone and Sempi in 1995 that the global geometry on $\mathscr{M}(X, \mathscr{X}, \mu)$ can be realized by an affine atlas whose charts are defined locally by the mappings $\mathscr{M}(X, \mathscr{X}, \mu) \supset \mathscr{U}_p \ni q \mapsto \log(q/p) + K(p, q) \in B_p$, where $\mathscr{U}_p$ is a suitable open set containing p, $K(p, q)$ is the Kullback–Leibler relative information and B_p is the vector space of centred and exponentially $(p \cdot \mu)$-integrable random variables. In the present paper we study the transformation of such an atlas and the related manifold structure under basic transformations, i.e. measurable transformation of the sample space. A generalization of the mixed parametrization method for exponential models is also presented.

Keywords: exponential families; exponential statistical manifolds; information; mean parameters; Orlicz spaces; orthogonality

1. Introduction

The present paper is devoted to mathematical developments connected to the so-called theory of statistical manifolds. As general references on the subject in book form we mention Amari (1985), Amari *et al*. (1987), Murray and Rice (1993) and Barndorff-Nielsen and Cox (1994). Other relevant references for the present paper are Rao (1945, 1949), Jeffreys (1946), Dawid (1975, 1977), Efron (1975, 1978), Madsen (1979), Amari (1982), Barndorff-Nielsen and Jupp (1989) and Kass (1989).

In the literature on statistical manifolds, the question of finding a suitable functional setting to a nonparametric extension of the geometric construction has been mentioned by many workers (see, for example, Dawid (1975, 1977), Amari (1982) and Murray and Rice (1993)). In those papers and books some fundamental ideas of the nonparametric theory

1350–7265 © 1999 ISI/BS

have been sketched but, as far as we know, no detailed formal construction has been published before that of Pistone and Sempi (1995).

Starting with unpublished seminars held in Lecce University by Pistone and Sempi in 1989, the following idea has been developed. The statistical object that induces the geometry is the exponential model with its particular form of the Fisher information (Efron 1975); so the starting point has to be a nonparametric definition of the exponential model. This definition in turn is related to the class of *exponentially integrable* random variables whose natural topology is given by the notion of Orlicz space for the exponential function. The present paper is dedicated to further developments of these ideas. Some of the results developed here in full detail were announced by Pistone and Rogantin (1994).

The content of the paper is as follows.

Sections 2 and 3 are mainly devoted to a new presentation of the previous results of Pistone and Rogantin (1990), Pistone and Rogantin (1994) and Pistone and Sempi (1995), where the proofs of the basic propositions have been given in detail. Some of those proofs rely on quite straightforward arguments from functional analysis, but nevertheless the programme sketched above is systematically developed in these references. A few new results are added here and also the presentation has been improved; we give a definition of tangent space, show the relevance of the Kullback information and recall the notion of submanifold.

The main results of the present paper are given in Sections 6 and 7, where we give a nonparametric version of the concept of *mixed parametrization* in exponential models, and finally the effect of space transformation on the exponential manifold. A paper by Rogantin (1996) has given further developments in the case of finite sample spaces, discussing the derivation of parameters' orthogonality for various classes of finite-state-space stochastic processes.

2. The exponential manifold

The definition of statistical manifold can be given in the framework of the theory of manifolds modelled on Banach spaces, as introduced for example by Lang (1995).

We consider a measure space $(X, \mathscr{X}, \mu)$, where μ is a reference measure, and the set $\mathscr{M}(X, \mathscr{X}, \mu)$ of the μ-almost surely strictly positive probability densities. We shall define on the set $\mathscr{M}(X, \mathscr{X}, \mu)$ a topology such that $\mathscr{M}(X, \mathscr{X}, \mu)$ is an Hausdorff space (i.e. points can be separated by open sets). Then we shall construct a covering of $\mathscr{M}(X, \mathscr{X}, \mu)$ with open sets $\mathscr{U}_p$, $p \in \mathscr{U}_p$, $p \in \mathscr{M}(X, \mathscr{X}, \mu)$, and a corresponding family of Banach spaces B_p, with norms $\|\cdot\|_p$, $p \in \mathscr{M}(X, \mathscr{X}, \mu)$, such that each density $q \in \mathscr{U}_p$ is represented with respect to p by a coordinate $s_p(q) \in B_p$.

We shall use the notation

$$s_p: \mathscr{U}_p \to \mathscr{V}_p \subset B_p, \tag{1}$$

$$e_p: \mathscr{V}_p \to \mathscr{U}_p \subset \mathscr{M}(X, \mathscr{X}, \mu), \tag{2}$$

to denote respectively the *charts*, i.e. the mappings from points to coordinates, and the *patches*, i.e. the mappings from coordinates to points (Figure 1).

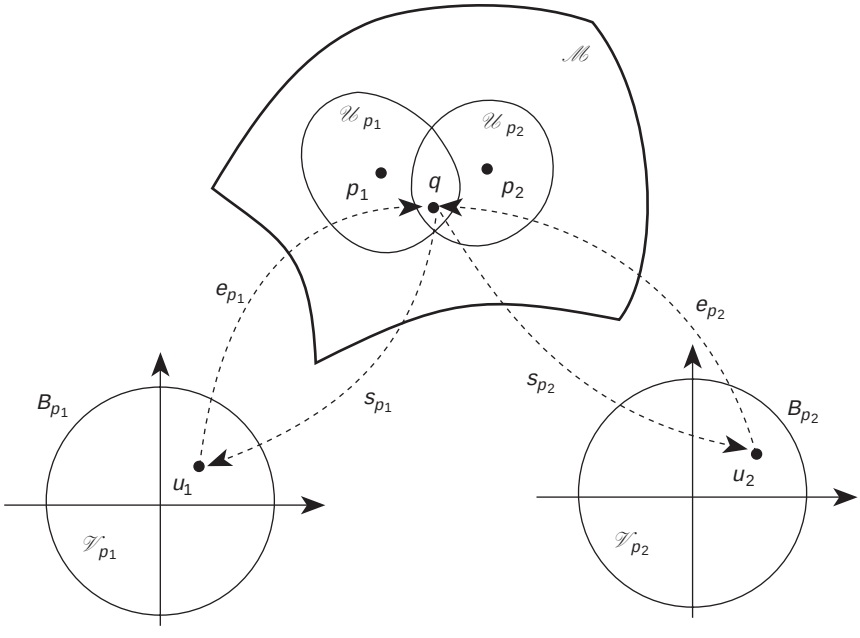

Figure 1. The charts and the patches of the atlas.

The idea of a chart on $U_p \subset \mathscr{M}(X, \mathscr{X}, \mu)$ is an abstraction of the parametrization for a family of probability densities ("statistical model"). Below we shall see that the condition of being a chart is actually more stringent than the condition of being a regular parametrization. Our model is nonparametric; so the coordinate mapping s_p cannot take values in a finite-dimensional vector space unless the sample space has a finite number of atoms.

As in differential geometry, we say that $\{(\mathscr{U}_p, s_p)\colon p \in \mathscr{M}(X, \mathscr{X}, \mu)\}$ is an *atlas* if all the space is covered by its charts. If each of the *change in coordinates* (see Figure 1)

$$s_{p_2} \circ e_{p_1}\colon s_{p_1}(\mathscr{U}_{p_1} \cap \mathscr{U}_{p_2}) \to s_{p_2}(\mathscr{U}_{p_1} \cap \mathscr{U}_{p_2})$$

is a diffeomorphism of some regularity, the atlas is said to have that regularity. In such a case the atlas, possibly augmented by all the compatible charts, defines the manifold. Here a new chart is said to be compatible if all the corresponding changes of coordinate are regular (Lang 1995).

In our case we shall introduce a very special manifold, such that the change of coordinates are actually affine functions (i.e. they differ from a linear function by a additive constant), but we shall keep a weaker regularity, namely the C^∞ regularity (differentiability of any order) for compatible charts.

We shall denote by $\mathrm{E}_{p\cdot\mu}[\cdot]$ the expectation with respect to the probability measure $p \cdot \mu$ (where $(p \cdot \mu)\,\mathrm{d}x = p(x)\mu\,\mathrm{d}x$); if there is no ambiguity we shall use the notation $\mathrm{E}_p[\cdot]$.

2.1. Functional framework

First we define the topology on $\mathscr{M}(X, \mathscr{X}, \mu)$ as follows. For simplicity we give only the definition of convergence of sequences.

Definition 1 (Exponential convergence). *The sequence* $(p_n)_{n\in\mathbb{N}}$ *in* $\mathscr{M}(X, \mathscr{X}, \mu)$ *is* e convergent *(exponentially convergent) to* p *if* $(p_n)_{n\in\mathbb{N}}$ *tends to* p *in* μ *probability as* $n \to \infty$ *and moreover the sequences* $(p_n/p)_{n\in\mathbb{N}}$ *and* $(p/p_n)_{n\in\mathbb{N}}$ *are eventually bounded in each* $L^\alpha(p)$, $\alpha > 1$, *i.e.*

$$\forall \alpha > 1 \quad \limsup_{n\to\infty} \mathrm{E}_p\left[\left(\frac{p_n}{p}\right)^\alpha\right] < +\infty, \quad \limsup_{n\to\infty} \mathrm{E}_p\left[\left(\frac{p}{p_n}\right)^\alpha\right] < +\infty.$$

Some properties of the topology associated with this notion of convergence have been given by Pistone and Sempi (1995).

Now we shall introduce the Banach spaces on which the statistical manifold is modelled. We give a definition that shows how they are connected with well-known statistical objects (Barndorff-Nielsen 1978a; Letac 1992).

Definition 2 (Cramér class). *For each density* $p \in \mathscr{M}(X, \mathscr{X}, \mu)$, *the* Cramér class at p *is the set of all random variables* u *on* $(X, \mathscr{X}, \mu)$ *such that the* moment generating function *of* u *with respect to the probability measure* $p \cdot \mu$ *given by*

$$\hat{u}_p(t) = \int \mathrm{e}^{tu}\, p \,\mathrm{d}\mu = \mathrm{E}_p[\mathrm{e}^{tu}], \qquad t \in \mathbb{R},$$

is finite in a neighbourhood of the origin 0.

If moreoever the expectation of u *is zero (the previous condition implies the existence of a finite expectation), then we shall call the set the* centred Cramér class at p.

Note that, if $p \in \mathscr{M}(X, \mathscr{X}, \mu)$ and $p(\theta)$ is a one-dimensional exponential model,

$$p(\theta) = \mathrm{e}^{\theta u - \psi(\theta)}\, p, \quad \theta \in I \text{ open real interval}, \quad 0 \in I,$$

then the sufficient statistic u belongs to the Cramér class at each density $p(\theta)$ in the model. In fact, for $t \in I$,

$$\hat{u}_{p(\theta)}(t) = \mathrm{E}_{p(\theta)}[\mathrm{e}^{tu}] = \mathrm{E}_p[\mathrm{e}^{tu}\,\mathrm{e}^{\theta u - \psi(\theta)}] = \frac{\mathrm{E}_p[\mathrm{e}^{(t+\theta)u}]}{\mathrm{e}^{\psi(\theta)}} = \frac{\mathrm{e}^{\psi(t+\theta)}}{\mathrm{e}^{\psi(\theta)}}$$

and $\mathrm{E}_{p(\theta)}[\mathrm{e}^{tu}]$ is finite for t in a neighbourhood of 0.

The following construction was presented by Pistone and Sempi (1995); we repeat it here for ease of presentation.

Proposition 3 (A norm on the Cramér class). *The Cramér class at* p *of Definition 2 is a vector space and a Banach space with the norm defined by*

$$\|u\|_p = \inf\left\{ r > 0 \colon \mathrm{E}_p\left[\cosh\left(\frac{u}{r}\right) - 1\right] \leqslant 1 \right\}. \tag{3}$$

The centred Cramér class, denoted by B_p and given by

$$B_p = \{u \in L^1(p \cdot \mu)\colon 0 \in \mathrm{dom}(\hat{u}_p)^\circ,\ \mathrm{E}_p[u] = 0\},$$

is a closed subspace.

Proof. It is clear that $\mathrm{E}_p[\mathrm{e}^{tu}] < +\infty$ in a neighbourhood of 0 if and only if $\mathrm{E}_p[\mathrm{e}^{u/r}] < +\infty$ and $\mathrm{E}_p[\mathrm{e}^{-u/r}] < +\infty$, i.e. $\mathrm{E}_p[\cosh(u/r)] < +\infty$. Moreover $\mathrm{E}_p[\cosh(|u|/r)] \to 1$ if $r \to \infty$. Consequently the Cramér class at p is defined by $\|u\|_p < +\infty$.

The Banach space property follows from general arguments on Orlicz spaces (Rao and Ren 1991; Krasnosel'skii and Rutickii 1961). □

Note that $\|u\|_p < 1$ if and only if there exists a real $\alpha > 1$ such that

$$\mathrm{E}_p[\cosh(\alpha u) - 1] \leqslant 1, \qquad \text{i.e. } \mathrm{E}_p[\cosh(\alpha u)] \leqslant 2.$$

Note also that, if u_n is a sequence of random variables, then $\|u_n - u\|_p \to u$ if and only if $\forall \epsilon > 0$, $\mathrm{E}_p[\cosh\{(u_n - u)/\epsilon\}] < 2$ eventually, $n \to \infty$.

Example 4. Figure 2 illustrates the construction of the norm $\|.\|_p$ when $X = \{x_1, x_2\}$ and μ is the counting measure. In such a case the space of the random variables u is $\mathbb{R}^2$. We consider the convex set $\overline{\mathscr{V}}_p$ of all the vectors $u = (u_1, u_2) \in \mathbb{R}^2$ such that

$$\overline{\mathscr{V}}_p = \{(u_1, u_2)\colon \{\cosh(u_1) - 1\}\theta + \{\cosh(u_2) - 1\}(1 - \theta) \leqslant 1\},$$

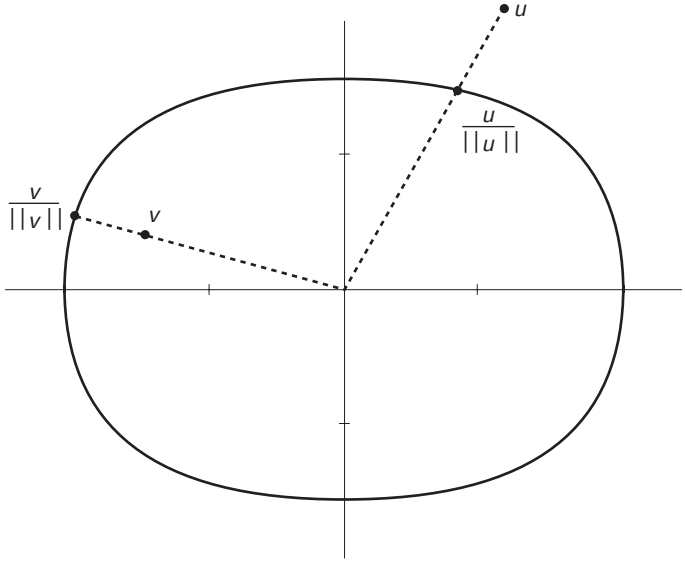

Figure 2. Construction of the norm $\|.\|_p$.

where $\theta = p(x_1)$ and $1 - \theta = p(x_2)$ (in Figure 2 we have assumed that $\theta = \frac{1}{3}$). For any vector $u \in \mathbb{R}^2$ this norm is the unique positive value r such that

$$\left\{\cosh\left(\frac{u_1}{r}\right) - 1\right\}\theta + \left\{\cosh\left(\frac{u_2}{r}\right) - 1\right\}(1 - \theta) = 1,$$

that is $(u_1/r, u_2/r)$ lies on the boundary of $\overline{\mathscr{V}}_p$.

This example shows the construction of the norm defined in (3) in a finite sample space, where actually all norms define the same topology because the vector space of random variables has finite dimension. The construction that we present is really needed only when the underlying space of random variables is infinite dimensional.

In the previous proposition the function $x \mapsto \cosh(x) - 1$ is a convex function that plays in the theory of the spaces B_p the same role as the function $x \mapsto |x|^\alpha/\alpha$ in the theory of Lebesgue spaces L^α, $\alpha > 1$. We cite Krasnosel'skii and Rutickii (1961) and Rao and Ren (1991) as general references.

We shall use various types of such convex functions:

$$\phi_1 : x \mapsto \cosh(|x|) - 1, \tag{4}$$

$$\phi_2 : x \mapsto \exp(|x|) - |x| - 1, \tag{5}$$

$$\phi_3 : x \mapsto (1 + |x|)\log(1 + |x|) - |x|. \tag{6}$$

For each of these functions it is possible to define a norm as in (3); we shall denote, for $i = 1, 2, 3$, the following.

(a) $\overline{\mathscr{V}}_{\phi_i, p}$ the convex set $\{u \in L^1(p \cdot \mu)\colon \mathrm{E}_p[\phi_i(u)] \leqslant 1\}$;
(b) $\|.\|_{\phi_i, p}$ the norm associated with ϕ_i:

$$\|u\|_{\phi_i, p} = \inf\left\{r > 0\colon \mathrm{E}_p\left[\phi_i\left(\frac{u}{r}\right)\right] \leqslant 1\right\};$$

(c) $L^{\phi_i}(p \cdot \mu)$ (or $L^{\phi_i}(p)$ if there is no ambiguity) the corresponding Banach spaces of non-centred random variables:

$$L^{\phi_i}(p) = \{u\colon \exists \alpha > 0 \text{ such that } \mathrm{E}_p[\phi_i(\alpha u)] < +\infty\} = \{u\colon \|u\|_{\phi_i, p} < +\infty\};$$

(d) $L_0^{\phi_i}(p)$ the corresponding space of centred random variables.

The classes of centred random variables are closed subspaces. The function ϕ_1 is the most important for us; above, we have denoted $\|.\|_{\phi_1, p}$ by $\|.\|_p$ and $L_0^{\phi_1}$ by B_p. We say that two ϕ functions are equivalent if the corresponding norms are equivalent, i.e. there exist two real constants a and b such that $a\|u\|_{\phi_i} \leqslant \|u\|_{\phi_j} \leqslant b\|u\|_{\phi_i}$.

As the ϕ functions in (4)–(6) are strictly convex and differentiable, it is possible to consider the inverse function $[\phi']^{-1}(x)$ and to define the *conjugate function* of ϕ as the function ψ, such that, for any $y = \phi'(x)$, $\psi'(y) = [\phi']^{-1}(x)$. This implies that $xy \leqslant \phi(x) + \psi(y)$, $x, y \in \mathbb{R}_+$, with equality if and only if $y = \phi'(x)$. If ϕ and ψ are conjugate, then $u \mapsto \sup_v\{\mathrm{E}_p[uv]\colon \mathrm{E}_p[\psi(v)] \leqslant 1\}$ is a norm equivalent to $\|.\|_{\phi, p}$ (Rao and Ren 1991, p. 61).

It can be shown that ϕ_1 and ϕ_2 are equivalent (and the corresponding Banach spaces coincide with the Cramer class at p) and that ϕ_2 and ϕ_3 are conjugate.

Conjugacy implies that the Banach spaces are in a duality relation almost in the same way the Lebesgue spaces L^α and L^β are in duality if $\alpha^{-1} + \beta^{-1} = 1$. Precisely the bilinear form

$$L^\phi(p) \times L^\psi(p) \ni (u, v) \mapsto \mathrm{E}_p[uv] \in \mathbb{R}$$

is continuous, but in general $L^\phi(p)$ and $L^\psi(p)$ are not dual.

The following proposition follows immediately from the main result of Pistone and Sempi (1995). Because of its importance, we give here a new direct proof.

Proposition 5. *Let p and q be two probability densities in $\mathscr{M}(X, \mathscr{X}, \mu)$ connected by a one-dimensional exponential model. Then*

$$L^{\phi_1}(p) = L^{\phi_1}(q).$$

Proof. Let $r \in \mathscr{M}(X, \mathscr{X}, \mu)$ be given and let $u \in L^{\phi_1}(r)$. Let $p(t) = \mathrm{e}^{tu - \psi(t)}\, r$ be the one-dimensional exponential model associated with r and u, where t belongs to the real interval I with $0 \in I$. Let $p(t_0) = p$ and $p(t_1) = q$ be two densities in the given model. We can assume that $t_0 < t_1$ because otherwise we change u with $-u$.

It is enough to show that $L^{\phi_1}(p) \subset L^{\phi_1}(q)$ because the relation that connects p and q is symmetric.

Let $w \in L^{\phi_1}(p)$ be given; we have to prove that there exists a $\beta > 0$ such that

$$\mathrm{E}_q[\cosh(\beta w)] < +\infty.$$

We have

$$\mathrm{E}_q[\cosh(\beta w)] = \mathrm{E}_r[\cosh(\beta w)\, \mathrm{e}^{t_1 u - \psi(t_1)}]$$

$$= \frac{\mathrm{E}_r[\cosh(\beta w)\, \mathrm{e}^{t_1 u}]}{\mathrm{E}_r[\mathrm{e}^{t_1 u}]}.$$

The previous equation involves the convex function

$$g\colon (\theta, t) \mapsto \mathrm{E}_r[\cosh(\theta w)\, \mathrm{e}^{tu}].$$

We study the domain in which g is finite. If $\theta = 0$ the value of the function is $g(0, t) = \mathrm{E}_r[\mathrm{e}^{tu}]$ which is finite for $t \in I$. If $t = t_0$ the value is

$$g(\theta, t_0) = \mathrm{E}_r[\cosh(\theta w)\, \mathrm{e}^{t_0 u}] = \mathrm{E}_p[\cosh(\theta w)]\mathrm{E}_r[\mathrm{e}^{t_0 u}]$$

and is finite for θ in some interval $]-\overline{\theta}, \overline{\theta}[$, because we have assumed that $w \in L^{\phi_1}(p)$. We chose $b > t_1$, such that $b \in I$ and $\gamma \in]0, \overline{\theta}[$. The function $t \mapsto g(0, t)$ is finite in the interval $[t_0, b]$ and the function $\theta \mapsto g(\theta, t_0)$ is finite in the interval $[-\gamma, \gamma]$. Then, because of the convexity, $(\theta, t) \mapsto g(\theta, t)$ is finite in the triangle with vertices at the points (t_0, γ), $(b, 0)$ and $(t_0, -\gamma)$.

We consider the straight line between the points (t_0, γ) and $(b, 0)$ and we denote by β the value of θ at the intersection point of the previous straight line and the straight line $\theta = t_1$, namely $\beta/\gamma = (b - t_1)/(b - t_0)$. It follows that

$$\mathrm{E}_q[\cosh(\beta w)]\,\mathrm{E}_r[\mathrm{e}^{t_1 u}] \leqslant \frac{t_1 - t_0}{b - t_0}\,\mathrm{E}_p[\cosh(\gamma w)]\,\mathrm{E}_r[\mathrm{e}^{t_0 u}] + \frac{b - t_1}{b - t_0}\,\mathrm{E}_r[\mathrm{e}^{t_0 u}],$$

$$\mathrm{E}_q[\cosh(\beta w)] \leqslant \frac{t_1 - t_0}{b - t_0}\,\mathrm{E}_p[\cosh(\gamma w)]\,\frac{\mathrm{E}_r[\mathrm{e}^{t_0 u}]}{\mathrm{E}_r[\mathrm{e}^{t_1 u}]} + \frac{b - t_1}{b - t_0}\frac{\mathrm{E}_r[\mathrm{e}^{t_0 u}]}{[\mathrm{e}^{t_1 u}]},$$

so that $\mathrm{E}_q[\cosh(\beta w)] < +\infty$. □

Note that the equality between the spaces $L^{\phi_1}(p)$ and $L^{\phi_1}(q)$ holds true for each pair of points p, q which are connected by a one-dimensional exponential model. This establishes an equivalence relation, as we show below.

If $p(t)$, $t \in I$, is a one-dimensional exponential model connecting p and q, and $r(t)$, $t \in J$, is a one-dimensional exponential model connecting q and r we can assume those models to be of the form

$$p(t) = \mathrm{e}^{tu - \psi_1(t)}\, q,$$

$$r(t) = \mathrm{e}^{tv - \psi_2(t)}\, q,$$

with $u, v \in L^{\phi_1}(q)$. Then by convexity all the densities in the two-dimensional model

$$q(t_1, t_2) = \mathrm{e}^{t_1 u + t_2 v - \phi_3(t_1, t_2)}\, q$$

are connected by a one-dimensional model; in particular p and r are connected.

Definition 6 (x log x class). *We shall denote by ${}^*B_p$ the Banach space of centred random variables in $L^{\phi_3}(p \cdot \mu)$, i.e. the centred random variable of the so-called $x \log x$ class.*

Proposition 7. *A $(p \cdot \mu)$ integrable random variable u belongs to the $x \log x$ class ${}^*B_p$ if and only if it is centred and $(1 + |u|)\log(1 + |u|)$ is $(p \cdot \mu)$ integrable.*

Proof. *If $u \in L^{\phi_3}(p)$*, there exists a constant α such that $\mathrm{E}_p[\phi_3(\alpha u)] < +\infty$. This implies that $\mathrm{E}_p[\phi_3(u)] < +\infty$; in fact there exists a constant k such that, for any n, $\phi_3(2^n u) < k^n\, \phi_3(u)$ (Rao and Ren 1991, p. 22). For any α there exists an n such that $1/\alpha \leqslant 2^n$ and, because ϕ_3 is even and increasing in the positive real values, $\phi_3(u) \leqslant \phi_3(2^n \alpha u) \leqslant k^n\, \phi_3(\alpha u)$. So $\mathrm{E}_p[\phi_3(\alpha u)] < +\infty$ implies that $\mathrm{E}_p[\phi_3(u)] < +\infty$. Finally both $u \in L^{\phi_3}(p)$ and the integrability of $(1 + |u|)\log(1 + |u|)$ imply that $u \in L^1(p)$, and the conclusion follows from $\mathrm{E}_p[u] = 0$. □

Now we give some details about the Banach spaces B_p and ${}^*B_p$ which will be useful in the construction of the statistical manifold.

Proposition 8.

(a) *All the elements* ${}^*u$ *in* ${}^*B_p$ *are identified with an element* u^* *of the dual space* B_p^* *of* B_p *by the formula*: $u^*(u) = \mathrm{E}_p[{}^*uu]$, *with* $u \in B_p$. *In general,* ${}^*B_p$ *is identified with a proper subset of* B_p^*. *The injection of* ${}^*B_p$ *into* B_p^* *is continuous; we write*

$$ {}^*B_p \sqsubseteq B_p^*. $$

(b) *All the elements* u *in* B_p *are identified with an element* $\overline{u}$ *of the dual space* $({}^*B_p)^*$ *of* ${}^*B_p$ *by the formula* $\overline{u}({}^*u) = \mathrm{E}_p[u\ {}^*u]$, *with* ${}^*u \in {}^*B_p$. *This identification is onto, i.e.* B_p *is identified with* $({}^*B_p)^*$; *we write*

$$ ({}^*B_p)^* \simeq B_p. $$

(c) *The following continuous injections hold true:*

$$ L_0^\infty(p\cdot\mu) \sqsubseteq B_p \sqsubseteq \bigcap_{\alpha>1} L_0^\alpha(p\cdot\mu) \sqsubseteq {}^*B_p \sqsubseteq B_p^*. $$

Proof. We recall some results about the Orlicz spaces; for further details see, for instance, Rao and Ren (1991).

Let ϕ and ψ be conjugate functions. If $f \in L^\phi(p)$ and $g \in L^\psi(p)$, then there exists a constant k such that $\int |fg|p\,\mathrm{d}\mu \leqslant k\|f\|_{\phi,p}\|g\|_{\psi,p}$ (Rao and Ren 1991, p. 58).

If $u \in L^{\phi_2}(p)$ and ${}^*u \in L^{\phi_3}(p)$, then $\mathrm{E}_p[u\ {}^*u] \leqslant k\|u\|_{\phi_2,p}\|{}^*u\|_{\phi_3,p}$. Thus for all ${}^*u \in L^{\phi_3}(p)$ the mapping $u \mapsto \mathrm{E}_p[u\ {}^*u]$ is always defined, linear and continuous; it is an element of $L^{\phi_2}(p)^*$. Similarly for all $u \in L^{\phi_2}(p)$ the mapping ${}^*u \mapsto \mathrm{E}_p[u\ {}^*u]$ is an element of $L^{\phi_3}(p)^*$.

ϕ_1 and ϕ_2 are equivalent. Then, for any ${}^*u \in L^{\phi_3}(p)$, $(u \mapsto \mathrm{E}_p[u\ {}^*u]) \in L^{\phi_1}(p)^*$ and for any $u \in L^{\phi_1}(p)$, $({}^*u \mapsto \mathrm{E}_p[u\ {}^*u]) \in L^{\phi_3}(p)^*$,

$$ L^{\phi_3}(p) \sqsubseteq L^{\phi_1}(p)^* \quad \text{and} \quad L^{\phi_1}(p) \sqsubseteq L^{\phi_3}(p)^*. $$

Moreover, if ϕ is such that there exists k such that $\phi(2x) \leqslant k\phi(x)$, then $L^\phi(p)^*$ is isometric to $L^\phi(p)$ (Rao and Ren 1991, p. 111). We apply this to ϕ_3; then

$$ L^{\phi_1}(p) \simeq L^{\phi_2}(p) = L^{\phi_3}(p)^*. $$

Now we have to show that the same properties extend to the centred spaces, i.e.

$$ {}^*B_p \sqsubseteq B_p^* \quad \text{and} \quad B_p \simeq ({}^*B_p)^*. $$

Let ${}^*u \in {}^*B_p$. Then ${}^*u \in L^{\phi_3}(p)$ and there exists $u^* \in L^{\phi_1}(p)^*$ such that $u^*(u) = \mathrm{E}_p[{}^*uu]$, for all $u \in L^{\phi_1}(p)$; if u^* is restricted to B_p, then it is an element of B_p^* such that $u^*(u) = \mathrm{E}_p[{}^*uu]$, $u \in B_p$. The mapping ${}^*u \to u^*$ is continuous from ${}^*B_p$ to B_p^* because the restriction is a contraction.

The same argument applies to show that $B_p \sqsubseteq ({}^*B_p)^*$.

Now let $\overline{u} \in ({}^*B_p)^*$. $\overline{u}$ extends to a continuous linear operator $\tilde{u}$ on $L^{\phi_3}(p)$ defined by $\tilde{u}({}^*u) = \overline{u}({}^*u - \mathrm{E}_p[{}^*u])$. Let $u \in L^{\phi_1}(p)$ be the representation of $\tilde{u}$; then $\tilde{u}({}^*u) = \mathrm{E}_p[u\ {}^*u]$, with ${}^*u \in L^{\phi_3}(p)$. In particular, for ${}^*u \in {}^*B_p$, $\tilde{u}({}^*u) = \overline{u}({}^*u) = \mathrm{E}_p[u^*u]$.

Proposition 8(c) follows from the existence of constants k_1, k_2 and k_3 such that

$$\|.\|_\infty \geqslant k_1\|.\|_{\phi_1} \geqslant k_2\|.\|_\alpha \geqslant k_3\|.\|_{\phi_3}.$$

Such inequalities depend on the different order of growth at ∞ of the ϕ functions involved (Rao and Ren 1991, p. 155). □

Proposition 9. *The multilinear mappings* $(u_1, \ldots, u_n) \mapsto \mathrm{E}_p[u_1 \ldots u_n]$ *with* $u_i \in B_p$, *are continuous; in particular the moments* $u \mapsto \mathrm{E}_p[u^n]$ *are continuous.*

Proof. This follows from the inclusion $B_p \sqsubseteq L^n(p)$; see Proposition 8(c). □

The Banach space B_p is an algebraic subspace of the Hilbert space $L_0^2(p)$, and its topology is stronger than the induced topology; in particular the scalar product of $L_0^2(p)$ is defined on it and it is continuous.

Definition 10 (Orthogonality in B_p). *The covariance induces a continuous scalar product on* B_p *defined as*

$$\langle u, v\rangle_p = \mathrm{E}_p[uv] = \mathrm{cov}_p[u, v] \text{ for all } u, v \in B_p.$$

We say that u *and* v *in* B_p *are* orthogonal *if* $\langle u, v\rangle_p = 0$.

The scalar product $\langle \cdot, \cdot \rangle_p$ extends to the usual scalar product in $L_0^2(p)$. A different extension is possible, as is shown in the following definition.

Definition 11 (Orthogonality on $B_p^* \times B_p$). *We shall denote by* $\langle ., . \rangle_{*,p}$ *the bilinear form between the Banach space* B_p *and its dual Banach space* B_p^*:

$$B_p^* \times B_p \ni (u^*, u) \mapsto u^*(u) = \langle u^*, u\rangle_{*,p}.$$

We say that $u^* \in B_p^*$ *is* orthogonal *to* $u \in B_p$ *if* $\langle u^*, u\rangle_{*,p} = 0$.

As $\langle \cdot, \cdot \rangle_{*,p}$ is not a scalar product, being defined on a product of different space, the previous definition is not consistent with the usual mathematical terminology. Nevertheless we suggest using the term "orthogonality" in this case because this notion is exactly what we need to discuss the notion of "orthogonal parametrization" in our framework; see below in Section 6.2.

We have seen in Proposition 8 that the Banach space ${}^*B_p$ can be identified with a subspace of the dual space B_p^*. Thus, if $u^* \in B_p^*$ is identified with ${}^*u \in {}^*B_p$, one has

$$\langle u^*, u\rangle_{*,p} = \mathrm{E}_p[{}^*uu].$$

2.2. Analytical prerequisites for the construction of the atlas

The patches of the atlas will be defined on the open ball of radius 1:

$$\mathscr{V}_p = \{u \in B_p \colon \|u\|_p < 1\},$$

where we have denoted $\|u\|_{\phi_1,p}$ by $\|u\|_p$.

Proposition 12. *If* $u \in \mathscr{V}_p$ *and* $q = \mathrm{e}^u\, p / \mathrm{E}_p[\mathrm{e}^u]$ *then*

(a) *the random variable* e^u *is* $(p \cdot \mu)$ *integrable and* q *is a probability density in* $\mathscr{M}(X, \mathscr{X}, \mu)$ *and*

(b) $L^{\phi_1}(p \cdot \mu) = L^{\phi_1}(q \cdot \mu)$.

Proof.

(a) We remark that the condition $\|u\|_p < 1$ is equivalent to the existence of an $\alpha > 1$ such that $\mathrm{E}_p[\cosh(\alpha u) - 1] \leqslant 1$, which in turn implies that $\mathrm{E}_p[\mathrm{e}^u] \leqslant 4$.

(b) The hypothesis implies that p and q are connected by a one-dimensional exponential model and the conclusion follows from Proposition 5. □

Definition 13 (Moment generating functional). *The* moment generating functional

$$G_p\colon L^{\phi_1}(p \cdot \mu) \to \overline{R}_+ = [0, +\infty]$$

is defined by

$$G_p(u) = \mathrm{E}_p[\mathrm{e}^u].$$

Proposition 14 (Properties of the moment generating functional). *The moment generating functional* G_p

(a) *takes the value 1 at 0; otherwise is strictly greater than 1, is convex and its proper domain* $\mathrm{dom}(G_p) = \{u \in L^{\phi_1}(p \cdot \mu)\colon G_p(u) < \infty\}$ *is a convex set which contains the open unit ball of* $L^{\phi_1}(p \cdot \mu)$;

(b) *is bounded and infinitely Fréchet differentiable on the open unit ball* $\mathscr{V}_p$ *with differential*

$$\mathrm{D}^n G_p(u)(v_1, \ldots, v_n) = \mathrm{E}_p[v_1 \ldots v_n\, \mathrm{e}^u].$$

Proof. See Pistone and Sempi (1995, Proposition 2.4). □

The previous abstract definition includes the usual definition of a (multivariate) moment generating function. In fact, let $u = (u_1, \ldots, u_n)$ be an n-dimensional random variable of the Cramer class under the density p. Then for each real vector $t = (t_1, \ldots, t_n)$ the linear combination $\sum_{i=1}^n t_i u_i$ belongs to the Cramer class and its multivariate moment generating function $G_p(t_1 \ldots t_n)$ is equal to $G_p(\sum_{i=1}^n t_i u_i)$. We remark that the multivariate moments of order $r_1, \ldots, r_n$ of the random variables $u_1, \ldots, u_n$ are the value of the derivative at 0 of order $r_1, \ldots, r_n$ of G_p in the direction $u_1, \ldots, u_n$:

$$\mathrm{E}_p[u_1^{r_1} \ldots u_n^{r_n}] = \mathrm{D}^{r_1 + \cdots + r_n} G_p(0)\ (u_1^{\circ r_1}, \ldots, u_n^{\circ r_n})$$

where $u^{\circ r}$ means $\overbrace{u, \ldots, u}^{r \text{ times}}$.

Definition 15 (Cumulant generating functional). *The* cumulant generating functional $K_p\colon B_p \to [0, +\infty]$ *is defined by*

$$K_p(u) = \log G_p(u).$$

We remark that we restrict the cumulant generating functional to be defined on *centred* random variables of the Cramér class at p.

Proposition 16 (Cumulant). *The cumulant generating functional K_p has proper domain $\mathrm{dom}(G_p) \cap B_p$. If $\mathscr{V}_p$ denotes the open ball of B_p of radius 1 then $\mathscr{V}_p \subset \mathrm{dom}(G_p) \cap B_p$. Moreover K_p satisfies the following properties.*

(a) *K_p is 0 at 0; otherwise is strictly positive, is convex and infinitely Fréchet differentiable on $\mathscr{V}_p$.*

(b) *$\forall u \in \mathscr{V}_p$, $q = \mathrm{e}^{u - K_p(u)}\, p$ is a probability density in $\mathscr{M}(X, \mathscr{X}, \mu)$. The value of the nth differential at u in the direction v ($\in B_p$) of K_p, that is the n-linear continuous form $\mathrm{D}^n K_p(u)$ applied to $(\overbrace{v, \ldots, v}^{n \text{ times}})$, is the nth cumulant of v under the probability density q:*

$$\mathrm{D}^n K_p(u) v^n = \frac{\mathrm{d}^n}{\mathrm{d}t^n} \log \mathrm{E}_q[\mathrm{e}^{tv}] \Big|_{t=0}.$$

(c) *For v, v_1 and v_2 in B_p, one has*

$$\begin{aligned} \mathrm{D}K_p(u)\, v &= \mathrm{E}_q[v], \\ \mathrm{D}^2 K_p(u)\, (v_1, v_2) &= \mathrm{E}_q[v_1 v_2] - \mathrm{E}_q[v_1]\, \mathrm{E}_q[v_2] = \mathrm{cov}_q[v_1, v_2]. \end{aligned} \tag{7}$$

(d) *$\forall u \in \mathscr{V}_p$ and $q = \mathrm{e}^{u - K_p(u)}\, p$, the random variable $q/p - 1$ belongs to ${}^*B_p$ and*

$$\mathrm{D}K_p(u)\, v = \mathrm{E}_p\left[\left(\frac{q}{p} - 1\right) v\right], \qquad v \in B_p.$$

In other words the differential of K_p at u, $\mathrm{D}K_p(u)$, is in B_p^* but actually is identified with an element of ${}^*B_p$, denoted by $\nabla K_p(u)$:

$$\nabla K_p(u) = \mathrm{e}^{u - K_p(u)} - 1 = \frac{q}{p} - 1.$$

(e) *The mapping $B_p \ni u \mapsto \nabla K_p(u) \in {}^*B_p$ is monotonic, and in particular one to one.*

(f) *The weak derivative of the map $B_p \ni u \mapsto \nabla K_p(u) \in {}^*B_p$ at u applied to $w \in B_p$ is given by*

$$\mathrm{D}(\nabla K_p(u))\, w = \frac{q}{p}(w - \mathrm{E}_q[w]),$$

and it is one to one at each point.

Proof.

(a) For the first point see Pistone and Sempi (1995, Proposition 2.5(a)).

(b) It is a consequence of the definition of $K_p(u)$ and its properties, see Proposition 14.

(c) Same proof as (b).

(d) We note that the random variable $q/p - 1$ is centred: $\mathrm{E}_p[q/p] - 1 = \mathrm{E}_q[1] - 1 = 0$.

To prove that $q/p - 1 \in {}^*B_p$ we show, using the Proposition 7, that $q/p \in L^{\phi_3}(p)$, i.e.

$$\mathrm{E}_p\left[\left(1+\frac{q}{p}\right)\log\left(1+\frac{q}{p}\right)\right] < +\infty.$$

We have

$$\begin{aligned}\mathrm{E}_p\left[\left(1+\frac{q}{p}\right)\log\left(1+\frac{q}{p}\right)\right] &= \mathrm{E}_p\left[\log\left(1+\frac{q}{p}\right)\right] + \mathrm{E}_p\left[\frac{q}{p}\log\left(1+\frac{q}{p}\right)\right] \\ &\leqslant \mathrm{E}_p\left[\frac{q}{p}\right] + \mathrm{E}_p\left[\frac{q}{p}\log\left(1+\frac{q}{p}\right)\right] \\ &= 1 + \mathrm{E}_p\left[\frac{q}{p}\log\left(1+\frac{q}{p}\right)\right].\end{aligned}$$

So we have to show that

$$\mathrm{E}_p\left[\frac{q}{p}\log\left(1+\frac{q}{p}\right)\right] < +\infty.$$

By the inequality $x\log(1+x) \leqslant (1+x)\log^+(x) + 1$, for $x > 0$, where $\log^+(x) = \max\{\log(x), 0\}$, we have

$$\begin{aligned}\mathrm{E}_p\left[\frac{q}{p}\log\left(1+\frac{q}{p}\right)\right] &\leqslant \mathrm{E}_p\left[\left(1+\frac{q}{p}\right)\log^+\left(\frac{q}{p}\right)\right] + 1 \\ &= \mathrm{E}_p\left[\log^+\left(\frac{q}{p}\right)\right] + \mathrm{E}_p\left[\frac{q}{p}\log^+\left(\frac{q}{p}\right)\right] + 1.\end{aligned} \tag{8}$$

We know that

$$\log\left(\frac{q}{p}\right) = u - K_p(u) \in L^{\phi_1}(p) \sqsubseteq L^1(p)$$

and also (see Proposition 12)

$$\log\left(\frac{q}{p}\right) \in L^{\phi_1}(q) \sqsubseteq L^1(q).$$

Then all terms on the right-hand side of (8) are bounded. Since $v \in B_p$ implies that $\mathrm{E}_p[v] = 0$, the first equality in (7) may be written as

$$\mathrm{D}K_p(u)\,v = \mathrm{E}_p\left[\left(\frac{q}{p} - 1\right)v\right].$$

By definition the gradient $\nabla K_p(u)$ is an element of B_p^* such that

$$\mathrm{D}K_p(u)\,v = \langle \nabla K_p(u),\, v\rangle_{*,p}.$$

By comparing the two previous equations, we conclude that the gradient can be identified with $q/p - 1 \in {}^*B_p$. By a change in notation we write $\nabla K_p(u) = q/p - 1 \in {}^*B_p$.

(e) Note that, if u and $\overline{u}$ are in B_p and $\theta \in [0, 1]$, the following relation holds:

$$\begin{aligned}
\mathrm{E}_p[(\nabla K_p(\overline{u}) - \nabla K_p(u))(\overline{u} - u)] &= \langle \nabla K_p(\overline{u}) - \nabla K_p(u), (\overline{u} - u) \rangle_{*,p} \\
&= \left\langle \int_0^1 \mathrm{d}\theta \frac{\mathrm{d}}{\mathrm{d}\theta} \nabla K_p((1-\theta)u + \theta\overline{u}), (\overline{u} - u) \right\rangle_{*,p} \\
&= \int_0^1 \mathrm{d}\theta \, \mathrm{D}^2 K_p((1-\theta)u + \theta\overline{u})(\overline{u} - u, \overline{u} - u) \\
&= \int_0^1 \mathrm{d}\theta \, \mathrm{var}_\theta(\overline{u} - u)
\end{aligned} \tag{9}$$

where var_θ is the variance with respect to the density $p_\theta = e_p((1-\theta)u + \theta\overline{u})$. As the mapping $\theta \mapsto \mathrm{D}^2 K_p((1-\theta)u + \theta\overline{u})$ is continuous and positive definite, then $\theta \mapsto \mathrm{var}_\theta(\overline{u} - u)$ is continuous and non-negative. If $\nabla K_p(\overline{u}) = \nabla K_p(u)$, then $\mathrm{var}_\theta(\overline{u} - u) = 0$ for any θ. This implies that $\overline{u} - u =$ constant and $\overline{u} = u$, because they are p-centred random variables.

(f) The map $f\colon B_p \ni u \mapsto \nabla K_p(u) \in^* B_p$ is weakly differentiable if, $\forall v \in B_p$, $u \mapsto \mathrm{E}_p[f(u)v]$ is differentiable. We have

$$\mathrm{E}_p[f(u)v] = \mathrm{E}_p[\nabla K_p(u)v] = \mathrm{D}K_p(u)\, v.$$

Now we consider the increment of $\nabla K_p(u)$ in the direction $w \in B_p$:

$$\begin{aligned}
\mathrm{E}_p[(f(u+w) - f(u))v] &= \mathrm{D}K_p(u+w)\, v - \mathrm{D}K_p(u)\, v \\
&= \mathrm{D}^2 K_p(u)(w, v) + R_p(u, v, w),
\end{aligned}$$

where $|R_p(u, v, w)| = o(\|w\|_p)$. We have shown that

$$\mathrm{D}^2 K_p(u)(w, v) = \mathrm{cov}_q[v, w] = \mathrm{E}_p\left[\frac{q}{p}(w - \mathrm{E}_q[w])\, v\right], \tag{10}$$

so that $\mathrm{E}_p[\mathrm{D}f(u)wv] = \mathrm{E}_p[(q/p)(w - \mathrm{E}_q[w])\, v]$. To check that $(q/p)(w - \mathrm{E}_q[w])$ is an element of ${}^*B_p$ we use (10) and the properties of the norms with respect to two conjugate functions. We want to prove that

$$\sup_{\|v\|_p < 1} \mathrm{E}_p\left[\frac{q}{p}(w - \mathrm{E}_q[w])\, v\right] < +\infty.$$

Now $\mathrm{cov}_q[v, w] = \mathrm{D}^2 K_p(u)(v, w) \leqslant k\|v\|_{\phi_2,p}\|w\|_{\phi_2,p}$ because K_p is twice differentiable at u. Finally $\sup_{\|v\|_{\phi_2,p} \leqslant 1} \|v\|_p < +\infty$ because the norms $\|\cdot\|_{\phi_2,p}$ and $\|\cdot\|_p$ are equivalent. The weak derivative is one to one; in fact, if there exists v_1 such that

$$\frac{q}{p}(v - \mathrm{E}_q[v]) = \frac{q}{p}(v_1 - \mathrm{E}_q[v_1]),$$

then $v - v_1 =$ constant and $v = v_1$ because they are centred random variables. □

Using the previous definitions and properties, it is possible to give a definition of the nonparametric exponential model as follows.

Definition 17 (Maximal exponential model). *For each p in $\mathscr{M}(X, \mathscr{X}, \mu)$ the maximal exponential model at p is the statistical model*

$$\mathscr{E}_p = \{\mathrm{e}^{u - K_p(u)}\, p\colon u \in \mathrm{dom}(K_p)^\circ,\ \mathrm{E}_p[u] = 0\}.$$

The function

$$B_p \supset \mathrm{dom}(K_p)^\circ \ni u \mapsto \mathrm{e}^{u - K_p(u)}\, p \in \mathscr{M}(X, \mathscr{X}, \mu)$$

is the likelihood function of the maximal exponential model; u plays the role of the "model parameter".

Any *parametric exponential* model generated by p is embedded into the maximal exponential model as follows. Let $(u_1, \ldots, u_d)$ be the sufficient statistic of the parametric exponential model, with $u_i \in B_p$, and let

$$\Theta = \left\{ \theta = (\theta_1, \ldots, \theta_d) \in \mathbb{R}^d\colon \sum \theta_i u_i \in \mathrm{dom}(K_p)^\circ \right\}. \tag{11}$$

Then the parametric exponential model is

$$p_\theta = \mathrm{e}^{\sum \theta_i u_i - \psi(\theta)}\, p \tag{12}$$

with $\psi(\theta) = K_p(\sum \theta_i u_i)$.

Note that the parametric exponential model is uniquely characterized by the linear subspace spanned by $(u_1, \ldots, u_d)$. Note also that

$$\frac{\partial}{\partial \theta_j}\psi(\theta) = \mathrm{D}K_p\left(\sum \theta_i u_i\right) u_j = \mathrm{E}_{p_\theta}[u_j] = \mathrm{E}_p\left[\frac{p_\theta}{p} u_j\right] = \mathrm{E}_p\left[\left(\frac{p_\theta}{p} - 1\right) u_j\right] = \mathrm{E}_p[{}^*u_\theta u_j],$$

where ${}^*u_\theta = p_\theta / p - 1 = \nabla K_p(u_\theta)$; the second equality follows from Proposition 16 and the last but one holds because the u_j are p-centred random variables.

We denote

$$\frac{\partial}{\partial \theta_j}\psi(\theta)$$

by η_j. The parameters $(\eta_1, \ldots, \eta_d)$ are the *mean parameters* of Barndorff-Nielsen and Cox (1994) or the *mixture coordinates* of Amari (1982).

2.3. The atlas

We now have all the elements for the definition of the atlas (see Figure 1). Let us consider the following map defined on a subset $\mathscr{V}_p$ of the proper domain of K_p:

$$e_p\colon \mathscr{V}_p \ni u \mapsto q = \mathrm{e}^{u - K_p(u)}\, p \in \mathscr{M}(X, \mathscr{X}, \mu), \tag{13}$$

where $K_p(u) = \log \mathrm{E}_p[\mathrm{e}^u] = \log G_p(u)$ is the cumulant generating functional computed at u.

This mapping is one to one because u is centred; in fact, if $u_1,\ u_2 \in \mathscr{V}_p$ and $e^{u_1 - K_p(u_1)} = e^{u_2 - K_p(u_2)}$, then $u_1 - K_p(u_1) = u_2 - K_p(u_2)$, and $u_1 - u_2$ is constant and this constant has to be 0.

According to (1) and (2) we shall denote by $\mathscr{U}_p$ the image of $\mathscr{V}_p$ by the mapping e_p and by s_p the inverse of e_p on $\mathscr{U}_p$. Such an inverse, $s_p\colon \mathscr{U}_p \to \mathscr{V}_p$, is easily computed as

$$s_p\colon \mathscr{U}_p \ni q \mapsto \log\left(\frac{q}{p}\right) - \mathrm{E}_p\left[\log\left(\frac{q}{p}\right)\right] \in \mathscr{V}_p. \tag{14}$$

The functions $s_p,\ p \in \mathscr{M}(X, \mathscr{X}, \mu)$, will be the coordinate mappings of our manifold in the sense that, locally around each $p \in \mathscr{M}(X, \mathscr{X}, \mu)$, each $q \in \mathscr{U}_p$ will be "parametrized" by its *centred log-likelihood*.

Let us now compute the change-in-coordinates formula; if p_1 and p_2 are two points in $\mathscr{M}(X, \mathscr{X}, \mu)$ such that $\mathscr{U}_{p_1} \cap \mathscr{U}_{p_2} \neq \emptyset$, then for all q in that intersection

$$\log\left(\frac{p_1}{p_2}\right) = \log\left(\frac{p_1}{q}\right) + \log\left(\frac{q}{p_2}\right)$$

belongs to $L^{\phi_1}(p_1) = L^{\phi_1}(q) = L^{\phi_1}(p_2)$. The composite transition mapping

$$s_{p_2} \circ e_{p_1}\colon s_{p_1}(\mathscr{U}_{p_1} \cap \mathscr{U}_{p_2}) \to s_{p_2}(\mathscr{U}_{p_1} \cap \mathscr{U}_{p_2})$$

simplifies to

$$s_{p_2} \circ e_{p_1}(u) = u + \log\left(\frac{p_1}{p_2}\right) - \mathrm{E}_{p_2}\left[u + \log\left(\frac{p_1}{p_2}\right)\right], \tag{15}$$

where the algebraic computations are done in the space of μ classes of measurable functions and the expectation is well defined as long as $\mathscr{U}_{p_1} \cap \mathscr{U}_{p_2} \neq \emptyset$.

Theorem 18. *The collection of pairs* $\{(\mathscr{U}_p, s_p)\colon p \in \mathscr{M}(X, \mathscr{X}, \mu)\}$ *is an affine* C^∞ *atlas on* $\mathscr{M}(X, \mathscr{X}, \mu)$. *The induced topology on sequences is equivalent to* e *convergence and the transition mappings are those defined in* (15).

Proof. This is the main result of Pistone and Sempi (1995, Theorem 3.6). □

Note that the derivative of the transition mapping defined in (15) is

$$B_{p_1} \ni u \mapsto u - \mathrm{E}_{p_2}[u] \in B_{p_2},$$

and this is an isomorphism between B_{p_1} and B_{p_2} as topological linear spaces (Lang 1995). This extends Proposition 12 (b).

Definition 19 (Manifold). *The* exponential (statistical) manifold *is the manifold defined by the property in Theorem 18 on the set* $\mathscr{M}(X, \mathscr{X}, \mu)$.

The maximal exponential model defined in Definition 17 has a precise place in the general framework, as the following theorem shows.

Theorem 20. *The maximal exponential model $\mathscr{E}_p$ is the connected component containing p of the exponential manifold $\mathscr{M}(X, \mathscr{X}, \mu)$.*

Proof. See Pistone and Sempi (1995, Theorem 4.1). □

The manifold structure that we have defined is special; many other types of atlas have been suggested in the literature, in particular the coordinates based on the mean parameters of Barndorff-Nielsen and Cox (1994) and the so-called Amari embeddings described by Amari (1982). In the infinite-dimensional case those different geometric structures are not equivalent to the exponential manifold, but in some restricted sense they are, because they induce the same manifold structure on finite-dimensional submanifolds (i.e. parametric statistical manifolds) (see again Amari (1982, 1985) and Murray and Rice (1993)).

3. Tangent space

A basic object of the theory of manifolds is the *tangent bundle*. In the case of the exponential statistical manifold it has been remarked from the very beginning (Dawid 1975) that there is a very natural identification between the tangent vectors and the exponential one-dimensional models around a point p. In fact each differentiable curve in $\mathscr{M}(X, \mathscr{X}, \mu)$, i.e. each one-dimensional statistical model $p(t)$, $t \in I \subseteq R$, such that $p(0) = p$, has a tangent model of the exponential form $e^{tu - K_p(tu)}\, p$. This prompts the identification of the tangent space with the set of one-dimensional exponential models.

Let $p_1(t)$ and $p_2(t)$ be two regular curves such that $p_1(t_0) = p_2(t_0) = p$ and let $u_1(t)$ and $u_2(t)$ be the corresponding representations by a chart s_q, i.e. $u_1(t) = s_q(p_1(t))$ and $u_2(t) = s_q(p_2(t))$. Then the curves p_1 and p_2 are equivalent at p if $\dot{u}_1(t_0) = \dot{u}_2(t_0)$. We denote by $T_p\mathscr{M}$ the set of equivalence classes of regular curves through p. In local coordinates determined by the chart s_q the equivalence class corresponding to $p(t)$ may be represented by the tangent vector to $u(t)$ at t_0, that is by $\dot{u}(t_0)$, if $p(t_0) = p$ and $u(t) = s_q(p(t))$. This definition does not depend on the chart chosen (Lang 1995).

Proposition 21 (Tangent space). *Let $p(t)$ be a regular curve in $\mathscr{M}(X, \mathscr{X}, \mu)$ with $p(t_0) = p$, and let $u(t) \in B_q$ be its representation by a chart s_q, where $t \in \{t\colon u(t) \in \mathrm{dom}(K_q)^\circ\}$. Then $p(t) = e^{u(t) - K_q(u(t))}\, q$.*

(a) *The relation between $\dot{u}(t_0)$, the tangent to $u(t)$ at t_0, and the* score function *of $p(t)$ with respect to the density p is*

$$\dot{u}(t_0) - \mathrm{E}_q[\dot{u}(t_0)] = \frac{\mathrm{d}}{\mathrm{d}t} \log\left(\frac{p(t)}{p}\right)\bigg|_{t=0}.$$

If $q = p$, i.e. if the chart is centred at the same point where the log-likelihood is calculated, we have

$$\dot{u}(t_0) = \frac{\mathrm{d}}{\mathrm{d}t} \log\left(\frac{p(t)}{p}\right)\Big|_{t=0}.$$

Then the space of the score function is a representation of the tangent space $T_p\mathscr{M}$.

(b) *The curve*

$$t \mapsto \frac{p(t)}{p} - 1$$

is in ${}^*B_p$ *and its weak derivative at* t_0 *is* $\dot{u}(t_0)$.

(c) *The score function of any one-dimensional exponential model through p, i.e.*

$$\mathrm{e}^{tu - K_p(tu)}\, p,$$

at $t = t_0$, *is* u *and vice versa any* $u \in B_p$ *has such a corresponding one-dimensional exponential model. Then the space of the one-dimensional exponential models is another representation of the tangent space* $T_p\mathscr{M}$.

Proof.

(a) We have

$$\frac{\mathrm{d}}{\mathrm{d}t} \log\left(\frac{p(t)}{p}\right)\Big|_{t=t_0} = \frac{\mathrm{d}}{\mathrm{d}t}\{u(t) - u(t_0) - K_q(u(t)) + K_q(u(t_0))\}\Big|_{t=t_0}$$

$$= \dot{u}(t_0) - \mathrm{E}_p[\dot{u}(t_0)].$$

If $p = q$, then $\mathrm{E}_p[\dot{u}(t_0)] = \mathrm{E}_q[\dot{u}(t_0)] = 0$.

(b) From Proposition 16 (c) we have

$$\frac{p(t)}{p} - 1 = \nabla K_p(u(t)).$$

Its weak derivative (see Proposition 16 (f)) calculated at $t = t_0$, is

$$\frac{\dot{p}(t_0)}{p} = \mathrm{D}\nabla K_p(u(t))\, \dot{u}(t)\Big|_{t=t_0} = \frac{p(t)}{p}(\dot{u}(t) - \mathrm{E}_{p(t)}[\dot{u}(t)])\Big|_{t=t_0} = \dot{u}(t_0)$$

and it is in ${}^*B_p$.

(c) It follows by direct computation that

$$\frac{\mathrm{d}}{\mathrm{d}t}\{tu - K_p(tu)\}|_{t=t_0} = u. \qquad \square$$

The tangent space inherits the structure of vector space and the topology from B_p.

As a scalar product is defined on B_p (see Definition 10), a scalar product is defined on $T_p\mathscr{M}$ together with the definition of orthogonality.

Let $v \in T_p\mathscr{M}$ and let $p(t)$ a regular curve whose tangent vector is v, i.e. a curve tangent at p to $p(t) = \mathrm{e}^{tv - K_p(tv)}\, p$. Let $\varphi\colon \mathscr{M}(X, \mathscr{X}, \mu) \to \mathbb{R}$. The differential of φ at p, denoted by $\mathrm{d}_p\varphi$, is a linear form on the tangent space:

$$\mathrm{d}_p\varphi(v) = \frac{\mathrm{d}}{\mathrm{d}t}\varphi(p(t))\bigg|_{t=0}.$$

If $p(t) = e_q(u(t))$, then

$$\mathrm{d}_p\varphi(v) = \mathrm{D}\{\varphi \circ e_q(u(t_0))\}\, v.$$

3.1. Regular parametrization

Now we give a definition of a parametrization and we shall show an example of a parametrization that is not a chart in our sense.

Definition 22. *Let A be an open set of the exponential statistical manifold $\mathscr{M}(X, \mathscr{X}, \mu)$, and let B be a Banach space. We shall say that $F\colon A \to B$ is a C^k* parametrization *of A, $k = 1, 2, \ldots, \infty$, if F is a one-to-one, k-times continuously differentiable parameter and the tangent mapping d_pF is one to one at each p.*

Note that the condition of being a C^∞ parametrization is weaker than the condition of being a chart (essentially because we do not require $(\mathrm{d}_pF)^{-1}$ to be continuous), unless the manifold is finite dimensional. In fact the regularity of the inverse mapping is not ensured by the conditions in Definition 22.

As an example, we consider the parameter shown in Proposition 16 based on the likelihood: $q \mapsto q/p - 1$. For any $u \in \mathscr{V}_p$ there exists an ${}^*u \in {}^*B_p$ such that ${}^*u = \nabla K_p(u) = q/p - 1$. The map $u \mapsto {}^*u = q/p - 1$ is one to one and ${}^*u$ belongs to $x - \log x$ class (see Proposition 16); then this is a reasonable parametrization but it does not define a chart.

To see this, we denote by $\mathscr{Z}_p$ the image of $\mathscr{U}_p$ in ${}^*B_p$ under the mapping $q \mapsto q/p - 1$. If $u \mapsto {}^*u = q/p - 1$, a chart $\mathscr{Z}_p$ would be an open set of ${}^*B_p$. This is false; in fact a base of the neighbourhood of ${}^*B_p$ is of the form $\{{}^*u\colon \mathrm{E}_p[\phi_3({}^*u)] < k\}$ while ${}^*u \in \mathscr{Z}_p$ satisfies the condition ${}^*u \geqslant -1$ and, if the space does not have a finite number of atoms, $\int({}^*u + 1)\log({}^*u + 1)\, p\, \mathrm{d}\mu < +\infty$ does not imply that ${}^*u \geqslant -1$. So in the nonparametric theory the mean parameters do not define a system of charts.

Another example is the global parametrization $p \mapsto p^{1/2} \in L^2(\mu)$. This case has been discussed in more detail by Brigo and Pistone (1996).

4. Information

We now describe briefly how the notion of information (or entropy) is connected to the notion of the exponential statistical manifold.

Definition 23 (Kullback–Leibler information). *Let the probability densities p and q in $\mathscr{M}(X, \mathscr{X}, \mu)$ be given. If $(q/p)\log(q/p)$ is $(p \cdot \mu)$ integrable, then the Kullback–Leibler relative information of q with respect to p is the number*

$$K(q,\ p) = \int \frac{q}{p} \log\left(\frac{q}{p}\right) p \,\mathrm{d}\mu = \mathrm{E}_p\left[\frac{q}{p} \log\left(\frac{q}{p}\right)\right] = \mathrm{E}_q\left[\log\left(\frac{q}{p}\right)\right].$$

Proposition 24. *Let $p \in \mathscr{M}(X, \mathscr{X}, \mu)$, let $q \in \mathscr{U}_p$ and let u be the s_p coordinate of q, i.e. $q = \mathrm{e}^{u - K_p(u)}\, p$. Let ${}^*u \in {}^*B_p$ defined as ${}^*u = q/p - 1$ and let*

$$H_p({}^*u) = \mathrm{E}_p[(1 + {}^*u)\log(1 + {}^*u)].$$

Then

(a) $$K_p(u) = K(p,\ q),$$

(b) $$H_p({}^*u) = K(q,\ p),$$

(c) $$\mathrm{E}_q[u] = K(p,\ q) + K(q,\ p) = K_p(u) + H_p({}^*u).$$

Proof. We have

(a) $$K(p,\ q) = -\mathrm{E}_p\left[\log\left(\frac{q}{p}\right)\right] = -\mathrm{E}_p[u] + K_p(u) = K_p(u),$$

(b) $$H_p({}^*u) = \mathrm{E}_p[(1 + {}^*u)\log(1 + {}^*u)] = \mathrm{E}_p\left[\frac{q}{p} \log\left(\frac{q}{p}\right)\right] = K(q,\ p),$$

(c) $$K(q,\ p) = \mathrm{E}_q[u - K_p(u)] = \mathrm{E}_q[u] - K_p(u) = \mathrm{E}_q[u] - K(p,\ q).$$ □

Note that the value $s_p(q)$ of the mapping $s_p(\cdot)$ defined in (14) is the log-likelihood of q with respect to p plus the Kullback–Leibler relative information $K(p, q)$ whose value is $\mathrm{E}_p[\log(p/q)]$

$$s_p(q) = \log\left(\frac{q}{p}\right) + K(p,\ q).$$

The mapping s_p is connected to the maximum-likelihood estimator as follows.

Proposition 25 (Maximum expected log-likelihood). *Let $p \in \mathscr{M}(X, \mathscr{X}, \mu)$, $u \in \mathscr{V}_p$ and $e_p(u) = \mathrm{e}^{u - K_p(u)}\, p$. Let $\hat{u} \in \mathscr{V}_p$ and $q = \mathrm{e}^{\hat{u} - K_p(\hat{u})}\, p$.*

The maximum expected (at q) log-likelihood, i.e. the maximum of the function,

$$\mathscr{V}_p \ni u \mapsto \mathrm{E}_q\left[\log\left(\frac{e_p(u)}{p}\right)\right]$$

is obtained at the point $\hat{u}$ and

$$\max_u \left\{\mathrm{E}_q\left[\log\left(\frac{e_p(u)}{p}\right)\right]\right\} = K(q,\ p).$$

Proof. We have

$$\mathrm{E}_q\left[\log\left(\frac{e_p(u)}{p}\right)\right] = \mathrm{E}_q[u] - K_p(u).$$

The function $u \mapsto \mathrm{E}_q[u] - K_p(u)$ is concave in u. Its derivative at u in the direction $v \in B_p$ is $\mathrm{E}_q[v] - \mathrm{D}K_p(u)\,v = \mathrm{E}_q[v] - \mathrm{E}_{e_p(u)}[v]$. Such a derivative is zero for all v if and only if $q = e_p(u)$; then the maximum is obtained at $\hat{u} = s_p(q)$ and consequently

$$\max_u \left\{\mathrm{E}_q\left[\log\left(\frac{e_p(u)}{p}\right)\right]\right\} = \mathrm{E}_q[\hat{u}] - K_p(\hat{u}) = K(q,\ p). \qquad \square$$

Proposition 26. *Let $u \in B_p$. Let ${}^*u$ and $H_p({}^*u)$ be as defined in Proposition 24.*

The functions $K_p(u)$ and $H_p({}^*u)$ are conjugate convex functions. The relations of conjugacy are

$$H_p({}^*u) = \max_u \{\mathrm{E}_p[{}^*uu] - K_p(u)\},$$

$$K_p(u) = \max_{{}^*u} \{\mathrm{E}_p[{}^*uu] - H_p({}^*u)\}.$$

Proof. The first relation follows from Proposition 24 (b) and from Proposition 25:

$$\begin{aligned} H_p({}^*u) &= K(q,\ p) \\ &= \max_u \{\mathrm{E}_q[u] - K_p(u)\} \\ &= \max_u \left\{\mathrm{E}_p\left[\frac{q}{p}u\right] - K_p(u)\right\} \\ &= \max_u \left\{\mathrm{E}_p\left[\left(\frac{q}{p} - 1\right)u\right] - K_p(u)\right\}. \end{aligned}$$

The last relation follows from the general theory of convex analysis (Ekeland and Temam 1974). $\square$

The results of this section develop standard arguments on the entropy function (Kullback and Leibler 1951; Donsker and Varadhan 1975; Amari 1985; Kullback 1997 (some 38 years ago)).

5. Submanifold

Definition 27 (Submanifold, submodel). *Let $\mathscr{N}$ be a subset of the exponential manifold $\mathscr{M}(X, \mathscr{X}, \mu)$ and, for each density $p \in \mathscr{N}$, let V_p^1 and V_p^2 be closed subspaces of B_p, such that there exist*

(a) *a homeomorphism (i.e. a linear invertible and bi-continuous mapping) between B_p and the direct product $V_p^1 \times V_p^2$ (we say that V_p^1 and V_p^2 are* split *in B_p) and*

(b) *a chart on a neighbourhood* $\mathscr{W}_p$ *of* p*:*

$$\sigma_p \colon \mathscr{W}_p \to B_p \simeq V_p^1 \times V_p^2,$$

where σ_p *maps* $\mathscr{W}_p$ *onto the product of two open sets* $\mathscr{V}_p^1 \times \mathscr{V}_p^2$ *(with* $\mathscr{V}_p^1 \subset V_p^1$ *and* $\mathscr{V}_p^2 \subset V_p^2$*) and maps* $\mathscr{N} \cap \mathscr{W}_p$ *onto* $\mathscr{V}_p^1 \times \{0\}$.

We shall say that $\mathscr{N}$ *is a* submodel *or a* submanifold *of the exponential statistical manifold* $\mathscr{M}(X, \mathscr{X}, \mu)$.

A submanifold $\mathscr{N}$ is a manifold whose charts are the restriction of the charts σ_p to $\mathscr{N}$ (Figure 3).

Frequently below we shall use this particular splitting; V_p^1 and V_p^2 are closed subspaces of B_p such that $V_p^1 \cap V_p^2 = \{0\}$ and $B_p = V_p^1 + V_p^2$. Then any element $u \in B_p$ can be written uniquely as $u = u_1 + u_2$, with $u_i \in V_p^i$, $i = 1, 2$.

In this case $\mathscr{N} = \{q = \mathrm{e}^{u_1 - K_p(u_1)}\, p \colon u_1 \in V_p^1 \cap \mathscr{V}_p\}$ is a submanifold, which we call "exponential submodel" of the maximal exponential model.

5.1. Some examples of submanifolds

We discuss two relevant examples of submanifolds.

5.1.1. (m, d)*-curved exponential models*

We consider the parametric exponential model defined above in (12). Let $u_1, \ldots, u_d \in B_p$ and $V_p^1 = \{\sum_{i=1}^d \theta_i u_i;\ \theta_i \in \mathbb{R}\}$ be given. Let $P_{u_1,\ldots,u_d}$ be the orthogonal projector from B_p (as a subspace of $L_0^2(p)$) onto the finite-dimensional Hilbert space V_p^1. For each $u \in B_p$, $u = P_{u_1,\ldots,u_d} u + (I - P_{u_1,\ldots,u_d})u$. The orthogonal projector P is characterized by

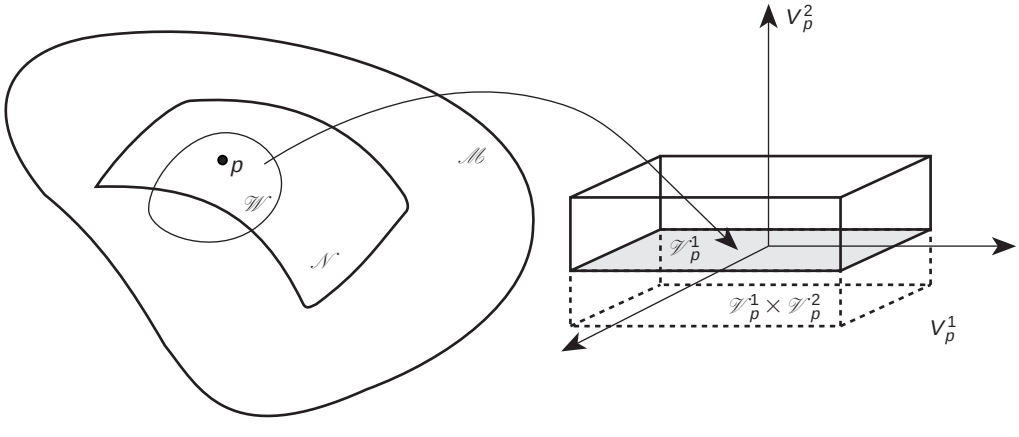

Figure 3. Submanifold: in this example the parallelepiped is $\mathscr{V}_p^1 \times \mathscr{V}_p^2$ and the emphasized rectangle is $\mathscr{V}_p^1$ (subset of horizontal plane).

$\langle u, u_j\rangle_p = \sum\hat{\theta}_i\langle u_i, u_j\rangle_p$ for $j = 1, \ldots, d$ and $\hat{\theta} \in \Theta$, with Θ as in (11). V^1_p is closed in B_p because it is finite dimensional. $P_{u_1,\ldots,u_d}$ and $I - P_{u_1,\ldots,u_d}$ are continuous functions in B_p; then $V^2_p = \ker(P_{u_1,\ldots,u_d})$ is a closed subset of B_p. This implies the existence of a splitting.

Then the d-dimensional parametric exponential model

$$\mathscr{N} = \left\{ \mathrm{e}^{\sum_i \theta_i u_i - \psi(\theta)}\, p \right\} \text{ with } \psi(\theta) = K_p\left(\sum_i \theta_i u_i\right)$$

is a submanifold of the nonparametric model $\mathscr{M}(X, \mathscr{X}, \mu)$. The set Θ, as in (11), is an open set of $\mathbb{R}^d$ and $\theta \leftrightarrow u = \sum\theta_i u_i$ is a C^∞ chart of $\mathscr{N}$ into Θ. If A is an m-dimensional submanifold of Θ the corresponding m-dimensional parametric submodel $\mathscr{S}$ is a submanifold of $\mathscr{N}$ and thus of $\mathscr{M}(X, \mathscr{X}, \mu)$. $\mathscr{S}$ is the (m, d)-curved exponential model of Barndorff-Nielsen and Cox (1994, p. 65).

5.1.2. Conditional expectation

Let $\mathscr{A}$ a sub-σ-algebra of $\mathscr{X}$. We consider $V^1_p = L^{\phi_1}_0(X, \mathscr{A}, \mu)$ and the conditional expectation

$$\mathrm{E}_p[.|\mathscr{A}]\colon B_p \to V^1_p.$$

This mapping is well defined; in fact any element $u \in B_p$ maps to an element of V^p_1: $\mathrm{E}_p[\phi_1(\alpha\mathrm{E}_p[u|\mathscr{A}])] \leqslant \mathrm{E}_p[\mathrm{E}_p[\phi_1(\alpha u|\mathscr{A})]] = \mathrm{E}_p[\phi_1(\alpha u)]$ (the Jensen inequality). It is surjective (any element of V^1_p maps to itself) and continuous (Neveu 1972). The subspace V^1_p is closed. We consider $V^2_p = \{u\colon \mathrm{E}_p[u|\mathscr{A}] = 0\}$. This subspace is closed because it is the *kernel* of a linear and continuous map.

V^1_p and V^2_p split in B_p because $u = \mathrm{E}_p[u|\mathscr{A}] + (u - \mathrm{E}_p[u|\mathscr{A}])$ is a unique decomposition. Then

$$\mathscr{M}(\mathscr{A}) = \{p \in \mathscr{M}(X, \mathscr{X}, \mu)\colon\ p \text{ is } \mathscr{A} \text{ measurable}\}$$

is a submanifold. In particular this applies to invariance with respect to measurable transformation of the sample space.

6. Splitting and orthogonality

What follows is a nonparametric version of results taken from Amari (1982); we extend to the statistical exponential manifold the method of mixed parametrization for exponential models (Barndorff-Nielsen 1978a,b).

6.1. Splitting

Let V^1_p and V^2_p be two closed subspaces of B_p, such that $V^1_p \cap V^2_p = \{0\}$ and $B_p = V^1_p + V^2_p$, i.e. any element $u \in B_p$ can be uniquely written as $u = u_1 + u_2$, with $u_i \in V^i_p$, $i = 1, 2$. We consider the linear projectors $P_i\colon B_p \to V^i_p$ defined by the splitting as $P_i(u) = u_i$

with $i = 1, 2$. The projectors are both continuous because of the Banach closed-graph theorem (Lang 1995, p. 4).

The mixed parametrization in exponential models is based on analytical computations that involve partial derivatives. Our next step is to describe how to compute the partial derivatives when the two components are not Cartesian components, but the two projections induced by a splitting. Such an extension is straightforward; nevertheless it is useful to have precise notation for the nonparametric case.

We consider the *annihilating subspaces* $(V^1_p)^0$ and $(V^2_p)^0$ of B^*_p, and ${}^0(V^1_p)$ and ${}^0(V^2_p)$ of ${}^*B_p$ defined by

$$(V^i_p)^0 = \{u^* \in B^*_p \colon \langle u^*, u_j\rangle_{*,p} = 0, \forall u_j \in V^j_p, j \neq i\},$$

$${}^0(V^i_p) = \{{}^*u \in {}^*B_p \colon \mathrm{E}_p[{}^*uu_j] = 0, \forall u_j \in V^j_p, j \neq i\},$$

with $i, j \in \{1, 2\}$. Such spaces can be considered the "orthogonal" spaces to the spaces of the splitting. This means, for example, that, if ${}^*u \in {}^0(V^2_p) \cap L^2_0(p)$, then ${}^*u$ belongs to $(V^1_p)^\perp$ in the Hilbert sense. Note that the numbering is such that sup-1 is "orthogonal" to sup-2 and vice versa for consistency with the natural L^2 notation.

The following poposition shows that the annihilating subspaces split the dual space and gives a characterization of the dual projections.

Proposition 28. *$(V^1_p)^0$ and $(V^2_p)^0$ split in B^*_p and any element of B^*_p can be written as $u^* = u^*_1 + u^*_2$, with $u^*_i \in (V^i_p)^0$, $i = 1, 2$. If $v \in B_p$, $v = v_1 + v_2$, $v_i \in V^i_p$, then $u^*_i(v) = u^* \circ P_i(v)$; so $u^*_i = u^* \circ P_i$, for $i = 1, 2$.*

Proof. This follows from the following properties of the subspaces.

(a) $(V^1_p)^0$ and $(V^2_p)^0$ are closed in B^*_p; in fact, if u_j is fixed in V^j_p, then the space $\{u^* \in B^*_p \colon \langle u^*, u_j\rangle_{*,p} = 0\}$ is closed because it is the kernel of a continuous map, and $(V^i_p)^0$, $i \neq j$, is the intersection of such closed subsets.

(b) $(V^1_p)^0 \cap (V^2_p)^0 = \{0\}$; in fact, if $u^* \in (V^1_p)^0 \cap (V^2_p)^0$, then, for any $v = v_1 + v_2 \in B_p$, $\langle u^*, v_1\rangle_{*,p} = 0$ and $\langle u^*, v_2\rangle_{*,p} = 0$; so $u^* = 0$ because the bilinear form of the duality is separating.

(c) $B^*_p = (V^1_p)^0 + (V^2_p)^0$; in fact, if $u^* \in B^*_p$, then $u^* = u^* \circ P_1 + u^* \circ P_2$; then, if $v \in B_p$, $v = v_1 + v_2$, we have $u^*(v) = u^* \circ P_1(v) + u^* \circ P_2(v) = u^*(v_1) + u^*(v_2)$ and u^*_i (with $u^*_i = u^*(P_i)) \in (V^i_p)^0$. The last assertion follows from the definition of $(V^i_p)^0$; in fact, if $v_j \in V^j_p$ and $i \neq j$, then

$$\langle u^*_i, v_j\rangle_{*,p} = u^*_i(v_j) = u^* \circ P_i(v_j) = u^*(0) = 0. \qquad \square$$

Note that properties (a) and (b) hold also for ${}^0(V^1_p)$ and ${}^0(V^2_p)$.

6.1.1. Partial derivatives of K_p

Given a splitting of B^*_p, we now look for a characterization of u^*_1 and u^*_2 for those elements

u^* of B_p^* (with $u^* = \mathrm{D}K_p(u)$) identified with an element ${}^*u$ of ${}^*B_p$ such that ${}^*u = \nabla K_p(u)$ (see Proposition 8 (a) and Proposition 16 (c)).

If ${}^*u = \nabla K_p(u) \in {}^*B_p$ is identified with $u^* = \mathrm{D}K_p(u) \in B_p^*$, $v = v_1 + v_2 \in B_p$ and $u^* = u_1^* + u_2^* \in B_p^*$, then

$$u_i^*(v) = u^*(v_i) = \mathrm{D}K_p(u)v_i = \mathrm{E}_p[\nabla K_p(u)\, v_i], \qquad i = 1, 2. \tag{16}$$

For $i = 1, 2$ we define the two partial derivative of $K_p(u)$ in the direction v as follows:

$$\partial_i K_p(u)\, v = \frac{\partial}{\partial t} K_p(u + P_i(tv))_{t=0} = \mathrm{D}K_p(u)\, P_i(v) = \mathrm{D}K_p(u)\, v_i = \mathrm{E}_p[\nabla K_p(u)\, v_i]. \tag{17}$$

Note that $\partial_i K_p(u)$ is an extension to B_p of the partial derivative of the function $(u_1, u_2) \mapsto K_p(u_1, u_2) = K_p(u_1 + u_2)$ defined on $V_p^1 \times V_p^2$; such a partial derivative takes the value 0 on V_p^j, with $i \neq j$ and $i, j \in \{1, 2\}$.

By (16) and (17), we have that $\partial_i K_p(u) = u_i^*$. In general, this element of the dual space will not be in ${}^*B_p$, even if the gradient ${}^*u = \nabla K_p(u)$ of K_p is; in the following we shall study special cases where such inclusion takes place; see Propositions 35 and 37 below.

The problem of finding general conditions which ensure that $u_i^* \in B_p^*$ may be identified with an element of ${}^*B_p$ remains open.

6.2. Orthogonality and mean parameters

In this section we show how to extend the notion of the Fisher information matrix and the corresponding notion of orthogonality (see, for example, Barndorff-Nielsen and Cox (1994)) to the exponential statistical manifold.

We assume that there exists a splitting V_p^1, V_p^2 of B_p. Let $q = \mathrm{e}^{u - K_p(u)}\, p \in \mathscr{C}_p$ be given. From the remark immediately after the proof of Theorem 18 it is seen that the derivative of the transition mapping, i.e.

$$B_p \ni v \mapsto \tilde{v} = v - \mathrm{E}_q[v] \in B_q$$

is a top-linear isomorphism (Lang 1995, Chapter 3). It follows also directly from Proposition 5. Here and in the following the $\tilde{v}$ and the v differ only by a constant. The subspaces V_q^1 and V_q^2 defined as

$$V_q^i = \{\tilde{v}_i = v_i - \mathrm{E}_q[v_i] \colon v_i \in V_p^i\}, \qquad i = 1, 2,$$

split in B_q. In such a case, for all $w_1, w_2 \in B_p$ and $\tilde{w}_i = w_i - \mathrm{E}_q[w_i]$, $i = 1, 2$, by Proposition 16 (b), we have

$$\begin{aligned} \mathrm{D}^2 K_p(u)(w_1, w_2) &= \mathrm{cov}_q[w_1, w_2] \\ &= \mathrm{E}_q[(w_1 - \mathrm{E}_q[w_1])(w_2 - \mathrm{E}_q[w_2])] \\ &= \langle \tilde{w}_1, \tilde{w}_2 \rangle_q. \end{aligned} \tag{18}$$

As in the previous section the cumulant generating functional K_p can be considered as a function of two variables:

$$K_p(\cdot,\cdot)\colon V^1_p \times V^2_p \ni (w_1, w_2) \mapsto K_p(w_1, w_2) = K_p(w_1 + w_2).$$

We can take partial derivatives that we shall denote by ∂_1, ∂_2, ∂_{11}, ∂_{12}, ∂_{21}, ∂_{22} by composing with the relevant projections; see (17).

Definition 29 (Fisher information operator). *The value at $q = e_p(u)$ of the Hessian linear operator from B_p to B^*_p of K_p at u will be denoted by $I(p, q)$:*

$$\langle I(p, q)w, v\rangle_{*,p} = \mathrm{D}^2 K_p(u)(w, v) \quad \textit{with } w, v \in B_p.$$

and it is called the Fisher information operator.

Given a splitting of B_p, if $v = v_1 + v_2$ and $w = w_1 + w_2$, we consider the partitioned operators I_{ij}, with $i, j \in \{1, 2\}$, restricted from V^j_p to $(V^k_p)^\circ$, $k \neq i$, and such that

$$\langle I_{ij}(p, q)w_j, v_i\rangle_{*,p} = \langle I(p, q)w_j, v_i\rangle_{*,p}.$$

In matrix form,

$$I(p, q) = \begin{bmatrix} I_{11}(p, q) & I_{12}(p, q) \\ I_{21}(p, q) & I_{22}(p, q) \end{bmatrix}\colon B_p \simeq V^1_p \times V^2_p \to (V^1_p)^\circ \times (V^2_p)^\circ \simeq B^*_p.$$

From (18) we get, for $\tilde{w} = w - \mathrm{E}_q[w]$, $\tilde{v} = v - \mathrm{E}_q[v]$,

$$\mathrm{cov}_q[w, v] = \langle \tilde{w}, \tilde{v}\rangle_q = \langle I(p, q)w, v\rangle_{*,p}.$$

We now consider the problem of finding an orthogonal projection of B_q onto V^1_q.

The space B_q is a pre-Hilbert space for the scalar product $\langle ., .\rangle_q$ as in Definition 10, i.e. the space is in general not complete under the norm induced by the scalar product. The existence of the orthogonal projection of $w \in B_q$ onto V^1_q is not ensured but, if the projection $w_{|1}$ exists, then it is unique; moreover, given a splitting such that $w = w_1 + w_2$, then $w_{2|1}$ defined as

$$w_{2|1} = w_{|1} - w_1$$

is the orthogonal projection of w_2 onto V^1_q.

The following proposition shows how to characterize $w_{2|1}$ with the partitioned Fisher information operator.

Proposition 30. *If $w = w_1 + w_2$ and $\tilde{w}_i = w_i - \mathrm{E}_q[w_i]$, for $i = 1, 2$, then $\tilde{w} = \tilde{w}_1 + \tilde{w}_2$. If there exists $w_{2|1} \in V^1_p$ such that the* normal equation

$$I_{11}(p, q)w_{2|1} = I_{12}(p, q)w_2 \tag{19}$$

is satisfied, then $\tilde{w}_{2|1} = w_{2|1} - \mathrm{E}_q[w_{2|1}]$ is the orthogonal projection of $\tilde{w}_2 \in V^2_q$ *onto* V^1_q.

Proof. If (19) holds true, then for any $w_1 \in V^1_p$ we have, from (18) and using the Fisher information operator,

$$\begin{aligned}\langle \tilde{w}_2 - \tilde{w}_{2|1}, \tilde{w}_1\rangle_q &= \langle \tilde{w}_2, \tilde{w}_1\rangle_q - \langle \tilde{w}_{2|1}, \tilde{w}_1\rangle_q \\ &= \langle I_{12}(p, q)w_2, w_1\rangle_{*,p} - \langle I_{11}(p, q)w_{2|1}, w_1\rangle_{*,p} \\ &= 0.\end{aligned}$$

This ends the proof. □

Definition 31. *Let B_1, B_2 be Banach spaces, and let F be a C^k parametrization of A, an open subset of $\mathscr{M}(X, \mathscr{X}, \mu)$, on $B_1 \times B_2$:*

$$A \ni q \mapsto F(q) = (F_1(q), F_2(q)) \in B_1 \times B_2.$$

We consider the pair of regular curves

$$]a, b[\ni t \mapsto q_i(t) \in A, \qquad i = 1, 2,$$

such that $0 \in]a, b[$ *and*

$$q_1(0) = q_2(0) = q \quad \text{and} \quad F_i(q_j(t)) = \text{constant}, \qquad i \neq j,\ t \in]a, b[. \tag{20}$$

We shall say that the two parameters F_1 and F_2 are orthogonal *at $q \in A$ if each such pair of curves is orthogonal at q, i.e. $\dot{q}_1(0)$ is orthogonal to $\dot{q}_2(0)$. If this is true for all $q \in A$, then we shall say that* the two parameters are orthogonal.

Theorem 32 (Mixed parametrization). *Given a splitting V^1_p, V^2_p of B_p, the mapping*

$$F\colon \mathscr{U}_p \ni q \mapsto (\eta_1, u_2) \in (V^1_p)^\circ \times V^2_p$$

with $q = \mathrm{e}^{u - K_p(u)}\, p$, $u = u_1 + u_2$, $u_i \in V^i_p$, $i = 1, 2$, and

$$\eta_1 = \partial_1 K_p(u)$$

is an orthogonal C^∞ parametrization of $\mathscr{U}_p$.

Note that, from (17) *and* (7)*,*

$$\eta_1(v) = \langle \eta_1, v\rangle_{*,p} = \mathrm{D}K_p(u)\, P_1(v) = \mathrm{D}K_p(u)\, v_1 = \mathrm{E}_q[v_1];$$

so the present parametrization will be called mixed parametrization*, as in the parametric case (Barndorff-Nielsen and Cox 1989; 1994, p. 62).*

Proof. First we show that the mapping $q = \mathrm{e}^{u - K_p(u)}\, p \mapsto (\partial_1 K_p(u), u_2)$ is one to one. In (9), if $\overline{u} = \overline{u}_1 + \overline{u}_2$ and $\overline{u}_2 = u_2$, then $\overline{u} - u$ equals $\overline{u}_1 - u_1$ and, denoting $\mathrm{e}^{\theta u - K_p(\theta u)}\, p$ by p_θ, we have

$$\int_0^1 \mathrm{d}\theta \,\mathrm{var}_{p_\theta}(\overline{u} - u) = \{\mathrm{D}K_p(\overline{u}) - \mathrm{D}K_p(u)\}(\overline{u}_1 - u_1)$$

$$= \{\partial_1 K_p(\overline{u}) - \partial_1 K_p(u)\}(\overline{u} - u)$$

$$= \langle \overline{\eta}_1 - \eta_1, \overline{u} - u \rangle_{*,p}.$$

If $\partial_1 K_p(\overline{u}) = \partial_1 K_p(u)$ and $\overline{u}_2 = u_2$, then $\overline{u} - u =$ constant and $\overline{u}_1 = u_1$, because $\overline{u}$ and u are centred random variables.

Let $u^{(i)}$, $i = 1, 2$, be the representation in the chart s_p of the regular curves through q in Definition 31, corresponding to the mixed parameter (η_1, u_2):

$$q_i(t) = \mathrm{e}^{u^{(i)}(t) - K_p(u^{(i)}(t))}\, p, \quad i = 1, 2.$$

Because of the second condition of (20) the first curve leaves the second parameter constant, $u_2^{(1)}(t) = u_2$, and the second curve leaves the first parameter constant, i.e., for any $w \in B_p$, $\partial_1 K_p(u^{(2)}(t))w =$ constant. If we take the derivative of the last equation with respect to t we get

$$\mathrm{D}^2 K_p(u^{(2)}(t))(P_1(w), \dot{u}^{(2)}(t)) = 0. \tag{21}$$

The condition on the first curve can be written as $P_2(u^{(1)}(t)) = u_2$ and derivation gives $P_2(\dot{u}^{(1)}(t)) = 0$. Consequently, $\dot{u}^{(1)}(t) = P_1(\dot{u}^{(1)}(t))$ and substituting w in (21) with $\dot{u}^{(1)}(t)$ we get

$$\mathrm{D}^2 K_p(u^{(2)}(t))(\dot{u}^{(1)}(t), \dot{u}^{(2)}(t)) = 0.$$

We denote by $\tilde{u}^{(i)}(t)$ the coordinate of $q_i(t)$ with respect to the chart s_q:

$$\tilde{u}^{(i)}(t) = u^{(i)}(t) - u^{(i)}(0) - \mathrm{E}_q[u^{(i)}(t) - u^{(i)}(0)];$$

then $\dot{\tilde{u}}^{(i)}(t) = \dot{u}^{(i)}(t) - \mathrm{E}_q[\dot{u}^{(i)}(t)]$. If $t = 0$, it follows from (17) and Definition 29 that

$$\langle \dot{\tilde{u}}^{(1)}(0), \dot{\tilde{u}}^{(2)}(0) \rangle_q = \langle I(p, q)\dot{u}^{(1)}(0), \dot{u}^{(2)}(0) \rangle_{*,p}$$

$$= \mathrm{D}^2 K_p(u^{(2)}(0))(\dot{u}^{(1)}(0), \dot{u}^{(2)}(0))$$

$$= 0.$$

So $q_1(t)$ and $q_2(t)$ are orthogonal at q. □

7. Transformation of the sample space

The problem that we consider in this section is the action of a transformation of the sample space on the manifold structure. Let us first introduce some notation. Let a measurable mapping on the sample space be given:

$$\varphi\colon (X, \mathscr{X}, \mu) \to (Y, \mathscr{Y}, \nu), \tag{22}$$

i.e.

$$\varphi\colon X \to Y, \quad \varphi^{-1}\colon \mathscr{Y} \to \mathscr{X}, \quad \nu = \mu \circ \varphi^{-1}.$$

If $p \in \mathscr{M}(X, \mathscr{X}, \mu)$, then the image under φ of the probability measure $p \cdot \mu$ is characterized by

$$(p \cdot \mu) \circ \varphi^{-1}(B) = \int_{\varphi^{-1}(B)} p \, \mathrm{d}\mu = \int_{\varphi^{-1}(B)} \mathrm{E}_\mu[p|\varphi^{-1}(\mathscr{Y})] \, \mathrm{d}\mu, \quad B \in \mathscr{Y}.$$

By the Doob lemma,

$$\mathrm{E}_\mu[p|\varphi^{-1}(\mathscr{Y})] = \hat{p} \circ \varphi, \tag{23}$$

where $\hat{p}$ is defined up to sets of ν-measure 0. We shall write $\hat{p} = \mathrm{E}_\nu[p|\varphi]$, so that

$$(p \cdot \mu) \circ \varphi^{-1}(B) = \int_{\varphi^{-1}(B)} \hat{p} \circ \varphi \, \mathrm{d}\mu = \int_B \mathrm{E}_\nu[p|\varphi] \, \mathrm{d}\nu, \quad B \in \mathscr{Y},$$

and the sample space transformation φ induces a transformation $\hat{\varphi}$ on $\mathscr{M}(X, \mathscr{X}, \mu)$ given by

$$\hat{\varphi}\colon \mathscr{M}(X, \mathscr{X}, \mu) \ni p \mapsto \hat{p} = \mathrm{E}_\nu[p|\varphi] \in \mathscr{M}(Y, \mathscr{Y}, \nu).$$

7.1. Transformation

The basic result about the kind of smoothness of the mapping $\hat{\varphi}\colon p \mapsto \hat{p}$ that we are able to prove is shown by the following Proposition 33. We refer to the book by Lang (1995) for an introduction to the basic notions regarding the differentiability and regularity of mappings between differentiable manifolds. In particular we shall mention the notion of *submersion*. This notion is connected with the stability of the rank of a mapping. Note that there is misuse of the language because the derivative mapping of the transformation is a linear continuous mapping between the tangent spaces, but we are able to prove the continuous differentiability in a weaker sense only.

The given regularity results are not connected in any way with the regularity of the mapping φ itself, nor with any algebraic or topological property of the spaces X or Y. The possible relations between the differentiable structure of the exponential statistical manifold and the regularity of the densities themselves are not dealt with at all in the present paper and suggest different research work. A special *ad hoc* condition is introduced in (27) to prove the Gateaux regularity of the mapping between the exponential statistical manifolds. We do not discuss examples of application here. We just mention the fact that this condition is easily shown to be true in trivial cases, such as constant or injective transformation, or Y with a finite number of atoms.

Differentiability of higher order could be proved under the analogous conditions. We shall not discuss this topic further as we are not going to use it.

Proposition 33 (Sample space transformation). *Let us consider a space transformation φ as in* (22) *and let*

$$\hat{\varphi}\colon \mathscr{M}(X, \mathscr{X}, \mu) \ni p \mapsto \hat{\varphi}(p) = \hat{p} \in \mathscr{M}(Y, \mathscr{Y}, \nu)$$

denote the action of φ on the exponential statistical manifolds. Then we have the following.

(a) *$\hat{\varphi}$ is onto and it is injective if and only if the mapping φ is μ-almost surely injective, i.e. it generates the σ-algebra $\mathscr{X}$ up to μ-null sets.*

(b) *The coordinate form*

$$\hat{\varphi}_p = s_{\hat{p}} \circ \hat{\varphi} \circ e_p, \quad \hat{p} = \hat{\varphi}(p),$$

of the mapping $\hat{\varphi}$ around the point p is given by

$$\hat{\varphi}_p(u) = K_{\hat{p}}(u|\varphi) - \mathrm{E}_{\hat{p}}[K_{\hat{p}}(u|\varphi)], \tag{24}$$

where

$$K_{\hat{p}}(u|\varphi) = \log \mathrm{E}_{\hat{p}}[e^u|\varphi] \tag{25}$$

in a neighbourhood of 0. Such a mapping is continuous from an open neighbourhood of 0 of B_p to $B_{\hat{p}}$; so the mapping $\hat{\varphi}$ is of class C^0 from the manifold $\mathscr{M}(X, \mathscr{X}, \mu)$ to the manifold $\mathscr{M}(Y, \mathscr{Y}, \nu)$. Moreover for all $\alpha \geqslant 1$ the mapping $\hat{\varphi}_p$ is continuously differentiable as a mapping in $L^\alpha(Y, \mathscr{Y}, \hat{p} \cdot \nu)$ and its derivative at $u \in \mathscr{V}_p$ in the direction $v \in B_p$ is given by

$$\mathrm{D}\hat{\varphi}_p(u)v = \mathrm{E}_{\hat{q}}[v|\varphi] - \mathrm{E}_{\hat{p}}[\mathrm{E}_{\hat{q}}][v|\varphi]], \quad q = e_p(u). \tag{26}$$

(c) *The derivative mapping in* (26) *is a linear continuous operator from B_p to $B_{\hat{p}}$. If moreover $u \in B_p$ is such that, for all $\alpha \geqslant 1$,*

$$\mathrm{E}_{\hat{p}}[\mathrm{e}^{u - \mathrm{E}_{\hat{p}}[u|\varphi]}|\varphi] \in L^\alpha(Y, \mathscr{Y}, \hat{p} \cdot \nu), \tag{27}$$

then the mapping $\hat{\varphi}_p$ is Gateaux differentiable from B_p to $B_{\hat{p}}$ in the direction u.

(d) *$u \mapsto \mathrm{D}\hat{\varphi}_p(u)$ is surjective and its kernel splits.*

Proof.

(a) The surjectivity follows from the condition formula (23); in fact for each density $\tilde{p} \in \mathscr{M}(Y, \mathscr{Y}, \nu)$, the density $\tilde{p} \circ \varphi$ in $\mathscr{M}(X, \mathscr{X}, \mu)$ and it is such that $\hat{\varphi}(\tilde{p} \circ \varphi) = \tilde{p}$. If $\varphi^{-1}(\mathscr{Y}) = \mathscr{X}$ μ-almost surely, then using (23)

$$\hat{p} \circ \varphi = \mathrm{E}_\mu[p|\varphi^{-1}(\mathscr{Y})] = p.$$

If $\hat{\varphi}(p_1) = \hat{\varphi}(p_2) = \hat{p}$, then $p_1 = p_2 = \hat{p} \circ \varphi$. On the other hand, if $\varphi^{-1}(\mathscr{Y}) \subset \mathscr{X}$ is a proper sub-σ-algebra, then there exists a density p which is not $\varphi^{-1}(\mathscr{Y})$ measurable, so that p and $\hat{p} \circ \varphi$ are different and have the same image $\hat{p}$.

(b) Let us compute the coordinate form of the mapping $\hat{\varphi}$. Consider a density p, the chart domain $\mathscr{U}_p$, and let q be a density representable in such a chart, $q = \mathrm{e}^{u - K_p(u)} p$, where $\|u\|_p < 1$. Then $\hat{\varphi}(q) = \hat{q}$ is given by

$$\hat{q} \circ \varphi = \mathrm{E}_{\mu}[\mathrm{e}^{u - K_p(u)}\, p|\varphi^{-1}(\mathscr{Y})].$$

Now we compute the likelihood with respect to $\hat{p}$:

$$\begin{aligned}\left(\frac{\hat{q}}{\hat{p}}\right) \circ \varphi &= \frac{\hat{q} \circ \varphi}{\hat{p} \circ \varphi} \\ &= \frac{\mathrm{E}_{\mu}[\mathrm{e}^{u - K_p(u)}\, p|\varphi^{-1}(\mathscr{Y})]}{\mathrm{E}_{\mu}[p|\varphi^{-1}(\mathscr{Y})]} \\ &= \mathrm{E}_p[\mathrm{e}^{u - K_p(u)}|\varphi^{-1}(\mathscr{Y})] \\ &= \mathrm{e}^{-K_p(u)}\, \mathrm{E}_p[\mathrm{e}^u|\varphi^{-1}(\mathscr{Y})].\end{aligned}$$

The previous formula simplifies to

$$\frac{\hat{q}}{\hat{p}} = \mathrm{e}^{-K_p(u)}\, \mathrm{E}_{\hat{p}}[\mathrm{e}^u|\varphi],$$

and we can write

$$\hat{q} = \hat{\varphi}(q) = \hat{\varphi} \circ e_p(u) = \mathrm{e}^{-K_p(u)}\, \mathrm{E}_{\hat{p}}[\mathrm{e}^u|\varphi]\, \hat{p}.$$

We have to find the coordinates of $\hat{q}$ with respect to the chart on $\mathscr{U}_{\hat{p}}$. If $\hat{q} \in \mathscr{U}_{\hat{p}}$, then

$$\hat{u} = s_{\hat{p}}(\hat{q}) = \log\left(\frac{\hat{q}}{\hat{p}}\right) - \mathrm{E}_{\hat{p}}\left[\log\left(\frac{\hat{q}}{\hat{p}}\right)\right] = \log \mathrm{E}_{\hat{p}}[\mathrm{e}^u|\varphi] - \mathrm{E}_{\hat{p}}[\log \mathrm{E}_{\hat{p}}[\mathrm{e}^u|\varphi]]. \tag{28}$$

Using the notation in (25), we get (24).

We shall show that $\hat{u}$ in (28) actually belongs to $\mathscr{U}_{\hat{p}}$ if u belongs to a suitable neighbourhood of zero. We start by showing that the mapping

$$\mathscr{V}_p \ni u \mapsto K_{\hat{p}}(u|\varphi) \in L^{\phi_1}(\hat{p} \cdot \nu) \tag{29}$$

is defined and norm decreasing. In fact, using the definition of norm in (3), we take α such that $\alpha\|u\|_p < 1$ and $\alpha = 1/r > 1$. From the convexity of the function

$$f(x) = \tfrac{1}{2}(x^{\alpha} + x^{-\alpha})$$

for $\alpha > 1$, and the Jensen inequality, it follows that the norm

$$\|K_{\hat{p}}(u|\varphi)\|_{\hat{p}} = \|\log \mathrm{E}_{\hat{p}}[\mathrm{e}^u|\varphi]\|_{\hat{p}}$$

can be bounded above. In fact,

$$\begin{aligned}\cosh(\alpha \log \mathrm{E}_{\hat{p}}[\mathrm{e}^u|\varphi]) - 1 &= \tfrac{1}{2}(\mathrm{E}_{\hat{p}}[\mathrm{e}^u|\varphi]^{\alpha} + \mathrm{E}_{\hat{p}}[\mathrm{e}^u|\varphi]^{-\alpha}) - 1 \\ &\leqslant \mathrm{E}_{\hat{p}}[\cosh(\alpha u) - 1|\varphi]\end{aligned}$$

and, taking the expected values, as $1 > 1/\alpha = r > \|u\|_p$,

$$\mathrm{E}_{\hat{p}}[\cosh(\alpha \log \mathrm{E}_{\hat{p}}[\mathrm{e}^u|\varphi]) - 1] \leqslant \mathrm{E}_p[\cosh(\alpha u) - 1] \leqslant 1. \tag{30}$$

As $\|u\|_p < 1$ and from (30), it follows that

$$\|K_{\hat{p}}(u|\varphi)\|_{\hat{p}} \leqslant \|u\|_p.$$

As the expectation is a contraction operator on the space of $L^{\phi_1}(\hat{p} \cdot \nu)$, then $\hat{u}$ in (28) belongs to the domain $\mathscr{U}_{\hat{p}}$ of the chart at $\hat{p}$ for u in a suitable neighbourhood of p.

To prove the continuity of $\hat{\varphi}_p$, consider the difference of the values of the function $K_{\hat{p}}(\cdot|\varphi)$ defined in (25) at two points u and v of the domain of $\hat{\varphi}_p$:

$$\begin{aligned} K_{\hat{p}}(v|\varphi) - K_{\hat{p}}(u|\varphi) &= \log \mathrm{E}_{\hat{p}}[\mathrm{e}^v|\varphi] - \log \mathrm{E}_{\hat{p}}[\mathrm{e}^u|\varphi] \\ &= \log\left(\frac{\mathrm{E}_{\hat{p}}[\mathrm{e}^v|\varphi]}{\mathrm{E}_{\hat{p}}[\mathrm{e}^u|\varphi]}\right) \\ &= \log \mathrm{E}_{\hat{q}}[\mathrm{e}^{v-u}|\varphi], \end{aligned}$$

where $q = e_p(u)$. The norm-decreasing property shown above, together with the equivalence of $B_{\hat{q}}$ and $B_{\hat{p}\cdot\nu}$ (see Proposition 5), now implies the continuity of $K_{\hat{p}}(\cdot|\varphi)$ and in turn the continuity of $\hat{\varphi}_p$. Finally $\hat{\varphi}$ is continuous because its chart representation is continuous.

For the differentiability the same argument as above shows that it is enough to show the differentiability of the mapping $u \mapsto K_{\hat{p}}(u|\varphi)$ defined in (25).

Let us check first the directional derivative. For $u,\ v \in B_p,\ u \in \mathscr{U}_p,\ q = \mathrm{e}^{u - K_p(u)}\mu$, $\hat{q} = \hat{\varphi}(q)$, and $t \to 0$, we want $R(u, v, t)$ to go to 0, where

$$\begin{aligned} R(u, v, t) &= t^{-1}(K_{\hat{p}}(u + tv|\varphi) - K_{\hat{p}}(u|\varphi)) - \mathrm{E}_{\hat{q}}[v|\varphi] \\ &= t^{-1}K_{\hat{q}}(tu|\varphi) - \mathrm{E}_{\hat{q}}[v|\varphi] \\ &= t^{-1}K_{\hat{q}}(t(v - \mathrm{E}_{\hat{q}}[v|\varphi])|\varphi). \end{aligned}$$

From the concavity of the log function we get

$$\begin{aligned} R(u, v, t) &= t^{-1} \log \mathrm{E}_{\hat{q}}[\mathrm{e}^{t(v - \mathrm{E}_{\hat{q}}[v|\varphi])}|\varphi] \\ &\geqslant t^{-1}\mathrm{E}_{\hat{q}}[\log \mathrm{e}^{t(v - \mathrm{E}_{\hat{q}\cdot\nu}[v|\varphi])}|\varphi] \\ &= \mathrm{E}_{\hat{q}}[v - \mathrm{E}_{\hat{q}}[v|\varphi]|\varphi] \\ &= 0. \end{aligned}$$

We use now a classical argument from Hardy *et al*. (1952). When t decreases to 0, the function

$$t \mapsto \mathrm{E}_{\hat{q}}[\mathrm{e}^{t(v - \mathrm{E}_{\hat{q}}[v|\varphi])}|\varphi]^{1/t}$$

is decreasing and the same is true for $t \mapsto R(u, v, t)$. From the inequality $\log x \leqslant x - 1$ we get

$$R(u, v, t) \leqslant t^{-1}(\mathrm{E}_{\hat{q}}[\mathrm{e}^{t(v-\mathrm{E}_{\hat{q}}[v|\varphi])}|\varphi] - 1)$$
$$= \mathrm{E}_{\hat{q}}\left[\left(\frac{\mathrm{e}^{t(v-\mathrm{E}_{\hat{q}\cdot v}[v|\varphi])} - 1}{t}\right)\middle|\varphi\right]. \tag{31}$$

The right-hand side of (31) goes to 0 as t goes to 0, and it is eventually bounded in all L^α. The same is true for negative t, as can be shown by changing v to $-v$. This and the addition of the expected value show directional (Gateaux) differentiability at any point of the domain of $\hat{\varphi}_p$ and the form of the derivative.

Finally we consider the continuity of the derivative. We consider the difference

$$\mathrm{E}_{\hat{q}}[v|\varphi] - \mathrm{E}_{\hat{p}}[v|\varphi] = \mathrm{E}_{\hat{p}}[v(\mathrm{e}^{u-K_{\hat{p}\cdot v}(u|\varphi)} - 1)|\varphi].$$

The L^α norm of this difference is bounded by the L^α norm of

$$v(\mathrm{e}^{u-K_p(u|\varphi^{-1}(\mathscr{Y}))} - 1),$$

where $K_p(u|\varphi^{-1}(\mathscr{Y})) = \log \mathrm{E}_p[\mathrm{e}^u|\varphi^{-1}(\mathscr{Y})]$. Now

$$\|v\mathrm{e}^{u-K_p(u|\varphi^{-1}(\mathscr{Y}))}\|^\alpha_{L^\alpha(p)} = \mathrm{E}_p[|v|^\alpha\, \mathrm{e}^{\alpha(u-K_p(u|\varphi^{-1}(\mathscr{Y})))}]$$
$$\leqslant \mathrm{E}_p[|v|^{2\alpha}]^{1/2}\, \mathrm{E}_p[\mathrm{e}^{2\alpha(u-K_p(u|\varphi^{-1}(\mathscr{Y})))}]^{1/2}. \tag{32}$$

However, $K_p(.|\varphi^{-1}(\mathscr{Y}))$ is a contraction; then on the open set of $\mathscr{V}_p$ where $2\alpha\|(u - K_p(u|\varphi^{-1}(\mathscr{Y}))\|_p < 1$ the second factor of (32) is bounded by 2. The first factor is bounded by a constant times $\|v\|^\alpha_p$.

(c) The linear mapping

$$v \mapsto \mathrm{E}_{\hat{q}\cdot v}[v|\varphi]$$

is a contraction from $L^{\phi_1}(q)$ to $L^{\phi_1}(\hat{q})$, because it is a conditional expectation. Then it is continuous from B_p to $B_{\hat{p}}$ because of the equivalence of the B spaces.

Assume now the condition (27). Then, for all $\alpha > 1$,

$$\cosh(\alpha R(u, v, t) - 1) = \tfrac{1}{2}(\mathrm{E}_{\hat{q}\cdot v}[\mathrm{e}^{t(v-\mathrm{E}_{\hat{q}}[v|\varphi])}|\varphi]^{\alpha/t} + \mathrm{E}_{\hat{q}}[\mathrm{e}^{t(v-\mathrm{E}_{\hat{q}}[v|\varphi])}|\varphi]^{-\alpha/t}) - 1.$$

As t decreases to 0, this quantity has been shown to decrease to zero. If the expected value at $t = 1$ is finite, then the bounded convergence theorem implies the convergence to zero of expected values. This is precisely what is implied by the condition (27).

(d) We have shown that the derivative of the local coordinate form of the transformation $\hat{\varphi}_p$ at $u \in \mathscr{V}_p$ in the direction $v \in B_p$ is given by (26). This linear mapping in surjective; in fact, for each $\hat{v} \in B_{\hat{p}}$ we have $v = \hat{v} \circ \varphi \in B_p$ and

$$\begin{aligned}\mathrm{D}\hat{\varphi}_p(u)\,v &= \mathrm{D}\hat{\varphi}_p(u)\,v \\ &= \mathrm{E}_{\hat{q}\cdot\nu}[\hat{v}\circ\varphi|\varphi] - \mathrm{E}_{\hat{p}}[\mathrm{E}_{\hat{q}\cdot\nu}[\hat{v}\circ\varphi|\varphi]] \\ &= \hat{v} - \mathrm{E}_p[v] \\ &= \hat{v}.\end{aligned}$$

Let us now study the kernel of $\mathrm{D}\hat{\varphi}_p(u)$. This linear mapping can be written as $\mathrm{D}\hat{\varphi}_p(u) = C\circ E(u)$, where

$$E = \mathrm{E}_{\hat{q}}[\cdot|\varphi]\colon B_p \to L^{\phi_1}(\hat{p}\cdot\nu)$$

and

$$C\colon L^{\phi_1}(\hat{p}\cdot\nu) \ni \hat{w} \mapsto \hat{w} - \mathrm{E}_{\hat{p}}[\hat{w}].$$

The kernel of C is the set of constant random variables in $L^{\phi_1}(\hat{p}\cdot\nu)$. The constant elements $v\in B_p$, such that this constant is 0, and we have shown that

$$\ker \mathrm{D}\hat{\varphi}_p(u) = \{v\in B_p\colon \mathrm{E}_p[v|\varphi^{-1}(\mathscr{Y})] = 0\}.$$

As we have shown in the previous paragraphs (see the second example of submanifold in Section 5.1), the kernel of the conditional expectation splits. □

Remark 34. If the mapping $\hat{\varphi}$ is of class C^n from $\mathscr{M}(X,\mathscr{X},\mu)$ to $\mathscr{M}(Y,\mathscr{Y},\nu)$, $n\geqslant 1$, then the previous splitting property implies that this mapping is a *submersion*, and, for all $\hat{p}\in\mathscr{M}(Y,\mathscr{Y},\nu)$, $\hat{\varphi}^{-1}(\hat{q})$ is a submanifold of $\mathscr{M}(X,\mathscr{X},\mu)$ (Lang 1995, p. 25). This general result cannot be applied directly, because we have not proved in general the C^1 regularity. Nevertheless it can be applied to special submanifolds. If we assume that $\mathscr{N}$ is a submanifold of $\mathscr{M}(X,\mathscr{X},\mu)$ such that the restriction of $\hat{\varphi}$ to $\mathscr{N}$ is C^1, and the splitting property holds on the submanifold, then the splitting result implies that $\hat{\varphi}^{-1}(\hat{q})\cap\mathscr{N}$ is a submanifold of $\mathscr{M}(X,\mathscr{X},\mu)$.

The following sections show a different direct approach to the differential structure of $\hat{\varphi}^{-1}(\hat{q})$, based on the idea of regular parametrization.

7.2. Bivariate densities with one marginal given

Let us assume in the remaining part of this section that the sample space $(X,\mathscr{X},\mu)$ is a product space $X = Y\times Z$, $\mathscr{X} = \mathscr{Y}\otimes\mathscr{Z}$, $\mu = \nu\otimes\pi$. Let us denote by $\varphi_1\colon Y\times Z\to Y$ and $\varphi_2\colon Y\times Z\to Z$ the two marginal mappings.

Proposition 35 (One marginal given).

(a) *Let $p_1 = \hat{\varphi}_1(p)$ be given. For each $q\in\mathscr{U}_p$, $q = \mathrm{e}^{u-K_p(u)}\,p$, let $q_1 = \hat{\varphi}_1(q)$. Then we have the following.*

(i) *The mapping $\mathscr{U}_p\ni q\mapsto q_1/p_1 - 1\in {}^*B_{p_1}$ is C^∞ from $\mathscr{M}(X,\mathscr{X},\mu)$ to $B^*_{p_1}$.*

(ii) *The mapping*

$$\mathscr{U}_p \ni q \mapsto \left(\frac{q_1}{p_1} - 1,\ u - \mathrm{E}_p[u|\varphi_1^{-1}(\mathscr{Y})]\right) \in {}^* B_{p_1} \times \ker \mathrm{E}_p[.|\varphi_1^{-1}(\mathscr{Y})]$$

is an orthogonal parametrization of $\mathscr{U}_p$ *on* $B^*_{p_1} \times \ker \mathrm{E}_p[.|\varphi_1^{-1}(\mathscr{Y})]$.

(b) *The set of all densities whose first marginal is* p_1, *i.e.*

$$\mathscr{N}_{p_1} = \{p \in \mathscr{M}(X, \mathscr{X}, \mu)\colon \hat{\varphi}_1(p) = p_1\},$$

has a C^∞ *parametrization on* $\mathscr{U}_p$ *given by*

$$\mathscr{N}_{p_1} \ni p \mapsto u - \mathrm{E}_p[u|\varphi_1^{-1}(\mathscr{Y})] \in \ker \mathrm{E}_p[.|\varphi_1^{-1}(\mathscr{Y})].$$

Proof.

(a) It is a particular case of Proposition 32 based on the splitting

$$B_p \ni u \leftrightarrow \mathrm{E}_p[u|\varphi_1^{-1}(\mathscr{Y})] + (u - \mathrm{E}_p[u|\varphi_1^{-1}(\mathscr{Y})]).$$

To identify the first component η_1 of the orthogonal parametrization we compute, as in Proposition 32, the partial derivative with respect to the first component of the splitting. We have

$$\begin{aligned}
\langle \eta_1, v\rangle_{*,p} &= \partial_1 K_p(u)v \\
&= \mathrm{D}K_p(u)\, P_1(v) \\
&= \mathrm{D}K_p(u)\, \mathrm{E}_p[v|\varphi_1^{-1}(\mathscr{Y})] \\
&= \mathrm{E}_q[\mathrm{E}_p[v|\varphi_1^{-1}(\mathscr{Y})]] \\
&= \mathrm{E}_q[\mathrm{E}_{p_1}[v|\varphi_1^{-1}(\mathscr{Y})] \circ \varphi_1] \\
&= \mathrm{E}_{q_1}[\mathrm{E}_{p_1}[v|\varphi_1]] \\
&= \mathrm{E}_{p_1}\left[\frac{q_1}{p_1}\mathrm{E}_{p_1}[v|\varphi_1]\right] \\
&= \mathrm{E}_p\left[\frac{q_1}{p_1} \circ \varphi_1 \mathrm{E}_p[v|\varphi_1^{-1}(\mathscr{Y})]\right] \\
&= \mathrm{E}_p\left[\frac{q_1}{p_1} \circ \varphi_1\, v\right] \\
&= \mathrm{E}_p\left[\left(\frac{q_1}{p_1} - 1\right) \circ \varphi_1\, v\right].
\end{aligned}$$

We have characterized η_1 as $(q_1/p_1 - 1) \circ \varphi_1$.

(b) $q \in \mathscr{N}_{p_1} \cap \mathscr{U}_p$ if and only if $q_1/p_1 - 1 = 0$. □

Note that there is a one-to-one mapping between the bivariate densities with a given

marginal and the set of conditional densities, i.e. the Bayes formula. This induces a parametrization on the set of the corresponding conditional densities $\{p/p_1 \colon p \in \mathscr{N}_{p_1}\}$.

7.3. Bivariate densities with two given marginals

Definition 36 (Fréchet class). *Using the previous notation, the set of all densities whose marginals are given, i.e.*

$$\mathscr{N}_{p_1 p_2} = \{p \in \mathscr{M}(X, \mathscr{X}, \mu) \colon \hat{\varphi}_i(p) = p_i,\ i = 1, 2\} \tag{33}$$

is called the Fréchet class *with marginals* p_1, p_2.

The following proposition shows that each Fréchet class has a regular parametrization and gives some details of its structure.

Proposition 37 (Fréchet manifold). *Let two marginal densities* p_1, p_2 *be given: we consider the Fréchet class* $\mathscr{N}_{p_1 p_2}$ *as in* (33).

(a) *Let* $p = p_1 p_2$ *be the product density, and let* $u \in B_p$ *be given. The formulae*

$$u_1 = \int u p_2 \, \mathrm{d}\pi,$$

$$u_2 = \int u p_1 \, \mathrm{d}\nu,$$

$$u_{12} = u - u_1 - u_2,$$

define a splitting of B_p *into the spaces* $(B_{p_1} \times B_{p_2})$ *and* B_p^{12}, *where*

$$B_p^{12} = \left\{ w \in B_p \colon \int w p_2 \, \mathrm{d}\pi = \int w p_1 \, \mathrm{d}\nu = 0 \right\}.$$

(b) *The mapping*

$$\mathscr{U}_p \ni q \mapsto \left(\left(\frac{q_1}{p_1} - 1, \frac{q_2}{p_2} - 1 \right), u_{12} \right) \in ({}^*B_{p_1} \times {}^*B_{p_2}) \times B_p^{12} \tag{34}$$

is an orthogonal parametrization of $\mathscr{U}_p$ *on* $(B_{p_1}^* \times B_{p_2}^*) \times B_p^{12}$.

(c) $\mathscr{N}_{p_1 p_2} \cap \mathscr{U}_p$ *has a* C^∞ *parametrization given by* $q \mapsto u_{12} \in B_p^{12}$.

Proof. It is an application of the Proposition 32 based on the following projections:

$$B_p \ni u \mapsto \left(\int u(\cdot, z) p_2(z) \, \mathrm{d}\pi \, \mathrm{d}z, \int u(y, \cdot) p_1(y) \nu \, \mathrm{d}y, u - \int u(\cdot, z) p_2(z) \pi \, \mathrm{d}z \right.$$

$$\left. + \int u(y, \cdot) p_1(y) \nu \, \mathrm{d}y \right).$$

□

Note that the term u_{12} has the character of an interaction term.

7.4. An example

Rogantin (1996) developed in detail the application of the previous result to the case of finite spaces and to Markov chains. In the present paper we give a very simple example aimed at showing how each of the abstract results in the previous propositions correspond to the usual objects of the statistical analysis of statistical dependence (Cox and Reid 1987).

Let us consider the sample space $\{0, 1\}^2$; let us take as a reference measure the uniform distribution, that is $\mu_{ij} = \frac{1}{4}$ for $i, j \in \{0, 1\}$, and consider a neighbourhood of the unit density, i.e. $p_{ij} = 1$ for $i, j \in \{0, 1\}$. The s_p coordinate of the generic density q_{ij}, with $i, j \in \{0, 1\}$, is given in matrix form by

$$u = \begin{bmatrix} \log q_{00} - \frac{1}{4}\sum_{ij} \log q_{ij} & \log q_{01} - \frac{1}{4}\sum_{ij} \log q_{ij} \\ \log q_{10} - \frac{1}{4}\sum_{ij} \log q_{ij} & \log q_{11} - \frac{1}{4}\sum_{ij} \log q_{ij} \end{bmatrix}$$

$$= \frac{1}{4}\begin{bmatrix} 3\log q_{00} - \log q_{01} - \log q_{10} - \log q_{11} & -\log q_{00} + 3\log q_{01} - \log q_{10} - \log q_{11} \\ -\log q_{00} - \log q_{01} + 3\log q_{10} - \log q_{11} & -\log q_{00} - \log q_{01} - \log q_{10} + 3\log q_{11} \end{bmatrix}.$$

We show in this example Proposition 34; the first marginal is given, i.e. $\varphi_1(i, j) = i$. The conditional mean value with respect to the first margin is

$$\mathrm{E}_p[u|\varphi_1] \circ \varphi_1 = \frac{1}{4}\begin{bmatrix} \log q_{00} + \log q_{01} - \log q_{10} - \log q_{11} & \log q_{00} + \log q_{01} - \log q_{10} - \log q_{11} \\ \log q_{10} + \log q_{11} - \log q_{00} - \log q_{01} & \log q_{10} + \log q_{11} - \log q_{00} - \log q_{01} \end{bmatrix}$$

$$= \frac{1}{4}\log\left(\frac{q_{00}q_{01}}{q_{10}q_{11}}\right)\begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix}$$

and

$$u - \mathrm{E}_p[u|\varphi_1] \circ \varphi_1 = \frac{1}{2}\begin{bmatrix} \log\left(\frac{q_{01}}{q_{00}}\right) \\ \log\left(\frac{q_{10}}{q_{11}}\right) \end{bmatrix}\begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}.$$

If we introduce the transition probabilities $\alpha = q_{01}/(q_{00} + q_{01})$ and $\beta = q_{10}/(q_{10} + q_{11})$, then we find that

$$u - \mathrm{E}_p[u|\varphi_1] \circ \varphi_1 = \frac{1}{2}\begin{bmatrix} \log\left(\frac{\alpha}{1-\alpha}\right) \\ \log\left(\frac{\beta}{1-\beta}\right) \end{bmatrix}\begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}.$$

Thus this two-dimensional parameter coincides with the log-odds of the transition probabilities and it is orthogonal to mean parameter $q_1 - 1$, where

$$q_1 = \mathrm{E}_\mu[q|\varphi_1] = \frac{1}{2}\begin{bmatrix} 4-(q_{10}+q_{11}) \\ q_{10}+q_{11} \end{bmatrix}.$$

Now we show Proposition 36; the marginals are given, i.e. $\varphi_1(i,j) = i$ and $\varphi_2(i,j) = j$ and the splitting is $u = u_1 + u_2 + u_{12}$, where

$$u_1 = \mathrm{E}_p[u|\varphi_1] \circ \varphi_1 = \frac{1}{4}\log\left(\frac{q_{00}q_{01}}{q_{10}q_{11}}\right)\begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix},$$

$$u_2 = \mathrm{E}_p[u|\varphi_2] \circ \varphi_2 = \frac{1}{4}\log\left(\frac{q_{00}q_{10}}{q_{01}q_{11}}\right)\begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix};$$

so u_{12} is given by

$$u_{12} = u - u_1 - u_2 = \frac{1}{4}\log\left(\frac{q_{01}q_{10}}{q_{00}q_{11}}\right)\begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}.$$

In terms of transition probabilities,

$$\log\left(\frac{q_{01}q_{10}}{q_{00}q_{11}}\right) = \log\left(\frac{\alpha}{1-\alpha}\frac{\beta}{1-\beta}\right).$$

Then this parameter is orthogonal to the mean parameters $q_1 - 1$ and $q_2 - 1$, where q_1 is as above and

$$q_2 = \mathrm{E}_\mu[q|\varphi_2] = \tfrac{1}{2}[4-(q_{01}+q_{11}) \quad q_{01}+q_{11}].$$

Acknowledgements

The present version of this paper has been greatly improved over the previous versions by working out many very useful constructive comments and meticulous remarks by the anonymous referees of the journal. The authors want to express to them and to the Editor their warmest thanks.

Some parts of the paper have been discussed with Damiano Brigo and Paolo Gibilisco who are currently working on further development of the theory of exponential statistical manifold (Brigo and Pistone 1996; Gibilisco and Pistone 1998). Their comments have contributed to the clarification of the presentation in many points.

References

Amari, S.-I. (1982) Differential geometry of curved exponential families–curvature and information loss. *Ann. Statist.*, **18**, 357–385.

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

Amari, S.-I., Barndorff-Nielsen, O.E., Kass, R., Lauritzen, S.L. and Rao, C.R. (1987) *Differential Geometry and Statistical Inference*. Hayward, CA: Institute of Mathematical Statistics.

Barndorff-Nielsen, O.E. (1978a) *Information and Exponential Families in Statistical Theory.* New York: Wiley.

Barndorff-Nielsen, O.E. (1978b) *Parametric Statistical Models and Likelihood*. Lecture Notes Statist., **50**. Berlin: Springer-Verlag.

Barndorff-Nielsen, O.E. and Cox, D.R. (1989) *Asymptotic Techniques for use in Statistics*. London: Chapman & Hall.

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

Barndorff-Nielsen, O.E. and Jupp, P.E. (1989) Approximating exponential models. *Ann. Inst. Statist. Math.*, **41**, 247–267.

Brigo, D. and Pistone, G. (1996) Projecting the Fokker–Planck equation onto a finite dimensional exponential family. Preprint 4, Dipartimento di Matematica Pura e Applicata Università di Padova.

Cox, D.R. and Reid, N. (1987) Parameter orthogonality and approximate conditional inference (with discussion). *J. Roy. Statist. Soc. B*, **49**, 1–39.

Dawid, A.P. (1975) Discussion of a paper by B. Efron. *Ann. Statist.*, **3**, 1231–1234.

Dawid, A.P. (1977) Further comments on a paper by Bradley Efron. *Ann. Statist.*, **5**, 1249.

Donsker, M.D. and Varadhan, S.R.S. (1975) Asymptotic evaluation of certain Markov processes expectations for large time I. *Commun. Pure Appl. Math.*, **28**, 1–47.

Efron, B. (1975) Defining the curvature of a statistical problem (with applications to second-order efficiency) (with discussion). *Ann. Statist.*, **3**, 1189–1242.

Efron, B. (1978) The geometry of exponential families. *Ann. Statist.*, **6**, 362–376.

Ekeland, I. and Temam, R. (1974) *Analyse Convexe et Problèmes Variationnels*. Paris: Dunod, Gauthier–Villars.

Gibilisco, P. and Pistone, G. (1998) Connections on non-parametric statistical manifolds by Orlicz space geometry, infinite dimensional analysis. *Quantum Probab. Related Topics*, **1**, 325–347.

Hardy, G.H., Littlewood, J.E. and Pólya, G. (1952) *Inequalities*, 2nd edn. London: Cambridge University Press.

Jeffreys, H. (1946) An invariant form of the prior probability in estimation problems. *Proc. Roy. Soc. London A*, **196**, 453–461.

Kass, R.E. (1989) The geometry of asymptotic inference (with discussion) *Statist. Sci.*, **4**, 188–234.

Krasnosel'skii, M.A. and Rutickii, Ya.B. (1961) *Convex Functions and Orlicz Spaces*. Groningen: Noordhoff. (Russian original (1958) Moscow: Fizmatgiz.)

Kullback, S. (1997) *Information Theory and Statistics*, reprint of 2nd (1968) edn., Mineola, NY: Dover.

Kullback, S. and Leibler, R.A. (1951) On information and sufficiency. *Ann. Math. Statist.*, **22**, 79–86.

Lang, S. (1995) *Differentiable Manifolds and Riemannian Manifolds*. New York: Springer-Verlag.

Letac, G. (1992) *Lectures on Natural Exponential Families and their Variance Functions*, Monogr. Mat., **50**. Rio de Janeiro: Istituto de Matemaática Pura e Aplicada.

Madsen, L.T. (1979) The geometry of statistical model—a generalization of curvature. Research Report 79-1, Statistics Research Unit, Danish Medical Research Council.

Murray, M.K. and Rice, J.W. (1993) *Differential Geometry and Statistics*, Monogr. Statist. Appl. Probab., **48**. London: Chapman & Hall.

Neveu, J. (1972) *Martingales à Temps Discrets*. Paris: Masson.

Pistone, G. and Rogantin, M.P. (1990) Gli strumenti della geometria differenziale nell'inferenza statistica. In P. Nastasi (ed.) *Memorial Beniamino Gulotta, Giornate di Lavoro di Probabilità e Statistica*, pp. 85–99. Palermo: Istituto Gramsci.

Pistone, G. and Rogantin, M.P. (1994) *The Transformation of the Non-Parametric Statistical Manifold under Conditioning and Sampling*. Proceedings of the 57th IMS Annual Meeting and Third World Congress of the Bernoulli Society, Chapel Hill, NC, 20–25 June 1994.

Pistone, G. and Sempi, C. (1995) An infinite dimensional geometric structure on the space of all the probability measures equivalent to a given one. *Ann. Statist.* **23**, 1543–1561.

Rao, C.R. (1945) Information and accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc*., **37**, 81–89.

Rao, C.R. (1949) On the distance between two populations. *Sankhyã*, **9**, 246–248.

Rao, M.M. and Ren, Z.D. (1991) *Theory of Orlicz Spaces*. New York: Dekker.

Rogantin, M.P. (1996) Geometrical modelling of Markovian dependence. *Metron*, **54**, 45–65.

Received April 1995 and revised October 1997