Informative Barycentres in Statistics

Bruno Pelletier
Institut de Mathématiques et de Modélisation de Montpellier
UMR CNRS 5149, Equipe de Probabilités et Statistique
Université Montpellier II
Place Eugène Bataillon
34095 Montpellier Cedex 5, France.

pelletier@math.univ-montp2.fr

Abstract

Barycentres of a discrete probability measure on a dually flat statistical manifold are introduced. They are shown to be unique and to behave as barycentres in Euclidean space. The estimation of these barycentres is studied. Potential applicative usefulness of informative barycentres include the problem of interpolating a statistical manifold valued map and the problem of model merging, which consists in merging several statistical models into a unique one. The results are illustrated on the exponential family, for which a projection theorem is proved.

Index Terms **— Barycentre, Density estimation, Entropy, Differentiable-geometrical methods, Statistical manifolds, Dual affine connections.**

1 Introduction

Since the work of Rao (1945), who turned a statistical structure into a Riemannian manifold when endowed with the Fisher information taken as a Riemannian metric, a great amount of work has been done to define geometrical objects of statistical relevance. Distances and divergences between two probability measures have been introduced (Csiszar 1975, Renyi 1961, Murray and Rice 1993, Kass and Vos 1997), together with the concepts, originating from differential geometry, of statistical curvature (Efron 1975), of dualistic structures and non-metric dual affine connections (Amari 1985, Amari and Nagaoka 2000, Amari *et al.* 1987), yielding the preferred-point geometry (Critchley *et al.* 1994) and, finally, the infinite-dimensional statistical

manifold (Pistone and Sempi 1995). This geometrical setting is especially useful for problems of statistical inference (Skovgaard 1984), quantum estimation (Fujiwara and Nagaoka 1995), and has led to intrinsic inference procedures used in hypothesis testing or estimation for example (Barndorff-Nielsen 1988, Oller and Corcuera 1995).

Within the framework of differential geometry in probability and statistics, the purpose of the present work is to introduce informative barycentres on a dually flat statistical manifold, and has been motivated by the geophysical problem known as the ocean color problem. The ocean color problem consists in estimating the concentrations of several oceanic products, such as the phytoplankton for example, from a vector $\mathbf{x}$ of remotely-sensed measurements. Those measurements depend on a vector $\mathbf{t}$ of three angular variables that are used to characterize the positions of the Sun and of the satellite, relatively to the observed point of the Earth' surface. In Pelletier (2002), a statistical estimation method of the phytoplankton concentration has been proposed. In this study, the measurements $\mathbf{x}$ have been considered as realizations of a family $\{\mathbf{X_t}\}$ of random vectors, indexed by the vector $\mathbf{t}$, with the assumption, relying on the physics of the problem, that for each $\mathbf{t}_0$, $\lim_{|\mathbf{t}-\mathbf{t}_0|\to 0} P_{\mathbf{t}} = P_{\mathbf{t}_0}$, where $P_{\mathbf{t}}$ is the probability law of $\mathbf{X_t}$. Given an estimate of the probability density function of $\mathbf{X_t}$, a nonlinear regression estimator of the phytoplankton concentration has been derived. This procedure has been repeated for several values $\mathbf{t}_1, ..., \mathbf{t}_n$ of $\mathbf{t}$, yielding a set $\{\hat{p}_{\mathbf{t}_1}, ..., \hat{p}_{\mathbf{t}_n}\}$ of estimates of the probability density functions of $\mathbf{X}_{\mathbf{t}_1}, ..., \mathbf{X}_{\mathbf{t}_n}$, respectively. Then given the estimates $\hat{p}_{\mathbf{t}_1}, ..., \hat{p}_{\mathbf{t}_n}$ the problem was to infer an estimate $\hat{p}_{\mathbf{t}}$ of the probability density function of $\mathbf{X_t}$. Next from the estimate $\hat{p}_{\mathbf{t}}$, a regression estimator of the phytoplankton concentration may be derived. A solution has been proposed in Pelletier (2002) which consists in building an interpolating map valued in a manifold of statistical models (*i.e.* probability density functions). More precisely, let $\mathcal{M}$ be a manifold of statistical models, let T be a set and let F be the map defined by:

$$F : \left| \begin{array}{l} T \to \mathcal{M} \\ \mathbf{t} \mapsto F(\mathbf{t}) = \hat{p}_{\mathbf{t}}. \end{array} \right.$$

Given n samples $(\mathbf{t}_k, F(\mathbf{t}_k) = \hat{p}_{\mathbf{t}_k})$, the problem is to build a continuous map $\hat{F} : T \to \mathcal{M}$ such that $\hat{F}(\mathbf{t}_k) = F(\mathbf{t}_k)$ for all k, i.e., which interpolates F at the $\mathbf{t}_k$. In this context of interpolation, barycentres, also called centers of mass, arise naturally.

The definition of a barycentre on a Riemannian manifold is based on the squared geodesic distance which, by definition, is symmetric in its arguments. For this fact, a Riemannian barycentre does not reflect the asymmetry that

arises naturally in statistics, and this motivated the introduction of another kind of barycentres.

The paper is organized as follows. In section 2, basic notions related to dually flat statistical manifolds are briefly reviewed. In section 3, we introduce two barycentres in a dually flat space, called informative barycentres in connection with the setting of information geometry, and show that they behave as barycentres in Euclidean space. We then study the estimation of the barycentres of M statistical models from M i.i.d. samples drawn from them. In the last paragraph of this section, two applications of the dual barycentres are presented, namely the interpolation problem, and the problem of model merging, which consists in merging, or aggregating, M models $p_1, ..., p_M$ into a unique model $\overline{p}$. Section 4 is devoted to informative barycentres in exponential families. For this particular manifold a projection theorem is proved. Finally, conclusions are given.

2 Background

Let $\mathcal{M}$ be a Riemannian manifold with Riemannian metric g. Let ∇ and ∇^* be two affine connections on $\mathcal{M}$. The connections ∇ and ∇^* are said to be dual with respect to the Riemannian metric g if for all vector fields $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ on $\mathcal{M}$ the following relation is satisfied:

$$\mathbf{X}g(\mathbf{Y},\mathbf{Z}) = g(\nabla_{\mathbf{X}}\mathbf{Y},\mathbf{Z}) + g(\mathbf{Y},\nabla^*_{\mathbf{X}}\mathbf{Z}).$$

Furthermore, if the connections ∇ and ∇^* are symmetric, or equivalently torsion-free, and if the curvature tensors with respect to ∇ and ∇^* vanish, then $\mathcal{M}$ is said to be flat with respect to ∇ and ∇^* and the triple $(\mathcal{M}, \nabla, \nabla^*)$ is called a dually flat space.

In such a case, the Riemannian manifold admits ∇-affine coordinates and ∇^*-affine coordinates. Thus there exist two coordinates neighborhoods (U, ν) and (U^*, ν^*) of some point $p_0 \in \mathcal{M}$ such that the ∇-geodesic and ∇^*-geodesic are given respectively by the curves γ and γ^* defined by equations (1) and (2).

$$\gamma : \left| \begin{array}{ccc} I \subset \mathbf{R} & \longrightarrow & \mathcal{M} \\ t & \mapsto & \nu^{-1}\left[(\nu(p_2) - \nu(p_1))\, t + \nu(p_1)\right] \end{array} \right. \tag{1}$$

$$\gamma^* : \left| \begin{array}{ccc} I \subset \mathbf{R} & \longrightarrow & \mathcal{M} \\ t & \mapsto & (\nu^*)^{-1}\left[(\nu^*(p_2) - \nu^*(p_1))\, t + \nu^*(p_1)\right]. \end{array} \right. \tag{2}$$

Let us denote by θ the ∇-affine coordinates, i.e. $\theta = \nu(p); p \in U$, and by η the ∇^*-affine coordinates, i.e. $\eta = \nu^*(p); p \in U^*$. Further, let $\{\partial_i = \frac{\partial}{\partial \theta_i}\}$

and $\{\partial^j = \frac{\partial}{\partial \eta^j}\}$ be two basis of $T_p(\mathcal{M})$. Then it has been shown, see Amari (1985) for example, that, first, parameters θ and η can be chosen in such a way that the following relation holds:

$$g(\partial_i, \partial^j) = \delta_i^j,$$

and second, that there exist two potential functions ψ and φ such that the following relations are satisfied:

$$\begin{aligned} \theta^i &= \partial^i \varphi(\eta), \\ \eta_i &= \partial_i \psi(\theta), \\ \psi(\theta) + \varphi(\eta) - \sum_i \theta^i \eta_i &= 0. \end{aligned}$$

The potential functions ψ and φ are respectively ∇-convex and ∇^*-convex, i.e. ψ is a convex function with respect to θ and φ is a convex function with respect to η. Moreover, ψ and φ are related by the Legendre transformation defined as follows:

$$\varphi(\eta) = \max_\theta \left(\sum_i \theta^i \eta_i - \psi(\theta) \right).$$

Using those potential functions, the ∇-divergence from $p_1 \in \mathcal{M}$ to $p_2 \in \mathcal{M}$ is defined as follows:

$$\mathcal{D}(p_1||p_2) = \psi(\theta_2) + \varphi(\eta_1) - \sum_i \theta_2^i \eta_{1,i}.$$

The ∇^*-divergence $\mathcal{D}^*(p_1||p_2)$ from $p_1 \in \mathcal{M}$ to $p_2 \in \mathcal{M}$ may be defined in a similar way, which yields:

$$\mathcal{D}^*(p_1||p_2) = \mathcal{D}(p_2||p_1).$$

At last, the ∇-divergence (∇^*-divergence) satisfy to:

$$\mathcal{D}^{(*)}(p_1||p_2) \geq 0,$$

where equality holds iff $p_1 = p_2$.

3 Barycentres in a dually flat space

3.1 Definition and properties

In the framework of differential geometry, a Riemannian barycentre $\overline{p}$ of a probability measure μ on a Riemannian manifold $\mathcal{M}$ may be defined as being

a minimizer of the energy functional $\mathcal{E}$ defined by:

$$\mathcal{E}(p) = \int_{\mathcal{M}} d^2(q,p) d\mu(q),$$

where $d(.,.)$ is the Riemannian distance on $\mathcal{M}$.

In the particular case where μ is a discrete measure concentrated on $p_1, ..., p_M$,the energy functional $\mathcal{E}(p)$ may be rewritten as:

$$\mathcal{E}(p) = \sum_{i=1}^{M} \alpha_i d^2(p, p_i),$$

where $\{\alpha_i\}$ are strictly positive scalars satisfying $\sum_{i=1}^{M} \alpha_i = 1$. Hence a minimizer $\hat{p}$ of $\mathcal{E}(p)$ defined above is a barycentre of M points $\{p_i\}$ with weights $\{\alpha_i\}$.

It is to be noted that a Riemannian barycentre may not be unique, as it is the case in the presence of cut-locus for example. For questions regarding uniqueness of Riemannian barycentres and their relations to geodesic convexity, we refer to Kobayashi and Nomizu (1969), Emery and Meyer (1989), and Corcuera and Kendall (1999).

Consider a finite-dimensional manifold $\mathcal{M}$ of probability measures which are absolutely continuous with respect to a measure P. Such a manifold may be turned into a Riemannian manifold when endowed with a Riemannian metric tensor taken as the Fisher information. Therefore, a Riemannian barycentre of M points of $\mathcal{M}$ could be defined as above. However its definition, based on the squared Riemannian geodesic distance, which is symmetric by definition, does not reflect the asymmetry arising naturally in statistics. Moreover from a practical point of view, the computation of the Riemannian distance between two points of a Riemannian manifold $\mathcal{M}$, irrespective to definition issues, is generally a complex procedure. These facts motivate the introduction of barycentres in statistical manifolds that first may be interpreted statistically, and second that may be computed rather simply.

Let $(\mathcal{M}, \nabla, \nabla^*)$ be a dually flat space. From now on, we shall assume that:

Assumption 3.1 *$\mathcal{M}$ may be covered by a single chart, and that,*

Assumption 3.2 *$\mathcal{M}$ is ∇-convex (resp. ∇^*-convex), i.e., for all points p_1 and p_2 in $\mathcal{M}$, there exists a unique ∇-geodesic (resp. a unique ∇^*-geodesic), connecting p_1 and p_2, and lying entirely in $\mathcal{M}$.*

Let ν and ν^* be the coordinates mappings that yield respectively the ∇-affine and the ∇^*-affine coordinates. Note that under assumption (3.1), the domain of the charts ν and ν^* is the whole $\mathcal{M}$, and that assumption (3.2) is equivalent to the assumption that $\nu(\mathcal{M})$ and $\nu^*(\mathcal{M})$ are convex subsets of a Euclidean space.

Let $p_1, ..., p_M$ be M points of $\mathcal{M}$ and let $\alpha_1, ..., \alpha_M$ be a sequence of strictly positive scalars such that $\sum_{i=1}^M \alpha_i = 1$. Define a ∇-barycentre of $\{(p_i, \alpha_i), i = 1...M\}$ as being a minimizer of

$$\mathcal{E}(p) = \sum_{i=1}^{M} \alpha_i \mathcal{D}(p||p_i),$$

where $\mathcal{D}(p||p_i)$ is the ∇-divergence from p to p_i. Similarly, define a ∇^*-barycentre of $\{(p_i, \alpha_i), i = 1...M\}$ as being a minimizer of

$$\mathcal{E}^*(p) = \sum_{i=1}^{M} \alpha_i \mathcal{D}^*(p||p_i),$$

where $\mathcal{D}^*(p||p_i)$ is the ∇^*-divergence from p to p_i. The following theorem shows that those barycentres are unique.

Theorem 3.3 *Assume that assumptions (3.1) and (3.2) hold. With the above notations, there exists a unique $\overline{p} \in \mathcal{M}$ such that $\overline{p} = \arg\min_p \mathcal{E}(p)$ and a unique $\overline{p}^*$ such that $\overline{p}^* = \arg\min \mathcal{E}^*(p)$. Moreover:*

$$\begin{aligned} \nu(\overline{p}) &= \sum_i \alpha_i \nu(p_i), \\ \nu^*(\overline{p}^*) &= \sum_i \alpha_i \nu^*(p_i). \end{aligned}$$

Proof Let $\theta_i = \nu(p_i)$, $\eta_i = \nu^*(p_i)$, $\theta = \nu(p)$ and $\eta = \nu^*(p)$. Then

$$\begin{aligned} \mathcal{E}(p) &= \sum_{i=1}^{M} \alpha_i \left(\psi(\theta_i) + \varphi(\eta) - \sum_k \theta_i^k \eta_k \right) \\ &= \sum_{i=1}^{M} \alpha_i \psi(\theta_i) + \varphi(\eta) - \sum_k \left(\sum_i \alpha_i \theta_i^k \right) \eta_k. \end{aligned}$$

Let $\overline{\theta} = \sum_i \alpha_i \theta_i$. Then

$$\begin{aligned} \mathcal{E}(p) &= \left(\psi(\overline{\theta}) + \varphi(\eta) - \sum_k \overline{\theta}^k \eta_k \right) + \left(\sum_i \alpha_i \psi(\theta_i) - \psi(\overline{\theta}) \right) \\ &= \mathcal{D}(p||\overline{p}) + C, \end{aligned}$$

where $C = \sum_i \alpha_i \psi(\theta_i) - \psi(\overline{\theta})$. Letting $\overline{p} = \nu^{-1}(\overline{\theta})$ and using the property of a divergence given by equation (12), it comes:

$$\overline{p} = \nu^{-1}(\overline{\theta}) = \arg\min_p \mathcal{E}(p).$$

The proof of the unicity and characterization formula of $\overline{p}^*$ may be obtained similarly. □

The ∇-barycentre and ∇^*-barycentre obey to the composition rule of barycentres in Euclidean space, as shown by the following proposition.

Proposition 3.4 *Let $p_1, ..., p_{M+N}$ be $M+N$ points of $\mathcal{M}$ and let $\alpha_1, ..., \alpha_{M+N}$ be $M+N$ strictly positive scalars such that $\sum_{i=1}^{M+N} \alpha_i = 1$. Let $A_M = \sum_{i=1}^{M} \alpha_i$ and $A_N = \sum_{i=M+1}^{N} \alpha_i$. Let $\overline{p}_M$ be the ∇-barycentre of $\{(p_i, \frac{\alpha_i}{A_M}), i = 1...M\}$ and let $\overline{p}_N$ be the ∇-barycentre of $\{(p_i, \frac{\alpha_i}{A_N}), i = M+1, ..., M+N\}$. Then the ∇-barycentre of $\{(p_i, \alpha_i), i = 1, ..., M+N\}$ is the ∇-barycentre of $(\overline{p}_M, A_M)$ and $(\overline{p}_N, A_N)$.*

Proof The proof is immediate. Let $\overline{p}$ be the ∇-barycentre of $\{(p_i, \alpha_i), i = 1, ..., M+N\}$. Then

$$\begin{aligned} \nu(\overline{p}) &= \sum_{i=1}^{M+N} \alpha_i \nu(p_i) \\ &= A_M \nu(\overline{p}_M) + A_N \nu(\overline{p}_N). \end{aligned}$$

□

The same result holds for ∇^*-barycentres and may be proved in a similar way.

3.2 Estimation of ∇ and ∇^* barycentres

We address now the estimation of the ∇ and ∇^*-barycentres of $(p_1, \alpha_1), ..., (p_M, \alpha_M)$ based on M independent i.i.d. samples $X_{k,1}, ..., X_{k,n_k}$ of size n_k drawn from the p_k, respectively. Let $\theta_1, ..., \theta_M$ and $\eta_1, ..., \eta_M$ be the ∇-affine and ∇^*-affine coordinates of $p_1, ..., p_M$, respectively, and let $\overline{\theta}$ and $\overline{\eta}^*$ be the ∇-affine and ∇^*-affine coordinates of the ∇ and ∇^*-barycentre, respectively, i.e.,

$$\begin{aligned} \overline{\theta} &= \sum_{k=1}^{M} \alpha_k \theta_k \\ \overline{\eta}^* &= \sum_{k=1}^{M} \alpha_k \eta_k. \end{aligned}$$

Let $\hat{\theta}_{k,n_k}$ be an estimator of θ_k, for all $k = 1, ..., M$. We consider the estimation of $\overline{\theta}$ by

$$\hat{\overline{\theta}}_n = \sum_{k=1}^{M} \alpha_k \hat{\theta}_{k,n_k},$$

where we have let $n = n_1 + ... + n_M$, i.e., the total number of observations. First, if the $\hat{\theta}_{k,n_k}$ are unbiased, then $\hat{\overline{\theta}}_n$ is also unbiased. Second if for all k, $\hat{\theta}_{k,n_k}$ converges in probability to θ_k, then $(\hat{\theta}_{1,n_1}, ..., \hat{\theta}_{M,n_M})$ converges in probability to $(\theta_1, ..., \theta_M)$, and so $\hat{\overline{\theta}}_n$ converges in probability to $\overline{\theta}$ by the continuous mapping theorem. Third, we show in the next proposition that if the $\hat{\theta}_{k,n_k}$ are asymptotically normal, then $\hat{\overline{\theta}}_n$ is also asymptotically normal.

Proposition 3.5 *Let $X_{k,1}, ..., X_{k,n_k}$ be i.i.d. samples of size n_k drawn from the p_k for $k = 1, ..., M$. Assume that:*

- *The M samples are independent;*
- *For all $k = 1, ..., M$, there exists an estimator $\hat{\theta}_{k,n_k}$ of θ_k, constructed on $X_{k,1}, ..., X_{k,n_k}$ such that $\sqrt{n_k}(\hat{\theta}_{n,n_k} - \theta_k)$ converges in distribution to a centered normal random vector with covariance matrix Σ_k.*

Let $n = n_1 + ... + n_M$, and let

$$\hat{\overline{\theta}}_n = \sum_{k=1}^{M} \alpha_k \hat{\theta}_{k,n_k}.$$

Then $\sqrt{n}(\hat{\overline{\theta}}_n - \overline{\theta})$ converges in distribution to a centered normal random vector with covariance matrix $\Sigma = \sum_{k=1}^{M} \frac{\alpha_k^2}{\lambda_k} \Sigma_k$, when all the n_k tend towards infinity in such a way that $\frac{n_k}{n}$ tends towards a number λ_k, with $0 < \lambda_k < 1$.

Proof By assumption, $\sqrt{n_k}(\hat{\theta}_{n,n_k} - \theta_k)$ converges in distribution to a centered normal random vector with covariance matrix Σ_k. Hence for all $k = 1, ..., M$, $\sqrt{n}(\hat{\theta}_{k,n_k} - \theta_k) = \sqrt{\frac{n}{n_k}}\sqrt{n_k}(\hat{\theta}_{k,n_k} - \theta_k)$ converges in distribution to a centered normal random vector with covariance matrix $\frac{1}{\lambda_k}\Sigma_k$, when the $n_k \to \infty$ with $\frac{n_k}{n} \to \lambda_k$. Since the samples for $k = 1, ..., M$ are independent by assumption, the $\hat{\theta}_{k,n_k}$ are also independent, and the characteristic function of $\sqrt{n}(\hat{\overline{\theta}}_n - \overline{\theta})$

writes as

$$\begin{aligned}
\Phi_{\sqrt{n}(\hat{\bar{\theta}}_n-\bar{\theta})}(t) &= \mathbb{E}\left[\exp\left(i<t,\sqrt{n}(\hat{\bar{\theta}}_n-\bar{\theta})>\right)\right] \\
&= \mathbb{E}\left[\exp\left(i<t,\sqrt{n}\sum_{k=1}^{M}\alpha_k(\hat{\theta}_{k,n_k}-\theta_k)>\right)\right] \\
&= \prod_{k=1}^{M}\mathbb{E}\left[\exp\left(i<t,\sqrt{n}\alpha_k(\hat{\theta}_{k,n_k}-\theta_k)\right)\right] \\
&= \prod_{k=1}^{M}\mathbb{E}\left[\exp\left(i<\alpha_k t,\sqrt{\frac{n}{n_k}}\sqrt{n_k}(\hat{\theta}_{k,n_k}-\theta_k)\right)\right],
\end{aligned}$$

from which the statement follows. □

The estimation of the ∇^*-barycentre can be addressed similarly in the ∇^*-affine coordinate system. If we assume that there exists estimators $\hat{\eta}_{k,n_k}$ of the η_k such that $\sqrt{n_k}(\hat{\eta}_{k,n_k}-\eta_k) \rightsquigarrow \mathcal{N}(0,\Sigma_k^*)$, then the ∇^*-affine coordinate $\overline{\eta}$ of the ∇^*-barycentre can be estimated by $\hat{\bar{\eta}}_n = \sum_{k=1}^{m}\hat{\eta}_{k,n_k}$ and, when $n_k \to \infty$ with $\frac{n_k}{n} \to \lambda_k$, $\sqrt{n}(\hat{\bar{\eta}}_n - \overline{\eta}) \rightsquigarrow \mathcal{N}(0,\sum_{k=1}^{M}\frac{\alpha_k^2}{\lambda_k}\Sigma_k^*)$.

3.3 Applications

3.3.1 Interpolation

Let $(\mathcal{M},\nabla,\nabla^*)$ be a finite-dimensional dually flat manifold covered by a single chart. Let F be a map defined, for simplicity but without loss of generality, on a real interval T and valued in $\mathcal{M}$:

$$F : \left| \begin{array}{l} T \longrightarrow \mathcal{M} \\ t \mapsto p(t). \end{array} \right.$$

Let be given N samples of F: $\{(t_k,p(t_k)),k=1...N\}$, with $t_1<...<t_N$. The problem is to build a piecewise linear, in some sense, map $\hat{F}$ interpolating F at $\{t_k\}$, *i.e.* such that $\hat{F}(t_k)=p(t_k)\forall k$. This may be achieved in a Riemannian manifold by connecting any two points $p(t_k)$ and $p(t_{k+1})$ by a geodesic, provided they belong to a geodesically convex subset of $\mathcal{M}$. If so, $\hat{F}$ is defined by $N-1$ maps $\hat{F}_k : [t_k;t_{k+1}] \longrightarrow \mathcal{M}$ such that $\hat{F}_k(t_k)=p(t_k)$, $\hat{F}(t_{k+1})=p(t_{k+1})$ and $\hat{F}_k$ is a geodesic on $\mathcal{M}$. Thus $\forall t\in[0;1]$, $\hat{F}_k(t)$ is the point which minimizes with respect to p the quantity $\frac{t_{k+1}-t}{t_{k+1}-t_k}d^2(p,p(t_k)) + \frac{t-t_k}{t_{k+1}-t_k}d^2(p,p(t_{k+1}))$, that is, $\hat{F}_k(t)$ is the Riemannian barycentre of $p(t_k)$ and $p(t_{k+1})$ with respective weights $(t_{k+1}-t)/(t_{k+1}-t_k)$ and $(t-t_k)/(t_{k+1}-t_k)$.

Returning to a dually flat manifold $\mathcal{M}$, it is possible to define the straightness of a curve on $\mathcal{M}$ with respect to the connections ∇ or ∇^*. This yields

two interpolating maps defined respectively by $N-1$ maps F_k and F_k^* such that $\forall t \in [t_k; t_{k+1}]$, $F_k(t)$ and $F_k^*(t)$ are respectively the ∇-barycentre and ∇^*-barycentre of $p(t_k)$ and $p(t_{k+1})$ with respective weights $(t_{k+1}-t)/(t_{k+1}-t_k)$ and $(t-t_k)/(t_{k+1}-t_k)$.

Remark 3.6 In dimension $d > 1$, the above results may be extended as follows. Assume T is a d-dimensional hypercube included in a Euclidean space E, with its usual metric, and that the sample points $\{\mathbf{t}_k\}$ are positioned on a regular grid of T. This leads to the map F_k (resp. F_k^*) defined on the interior of a d-dimensional cube $C(\mathbf{t}_{k_1}, ..., \mathbf{t}_{k_{2^d}})$ with corners $\mathbf{t}_{k_1}, ..., \mathbf{t}_{k_{2^d}}$ where $F_k(\mathbf{t})$ (resp. $F_k^*(\mathbf{t})$) is the ∇-barycentre (resp. ∇^*-barycentre) of $\{(p(\mathbf{t}_{k_1}), \alpha_1(\mathbf{t})), ..., \big(p(\mathbf{t}_{k_{2^d}}), \alpha_{2^d}(\mathbf{t})\big)$. In this case, the coefficients $\alpha_1(\mathbf{t}), ..., \alpha_{2^d}(\mathbf{t})$ are the coefficients of the standard multilinear interpolation scheme (linear when $d=1$, bilinear when $d=2$ and so on). They are defined as follows. Assume without loss of generality that $C(\mathbf{t}_{k_1}, ..., \mathbf{t}_{k_{2^d}})$ is $[0;1]^d$. The coefficient $\alpha_i(\mathbf{t})$ is defined by $\alpha_i(\mathbf{t}) = \prod_{j=1}^d (1 - |\mathbf{t}^j - \mathbf{t}^j_{k_i}|)$. Multilinear interpolation involves the processing of 2^d data points, which becomes computationaly extensive as d increases. One alternative to multilinear interpolation is interpolation on simplicial complexes, where the interpolated value is expressed as an appropriate convex combination of $d+1$ values.

We consider now the situation where the $p(t_k)$ are estimated by $\hat{p}(t_k)$ constructed on i.i.d. samples $X_{k,1}, ..., X_{k,n_k}$, as in Section 3.2, and where the interpolating map $\hat{F}$ is such that $\hat{F}(t_k) = \hat{p}(t_k)$, for all k. Then $F_k(t)$ and $F_k^*(t)$ are estimators of respectively the ∇-barycentre and ∇^*-barycentre of $p(t_k)$ and $p(t_{k+1})$, with associated weitghts $(t_{k+1}-t)/(t_{k+1}-t_k)$ and $(t-t_k)/(t_{k+1}-t_k)$. In fact, there are two problems in this context of interpolation. The first one concerns the statistical properties of the map F_k and F_k^* which, as estimators of the ∇ and ∇^*-barycentres, are given above. The second one is related to the choice of the connection to perform the interpolation. This question is difficult and does not seem to admit a general answer. For instance, the choice could be based on a comparison of the regularity of F_k and F_k^*, but which in turn would depend on how regularity is defined. In effect, F_k which is straight w.r.t. the ∇ connection is, in this sense, the most regular curve joining $p(t_k)$ and $p(t_{k+1})$, and the same holds for F_k^* w.r.t. the ∇^* connection. This question is discussed in Section 4 in the case of the gaussian family.

3.3.2 Model merging

The problem addressed herein is the one which consists in merging or aggregating together M models $p_1, ..., p_M$ in a unique model $\overline{p}$, which occurs for instance in modeling with finite mixtures (McLachlan and Peel 2000). Consider the problem of fitting a gaussian mixture model, i.e., a convex combination of N gaussians. One approach is to compute the maximum likelihood estimator (MLE) of the mixture parameters (e.g. by using the expectation maximization (EM) algorithm (Dempster *et al.* 1977)) for several values of N, and next to select the more appropriate number of components according to some model selection criterion. This approach, which requires multiple MLE computations, becomes intractable as the dimension of the data increases. One alternative is the iterative pairwise replacement algorithm (Scott and Szewczyk 2001), which consists in starting with a nonparametric kernel density estimator, and next in iteratively reducing the number of components by merging the two most similar kernels (in the sense of a similarity measure, e.g. Hellinger distance) into a single one. This algorithm yields a sequence of mixture models indexed by N, among which one of them may be selected according to a goodness-of-fit measure, or a model selection criterion as above. Also, it provides a simple yet powerful method for estimating the number of components, prior to computation of the MLE, thus reducing the burden of several MLE computations for different N. We focus here on the merging operation, which is performed in Scott and Szewczyk (2001) according to the method of moments (Everitt and Hand 1981, Furman and Lindsay 1994). Let p_1 and p_2 be two gaussian densities with means μ_1 and μ_2, variances σ_1^2 and σ_2^2, and associated weights ω_1 and ω_2, respectively. Let $\alpha = \omega_2/(\omega_1+\omega_2)$. By the method of moments, p_1 and p_2 are merged into the gaussian $\overline{p}^*$ with mean $(1-\alpha)\mu_1+\alpha\mu_2$ and variance $(1-\alpha)\sigma_1^2+\alpha\sigma_2^2+\alpha(1-\alpha)(\mu_1-\mu_2)^2$, which receives the weight $\omega_1+\omega_2$ in the new mixture model. As shown in the next section, $\overline{p}^*$ is in fact the ∇^*-barycentre of $(p_1, (1-\alpha))$ and (p_2, α). Assume that the sequence of merging operations that led to $\overline{p}^*$ involves M initial gaussians $p_1, ..., p_M$ centered on M data points. Then by the composition rule of barycentres, $\overline{p}^*$ is the ∇^*-barycentre of $p_1, ..., p_M$ with all associated weights equal to $\frac{1}{M}$.

4 Informative barycentres in full exponential families

4.1 Preliminaries : dual geometry of full exponential families

In this section, the focus is on informative barycentres in full exponential families. A full exponential family $\mathcal{M}$ is composed of the probability density functions $p(\mathbf{x};\theta)$ on a set Ω, depending on a parameter vector $\theta \in \Theta \subset \mathbf{R}^d$, whose log-likelihood $l(\mathbf{x};\theta)$ is of the following type:

$$l(\mathbf{x},\theta) = C(\mathbf{x}) + \sum_{i=1}^{d} \theta^i F_i(\mathbf{x}) - \psi(\theta),$$

where

$$\psi(\theta) = \log \int \exp\left(C(\mathbf{x}) + \sum_{i=1}^{d} \theta^i F_i(\mathbf{x}) \right) d\mathbf{x}.$$

Under some regularity conditions, $\mathcal{M}$ may be endowed with the Fisher information metric. The components g_{ij} of the Fisher information metric tensor are defined by the following equation :

$$g_{ij} = E_{p(.;\theta)} \left[\partial_i \log p(.;\theta) \partial_j \log p(.;\theta) \right],$$

where $E_{p(.;\theta)}[.]$ is the expectation operator with respect to $p(.;\theta)$. It has been shown, first, that a full exponential family is flat with respect to the exponential connection ∇ and the mixture connection ∇^* defined by the sets of Christoffel symbols given by:

$$\Gamma_{ijk} = \tilde{\Gamma}_{ijk} - \frac{1}{2} T_{ijk}, \tag{3}$$

$$\Gamma^*_{ijk} = \tilde{\Gamma}_{ijk} + \frac{1}{2} T_{ijk}. \tag{4}$$

In those equations, $\tilde{\Gamma}_{ijk}$ are the Christoffel symbols of the Levi-Civita connection associated with the Fisher information metric and $\{T_{ijk}\}$ are the skewness tensor components, defined by:

$$T_{ijk} = E_p \left[\partial_i l \partial_j l \partial_k l \right].$$

It has been shown, second, that the ∇-divergence is none other than the Kullback-Leibler (KL) divergence $\mathcal{D}(p_1||p_2)$ from $p_1 \in \mathcal{M}$ to $p_2 \in \mathcal{M}$ defined by:

$$\mathcal{D}(p_1||p_2) = \int_\Omega p_1(\mathbf{x}) \log \frac{p_1(\mathbf{x})}{p_2(\mathbf{x})}.$$

In fact, Amari defined a one-parameter class of statistical connections, the Christoffel symbols of which are given by:

$$\Gamma_{ijk}^{(\alpha)} = \tilde{\Gamma}_{ijk} - \frac{\alpha}{2} T_{ijk}.$$

Those connections are the only statistical connections derived from a divergence of the type $E_p\left[F(\frac{p}{q})\right]$, where F is a convex function such that $F(1) = 0$ and $F''(1) = 1$ (Amari 1985, p. 99). This divergence is invariant, by construction, with respect to an invertible differentiable transformation of the sample space. The mixture and exponential connections correspond respectively to $\alpha = -1$ and $\alpha = +1$. The exponential family is therefore said to be ± 1flat.

4.2 Barycentres in full exponential families

The following theorem relates the ∇^*-barycentre of $(p_i, \alpha_i); i = 1...M$ with the mixture distribution $\tilde{p} = \sum_{i=1^k} \alpha_k p_k$.

Theorem 4.1 *Let $\mathcal{M}$ be a full exponential family. Let $p_1, ..., p_M$ be M points of $\mathcal{M}$ and let $\theta_k = \nu(p_k)$ and $\eta_k = \nu^*(p_k)$. Let $\{\alpha_k\}_k$ be a sequence of strictly positive scalars such that $\sum_{k=1}^M \alpha_k = 1$. Then the ∇^*-barycentre $\overline{p}^*$ of $\{(p_k, \alpha_k)\}_k$ is the unique point which minimizes with respect to $p \in \mathcal{M}$ the KL-divergence from $\tilde{p} = \sum_{k=1}^M \alpha_k p_k$ to p.*

Proof Let $\mathcal{D}^*(p||\tilde{p}) = \mathcal{D}(\tilde{p}||p)$, i.e., the KL-divergence from $\tilde{p}$ to p. We have:

$$\begin{aligned}
\mathcal{D}^*(p||\tilde{p}) &= E_{\tilde{p}}[\log \tilde{p}] - E_{\tilde{p}}[\log p] \\
&= E_{\tilde{p}}[\log \tilde{p}] - E_{\tilde{p}}[C(x)] - \sum_i \theta^i E_{\tilde{p}}[F_i(\mathbf{x})] + \psi(\theta) \\
&= E_{\tilde{p}}[\log \tilde{p}] - E_{\tilde{p}}[C(x)] - \sum_i \theta^i \sum_k \alpha_k E_{p_k}[F_i(\mathbf{x})] + \psi(\theta) \\
&= E_{\tilde{p}}[\log \tilde{p}] - E_{\tilde{p}}[C(x)] - \sum_i \theta^i \sum_k \alpha_k \eta_k^i + \psi(\theta).
\end{aligned}$$

Let $\overline{\eta}^* = \nu^*(\overline{p}^*)$. Then:

$$\begin{aligned}
\mathcal{D}^*(p||\tilde{p}) &= E_{\tilde{p}}[\log \tilde{p}] - E_{\tilde{p}}[C(x)] - \sum_i (\overline{\theta}^*)^i \overline{\eta}_i^* + \psi(\overline{\theta}^*) \\
&\quad + \sum_i (\overline{\theta}^*)^i \overline{\eta}_i^* - \psi(\overline{\theta}^*) - \sum_i \theta^i \overline{\eta}_i^* + \psi(\theta) \\
&= \mathcal{D}^*(\overline{p}^*||\tilde{p}) + \sum_i (\overline{\theta}^*)^i \overline{\eta}_i^* - \psi(\overline{\theta}^*) - \sum_i \theta^i \overline{\eta}_i^* + \psi(\theta) \\
&= \mathcal{D}^*(p||\overline{p}^*) + \mathcal{D}^*(\overline{p}^*||\tilde{p}).
\end{aligned}$$

□

This theorem states that the ∇^*-barycentre of M points $p_1, ..., p_M$ of $\mathcal{M}$ with associated weights $\alpha_1, ..., \alpha_M$ is the unique closest point of $\mathcal{M}$, in the sense of the KL-divergence, to the mixture $\tilde{p} = \sum_{k=1}^{M} \alpha_k p_k$. Therefore the ∇^*-barycentre $\overline{p}^*$ is the projection, in the sense of minimum KL-divergence, of the mixture $\tilde{p}$ onto $\mathcal{M}$. It may be noted that the mixture $\tilde{p}$ does not belong to $\mathcal{M}$, but that the KL-divergence from $\tilde{p}$ to $p \in \mathcal{M}$ is well defined, contrary to the ∇^*-divergence from p to $\tilde{p}$ which makes sense only on $\mathcal{M}$ where it coincides with the KL-divergence from $\tilde{p}$ to p. Hence $\mathcal{M}$ is implicitely considered as a submanifold of a larger manifold containing $\tilde{p}$. To make geometricaly precise the notion of projection, one can define an enlarged model of $\mathcal{M}$ containing $\tilde{p}$, at least in a neighborhood of $\overline{p}^*$, by using the mixture connection, in directions orthogonal to $\mathcal{M}$, and this would lead to projection in the sense of the ∇^*-divergence. This construction is introduced by Komaki (1996) to improve an estimative distribution to a predictive distribution, by shifting it in a direction orthogonal to the model.

There do not exists such a natural projection theorem for the ∇-barycentre in full exponential families. In fact, the ∇-barycentre $\overline{p}$ of $(p_i, \alpha_i)_i$ may be expressed as follows:

$$\overline{p} = C \prod_i p_i^{\alpha_i},$$

where C is a normalizing constant. This is immediate from the definition of a full exponential family. In the next section, these results are illustrated on the gaussian family on $\mathbf{R}$.

4.3 Barycentres in the gaussian family

Let $\mathcal{M}$ be the gaussian family on $\mathbf{R}$ parameterized by the mean μ and the variance σ^2. So $\mathcal{M}$ is a two-dimensional manifold, and is flat with respect to the ∇ and ∇^* connections defined by equations (3) and (4). Let θ and η be respectively the ∇-affine and ∇^*-affine coordinates. They are related to (μ, σ) by

$$\begin{aligned} \theta &= (\frac{\mu}{\sigma^2}, \frac{1}{2\sigma^2}), \\ \eta &= (\mu, -(\mu^2 + \sigma^2)). \end{aligned}$$

Let p_1 and p_2 be two points of $\mathcal{M}$ of parameters (μ_1, σ_1) and (μ_2, σ_2). Let $\alpha \in [0; 1]$. Let $\overline{\theta}$ and $\overline{\eta}^*$ be the coordinates of respectively the ∇-barycentre

and ∇^*-barycentre of $(p_1,(1-\alpha))$ and (p_2,α). They are given by

$$\begin{aligned}\overline{\theta} &= \left((1-\alpha)\frac{\mu_1}{\sigma_1^2}+\alpha\frac{\mu_2}{\sigma_2^2};\frac{1-\alpha}{2\sigma_1^2}+\frac{\alpha}{2\sigma_2^2}\right),\\ \overline{\eta}^* &= \left((1-\alpha)\mu_1+\alpha\mu_2;-\left[(1-\alpha)(\mu_1^2+\sigma_2^2)+\alpha(\mu_2^2+\sigma_2^2)\right]\right).\end{aligned}$$

Returning to the coordinates μ and σ, the mean $\overline{\mu}$ and the variance $\overline{\sigma}^2$ of the ∇-barycentre may be expressed as

$$\begin{aligned}\overline{\mu} &= \frac{(1-\alpha)\mu_1\sigma_2^2+\alpha\mu_2\sigma_1^2}{\alpha\sigma_1^2+(1-\alpha)\sigma_2^2},\\ \overline{\sigma}^2 &= \frac{\sigma_1^2\sigma_2^2}{(1-\alpha)\sigma_2^2+\alpha\sigma_1^2}.\end{aligned}$$

Similarly, the mean $\overline{\mu}^*$ and the variance $(\overline{\sigma}^*)^2$ of the ∇^*-barycentre may be expressed as

$$\overline{\mu}^* = (1-\alpha)\mu_1+\alpha\mu_2, \tag{5}$$

$$(\overline{\sigma}^*)^2 = (1-\alpha)\sigma_1^2+\alpha\sigma_2^2+\alpha(1-\alpha)(\mu_1-\mu_2)^2. \tag{6}$$

From the expression of $\overline{\sigma}^2$, it results that for all $\alpha\in[0;1]$,

$$\min(\sigma_1^2,\sigma_2^2)\leq\overline{\sigma}^2\leq\max(\sigma_1^2,\sigma_2^2). \tag{7}$$

In the case where $\mu_1\neq\mu_2$, $(\overline{\sigma}^*)^2$ does not satisfy to a similar relation. Indeed, the expression of $(\overline{\sigma}^*)^2$ is a parabola in α, with extremal values σ_1^2 and σ_2^2, and attains its maximal value at $\alpha=\frac{1}{2}\left[1-\frac{\sigma_1^2-\sigma_2^2}{(\mu_1-\mu_2)^2}\right]$. Therefore, there exists an interval $I\subset[0;1]$ of values of α such that $(\overline{\sigma}^*)^2$ is greater than $\max(\sigma_1^2,\sigma_2^2)$. Consequently the first inequality in (7) holds for $\overline{\sigma}^*$ but the second one does not.

Consider again the problem of model merging in the context of fitting a finite gaussian mixture. Let $x_1,...,x_n$ be a realization of an i.i.d. sample $X_1,...,X_n$ drawn from an unknown density f. Let $\hat{f}_{n,K}$ be the kernel density estimator of f with kernel $K(x)=\frac{1}{\sqrt{(2\pi)}}e^{-\frac{1}{2}x^2}$ and smoothing parameter h, i.e.,

$$f_{n,K}(x)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}e^{-\frac{1}{2}\left(\frac{x-X_i}{h}\right)^2}.$$

The iterative pairwise replacement algorithm mentioned above merges two selected kernels into a unique one. If the ∇-barycentre is used to merge two initial kernels, then the resulting kernel will have variance h^2. By recurrence,

such an approach leads to a sequence of finite mixture of gaussians with all variances equal to h^2. So clearly the ∇-barycentre is not appropriate for this purpose contrary to the ∇^*-barycentre which yields the same sequence of finite mixtures as the one obtained by using the method of moments, as may be seen by comparing equations (5) and (6) with the ones given in Section 3.3.2.

Consider the problem of interpolating a continuous map defined on a real interval $[a; b]$. Let $a = t_0 < t_1 < ... < t_n = b$ be $n + 1$ interpolating points. Let $p(t_0), ..., p(t_n)$ and $\hat{p}(t_0), ..., \hat{p}(t_n)$ be respectively the true and estimated models (constructed independently from i.i.d. samples drawn from the $p(t_k)$, eventually of different sizes) at $t_0, ..., t_n$. For each t in $[t_k, t_{k+1}]$, we can take as an approximation to $p(t)$ either the ∇ or the ∇^*-barycentre of $\hat{p}(t_k)$ and $\hat{p}(t_{k+1})$, with associated weights $\frac{t_{k+1}-t}{t_{k+1}-t_k}$ and $\frac{t-t_k}{t_{k+1}-t_k}$. To choose between these barycentres, one needs a criterion or principle, as it is the case for instance in the regularization of an inverse ill-posed problem. One principle can be to select the barycentre such that the estimated variability of the underlying phenomenon at t is nor lower nor greater than the variabilities that have been observed at t_k and t_{k+1}, and this one leads to choosing the ∇-barycentre.

5 Conclusion

Dual informative barycentres on a dually flat statistical manifold $(\mathcal{M}, \nabla, \nabla^*)$ have been introduced. The major concern of this research has been to provide a solution to the problem of interpolating a statistical manifold valued map, which originates from the ocean color problem described in the introduction. One close problem is the one of model merging, which occurs for instance in modeling with finite mixtures. The important question that arises when using informative barycentres for these problems concerns the choice of the connection and its associated divergence. A piece of answer has been given above for the two examples and when $\mathcal{M}$ is the manifold of normal densities on $\mathbf{R}$. However it seems that the choice depends both on the family of statistical models and on the application for which they are used. So the question is especialy difficult, and also arises in another statistical context, namely the estimation by minimum ϕ-divergence. As discussed in Pardo *et al.* (2002), choosing the "best" function ϕ for this purpose depends on the family of models under consideration, and this important question remains open.

It has been shown in the case where the manifold of statistical models is an exponential family $\mathcal{M}$ that the ∇^*-barycentre of $\{p_i, \alpha_i\}$ is the unique closest point of the manifold to the mixture $\tilde{p} = \sum_i \alpha_i p_i$, in the sense of

the KL-divergence. As mentioned above, the mixture $\tilde{p}$ does not belong to $\mathcal{M}$ but the KL-divergence from the mixture $\tilde{p}$ to a point $p \in \mathcal{M}$ is well defined. Hence $\mathcal{M}$ has been implicitly considered as a subset of the set of all probability density functions. It would be interesting to pursue the work by studying the question of whether similar results may be obtained for other families and statistical connections and, more generally, for submanifolds of the infinite-dimensional statistical manifold.

Acknowledgements

The author is grateful to the anonymous referees for their detailed comments that significantly helped improve the paper.

References

Amari, S. (1985), *Differential-geometrical methods in statistics*, Vol. 28 of *Lecture Notes in Statistics*, Springer-Verlag, Berlin.

Amari, S. and Nagaoka, H. (2000), *Methods of Information Geometry*, AMS and Oxford University Press.

Amari, S., Barndorff-Nielsen, O., Kass, R., Lauritzen, S. and Rao, C. (1987), *Differential geometry and statistical inference*, IMS.

Barndorff-Nielsen, O. (1988), *Parametric Statistical Models and Likelihood*, Vol. 50 of *Lecture Notes in Statistics*, Springer.

Corcuera, J. and Kendall, W. (1999), Riemannian barycentres and geodesic convexity, *Mathematical Proceedings of the Cambridge Philosophical Society* **127**(2), 253–269.

Critchley, F., Mariott, P. and Salmon, M. (1994), Preferred point geometry and the local differential geometry of the kullback-leibler divergence, *The Annals of Statistics* **22**(3), 1587–1602.

Csiszar, I. (1975), I-divergence geometry of probability distributions and minimization problems, *The Annals of Probability* **3**, 146–158.

Dempster, A., Laird, N. and Rubin, D. (1977), Maximum likelihood from incomplete data via the em algorithm (with discussion), *Journal of the Royal Statistical Society* **39**(1), 1–38.

Efron, B. (1975), Defining the curvature of a statistical problem (with applications to second order efficiency), *The Annals of Statistics* **3**, 1189–1242.

Emery, M. and Meyer, P. (1989), *Stochastic Calculus in Manifolds*, Springer Verlag.

Everitt, B. and Hand, D. (1981), *Finite Mixture Distributions*, Chapmann and Hall, New York.

Fujiwara, A. and Nagaoka, H. (1995), Quantum fisher metric and estimation for pure state models, *Physics Letters* **201A**, 119–124.

Furman, W. and Lindsay, B. (1994), Testing for the number of components in a mixture of normal distributions using moment estimators, *Computational Statistics and Data Analysis* **17**(5), 473–492.

Kass, R. and Vos, P. (1997), *Geometrical Foundations of Asymptotix Inference*, John Wiley, New York.

Kobayashi, S. and Nomizu, K. (1969), *Foundations of differential geometry*, Interscience.

Komaki, F. (1996), On asymptotic properties of predictive distributions, *Biometrika* **83**(2), 299–313.

McLachlan, G. and Peel, D. (2000), *Finite Mixture Models*, Wiley, New York.

Murray, M. and Rice, J. (1993), *Differential Geometry and Statistics*, Chapman and Hall.

Oller, J. and Corcuera, J. (1995), Intrinsic analysis of statistical estimation, *The Annals of Statistics* **23**(5), 1562–1581.

Pardo, J., Pardo, L. and Zografos, K. (2002), Minimum ϕ-divergence estimators with constraints in multinomial populations, *Journal of Statistical Planning and Inference* **104**, 221–237.

Pelletier, B. (2002), Remote sensing of phytoplancton with neural networks, *in* Remote Sensing of the Ocean and Sea Ice, Vol. 4880. Proc. SPIE Int. Symp. on Remote Sensing.

Pistone, G. and Sempi, C. (1995), An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one, *The Annals of Statistics* **23**(5), 1543–1561.

Rao, C. (1945), Information and the accuracy attainable in the estimation of statistical parameters, *Bulletin of the Calcutta Mathematical Society* **37**, 81–91.

Renyi, A. (1961), On measures of entropy and information, *in* 'Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability'.

Scott, D. and Szewczyk (2001), From kernels to mixtures, *Technometrics* **43**(3), 323–335.

Skovgaard, L. (1984), A riemannian geometry of the multivariate normal model, *Scandinavian Journal of Statistics* **11**, 211–223.