Electronic Journal of Statistics
Vol. 12 (2018) 2401–2429
ISSN: 1935-7524
https://doi.org/10.1214/18-EJS1454

Asymptotic minimum scoring rule prediction

Federica Giummolè and Valentina Mameli

Dept. of Environmental Sciences, Informatics and Statistics
Ca' Foscari University of Venice, Italy
e-mail: giummole@unive.it; valentina.mameli@unive.it

Abstract: Most of the methods nowadays employed in forecast problems are based on scoring rules. There is a divergence function associated to each scoring rule, that can be used as a measure of discrepancy between probability distributions. This approach is commonly used in the literature for comparing two competing predictive distributions on the basis of their relative expected divergence from the true distribution.

In this paper we focus on the use of scoring rules as a tool for finding predictive distributions for an unknown of interest. The proposed predictive distributions are asymptotic modifications of the estimative solutions, obtained by minimizing the expected divergence related to a general scoring rule.

The asymptotic properties of such predictive distributions are strictly related to the geometry induced by the considered divergence on a regular parametric model. In particular, the existence of a global optimal predictive distribution is guaranteed for invariant divergences, whose local behaviour is similar to well known α-divergences.

We show that a wide class of divergences obtained from weighted scoring rules share invariance properties with α-divergences. For weighted scoring rules it is thus possible to obtain a global solution to the prediction problem. Unfortunately, the divergences associated to many widely used scoring rules are not invariant. Still for these cases we provide a locally optimal predictive distribution, within a specified parametric model.

MSC 2010 subject classifications: Primary 62M20; secondary 60G25.
Keywords and phrases: α-connection, Fisher metric, Kullback-Leibler divergence, monotone and regular divergence, predictive density, scoring rule, weighted scoring rule.

Received June 2017.

Contents

1. Introduction

Recent years have seen growing interest in the use of scoring rules for statistical estimation and prediction. These emerge in forecast problems as means to assess the relative quality of a probabilistic forecast; see [7], [24] and [40]. In particular, scoring rules provide a measure of the loss suffered by a forecaster in view of a certain outcome. The notion of proper scoring rule, which motivates a forecaster to be honest in his predictions, is an attractive property for scoring rules in both the contexts of prediction and estimation. Indeed, from the estimation point of view, proper scoring rules lead to unbiased estimating equations; see for instance [15]. Therefore, estimators for an unknown parameter based on proper scoring rules can be obtained by resorting to results from M-estimation; see [16], [34] and [35]. In the prediction framework, the use of scoring rules is almost exclusively restricted to measure the relative quality of an available probability distribution in comparison to other distributions. Indeed, proper scoring rules furnish summary measures to simultaneously evaluate calibration and sharpness of probability forecasts; see [22].

There exists a great variety of scoring rules, such as the logarithmic score, the Tsallis score and the Bregman score. Moreover, [30] introduce the alternative class of weighted scoring rules that generalise already known score functions. Weighted scoring rules evaluate probability forecasts on the basis of a non-uniform distribution that represents the available information at the time of prediction. One general family of weighted scoring rules, which encompasses the class of [30], is the family of quasi-Bregman weighted scoring rules; see [19]. It should be noted that there are several different notions of weighted scoring rules in the literature, which differ from the one considered in this paper and in [30]

and [19]. For instance, the weighted scoring rules introduced in [27] evaluate forecasts only on a restricted domain of the outcome.

A divergence function can be naturally associated to each proper or weighted scoring rule, thus providing an instrument for comparing different predictive distributions. Divergence functions derived from proper scoring rules are called score divergences. Instead, we can name weighted score divergences those derived from weighted scores. The most famous score divergence is probably the Kullback-Leibler divergence that can be obtained from the logarithmic score.

The use of measures of discrepancy between probability distributions is widely spread in statistics; see [39] and references therein. Many authors have considered the problem of finding optimal predictive distributions with respect to some suitable divergence; see, among others, [1], [10], [29] and [37]. As shown in [11], the existence of asymptotically optimal predictive distributions depends on the geometric properties of the considered divergence. In particular, the class of monotone and regular divergences, introduced in [8] and studied in [9] as a wide class of invariant divergences, leads to asymptotically optimal predictive distributions.

In this work we propose a wider and more complete use of scoring rules for prediction, that goes beyond the simple comparison of two competing predictive distributions. In particular, borrowing from [11], we consider the existence of predictive distributions that asymptotically minimize score and weighted score divergences. To this aim, we first study the local behaviour of score and weighted score divergences up to third order. This clarifies the relationship among commonly used score divergences, weighted score divergences and well known classes of invariant divergences, such as α-divergences and ϕ-divergences. We show that, for the most part, proper scoring rules used in practice lead to score divergences which are not monotone and regular. As a consequence, the geometry these divergences induce on a statistical manifold is non-invariant. Instead, quasi-Bregman weighted divergences share the same invariant properties of monotone and regular divergences.

From the predictive viewpoint, we provide asymptotically optimal predictive distributions for both score and weighted score divergences. The lack of invariance of score divergences implies that the proposed predictive distribution can only be found within a specified parametric model. On the other hand, for quasi-Bregman weighted divergences, the optimal predictive distribution does not depend on the considered parametric model, thus constituting a complete and global solution to the prediction problem.

The paper is organised as follows. In Section 2 we aim to introduce the basic notions of differential geometry in statistical theory, which will be used in the paper. Then, in Section 3, we study the geometric properties of score divergences and weighted score divergences and we characterise their Taylor expansion up to third order. In Section 4 we apply these new results in the context of prediction. We discuss the existence of an optimal predictive distribution that asymptotically minimizes the expected divergence from the true distribution. Section 5 is dedicated to some touchstone examples covering the range of possible situations that may be encountered in practice.

2. Basic concepts in differential geometry of statistical models

This section is devoted to recall some basic notions on statistical manifolds. The reader wishing to deepen his knowledge in the field, can refer to [3], [28], [38] and references therein.

Let χ be a set and $\mathcal{A}$ be a σ-algebra of subsets of χ. Let $\mathcal{P}$ be a parametric model, $\mathcal{P} = \{p(x;\theta) : \theta \in \Theta \subseteq \mathbb{R}^d\}$, with $p(x;\theta)$ densities with respect to some dominating measure μ on $(\chi, \mathcal{A})$. Under suitable regularity conditions (see [3], p. 16), $\mathcal{P}$ can be regarded as a d-dimensional manifold embedded in the set of all probability distributions on $(\chi, \mathcal{A})$. In this context, the parametrisation $\theta = (\theta^1, \dots, \theta^d)$ plays the role of a system of local coordinates, so that any density $p(x;\theta)$ in $\mathcal{P}$ corresponds to a point in the manifold, univocally determined by the value of θ.

The tangent space T_θ at a point $p(x;\theta)$ of $\mathcal{P}$ is a d–dimensional vector space including the tangent vectors to all smooth curves passing through $p(x;\theta)$. It can be represented as the linear space spanned by the basis vectors $\ell_i(x;\theta) = \partial_i \ell(x;\theta)/\partial\theta^i$, $i = 1, \dots, d$, where $\ell(x;\theta) = \log p(x;\theta)$. Thus, any tangent vector $A(x) \in T_\theta$ can be represented as a linear combination of ℓ_i, i.e.

$$A(x) = \sum_{i=1}^{d} A^i \ell_i(x;\theta) = A^i \ell_i(x;\theta),$$

where A^i are the components of $A(x)$ with respect to the basis ℓ_i, $i = 1, \dots, d$. Here and in the sequel, we make use of Einstein summation convention, so that when an index appears twice in an expression, summation on that index is intended.

When the tangent space T_θ is provided with an inner product $\langle \cdot, \cdot \rangle$, the manifold is said to be a *Riemannian space*. An inner product between two vectors $A(x)$ and $B(x)$ belonging to T_θ can always be expressed in terms of inner products between pairs of basis vectors:

$$\langle A(\cdot), B(\cdot) \rangle = A^i B^j \langle \ell_i(\cdot;\theta), \ell_j(\cdot;\theta) \rangle,$$

where B^j are the components of $B(x)$ with respect to the basis vectors. The d^2 quantities $\langle \ell_i(\cdot;\theta), \ell_j(\cdot;\theta) \rangle$, $i, j = 1, \dots, d$, define a *Riemannian metric* in $\mathcal{P}$.

An inner product can be naturally defined on the tangent spaces of a statistical manifold by letting

$$\langle \ell_i(\cdot;\theta), \ell_j(\cdot;\theta) \rangle = E_\theta[\ell_i(X;\theta)\ell_j(X;\theta)] = g_{ij}(\theta),$$

where E_θ denotes the expected value with respect to the distribution $p(x;\theta)$. Notice that $g_{ij}(\theta)$ are the components of the Fisher information matrix and constitute the so-called *Fisher metric*.

In order to compare vectors belonging to different tangent spaces, we need to establish a correspondence between two adjacent tangent spaces. This is done by means of an *affine connection*. An affine connection is a sort of derivative, called *covariant derivative* and denoted by ∇, expressing the intrinsic change in

a tangent vector as the point θ changes to $\theta' = \theta + d\theta$ in some direction along the surface $\mathcal{P}$. In particular, d^3 quantities are needed to define the k-th component of the change in the basis vector ℓ_j when we move along the direction of ℓ_i, $i, j, k = 1, \ldots, d$:

$$\langle \nabla_{\ell_i(\cdot;\theta)} \ell_j(\cdot;\theta), \ell_k(\cdot;\theta) \rangle = \Gamma_{ijk}(\theta).$$

The quantities Γ_{ijk} are said to be the coefficients or the Christoffel symbols of the affine connection. Affine connections allow to describe important properties of a statistical manifold embedded in the set of all probability distributions on $(\chi, \mathcal{A})$, such as curvature, torsion and flatness.

A whole class of affine connections is naturally introduced in a statistical manifold by considering the so-called *Amari's α-connections*, whose Christoffel symbols are defined as

$$\overset{\alpha}{\Gamma}_{ijk}(\theta) = E_\theta \left[\left(\ell_{ij}(X;\theta) + \frac{1-\alpha}{2} \ell_i(X;\theta) \ell_j(X;\theta) \right) \ell_k(X;\theta) \right],$$

where α is a scalar parameter and $\ell_{ij}(x;\theta) = \partial_i \partial_j \ell(x;\theta) / \partial\theta^i \partial\theta^j$, $i, j = 1, \ldots, d$.

Consider now a parametric model $\mathcal{M} = \{p(x;\omega) : \omega \in \Omega \subseteq \mathbb{R}^r\}$, $r > d$, including $\mathcal{P}$. We say that $\mathcal{P}$ is a submanifold embedded in $\mathcal{M}$ if there exists a smooth and full rank mapping $\omega = \omega(\theta)$ from $\mathcal{P}$ to $\mathcal{M}$. As a consequence, the tangent space $T_\theta^{\mathcal{P}}$ at point $p(x;\theta)$ in $\mathcal{P}$ is a subspace of the tangent space $T_{\omega(\theta)}^{\mathcal{M}}$ at point $p(x;\omega(\theta))$ in $\mathcal{M}$.

Assume that both $\mathcal{P}$ and $\mathcal{M}$ are provided with a Riemannian metric and a covariant derivative, that we denote by $\langle \cdot, \cdot \rangle$ and ∇ with superscript $\mathcal{P}$ and $\mathcal{M}$, respectively. Notice that $\mathcal{P}$ naturally inherits the geometric structure defined in $\mathcal{M}$, by projection of $\langle \cdot, \cdot \rangle^{\mathcal{M}}$ and $\nabla^{\mathcal{M}}$. Anyway, the metric and covariant derivative induced by $\mathcal{M}$ in $\mathcal{P}$ do not necessarily coincide with those previously defined in $\mathcal{P}$. When they coincide, we say that $\langle \cdot, \cdot \rangle$ and ∇ are *embedding invariant* metric and covariant derivative, respectively.

Fisher metric and Amari's α-connections play a fundamental role in the theory of statistical manifolds, due to their property of invariance with respect to one-to-one transformations in the sample space (see [3]). Moreover, as [8] proved, they are the unique invariant metric and connections with respect to Markov embeddings on finite sample spaces.

3. The geometry of divergences

In this section we recall the concepts of divergence and geometry that a divergence induces on a parametric model. We show that the Riemannian metric and affine connections associated to a divergence, characterise its local behaviour up to third order. Moreover, we discuss in detail the geometric properties of divergences obtained from proper and weighted scoring rules. We refer the reader to [3], [4], [8], [9], [11], [14], [17], and [18] for deeper discussions on the subject.

Let χ be a set and $\mathcal{A}$ be a σ-algebra of subsets of χ. A *divergence* D is a non-negative function, whose arguments are two probability measures defined

on the measurable space $(\chi, \mathcal{A})$, such that

$$D(P,Q) = 0 \quad \Longleftrightarrow \quad P = Q.$$

Let $\mathcal{P}$ be a parametric model, $\mathcal{P} = \{p(x;\theta) : \theta \in \Theta \subseteq \mathbb{R}^d\}$, with $p(x;\theta)$ densities with respect to some dominating measure μ on $(\chi, \mathcal{A})$. As we have seen in the previous section, $\mathcal{P}$ can be regarded as a d-dimensional manifold embedded in the set of all probability distributions on $(\chi, \mathcal{A})$, and the parametrisation $\theta = (\theta^1, \ldots, \theta^d)$ plays the role of a system of local coordinates. Thus, the divergence between $p(x;\theta)$ and $p(x;\theta')$ can be written as a function of θ and θ', $D(\theta, \theta')$.

We can derive the geometry induced by a divergence D on $\mathcal{P}$, by differentiating $D(\theta, \theta')$ with respect to θ and θ'. Indices on D indicate derivatives of D with respect to the components of the two arguments θ and θ' separated by a semicolon, so that $D_{;i}(\theta, \theta') = \partial D(\theta, \theta')/\partial \theta'^{\,i}$, $D_{;ij}(\theta, \theta') = \partial^2 D(\theta, \theta')/\partial \theta'^{\,i} \partial \theta'^{\,j}$, $D_{ij;k}(\theta, \theta') = \partial^3 D(\theta, \theta')/\partial \theta^i \partial \theta^j \partial \theta'^{\,k}$, and so on.

Now, we can define a Riemannian metric as

$$g^D_{ij}(\theta) = D_{;ij}(\theta, \theta),$$

and a family of affine connections whose Christoffel symbols are given by

$$\overset{\beta}{\Gamma}{}^{D}_{ijk}(\theta) = -\frac{1-\beta}{2} D_{ij;k}(\theta, \theta) - \frac{1+\beta}{2} D_{k;ij}(\theta, \theta), \quad \beta \in \mathbb{R}.$$

Notice that

$$\overset{-1}{\Gamma}{}^{D}_{ijk}(\theta) = -D_{ij;k}(\theta, \theta) \quad \text{and} \quad \overset{1}{\Gamma}{}^{D}_{ijk}(\theta) = -D_{k;ij}(\theta, \theta).$$

The 1 and -1-connections play an important role in the geometrical approach to inference, as pointed out in [14] and [17].

The local behaviour of D in a neighbourhood of θ can be expressed in terms of the metric and affine connections induced by D on $\mathcal{P}$. Indeed,

$$\begin{aligned} D(\theta, \theta') &= D(\theta, \theta) + D_{;i}(\theta, \theta)(\theta' - \theta)^i + \frac{1}{2} D_{;ij}(\theta, \theta)(\theta' - \theta)^{ij} \\ &\quad + \frac{1}{6} D_{;ijk}(\theta, \theta)(\theta' - \theta)^{ijk} + o(|\theta' - \theta|^3) \\ &= \frac{1}{2} g^D_{ij}(\theta)(\theta' - \theta)^{ij} \\ &\quad + \frac{1}{6} \left[2\, \overset{1}{\Gamma}{}^{D}_{ijk}(\theta) + \overset{-1}{\Gamma}{}^{D}_{ijk}(\theta) \right] (\theta' - \theta)^{ijk} + o(|\theta' - \theta|^3) \\ &= \frac{1}{2} \left[g^D_{ij}(\theta)(\theta' - \theta)^{ij} + \overset{1/3}{\Gamma}{}^{D}_{ijk}(\theta)(\theta' - \theta)^{ijk} \right] + o(|\theta' - \theta|^3), \end{aligned} \tag{3.1}$$

where $(\theta' - \theta)^i$ is the i-th component of $(\theta' - \theta)$ and $(\theta' - \theta)^{ij}$ and $(\theta' - \theta)^{ijk}$ are the product of two and three components of $(\theta' - \theta)$, respectively.

3.1. Monotone and regular divergences

Monotone and regular divergences constitute a wide class of discrepancy functions useful in statistical applications. Important invariance properties of the geometry induced by monotone and regular divergences on a parametric model $\mathcal{P}$ have been derived in [8] and [9]. In particular, the Riemannian metric and the affine connections associated to such divergences are equivalent to the Fisher metric and Amari's α-connections. Indeed, monotonicity implies invariance with respect to Markov embeddings and allows the characterisation of the geometry induced by monotone divergences on statistical manifolds defined in finite dimensional sample spaces. Regularity extends this characterisation to any regular parametric model of absolutely continuous probability distributions.

Using these geometric properties, the local behaviour of monotone and regular divergences has been characterised in [9]. A Taylor expansion up to third order for a monotone and regular divergence is

$$D(\theta,\theta') = A\left[g_{ij}(\theta)(\theta'-\theta)^{ij} + \overset{-B/3}{\Gamma}_{ijk}(\theta)(\theta'-\theta)^{ijk}\right] + o(|\theta'-\theta|^3), \qquad (3.2)$$

for some constants A and B, where g_{ij} and $\overset{\alpha}{\Gamma}_{ijk}$ are respectively the components of the Fisher metric and the Christoffel symbols of Amari's α-connections in $\mathcal{P}$. The preceding expansion is a particular case of the general equation (3.1), based on the fact that monotone and regular divergences induce an invariant geometric structure on the parametric family where they are defined.

3.1.1. Examples

Well known examples of monotone and regular divergences are Csiszar's ϕ-*divergences*, defined as

$$D^{\phi}(\theta,\theta') = \int \phi\left(\frac{p(x;\theta')}{p(x;\theta)}\right) p(x;\theta)\mu(dx),$$

where ϕ is a strictly convex function with $\phi(1) = 0$; see [12]. Notice that for a ϕ-divergence the constants characterising Taylor expansion (3.2) are $A = A^{\phi} = \phi''(1)/2$ and $B = B^{\phi} = 2\phi'''(1) + 4\phi''(1) - 1$.

By letting

$$\phi(z) = \phi_{\alpha}(z) = \begin{cases} \frac{4}{1-\alpha^2}[1 - z^{\frac{1+\alpha}{2}}] & \alpha \neq \pm 1 \\ z\log z & \alpha = 1 \\ -\log z & \alpha = -1, \end{cases}$$

we obtain the family of α-*divergences* that include the Kullback-Leibler divergence ($\alpha = -1$) and twice the Hellinger distance ($\alpha = 0$) as special cases. For an α-divergence the constants in (3.2) reduce to $A = A^{\phi_\alpha} = 1/2$ and $B = B^{\phi_\alpha} = \alpha$. So the -1-connection derived from an α-divergence is exactly Amari's α-connection.

It may also be useful to consider a general family of divergences obtained by transforming a ϕ-divergence with a differentiable and increasing function h. This is the class of (h,ϕ)-*divergences* which has already been studied; see for instance [36] and [39]. It is easy to see that (h,ϕ)-divergences are monotone and regular.

Proposition 3.1. *(h,ϕ)-divergences are monotone and regular divergences.*

Proof. Let $D^{\phi}(P,Q)$ be a ϕ-divergence and $D^{h\phi}(P,Q)=h\left(D^{\phi}(P,Q)\right)$ an (h,ϕ)-divergence. Monotonicity of a ϕ-divergence means that, for every Markov embedding K and every P and Q probability measures on $(\chi,\mathcal{A})$, we can write $D^{\phi}(K(P),K(Q))\leq D^{\phi}(P,Q)$; see [9]. Then monotonicity of $D^{h\phi}$ follows from

$$D^{h\phi}(K(P),K(Q))=h(D^{\phi}(K(P),K(Q)))\leq h(D^{\phi}(P,Q))=D^{h\phi}(P,Q),$$

since h is an increasing function. Moreover, $D^{h\phi}(P,Q)$ is regular too, being a continuous transformation of a regular functional; see [8]. □

As a consequence, on a regular parametric model $\mathcal{P}$, (h,ϕ)-divergences can be expanded in a neighbourhood of θ as in formula (3.2). The constants $A=A^{h\phi}$ and $B=B^{h\phi}$ can be easily calculated by observing that second and third order derivatives of $D^{h\phi}$ calculated on the diagonal are respectively

$$D^{h\phi}_{;ij}(\theta,\theta)=h'(0)D^{\phi}_{;ij}(\theta,\theta),\qquad D^{h\phi}_{ij;k}(\theta,\theta)=h'(0)D^{\phi}_{ij;k}(\theta,\theta)$$

and

$$D^{h\phi}_{k;ij}(\theta,\theta)=h'(0)D^{\phi}_{k;ij}(\theta,\theta).$$

Thus, the metric associated to a (h,ϕ)-divergence has components

$$g^{h\phi}_{ij}=D^{h\phi}_{;ij}(\theta,\theta)=2h'(0)A^{\phi}g_{ij}(\theta),$$

where g_{ij} is the Fisher metric and $A^{\phi}=\phi''(1)/2$ is the constant A in (3.2) associated to the ϕ-divergence. Moreover, the 1/3 affine connection associated to a (h,ϕ)-divergence has components

$$\overset{1/3}{\Gamma}{}^{h\phi}_{ijk}=h'(0)\overset{1/3}{\Gamma}{}^{\phi}_{ijk}=2h'(0)A^{\phi}\overset{-B^{\phi}/3}{\Gamma}{}_{ijk},$$

where $\overset{\alpha}{\Gamma}_{ijk}$ denotes the Christoffel symbols of Amari's α-connections and $B^{\phi}=2\phi'''(1)+4\phi''(1)-1$ is the constant B associated to the ϕ-divergence. By substituting in (3.1) we obtain expansion (3.2) with $A=A^{h\phi}=h'(0)A^{\phi}$ and $B=B^{h\phi}=B^{\phi}$.

3.2. Score divergences

Scoring rules are loss functions that measure the quality of a proposed probability distribution Q for a random variable X taking values on χ, in view of the outcome x of X. Specifically, if a forecaster quotes a predictive distribution Q

for X and the event $X = x$ realises, then we can measure the loss using a score function $S(x, Q)$; see [22].

The expected value of $S(X, Q)$ when X has distribution P is called the *expected score*: $S(P, Q) = E_P[S(X, Q)]$. In a frequentist approach, $S(P, Q)$ suggests a way for evaluating the performance of Q when P is the true distribution for X: the smaller $S(P, Q)$, the better Q as an estimate of P. From this viewpoint, a natural feature for a good scoring rule is that of propriety; see for example [15] and references therein. A scoring rule S is proper relative to $\mathcal{P}$ if $S(P, Q) \geq S(P, P)$, for all P, $Q \in \mathcal{P}$. It is strictly proper if equality holds only when $P = Q$.

Each proper scoring rule is associated with two functions, the *entropy function* $H(P) = \inf_{Q \in \mathcal{P}} S(P, Q)$ and the *divergence function*

$$D(P, Q) = S(P, Q) - S(P, P).$$

Indeed, it is easy to see that S is proper if and only if $D(P, Q) \geq 0$ for all $P, Q \in \mathcal{P}$, and S is strictly proper if and only if in addition $D(P, Q) = 0$ implies $P = Q$.

A divergence D induced by a proper scoring rule is called a *score divergence*.

3.2.1. Examples

Assume that P and Q have respectively densities p and q with respect to some dominating measure μ on $(\chi, \mathcal{A})$. Notice that the majority of scoring rules presented in this section deals with continuous random variables, however for each of these scoring rules we can also define a categorical counterpart by considering μ as a discrete measure; see for example the Brier score, the discrete counterpart of the quadratic score.

The *logarithmic score* is a well-known scoring rule (see [24]), defined as

$$S(x, Q) = -\log q(x).$$

It is easily seen that the corresponding divergence is the *Kullback-Leibler divergence*

$$D(P, Q) = \int \log\left(\frac{p(x)}{q(x)}\right) p(x)\mu(dx).$$

The *Tsallis score* (see [41]), also known as the generalised power score, is given by

$$S(x, Q) = (\gamma - 1) \int q(y)^\gamma \mu(dy) - \gamma q(x)^{\gamma - 1}, \quad \gamma > 1.$$

The associated divergence is

$$D(P, Q) = \int p(x)^\gamma \mu(dx) + (\gamma - 1) \int q(x)^\gamma \mu(dx) - \gamma \int p(x) q(x)^{\gamma - 1} \mu(dx).$$

For $\gamma = 2$, the Tsallis score reduces to the *quadratic score*

$$S(x, Q) = \int q(y)^2 \mu(dy) - 2q(x).$$

In the special case of a finite sample space χ, an equivalent rule is the *Brier score* (see [7])

$$S(x, Q) = (1 - q(x))^2 + \sum_{y \neq x} q(y)^2.$$

The associated divergence is the square of the Euclidean distance between two probability functions

$$D(P, Q) = \sum_x (p(x) - q(x))^2 .$$

The *Bregman score* is among the most used families of scoring rules; see [14] and [15]. It is defined as

$$S(x, Q) = -\psi'(q(x)) - \int [\psi(q(y)) - q(y)\psi'(q(y))]\, \mu(dy),$$

where ψ is a convex and differentiable function. The associated divergence can be written as

$$D(P, Q) = \int [\psi(p(x)) - \psi(q(x)) - \psi'(q(x))(p(x) - q(x))]\, \mu(dx). \tag{3.3}$$

Notice that the logarithmic, the Tsallis and the Brier scores are all special cases of the Bregman score with $\psi(p) = p \log p$, $\psi(p) = p^\gamma$ and $\psi(p) = (2p^2 - 1)/4$, respectively.

The *pseudo-spherical score* (see [24]) is an example of scoring rule which does not belong to the class of Bregman scores. Let $||p||_\gamma = \left(\int p(x)^\gamma \mu(dx)\right)^{1/\gamma}$. The pseudo-spherical score is defined as

$$S(x, Q) = -\left(\frac{q(x)}{||q||_\gamma}\right)^{\gamma - 1}, \quad \gamma > 1.$$

The divergence associated to the pseudo-spherical score is

$$\begin{aligned} D(P, Q) &= ||q||_\gamma^{1-\gamma} \left(||q||_\gamma^{\gamma-1} ||p||_\gamma - \int p(x) q(x)^{\gamma-1} \mu(dx) \right) \\ &= ||q||_\gamma^{1-\gamma} \left(||q||_\gamma^{\gamma-1} ||p||_\gamma - E_p[q^{\gamma-1}(X)] \right). \end{aligned} \tag{3.4}$$

For $\gamma = 2$ we obtain, as a special case, the *spherical score*

$$S(x, Q) = -\frac{q(x)}{||q||_2},$$

with associated divergence

$$D(P, Q) = ||q||_2^{-1} \left(||q||_2 ||p||_2 - E_p[q(X)] \right).$$

3.2.2. The geometry of score divergences

Now we present some new results on the geometry of score divergences. These results allow to characterise the local behaviour of such divergences, using expansion (3.1).

The geometry associated with score divergences has already been considered in [14] in the context of a decision problem, with the name of decision geometry. Here we extend some of those results to the multi-parametric case and we characterise the Taylor expansion of score divergences up to a higher order.

In particular, we derive the geometry induced by the Bregman and the pseudo-spherical divergences. We will see that it does not coincide with the geometry of the Fisher metric and Amari's α-connections, except for the special case of Kullback-Leibler divergence. This proves that, in fact, these divergences are not monotone and regular.

In the rest of this paragraph, we consider a regular parametric model $\mathcal{P} = \{p(x;\theta) : \theta \in \Theta \subseteq \mathbb{R}^d\}$. Moreover we use the same notation previously introduced, so that $\ell(x;\theta) = \log p(x;\theta)$ and indices mean derivatives with respect to the different components of $\theta = (\theta^1, \dots, \theta^d)$.

Bregman divergences By differentiating (3.3), we obtain the Riemannian metric associated to the Bregman divergence:

$$g^D_{ij}(\theta) = \int \psi''(p(x;\theta))\ell_i(x;\theta)\ell_j(x;\theta)p^2(x;\theta)\,\mu(dx),$$

and the corresponding affine connections:

$$\begin{aligned}\overset{\beta}{\Gamma}{}^D_{ijk}(\theta) &= \frac{1+\beta}{2}\int \psi'''(p(x;\theta))\ell_i(x;\theta)\ell_j(x;\theta)\ell_k(x;\theta)p^3(x;\theta)\,\mu(dx) + \\ &+ \int \psi''(p(x;\theta))\ell_k(x;\theta)\left(\ell_{ij}(x;\theta) + \ell_i(x;\theta)\ell_j(x;\theta)\right)p^2(x;\theta)\,\mu(dx).\end{aligned}$$

The metric coincides with the decision metric given in [14] for the case of a scalar parameter.

Proposition 3.2. *The Kullback-Leibler divergence is the only Bregman divergence that is also monotone and regular.*

Proof. The proof easily follows by noticing that the above metric and affine connections recover the Fisher metric and Amari's α-connections, respectively, only when $\psi(p) = p\log p$, i.e. when the Bregman score reduces to the logarithmic score. □

As a special case of the Bregman divergence, we consider the Tsallis divergence which is obtained for $\psi(p) = p^\gamma$. In the following, to simplify the notation, we write

$$\begin{aligned}P_i(\theta) &= \int p^{\gamma-1}(x;\theta)p_i(x;\theta)\,\mu(dx),\\ P_{ij}(\theta) &= \int p^{\gamma-1}(x;\theta)p_{ij}(x;\theta)\,\mu(dx),\end{aligned}$$

$$P_{i,j}(\theta) = \int p^{\gamma-2}(x;\theta)p_i(x;\theta)p_j(x;\theta)\,\mu(dx),$$
$$P_{ij,k}(\theta) = \int p^{\gamma-2}(x;\theta)p_{ij}(x;\theta)p_k(x;\theta)\,\mu(dx),$$
$$P_{i,j,k}(\theta) = \int p^{\gamma-3}(x;\theta)p_i(x;\theta)p_j(x;\theta)p_k(x;\theta)\,\mu(dx).$$

The metric is

$$g^D_{ij}(\theta) = \gamma(\gamma-1)P_{i,j}(\theta)$$

and the affine connections have components

$$\overset{\beta}{\Gamma}{}^D_{ijk}(\theta) = \gamma(\gamma-1)\left[\left(\frac{(1+\beta)\gamma}{2}-\beta\right)P_{i,j,k}(\theta)+P_{ij,k}(\theta)\right].$$

Then, the Taylor expansion up to third order of Tsallis divergence is

$$\begin{aligned} D(\theta,\theta') &= \frac{\gamma(\gamma-1)}{2}\{P_{i,j}(\theta)(\theta'-\theta)^{ij} \\ &\quad + \left(\frac{2\gamma-1}{3}P_{i,j,k}(\theta)+P_{ij,k}(\theta)\right)(\theta'-\theta)^{ijk}\}+o(|\theta'-\theta|^3). \end{aligned}$$

In a similar way we can find the expansion for the Brier divergence.

Pseudo-spherical divergences We now present the geometry associated to the pseudo-spherical divergence, using the same notation of the previous case. Moreover, let $||p(\theta)||_\gamma = \left(\int p(x;\theta)^\gamma \mu(dx)\right)^{1/\gamma}$. By straightforward differentiation of (3.4), we obtain the metric, which coincides with [14] for a scalar parameter,

$$g^D_{ij}(\theta) = (\gamma-1)||p(\theta)||_\gamma^{1-2\gamma}\left(||p(\theta)||_\gamma^\gamma P_{i,j}(\theta)-P_i(\theta)P_j(\theta)\right),$$

and the affine connections

$$\begin{aligned} \overset{\beta}{\Gamma}{}^D_{ijk}(\theta) &= (\gamma-1)||p(\theta)||_\gamma^{1-2\gamma}\{-\frac{(1-\beta)}{2}\left(P_{ij}(\theta)P_k(\theta)-P_{ij,k}(\theta)\right) \\ &\quad -\frac{(1+\beta)}{2}\left[||p(\theta)||_\gamma^{-\gamma}(1-2\gamma)P_i(\theta)P_j(\theta)P_k(\theta)\right. \\ &\quad \left.-||p(\theta)||_\gamma^\gamma\left((\gamma-2)P_{i,j,k}(\theta)+P_{ij,k}(\theta)\right)\right] \\ &\quad -\frac{(1+\beta)}{2}\left[(\gamma-1)\left(P_i(\theta)P_{j,k}(\theta)+P_j(\theta)P_{k,i}(\theta)+P_k(\theta)P_{i,j}(\theta)\right)\right. \\ &\quad \left.+P_{ij}(\theta)P_k(\theta)\right]\}. \end{aligned}$$

We can thus state the following result.

Theorem 3.1. *Pseudo-spherical divergences are not monotone and regular.*

Proof. The proof easily follows by noticing that the above metric and affine connections always differ from the Fisher metric and Amari's α-connections, respectively. □

Now, the Taylor expansion up to third order of the pseudo-spherical divergence is

$$\begin{aligned} D(\theta,\theta') &= (\gamma-1)||p(\theta)||_\gamma^{1-2\gamma}\left\{\left(||p(\theta)||_\gamma^\gamma P_{i,j}(\theta)-P_i(\theta)P_j(\theta)\right)(\theta'-\theta)^{ij}\right. \\ &\quad+\left[-\frac{1}{3}\left(P_{ij}(\theta)P_k(\theta)-P_{ij,k}(\theta)\right)\right. \\ &\quad-\frac{2}{3}\left((\gamma-1)\left(P_i(\theta)P_{j,k}(\theta)+P_j(\theta)P_{k,i}(\theta)+P_k(\theta)P_{i,j}(\theta)\right)\right. \\ &\quad\left.+P_{ij}(\theta)P_k(\theta)\right) \\ &\quad+\frac{2}{3}||p(\theta)||_\gamma^\gamma\left((\gamma-2)P_{i,j,k}(\theta)+P_{ij,k}(\theta)\right) \\ &\quad\left.-\frac{2}{3}||p(\theta)||_\gamma^{-\gamma}\left(1-2\gamma\right)P_i(\theta)P_j(\theta)P_k(\theta)\right](\theta'-\theta)^{ijk}\Big\}+o(|\theta'-\theta|^3). \end{aligned}$$

3.2.3. The relationship between monotone and regular divergences and score divergences

In the previous section, we have proved that pseudo-spherical divergences and Bregman divergences, except for the Kullback-Leibler divergence, are not monotone and regular. This has been done by showing that the geometry induced by these divergences on a regular parametric model is different from the geometry defined by the Fisher metric and Amari's α-connections.

Now we prove that (h,ϕ)-divergences are not score divergences, i.e. it does not exist a proper scoring rule with associated divergence in the class of (h,ϕ)-divergences. The only exception to this statement is the logarithmic score with the Kullback-Leibler divergence. This result has already been conjectured by Dawid in [14] for α-divergences, $\alpha\neq-1$, but not proved.

Theorem 3.2. *Increasing linear transformations of the Kullback-Leibler divergence are the only (h,ϕ)-divergences that are also score divergences.*

Proof. Consider P, Q and R probability distributions defined on $(\chi,\mathcal{A})$, with densities p, q and r with respect to some dominating measure μ.

It is shown in [13], Section 11, that if D is a score divergence then $D(P,Q)-D(P,R)$ can be written as an affine function of P for any fixed Q and R. This means that there exists a function $f(x;Q,R)$ such that $D(P,Q)-D(P,R)=E_P[f(X;Q,R)]$.

Now, let $D^{h\phi}$ be a (h,ϕ)-divergence. We can write

$$D^{h\phi}(P,Q)-D^{h\phi}(P,R) = h\left(E_p[\phi(q(X)/p(X))]\right)-h\left(E_p[\phi(r(X)/p(X))]\right).$$

It is easy to see that this expression is an affine function of P if and only if h is an increasing linear function and ϕ is a linear transformation of the logarithm. Indeed, note that if

$$E_p[\phi(q(X)/p(X))] - E_p[\phi(r(X)/p(X))] = E_p[f(X; q(X), r(X))]$$

then the function $\phi(q/p) - \phi(r/p)$ must be independent of p. Thus, $\partial\phi(q/p)/\partial p = \partial\phi(r/p)/\partial p$, so that $\partial\phi(q/p)/\partial p$ does not depend on q. These considerations lead to the following differential equation: $\phi''(t)t + \phi'(t) = 0$, which admits the solution $\phi(t) = A\log(t) + B$, with A and B two suitable constants. □

Finally, as a consequence of the previous results, we can state that the Kullback-Leibler divergence is essentially the unique intersection between the classes of (h, ϕ)-divergences and Bregman divergences. Actually, we believe that this result is even more general, but this is only a conjecture.

Conjecture. *The Kullback-Leibler divergence is the unique intersection between the class of monotone and regular divergences and the class of score divergences.*

3.3. Weighted score divergences

A different way for deriving divergences from scoring rules relies on the use of weighted scoring rules that have been introduced in [30] and represent an interesting generalisation of the above defined (unweighted) scoring rules. Weighted scoring rules depend on a baseline probability distribution P, which represents the available information at the time of prediction. Thus, the expected weighted scoring rule allows to compare two forecasts R and Q, taking P as the reference distribution. It is a function of three arguments: $S(Q, R||P)$. The entropy function associated to a weighted scoring rule, $H(Q||P) = S(Q, Q||P)$ can be used to define a *weighted divergence* between P and Q:

$$D^w(P, Q) = H(Q||P) - H(P||P). \tag{3.5}$$

As stated in [19], the entropy $H(Q||P)$ is minimized only when the forecaster's true belief Q is exactly the baseline distribution P. More formally, a baseline distribution P is defined as the unique distribution which minimizes the entropy, i.e. $H(P||P) \leq H(Q||P)$, for all $Q \in \mathcal{P}$. Thus, it follows that $D^w(P, Q) \geq 0$, with equality holding if and only if $P = Q$.

It is important noticing that weighted divergences are defined from weighted scoring rules by means of the corresponding entropy function. This construction is different from the one used for score divergences.

The dependence of the scoring rule on the non-uniform baseline distribution P can be introduced in different ways, depending on the problem under consideration. Here we focus on scoring rules that depend on the ratio between the considered probability distributions and the baseline P; see [30]. Such a dependence implies that the associated divergence (3.5) measures the distance between two probability distributions P and Q in terms of the ratio Q/P. From our point of view, this condition results in divergences that are invariant with respect to one-to-one transformations of the sample space.

3.3.1. Examples

This section deals with some examples of weighted scoring rules and associated divergences. The *weighted power score* (see [31]), also known as the weighted Tsallis score, is a generalisation of the power score, defined as

$$S^w(x,Q||P) = \frac{(q(x)/p(x))^{\gamma-1}-1}{\gamma-1} - \frac{E_p[(q(X)/p(X))^{\gamma}]-1}{\gamma}.$$

The associated weighted power divergence is

$$D^w(P,Q) = \frac{E_p[(q(X)/p(X))^{\gamma}]-1}{\gamma(\gamma-1)},$$

which is a case of ϕ-divergence with $\phi(x) = (x^{\gamma}-1)/(\gamma(\gamma-1))$.

When $\gamma \to 1$ the weighted power score reduces to the weighted logarithmic score. For $\gamma = 2$, we obtain the weighted Brier scoring rule. For more examples and details we refer to [30] and [31].

As already pointed out, for $\alpha \neq -1$ α-divergences are not score divergences, that is they are not associated to an unweighted scoring rule. Anyway, for $\alpha \neq \pm 1$ they can be easily obtained from the following class of weighted scoring rules:

$$S^w(x,Q||P) = \frac{4}{1-\alpha^2}\left(1-(q(x)/p(x))^{\frac{\alpha-1}{2}}\right),$$

which is equivalent, up to a linear transformation, to the weighted power score with $\gamma = (1+\alpha)/2$.

The *weighted pseudo-spherical score* (see [31]) is defined as

$$S^w(x,Q||P) = \frac{1}{\gamma-1}\left[\frac{(q(x)/p(x))^{\gamma-1}}{(E_p[(q(X)/p(X))^{\gamma}])^{(\gamma-1)/\gamma}} - 1\right].$$

The associated divergence is

$$D^w(P,Q) = \frac{(E_p[(q(X)/p(X))^{\gamma}])^{1/\gamma}-1}{\gamma-1}.$$

It can be easily shown that it is monotonically related to the weighted power divergence through the function $h(x) = [((\gamma-1)x+1)^{\gamma}-1]/[\gamma(\gamma-1)]$.

Note that for $\gamma \to 1$, the weighted logarithmic score represents a limiting case for the weighted pseudo-spherical score too.

The family of *quasi-Bregman weighted scoring rules* has recently been defined by [19] for a random variable X taking values in a finite sample space $\chi = \{x_1,\ldots,x_n\}$. It is associated to the following class of divergences:

$$\begin{aligned} D^w(P,Q) &= l\left(\sum_{i=1}^n f(p_i)g\left(\frac{q_i}{p_i}\right)\right) - g'(1)\sum_{i=1}^n \frac{q_i f(p_i)}{p_i} l'\left(g(1)\sum_{i=1}^n f(p_i)\right) \\ &- l\left(g(1)\sum_{i=1}^n f(p_i)\right) + g'(1)\sum_{i=1}^n f(p_i) l'\left(g(1)\sum_{i=1}^n f(p_i)\right), \end{aligned}$$

where $p = (p_1, \ldots, p_n)$, $q = (q_1, \ldots, q_n)$, $p_i = P(\{x_i\})$ and $q_i = Q(\{x_i\})$, $i = 1, \ldots, n$. Moreover, the function f is positive, g is twice differentiable and strictly convex, l is twice differentiable and strictly increasing and g' and l' denote the derivatives of g and l, respectively.

The family of quasi-Bregman weighted scoring rules has been defined only for finite sample spaces, but in some cases it can be extended to continuous random variables. In fact, notice that the weighted power and pseudo-spherical scores are special instances of the quasi-Bregman weighted scoring rules with $f(x) = x$.

3.3.2. The geometry of quasi-Bregman weighted score divergences

In this section we consider the divergences associated to the quasi-Bregman weighted scoring family characterised by $f(x) = x$, namely

$$D^w(P,Q) = l\left(E_p\left[g\left(\frac{q(X)}{p(X)}\right)\right]\right) - l\left(g(1)\right). \tag{3.6}$$

We show that, under suitable conditions on g and l, they are monotone and regular. As a consequence, their local behaviour is characterised by (3.2) with the Fisher metric, Amari's α-connections and suitable constants A and B.

Theorem 3.3. *Let g be a twice differentiable and strictly convex function and let $l(x) = (x - g(1))/g''(1)$. Then the divergence $D^w(P,Q)$ defined in (3.6) is a ϕ-divergence, with $\phi(x) = (g(x) - g(1))/g''(1)$.*

Proof. First, notice that $l(x) = (x - g(1))/g''(1)$ is a strictly increasing linear function, since $g''(x) > 0$ for all x. Thus,

$$\begin{aligned} D^w(P,Q) &= l\left(E_p\left[g\left(\frac{q(X)}{p(X)}\right)\right]\right) - l(g(1)) \\ &= \frac{1}{g''(1)}\left(E_p\left[g\left(\frac{q(X)}{p(X)}\right)\right] - g(1)\right) \\ &= E_p\left[\left(\frac{g\left(q(X)/p(X)\right) - g(1)}{g''(1)}\right)\right] = E_p\left[\phi\left(\frac{q(X)}{p(X)}\right)\right]. \end{aligned}$$

Since $\phi(x) = (g(x) - g(1))/g''(1)$ is a convex function such that $\phi(1) = 0$, it follows that $D^w(P,Q)$ is a ϕ-divergence. □

The previous result can be generalised as follows.

Theorem 3.4. *Let l be strictly increasing, g strictly convex and both twice differentiable, such that $l(g(1)) = 0$. Then the divergence $D^w(P,Q)$ defined in (3.6) is a (h,ϕ)-divergence, with $h(x) = l(g''(1)x + g(1))$ and $\phi(x) = (g(x) - g(1))/g''(1)$.*

Proof. Note that the function $h(x) = l(g''(1)x + g(1))$ is strictly increasing, twice differentiable and such that $h(0) = 0$. Then, the divergence in (3.6) can be rewritten as

$$\begin{aligned} D^w(P,Q) &= l\left(E_p\left[g\left(\frac{q(X)}{p(X)}\right)\right]\right) - l(g(1)) \\ &= l\left(E_p\left[g\left(\frac{q(X)}{p(X)}\right)\right] - g(1) + g(1)\right) \\ &= l\left(g''(1)E_p\left[\frac{g(q(X)/p(X)) - g(1)}{g''(1)}\right] + g(1)\right) \\ &= l\left(g''(1)E_p\left[\phi\left(\frac{q(X)}{p(X)}\right)\right] + g(1)\right) = h\left(E_p\left[\phi\left(\frac{q(X)}{p(X)}\right)\right]\right). \end{aligned}$$

Since $\phi(x) = (g(x) - g(1))/g''(1)$ is a convex function such that $\phi(1) = 0$, it follows that $D^w(P,Q)$ is a (h,ϕ)-divergence. □

Notice that the weighted power and pseudo-spherical scores satisfy the conditions of Theorem 3.4 on l and g, thus being equivalent to (h,ϕ)-divergences.

It is easy to verify that the constants characterising expansion (3.2) are the same for the weighted pseudo-spherical and the weighted power divergence, namely $A = 1/2$ and $B = 2\gamma - 1$. This is due to the fact that, as already said, the weighted pseudo-spherical divergence is an increasing transformation of the weighted power divergence with $h(x) = [((\gamma-1)x+1)^\gamma - 1]/[\gamma(\gamma-1)]$ such that $h'(0) = 1$; see also [31].

4. The prediction problem

An important application of the new results presented in the previous sections is in the context of prediction.

Let x be an observed random sample of size n from the random vector $X = (X_1, \ldots, X_n)$ with joint distribution $p_X(x;\theta)$, where $\theta \in \Theta \subseteq \mathbb{R}^d$ is an unknown parameter. Notice that the components of X may be dependent and with different distributions, as in the examples of Section 5.2 and 5.3. The problem of prediction focuses on a future as yet unobserved random variable Y whose distribution is related to the distribution of X. In general, X and Y are dependent and we need to specify the joint distribution $p(x,y;\theta)$ or, equivalently, the conditional distribution of $Y|X$, $p(y;\theta|x)$. To simplify the analysis we confine ourselves to the case of independence where the distribution of Y belongs to a regular parametric model $\mathcal{P} = \{p(y;\theta), \theta \in \Theta \subseteq \mathbb{R}^d\}$ indexed by the same unknown parameter as $p_X(x;\theta)$. Nevertheless, the following results still apply to the case of dependence, after substituting the marginal density of Y with the conditional density of $Y|X$, as shown in the example of Section 5.2. Here we assume that the model $\mathcal{P}$ is correctly specified. An extension to the case of misspecification is possible by considering M-estimators, as suggested in [21].

A complete solution to the prediction problem can be expressed as a whole distribution for Y, depending on the observed sample x. This is called a *predictive distribution* and can be denoted by $\hat{p}(y;x)$. Notice that, in general, $\hat{p}(y;x)$ may also lie outside the model $\mathcal{P}$.

In a Bayesian framework, the conditional distribution of Y given $X = x$ can be obtained by integrating $p(y;\theta)$ with respect to the posterior distribution of θ. Though natural and appealing, the Bayesian approach may be computationally demanding for high dimensional parameters.

Following the classic approach, a predictive distribution can be obtained by replacing the parameter θ with an efficient estimate $\hat{\theta} = \hat{\theta}_n(x)$, in the density of Y. The resulting density $p(y;\hat{\theta})$ is called *estimative* or *plug-in*, and belongs to the parametric model $\mathcal{P}$. For instance, in the review paper [23] and references therein, regression models are used to obtain probabilistic forecasts in the form of full probability distributions. These papers consider estimative solutions to the prediction problem obtained by substituting the unknown regression parameters with minimum scoring rule estimates. Unfortunately, it is well known that, because of the uncertainty introduced by assuming $\theta = \hat{\theta}$, the estimative solution may give rise to inaccurate predictions especially when the dimension of the parameter θ is large compared to the sample size.

The goodness of a predictive distribution can be evaluated by studying its long run properties, in a frequentist perspective. In this respect, basically two different criteria have been considered in the literature. One considers the coverage probability of prediction intervals obtained from a predictive distribution; see among others [5], [20], [25], [32], [43]. The other measures the performance of a predictive distribution by means of a divergence from the true distribution; see for instance [10], [26], [29], [42]. Other solutions based on the concept of *predictive likelihood* have also been considered; see [6] for a comprehensive review.

Here we focus on the approach to prediction based on divergence functions. In particular, we consider divergences obtained from scoring rules.

To our knowledge, up to now scoring rules have been used only to measure the relative quality of a proposed probability distribution. Such approach allows for comparison of already known predictive distributions. In this paper we propose to use proper scoring rules as a tool for constructing new predictive distributions for the unknown of interest. The proposed predictive distributions turn out to be asymptotic modifications of the estimative distribution, obtained by minimizing divergences associated to proper and weighted scoring rules. Borrowing from [11] and using the results presented in the previous sections, we analyse the asymptotic properties of the resulting predictive distributions.

In the following sections we review some general concepts on predictive distributions that asymptotically minimize an expected divergence to the true density. First we consider the asymptotically optimal predictive distribution within a parametric model $\mathcal{M}$ containing $\mathcal{P}$. Then we discuss the existence of a global solution, independent on the particular enlargement $\mathcal{M}$ that we may consider. We will try to stress the attention on the main underlying ideas more than on complex details. For an extensive treatment, see [11] and references therein.

4.1. A locally optimal predictive distribution

A predictive distribution $\hat{p}(y;x)$ can be obtained by minimizing the expected divergence from the true distribution,

$$E_\theta[D(p(y;\theta),\hat{p}(y;X))] = \int D(p(y;\theta),\hat{p}(y;x))p_X(x;\theta)\mu(dx), \qquad (4.1)$$

in the set of all probability distributions for Y. Since the true value of the parameter θ is not known, minimization has to be carried over uniformly in θ. Unfortunately, this is not an easy task to achieve, except in some very special cases.

A restricted solution to the problem can be found by minimizing the leading terms of the asymptotic expansion of (4.1), within a parametric model $\mathcal{M}$ including $\mathcal{P}$ as a sub-model. The resulting predictive distribution belongs to the enlarged model $\mathcal{M}$.

For fixing the notation, let $\mathcal{M} = \{p(y;\omega), \omega \in \Omega \subseteq \mathbb{R}^r\}$ be any regular parametric model containing $\mathcal{P}$, with $\Omega \subseteq \mathbb{R}^r$, $r > d$. We can consider on $\mathcal{M}$ the coordinate system $\omega = (\theta, s)$, where θ^i, with $i = 1, \dots, d$ are the old coordinates on $\mathcal{P}$, and s^I, with $I = d+1, \dots, r$, are new coordinates in $\mathcal{M}$. We use indices A, B, C to indicate the coordinate system $\omega = (\theta, s)$ in $\mathcal{M}$; i, j, k ... for the components of θ in $\mathcal{P}$ and I, J, K ... for the components of s. We suppose that $s = 0$ for the points in $\mathcal{P}$ and that θ and s are orthogonal in $\mathcal{P}$, i.e. the mixed components iI of the metric tensor defined by the divergence D in $\mathcal{M}$, calculated at the points in $\mathcal{P}$, are zero. Moreover, we use the same notation of Section 2, so that $\ell(y;\omega) = \log p(y;\omega)$ and indices denote derivatives of ℓ with respect to the components of ω.

Assuming that the true value of the parameter ω in $\mathcal{M}$ is $\omega_0 = (\theta_0, 0)$ with $\hat{\theta} - \theta_0 = O_p(n^{-1/2})$, where n is the sample size, the optimal predictive density in $\mathcal{M}$ which asymptotically minimizes the expected divergence is given by

$$\hat{p}^{\mathcal{M}}(y;X) = p(y;\hat{\theta})\left[1 + \frac{1}{2n}\mathbf{i}^{ij}(\hat{\theta})\overset{-1}{\Gamma}_{ijK}(\hat{\theta},0)g^{KI}(\hat{\theta},0)h_I(y;\hat{\theta})\right] + o_p(n^{-1}), \quad (4.2)$$

where $\mathbf{i}^{ij}(\theta) = \lim_{n\to\infty} nE_\theta\left[(\hat{\theta}-\theta)^i(\hat{\theta}-\theta)^j\right]$ and $h_I(y;\theta) = \ell_I(y;\theta,s)|_{s=0}$; see [11], formula (12). Note that g and Γ refer to the geometry induced by the divergence D in $\mathcal{M}$ and g^{KI} are the components of the inverse of the matrix with components g_{KI}.

The predictive distribution (4.2) is obtained by shifting the estimative density $p(y;\hat{\theta})$ in a direction orthogonal to $\mathcal{P}$. It does not constitute a global solution to the problem of prediction, since its expression depends on the parametric model $\mathcal{M}$. Thus, a better solution could be obtained by further enlarging $\mathcal{M}$.

4.2. A global solution to the problem of prediction

The correction in (4.2) to the estimative density can be interpreted as an optimal shift of the estimative density along a direction in $\mathcal{M}$ which is orthogonal to $\mathcal{P}$.

This orthogonal correction depends on the geometry induced by the divergence D on $\mathcal{M}$.

Now, as mentioned at the end of Section 2, it should be pointed out that the Riemannian metric and the family of affine connections defined by a divergence on $\mathcal{M}$ are not necessarily compatible with those induced on $\mathcal{P}$. In fact, as shown in [3], such an invariance property characterises the Fisher metric and Amari's α-connections.

It turns out that a sufficient condition for obtaining a global solution to the prediction problem is the invariance of the metric and the connections induced by the considered divergence D. Under this condition, expression (4.2) can be written independently of $\mathcal{M}$, as

$$\begin{aligned}\hat{p}(y;X) &= p(y;\hat{\theta})\left[1+\frac{1}{2n}\mathbf{i}^{ij}(\hat{\theta})\left(\ell_{ij}(y;\hat{\theta})+\frac{1-\alpha}{2}\ell_i(y;\hat{\theta})\ell_j(y;\hat{\theta})+\right.\right.\\ &\quad\left.\left.+\frac{1+\alpha}{2}g_{ij}(\hat{\theta})-\overset{\alpha}{\Gamma}_{ijr}(\hat{\theta})g^{rk}(\hat{\theta})\ell_k(y;\hat{\theta})\right)\right]+o_p(n^{-1}) \qquad (4.3)\end{aligned}$$

(see [11], Proposition 6.1). Here g_{ij} and $\overset{\alpha}{\Gamma}_{ijk}$ are the components of the Fisher metric and Amari's α-connections and α is a constant depending on the divergence D.

Thus, when the geometry induced by the divergence D is invariant, (4.3) constitutes a global solution to the problem of prediction, in that it asymptotically minimizes the expected divergence from the true distribution in all parametric models including $\mathcal{P}$.

In Section 3, the geometric properties of divergences derived from proper scoring rules and weighted scoring rules have been investigated. We have seen that the geometry induced by both the Bregman and the pseudo-spherical divergences is not invariant, except for the only case of the Kullback-Leibler divergence. As a consequence, there is no global solution to the prediction problem based on these classes of loss functions. Anyway, using (4.2), an improvement on the estimative distribution can still be proposed in cases when the original parametric model can be embedded in a larger one. When the situation is too complicated, the Bregman and the pseudo-spherical divergences can only be used for comparison of competing predictive distributions, as shown in the numerical examples of the following section. Instead, under suitable conditions on the functions f, g and l, quasi-Bregman weighted divergences are in fact monotone and regular divergences. Their use in the context of prediction is thus equivalent to that of α-divergences, for some suitable value of α. The existence of an asymptotically optimal predictive distribution is guaranteed, using formula (4.3).

5. Examples

In this section we first consider three toy examples involving Gaussian distributions with unknown mean and known variance. Though non very useful in the

practice, they allow to analytically calculate a global optimal predictive distribution for invariant divergences. Moreover, using non-invariant divergences, a locally optimal predictive distribution can still be analytically obtained, improving on the simple estimative within the enlarged class of Gaussian distributions with unknown mean and variance.

5.1. The normal model with unknown mean

Let x be a random sample from $X = (X_1, \cdots, X_n)$ and assume that we want to predict the random variable $Y = \frac{1}{m}\sum_{i=1}^{m} X_{n+i}$, where $X_1, \ldots, X_{n+m}$ are independent and identically distributed as $N(\mu, \sigma_0^2)$, with σ_0^2 known. The maximum likelihood estimator for μ is $\hat{\mu} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, which is normal with mean μ and variance σ_0^2/n. Indices 1 and 2 refer to μ and σ, respectively. Since $Y \sim N(\mu, \sigma_0^2/m)$, it easily follows that:

$$\ell_1(y;\mu) = \frac{m(y-\mu)}{\sigma_0^2}, \quad \ell_{11}(y;\mu) = -\frac{m}{\sigma_0^2}.$$

The asymptotically optimal predictive density with respect to an α-divergence has already been derived in [11].

The weighted power and weighted pseudo-spherical divergences are asymptotically equivalent up to the second order, so that the optimal predictive density associated to both divergences is

$$\hat{p}(y;X) = \phi\left(y - \hat{\mu}, \frac{\sigma_0^2}{m}\right)\left[1 + \frac{(1-\gamma)m}{2n}\left(\frac{m(y-\hat{\mu})^2}{\sigma_0^2} - 1\right)\right] + o_p(n^{-1}),$$

which corresponds to the optimal predictive distribution obtained for an α-divergence with $\alpha = 2\gamma - 1$; see [11]. In the preceding formula and in the following examples, $\phi(\cdot, \sigma^2)$ denotes the density of a centered normal distribution with variance σ^2.

Now we want to derive the optimal predictive density associated to the Tsallis divergence. Since the geometry induced by the Tsallis divergence is not invariant, we are only able to find the optimal predictive density within an enlarged model $\mathcal{M}$. Let $\mathcal{M} = \left\{N(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma > 0\right\}$ be the extended manifold which contains $\mathcal{P} = \left\{N(\mu, \sigma_0^2) : \mu \in \mathbb{R}\right\}$. The components of the metric and the connections induced by the Tsallis divergence on $\mathcal{M}$ have already been obtained in Section 3.2.2. Here we consider only the components useful to derive the predictive distribution, specifically,

$$g_{22} = \left(\frac{2\pi\sigma^2}{m}\right)^{\frac{1-\gamma}{2}} \frac{(\gamma^2 - 2\gamma + 3)(\gamma - 1)}{\sigma^2\gamma\sqrt{\gamma}},$$
$$\overset{-1}{\Gamma}_{112} = \left(\frac{2\pi\sigma^2}{m}\right)^{\frac{1-\gamma}{2}} \frac{m(\gamma^2 - 2\gamma + 3)(\gamma - 1)}{\gamma\sigma^3\sqrt{\gamma}}.$$

The optimal predictive density in $\mathcal{M}$ with respect to the Tsallis divergence is given by

$$\hat{p}(y;X) = \phi\left(y-\hat{\mu}, \frac{\sigma_0^2}{m}\right)\left[1+\frac{m}{2n}\left(\frac{m(y-\hat{\mu})^2}{\sigma_0^2}-1\right)\right]+o_p(n^{-1}).$$

Note that it does not depend on the parameter γ. Moreover, it coincides with the global optimal solution obtained by minimizing the Kullback-Liebler divergence.

Next we turn our attention to the pseudo-spherical divergence. Let $\mathcal{M}$ be the extended manifold which contains $\mathcal{P}$ as in the case of the Tsallis divergence. The components of the metric and of the connections can be found in Section 3.2.2. As before we consider only the components useful to derive the predictive distribution, i.e.

$$g_{22} = \frac{2(\gamma-1)}{\sigma^2}\gamma^{-\frac{1}{2\gamma}-2}\left(\frac{2\pi\sigma^2}{m}\right)^{\frac{1-\gamma}{2\gamma}},$$
$$\overset{-1}{\Gamma}_{112} = \frac{3m(\gamma-1)^2}{\sigma^3}\gamma^{-\frac{1}{2\gamma}-2}\left(\frac{2\pi\sigma^2}{m}\right)^{\frac{1-\gamma}{2\gamma}}.$$

Hence the asymptotically optimal predictive density in $\mathcal{M}$ with respect to the pseudo-spherical divergence is

$$\hat{p}(y;X) = \phi\left(y-\hat{\mu}, \frac{\sigma_0^2}{m}\right)\left[1+\frac{3m(\gamma-1)}{4n}\left(\frac{m(y-\hat{\mu})^2}{\sigma_0^2}-1\right)\right]+o_p(n^{-1}).$$

It should be pointed out that it coincides with the global solution obtained by the Kullback-Liebler divergence when $\gamma = \frac{5}{3}$.

5.2. The autoregressive model

Let $\{X_n\}_{n\geq 0}$ be an autoregressive process of order 1, with $X_0 = 0$ and $X_n|X_{n-1} \sim N(\rho X_{n-1}, \sigma_0^2)$, for $n \geq 1$, where σ_0^2 is known and $|\rho| < 1$. Suppose that $X = (X_0, \ldots, X_n)$ and we want to estimate the value of the future random variable $Y = X_{n+1}$. Since X and Y are dependent, we work with the conditional density of $Y|X$, $p(y;\rho|x)$, and we denote the predictive density as $\hat{p}(y|x)$. As it is well known $\hat{\rho} = \frac{\sum_{i=1}^{n} X_{i-1}X_i}{\sum_{i=1}^{n+1} X_{i-1}^2}$ is asymptotically normal with asymptotic variance $1-\rho^2$. Let us denote by 1 and 2 the indices corresponding to ρ and σ, respectively. The derivatives of the conditional log-likelihood $l(y;\rho|x)$ with respect to ρ are

$$\ell_1(y;\rho|x) = \frac{x_n(y-\rho x_n)}{\sigma_0^2} \quad \text{and} \quad \ell_{11}(y;\rho|x) = -\frac{x_n^2}{\sigma_0^2}.$$

The optimal predictive density associated to the α-divergence is given in [11]. Our aim here is to derive the optimal predictive density associated to the Tsallis divergence within the enlarged model $\mathcal{M}$ with unknown ρ and σ^2. For this

purpose, we need the following components of the metric and the connections induced by the Tsallis divergence:

$$g_{22} = \frac{(\gamma-1)}{\sigma^2\sqrt{\gamma}}\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{\gamma-1}\frac{\gamma^2-2\gamma+3}{\gamma},$$
$$\overset{-1}{\Gamma}_{112} = \frac{x_n^2(\gamma-1)}{\sigma^3\sqrt{\gamma}}\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{\gamma-1}\frac{\gamma^2-2\gamma+3}{\gamma}.$$

We find that the optimal predictive distribution in $\mathcal{M}$ associated to the Tsallis divergence is

$$\hat{p}(y|X) = \phi(y-\hat{\rho}x_n,\sigma_0^2)\left[1+\frac{x_n^2(1-\hat{\rho}^2)}{2n\sigma_0^2}\left(\frac{(y-\hat{\rho}x_n)^2}{\sigma_0^2}-1\right)\right]+o_p(n^{-1}).$$

It is worth pointing out that even for this case it does not depend on γ and it reduces to the global optimal predictive distribution with respect to the Kullback-Liebler divergence.

5.3. The normal non-linear model

Let y be an observation from a random vector $Y=(Y_1,\ldots,Y_n)$ such that

$$Y_i = \mu(x_i,\beta)+\epsilon_i, \qquad i=1,\ldots,n,$$

where x_i is a vector of m predictors and $\beta=(\beta_1,\ldots,\beta_m)'$ is an unknown vector of parameters. We assume that the random variables ϵ_i are independent and normally distributed, with $E(\epsilon_i)=0$ and known $Var(\epsilon_i)=\sigma_0^2$. Suppose further that $\left(\frac{\partial\mu(x,\beta)}{\partial\beta}\frac{\partial\mu(x,\beta)}{\partial\beta}^T\right)^{-1} = \frac{\Sigma}{n}+O(n^{-2})$, with Σ known. Our purpose here is to predict the future observation y_{n+1} from the random variable Y_{n+1} independent of Y and defined as

$$Y_{n+1} = \mu(x_{n+1},\beta)+\epsilon_{n+1},$$

with $\epsilon_{n+1}\sim N(0,\sigma_0^2)$.

As it is well known, the maximum likelihood estimator $\hat{\beta}$ for the regression parameters β is asymptotically normal with mean β and covariance matrix $Var(\hat{\beta}) = \sigma_0^2\left(\frac{\partial\mu(x,\beta)}{\partial\beta}\frac{\partial\mu(x,\beta)}{\partial\beta}^T\right)^{-1}$, such that $\lim_{n\longrightarrow\infty} nVar(\hat{\beta}) = \sigma_0^2\Sigma$. Since $Y_{n+1}\sim N(\mu(x_{n+1},\beta),\sigma_0^2)$, it can be easily shown that

$$\begin{aligned}\ell_i(y_{n+1};\beta) &= \frac{y_{n+1}-\mu(x_{n+1},\beta)}{\sigma_0^2}\mu_i(x_{n+1},\beta),\\ \ell_{ij}(y_{n+1};\beta) &= -\frac{\mu_i(x_{n+1},\beta)\mu_j(x_{n+1},\beta)}{\sigma_0^2},\end{aligned}$$

where $\mu_i(x_{n+1},\beta)=\partial\mu(x_{n+1},\beta)/\partial\beta_i$, $i=1,\ldots,m$.

The optimal predictive density for Y_{n+1} with respect to an α-divergence can be explicitly determined using formula (4.3):

$$\begin{aligned}\hat{p}(y_{n+1};Y) &= \phi(y_{n+1}-\mu(x_{n+1},\hat{\beta}),\sigma_0^2)\\ &\quad\left[1+\frac{1-\alpha}{4n}w^T\Sigma w\left(\frac{(y_{n+1}-\mu(x_{n+1},\hat{\beta}))^2}{\sigma_0^2}-1\right)\right]+o_p(n^{-1}),\end{aligned}$$

where $w=\partial\mu(x_{n+1},\beta)/\partial\beta\Big|_{\beta=\hat{\beta}}$.

Notice that, if μ is a linear function, it reduces to the solution found in [11] for the classical linear model. This result represents a global solution to the problem of prediction with respect to α-divergences and other asymptotically equivalent divergences as those derived from weighted scoring rules in Section 3.3.

Instead, for non-invariant divergences we can only find a partial solution within some enlarged parametric model. For instance, consider the Tsallis divergence in the enlarged parametric model with both β and σ^2 unknown. The essential terms for the derivation of the optimal predictive distribution are:

$$g_{(m+1)(m+1)}=\frac{(\gamma-1)(\gamma^2-2\gamma+3)}{\gamma^{3/2}\sigma_0^{\gamma+1}(2\pi)^{(\gamma-1)/2}},$$

$$\overset{-1}{\Gamma}_{ij(m+1)}=\frac{(\gamma-1)(\gamma^2-2\gamma+3)\mu_j(x_{n+1},\hat{\beta})\mu_i(x_{n+1},\hat{\beta})}{\gamma^{3/2}\sigma_0^{\gamma+2}(2\pi)^{(\gamma-1)/2}},$$

where the index $m+1$ refers to σ^2. The local optimal predictive density for Y_{n+1} with respect to the Tsallis divergence, is

$$\begin{aligned}\hat{p}(y_{n+1};Y) &= \phi(y_{n+1}-\mu(x_{n+1},\hat{\beta}),\sigma_0^2)\\ &\quad\left[1+\frac{1}{2n}w^T\Sigma w\left(\frac{(y_{n+1}-\mu(x_{n+1},\hat{\beta}))^2}{\sigma_0^2}-1\right)\right]+o_p(n^{-1}),\end{aligned}$$

where $w=\partial\mu(x_{n+1},\beta)/\partial\beta\Big|_{\beta=\hat{\beta}}$.

As in the previous examples, the local optimal predictive density with respect to the Tsallis divergence does not depend on the parameter γ and reduces to the global optimal predictive density with respect to the Kullback-Liebler divergence.

As we have seen in the previous examples, a predictive density that improves on the estimative density with respect to a non invariant divergence can be easily found within a parametric model including the original one. Unfortunately, such an enlargement is usually difficult to specify and the analytical computation of the predictive density can involve cumbersome calculations. Anyway, we can still compare two competing predictive densities by numerically estimating the corresponding expected divergence.

In the next sections we consider two special examples for which it is possible to derive an exact predictive solution by means of predictive pivotal quantities.

The aim is to compare the performance of the estimative solution with that of the pivotal solution by using both the Tsallis and the pseudo-spherical scoring rules. We remind that the lower the score the better the prediction, since we have considered negatively oriented scoring rules. In this comparison we have taken into account also the Brier and the spherical scoring rules, both corresponding to $\gamma = 2$.

5.4. The normal model with unknown mean and variance

The following example considers a normal distribution with unknown mean and variance. Let x be a random sample from $X = (X_1, \cdots, X_n)$, where the X_i's are independent and with the same distribution $N(\mu, \sigma^2)$, $i = 1, \cdots, n$, with μ and σ^2 unknown. We want to estimate the value of a future random variable $Y \sim N(\mu, \sigma^2)$, independent of X. The maximum likelihood estimators for μ and σ^2 are $\hat{\mu} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$, respectively.

A predictive solution giving exact predictive intervals, can be obtained by means of the pivotal quantity $(Y-\hat{\mu})/\hat{\sigma}$, which has a Student T distribution with $n-1$ degrees of freedom. It is asymptotically equivalent to the optimal predictive solution obtained from the Kullback-Liebler divergence; see for instance [37]. More in general, we remind that the asymptotically optimal predictive solution with respect to an α-divergence, is a Student T distribution with $\frac{2n-2}{1-\alpha}$ degrees of freedom; [11]. A similar result also holds for the weighted power score and the weighted pseudo-spherical divergences.

We now compare the estimative and the pivotal solution by using the Tsallis and the pseudo-spherical divergences for different values of the parameter γ; the results are in given in Tables 1 and 2, respectively. As it can be seen, both the Tsallis and the pseudo-spherical scoring rules prefer the pivotal solution for every value of γ.

TABLE 1

Normal model with $\mu = 2$, $\sigma = 2$, $n = 30$. Expected Tsallis score for the estimative and the pivotal solution. Estimates are based on 10,000 Monte Carlo replications, with standard errors in parentheses.

γ	estimative	pivotal
1.05	−0.9060 (0.0001)	−0.9548 (0.0001)
1.1	−0.8226 (0.0002)	−0.9137 (0.0002)
1.2	−0.6866 (0.0003)	−0.8442 (0.0004)
1.4	−0.5081 (0.0005)	−0.7408 (0.0006)
1.6	−0.4025 (0.0005)	−0.6637 (0.0007)
1.8	−0.3277 (0.0004)	−0.5969 (0.0007)
2	−0.2644 (0.0004)	−0.5327 (0.0007)

5.5. The exponential model

The following example considers an exponential distribution with unknown mean. Let x be a random sample from $X = (X_1, \ldots, X_n)$, where the X_i's

TABLE 2
Normal model with $\mu = 2$, $\sigma = 2$, $n = 30$. Expected pseudo-spherical score for the estimative and the pivotal solution. Estimates are based on 10,000 Monte Carlo replications, with standard errors in parentheses.

γ	estimative	pivotal
1.05	−0.9104 (0.0001)	−0.9652 (0.0001)
1.1	−0.8361 (0.0002)	−0.9353 (0.0002)
1.2	−0.7220 (0.0003)	−0.8884 (0.0004)
1.4	−0.5757 (0.0004)	−0.8259 (0.0006)
1.6	−0.6637 (0.0005)	−0.7866 (0.0007)
1.8	−0.4281 (0.0005)	−0.7607 (0.0008)
2	−0.3865 (0.0005)	−0.7426 (0.0010)

are independent and with the same distribution $Exp(1/\theta)$, $i = 1, \ldots, n$, with θ unknown. We want to estimate the value of a further random variable $Y \sim Exp(1/\theta)$, independent of X. The maximum likelihood estimator for θ is $\hat{\theta} = \overline{X}$. A predictive density giving exact predictive intervals could be obtained using the pivotal quantity $Y/\hat{\theta}$, which has a Fisher F distribution with $(2, 2n)$ degrees of freedom; see for instance [32].

We have performed a simulation study for comparing the estimative and the pivotal solution using the Tsallis and the pseudo-spherical divergences, for various values of the parameter γ; the results are given in Tables 3 and 4, respectively. Again, both the Tsallis and the pseudo-spherical divergences prefer the pivotal solution for every value of γ.

TABLE 3
Exponential model with $\theta = 2$, $n = 30$. Expected Tsallis score for the estimative and the pivotal solution. Estimates are based on 10,000 Monte Carlo replications, with standard errors in parentheses.

γ	estimative	pivotal
1.05	−0.9144 (0.0002)	−0.9509 (0.0002)
1.1	−0.8291 (0.0003)	−0.9064 (0.0003)
1.2	−0.6595 (0.0005)	−0.8292 (0.0005)
1.4	−0.3294 (0.0007)	−0.709 (0.0010)
1.6	−0.0171 (0.0009)	−0.6181 (0.0013)
1.8	0.2697 (0.0010)	−0.5499 (0.0016)
2	0.5321 (0.0011)	−0.4887 (0.0018)

TABLE 4
Exponential model with $\theta = 2$, $n = 30$. Expected pseudo-spherical score for the estimative and the pivotal solution. Estimates are based on 10,000 Monte Carlo replications, with standard errors in parentheses.

γ	estimative	pivotal
1.05	−0.9186 (0.0001)	−0.9532 (0.0001)
1.1	−0.8451 (0.0003)	−0.9145 (0.0003)
1.2	−0.7187 (0.0004)	−0.8555 (0.0005)
1.4	−0.5311 (0.0005)	−0.7822 (0.0008)
1.6	−0.4029 (0.0005)	−0.7403 (0.0010)
1.8	−0.3138 (0.0005)	−0.7173 (0.0011)
2	−0.2482 (0.0005)	−0.6991 (0.0013)

Aknowledgements

The authors are grateful to the Associate Editor and two anonymous referees for helpful comments and suggestions that led to an improved version of the paper.

This research was partially supported by the Italian Ministry for University and Research under the PRIN2015 grant No. 2015EASZFS 003.

References

[1] Aitchison, J. (1975). Goodness of prediction fit. *Biometrika* **62** 547–554. MR0391353

[2] Aitchison, J. and Dunsmore, I.R. (1975). *Statistical prediction analysis*. Cambridge University Press, Cambridge. MR0408097

[3] Amari, S. (1985). *Differential Geometric Methods in Statistics. Lecture Notes in Statistics*, 28. New York: Springer-Verlag. MR0788689

[4] Amari, S. (2010). Divergence function, information monotonicity and information geometry. *Bulletin of the Polish Academy of Sciences: Technical Sciences* **58** 183–195.

[5] Barndorff-Nielsen, O.E. and Cox, D.R. (1996). Prediction and asymptotics. *Bernoulli* **2** 319–340. MR1440272

[6] Bjørnstad, J.F. (1990). Predictive likelihood: A review. *Statistical Sciences* **5** 242–265. MR1062578

[7] Brier, G.W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review* **78** 1–3.

[8] Čencov, N.N. (1982). *Statistical decision rules and optimal inference*. Translations of Mathematical Monographs 53. AMS, Providence, RI. MR0645898

[9] Corcuera, J.M. and Giummolè, F. (1998). A characterization of monotone and regular divergences. *Annals of the Institute of Statistical Mathematics* **50** 433–450. MR1664587

[10] Corcuera, J.M. and Giummolè, F. (1999). On the relationship between α-connections and the asymptotic properties of predictive distributions. *Bernoulli* **5** 163–176. MR1673576

[11] Corcuera, J.M. and Giummolè, F. (2000). First order optimal predictive densities. In Marriott, P., Salmon, M., *Applications of differential geometry to econometrics*, Cambridge University Press. MR1789300

[12] Csiszár, I. (1967). Information-type measure of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica* **2** 299–318. MR0219345

[13] Dawid, A.P. (1998). Coherent measures of discrepancy, uncertainty and dependence, with applications to Bayesian predictive experimental design. *Technical Report* 139, Department of Statistical Science, University College London. http://www.ucl.ac.uk/Stats/research/pdfs/139b.zip

[14] Dawid, A.P. (2007). The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics* **59** 77–93. MR2396033

[15] Dawid, A.P. and Musio, M. (2014). Theory and applications of proper scoring rules. *Metron* **72** 169–183. MR3233147
[16] Dawid, A.P., Musio, M. and Ventura, L. (2016). Minimum scoring rule inference. *Scandinavian Journal of Statistics* **43** 123–138. MR3466997
[17] Eguchi, S. (1992). Geometry of minimum contrast. *Hiroshima Mathematical Journal* **22** 631–647. MR1194056
[18] Eguchi, S. (2006). Information geometry and statistical pattern recognition. *Sugaku Exposition*. AMS, Providence, RI. MR2279777
[19] Forbes, P.G.M. (2012). Compatible weighted proper scoring rules. *Biometrika* **99** 989–994. MR2999174
[20] Fonseca, G., Giummolè, F. and Vidoni, P. (2014). Calibrating predictive distributions. *Journal of Statistical Computation and Simulation* **84** 373–383. MR3169333
[21] Giummolè, F., Ventura, L. (2006). Robust prediction limits based on M-estimators. *Statistics and Probability Letters* **76** 1735–1740. MR2274134
[22] Gneiting, T. and Raftery, A.E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102** 359–378. MR2345548
[23] Gneiting,T. and Katzfuss, M. (2014). Probabilistic Forecasting. *Annual Review of Statistics and Its Application* **1** 125–151.
[24] Good, I.J. (1952). Rational decisions. *Journal of the Royal Statistical Society, Series B* **14** 107–114. MR0077033
[25] Hall, P., Peng, L. and Tajvidi, N. (1999). On prediction intervals based on predictive likelihood or bootstrap methods. *Biometrika* **86** 871–880. MR1741983
[26] Harris, I.R. (1989). Predictive fit for natural exponential families. *Biometrika* **76** 675–684. MR1041412
[27] Holzmann, H. and Klar, B. (2017). Focusing on regions of interest in forecast evaluation. *The Annals of Applied Statistics* **11(4)**, 2404–2431. MR3743302
[28] Kass, S. and Vos, P.W. (1997). *Geometrical Foundations of Asymptotic Inference, Wiley Series in Probability and Statistics*. New York: John Wiley & Sons, Inc. MR1461540
[29] Komaki, F. (1996). On asymptotic properties of predictive distributions. *Biometrika* **83** 299–313. MR1439785
[30] Jose, V.R.R., Nau, R.F. and Winkler, R.L. (2009). Scoring rules, generalized entropy, and utility maximization. *Operations Research* **56** 1146–1157. MR2468903
[31] Jose, V.R.R. (2008). *The Verification of Probabilistic Forecasts in Decision and Risk Analysis*. Phd Thesis, Department of Business Administration, Duke University.
[32] Lawless, J.F. and Fredette, M. (2005). Frequentist prediction intervals and predictive distributions. *Biometrika* **92** 529–542. MR2202644
[33] Machete, R. (2013). Contrasting probabilistic scoring rules. *Journal of Statistical Planning and Inference* **143** 1781–1790. MR3082233
[34] Mameli, V. and Ventura, L. (2015). Higher-order asymptotics for

scoring rules. *Journal of Statistical Planning and Inference* **165** 13–26. MR3498291

[35] Mameli, V., Musio, M. and Ventura, L. (2018). Bootstrap adjustments of signed scoring rule root statistics. *Communications in Statistics – Simulation and Computation* **47** 1204–1215. MR3812407

[36] Mendenez, M.L., Morales, D., Pardo, L. and Salicrù, M. (1997). (h, ϕ)-entropy differential metric. *Applications of Mathematics* **42** 81–98. MR1430403

[37] Murray, G.D. (1977). A note on the estimation of probability density functions. *Biometrika* **64** 150–152. MR0448690

[38] Murray, M.K. and Rice, J.W. (1993). *Differential Geometry and Statistics, Monographs on Statistics and Applied Probability*, 48. London: Chapman & Hall. MR1293124

[39] Pardo, L. (2006). *Statistical inference based on divergence measure*. Florida: Boca Raton, Taylor & Francis. MR2183173

[40] Savage, L.J. (1971). Elicitation of Personal Probabilities and Expectations. *Journal of the American Statistical Association* **66** 783–801. MR0331571

[41] Tsallis, C. (1988). Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics* **52** 479–487. MR0968597

[42] Vidoni, P. (1995). A simple predictive density based on the p^*-formula. *Biometrika* **82** 855–863. MR1380820

[43] Vidoni, P. (2009). Improved prediction intervals and distribution functions. *Scandinavian Journal of Statistics* **36** 735–748. MR2573305