Noname manuscript No.
(will be inserted by the editor)

Bregman divergences based on optimal design criteria and simplicial measures of dispersion

Luc Pronzato · Henry P. Wynn · Anatoly Zhigljavsky

Received: date / Accepted: date

Abstract In previous work the authors define the k-th order simplicial distance between probability distributions which arises naturally from a measure of dispersion based on the squared volume of random simplices of dimension k. This theory is embedded in the wider theory of divergences and distances between distributions which includes Kullback-Leibler, Jensen-Shannon, Jeffreys-Bregman divergence and Bhattacharyya distance. A general construction is given based on defining a directional derivative of a function ϕ from one distribution to the other whose concavity or strict concavity influences the properties of the resulting divergence. For the normal distribution these the divergences can be expressed as matrix formula for the (multivariate) means and covariances. Optimal experimental design criteria contribute a range of functionals applied to non-negative, or positive definite, information matrices. Not all can distinguish normal distributions but sufficient conditions are given. The k-th order simplicial distance is revisited from this aspect and the results are used to test empirically the identity of means and covariances.

Keywords Simplicial distances · Bregman divergence · optimal design criteria · Burbea-Rao divergence · energy statistic

Mathematics Subject Classification (2000) 62H30 · 62K05

1 Introduction

There are close connections between divergences and distances between probability distributions, and certain Frechet-type derivatives. Moreover, for the normal distributions and for the information matrices which dominate the theory of optimal experimental design, the distances can be expressed in matrix form. A natural question that the paper explores is which distance, or which type of experimental design

Luc Pronzato
CNRS, UCA, Laboratoire I3S, UMR 7172; 2000, route des Lucioles, Les Algorithmes, bât. Euclide B, 06900 Sophia Antipolis, France
Tel.: +33-4-89154345
E-mail: Luc.Pronzato@cnrs.fr

H.P. Wynn
London School of Economics, Houghton Street, London, WC2A 2AE, UK
E-mail: H.Wynn@lse.ac.uk

A. Zhigljavsky
School of Mathematics, Cardiff University, Senghennydd Road, Cardiff, CF24 4YH, UK
E-mail: ZhigljavskyAA@cf.ac.uk

criterion, is best able to distinguish between two normal distributions, in particular when the covariance matrices are close to singularity.

Many divergences and distances between probability distributions are constructed from concave functionals ϕ defined on the set of probability measures, with the symmetrized Kullback-Leibler divergence, the Jensen-Shannon divergence and Bhattacharyya distance as typical examples; see, e.g., [2], [12]. Note that we shall call them distances also in the case when they only define semi-metrics; that is, when they do not satisfy the triangular inequality. Distances between two normal distributions only depend on their first two moments, which yields simple statistical criteria for testing the identity of means and covariances matrices of two distributions based on two samples, using empirical estimates.

In the same way, design of optimal experiments relies on the maximization of concave functions Φ of information matrices, see the abundant literature on the subject [1], [5], [6], [7], [14], [15], [19], [24]. Such concave, sometimes strictly concave, design criteria form natural candidates for the definition of distances between two normal distributions. Also, in a recent paper [18] we considered simplicial distances induced by the dispersion functionals

$$\phi_k(\mu) = \mathsf{E}_\mu\{\mathscr{V}_k^2(X_0,\ldots,X_k)\}\,,$$

where $\mathscr{V}_k(X_0,X_1,\ldots,X_k)$ is the volume of the k-dimensional simplex (its length when $k=1$ and area when $k=2$) formed by the $k+1$ vertices $X_0,X_1,\ldots,X_k$ assumed to be i.i.d. with μ in $\mathbb{R}^d$. The functional $\phi_k^{1/k}$ is concave [17], and may thus also be considered for the construction of distances between distributions.

The paper explores the connections between the various notions of distances induced by these approaches. In particular, we show that the construction of ϕ_k, based on volumes of k-dimensional simplices, makes the associated distances more sensitive to the dimensionality of the data than other, more usual, distances between normal distributions, Bhattacharyya distance for instance. We also show that Kiefer's family of design criteria φ_p with $p>0$, which are rather insensitive to the presence of small eigenvalues, may conveniently be used to define distances between normal distributions, in particular for measures concentrated in small dimension subspaces.

2 Distances defined from concave functionals

Let ϕ denote a twice-continuously Fréchet-differentiable real-valued concave functional defined on the set $\mathscr{M}$ of probability measures on Borel sets of $\mathbb{R}^d$. For any $\mu,\zeta\in\mathscr{M}$, denote by $F_\phi(\mu,\zeta)$ the directional derivative of ϕ at μ in the direction ζ,

$$F_\phi(\mu,\zeta) = \lim_{\alpha\to 0^+} \frac{\phi[(1-\alpha)\mu+\alpha\xi]-\phi(\mu)}{\alpha}\,, \tag{1}$$

that is, the Fréchet derivative of ϕ at μ in the direction $\xi-\mu$, see for instance [8]. The Bregman divergence between μ and ζ associated with ϕ is then

$$D_{\phi,B}(\mu,\zeta) = \phi(\mu) + F_\phi(\mu,\zeta) - \phi(\zeta)\,,$$

and the strict concavity of ϕ implies that $D_{\phi,B}(\mu,\xi)\geq 0$ with $D_{\phi,B}(\mu,\xi)=0$ if and only if $\zeta=\mu$. When ϕ is strictly concave on $\mathscr{M}$, the Jeffreys-Bregman divergence

$$D_{\phi,JB}(\mu,\zeta) = \frac{1}{2}\left[D_{\phi,B}(\mu,\zeta)+D_{\phi,B}(\zeta,\mu)\right] = \frac{1}{2}\left[F_\phi(\mu,\zeta)+F_\phi(\zeta,\mu)\right]\,, \tag{2}$$

obtained by symmetrization, and the Burbea-Rao divergence

$$D_{\phi,BR}(\mu,\zeta) = \phi\left(\frac{\mu+\zeta}{2}\right) - \frac{\phi(\mu)+\phi(\zeta)}{2}\,, \tag{3}$$

which does not require ϕ to be Fréchet-differentiable, thus define a semi-metric on $\mathscr{M}$; see for instance [2], [12]. A classical example in the case when μ and ζ have densities ϖ_μ and ϖ_ζ on $\mathbb{R}^d$ with respect to the Lebesgue measure, is given by ϕ equal to the Shannon entropy H_0, with

$$H_0(\mu) = -\int \log[\varpi_\mu(x)]\,\varpi_\mu(x)\,\mathrm{d}x\,.$$

The Jeffreys-Bregman divergence $D_{H_0,JB}$ is then simply the symmetrized Kullback-Leibler divergence

$$D_{KL}(\mu,\zeta) = \frac{1}{2}\,[KL(\mu\|\zeta) + KL(\zeta\|\mu)]\,,$$

and $D_{H_0,BR}(\mu,\zeta)$ coincides with the Jensen-Shannon divergence,

$$D_{JS}(\mu,\zeta) = \frac{1}{2}\left\{KL\left[\mu\|\left(\frac{\mu+\zeta}{2}\right)\right] + KL\left[\zeta\|\left(\frac{\mu+\zeta}{2}\right)\right]\right\}\,,$$

where $KL(\mu\|\zeta) = \int \log[\varpi_\mu(x)/\varpi_\zeta(x)]\,\varpi_\mu(x)\,\mathrm{d}x$. More generally, one can define $KL(\mu\|\zeta) = \int \log[\mathrm{d}\mu/\mathrm{d}\zeta](x)\,\mathrm{d}\mu(x)$ if $\mu \ll \zeta$ (i.e., if ζ dominates μ) and $KL(\mu\|\zeta) = +\infty$ otherwise; see [23, Sect. III.9]. Also, the Hellinger integral $H(\mu,\zeta)$ is defined by

$$H(\mu,\zeta) = \int \sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x)\frac{\mathrm{d}\zeta}{\mathrm{d}\nu}(x)}\,\mathrm{d}\nu(x)\,,$$

with ν denoting any dominating measure for μ and ζ, and $\rho(\mu,\zeta) = \sqrt{1 - H(\mu,\zeta)}$ defining a metric on $\mathscr{M}$. The Bhattacharyya distance $D_B(\mu,\zeta) = -\log H(\mu,\zeta)$ defines a semi-metric on $\mathscr{M}$; see [3], [23, Sect. III.9].

When μ and ζ correspond to normal distributions, the distances defined above only depend on their respective means a_μ and a_ζ and covariances $\varSigma_\mu$ and $\varSigma_\zeta$ (we assume that $\varSigma_\mu$ and $\varSigma_\zeta$ have full rank d). In particular, D_{KL}, D_{JS} and D_B take simple expressions:

$$\begin{aligned}
D_{KL}(\mu,\zeta) &= \frac{1}{4}\left[\mathrm{trace}(\varSigma_\mu^{-1}\varSigma_\zeta) + \mathrm{trace}(\varSigma_\zeta^{-1}\varSigma_\mu)\right] \\
&\quad + \frac{1}{4}\,(a_\zeta - a_\mu)^\top(\varSigma_\mu^{-1} + \varSigma_\zeta^{-1})(a_\zeta - a_\mu) - \frac{d}{2} && (4)\\
D_{JS}(\mu,\zeta) &= \frac{1}{2}\,\log\left[\frac{\det\left(\frac{\varSigma_\mu+\varSigma_\zeta}{2}\right)}{\sqrt{\det(\varSigma_\mu)\det(\varSigma_\zeta)}}\right] \\
&\quad + \frac{1}{2}\,\log\left[1 + \frac{1}{2}(a_\mu - a_\zeta)^\top(\varSigma_\mu + \varSigma_\zeta)^{-1}(a_\mu - a_\zeta)\right] && (5)\\
D_B(\mu,\zeta) &= \frac{1}{2}\,\log\left[\frac{\det\left(\frac{\varSigma_\mu+\varSigma_\zeta}{2}\right)}{\sqrt{\det(\varSigma_\mu)\det(\varSigma_\zeta)}}\right] \\
&\quad + \frac{1}{4}(a_\mu - a_\zeta)^\top(\varSigma_\mu + \varSigma_\zeta)^{-1}(a_\mu - a_\zeta)\,. && (6)
\end{aligned}$$

Note that $D_B(\mu,\zeta) \geq D_{JS}(\mu,\zeta)$, with equality when $a_\zeta = a_\mu$, and that D_{KL}, D_{JS} and D_B satisfy the following invariance property

$$D(\mu,\zeta) = D(\mu_0,\zeta[\mu])\,, \tag{7}$$

where μ_0 has zero mean and covariance I_d, the d-dimensional identity matrix, and $\zeta[\mu]$ has mean $\varSigma_\mu^{-1/2}(a_\zeta - a_\mu)$ and covariance $\varSigma_\mu^{-1/2}\varSigma_\zeta\varSigma_\mu^{-1/2}$. One may refer to [13] for more detailed developments on connections between Bhattacharyya distance and other divergence measures.

For each of these distances, $D_{KL}(\mu,\zeta)$, $D_{JS}(\mu,\zeta)$ and $D_B(\mu,\zeta)$, equality to zero is obtained if and only if $a_\zeta = a_\mu$ and $\varSigma_\zeta = \varSigma_\mu$. When a distance D satisfies this property, we shall say that D distinguishes normal distributions.

3 Distances based on optimal design criteria

Optimal design of experiments rely on the maximization of a concave functional $\varPhi$ of the information matrix. Below we show that some classical optimality criteria, such as A- and D-optimality, yield distance measures that are able to distinguish normal distributions, but that the usual notion of strict concavity used in optimal design theory is not enough to obtain this property.

3.1 Construction

Denote by $\mathbb{M}^>$ (respectively, $\mathbb{M}^\geq$) the set of $d \times d$ symmetric positive definite (respectively, non-negative definite) matrices. In Section 4 we shall consider functions $\varPhi$ whose properties depend on the rank on the matrices involved; $\mathbb{M}^*$ will denote a general matrix cone included in $\mathbb{M}^\geq$ such that $\mathbf{M}_1 + \mathbf{M}_2 \in \mathbb{M}^*$ for any $\mathbf{M}_1 \in \mathbb{M}^*$ and $\mathbf{M}_2 \in \mathbb{M}^\geq$. We shall consider two particular cases: $\mathbb{M}^* = \mathbb{M}^>$, and $\mathbb{M}^* = \mathbb{M}^r$, the subset of $\mathbb{M}^\geq$ containing matrices of rank at least $r \leq d$. We denote by $\mathscr{M}^*$ the subset of $\mathscr{M}$ containing distributions with finite covariances in $\mathbb{M}^*$, with $\mathscr{M}^>$ and $\mathscr{M}^r$ as particular cases associated with $\mathbb{M}^* = \mathbb{M}^>$ and $\mathbb{M}^* = \mathbb{M}^r$.

Let $\varPhi$ be a function defined on $\mathbb{M}^\geq$, isotonic on $\mathbb{M}^*$ relative to the Loewner ordering ($\varPhi(M_1) \geq \varPhi(M_2)$ when $M_2 \in \mathbb{M}^*$ and $M_1 - M_2 \in \mathbb{M}^\geq$) and concave on $\mathbb{M}^*$ ($\varPhi[(1-\alpha)M_1 + \alpha M_2] \geq (1-\alpha)\varPhi(M_1) + \alpha\varPhi(M_2)$ for all $\alpha \in (0,1)$ and $M_1 \in \mathbb{M}^*$, $M_2 \in \mathbb{M}^\geq$). Consider two probability measures μ and ζ with respective means a_μ and a_ζ and covariances $\varSigma_\mu = \mathsf{var}(\mu)$ and $\varSigma_\zeta = \mathsf{var}(\zeta)$, with $\varSigma_\mu, \varSigma_\zeta \in \mathbb{M}^*$. Following (3), the Burbea-Rao divergence $D_{\varPhi,BR}(\mu,\zeta)$ associated with $\varPhi$ is defined by

$$D_{\varPhi,BR}(\mu,\zeta) = \varPhi\left[\mathsf{var}\left(\frac{\mu+\zeta}{2}\right)\right] - \frac{\varPhi(\varSigma_\mu) + \varPhi(\varSigma_\zeta)}{2}\,.$$

Direct calculation gives

$$\mathsf{var}[(\mu+\zeta)/2] = \frac{1}{2}\,(\varSigma_\mu + \varSigma_\zeta) + \frac{1}{4}\,(a_\zeta - a_\mu)(a_\zeta - a_\mu)^\top$$

so that

$$D_{\varPhi,BR}(\mu,\zeta) = \varPhi\left[\frac{1}{2}\,(\varSigma_\mu + \varSigma_\zeta) + \frac{1}{4}\,(a_\zeta - a_\mu)(a_\zeta - a_\mu)^\top\right] - \frac{\varPhi(\varSigma_\mu) + \varPhi(\varSigma_\zeta)}{2}\,, \tag{8}$$

with $D_{\varPhi,BR}(\mu,\zeta) \geq\ 0$ from the isotonicity and concavity of $\varPhi$ on $\mathbb{M}^*$.

Denote now

$$F_\varPhi(\mu,\zeta) = \lim_{\alpha\to 0^+} \frac{\varPhi\{\mathsf{var}[(1-\alpha)\mu + \alpha\xi]\} - \varPhi[\mathsf{var}(\mu)]}{\alpha}\,,$$

see (1). For $\alpha \in [0,1]$, define $\mu_{x,\alpha} = (1-\alpha)\mu + \alpha\delta_x$, with δ_x the Dirac delta measure at x. Straightforward calculation gives

$$\left.\frac{\partial \mathsf{var}[\mu_{x,\alpha}]}{\partial\alpha}\right|_{\alpha=0} = (x - a_\mu)(a - a_\mu)^\top - \varSigma_\mu\,,$$

so that, when Φ is differentiable at Σ_μ, with gradient $\nabla_\Phi(\Sigma_\mu)$,

$$
\begin{aligned}
F_\Phi(\mu,\zeta) = \int F_\Phi(\mu,\delta_x)\,\mathrm{d}\zeta(x) &= \int \mathrm{trace}\left[\nabla_\Phi(\Sigma_\mu)\left.\frac{\partial \mathsf{var}[\mu_{x,\alpha}]}{\partial\alpha}\right|_{\alpha=0}\right]\mathrm{d}\zeta(x) \\
&= \mathrm{trace}[\nabla_\Phi(\Sigma_\mu)(\Sigma_\zeta-\Sigma_\mu)] + (a_\zeta-a_\mu)^\top\nabla_\Phi(\Sigma_\mu)(a_\zeta-a_\mu)\,. \qquad (9)
\end{aligned}
$$

Similarly to (2), the Jeffreys-Bregman divergence $D_{\Phi,JB}(\mu,\zeta)$ associated with Φ is then defined as

$$
\begin{aligned}
D_{\Phi,JB}(\mu,\zeta) &= \frac{1}{2}\left[F_\Phi(\mu,\zeta)+F_\Phi(\zeta,\mu)\right] \\
&= \frac{1}{2}\,\Big[\mathrm{trace}\{[\nabla_\Phi(\Sigma_\mu)-\nabla_\Phi(\Sigma_\zeta)](\Sigma_\zeta-\Sigma_\mu)\} \\
&\quad + (a_\zeta-a_\mu)^\top[\nabla_\Phi(\Sigma_\mu)+\nabla_\Phi(\Sigma_\zeta)](a_\zeta-a_\mu)\Big]\,. \qquad (10)
\end{aligned}
$$

For any $z\in\mathbb{R}^d$ with $\|z\|=1$, we have $\Phi(\Sigma_\mu+zz^\top)\leq\Phi(\Sigma_\mu)+z^\top\nabla_\Phi(\Sigma_\mu)z$ from the concavity of Φ on $\mathbb{M}^*$ and $\Phi(\Sigma_\mu+zz^\top)\geq\Phi(\Sigma_\mu)$ from its isotonicity. Therefore, $z^\top\nabla_\Phi(\Sigma_\mu)z\geq 0$, and $\nabla_\Phi(\Sigma_\mu)\in\mathbb{M}^{\geq}$. Similarly, $\nabla_\Phi(\Sigma_\zeta)\in\mathbb{M}^{\geq}$, showing that the second term in (10) is non-negative. Concavity on $\mathbb{M}^*$ also implies $\Phi(\Sigma_\zeta)\leq\Phi(\Sigma_\mu)+\mathrm{trace}[\nabla_\Phi(\Sigma_\mu)(\Sigma_\zeta-\Sigma_\mu)]$ and $\Phi(\Sigma_\mu)\leq\Phi(\Sigma_\zeta)+\mathrm{trace}[\nabla_\Phi(\Sigma_\zeta)(\Sigma_\mu-\Sigma_\zeta)]$, which gives $\mathrm{trace}\{[\nabla_\Phi(\Sigma_\mu)-\nabla_\Phi(\Sigma_\zeta)](\Sigma_\zeta-\Sigma_\mu)\}\geq 0$. Therefore, $D_{\Phi,JB}(\mu,\zeta)\geq 0$.

Below we investigate which additional conditions must be imposed on Φ to ensure that $D_{\Phi,BR}$ (8) and $D_{\Phi,JB}$ (10) distinguish normal distributions in $\mathscr{M}^*$.

3.2 Sufficient conditions for distinguishability

We say that Φ is positively homogeneous when

$$\Phi(\alpha M)=\alpha\,\Phi(M) \text{ for any } \alpha>0 \text{ and } M\in\mathbb{M}^{\geq}\,,$$

and we shall say that Φ is strictly isotonic on $\mathbb{M}^*$ when

$$\Phi(M_1)>\Phi(M_2) \text{ for any } M_1,M_2 \text{ such that } M_1-M_2\in\mathbb{M}^{\geq},\ M_2\in\mathbb{M}^*,\ M_2\neq M_1\,.$$

In optimal design of experiments, a function Φ is said to be strictly concave on the cone $\mathbb{M}^*\subset\mathbb{M}^{\geq}$ when

$$
\begin{aligned}
\Phi[(1-\alpha)M_1+\alpha M_2] > (1-\alpha)&\Phi(M_1)+\alpha\,\Phi(M_2) \\
&\text{for all } \alpha\in(0,1)\,,\ M_1\in\mathbb{M}^* \text{ and } M_2\in\mathbb{M}^{\geq} \qquad (11)\\
&\text{with } M_2\neq 0 \text{ and } M_2 \text{ not proportional to } M_1\,.
\end{aligned}
$$

When $\mathbb{M}^*=\mathbb{M}^>$, this definition coincides with that in [19, Sect. 5.2]. The usual definition in convex analysis is stronger and requires the inequality to be valid for a wider class of matrices M_2. We shall call *strongly strictly* concave on $\mathbb{M}^*$ a function Φ such that

$$
\begin{aligned}
\Phi[(1-\alpha)M_1+\alpha M_2] > (1-\alpha)&\Phi(M_1)+\alpha\,\Phi(M_2) \\
&\text{for all } \alpha\in(0,1)\,,\ M_1\in\mathbb{M}^* \text{ and } M_2\in\mathbb{M}^{\geq} \qquad (12)\\
&\text{with } M_2\neq 0 \text{ and } M_2\neq M_1\,.
\end{aligned}
$$

The following property shows that $D_{\Phi,BR}$ (8) and $D_{\Phi,JB}$ (10) distinguish normal distributions in $\mathscr{M}^*$ when Φ is strictly isotonic and strongly strictly concave on $\mathbb{M}^*$.

Lemma 1 *Let Φ be a strictly isotonic and strongly strictly concave function on $\mathbb{M}^*$. Then, for μ and ζ two probability measures with respective means a_μ and a_ζ and covariances $\Sigma_\mu = \mathsf{var}(\mu)$ and $\Sigma_\zeta = \mathsf{var}(\zeta)$, $\Sigma_\mu, \Sigma_\zeta \in \mathbb{M}^*$, we have*

$$D_{\Phi,BR}(\mu,\zeta) = 0 \Rightarrow a_\mu = a_\zeta \text{ and } \Sigma_\mu = \Sigma_\zeta \,, \tag{13}$$

with $D_{\Phi,BR}$ given by (8)*, and, when Φ is differentiable at Σ_μ and Σ_ζ,*

$$D_{\Phi,JB}(\mu,\zeta) = 0 \Rightarrow a_\mu = a_\zeta \text{ and } \Sigma_\mu = \Sigma_\zeta \,, \tag{14}$$

where $D_{\Phi,JB}$ is defined by (10)*.*

Proof We first prove that $D_{\Phi,BR}(\mu,\zeta) = 0$, or $D_{\Phi,JB}(\mu,\zeta) = 0$, implies $a_\mu = a_\zeta$. Suppose that $a_\mu \neq a_\zeta$. The strict isotonicity of Φ on $\mathbb{M}^*$ implies $\Phi[(\Sigma_\mu+\Sigma_\zeta)/2+(a_\zeta-a_\mu)(a_\zeta-a_\mu)^\top/4] > \Phi[(\Sigma_\mu+\Sigma_\zeta)/2]$, and therefore $D_{\Phi,BR}(\mu,\zeta) > 0$ from concavity. Take any $z \in \mathbb{R}^d$ with $\|z\| = 1$. We have $\Phi(\Sigma_\mu + zz^\top) \leq \Phi(\Sigma_\mu) + z^\top \nabla_\Phi(\Sigma_\mu) z$ from the concavity of Φ and $\Phi(\Sigma_\mu + zz^\top) > \Phi(\Sigma_\mu)$ from its strict isotonicity. Therefore, $z^\top \nabla_\Phi(\Sigma_\mu) z > 0$, and $\nabla_\Phi(\Sigma_\mu) \in \mathbb{M}^>$, showing that $(a_\zeta - a_\mu)^\top [\nabla_\Phi(\Sigma_\mu) + \nabla_\Phi(\Sigma_\zeta)](a_\zeta - a_\mu)$ in (10) is strictly positive.

We consider now two distributions such that $a_\mu = a_\zeta$. Since Φ is strongly strictly concave on $\mathbb{M}^*$, $\Phi[(\Sigma_\mu+\Sigma_\zeta)/2] > [\Phi(\Sigma_\mu)+\Phi(\Sigma_\zeta)]/2$ for $\Sigma_\zeta \neq \Sigma_\mu$, which concludes the proof of (13). Also, for $\Sigma_\zeta \neq \Sigma_\mu$ we have $\Phi(\Sigma_\zeta) < \Phi(\Sigma_\mu) + \mathrm{trace}[\nabla_\Phi(\Sigma_\mu)(\Sigma_\zeta - \Sigma_\mu)]$ and $\Phi(\Sigma_\mu) < \Phi(\Sigma_\zeta) + \mathrm{trace}[\nabla_\Phi(\Sigma_\zeta)(\Sigma_\mu - \Sigma_\zeta)]$, so that $\mathrm{trace}\{[\nabla_\Phi(\Sigma_\mu) - \nabla_\Phi(\Sigma_\zeta)](\Sigma_\zeta - \Sigma_\mu)\} > 0$ and $D_{\Phi,JB}(\mu,\zeta) > 0$, which proves (14). ■

A positively homogeneous function Φ is not strongly strictly concave. Indeed, take $M_2 = \beta M_1$, with $M_1 \in \mathbb{M}^*$, $\beta > 0$ and $\beta \neq 1$. We have $\Phi[(1-\alpha)M_1 + \alpha M_2] = \Phi[(1-\alpha+\alpha\beta)M_1] = (1-\alpha+\alpha\beta)\Phi(M_1) = (1-\alpha)\Phi(M_1) + \alpha\,\Phi(M_2)$. An important consequence is that the Burbea-Rao and Jeffreys-Bregman divergences associated with a strictly concave (in the sense of (11)) and positively homogeneous function Φ are unable to distinguish normal distributions. Take μ and ζ such that $a_\mu = a_\zeta$ and $\Sigma_\zeta = \beta\,\Sigma_\mu$, $\beta > 0$ and $\beta \neq 1$. One can readily check that $D_{\Phi,BR} = 0$, see (8). Also, when Φ is differentiable at Σ_μ, then $\nabla_\Phi(\Sigma_\mu) = \nabla_\Phi(\Sigma_\zeta)$ and $D_{\Phi,JB} = 0$, see (10). In contrast, the following property shows that $D_{\Phi,BR}$ and $D_{\Phi,JB}$ do distinguish normal distributions when using $\log\Phi$ instead of Φ.

Lemma 2 *Let Φ be a function positively homogeneous, non identically zero, strictly isotonic on $\mathbb{M}^>$, and strictly concave in the sense of* (11)*. Then, for μ and ζ two probability measures with respective means a_μ and a_ζ and covariances $\Sigma_\mu = \mathsf{var}(\mu)$ and $\Sigma_\zeta = \mathsf{var}(\zeta)$, $\Sigma_\mu, \Sigma_\zeta \in \mathbb{M}^>$, we have*

$$D_{\log\Phi,BR}(\mu,\zeta) = 0 \Rightarrow a_\mu = a_\zeta \text{ and } \Sigma_\mu = \Sigma_\zeta \,,$$

and when Φ is differentiable at Σ_μ and Σ_ζ,

$$D_{\log\Phi,JB}(\mu,\zeta) = 0 \Rightarrow a_\mu = a_\zeta \text{ and } \Sigma_\mu = \Sigma_\zeta \,.$$

Proof First note that $\Phi(\Sigma_\mu) > 0$ and $\Phi(\Sigma_\zeta) > 0$ since $\Sigma_\mu, \Sigma_\zeta \in \mathbb{M}^>$, see [19, Chap. 5], so that $\log\Phi(\Sigma_\mu)$ and $\log\Phi(\Sigma_\zeta)$ are well defined. Also, when Φ is differentiable at Σ, $\log\Phi$ is differentiable too, with $\nabla_{\log\Phi}(\Sigma) = \nabla_\Phi(\Sigma)/\Phi(\Sigma)$.

Using Lemma 1, we only need to show that $\log\Phi$ is strictly isotonic and strongly strictly concave function on $\mathbb{M}^>$. Strict isotonicity follows from the fact that the logarithm is increasing. Consider now (12). Take any $M_1 \in \mathbb{M}^>$ and $M_2 \in \mathbb{M}^\geq$, $M_2 \neq 0$, and any $\alpha \in (0,1)$. We can write

$$\begin{aligned}\log\Phi[(1-\alpha)M_1 + \alpha M_2] &\geq \log[(1-\alpha)\Phi(M_1) + \alpha\,\Phi(M_2)] \\ &\geq (1-\alpha)\log\Phi(M_1) + \alpha\,\log\Phi(M_2)\,,\end{aligned} \tag{15}$$

where the first inequality follows from the concavity of Φ and the second from the concavity of logarithm. From the monotonicity of logarithm and the strict concavity of Φ in the sense of (11), equality between the two extreme terms implies $M_2 = \beta\, M_1$ for some $\beta > 0$. Since Φ is positively homogeneous, $\log \Phi[(1-\alpha)M_1 + \alpha M_2] = (1-\alpha)\log\Phi(M_1) + \alpha\,\log\Phi(M_2)$ then gives $f(\beta) = \log(1-\alpha+\alpha\beta) - \alpha\log(\beta) = 0$. Direct calculation gives $\mathrm{d}f(\beta)/\mathrm{d}\beta = \alpha[1/(1-\alpha+\alpha\beta) - 1/\beta]$, showing that, for any $\alpha \in (0,1)$, $f(\beta)$ has a unique minimum at $\beta = \beta_* = 1$, with $f(\beta_*) = 0$. Equality in (15) thus implies $M_2 = \beta_*\, M_1 = M_1$, which proves (12). ■

3.3 Optimal-design criteria

Consider Kiefer's [9] φ_p-class of functions, $p \in \mathbb{R} \cup \{-\infty, +\infty\}$, which defines a family of design criteria widely used in optimal design. For any $M \in \mathbb{M}^{\geq}$, $\varphi_p(M)$ is defined by

$$\varphi_p(M) = \begin{cases} \lambda_{\max}(M) & \text{for } p = \infty\,, \\ \left[\frac{1}{d}\,\mathrm{trace}(M^p)\right]^{1/p} & \text{for } p \neq 0 \text{ and } p \neq \pm\infty\,, \\ \det^{1/d}(M) & \text{for } p = 0\,, \\ \lambda_{\min}(M) & \text{for } p = -\infty\,, \end{cases}$$

with $\varphi_p(M) = 0$ if M is singular when $p \leq 0$. A-optimal design corresponds to $p = -1$, D-optimal design to $p = 0$ and E-optimal design to $p = -\infty$; $\varphi_p(I_d) = 1$ for all p. All φ_p are positively homogeneous; for $p \in (-\infty, 1)$, φ_p is differentiable and strictly isotonic on $\mathbb{M}^{>}$, and strictly concave in the sense of (11), see [19, Sect. 6.13]. Lemma 2 applies, and the Burbea-Rao and Jeffreys-Bregman divergences associated with $\log\varphi_p$, $p \in (-\infty, 1)$, distinguish normal distributions in $\mathscr{M}^{>}$. However, as Section 5 will illustrate, distances associated with negative p are very sensitive to the presence of small eigenvalues in the spectrum of covariances matrices, and are therefore not recommended. In contrast, the presence of zero eigenvalues $\lambda_i(M)$ has little influence when $p > 0$ as $\varphi_p(M) = [(1/d)\sum_{i:\lambda_i(M)>0} \lambda_i^p(M)]^{1/p}$.

The φ_p are information functions and therefore satisfy trace $\left[\nabla_{\varphi_p}(M) M\right] = \varphi_p(M)$ for $M \in \mathbb{M}^{>}$, see [19, p. 168], and we obtain

$$\begin{aligned} D_{\log\varphi_p,JB}(\mu,\zeta) &= \frac{1}{2}\left\{\mathrm{trace}\left[\frac{\nabla_{\varphi_p}(\Sigma_\mu)}{\varphi_p(\Sigma_\mu)}\Sigma_\zeta\right] + \mathrm{trace}\left[\frac{\nabla_{\varphi_p}(\Sigma_\zeta)}{\varphi_p(\Sigma_\zeta)}\Sigma_\mu\right]\right\} \\ &\quad + \frac{1}{2}\left\{(a_\zeta - a_\mu)^\top \left[\frac{\nabla_{\varphi_p}(\Sigma_\mu)}{\varphi_p(\Sigma_\mu)} + \frac{\nabla_{\varphi_p}(\Sigma_\zeta)}{\varphi_p(\Sigma_\zeta)}\right](a_\zeta - a_\mu)\right\} - 1\,. \end{aligned}$$

For $p \neq 0$, we get

$$\begin{aligned} D_{\log\varphi_p,JB}(\mu,\zeta) &= \frac{1}{2}\left[\frac{\mathrm{trace}(\Sigma_\mu^{p-1}\Sigma_\zeta)}{\mathrm{trace}(\Sigma_\mu^p)} + \frac{\mathrm{trace}(\Sigma_\zeta^{p-1}\Sigma_\mu)}{\mathrm{trace}(\Sigma_\zeta^p)}\right] \\ &\quad + \frac{1}{2}\,(a_\zeta - a_\mu)^\top \left(\frac{\Sigma_\mu^{p-1}}{\mathrm{trace}(\Sigma_\mu^p)} + \frac{\Sigma_\zeta^{p-1}}{\mathrm{trace}(\Sigma_\zeta^p)}\right)(a_\zeta - a_\mu) - 1\,. \qquad (16) \end{aligned}$$

This expression is also valid when $p = 0$ with the convention $\mathrm{trace}(M^0) = d$. In general, (8) does not yield a simple expression for $D_{\log\varphi_p,BR}(\mu,\zeta)$. For $p = 0$, $\varphi_0(\Sigma) = \det^{1/d}(\Sigma)$ is directly related to the Shannon entropy H_0 of a normal distribution with covariance Σ, and we have

$$D_{\log\varphi_0,JB}(\mu,\zeta) = \frac{2}{d}\,D_{KL}(\mu,\zeta) \quad\text{and}\quad D_{\log\varphi_0,BR}(\mu,\zeta) = \frac{2}{d}\,D_{JS}(\mu,\zeta)\,,$$

with D_{KL} and D_{JS} respectively given by (4) and (5). In general, $D_{\log\varphi_p,JB}$ and $D_{\log\varphi_p,BR}$ with $p \neq 0$ do not satisfy the invariance property (7).

4 k-th order simplicial distances

4.1 Squared volumes of k-dimensional simplices

In a recent paper [18], we considered simplicial distances induced by the dispersion functionals

$$\phi_k(\mu) = \mathsf{E}_\mu\{\mathscr{V}_k^2(X_0,\ldots,X_k)\}\,, \tag{17}$$

where $\mathscr{V}_k(X_0,X_1,\ldots,X_k)$ is the volume of the k-dimensional simplex (its length when $k=1$ and area when $k=2$) formed by the $k+1$ vertices $X_0,X_1,\ldots,X_k$ assumed to be i.i.d. with μ in $\mathbb{R}^d$. In particular, for $k=1$ we have

$$\phi_1(\mu) = \int\int \|x_1-x_2\|^2\,\mu(\mathrm{d}x_1)\mu(\mathrm{d}x_2) = 2\,\mathrm{trace}[\varSigma_\mu]\,,$$

twice the trace of the covariance matrix of μ. As shown below, when $k=d$ we get $\phi_d(\mu) = (d+1)/d!\,\det(\varSigma_\mu)$, which is proportional to the generalised variance widely used in multivariate statistics.

For any $M\in\mathbb{M}^{\geq}$, define

$$\varPhi_k(M) = \frac{k+1}{k!}\,e_k[\varLambda(M)] \tag{18}$$

with $\varLambda(M)$ the set of eigenvalues of M and e_k the elementary symmetric function of degree k (with $e_0=1$). The following theorem is proved in [17].

Theorem 1 *For any $k\in\{1,\ldots,d\}$ and $\mu\in\mathscr{M}$, we have $\phi_k(\mu)=\varPhi_k[\mathsf{var}(\mu)]$. Moreover, the functional $\phi_k^{1/k}$ is concave on $\mathscr{M}$.*

The $\varPhi_k^{1/k}$, $k\in\{1,\ldots,d\}$, form a family of criteria between $\varphi_1=\varPhi_1/(2d)$ and $\varphi_0=[d!/(d+1)]^{1/d}\,\varPhi_d^{1/d}$. On the one hand, similarly to φ_p with positive p, $\varPhi_k(M)$ with k small enough is relatively insensitive to the presence of small eigenvalues in $\varLambda(M)$. On the other hand, $\varPhi_k(M)>0$ if and only if $M\in\mathbb{M}^k$ (i.e., $\mathrm{rank}(M)\geq k$), which makes the $\varPhi_k$ more sensitive to the true dimensionality of the data than the φ_p for $p\in[0,1]$.

The expressions of $\varPhi_k(M)$ and its gradient $\nabla_{\varPhi_k}(M)$ at $M\in\mathbb{M}^{\geq}$ are given by

$$\varPhi_k(M) = \frac{k+1}{k\,k!}\sum_{i=0}^{k-1}(-1)^{i-1}\,e_{k-i}[\varLambda(M)]\,\mathrm{trace}(M^i)\,,$$

$$\nabla_{\varPhi_k}(M) = \frac{k+1}{k!}\sum_{i=0}^{k-1}(-1)^{i}\,e_{k-i-1}[\varLambda(M)]\,M^i\,,$$

see [10], [20]. In [17], we show that the directional derivative of ϕ_k at μ in the direction ζ is given by (9), with the additional property $\mathrm{trace}[\nabla_{\varPhi_k}(M)M]=k\,\varPhi_k(M)$, $M\in\mathbb{M}^{\geq}$, which gives

$$F_{\varPhi_k}(\mu,\zeta) = \mathrm{trace}[\nabla_{\varPhi_k}(\varSigma_\mu)\varSigma_\zeta] + (a_\zeta-a_\mu)^\top\nabla_{\varPhi_k}(\varSigma_\mu)(a_\zeta-a_\mu) - k\,\varPhi_k(\varSigma_\mu)\,. \tag{19}$$

One can readily check that $\varPhi_k^{1/k}$ is positively homogeneous, it is therefore not strongly strictly concave, see Section 3. However, $\varPhi_k^{1/k}$ is strictly isotonic on $\mathbb{M}^k$ [18, Lemma 3] and strictly concave in the sense of (11) for $k\geq 2$ [18, Lemma 6]. Arguments similar to those in the proof of Lemma 2 indicate that $\log\varPhi_k$ is strictly isotonic and strongly strictly concave on $\mathbb{M}^k$ for $k\geq 2$. The following property is then a consequence of Lemma 1.

Theorem 2 *Let μ and ζ be two probability measures with respective means a_μ and a_ζ and covariances $\varSigma_\mu = \mathsf{var}(\mu)$ and $\varSigma_\zeta = \mathsf{var}(\zeta)$, $\varSigma_\mu, \varSigma_\zeta \in \mathbb{M}^k$. Then,*

$$D_{\log \varPhi_k, BR}(\mu,\zeta) = 0 \Rightarrow a_\mu = a_\zeta \text{ and } \varSigma_\mu = \varSigma_\zeta \,,$$
$$D_{\log \varPhi_k, JB}(\mu,\zeta) = 0 \Rightarrow a_\mu = a_\zeta \text{ and } \varSigma_\mu = \varSigma_\zeta \,.$$

Using (19), we obtain that $D_{\log \varPhi_k, JB}(\mu,\zeta)$ corresponds to the simplicial distance between μ and ζ introduced in [18],

$$\begin{aligned} D_{\log \varPhi_k, JB}(\mu,\zeta) &= \frac{1}{2}\left\{\mathrm{trace}\left[\frac{\nabla_{\varPhi_k}(\varSigma_\mu)}{\varPhi_k(\varSigma_\mu)}\varSigma_\zeta\right] + \mathrm{trace}\left[\frac{\nabla_{\varPhi_k}(\varSigma_\zeta)}{\varPhi_k(\varSigma_\zeta)}\varSigma_\mu\right]\right\} \\ &+\frac{1}{2}\left\{(a_\zeta - a_\mu)^\top \left[\frac{\nabla_{\varPhi_k}(\varSigma_\mu)}{\varPhi_k(\varSigma_\mu)} + \frac{\nabla_{\varPhi_k}(\varSigma_\zeta)}{\varPhi_k(\varSigma_\zeta)}\right](a_\zeta - a_\mu)\right\} - k\,. \qquad (20)\end{aligned}$$

When $k = d$, we have $D_{\log \varPhi_d, JB}(\mu,\zeta) = d\, D_{\log \varphi_0, JB}(\mu,\zeta)$ and $D_{\log \varPhi_d, BR}(\mu,\zeta) = d\, D_{\log \varphi_0, BR}(\mu,\zeta)$, see Section 3. In general, $D_{\log \varPhi_k, JB}$ and $D_{\log \varPhi_k, BR}$ with $k \neq d$ do not satisfy the invariance property (7). In [18], we show that the gradient matrix $\nabla_{\varPhi_k}(M)$ is non-negative definite for any $M \in \mathbb{M}^k$ and any $k \in \{1, \ldots, d\}$, and is positive definite when $\mathrm{rank}(M) \geq k$. Moreover, $\nabla_{\varPhi_k}(M)/\varPhi_k(M)$ is the inverse of M when $\mathrm{rank}(M) = k = d$ and is a generalized inverse of M when $\mathrm{rank}(M) = k < d$. If we write the characteristic polynomial of M as

$$\det(\lambda I_d - M) = c_1\,\lambda^d + c_2\,\lambda^{d-1} + \cdots + c_d\,\lambda + c_{d+1}\,,$$

with $c_1 = 1$, then

$$\begin{aligned} \varPhi_k(M) &= (-1)^k\,\frac{k+1}{k!}\,c_{k+1}\,, \\ \nabla_{\varPhi_k}(M) &= (-1)^{k-1}\frac{k+1}{k!}\,(M^{k-1} + c_2\,M^{k-2} + \cdots + c_k\,I_d)\,. \end{aligned}$$

4.2 Other simplicial functionals

By considering other powers than 2 in (17), we can obtain simplicial functionals that depend on the full measure μ and not only on its covariance matrix $\varSigma_\mu$. In particular, we may obtain divergence measures that define semi-metrics, i.e., that satisfy

$$\text{for any } \mu,\zeta \in \mathscr{M},\ D(\mu,\zeta) = 0 \Leftrightarrow \mu = \zeta\,. \qquad (21)$$

Consider in particular the dispersion measure

$$\phi_{1,\delta}(\mu) = \mathsf{E}_\mu\{\mathscr{V}_1^\delta(X_0,X_1)\} = \int\int \|x_0 - x_1\|^\delta\,\mu(\mathrm{d}x_0)\mu(\mathrm{d}x_1)\,,$$

see [4], [16]. Direct calculation shows that its directional derivative $F_{\phi_{1,\delta}}(\mu,\zeta)$ is

$$\begin{aligned} F_{\phi_{1,\delta}}(\mu,\zeta) &= 2\int\int \|x_0 - x_1\|^\delta\,(\xi - \mu)(\mathrm{d}x_0)\mu(\mathrm{d}x_1) \\ &= 2\left[\int\int \|x_0 - x_1\|^\delta\,\xi(\mathrm{d}x_0)\mu(\mathrm{d}x_1) - \phi_{1,\delta}(\mu)\right]\,, \end{aligned}$$

where the first term on the right-hand side, $\int\int \|x_0 - x_1\|^\delta\,\xi(\mathrm{d}x_0)\mu(\mathrm{d}x_1)$, corresponds to Łukaszyk-Karmowski metric, see [11]. The corresponding Jeffreys-Bregman divergence is

$$D_{\phi_{1,\delta},JB}(\mu,\zeta) = -\int\int \|x_0 - x_1\|^\delta\,(\zeta - \mu)(\mathrm{d}x_0)(\zeta - \mu)(\mathrm{d}x_1)\,.$$

It is called energy distance for $\delta = 1$ and generalized energy distance [26] for $\delta \in (0, 2]$. The functional $\phi_{1,\delta}$ is concave for $\delta \in (0, 2]$, strictly concave for $\delta \in (0, 2)$, and the kernel $K(x_0, x_1) = -\|x_0 - x_1\|^\delta$ is conditionally integrally strictly positive definite for $\delta \in (0, 2)$; see [21], [25]. Then, $D_{\phi_{1,\delta},JB}(\mu, \zeta) > 0$ for two probability measures $\mu \neq \zeta$ having finite energy, i.e., such that $\int\int -\|x_0 - x_1\|^\delta \, \mu(\mathrm{d}x_0)\mu(\mathrm{d}x_1) < +\infty$ and $\int\int -\|x_0 - x_1\|^\delta \, \zeta(\mathrm{d}x_0)\zeta(\mathrm{d}x_1) < +\infty$.

Other conditionally integrally strictly positive definite kernels $K(\cdot, \cdot)$ yield strictly concave measures of dispersion $\phi_K(\mu) = \int\int -K(x_0, x_1) \, \mu(\mathrm{d}x_0)\mu(\mathrm{d}x_1)$ for probability measures, and the associated Jeffreys-Bregman divergence is

$$D_{\phi_K,JB}(\mu, \zeta) = \int\int K(x_0, x_1) \, (\zeta - \mu)(\mathrm{d}x_0)(\zeta - \mu)(\mathrm{d}x_1) \, ,$$

which corresponds to the (squared) maximum mean discrepancy between ζ and μ, as defined in [22]. Uniformly bounded kernels are conditionally integrally strictly positive definite if and only if they are characteristic, i.e., satisfy (21); see [25]. The question of whether simplicial dispersion functionals $\phi_{k,\delta}(\mu) = \left[\mathsf{E}_\mu\{\mathscr{V}_k^\delta(X_0, \ldots, X_k)\}\right]$ with $k \geq 2$ and $\delta \in (0, 2)$ may define characteristic kernels remains an open issue.

5 Application: testing the equality between means and covariances

We illustrate the behaviour of the distances presented in previous sections by considering the situation where one wishes to test whether two distributions μ and ζ have the same mean and covariance, using empirical data. We denote

$$\hat{a}_{\mu,n}^{(i)} = \frac{1}{n} \sum_{k=1}^{n} X_k^{(i)} \text{ and } \hat{\Sigma}_{\mu,n}^{(i)} = \frac{1}{n-1} \sum_{k=1}^{n} (X_k^{(i)} - a_{\mu,n}^{(i)})(X_k^{(i)} - a_{\mu,n}^{(i)})^\top$$

the sample mean and covariance matrix for a sample $\mathbf{X}_n^{(i)} = \{X_1, \ldots, X_n\}$ of n independent d-dimensional vectors distributed with μ, and similarly $\hat{a}_{\zeta,m}^{(i)}$ and $\hat{\Sigma}_{\zeta,m}^{(i)}$ for a sample $\mathbf{Y}_m^{(i)}$ of m independent vectors distributed with ζ. We denote by $D(\hat{\mu}_n^{(i)}, \hat{\zeta}_m^{(i)})$ the distance D computed with the empirical values $\hat{a}_{\mu,n}^{(i)}$, $\hat{\Sigma}_{\mu,n}^{(i)}$, $\hat{a}_{\zeta,m}^{(i)}$ and $\hat{\Sigma}_{\zeta,m}^{(i)}$.

5.1 ROC curves

Suppose we have N pairs of independent samples $\mathbf{X}_n^{(i)}$ and $\mathbf{Y}_m^{(i)}$, $i = 1, 2, \ldots, N$, respectively distributed $\mathscr{N}(a_\mu, \Sigma_\mu)$ and $\mathscr{N}(a_\zeta, \Sigma_\zeta)$. Each pair $(\mathbf{X}_n^{(i)}, \mathbf{Y}_m^{(i)})$ yields an empirical distance $D(\hat{\mu}_n^{(i)}, \hat{\zeta}_m^{(i)})$, with D one of the distances considered above, and the N pairs give an empirical estimate of the c.d.f. F_1 of $D(\hat{\mu}_n, \hat{\zeta}_m)$. Similarly, pairs $(\mathbf{X}_n^{(i)}, \mathbf{X}_n^{(j)})$ yield an empirical estimate of the c.d.f. F_0 of $D(\hat{\mu}_n, \hat{\mu}_n)$.

Denote H0 the hypothesis that two given samples $\mathbf{X}_n$ and $\mathbf{Y}_m$ have the same mean and covariance and H1 the hypothesis that they have different means and/or covariances. A standard statistical test based on D would compare the distance calculated for the empirical estimates $\hat{a}_{\mu,n}$, $\hat{\Sigma}_{\mu,n}$, $\hat{a}_{\zeta,m}$ and $\hat{\Sigma}_{\zeta,m}$ to some critical value τ. A plot of $1 - \mathsf{F}_1$ against $1 - \mathsf{F}_0$ gives the Receiver Operating Characteristic (ROC) curve for the test. It shows the value of the true positive rate against the false positive rate at various threshold settings τ, and the power of the test as a function of the type-1 error of the decision rule. The Area Under the ROC Curve (AUC) gives a scalar figure of merit for the performance of the test considered.

Example 1 We use pairs of samples with equal size $n = m = 200$ in dimension $d = 20$. Detection of different means is far easier than detection of slightly different covariances, and we take $a_\mu = a_\zeta = (1, \ldots, 1)^\top$. The covariances are

$$\Sigma_\mu = \begin{pmatrix} A & 0 \\ 0 & 10^{-3}\, I_{d-2} \end{pmatrix} \quad \text{and} \quad \Sigma_\zeta = \begin{pmatrix} \alpha\, A & 0 \\ 0 & 10^{-3}\, I_{d-2} \end{pmatrix} , \tag{22}$$

with $A = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$ and I_{d-2} the $(d-2)$-dimensional identity matrix. The empirical estimates of the c.d.f. F_0 and F_1 are built from $N = 1,000$ pairs of normal samples.

The left panel of Figure 1 presents the ROC curve obtained when $\alpha = 1.4$ in (22), for Bhattacharyya distance D_B (6) (dashed line, bottom), $D_{\log \varphi_p, JB}$ (16) with $p = 1/2$ (dotted line) and $D_{\log \Phi_k, JB}$ (20) with $k = 3$ (solid line). The right panel of Figure 1 shows the AUC as α varies between 1 and 2 for these three distances. The curves obtained with Burbea-Rao divergence $D_{\log \Phi_3, BR}$ cannot be visually distinguished from those obtained with $D_{\log \Phi_3, JB}$. Note the similar behaviours observed for $D_{\log \varphi_{1/2}, JB}$ and $D_{\log \Phi_3, JB}$ on this example, both performing much better than D_B. The curves obtained with $D_{\log \varphi_0, JB}$ or $D_{\log \Phi_d, JB}$ (not shown) are hardly distinguishable form those with D_B; distances $D_{\log \varphi_p, JB}$ with $p < 0$ perform very poorly due to the high sensitivity to the presence of small eigenvalues in the spectra of Σ_μ and Σ_ζ. □

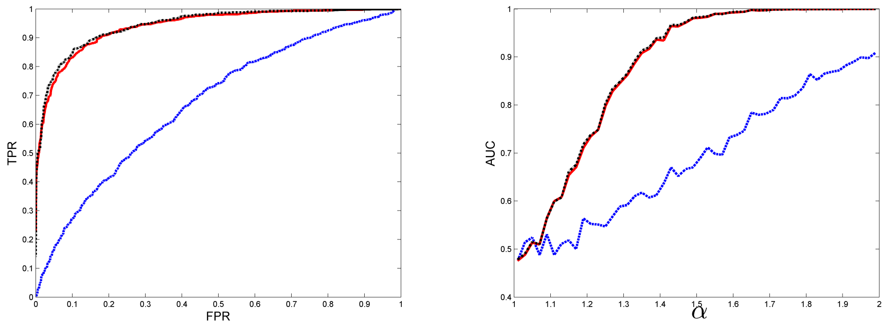

Fig. 1 Left: ROC curve for Bhattacharyya distance D_B (6) (dashed line, bottom), $D_{\log \varphi_{1/2}, JB}$ (16) (dotted line) and $D_{\log \Phi_3, JB}$ (20) (solid line) when $\alpha = 1.4$ in (22). Right: Area Under the ROC Curve as a function of α.

As the next example illustrate, the ranking of the different methods is not always the same as in Example 1.

Example 2 We slightly modify the setting of Example 1, and consider now covariances given by

$$\Sigma_\mu = \begin{pmatrix} A & 0 \\ 0 & I_{d-2} \end{pmatrix} \quad \text{and} \quad \Sigma_\zeta = \begin{pmatrix} R_\theta A R_\theta^\top & 0 \\ 0 & I_{d-2} \end{pmatrix} , \tag{23}$$

with A as in (22) and R_θ the rotation matrix

$$R_\theta = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix} .$$

We still have $a_\mu = a_\zeta = (1, \ldots, 1)^\top$, $n = m = 200$ and $d = 20$. The left panel of Figure 2 presents the ROC curve obtained when $\theta = \pi/16$ in (23), for Bhattacharyya distance D_B, $D_{\log\varphi_{1/2},JB}$ and $D_{\log\Phi_3,JB}$, with the same colour code as in Figure 1. The right panel of Figure 2 shows the AUC as θ varies between 0 and $\pi/2$ for these three distances. Again, the curves obtained with Burbea-Rao divergence $D_{\log\Phi_3,BR}$ cannot be visually distinguished from those obtained with $D_{\log\Phi_3,JB}$, and the curves obtained with $D_{\log\varphi_0,JB}$ (or $D_{\log\Phi_d,JB}$) are hardly distinguishable form those with D_B. The three distances D_B, $D_{\log\varphi_{1/2},JB}$ and $D_{\log\Phi_3,JB}$ yield now different performances, with D_B performing best, notably better than $D_{\log\Phi_3,JB}$ in particular. □

Examples 1 and 2 show the importance of being able to choose a suitable k in $\{2,\ldots,d\}$ for $\log\Phi_k$, or a suitable p in $[0,1)$ for $\log\varphi_p$. This is considered in the next section.

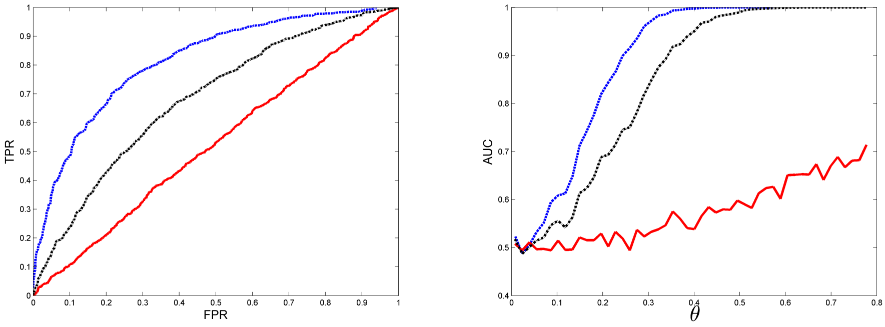

Fig. 2 Left: ROC curve for Bhattacharyya distance D_B (6) (dashed line, top), $D_{\log\varphi_{1/2},JB}$ (16) (dotted line, middle) and $D_{\log\Phi_3,JB}$ (20) (solid line, bottom) when $\theta = \pi/16 \approx 0.196$ in (23). Right: Area Under the ROC Curve as a function of θ.

5.2 Choosing k in $\log\Phi_k$ and p in $\log\varphi_p$

Ideally, for testing identity between means and covariance matrices of two distributions from one sample of each, $\mathbf{X}_n$ and $\mathbf{Y}_m$ say, one may use different distances and combine the test statistics obtained, p-values for instance. Here, we shall consider a naive approach where we first select a value k_* for k for distances based on $\log\Phi_k$, or p_* for p for distances based on $\log\varphi_p$, and then use the corresponding k_*, or p_*, in the testing procedure. A consequence of using such a simple approach is that we shall have little control of the type-I error. However, the implementation of a more precise and rigorous method would require sophisticated developments out of the scope of this paper.

When only one pair of samples, $\mathbf{X}_n$ and $\mathbf{Y}_m$, is available, we can nevertheless generate N pairs of pseudo samples from $(\mathbf{X}_n, \mathbf{Y}_m)$ and use the approach of Section 5.1 to evaluate the AUC under the ROC curve for each distance considered, for several choices of k and p. For a distance based on $\log\Phi_k$ (respectively, $\log\varphi_p$), the value k_* (respectively, p_*) that yields the largest AUC is then selected for testing the identity between the distributions that generated the two samples $\mathbf{X}_n$ and $\mathbf{Y}_m$.

For instance, we may generate pairs of pseudo samples by bootstrap. For the estimation of F_1, each $\mathbf{X}_n^{(i,1)}$ (respectively, $\mathbf{Y}_m^{(i,1)}$) is obtained by sampling with replacement within $\mathbf{X}_n$ (respectively, $\mathbf{Y}_m$). For the estimation of F_0, for each i we first merge $\mathbf{X}_n$ and $\mathbf{Y}_n$ into $\mathbf{Z}_n = \{\mathbf{X}_n, \mathbf{Y}_n\}$; then we randomly select n points from $\mathbf{Z}_n$, within which we sample with replacement to obtain $\mathbf{X}_n^{(i,0)}$ and sample with replacement within the m other points of $\mathbf{Z}_n$ to obtain $\mathbf{Y}_m^{(i,0)}$. This construction ensures that there is no intersection between $\mathbf{X}_n^{(i,0)}$ and $\mathbf{Y}_m^{(i,0)}$, so that the pairs $(\mathbf{X}_n^{(i,0)}, \mathbf{Y}_n^{(i,0)})$ do not look artificially too similar compared to the $(\mathbf{X}_n^{(i,1)}, \mathbf{Y}_n^{(i,1)})$.

We experimentally found that sampling without replacement, as described hereafter, gives better results. Take $n' = n - r$ and $m' = m - r$, with r sufficiently large to induce enough variability among pseudo samples. For the estimation of F_1, each $\mathbf{X}_{n'}^{(i,1)}$ (respectively, $\mathbf{Y}_{m'}^{(i,1)}$) is given by n' points randomly selected within $\mathbf{X}_n$ (respectively, of m' points selected within $\mathbf{Y}_m$). For the estimation of F_0, we first merge $\mathbf{X}_n$ and $\mathbf{Y}_n$ into $\mathbf{Z}_n = \{\mathbf{X}_n, \mathbf{Y}_n\}$; then we randomly select n' points from $\mathbf{Z}_n$ to form $\mathbf{X}_{n'}^{(i,0)}$ and select m' points from the remaining $n + m - n'$ points of $\mathbf{Z}_n$ to form $\mathbf{Y}_{m'}^{(i,0)}$. This construction ensures that there are no repetitions of points within $\mathbf{X}_{n'}^{(i,0)}$ and $\mathbf{Y}_{m'}^{(i,0)}$ and no intersection between them. The value of r does not need to be large: with $n = 100$, $r = 5$ already gives more that $75\ 10^6$ different choices for $\mathbf{X}_{n'}^{(i,0)}$.

Examples 1 and 2 (continued) We consider again the situation of Example 1, and draw two samples $\mathbf{X}_n$ and $\mathbf{Y}_n$ from $\mathscr{N}(a_\mu, \Sigma_\mu)$ and $\mathscr{N}(a_\zeta, \Sigma_\zeta)$, respectively, with $n = 200$ and $\alpha = 1.4$ in (22). The left panel of Figure 3 shows the AUC under the ROC curve for $D_{\log \Phi_k, BR}$ as a function of k, constructed according to the procedure above with $N = n$ and $r = 5$. The optimal k is here $k_* = 5$; the value of k_* fluctuates depending on the random samples $\mathbf{X}_n$ and $\mathbf{Y}_n$ that are drawn, with $k_* \leq 6$ in about 90% of the cases. The right panel of Figure 3 shows the AUC under the ROC curve for $D_{\log \varphi_p, BR}$ as a function of p. The optimal p_* varies with $\mathbf{X}_n$ and $\mathbf{Y}_n$, but remains larger than 1/2 in about 90% of the cases. These observations suggest that in this example distances based on $\log \Phi_k$ (respectively, based on $\log \varphi_p$) perform better with small k than with large k (respectively, with large p than with small p), which is confirmed by Figure 1.

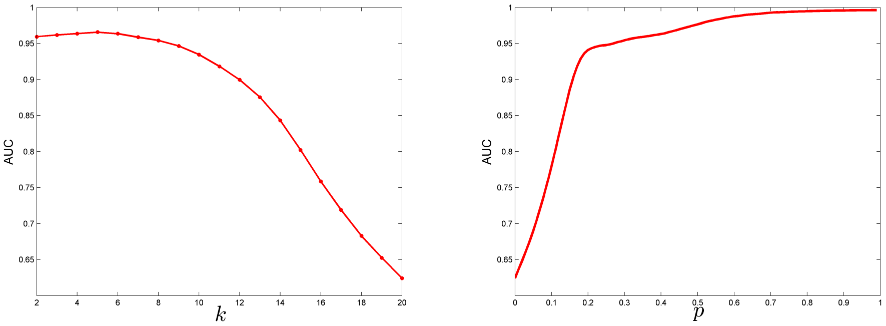

Fig. 3 AUC under the ROC curve for $D_{\log \Phi_k, BR}$ as a function of k (Left) and for $D_{\log \varphi_p, BR}$ as a function of p (Right) in the situation of Example 1 with $\alpha = 1.4$. We used $N = n = m = 200$ pseudo samples of size $n - 5 = 195$.

We repeat now the same exercice for the situation of Example 2, with $\theta = \pi/16$. The left panel of Figure 4 shows the AUC under the ROC curve for $D_{\log \Phi_k, BR}$ as a function of k, and the right panel shows the AUC under the ROC curve for $D_{\log \varphi_p, BR}$ as a function of p. We obtain $k_* = d$ and $p_* = 0$, with $D_{\log \Phi_k, BR}$ and $D_{\log \varphi_p, BR}$ being equivalent to D_{JS}, see (5), which coincides with Bhattacharyya distance D_B when the distributions have the same mean. Figure 2 confirms that D_B is indeed a good choice for this example. □

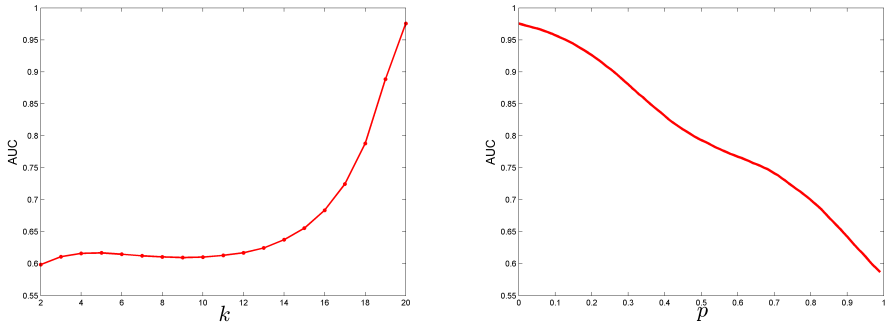

Fig. 4 AUC under the ROC curve for $D_{\log \Phi_k, BR}$ as a function of k (Left) and for $D_{\log \varphi_p, BR}$ as a function of p (Right) in the situation of Example 2 with $\theta = \pi/16$. $N = n = m = 200$, $n' = m' = 195$.

	Example 1		Example 2	
	FP	TP	FP	TP
D_B	4.5	13.6	4.5	33.0
$D_{\log \varphi_p, BR}$	12.0	79.9	9.7	35.3
$D_{\log \phi_k, BR}$	12.1	77.6	9.9	34.9

Table 1 Percentage of false positive FP (type-I error) and true positive TP obtained in 1000 repetitions for a targeted significance level of 5%.

5.3 Adjusting the critical value τ

We consider a simple (and incorrect) approach, where the N pairs $(\mathbf{X}_{n'}^{(i,0)}, \mathbf{Y}_{m'}^{(i,0)})$ of pseudo samples generated to select k or p, see Section 5.2, are also used to adjust the critical value τ of the threshold for the test statistic. Since pseudo samples have sizes $n' = n - r$ and $m' = m - r$ respectively, and distances are not invariant with respect to the sample size, we shall discard (randomly) r points from $\mathbf{X}_n$ and $\mathbf{Y}_m$ to compute the test statistic $D(\hat{\mu}_{n'}, \hat{\zeta}_{m'})$.

Generation of bootstrap samples can be considered too. In that case, we first merge $\mathbf{X}_n$ and $\mathbf{Y}_n$ into $\mathbf{Z}_n = \{\mathbf{X}_n, \mathbf{Y}_n\}$, then sample with replacement within $\mathbf{Z}_n$, the first n points give $\mathbf{X}_n^{(i,0)}$, the m next points give $\mathbf{Y}_m^{(i,0)}$ and we do not need to discard any data from from $\mathbf{X}_n$ and $\mathbf{Y}_m$ (the test statistic is $D(\hat{\mu}_n, \hat{\zeta}_m)$).

Examples 1 and 2 (continued) Empirical results (false positive FP and true positive TP) for the situation in Example 1 are given in the left part of Table 1. For H0, $\mathbf{X}_n$ and $\mathbf{Y}_n$ are normal samples generated with μ; for H1, $\mathbf{X}_n$ is generated with μ and $\mathbf{Y}_n$ with ζ. The experiment is repeated 1000 times, the significance level is set at 5%. The value of k_* is searched within $\{1, \ldots, d\}$ and that of p_* within $\{0, 0.01, \ldots, 0.99\}$. Results for Example 2 are indicated in the right part of the table.

In both examples, the percentage of false positive is notably larger than the targeted significance level of 5%, pointing out the weakness of the naive plug-in approach based on a selection of the best values k_* and p_* for k and p. Nevertheless, the percentage of true positives with a distance based on $\log \varphi_p$ or $\log \phi_k$ is much higher than for Bhattacharyya distance in Example 1 and is similar to the one with Bhattacharyya distance in Example 2. These promising results confirm what can be observed in Figures 1 and 2. □

5.4 Example 3: comparison of means and covariances for the Wine Recognition Data

We consider the wine data-set of the machine-learning repository, see `www.mlr.cs.umass.edu/ml/datasets/Wine`, widely used in particular as a test-bed for comparing classifiers. Here we simply consider the three classes of the data-set as three different data-sets **X**, **Y** and **Z** and wish to test whether they significantly differ in their means and/or covariances. The data have dimension $d = 14$ and the sample sizes are 59, 71 and 48. The empirical covariance have very large leading eigenvalue (larger than 10^4) but also several eigenvalues smaller than one.

The left panel of Figure 5 shows the value of distance $D_{\log\phi_k,BR}$ computed for the empirical measures associated with the second and third data sets, **Y** and **Z** as a function of $k \in \{2,\ldots,d\}$. The curve in solid line (bottom) is when all data points are used, the one in dashed line (top) is when r points are removed from each sample, see Section 5.3; we use $r = 5$. The right panel of Figure 5 shows (a kernel approximation of) the pdf of $D_{\log\phi_{10},BR}$ obtained from 1000 bootstrap samples under H0, see Section 5.3; the observed distance (corresponding to the value for $k = 10$ on the curve in solid line on the left panel) is indicated by a vertical dashed line. The hypothesis H0 that both samples come from distributions having the same mean and covariance is clearly rejected. The figure obtained is similar when using sampling without replacement with $r = 5$, see Section 5.2. Similarly, H0 is also rejected for all other $k \in \{2,\ldots,d\}$, and when using $D_{\log\varphi_p,BR}$ for all $p = 0, 0.01, \ldots, 0.99$. The same conclusions are obtained when comparing the distributions of **X** and **Y**, and **X** and **Z**. They also remain unchanged when the three samples are first centered, indicating that they all have different covariances.

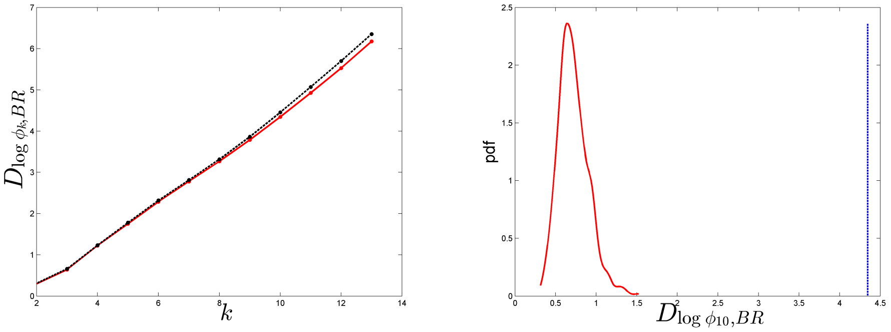

Fig. 5 Left: $D_{\log\phi_k,BR}$ for the empirical measures associated with **Y** and **Z** as a function of k; solid line (bottom) when all data points are used, dashed line (top) when $n_2 - r$ and $n_3 - r$ points are used ($r = 5$). Right: kernel approximation of the pdf of $D_{\log\phi_{10},BR}$ using 1000 bootstrap samples under H0 and observed value of $D_{\log\phi_{10},BR}$ (vertical dashed line).

Since the data-sets **X**, **Y** and **Z** have different sizes, we may exploit the fact that

$$(\widehat{\phi}_k)_n = \frac{(n-k-1)!(n-1)^k}{(n-1)!}\,\phi_k(\widehat{\Sigma}_{\mu,n})$$

forms an unbiased estimator of $\phi_k(\Sigma_\mu)$ with minimum variance among all unbiased estimators, see Theorem 3.2 in [17]. This modification in the calculation of ϕ_k does not change the conclusions above for this example.

References

1. A.C. Atkinson, A.N. Donev, and R.D. Tobias. *Optimum Experimental Designs, with SAS*. Oxford University Press, 2007.
2. M. Basseville. Divergence measures for statistical data processing — An annotated bibliography. *Signal Processing*, 93(4):621–633, 2013.
3. A. Bhattacharyya. On a measure of divergence between two multinomial populations. *Sankhyā: The Indian Journal of Statistics*, 7(4):401–406, 1946.
4. G. Björck. Distributions of positive mass, which maximize a certain generalized energy integral. *Arkiv för Matematik*, 3(21):255–269, 1956.
5. V.V. Fedorov. *Theory of Optimal Experiments*. Academic Press, New York, 1972.
6. V.V. Fedorov and P. Hackl. *Model-Oriented Design of Experiments*. Springer, Berlin, 1997.
7. V.V. Fedorov and S.L. Leonov. *Optimal Design for Nonlinear Response Models*. CRC Press, Boca Raton, 2014.
8. B.A. Frigyik, S. Srivastava, and M.R. Gupta. Functional Bregman divergence and Bayesian estimation of distributions. *IEEE Transactions on Information Theory*, 54(11):5130–5139, 2008.
9. J. Kiefer. General equivalence theory for optimum designs (approximate theory). *Annals of Stat.*, 2(5):849–879, 1974.
10. J. López-Fidalgo and J.M. Rodríguez-Díaz. Characteristic polynomial criteria in optimal experimental design. In A.C. Atkinson, L. Pronzato, and H.P. Wynn, editors, *Advances in Model–Oriented Data Analysis and Experimental Design, Proceedings of MODA'5, Marseilles, June 22–26, 1998*, pages 31–38. Physica Verlag, Heidelberg, 1998.
11. S. Łukaszyk. A new concept of probability metric and its applications in approximation of scattered data sets. *Computational Mechanics*, 33(4):299–304, 2004.
12. F. Nielsen and S. Boltz. The Burbea-Rao and Bhattacharyya centroids. *IEEE Transactions on Information Theory*, 57(8):5455–5466, 2011.
13. F. Nielsen and R. Nock. Generalizing Jensen and Bregman divergences with comparative convexity and the statistical Bhattacharyya distances with comparable means. *arXiv preprint arXiv:1702.04877*, 2017.
14. A. Pázman. *Foundations of Optimum Experimental Design*. Reidel (Kluwer group), Dordrecht (co-pub. VEDA, Bratislava), 1986.
15. L. Pronzato and A. Pázman. *Design of Experiments in Nonlinear Models. Asymptotic Normality, Optimality Criteria and Small-Sample Properties*. Springer, LNS 212, New York, 2013.
16. L. Pronzato, H.P. Wynn, and A. Zhigljavsky. Extremal measures maximizing functionals based on simplicial volumes. *Statistical Papers*, 57(4):1059–1075, 2016. hal-01308116.
17. L. Pronzato, H.P. Wynn, and A. Zhigljavsky. Extended generalised variances, with applications. *Bernoulli*, 23(4A):2617–2642, 2017.
18. L. Pronzato, H.P. Wynn, and A.A. Zhigljavsky. Simplicial variances, potentials and Mahalanobis distances. *Journal of Multivariate Analysis*, 2018. to appear.
19. F. Pukelsheim. *Optimal Experimental Design*. Wiley, New York, 1993.
20. J.M. Rodríguez-Díaz and J. López-Fidalgo. A bidimensional class of optimality criteria involving ϕ_p and characteristic criteria. *Statistics*, 37(4):325–334, 2003.
21. R.L. Schilling, R. Song, and Z. Vondracek. *Bernstein Functions: Theory and Applications*. de Gruyter, Berlin/Boston, 2012.
22. S. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5):2263–2291, 2013.
23. A.N. Shiryaev. *Probability*. Springer, Berlin, 1996.
24. S.D. Silvey. *Optimal Design*. Chapman & Hall, London, 1980.
25. B.K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G.R.G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(Apr):1517–1561, 2010.
26. G.J. Székely and M.L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013.