Simplicial variances, potentials and Mahalanobis distances

Luc Pronzato[1] , Henry P. Wynn[2] and Anatoly A. Zhigljavsky[3]

Keywords dispersion; generalised variance; potential; scatter, Mahalanobis distance; Bregman divergence; characteristic polynomial

MSC 94A17, 62B10, 62K05

1. Introduction

A rather common problem in multivariate statistical data analysis involves measuring the scatter of a data-set. Classical approaches rely on the empirical covariance matrix (or a robust version of it). Most frequently, this matrix is close to being degenerate, with several small eigenvalues. In such situations, many standard methods, including analysis via the generalised variance, may not be applicable. Hence, the need of methods that concentrate their attention on subspaces of appropriate dimensions. In [17], the authors introduced a class of extended generalised k-variances for a probability measure μ on $\mathbb{R}^d$ with covariance matrix $\Sigma = \Sigma_\mu$. These measures of dispersion are indexed by an integer parameter $k \in \{1, \ldots, d\}$. When $k = 1$ the generalised k-variance becomes $\mathrm{Tr}(\Sigma)$ and when $k = d$ we obtain the usual generalised variance $\det(\Sigma)$. For general $1 \leq k \leq d$, the k-variance is the sum of the determinants of all the $k \times k$ principal minors of Σ; that is, the sum of generalised variances for all k-dimensional minors.

The simplicial nature of the results stems from a theorem which, up to a circumstantial multiplier, equates the extended generalised variance to the expected squared volume of simplices formed from independent copies of the random vector associated with μ; for the value k we take $k + 1$ copies.

A main idea of this paper is that an integral measure of dispersion generates a notion of potential at a general point x and dependent on μ. A main result relates the notion of simplicial potential obtained here to a generalised Mahalanobis distance, expressed as a weighted sum of such distances in every k-margin. We show also that the potential arises from the directional derivative, towards x, of the simplicial variance, and that the matrix involved in the generalised Mahalanobis distance is a particular generalised inverse of Σ, constructed from its characteristic polynomial, when $k = \mathrm{rank}(\Sigma)$. Finally, simplicial potentials yield simplicial distances between two distributions, depending on their means and covariances, which are particular Jeffreys-Bregman divergences, with interesting features when the distributions are close to being singular.

The paper is organised as follows. Section 2 sets the notation and introduces the main notions of simplicial variance and potential. The construction of empirical generalised k-variances is provided and the choice of k is discussed. The generalised Mahalanobis distance and the simplicial distance between two distributions are developed and studied in Section 3. Three examples are presented in Section 4, including a real-life example used to illustrate the importance of the choice of an appropriate k.

2. Simplicial variances and potentials

2.1. Notation

- $\mathscr{M}$ is the set of non-degenerate probability measures on Borel sets of $\mathbb{R}^d$ with finite mean a_μ and finite non-zero covariance matrix Σ_μ.
- Λ_μ is the set of eigenvalues of Σ_μ.

[1]Université Côte d'Azur, CNRS, I3S, Sophia Antipolis, France; Luc.Pronzato@cnrs.fr (Corresponding author)
[2]London School of Economics, London, UK; H.Wynn@lse.ac.uk
[3]School of Mathematics, Cardiff University, Cardiff, UK; ZhigljavskyAA@cf.ac.uk

- $\Lambda(\Sigma)$ is the set of eigenvalues of a square matrix Σ.
- k is an integer, $k \in \{1, \dots, d\}$.
- $\mathscr{V}_k(x_0, \dots, x_k)$ is the volume of the k-dimensional simplex (its length when $k = 1$ and area when $k = 2$) formed by the $k+1$ vertices $x_0, \dots, x_k \in \mathbb{R}^d$.
- $e_k(L)$ is the elementary symmetric function of degree k of a set $L = \{\ell_1, \dots, \ell_d\}$, defined as

$$e_k(L) = \sum_{1 \le i_1 < \cdots < i_k \le d} \ell_{i_1} \dots \ell_{i_k} \,. \tag{1}$$

- $\mathrm{adj}(C)$ is the adjoint of a $k \times k$ matrix C: if $\det(C) \neq 0$ then $\mathrm{adj}(C) = \det(C) \cdot C^{-1}$, otherwise $\mathrm{adj}(C)$ is the zero matrix of size $k \times k$.
- $b_{i_1,\dots,i_k} = (b_{i_1}, \dots, b_{i_k})^\top$ is the vector in $\mathbb{R}^k$ formed by extracting components from the vector $b = (b_1, \dots, b_d)^\top \in \mathbb{R}^d$, with $1 \le i_1 < \dots < i_k \le d$.
- $B_{i_1,\dots,i_k}$ is the principal $k \times k$ submatrix of a matrix B of size $d \times d$ formed by picking up rows and columns with indices $i_1, \dots, i_k$, with $1 \le i_1 < \dots < i_k \le d$.
- $\overline{\mathrm{adj}\,(\Sigma_{i_1,\dots,i_k})}$ is the $d \times d$ matrix formed from the $k \times k$ matrix $\mathrm{adj}\,(\Sigma_{i_1,\dots,i_k})$ by inserting zeroes for all pairs of indices $(u, v) \in \{1, \dots, d\} \times \{1, \dots, d\}$ such that u or v is not in $\{i_1, \dots, i_k\}$, with $1 \le i_1 < \dots < i_k \le d$.

2.2. *Integral measure of dispersion, directional derivative and potential*

Consider any general functional $\psi(\mu)$ defined on $\mathscr{M}$. From [7], $\psi(\mu)$ admits an unbiased estimator if and only if it takes the form

$$\psi(\mu) = \int \dots \int \phi(x_0, \dots, x_k)\, \mu(\mathrm{d}x_0) \dots \mu(\mathrm{d}x_k) \tag{2}$$

for some function ϕ. Without loss of generality, we can assume that the kernel ϕ is symmetric. From [8, Th. 2, p. 2], there exists a unique symmetric unbiased estimator of $\psi(\mu)$, which is given by

$$\hat{\psi}^n(X_1, \dots, X_n) = \frac{(n-k-1)!}{n!} \sum \phi(X_{i_0}, \dots, X_{i_k}) \,, \tag{3}$$

where the sum extends over all $n!/(n-k-1)!$ permutations of the sample $\mathbb{X}_n = \{X_1, \dots, X_n\}$. Moreover, $\hat{\psi}^n(X_1, \dots, X_n)$ has minimum variance over all unbiased estimators of $\psi(\mu)$ [8, Th. 3, p. 3].

This paper investigates properties of particular measures of dispersion, or scatter, having the integral form (2) with ϕ non negative (and non identically zero). A fundamental property here is that for any functional of this form we can derive a potential which naturally arises from the notion of directional derivative.

The potential of μ at x for the functional $\psi(\cdot)$ in (2) is obtained by considering $x_0 = x$ as fixed:

$$P_\mu(x) = \int \dots \int \phi(x, x_1, \dots, x_k)\, \mu(\mathrm{d}x_1) \dots \mu(\mathrm{d}x_k) \,.$$

Clearly, $\psi(\mu) = \int P_\mu(x)\, \mu(\mathrm{d}x)$. We show that the potential $P_\mu(x)$ is strongly related to the notion of directional derivative of $\psi(\cdot)$ at μ in the direction of the delta-measure δ_x at x, defined as follows:

$$F(\mu, x) = \frac{\partial \psi[(1-\alpha)\mu + \alpha\delta_x]}{\partial \alpha}\Big|_{\alpha=0+} \,.$$

Theorem 1. *Potentials $P_\mu(x)$ are expressed through the directional derivatives $F(\mu, x)$ as*

$$P_\mu(x) = \frac{1}{k+1} F(\mu, x) + \psi(\mu) \,. \tag{4}$$

Proof. We have

$$
\begin{aligned}
F(\mu, x) &= \lim_{\alpha \to 0+} \frac{1}{\alpha} \left\{ \int \dots \int \phi(x_0, \dots, x_k) \left[\prod_{i=0}^{k} (\mu + \alpha(\delta_x - \mu))(\mathrm{d}x_i) \right] - \psi(\mu) \right\} \\
&= (k+1) \int \dots \int \phi(x_0, \dots, x_k)(\delta_x - \mu)(\mathrm{d}x_0) \left[\prod_{i=1}^{k} \mu(\mathrm{d}x_i) \right] \\
&= (k+1) \left[P_\mu(x) - \psi(\mu) \right],
\end{aligned}
$$

which yields (4). ∎

Of particular interest are situations where the potential $P_\mu(x)$ is a convex function of x for any μ. In this case, the potential $P_\mu(\cdot)$ can be considered as an outlyingness function (perhaps with some normalisation), measuring how far a point is from the core of the distribution, and $P_\mu(x)$ can also be considered as a measure of scatter of μ around x; see, e.g., [22]. Any point $\bar{x}_\mu$ minimizing $P_\mu(x)$ (unique if $P_\mu(\cdot)$ is strictly convex) can be considered as a central point for μ and defines a generalised median for μ associated with ψ. If $P_\mu(\bar{x}_\mu) > 0$, it can be considered as a central measure of scatter (around $\bar{x}_\mu$), alternative to $\psi(\mu)$. The two measures of scatter $P_\mu(\bar{x}_\mu)$ and $\psi(\mu)$ may coincide in some cases; see [21, 22] and Section 2.6. Of course, all this is of special interest when μ is the empirical measure of some sample.

2.3. Simplicial variances, directional derivatives and potentials

In the rest of the paper we consider the special case where $\phi(x_0, \dots, x_k) = \mathscr{V}_k^2(x_0, \dots, x_k)$ in (2), with $\mathscr{V}_k(x_0, \dots, x_k)$ the volume of the k-dimensional simplex formed by the $k+1$ vertices $x_0, \dots, x_k$. We denote by $\psi_k(\mu)$ the corresponding functional, that is

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

which we call the simplicial k-variance of μ, extending the interpretation of the generalised variance of [1, Th. 7.5.2, p. 268] to simplices of dimension smaller than d. In particular, for $k = 1$ we have

$$\psi_1(\mu) = \int \int \|x_1 - x_2\|^2 \mu(\mathrm{d}x_1)\mu(\mathrm{d}x_2) = 2\,\mathrm{Tr}[\Sigma_\mu],$$

twice the trace of the covariance matrix of μ. The potential of μ at x is then

$$P_{k,\mu}(x) = \mathsf{E}_\mu\{\mathscr{V}_k^2(x, X_1, \dots, X_k)\}.$$

Geometrically, this is the expected squared volume of k-simplices formed by x and k random vectors independently distributed with μ.

In [17], the authors have proved the following theorem and lemma.

Theorem 2. *For any $k \in \{1, \dots, d\}$ and $\mu \in \mathscr{M}$, we have*

$$\psi_k(\mu) = \frac{k+1}{k!}\, e_k(\Lambda_\mu), \tag{5}$$

with Λ_μ the set of eigenvalues of Σ_μ, the covariance matrix of μ, and $e_k(\cdot)$ the elementary symmetric function of degree k. Moreover, the functional $\psi_k^{1/k}(\cdot)$ is concave on $\mathscr{M}$.

In the following, we shall denote

$$\Psi_k(\Sigma) = \frac{k+1}{k!}\, e_k[\Lambda(\Sigma)],$$

with $\Lambda(\Sigma)$ the set of eigenvalues of the matrix Σ, so that $\psi_k(\mu) = \Psi_k(\Sigma_\mu)$. In particular, when $k = d$ we get $\psi_d(\mu) = (d+1)/d!\, \det(\Sigma_\mu)$, which is proportional to the generalised variance widely used in multivariate statistics.

Lemma 1. *The directional derivative of $\psi_k(\cdot)$ at μ in the direction δ_x is*

$$F_k(\mu, x) = (x - a_\mu)^\top \nabla_k(\mu)(x - a_\mu) - k\psi_k(\mu),$$

where $a_\mu = \mathsf{E}_\mu\{X\}$ and $\nabla_k(\mu)$ is the $d \times d$ gradient matrix

$$\nabla_k(\mu) = \partial \Psi_k(A)/\partial A\big|_{A=\Sigma_\mu}.$$

Using Lemma 1 and (4), we obtain

$$P_{k,\mu}(x) = \frac{1}{k+1}\left[(x-a_\mu)^\top \nabla_k(\mu)(x-a_\mu) + \psi_k(\mu)\right], \tag{6}$$

where, using Lemma 2 in the Appendix, the gradient matrices $\nabla_k(\mu)$ can be computed as follows:

$$\nabla_k(\mu) = \frac{k+1}{k!}\sum_{i=0}^{k-1}(-1)^i e_{k-i-1}(\Lambda_\mu)\Sigma_\mu^i . \tag{7}$$

We obtain in particular

$$\begin{aligned}
\nabla_1(\mu) &= 2\, I_d\,,\\
\nabla_2(\mu) &= \frac{3}{2}\,[\mathrm{Tr}(\Sigma_\mu)\, I_d - \Sigma_\mu]\,,\\
\nabla_3(\mu) &= \frac{1}{3}\,[\mathrm{Tr}^2(\Sigma_\mu) - \mathrm{Tr}(\Sigma_\mu^2)]\, I_d - \frac{2}{3}\,\mathrm{Tr}(\Sigma_\mu)\,\Sigma_\mu + \frac{2}{3}\,\Sigma_\mu^2\,,\\
\nabla_d(\mu) &= \frac{d+1}{d!}\,\mathrm{adj}(\Sigma_\mu)\,.
\end{aligned}$$

Note that $\mathsf{E}_\mu\{P_{k,\mu}(X)\} = \psi_k(\mu)$ and (6) imply

$$\mathrm{Tr}[\Sigma_\mu \nabla_k(\mu)] = k\,\psi_k(\mu)\,. \tag{8}$$

Also, Lemma 3 in the Appendix indicates that the gradient matrix $\nabla_k(\mu)$ is positive definite when $\mathrm{rank}(\Sigma_\mu) \geq k$.

2.4. Empirical simplicial variances

Let $\mathbb{X}_n = \{x_1, \ldots, x_n\}$ be a sample of n vectors of $\mathbb{R}^d$, i.i.d. with the measure μ, and denote the sample mean and variance-covariance matrix by

$$\widehat{x}_n = \frac{1}{n}\sum_{i=1}^n x_i \quad \text{and} \quad \widehat{\Sigma}_n = \frac{1}{n-1}\sum_{i=1}^n (x_i - \widehat{x}_n)(x_i - \widehat{x}_n)^\top .$$

For $k \geq 1$, consider the empirical estimate

$$(\widehat{\psi_k})_n = \binom{n}{k+1}^{-1} \sum_{1\leq j_1<j_2<\cdots<j_{k+1}\leq n} \mathscr{V}_k^2(x_{j_1}, \ldots, x_{j_{k+1}})\,,$$

see (3). The following theorem is proved in [17].

Theorem 3. *For $\{x_1, \ldots, x_n\}$ a sample of n vectors of $\mathbb{R}^d$, i.i.d. with the measure μ, and for any $k \in \{1, \ldots, d\}$, we have*

$$(\widehat{\psi_k})_n = \frac{(n-k-1)!(n-1)^k}{(n-1)!}\,\Psi_k(\widehat{\Sigma}_n)\,, \tag{9}$$

and $(\widehat{\psi_k})_n$ forms an unbiased estimator of $\psi_k(\mu)$ with minimum variance among all unbiased estimators.

The value of $(\widehat{\psi_k})_n$ only depends on $\widehat{\Sigma}_n$, with $\mathsf{E}\{(\widehat{\psi_k})_n\} = \psi_k(\Sigma_\mu)$. From [20, Lemma A, p. 183], if $\mathsf{E}_\mu\{\mathscr{V}_k^4(X_1, \ldots, X_{k+1})\} < \infty$, then the variance of $(\widehat{\psi_k})_n$ satisfies

$$\mathrm{var}[(\widehat{\psi_k})_n] = \frac{(k+1)^2}{n}\,\mathrm{var}_\mu[P_{k,\mu}(X)] + O(n^{-2})\,.$$

Other properties of U-statistics apply to the estimator $(\widehat{\psi_k})_n$, including almost-sure consistency and the classical law of the iterated logarithm, see [20, Section 5.4]. In particular, $(\widehat{\psi_k})_n$ is asymptotically normal, $\sqrt{n}[(\widehat{\psi_k})_n - \psi_k(\mu)] \xrightarrow{\mathrm{d}} \mathcal{N}(0, (k+1)^2\,\mathrm{var}_\mu[P_{k,\mu}(X)])$. One may refer for instance to [15] for a comprehensive survey of results on the asymptotic distribution of eigenvalues of empirical covariance matrices and the asymptotic moments of associated elementary symmetric functions; see also [1, Chap. 7] and [5, Chap. 10]. The variance of $(\widehat{\psi_k})_n$ can also be estimated by jackknife or bootstrap methods, see [8, Chap. 5].

2.5. Alternative representations of simplicial potentials

Refining the arguments used in [17] for proving Theorem 2, we establish the following property.

Theorem 4. *For any $\mu \in \mathscr{M}$, any $k \in \{1, \ldots, d\}$ and any $x \in \mathbb{R}^d$, we have*

$$P_{k,\mu}(x) = \frac{1}{k!}\, e_k \left[\Lambda\left(\Sigma_\mu + (x - a_\mu)(x - a_\mu)^\top\right)\right] . \tag{10}$$

Proof. Consider the squared volume $\mathscr{V}_k^2(x, x_1, \ldots, x_k)$. By the Binet-Cauchy formula, see, e.g., [6, vol. 1, p. 9], we obtain

$$\begin{aligned}
\mathscr{V}_k^2(x, x_1, \ldots, x_k) &= \frac{1}{(k!)^2} \det\left(\begin{bmatrix} (x_1 - x)^\top \\ (x_2 - x)^\top \\ \vdots \\ (x_k - x)^\top \end{bmatrix} [(x_1 - x)\, (x_2 - x)\, \cdots (x_k - x)] \right) \\
&= \frac{1}{(k!)^2} \sum_{1 \le i_1 < i_2 < \cdots < i_k \le d} \det^2 \begin{bmatrix} \{x_1 - x\}_{i_1} & \cdots & \{x_k - x\}_{i_1} \\ \vdots & \vdots & \vdots \\ \{x_1 - x\}_{i_k} & \cdots & \{x_k - x\}_{i_k} \end{bmatrix} \\
&= \frac{1}{(k!)^2} \sum_{1 \le i_1 < i_2 < \cdots < i_k \le d} \det\left[\sum_{i=1}^k (x_i - x)_{i_1,\ldots,i_k} (x_i - x)^\top_{i_1,\ldots,i_k}\right] .
\end{aligned}$$

From the definition of the potential $P_{k,\mu}(x)$, we obtain

$$\begin{aligned}
P_{k,\mu}(x) &= \int \cdots \int \mathscr{V}_k^2(x, x_1, \ldots, x_k) \mu(\mathrm{d}x_1) \ldots \mu(\mathrm{d}x_k) \\
&= \frac{1}{(k!)^2} \sum_{1 \le i_1 < i_2 < \cdots < i_k \le d} \int \cdots \int \det\left[\sum_{i=1}^k (x_i - x)_{i_1,\ldots,i_k} (x_i - x)^\top_{i_1,\ldots,i_k}\right] \mu(\mathrm{d}x_1) \ldots \mu(\mathrm{d}x_k) .
\end{aligned}$$

From Lemma 4 in the Appendix, with $Z_i = (x_i - x)_{i_1,\ldots,i_k}$, we get

$$\begin{aligned}
P_{k,\mu}(x) &= \frac{1}{k!} \sum_{1 \le i_1 < i_2 < \cdots < i_k \le d} \det\left[\mathsf{E}_\mu\{(X - x)_{i_1,\ldots,i_k} (X - x)^\top_{i_1,\ldots,i_k}\}\right] \\
&= \frac{1}{k!} \sum_{1 \le i_1 < i_2 < \cdots < i_k \le d} \det\left[\mathsf{E}_\mu\{(X - a_\mu)_{i_1,\ldots,i_k} (X - a_\mu)^\top_{i_1,\ldots,i_k} + (a_\mu - x)_{i_1,\ldots,i_k} (a_\mu - x)^\top_{i_1,\ldots,i_k}\}\right] \\
&= \frac{1}{k!} \sum_{1 \le i_1 < i_2 < \cdots < i_k \le d} \det\left\{\left[\Sigma_\mu + (x - a_\mu)(x - a_\mu)^\top\right]_{i_1,\ldots,i_k}\right\} . \qquad (11)
\end{aligned}$$

Lemma 5 in the Appendix completes the proof. ∎

When all $k \times k$ principal minors of Σ_μ have rank at least k, Theorem 4 provides an alternative representation for $P_{k,\mu}(x)$.

Corollary 1. *When* $\mathrm{rank}(\Sigma_{i_1,\ldots,i_k}) \ge k$ *for all* $1 \le i_1 < i_2 < \cdots < i_k \le d$*, the gradient matrix* $\nabla_k(\mu)$ *in* (6) *can be expressed as*

$$\nabla_k(\mu) = \frac{k+1}{k!} \sum_{1 \le i_1 < i_2 < \cdots < i_k \le d} \overline{\mathrm{adj}\,(\Sigma_{i_1,\ldots,i_k})} , \tag{12}$$

where we have denoted $\Sigma = \Sigma_\mu$*.*

Proof. Each determinant $\det\left\{\left[\Sigma + (x - a_\mu)(x - a_\mu)^\top\right]_{i_1,\ldots,i_k}\right\}$ in (11) can be represented as

$$\begin{aligned}
\det\left\{\left[\Sigma + (x - a_\mu)(x - a_\mu)^\top\right]_{i_1,\ldots,i_k}\right\} &= \left[1 + (x - a_\mu)^\top_{i_1,\ldots,i_k} \Sigma^{-1}_{i_1,\ldots,i_k} (x - a_\mu)_{i_1,\ldots,i_k}\right] \det(\Sigma_{i_1,\ldots,i_k}) \\
&= \det(\Sigma_{i_1,\ldots,i_k}) + (x - a)^\top_{i_1,\ldots,i_k} \mathrm{adj}\,(\Sigma_{i_1,\ldots,i_k})\, (x - a)_{i_1,\ldots,i_k} .
\end{aligned}$$

By Lemma 5 in the Appendix and Theorem 2, we have

$$\frac{1}{k!} \sum_{1 \le i_1 < i_2 < \cdots < i_k \le d} \det\left(\Sigma_{i_1,\dots,i_k}\right) = \frac{1}{k!} e_k(\Lambda_\mu) = \frac{1}{k+1} \psi_k(\mu)\,.$$

Therefore, formula (11) yields

$$P_{k,\mu}(x) = \frac{1}{k+1}\psi_k(\mu) + \frac{1}{k!} \sum_{1 \le i_1 < i_2 < \cdots < i_k \le d} (x-a)^\top_{i_1,\dots,i_k} \operatorname{adj}\left(\Sigma_{i_1,\dots,i_k}\right)(x-a)_{i_1,\dots,i_k}\,. \tag{13}$$

The statement of the corollary follows from (13) and (6). ■

2.6. A generalisation of results of Wilks and van der Vaart

Equation (6) and Lemma 3 show that the potential $P_{k,\mu}(x)$ is a quadratic convex function of x, with minimum value $\psi_k(\mu)/(k+1) \ge 0$ attained at $x = a_\mu$. As mentioned in Section 2.2, when $P_{k,\mu}(a_\mu) > 0$, it can be considered as a central measure of scatter. The relations between $P_{k,\mu}(a_\mu)$ and $\psi_k(\mu)$ have been investigated by Wilks [22] and van der Vaart [21] for the case $k = d$ where $\psi_d(\mu) = (d+1)/d!\ \det(\Sigma_\mu)$. The following theorem extends their results to general $k \in \{1, \cdots, d\}$. Note that the case $k = 1$, with $\psi_1(\mu) = 2\operatorname{Tr}[\Sigma_\mu]$, is classical.

Theorem 5. *For any $\mu \in \mathscr{M}$ and any $k \in \{1, \dots, d\}$, we have*

$$a_\mu = \arg\min_x P_{k,\mu}(x)\,. \tag{14}$$

Moreover,

$$P_{k,\mu}(x) > P_{k,\mu}(a_\mu) = \frac{1}{k!}\, e_k(\Lambda_\mu) = \frac{\psi_k(\mu)}{k+1} > 0$$

for all $x \neq a_\mu$ if and only if $\operatorname{rank}(\Sigma_\mu) \ge k$.

Proof. Equation (14) is a direct consequence of (6) and of the fact that the gradient matrix $\nabla_k(\mu)$ is non-negative definite, see Lemma 3.

Assume first that $\operatorname{rank}(\Sigma_\mu) \ge k$; then $e_k(\Lambda_\mu) > 0$ and $\nabla_k(\mu)$ is positive definite from Lemma 3. Therefore $P_{k,\mu}(x) > P_{k,\mu}(a_\mu) > 0$ for $x \neq a_\mu$.

Assume now that $\operatorname{rank}(\Sigma_\mu) < k$, which implies $P_{k,\mu}(a_\mu) = 0$. Choose any $z \neq 0$ in the subspace spanned by the eigenvectors of Σ_μ corresponding to the non-zero eigenvalues of Σ_μ and consider the form (10) for $P_{k,\mu}(x)$. Since the ranks of the matrices Σ_μ and $\Sigma_\mu + zz^\top$ coincide, $P_{k,\mu}(x) = 0$ for $x = z + a_\mu$. ■

2.7. Choosing k

Since the simplicial k-variance $\psi_k(\mu)$ is constructed from volumes of k-dimensional simplices, its standardised version $\psi_k^{1/k}(\cdot)$ allows us to compare scatters of different dimensional distributions, similarly to the standardised generalised variance used by SenGupta [19] which corresponds to the case $k = d$. Newton inequalities for symmetric functions indicate that

$$\left(\frac{e_k(\Lambda_\mu)}{\binom{d}{k}}\right)^{1/k} > \left(\frac{e_{k+1}(\Lambda_\mu)}{\binom{d}{k+1}}\right)^{1/(k+1)}$$

for all $k = 1, \dots, d-1$ unless all eigenvalues in Λ_μ coincide, see [13, p. 213]. Also, one can check that, for any d, $(k!/[(k+1)\binom{d}{k}])^{1/k}$ increases with k, $1 \le k \le d$. This implies that $\psi_k^{1/k}(\mu)$ is strictly decreasing with k, also when all eigenvalues in Λ_μ coincide. This remains true when considering the empirical version (9) with a large enough n, since the correcting factor satisfies $(n-k-1)!(n-1)^k/(n-1)! = 1 + k(k-1)/(2n) + O(1/n^2)$.

The consideration of $\psi_k^{1/k}(\cdot)$ does not allow us to make a recommendation concerning the most appropriate k. We can simply notice that $\psi_k^{1/k}(\mu) = 0$ when μ is concentrated in a d'-dimensional subspace with $d' < k$. However, numerical experimentation indicates that the estimation of the approximate dimensionality of a data-set is easier by simple inspection of the eigenvalues of the empirical covariance matrix $\widehat{\Sigma}_n$ than by setting a threshold on values of $\Psi_k(\widehat{\Sigma}_n)$.

By extending the definition of $\psi(\cdot)$ in (2) to arbitrary positive measures, we may consider the variation of $\psi_k(\mu)$ when μ is changed into $\mu + \alpha\delta_x$ for a small α. This corresponds to considering the influence function

$$G_k(\mu, x) = \frac{\partial \psi_k[\mu + \alpha\delta_x]}{\partial \alpha}\Big|_{\alpha=0+} .$$

An appropriate k should then yield large values of $G_k(\mu, x)$ to achieve high sensitivity of the measure of scatter of μ to deviations from μ. Similarly to Lemma 1, we obtain $G_k(\mu, x) = (x - a_\mu)^\top \nabla_k(\mu)(x - a_\mu)$. Averaging $G_k(\mu, X)$ with $X \sim \mu$, we get from (8)

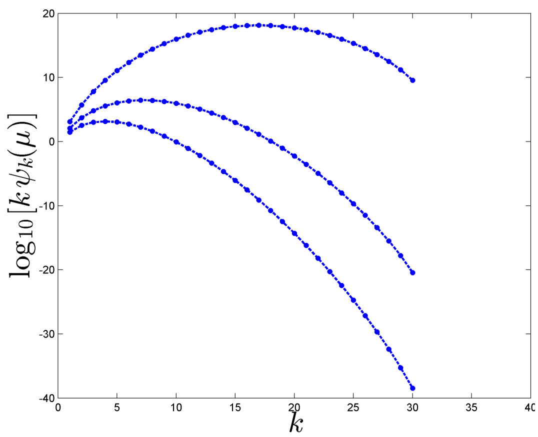
$$\mathsf{E}_\mu\{G_k(\mu, X)\} = \mathrm{Tr}[\Sigma_\mu \nabla_k(\mu)] = k\,\psi_k(\mu) .$$

As a result, choosing k_* that maximises $k\,\psi_k(\mu)$ (or $k\,(\widehat{\psi_k})_n$ given by (9) for empirical data) appears most appropriate.

The value of k_* depends on the scale of the data. As an example, assume that Σ_μ has $d' \le d$ eigenvalues equal to β and $d - d'$ equal to zero. Then,

$$k\,\psi_k(\mu) = \frac{k+1}{(k-1)!}\binom{d'}{k}\beta^k , \; k \le d' , \tag{15}$$

and $k\,\psi_k(\mu) = 0$ for $k > d'$. To determine the associated k_*, we compute the ratio $\rho(k) = k\,\psi_k(\mu)/[(k+1)\,\psi_{k+1}(\mu)] = k(k+1)^2/[\beta(k+2)(d'-k)]$. If $\beta < 4/[3(d'-1)]$, then $\rho(1) > 1$ and $k_* = 1$, if $\beta > d'^2(d'-1)/(d'+1)$ then $\rho(d'-1) < 1$ and therefore $k_* = d'$. Otherwise, we find t_* as the solution of the cubic equation $\rho(t) = 1$ and get $k_* = \lceil t_* \rceil$. Figure 1-left presents the evolution of $k\,\psi_k(\mu)$ (in log scale) as a function of k for $\beta = 20$, 2 and 0.5, from top to bottom, when $d' = 30$; the corresponding values of k_* are 17, 7 and 4, respectively. Figure 1-right shows $k\,\psi_k(\mu)$ (log scale) when Σ_μ has eigenvalues $\Lambda_\mu = \{\beta, \beta/2, \ldots, \beta/30, 0, \ldots, 0\}$, also for $\beta = 20$ (top), 2 and 0.5 (bottom), with associated k_* equal to 7, 3 and 2. Both figures indicate that a small k is preferable when Σ_μ has small eigenvalues and illustrate the difficulty of estimating the true dimensionality of the data when there are several eigenvalues smaller than one, due to the fast decrease of $\psi_k(\mu)$ as a function of k. This point is further illustrated in the example of Section 4.3.



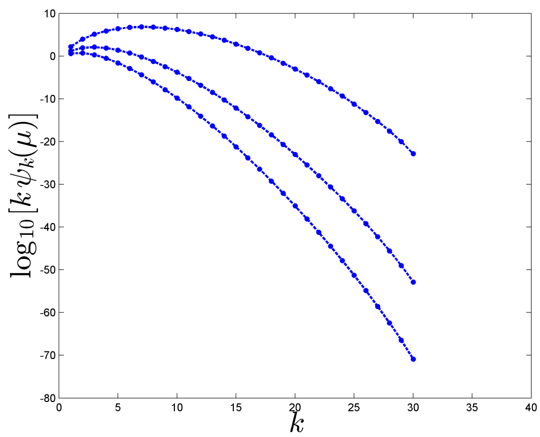


Figure 1: $k\,\psi_k(\mu)$ for $k = 1, \ldots, d' = 30$, for $\beta = 20$ (top), 2 (middle) and 0.5 (bottom). Left: $\Lambda_\mu = \{\beta, \ldots, \beta, 0, \ldots, 0\}$ and $k\,\psi_k(\mu)$ is given by (15); Right: $\Lambda_\mu = \{\beta, \beta/2, \ldots, \beta/30, 0, \ldots, 0\}$.

3. Simplicial Mahalanobis distances

3.1. From simplicial potentials to Mahalanobis distances

Consider a measure $\mu \in \mathscr{M}$ such that $P_{k,\mu}(a_\mu) = e_k(\Lambda_\mu)/k! > 0$. For this measure we define the function

$$O_{k,\mu}(x) = \frac{P_{k,\mu}(x)}{P_{k,\mu}(a_\mu)} - 1 = (x - a_\mu)^\top S_{k,\mu}(x - a_\mu) , \tag{16}$$

with

$$S_{k,\mu} = \frac{\nabla_k(\mu)}{\psi_k(\mu)}, \tag{17}$$

where the second equality follows from (6). In the special case $k = d$, (12) gives $\nabla_d(\mu) = (d+1)/d!\ \det(\Sigma_\mu) \cdot \Sigma_\mu^{-1}$ and (5) gives $\psi_d(\mu) = (d+1)/d!\ \det(\Sigma_\mu)$, so that

$$O_{d,\mu}(x) = (x - a_\mu)^\top \Sigma_\mu^{-1} (x - a_\mu),$$

which is exactly the original Mahalanobis distance [10]. We will call $O_{k,\mu}(\cdot)$ the k-simplicial outlyingness function, which can also be thought of as a simplicial Mahalanobis distance between x and μ. Geometrically, it is a suitably normalised version of the expected squared volume of k-simplices formed by x and k random vectors independently distributed according to μ, and measures the distance from x to the central point a_μ.

The definition (16) of $O_{k,\mu}(x)$ implies that $O_{k,\mu}(x) \geq 0$ for any $\mu \in \mathscr{M}$, any $k \in \{1, \ldots, d\}$ and any x. Also, $\mathsf{E}_\mu\{P_{k,\mu}(X)\} = \psi_k(\mu)$ implies that $\mathsf{E}_\mu\{O_{k,\mu}(X)\} = k$, and therefore

$$\max_{x \in \mathscr{X}} O_{k,\mu}(x) \geq k \tag{18}$$

for any set $\mathscr{X}$ having full measure, i.e., such that $\mu(\mathscr{X}) = 1$. On the other hand, Theorem 4.1 in [17] gives a necessary and sufficient condition on μ to have equality in (18) for a given set $\mathscr{X}$: μ must maximise $\psi_k(\cdot)$ over the set of all measures supported on $\mathscr{X}$.

In view of Theorem 5,

$$\min_x O_{k,\mu}(x) = O_{k,\mu}(a_\mu) = 0$$

and $O_{k,\mu}(x) > 0$ for all $x \neq a_\mu$. The quadratic form in (16) defines a metric on $\mathbb{R}^d$, and we define the *k-th order simplicial Mahalanobis distance* relative to μ (or simply k-distance) between z_1 and z_2 in $\mathbb{R}^d$ as

$$\delta_{k,\mu}(z_1, z_2) = O_{k,\mu}(z_1 - z_2 + a_\mu) = (z_1 - z_2)^\top S_{k,\mu} (z_1 - z_2).$$

The geometric interpretation of $\delta_{k,\mu}(z_1, z_2)$ when $\mu = \mu_n$ is a centralised empirical measure of a sample $\mathbb{X}_n$ is that $1 + \delta_{k,\mu}(z_1, z_2)$ is proportional to the sum of squared volumes of all simplices formed by $z_1 - z_2$ and all k-tuples of the sample $\mathbb{X}_n$.

As already mentioned, when $k = d$ we get $O_{d,\mu}(x) = (x - a_\mu)^\top \Sigma_\mu^{-1} (x - a_\mu)$. For $k = 1$, we obtain

$$O_{1,\mu}(x) = \|x - a_\mu\|^2 / \mathrm{Tr}(\Sigma_\mu),$$

which is the usual squared Euclidean distance between x and a_μ normalised by the trace of Σ_μ.

For general k, when all $k \times k$ principal minors of Σ_μ have rank at least k, from (5) and (12) we have

$$\begin{aligned}
O_{k,\mu}(x) &= \frac{1}{e_k(\Lambda_\mu)} \sum_{1 \leq i_1 < i_2 < \cdots < i_k \leq d} (x - a_\mu)_{i_1,\ldots,i_k}^\top \operatorname{adj}(\Sigma_{i_1,\ldots,i_k})\, (x - a_\mu)_{i_1,\ldots,i_k} \\
&= \frac{1}{e_k(\Lambda_\mu)} \sum_{1 \leq i_1 < i_2 < \cdots < i_k \leq d} \det(\Sigma_{i_1,\ldots,i_k}) \cdot (x - a_\mu)_{i_1,\ldots,i_k}^\top \Sigma_{i_1,\ldots,i_k}^{-1} (x - a_\mu)_{i_1,\ldots,i_k},
\end{aligned}$$

where $\Sigma = \Sigma_\mu$. Since $e_k(\Lambda_\mu) = \sum \det(\Sigma_{i_1,\ldots,i_k})$, see Lemma 5, the simplicial Mahalanobis distance of order k, $\delta_{k,\mu}(z_1, z_2)$, is then the weighted sum of the usual Mahalanobis distances of all k-th marginal vectors, with weights given by the corresponding determinants.

3.2. Construction through characteristic polynomial and generalised inverse

The expression (7) of the gradient matrix $\nabla_k(\mu)$ allows us to express the matrix $S_{k,\mu}$ in (16) in terms of the characteristic polynomial of Σ_μ, and to show that $S_{k,\mu}$ is a generalised inverse of Σ_μ when Σ_μ has rank k.

Denote by $p_\mu(\cdot)$ the characteristic polynomial of the $d \times d$ matrix Σ_μ,

$$p_\mu(\lambda) = \sum_{i=0}^{d} (-1)^i e_i(\Lambda_\mu) \lambda^{d-i}.$$

For any $k \in \{1, \ldots, d\}$, we introduce a truncated version $p_{k,\mu}(\lambda)$ of $p_\mu(\lambda)$ which only contains terms of degree at least $d-k$,

$$p_{k,\mu}(\lambda) = \lambda^{d-k} \sum_{i=0}^{k} (-1)^{k-i} e_{k-i}(\Lambda_\mu) \lambda^i \,,$$

which we rewrite as

$$p_{k,\mu}(\lambda) = \lambda^{d-k} (-1)^{k+1} \left[\lambda q_{k,\mu}(\lambda) - e_k(\Lambda_\mu) \right] \,, \tag{19}$$

where

$$q_{k,\mu}(\lambda) = \sum_{i=0}^{k-1} (-1)^i e_{k-i-1}(\Lambda_\mu) \lambda^i \,. \tag{20}$$

Comparing (7) with (20), we obtain $\nabla_k(\mu) = (k+1)/k!\, q_{k,\mu}(\Sigma_\mu)$. Therefore, $S_{k,\mu}$ in (16) becomes

$$S_{k,\mu} = \frac{q_{k,\mu}(\Sigma_\mu)}{e_k(\Lambda_\mu)} \,.$$

Theorem 6. *If* $\mathrm{rank}(\Sigma_\mu) = k \leq d$, *then the matrix* $S_{k,\mu}$ *is a generalised inverse of* Σ_μ *(inverse if* $k = d$*). When* Σ_μ *has eigenvalues* $\lambda_1 \geq \cdots \lambda_k > \lambda_{k+1} = \cdots = \lambda_d = 0$, $S_{k,\mu}$ *has eigenvalues* $\zeta_j = 1/\lambda_j$ *for* $j = 1, \ldots, k$ *and* $\zeta_j = \sum_{i=1}^{k} 1/\lambda_i$ *for* $j = k+1, \ldots, d$*; moreover,* $S_{k,\mu}$ *and* Σ_μ *have the same eigenspaces.*

Proof. Assume $\mathrm{rank}(\Sigma_\mu) = k \leq d$. We need to verify the generalised inverse condition $\Sigma_\mu S_{k,\mu} \Sigma_\mu = \Sigma_\mu$. We have:

$$\Sigma_\mu S_{k,\mu} \Sigma_\mu - \Sigma_\mu = \Sigma_\mu \frac{q_{k,\mu}(\Sigma_\mu)}{e_k(\Lambda_\mu)} \Sigma_\mu - \Sigma_\mu = \frac{1}{e_k(\Lambda_\mu)} \Sigma_\mu \left[\Sigma_\mu q_{k,\mu}(\Sigma_\mu) - e_k(\Lambda_\mu) I_d \right] \,. \tag{21}$$

Since $\mathrm{rank}(\Sigma_\mu) = k$, the characteristic polynomial $p_\mu(\cdot)$ of the matrix Σ_μ is equal to $p_{k,\mu}(\cdot)$. The matrix Σ_μ satisfies its own characteristic equation, and therefore $p_{k,\mu}(\Sigma_\mu) = 0$. In view of (19), this gives

$$\Sigma_\mu^{d-k} \left[\Sigma_\mu q_{k,\mu}(\Sigma_\mu) - e_k(\Lambda_\mu) I_d \right] = 0 \,. \tag{22}$$

If $k = d$ or $k = d-1$ this implies $\Sigma_\mu S_{k,\mu} \Sigma_\mu = \Sigma_\mu$, see (21).

Let us assume $k < d-1$. From (22), all eigenvalues λ_i of the matrix Σ_μ satisfy

$$\lambda_i^{d-k} \left[\lambda_i q_{k,\mu}(\lambda_i) - e_k(\Lambda_\mu) \right] = 0 \,. \tag{23}$$

For each $i = 1, \ldots, d$ this implies that either $\lambda_i = 0$ or $\left[\lambda_i q_{k,\mu}(\lambda_i) - e_k(\Lambda_\mu) \right] = 0$. In either case we obtain $\lambda_i \left[\lambda_i q_k(\lambda_i) - e_k(\Lambda_\mu) \right] = 0$, which yields $\Sigma_\mu S_{k,\mu} \Sigma_\mu = \Sigma_\mu$.

The fact that $S_{k,\mu}$ is a polynomial in Σ_μ implies that they have the same eigenspaces. The eigenvalues ζ_j of $S_{k,\mu}$ are $q_{k,\mu}(\lambda_j)/e_k(\Lambda_\mu)$. If $\lambda_j \neq 0$, then (23) implies $\zeta_j = 1/\lambda_j$. If $\lambda_j = 0$, then (20) gives $\zeta_j = e_{k-1}(\Lambda_\mu)/e_k(\Lambda_\mu) = \sum_{i=1}^{k} 1/\lambda_i$. ■

3.3. A simplicial distance between two distributions

Let μ_1 and μ_2 be two probability measures in $\mathscr{M}$. The average squared volume of a k-simplex with one vertex coming from measure μ_1 and k vertices i.i.d. from μ_2 equals $\mathsf{E}_{\mu_1}\{P_{k,\mu_2}(X)\}$. Symmetrising and normalising, we naturally arrive at the following expression

$$\Delta_k(\mu_1, \mu_2) = \frac{1}{2} \left[\mathsf{E}_{\mu_1}\{O_{k,\mu_2}(X)\} + \mathsf{E}_{\mu_2}\{O_{k,\mu_1}(X)\} \right] - k \,,$$

where $O_{k,\mu}(\cdot)$ is the outlyingness function defined in (16). We shall informally consider $\Delta_k(\mu_1, \mu_2)$ as a measure of distance between μ_1 and μ_2, although $\Delta_k(\mu_1, \mu_2)$ does not in general satisfy the triangular inequality and only depends on the means and covariance matrices of μ_1 and μ_2. Expanding $\mathsf{E}_{\mu_2}\{O_{k,\mu_1}(X)\}$, and denoting $\Sigma_i = \Sigma_{\mu_i}$ and $a_i = a_{\mu_i}$ for $i = 1, 2$, we get

$$\mathsf{E}_{\mu_2}\{O_{k,\mu_1}(X)\} = \mathrm{Tr}(S_{k,\mu_1} \Sigma_2) + (a_2 - a_1)^\top S_{k,\mu_1} (a_2 - a_1) \,,$$

therefore,

$$\Delta_k(\mu_1,\mu_2) = \frac{1}{2}\left[\mathrm{Tr}(S_{k,\mu_1}\Sigma_2) + \mathrm{Tr}(S_{k,\mu_2}\Sigma_1)\right] + (a_2-a_1)^\top \frac{S_{k,\mu_1}+S_{k,\mu_2}}{2}(a_2-a_1) - k\,. \tag{24}$$

Note that the substitution of the Moore-Penrose pseudo inverses Σ_i^+ for S_{k,μ_i} in (24) would lead to negative distance values for some measures with singular Σ_i.

Direct calculation shows that $\Delta_k(\mu_1,\mu_2)$ corresponds to the Jeffreys-Bregman divergence between μ_1 and μ_2 (see [2, 14]) for $\ln\psi_k(\cdot)$, that is,

$$\Delta_k(\mu_1,\mu_2) = \frac{1}{2}\left[F_{\ln\psi_k}(\mu_1,\mu_2) + F_{\ln\psi_k}(\mu_2,\mu_1)\right],$$

with $F_{\ln\psi_k}(\mu,\nu) = F_k(\mu,\nu)/\psi_k(\mu)$ the directional derivative of $\ln\psi_k(\cdot)$ at μ in the direction ν.

In the particular case when $k = d$ and both matrices Σ_1 and Σ_2 are invertible, we obtain

$$\Delta_d(\mu_1,\mu_2) = \frac{1}{2}\left[\mathrm{Tr}(\Sigma_1^{-1}\Sigma_2) + \mathrm{Tr}(\Sigma_2^{-1}\Sigma_1)\right] + (a_2-a_1)^\top \frac{\Sigma_1^{-1}+\Sigma_2^{-1}}{2}(a_2-a_1) - d\,,$$

which is non negative since $A + A^{-1} \geq 2\,I_d$ for any $d\times d$ matrix $A > 0$, with equality if and only if $A = I_d$. Therefore, $\Delta_d(\mu_1,\mu_2) = 0$ implies $a_1 = a_2$ and $\Sigma_1 = \Sigma_2$. It resembles the Bhattacharyya distance between two normal distributions,

$$\Delta_B(\mu_1,\mu_2) = \frac{1}{2}\ln\left[\frac{\det(\Sigma_1+\Sigma_2)}{\sqrt{\det(\Sigma_1)\det(\Sigma_2)}}\right] + \frac{1}{4}(a_2-a_1)^\top(\Sigma_1+\Sigma_2)^{-1}(a_2-a_1) - \frac{d}{2}\ln(2)\,,$$

but is not equivalent to it. In particular, $\Delta_B(\mu_1,\mu_2)$ cannot be used when at least one of the distributions is singular, whereas $\Delta_k(\mu_1,\mu_2)$ can, see (24). The example in Section 4.2 gives an illustration with distributions close to singularity.

When $k = 1$, $S_{1,\mu} = I_d/\mathrm{Tr}(\Sigma_\mu)$, and therefore

$$\Delta_1(\mu_1,\mu_2) = \frac{1}{2}\left[\frac{\mathrm{Tr}(\Sigma_1)}{\mathrm{Tr}(\Sigma_2)} + \frac{\mathrm{Tr}(\Sigma_2)}{\mathrm{Tr}(\Sigma_1)}\right] + \frac{1}{2}\|a_2-a_1\|^2\left[\frac{1}{\mathrm{Tr}(\Sigma_1)} + \frac{1}{\mathrm{Tr}(\Sigma_2)}\right] - 1\,,$$

which is clearly non negative. However, $\Delta_1(\mu_1,\mu_2) = 0$ only implies $a_1 = a_2$ and $\mathrm{Tr}(\Sigma_1) = \mathrm{Tr}(\Sigma_2)$, showing that Δ_1 is arguably a less interesting measure of discrepancy between distributions. On the other hand, when $k > 1$ we have the following property.

Theorem 7. *For any $k \in \{2,\ldots,d\}$, $\Delta_k(\mu_1,\mu_2) \geq 0$ for any two measures μ_1 and μ_2 in $\mathscr{M}$ such that $\mathrm{rank}(\Sigma_1) \geq k$ and $\mathrm{rank}(\Sigma_2) \geq k$; moreover $\Delta_k(\mu_1,\mu_2) = 0$ implies $a_1 = a_2$ and $\Sigma_1 = \Sigma_2$.*

Proof. The proof relies on the strict concavity of $\Psi_k^{1/k}(\cdot)$, see Lemma 6 in the Appendix. Concavity implies that

$$\psi_k^{1/k}(\mu_1) + \frac{1}{k}\,\frac{\mathrm{Tr}\left\{\nabla_k(\mu_1)[\Sigma_2-\Sigma_1]\right\}}{\psi_k^{1-1/k}(\mu_1)} \geq \psi_k^{1/k}(\mu_2)\,,$$

that is, $\mathrm{Tr}\left[S_{k,\mu_1}\Sigma_2\right] \geq k\,\psi_k^{1/k}(\mu_2)/\psi_k^{1/k}(\mu_1)$, see (8) and (17), with equality when $\Sigma_2 = \gamma_2\,\Sigma_1$ for some $\gamma_2 > 0$ ($\gamma_2 \neq 0$ since $\mu_2 \in \mathscr{M}$). Similarly, $\mathrm{Tr}\left[S_{k,\mu_2}\Sigma_1\right] \geq k\,\psi_k^{1/k}(\mu_1)/\psi_k^{1/k}(\mu_2)$, with equality implying $\Sigma_1 = \gamma_1\,\Sigma_2$ for some $\gamma_1 > 0$. Therefore, (24) gives

$$\Delta_k(\mu_1,\mu_2) \geq \frac{k}{2}\left\{\left[\frac{\psi_k^{1/k}(\mu_2)}{\psi_k^{1/k}(\mu_1)} + \frac{\psi_k^{1/k}(\mu_1)}{\psi_k^{1/k}(\mu_2)}\right] - 2\right\} + (a_2-a_1)^\top\frac{S_{k,\mu_1}+S_{k,\mu_2}}{2}(a_2-a_1)\,.$$

Since, from Lemma 3, S_{k,μ_1} and S_{k,μ_2} are positive definite, $\Delta_k(\mu_1,\mu_2) \geq 0$, and equality implies that $a_1 = a_2$ and $\psi_k(\mu_1) = \psi_k(\mu_2)$. When $k \geq 2$, equality also implies that $\Sigma_2 = \gamma\,\Sigma_1$ for some $\gamma > 0$, and $\gamma = 1$ since $\psi_k(\mu_1) = \psi_k(\mu_2)$. ∎

4. Examples

4.1. Clustering with the simplicial Mahalanobis distance

We consider a clustering problem for which we apply a k-means algorithm with Lloyd's type iterations [9], with three different intra-class distances: the Euclidean distance (leading to the usual k-means algorithm), Mahalanobis distance with Moore-Penrose pseudo-inverse if needed, and the k-simplicial Mahalanobis distance with $k = 3$.

We consider two examples, each with two clusters of $n/2 = 50$ points, respectively with $d = 50$ and $d = 100$. The 50 points of cluster i are normally distributed $\mathcal{N}(b_i, W_i)$, with $b_1 = 0$, $b_2 = (1, 1, 0.5, 0.5, \ldots, 0.5)^\top$ for $d = 50$, $b_2 = (1, 1, 0.1, 0.1, \ldots, 0.1)^\top$ for $d = 100$, and

$$W_1 = \begin{pmatrix} \begin{pmatrix} 5 & -4 \\ -4 & 5 \end{pmatrix} & 0 \\ 0 & \sigma_d^2 I_{d-2} \end{pmatrix}, \quad W_2 = \begin{pmatrix} \begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix} & 0 \\ 0 & \sigma_d^2 I_{d-2} \end{pmatrix}.$$

We used $\sigma_{50} = 10^{-2}$ and $\sigma_{100} = 10^{-9}$ and performed 1,000 runs of each algorithm (initialised in the same way for each of the 1,000 samples, with 100 iterations every time). The performances of the algorithms are summarised in Figure 2. We plot the empirical cdf, over the 1,000 runs, of the classification error rate introduced by Chipman and Tibshirani [3], which gives the proportion of misclassified pairs in one run of the algorithm.

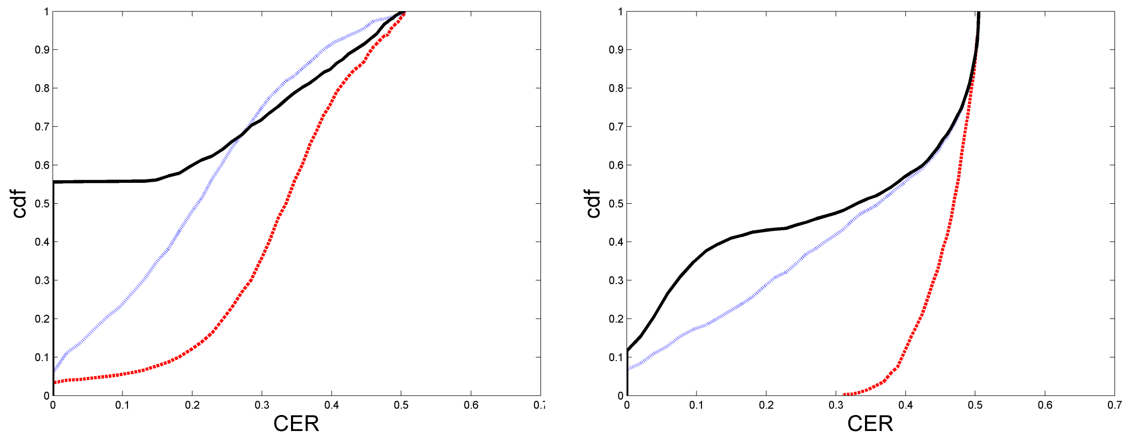


Figure 2: Empirical cdf, over the 1,000 runs, of the Classification Error Rates (CER) when clustering $n = 100$ points with the Euclidean distance (dashed line), the Mahalanobis distance (dotted line) and the 3-simplicial Mahalanobis distance (solid line); Left: $d = 50$; Right: $d = 100$.

The results illustrate the property that the substitution of the k-simplicial Mahalanobis distance for the usual one may significantly improve performance of some classical algorithms of multivariate statistics, in cases when the data are high-dimensional but lie very close to a subspace of much lower dimension. Choosing $k = k_*$ as suggested in Section 2.7 at each iteration makes the algorithm slightly more complicated than when k is fixed at 3 and does not yield any visible improvement in performance. When $d = 100$, in addition to the presence of a delta measure at zero (which also exists for clustering with the Mahalanobis distance), the distribution of classification error rates also has a mode at low error rates for the 3-simplicial Mahalanobis distance. For all three methods, the worst misclassification occurs when all points are assigned to one cluster or when one cluster only contains two points that should belong to different clusters (which gives here a CER value of $n/[2(n-1)] \simeq 0.505$). The performance of clustering with k-simplicial Mahalanobis distance significantly improves when increasing the number of points in each cluster: for example, with 400 points in each cluster in the setting above with $d = 50$, perfect classification is obtained in 1,000 repetitions for $k = 3, 4, 5, 6$. On the other hand, for clustering with Euclidean and Mahalanobis distances, the CER remains similar to the case with 50 points per cluster depicted in Figure 2-left.

4.2. Comparison between Bhattacharyya and simplicial distances

Consider two d-dimensional distributions μ_1 and μ_2 with means a_1 and a_2 and covariance matrices

$$\Sigma_1 = \left(\begin{array}{cc} \left(\begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array} \right) & 0 \\ 0 & \alpha^2 I_{d-2} \end{array} \right), \ \Sigma_2 = \left(\begin{array}{cc} \left(\begin{array}{cc} 5 & 4 \\ 4 & 5 \end{array} \right) & 0 \\ 0 & \beta^2 I_{d-2} \end{array} \right).$$

First, we set $a_1 = a_2 = 0$, $\alpha = 0.01$ and $\beta = 0.001$. Figure 3-left shows that Bhattacharyya distance $\Delta_B(\mu_1, \mu_2)$ between the two distributions increases linearly with d, although intuitively the distributions look more similar as d increases. Figure 3-right shows that the behaviour of simplicial distance $\Delta_3(\mu_1, \mu_2)$ is consistent with this intuition. For illustration, we have considered $k = 3$, but other values of $k \geq 2$ yield similar behaviours.

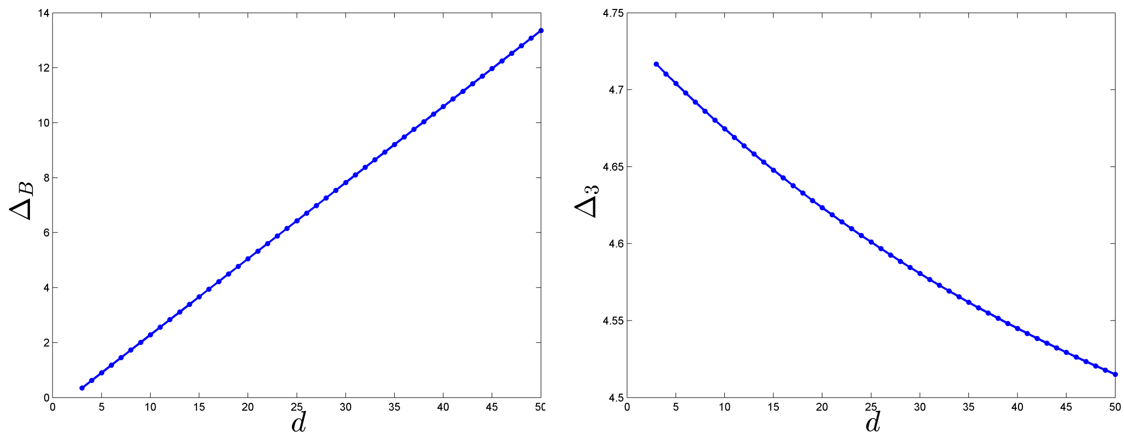


Figure 3: Distance between μ_1 and μ_2 as a function of $d \in \{3, \ldots, 50\}$. Left: Bhattacharyya distance $\Delta_B(\mu_1, \mu_2)$; Right: simplicial distance $\Delta_3(\mu_1, \mu_2)$.

Now take $\beta = \alpha$, but $a_2(1) = a_2(2) = 1$, the other components of a_2 being left equal to zero, like all components of a_1. Again, intuitively the distributions are getting more similar as d increases, but $\Delta_B(\mu_1, \mu_2)$ remains constant, whereas $\Delta_k(\mu_1, \mu_2)$ decreases with d for $k \geq 2$.

4.3. Comparing scatters of Wine Recognition Data

In this section we illustrate the use of simplicial k-variances ψ_k for comparing scatters of different data-sets. We consider the wine data-set of the machine-learning repository, see `www.mir.cs.umass.edu/ml/datasets/Wine`, widely used in particular as a test-bed for comparing classifiers. Here we use the class labels and consider the three classes of the data-set as three different data-sets. The data have dimension $d = 14$ and the sample sizes are 59, 71 and 48. The eigenvalues of the three empirical covariance matrices are plotted in Figure 4-left (in log scale). For each data-set, the leading eigenvalue is very large and several of them are much smaller than one. Figure 4-right shows the values of the standardised empirical simplicial k-variances $(\widehat{\psi_k})_n^{1/k}$ obtained using (9) and the corresponding 2σ-confidence intervals computed by jackknifing as explained in [8, Chap. 5]. As already mentioned in Section 2.7, $\psi_k^{1/k}$ is a decreasing function of k, and the decrease is very fast due to the presence of small eigenvalues. Non-standardised values of $(\widehat{\psi_k})_n$ are shown in Figure 5-left, along with their 2σ-confidence intervals (also computed with the jackknife). These two figures suggest that measuring scatter through $\psi_k^{1/k}$ (or ψ_k) with a large k is doubtful in the presence of small eigenvalues. This true in particular for the generalised variance for which $k = d$. Figure 5-right presents the values of $(\widehat{\psi_k})_n$ for $k = 1, \ldots, 5$ together with their confidence intervals (in log scale). The figure suggests that scatters of the three data-sets are slightly different.

Appendix

The Newton equations for symmetric functions and straightforward calculation yield the following properties.

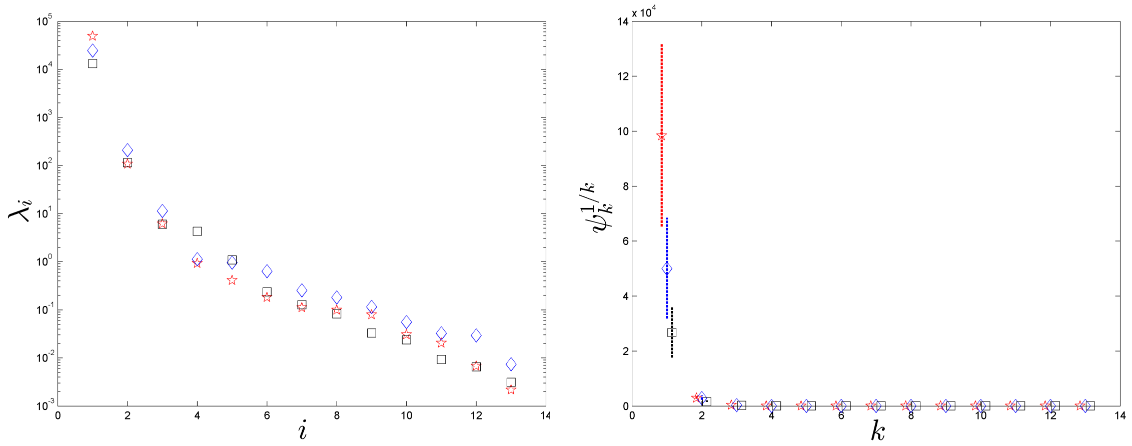


Figure 4: Left: eigenvalues (log scale) of the three empirical covariance matrices. Right: standardised empirical simplicial k-variances $(\widehat{\psi}_k)_n^{1/k}$ and 2σ-confidence intervals.

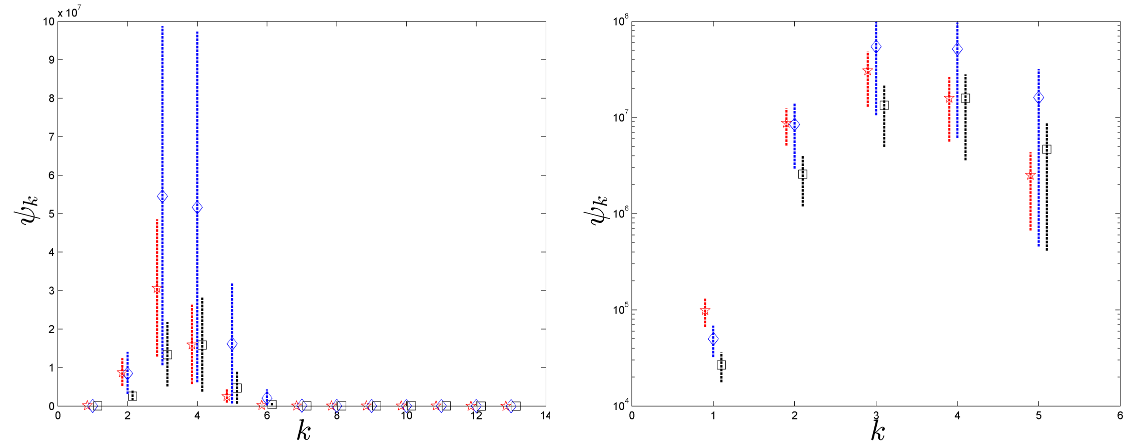


Figure 5: Left: non-standardised empirical simplicial k-variances $(\widehat{\psi}_k)_n$ and 2σ-confidence intervals. Right: values of $(\widehat{\psi}_k)_n$ for $k = 1, \ldots, 5$ and 2σ-confidence intervals (log scale).

Lemma 2. *Let* $\mathrm{V}_i(A) = e_i(\Lambda(A))$*, where* $\Lambda(A)$ *is the set of eigenvalues of a square matrix* A *(not necessarily symmetric). Then*

$$\mathrm{V}_k(A) = \frac{1}{k}\sum_{i=0}^{k-1}(-1)^{i-1}\mathrm{V}_{k-i}(A)\mathrm{Tr}(A^i) \;\; \textit{and} \;\; \frac{\partial \mathrm{V}_k(A)}{\partial A} = \sum_{i=0}^{k-1}(-1)^i \mathrm{V}_{k-i-1}(A)(A^i)^\top .$$

The next lemma indicates that $\nabla_k(\mu)$ is non-negative definite for any $\mu \in \mathscr{M}$ and is positive definite when Σ_μ has rank at least k.

Lemma 3. *For any probability measure* μ *in* $\mathscr{M}$ *and any* k *in* $\{1, \ldots, d\}$*, the gradient matrix* $\nabla_k(\mu)$ *is non-negative definite. When the covariance matrix* Σ_μ *is such that* $\mathrm{rank}(\Sigma_\mu) \geq k$*, then* $\nabla_k(\mu)$ *is positive definite.*

Proof. The proof follows the same lines as in [18, Th. 7.5]. The function $\Psi_k^{1/k}(\cdot)$ is concave, see [11, p. 116]. Therefore, the function $\ln \Psi_k(\cdot)$ is concave on the set of non-negative definite matrices, with gradient at $\Sigma = \Sigma_\mu$

given by $\nabla_k(\mu)/\psi_k(\mu)$. Concavity implies that

$$\ln \Psi_k(\Sigma_\mu + zz^\top) \leq \ln \psi_k(\mu) + \mathrm{Tr}\left[\frac{zz^\top \nabla_k(\mu)}{\psi_k(\mu)}\right].$$

By the monotonicity of the eigenvalues, for all $1 \leq i \leq d$, the i-th largest eigenvalue of $\Sigma_\mu + zz^\top$ is larger than or equal to the i-th largest eigenvalue of Σ_μ, the inequality being strict for at least one pair of eigenvalues. Therefore, $\ln \Psi_k(\Sigma_\mu + zz^\top) \geq \ln \psi_k(\mu)$, and $\mathrm{Tr}[zz^\top \nabla_k(\mu)] = z^\top \nabla_k(\mu) z \geq 0$ for any z since $\psi_k(\mu) \geq 0$, showing that $\nabla_k(\mu)$ is non-negative definite.

Suppose now that $\mathrm{rank}(\Sigma_\mu) \geq k \in \{1, \ldots, d\}$ and take $z \neq 0$. This implies $\ln \Psi_k(\Sigma_\mu + zz^\top) > \ln \psi_k(\mu)$, and therefore $z^\top \nabla_k(\mu) z > 0$ since $\psi_k(\mu) = (k+1)/k!\, e_k(\Lambda_\mu) > 0$, which completes the proof. ■

Next lemma follows from [16, Th. 1].

Lemma 4. *Let the k vectors $Z_1, \ldots, Z_k \in \mathbb{R}^k$ be i.i.d. with some probability measure μ, $k \geq 2$. Then*

$$\mathsf{E}_\mu \left\{ \det\left[\sum_{i=1}^{k} Z_i Z_i^\top\right]\right\} = k!\ \det\left[\mathsf{E}_\mu\{Z_1 Z_1^\top\}\right].$$

The following property is proved in [11, p. 22].

Lemma 5. *Let B be a non-negative definite $d \times d$ matrix with eigenvalues $\Lambda_B = (\lambda_{1,B}, \ldots, \lambda_{d,B})$. Then*

$$\sum_{1\leq i_1<i_2<\cdots<i_k\leq d} \det[\{B\}_{(i_1,\ldots,i_k)\times(i_1,\ldots,i_k)}] = \sum_{1\leq i_1<i_2<\cdots<i_k\leq d} \lambda_{i_1,B} \times \cdots \times \lambda_{i_k,B} = e_k(\Lambda_B).$$

Lemma 6. *For any probability measure $\mu \in \mathscr{M}$, the function $\Psi_k^{1/k}(\cdot)$ is strictly concave at Σ_μ for $k \geq 2$ when* $\mathrm{rank}(\Sigma_\mu) \geq k$, *that is,*

$$\Psi_k^{1/k}[(1-\alpha)\Sigma_\mu + \alpha\Sigma] > (1-\alpha)\Psi_k^{1/k}(\Sigma_\mu) + \alpha\Psi_k^{1/k}(\Sigma)$$

for any $\alpha \in (0,1)$ *and any symmetric non-negative definite matrix* $\Sigma \neq 0$ *not proportional to* Σ_μ.

Proof. The function $\Psi_k^{1/k}(\cdot)$ is concave, see Lemma 3. Suppose that

$$\Psi_k^{1/k}[(1-\beta)\Sigma_\mu + \beta\Sigma] = (1-\beta)\Psi_k^{1/k}(\Sigma_\mu) + \beta\Psi_k^{1/k}(\Sigma) \tag{.1}$$

for some $\beta > 0$. We show that (.1) implies that $\Sigma = \gamma\,\Sigma_\mu$ for some $\gamma \geq 0$.

Due to the concavity of $\Psi_k^{1/k}(\cdot)$, (.1) implies

$$\Psi_k^{1/k}[(1-\alpha)\Sigma_\mu + \alpha\Sigma] = (1-\alpha)\Psi_k^{1/k}(\Sigma_\mu) + \alpha\Psi_k^{1/k}(\Sigma),\ \alpha \in (0,\beta), \tag{.2}$$

that is

$$e_k^{1/k}\{\Lambda[(1-\alpha)\Sigma_\mu + \alpha\Sigma]\} = (1-\alpha)\, e_k^{1/k}[\Lambda(\Sigma_\mu)] + \alpha\, e_k^{1/k}[\Lambda(\Sigma)],\ \alpha \in (0,\beta).$$

Now, $\Lambda[(1-\alpha)\Sigma_\mu + \alpha\Sigma] \prec \Lambda[(1-\alpha)\Sigma_\mu] + \Lambda[\alpha\Sigma]$, with $\prec$ denoting majorisation, see [4]. The strict Shur-concavity of $e_k(\cdot)$ for $k > 1$ [12, p. 115] then implies

$$e_k\{\Lambda[(1-\alpha)\Sigma_\mu + \alpha\Sigma]\} \geq e_k\{\Lambda[(1-\alpha)\Sigma_\mu] + \Lambda[\alpha\Sigma]\} = e_k[(1-\alpha)\Lambda(\Sigma_\mu) + \alpha\Lambda(\Sigma)],$$

with equality when $\Lambda[(1-\alpha)\Sigma_\mu + \alpha\Sigma] = (1-\alpha)\Lambda(\Sigma_\mu) + \alpha\Lambda(\Sigma)$. Therefore, (.2) implies

$$e_k^{1/k}[(1-\alpha)\Lambda(\Sigma_\mu) + \alpha\Lambda(\Sigma)] = (1-\alpha)\, e_k^{1/k}[\Lambda(\Sigma_\mu)] + \alpha\, e_k^{1/k}[\Lambda(\Sigma)],\ \alpha \in (0,\beta),$$

and the strict concavity of $e_k^{1/k}(\cdot)$ for $k > 1$ [12, p. 116] implies that $\Lambda(\Sigma) = \gamma\,\Lambda(\Sigma_\mu)$ for some $\gamma \geq 0$.

We thus obtain $\Lambda[(1-\alpha)\Sigma_\mu + \alpha\Sigma] = (1-\alpha+\alpha\gamma)\Lambda(\Sigma_\mu)$, $\alpha \in (0,\beta)$. Take any z with $\|z\| = 1$ in the eigenspace of the largest eigenvalue λ of $(1-\alpha)\Sigma_\mu + \alpha\Sigma$. We have $\lambda = (1-\alpha+\alpha\gamma)\lambda'$, with λ' the largest eigenvalue of Σ_μ, and

$$\begin{aligned}\lambda = z^\top[(1-\alpha)\Sigma_\mu + \alpha\Sigma]z &= (1-\alpha)z^\top\Sigma_\mu z + \alpha z^\top\Sigma z \\ &\leq (1-\alpha)\sup_{\|z\|=1} z^\top\Sigma_\mu z + \alpha \sup_{\|z\|=1} z^\top\Sigma z = (1-\alpha+\alpha\gamma)\lambda',\end{aligned}$$

implying that z is in the eigenspace of the largest eigenvalues λ' and $\gamma\lambda'$ of Σ_μ and Σ. By repeating the same argument, we obtain that Σ_μ and Σ have the same eigenspaces, and therefore $\Sigma = \gamma\,\Sigma_\mu$. ■

Acknowledgments

The authors are grateful to an anonymous referee for his very careful reading of the paper and his many suggestions that helped us to clarify the exposition of our results.

References

[1] Anderson, T. (2003). *An Introduction to Multivariate Statistical Analysis*. Wiley, New York. Third Edition.
[2] Basseville, M. (2013). Divergence measures for statistical data processing — An annotated bibliography. *Signal Processing*, 93(4):621–633.
[3] Chipman, H. and Tibshirani, R. (2005). Hybrid hierarchical clustering with applications to microarray data. *Biostatistics*, 7:286–301.
[4] Fan, K. (1949). On a theorem of Weyl concerning eigenvalues of linear transformations. *Int. Proc. Nat. Acad. Sci. U.S.A.*, 35:652–655.
[5] Fujikoshi, Y., Ulyanov, V., and Shimizu, R. (2010). *Multivariate Statistics: High-Dimensional and Large-Sample Approximations*. Wiley, New Jersey.
[6] Gantmacher, F. (1966). *Théorie des Matrices*. Dunod, Paris.
[7] Halmos, P. (1946). The theory of unbiased estimation. *Ann. Math. Stat.*, 17:34–43.
[8] Lee, A. (1990). *U-Statistics. Theory and Pratice*. CRC Press, Boca Raton.
[9] Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Trans. on Information Theory*, 28(2):129–137.
[10] Mahalanobis, P. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2(1):49–55.
[11] Marcus, M. and Minc, H. (1964). *A Survey of Matrix Theory and Matrix Inequalities*. Dover, New York.
[12] Marshall, A., Olkin, I., and Arnold, B. (1979). *Inequalities: Theory of Majorization and its Applications*. Springer.
[13] Niculescu, C. and Persson, L.-E. (2006). *Convex Functions and Their Applications — A contemporary Approach*. Springer/Canadian Mathematical Society.
[14] Nielsen, F. and Boltz, S. (2011). The Burbea-Rao and Bhattacharyya centroids. *IEEE Transactions on Information Theory*, 57(8):5455–5466.
[15] Pillai, K. (1977). Distributions of characteristic roots in multivariate analysis Part II. Non-null distributions. *The Canadian Journal of Statistics*, 5(1):1–62.
[16] Pronzato, L. (1998). On a property of the expected value of a determinant. *Stat. & Prob. Lett.*, 39:161–165.
[17] Pronzato, L., Wynn, H., and Zhigljavsky, A. (2017). Extended generalised variances, with applications. *Bernoulli*, 23(4A):2617–2642.
[18] Pukelsheim, F. (1993). *Optimal Experimental Design*. Wiley, New York.
[19] SenGupta, A. (1987). Tests for standardized generalized variances of multivariate normal populations of possibly different dimensions. *Journal of Multivariate Analysis*, 23:2019–219.
[20] Serfling, R. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.
[21] van der Vaart, H. (1965). A note on Wilks' internal scatter. *Ann. Math. Statist.*, 36(4):1308–1312.
[22] Wilks, S. (1960). Multidimensional statistical scatter. In Olkin, I., Ghurye, S., Hoeffding, W., Madow, W., and Mann, H., editors, *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, pages 486–503. Stanford University Press, Stanford.