An extended Generalised Variance, with Applications

Luc Pronzato, Henry Wynn, Anatoly Zhigljavsky

An extended Generalised Variance, with Applications

Luc Pronzato*, Henry P. Wynn†and Anatoly A. Zhigljavsky‡

CNRS, London School of Economics and Cardiff University

November 24, 2014

Abstract

We consider a measure ψ_k of dispersion which extends the notion of Wilk's generalised variance, or entropy, for a d-dimensional distribution, and is based on the mean squared volume of simplices of dimension $k \leq d$ formed by $k+1$ independent copies. We show how ψ_k can be expressed in terms of the eigenvalues of the covariance matrix of the distribution, also when a n-point sample is used for its estimation, and prove its concavity when raised at a suitable power. Some properties of entropy-maximising distributions are derived, including a necessary and sufficient condition for optimality. Finally, we show how this measure of dispersion can be used for the design of optimal experiments, with equivalence to A and D-optimal design for $k = 1$ and $k = d$ respectively. Simple illustrative examples are presented.

MSC: Primary 94A17; secondary 62K05

keyword: dispersion, generalised variance, ,quadratic entropy, ,maximum-entropy measure, design of experiments, optimal design

1 Introduction

The idea of dispersion is fundamental to statistics and with different terminology, such as potential, entropy, information and capacity, stretches over a wide area. The variance and standard deviation are the most prevalent for a univariate distribution, and Wilks generalised variance is the term usually reserved for the determinant of the covariance matrix, V, of a multivariate distribution. Many other measures of dispersion have been introduced and a rich area comprises those that are order-preserving with respect to a dispersion ordering; see [21, 12, 5]. These are sometimes referred to as *measures of peakness* and *peakness ordering*, and are related to the large literature on dispersion measures which grew out of the Gini coefficient, used to measure income inequality [4] and diversity in biology, see [15], which we will discuss briefly below.

In the definitions there are typically two kinds of dispersion, those measuring some kind of mean distance, or squared distance, from a central value, such as in the usual definition of variance, and those based on the expected distance, or squared distance, between two independent copies from the same distribution, such

*pronzato@i3s.unice.fr, Laboratoire I3S, UMR 7172, UNS/CNRS, 2000 route des Lucioles, Les Algorithmes, bât. Euclide B, 06900 Sophia Antipolis, France

†H.Wynn@lse.ac.uk, London School of Economics, Houghton Street, London, WC2A 2AE, UK

‡ZhigljavskyAA@cf.ac.uk, School of Mathematics, Cardiff University, Senghennydd Road, Cardiff, CF24 4YH, UK

as the Gini coefficient. It is this second type that will concern us here and we will generalise the idea in several ways by replacing distance by volumes of simplices formed by k independent copies and by transforming the distance, both inside the expectation and outside.

The area of optimal experimental design is another which has provided a range of dispersion measures. Good designs, it is suggested, are those whose parameter estimates have low dispersion. Typically, this means that the design measure, the spread of the observation sites, *maximises* a measure of dispersion and we shall study this problem.

We think of a dispersion measure as a functional directly on the distribution. The basic functional is an integral, such as a moment. The property we shall stress for such functionals most is concavity: that a functional does not decrease under mixing of the distributions. A fundamental theorem in Bayesian learning is that we expect concave functionals to decrease through taking of observations, see Section 2.2 below.

Our central result (Section 3) is an identity for the mean squared volume of simplices of dimension k, formed by $k+1$ independent copies, in terms of the eigenvalues of the covariance matrices or equivalently in terms of sums of the determinants of k-marginal covariance matrices. Second, we note that after an appropriate (exterior) power transformation the functional becomes concave. We can thus (i) derive properties of measures that maximise this functional (Section 4.1), (ii) use this functional to measure the dispersion of parameter estimates in regression problems, and hence design optimal experiments which minimise this measure of dispersion (Section 4.2).

2 Dispersion measures

2.1 Concave and homogeneous functionals

Let $\mathscr{X}$ be a compact subset of $\mathbb{R}^d$, $\mathscr{M}$ be the set of all probability measures on the Borel subsets of $\mathscr{X}$ and $\phi : \mathscr{M} \longrightarrow \mathbb{R}^+$ be a functional defined on $\mathscr{M}$. We will be interested in the functionals $\phi(\cdot)$ that are (see Appendix for precise definitions)

(a) shift-invariant,

(b) positively homogeneous of a given degree q, and

(c) concave: $\phi[(1-\alpha)\mu_1 + \alpha\mu_2] \geq (1-\alpha)\phi(\mu_1) + \alpha\phi(\mu_2)$ for any $\alpha \in (0,1)$ and any two measures μ_1, μ_2 in $\mathscr{M}$.

For $d = 1$, a common example of a functional satisfying the above properties, with $q = 2$ in (b), is the variance

$$\sigma^2(\mu) = E^{(2)}_\mu - E^2_\mu = \frac{1}{2}\int\int (x_1 - x_2)^2 \, \mu(\mathrm{d}x_1)\, \mu(\mathrm{d}x_2)\,,$$

where $E_\mu = \mathsf{E}_\mu(x) = \int x\, \mu(\mathrm{d}x)$ and $E^{(2)}_\mu = \int x^2\, \mu(\mathrm{d}x)$. Concavity follows from linearity of $E^{(2)}_\mu$, that is, $E^{(2)}_{(1-\alpha)\mu_1+\alpha\mu_2} = (1-\alpha)E^{(2)}_{\mu_1} + \alpha E^{(2)}_{\mu_2}$, and Jensen's inequality which implies $E^2_{(1-\alpha)\mu_1+\alpha\mu_2} \leq (1-\alpha)E^2_{\mu_1} + \alpha E^2_{\mu_2}$.

Any moment of $\mu \in \mathscr{M}$ is a homogeneous functional of a suitable degree. However, the variance is the only moment which satisfies (a) and (c). Indeed, the shift-invariance implies that the moment should be central, but the variance is the only concave functional among the central moments, see Appendix. In this sense, one of the aims of this paper is a generalisation of the concept of variance.

In the general case $d \geq 1$, the double variance $2\sigma^2(\mu)$ generalises to

$$\phi(\mu) = \int\int \|x_1 - x_2\|^2 \, \mu(\mathrm{d}x_1)\, \mu(\mathrm{d}x_2) = 2 \int \|x - E_\mu\|^2 \, \mu(\mathrm{d}x) = 2\,\mathrm{trace}(V_\mu)\,, \quad (2.1)$$

where $\|\cdot\|$ is the L_2-norm in $\mathbb{R}^d$ and V_μ is the covariance matrix of μ. This functional, like the variance, satisfies the conditions (a)-(c) with $q = 2$.

The functional (2.1) is the double integral of the squared distance between two random points distributed according to the measure μ. Our main interest will be concentrated around the general class of functionals defined by

$$\phi(\mu) = \phi_{[k],\delta,\tau}(\mu) = \left(\int \dots \int \mathscr{V}_k^\delta(x_1, \dots, x_{k+1})\, \mu(\mathrm{d}x_1) \dots \mu(\mathrm{d}x_{k+1}) \right)^\tau , \ k \geq 2 \quad (2.2)$$

for some δ and τ in $\mathbb{R}^+$, where $\mathscr{V}_k(x_1, \dots, x_{k+1})$ is the volume of the k-dimensional simplex (its area when $k = 2$) formed by the $k+1$ vertices $x_1, \dots, x_{k+1}$ in $\mathbb{R}^d$, with $k = d$ as a special case. Property (a) for the functionals (2.2) is then a straightforward consequence of the shift-invariance of $\mathscr{V}_k$, and positive homogeneity of degree $q = k\,\delta\tau$ directly follows from the positive homogeneity of $\mathscr{V}_k$ with degree k. Concavity will be proved to hold for $\delta = 2$ and $\tau \leq 1/k$ in Section 3. There, we show that this case can be considered as a natural extension of (2.1) (which corresponds to $k = 1$), with $\phi_{[k],2,\tau}(\mu)$ being expressed as a function of V_μ, the covariance matrix of μ. The concavity for $k = \tau = 1$ and all $0 < \delta \leq 2$, follows from the Schoenberg theory, which will be discussed briefly below. The functionals (2.2) with $\delta = 2$ and $\tau > 0$, $1 \leq k \leq d$, can be used to define a family of criteria for optimal experimental design, concave for $\tau \leq 1/k$, for which an equivalence theorem can be formulated.

2.2 Quadratic entropy and learning

In a series of papers [15, 16, 17] C.R. Rao and co-workers have introduced a *quadratic entropy* which is a generalised version of the $k = 2$ functional of this section but with a general kernel $K(x_1, x_2)$ in $\mathbb{R}^d$:

$$Q_R = \int\int K(x_1, x_2) \mu(\mathrm{d}x_1) \mu(\mathrm{d}x_2)\,. \quad (2.3)$$

For the discrete version

$$Q_R = \sum_{i=1}^N \sum_{j=1}^N K(x_i, x_j)\, p_i\, p_j,$$

Rao and co-workers developed a version of the Analysis of Variance (ANOVA), which they called Anaysis of Quadratic Entropy (ANOQE). The Gini coefficient, also used in the continuous and discrete form is a special case with $d = 1$ and $K(x_1, x_2) = |x_1 - x_2|$.

As pointed in [17, Chap. 3], a necessary and sufficient condition for the functional Q_R to be concave is

$$\int\int K(x_1, x_2) \nu(\mathrm{d}x_1) \nu(\mathrm{d}x_2) \leq 0 \quad (2.4)$$

for all measures ν with $\int \nu(\mathrm{d}x) = 0$. The discrete version of this is

$$\sum_{i=1}^N \sum_{j=1}^N K(x_i, x_j)\, q_i\, q_j \leq 0$$

for any choice of real numbers $q_1, \ldots, q_N$ such that $\sum_{i=1}^N q_i = 0$. Schoenberg [19] solves the general problem of finding for what class of functions $B(\cdot)$ on $\|x_1 - x_2\|^2$ does the kernel $K(x_1, x_2) = B\left(\|x_1 - x_2\|^2\right)$ satisfy (2.4). The solution is that B must be a so-called Bernstein function, see [18]. We do not develop these ideas here, but note that $B(\lambda) = \lambda^\alpha$ is a Bernstein function for all $0 < \alpha \leq 1$. This is the reason that, above, we can claim concavity for $k = 1$ and all $0 < \delta \leq 2$ in (2.2).

Hainy *et al* [6] discuss the link to embedding and review some basic results related to Bayesian learning. One asks what is the class of functionals ψ on a distribution $\mu(\theta)$ of a parameter in the Bayesian statistical learning such that for all $\mu(\theta)$ and all sampling distributions $\pi(x|\theta)$ one expects to learn, in the preposterior sense: $\psi(\mu(\theta)) \leq \mathsf{E}_\nu \psi(\pi(\theta|X))$, with $X \sim \nu$. The condition is that ψ is convex, a result which has a history but is usually attributed to De Groot [2]. This learning is enough to justify calling such a functional a generalised information functional, or a general learning functional. Shannon information falls in this class, and earlier versions of the result were for Shannon information. It follows that wherever, in this paper, we have a concave functional then its *negative* is a learning functional.

3 Functionals based on squared volume

In the rest of the paper we focus our attention on the functional

$$\mu \in \mathscr{M} \longrightarrow \psi_k(\mu) = \phi_{[k],2,1}(\mu) = \mathsf{E}_\mu\{\mathscr{V}_k^2(x_1, \ldots, x_{k+1})\}\,,$$

which corresponds to the mean squared volume of simplices of dimension k formed by $k+1$ independent samples from μ. For instance,

$$\psi_2(\mu) = \int\int\int \mathscr{V}_2^2(x_1, x_2, x_3)\, \mu(\mathrm{d}x_1)\, \mu(\mathrm{d}x_2)\, \mu(\mathrm{d}x_3)\,, \tag{3.1}$$

with $\mathscr{V}_2(x_1, x_2, x_3)$ the area of the triangle formed by the three points with coordinates x_1, x_2 and x_3 in $\mathbb{R}^d$, $d \geq 2$. Functionals $\phi_{[k],\delta,\tau}(\mu)$ for $\delta \neq 2$ will be considered elsewhere, including the case of negative δ and τ in connection with space-filling design for computer experiments.

Theorem 3.1 of Section 3.1 indicates how $\psi_k(\mu)$ can be expressed as a function of V_μ, the covariance matrix of μ, and shows that $\phi_{[k],2,1/k}(\cdot)$ satisfies properties (a), (b) and (c) of Section 2.1. The special case of $k = d$ was known to Wilks [25, 26] in his introduction of generalised variance, see also [24]. The connection with U-statistics is exploited in Section 3.2, where an unbiased minimum-variance estimator of $\psi_k(\mu)$ based on a sample $x_1, \ldots, x_n$ is expressed in terms of the empirical covariance matrix of the sample.

3.1 Expected squared k-simplex volume

Theorem 3.1. *Let the x_i be i.i.d. with the probability measure $\mu \in \mathscr{M}$. Then, for any $k \in \{1, \ldots, d\}$, we have*

$$\psi_k(\mu) = \frac{k+1}{k!} \sum_{i_1 < i_2 < \cdots < i_k} \det[\{V_\mu\}_{(i_1,\ldots,i_k)\times(i_1,\ldots,i_k)}] \tag{3.2}$$

$$= \frac{k+1}{k!} \sum_{i_1 < i_2 < \cdots < i_k} \lambda_{i_1}[V_\mu] \times \cdots \times \lambda_{i_k}[V_\mu]\,, \tag{3.3}$$

where $\lambda_i[V_\mu]$ is the i-th eigenvalue of the covariance matrix V_μ and all i_j belong to $\{1, \ldots, d\}$. Moreover, the functional $\psi_k^{1/k}(\cdot)$ is shift-invariant, homogeneous of degree 2 and concave on $\mathscr{M}$.

The proof uses the following two lemmas.

Lemma 3.1. *Let the $k+1$ vectors $x_1, \dots, x_{k+1}$ of $\mathbb{R}^k$ be i.i.d. with the probability measure μ, $k \geq 2$. For $i = 1, \dots, k+1$, denote $z_i = (x_i^\top \ 1)^\top$. Then*

$$\mathsf{E}_\mu \left\{ \det \left[\sum_{i=1}^{k+1} z_i z_i^\top \right] \right\} = (k+1)! \, \det[V_\mu] \, .$$

Proof. We have

$$\mathsf{E}_\mu \left\{ \det \left[\sum_{i=1}^{k+1} z_i z_i^\top \right] \right\} = (k+1)! \, \det \begin{bmatrix} \mathsf{E}_\mu(x_1 x_1^\top) & E_\mu \\ E_\mu^\top & 1 \end{bmatrix} = (k+1)! \det[V_\mu] \, ,$$

see for instance [13, Theorem 1]. □

Lemma 3.2. *The matrix functional $\mu \mapsto V_\mu$ is Loewner-concave on $\mathscr{M}$, in the sense that, for any μ_1, μ_2 in $\mathscr{M}$ and any $\alpha \in (0,1)$,*

$$V_{(1-\alpha)\mu_1 + \alpha\mu_2} \succeq (1-\alpha) V_{\mu_1} + \alpha V_{\mu_2}, \tag{3.4}$$

where $A \succeq B$ means that $A - B$ is nonnegative definite.

Proof. Take any vector z of the same dimension as x. Then $z^\top V_\mu z = \mathsf{var}_\mu(z^\top x)$, which is a concave functional of μ, see Section 2.1. This implies that $z^\top V_{(1-\alpha)\mu_1+\alpha\mu_2} z = \mathsf{var}_{(1-\alpha)\mu_1+\alpha\mu_2}(z^\top x) \geq (1-\alpha)\mathsf{var}_{\mu_1}(z^\top x) + \alpha \mathsf{var}_{\mu_2}(z^\top x) = (1-\alpha) z^\top V_{\mu_1} z + \alpha z^\top V_{\mu_2} z$, for any μ_1, μ_2 in $\mathscr{M}$ and any $\alpha \in (0,1)$ (see Section 2.1 for the concavity of var_μ). Since z is arbitrary, this implies (3.4). □

Proof of Theorem 3.1. When $k = 1$, the results follow from $\psi_1(\mu) = 2\,\mathrm{trace}(V_\mu)$, see (2.1). Using Binet-Cauchy formula, see, e.g., [3, vol. 1, p. 9], we obtain

$$\begin{aligned}
\mathscr{V}_k^2(x_1, \dots, x_{k+1}) = & \\
& \frac{1}{(k!)^2} \det \left(\begin{bmatrix} (x_2 - x_1)^\top \\ (x_3 - x_1)^\top \\ \vdots \\ (x_{k+1} - x_1)^\top \end{bmatrix} \begin{bmatrix} (x_2 - x_1) & (x_3 - x_1) & \cdots (x_{k+1} - x_1) \end{bmatrix} \right) \\
= & \frac{1}{(k!)^2} \sum_{i_1 < i_2 < \cdots < i_k} \det^2 \begin{bmatrix} \{x_2 - x_1\}_{i_1} & \cdots & \{x_{k+1} - x_1\}_{i_1} \\ \vdots & \vdots & \vdots \\ \{x_2 - x_1\}_{i_k} & \cdots & \{x_{k+1} - x_1\}_{i_k} \end{bmatrix} \\
= & \frac{1}{(k!)^2} \sum_{i_1 < i_2 < \cdots < i_k} \det^2 \begin{bmatrix} \{x_1\}_{i_1} & \cdots & \{x_{k+1}\}_{i_1} \\ \vdots & \vdots & \vdots \\ \{x_1\}_{i_k} & \cdots & \{x_{k+1}\}_{i_k} \\ 1 & \cdots & 1 \end{bmatrix},
\end{aligned}$$

where $\{x\}_i$ denotes the i-th component of vector x. Also, for all $i_1 < i_2 < \cdots < i_k$,

$$\det^2 \begin{bmatrix} \{x_1\}_{i_1} & \cdots & \{x_{k+1}\}_{i_1} \\ \vdots & \vdots & \vdots \\ \{x_1\}_{i_k} & \cdots & \{x_{k+1}\}_{i_k} \\ 1 & \cdots & 1 \end{bmatrix} = \det \left(\sum_{j=1}^{k+1} z_j z_j^\top \right)$$

where we have denoted by z_j the $k+1$-dimensional vector with components $\{x_j\}_{i_\ell}, \ell = 1, \dots, k$, and 1. When the x_i are i.i.d. with the probability measure μ, using Lemma 3.1 we obtain (3.2), (3.3). Therefore

$$\psi_k(\mu) = \Psi_k[V_\mu] = \frac{k+1}{k!} \, \mathcal{E}_k\{\lambda_1[V_\mu], \dots, \lambda_d[V_\mu]\} \, , \tag{3.5}$$

with $\mathcal{E}_k\{\lambda_1[V_\mu], \ldots, \lambda_d[V_\mu]\}$ the elementary symmetric function of degree k of the d eigenvalues of V_μ, see, e.g., [11, p. 10]. Note that

$$\mathcal{E}_k[V_\mu] = \mathcal{E}_k\{\lambda_1[V_\mu], \ldots, \lambda_d[V_\mu]\} = (-1)^k a_{d-k}\,,$$

with a_{d-k} the coefficient of the monomial of degree $d-k$ of the characteristic polynomial of V_μ; see, e.g., [11, p. 21]. We have in particular $\mathcal{E}_1[V_\mu] = \text{trace}[V_\mu]$ and $\mathcal{E}_d[V\mu)] = \det[V_\mu]$. The shift-invariance and homogeneity of degree 2 of $\psi_k^{1/k}(\cdot)$ follow from the shift-invariance and positive homogeneity of $\mathscr{V}_k$ with degree k. Concavity of $\Psi_k^{1/k}(\cdot)$ follows from [11, p. 116] (take $p = k$ in eq. (10), with $\mathcal{E}_0 = 1$). From [9], the $\Psi_k^{1/k}(\cdot)$ are also Loewner-increasing, so that from Lemma 3.2, for any μ_1, μ_2 in $\mathscr{M}$ and any $\alpha \in (0,1)$,

$$\begin{aligned}
\psi_k^{1/k}[(1-\alpha)\mu_1 + \alpha\mu_2] &= \Psi_k^{1/k}\{V_{(1-\alpha)\mu_1+\alpha\mu_2}\} \\
&\geq \Psi_k^{1/k}[(1-\alpha)V_{\mu_1} + \alpha V_{\mu_2}] \\
&\geq (1-\alpha)\Psi_k^{1/k}[V_{\mu_1}] + \alpha\Psi_k^{1/k}[V_{\mu_2}] \\
&= (1-\alpha)\psi_k^{1/k}(\mu_1) + \alpha\psi_k^{1/k}(\mu_2)\,.
\end{aligned}$$

□

The functionals $\mu \longrightarrow \phi_{[k],2,\tau}(\mu) = \psi_k^\tau(\mu)$ are thus concave for $0 < \tau \leq 1/k$, with $\tau = 1/k$ yielding positive homogeneity of degree 2. The functional $\psi_1(\mu)$ is a quadratic entropy Q_R, see (2.3), $\psi_d(\mu)$ is proportional to Wilks generalised variance, and $\psi_2^{1/2}(\mu)$, see (3.1), and $\psi_3^{1/3}(\mu)$ can be respectively considered as particular versions of cubic and quartic entropies.

From the well-known expression of the coefficients of the characteristic polynomial of a matrix V, we have

$$\begin{aligned}
\Psi_k(V) &= \frac{k+1}{k!}\mathcal{E}_k(V) \qquad (3.6)\\
&= \frac{k+1}{(k!)^2}\det\begin{bmatrix}
\text{trace}(V) & k-1 & 0 & \cdots \\
\text{trace}(V^2) & \text{trace}(V) & k-2 & \cdots \\
\cdots & \cdots & \cdots & \cdots \\
\text{trace}(V^{k-1}) & \text{trace}(V^{k-2}) & \cdots & 1 \\
\text{trace}(V^k) & \text{trace}(V^{k-1}) & \cdots & \text{trace}(V)
\end{bmatrix},
\end{aligned}$$

see, e.g., [10, p. 28], and the $\mathcal{E}_k(V)$ satisfy the recurrence relations (Newton identities):

$$\mathcal{E}_k(V) = \frac{1}{k}\sum_{i=1}^{k}(-1)^{i-1}\,\mathcal{E}_{k-i}(V)\,\mathcal{E}_1(V^i)\,, \qquad (3.7)$$

see, e.g., [3, Vol. 1, p. 88] and [9]. Particular forms of $\psi_k(\cdot)$ are

$$\begin{aligned}
k=1: &\quad \psi_1(\mu) = 2\,\text{trace}(V_\mu)\,,\\
k=2: &\quad \psi_2(\mu) = \frac{3}{4}\,[\text{trace}^2(V_\mu) - \text{trace}(V_\mu^2)]\,,\\
k=3: &\quad \psi_3(\mu) = \frac{1}{9}\,[\text{trace}^3(V_\mu) - 3\,\text{trace}(V_\mu^2)\text{trace}(V_\mu) + 2\,\text{trace}(V_\mu^3)]\,,\\
k=d: &\quad \psi_d(\mu) = \frac{d+1}{d!}\,\det(V_\mu)\,.
\end{aligned}$$

3.2 The empirical version and unbiased estimates

Let $x_1, \dots, x_n$ be a sample of n vectors of $\mathbb{R}^d$, i.i.d. with the measure μ. This sample can be used to obtain an empirical estimate $(\widehat{\psi}_1)_n$ of $\psi_k(\mu)$, through the consideration of the $\binom{n}{k+1}$ k-dimensional simplices that can be constructed with the x_i. Below we show how a much simpler (and still unbiased) estimation of $\psi_k(\mu)$ can be obtained through the empirical variance-covariance matrix of the sample. See also [25, 26].

Denote

$$\begin{aligned} \widehat{x}_n &= \frac{1}{n}\sum_{i=1}^n x_i\,, \\ \widehat{V}_n &= \frac{1}{n-1}\sum_{i=1}^n (x_i-\widehat{x}_n)(x_i-\widehat{x}_n)^\top = \frac{1}{n(n-1)}\sum_{i<j}(x_i-x_j)(x_i-x_j)^\top\,, \end{aligned}$$

respectively the empirical mean and variance-covariance matrix of x_1. Note that both are unbiased. We thus have

$$(\widehat{\psi}_1)_n = \frac{2}{n(n-1)}\sum_{i<j}\|x_i-x_j\|^2 = 2\,\mathrm{trace}[\widehat{V}_n] = \Psi_1(\widehat{V}_n) = \psi_1(\mu_n)\,,$$

with μ_n the empirical measure of the sample, and the estimator $(\widehat{\psi}_1)_n$ is unbiased. More generally, for $k \geq 1$ we have the following.

Theorem 3.2. *For $x_1, \dots, x_n$ a sample of n vectors of $\mathbb{R}^d$, i.i.d. with the measure μ, and for any $k \in \{1, \dots, d\}$, the quantity*

$$(\widehat{\psi}_k)_n = \frac{(n-k-1)!(n-1)^k}{(n-1)!}\,\Psi_k(\widehat{V}_n) = \frac{(n-k-1)!(n-1)^k}{(n-1)!}\,\psi_k(\mu_n)\,, \tag{3.8}$$

forms an unbiased estimator of $\psi_k(\mu)$ and has minimum variance among all unbiased estimators.

Proof. Denote

$$(\widehat{\psi}_k)_n = \binom{n}{k+1}^{-1} \sum_{j_1<j_2<\cdots<j_{k+1}} \mathscr{V}_k^2(x_{j_1}, \dots, x_{j_{k+1}})\,. \tag{3.9}$$

It forms a U-statistics for the estimation of $\psi_k(\mu)$ and is thus unbiased and has minimum variance, see, e.g., [20, Chap. 5]. We only need to show that it can be written as (3.8).

We can write

$$\begin{aligned} (\widehat{\psi}_k)_n &= \binom{n}{k+1}^{-1} \\ &\quad \times \sum_{j_1<j_2<\cdots<j_{k+1}} \frac{1}{(k!)^2} \sum_{i_1<i_2<\cdots<i_k} \det{}^2 \begin{bmatrix} \{x_{j_1}\}_{i_1} & \cdots & \{x_{j_{k+1}}\}_{i_1} \\ \vdots & \vdots & \vdots \\ \{x_{j_1}\}_{i_k} & \cdots & \{x_{j_{k+1}}\}_{i_k} \\ 1 & \cdots & 1 \end{bmatrix}, \\ &= \binom{n}{k+1}^{-1} \frac{1}{(k!)^2} \sum_{i_1<i_2<\cdots<i_k} \det\left(\sum_{j=1}^n \{z_j\}_{i_1,\dots,i_k}\{z_j\}_{i_1,\dots,i_k}^\top\right), \end{aligned}$$

where we have used Binet-Cauchy formula and where $\{z_j\}_{i_1,\ldots,i_k}$ denotes the $k+1$ dimensional vector with components $\{x_j\}_{i_\ell}$, $\ell = 1,\ldots,k$, and 1. This gives

$$\begin{aligned}(\widehat{\psi}_k)_n &= \binom{n}{k+1}^{-1} \frac{n^{k+1}}{(k!)^2} \sum_{i_1<i_2<\cdots<i_k} \det\left(\frac{1}{n}\sum_{j=1}^n \{z_j\}_{i_1,\ldots,i_k}\{z_j\}_{i_1,\ldots,i_k}^\top\right), \\ &= \binom{n}{k+1}^{-1} \frac{n^{k+1}}{(k!)^2} \\ &\times \sum_{i_1<i_2<\cdots<i_k} \det\begin{bmatrix} (1/n)\{\sum_{j=1}^n x_j x_j^\top\}_{(i_1,\ldots,i_k)\times(i_1,\ldots,i_k)} & \{\widehat{x}_n\}_{i_1,\ldots,i_k} \\ \{\widehat{x}_n\}_{i_1,\ldots,i_k}^\top & 1 \end{bmatrix}, \\ &= \binom{n}{k+1}^{-1} \frac{n^{k+1}}{(k!)^2} \sum_{i_1<i_2<\cdots<i_k} \det\left[\frac{n-1}{n}\{\widehat{V}_n\}_{(i_1,\ldots,i_k)\times(i_1,\ldots,i_k)}\right],\end{aligned}$$

and thus (3.8). □

Using the notation of Theorem 3.1, since $\mathcal{E}_k(V) = (-1)^k a_{d-k}(V)$, with $a_{d-k}(V)$ the coefficient of the monomial of degree $d-k$ of the characteristic polynomial of V, for a nonsingular V we obtain

$$\mathcal{E}_k(V) = \det(V)\,\mathcal{E}_{d-k}(V^{-1}), \tag{3.10}$$

see also [9, Eq. 4.2]. Therefore, we also have

$$(\widehat{\psi}_{d-k})_n = \frac{(n-d+k-1)!(n-1)^{d-k}}{(n-1)!}\,\frac{(d-k+1)k!}{(k+1)(d-k)!}\det(\widehat{V}_n)\,\Psi_k(\widehat{V}_n^{-1}), \tag{3.11}$$

which forms an unbiased and minimum-variance estimator of $\psi_{d-k}(\mu)$. Note that the estimation of $\psi_k(\mu)$ is much simpler through (3.8) or (3.11) than using the direct construction (3.9).

One may notice that $\psi_1(\mu_n)$ is clearly unbiased due to the linearity of $\Psi_1(\cdot)$, but it is remarkable that $\psi_k(\mu_n)$ becomes unbiased after a suitable scaling, see (3.8). Since $\Psi_k(\cdot)$ is highly nonlinear for $k > 1$, this property would not hold if $\widehat{V}_n$ were replaced by another unbiased estimator of V_μ.

The value of $(\widehat{\psi}_k)_n$ only depend on $\widehat{V}_n$, with $\mathsf{E}\{(\widehat{\psi}_k)_n\} = \psi_k(V_\mu)$, but its variance depends on the distribution itself. From [20, Lemma A, p. 183], the variance of $(\widehat{\psi}_k)_n$ satisfies

$$\mathsf{var}[(\widehat{\psi}_k)_n] = \frac{(k+1)^2}{n}\,\omega + O(n^{-2}),$$

where $\omega = \mathsf{var}[h(x)]$, with $h(x) = \mathsf{E}\{\mathscr{V}_k^2(x_1, x_2, \ldots, x_{k+1})|x_1 = x\}$. Obviously, $\mathsf{E}[h(x)] = \psi_k(\mu)$ and calculations similar to those in the proof of Theorem 3.1 give

$$\begin{aligned}\omega = \frac{1}{(k!)^2}\sum_{I,J}\det[\{V_\mu\}_{I\times I}]\,\det[\{V_\mu\}_{J\times J}] \\ \times\left[\mathsf{E}\left\{(E_\mu - x)_I^\top\{V_\mu\}_{I\times I}^{-1}(E_\mu - x)_I(E_\mu - x)_J^\top\{V_\mu\}_{J\times J}^{-1}(E_\mu - x)_J\right\} - k^2\right],\end{aligned} \tag{3.12}$$

where I and J respectively denote two sets of indices $i_1 < i_2 < \cdots i_k$ and $j_1 < j_2 < \cdots j_k$ in $\{1,\ldots,k+1\}$, the summation being over all possible such sets. Simplifications occur in some particular cases. For instance, when μ is a normal measure, then

$$\begin{aligned}\omega &= \frac{2}{(k!)^2}\sum_{I,J}\det[\{V_\mu\}_{I\times I}]\,\det[\{V_\mu\}_{J\times J}] \\ &\quad\times \mathrm{trace}\left[\{V_\mu\}_{J\times J}^{-1}\{V_\mu\}_{J\times I}\{V_\mu\}_{I\times I}^{-1}\{V_\mu\}_{I\times J}\right].\end{aligned}$$

If, moreover, V_μ is the diagonal matrix $\mathrm{diag}\{\lambda_1, \ldots, \lambda_d\}$, then

$$\omega = \frac{2}{(k!)^2} \sum_{I,J} \beta(I,J) \prod_I \lambda_i \prod_J \lambda_j \,,$$

with $\beta(I,J)$ denoting the number of coincident indices between I and J (i.e., the size of $I \cap J$). When μ is such that the components of x are i.d.d. with variance σ^2, then $V_\mu = \sigma^2 I_d$, with I_d the d-dimensional identity matrix, and

$$\mathsf{E}\left\{(E_\mu - x)_I^\top \{V_\mu\}_{I\times I}^{-1} (E_\mu - x)_I (E_\mu - x)_J^\top \{V_\mu\}_{J\times J}^{-1} (E_\mu - x)_J\right\} = \mathsf{E}\left\{\left(\sum_{i\in I} z_i^2\right)\left(\sum_{j\in J} z_j^2\right)\right\},$$

where the $z_i = \{x - E_\mu\}_i/\sigma$ are i.i.d. with mean 0 and variance 1. We then obtain

$$\omega = \frac{\sigma^{4k}}{(k!)^2} \left(\mathsf{E}\{z_i^4\} - 1\right) \beta_{d,k} \,,$$

where

$$\begin{aligned} \beta_{d,k} = \sum_{I,J} \beta(I,J) &= \sum_{i=1}^{k} i \binom{d}{i} \binom{d-i}{k-i} \binom{d-i-(k-i)}{k-i} \\ &= \frac{(d-k+1)^2}{d} \binom{d}{k-1}^2 . \end{aligned}$$

Example 1 We generate 1,000 independent samples of n points for different measures μ. Figure 1 presents a box-plot of the ratios $(\widehat{\psi}_k)_n/\psi_k(\mu)$ for various values of k and $n = 100$ (left), $n = 1,000$ (right), when $\mu = \mu_1$ uniform in $[0,1]^{10}$. Figure 2 presents the same information when $\mu = \mu_2$ which corresponds to the normal distribution $\mathcal{N}(0, I_{10}/12)$ in $\mathbb{R}^{10}$. Note that $V_{\mu_1} = V_{\mu_2}$ but the dispersions are different in the two figures. Table 1 gives $(10^3 \times)$ the values of $\widehat{\mathsf{E}}\{(\widehat{\psi}_k)_n\}/\psi_k(\mu) - 1$ for $\mu = \mu_1$, μ_2 and the same series of values for k, with $\widehat{\mathsf{E}}\{(\widehat{\psi}_k)_n\}$ denoting the empirical mean over the 1,000 independent repetitions.

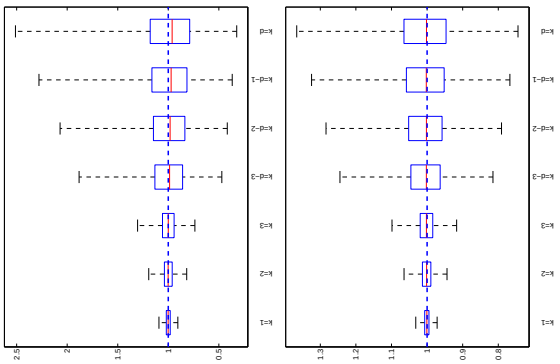

Figure 1: Box-plot of $(\widehat{\psi}_k)_n/\psi_k(\mu)$ for different values of k: μ is uniform in $[0,1]^{10}$, $n = 100$ (Left) and $n = 1,000$ (Right) — 1,000 repetitions; minimum, median and maximum values are indicated, together with 25% and 75% quantiles.

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].

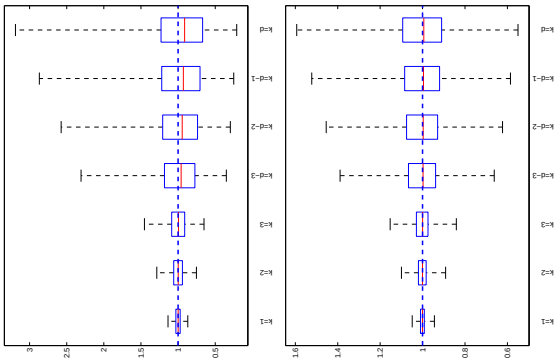

Figure 2: Same as in Figure 1 but for μ normal $\mathscr{N}(0, I_{10}/12)$.

Table 1: $\delta(n,k) = 10^3 \times [\widehat{\mathsf{E}}\{(\widehat{\psi}_k)_n\}/\psi_k(\mu) - 1]$ for different values of n and k; μ_1 is uniform in $[0,1]^{10}$, μ_2 is normal $\mathscr{N}(0, I_{10}/12)$.

	k	1	2	3	$d-3$	$d-2$	$d-1$	d
μ_1	$n = 100$	0.9	1.9	2.9	7.7	9.0	10.3	11.6
	$n = 1,000$	0.7	1.4	2.1	4.9	5.6	6.3	7.0
μ_2	$n = 100$	1.2	2.3	3.2	4.3	3.9	3.0	1.9
	$n = 1,000$	0.5	1.0	1.5	3.6	4.1	4.7	5.2

In particular, $(\widehat{\psi}_k)_n$ is asymptotically normal $\mathscr{N}(\psi_k(\mu), (k+1)^2\omega/n)$, with ω given by (3.12). This is illustrated in Figure 3-left below for μ uniform in $[0,1]^{10}$, with $n = 1,000$ and $k = 3$. The distribution is already reasonably close to normality for small values of n, see Figure 3-right for which $n = 20$.

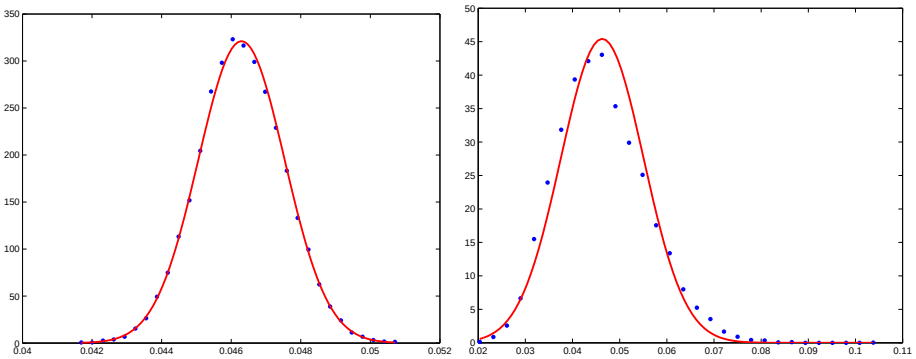

Figure 3: Dots: empirical distribution of $(\widehat{\psi}_k)_n$ (histogram for 10,000 independent repetitions); solid line: asymptotic normal distribution $\mathscr{N}(\psi_k(\mu), (k+1)^2\omega/n)$; μ is uniform in $[0,1]^{10}$ and $k = 3$; left: $n = 1,000$; right: $n = 20$.

4 Maximum-entropy measures and optimal designs

In this section we consider two types of optimisation problems on $\mathscr{M}$ related to the functions $\Psi_k(\cdot)$ introduced in Theorem 3.1. First, in Section 4.1, we are interested in the characterisation and construction of maximum-entropy measures; that is, measures $\mu_k^* \in \mathscr{M}$ which maximize $\psi_k(\mu) = \Psi_k(V_\mu)$. The existence of an optimal measure follows from the compactness of $\mathscr{X}$ and continuity of $\mathscr{V}_k(x_1, \ldots, x_{k+1})$ in

each x_i, see [1, Th. 1]; the concavity and differentiability of the functional $\psi_k^{1/k}(\cdot)$ allow us to derive a necessary and sufficient condition for optimality.

In Section 4.2 we consider the problem of optimal design of experiments, where the covariance matrix V is the inverse of the information matrix $M(\xi)$ for some regression model.

4.1 Maximum-entropy measures

4.1.1 Necessary and sufficient condition

Since the functionals $\psi_k^{1/k}(\cdot)$ are concave and differentiable, for all $k = 1, \ldots, d$, we can easily derive a necessary and sufficient condition for a probability measure μ_k^* on $\mathscr{X}$ to maximise $\psi_k(\mu)$, in the spirit of the celebrated Equivalence Theorem of Kiefer and Wolfowitz [8].

Denote by $\nabla_{\Psi_k}[V]$ the gradient of $\Psi_k(\cdot)$ at matrix V (a matrix of the same size as V) and by $F_{\psi_k}(\mu;\nu)$ the directional derivative of $\psi_k(\cdot)$ at μ in the direction ν;

$$F_{\psi_k}(\mu;\nu) = \lim_{\alpha\to 0^+} \frac{\psi_k[(1-\alpha)\mu + \alpha\nu] - \psi_k(\mu)}{\alpha}\,.$$

From the expression (3.6) of $\Psi_k(V)$, we have

$$\nabla_{\Psi_k}[V] = \frac{k+1}{k!}\nabla_{\mathcal{E}_k}[V])\,,$$

where $\nabla_{\mathcal{E}_k}[V]$ denotes the gradient of $\mathcal{E}_k(\cdot)$ at V, which, using (3.7), can be shown by induction to satisfy

$$\nabla_{\mathcal{E}_k}[V] = \sum_{i=0}^{k-1}(-1)^i\,\mathcal{E}_{k-i-1}(V)\,V^i\,, \tag{4.1}$$

see [9]. We thus obtain in particular

$$\begin{aligned}
k=1: &\qquad \nabla_{\Psi_1}[V] = 2\,I_d\,,\\
k=2: &\qquad \nabla_{\Psi_2}[V] = \frac{3}{2}\,[\mathrm{trace}(V)\,I_d - V]\,,\\
k=3: &\qquad \nabla_{\Psi_3}[V] = \frac{1}{3}\,[\mathrm{trace}^2(V) - \mathrm{trace}(V^2)]\,I_d - \frac{2}{3}\,\mathrm{trace}(V)\,V + \frac{2}{3}\,V^2\,,\\
k=d: &\qquad \nabla_{\Psi_d}[V] = \frac{d+1}{d!}\,\det(V)\,V^{-1}\,.
\end{aligned}$$

Using the differentiability of $\Psi_k(\cdot)$, direct calculation gives

$$F_{\psi_k}(\mu;\nu) = \mathrm{trace}\left\{\nabla_{\Psi_k}[V_\mu]\,\frac{\mathrm{d}V_{(1-\alpha)\mu+\alpha\nu}}{\mathrm{d}\alpha}\bigg|_{\alpha=0}\right\}\,,$$

with

$$\frac{\mathrm{d}V_{(1-\alpha)\mu+\alpha\nu}}{\mathrm{d}\alpha}\bigg|_{\alpha=0} = \int [xx^\top - (E_\mu x^\top + xE_\mu^\top)]\,\nu(\mathrm{d}x) - \int xx^\top\,\mu(\mathrm{d}x) + 2E_\mu E_\mu^\top\,. \tag{4.2}$$

Notice that $\mathrm{d}V_{(1-\alpha)\mu+\alpha\nu}/\mathrm{d}\alpha\big|_{\alpha=0}$ is linear in ν.

Then, from the concavity of $\psi_k^{1/k}(\cdot)$, μ_k^* maximises $\psi_k(\mu)$ with respect to $\mu \in \mathscr{M}$ if and only if $\psi_k(\mu_k^*) > 0$ and $F_{\psi_k}(\mu_k^*;\nu) \le 0$ for all $\nu \in \mathscr{M}$, that is

$$\mathrm{trace}\left\{\nabla_{\Psi_k}[V_{\mu_k^*}]\,\frac{\mathrm{d}V_{(1-\alpha)\mu_k^*+\alpha\nu}}{\mathrm{d}\alpha}\bigg|_{\alpha=0}\right\} \le 0\,,\ \forall\nu\in\mathscr{M}\,. \tag{4.3}$$

We obtain the following.

Theorem 4.1. *The probability measure μ_k^* such that $\psi_k(\mu_k^*) > 0$ is ψ_k-optimal, that is, maximises $\psi_k(\mu)$ with respect to $\mu \in \mathscr{M}$, $k \in \{1, \ldots, d\}$, if and only if*

$$\max_{x \in \mathscr{X}} (x - E_{\mu_k^*})^\top \frac{\nabla_{\Psi_k}[V_{\mu_k^*}]}{\Psi_k(V_{\mu_k^*})} (x - E_{\mu_k^*}) \leq k \,. \tag{4.4}$$

Moreover,

$$(x - E_{\mu_k^*})^\top \frac{\nabla_{\Psi_k}[V_{\mu_k^*}]}{\Psi_k(V_{\mu_k^*})} (x - E_{\mu_k^*}) = k \tag{4.5}$$

for all x in the support of μ_k^*.

Proof. First note that the Newton equations (3.7) and the recurrence (4.1) for $\nabla_{\mathcal{E}_k}[\cdot]$ imply that $\operatorname{trace}(V \nabla_{\Psi_k}[V]) = k \Psi_k(V)$ for all $k = 1, \ldots, d$.

The condition (4.4) is sufficient. Indeed, suppose that μ_k^* such that $\psi_k(\mu_k^*) > 0$ satisfies (4.4). We obtain

$$\int (x - E_{\mu_k^*})^\top \nabla_{\Psi_k}[V_{\mu_k^*}] (x - E_{\mu_k^*}) \, \nu(\mathrm{d}x) \leq \operatorname{trace} \left\{ V_{\mu_k^*} \nabla_{\Psi_k}[V_{\mu_k^*}] \right\}$$

for any $\nu \in \mathscr{M}$, which gives (4.3) when we use (4.2). The condition is also necessary since (4.3) must be true in particular for δ_x, the delta measure at any $x \in \mathscr{X}$, which gives (4.4). The property (4.5) on the support of μ_k^* follows from the observation that $\int (x - E_{\mu_k^*})^\top \nabla_{\Psi_k}[V_{\mu_k^*}] (x - E_{\mu_k^*}) \, \mu_k^*(\mathrm{d}x) = \operatorname{trace} \left\{ V_{\mu_k^*} \nabla_{\Psi_k}[V_{\mu_k^*}] \right\}$. □

Note that for $k < d$, the covariance matrix $V_{\mu_k^*}$ of a ψ_k-optimal measure μ_k^* is not necessarily unique and may be singular; see, e.g., Example 1 below. Also, $\psi_k(\mu) > 0$ implies that $\psi_{k-1}(\mu) > 0$, $k = 2, \ldots, d$.

Remark 4.1. *As a natural extension of the concept of potential in case of order-two interactions ($k = 1$), we call $P_{k,\mu}(x) = \psi_k(\mu, \ldots, \mu, \delta_x)$ the potential of μ at x, where*

$$\psi_k(\mu_1, \ldots, \mu_{k+1}) = \int \ldots \int \mathscr{V}_k^2(x_1, \ldots, x_{k+1}) \, \mu_1(\mathrm{d}x_1) \ldots \mu_{k+1}(\mathrm{d}x_{k+1}) \,.$$

This yields $F_{\psi_k}(\mu; \nu) = (k+1) \, [\psi_k(\mu, \ldots, \mu, \nu) - \psi_k(\mu)]$, where μ appears k times in $\psi_k(\mu, \ldots, \mu, \nu)$. Therefore, Theorem 4.4 states that μ_k^* is ψ_k-optimal if and only if $\psi_k(\mu_k^*, \ldots, \mu_k^*, \nu) \leq \psi_k(\mu_k^*)$ for any $\nu \in \mathscr{M}$, or equivalently $P_{k,\mu_k^*}(x) \leq \psi_k(\mu_k^*)$ for all $x \in \mathscr{X}$.

Remark 4.2. *Consider Kiefer's Φ_p-class of orthogonally invariant criteria and their associated functional $\varphi_p(\cdot)$, defined by*

$$\varphi_p(\mu) = \Phi_p(V_\mu) = \begin{cases} \lambda_{\max}(V_\mu) & \text{for } p = \infty \,, \\ \{\frac{1}{d} \operatorname{trace}(V_\mu^p)\}^{1/d} & \text{for } p \neq 0, \pm\infty \,, \\ \det{}^{1/d}(V_\mu) & \text{for } p = 0 \,, \\ \lambda_{\min}(V_\mu) & \text{for } p = -\infty \,, \end{cases}$$

where V_μ is a $d \times d$ matrix; see, e.g., [14, Chap. 6]. From a result in [7], if a measure μ_p optimal for some $\varphi_p(\cdot)$ with $p \in (-\infty, 1]$ is such that V_{μ_p} is proportional to the identity matrix I_d, then μ_p is simultaneously optimal for all orthogonally invariant criteria. A measure μ_p having this property is therefore ψ_k-optimal for all $k = 1, \ldots, d$. Notice that $\psi_1(\cdot)$ and $\psi_d^{1/d}(\cdot)$ respectively coincide with $\varphi_1(\cdot)$ and $\varphi_0(\cdot)$ (up to a multiplicative scalar).

Remark 4.3. *Using (3.10), when V is nonsingular we obtain the property*

$$\Psi_k(V) = \frac{(k+1)(d-k)!}{(d-k+1)k!}\,\det(V)\,\Psi_{d-k}(V^{-1})$$

which implies that maximising $\Psi_k(V)$ is equivalent to maximising $\log\det(V) + \log\Psi_{d-k}(V^{-1})$. Therefore, Theorem 4.4 can be reformulated as: μ_k^* maximises $\psi_k(\mu)$ if and only if

$$\max_{x\in\mathscr{X}}(x - E_{\mu_k^*})^\top \left[V_{\mu_k^*}^{-1} - V_{\mu_k^*}^{-1}\,\frac{\nabla_{\Psi_{d-k}}[V_{\mu_k^*}^{-1}]}{\Psi_{d-k}(V_{\mu_k^*}^{-1})}\,V_{\mu_k^*}^{-1}\right](x - E_{\mu_k^*}) \leq d-k\,,$$

with equality for x in the support of μ_k^*. When k is large (and $d-k$ is small), one may thus check the optimality of μ_k^* without using the complicated expressions of $\Psi_k(V)$ and $\nabla_{\Psi_k}[V]$.

4.1.2 A duality property

The characterisation of maximum-entropy measures can also be approached from the point of view of duality theory.

When $k=1$, the determination of a ψ_1-optimal measure μ_1^* is equivalent to the dual problem of constructing the minimum-volume ball $\mathscr{B}_d^*$ containing $\mathscr{X}$. If this ball has radius ρ, then $\psi_1(\mu_1^*) = 2\rho^2$, and the support points of μ_1^* are the points of contact between $\mathscr{X}$ and $\mathscr{B}_d^*$; see [1, Th. 6]. Moreover, there exists an optimal measure with no more than $d+1$ points.

The determination of an optimal measure μ_d^* is also dual to a simple geometrical problem: it corresponds to the determination of the minimum-volume ellipsoid $\mathscr{E}_d^*$ containing $\mathscr{X}$. This is equivalent to a D-optimal design problem in $\mathbb{R}^{d+1}$ for the estimation of $\beta = (\beta_0, \beta_1^\top)^\top$, $\beta_1 \in \mathbb{R}^d$, in the linear regression model with intercept $\beta_0 + \beta_1^\top x$, $x \in \mathscr{X}$, see [23]. Indeed, denote

$$W_\mu = \int_{\mathscr{X}} (1 \;\; x^\top)^\top (1 \;\; x^\top)\,\mu(\mathrm{d}x)\,.$$

Then $\mathscr{E}_{d+1}^* = \{z \in \mathbb{R}^{d+1} : z^\top W_{\mu_d^*}^{-1} z \leq d+1\}$, with μ_d^* maximising $\det(W_\mu)$, is the minimum-volume ellipsoid centered at the origin and containing the set $\{z \in \mathbb{R}^{d+1} : z = (1 \;\; x^\top)^\top,\; x \in \mathscr{X}\}$. Moreover, $\mathscr{E}_d^*$ corresponds to the intersection between $\mathscr{E}_{d+1}^*$ and the hyperplane $\{z\}_1 = 1$; see, e.g., [22]. This gives $\psi_d(\mu_d^*) = (d+1)/d!\,\det(W_{\mu_d^*})$. The support points of μ_d^* are the points of contact between $\mathscr{X}$ and $\mathscr{E}_d^*$, there exists an optimal measure with no more than $d(d+3)/2+1$ points, see [23].

The property below generalises this duality property to any $k \in \{1, \ldots, d\}$.

Theorem 4.2.

$$\max_{\mu\in\mathscr{M}} \Psi_k^{1/k}(V_\mu) = \min_{M,c:\;\mathscr{X}\subset\mathscr{E}(M,c)} \frac{1}{\phi_k^\infty(M)}\,,$$

where $\mathscr{E}(M,c)$ denotes the ellipsoid $\mathscr{E}(M,c) = \{x \in \mathbb{R}^d : (x-c)^\top M(x-c) \leq 1\}$ and $\phi_k^\infty(M)$ is the polar function

$$\phi_k^\infty(M) = \inf_{V\succeq 0:\;\mathrm{trace}(MV)=1} \frac{1}{\Psi_k^{1/k}(V)}\,. \tag{4.6}$$

The proof is given in Appendix. The polar function $\phi_k^\infty(\cdot)$ possesses the properties of what is called an information function in [14, Chap. 5]; in particular, it is concave on the set of symmetric non-negative definite matrices. This duality property has the following consequence.

Corollary 4.1. *The determination of a covariance matrix V_k^* that maximises $\Psi_k(V_\mu)$ with respect to $\mu \in \mathscr{M}$ is equivalent to the determination of an ellipsoid $\mathscr{E}(M_k^*, c_k^*)$ containing $\mathscr{X}$, minimum in the sense that M_k^* maximizes $\phi_k^\infty(M)$. The points of contact between $\mathscr{E}(M_k^*, c_k^*)$ and $\mathscr{X}$ form the support of μ_k^*.*

For any $V \succeq 0$, denote by $M_*(V)$ the matrix

$$M_*(V) = \frac{\nabla_{\Psi_k}[V]}{k\,\Psi_k(V)]} = \frac{1}{k}\,\nabla_{\log \Psi_k}[V]\,. \tag{4.7}$$

Note that $M_*(V) \succeq 0$, see [14, Lemma 7.5], and that

$$\mathrm{trace}[V M_*(V)] = 1\,,$$

see the proof of Theorem 4.4. The matrix $V \succeq 0$ maximises $\Psi_k(V)$ under the constraint $\mathrm{trace}(MV) = 1$ for some $M \succeq 0$ if and only if $V[M_*(V) - M] = 0$. Therefore, if M is such that there exists $V_* = V_*(M) \succeq 0$ such that $M = M_*[V_*(M)]$, then $\phi_k^\infty(M) = \Psi_k^{-1/k}[V_*(M)]$. When $k < d$, the existence of such a V_* is not ensured for all $M \succeq 0$, but happens when $M = M_k^*$ which maximises $\phi_k^\infty(M)$ under the constraint $\mathscr{X} \in \mathscr{E}(M, c)$. Moreover, in that case there exists a $\mu_k^* \in \mathscr{M}$ such that $M_k^* = M_*(V_{\mu_k^*})$, and this μ_k^* maximises $\psi_k(\mu)$ with respect to $\mu \in \mathscr{M}$.

Consider in particular the case $k = 1$. Then, $M_*(V) = I_d/\mathrm{trace}(V)$ and $\phi_1^\infty(M) = \lambda_{\min}(M)/2$. The matrix M_k^* of the optimal ellipsoid $\mathscr{E}(M_k^*, c_k^*)$ is proportional to the identity matrix and $\mathscr{E}(M_k^*, c_k^*)$ is the ball of minimum-volume that encloses $\mathscr{X}$.

When $k = 2$ and $I_d \succeq (d-1)M/\mathrm{trace}(M)$, direct calculations show that $\phi_2^\infty(M) = \Psi_2^{-1/2}[V_*(M)]$, with

$$V_*(M) = [I_d\,\mathrm{trace}(M)/(d-1) - M]\,[\mathrm{trace}^2(M)/(d-1) - \mathrm{trace}(M^2)]^{-1}\,;$$

the optimal ellipsoid is then such that $\mathrm{trace}^2(M)/(d-1) - \mathrm{trace}(M^2)$ is maximised.

4.1.3 Examples

Example 2 Take $\mathscr{X} = [0,1]^d$, $d \geq 1$ and denote by v_i, $i = 1, \ldots, 2^d$ the 2^d vertices of $\mathscr{X}$. Consider $\mu^* = (1/2^d)\sum_{i=1}^{2^d} \delta_{v_i}$, with δ_v the Dirac delta measure at v. Then, $V_{\mu^*} = I_d/4$ and one can easily check that μ^* is ψ_1-optimal. Indeed, $E_{\mu^*} = \mathbf{1}_d/2$, with $\mathbf{1}_d$ the d-dimensional vector of ones, and $\max_{x\in\mathscr{X}}(x - \mathbf{1}_d/2)^\top(2\,I_d)(x - \mathbf{1}_d/2) = d/2 = \mathrm{trace}\{V_{\mu^*}\nabla_{\Psi_1}[V_{\mu^*}]\}$. From Remark 4.2, the measure μ^* is ψ_k-optimal for all $k = 1, \ldots, d$.

Note that the two-point measure $\mu_1^* = (1/2)[\delta_{\mathbf{0}} + \delta_{\mathbf{1}_d}]$ is such that $V_{\mu_1^*} = (\mathbf{1}_d\,\mathbf{1}_d^\top)/4$ and $\psi_1(\mu_1^*) = d/2 = \psi_1(\mu^*)$, and is therefore ψ_1-optimal too. It is not ψ_k-optimal for $k > 1$, since $\psi_k(\mu_1^*) = 0$, $k > 1$.

Example 3 Take $\mathscr{X} = \mathscr{B}_d(\mathbf{0}, \rho)$, the closed ball of $\mathbb{R}^d$ centered at the origin $\mathbf{0}$ with radius ρ. Let μ_0 be the uniform measure on the sphere $\mathscr{S}_d(\mathbf{0}, \rho)$ (the boundary of $\mathscr{B}_d(\mathbf{0}, \rho)$). Then, V_{μ_0} is proportional to the identity matrix I_d, and $\mathrm{trace}[V_{\mu_0}] = \rho^2$ implies that $V_{\mu_0} = \rho^2 I_d/d$. Take $k = d$. We have $E_{\mu_0} = 0$ and

$$\max_{x\in\mathscr{X}}(x - E_{\mu_0})^\top \nabla_{\Psi_d}[V_{\mu_0}](x - E_{\mu_0}) = \frac{(d+1)\rho^{2d}}{d^{d-1}d!} = \mathrm{trace}\{V_{\mu_0}\nabla_{\Psi_d}[V_{\mu_0}]\}\,,$$

so that μ_0 is ψ_d-optimal from (4.4).

Let μ_d be the measure that allocates mass $1/(d+1)$ at each vertex of a d regular simplex having its $d+1$ vertices on $\mathscr{S}_d(\mathbf{0}, \rho)$, with squared volume $\rho^{2d}(d+$

$1)^{d+1}/[d^d(d!)^2]$. We also have $V_{\mu_d} = \rho^2 I_d/d$, so that μ_d is ψ_d-optimal too. In view of Remark 4.2, μ_0 and μ_d are ψ_k-optimal for all k in $\{1, \ldots, d\}$.

Let now μ_k be the measure that allocates mass $1/(k+1)$ at each vertex of a k regular simplex $\mathscr{P}_k$, centered at the origin, with its vertices on $\mathscr{S}_d(\mathbf{0}, \rho)$. The squared volume of $\mathscr{P}_k$ equals $\rho^{2k}\,(k+1)^{k+1}/[k^k(k!)^2]$. Without any loss of generality, we can choose the orientation of the space such that V_{μ_k} is diagonal, with its first k diagonal elements equal to ρ^2/k and the other equal to zero. Note that $\psi_{k'}(\mu_k) = 0$ for $k' > k$. Direct calculations based on (3.6) give

$$\psi_k(\mu_k) = \frac{k+1}{k!}\,\frac{\rho^{2k}}{k^k} \leq \psi_k(\mu_0) = \frac{k+1}{k!}\binom{d}{k}\frac{\rho^{2k}}{d^k}\,,$$

with equality for $k = 1$ and $k = d$, the inequality being strict otherwise. Figure 4 presents the efficiency $[\psi_k(\mu_k)/\psi_k(\mu_0)]^{1/k}$ as a function of k when $d = 20$.

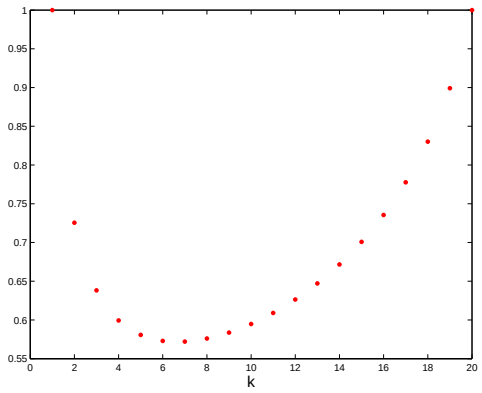

Figure 4: Efficiency $[\psi_k(\mu_k)/\psi_k(\mu_0)]^{1/k}$ as a function of k when $d = 20$ in Example 3.

4.2 Optimal design in regression models

In this section we consider the case when $V = M^{-1}(\xi)$, where $M(\xi)$ is the information matrix

$$M(\xi) = \int_{\mathbb{T}} f(t) f^\top(t)\, \xi(\mathrm{d}t)$$

in a regression model $Y_j = \theta^\top f(t_j) + \varepsilon_j$ with parameters $\theta \in \mathbb{R}^d$, for a design measure $\xi \in \Xi$. Here Ξ denotes the set of probability measures on a set $\mathbb{T}$ such that $\{f(t) : t \in \mathbb{T}\}$ is compact, and $M^{-1}(\xi)$ is the (asymptotic) covariance matrix of an estimator $\hat{\theta}$ of θ when the design variables t are distributed according to ξ. The value $\psi_k(\mu)$ of Theorem 3.1 defines a measure of dispersion for $\hat{\theta}$, that depends on ξ through $V_\mu = M^{-1}(\xi)$. The design problem we consider consists in choosing ξ that minimises this dispersion, as measured by $\Psi_k[M^{-1}(\xi)]$, or equivalently that maximises $\Psi_k^{-1}[M^{-1}(\xi)]$.

4.2.1 Properties

It is customary in optimal design theory to maximise a concave and Loewner-increasing function of $M(\xi)$, see [14, Chap. 5] for desirable properties of optimal design criteria. Here we have the following.

Theorem 4.3. *The functions $M \longrightarrow \Psi_k^{-1/k}(M^{-1})$, $k = 1, \ldots, d$, are Loewner-increasing, concave and differentiable on the set $\mathbb{M}^+$ of $d \times d$ symmetric positive-definite matrices. The functions $\Psi_k(\cdot)$ are also orthogonally invariant.*

Proof. The property (3.10) yields

$$\Psi_k^{-1/k}(M^{-1}) = \left(\frac{k+1}{k!}\right)^{-1/k} \frac{\det^{1/k}(M)}{\mathcal{E}_{d-k}^{1/k}(M)} \tag{4.8}$$

which is a concave function of M, see Eq. (10) of [11, p. 116]. Since $\Psi_k(\cdot)$ is Loewner-increasing, see [9], the function $M \longrightarrow \Psi_k^{-1/k}(M^{-1})$ is Loewner-increasing too. Its orthogonal invariance follows from the fact that it is defined in terms of the eigenvalues of M. □

Note that Theorems 3.1 and 4.3 imply that the functions $M \longrightarrow -\log \Psi_k(M)$ and $M \longrightarrow \log \Psi_k(M^{-1})$ are convex for all $k = 1, \ldots, d$, a question which was left open in [9].

As a consequence of Theorem 4.3, we can derive a necessary and sufficient condition for a design measure ξ_k^* to maximise $\Psi_k^{-1/k}[M^{-1}(\xi)]$ with respect to $\xi \in \Xi$, for $k = 1, \ldots, d$.

Theorem 4.4. *The design measure ξ_k^* such that $M(\xi_k^*) \in \mathbb{M}^+$ maximises $\tilde{\psi}_k(\xi) = \Psi_k^{-1/k}[M^{-1}(\xi)]$ with respect to $\xi \in \Xi$ if and only if*

$$\max_{t \in \mathbb{T}} f^\top(t) M^{-1}(\xi_k^*) \frac{\nabla_{\Psi_k}[M^{-1}(\xi_k^*)]}{\Psi_k[M^{-1}(\xi_k^*)]} M^{-1}(\xi_k^*) f(t) \le k \tag{4.9}$$

or, equivalently,

$$\max_{t \in \mathbb{T}} \left\{ f^\top(t) M^{-1}(\xi_k^*) f(t) - f^\top(t) \frac{\nabla_{\Psi_{d-k}}[M(\xi_k^*)]}{\Psi_{d-k}[M(\xi_k^*)]} f(t) \right\} \le d - k \,. \tag{4.10}$$

Moreover, there is equality in (4.9) and (4.10) for all t in the support of ξ_k^*.

Proof. From (4.8), the maximisation of $\tilde{\psi}_k(\xi)$ is equivalent to the maximisation of $\tilde{\phi}_k(\xi) = \log\det[M(\xi)] - \log \Psi_{d-k}[M(\xi)]$. The proof is similar to that of Theorem 4.4 and is based on the following expressions for the directional derivatives of these two functionals at ξ in the direction $\nu \in \Xi$,

$$F_{\tilde{\psi}_k}(\xi; \nu) = \operatorname{trace}\left(\frac{1}{k} M^{-1}(\xi) \frac{\nabla_{\Psi_k}[M^{-1}(\xi)]}{\Psi_k[M^{-1}(\xi)]} M^{-1}(\xi) \left[M(\nu) - M(\xi) \right] \right)$$

and

$$F_{\tilde{\phi}_k}(\xi; \nu) = \operatorname{trace}\left(\left\{ M^{-1}(\xi) - \frac{\nabla_{\Psi_{d-k}}[M(\xi)]}{\Psi_{d-k}[M(\xi)]} \right\} \left[M(\nu) - M(\xi) \right] \right) ,$$

and on the property $\operatorname{trace}\{M \nabla_{\Psi_j}[M]\} = j\, \Psi_j(M)$. □

In particular, consider the following special cases for k (note that $\Psi_0(M) = \mathcal{E}_0(M) = 1$ for any M).

$$\begin{aligned}
k = d: \quad & \tilde{\psi}_d(\xi) = \log\det[M(\xi)] \,, \\
k = d-1: \quad & \tilde{\psi}_{d-1}(\xi) = \log\det[M(\xi)] - \log \operatorname{trace}[M(\xi)] - \log 2 \,, \\
k = d-2: \quad & \tilde{\psi}_{d-2}(\xi) = \log\det[M(\xi)] \\
& \qquad - \log\left\{ \operatorname{trace}^2[M(\xi)] - \operatorname{trace}[M^2(\xi)] \right\} - \log(3/4) \,.
\end{aligned}$$

The necessary and sufficient condition (4.10) then takes the following form:

$$
\begin{aligned}
k = d: \quad & \max_{t\in\mathbb{T}} f^\top(t) M^{-1}(\xi_k^*) f(t) \le d\,, \\
k = d-1: \quad & \max_{t\in\mathbb{T}} \left\{ f^\top(t) M^{-1}(\xi_k^*) f(t) - \frac{f^\top(t) f(t)}{\mathrm{trace}[M(\xi_k^*)]} \right\} \le d-1\,, \\
k = d-2: \quad & \max_{t\in\mathbb{T}} \Big\{ f^\top(t) M^{-1}(\xi_k^*) f(t) \\
& \qquad -2\,\frac{\mathrm{trace}[M(\xi_k^*)] f^\top(t) f(t) - f^\top(t) M(\xi_k^*) f(t)}{\mathrm{trace}^2[M(\xi_k^*)] - \mathrm{trace}[M^2(\xi_k^*)]} \Big\} \le d-2\,.
\end{aligned}
$$

Also, for $k = 1$ condition (4.9) gives

$$
\max_{t\in\mathbb{T}} f^\top(t) \frac{M^{-2}(\xi_1^*)}{\mathrm{trace}[M^{-1}(\xi_1^*)]} f(t) \le 1
$$

(which corresponds to A-optimal design), and for $k = 2$

$$
\max_{t\in\mathbb{T}} \frac{\mathrm{trace}[M^{-1}(\xi_2^*)] f^\top(t) M^{-2}(\xi_2^*) f(t) - f^\top(t) M^{-3}(\xi_2^*) f(t)}{\mathrm{trace}^2[M^{-1}(\xi_2^*)] - \mathrm{trace}[M^{-2}(\xi_2^*)]} \le 1\,.
$$

Finally, note that a duality theorem, in the spirit of Theorem 4.2, can be formulated for the maximisation of $\Psi_k^{-1/k}[M^{-1}(\xi)]$; see [14, Th. 7.12] for the general form a such duality properties in optimal experimental design.

4.2.2 Examples

Example 4 For the linear regression model on $\theta_0 + \theta_1 x$ on $[-1, 1]$, the optimal design for $\tilde\psi_k(\cdot)$ with $k = d = 2$ or $k = 1$ is

$$
\xi_k^* = \left\{ \begin{array}{cc} -1 & 1 \\ 1/2 & 1/2 \end{array} \right\},
$$

where the first line corresponds to support points and the second indicates their respective weights.

Example 5 For linear regression with the quadratic polynomial model $\theta_0 + \theta_1 t + \theta_2 t^2$ on $[-1, 1]$, the optimal designs for $\tilde\psi_k(\cdot)$ have the form

$$
\xi_k^* = \left\{ \begin{array}{ccc} -1 & 0 & 1 \\ w_k & 1-2w_k & w_k \end{array} \right\},
$$

with $w_3 = 1/3$, $w_2 = (\sqrt{33}-1)/16 \simeq 0.2965352$ and $w_1 = 1/4$. Define the efficiency $\mathrm{Eff}_k(\xi)$ of a design ξ as

$$
\mathrm{Eff}_k(\xi) = \frac{\tilde\psi_k(\xi)}{\tilde\psi_k(\xi_k^*)}\,.
$$

Table 2 gives the efficiencies $\mathrm{Eff}_k(\xi_j^*)$ for $j, k = 1, \ldots, d = 3$. The design ξ_2^*, optimal for $\tilde\psi_2(\cdot)$, appears to make a good compromise between A-optimality (which corresponds to $\tilde\psi_1(\cdot)$) and D-optimality (which corresponds to $\tilde\psi_3(\cdot)$).

Example 6 For linear regression with the cubic polynomial model $\theta_0 + \theta_1 t + \theta_2 t^2 + \theta_3 t^3$ on $[-1, 1]$, the optimal designs for $\tilde\psi_k(\cdot)$ have the form

$$
\xi_k^* = \left\{ \begin{array}{cccc} -1 & -z_k & z_k & 1 \\ w_k & 1/2 - w_k & 1/2 - w_k & w_k \end{array} \right\},
$$

Table 2: Efficiencies $\mathrm{Eff}_k(\xi_j^*)$ for $j,k=1,\ldots,d$ in Example 5.

	Eff_1	Eff_2	Eff_3
ξ_1^*	1	0.9770	0.9449
ξ_2^*	0.9654	1	0.9886
ξ_3^*	0.8889	0.9848	1

Table 3: Efficiencies $\mathrm{Eff}_k(\xi_j^*)$ for $j,k=1,\ldots,d$ in Example 6.

	Eff_1	Eff_2	Eff_3	Eff_4
ξ_1^*	1	0.9785	0.9478	0.9166
ξ_2^*	0.9694	1	0.9804	0.9499
ξ_3^*	0.9180	0.9753	1	0.9897
ξ_4^*	0.8527	0.9213	0.9872	1

where

$$\begin{aligned} z_4 &= 1/\sqrt{5} \simeq 0.4472136\,, & w_4 &= 0.25\,,\\ z_3 &\simeq 0.4350486\,, & w_3 &\simeq 0.2149859\,,\\ z_2 &\simeq 0.4240013\,, & w_2 &\simeq 0.1730987\,,\\ z_1 &= \sqrt{3\sqrt{7}-6}/3 \simeq 0.4639509\,, & w_1 &= (4-\sqrt{7})/9 \simeq 0.1504721\,, \end{aligned}$$

with z_3 satisfying the equation $2z^6-3z^5-45z^4+6z^3-4z^2-15z+3=0$ and

$$w_3 = \frac{5\,z^6+5\,z^4+5\,z^2+1-\sqrt{z^{12}+2\,z^{10}+3\,z^8+60\,z^6+59\,z^4+58\,z^2+73}}{12(z^6+z^4+z^2-3)}\,,$$

with $z=z_3$. For $k=d-2=2$, the numbers z_2 and w_2 are too difficult to express analytically. Table 3 gives the efficiencies $\mathrm{Eff}_k(\xi_j^*)$ for $j,k=1,\ldots,d$. Here again the design ξ_2^* appears to make a good compromise: it maximises the minimum efficiency $\min_k \mathrm{Eff}_f(\cdot)$ among the designs considered.

Appendix

Shift-invariance and positive homogeneity Denote by $\mathscr{M}$ the set of probability measures defined on the Borel subsets of $\mathscr{X}$, a compact subset of $\mathbb{R}^d$. For any $\mu\in\mathscr{M}$, any $\theta\in\mathbb{R}^d$ and any $\lambda\in\mathbb{R}^+$, respectively denote by $T_{-\theta}[\mu]$ and $H_{\lambda^{-1}}[\mu]$ the measures defined by:

$$\text{for any } \mu\text{-measurable } \mathscr{A}\subseteq\mathscr{X}\,,\ T_{-\theta}[\mu](\mathscr{A}+\theta)=\mu(\mathscr{A})\,,\ H_{\lambda^{-1}}[\mu](\lambda\mathscr{A})=\mu(\mathscr{A})\,,$$

where $\mathscr{A}+\theta=\{x+\theta : x\in\mathscr{A}\}$ and $\lambda\mathscr{A}=\{\lambda\,x : x\in\mathscr{A}\}$. The shift-invariance of $\phi(\cdot)$ then means that $\phi(T_{-\theta}[\mu])=\phi(\mu)$ for any $\mu\in\mathscr{M}$ and any $\theta\in\mathbb{R}^d$, positive homogeneity of degree q means that $\phi(H_{\lambda^{-1}}[\mu])=\lambda^q\,\phi(\mu)$ for any $\mu\in\mathscr{M}$ and any $\lambda\in\mathbb{R}^+$.

The variance is the only concave central moment For $q\neq 2$, the q-th central moment $\Delta_q(\mu)=\int\|x-E_\mu\|^q\,\mu(\mathrm{d}x)$ is shift-invariant and homogeneous of degree q, but it is not concave on $\mathscr{M}(\mathscr{X})$. Indeed, consider for instance the two-point probability measures

$$\mu_1=\left\{\begin{array}{cc}0&1\\1/2&1/2\end{array}\right\}\ \text{and}\ \mu_2=\left\{\begin{array}{cc}0&101\\w&1-w\end{array}\right\}\,,$$

where the first line denotes the support points and the second one their respective weights. Then, for

$$w=1-\frac{1}{404}\,\frac{201^{q-1}-202q+405}{201^{q-1}-101q+102}$$

one has $\partial^2 \Delta_q[(1-\alpha)\mu_1 + \alpha\mu_2]/\partial\alpha^2\big|_{\alpha=0} \geq 0$ for all $q \geq 1.84$, the equality being obtained at $q = 2$ only. Counterexamples are easily constructed for values of q smaller than 1.84.

Proof of Theorem 4.2
(i) The fact that $\max_{\mu\in\mathscr{M}} \Psi_k^{1/k}(V_\mu) \geq \min_{M,c:\ \mathscr{X}\subset\mathscr{E}(M,c)} 1/\phi_k^\infty(M)$ is a consequence of Theorem 4.4. Indeed, the measure μ_k^* maximises $\Psi_k^{1/k}(V_\mu)$ if and only if

$$(x - E_{\mu_k^*})^\top M_*(V_{\mu_k^*})(x - E_{\mu_k^*}) \leq 1 \quad \text{for all } x \text{ in } \mathscr{X}. \tag{4.11}$$

Denote $M_k^* = M_*(V_{\mu_k^*})$, $c_k^* = E_{\mu_k^*}$, and consider the Lagrangian $L(V,\alpha;M)$ for the maximisation of $(1/k)\log\Psi_k(V)$ with respect to $V \succeq 0$ under the constraint $\mathrm{trace}(MV) = 1$: $L(V,\alpha;M) = (1/k)\log\Psi_k(V) - \alpha[\mathrm{trace}(MV) - 1]$. We have

$$\frac{\partial L(V,1;M_k^*)}{\partial V}\bigg|_{V=V_{\mu_k^*}} = M_k^* - M_k^* = 0$$

and $\mathrm{trace}(M_k^* V_{\mu_k^*}) = 1$, with $V_{\mu_k^*} \succeq 0$. Therefore, $V_{\mu_k^*}$ maximises $\Psi_k(V)$ under the constraint $\mathrm{trace}(M_k^* V) = 1$, and, moreover, $\mathscr{X} \subset \mathscr{E}(M_k^*, c_k^*)$ from (4.11). This implies

$$\begin{aligned}\Psi_k^{1/k}(V_{\mu_k^*}) &= \max_{V\succeq 0:\ \mathrm{trace}(M_k^* V)=1} \Psi_k^{1/k}(V) \\ &\geq \min_{M,c:\ \mathscr{X}\subset\mathscr{E}(M,c)} \ \max_{V\succeq 0:\ \mathrm{trace}(MV)=1} \Psi_k^{1/k}(V) = \min_{M,c:\ \mathscr{X}\subset\mathscr{E}(M,c)} \frac{1}{\phi_k^\infty(M)}\,.\end{aligned}$$

(ii) We prove now that $\min_{M,c:\ \mathscr{X}\subset\mathscr{E}(M,c)} 1/\phi_k^\infty(M) \geq \max_{\mu\in\mathscr{M}} \Psi_k^{1/k}(V_\mu)$. Note that we do not have an explicit form for $\phi_k^\infty(M)$ and that the infimum in (4.6) can be attained at a singular V, not necessarily unique, so that we cannot differentiate $\phi_k^\infty(M)$. Also note that compared to the developments in [14, Chap. 7], here we consider covariance matrices instead of moment matrices.

Consider the maximisation of $\log\phi_k^\infty(M)$ with respect to M and c such that $\mathscr{X} \subset \mathscr{E}(M,c)$, with Lagrangian

$$L(M,c,\beta) = \log\phi_k^\infty(M) + \sum_{x\in\mathscr{X}} \beta_x[1 - (x-c)^\top M(x-c)]\,,\ \beta_x \geq 0 \text{ for all } x \text{ in } \mathscr{X}.$$

For the sake of simplicity we consider here $\mathscr{X}$ to be finite, but β may denote any positive measure on $\mathscr{X}$ otherwise. Denote the optimum by $T^* = \max_{M,c:\ \mathscr{X}\subset\mathscr{E}(M,c)} \log\phi_k^\infty(M)$. It satisfies $T^* = \max_{M,c}\min_{\beta\geq 0} L(M,c,\beta) \leq \min_{\beta\geq 0}\max_{M,c} L(M,c,\beta)$, and $\max_{M,c} L(M,c,\beta)$ is attained for any c such that $Mc = M\sum_{x\in\mathscr{X}}\beta_x\, x/(\sum_{x\in\mathscr{X}}\beta_x)$, that is, in particular for

$$c^* = \frac{\sum_{x\in\mathscr{X}}\beta_x\, x}{\sum_{x\in\mathscr{X}}\beta_x}\,,$$

and for M^* such that $0 \in \partial_M L(M,c^*,\beta)\big|_{M=M^*}$, the subdifferential of $L(M,c^*,\beta)$ with respect to M at M^*. This condition can be written as

$$\sum_{x\in\mathscr{X}} \beta_x\,(x-c^*)(x-c^*)^\top = \tilde{V} \in \partial\log\phi_k^\infty(M)\big|_{M=M^*}\,,$$

with $\partial\log\phi_k^\infty(M)$ the subdifferential of $\log\phi_k^\infty(M)$,

$$\partial\log\phi_k^\infty(M) = \{V \succeq 0 :\ \Psi_k^{1/k}(V)\phi_k^\infty(M) = \mathrm{trace}(MV) = 1\}\,,$$

see [14, Th. 7.9]. Since $\mathrm{trace}(MV) = 1$ for all $V \in \partial \log \phi_k^\infty(M)$, $\mathrm{trace}(M^* \tilde{V}) = 1$ and thus $\sum_{x \in \mathscr{X}} \beta_x \, (x - c^*)^\top M^* (x - c^*) = 1$. Also, $\Psi_k^{1/k}(\tilde{V}) = 1/\phi_k^\infty(M^*)$, which gives

$$L(M^*, c^*, \beta) = -\log \Psi_k^{1/k} \left[\sum_{x \in \mathscr{X}} \beta_x \, (x - c^*)(x - c^*)^\top \right] + \sum_{x \in \mathscr{X}} \beta_x - 1 \, .$$

We obtain finally

$$\begin{aligned}
&\min_{\beta \geq 0} L(M^*, c^*, \beta) \\
&\quad = \min_{\gamma > 0, \ \alpha \geq 0} \left\{ -\log \Psi_k^{1/k} \left[\sum_{x \in \mathscr{X}} \alpha_x \, (x - c^*)(x - c^*)^\top \right] + \gamma - \log(\gamma) - 1 \right\} , \\
&\quad = \min_{\alpha \geq 0} -\log \Psi_k^{1/k} \left[\sum_{x \in \mathscr{X}} \alpha_x \, (x - c^*)(x - c^*)^\top \right] = -\log \Psi_k^{1/k}(V_k^*) \, ,
\end{aligned}$$

where we have denoted $\gamma = \sum_{x \in \mathscr{X}} \beta_x$ and $\alpha_x = \beta_x / \gamma$ for all x. Therefore $T^* \leq -\log \Psi_k^{1/k}(V_k^*)$, that is, $\log \left[\min_{M, c: \ \mathscr{X} \subset \mathscr{E}(M,c)} 1/\phi_k^\infty(M) \right] \geq \log \Psi_k^{1/k}(V_k^*)$.

References

[1] G. Björck. Distributions of positive mass, which maximize a certain generalized energy integral. *Arkiv för Matematik*, 3(21):255–269, 1956.

[2] M.H. DeGroot and M.M. Rao. Bayes estimation with convex loss. *Ann. Math. Statist.*, 34(3):839–846, 1963.

[3] F.R. Gantmacher. *Théorie des Matrices*. Dunod, Paris, 1966.

[4] C. Gini. Measurement of inequality of incomes. *Economic J.*, 31(121):124–126, 1921.

[5] A. Giovagnoli and H.P. Wynn. Multivariate dispersion orderings. *Stat. & Prob. Lett.*, 22(4):325–332, 1995.

[6] M. Hainy, W.G. Müller, and H.P. Wynn. Learning functions and approximate bayesian computation design: ABCD. *Entropy*, 16(8):4353–4374, 2014.

[7] R. Harman. Lower bounds on efficiency ratios based on ϕ_p-optimal designs. In A. Di Bucchianico, H. Läuter, and H.P. Wynn, editors, *mODa'7 – Advances in Model–Oriented Design and Analysis, Proceedings of the 7th Int. Workshop, Heeze (Netherlands)*, pages 89–96, Heidelberg, 2004. Physica Verlag.

[8] J. Kiefer and J. Wolfowitz. The equivalence of two extremum problems. *Canad. J. Math.*, 12:363–366, 1960.

[9] 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.

[10] I.G. Macdonald. *Symmetric functions and Hall polynomials*. Oxford University Press, Oxford, 1995. [2nd ed.].

[11] M. Marcus and H. Minc. *A Survey of Matrix Theory and Matrix Inequalities*. Dover, New York, 1964.

[12] H. Oja. Descriptive statistics for multivariate distributions. *Stat. & Prob. Lett.*, 1(6):327–332, 1983.

[13] L. Pronzato. On a property of the expected value of a determinant. *Stat. & Prob. Lett.*, 39:161–165, 1998.

[14] F. Pukelsheim. *Optimal Experimental Design*. Wiley, New York, 1993.

[15] C.R. Rao. Diversity and dissimilarity coefficients: a unified approach. *Theoret. Popn Biol.*, 21(1):24–43, 1982.

[16] C.R. Rao. Diversity: Its measurement, decomposition, apportionment and analysis. *Sankhyā: Indian J. Statist., Series A*, 44(1):1–22, 1982.

[17] C.R. Rao. Convexity properties of entropy functions and analysis of diversity. In *Inequalities in Statistics and Probability*, volume 5, pages 68–77. Lecture Notes-Monograph Series, IMS, Hayward, CA, 1984.

[18] R.L. Schilling, R. Song, and Z. Vondracek. *Bernstein Functions: Theory and Applications*. de Gruyter, Berlin/Boston, 2012.

[19] I.J. Schoenberg. Metric spaces and positive definite functions. *Trans. Amer. Math. Soc.*, 44(3):522–536, 1938.

[20] R.J. Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley, New York, 1980.

[21] M. Shaked. Dispersive ordering of distributions. *J. Appl. Prob.*, pages 310–320, 1982.

[22] N.Z. Shor and O.A. Berezovski. New algorithms for constructing optimal circumscribed and inscribed ellipsoids. *Optim. Meth. Soft.*, 1:283–299, 1992.

[23] D.M. Titterington. Optimal design: some geometrical espects of D-optimality. *Biometrika*, 62(2):313–320, 1975.

[24] H.R. van der Vaart. A note on Wilks' internal scatter. *Ann. Math. Statist.*, 36(4):1308–1312, 1965.

[25] S.S. Wilks. Certain generalizations in the analysis of variance. *Biometrika*, 24:471–494, 1932.

[26] S.S. Wilks. Multidimensional statistical scatter. In I. Olkin, S.G. Ghurye, W. Hoeffding, W.G. Madow, and H.B. Mann, editors, *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, pages 486–503. Stanford University Press, Stanford, 1960.