Shape of a distribution through the L_2-Wasserstein Distance *

Juan A. Cuesta-Albertos (cuesta@matesco.unican.es)
Departamento de Matemáticas, Universidad de Cantabria, España

Carlos Matrán Bea[a] (matran@eio.uva.es)
Departamento de Estadística e I.O., Universidad de Valladolid, España

Jesús M. Rodríguez Rodríguez[a] (jesusmro@eio.uva.es)
Departamento de Estadística e I.O., Universidad de Valladolid, España

September 28, 2000

Abstract. Let Q be a probability measure on $\mathbb{R}^d$ and let $\Im$ be a family of probability measures on $\mathbb{R}^d$ which will be considered as a pattern. For suitable patterns we consider the closest law to Q in $\Im$, through the L_2-Wasserstein distance, as a descriptive measure associated to Q. The distance between Q and $\Im$ is a natural measure of the fit of Q to the pattern.

We analyze this approach via the consideration of different patterns. Some of them generalize usual location and dispersion measures. Special attention will be paid to patterns based on uniform distributions on suitable families of sets, like the intervals in the unidimensional case (which allows us to analyze the flatness of the one-dimensional distributions) or the ellipsoids for the multivariate distributions.

Keywords: Wasserstein distance, flatness, shape of a distribution, multivariate distributions

AMS 1991 subject classification: Primary 62F05, 62E20; secondary 60F25.

1. Introduction

Let P and Q be two probability measures defined on the euclidean d-dimensional space with finite second order moment. Let $M(P,Q)$ be the set of the distributions on $\mathbb{R}^{2d}$ with P and Q as marginals. The so-called L_2-Wasserstein distance between them is

$$W(P,Q) := \inf\left\{ \left[\int \|x-y\|^2 \nu(dx,dy)\right]^{1/2} : \quad \nu \in M(P,Q) \right\}.$$

* Research partially supported by DGESIC, grants PB98-0369-C02-00,01 and 02.
[a] These authors have also been supported by PAPIJCL grant VA08/97.

The interest of this metric is often made apparent by the fact that it metrizes the convergence in distribution plus the convergence of moments (see e.g. (Rachev and Rüschendorf, 1998)).

The framework of this paper is that of the following problem:

Given a family of distributions $\Im$ and a distribution Q, find the nearest distribution to Q on $\Im$ according to the L_2-Wasserstein distance.

Our point of view relies on the consideration of $\Im$ as a pattern. Thus, the closest distribution to Q in $\Im$ can be considered as an indicator of a certain characteristic of Q, and the value of the distance as a measure of the fit of Q to the pattern.

We consider several patterns of interest. The first ones are discrete and generalize the mean, the variance or the k-means of a distribution. We present them in Section 2 within the framework of the one-dimensional distributions because of the simplicity of this case, but we want to state that every exposed result for these patterns, with the obvious changes, remains valid for the multidimensional case.

In Section 2 we also consider the pattern composed by the uniform distributions on intervals. This pattern is employed to define the flatness of a distribution.

The multivariate case is discussed in Section 3. We begin by making some comments on the difficulty of computing optimal couplings (see Definition 1 below) in this situation and the possibilities of obtaining approximate solutions. We also analyze patterns spherically equivalent (see Definition 4) to the distribution under consideration as a special case in which it is possible to find the closest distribution. As a particular case, we analyze the case in which the distribution under consideration is spherically equivalent to the uniform distribution on a sphere.

We finalize the paper with a brief analysis of the possibility of employing patterns composed by uniform distributions whose support have a lower dimension than the support of Q and their possible relation with the principal components analysis.

A key tool will be the following representation for W in terms of random variables (r.v.'s).

DEFINITION 1. Given X, Y two $\mathbb{R}^{2d}$-valued r.v.'s defined on some probability space, we will say that the pair (X, Y) is an optimal coupling (o.c.) for (P, Q) if the marginal distributions of (X, Y) are P and Q and $W^2(P, Q) = E\|X - Y\|^2$.

Concerning the notation, given the distribution Q (resp. the r.v. Y), we will denote by F_Q (resp. by F_Y) its distribution function and by σ_Q

(resp. σ_Y)its standard deviation. Given the distribution function F its quantile function F^{-1} is defined, as usuall, by $F^{-1}(t) = \inf\{x : F(x) \geq t\}$. Given the Borel set A, I_A will denote the indicator of A. Finally, the distribution of the r.v. Y will be denoted by P_Y.

2. Patterns in the univariate case

For one-dimensional distributions P and Q, $W(P,Q)$ can be explicitly calculated (see e.g. Bickel and Freedman, 1981) by

$$W(P,Q) = \left[\int_0^1 \left(F_P^{-1}(t) - F_Q^{-1}(t)\right)^2 dt\right]^{\frac{1}{2}}. \tag{1}$$

From (1) it can be deduced that if P^* and Q^* are the centered in mean translations of P and Q, then

$$W^2(P,Q) = W^2(P^*,Q^*) + (m_P - m_Q)^2 \tag{2}$$

where m_P and m_Q are respectively the means of P and Q.

2.1. Discrete patterns.

If F_Q is continuous, then an o.c. between P and Q is $\left(F_P^{-1}\left(F_Q\left(Y\right)\right), Y\right)$ where Y is any r.v. with distribution Q. Thus, if P is discrete, with support $\{x_1, ..., x_k\}$, where we suppose $x_1 < ... < x_k$, and $p_i = P[x_i]$, $i = 1, ..., k$, the o.c. between P and Q is obtained by dividing the support of Q by the corresponding quantiles, $\{-\infty, F_Q^{-1}(p_1), ..., F_Q^{-1}(\sum_{i=1}^{k-1} p_i), \infty\}$, and assigning to each interval the corresponding x_i. Then

$$W^2(P,Q) = \sum_{i=1}^{k} E\left[(Y - x_i)^2 I_{\left\{Y \in \left(F_Q^{-1}(\overline{p}_{i-1}), F_Q^{-1}(\overline{p}_i)\right)\right\}}\right],$$

where $\overline{p}_0 = 0$, and $\overline{p}_i = \sum_{j=1}^{i} p_j$, $i = 1, ..., k$.

In consequence, if we fix $p_1, ..., p_k > 0$, $\sum_i p_i = 1$ and we consider the pattern given by the family $\Im = \{P :$ discrete with probabilities $p_1, p_2, ..., p_k\}$, then the support of the nearest distribution to Q in $\Im$ is composed by the points

$$x_i = E\left[Y \middle/ Y \in \left(F_Q^{-1}(\overline{p}_{i-1}), F_Q^{-1}(\overline{p}_i)\right)\right],\ i = 1, ...k,$$

Table I. Support of the uniform distribution on the 5 points closest to the chosen distribution.

Distribution	x_1	x_2	x_3	x_4	x_5
Exponencial(1)	0.1074	0.3601	0.6998	1.2231	2.6094
Normal(0,1)	-1.3998	-0.5319	0	0.5319	1.3998
Student T(5)	-1.7358	-0.5704	0	0.5704	1.7358
Student T(10)	-1.5464	-0.55064	0	0.55064	1.5464
Student T(20)	-1.4687	-0.54114	0	0.54114	1.4687
Student T(40)	-1.4332	-0.53649	0	0.53649	1.4332

or, also (see (Tarpey and Flury, 1996)), if (X, Y) is an o.c. for (P, Q), then X is self-consistent for Y .

An interesting possibility is to choose $p_1 = p_2 = ... = p_k = k^{-1}$, which would give a discrete uniform distribution as close to Q as possible. Moreover, if $k = 1$, then the support of the nearest distribution is the mean of Q and the distance is σ_Q.

As an illustration we next include the values supporting the uniform distribution on five points for several distributions. Notice the difference between the gaussian and the t-distributions. Those values could be employed to construct a goodness test of fit between them but we have not yet explored this possibility.

Things behave very differently if we only fix $k \in \mathbb{N}$. I.e. if we consider the pattern $\Im = \{P : \text{supported by } k \text{ points}\}$. In this case the closest distribution to Q is not necessarily unique, but the support of every closest distribution to Q gives a k-mean of Q and viceversa.

2.2. Flatness

The pattern $\Im$ composed by the uniform distributions on intervals is a suitable reference for obtaining a measure of flatness of a given distribution. From (2), we have that the closest distribution to Q in $\Im$ has the same mean as Q and that the distance is mean-invariant. Therefore, w.l.o.g., we can assume that Q is centered. Let $a > 0$ and P be uniform on $(-a, a)$. From (1) we have that

$$\begin{aligned} W^2(P,Q) &= \int_0^1 \left(2at - a - F_Q^{-1}(t)\right)^2 dt \\ &= \frac{a^2}{3} + \int_0^1 \left(F_Q^{-1}(t)\right)^2 dt - 4a \int_0^1 tF_Q^{-1}(t)dt. \end{aligned} \tag{3}$$

The minimum of this expression is reached at

$$a = 6 \int_0^1 t F_Q^{-1}(t) dt, \tag{4}$$

and then

$$\inf_{P \in \Im} W^2(P, Q) = \sigma_Q^2 - \frac{a^2}{3}.$$

If this value is divided by σ_Q^2, we obtain a measure of flatness of Q, which is independent of the location and dispersion of Q.

DEFINITION 2. Given a distribution Q with finite second order moment and $\sigma_Q > 0$, we define its flatness by

$$\mathcal{F}(Q) := \frac{\inf_{P \in \Im} W(P, Q)}{\sigma_Q} = \sqrt{1 - \frac{1}{12} \frac{c^2}{\sigma_Q^2}}$$

where $c = 12 \int_0^1 t F_Q^{-1}(t) dt$ is the length of the interval supporting the uniform distribution closest to Q according to the L_2-Wasserstein distance.

From the definition it becomes obvious that $0 \leq \mathcal{F}(Q) \leq 1$. Notice that $\mathcal{F}(Q) = 0$ happens only when $Q \in \Im$. The next proposition shows that the extreme value 1 would be only attained as a limit case for degenerated distributions.

PROPOSITION 3. *Given a probability measure Q with finite second order moment, then $\sigma_Q = 0$ if and only if $\inf_{P \in \Im} W(P, Q) = \sigma_Q$.*

PROOF.- The necessity of the condition is obvious. On the other hand, let us assume that $\sigma_Q > 0$ and that $\inf_{P \in \Im} W(P, Q) = \sigma_Q$. In a first step we will assume that Q is symmetric and centered. Given $k \in \mathbb{R}$ we will denote by $\delta(k)$ the k-quantil of Q.

Let $k > 0$ be a continuity point of Q. Let $a \in (0, k)$ and let Y be a r.v. with uniform distribution on $(-a, a)$. If H_a is the map defined by $H_a(y) := \delta^{-1}(0.5 + y/(2a)), y \in (-a, a)$, then the distribution of $H_a(Y)$ is Q. The symmetry of Q implies that if $y > 0$ (resp. $y < 0$), then $H_a(y) \geq 0$ (≤ 0).

Let $y \in (-a, a)$ such that $|H_a(y)| > k$. Then $|y| > 2a(\delta(k) - 0.5)$ and

$$\begin{aligned} H_a^2(y) - (H_a(y) - y)^2 &> H_a^2(y) - (|H_a(y)| - 2a(\delta(k) - 0.5))^2 \\ &= 4a(\delta(k) - 0.5)|H_a(y)| - (2a(\delta(k) - 0.5))^2 \\ &\geq 2a(\delta(k) - 0.5)|H_a(y)|. \end{aligned}$$

Now let us define $f_a(y) = H_a^2(y) - (y - H_a(y))^2$, $y \in (-a, a)$. Then

$$\begin{aligned}\int_{\{y:|H_a(y)|\leq k\}} f_a(y)(2a)^{-1}dy &= \int_{\{y:|H_a(y)|\leq k\}} \left(2H_a(y)y - y^2\right)(2a)^{-1}dy \\ &\geq -\int_{\{y:|H_a(y)|\leq k\}} y^2(2a)^{-1}dy \\ &= -\frac{8}{3}a^2(\delta(k)-0.5)^3;\end{aligned}$$

and, in consequence,

$$\begin{aligned}&\sigma_Q^2 - E[(Y^2 - H_a(Y))^2] = E[H_a^2(Y)] - E[(Y^2 - H_a(Y))^2] \\ &= \int_{\{y:|H_a(y)|\leq k\}} f_a(y)(2a)^{-1}dy + \int_{\{y:|H_a(y)|> k\}} f_a(y)(2a)^{-1}dy \\ &\geq -\frac{8}{3}a^2(\delta(k)-0.5)^3 + 2a(\delta(k)-0.5)\int_{\{|H_a(y)|>k\}} |H_a(y)|(2a)^{-1}dt \\ &= 4a(\delta(k)-0.5)\left[k(1-\delta(k)) - \frac{2a}{3}(\delta(k)-0.5)^2\right],\end{aligned}$$

and this expression is positive if we take small enough $a > 0$. In consequence, $W(P_Y, Q) < \sigma_Q$ and the result is proved for symmetric distributions.

Now let us assume that Q is a nondegenerated distribution such that $\inf_{P\in\Im} W(P,Q) = \sigma_Q$. Without loss of generality we can assume that Q is centered. Let Y, B be two independent r.v.'s such that $P_Y = Q$ and B takes the values $+1$ and -1 with probability a half each. Let $Y^* = BY$ and let X be a r.v. whose distribution is uniform on an interval centered at 0. We have

$$\begin{aligned}E\left[(Y^*-X)^2\right] &= E\left[(Y-X)^2 I_{\{B=1\}}\right] + E\left[(-Y-X)^2 I_{\{B=-1\}}\right] \\ &\geq \frac{1}{2}\left(E[Y^2] + E[Y^2]\right) \\ &= E\left[Y^2 I_{\{B=1\}}\right] + E\left[Y^2 I_{\{B=-1\}}\right] \\ &= E\left[(Y^*)^2\right],\end{aligned}$$

where the inequality comes from the independence between Y and B and from $\inf_{P\in\Im} W(P,Q) = \sigma_Q$ (thus also $\inf_{P\in\Im} W(P, P_{-Y}) = \sigma_Q$).

Finally, if we minimize the previous expression on X we would obtain that $\inf_{P\in\Im} W(P, P_{Y^*}) = \sigma_{P_{Y^*}}$, which is not possible because P_{Y^*} is symmetric.

Table II. Variance, kurtosis and squared flatness of some distributions.

Distribution	Density Function	Variance	Kurtosis	Flatness^2
Normal	$\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}$	1.000	3.000	0.045
Uniform	$\frac{1}{\sqrt{12}}$	1.000	1.800	0.000
Double exponential	$\frac{1}{2}e^{-\|x\|}$	2.000	6.000	0.156
Triangular	$1-\|x\|$	0.167	2.400	0.0199
Triangular inverted	$\|x\|$	0.500	1.333	0.040
Exponential	$e^{-(x+1)}$	1.000	9.000	0.250
Mixture of Normal	$c\left(e^{-\frac{1}{2}(x+1)^2}+e^{-\frac{1}{2}(x-1)^2}\right)$	2.000	2.500	0.022
Mixture of Normal	$c\left(e^{-\frac{1}{2}(x+3)^2}+e^{-\frac{1}{2}(x-3)^2}\right)$	10.00	1.380	0.047
Discrete	$\begin{cases} -0.5 & con \quad p=0.5 \\ 0.5 & con \quad p=0.5 \end{cases}$	0.250	1.000	0.250

Notice that if F_Q is continuous and Y is a r.v. with distribution Q, then (4) becomes

$$a = 6 \ \mathrm{Cov}(Y, F_Q(Y)),$$

and, that if Q is discrete, not necesarily centered, then

$$a = 3\left(\sum_{i=1}^{k} y_i\left(q_i^2 - q_{i-1}^2\right)\right) + \int tQ(dt),$$

where $Q(y_i) = q_i, i = 1, ..., k$ and $\sum_i q_i = 1$.

The flatness of a distribution measures a kind of shape which is in some sense similar to the kurtosis, however its behaviour can be very different as it is showed in the next table, where we include, for sake of comparison, the kurtosis and squared flatness for some distributions.

3. Patterns in the Multivariate case

In the multivariate case the computation of the o.c. between two distributions is only possible for some particular distributions (see, for instance, Rachev and Rüschendorf, 1998). Therefore the search for the closest distribution on a pattern is not possible for most of the cases.

However, when the explicit calculation of the Wasserstein distance between two distributions is not feasible, it is still possible to obtain an approximation by simulation (see (Cuesta-Albertos et al., 1997) for a justification) and to compute the distance between the samples by using an algorithm similar to the one proposed in Aurenhammer et al., 1998.

In order to describe the shape of a distribution some interesting models to compare them with are those formed by uniform distributions in some peculiar type of sets, like balls, ellipsoids or convex sets in general. Notice that it is easy to obtain samples of these distributions by using methods of acceptance and rejection.

It is not too dificult to find an approximation to the uniform distribution on an ellipsoid closest to a discrete distribution by using a procedure based on the sequential search for the directions of the axis. In the following figure a case is shown.

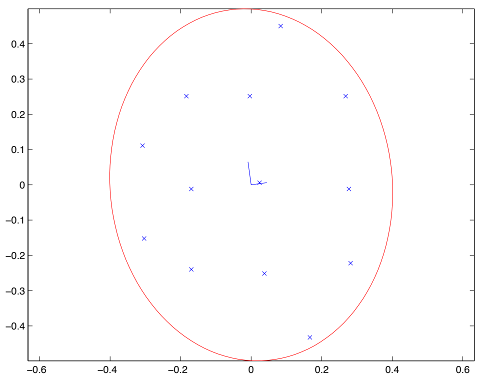

Figure 1. Uniform on an ellipse nearest to a set of points. The axes of the ellipse are shown.

However, in some particular cases an exact computation is possible. This happens, for instance, if P and Q are spherically equivalent distributions (s.e.d.).

DEFINITION 4. P and Q are spherically equivalent distributions if they can be represented as rD and sD, where r and s are $[0, \infty)$-valued r.v.'s and D is a random vector, independent of r and s, supported by the sphere of radius 1.

The next proposition, which can be found in Cuesta-Albertos et al., 1993, leads to an expression for the o.c. between two s.e.d.

PROPOSITION 5. *Let P and Q be two s.e.d. and consider the representation given in Definition 4. Then (r, s) is an o.c. between their distributions if and only if (rD, sD) is an o.c. between P and Q.*

From this proposition we obtain the following corollary.

COROLLARY 6. *Under the same hypotheses and notation as in Proposition 5, we have that*

1. *If* (r,s) *is an o.c., then*
$$W^2(P,Q) = E(r^2) + E(s^2) - 2E(rF_s^{-1}(F_r(r))).$$

2. *If there exists* $c \in \mathbb{R}$ *such that* $r = c$ *a.s., then*
$$W^2(P,Q) = Var(s) + (c - E(s))^2.$$

Therefore, if we choose the pattern $\Im$ composed of all the s.e.d. to Q (if any) which are supported by a sphere, then the closest distribution to Q in $\Im$ is that one supported by the sphere with radius $\int \|y\| Q(dy)$.

In a similar scheme, if we consider the pattern $\Im = \{P$: uniform on a ball$\}$ and we assume that Q can be represented as rU where r and U are independent, r is $[0,\infty)$-valued and U is uniform on the unit sphere, then an elementary computation shows that the radius of the ball giving the closest distribution to Q is
$$\frac{d+2}{d} E\left[rF_r^{1/d}(r)\right],$$

Notice that this distribution is related to the concentration ellipsoid. However, the concentration ellipsoid depends only on second order moments of Q while this distribution depends on the whole distribution. Consequently it could be a better indicator of the shape of the distribution.

To end the paper we want to remark that an interesting possibility to look for the shape of a multidimensional distribution Q is to select the pattern composed by the uniform distributions on certain kind of convex sets supported by a subspace with a dimension less than d. Within this framework we have been able to obtain two results. The first one is related to the following representation theorem.

THEOREM 7. (Cuesta-Albertos and Matrán Bea, 1997) *Let* P, Q *be two probabilities on* $\mathbb{R}^d$ *with finite second order moment. Let us assume that* P *is absolutely continuous with respect to the Lebesgue measure. Then there exists a map* $H : \mathbb{R}^d \to \mathbb{R}^d$ *such that, if* X *is a r.v. with distribution* P*, then* $(X, H(X))$ *is an o.c. between* P *and* Q*.*

The proposition we have obtained is the following:

PROPOSITION 8. *Let* P, Q *be two probabilities on* $\mathbb{R}^d$ *with finite second order moment. Let us assume that* P *is absolutely continuous with* *respect to the Lebesgue measure and that Q is supported on a subspace with dimension $q < p$.*

Then the map H obtained in Theorem 7 can be chosen in such a way that, if $x, y \in \mathbb{R}^d$ have the same projection on the subspace supporting Q, then $H(x) = H(y)$.

Let us consider a basis in $\mathbb{R}^d$ such that the first q vectors generate the subspace supporting Q. We will denote by P_q the marginal distribution of P on this subspace.

Let $X = (X_1, ..., X_p)$ a random vector with distribution P. Notice that P_q is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^q$ and, by applying Theorem 7, we have that there exists a map H_q such that $[(X_1, ..., X_q), H_q(X_1, ..., X_q)]$ is an o.c. between P_q and Q.

Let $H : \mathbb{R}^d \to \mathbb{R}^d$ be the map defined by $H(x_1, ..., x_p) = H^q(x_1, ..., x_q)$. We have that

$$\begin{aligned}
W^2(P,Q) &= \inf_{\nu \in M(P,Q)} \int \sum_{i=1}^{p} (x_i - y_i)^2 \nu(dx, dy) \\
&= \sum_{i=q+1}^{p} E[X_i^2] + \inf_{\nu \in M(P_q,Q)} \int \sum_{i=1}^{q} (x_i - y_i)^2 \nu(dx, dy) \\
&= \sum_{i=q+1}^{p} E[X_i^2] + W^2(P_q, Q) \\
&= \sum_{i=q+1}^{p} E[X_i^2] + E\left[((X_1, ..., X_q) - H_q(X_1, ..., X_q))^2\right] \\
&= E\left[(X - H(X))^2\right] \\
&= W^2(P,Q).
\end{aligned}$$

Thus, $(X, H(X))$ is an o.c. between P and Q and the proposition is proved.

The second result refers to the following problem. Let us assume we are considering the pattern given by the uniform distributions supported by an ellipsoid contained in a q-dimensional subspace, $q < p$. It is quite reasonable to assume that, analogously to what happens in many situations like the principal components analysis, the supporting subspace is determined by the covariance matrix of Q and that it coincides with the subspace generated by the first q principal axis.

However, with the following counterexample we prove that this is, in general, false; thus giving rise to the unsolved question of the search for sufficient conditions for this property.

EXAMPLE 9. Let (X_1, X_2) be a random vector with uniform distribution on an ellipse. Let $\Sigma_{1,2} = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma \end{pmatrix}$ be its covariance matrix where we assume $\sigma_1 < \sigma$. Let X_3 a r.v. with distribution $N(0, \sigma^2)$ independent from (X_1, X_2) and let Q the distribution of the random vector (X_1, X_2, X_3).
Let us consider as pattern $\Im$ the family of uniform distributions which are supported by two-dimensional ellipsoids.
Let us denote by $\Im^{1,2}$ the family of uniform probabilities supported by an ellipse generated by the first two principal vectors. Then we have that

$$\inf_{P \in \Im^{1,2}} W^2(P, Q) = \sigma^2 + \inf_{P \in \Im^{1,2}} W^2(P, P_{(X_1, X_2)}) = \sigma^2. \tag{5}$$

However, let $\Im^{2,3}$ be the pattern composed by all uniform probabilities supported by an ellipse generated by the last two principal vectors and $\Im^2$ (resp. $\Im^3$) the pattern composed by the first (resp. the second) marginals of the distributions in $\Im^{2,3}$. If we denote by K the distance between the $N(0, 1)$-distribution and the pattern $\Im^3$, we have that

$$\begin{aligned} \inf_{P \in \Im^{2,3}} W^2(P, Q) &= \sigma_1^2 + \inf_{P \in \Im^{2,3}} W^2(P, P_{(X_2, X_3)}) \\ &\geq \sigma_1^2 + \inf_{P \in \Im^2} W^2(P, P_{X_2}) + \inf_{P \in \Im^3} W^2(P, P_{X_3}) \\ &= \sigma_1^2 + \sigma^2 K, \end{aligned}$$

where previous inequality comes from the fact (shown in Cuesta-Albertos et al., 1996) that the square of the L_2-Wasserstein distance between two probabilities is greater than or equal to the sum of the squares of the distances between their marginals.
Obviously $K > 0$ and, therefore, if we choose σ close enough to σ_1 we have that the last expression is greater than the one in (5) and, in consequence, the closest distribution in $\Im$ to Q is not supported by the subspace generated by the first two principal axis of P.

References

Aurenhammer, F., F. Hoffmann, and B. Aronov: 1998, 'Minkowski-type theorems and least-squares clustering'. *Algorithmica* **20**, 61–76.

Bickel, P. and D. Freedman: 1981, 'Some asymptotic theory for the bootstrap'. *Ann. Statist.* **9**, 1196–1217.

Cuesta-Albertos, J. A. and C. Matrán Bea: 1989, 'Notes on the Wasserstein metric in Hilbert spaces'. *Ann. Probab.* **17**, 1264–1276.

Cuesta-Albertos, J. A., C. Matrán Bea, and A. Tuero Díaz: 1996, 'On lower bounds for the L^2-Wasserstein metric in a Hilbert space'. *J. Theor. Probab.* **9**, 263–283.

Cuesta-Albertos, J. A., C. Matrán Bea, and A. Tuero Díaz: 1997, 'Optimal transportation plans and convergence in distribution'. *J. Multivariate Anal.* **60**, 72–83.

Cuesta-Albertos, J. A., L. Rüschendorf, and A. Tuero Díaz: 1993, 'Optimal coupling of multivariate distributions and stochastic processes'. *J. Multivariate Anal.* **46**, 355–361.

Rachev, S. T. and L. Rüschendorf: 1998, *Mass transportation problems.* Vol. I and II. Springer-Verlag, New York.

Tarpey, T. and B. Flury: 1996, 'Self-Consistency: A Fundamental Concept in Statistics'. *Statist. Sci.* **11**, 229–243.