JOURNAL OF OPTIMIZATION THEORY AND APPLICATIONS: Vol. 43, No. 1, MAY 1984

On the Optimal Mapping of Distributions

M. KNOTT[1] AND C. S. SMITH[2]

Communicated by D. Q. Mayne

Abstract. We consider the problem of mapping $X \to Y$, where X and Y have given distributions, so as to minimize the expected value of $|X-Y|^2$. This is equivalent to finding the joint distribution of the random variable (X, Y), with specified marginal distributions for X and Y, such that the expected value of $|X-Y|^2$ is minimized. We give a sufficient condition for the minimizing joint distribution and supply numerical results for two special cases.

Key Words. Inequalities, marginal distributions, Fréchet derivatives.

1. Introduction

If the random variable X has a continuous distribution function F and the random variable Y has distribution function G, what mapping, from values of X to values of Y will maximize the correlation of X and Y? It is well known that the solution is

$$Y = G^{-1}F(x),$$

where G^{-1} is defined as

$$G^{-1}(u) = \inf\{y: G(y) > u\}.$$

An equivalent problem is this. What is the bivariate distribution function H which has marginal distributions given by F and G and which has maximum correlation between its two variables X, Y? This problem was investigated by Fréchet, Hoeffding, and others (see Ref. 1).

[1] Senior Lecturer in Statistics, Department of Statistical and Mathematical Sciences, London School of Economics and Political Science, London, England.
[2] Lecturer in Statistics, Department of Statistical and Mathematical Sciences, London School of Economics and Political Science, London, England.

0022-3239/84/0500-0039$03.50/0

A typical result is that the correlation is maximized for

$$H(x, y) = \min(F(x), G(y)).$$

If the marginal distribution F has a density function, then H is the distribution function of $(X, G^{-1}F(X))$. The function $G^{-1}F$ is nondecreasing, so it maintains the ordering of the values of X.

There are many ways to prove such results. We shall in this paper be mainly using an approach derived from transportation algorithms. To maximize the correlation between X and Y is to minimize the average squared distance $E[(X-Y)^2]$, so the problem may be thought of as that of moving the mass distributed as F to that distribution given by G in such a way that the average squared distance moved is as small as possible. The connection with transportation or transhipment problems has long been realized; see Berge and Ghouila-Houri (Ref. 2, p. 155). One may note in passing that, if F and G each give probability $1/n$ to n points, there are strong links to work with rankings or rearrangements.

When F and G are defined on R^1, the problem is, as in the result above, completely solved. If instead we define F and G on R^2, and try to find a quadrivariate distribution minimizing the expected squared Euclidean distance $E|X-Y|^2$, there are very few known results. We shall discuss the more general problem when F and G are defined on R^m. The main difference of the case $m > 1$ from $m = 1$ is that there is no obvious ordering of the values of X and Y. In Section 2, we give our main results. Section 3 specializes to distributions uniform over regions of R^2, and Section 4 contains numerical work allied to Section 3.

2. Problem in *m* Dimensions

Let F, G be distribution functions for random variables X and Y which take values in R^m. What is the distribution function H for (X, Y), defined on R^{2m}, which minimizes the expected squared Euclidean distance $E|X-Y|^2$?

There is related work due to Monge (Ref. 3), Appell (Ref. 4), and others referred to by Appell, but they consider mainly $m=2$ or 3, take $E|X-Y|$ as the distance to minimize, and obtain necessary conditions on H. Appell (Ref. 5) obtains necessary conditions for a distance $Ef(|X-Y|)$ and $m=3$, but does not give the sufficient conditions of our main result.

None of these authors have attempted to give any method of finding H, except in the most special and trivial cases. In Sections 3 and 4, we present less trivial examples.

It is clear from the results of linear programming which apply to transportation algorithms that the distribution function H which we wish to find will concentrate its probability on a set of Lebesgue measure zero in R^{2m}. We shall consider only cases where H is the distribution function of $(X, \phi(X))$, $\phi(X)$ being a suitably smooth invertible mapping from R^m to R^m. We believe that this covers many cases of interest. When $m=1$, we have seen in Section 1 that H is often of this form. Our main theoretical result is given below. Treat $x, y, \phi(x), \psi(y) \in R^m$ as column vectors, where ψ is inverse to ϕ. We shall always assume that the elements of $\partial\phi(x)/\partial x'$ and $\partial\psi(y)/\partial y'$ are continuous on R^m. This implies that we assume ϕ is a diffeomorphism between R^m and R^m. In a fairly standard notation, we would say that ϕ and ψ are of class $C^{(2)}$.

Theorem 2.1. Let F and G be known distribution functions of random variables taking values on R^m. A sufficient condition that, among all distributions of (X, Y) on R^{2m} with $E|X-Y|^2$ finite, that of $(X, \phi(X))$ minimizes $E|X-Y|^2$ is that:

(i) $\phi(X)$ has distribution G;
(ii) $\partial\phi(x)/\partial x'$ is symmetric and positive semidefinite.

Proof. If $u \in R^m$, then the differential form

$$2du'(u-\phi(u)) = d\zeta,$$

is exact because $\partial\phi(x)/\partial x'$ is symmetric by condition (ii) and has continuous elements by assumption. So the line integral along a piecewise smooth path γ from 0 to x in R^m, given by

$$\lambda(x) = \int_\gamma d\zeta, \tag{1}$$

depends, as the notation suggests, only on x and not on γ. Similarly, for $v \in R^m$, the differential form

$$-2dv'(\psi(v)-v) = d\eta,$$

where

$$\psi(\phi(u)) = u, \qquad \text{for all } u \in R^m,$$

is exact. This is, as before, implied by condition (ii) and the well-known result

$$\partial\psi(v)/\partial v' = [\partial\phi(u)/\partial u']^{-1}|_{\phi(u)=v}.$$

So, the line integral along the piecewise smooth path δ from $\phi(0)$ to y in R^m, given by

$$\mu(y)=\int_\delta d\eta, \tag{2}$$

depends only on y and not on δ.

We shall define the constant

$$c=|0-\phi(0)|^2.$$

The line integral from $(0, \phi(0))$ to $(x, \phi(x))$ of the differential form $d\zeta+d\eta$ along the path in R^{2m} given by $(u, \phi(u))$, where u moves along path γ in R^m, is not dependent on path γ, and is $|x-\phi(x)|^2-c$. Thus, for $x\in R^m$,

$$\lambda(x)+\mu(\phi(x))=|x-\phi(x)|^2-c. \tag{3}$$

Also, for $(x, y)\in R^{2m}$,

$$\lambda(x)+\mu(y)=|\psi(y)-y|^2+\int_\nu d\zeta-c,$$

where ν may be taken as the straight line in R^m from $\psi(y)$ to x. Putting

$$u=\psi(y)+\alpha(x-\psi(y)),$$

we have

$$\begin{aligned}\lambda(x)+\mu(y)+c&=|\psi(y)-y|^2+2\int_0^1(x-\psi(y))'(u-\phi(u))\,d\alpha\\&=|x-y|^2+2\int_0^1(x-\psi(y))'[(u-\phi(u))-(u-y)]\,d\alpha\\&=|x-y|^2-2\int_0^1\int_0^1(x-\psi(y))'[\partial\phi(v)/\partial v'](x-\psi(y))\alpha\,d\beta\,d\alpha,\end{aligned}$$

where

$$v=\psi(y)+\beta(u-\psi(y)).$$

Since the quadratic form in the integral is nonnegative by condition (ii), we deduce that, for all $(x, y)\in R^{2m}$,

$$\lambda(x)+\mu(y)\leq|x-y|^2-c. \tag{4}$$

It will be noticed that $\lambda(x)$ and $\mu(y)$ are the *shadow costs* in the programming formulation of our problem. Using (3) and (4) leads to the theorem. If H_0 is any distribution on R^{2m} with the required marginal distributions

F and G, and if H is the distribution of $(X, \phi(X))$ where conditions (i) and (ii) are satisfied, then by (4) and (3),

$$\begin{aligned}
\underset{H_0}{E}|X-Y|^2 &\geq \underset{H_0}{E}[\lambda(X)+\mu(Y)]+c = \underset{H_0}{E}[\lambda(X)]+\underset{H_0}{E}[\mu(Y)]+c \\
&= \underset{H}{E}[\lambda(X)]+\underset{H}{E}[\mu(Y)]+c = \underset{H}{E}[\lambda(X)+\mu(Y)]+c \\
&= \underset{H}{E}[\lambda(X)+\mu(\phi(X)]+c = \underset{H}{E}|X-\phi(X)|^2 \\
&= \underset{H}{E}|X-Y|^2. \qquad \square
\end{aligned}$$

Comments

(a) Condition (ii) ensures that pairwise interchanges and local rotations cannot produce a better ϕ.

(b) If ϕ satisfies (i) and (ii) for F, G, then $\tilde{\phi}$ satisfies (i) and (ii) for $\tilde{F}$, G, where

$$\tilde{\phi}(x+a)=\phi(x)+a \quad \text{and} \quad \tilde{F}(x+a)=F(x),$$

for fixed $a \in R^m$ and all $x \in R^m$. A linear transformation ϕ will satisfy (ii) if and only if it is self-adjoint.

(c) The inequality (4) reduces to Young's inequality when $m=1$ and is a generalization of this inequality to many diemensions when there are suitable restrictions on the derivatives of ϕ.

(d) If $m=1$, the Hessian $\partial\phi(v)/\partial v'$ is automatically symmetric, and condition (ii) reduces to

$$d\phi(x)/dx \geq 0.$$

This fits with the known results, but adds nothing new.

(e) Condition (ii) of the theorem may be interpreted as requiring that ϕ is the gradient of a convex function on R^m.

(f) If F and G have density functions f, g, then, when $m=2$,

$$r(x)=(x_1^2+x_2^2-\lambda(x))$$

satisfies a Monge–Ampère differential equation,

$$(\partial^2 r/\partial x_1^2)(\partial^2 r/\partial x_2^2)-(\partial^2 r/\partial x_1\,\partial x_2)^2 = f(x)/g(\partial r/\partial x_1, \partial r/\partial x_2),$$

obtained by writing the Jacobian of the mapping ϕ and using conditions (i) and (ii) of the theorem. Such equations are hard to solve, but led us to the results of Section 3. We are indebted to Professor Pirani of King's

College, London, for referring us to literature on the Monge–Ampère equation.

Simple Example. Suppose that

$$X \sim MN(0, \Sigma_1) \text{ and } Y \sim MN(0, \Sigma_2).$$

Then, the transformation ϕ giving the smallest expected squared distance is

$$\phi(X) = \Sigma_2^{1/2}(\Sigma_2^{1/2}\Sigma_1\Sigma_2^{1/2})^{-1/2}\Sigma_2 X, \tag{5}$$

where the square roots are all taken positive semidefinite. The maximum possible value for $EX'Y$ is trace $(\Sigma_1 \Sigma_2)$. Notice that the special case $\Sigma_1 = I$ is particularly easy.

3. Case $m = 2$, with F and G Giving a Uniform Distribution over Bounded Sets A, B

In the case $m = 2$, when X, Y are distributed uniformly over bounded sets A and B, respectively, we can simplify the problem of finding an optimal ϕ by using complex variables. Suppose that $\psi(w, z)$ is a polynomial in the complex variables w and z, and put

$$\psi_1 = \partial\psi/\partial w \quad \text{and} \quad \psi_2 = \partial\psi/\partial z.$$

Then,

$$\psi(\sqrt{q}x_1 + i\phi_2, \phi_1 + i\sqrt{q}x_2) = 0, \tag{6}$$

for suitable ψ, defines a mapping

$$(x_1, x_2) \to (\phi_1, \phi_2),$$

where we take q as the ratio of the area of set B to that of set A. Differentiation w.r.t. x_1 and x_2 leads to

$$\psi_1(\sqrt{q} + i\partial\phi_2/\partial x_1) + \psi_2(\partial\phi_1/\partial x_1] = 0,$$

$$\psi_1(i\partial\phi_2/\partial x_2) + \psi_2(\partial\phi_1/\partial x_2 + i\sqrt{q}) = 0.$$

Assuming $\psi_1\psi_2 \neq 0$ gives

$$\partial\phi_1/\partial x_2 = \partial\phi_2/\partial x_1 \quad \text{and} \quad (\partial\phi_1/\partial x_1)(\partial\phi_2/\partial x_2) - (\partial\phi_1/\partial x_2)^2 = q.$$

So, if we find a suitable ψ which implies that ϕ maps A to B, then the requirements of our theorem are satisfied.

As a simple example, consider

$$\psi(w, z) = z - w^2.$$

Solving gives

$$\phi_1 = (x_1 + x_2^2), \qquad \phi_2 = 2x_2(x_1 + x_2^2).$$

This ϕ will give an optimal mapping from

$$A = \{x: x_1 \geq 0, x_1 + x_2^2 \leq 2\}$$

to

$$B = \{\phi: 0 \leq \phi_1 \leq c, |\phi_2| \leq 2\phi_1^2\},$$

where $c > 0$. For $x_1 < 0$, ϕ may be defined in any convenient way. For instance, ϕ might be taken as the gradient of the convex function

$$\lambda(x) = \begin{cases} 2/3(x_1 + x_2^2)^{3/2}, & x_1 + x_2^2 \geq 0, \\ 0, & x_1 + x_2^2 < 0, \end{cases}$$

confirming comment (d) to the theorem.

In the next section, we use (6) to obtain approximations to ϕ for more interesting shapes of A and B. The idea is to approximate ψ by a polynomial, and to make use of the fact that, for ϕ satisfying the conditions of our theorem, the boundary of A will map to the boundary of B.

4. Numerical Investigations

Suppose that we are given sets A, B in R^2, and we want to find $\psi(w, z)$, a polynomial of degree sufficiently high, such that, on solving (6), the resulting ϕ maps A to B at least to a good approximation. If ψ is chosen so that ϕ maps the boundary of A to the boundary of B, then the same ϕ will optimal throughout the interior of A. To find suitable polynomials, guess a correspondence between K fixed points on the boundary of A and K points on the boundary of B, to give K pairs (w_i, z_i). A simple linear least-square algorithm allows the determination of the coefficients of the polynomial so that $\sum_{i=1}^{k} |\psi(w_i, z_i)|^2$ is minimized. Now, use the polynomial obtained to recalculate K points close to the boundary of B corresponding to the fixed points on the boundary of A. Each point will require the solution of a polynomial equation. Use the recalculated points to obtain K new points close to them exactly on the boundary of B. Repeat the whole process until (6) produces a ϕ which satisfactorily maps the boundary of A to the boundary of B.

Though this procedure seems rather a roundabout approach, it gives better results than using transportation algorithms on a discrete approximation.

Example 4.1

$$A = \{x: x_1 \geq 0, x_2 \geq 0, x_1 + x_2 \leq \sqrt{2}\},$$

$$B = \{y: 0 \leq y_1 \leq 1, 0 \leq y_2 \leq 1\}.$$

Here, A and B are equal in area.

Taking account of the symmetries of the sets A, B, a polynomial approximation

$$\psi(w, z) = w - z - \sum_{1}^{N_0} a_r (w+z)^{4r-1},$$

where the a_i's are real, is likely to be adequate. We started with 11 points

$$x_{1j} = 1/\sqrt{2}(1 + j/10), \quad x_{2j} = \sqrt{2} - x_{1j}, \qquad j = 0(1)10.$$

We guessed $y_{1j} = 1$ and intuitively reasonable values for y_{2j}. The polynomial

$$\psi(w, z) = w - z - 0.028(w+z)^3 - 0.000001(w+z)^{11j}$$

was found on iteration to give good results. We used a VDU screen to display results and to guide the choice of new points on the boundary at each stage. A few results are given in Table 1, and a picture of the mapping is given in Figs. 1a and 1b.

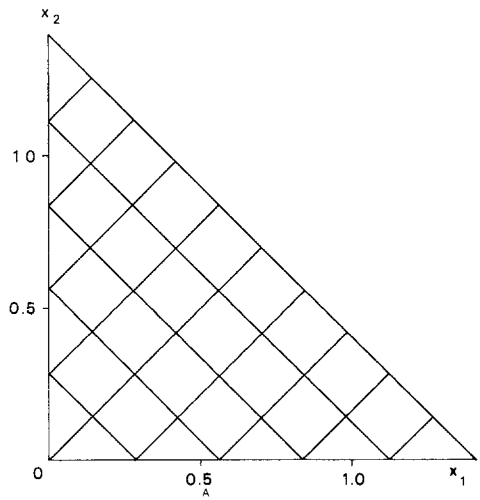

Fig. 1a. Initial grid for Example 4.1.

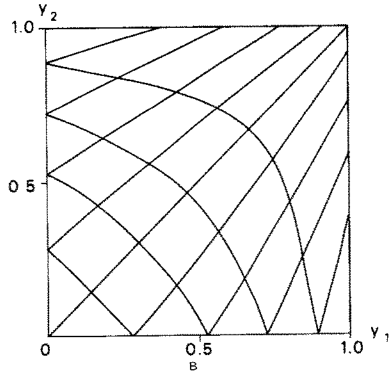

Fig. 1b. Final grid for Example 4.1 after optimal mapping.

Table 1. Values for approximately optimal mapping in Example 4.1.

x_1	x_2	ϕ_1	ϕ_2
0.000	0.000	0.000	0.000
0.142	0.142	0.142	0.142
0.283	0.283	0.294	0.294
0.424	0.424	0.464	0.464
0.566	0.566	0.672	0.672
0.707	0.707	0.992	0.992
0.283	0.000	0.278	0.000
0.424	0.142	0.414	0.158
0.566	0.283	0.563	0.343
0.707	0.424	0.743	0.573
0.849	0.566	1.005	0.914
0.566	0.000	0.529	0.000
0.707	0.142	0.650	0.192
0.849	0.283	0.795	0.434
0.990	0.424	0.997	0.768
0.849	0.000	0.737	0.000
0.990	0.142	0.842	0.251
1.131	0.283	1.000	0.589
1.131	0.000	0.896	1.002
1.273	0.142	1.001	0.365
1.414	0.000	1.003	0.002

Example 4.2

$A = \{x\colon 0 \leq x_1 \leq 1, 0 \leq x_2 \leq 1\}$,

$B = \{y\colon y_1 \geq 0, y_2 \geq 0, y_1^2 + y_2^2 \leq 4/\pi\}$.

A and B have equal area. A good approximation is

$$\psi(w, z) = (z - w) - 0.146(w + z)^3 + 0.00004(w + z)^7.$$

Table 2 gives a few of the mappings for points on the boundary. A picture of the mapping is given in Figs. 2a and 2b.

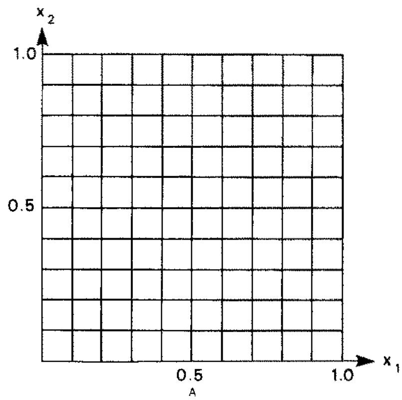

Fig. 2a. Initial grid for Example 4.2.

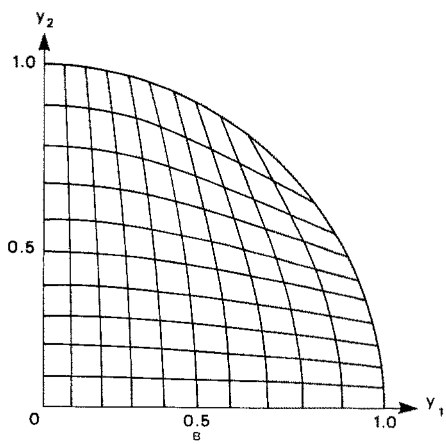

Fig. 2b. Final grid for Example 4.2 after optimal mapping.

Table 2. Values for approximately optimal mapping in Example 4.2.

x_1	x_2	ϕ_1	ϕ_2
0.1	1	0.071	1.134
0.2	1	0.140	1.129
0.3	1	0.211	1.116
0.4	1	0.282	1.100
0.5	1	0.353	1.077
0.6	1	0.426	1.046
0.7	1	0.504	1.003
0.8	1	0.590	0.951
0.9	1	0.688	0.886
1.0	1	0.803	0.810

References

1. TCHEN, A. H., *Inequalities for Distributions with Given Marginals*, Annals of Probability, Vol. 8, pp. 814–827, 1980.
2. BERGE, C., and GHOUILA-HOURI, A., *Programming, Games, and Transportation Networks*, Methuen and Company, London, England, 1965.
3. MONGE, G., *Deblai et Remblai*, Mémoires de l'Académie des Sciences, 1781.
4. APPELL, P., *Le Problème Géométrique des Déblais et Remblais*, Mémorial des Sciences Mathématiques, Vol. 27, pp. 1–34, 1928.
5. APPELL, P. *Sur un Théorème de Monge et sur une Généralization de Ce Théorème*, Acta Mathematica, Vol. 47, pp. 7–14, 1926.