Proc. Estonian Acad. Sci. Phys. Math., 1996, **45**, 2/3, 250–258
https://doi.org/10.3176/phys.math.1996.2/3.16

ON THE RESTORATION OF DISTRIBUTION FUNCTIONS

Jüri LEMBER

Tartu Ülikooli matemaatilise statistika instituut (Institute of Mathematical Statistics, University of Tartu), Liivi 2, EE-2400 Tartu, Eesti (Estonia)

Received 14 December 1995, accepted 4 June 1996

Abstract. The k-centres of a random variable X are defined as the points minimizing the (quadratic) loss function. In the paper an equivalent clustering criterion, the R-function, is defined. Under some assumptions the R-function of X uniquely determines the distribution of X and can be regarded as a characterization of the distribution. The possibilities of restoring the distribution function of X by the given R-function of X are discussed.

Key words: quadratic loss function, optimal partitions, R-function, distribution function.

1. PRELIMINARIES

Let X be a m-variate random vector with the law P and the distribution function F. Whenever needed, it will be implicitly assumed that the second moment of X is finite. Let $S = \{S_1, \ldots, S_k\}$ be a Borel-measurable k-partition of $\Re^m$, i.e. $\bigcup_{i=1}^k S_i = \Re^m$, $S_i \bigcap S_j = \emptyset$, $\forall i \neq j$.

Let $a(S_i)$ be the conditional expectation of X on S_i, i.e. $a(S_i) = E(X|S_i)$. The ordered set $A(S) = \{a(S_1), \ldots, a(S_k)\}$ is called *the conditional means set of X given S*.

Definition 1.1. *The function*

$$W_k(S, F) = \sum_{i=1}^{k} \int_{S_i} \|x - a(S_i)\|^2 \, dF$$

is called the loss function of the distribution F by the partition S.

Using the definition of conditional expectation, it is easy to check the equality (see, e.g. [1])

$$W_k(S, F) = E\|X\|^2 - R_k(S, F), \tag{1.1}$$

where

$$R_k(S, F) = \sum_{i=1}^{k} \|a(S_i)\|^2 P(S_i) \tag{1.2}$$

is the second moment of the distribution which is concentrated on the set $A(S)$ with the probability $P(S_i)$ at $a(S_i)$ (see [2]).

Definition 1.2. *The function* (1.2) *is called the R-function of the distribution F by the partition S.*

Definition 1.3. *A k-partition* $S^* = \{S_1^*, \ldots, S_k^*\}$ *is called optimal for the distribution F if it minimizes* $W_k(S, F)$ *over all k-partitions of* $\Re^m$.

From (1.1) it follows that the optimal partition maximizes the R-function and for distributions with a finite second moment the minimization of the loss function is equivalent to the maximization of the R-function (see also [3]).

2. THE SPACE $\Re^1$

Definition 2.1. *A k-partition* $S = \{S_1, \cdots, S_k\}$ *is called convex if all regions in S are convex.*

In the following only the convex partitions in the space $\Re^1$ are considered. Since the value of the R-function by an arbitrary k-partition is not greater than the value of the R-function for a suitably chosen convex partition, such assumption is not restrictive. In the space $\Re^1$ every convex k-partition is obtained by using exactly $k-1$ cutting points $t_1 < t_2 <, \cdots, < t_{k-1}$. Therefore in the space $\Re^1$ (1.2) reduces to

$$R_k(t_1, \cdots, t_{k-1}) := R_k(S, F)$$

$$= \sum_{i=1}^{k} a^2((t_{i-1}, t_i])(F(t_i) - F(t_{i-1})), \tag{2.1}$$

where $t_1, \cdots, t_{k-1}$ are the cutting points to get S, and $t_0 := -\infty$, $t_k := \infty$. Obviously, the convergence $\lim_{t_k \to \infty} R_{k+1}(t_1, \cdots, t_k) = R_k(t_1, \cdots, t_{k-1})$ holds, and, in case $k = 1$, $\lim_{t \to \infty} R_2(t) = (EX)^2$.

Let us define the b-function of F by $b(t, F) = \int_{-\infty}^{t} x dF(x)$. If F has a finite expectation, the b-function can be represented in the following way (the argument F will be usually dropped):

$$b(t) = tF(t) - S(t), \tag{2.2}$$

where $S(t) = \int_{-\infty}^{t} F(x)\, dx < \infty$.

Obviously, the *b-function of* F is continuous iff F is continuous and the distribution function F is uniquely determined by the b-function.

Using b-functions, we get for (2.1) the form

$$R_k(t_1, \cdots, t_{k-1}) = \sum_{i=1}^{k} \frac{(b(t_i) - b(t_{i-1}))^2}{F(t_i) - F(t_{i-1})}, \tag{2.3}$$

where $b(t_0) := 0$, $b(t_k) := EX$, $F(t_0) := 0$, $F(t_k) := 1$, and $(b(t_i) - b(t_{i-1}))^2/(F(t_i) - F(t_{i-1})) := 0$ if $F(t_i) - F(t_{i-1}) = 0$.

Note that if F has a finite second moment and mean μ, the convergences

$$\lim_{t\to\infty} \frac{(\mu - b(t))^2}{1 - F(t)} = \lim_{t\to-\infty} \frac{b^2(t)}{F(t)} = 0 \tag{2.4}$$

take place. If F is continuous, the b-function of F and thus the R-function of F can be regarded as functions depending on the values of F. For this end a change of variables is needed, namely

$$b(x) = \int_0^x F^{-1}(u)du, \tag{2.5}$$

where $0 \le x \le 1$ and $F^{-1}(u) = \inf\{t|F(t) \ge u\}$. (Of course, the inverse of F can be defined in some other way, but because the set where F^{-1} is not unique has the P-measure zero, all different definitions of F^{-1} yield the same function $b(x)$).

Regarding the R-function as a function depending on the values of F, we get from (2.3)

$$R_k(x_1, \cdots, x_{k-1}) = \sum_{i=1}^{k} \frac{b(x_i) - b(x_{i-1})}{x_i - x_{i-1}}, \tag{2.6}$$

where $x_0 := 0$, $x_k := 1$, $b(x_0) := 0$, $b(x_k) := EX$.

Given EX, a continuous distribution function is uniquely determined by its R-function, i.e. the following proposition holds.

Proposition 2.1. *Let $X \sim F$, $Y \sim G$ be the continuous random variables such that $EX = EY = \mu$, and $R_k(t_1, \cdots, t_{k-1}, F) = R_k(t_1, \cdots, t_{k-1}, G) = R_k(t_1, \cdots, t_{k-1})$. Then $F = G$.*

Proof. Without loss of generality assume $\mu = 0$. Consider first the case $k = 2$. Let $R_2(t) = R_2(t, F) = R_2(t, G)$ be the common R-function. From (2.2) and (2.3)

$$R_2(t) = \frac{b^2(t, F)}{F(t)(1 - F(t))} = \frac{(tF(t) - S(t))^2}{F(t)(1 - F(t))}$$
$$= \frac{b^2(t, G)}{G(t)(1 - G(t))} = \frac{(tG(t) - T(t))^2}{G(t)(1 - G(t))},$$

where $T(t) = \int_{-\infty}^{t} G(x)dx$. Because $EX = EY$, there exists a point t_0 such that $F(t_0) = G(t_0)$. At any point t the equality $F(t) = G(t)$ implies $S(t) = T(t)$, therefore at the point t_0 $S(t_0) = T(t_0)$. Since F and G are continuous, the functions S and T can be equal at the point t_0 only if there

exists at least one point $t_1 < t_0$ such that $F(t_1) = G(t_1)$ and $S(t_1) = T(t_1)$. Now the continuity of the distribution function implies $F(t) = G(t), t \in [t_1, t_0]$. Similarly we get $F(t) = G(t), t \in (-\infty, t_0]$. The repeating of the argument presented above in the set (t_0, ∞) completes the proof in the case $k = 2$.

For the general case just note that $R_k(\cdot, F) = R_k(\cdot, G)$ implies $R_{k-1}(\cdot, F) = R_{k-1}(\cdot, G)$ and the proof is complete. □

If the arguments of the R-function are the probabilities, i.e. $R = R_k(x_1, \cdots, x_{k-1})$, the proof of Proposition 2.1 is even simpler. Indeed, for the case $k = 2$ we have

$$R_2(x, F) = \frac{b^2(x, F)}{x(1-x)} = \frac{b^2(x, G)}{x(1-x)} = R_2(x, G)$$

and the uniqueness follows from the definition of $b(x)$.

Proposition 2.1 does not hold for discontinuous distribution functions. Many different discrete distributions with the same R-function can be constructed. As a counterexample consider the following discrete distributions:

-20	-20/6	-3.884	0	5.275	20
.1	.1	.1	.4	.2	.1

-20	-5.275	0	3.884	20/6	20
.1	.2	.4	.1	.1	.1

In the case $k = 2$ both distributions have the same R-function which is equal to $400/9$ if $-20 < t < 20$ and zero elsewhere.

3. RESTORING A CONTINUOUS DISTRIBUTION BY THE R-FUNCTION

From Proposition 2.1 it follows that, given expectation, the R-function determines the corresponding continuous distribution uniquely (provided that such a distribution exists). Therefore the R-function can be regarded as a characterization of a continuous distribution. Now the question of finding the distribution function of a given R-function (or loss function) arises.

Suppose we are given the R-function $R_k(\cdot)$. We are going to find the corresponding distribution function. Since for every k the function $R_k(\cdot)$ determines also the function $R_2(\cdot)$, we are able to restore the distribution function if we can do it for the case $k = 2$. Therefore only the latter case as the most important and general one is considered.

Let $R(\cdot) = R_2(\cdot)$ denote the given R-function, let μ be the given expectation. We are looking for a distribution function F such that $R_2(t, F) = R(t)$, and $EX = \mu$, where $X \sim F$. Note that we cannot skip the fixed expectation assumption, because different distributions having different means can have the same R-function.

In the paper it is implicitly assumed that every R-function has the corresponding distribution function, i.e. there exists a distribution function F such that $R(t) = R_2(t, F)$.

Without loss of generality we may assume $\mu = 0$. From (2.3) it follows that the function F is the solution of the equality

$$\frac{-b(t,F)}{\varphi(F(t))} = \sqrt{R(t)}, \tag{3.1}$$

where $\varphi(F(t)) = \sqrt{F(t)(1-F(t))}$. Note that $\mu = 0$ implies $b(t) \leq 0$ for every t. Using (2.2), we get the following differential equation

$$S(t) = tF(t) + \varphi(F(t))\sqrt{R(t)}$$

or, equivalently,

$$S = tS' + \sqrt{R(t)}\varphi(S'). \tag{3.2}$$

There is no direct way to solve (3.2). The parametrizing $t = uS' = vS = uv + \varphi(v)\sqrt{(R(u)}$ yields

$$\sqrt{v(1-v)}\frac{R'(u)}{2\sqrt{R(u)}}du + \left[u + \sqrt{R(u)}\frac{1-2v}{2\sqrt{v(1-v)}}\right]dv = 0. \tag{3.3}$$

Equation (3.3) has a symmetric form and, depending on $R(t)$, may be explicitly solvable. The desired distribution function is the solution in the form $v = v(u)$.

If the R-function depends on t (as in the case we are considering now), the change of variables (2.5) does not facilitate our task. Indeed, from (2.5) $t = b'(x)$ (not necessarily continuous) and Eq. (3.2) takes the form

$$b = -\sqrt{R(b')x(1-x)}. \tag{3.4}$$

Now the parametrization $b' = u$, $x = v$, $b = -\sqrt{R(u)v(1-v)}$ yields again the equality (3.3) and the solution in the form $v = v(u)$ is the desired distribution function.

Since (3.4) implies

$$x = \frac{1}{2}\left(1 \pm \sqrt{1 - 4\frac{b^2}{R(b')}}\right), \tag{3.5}$$

the parametrizing $b = u$, $b' = v$ yields the symmetric differential equation

$$\left(\pm 1 + \frac{2uv}{\sqrt{R^2(v) - 4u^2R(v)}}\right)du = \frac{u^2vR'(v)}{R(v)\sqrt{R^2(v) - 4u^2R(v)}}dv. \tag{3.6}$$

Equation (3.6) is symmetric and may be explicitly solvable. The desired distribution function can be directly obtained by substituting the solution $u = u(v)$ into (3.5) (the particular solution must be chosen to ensure the

required properties of the distribution function).

$$x = \frac{1}{2}\left(1 \pm \sqrt{1 - 4\frac{u(v)^2}{R(v)}}\right) = \frac{1}{2}\left(1 \pm \sqrt{1 - 4\frac{(u(t))^2}{R(t)}}\right) = F(t).$$

Let us consider the case where the argument of the R-function is the probability, i.e. given R-function has the form $R(x) = R_2(x)$. Given such R-function, determination of the distribution function is relatively easy. Indeed, Eq. (3.2) yields the Clairaut' differential equation $S - S't = \sqrt{R(S')}\varphi(S') = -\psi(S')$, and after the parametrization $v = S'$, $u = t$, we have $[u - \psi'(v)]dv = 0$. Now $t = \psi'(S') = \psi'(F(t))$; from this equation the distribution function can be obtained.

Obviously, in the case $R_2(x)$ the change of variables (2.5) is most useful. From (2.6) we get

$$\frac{b^2(x)}{x(1-x)} = R(x), \tag{3.7}$$

implying $b = -\sqrt{R(x)(1-x)x} = \psi(x)$. The differentiation $t = b' = \psi'(x)$ yields the equation $t = \psi'(F(t))$, from which the distribution function can be obtained.

All the described possibilities of determining the distribution function are easy to use in the case when R is constant. For example, by (3.7) $b(x) = \psi(x) = -\sqrt{Rx(1-x)}$ and the equation for the distribution function is therefore $b'(x) = \sqrt{R}(2x-1)/2\sqrt{x(1-x)}$.

The solution of the obtained equation, the distribution with a constant R-function is

$$x = \frac{1}{2}\left(1 + \frac{b'}{\sqrt{R + (b')^2}}\right) = \frac{1}{2}\left(1 + \frac{t}{\sqrt{R + t^2}}\right) = F(t). \tag{3.8}$$

By Proposition 2.1 the distribution (3.8) is the only continuous distribution having the R-function equal to the constant R and mean zero. Obviously, the second moment of (3.8) is not finite, otherwise the convergences (2.4) would lead to a contradiction with the assumption $R_2(t) = R$ for every t. In other words: there is no distribution with a constant loss function (1.1) (this is obvious by intuition), but there do exist distributions with a constant (and finite) R-function having an infinite loss function. For such distributions with mean zero the convergences (2.4) are replaced by the convergences

$$\lim_{t\to\infty} \frac{b^2(t)}{1 - F(t)} = \lim_{t\to-\infty} \frac{b^2(t)}{F(t)} = R.$$

4. RESTORING A DISCRETE DISTRIBUTION BY R-FUNCTION

The R-function and the mean do not determine a discrete distribution P uniquely. Therefore, in order to determine a discrete distribution by its R-function, additional information is needed.

Every discrete distribution can be considered as consisting of two parts: a set of atoms and the corresponding masses (a discrete distribution P can be presented as the sum of Dirac' measures $P = \sum p_i \delta_{t_i}$; the set of points $\{t_i\}$ designates atoms, the set of the corresponding probabilities $\{p_i\}$ denotes the masses). A discrete distribution is uniquely determined, if besides the R-function and the mean one of the described components is given.

In the following some rules for calculating the discrete distribution, given the R-function and one of the mentioned components, are discussed. As before we consider the case $k = 2$ and $\mu = 0$.

Let $R = R_2(\cdot)$ be the given R-function. Suppose at first that atoms $\{t_i\}$ are given. It means that the R-function is given in the form $R = R_2(t)$. Since the R-function is constant between atoms, only the values of the R-function $R_i := R(t_i)$ are important. We are looking for a set of probabilities $\{p_i\}$ or, equivalently, for the distribution function $F(t)$. Since $b(t_{i+1}) = b(t_i) + p_{i+1}t_{i+1}$, (3.1) implies

$$\sqrt{R_{i+1}}\varphi(F_{i+1}) = \sqrt{R_i}\varphi(F_i) - p_{i+1}t_{i+1}, \tag{4.1}$$

where $F_{i+1} = F(t_{i+1})$. Denoting $F_{i+1} - F_i = \Delta F_i$ and $\Delta t_i = t_{i+1} - t_i$, (4.1) we get

$$\sqrt{R_{i+1}}\varphi(F_{i+1}) - \sqrt{R_i}\varphi(F_i) + t_i \Delta F_i = -\Delta F_i \Delta t_i,$$

implying

$$-p_{i+1} = -\Delta F_i = \frac{\Delta\sqrt{R(t_{i+1})}\varphi(F_{i+1})}{\Delta t_i} + \frac{t_i \Delta F_i}{\Delta t_i}, \tag{4.2}$$

where $\Delta\sqrt{R(t_{i+1})}\varphi(F_{i+1}) = \sqrt{R_{i+1}}\varphi(F_{i+1}) - \sqrt{R_i}\varphi(F_i)$. Equation (4.2) is the discrete counterpart of (3.3). If F is continuous, (4.2) yields (3.3) in the process $\Delta t_i \to 0$.

As in the continuous case, there is no direct way for solving (4.2), and therefore also (4.1). However, if F_i is known, from (4.1) F_{i+1} can be calculated. It means that the value of p_{i+1} can be calculated if all the previous masses up to the p_i are known. That suggests the use of the step-by-step calculation starting from the first atom t_1. Of course, the latter is possible only for the distributions with a finitely large number of atoms.

Assume now that the value F_i is known. Denoting $F_{i+1} = x$, $t_{i+1} = t$ and $B = \sqrt{R_i}\varphi(F_i) + t_{i+1}F_i$, from (4.1) we obtain

$$(R_{i+1} + t^2)x^2 - (R_{i+1} + 2Bt)x + B^2 = 0. \tag{4.3}$$

After solving (4.3) we get the rule for calculating F_{i+1}:

$$F_{i+1} = \frac{R_{i+1} + 2Bt \pm \sqrt{R_{i+1}^2 + 4R_{i+1}B(t-B)}}{2(R_{i+1} + t^2)}, \tag{4.4}$$

where the minus sign must be used if $R_{i+1} - R_i < 0$.

Thus we have found an algorithm for determining the distribution function by its R-function, if the R-function is in the form $R = R_k(t)$ and the distribution has only finitely many atoms.

Suppose now that we want to restore a discrete distribution by its R-function and by the set of masses $\{p_i\}$. It means that the R-function is in the form $R = R_k(x)$. As usual the case $k = 2$ is considered.

From (4.1) and (4.2)

$$-t_{i+1} = \frac{\sqrt{R_{i+1}}\varphi(F_{i+1}) - \sqrt{R_i}\varphi(F_i)}{F_{i+1} - F_i} = \frac{\Delta\sqrt{R_{i+1}}\varphi(F_{i+1})}{\Delta F_i}. \tag{4.5}$$

Since all components on the right-hand side of (4.5) are known, the atoms and therefore the distribution can be completely reconstructed. Note that in the continuous case (4.5) yields the equation $t = \psi'(F(t))$, where $\psi(x) = -\sqrt{R(x)x(1-x)}$.

Using (4.5) we are able to determine the distribution function by its R-function in the form $R = R_k(x)$. Note that the capacity of the set of atoms is of no importance in the present case.

Example 4.1. The algorithm (4.5) can be used to generate a sample having a constant R-function (and therefore a constant loss function) between the smallest and the largest observation. Let the given masses be $p_i = n^{-1}$, where n is the sample size; the given R-function have the form $R_i = R$, $i = 1, \cdots, n$, where the constant R is known. If $n = 20$ and $R = 1$, such a sample is the following:

+4.358	+1.641	+1.141	+0.858	+0.660	+0.504	+0.374	+0.258	+0.152	+0.050
−0.050	−0.152	−0.285	−0.374	−0.504	−0.660	−0.858	−1.141	−1.641	−4.358

Since the loss function is constant between the largest and the smallest sample point, every such convex partition, which separates at least one of the points from the others, has the same value of the loss function and is therefore optimal.

REFERENCES

1. Pärna, K. On the stability of k-means clustering in metric spaces. *Tartu Riikl. Ülik. Toimetised*, 1988, **798**, 19–36.
2. Pärna, K. and Lember, J. On some properties of k-means. In *Proceedings of Fifth Taru Conference on Multivariate Analysis* (Tiit, E.-M., Kollo, T., and Niemi, H., eds.). VSP/TEV, Vilnius, 1994, 267–278.
3. Flury, B. A. Principal points. *Biometrika*, **77**, 33–41.

JAOTUSFUNKTSIOONIDE RESTAUREERIMISEST

Jüri LEMBER

Reaaltelje k-tükeldust nimetatakse antud tõenäosusjaotuse jaoks optimaalseks, kui ta minimiseerib teatava kriteeriumfunktsiooni – ruutkaofunktsiooni. Käesolevas töös on uuritud ruutkaofunktsiooniga mõnes mõttes ekvivalentset kriteeriumi – R-funktsiooni. Kindlatel tingimustel võib R-funktsiooni vaadelda tõenäosusjaotuse karakteristikuna. Seetõttu pakub huvi tõenäosusjaotuse rekonstrueerimine etteantud R-funktsiooni (ruutkaofunktsiooni) põhjal. On käsitletud jaotusfunktsiooni rekonstrueerimise võimalusi nii pidevate kui ka diskreetsete jaotuste korral.