Metrika (2002) 56: 97–111

Metrika
© Springer-Verlag 2002

Least squares fitting of an affine function and strength of association for interval-valued data[1]

María Angeles Gil, María Asunción Lubiano, Manuel Montenegro, María Teresa López

Dpto. de Estadística, I.O. y D.M., Universidad de Oviedo, C/Calvo Sotelo, s/n, 33007 Oviedo, Spain (e-mail: angeles@pinon.ccu.uniovi.es)

Abstract. The ultimate goal of this paper is to determine a measure of the degree of dependence between two interval-valued random sets, when the dependence is intended in the sense of an affine function relating these random elements. For this purpose, a general study on the least squares fitting of an affine function for interval-valued data is first carried out, where the least squares method we will present considers that squared residuals are based on a generalized metric on the space of nonempty compact intervals, and output and input random mechanisms are modelled by means of convex compact random sets. For the general case of nondegenerate convex compact random sets, solutions are presented in an algorithmic way, and the few cases leading to nonunique solutions are characterized. On the basis of this regression study we later introduce and analyze a well-defined determination coefficient of two interval-valued random sets, which will allow us to quantify the strength of association between them, and an algorithm for the computation of the coefficient has been also designed. Finally, a real-life example illustrates the study developed in the paper.

Key words: Aumann's integral; convex compact random set; linear correlation; linear regression; *W*-metric between compact intervals

1 Introduction

In many experimental problems, the variation in the measurements of a random element is caused to a certain extent by other related random element.

[1] The research in this paper has been partially supported by DGESIC Grants DGE-99-PB98-1534 and DGE-98-PB97-1282 of the Spanish MEC. Their financial support is gratefully acknowledged.

An interesting point for discussion is that of assessing the nature and strength of the relationship between these two elements, and then use it to describe or predict the random element of primary interest from observations on the other one. Descriptive and inferential regression and correlation analyses focussed on such a discussion have been widely studied in the statistical literature.

When random elements correspond to random sets, regression and correlation analyses become more complex due to the interpretation of the relationships and to the difficulties in finding optimal solutions and measures of their adequacy.

Diamond (see [4]) has studied the case in which the values the random elements take on are nonempty compact intervals, and he has established a sufficient condition for nondegenerate elements to admit a unique optimal solution when an affine function is employed to model the relationship between them, the optimality criterion being intended as an extension of the least squares one. Interval-valued data arise in some real-life situations from different areas, especially in Economics, Medicine, etc., where random elements quantifying (economical, time, electric, and so on) ranges are frequently considered (see, for instance, [4], [21], [12]).

The first analysis in this paper generalizes Diamond's study in two ways, namely, by extending the least squares method in terms of a generalized metric on the space of nonempty compact intervals, and by finding all the optimal solutions for the general case of nondegenerate interval-valued random sets with the statement of the necessary and sufficient conditions for the nonuniqueness. On the basis of the conclusions from this general regression problem, we will later introduce a well-defined extended determination coefficient for the affine connection between the two random sets.

Computations for the optimal solutions and the extended determination coefficient will be presented in an algorithmic way, and the application of these algorithms is illustrated by means of a real-life example.

Finally, some future directions related to the study developed in this paper are commented.

2 Preliminaries

Random elements in this paper will be assumed to take on interval values, and they will be modelled by means of certain convex compact random sets.

Let $\mathscr{K}_c(\mathbb{R})$ be the class of nonempty compact intervals. $\mathscr{K}_c(\mathbb{R})$ can be endowed with a semilinear structure induced by the product by a scalar and the Minkowski addition. The well-known Hausdorff metric can be defined on $\mathscr{K}_c(\mathbb{R})$, leading in this case to the simple expression $d_H(A,B) = \max\{|\sup A - \sup B|, |\inf A - \inf B|\}$ for $A, B \in \mathscr{K}_c(\mathbb{R})$.

Given a probability space $(\Omega, \mathscr{A}, P)$, a mapping $X : \Omega \to \mathscr{K}_c(\mathbb{R})$ being $(\mathscr{A}, \mathscr{B}_{d_H})$-measurable is said to be an *interval-valued random set* (or convex compact random set of $\mathbb{R}$) associated with $(\Omega, \mathscr{A}, P)$, $\mathscr{B}_{d_H}$ denoting the σ-field generated by the topology induced by d_H on $\mathscr{K}_c(\mathbb{R})$.

If $X : \Omega \to \mathscr{K}_c(\mathbb{R})$ is an interval-valued random set associated with $(\Omega, \mathscr{A}, P)$, and $E[|X| \,|\, P] < \infty$ (condition which is often referred to as the *integrable boundedness of X*), with $|X|(\omega) = \sup\{|x| \,|\, x \in X(\omega)\}$ for all $\omega \in \Omega$, then the *expected value of X in Aumann's sense* [1] is defined as the set $E^A[X|P] = \{E(f|P) \,|\, f : \Omega \to \mathbb{R}, f \in L^1(\Omega, \mathscr{A}, P), f \in X \ a.s.\ [P]\}$, which in the case of X being interval-valued is given by the compact interval $E^A[X|P] = [E(\inf X|P), E(\sup X|P)]$.

Random sets are widely applicable in many areas (see, for instance, [15], [4], [3], [16], [22], [21]).

To develop an extension of the classical least squares method to the case in which we deal with interval-valued data, we will make use of a metric on $\mathscr{K}_c(\mathbb{R})$ extending the Euclidean one, and being easy to handle and interpret. For this purpose, we will consider the W-distance on $\mathscr{K}_c(\mathbb{R})$ which is defined for $A, B \in \mathscr{K}_c(\mathbb{R})$ as follows:

$$d_W(A,B) = \sqrt{\int_{[0,1]} [f_A(\lambda) - f_B(\lambda)]^2 \, dW(\lambda)}$$

with $f_A(\lambda) = \lambda \sup A + (1-\lambda) \inf A$ for all $\lambda \in [0,1]$, and W being formalized by means of a probability measure on the measurable space $([0,1], \mathscr{B}_{[0,1]})$ associated with a nondegenerate symmetric probability distribution on $[0,1]$ ($\mathscr{B}_{[0,1]}$ being the Borel σ-field on $[0,1]$) .

The W-distance is a particularization of a metric recently introduced (although in a more general space) by Körner and Näther [10]. On the basis of Radström Theorem (see, for instance, [5]), any convex set $A \in \mathscr{K}_c(\mathbb{R})$ can be embedded isometrically via its support function s_A into a cone of a Hilbert space of functions. As a consequence, an interval-valued random set can be viewed as a random function which takes on values in a Hilbert space, and any L_2-distance between the support functions of two elements $A, B \in \mathscr{K}_c(\mathbb{R})$ could be expressed as follows:

$$D_K(s_A, s_B) = \sqrt{\sum_{(u,v) \in S^0 \times S^0} (s_A(u) - s_B(u))(s_A(v) - s_B(v)) K(u,v)}$$

for some $K : S^0 \times S^0 \to \mathbb{R}$ (S^0 being the unit sphere in $\mathbb{R}$, i.e., $S^0 = \{-1, 1\}$) where K represents a symmetric and positive definite kernel, that is, $K(1,1) > 0$, $K(1,-1) = K(-1,1)$ and $K(1,1)K(-1,-1) > K(1,-1)K(-1,1)$.

The square of the distance above could be expressed alternatively by

$$\begin{aligned}[D_K(s_A, s_B)]^2 &= (K(1,1) - K(1,-1))[\sup A - \sup B]^2 \\ &\quad + (K(-1,-1) - K(1,-1))[\inf A - \inf B]^2 \\ &\quad + 4K(1,-1)[\operatorname{mid} A - \operatorname{mid} B]^2\end{aligned}$$

with $\operatorname{mid} A = [\sup A + \inf A]/2$ denoting the centre of interval A. If we slightly constrains the kernel K to assess the same "weight" to the squared Euclidean distance between the suprema and the squared Euclidean distance between the infima (i.e., $K(1,1) = K(-1,-1)$) and to assess a nonnegative "weight" to the squared Euclidean distance between the mid-points (i.e., $K(1,-1) \geq 0$), then, for $A, B \in \mathscr{K}_c(\mathbb{R})$ we have that $D_K(s_A, s_B) = d_W(A,B)$.

Remark 1. It should be remarked that the metric employed by Diamond corresponds to d_W for $W(0) = W(1) = .5$. It is obvious that, in the conditions assumed for K, the generalized distance d_W is equivalent to the generalized metric $d_{\boldsymbol{\lambda}}$ by Bertoluzza, Corral, and Salas [2] (see also [13]), with $\boldsymbol{\lambda} = (W(0), W(.5), W(1))$, but frequently choosing W on $[0, 1]$ is more intuitive and easier in practice than choosing $\boldsymbol{\lambda}$. On the other hand, although the measure W has no stochastic meaning, we can formally deal with it in a probabilistic context and hence we can work if required with the probability space $(\Omega \times [0, 1], \mathscr{A} \otimes \mathscr{B}_{[0,1]}, P \otimes W)$. The mapping $f_{[0,1]} : \Omega \times [0, 1] \to \mathbb{R}$ can be treated as a real-valued random variable which is constant w.r.t. P, and $\sigma^2_{f_{[0,1]}} = \mathrm{Var}[f_{[0,1]} \mid P \otimes W] = \int_{[0,1]} \lambda^2 \, dW(\lambda) - .25 > 0$. We can easily prove for the arbitrary $A, B \in \mathscr{K}_c(\mathbb{R})$ that, due to the symmetry assumed for W, $[d_W(A, B)]^2 = [\mathrm{mid}\, A - \mathrm{mid}\, B]^2 + 4\sigma^2_{f_{[0,1]}} [\mathrm{spr}\, A - \mathrm{spr}\, B]^2$ (with the spread being given by $\mathrm{spr}\, A = [\sup A - \inf A]/2$), whence the greater $\sigma^2_{f_{[0,1]}}$ the greater the influence of the Euclidean distance between the spreads of A and B on $d_W(A, B)$, this influence being the greatest possible one for the metric used by Diamond.

As we have just indicated, from now on we will consider the generalized metric d_W, since choosing W and the interpretation of d_W are usually more intuitive than assigning weigths to the extreme and mid-points of interval-valued data (or, alternatively, assigning the value of $\sigma^2_{f_{[0,1]}}$).

3 Extended least squares method for interval-valued random sets

Suppose that X and Y are two nondegenerate interval-valued random sets associated with a probability space $(\Omega, \mathscr{A}, P)$. If an interval $X(\omega)$ is observed and we want to "estimate" the corresponding interval value $Y(\omega)$ (that is, X is an independent or predictor interval-valued random set, and Y is a dependent or response interval-valued random set), we can try to approximate Y as an affine function of either X (or, more generally, of $g(X)$, g being a well-defined measurable function). Figure 1 shows the graphical representation of the esti-

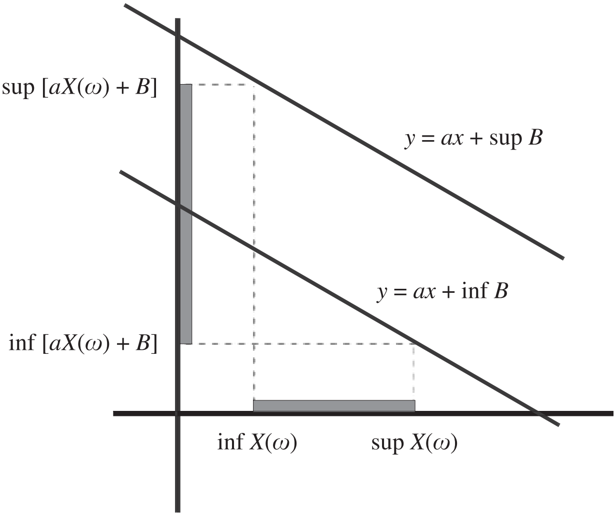

Fig. 1. Graphical representation of $aX(\omega) + B$

mate $aX(\omega) + B$, $+$ denoting the Minkowski addition on $\mathcal{K}_c(\mathbb{R})$ and $aX(\omega)$ meaning the product of the interval $X(\omega)$ by the scalar a.

In order to determine the best approximation of Y by an affine function of X, we are going to consider an *extension of the least squares method* based on the W-distance in Section 2. Thus, the sum (or, more generally, the mean) of squares of residuals is now extended by the sum (or the mean) of the W-distances between the observed and the estimated values of Y by the affine function.

Therefore, the aim is to minimize the objective function $\phi : \mathbb{R} \times \mathcal{K}_c(\mathbb{R}) \to [0, +\infty)$ such that $\phi(a, B) = E([d_W(Y, aX + B)]^2 \mid P)$ for $a \in \mathbb{R}$, $B \in \mathcal{K}_c(\mathbb{R})$.

3.1 Objective function of the problem

In virtue of the definition of the product of elements in $\mathcal{K}_c(\mathbb{R})$ by a scalar, the minimization of function ϕ is equivalent to the minimization of an alternative *objective function* defined in terms of expected values over the probability space $(\Omega \times [0,1], \mathcal{A} \otimes \mathcal{B}_{[0,1]}, P \otimes W)$. This alternative function is given by the mapping $\psi : \mathbb{R} \times [0, +\infty) \times \mathbb{R} \to [0, +\infty)$ such that

$$\psi(a,b,c) = \begin{cases} E[(f_Y - af_X - bf_{[0,1]} - c)^2 \mid P \otimes W] & \text{if } a \geq 0 \\ E[(f_Y - a\tilde{f}_X - bf_{[0,1]} - c)^2 \mid P \otimes W] & \text{if } a \leq 0 \end{cases}$$

with $\tilde{f}_A(\lambda) = f_A(1-\lambda)$ for all $\lambda \in [0,1]$, $b = 2\,\mathrm{spr}\, B \in [0, +\infty)$, and $c = \inf B \in \mathbb{R}$.

In particular, if we consider a random sample of n pairs of interval-valued data, $(X_1, Y_1), \ldots, (X_n, Y_n)$, then

$$\psi(a,b,c) = \begin{cases} \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \int_{[0,1]} (f_{Y_i}(\lambda) - af_{X_i}(\lambda) - b\lambda - c)^2 \, dW(\lambda) & \text{if } a \geq 0 \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \int_{[0,1]} (f_{Y_i}(\lambda) - af_{X_i}(1-\lambda) - b\lambda - c)^2 \, dW(\lambda) & \text{if } a \leq 0. \end{cases}$$

As a consequence, the problem we want to solve can be viewed as a mixture of two (according to the sign of the "slope" a) multiple linear regression problems, each of them involving three (real-valued) random variables associated with $(\Omega \times [0,1], \mathcal{A} \otimes \mathcal{B}_{[0,1]}, P \otimes W)$. These variables are f_X, f_Y and $f_{[0,1]}$ if $a \geq 0$, and $\tilde{f}_X, f_Y$ and $f_{[0,1]}$ if $a < 0$, with the constraint $b \geq 0$.

Note that there is an essential difference between $\psi(a,b,c)$ in the theoretical and sample cases. Thus, the first one corresponds to a criterion for approximating a theoretical function Y and trying to find a proper mathematical solution, whereas the second one is a data-analytic fitting problem which always can be solved by suitable numerical procedures.

Remark 2. The problem above could be also presented by considering an approach involving a second alternative objective function defined (but for the moment $\sigma^2_{f_{[0,1]}}$ which only depends on W) in terms of expected values over the probability space $(\Omega, \mathcal{A}, P)$, and involving the real-valued random variables $\mathrm{mid}\, X$, $\mathrm{mid}\, Y$, $\mathrm{spr}\, X$, and $\mathrm{spr}\, Y$, as follows:

$$\varphi(a,b,c) = E[(\text{mid}\, Y - a\,\text{mid}\, X - c)^2 + 4\sigma^2_{f_{[0,1]}}(\text{spr}\, Y - |a|\,\text{spr}\, X - b)^2 \mid P]$$

with $a \in \mathbb{R}$, $b = \text{spr}\, B \in [0, +\infty)$, and $c = \text{mid}\, B \in \mathbb{R}$. The minimization of φ over $\mathbb{R} \times [0, +\infty) \times \mathbb{R}$ can be solved by fitting two parallel lines twice: once for $(\text{mid}\, X, \text{mid}\, Y)$ and $(2\sigma_{f_{[0,1]}}\,\text{spr}\, X, 2\sigma_{f_{[0,1]}}\,\text{spr}\, Y)$ (which will be valid whenever the slope a is nonnegative), and then for $(\text{mid}\, X$, $\text{mid}\, Y)$ and $(2\sigma_{f_{[0,1]}} \cdot [-\text{spr}\, X], 2\sigma_{f_{[0,1]}}\,\text{spr}\, Y)$ (which will be valid whenever the slope a is negative). In this approach to the problem, we would make use of standard linear algebra studies to get the solutions, to compute the residual sums of squares, and to give the conditions leading to nonunique optimal solutions.

3.2 Notations and preliminary computations

Before analyzing the solutions of the problem subject to the mixture and constraint pointed out above, we are going to describe some notations to be used in such an analysis, as well as to indicate that (as already commented in Remark 2) all the values employed to formalize the solutions can be expressed in terms of moments of the random variables mid X, mid Y, spr X, and spr Y. To better distinguish these moments over the probability space $(\Omega, \mathcal{A}, P)$, we will make use of the standard linear model notation when we deal with these four variables. An additional reason to do it is that in statistical applications (like in the example below) we will consider the sample analogue estimates instead of the theoretical parameters. Thus,

$$\mu_{f_{[0,1]}} = E[f_{[0,1]} \mid P \otimes W] = .5,$$

$$\mu_{f_X} = E[f_X \mid P \otimes W] = E[\tilde{f}_X \mid P \otimes W] = E(\text{mid}\, X|P) = \text{mid}\, E^A[X|P],$$

$$\sigma^2_{f_X} = \text{Var}[f_X \mid P \otimes W] = \text{Var}[\tilde{f}_X \mid P \otimes W] = 4\sigma^2_{f_{[0,1]}} E((\text{spr}\, X)^2 \mid P) + \text{Var}(\text{mid}\, X|P),$$

$$\rho_{f_X f_Y} = \frac{\sigma_{f_X f_Y}}{\sqrt{\sigma^2_{f_X}\sigma^2_{f_Y}}} = \frac{\text{Cov}[f_X, f_Y \mid P \otimes W]}{\sqrt{\sigma^2_{f_X}\sigma^2_{f_Y}}} = \frac{\text{Cov}(\text{mid}\, X, \text{mid}\, Y|P) + 4\sigma^2_{f_{[0,1]}} E(\text{spr}\, X \cdot \text{spr}\, Y|P)}{\sqrt{\sigma^2_{f_X}\sigma^2_{f_Y}}},$$

$$\rho_{\tilde{f}_X f_Y} = \frac{\text{Cov}(\text{mid}\, X, \text{mid}\, Y|P) - 4\sigma^2_{f_{[0,1]}} E(\text{spr}\, X \cdot \text{spr}\, Y|P)}{\sqrt{\sigma^2_{f_X}\sigma^2_{f_Y}}} \leq \rho_{f_X f_Y},$$

$$\rho_{f_X f_{[0,1]}} = -\rho_{\tilde{f}_X f_{[0,1]}} = 2\sqrt{\frac{\sigma^2_{f_{[0,1]}}}{\sigma^2_{f_X}}} E(\text{spr}\, X|P).$$

If X and Y are assumed to be nondegenerate, then $\sigma^2_{f_X} > 0$, $\sigma^2_{f_Y} > 0$, $\rho_{f_X f_{[0,1]}} \neq 1$, and $\rho_{f_Y f_{[0,1]}} \neq 1$.

3.3 *Optimal solutions and discussion on the uniqueness of the solution*

In accordance with the first approach to the problem we are considering, to look for $(a^*, B^*) \in \mathbb{R} \times \mathscr{K}_c(\mathbb{R})$ minimizing ϕ is equivalent to looking for the vectorial value $(a^*, b^*, c^*) \in \mathbb{R} \times [0, +\infty) \times \mathbb{R}$ minimizing ψ, which can be proven to be given by

$$(a^*, b^*, c^*) = \begin{cases} (a_1, b_1, c_1) & \text{if CASE 1} \\ (a_2, b_2, c_2) & \text{if CASE 2} \\ (a'_1, 0, c'_1) & \text{if CASE 3} \\ (a'_2, 0, c'_2) & \text{if CASE 4} \\ (0, b'', c'') & \text{if CASE 5,} \end{cases}$$

that is, ϕ is minimized for $(a^*, B^*) \in \mathbb{R} \times \mathscr{K}_c(\mathbb{R})$ such that

$$(a^*, B^*) = \begin{cases} (a_1, [c_1, c_1 + b_1]) & \text{if CASE 1} \\ (a_2, [c_2, c_2 + b_2]) & \text{if CASE 2} \\ (a'_1, \{c'_1\}) & \text{if CASE 3} \\ (a'_2, \{c'_2\}) & \text{if CASE 4} \\ (0, E^A[X|P]) & \text{if CASE 5,} \end{cases}$$

where

- CASE 1 holds iff either $a_1 > 0$, $a_2 \geq 0$, $b_1 > 0$, or $a_1^2 \geq a_2^2$, $a_1 > 0$, $a_2 < 0$, $b_1 > 0$, $b_2 > 0$, or $a_1^2[1 - \rho^2_{f_X f_{[0,1]}}] \geq {a'_2}^2 - \sigma^2_{f_Y} \rho^2_{f_Y f_{[0,1]}} / \sigma^2_{f_X}$, $a_1 > 0$, $a_2 < 0$, $b_1 > 0$, $b_2 \leq 0$,
- CASE 2 holds iff either $a_1 \leq 0$, $a_2 < 0$, $b_2 > 0$, or $a_1^2 \leq a_2^2$, $a_1 > 0$, $a_2 < 0$, $b_1 > 0$, $b_2 > 0$, or $a_2^2[1 - \rho^2_{f_X f_{[0,1]}}] \geq {a'_1}^2 - \sigma^2_{f_Y} \rho^2_{f_Y f_{[0,1]}} / \sigma^2_{f_X}$, $a_1 > 0$, $a_2 < 0$, $b_1 \leq 0$, $b_2 \leq 0$,
- CASE 3 holds iff either $a_1 > 0$, $a_2 \geq 0$, $b_1 \leq 0$, or ${a'_1}^2 \geq {a'_2}^2$, $a_1 > 0$, $a_2 < 0$, $b_1 \leq 0$, $b_2 > 0$, or $a_2^2[1 - \rho^2_{f_X f_{[0,1]}}] \leq {a'_1}^2 - \sigma^2_{f_Y} \rho^2_{f_Y f_{[0,1]}} / \sigma^2_{f_X}$, $a_1 > 0$, $a_2 < 0$, $b_1 \leq 0$, $b_2 \leq 0$,
- CASE 4 holds iff either $a_1 \leq 0$, $a_2 < 0$, $b_2 \leq 0$, or ${a'_1}^2 \leq {a'_2}^2$, $a_1 > 0$, $a_2 < 0$, $b_1 \leq 0$, $b_2 > 0$, or $a_1^2[1 - \rho^2_{f_X f_{[0,1]}}] \leq {a'_2}^2 - \sigma^2_{f_Y} \rho^2_{f_Y f_{[0,1]}} / \sigma^2_{f_X}$, $a_1 > 0$, $a_2 < 0$, $b_1 > 0$, $b_2 \leq 0$,
- CASE 5 holds iff $a_1 \leq 0 \leq a_2$,

and with

$$a_1 = \sqrt{\frac{\sigma^2_{f_Y}}{\sigma^2_{f_X}} \cdot \frac{\rho_{f_X f_Y} - \rho_{f_X f_{[0,1]}} \rho_{f_Y f_{[0,1]}}}{1 - \rho^2_{f_X f_{[0,1]}}}},$$

$$a_2 = \sqrt{\frac{\sigma^2_{f_Y}}{\sigma^2_{f_X}} \cdot \frac{\rho_{\tilde{f}_X f_Y} + \rho_{f_X f_{[0,1]}} \rho_{f_Y f_{[0,1]}}}{1 - \rho^2_{f_X f_{[0,1]}}}},$$

$$b_1 = \sqrt{\frac{\sigma^2_{f_Y}}{\sigma^2_{f_{[0,1]}}} \cdot \frac{\rho_{f_Y f_{[0,1]}} - \rho_{f_X f_{[0,1]}} \rho_{f_X f_Y}}{1 - \rho^2_{f_X f_{[0,1]}}}},$$

$$b_2 = \sqrt{\frac{\sigma^2_{f_Y}}{\sigma^2_{f_{[0,1]}}} \cdot \frac{\rho_{f_Y f_{[0,1]}} + \rho_{f_X f_{[0,1]}} \rho_{\tilde{f}_X f_Y}}{1 - \rho^2_{f_X f_{[0,1]}}}},$$

$$c_1 = \mu_{f_Y} - a_1 \mu_{f_X} - .5 b_1, \quad c_2 = \mu_{f_Y} - a_2 \mu_{f_X} - .5 b_2,$$

$$a'_1 = \sqrt{\frac{\sigma^2_{f_Y}}{\sigma^2_{f_X}}} \cdot \rho_{f_X f_Y}, \quad a'_2 = \sqrt{\frac{\sigma^2_{f_Y}}{\sigma^2_{f_X}}} \cdot \rho_{\tilde{f}_X f_Y},$$

$$c'_1 = \mu_{f_Y} - a'_1 \mu_{f_X}, \quad c'_2 = \mu_{f_Y} - a'_2 \mu_{f_X},$$

$$b'' = \sqrt{\frac{\sigma^2_{f_Y}}{\sigma^2_{f_{[0,1]}}}} \cdot \rho_{f_Y f_{[0,1]}}, \quad c'' = \mu_{f_Y} - .5 b''.$$

Obviously, in case $a_1 > 0 > a_2$, the solution is not necessarily unique, but for the cases of nonunique solution (in fact, for obtaining two possible solutions, one in the cone corresponding to $a > 0$ and the other one in the cone for $a < 0$) the following conditions are necessary and sufficient:

- if $a_1 > 0, b_1 > 0$ and $a_2 < 0, b_2 > 0$, then the optimal solution is nonunique if, and only if, $a_1 + a_2 = 0$, $b_1 = b_2$, $c_2 = c_1 + 2a_1 \mu_{f_X}$;
- if $a_1 > 0$, $b_1 \leq 0$ and $a_2 < 0, b_2 \leq 0$, then the optimal solution is nonunique if, and only if, $a'_1 + a'_2 = 0$, $c'_2 = c'_1 + 2a'_1 \mu_{f_X}$ (in such a case we have that $b'_1 = b'_2 = 0$);
- if $a_1 > 0$, $b_1 > 0$ and $a_2 < 0$, $b_2 \leq 0$, then the optimal solution is nonunique if, and only if, ${a'_2}^2 - a_1^2[1 - \rho^2_{f_X f_{[0,1]}}] = \frac{\sigma^2_{f_Y}}{\sigma^2_{f_X}} \rho^2_{f_Y f_{[0,1]}}$;
- if $a_1 > 0$, $b_1 \leq 0$ and $a_2 < 0$, $b_2 > 0$, then the optimal solution is nonunique if, and only if, ${a'_1}^2 - a_2^2[1 - \rho^2_{f_X f_{[0,1]}}] = \frac{\sigma^2_{f_Y}}{\sigma^2_{f_X}} \rho^2_{f_Y f_{[0,1]}}$.

Remark 3. In accordance with the notations we are using in this section, the cases Diamond (see [4]) has formally examined and for which the uniqueness of the optimal solution has been proven (which are referred to as situations involving *coherent* interval-valued data) correspond to $a_1 \geq a_2 \geq 0$ and $a_1 \leq a_2 \leq 0$. As indicated by Diamond [4], nonuniqueness of the optimal solution would be undesirable, and we would, therefore, want to fit another type of function.

3.4 Algorithm for the optimal solution

Step 1. Compute the values $a_1, a_2, a'_1, a'_2, b_1, b_2, \mu_{f_X}, \mu_{f_Y}$ and the Aumann expected value $E^A[Y|P] = [c'', c'' + b'']$, and go to Step 2.

Step 2. IF $a_1 \leq 0$ THEN go to Step 3, ELSE go to Step 5.

Step 3. IF $a_2 \geq 0$ THEN the optimal solution is the one for CASE 5, ELSE go to Step 4.

Step 4. IF $b_2 > 0$ THEN the optimal solution is the one for CASE 2, ELSE the optimal solution is the one for CASE 4.

Step 5. IF $a_2 \geq 0$ THEN go to Step 6, ELSE go to Step 7.

Step 6. IF $b_1 > 0$ THEN the optimal solution is the one for CASE 1, ELSE the optimal solution is the one for CASE 3.

Step 7. IF $b_1 > 0$ THEN go to Step 8, ELSE go to Step 11.

Step 8. IF $b_2 > 0$ THEN go to Step 9, ELSE go to Step 10.

Step 9. IF $a_1^2 > a_2^2$ THEN the optimal solution is the one for CASE 1, ELSE
IF $a_1^2 < a_2^2$ THEN the optimal solution is the one for CASE 2,
ELSE there are two optimal solutions, namely, those for CASE 1 and CASE 2.

Step 10. IF $a_1^2[1 - \rho^2_{f_X f_{[0,1]}}] > {a'_2}^2 - \frac{\sigma^2_{f_Y}}{\sigma^2_{f_X}} \rho^2_{f_Y f_{[0,1]}}$ THEN the optimal solution the one for CASE 1, ELSE
IF $a_1^2[1 - \rho^2_{f_X f_{[0,1]}}] < {a'_2}^2 - \frac{\sigma^2_{f_Y}}{\sigma^2_{f_X}} \rho^2_{f_Y f_{[0,1]}}$ THEN the optimal solution is the one for CASE 4,
ELSE there are two optimal solutions, namely, those for CASE 1 and CASE 4.

Step 11. IF $b_2 > 0$ THEN go to Step 12, ELSE go to Step 13.

Step 12. IF ${a'_1}^2 > {a'_2}^2$ THEN the optimal solution is the oNe for CASE 3, ELSE
IF ${a'_1}^2 < {a'_2}^2$ THEN the optimal solution is the one for CASE 4,
ELSE there are two optimal solutions, namely, those for CASE 3 and CASE 4.

Step 13. IF $a_2^2[1 - \rho^2_{f_X f_{[0,1]}}] > {a'_1}^2 - \frac{\sigma^2_{f_Y}}{\sigma^2_{f_X}} \rho^2_{f_Y f_{[0,1]}}$ THEN the optimal solution is the one for CASE 2, ELSE
IF $a_2^2[1 - \rho^2_{f_X f_{[0,1]}}] < {a'_1}^2 - \frac{\sigma^2_{f_Y}}{\sigma^2_{f_X}} \rho^2_{f_Y f_{[0,1]}}$ THEN the optimal solution is the one for CASE 3,
ELSE there are two optimal solutions, namely, those for CASE 2 and CASE 3.

Remark 4. Since $a_1 \geq 0$ implies that $a'_1 \geq 0$, $a_1 > 0$ implies that $a'_1 > 0$, $a_2 \leq 0$ implies that $a'_2 \leq 0$, and $a_2 < 0$ implies that $a'_2 < 0$, we can conclude that $a^* = 0$ (and, consequently, that the optimal solution corresponds to the affine function $Y = E^A[Y|P]$) can only happen in cases in which the optimal solution is unique. In fact, the "affine independence", intended to occur whenever $a^* = 0$, will happen if, and only if, $a_1 \leq 0$ and $a_2 \geq 0$. In particular, when X and Y are *independent random sets*, then $\operatorname{mid} X$ and $\operatorname{mid} Y$ are independent random variables, and also $\operatorname{spr} X$ and $\operatorname{spr} Y$ are independent random variables, whence $a_1 = a_2 = 0$, and hence $a^* = 0$.

4 Linear correlation analysis

In this section we are confronted with the ultimate goal of this paper, which is that of quantiying how well the optimal affine function of the interval-valued random set X explains the variation in the dependent interval-valued random set Y.

As we have just seen in Remark 4, the optimal "slope" a^* equals 0 in a few situations, in all of which the optimal affine function being $Y = E^A[Y|P]$. In these situations, the unexplained W-variation equals the highest possible one, $\sigma^2_{f_Y}[1 - \rho^2_{f_Y f_{[0,1]}}]$, that is, the total W-variation. Following the ideas in the real-valued case, we now introduce the *extended determination coefficient* of X and Y, which is defined by means of the quotient

$$R^2_{XY} = \frac{\text{Total variation} - \text{Unexplained variation}}{\text{Total variation}} = 1 - \frac{\phi(a^*, B^*)}{\sigma^2_{f_Y}[1 - \rho^2_{f_Y f_{[0,1]}}]}$$

$$= \begin{cases} \dfrac{a_1^2\sigma^2_{f_X}[1 - \rho^2_{f_X f_{[0,1]}}]}{\sigma^2_{f_Y}[1 - \rho^2_{f_Y f_{[0,1]}}]} & \text{if CASE 1} \\[2ex] \dfrac{a_2^2\sigma^2_{f_X}[1 - \rho^2_{f_X f_{[0,1]}}]}{\sigma^2_{f_Y}[1 - \rho^2_{f_Y f_{[0,1]}}]} & \text{if CASE 2} \\[2ex] \dfrac{\rho^2_{f_X f_Y} - \rho^2_{f_Y f_{[0,1]}}}{1 - \rho^2_{f_Y f_{[0,1]}}} & \text{if CASE 3} \\[2ex] \dfrac{\rho^2_{\tilde{f}_X f_Y} - \rho^2_{f_Y f_{[0,1]}}}{1 - \rho^2_{f_Y f_{[0,1]}}} & \text{if CASE 4} \\[2ex] 0 & \text{if CASE 5.} \end{cases}$$

R^2_{XY} is a well-defined dimensionless coefficient, whose algorithmic computation is gathered in Subsection 4.2.

4.1 *Discussing the strength of association in terms of the extended determination coefficient*

Since $0 \leq \phi(a^*, B^*) \leq \sigma^2_{f_Y}[1 - \rho^2_{f_Y f_{[0,1]}}][1 - R^2_{XY}]$, then, R^2_{XY} is a symmetric measure varying from 0 to 1, and the smaller R^2_{XY} the greater the unexplained variation. In other words, R^2_{XY} is a measure of the portion of the total variation $E([d_W(Y, E^A[Y|P])]^2 \,|\, P) = \sigma^2_{f_Y}[1 - \rho^2_{f_Y f_{[0,1]}}]$ of random set Y that is explained by the optimal affine relation with X.

Obviously, $R^2_{XY} = 0$ is equivalent to the "affine independence" of X and Y (i.e., it corresponds to the case in which $a^* = 0$), and the optimal relation $Y = E^A[Y|P]$ does not depend on X. Consequently, the independence of X and Y entails the "affine independence" of Y with respect to X.

On the other hand, $R^2_{XY} = 1$ if, and only if, $\phi(a^*, B^*) = 0$, that is, there exist $a^* \in \mathbb{R}$ and $B^* \in \mathscr{K}_c(\mathbb{R})$ such that $d_W(Y, a^*X + B^*) = 0$ *a.s.* $[P]$, which is equivalent to $Y = a^*X + B^*$ *a.s.* $[P]$. Furthermore, since Y is nondegenerate, then $R^2_{XY} = 1$ entails that $a^* \neq 0$.

4.2 Algorithm for computing the determination coefficient

Step 1. Compute values $a_1, a_2, b_1, b_2, \sigma^2_{f_X}, \sigma^2_{f_Y}, \rho^2_{f_X f_{[0,1]}}, \rho^2_{f_Y f_{[0,1]}}, \rho^2_{f_X f_Y}$ and $\rho^2_{\tilde{f}_X f_{[0,1]}}$, and go to Step 2.

Step 2. IF $a_1 \leq 0$ THEN go to Step 3, ELSE go to Step 5.

Step 3. IF $a_2 \geq 0$ THEN the extended determination coefficient is the one for CASE 5, ELSE go to Step 4.

Step 4. IF $b_2 > 0$ THEN the extended determination coefficient is the one for CASE 2, ELSE the extended determination coefficient is the one for CASE 4.

Step 5. IF $a_2 \geq 0$ THEN go to Step 6, ELSE go to Step 7.

Step 6. IF $b_1 > 0$ THEN the extended determination coefficient is the one for CASE 1, ELSE the extended determination coefficient is the one for CASE 3.

Step 7. IF $b_1 > 0$ THEN go to Step 8, ELSE go to Step 9.

Step 8. IF $b_2 > 0$ THEN the extended determination coefficient is given by the maximum of the values of R^2_{XY} for CASE 1 and CASE 2, ELSE the extended determination coefficient is given by the maximum of the values of R^2_{XY} for CASE 1 and CASE 4.

Step 9. IF $b_2 > 0$ THEN the extended determination coefficient is given by the maximum of the values of R^2_{XY} for CASE 2 and CASE 3, ELSE the extended determination coefficient is given by the maximum of the values of R^2_{XY} for CASE 3 and CASE 4.

5 Illustrative example

The following example illustrates the application of the studies above by means of a real-life case. Data have been supplied by the Servicio de Nefrología of the Hospital Valle del Nalón in Langreo (Asturias, Spain).

Example 1. The paired data in Table 1 correspond to the "values" (observed on a population Ω of 59 patients who are hospitalized) for the interval-valued random sets $X =$ "range of systolic blood pressure over a day" and $Y =$ "range of diastolic blood pressure over the same day".

If we want to look for the optimal affine relation between Y and X, in which the first one is expressed as an affine function of the second one, and we use the results in Section 3 by choosing as the measure W the Lebesgue measure on $[0, 1]$, then the optimal relation is given by

$$Y = .4384X + [9.6436, 27.5856],$$

and the extended linear determination coefficient of X and Y is given by $R^2_{XY} = .3699$. Figure 2 shows the scatter diagram and the corresponding optimal solution.

Table 1. Data on the ranges of the systolic (X) and diastolic (Y) blood pressure

X	Y	X	Y	X	Y
118–173	63–102	119–212	47–93	98–160	47–108
104–161	71–108	122–178	73–105	97–154	60–107
131–186	58–113	127–189	74–125	87–150	47–86
105–157	62–118	113–213	52–112	141–256	77–158
120–179	59–94	141–205	69–133	108–147	62–107
101–194	48–116	99–169	53–109	115–196	65–117
109–174	60–119	126–197	60–98	99–172	42–86
128–210	76–125	99–201	55–121	113–176	57–95
94–145	47–104	88–221	37–94	114–186	46–103
148–201	88–130	113–183	55–85	145–210	100–136
111–192	52–96	94–176	56–121	120–180	59–90
116–201	74–133	102–156	50–94	100–161	54–104
102–167	39–84	103–159	52–95	159–214	99–127
104–161	55–98	102–185	63–118	138–221	70–118
106–167	45–95	111–199	57–113	87–152	50–95
112–162	62–116	130–180	64–121	120–188	53–105
136–201	67–122	103–161	55–97	95–166	54–100
90–177	52–104	125–192	59–101	92–173	45–107
116–168	58–109	97–182	54–104	83–140	45–91
98–157	50–111	127–226	57–101		

6 Concluding remarks and open problems

The study of approximating X by an affine function of Y would lead also to unique optimal solutions for, and only for, the cases in Subsection 3.3. Actually, if in accordance with the optimal affine relation of Y w.r.t. X the "slope" is positive, negative or null, the same happens for the "slope" of the optimal affine relation of X w.r.t. Y.

If, in particular, Y is almost surely real-valued and X is interval-valued, the optimal affine function of Y w.r.t. X is always unique and it is given by

$$Y - \mu_{f_Y} = \frac{\sigma_{f_X f_Y}}{\sigma^2_{f_X}}(X - \mu_{f_X}),$$

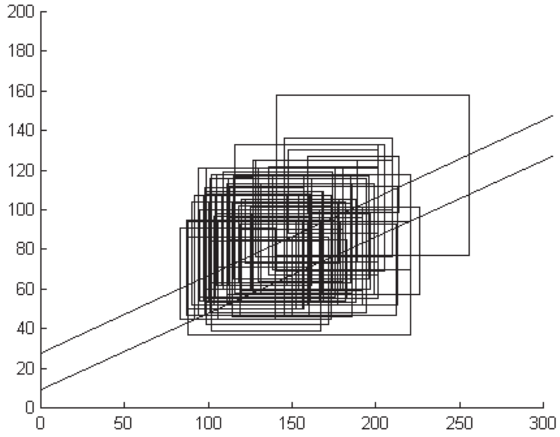

Fig. 2. Scatter diagram of rectangular-valued data in Example 1

that is, the "slope" and "intercept" coincides with those for the optimal linear relation of f_Y w.r.t. f_X.

As Diamond [4] has indicated, if Y is interval-valued and X is almost surely real-valued, then a more natural and convenient relation will be that given by the affine function $Y = AX + B$ with $A, B \in \mathscr{K}_c(\mathbb{R})$, for which some interesting discussions have been developed by Diamond.

The studies in this paper should be complemented with inferential procedures for testing/estimating the real- or interval-valued parameters in the affine function, as well as the determination coefficient. Since the optimal solutions and determination coefficient in the study in this paper depend basically on moments of the real-valued random variables $\operatorname{mid} X$, $\operatorname{mid} Y$, $\operatorname{spr} X$, and $\operatorname{spr} Y$, then the computation of these terms will become quite simple in practice. On the other hand, and due to such a dependence, inferential analyses could be certainly based either on the normality of the random vector $(\operatorname{mid} X, \operatorname{mid} Y, \operatorname{spr} X, \operatorname{spr} Y)$ (in particular, a first study could be that based on normal random sets [14]), or on Large Sample Theory techniques applied for this random vector (a first approach in this sense would be obtained by following ideas similar to those by Körner [8] and Montenegro et al. [17]).

Although the assumption that W is symmetric looks quite natural in the context in this paper, the problem in which this assumption is removed can also be solved by considering, for instance, the approach in Remark 2, where the objective function would be now given by

$$\begin{aligned}\varphi(a,b,c) = E[((&\operatorname{mid} Y - a \operatorname{mid} X - c)\\ &+ (2\mu_{f_{[0,1]}} - 1)(\operatorname{spr} Y - |a| \operatorname{spr} X - b))^2\\ &+ 4\sigma^2_{f_{[0,1]}}(\operatorname{spr} Y - |a| \operatorname{spr} X - b)^2 \mid P]\end{aligned}$$

with $a \in \mathbb{R}$, $b = \operatorname{spr} B \in [0, +\infty)$, and $c = \operatorname{mid} B \in \mathbb{R}$, and the solution can be obtained by using traditional techniques for linear regression/correlation problems.

The study in this paper could be extended to more complex situations like multiple regression of the type $Y = a_1 X_1 + \cdots + a_k X_k + B$, with $a_1, \ldots, a_k \in \mathbb{R}$ and $B \in \mathcal{K}_c(\mathbb{R})$, where $X_1, \ldots, X_k$ are assumed to be the independent interval-valued random sets. In this case, the discussion on the optimal solution and the associated degree of strength would require to consider 2^k cones, whence the complexity of the problem will increase extremely.

It would be interesting to extend the conclusions in this paper to the case in which interval-valued random sets are replaced by the so-called fuzzy random variables or random fuzzy sets in accordance with Puri and Ralescu ([19], [20]) (see also, [11] for some statistical studies concerning these random elements). Several studies have been already carried out in this respect (see, for instance, the recent review [6] and [18], [9], [10]).

Finally, it should be emphasized that the regression study and conclusions in this paper generalize those by Gil et al. [7], whence given two interval-valued random sets X and Y, the unexplained variation for the relation in this paper is lower than or equal to that for the relation in [7]. However, for an arbitrary reponse random set Y, the supremum (actually, the maximum) of the unexplained variation of the relation considered in this paper equals $\sigma^2_{f_Y}[1 - \rho^2_{f_Y f_{[0,1]}}]$, whereas the supremum of the unexplained variation of the relation considered in the previous one is $\sigma^2_{f_Y}$. Consequently, the correlation analysis in this paper does not represent a real generalization of that in [7].

Acknowledgements. The authors are deeply grateful to Professor Wolfgang Näther from the University of Freiberg and their colleagues Professors Norberto Corral, Miguel López-Díaz and Ignacio MartVnez for their valuable comments and suggestions in connection with this paper. The real-life data which illustrate the study in this paper have been supplied by Aurora Fonseca, Head of the Servicio de Nefrología, and Miguel de Zárraga, General Medical Manager, from the Hospital Valle del Nalón de Langreo in Asturias; we want to thank them sincerely for their collaboration and support.

References

1. Aumann RJ (1965) Integrals of set-valued functions. J Math Anal Appl 12:1–12
2. Bertoluzza C, Corral N, Salas A (1995) On a new class of distances between fuzzy numbers. Mathware & Soft Computing 2:71–84
3. Cressie NAC (1993) Statistics for spatial data. J Wiley & Sons, New York
4. Diamond P (1990) Least squares fitting of compact set-valued data. J Math Anal Appl 147:531–544
5. Diamond P, Kloeden PE (1994) Metric spaces of fuzzy sets. World Sci, New Jersey
6. Diamond P, Tanaka H (1998) Fuzzy regression analysis. In: Słowiński R (ed) Fuzzy Sets in Decision Analysis, Operations Research and Statistics. Kluwer Acad Pub, Norwell 11:349–387
7. Gil MA, López-García MT, Lubiano MA, Montenegro M (2001) Regression and correlation analyses of a linear relation between random intervals. Test 10:183–201
8. Körner R (2000) An asymptotic α-test for the expectation of random fuzzy variables. J Statist Plan Infer 83:331–346
9. Körner R, Näther W (1998) Linear regression with random fuzzy variables: extended classical estimates, best linear estimates, least squares estimates. Inform Sci 109:95–118
10. Körner R, Näther W (2001) On the variance of random fuzzy variables. In: Bertoluzza C, Gil MA, Ralescu DA (eds) Statistical Modeling, Analysis and Management of Fuzzy Data. Physica-Verlag, 22–39

11. López-Díaz M, Gil MA (1998) Reversing the order of integration in iterated expectations of fuzzy random variables, and statistical applications. J Stat Plan Infer 74:11–29
12. López-García H, López-Díaz M, Gil MA (2000) Interval-valued quantification of the inequality associated with a random set. Statist Probab Lett 46:149–159
13. Lubiano MA, Gil MA (1999) Estimating the expected value of fuzzy random variables in random samplings from finite populations. Statistical Papers 40:277–295
14. Lyashenko NN (1980) The statistics of random compacts in the Euclidean space. Zap Nauchn Semin Leningr Otd Mat Inst Steklova 98:115–139
15. Matheron G (1975) Random sets and integral Geometry. J Wiley & Sons, New York
16. Molchanov IS (1993) Limit theorems for unions of random closed sets. Lect Notes Math 1561. Springer-Verlag, Berlin
17. Montenegro M, Casals MR, Lubiano MA, Gil MA (2001) Two-sample hypothesis tests of means of a fuzzy random variable. Inform Sci 133:89–100
18. Näther W (1997) Linear statistical inference for random fuzzy data. Statistics 29:221–240
19. Puri ML, Ralescu DA (1985) The concept of normality for fuzzy random variables. Ann Probab 13:1373–1379
20. Puri ML, Ralescu D (1986) Fuzzy random variables. J Math Anal Appl 114:409–422
21. Stoyan D (1998) Random sets: Models and statistics. Int Stat Rev 66:1–27
22. Stoyan D, Kendall WS, Mecke J (1995) Stochastic Geometry and its applications. J Wiley & Sons, Chichester