On a Linear Independence Test for Interval-Valued Random Sets

Ángela Blanco[1,2], Ana Colubi[1], Norberto Corral[1], and Gil González-Rodríguez[2]

[1] Departamento de Estadística e I.O. y D.M., Universidad de Oviedo, Oviedo, Spain
[2] European Centre for Soft Computing, Mieres, Spain

Abstract. The linear relationship between interval-valued random sets can arise in different ways. Recently, a linear model based on the natural arithmetic for intervals has been studied. In order to test whether the explanatory random set contributes significantly to explain the response random set through that linear model, an asymptotic testing procedure is here proposed. The empirical size of the test is illustrated by means of some simulations. The approach is also applied to a case-study.

1 Introduction

The linear regression problem between interval-valued random sets has been previously considered in the literature from different viewpoints (see, for instance, [4, 5, 6, 9], [8, 12]).

In [8] a linear regression model for compact and convex random sets based on a set-arithmetic approach has been established, and the estimators for the parameters have been obtained by applying the least-squares criterion based on a generalized L_2-type metric (see also [7]). In this communication we propose to complement those studies by proposing a linear independence test in the same context.

The organization of the paper is as follows. In Section 2 some preliminary concepts about interval-valued random sets and the considered linear regression model are presented. In Section 3 we suggest a test statistic for the linear independence. The asymptotic distribution of the statistic in some particular cases is used to state the asymptotic testing procedure. In Section 4 we show the results of some simulations in connection with the empirical significance level. The test is applied to a case-study in Section 5. Finally, in Section 6 some concluding remarks are commented.

2 Preliminaries

Let $\mathscr{K}_c(\mathbb{R})$ denote the class of nonempty compact intervals endowed with the natural interval-arithmetic induced by the Minkowski addition and the product by a scalar; namely, $A+B=\{a+b : a\in A, b\in B\}$ and $\lambda A=\{\lambda a : a\in A\}$, for all $A,B\in\mathscr{K}_c(\mathbb{R})$ and $\lambda\in\mathbb{R}$.

Due to the lack of symmetric element with respect to the addition, the space $(\mathscr{K}_c(\mathbb{R}),+,\cdot)$ is not linear, but semilinear, so it is useful to consider the *Hukuhara*

D. Dubois et al. (Eds.): Soft Methods for Hand. Var. and Imprecision, ASC 48, pp. 111–117, 2008.
springerlink.com

difference between A and B, defined as the interval C so that $A = B + C$ (if it exists) and denoted in this case as $C = A -_H B$ (see [11]). It is possible to assure the existence of $A -_H B$ if, and only if, $\inf A - \inf B \leq \sup A - \sup B$; moreover, in this case $A -_H B = [\inf A - \inf B, \sup A - \sup B]$.

The space $(\mathscr{K}_c(\mathbb{R}), +, \cdot)$ can be embedded onto a convex cone of the square integrable functions $\mathscr{L}(\mathbb{R})$ via the mapping $s : \mathscr{K}_c(\mathbb{R}) \longrightarrow \mathscr{L}(\mathbb{R})$ defined by $s(A) = s_A$ for all $A \in \mathscr{K}_c(\mathbb{R})$, where s_A denotes the *support function* of the interval A, namely, $s_A : \mathbb{R} \to \mathbb{R}$ such that $s_A(u) = \sup_{a \in A} \langle a, u \rangle$ for every $u \in \mathbb{R}$, $\langle \cdot, \cdot \rangle$ being the usual inner product on $\mathbb{R}$. The support function is semilinear, that is, $s_{A+B} = s_A + s_B$ and $s_{\lambda A} = \lambda s_A$, for $A, B \in \mathscr{K}_c(\mathbb{R})$ and $\lambda \geq 0$. Furthermore, if $A -_H B$ exists, then $s_{A -_H B} = s_A - s_B$. The function s allows us to deal with the space $\mathscr{L}(\mathbb{R})$, which can be endowed with an inner product which entails a Hilbertian structure.

The least square method considered in [8] for the estimation process is based on a generalized metric on $\mathscr{K}_c(\mathbb{R})$ via support functions (see [14]), which is defined for any $A, B \in \mathscr{K}_c(\mathbb{R})$ as

$$d_K(A,B) = \Big(\int_{\mathbb{S}^0} (s_A(u) - s_B(u))(s_A(v) - s_B(v)) dK(u,v) \Big)^{1/2}.$$

where $\mathbb{S}^0$ is the unit sphere in $\mathbb{R}$ and $K : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is a positive definite and symmetric kernel such that $K(u,v) = K(-u,-v)$ for any $u, v \in \mathbb{S}^0$. The support function s is an isometry between $\mathscr{K}_c(\mathbb{R})$ and a cone of the Hilbert subspace $\mathscr{L}(\mathbb{S}^0) \subset \mathscr{L}(\mathbb{R})$ endowed with the generic L_2-type distance w.r.t. K. Thus, if $\langle \cdot, \cdot \rangle_K$ denotes the corresponding inner product, it is possible to express the d_K metric on $\mathscr{K}_c(\mathbb{R})$ as $d_K(A,B) = \langle s_A - s_B, s_A - s_B \rangle_K$.

Given a probability space $(\Omega, \mathscr{A}, P)$, a mapping $X : \Omega \to \mathscr{K}_c(\mathbb{R})$ is said to be an *interval-valued random set* associated with $(\Omega, \mathscr{A}, P)$ if the corresponding variables $\inf X$ and $\sup X$ are real random variables. It can be shown that this condition is equivalent to the $\mathscr{A}$-β_{d_H} measurability, where β_{d_H} denotes the σ-field generated by the topology induced by Hausdorff metric d_H on $\mathscr{K}_c(\mathbb{R})$. X can be also characterized by means of the random vector $(\mathrm{mid}X, \mathrm{spr}X)$ where $\mathrm{mid}X = (\sup X + \inf X)/2$ and $\mathrm{spr}X = (\sup X - \inf X)/2$ denote the mid-point and the spread of X, respectively.

If $E(|X|) < \infty$, where $|X|(\omega) = \sup\{|x| : x \in X(\omega)\}$ for any $\omega \in \Omega$, the *expected value of X in Kudō-Aumann's sense* (see [2]), is given by the expression

$$E(X) = \Big\{ E(f) | f : \Omega \to \mathbb{R}, f \in \mathscr{L}^1(\Omega), f \in X a.s.(P) \Big\}.$$

The expected value of an interval-valued random set is an element of $\mathscr{K}_c(\mathbb{R})$, that can be expressed in terms of the classical expectations of the real random variables $\inf X$ and $\sup X$ as $\big[E(\inf X), E(\sup X)\big]$. Furthermore, if $E(|X|^2) < \infty$, the *variance* of X is defined as $\sigma_X^2 = E\Big(\big(d_K(X, E[X])\big)^2 \Big)$ (see [10], [13]). It can be also expressed in terms of the inner product in $\mathscr{L}(\mathbb{S}^0)$ as $\sigma_X^2 = E\Big(\langle s_X - E(s_X), s_X - E(s_X) \rangle_K \Big)$. Finally, the covariance between two random sets X and Y can be defined via support functions as $\sigma_{X,Y} = E\Big(\big\langle s_X - E(s_X), s_Y - E(s_Y) \big\rangle_K \Big)$ whenever this expectation exists.

Let $X, Y : \Omega \longrightarrow \mathscr{K}_c(\mathbb{R})$ be two interval-valued random sets, and $\{X_i, Y_i\}_{i=1}^n$ a simple random sample obtained from (X, Y). The *sample mean* of X is defined by

$\overline{X} = (X_1 + X_2 + \ldots + X_n)/n$. It should be remarked that the Aumann expected value for a random set is coherent with the interval-arithmetic in the sense of the Strong Law of Large Numbers, which means that the preceding concept of sample mean converge a.s.-[P] to the Aumann expectation (see, for instance, [1]). The *sample variance* of X is given by $\hat{\sigma}_X^2 = \overline{d_K(X,\overline{X})^2}$ (analogously $\overline{Y}$ and $\hat{\sigma}_Y^2$). Finally, $\hat{\sigma}_{X,Y}$ denotes the *sample covariance* of X and Y, and it is defined as $\hat{\sigma}_{X,Y} = \overline{\langle s_X - s_{\overline{X}}, s_Y - s_{\overline{Y}} \rangle_K}$.

2.1 Simple Linear Regression Model

The *Simple Linear Regression Model* between X and Y on the basis of the interval-arithmetic approach is formalized as $Y = aX + \varepsilon$, where $a \in \mathbb{R}$ and $\varepsilon : \Omega \longrightarrow \mathscr{K}_c(\mathbb{R})$ is a random set such that $E(\varepsilon|X) = B \in \mathscr{K}_c(\mathbb{R})$ and $\sigma_{\varepsilon,X} = 0$ (see [9], [8]). The population *linear regression function* associated with this model is given by $E(Y|x) = ax + B$ for any $x \in \mathscr{K}_c(\mathbb{R})$.

The theoretical constants of the linear regression function can be expressed in terms of the moments of X and Y as $B = E(Y) -_H aE(X)$ and

$$a = \begin{cases} \dfrac{\sigma_{X,Y}}{\sigma_X^2} & \text{if } a \geq 0 \\ \dfrac{-\sigma_{-X,Y}}{\sigma_X^2} & \text{if } a \leq 0 \end{cases} \tag{1}$$

The estimates for the regression parameters have been obtained in [8]. In this communication we restrict ourselves to the case $a \geq 0$ as a first step. Note that in this way some of the difficulties that entail the lack of linearity of the space $\mathscr{K}_c(\mathbb{R})$ are avoided.

Following the ideas in [8] for the estimation process, we can obtain the corresponding estimates for the particular situation in which $a \geq 0$.

Let (X,Y) be two interval-valued random sets satisfying the considered linear model $Y = aX + \varepsilon$, with $a \geq 0$, and let $\{X_i, Y_i\}_{i=1}^n$ be a simple random sample obtained from (X,Y). Since $Y_i = aX_i + \varepsilon_i$, we have that $Y_i -_H aX_i$ exists for all $i = 1, \ldots, n$, then the estimator of a should be searched within the set

$$\widetilde{A} = \{c \geq 0 : \exists Y_i -_H cX_i, \text{ for all } i = 1 \ldots n\}. \tag{2}$$

The set of feasible solutions $\widetilde{A}$ can be represented by means of a non-empty compact real interval as $[0, \hat{a}^0]$, with $\hat{a}^0 \geq 0$.

The least squares estimation problem is expressed as

$$\text{Minimize } \tfrac{1}{n} \textstyle\sum_{i=1}^n d_K(Y_i, aX_i + B)^2$$

$$\text{subject to } a \in \widetilde{A}.$$

The solutions for this minimization problem, and then, the estimators for the regression model parameters, can be expressed in terms of moments of X and Y as

$$\widehat{a} = \min\left\{\hat{a}^0, \max\left\{0, \frac{\hat{\sigma}_{X,Y}}{\hat{\sigma}_X^2}\right\}\right\} \tag{3}$$

and $\widehat{B} = \overline{Y} -_H \widehat{a}\overline{X}$.

3 Linear Independence Test

Let $X, Y : \Omega \to \mathscr{K}_c(\mathbb{R})$ be two interval-valued random sets such that $Y = aX + \varepsilon$, with $a \geq 0$ and $\varepsilon : \Omega \to \mathscr{K}_c(\mathbb{R})$ fulfilling $E(\varepsilon|X) = B \in \mathscr{K}_c(\mathbb{R})$ and $\sigma_{X,\varepsilon} = 0$.

The aim in this work is to develop a test to determine whether X contributes to explain Y through the linear model or not. Since we have assumed that $a \geq 0$, this is equivalent to test

$$\begin{array}{l} H_0 : a = 0 \\ H_1 : a > 0 \end{array} \tag{4}$$

In this work we propose testing H_0 by means of the statistic

$$T_n = \sqrt{n} \max\left\{0, \frac{\hat{\sigma}_{X,Y}}{\hat{\sigma}_X^2}\right\} \tag{5}$$

Remark 1. From (3), the intuitive statistic for the test would be

$$\widetilde{T}_n = \sqrt{n} \min\left\{\hat{a}^0, \max\left\{0, \frac{\hat{\sigma}_{X,Y}}{\hat{\sigma}_X^2}\right\}\right\}$$

because it uses the information given by the linear model. Unfortunately, the asymptotic behaviour of $\widetilde{T}_n$ is not easy to find, because the term $\hat{a}^0$ is difficult to handle. Nonetheless, given a significance level α and $k \geq 0$ such that $P(T_n > k|H_0) \to \alpha$ as $n \to \infty$, it is possible to check that $P(\widetilde{T}_n > k|H_0)$ is asymptotically lower or equal to α. Thus, the critical region $\{\widetilde{T}_n > k\}$ allows us to solve asymptotically the test (4) by using the statistic $\widetilde{T}_n$ with a significance level $\beta \leq \alpha$.

Remark 2. Both statistics $\widetilde{T}_n$ and T_n depend on $\hat{\sigma}_{X,Y}$, that converges almost-sure to zero under H_0. Indeed, since the random intervals X and Y are linear independent under H_0, then $\sigma_{X,Y} = 0$, and the strong consistency of the covariance guarantees the convergence.

If $0 < \sigma_X, \sigma_Y, \sigma_{X,Y} < \infty$, the asymptotic distribution of $\sqrt{n}\hat{\sigma}_{X,Y}$ under H_0 can be shown to be a normal distribution, with mean value 0 and variance σ_η, where η is the real-valued random variable defined as

$$\eta = \langle s_X - s_{E(X)}, s_\varepsilon - s_{E(\varepsilon)} \rangle_K.$$

Since the sample variance $\hat{\sigma}_X^2$ is consistent w.r.t. σ_X^2, by means of the Slutsky Theorem we obtain that $\sqrt{n}\hat{\sigma}_{X,Y}/\hat{\sigma}_X^2$ converges in law to a distribution $N(0, \sigma_\eta/\sigma_X^2)$.

Finally, since the function $\max\{0, \_\}$ is continuous, by means of the Continuous Function Theorem we can assure that T_n converges in law to the corresponding function of the normal distribution above, that is,

$$T_n \xrightarrow{\mathscr{L}} \max\left\{0, N(0, \sigma_\eta/\sigma_X^2)\right\}.$$

Remark 3. The population variance σ_X^2 is often unknown, so it would be necessary to estimate it by $\hat{\sigma}_X^2$ and then, the obtained asymptotic distribution corresponds to

$$\hat{\sigma}_X^2 T_n \xrightarrow{\mathscr{L}} \max\left\{0, N(0, \sigma_\eta)\right\}$$

For this reason, we could solve the test equivalently with the statistic

$$T_n' = \hat{\sigma}_X^2 T_n = \sqrt{n}\max\{0, \hat{\sigma}_{X,Y}\}$$

whose asymptotic distribution under H_0 does not depend on σ_X^2.

As a result, we can conclude that to test (4) at the nominal significance level α, H_0 should be asymptotically rejected whenever

$$T_n' > \max\{0, z_\alpha\}, \tag{6}$$

where z_α is the $100(1-\alpha)$ fractile of the normal distribution $N(0, \sigma_\eta)$.

Remark 4. In practice, the population variance σ_η^2 is usually unknown, so we should approximate this parameter by its estimator, $\hat{\sigma}_\eta^2$.

4 Simulation Studies

To illustrate the empirical behaviour of the asymptotic procedure suggested in Section 3, some simulations have been carried out. Let X and Y be two interval-valued random sets such that $\mathrm{mid}X, \mathrm{mid}Y \sim N(0,1)$, $\mathrm{spr}X, \mathrm{spr}Y \sim \chi_1^2$ are independent random variables.

Samples of intervals $\{(x_i, y_i)\}_{i=1}^n$ for different sizes n have been generated in order to apply the suggested test. We have developed two different tests. T_1' represents the theoretical test in which the variance of η is known, and T_2' denotes the test in which the population variance of η is aproximated by $\hat{\sigma}_\eta$. In Table 1 we present the percentage of rejections of H_0 at a significance level $\alpha = 0.05$ in 10,000 iterations for each different sample size and each test. The results indicate that the test T_2' is conservative. As expected, in both tests the empirical size is closer to the theoretical one as the sample size increases, although large sample sizes are required in order to obtain suitable results. In addition, T_1' seems to be more accurate than T_2', because T_1' uses the population information instead of the sample one.

Remark 5. In the case of dealing with small samples, asymptotic procedures do not apply. In these situations, alternative techniques should be developed in order to solve the linear independence test considered in this work. For instance, conditions to find the exact distribution of the statistic may be investigated. However, in general they mean

Table 1. Simulation results: empirical size at $\alpha = 0.05$

Sample size	T_1'	T_2'
100	5.25	4.56
200	5.24	4.6
500	5.18	4.62
1000	5.11	4.72
5000	5.03	4.75

the addition of important restrictions to the problem. Bootstrap procedures are another possible way to solve the test more widely applicable.

5 Case-Study: The Blood Pressure Data-Set

In order to show the application of the asymptotic procedure to test the linear independence, we have applied the suggested procedure to a real-life sample data set. Data have been previously used in some works (see, for instance, [6]). They have been supplied by the Hospital Valle del Nalón in Asturias (Spain), and correspond to the range of the systolic X and diastolic Y blood pressure over a day for 59 patients. In Table 2 some of the sample data are presented (full sample data set is available at [6]).

Table 2. Some data of the ranges of systolic (X) and diastolic (Y) blood pressure

X	11.8-17.3	10.4-16.1	13.1-18.6	10.5-15.7	12-17.9	10.1-19.4	...
Y	6.3-10.2	7.1-11.8	5.8-11.3	6.2-11.8	5.9-9.4	4.8-11.6	...

If we test the linear independence between X and Y by using the asymptotic test suggested in Section 3 at nominal significance level $\alpha = 0.05$, we obtain that the value of the typified statistic is $T^* = 6.027$, which is greater than $\max\{0, z_{0.05}\} = 1.645$. Thus, the null hypothesis should be rejected, and we conclude that there is a linear relationship between the fluctuation of the systolic and the diastolic blood pressure in terms of the model considered in this communication.

6 Concluding Remarks

In this communication, an asymptotic procedure for testing the linear independence between two interval-valued random sets by considering a particular case has been suggested. Furthermore, its suitability for large samples has been demonstrated by means of some simulations. It should be underlined that the results are not accurate for moderate and small sample sizes. We are analyzing currently other techniques, like bootstrap procedures, which are often better in these cases.

In the particular case we have analyzed, only positive coefficients for X have been considered. In this way some difficulties due to the lack of linearity of $\mathscr{K}_c(\mathbb{R})$ are avoided. We are also analyzing at present the general case.

Acknowledgement. The research in this paper has been partially supported by the Spanish Ministry of Education and Science Grant MTM2006–07501 and by the Government of Asturias through the Innovation, Technology and Science Plan (PCTI) 2006–2009, PhD Grant BP06–010. Their financial support is gratefully acknowledged.

References

1. Arstein, Z., Vitale, R.A.: A strong law of large numbers for random compact sets. Ann. Probab. 5, 879–882 (1975)
2. Aumann, R.J.: Integrals of set-valued functions. J. Math. Anal. Appl. 12, 1–12 (1965)
3. Blanco, A., Colubi, A., González-Rodríguez, G., Corral, N.: On the consistency of the least-squares estimators for a linear regression model between interval-valued random sets. In: Abstracts of the 56th Session of the International Statistical Institute (ISI 2007, Lisbon, Portugal), p. 406 (2007)
4. De Carvalho, F.A.T., Lima Neto, E.A., Tenorio, C.P.: A new method to fit a linear regression model for interval-valued data. In: Biundo, S., Frühwirth, T., Palm, G. (eds.) KI 2004. LNCS (LNAI), vol. 3238, pp. 295–306. Springer, Heidelberg (2004)
5. Gil, M.A., López, M.T., Lubiano, M.A., Montenegro, M.: Regression and correlation analyses of a linear relation between random intervals. Test 10(1), 183–201 (2001)
6. Gil, M.A., Lubiano, M.A., Montenegro, M., López, M.T.: Least squares fitting of an affine function and strength of association for interval valued data. Metrika 56(2), 97–111 (2002)
7. González-Rodríguez, G., Blanco, A., Colubi, A., Lubiano, M.A.: Estimation of a simple linear regression model for fuzzy random variables. In: Abstracts of the 28th Linz Seminar on Fuzzy Set Theory: Fuzzy sets, Probabilities and Statistics - Gaps and Bridges (LINZ 2007, Linz, Austria), pp. 56–59 (2008)
8. González-Rodríguez, G., Blanco, A., Corral, N., Colubi, A.: Least squares estimation of linear regression models for convex compact random sets. Adv. Data Anal. Classif. 1(1), 67–81 (2007)
9. González-Rodríguez, G., Colubi, A., Coppi, R., Giordani, P.: On the estimation of linear models with interval-valued data. In: Proceedings of the 17th Conference of IASC-ERS (COMPSTAT 2006, Roma, Italy), pp. 697–704 (2006)
10. Körner, R., Näther, W.: On the variance of random fuzzy variables. In: Bertoluzza, C., Gil, M., Ralescu, D. (eds.) Statistical Modeling, Analysis and Management of Fuzzy Data, pp. 22–39. Physica-Verlag, Heildelberg (2002)
11. Hukuhara, M.: Intégration des applications measurable dont la valeur est un compact convexe. Funkcial Ekvac 10, 205–223 (1967)
12. Lima Neto, E.A., De Carvalho, F.A.T., Tenorio, C.P.: Univariate and multivariate linear regression methods to predict interval-valued features. In: Webb, G.I., Yu, X. (eds.) AI 2004. LNCS (LNAI), vol. 3339, pp. 526–537. Springer, Heidelberg (2004)
13. Lubiano, M.A., Gil, M.A., López-Díaz, M., López, M.T.: The $\overrightarrow{\lambda}$-mean squared dispersion associated with a fuzzy random variable. Fuzzy Sets Syst. 111(3), 307–317 (2000)
14. Näther, W.: On random fuzzy variables of second order and their application to linear statistical inference with fuzzy data. Metrika 51(3), 201–221 (2000)