Some random projection based hypothesis tests.

Fraiman Ricardo, Udelar, Uruguay. (rfraiman@cmat.edu.uy)
Moreno Leonardo*, Udelar, Uruguay. (mrleo@iesta.edu.uy)
Vallejo Sebastian, Udelar, Uruguay. (sebastian.vallejo@gmail.com)

Abstract

We propose two new non parametric tests based on continuous one dimensional random projections. The first one deals with central symmetry and the second one with independence. These tests are implemented for finite and infinite dimensional (functional) data sets. Both tests are distribution free and universally consistent. We also propose different techniques to improve the power of the tests. Promising results are obtained comparing the new tests with existing ones through a simulation study. We develop an application with real data in Banach Spaces.

keywords: Random projections, central symmetry, independence tests, universal consistency.

1 Introduction.

In recent years important advances have been obtained on statistics for high dimensional and functional data analysis, in theory development as well as in applications to real data. However, together with the development of the theory many difficulties have emerged regarding the space dimensionality and the extension of classic methods to this paradigm.

At present, a challenging problem is to find statistical methods with enough theoretical background, low computational complexity and easy to apply that behaves efficiently for high or infinite dimensional spaces.

The projection pursuit technique have been increasingly used in the last decades as a way to handle these problems. In particular [1] show an application to symmetry tests. But while this technique extends univariate procedures to the multivariate setting, it is highly sensible to the space dimensionality and it is supported only by empirical results.

One way to solve this problem is provided by [4]. The solution is to work on the basis of a generalization of the Cramér-Wold theorem, through random one dimensional projections. The same authors proposed an application to goodness of fit tests in Hilbert spaces, see [3]. Later [5] extend the results to Banach spaces. The main notion, unlike the pursuit technique, is that is enough to know to determine a probability measure in a multivariate or infinite dimensional space only the distribution of the measure of one projections, under certain assumptions, if it is chosen randomly.

The goal of this work is to propose hypothesis tests for two important problems: *symmetry and independence tests,* using random projections in an efficient way for high and infinite dimensional problems.

2 Basic Concepts and Results.

There are several types of symmetry definitions, which can be found for instance in [10] together with their main properties.

Definition 2.1 (Central Symmetry)**.** *Let E a Banach space. A random element* $\boldsymbol{X}$ *from E has a centrally symmetric distribution around the origin if and only if* $\boldsymbol{X}$ *and* $-\boldsymbol{X}$ *have the same distribution.*

Both problems, symmetry and independence, can be determined through the one dimensional projections of the underlying distribution.

Indeed, if π_h denotes the orthogonal projection of $\mathbb{R}^d$ over the subspace spanned by h (with norm one), and B a Borel set in this subspace, then the induced measure in the subspace is given by,

$$P_{\langle h\rangle}(B) = \mathbb{P}\left[\pi_h^{-1}(B)\right]. \tag{2.1}$$

The following set will be crucial in our approach. Denote by $\mathscr{E}(P,Q) = \{h \in \mathbb{R}^d / P_{\langle h\rangle} = Q_{\langle h\rangle}\}$. First observe that we can rewrite the Cramér Wold Theorem [2], in terms of the set $\mathscr{E}(P,Q)$ by,

$$\mathscr{E}(P,Q) = \mathbb{R}^n \Leftrightarrow P = Q. \tag{2.2}$$

If $\mathbf{X}$ e $\mathbf{Y}$ are random elements in a separable Banach space E, let E^* stands for the dual space and $\mathscr{E}(\mathbf{X},\mathbf{Y}) = \{f \in E^* / \, f(\mathbf{X})$ and $f(\mathbf{Y})$ have the same distribution$\}$, it can be stated a functional version of the Cramér-Wold theorem, see [8],

$$\mathscr{E}(\mathbf{X},\mathbf{Y}) = E^* \Leftrightarrow \mathbf{X} \text{ e } \mathbf{Y} \text{ has the same distribution.} \tag{2.3}$$

Based on this facts, it can be characterized the central symmetry and the independence of the random elements in terms of the distributions of the one dimensional linear functionals.

Theorem 2.1 (Characterization of Central Symmetry)**.** *Let* X *a random element defined over a Banach space E.* X *is centrally symmetric if and only if* $f(X)$ *y* $-f(X)$ *are random variables with the same distribution for any* $f \in E^*$

So is necessary and sufficient to have central symmetry in the original space that the distributions of all univariate projections are symmetrical.

Theorem 2.2 (Characterization of Independence)**.** *Let E a separable Banach space. Two random elements* $\boldsymbol{X}$ *e* $\boldsymbol{Y}$ *in E are independent if and only if* $f(\boldsymbol{X})$ *y* $g(\boldsymbol{Y})$ *are independent for all* $f, g \in E^*$.

Both theorems can be expressed in a Hilbert space $\mathscr{H}$, including the finite dimension case, from a Riesz representation, where the characterization is through orthogonal projections of the random vectors over all the possible directions in $\mathscr{H}$.

3 Random Projections.

In this section it is characterized the symmetry and independence through random projections.

Definition 3.1. *S is a projective hyper-surface in $\mathbb{R}^d$ if and only if exists an homogeneous polynomial $p(x)$ in $\mathbb{R}^d$ such that*

$$S = \{x \in \mathbb{R}^d / p(x) = 0\}. \tag{3.1}$$

Theorem 3.1 (Cuesta, Fraiman and Ransford, 2007)**.** *Let P y Q Borel measures in $\mathbb{R}^d$ where $d \geq 2$. If it holds*

- *P is determined by its moments.*
- *$\mathscr{E}(P,Q)$ it is not contained in a projective hyper-surface in $\mathbb{R}^d$.*

Then $P = Q$

In particular, if the set $\mathscr{E}(P,Q)$ has positive H-measure in $\mathbb{R}^d$, being H an absolutely continuous measure with respect the Lebesgue measure, this set is not contained in any projective hyper-surface. In [5] is generalized the above statement for random elements defined in separable Banach spaces E.

3.1 The Symmetry and Independence Problems.

We derive from the previous results conditions for the symmetry and independence problems of random elements.

Theorem 3.2. *Let μ a Gaussian non degenerate Radon measure in E^*. Let $\boldsymbol{X}$ a random element in E such that,*

- *the absolute moments are finite and verify Carleman's condition,*
- *the set $\mathscr{E}(\boldsymbol{X}) = \{f \in E^* / \ f(\boldsymbol{X})$ is a symmetric random variable$\}$ has positive μ-measure.*

Then $\boldsymbol{X}$ is a central symmetric random element.

Theorem 3.3. *Let $(\Omega, \mathscr{B}, \mathbb{P})$ a probability space, $\boldsymbol{X}$ and $\boldsymbol{Y}$ two random elements taking values on a separable Banach space E, and μ a Radon non degenerate Gaussian measure in $(E \times E)^*$. Assume that the absolute moments $m_X(n)$ and $m_Y(n)$ of $\boldsymbol{X}$ and $\boldsymbol{Y}$ are finite and that the following series diverge,*

$$\sum_{n\geq 1} \min\left\{m_X^{-1/n}(n), m_Y^{-1/n}(n)\right\} = \infty, \tag{3.2}$$

Then, if the set,

$$\mathscr{E}(\boldsymbol{X},\boldsymbol{Y}) = \{h \in (E \times E)^* / \ h(\boldsymbol{X},0) \text{ and } h(0,\boldsymbol{Y}) \text{ are random independent variables}\} \tag{3.3}$$

has positive μ-measure, then $\boldsymbol{X}$ e $\boldsymbol{Y}$ are independent.

4 A Central Symmetry Test.

Let $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n\}$ a set of i.i.d. random elements in a separable Banach space E, whose distribution is determined by its moments. We want to perform a central symmetry test in E, namely

$$\begin{array}{l} H_0)\ \mathbf{X} \text{ and } -\mathbf{X} \text{ have the same distribution} \\ H_1)\ \mathbf{X} \text{ and } -\mathbf{X} \text{ dont have the same distribution} \end{array} \tag{4.1}$$

The proposed methodology is the following:

- It is randomly drawn $h \in E^*$ with a gaussian μ-measure defined as in Theorem 3.2. For the finite dimensional case it suffices to draw a direction h generated by a measure H in $\mathbb{R}^d$ (H absolutely continuous with respect to the Lebesgue measure). In both cases que take $||h|| = 1$.
- Fixed h, we built up the sample of i.i.d. random variables induced by h,

  $\{h(\mathbf{X}_1), h(\mathbf{X}_2), \ldots, h(\mathbf{X}_n)\}$. For finite dimensional spaces, it is orthogonally projected the i.i.d. sample $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n\}$ onto the one dimensional space generated by h, thereby to obtain a sample of random variables in $\mathbb{R}$. We denote $h(\mathbf{X}) = \langle \mathbf{X}, h \rangle$.
- It is performed with this projected data, for a given significance level α, a symmetry test in $\mathbb{R}$ of the Kolmogorov-Smirnov class, developed by [9].

Let $\mathscr{F}_0$ the set of all the symmetric distributions in E. We determine the exact distribution of the statistic under H_0, verifying that it has free distribution, namely, the distribution does not depend on the distribution H used to choose the direction h neither the distribution $F \in \mathscr{F}_0$ of the data. It is also possible to obtain the asymptotic distribution of the statistic under the null hypothesis H_0. The proposed test is universally consistent under any non symmetrical alternative. But for finite sample sizes, in general the test has low power. This problem was already analyzed in [3]. We give some possible solutions to this problem.

5 An Independence Test.

Let $\{(\mathbf{X}, \mathbf{Y}), (\mathbf{X}_1, \mathbf{Y}_1), (\mathbf{X}_2, \mathbf{Y}_2), \ldots, (\mathbf{X}_n, \mathbf{Y}_n)\}$ a set of i.i.d. random elements in $E \times E$, E a separable Banach space, where the distribution of $(\mathbf{X}_1, \mathbf{Y}_1)$ is determined by its moments. We want to develop a test to contrast,

$$\begin{array}{l} H_0)\ \ \mathbf{X} \text{ and } \mathbf{Y} \text{ are independent} \\ H_1)\ \ \mathbf{X} \text{ and } \mathbf{Y} \text{ are not independent.} \end{array} \tag{5.1}$$

The goal, as in the symmetry tests, is to extrapolate statistical methods that are developed for finite dimension. In particular we will use a independence test for infinite dimensional data based in copulas, whose theory was intrinsically conceived for finite dimension. The results are obtained for the space $E \times E$, being E a separable Banach space, but is easy to extend it for the case $E_1 \ldots \times E_k$.

Let $(\mathbf{X}, \mathbf{Y}), (\mathbf{X}_1, \mathbf{Y}_1), \ldots, (\mathbf{X}_n, \mathbf{Y}_n)$ a sample of i.i.d. random elements in $E \times E$, we are interested in contrast if $\mathbf{X}$ e $\mathbf{Y}$ are independent elements.

Under some conditions for the moments and choosing a non degenerate Gaussian measure μ in the dual space $(E \times E)^*$, we will show that if the set,

$$\mathscr{E}(X, Y) = \left\{h \in (E \times E)^* / h(X, 0) \text{ and } h(0, Y) \text{ are independent random variables}\right\},$$

has positive μ-measure, being μ a Gaussian Radon measure in $(E \times E)^*$, then $\mathbf{X}$ e $\mathbf{Y}$ are independent.

The proposed methodology is the following,

- It is drawn $h \in (E \times \ E)^*$, generated by μ. We denote $f(x) = h(x, 0)$ y $g(y) = h(0, y)$. Observe that $f, g \in E^*$.

- Fixed f and g we implement an Independence Test from the empirical copula, $C_n^{f(X), g(Y)}$, of the random variables $f(X)$ y $g(Y)$. We will use the statistic proposed by [7].

  We show that is has asymptotic free distribution and that is consistent under any alternative, based on the work of [6].

We compare by simulations the efficiency of the proposed test with respect to a recent test, very relevant in the study of independence of random vectors, introduced by [12]. This test was adjusted by [11] to perform better in high dimensions.

6 Real Data: Neuronal Activity in Alcoholics.

The database is about EEG waves registered from 120 patients. There were two groups of subjects: alcoholic and control. In our work to have a balanced sample we use 44 subjects of each group.

The objective is to determine the loss of association or synchronization between both hemispheres because of alcohol consumption, and in which areas of the brain the loss is greater.

We used as an association or synchrony measure the p-value of our independence random projection based test. For greater p-value we get more loss of association.

7 Conclusions.

He have developed two tests, one for central symmetry and the other of independence for Banach spaces, obtaining an improved efficiency over the competitors.

Both tests are simply implemented and fast computational performance. In the case of central symmetry test it has been developed a more general than existing for sphericity and ellipticity of the data. Referring to the independence test we could measure the association of more than one random element, as we show in the example with the EEG signals. In this case it has been proved the loss of synchronicity between signals of opposite hemispheres in alcoholic patients.

References

[1] D. K. Blough. Multivariate symmetry via projection pursuit. *Annals of the Institute of Statistical Mathematics*, 41:461–475, 1989.

[2] H. Cramér and H. Wold. Some theorems on distribution functions. *Journal London Mathematical Society*, 11:290–295, 1936.

[3] J. A. Cuesta-Albertos, R. Fraiman, and T. Ransford. Random projections and goodness of fit tests in infinite-dimensional spaces. *Bulletin of the Brazilian Mathematical Society*, 37:477–501, 2006.

[4] J. A. Cuesta-Albertos, R. Fraiman, and T. Ransford. A sharp form of the Cramer–Wold theorem. *Journal of Theoretical Probability*, 20:201–209, 2007.

[5] A. Cuevas and R. Fraiman. On depth measures and dual statistics. A methodology for dealing with general data. *Journal of Multivariate Analysis*, 100:753–766, 2009.

[6] J. Fermaninan. Goodness-of-fit tests for copulas. *Journal of Multivariate Analysis*, 95:119–152, 2005.

[7] C. Genest, J-F. Quessy, and B. Rémillard. Asymptotic Local Efficiency of Cramér-Von Mises Tests for Multivariate Independence. *The Annals of Statistics*, 35:166–191, 2007.

[8] W. J. Padgett and R. L. Taylor. *Laws of Large Number for Normed Linear Spaces and Certain Fréchet Spaces*. Springer-Verlag, 1973.

[9] P. K. Sen and S. K. Chatterjee. On Kolmogorov-Smirnov type test for symmetry. *Annals of the Institute of Statistical Mathematics*, 25:288–300, 1973.

[10] R. Serfling. Multivariate symmetry and asymmetry. In *Encyclopedia of Statistical Sciences, Second Edition*. 2006.

[11] G. J. Székely and M. L. Rizzo. The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis*, 117:193–213, 2013.

[12] G. J. Székely, M. L. Rizzo, and N. Bakirov. Measuring and testing dependence by correlation of distances. *Annals of Statistics*, 35:2769–2794, 2007.