Independence tests for financial variables

Sergio Ortobelli Lozza, Tommaso Lando

***Abstract*— T**his paper proposes an alternative method to evaluate the independence between random variables. The new method is particularly useful when the tested random variables are continuous, because the most used tests for independence are not able to give precise evaluations. In particular, we analyze and compare two different methods to test the independence among financial variables. The first is the classical chi-squared test generally used to evaluate the independence of historical observations in the portfolio risk valuation. The new alternative method is based on a conditional expectation estimator. Thus, we can compare the results of the two methods by evaluating the performance in terms of goodness-of-fit tests.

***Keywords*—**test of independence, conditional expectation, Kernel, Non Parametric test.

I. Introduction

This paper discusses two different methods to test the independence among random variables. On the one hand, several well known methods test independence between random variables by evaluating the independence between their realizations. Clearly, if the events are not independent, this criterion is sufficient to guarantee that the random variables are not independent. Thus, these methods can be properly used for discrete random variables. However, when random variables are continuous, we cannot guarantee that the random variables are independent only if a few events are independent. Moreover, in several financial applications, tests of these kinds are generally used although the financial random variables are assumed to be continuous. For example, when we evaluate the risk interval forecasts, with reference to the information available at each time, we use the tests proposed by [1], [2], and with a chi-squared test we also evaluate the time independence. In this paper, we propose an alternative method to test the independence among random variables, based on the conditional expectation between random variables. As observed by [3], the conditional expectation between two random variables $E(Y|X)$ can be estimated using different methodologies: the Kernel method and the OLP method. On the one hand, the kernel non-parametric regression (see [4] and [5]) allows to estimate $E(Y|X=x)$ as a locally weighted average, based on the choice of an appropriate kernel function: the method yields consistent estimators, provided that the kernel functions and the random variable Y satisfy some conditions, described in Section II. On the other hand, an alternative methodology was recently introduced by [6] for estimating the random variable $E(Y|X)$: this method has been proved to be consistent without requiring any regularity assumption. In this paper we use both methods to evaluate the difference between tests based on the conditional expectation and the classic chi squared test for the independence. In order to compare the effects of the two tests we discuss and examine the case of some financial variables using both alternative methodologies for estimating the random variable $E(Y|X)$. Then, we can perform a simulation analysis, drawing a bivariate random sample from (X,Y), and finally investigate which test better fits to the true case.

The paper is organized as follows: in Section II we present the different methodologies and their properties; in Section III we examine a method to compare the two tests; in Section IV we briefly illustrate the financial interpretation and possible application of the tests of independence.

II. Tests of Independence

In this section, we describe two different procedures to evaluate the independence among random variables. First, we present the well know Pearson chi-square test, used to test independence between random variables and events. Then, the second alternative test is based on the conditional expectation and thereby differs from several other tests which have been proposed for continuous random variables (see [7],[8] and [9]):

The chi-squared independence test

Two random variables X and Y are independent if for any couple of Borel sets A and B then

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B).$$

Therefore, if $X = \sum_{i=1}^{n} a_i I_{[X \in A_i]}$ and $Y = \sum_{j=1}^{m} b_j I_{[Y \in B_j]}$ are discrete random variables (where the collections $\{A_i\}_{i=1,\dots,n}$ and $\{B_j\}_{j=1,\dots,m}$ are partitions of the real line) we can easily test independence using the chi squared test. As a matter of fact, in order to prove the independence of X and Y it is sufficient to show that, for any i=1,…,n and j=1,…,m, we have that

$$p_{i,j} = P(X \in A_i, Y \in B_j) = P(X \in A_i)P(Y \in B_j) = p_i p_j.$$

Hence, we can use the statistic

This paper has been supported by the Italian funds ex MURST 60% 2014, 2015 and MIUR PRIN MISURA Project, 2013–2015, and ITALY project (Italian Talented Young researchers). The research was also supported through the Czech Science Foundation (GACR) under project 13-13142S and through SP2013/3, an SGS research project of VSB-TU Ostrava, and furthermore by the European Regional Development Fund in the IT4Innovations Centre of Excellence, including the access to the supercomputing capacity, and the European Social Fund in the framework of CZ.1.07/2.3.00/20.0296 (to S.O.) and CZ.1.07/2.3.00/30.0016 (to T.L.).

S.O. L. Author is with University of Bergamo, via dei Caniana, 2, Bergamo, Italy; and VŠB -TU Ostrava, Sokolská třída 33, Ostrava, Czech republic; e-mail: sergio.ortobelli@unibg.it.

T.L. Author is with University of Bergamo, via dei Caniana, 2, Bergamo, Italy; and VŠB -TU Ostrava, Sokolská třída 33, Ostrava, Czech republic; e-mail: tommaso.lando@unibg.it.

$$\chi^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(f_{ij}-e_{ij})^2}{e_{ij}} = N\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(p_{ij}-p_i p_j)^2}{p_i p_j} \quad (1)$$

where N is the sample size, p_{ij}, p_i, p_j are the estimated probabilities and similarly f_{ij} is the observed frequency count of the events belonging to both the *i*-th category of *X* and *j*-th category of *Y*, while e_{ij} is the expected count when *X* and *Y* are independent. Thus, the null hypothesis of the independence assumption must be rejected when the p-value of the chi squared statistic (1) (that is chi squared distributed with (m-1)(n-1) degrees of freedom) is less than a given significance level α. Observe that this statistic can also be used to test the independence of continuous random variables. However, in this case, the statistic cannot be applied in order to evaluate whether the random variables are independent, indeed we can only guarantee that the random variables are not independent if the null hypothesis is rejected.

Independence test based on the conditional expected value

Let $X:\Omega\to\mathbb{R}$ and $Y:\Omega\to\mathbb{R}$ be integrable random variables in the probability space $(\Omega,\Im,P)$. As observed by [3] when two integrable random variables *X* and *Y* are independent, then $E(Y|X)$=$E(Y)$ and generally the converse is not true, except in the case that *Y* is positive (negative). Thus, given a positive non constant measurable function g such that $E(g(Y))<\infty$ we can easily test the independence of two integrable random variables *X* and *Y* by considering the variance of $E(g(Y)|X)$. As a matter of fact, the variance of $E(g(Y)|X)$ is equal to zero if and only if *Y* is independent from *X*. Hence, assume that "*Y* is independent from *X*" represents the null hypothesis of the test. We reject the null hypothesis anytime the variance of $E(g(Y)|X)$ is significantly greater than a given positive benchmark value. We call this test *conditional test*. Typically, we consider the function g(x)=|x|. Let $(x_1,y_1),(x_2,y_2),\dots,(x_n,y_n)$ be a random sample of independent observations from the bi-dimensional variable (X,Y). Next, we need an estimator of $E(|Y||X)$. In particular we recall that [3] proposed two alternative estimators of the conditional expected value: the first one is based on the Kernel non-parametric regression, and the other is based on the approximation of the sigma algebra generated by *X*. Thus the first procedure is aimed at estimating the conditional expectation of $|Y|$ given $X=x$, which is a mathematical function of X; the second method yields an unbiased and consistent estimator of the random variable $E(|Y||X)$.

The kernel non-parametric regression. It is well known that, if we know the form of the function $f(x)=E(|Y||X=x)$ (e.g. polynomial, exponential, etc.), then we can estimate the unknown parameters of $f(x)$ with several methods (e.g. least squares). In particular, if we do not know the general form of $f(x)$, except that it is a continuous and smooth function, then we can approximate it with a non-parametric method, as proposed by [4] and [5]. Thus, $f(x)$ can be estimated by:

$$\hat{f}_n(x)=\frac{\sum_{i=1}^{n}|y_i|K\left(\frac{x-x_i}{h(n)}\right)}{\sum_{i=1}^{n}K\left(\frac{x-x_i}{h(n)}\right)}, \quad (2)$$

where $K(x)$ is a density function such that i) $K(x)<C<\infty$; ii) $\lim_{x\to\pm\infty}|xK(x)|=0$; iii) $h(n)\to 0$ when $n\to\infty$. The function $K(x)$ is denoted by *kernel*, observe that kernel functions are generally used for estimating probability densities non-parametrically (see [10]). It was proved in [10] that if $|Y|$ is quadratically integrable then $\hat{f}_n(x)$ is a consistent estimator for $f(x)$.

The OLP method. We now describe an alternative non-parametric approach [6] for approximating the conditional expectation, the method is denoted by "*OLP*", which is an acronym of the authors' names. Define by $\Im_X$ the σ-algebra generated by X (that is, $\Im_X=\sigma(X)=X^{-1}(\mathcal{B})=\{X^{-1}(B):B\in\mathcal{B}\}$, where $\mathcal{B}$ is the Borel σ-algebra on $\mathbb{R}$). Observe that the regression function is just a "pointwise" realization of the random variable $E(|Y||\Im_X)$, which can equivalently be denoted by $E(|Y||X)$. $\Im_X$ can be approximated by a σ-algebra generated by a suitable partition of Ω. In particular, for any $k\in\mathbb{N}$, we consider the partition $\{A_j\}_{J=1}^{b^k}=\{A_1,\dots,A_{b^k}\}$ of Ω in b^k subsets, where b is an integer number greater than 1 and:

- $A_1=\left\{\omega:X(\omega)\le F_X^{-1}\left(\frac{1}{b^k}\right)\right\}$,
- $A_h=\left\{\omega:F_X^{-1}\left(\frac{h-1}{b^k}\right)<X(\omega)\le F_X^{-1}\left(\frac{h}{b^k}\right)\right\}$, $for\ h=2,\dots,b^k-1$
- $A_{b^k}=\Omega-\bigcup_{J=1}^{b^k-1}A_j=\{\omega:X(\omega)>F_X^{-1}\left(\frac{b^k-1}{b^k}\right)\}$.

Starting with the trivial sigma algebra $\Im_0=\{\emptyset,\Omega\}$, we can obtain a sequence of sigma algebras generated by these partitions, for different values of *k* (*k*=1,…,*m*,…). For instance, $\Im_1=\sigma\{\emptyset,\Omega,A_1,\dots,A_b\}$ is the sigma algebra generated by $A_1=\{\omega:X(\omega)\le F_X^{-1}(1/b)\}$, $A_s=\left\{\omega:F_X^{-1}\left(\frac{s-1}{b}\right)<X(\omega)\le F_X^{-1}\left(\frac{s}{b}\right)\right\}$, *s*=1,...,*b*-1 and $A_b=\{\omega:X(\omega)>F_X^{-1}((b-1)/b)\}$. Generally:

$$\Im_k=\sigma\left(\{A_j\}_{J=1}^{b^k}\right),k\in\mathbb{N}. \quad (3)$$

Hence, it is possible to estimate the random variable $E(Y|\Im_X)$ by

$$E(|Y||\Im_k)(\omega)=\sum_{j=1}^{b^k}\frac{1_{A_j}(\omega)}{P(A_j)}\int_{A_j}|Y|dP=\sum_{j=1}^{b^k}E(|Y||A_j)1_{A_j}(\omega), \quad (4)$$

where $1_A(\omega)=\begin{cases}1 & \omega\in A\\ 0 & \omega\notin A\end{cases}$. It is proved in [6] that $E(|Y||\Im_k)$ is a consistent estimator of the random variable $E(|Y||X)$, that is, $\lim_{k\to\infty}E(|Y||\Im_k)=E(|Y||X)$ a.s.

From a practical point of view, given *n* i.i.d. observations of Y, if we know the probability p_i corresponding to the *i*-th outcome y_i, we obtain:

$$E(|Y||A_j)=\sum_{y_i\in A_j}|y_i|p_i/P(A_j). \quad (5)$$

Otherwise, we can give uniform weight to each observation, which yields the following consistent estimator of $E(|Y||A_j)$:

$$\frac{1}{n_{A_j}}\sum_{y_i\in A_j}|y_i|, \quad (6)$$

where n_{A_j} is the number of elements of A_j. Therefore, we are always able to estimate $E(|Y||\Im_k)$, which in turn is a consistent estimator of the conditional expected value $E(|Y||X)$.

A simple proof of the potentiality of the test can be given when we compare uncorrelated but dependent random variables as in the following section.

III. A COMPARISON AMONG TWO PORTFOLIOS

Let us consider two portfolios of daily returns X and Y, taken from the NYSE, which are empirically uncorrelated.[1] Consider that we have about three years of historical daily joint observations (750 trading days). First of all, we want to test if the losses and gains of the two portfolios are independent. Using the chi square test with one degree of freedom we could not reject the independence of the two portfolios at 95% significance level. Secondly we want to test if the two portfolios are independent. Thus, we apply the conditional test to the standardized random variables $\tilde{X}$ and $\tilde{Y}$ of X and Y. We get a variance of $E(|\tilde{Y}|/\tilde{X})$. equal to 0.0512 with the OLP estimator and 0.0445 with the Kernel estimator. We observe that the joint distribution of the two standardized portfolios can be well approximated by a bivariate t-student with 5 degrees of freedom. Thus, with a bootstrap technique based on bivariate t-student, we estimated the variance obtained for a sample of the same dimension (750 observations) under two different hypotheses: X and Y are independent t- distributed or X and Y are dependent but uncorrelated. For independent t distributed random variables we get an average variance of $E(|\tilde{Y}|/\tilde{X})$ equal to 0.0082, while for uncorrelated dependent t distributed random variables we get an average variance of $E(|\tilde{Y}|/\tilde{X})$ equal to 0.0431. This simple observation suggests to reject the independence hypothesis even if the two portfolio are uncorrelated.

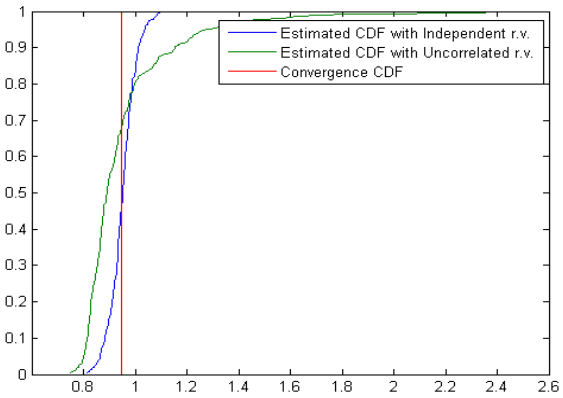

FIG. 1 Distributions of $E(|Y|/X)$ for uncorrelated or independent t-student random variables.

One further example of this analysis is given in Fig. 1 where we report the distributions of $E(|\tilde{Y}|)$ and of $E(|\tilde{Y}|/\tilde{X})$ (estimated with the OLP method) assuming $\tilde{X}$ and $\tilde{Y}$ to be uncorrelated or independent t distributed with 5 degrees of freedom.

[1] The procedure to get portfolio uncorrelated is very simple and can be useful in several practices for example in the PCA to reduce the dimensionality of the problem.

IV. CONCLUSION

In this paper, we deal with tests of independence among random variables. In particular, we show that the well known chi squared test for independence is not always able to evaluate correctly the independence between random variables. On the other hand, a newly proposed test is able to capture the dependence of random variables even when they are uncorrelated. In particular, we show that the new test could be based on two different methodologies for estimating the conditional expectation, namely the kernel method and the OLP method recently proposed by [6].

REFERENCES

[1] Christoffersen P. Evaluating interval forecasts. International Economic Review 1998; 39; 841-862.
[2] Kupiec P. Techniques for verifying the accuracy of risk measurement models, Journal of Derivatives 1995; 3; 73-84
[3] Lando T., Ortobelli S. (2015) "On the Approximation of a Conditional Expectation" WSEAS Transactions on Mathematics, Volume 14, pp. 237-247
[4] E. A. Nadaraya, "On estimating regression," *Theory of Probability and its Applications*, vol. 9, no. 1, pp. 141-142, 1964.
[5] G. S. Watson, "Smooth regression analysis," *Sankhya, Series A*, vol. 26, no. 4, pp. 359-372, 1964.
[6] S. Ortobelli, F. Petronio, T. Lando, "A portfolio return definition coherent with the investors preferences," under revision in IMA-Journal of Management Mathematics.
[7] K. Sricharan,R. Raich, and A. Hero. 2012 Empirical estimation of entropy functionals with confidence. *Technical report* arXiv:1012.4188
[8] A. Gretton, O. Bousqet, A. Smola, and B. Schölkopf, (2005) Measuring Statistical Independence with Hilbert-Schmidt Norms, *Algorithmic Learning Theory Lecture Notes in Computer Science* 3734, pp 63-77 .
[9] B. Schweizer and E.F. Wolff, (1981) On Nonparametric Measures of Dependence for Random Variables, *Annals of Statistics* 9(4), 879-885.
[10] V. A. Epanechnikov, "Non-parametric estimation of a multivariate probability density," *Theory of Probability and its Applications*, vol. 14, no. 1, pp. 153-158, 1965.