On the Approximation of a Conditional Expectation

TOMMASO LANDO
Dipartimento di Scienze aziendali, economiche e metodi quantitativi
University of Bergamo
Via dei Caniana 2, Bergamo
ITALY
tommaso.lando@unibg.it

SERGIO ORTOBELLI
Dipartimento di Scienze aziendali, economiche e metodi quantitativi
University of Bergamo
Via dei Caniana 2, Bergamo
ITALY
sergio.ortobelli@unibg.it

Abstract: - In this paper, we discuss how to approximate the conditional expectation of a random variable *Y* given a random variable *X*, i.e. *E*(*Y*|*X*). We propose and compare two different non parametric methodologies to approximate *E*(*Y*|*X*). The first approach (namely the OLP method) is based on a suitable approximation of the σ-algebra generated by *X*. A second procedure is based on the well known kernel non-parametric regression method. We analyze the convergence properties of the OLP estimator and we compare the two approaches with a simulation study.

Key-Words: - Conditional Expectation, Kernel, Non Parametric methods, Regression, Approximation, Simulation.

1 Introduction

This paper discusses different methods to estimate the conditional expected value. Let Y be a random variable with finite mean, and let X be some other random variable defined on the same probability space. The conditional expectation $E(Y|X)$ is a random variable and it is a function of X. In particular, $E(Y|X)$ can be intuitively interpreted as the function of X that "best" approximates Y, in that it represents the best approximation as to the value of Y, given the only value of the random variable X. On the one hand, several well known methods are aimed at estimating the regression function $g(x) = E(Y|X = x)$, which represents just a realization of $E(Y|X)$, namely: parametric regression methods, semi-parametric and non-parametric regression methods such as kernel regression [1],[2], smoothing splines [3], or various generalizations of these models, see e.g. [4]. On the other hand, in many real world problems (see e.g.[5]) we are interested in approximating the random variable $g(X) = E(Y|X)$, and estimating its distribution function. Provided that $g(X)$ has a density with respect to the Lebesgue measure, a method to estimate the density function of $g(X)$ have been recently introduced in the literature by Steckey and Henderson [6] (see also [7]). In particular, this method is based on a sort of conditional sampling which consists in i) sample X; and ii) sample Y from the conditional distribution of Y given $X = x$. Then, it is possible to estimate the density of $g(X)$ with the kernel method. Nevertheless, we observe that in several situations it would not be possible to satisfy these sampling assumption, as we only have available a bivariate random sample from (X, Y). Therefore, in this paper we attempt to estimate the distribution of $g(X)$ simply using a random sample of independent observations $(x_1, y_1), \dots, (x_n, y_n)$ from the bi-dimensional variable (X, Y). Obviously, if we had available a sample of independent and identically distributed random variables from $g(X)$, then it would be trivial to estimate the distribution of $g(X)$. Hence, our idea is that we can use the observations $(x_1, y_1), \dots, (x_n, y_n)$ form (X, Y) in order to generate vector of outcomes $(g_1, \dots, g_n)$, which approximate the realizations of $g(X)$. For this aim, we propose to use two different methods, namely the OLP method, recently introduced by [8], and the well known kernel method, as recently suggested by [9]. The OLP method consists in approximating the sigma algebra generated by X (denoted by $\sigma(X)$) with a sigma algebra generated by a suitable partition of the sample space, according to a given number $(k - 1)$ of percentiles of X. Hence, by averaging the observed values of X over the above defined intervals, we can approximate the random variable $g(X)$ and thereby its distribution function. Differently, the kernel non-parametric regression (see [1] and [2]) allows to estimate $E(Y|X = x)$ as a locally weighted

average, based on the choice of an appropriate kernel function. Therefore, by applying the kernel method to each observation of X ($X = x_1, X = x_2$, etc.), we obtain n outcomes, which can be similarly used to estimate the distribution of the random variable $g(X)$. In this paper we compare the two methods with a simulation analysis. Then we study the properties of the OLP estimator and propose some practical rules to enhance its performance. In particular, while it is well known that the kernel method depends on the choice of the kernel function and the bandwidth parameter, the OLP method depends on the choice of the number of intervals k, used for approximating $\sigma(X)$. The choice of k is crucial in order to obtain an accurate approximation. First, we propose a rule for determining k under general assumptions, and then we compare the kernel and OLP methods with a simulation study, under assumption of normality. Indeed, if we know the joint distribution of (X,Y) and the true distribution F of $E(Y|X)$ (which, for instance, can be easily computed in the Gaussian case), then we can generate a bivariate random sample from (X,Y), and finally investigate which estimated distribution better fits to F. In the Gaussian case, the performance of the kernel method can be optimized quite easily (in terms of kernel density function and bandwidth parameter). Thus, we compare the "optimal" kernel method with the OLP method, where the number of intervals k is determined without using any information on the joint distribution. The results show that, even in this "adverse" situation for the OLP method, the two methods provide comparable outputs. In the last section we study further properties the OLP estimator in the Gaussian case. In particular, we argue that the performance of the OLP estimator can be further enhanced if the number of intervals used for approximating $\sigma(X)$ is determined according to the correlation between the variables. In the last section we briefly summarize the paper and we propose some possible financial applications.

2 Problem Formulation

Let Y be an integrable random variable on the probability space $(\Omega, \Im, P)$ and let $\Im'$ be a sub-sigma-algebra of $\Im$ (i.e. $\Im' \subseteq \Im$). The conditional expectation of Y given $\Im'$ is the unique (P a.s.) random variable $E(Y|\Im')$ such that:

i) $E(Y|\Im')$ is $\Im'$-measurable;

ii) $\forall A \in \Im', \int_A E(Y|\Im')dP = \int_A YdP$.

Let $X: \Omega \to \mathbb{R}$ and $Y: \Omega \to \mathbb{R}$ be integrable random variables in the probability space $(\Omega, \Im, P)$. When $\Im' = \sigma(X)$ is the sigma algebra generated by X we write $E(Y|\sigma(X)) = E(Y|X) = g(X)$. Generally, the distribution of $g(X)$ is unknown, unless the joint distribution of the random vector (X,Y) follows some special distribution, e.g. the Gaussian distribution or the multivariate t distribution (note that, except for trivial text book examples, an exact expression for $g(X)$ is rare). However, if we assume that X and Y are jointly normally distributed, i.e. $(X,Y) \sim N(\mu, \Sigma)$, (where obviously $\mu = (\mu_X, \mu_Y)$ is the vector of the means, and $\Sigma = \left((\sigma_X{}^2, \rho_{XY}\sigma_X\sigma_Y), (\rho_{XY}\sigma_X\sigma_Y, \sigma_Y{}^2)\right)$ is the variance-covariance matrix[1]) we can obtain the distribution of the random variable $E(Y|X)$ quite easily. Indeed, it is well known that $g(x) = E(Y|X = x) = \mu_Y + \rho_{XY}\frac{\sigma_Y}{\sigma_X}(x - \mu_X)$, and thus,

$$E(Y|X) = \mu_Y + \rho_{XY}\frac{\sigma_Y}{\sigma_X}(X - \mu_X) \qquad (1)$$

is still Gaussian distributed with mean μ_Y and variance $\rho_{XY}^2\sigma_Y^2$. Clearly, when the correlation of a couple of random variables (X,Y) is $\rho_{XY} = \pm 1$, then $Y = \mu_Y + \rho_{XY}\frac{\sigma_Y}{\sigma_X}(X - \mu_X)$ P almost surely and equation (1) holds for any joint distribution of the vector (X,Y). Equation (1) holds also for joint Student's t bivariate vector, as pointed out by [10].

Basically, if the bivariate random vector (X,Y) is Gaussian or t-distributed, we also know the general form of the distribution of $g(X)$, and therefore we can estimate it quite easily. For instance, we could approximate $g(X)$ by estimating the unknown parameters μ_Y, σ_{XY} and σ_X^2 respectively with the sample mean, the sample covariance coefficient and the sample variance.

Unfortunately, in most of the cases we do not know the form of the distribution of $g(X)$, thus we cannot estimate it with parametric methods. Moreover, we cannot even use non parametric methods, unless having available a random sample drawn from the random variable $g(X)$, which is surely uncommon and difficult to obtain. For these reasons, the aim of this paper is to provide a method for estimating $g(X)$ and its distribution using a "standard" bivariate random sample $(x_1, y_1), \ldots, (x_n, y_n)$ of independent observations from the bi-dimensional

[1] For simplicity, in this paper we write, with a little abuse of notation, the dispersion matrix of (X,Y) as:

$\Sigma = \begin{bmatrix} \sigma_X{}^2 & \sigma_{XY} \\ \sigma_{YX} & \sigma_Y{}^2 \end{bmatrix} = \left((\sigma_X{}^2, \sigma_{XY}), (\sigma_{YX}, \sigma_Y{}^2)\right)$.

variable (X,Y). For this purpose, we use two different methods, namely the OLP method, recently introduced by [8], and the kernel non parametric regression method.

2.1 The OLP Method

The OLP method has been recently introduced by [8] (see also [11]) to approximate the conditional expectation, based on an appropriate partition of the sample space. The method, as defined in [8], requires the knowledge of the probability measure P: in this paper we do not rely on this assumption and we propose an "estimator" for the random variable $g(X)$.

Define by $\sigma(X)$ the σ-algebra generated by X (that is, $\sigma(X) = X^{-1}(\mathcal{B}) = \{X^{-1}(B): B \in \mathcal{B}\}$, 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 $E(Y|\sigma(X))$. The following methodology is aimed at approximating $E(Y|X)$ rather than estimating $g(x)$. The σ-algebra $\sigma(X)$ 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 $\{A_j\}_{J=1}^{k} = \{A_1, \dots, A_k\}$ of Ω in k subsets, described as follows:

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

The partition $\{A_j\}_{J=1}^{k}$ is practically determined by a number $(k-1)$ of percentiles of X. Furthermore, note that, by definition of percentile, each interval A_j have equal probability, that is, $P(A_j) = 1/k$, for $j = 1, \dots, k$. Starting with the trivial sigma algebra $\Im_1 = \{\emptyset, \Omega\}$, we can obtain a sequence of sigma algebras generated by these partitions, for different values of *k*. Generally:

$$\Im_k = \sigma\left(\{A_j\}_{J=1}^{k}\right), k \in \mathbb{N}. \quad (2)$$

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

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

$$= \sum_{j=1}^{k} E(Y|A_j) 1_{A_j}(\omega), \quad (3)$$

where $1_A(\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \omega \notin A \end{cases}$. Indeed, by definition of the conditional expectation, observe that $E(Y|\Im_k)$ is a $\Im_k$-measurable function such that, for any set $A \in \Im_k$, (that can be seen as a union of disjoint sets, in particular $A = \cup_{A_j \subseteq A} A_j$) we obtain the equality

$$\int_A E(Y|\Im_k) dP = \int_A Y(\omega) dP\,(\omega). \quad (4)$$

It is proved in [8] that $E(Y|\Im_{k(h)})$ converges almost certainly to the random variable $E(Y|X)$, that is:

$$\lim_{k\to\infty} E(Y|\Im_k) = E(Y|X) \text{ a.s.}. \quad (5)$$

Hence, if we approximate $E(Y|\Im_k)$, then we also approximate $g(X)$, for sufficiently large k. However, in practical situations, we do not know the probability measure P used to approximate $E(Y|A_j)$ in (3). Hence, in this paper, we propose to approximate the random variable $E(Y|\Im_k)$, which in turns approximates $E(Y|X)$, based on the observations of a random sample. 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). First, as we generally do not know the marginal distribution of X, we can determine the partition $\{\hat{A}_j\}_{J=1}^{k}$ using the percentiles of the empirical distribution, obtained from the observations $(x_1, \dots, x_n)$. The number of intervals k should be basically an increasing function of the number of observations n, as discussed below. Then, if we assume to know the probability p_i, corresponding to the *i*-th outcome y_i, we obtain:

$$E(Y|\hat{A}_j) = \sum_{x_i \in \hat{A}_j} y_i p_i / P(\hat{A}_j). \quad (6)$$

Otherwise, we can give uniform weight to each observation, and thus we can use the following estimator of $E(Y|A_j)$:

$$\hat{a}_j = \frac{1}{n_{\hat{A}_j}} \sum_{x_i \in \hat{A}_j} y_i, \quad (7)$$

where n_{A_j} is the number of observations in $\hat{A}_j$, that is, $n_{A_j} = \#\{x_i: x_i \in A_j, i = 1, .., n\} \cong n/k$ (to clarify the explanation, for $k = 4$ we obtain the three quartiles, and therefore $n_{A_j} \cong \frac{n}{4}$ and similarly $P(A_j)$ can be estimated by $\frac{1}{4}$). Note that, fixed k, as the number of observations n grows, $P(\hat{A}_j) \xrightarrow{n\to\infty} P(A_j) = 1/k$ and $\hat{a}_j$ is an asymptotically unbiased estimator of $E(Y|A_j)$:

$$E(\hat{a}_j) = \frac{1}{n_{\hat{A}_j}} \sum_i E(Y_i \, 1_{X_i \in \hat{A}_j}) =$$

$$= \frac{\int_{X \in \hat{A}_j} y dP}{1/k} \xrightarrow{n \to \infty} E(Y|A_j). \quad (8)$$

Therefore, we are always able to approximate $E(Y|\Im_k)$, and thereby the conditional expectation $E(Y|X)$, by using the following estimator :

$$\hat{g}_{OLP}(X) = \sum_{j=1}^{k} 1_{X \in \hat{A}_j} \sum_{x_i \in \hat{A}_j} y_i \frac{1}{n_{\hat{A}_j}} =$$

$$= \sum_{j=1}^{k} 1_{X \in \hat{A}_j} \hat{a}_j. \quad (9)$$

where X is assumed independent from the i.i.d. observations (x_i, y_i). Note that $\hat{g}_{OLP}$ is a simple $\Im_k$ measurable function, and it is conceptually different from the classical estimators, which are generally aimed at estimating an unknown parameter rather than a random variable. A further property of the OLP estimator is that $E(\hat{g}_{OLP}(X)) = E(Y)$:

$$E(\hat{g}_{OLP}(X)) = \sum_{j=1}^{k} E\left(1_{X \in \hat{A}_j} \hat{a}_j\right) =$$

$$= P(\hat{A}_j) \sum_{j=1}^{k} E\left(\hat{a}_j\right) =$$

$$= P\left(\hat{A}_j\right) \sum_{j=1}^{k} E(Y|\hat{A}_j) = E(Y), \quad (10)$$

because it satisfies the basic properties of the conditional expectation, that is, $E(E(Y|\Im_k)) = E(Y)$.

Observe that, given a bivariate sample of size n, the OLP estimator yields k distinct values for $\hat{g}_{OLP}(x_i)$,, i.e. the $\hat{a}_j$'s, where each one has frequency $n_{\hat{A}_j} \cong n/k$, for $j = 1, \dots, k$. These outcomes can be used to estimate the unknown distribution function of $g(X)$.

2.2 The Kernel Method

The kernel method, typically used to estimate the probability density of an unknown random variable (see, for instance, [11]), can also be applied to estimate the regression function $g(x) = E(Y|X = x)$. In particular, if we do not know the general form of $g(x)$, except that it is a continuous and smooth function, then we can consider the following kernel estimator:

$$\hat{g}_n(x) = \frac{\sum_{i=1}^{n} y_i K\left(\frac{x - x_i}{h(n)}\right)}{\sum_{i=1}^{n} K\left(\frac{x - x_i}{h(n)}\right)}, \quad (11)$$

where $K(x)$, denoted by *kernel*, is a density function (typically unimodal and symmetric around zero) such that i) $K(x) < C < \infty$; ii) $\lim_{x \to \pm\infty} |xK(x)| = 0$ (see, among others, [1] and [2]). Moreover, $h(n)$ is the smoothing parameter, often referred to as the *bandwidth* of the kernel, and it is a positive number such that $h(n) \to 0$ when $n \to \infty$. When the kernel K is the probability density function of a standard normal distribution, then the bandwidth is the standard deviation. It was proved in [1] that if Y is quadratically integrable (see also [12]) then $\hat{g}_n(x)$ is a consistent estimator for $g(x)$. In particular, observe that, if we denote by $f(x, y)$ the joint density of (X, Y), the denominator of (11) converges to the marginal density of X, while the numerator converges to $\int_{-\infty}^{\infty} \int_{\{X=x\}} yP(dx, dy)$. As a consequence, we know that $\hat{g}_n(X) \to_{a.s.} g(X)$. From a practical point of view, if we apply the kernel estimator to the bi-variate random sample $(x_1, y_1), \dots, (x_n, y_n)$ we obtain the vector $(g_1, \dots, g_n) = (\hat{g}_n(x_1), .., \hat{g}_n(x_n))$. In other words, each value g_i is a weighted average of kernels, centered at each sample observation x_i. Since we know that $g_i \to E(Y|X = x_i)$ when $n \to \infty$, then we can also estimate the distribution function (say $F(x) = P(g(X) \leq x)$) of $g(X)$ with any parametric or non-parametric method, based on the outcomes $(g_1, \dots, g_n)$.

3 A Simulation Comparison

In this section, we compare the OLP and the Kernel method with a simulation study. It is worth noting that the comparison between these methods is not really balanced. Indeed, the main difference between the two procedures is that the OLP method generates k distinct outputs, each one with frequency n/k, while the kernel method generally yields n different outputs, one for each observation of X. Hence, if the kernel density and the bandwidth parameter are suitably chosen, then the kernel method should outperform the OLP method in terms of accuracy. On the other hand, the OLP estimator yields a set of distinct outputs $(g_1, \dots, g_k)$ where each value g_j is a conditional average over the set $\hat{A}_j$. Therefore, the values g_j's are generally robust estimates.

As pointed out in section 2, if we know that the random vector (X, Y) is jointly Gaussian (or *t*-

distributed), then we also know the true distribution of $g(X)$. The main motivation of this study is that, if we show that a method provides a good estimate in the normal case, when $g(X)$ is known, then we argue that it can provide similar results in many other cases, when the distribution of $g(X)$ is unknown. This is especially true for the OLP method, which does not depend on any particular specification except from the choice of the number of intervals k. Hence, assuming that $(X,Y) \sim N(\mu, \Sigma)$, (where $\mu = (\mu_X, \mu_Y)$ and $\Sigma = \left((\sigma_X{}^2, \rho\sigma_X\sigma_Y), (\rho\sigma_X\sigma_Y, \sigma_Y{}^2)\right)$) we propose to simulate a bivariate random sample from (X,Y) and to apply the OLP and the Kernel methods to the observed data, just as described in section 2, in order to evaluate which method yields a better approximation of the distribution of $g(X)$. Indeed, in both cases we obtain n outcomes $(g_1, \dots, g_n)$, which are used to estimate the probability distribution $F(x) = P(g(X) \le x)$ of the r.v. $g(X)$ (i.e. the Gaussian distribution $N(\mu_Y, |\rho|\sigma_Y)$ in this particular case). For this purpose, we simply apply the empirical distribution function to the vector $(g_1, \dots, g_n)$. The empirical distribution is actually the natural consistent estimator of F (see e.g. [13]), and it is defined by

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{g_i \le x\}}, \qquad (12)$$

where 1_A is the indicator function for the set A. In this paper we propose to use $\hat{F}_n$ as a non-parametric estimator for the distribution of $g(X)$. Obviously, according to how the outcomes $(g_1, \dots, g_n)$ are generated (OLP or Kernel) we obtain two different empirical distributions $\hat{F}_n$'s, which approximate the true distribution, given by $N(\mu_Y, |\rho|\sigma_Y)$. Then, in order to evaluate which method provides the best fitting distribution, we compute two different descriptive measures of fit based on probability distances, namely the Kolmogorov-Smirnov (or uniform) metric, defined by

$$D\left(\hat{F}_n, F\right) = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)|, \qquad (13)$$

and the Kantorovich metric [14] (i.e. the L^1 metric for distribution functions), defined by

$$K\left(\hat{F}_n, F\right) = \int_{-\infty}^{\infty} |\hat{F}_n(x) - F(x)| dx. \qquad (14)$$

Generally, provided that both methods can capture the shape of F, we would expect that the kernel method yields a better fit, because of its larger number of distinct outcomes g_i's. Indeed, a continuous distribution is typically better estimated by a large number of distinct observations. However, it should be stressed that the OLP method has several other advantages compared to the Kernel method. While the OLP method only requires that Y is an integrable random variable, the kernel method is suitable only for continuous random variables and requires also the assumption of finite variance. Moreover, for the OLP method we only need to specify how to determine the number of intervals k, while, for the kernel method, we have to choose the "best" kernel density and bandwidth parameter. In particular, for the proposed analysis we used the following specifications.

i) OLP - number of intervals

Obviously, the selected number of intervals k can vary between 1 and n and, in order to improve the accuracy of the estimate it must generally be an increasing function of n (we shall discuss this point in the next section). Note that, for $k = 1$ we approximate the random variable $g(X)$ with a number, i.e. the sample mean $\bar{y}$, which is obviously not appropriate. On the other hand, for $k = n$ we approximate $g(X)$ with the marginal distribution of Y, given by $y_1, .., y_n$, which is also generally inappropriate. Hence, in order to maximize i) the number of intervals, and ii) the number of observations in each interval $(n_{\hat{A}_j})$, in this analysis we propose to use:

$$k = \lfloor \sqrt{n} \rfloor, \qquad (15)$$

where $\lfloor x \rfloor$ is the integer part of x. By doing so, we obtain k intervals containing (approximately) k observations. If we do not have any information about the dependence between X and Y, this method is actually the most robust, in that it provides the largest possible number of conditional averages $\hat{a}_j$, where each $\hat{a}_j$ is computed based on the largest possible number of values $(n_{\hat{A}_j})$. In the next section we prove that the rule identified by (14) is actually appropriate and yields a consistent estimator. Moreover, in section 4 we propose a new rule for the choice of k, based on the correlation value between X and Y, that might further enhance the performance of the OLP estimator.

ii) Kernel – density and bandwidth

Generally, the choice of the kernel density and especially the bandwidth parameter can be really troublesome and this could be a drawback of the method. Indeed, there are several sophisticated techniques to choose the

optimal bandwidth, which is still an open problem in the literature (see e.g. [15]). Since we know that (X,Y) is jointly normally distributed, we simply use the normal kernel, which is obviously the most appropriate choice in this particular case. As for the bandwidth parameter, we can use the Sturge's or the Scott's rule (see [16] and [17]) which are especially suitable under normality assumptions (see also [18]). In particular, in what follows we shall show just the results obtained by applying the Scott's rule, because it provided better approximations of F in our analyses. We recall that the optimal bandwidth, according to the Scott's rule, is given by $3.5\sigma_X n^{-1/3}$.

Note that, in the Gaussian case, the distribution of the conditional expectation, that is, $g(X)\sim N(\mu_Y, |\rho|\sigma_Y)$, mainly depends on the correlation between the variables. Hence, we generated several random samples of different sizes from $(X,Y)\sim N(\mu,\Sigma)$ (where we assume the marginals be standard normal random variables, i.e., $\mu_X=\mu_Y=0,\quad \sigma_X=\sigma_Y=1$), for different (positive) values of ρ, and analyzed the results accordingly. The Table 1 shows the results in terms of the probability metrics defined above. First, note that the K-S distance is quite high (about 0.5) for $\rho=0$. The obvious reason is that $\rho=0$ yields $g(X)=_d E(Y)$, which is a degenerate distribution (at 0, in this case) and therefore the Kolomogorov-Smirnov metric $D\left(\hat{F}_n, 1_{\{x\geq 0\}}\right)$ is generally close to 0.5, while the Kantorovich metric, which is based on the area between the functions, better captures the distance between the distributions in this particular case. Note also that the consistency of both methods is apparent from tables 1 and 2, in that increasing the sample size (from $n=500$ in Table 1 to $n=10^5$ in Table 2) the distance between the true and the estimated distribution approaches zero, for any fixed value of ρ (except for the Kolomogorov-Smirnov distance at $\rho=0$, as explained above). However, we observe that the kernel method generally outperforms the OLP method. Although in several cases the results are similar (the OLP method is better only in some rare cases), as expected and discussed above. In particular, note that the kernel method generally provides more accurate estimates than the OLP for small or large values of ρ: indeed, the value of ρ will be critical for the optimal choice of k, as discussed in the next section. Nevertheless, if we consider that the kernel method has been calibrated just to provide the best possible estimates under assumption of normality, the results of the OLP method are surprisingly valiant.

	Kernel		*OLP*	
ρ	*D*	*K*	*D*	*K*
0	0.784	0.059	0.565	0.207
0.09	0.894	0.070	0.609	0.203
0.18	0.255	0.044	0.284	0.088
0.27	0.227	0.073	0.122	0.060
0.36	0.225	0.105	0.177	0.085
0.45	0.059	0.031	0.118	0.085
0.54	0.076	0.057	0.111	0.069
0.63	0.111	0.103	0.169	0.122
0.72	0.103	0.106	0.072	0.078
0.81	0.098	0.129	0.094	0.091
0.9	0.067	0.132	0.077	0.082

Table 1: simulations with $n=500$ from a multivariate normal distribution

	Kernel		*OLP*	
ρ	*D*	*K*	*D*	*K*
0	0.601	0.009	0.502	0.045
0.09	0.067	0.008	0.062	0.015
0.18	0.038	0.011	0.030	0.009
0.27	0.034	0.013	0.033	0.011
0.36	0.017	0.008	0.029	0.010
0.45	0.017	0.009	0.017	0.010
0.54	0.012	0.010	0.017	0.013
0.63	0.013	0.010	0.015	0.013
0.72	0.009	0.010	0.014	0.012
0.81	0.007	0.009	0.015	0.013
0.9	0.005	0.008	0.014	0.014

Table 2: simulations with $n=10^5$ from a multivariate normal distribution

Similarly, we performed a simulation analysis also for the Student's t distribution. We generated several random samples of different sizes from $(X,Y)\sim St(\mu,\Sigma,4)$ (with $\mu=(0,2)$ and $\sigma_Y=\sigma_Y=1$): in this case, we know that $E(Y|X)\sim St(0,|\rho|,4)$. The results confirms what discussed above, as shown in Table 3, for $n=10^5$. In particular, observe that the OLP estimator outperforms the kernel estimator for values of ρ between 0.18 and 0.45, but in the other cases the kernel estimator is more accurate.

Finally, Fig. 1 and Fig. 2 show that the OLP estimator well captures the shape of the distribution

but approximates it with a simple function which has an inferior number of addends. Differently in Fig.3 we observe that the kernel method yields more accurate results for small values of ρ.

	Kernel		*OLP*	
ρ	*D*	*K*	*D*	*K*
0	0.670	0.023	0.590	0.057
0.09	0.079	0.015	0.069	0.017
0.18	0.056	0.019	0.039	0.014
0.27	0.025	0.019	0.027	0.016
0.36	0.021	0.017	0.020	0.015
0.45	0.019	0.018	0.021	0.017
0.54	0.011	0.013	0.025	0.019
0.63	0.014	0.018	0.017	0.024
0.72	0.011	0.016	0.024	0.023
0.81	0.005	0.010	0.012	0.022
0.9	0.006	0.012	0.013	0.023

Table 3: simulations with $n = 10^5$ from a multivariate t distribution

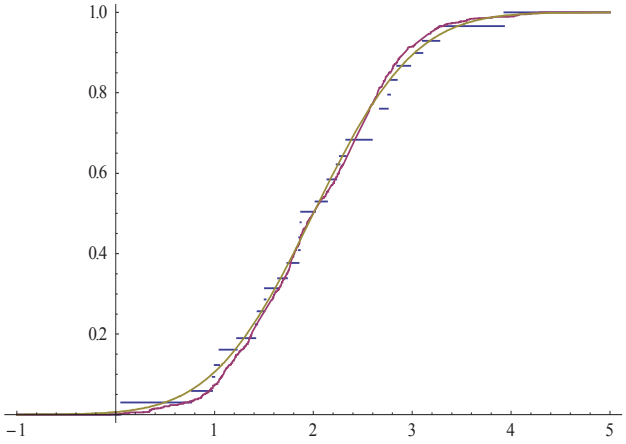

Fig. 1. $(X,Y)\sim N(\mu,\Sigma)$, $\mu = (0,2)$, $\sigma_Y = \sigma_Y = 1$ and $\rho = 0.8$. Green=true dist ,Red= estimated dist (Kernel), Blue=estimated dist (OLP)

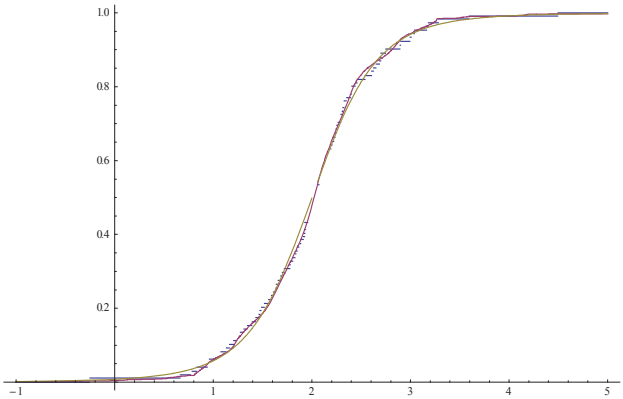

Fig. 2. $(X,Y)\sim St(\mu,\Sigma,4)$, $\mu = (0,2)$, $\sigma_Y = \sigma_Y = 1$ and $\rho = 0.5$. n=10000, Green=true dist ,Red= estimated dist (Kernel), Blue=estimated dist (OLP)

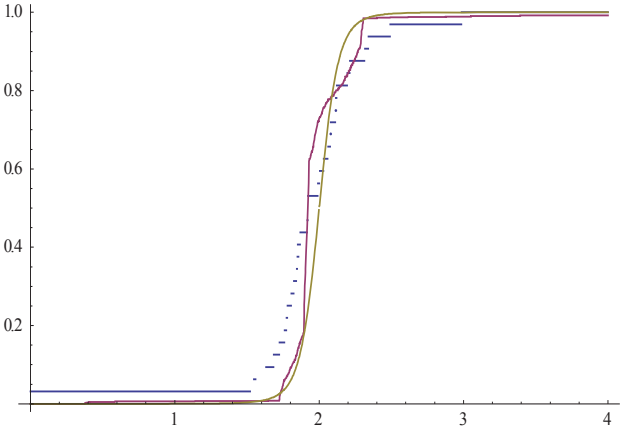

Fig. 3$(X,Y)\sim St(\mu,\Sigma,4)$, $\mu = (0,2)$, $\sigma_Y = \sigma_Y = 1$ and $\rho = 0.1$. n=1000, Green=true dist ,Red= estimated dist (Kernel), Blue=estimated dist (OLP)

4 On the Optimal Number of Intervals

The simulation comparison in section 3 was apparently "rigged" in favor of the kernel method. Indeed, we used the information about the distributional assumptions (normality) to improve the results of the kernel method as much as possible, i.e. using the normal kernel and the optimal bandwidth. On the other hand, this information was not used also to enhance the performance of the OLP estimator. However, in this adverse situation, we obtained that the OLP estimator yields surprising results. Differently, in this section we propose to use the information about the joint distribution in order to further improve the OLP estimator, under some particular conditions.

As discussed in section 3, the number of intervals k for the OLP method can vary between 1 and n.

However, if we assume that the random vector (X,Y) is jointly normally distributed, then we know that, for $\rho = 0$, we obtain $g(X) =_d E(Y)$, and therefore the distribution of $g(X)$ can be better estimated by a number, that is, the sample mean $\bar{y}$. Hence, in this particular case, the optimal value of k is exactly 1, rather than $\lceil\sqrt{n}\rceil$. On the other hand, for $|\rho| = 1$ (i.e. $Y = a + bX$), we obtain $g(X) = E(a + bX|X) = a + bX = Y$, and therefore the distribution of $g(X)$ can be estimated with the marginal observations of Y, $(y_1, \dots, y_n)$. Thus, when $|\rho| = 1$ the optimal value of k is exactly n: in this case we would get the maximum possible number ($n \geq \lceil\sqrt{n}\rceil$) of distinct observations,

From these considerations, we argue that, also in the case that $\rho \in (0,1)$, the dependence between X and Y should influence the choice of the number of intervals k. In particular, k should be chosen according to the mean-dependence structure between the random variables. We recall that generally stochastic independence implies mean-independence, which in turn implies uncorrelation (nevertheless, if (X,Y) is jointly Gaussian the three conditions are equivalent). In the following proposition we derive the formula of the mean squared error (MSE) between the OLP estimator and the conditional expectation $g(X)$ in the Gaussian case. It should be stressed, that generally the MSE is intended as the expectation of the squared error between an estimator, that is, a random variable, and a number. Interestingly, in this special case the MSE is based on the difference between two random variables. We show that the MSE is a mathematical function of the correlation coefficient, and therefore we can provide a simple rule of thumb for determining the optimal number of intervals k, also in the case $|\rho| \in (0,1)$. Without loss of generality, we focus on the special case that $(X,Y) \sim N(0,\Sigma)$ with $\sigma_Y = \sigma_Y = 1$, to simplify the computation. Obviously, if $\rho = 0$ then $E(Y|X) = 0$ and $k = 1$ is the optimal choice, if $|\rho| = 1$ then $E(Y|X) = \pm X = Y$ and $k = n$ is the optimal choice, as discussed above. For the proof of the following proposition we assume to know the true percentiles of X and thereby the true intervals A_j.

Proposition 1. *Let (X,Y) be a bivariate Gaussian vector, $(X,Y) \sim N(0,\Sigma)$ with $\Sigma = \big((1,\rho),(\rho,1)\big)$ (i.e. $\sigma_Y = \sigma_Y = 1$), and let $(x_1,y_1), \dots, (x_n,y_n)$ be a random sample of independent observations from (X,Y). Assume to know the $k-1$ percentiles $F_X^{-1}\left(\frac{j}{k}\right)$ of X, for $j = 1, \dots, k-1$, and thereby the intervals A_j, $j = 1, \dots, k$. Then, the mean squared error of the OLP estimator is given by:*

$$E\left[\left(\hat{g}_{OLP}(X) - g(X)\right)^2\right] = \\ = \frac{k}{n} + \rho^2\left\{1 - k\sum_{j=1}^{k}\left[f_X\left(F_X^{-1}\left(\frac{j-1}{k}\right)\right) + \right.\right. \\ \left.\left. - f_X\left(F_X^{-1}\left(\frac{j}{k}\right)\right)\right]^2\left(1 + \frac{1}{n}\right)\right\}, \qquad (16)$$

where f_X is the (Gaussian) marginal density of X.

Proof

We know that: $E(Y|X) = \rho X \sim N(0,|\rho|)$. Moreover, $E\left[\left(\hat{g}_{OLP}(X) - g(X)\right)^2\right] = E[(\hat{g}_{OLP}(X) - \rho X)^2] = E(\hat{g}_{OLP}(X))^2 + \rho^2 - 2\rho E(X\hat{g}_{OLP}(X))$, where

$$E(\hat{g}_{OLP}(X))^2 = E\left(\sum_{j=1}^{k} 1_{X\in A_j}\hat{a}_j\right)^2 = \\ = E\left(\sum_{j=1}^{k} 1_{X\in A_j}(\hat{a}_j)^2\right) = \\ = \frac{1}{k}\sum_{j=1}^{k} E\left(\hat{a}_j{}^2\right)$$

because $\sum_{i\neq j}\hat{a}_j\hat{a}_j 1_{X\in A_i\cap A_j} = 0$.

Note that

$$E(\hat{a}_j{}^2) = \left(\frac{k}{n}\right)^2 E\left(\left[\sum_i Y_i\, 1_{X\in A_j}\right]^2\right) \\ = \left(\frac{k}{n}\right)^2\left(\begin{matrix} nE\left(Y_i{}^2 1_{X\in A_j}\right) + \\ +\sum_{i\neq h} E\left[\left(Y_i 1_{X\in A_j}\right)\left(Y_h 1_{X\in A_j}\right)\right]\end{matrix}\right) = \\ = \left(\frac{k}{n}\right)^2\left(nE\left(Y^2 1_{X\in A_j}\right) + n(n-1)\left[E\left(Y 1_{X\in A_j}\right)\right]^2\right) \\ = \frac{k^2}{n}\left(\begin{matrix} E\left(Y^2 1_{X\in A_j}\right) + \\ +(n-1)\left[\int_{X\in A_j} f_X(x)\int_{-\infty}^{\infty} y\frac{f_{XY}(x,y)}{f_X(x)}dydx\right]^2\end{matrix}\right) \\ = \frac{k^2}{n}\left(\begin{matrix} E\left(Y^2 1_{X\in A_j}\right) + (n-1)\cdot \\ \cdot\left[\frac{\rho}{\sqrt{2\pi}}\left(\begin{matrix}\exp\left\{-\left(F_X^{-1}\left(\frac{j-1}{k}\right)\right)^2/2\right\} + \\ -\exp\left\{-\left(F_X^{-1}\left(\frac{j}{k}\right)\right)^2/2\right\}\end{matrix}\right)\right]^2\end{matrix}\right)$$

where

$$\int_{-\infty}^{\infty} y\frac{f_{XY}(x,y)}{f_X(x)}dy = E(Y|X) = \rho X, \\ \int xe^{-\frac{x^2}{2}} = -e^{-\frac{x^2}{2}},$$

and the equality $E\left[\left(Y_i 1_{X\in A_j}\right)\left(Y_h 1_{X\in A_j}\right)\right] = \left[E\left(Y_i 1_{X\in A_j}\right)\right]^2$ holds for the independence between the observations. Observe also that

$\sum_{j=1}^{k} E\left(Y^2 1_{X\in A_j}\right) = \sum_{j=1}^{k} \int_{X\in A_j} y^2 f_Y(y) = V(Y) = 1$. Then:

$$\frac{1}{k}\sum_{j=1}^{k} E\left(\hat{a}_j{}^2\right) =$$

$$= \frac{k}{n}\left(1 + (n-1)\frac{\rho^2}{2\pi}\sum_{j=1}^{k}\left[\exp\left\{-\left(F_X^{-1}\left(\frac{j-1}{k}\right)\right)^2/2\right\} - \exp\left\{-\left(F_X^{-1}\left(\frac{j}{k}\right)\right)^2/2\right\}\right]^2\right).$$

Furthermore, since

$$E(\hat{a}_j) = \frac{k\rho}{\sqrt{2\pi}}\left(\exp\left\{-\left(F_X^{-1}\left(\frac{j}{k}\right)\right)^2/2\right\} - \exp\left\{-\left(F_X^{-1}\left(\frac{j+1}{k}\right)\right)^2/2\right\}\right),$$

we obtain

$$E\left(X\hat{g}_{OLP}(X)\right) = \sum_{j=1}^{k} E\left(\hat{a}_j\right) E\left(X 1_{X\in A_j}\right) = \frac{k\rho}{2\pi}\left(\exp\left\{-\left(F_X^{-1}\left(\frac{j}{k}\right)\right)^2/2\right\} - \exp\left\{-\left(F_X^{-1}\left(\frac{j+1}{k}\right)\right)^2/2\right\}\right)^2,$$

which yields the thesis.

Obviously, in practical cases we do not know the true percentiles, but we estimate them with the empirical distribution: these estimates are consistent, that is, for fixed k and for $n \to \infty$ the sample percentiles converge to the true ones (as an obvious consequence of the law of large numbers). Nevertheless, we also need that $k \to \infty$ in order to obtain a consistent estimator of $g(X)$, therefore if the number of estimands grows as fast as n does, then the MSE of OLP method will not converge to 0. In view of Proposition 1, it is apparent that, in the case $|\rho| \neq 1$, the necessary and sufficient conditions for the convergence of the OLP estimator are: i) $k(n) \to \infty$; iii) $k/n \to 0$. This is stated in the following corollary, which is a straightforward consequence of Proposition 1.

Corollary 2. *Let (X,Y) be a bivariate Gaussian vector, i.e. $(X,Y)\sim N(\mu,\Sigma)$, where $|\rho| \in (0,1)$, and let $(x_1,y_1),\dots,(x_n,y_n)$ be a random sample of independent observations from (X,Y). A necessary and sufficient condition for $MSE\left(\hat{g}_{OLP}(X)\right) \to 0$ is that $k \to \infty$ and $k/n \to 0$.*

In other words, the general rule is that the number of percentiles (which have to be estimated) must grow slower than the number of observations. Corollary 1 gives necessary and sufficient conditions for the consistency of the OLP estimator in the Gaussian case. We argue that the same rule holds also if the joint distribution is not normal. Obviously, the rule $k = \lceil n^a \rceil$ where $a \in (0,1)$ (e.g. $k = \lceil n^{0.5} \rceil$, used in the simulation analysis) satisfies the conditions of Corollary 1. In order to further increase the convergence rate of the OLP estimator, we argue that, when $|\rho|$ is close to 0 the exponent a should be close to 0, and when $|\rho|$ is close to 1 the exponent a should be close to 1. Indeed, observe that the second term in the MSE expression approaches zero only as k tends to infinity, but it can be negligible for small values of $|\rho|$, while in this case the asymptotic behavior of the first term is critical. On the other hand, we obtain the exactly opposite situation when the variables are highly correlated, thus, in this case, a larger value of k would increase the convergence rate. Hence, as a rule of thumb, we propose to use

$$k = k^* = \left\lceil n^{|r|^{2/3}} \right\rceil \quad (17)$$

(where r is the value of the empirical correlation between the data) which yields $k = 1$ for $|r| \cong 0$, $k \cong n$ for $|r| \cong 1$, and ensures that $MSE\left(\hat{g}_{OLP}(X)\right) \to 0$ for any different value of ρ. The possible usefulness of this simple rule is well described by the following examples.

Example

Let $(X,Y)\sim N(\mu,\Sigma)$ with $\mu = (0,2)$, $\sigma_Y = \sigma_Y = 1$ and $\rho = 0.7$, which yields that $g(X)\sim N(2,0.7)$. We generate a bivariate sample of size 1000 from the random vector (X,Y) and estimate the distribution function F of $g(X)$ with the OLP method, using the following values of k: 1) $k = k^* = \left\lceil n^{|r|^{2/3}} \right\rceil = 223$ (where $r = 0.69$); 2) $k = \lceil n^{0.3} \rceil = 7$; 3) $k = \lceil \sqrt{n} \rceil = 31$; and 4) $k = \lceil n^{0.95} \rceil = 707$. Fig. 4 below shows that the best performance of the OLP estimator is obtained for 1) $k = \lceil n^{|r|} \rceil$, as in this case the estimated distribution is incredibly well-fitting to the true one. The other methods surely provide inferior performances. Indeed, the estimated distributions yielded by 2) and 3) seem to capture the shape of the reference distribution F (that is, $N(2,0.7)$) but approximate it with a "lower resolution", in that the number of intervals (i.e. distinct values $\hat{a}_j$) is quite poor (especially in case 2)). Conversely, in 4) we have a higher value of distinct values of $\hat{a}_j$ and consequently a "higher resolution" in the plot, but the estimated distribution has apparently a different shape from the true one. Nevertheless these results also confirm that with

$k = \lceil \sqrt{n} \rceil$ we generally obtain a good compromise between closeness to the true distribution and number of distinct observations $\hat{a}_j$.

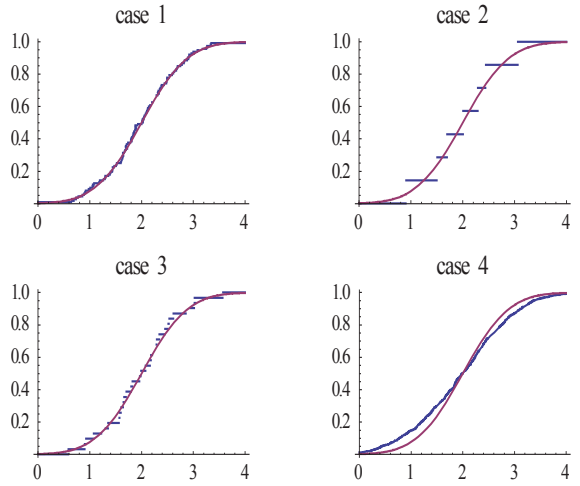

Fig. 4. Estimated (blue) and true (red) distribution functions for different values of k in the Gaussian case.

Furthermore, we repeat the same experiment for a bivariate t distribution, that is $(X,Y) \sim St(\mu, \Sigma, 5)$ with $\mu = (0,2)$, $\sigma_Y = \sigma_Y = 1$ and $\rho = 0.3$. In this case we obtain $g(X) \sim St(2,0.3,5)$. We generate a bivariate sample of size 10000 from the random vector (X,Y) and estimate the distribution function F of $g(X)$ with the OLP method, using the following values of k: 1) $k = k^* = 26$ (where $r = 0.32$); 2) $k = \lceil n^{0.1} \rceil = 1$; 3) $k = \lceil \sqrt{n} \rceil = 100$; and 4) $k = \lceil n^{0.7} \rceil = 125$. Fig. 5. shows that in case 1) and 3) we apparently obtain the best approximations. However, the Kantorovich metric (0.168 for case 1 and 0.18 for case 3) confirms that the best choice is k^*.

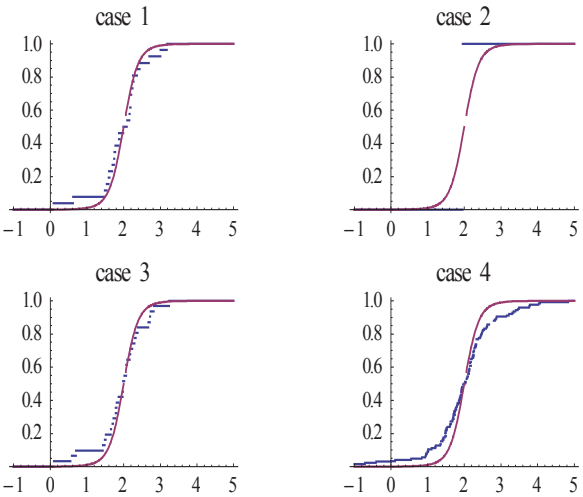

Fig. 5. Estimated (blue) and true (red) distribution functions for different values of k in the Student's t case.

We finally argue that the proposed rule can be appropriate for dealing with distributions which are approximately Gaussian, and therefore can be usefully applied to several kinds of data, because of the central limit theorem for random vectors. Indeed, the multidimensional central limit theorem states that the standardized sum of i.i.d random vectors (with finite variance) converges to a multivariate normal distribution. Therefore, just based on the assumption that (X,Y) have a joint distribution (i.e. they are defined on the same probability space), we can generally use the empirical correlation between the observations of X and Y in order to get information about the dependence between the variables, and thereby to determine the optimal number of intervals for the OLP estimator.

5 Conclusion

In this paper, we proposed two different procedures (OLP and kernel) to estimate the distribution function of a conditional expectation, based on a bivariate random sample. In particular, the properties of the OLP estimator have been studied thoroughly. It has been shown that the method can provide a consistent approximation of the random variable $E(Y|X)$, based on a suitable choice of a parameter k. Both the OLP and kernel methods make it possible to estimate the distribution function of $E(Y|X)$ non parametrically. Our simulation results show that the methods are comparable. However, it should be stressed that the OLP method presents several advantages compared to the kernel, in that it does not require any particular assumption in order to be applied. Conversely, the kernel method requires on many restrictive conditions, such as continuity and finite variance, and its performance depends on a suitable choice of the kernel function and bandwidth parameter. Finally, since the performance of the OLP estimator depends just on the chosen number of intervals k, we provided some general criteria for the choice of k under normal assumptions. Consequently, we proposed a practical rule in order to optimize the performance of the method.

Both estimators (kernel and OLP) can be used in optimization procedures as required in several financial applications. In particular, we can use the conditional expectation estimators to i) order the investors' choices or ii) evaluate and exercise arbitrage strategies in the market (see [10]). In the first case, we know that any non satiable risk averse investor prefers the future wealth W_T at time T with respect the wealth W_t at time t (t<T), only if

$E(W_t|W_T) \leq W_T$ a. s.. Thus, by using the proposed approximation of the conditional expected value, we can attempt to order and optimize the choices of non satiable risk averse investors, as suggested by [8]. In the second case, as a consequence of the fundamental theorem of arbitrage, we know that there exists no arbitrage opportunity in the market if there exists a risk neutral martingale measure under which the discounted price process results to be a martingale. So, if we consider the augmented filtration $\{\Im_t\}_{t\geq 0}$ associated to the Markov price process $\{X_t\}_{t\geq 0}$, then we obtain $\forall s \leq t$ that $E(X_t|\Im_s) = E(X_t|X_s)$. Therefore, the conditional expected value estimator and the fundamental theorem of arbitrage can be used to estimate the risk neutral measure and to optimize arbitrage strategies in the market.

Acknowledgements

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 supported through the Czech Science Foundation (GACR) under project 15-23699S and through SP2015/15, an SGS research project of VSB-TU Ostrava, and furthermore by the European Social Fund in the framework of CZ.1.07/2.3.00/20.0296.

References:

[1] E. A. Nadaraya, On estimating regression, *Theory of Probability and its Applications*, vol. 9, no. 1, 1964, pp. 141-142.

[2] G. S. Watson, Smooth regression analysis, *Sankhya, Series A*, vol. 26, no. 4, 1964, pp. 359-372.

[3] D. Ruppert, M. P. Wand, R. J. Carroll, *Semiparametric Regression*. Cambridge University Press, 2003.

[4] G, Mzyk, Generalized kernel regression for the identification of Hammerstein series, *International journal of applied mathematics and computer science*, Vol. 17., No. 2, 2007, pp. 189.197.

[5] S. Ortobelli, F. Petronio, T. Lando, Portfolio problems based on returns consistent with the investor's Preferences, *Advances in Applied and Pure Mathematics*, ISBN: 978-960-474-380-3, 2014, pp. 340-347.

[6] S. G. Steckey, S.G. Henderson, A kernel approach to estimating the desity of a conditional expectation, *Proceedings of the 2003 Winter Simulation Conference*, S. Chick, P. J. Sánchez, D. Ferrin, and D. J. Morrice, eds, 1003, pp.383-391.

[7] A.N. Avramidis, A cross-validation approach to bandwidth selection for a kernel-based estimate of the density of a conditional expectation, *Proceedings of the 2011 Winter Simulation Conference*, S. Jain, R. R. Creasey, J. Himmelspach, K. P. White, and M. Fu, eds., 2011, pp.439-443.

[8] S. Ortobelli, F. Petronio, T. Lando, A portfolio return definition coherent with the investors preferences, under revision in *IMA-Journal of Management Mathematics*.

[9] M. Roth, *On the multivariate t distribution*, Technical report from Automatic Control at Linköpings universitet, 2013.

[10] S. Ortobelli, T. Lando, On the use of conditional expectation estimators, *New Developments in Pure and Applied Mathematics*, ISBN: 978-1-61804-287-3, 2015, pp. 244-246.

[11] W. Hardle, M. Muller, S. Sperlich, A. Werwatz, *Nonparametric and semiparametric models*. Springer Verlag, Heidelberg, 2004.

[12] V. A. Epanechnikov, Non-parametric estimation of a multivariate probability density, *Theory of Probability and its Applications*, vol. 14, no. 1, 1965, pp. 153-158.

[13] T. Lando, L., Bertoli-Barsotti, Statistical Functionals Consistent with a Weak Relative Majorization Ordering: Applications to the Mimimum Divergence Estimation, *WSEAS Transactions on Mathematics*, Vol 13, 2014, pp. 666-675.

[14] S. T. Rachev, *Probability metrics and the stability of stochastic models*. Wiley, New York, 1991.

[15] C. Raykar, R. Duraiswami, Fast optimal bandwidth selection for kernel density estimation, *Proceedings of the Sixth SIAM International Conference on Data Mining*, 2006, pp. 522-526.

[16] D.V. Scott, On Optimal and Data-Based Histograms, *Biometrika*,Vol. 66, No. 3, 1979, pp. 605-610.

[17] H. A. Sturges, The Choice of a Class Interval, *Journal of the American Statistical Association*, Vol. 21, No. 153, 1926, pp. 65-66.

[18] E. Bura, A. Zhmurov, V. Barsegov, Nonparametric density estimation and optimal bandwidth selection for protein unfolding and unbinding data, *The Journal of chemical physics*, Vol. 130, 2009, article no. 015102.