Kernel density estimation using local cubic polynomials through option prices applied to intraday data

ANA MARGARIDA MONTEIRO

ANTÓNIO ALBERTO FERREIRA SANTOS

Faculty of Economics, Centre for Business and Economics Research (CeBER), Monetary and Financial Research Group (GEMF), University of Coimbra

CeBER Working Papers

No. 2 2019

Kernel density estimation using local cubic polynomials through option prices applied to intraday data

A. M. Monteiro* and Antonio A. F. Santos†
Faculty of Economics, Centre for Business and
Economics Research (CeBER),
Monetary and Financial Research Group (GEMF)
University of Coimbra, Portugal

*ammm@fe.uc.pt
†aasantos@fe.uc.pt

Abstract

A new approach is considered to estimate risk-neutral densities (RND) within a kernel regression framework, through local cubic polynomial estimation using intraday data. There is a new strategy for the definition of a criterion function used in nonparametric regression that includes calls, puts, and weights in the optimization problem associated with parameters estimation. No-arbitrage restrictions are incorporated in the problem through equality and bound constraints. This yields directly density functions of interest with minimum requirements needed. Within a simulation framework, it is demonstrated the robustness of proposed procedures. Additionally, RNDs are estimated through option prices associated with two indices, S&P500 and VIX.

1 Introduction

Risk-neutral densities (RND) are determinant when dealing with risk management and pricing of new derivative products. In recent years, the amount of information coming from intraday data allows improving and developing of existing and new approaches. Nowadays, option prices are intensively traded in several markets, these transactions can reveal market expectations on the underlying asset, that are reflected in the corresponding RND. In fact, observed option prices have been used to extract information about behaviour of the underlying asset since they give insights about risk factors associated with it. RNDs can be used for different purposes, namely to infer about risk in the market, and to price complex option contracts. A main feature that must be considered is that RND estimation is an indirect method as no risk-neutral prices can be observed.

We propose a novel approach for risk-neutral density estimation within the framework of a nonparametric regression setting (Aït-Sahalia and Lo, 1998; Aït-Sahalia and Duarte, 2003; Yatchew, 2003; Monteiro et al., 2008). This is able to address some main problems presented in previous attempts found in the literature, such as non-monotonicity and non-convexity of call and put pricing estimated functions. There are also problems related to violation of basic requirements associated with a density or distribution function, its limits, and non-negativity associated with the density. Additionally, we would expect some degree of smoothness for density estimation since a lack of regularity is not intuitive.

The novel approach is based on a local cubic polynomial kernel regression applied to intraday data. First, it yields directly the density of interest without the need of further transformations. Usually, the literature presents estimation processes that retrieve the call pricing function or its first derivative, and it is necessary to differentiate in order to obtain the density (Aït-Sahalia and Lo, 1998; Song and Xiu, 2016). Second, it uses compatible information contained in observed call and put intraday prices without using put-call parity. When using intraday data, the difficulty in synchronizing call and put option prices with the underlying asset price can lead to errors (Aït-Sahalia and Lo, 1998; Fengler and Hin, 2015), our approach prevent these to occur. This allows to incorporate both prices in the optimization problem, increasing the amount of information that is retrieved from the market, and avoid errors from lack of synchronicity. Third, it includes in a smoothly and intuitive way no-arbitrage constraints in the optimization problem. Smoothness is mainly a result of including no-arbitrage constraints, and two distinct bandwidths for calls and puts. This allows to adapt the optimization problem to different data sets. Also, it takes into account that first derivatives of

call and put pricing functions differ from a constant and second derivatives are equal. Fourth, in order to account for the relevance of each option price, we introduce open interest data as weights in the criterion function associated with the problem. By this way, we add more information on market perspectives and beliefs.

The approach is tested against simulated data, which can confirm that the method is able to recover density functions accurately. When applied to real data sets, constituted by intraday data with a less uniform behaviour, also gives results that are robust and easily interpreted.

The remainder of the paper is organized as follows, Section 2 discusses the technical details of the risk-neutral density theory. Section 3 introduces the nonparametric estimation based on kernel approaches. Monte Carlo simulated experiments and their discussion are presented in Section 4. Section 5 presents the estimations from market data. In the final section concluding remarks are exposed.

2 Risk-neutral density through option prices

Cox and Ross (1976) presented the risk-neutral density in the context of no-arbitrage based models assuming that investors are risk-neutral. Breeden and Litzenberger (1978) and Banz and Miller (1978) proposed a way of estimating these densities from prices of financial options by considering second derivatives of option pricing functions.

There are several approaches used to estimate RNDs from option prices. They are extensive, and can be divided in two main groups: structural and non-structural models. The former specify a process for the underlying price, and sometimes for volatility. Non-structural models describe the density

behaviour without prescribing a stochastic process for the underlying asset. In this last scenario, nonparametric methods and estimation procedures have been developed. They are more flexible, usually allow for a wide set of shapes for densities but require, in general, larger sample sizes. By relaxing assumptions on the underlying process, these models try to achieve a function that describes the data. Utilization of nonparametric estimation methods, based on kernel approaches, to obtain a RND implicit in option prices dates back to the seminal paper of Aït-Sahalia and Lo (1998), revisited by Aït-Sahalia and Lo (2000) and Aït-Sahalia et al. (2001). The main idea was to overcome some drawbacks associated with a parametric setting.

Usually, this kind of nonparametric estimators present rates of convergence substantially lower than their parametric counterparts, and to obtain similar degrees of accuracy far larger sample sizes are needed. This fact is even more relevant when we are considering estimators for first or second derivatives, which is fully addressed by general references for nonparametric methods as Fan and Gijbels (1996), Yatchew (2003), Härdle (1990), Li and Racine (2007), and for the specific application to RND estimation by Aït-Sahalia and Duarte (2003).

There are different goals when considering kernel estimation procedures dealing with no-arbitrage constraints. Jackwerth (2000) computed a subjective distribution based on a kernel estimator. Rosenberg and Engle (2002) estimated risk aversion by considering a kernel function depending on the maturity. Aït-Sahalia and Duarte (2003) proposed shape restrictions for the optimization problem considering local polynomial estimation. Yatchew and Härdle (2006) and Härdle and Hlávka (2009), using smoothing splines, also imposed constraints on the estimation problem, in order to guarantee convexity and monotonicity for call pricing functions. Monteiro et al. (2008)

proposed a nonparametric approach with no-arbitrage constraints based on cubic splines through a semidefinite programming problem. Zhang et al. (2009) considered a local polynomial estimator together with a method based on Gram-Charlier series expansion to obtain RNDs. Song and Xiu (2016) considered a nonparametric kernel approach for estimate RND including volatility factors. They used local linear estimators for first derivatives of option pricing functions using end-of-day data from S&P500 and VIX.

Data length is an important issue in any estimation process. Aït-Sahalia and Lo (1998), and most of the subsequent studies, considered extensive time series. Most papers from the literature consider data from a large time period and assume the estimated risk-neutral density as an average of densities. Recently, intraday data has become more accessible, and the amount of data collected in a few days gives enough information to infer RNDs. Dalderop (2018) estimates time-varying RNDs by considering a kernel estimator as a function of time and moneyness applied to intraday data. The author uses different order approximations: local constant for time dimension and local cubic for moneyness.

Several frameworks have been used to derive call option prices, which by no-arbitrage arguments must be associated with a portfolio without risk and risk-free interest rate. It has been found that corresponds to calculate the expected future option value at expiration, computed through a risk-neutral density measure, $\mathbb{Q}$, discounted by the risk-free interest rate, which can be expressed as

$$\begin{aligned} C(S_t, X, \sigma_t, \tau, r, \delta) &= e^{-r\tau} E^{\mathbb{Q}} \left[(S_T - X)^+ \right] \\ &= S_t \mathbb{Q}^S (S_T > X) - X e^{-r\tau} \mathbb{Q}(S_T > X) \\ &= S_t \, P_1 - X e^{-r\tau} \, P_2. \end{aligned}$$

where t represents the current date, X the strike price, r the risk-free interest rate, S_t the current underlying asset price, S_T the asset price at maturity, $\tau = T - t$ the time-to-maturity, and δ the dividend yield. A main difference from Black-Scholes' formula is related to probabilities P_1 and P_2 measured through risk-neutral densities. The formula reveals that option price depends mainly on the underlying stock and strike prices. Different variations can be considered by changing the structure of P_1 and P_2, which allows the use of mean pricing functions in different contexts through minor changes.

Consider an economy with two state variables, the price of S&P 500 index S, and an unobserved volatility V. Since the volatility is determinant for pricing S, consider also the VIX option market and denote the variable by Z. As there are no contingent claims written on V, we will consider option contracts on the volatility index VIX in order to estimate the risk-neutral densities of S and Z.

The call option price for contracts on S can be given by

$$C(S_t, X, \sigma_t, \tau, r, \delta) = e^{-r\,\tau} \int_X^\infty (S_T - X)\, g(S_T|S_t, \sigma_t, r, \delta, \tau)\, dS_T,$$

where $g(\cdot)$ represents the conditional risk-neutral density for the underlying asset at expiration T. Considering the same assumptions, the price of a European put option is

$$P(S_t, X, \sigma_t, \tau, r, \delta) = e^{-r\,\tau} \int_0^X (X - S_T)\, g(S_T|S_t, \sigma_t, r, \delta, \tau)\, dS_T.$$

The price of a VIX call option with strike L can be given in a similar way by

$$J(Z_t, L, \sigma_t, \tau, r, \delta) = e^{-r\,\tau} \int_X^\infty (Z_T - L)\, h(Z_T|Z_t, \sigma_t, r, \delta, \tau)\, dZ_T,$$

where Z_T is the VIX price index on the maturity, and $h(\cdot)$ is the conditional risk-neutral density for Z_T.

Breeden and Litzenberger (1978) and Banz and Miller (1978) proposed a relation between second derivative of the call option price, with respect to the strike price, and the risk-neutral density:

$$\begin{aligned}\frac{\partial C(S_t, X, \sigma_t, \tau, r, \delta)}{\partial X} &= \frac{\partial \left(e^{-r\,\tau}\int_X^\infty (X - S_T)\, g(S_T|S_t, \sigma_t, r, \delta, \tau)\, dS_T\right)}{\partial X} \quad (1)\\ &= e^{-r\,\tau}(G(X|S_t, \sigma_t, r, \delta, \tau) - 1), \quad (2)\end{aligned}$$

where $G(\cdot)$ is the respective distribution function associated with $g(\cdot)$. The second derivative is then expressed as

$$\frac{\partial^2 C(S_t, X, \sigma_t, \tau, r, \delta)}{\partial X^2} = e^{-r\,\tau}\, g(X|S_t, \sigma_t, r, \delta, \tau) \quad (3)$$

and the risk-neutral density at expiration is

$$g(X|S_t, \sigma_t, r, \delta, \tau) = e^{r\tau}\, \frac{\partial^2 C(S_t, X, \sigma_t, \tau, r, \delta)}{\partial X^2}|_{X=S_T}.$$

The risk-neutral density can be established in an equivalent form using puts as was considered for calls,

$$\frac{\partial P(S_t, X, \sigma_t, \tau, r, \delta)}{\partial X} = e^{-r\,\tau}\, G(X|S_t, \sigma_t, r, \delta, \tau) \quad (4)$$

and

$$\frac{\partial^2 P(S_t, X, \sigma_t, \tau, r, \delta)}{\partial X^2} = e^{-r\,\tau} g(X|S_t, \sigma_t, r, \delta, \tau). \quad (5)$$

$$g(X|S_t, \sigma_t, r, \delta, \tau) = e^{r\tau}\, \frac{\partial^2 P(S_t, X, \sigma_t, \tau, r, \delta)}{\partial X^2}|_{X=S_T}.$$

Considering the VIX options

$$\begin{aligned}\frac{\partial J(Z_t, L, \sigma_t, \tau, r, \delta)}{\partial L} &= \frac{\partial \left(e^{-r\,\tau}\int_X^\infty (L - Z_T)\, h(Z_T|Z_t, \sigma_t, r, \delta, \tau)\, dZ_T\right)}{\partial L}\\ &= e^{-r\,\tau}(H(L|Z_t, \sigma_t, r, \delta, \tau) - 1),\end{aligned}$$

where L represents the strike price, and $H(\cdot)$ is the respective distribution function associated with $h(\cdot)$

$$h(Z_T|Z_t, \sigma_t, r, \delta, \tau) = e^{r\tau}\, \frac{\partial^2 J(Z_t, L, \sigma_t, \tau, r, \delta)}{\partial L^2}|_{L=Z_T}.$$

By combining first and second derivatives of call and put pricing functions, we obtain some constraints to be imposed to our optimization problem. Since pricing functions in this context are homogeneous of degree one in the strike (Fengler and Hin, 2015; Song and Xiu, 2016), we can scale strikes and prices without changing the relation between variables.

3 Nonparametric estimation

A cornerstone for nonparametric methods applied to RND estimation is given by the seminal paper of Aït-Sahalia and Lo (1998). It is stressed out its importance, and how with sufficient amount of data, it is possible to get rid of constraints imposed by some difficult to justify parametric approaches.

Nonparametric approaches offer more flexible methods for modelling mean function option prices on strikes and other relevant variables. The second derivative has been shown to be related with RNDs. Using Aït-Sahalia and Lo (1998) notation, suppose there is a smooth function $H(\cdot)$ that can be seen as an option pricing function depending on a vector Z, set as $Z = (S_t, X, \tau, r, \delta)$. A possible nonlinear relationship is established as

$$H_i = H(Z_i) + \varepsilon_i, \quad i = 1, \ldots, n$$

assuming ε_i as a white noise.

Nadaraya-Watson (NW) estimator is commonly used (Nadaraya, 1964; Watson, 1964), and assumes the form

$$\hat{H}(Z_i) = \frac{\sum_{i=1}^{n} K_h\left(Z_i - Z\right) H_i}{\sum_{i=1}^{n} K_h\left(Z_i - Z\right)},$$

which can be seen as a weighted mean average of the $H_i's$. A kernel function $K_h(\cdot)$, depending on a bandwidth h, for a given point Z, defines the weights.

The application of this estimator can be challenging. As nonparameteric methods are data intensive, their effectiveness rapidly decreases as problem's dimension increases (number of explanatory variables). Without imposing adequate constraints, estimates can contradict basic economic principles and even common sense. To overcome this problem Aït-Sahalia and Lo (1998) proposed some justifications for dimension reduction, and estimated RNDs by an indirect way, through an estimator for implied volatility, which is plugged-in Black-Scholes' formula. By this approach, it is natural that RNDs inherit most characteristics obtained using a parametric model as the one referred.

To address some problems associated with estimation procedures proposed in Aït-Sahalia and Lo (1998), Aït-Sahalia and Duarte (2003) revisited the problem, and a new approach was proposed that has been followed in subsequent literature. Instead of local constant kernel regression, a more general setting was proposed based on local polynomial regressions, but more importantly, it was highlighted the importance of shape restrictions on mean pricing functions, and respective derivatives for obtaining meaningful results. A univariate setting is adopted making option prices depending only on strikes, and obtaining similar accuracy using fewer observations. Estimators were subjected to a series of shape constraints, and a kind of double smoothing. They devised a two-step procedure which incorporates an intricate optimization problem, followed by a kernel smoothing estimation for obtaining the desired densities.

NW estimator can seen as a local constant kernel regression type estimator. Some drawbacks associated with it can be softened by a more general approximation. Let us consider for simplicity a general formulation $y_i = m(x_i)+\varepsilon_i$. For a local polynomial of order p, kernel regression estimators

are obtained by solving the problem

$$\text{minimize} \quad \sum_{i=1}^{n} \left(y_i - \sum_{k=0}^{p} \beta_k(x) \frac{(x_i - x)^k}{k!} \right)^2 K\left(\frac{x_i - x}{h} \right),$$

where the decision vector is $\beta(x)$. This general formulation encompasses constant ($p = 0$), linear ($p = 1$), quadratic ($p = 2$), and cubic ($p = 3$) orders. Compared with NW estimator, nonparametric local linear polynomial approximation ($p = 1$) represents an important improvement in terms of flexibility and estimator's properties (Fan and Gijbels, 1996). Applying a local p-order polynomial criterion, the mean function estimator and respective derivatives are given directly by $\hat{m}(x) = \hat{\beta}_0(x)$, and $\hat{m}^{(k)}(x) = k!\,\hat{\beta}_k(x)$.

Several authors accommodate no-arbitrage constraints in the definition of nonparametric estimates for RNDs, as already mentioned. Aït-Sahalia and Duarte (2003), but also Yatchew and Härdle (2006), used constrained nonparametric least squares, where constraints are defined through a penalty component, expressed by the Sobolev norm, which needs a function called *representor*, making this approach less intuitive, and not easy to implement. Birke and Pilz (2008) address the problem using an auxiliary inverse function associated with call pricing function first derivative, that needs to be integrated or differentiated for obtaining call pricing functions or risk-neutral densities, respectively.

We devise an alternative procedure in comparison with aforementioned. In contrast, a simple and intuitive framework to include no-arbitrage constraints directly in a criterion function is developed. A fact not fully explored in literature is related to derivation of risk-neutral densities using information contained directly in both calls and puts. As functions of strikes, call and put prices move in opposite directions, and also the variability. In left tail call prices vary more than puts, and vice versa in the right tail. Expressing

parameters, in a criterion function, as values of risk-neutral distributions and densities (the latter coincides for calls and puts) allows easily to impose no-arbitrage constraints. The contrast (variability) between call and put prices can be used as a valuable source of information to define robust estimation procedures.

3.1 Nonparametric with no-arbitrage constraints

When RNDs are estimated implicitly through option prices, it is desired to obtain a smooth function. Most area must be associated with a neighbourhood around current value of the underlying asset, and on tails direction, density values must tend to zero. A fundamental problem is how tails behave, their rate of convergence to zero, and comparisons between left and right tails.

No-arbitrage constraints are intimately related to monotonicity and convexity that are established characteristics of call and put pricing functions. Following Birke and Pilz (2008), no-arbitrage constraints assume the form

$$\begin{aligned} -e^{-r\,\tau} \leq \frac{\partial C}{\partial X}(X) &\leq 0 \\ \frac{\partial^2 C}{\partial X^2}(X) &\geq 0 \\ C(X) &\geq 0, \quad \forall\, X \in [0, \infty[. \end{aligned}$$

Using put-call parity, the same kind of constraints can be associated with put pricing functions,

$$\begin{aligned} 0 \leq \frac{\partial P}{\partial X}(X) &\leq e^{-r\,\tau} \\ \frac{\partial^2 P}{\partial X^2}(X) &\geq 0 \\ P(X) &\geq 0 \;\; \forall\, X \in [0, \infty[. \end{aligned}$$

Using these constraints, we define an extended criterion function within a constraint nonparametric regression framework.

Let us designate c_i and p_j observed prices for calls and puts, with respective strikes x_i and x_j for $i = 1, \ldots, n$ and $j = 1, \ldots, m$. The proposed criterion function is an extension of a local cubic polynomial approximation within a nonparametric regression setting. The extension accounts jointly call and put prices. This has the advantage of representing a contrast of information, and also allows no-arbitrage constraints straightforwardly.

In a kernel regression framework, where local approximations are defined around x, kernel functions serve to weight the distance of sample observations to x. Two kernel functions are considered, $K((x_i - x)/h_c)$ for calls, and $K((x_j - x)/h_p)$ for puts, accounting for different bandwidths, h_c and h_p. As it is well documented in the literature (Härdle, 1990; Fan and Gijbels, 1996; Yatchew, 2003; Li and Racine, 2007), these parameters have a major influence for the adjustment in a kernel regression framework, namely, in comparison with the choice of kernel function. This fact leads to consider only Gaussian kernels, although different kernel functions were tested without significant changes.

By relaxing the assumption of a white noise for error terms in the mean function, different weights are associated with each observation. More informative observations are ones represented by at-the-money prices, deep-in-the-money or deep-out-the-money are less informative. These can be weighted by volume or open interest values, represented for calls by $w_{i,c}$, and for puts by $w_{j,p}$.

The estimation is performed by minimizing a criterion function subject

to a set of linear and bound constraints,

$$\text{minimize} \sum_{i=1}^{n} w_{i,c} \left(c_i - \beta_{0,c}(x) - \sum_{k=1}^{3} \beta_{k,c}(x) \frac{(x_i - x)^k}{k!} \right)^2 K\left(\frac{x_i - x}{h_c} \right) + \sum_{j=1}^{n} w_{j,p} \left(p_j - \beta_{0,p}(x) - \sum_{k=1}^{3} \beta_{k,p}(x) \frac{(x_j - x)^k}{k!} \right)^2 K\left(\frac{x_j - x}{h_p} \right) \quad (6)$$

subject to

$$\begin{aligned}
-\beta_{1,c}(x) + \beta_{1,p}(x) &= e^{-r\tau} && (7)\\
\beta_{2,c}(x) - \beta_{2,p}(x) &= 0 && (8)\\
\max(0, S_t - x\, e^{-r\,\tau}) \le \beta_{0,c}(x) &\le S_t && (9)\\
\max(0, x\, e^{-r\,\tau} - S_t) \le \beta_{0,p}(x) &\le \infty && (10)\\
-e^{-r\tau} < \beta_{1,c}(x) &< 0 && (11)\\
0 \le \beta_{1,p}(x) &\le e^{-r\tau} && (12)\\
\beta_{2,c}(x) &\ge 0 && (13)\\
\beta_{2,p}(x) &\ge 0. && (14)
\end{aligned}$$

For each local approximation at x, the problem can be characterized as a Generalized Least Squares (GLS) problem with constraints, which can be solved as a Quadratic Programming (QP) problem. Let us designate y as the observations vector for call and put prices, and consider matrices $X_c(x)$ and $X_p(x)$, with typical rows i and j given by

$$\begin{aligned}
X_{i,c}(x) &= [\ 1 \quad (x_i - x) \quad (1/2)(x_i - x)^2 \quad (1/6)(x_i - x)^3]\\
X_{j,p}(x) &= [\ 1 \quad (x_j - x) \quad (1/2)(x_j - x)^2 \quad (1/6)(x_j - x)^3]
\end{aligned}$$

for the vectors and matrix defined as

$$y = \begin{bmatrix} c \\ p \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_c \\ \beta_p \end{bmatrix}, \quad X = \begin{bmatrix} X_c(x) & \mathbf{0} \\ \mathbf{0} & X_p(x) \end{bmatrix}.$$

Given a matrix of weights represented by W, and a "kernel matrix" $K = K(h_c, h_p)$, the minimization problem is expressed as a quadratic optimization problem

$$\begin{aligned} \text{minimize} \quad & (y - X\,\beta)^\top W^{1/2} K W^{1/2} (y - X\,\beta) \\ \text{subject to} \quad & \beta \in \mathscr{B} \end{aligned}$$

where $\mathscr{B}$ is the set of constraints. By considering

$$\begin{aligned} y^* &= W^{1/2} K^{1/2}\, y \\ X^* &= W^{1/2} K^{1/2}\, X \end{aligned}$$

the latter can be rewritten as a norm minimization problem subject to convex constraints,

$$\begin{aligned} \text{minimize} \quad & ||y^* - X^*\,\beta|| \\ \text{subject to} \quad & \beta \in \mathscr{B} \end{aligned}$$

which can be translated to a QP optimization problem

$$\begin{aligned} \text{minimize} \quad & \beta^\top\, H\,\beta + f^\top \beta \\ \text{subject to} \quad & \beta \in \mathscr{B} \end{aligned}$$

where $H = {X^*}^\top X^*$ and $f = -{X^*}^\top\, y^*$.

By applying a local cubic polynomial approximation the estimates at each point x are obtained for $\beta_{i,j}(x)$, with $i = 0, 1, 2, 3$, $j = c, p$. The main aim are the estimates $\hat{\beta}_{2,c}(x)$ and $\hat{\beta}_{2,p}(x)$, with $\hat{\beta}_{2,c}(x) = \hat{\beta}_{2,p}(x)$, that represent the risk-neutral-density's value at x. Several sources of information are considered, namely the contrast between call and put prices evolution. Equally important are constraints implied by no-arbitrage arguments that act as smoothing components, to obtain more reliable and intuitive RND estimates.

3.2 Bandwiths and weights selection

The distance to a point x, defining locality, must be taken into account, and in nonparametric regression methods this is done by a kernel function, symmetric around x, and that integrates one. Kernel functions depend on the distance of observations to x, scaled by a bandwidth parameter, which is recognized as the most relevant factor in terms of characteristics and quality of model fitting.

When approximating a mean function, using the Mean-Square Error (MSE) criterion, an optimal bandwidth is chosen through a min-max optimization problem, which is related to a trade-off between bias and variance. An MSE criterion allows the definition of a local optimal bandwidth, which depends on many factors, for example, sample size, curvature of the mean function, distribution of design variables, and their respective variance. Usually these quantities are unknown.

Local optimal bandwidths are difficult to define. A common approach tries to approximate a global optimal bandwidth, which is defined through minimization of the Mean Integrated Square Error (MISE). In some cases, it is possible to define an analytic expression to the global bandwidth, however, it depends on unknown quantities. In practical terms, to define the optimal global bandwidth, Cross Validation (CV) methods are used. An obtained value is asymptotically optimal through the MISE criterion.

In developed results, two bandwidths were considered, h_c related to call observations, and h_p to puts. We apply CV to obtain a first approximation as only observed values and estimated means can be compared. Considering again a general setting, $y_i = m(x_i) + \varepsilon_i$, $i = 1, \ldots, n$, and using the common

approach for CV, which is leave-one-out, h is chosen by

$$\text{minimize} \;\; CV(h) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{m}_{-i}(x_i))^2 W(x_i),$$

where $\hat{m}_{-i}(x_i)$ is the leave-one-out kernel estimator of $m(x_i)$, and $W(\cdot)$ is a weight function, see Li and Racine (2007).

Adapting CV to define h_c and h_p, we take into account a fixed design framework, where for each x_i, $i = 1, \ldots, n$, k_i observations for y are available, which means that we have to implement the procedure leave-k_i-out. The CV criterion is modified, and bandwidths are chosen by

$$\begin{aligned}\text{minimize } CV(h_c, h_p) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k_i}(y_{c,j} - \hat{m}^c_{-k_i}(x_i))^2 \, W_c(x_i) + \\ \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k_i}(y_{p,j} - \hat{m}^p_{-k_i}(x_i))^2 \, W_p(x_i),\end{aligned}$$

where kernel estimator for calls $\hat{m}^c_{-k_i}(x_i)$ depends on h_c, and for puts $\hat{m}^p_{-k_i}(x_i)$ depends on h_p. Weight functions used, $W_c(x_i)$ and $W_p(x_i)$, will also be different for calls and puts, reflecting choices done when the criterion function for parameters estimation was defined.

Weights used for parameters estimation and CV account for the distance between observations and a point representing at-the-money prices. As h_c and h_p define a neighbourhood around x, elements of $W_c(\cdot)$ and $W_p(\cdot)$ represent distance to a point where observations carry more information (at-the-money). In this paper, the proxies considered for these weight functions are open interest values associated with call and put option contracts.

4 Monte Carlo analysis

In this section, we present a simulation analysis that demonstrates effectiveness of the methods proposed in this study. Risk-neutral prices are not directly observable but can be inferred indirectly through option prices. Except in the case of simulated data, no confrontation between true and estimate RND can be done.

It is assumed that the stochastic process associated with an underlying asset is given by a diffusion process subjected to stochastic volatility, which can be represented as

$$\begin{aligned} dS_t &= \mu dt + \sqrt{v_t}\, S_t\, dW_t \\ dv_t &= \kappa(\theta - v_t)\, dt + \sigma\sqrt{v_t}\, dZ_t \end{aligned}$$

where W_t and Z_t are two standard Brownian motion processes with $E(dW_t\, dZ_t) = \rho\, dt$, κ represents the mean-reverting volatility parameter, θ the long-run volatility, and σ the volatility of volatility. Under certain assumptions, given in Heston (1993), there is a closed form solution for European-type option prices. The assumptions are related to a risk-premium function, and the current value of volatility. Henceforth, assuming a given value for the current volatility, prices of calls and puts for different strike values, are generated by what we refer as Heston (1993) model.

The parameters adopted are $\kappa = 5$, $\theta = 0.03$, $\sigma = 0.3$, and $\rho = -0.7$. In the simulation a zero dividend yield, a risk-free rate $r = 0.02$, and a time to maturity of 3-months ($\tau = 0.25$), are adopted. It is assumed that, at t the price of the underlying asset is $S_t = 50$, and the range of strikes is given by the interval $[35, 62]$. For prices obtained using the model, random noise was added to mimic observed market prices. Using these perturbed prices, we illustrate the performance of nonparametric methods developed in

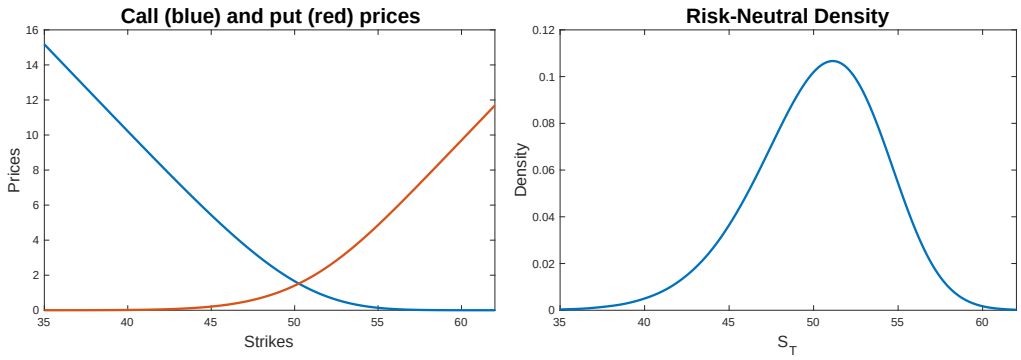

Figure 1: Heston's model prices and RND

this paper. With Heston's model, no analytic formula is available to express RNDs. Nowadays, we can generate Heston's prices for calls and puts within a fine grid of strikes, and approximate RNDs smoothly and accurately using second derivatives calculated numerically. For parameters' values defined above this kind of calculations were performed and are depicted in Figure 1. An equally-spaced of fine strike prices was considered and numerical second derivatives were calculated, presenting a great stability. True RND can be obtained with high-accuracy levels, which is a positive function that integrates one, slightly negatively asymmetric corresponding to a negative value assumed by the leverage effect $\rho = -0.7$.

This kind of data is corrupted with noise, considering the example of intraday data for a given strike, different option prices values can be observed. To reproduce market data, for each strike a theoretical price is calculated, and a series of observations are simulated adding some noise to the prices. Following Yatchew and Härdle (2006), and assuming already a nonparametric framework, we used the formula $y_i = m(x_i) + 0.03\, m(x_i)\varepsilon_i$, with $\varepsilon_i \sim N(0,1)$, where for each strike one thousand observations were simulated. Resulting prices and comparisons between true and estimated RND are depicted

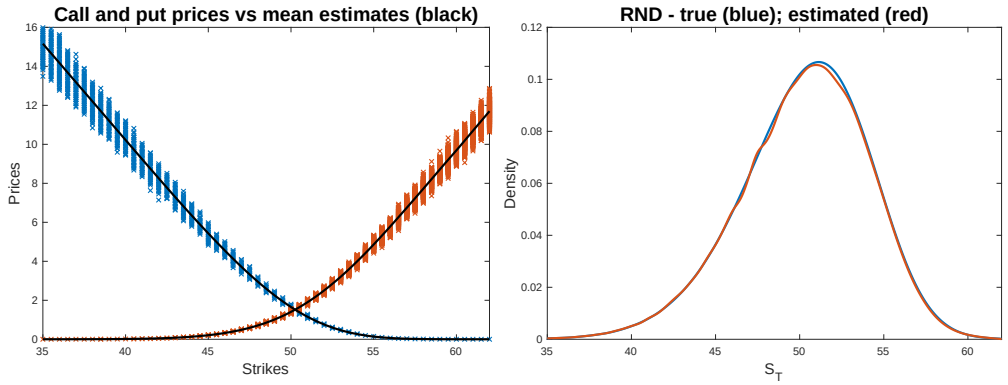

Figure 2: Simulated prices; mean and RND estimates

in Figure 2. We obtain a good overall fitting and the expected behaviour for density tails. This is a very controlled environment with equally-spaced strikes, and equal number of observations for each strike, however, prices are highly perturbed. As it is also demonstrated in the literature associated with RND estimation, obtaining a good fit for the mean pricing function through noparametric estimation methods does not constitute a great challenge, in opposition to its second derivative estimation (Aït-Sahalia and Duarte, 2003; Yatchew and Härdle, 2006; Birke and Pilz, 2008; Grith et al., 2012).

5 Empirical demonstration

In this section, we provide nonparametric density estimations using S&P500 and VIX options. Each underlying asset gives raise to a huge number of option contracts. In fact, there are different maturities available, and also many strikes. Considering only vanilla options, significant amount of put and call contracts can make difficult data retrieving, treatment and application of common data-cleaning procedures. For these tasks it is necessary to build

data structures able to accommodate the diversity and complexity in data. The number of contracts and maturity dates available for a given stock can be different for calls and puts. Finally, most trades are done near-the-money, which varies with the underlying asset price. Option contracts that are deep-in-the-money or deep-out-the-money are rarely traded, which means that observed prices can carry different amounts of information.

5.1 Data description

We use intraday data for options associated with two indices, S&P500 and VIX, from CBOE. Options related to S&P500 index correspond to a SPXW version of contracts. Data was collected from the publicly available site `YahooFinance` using tailored software to record observations for every contract (strikes, maturities, etc.), during regular daily negotiation time. The sample corresponds to observations obtained from April 16 to April 20, 2018.

Despite the short period represented by one week, due to aforementioned diversity the number of observed contracts are of orders 10^4 and 10^3. Added to this fact, the frequency of observed data is 5 minutes, and each contract was considered 390 times in the sample. For SPXW with time-to-maturity of around one month, 15 964 and 21 508 observations were considered for calls and puts, respectively, whereas, for VIX, 1 465 and 1 196 observations are available. These numbers are comparable with ones considered in literature that uses nonparametric methods to estimate RND functions, see e.g. Aït-Sahalia and Lo (1998), Song and Xiu (2016).

Following the literature, which can be clearly understood due to heterogeneity associated with options data, data-cleaning procedures are needed. The main aim is to remove troublesome data points in terms of compatibility with theoretical results, e.g. no-arbitrage constraints, and points that are

irrelevant as correspond to contracts that have never been traded. In our approach, data-cleaning is reduced to minimal procedures. First, we eliminate duplicate observations that result from working with high-frequency data. Second, option contracts with bid or open interest equal to zero are also eliminated. This last step assumes less importance since we use open interest as a weight in the estimation procedures which eliminates automatically such observation.

In the literature, out-of-the-money and in-the-money call and put prices are considered separately because they accommodate different information. This is addressed by using the well-known put-call parity formula, converting put into call prices. This conversion is not absent of difficulties, especially because lack of synchronization, which can only be softened using end-of-day data (Song and Xiu, 2016), but not totally resolved. We include directly in estimation procedures call and put contract prices. By taking advantage of this inclusion, avoiding using put-call parity conversions, new information is added allowing better estimates to be obtained.

As the sample period considered is short, less variation of elements that determine option prices need to be accounted for. Examples are risk-free interest rate, dividend yield and underlying stock prices. Long sample periods may rise concerns about structure maintenance of prices. In contrasting with Aït-Sahalia and Lo (1998) that have considered around one year of daily data, and Song and Xiu (2016) with seventeen years, we use a short period of intraday data that allow us to compare results with the referred ones in terms of data dimension. As we deal with a short sample period, we only considered contracts with a fixed short maturity date. Time-to-maturity is fixed, and in RND estimation setting, this resembles a cross-section approach.

Aït-Sahalia and Lo (1998) and Song and Xiu (2016) performed an analysis

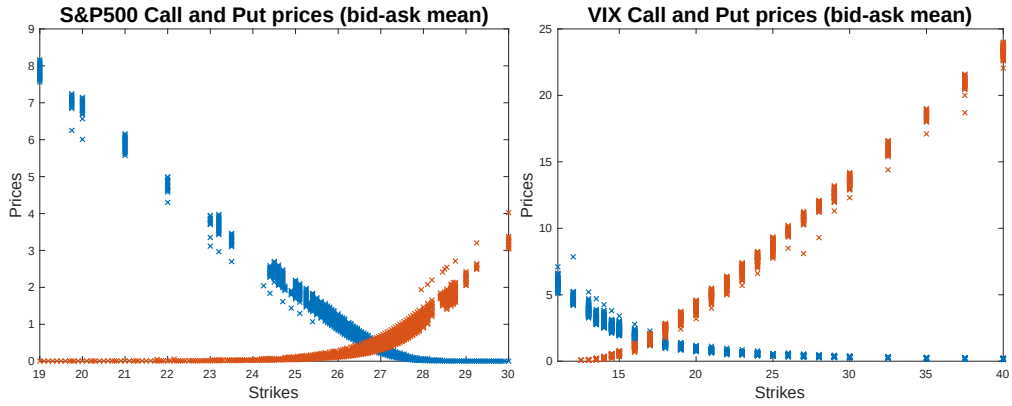

Figure 3: Call and put bid-ask mean prices with expiration date May 18, 2018 ($\tau = 1/12$) on S&P 500 index, and May 16, 2018 ($\tau = 1/12$) on VIX index.

by rolling forward contracts, and defined results for a mean maturity period. Due to using intraday data, we do not need such data manipulation procedure. Considering options in aforementioned sample, a time-to-maturity of around one month was considered, May 18 for S&P500, and May 16 for VIX. Bid-ask mean prices were used to represent observed prices, and for each strike, a series was obtained, for calls and puts. Data is depicted in Figure 3, where graphics were truncated to show most relevant parts. For S&P500, strikes and prices were scaled by a 10^-2 factor, representing a change of unity justified by pricing functions homogeneity, and allowing a better visualization of data.

5.2 Estimation results

Using intraday data, for each strike a series of prices is obtained. A mean function can be approximated for each strike by simple averaging option

prices. Considering yet a nonparametric approach, even with a local constant estimator (Nadaraya-Watson), mean function estimates cannot be ruled out by any economic or statistical argument. It is assumed that option pricing functions are twice differentiable, and are expected to be monotone and convex, consequently for an interior point the estimator is consistent. However, when first and second derivatives are estimated, statistical properties degrade with a substantial decrease of convergence rates. More importantly, estimates start to lack economic sense, and go against established theoretical results. These facts are well-established in Aït-Sahalia and Duarte (2003) and Yatchew and Härdle (2006), which constitute main motivations for presenting new methods capable of dealing with such drawbacks.

Considering the estimation applied to S&P500 data set, and as we used one hundred units to refer S&P500 data, the less troublesome region is defined by strikes between 24 and 28, for calls and puts. Below strike 24, for calls, strikes grid is sparser, the same happens for puts with values above 28. Regions where information for calls is lacking are compensated by information from puts, and vice versa. For the performed estimation, time-to-maturity was set to $\tau = 1/12$ and, based on Treasury Bills data, a value of 2% was considered to be the risk-free interest rate r. Our main contribution was to devise a method able to cope with such different sets within the estimation of a unique RND.

Results for S&P500 are depicted in Figure 4, and reveal the effectiveness of proposed methods. Mean pricing functions are not difficult to estimate, however, if we consider only call prices (put prices) a greater variability can be observed for mean pricing functions. This affects RND estimation for the left tail using just calls, and right tail for puts. We use calls and puts within a unique criterion function, imposing no-arbitrage constraints. A smooth

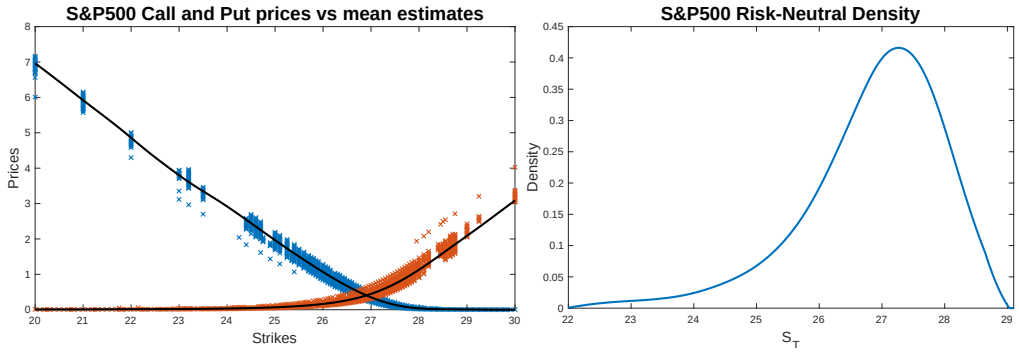

Figure 4: Estimated mean function and RND for call and put bid-ask mean prices for contracts with expiration date of May 18, 2018 ($\tau = 1/12$) on the S&P500 index.

and reliable RND is estimated guaranteeing no-negative density values, an expected behaviour for tails, and an area under the curve near one. Estimation procedures were also applied to VIX, with similar data problems as found for S&P500. The same effect of information compatibility and smoothness is reflected for RND estimation, which is depicted in Figure 5. Performance of different estimators from Nadaraya-Watson (local constant) though local linear, local quadratic, and local cubic is substantially different. The improvements in results are significative. Nadaraya-Watson estimator gives a less acceptable estimated mean function, and for the density, severe problems at tails (Aït-Sahalia and Duarte, 2003; Yatchew and Härdle, 2006). This effect is softened by considering higher order polynomials, but meaningful results can be obtained only by imposing a set of no-arbitrage constraints. We are able to define such in a very natural manner using a unique criterion function.

Finally, we have to highlight the difference in shape of both RND estimates (Figures 4 and 5). Results are intuitive and confirm what seems to

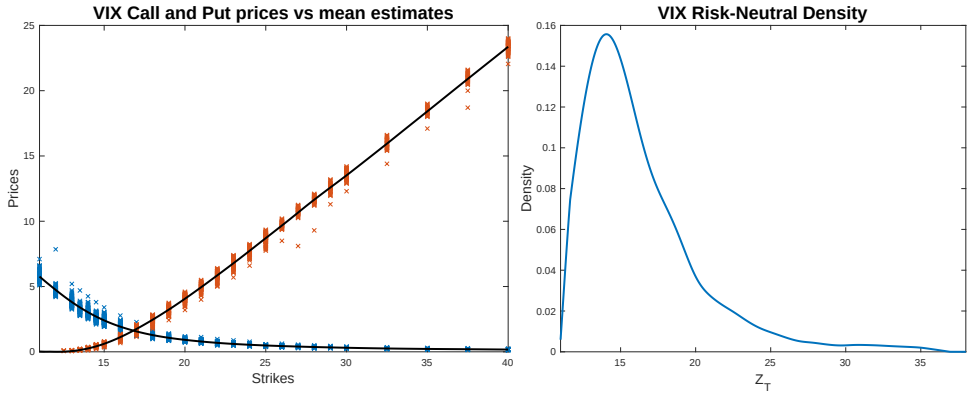

Figure 5: Estimated mean function and RND for call and put bid-ask mean prices for contracts with expiration date of May 16, 2018 ($\tau = 1/12$) on the VIX index.

be expected considering the different nature of the underlying assets. When subscribing S&P500 option options, the main interest is the left tail, since it reveals a possibility of drop in prices, or eventually a default. For VIX, the main interest is the right tail, as it is related to a possibility of an increase in volatility. This difference is clearly reflected in the estimation performed is this paper, and seems to reinforce the meaningfulness of proposed extension.

6 Concluding remarks

This paper has developed and tested a new nonparametric approach for estimating RNDs from European option prices, using intraday data. The resulting problem is a quadratic programming problem, with a convex objective function, linear constraints, upper and lower bounds on variables. This is a challenging problem since RNDs are obtained through estimates for second derivatives. Naive approaches do not guarantee non-negativity, integration

to one, and RND smoothness. Although by defining a problem that includes calls, puts, and respective weights, it is guaranteed to obtain aforementioned features. Using simulated data we demonstrated that the method is able to recover, with acceptable accuracy, true RNDs. We applied the method to S&P500 and VIX options with results that are robust and easily interpretable. Comparison between both RNDs reveals main motivations for subscribing such securities: protection against decreases (S&P500) and increases (VIX) on values of respective underlying assets.

References

Aït-Sahalia, Y. and J. Duarte (2003). Nonparametric option pricing under shape restrictions. *Journal of Econometrics 116*(1), 9–47.

Aït-Sahalia, Y. and A. W. Lo (1998). Nonparametric estimation of state-price densities implicit in financial asset prices. *The Journal of Finance 53*(2), 499–547.

Aït-Sahalia, Y. and A. W. Lo (2000). Nonparametric risk management and implied risk aversion. *Journal of Econometrics 94*(1), 9–51.

Aït-Sahalia, Y., Y. Wang, and F. Yared (2001). Do option markets correctly price the probabilities of movement of the underlying asset? *Journal of Econometrics 102*(1), 67–110.

Banz, R. W. and M. H. Miller (1978). Prices for state-contingent claims: Some estimates and applications. *Journal of Business*, 653–672.

Birke, M. and K. F. Pilz (2008). Nonparametric option pricing with no-arbitrage constraints. *Journal of Financial Econometrics 7*(2), 53–76.

Breeden, D. T. and R. H. Litzenberger (1978). Prices of state-contingent claims implicit in option prices. *Journal of Business*, 621–651.

Cox, J. C. and S. A. Ross (1976). The valuation of options for alternative stochastic processes. *Journal of Financial Economics 3*(1-2), 145–166.

Dalderop, J. (2018). Nonparametric filtering of conditional state-price densities. Technical report.

Fan, J. and I. Gijbels (1996). *Local Polynomial Modelling and Its Applications*, Volume 66. CRC Press.

Fengler, M. R. and L.-Y. Hin (2015). Semi-nonparametric estimation of the call-option price surface under strike and time-to-expiry no-arbitrage constraints. *Journal of Econometrics 184*(2), 242–261.

Grith, M., W. K. Härdle, and M. Schienle (2012). Nonparametric estimation of risk-neutral densities. In *Handbook of Computational Finance*, pp. 277–305. Springer.

Härdle, W. (1990). *Applied Nonparametric Regression*. Number 19. Cambridge university press.

Härdle, W. and Z. Hlávka (2009). Dynamics of state price densities. *Journal of Econometrics 150*(1), 1–15.

Heston, S. L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies 6*(2), 327–343.

Jackwerth, J. C. (2000). Recovering risk aversion from option prices and realized returns. *The Review of Financial Studies 13*(2), 433–451.

Li, Q. and J. S. Racine (2007). *Nonparametric Econometrics: Theory and Practice*. Princeton University Press.

Monteiro, A. M., R. H. Tütüncü, and L. N. Vicente (2008). Recovering risk-neutral probability density functions from options prices using cubic splines and ensuring nonnegativity. *European Journal of Operational Research 187*(2), 525–542.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications 9*(1), 141–142.

Rosenberg, J. V. and R. F. Engle (2002). Empirical pricing kernels. *Journal of Financial Economics 64*(3), 341–372.

Song, Z. and D. Xiu (2016). A tale of two option markets: Pricing kernels and volatility risk. *Journal of Econometrics 190*(1), 176–196.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 359–372.

Yatchew, A. (2003). *Semiparametric Regression for the Applied Econometrician*. Cambridge University Press.

Yatchew, A. and W. Härdle (2006). Nonparametric state price density estimation using constrained least squares and the bootstrap. *Journal of Econometrics 133*(2), 579–599.

Zhang, X., R. D. Brooks, and M. L. King (2009). A bayesian approach to bandwidth selection for multivariate kernel regression with an application to state-price density estimation. *Journal of Econometrics 153*(1), 21–32.

CEBER WORKING PAPERS

(Available on-line at www.uc.pt/go/ceber)

2019-02 *Kernel density estimation using local cubic polynomials through option prices applied to intraday data – Ana Margarida Monteiro & António Alberto Santos*

2019-01 *High Performance Work Systems and Employee Outcomes: A Meta-analysis for Future Research – Pedro Diogo & José Fontes da Costa*

2018-15 *Endogeneity Issues in the Empirical Assessment of the Determinants of Loan Renegotiation – José Valente, José Murteira & Mário Augusto*

2018-14 *Why are credit booms sometimes sweet and sometimes sour? - Vítor Castro & Rodrigo Martins*

2018-13 *Copula-based Tests for Nonclassical Measurement Error – The Case of Fractional Random Variables - José M. R. Murteira*

2018-12 *Differentiated Impact of Spread Determinants by Personal Loan Category: Evidence from the Brazilian Banking Sector – José Valente, Mário Augusto & José Murteira*

2018-11 *The Effect of Family Ownership, Control and Management on Corporate Debt Structure – Evidence from Panel Fractional Data – Mário Augusto, José Murteira & António Pedro Pinto*

2018-10 *Predictability of stock returns and dividend growth using dividend yields: An international approach – Ana Sofia Monteiro, Hélder Sebastião & Nuno Silva*

2018-09 *Political and institutional determinants of credit booms – Vítor Castro & Rodrigo Martins*

2018-08 *An Aggregate View of Portuguese Exports and Competitiveness – Pedro Bação, António Portugal Duarte & Diogo Viveiros*

2018-07 *The Cycle of recycling and sustainable development. Evidence from the OECD Countries – Pedro André Cerqueira, Elias Soukiazis & Sara Proença*

2018-06 *Information Transmission Between Cryptocurrencies: Does Bitcoin Rule the Cryptocurrency World? – Pedro Bação, António Portugal Duarte, Hélder Sebastião & Srdjan Redzepagic*

2018-05 *Endogenous Growth and Entropy – Tiago Neves Sequeira, Pedro Mazeda Gil & Óscar Afonso*

2018-04 *Determinants of overall and sectoral entrepreneurship: evidence from Portugal – Gonçalo Brás & Elias Soukiazis*

2018-03 *Young and healthy but reluctant to donate blood: An empirical study on attitudes and motivations of university students – Tiago Henriques & Carlota Quintal*

2018-02 *The Iberian electricity market: Price dynamics and risk premium in an illiquid market – Márcio Ferreira & Hélder Sebastião*

2018-01 *Health Investment and Long run Macroeconomic Performance: a quantile regression approach – Francisca Silva, Marta Simões & João Sousa Andrade*

2017-12 *Deflation in the Euro Zone: Overview and Empirical Analysis – Pedro Bação & António Portugal Duarte*

2017-11 *Fiscal Consolidation Programs and Income Inequality – Pedro Brinca, Miguel H. Ferreira, Francesco Franco, Hans A. Holter & Laurence Malafry*

2017-10 *The interconnections between Renewable Energy, Economic Development and Environmental Pollution. A simultaneous equation system approach - Elias Soukiazis, Sara Proença & Pedro André Cerqueira*

2017-09 *The Renminbi: A Warrior for Competitiveness? – Pedro Bação, António Portugal Duarte & Matheus Santos*

2017-08 *Le Portugal et l'Euro – João Sousa Andrade*

2017-07 *The Effect of Public Debt on Growth in Multiple Regimes in the Presence of Long-Memory and Non-Stationary Debt Series - Irina Syssoyeva-Masson & João Sousa Andrade*

2017-06 *The Blank and the Null: An examination of non-conventional voting choices – Rodrigo Martins*

2017-05 *Where is the information on USD/Bitcoins hourly price movements? - Helder Sebastião, António Portugal Duarte & Gabriel Guerreiro*

2017-04 *The response of non-price competitiveness and productivity due to changes in passed income gaps. Evidence from the OECD countries - Pedro André Cerqueira, Micaela Antunes & Elias Soukiazis*

2017-03 *Dutch Disease in Central and Eastern European Countries - João Sousa Andrade & António Portugal Duarte*

2017-02 *On the gains of using high frequency data and higher moments in Portfolio Selection- Rui Pedro Brito, Hélder Sebastião & Pedro Godinho*

2017-01 Growth adjustments through non-price competitiveness and productivity. A cumulative causation approach- Elias Soukiazis, Micaela Antunes & Pedro André Cerqueira

A série CeBER Working Papers foi iniciada em 2017.