Université de Montréal

Nonparametric Estimation of Risk Neutral Density

Présenté par
Adjimon Marcel Djossaba

Évalué par un jury composé de :

René Garcia
Président-Rapporteur

Marine Carrasco
Directrice de mémoire

Benoit Perron
Membre du jury

Département de sciences économiques
Faculté des arts et des sciences

Mémoire présenté à la Faculté des arts et des sciences en vue de l'obtention du grade de maitrise en sciences économiques, option recherche

15 Octobre 2023

Résumé

Ce mémoire vise à estimer la densité neutre au risque (Risk neutral density (RND) en anglais) par une approche non paramétrique tout en tenant compte de l'endogénéité. Les prix transversaux des options européennes sont utilisés pour l'estimation. Le modèle principal considéré est la régression linéaire fonctionnelle. Nous montrons comment utiliser des variables instrumentales dans ce modèle pour corriger l'endogénéité. En outre, nous avons intégré des variables instrumentales dans le modèle approximant le RND par l'utilisation des fonctions d'Hermite à des fins de comparaison des résultats. Pour garantir un estimateur stable, nous utilisons la technique de régularisation de Tikhonov. Ensuite, nous effectuons des simulations de Monte-Carlo pour étudier l'impact des différents types de distribution RND sur les résultats obtenus. Plus précisément, nous analysons une distribution de mélange lognormale et une distribution de smile de Black-Scholes. Les résultats des simulations démontrent que l'estimateur utilisant des variables instrumentales pour corriger l'endogénéité est plus performant que l'alternative qui ne les utilise pas. En outre, les résultats de la distribution de smile de Black-Scholes sont plus performants que ceux de la distribution de mélange log-normale. Enfin, $S\&P$ 500 options sont utilisées pour une application de l'estimateur.

Mots clés : Densité neutre au risque, Tarification des options, Variable instrumentale, Régression fonctionnelle, Méthode de régularisation.

Abstract

This thesis aims to estimate the risk-neutral density (RND) through a non-parametric approach while accounting for endogeneity. The cross-sectional prices of European options are used for the estimation. The primary model under consideration is functional linear regression. We have demonstrated the use of instrumental variables in this model to address endogeneity. Additionally, we have integrated instrumental variables into the model approximating RND through the use of Hermite functions for the purpose of result comparison. To ensure a stable estimator, we employ the Tikhonov regularization technique. Following this, we conduct Monte-Carlo simulations to investigate the impact of different RND distribution types on the obtained results. Specifically, we analyze a lognormal mixture distribution and a Black-Scholes smile distribution. The simulation results demonstrate that the estimator utilizing instrumental variables to adjust for endogeneity outperforms the non-adjusted alternative. Additionally, outcomes from the Black-Scholes smile distribution exhibit superior performance compared to those from the log-normal mixture distribution. Finally, $S\&P$ 500 options are used for an application of the estimator.

Keywords : Risk neutral density, Option pricing, Instrumental variable, Functional regression, Regularization method.

Table des matières

Remerciements

Je tiens à remercier ma directrice de recherche, Marine Carrasco, pour sa disponibilité durant tout ce moment passé sous son encadrement. Merci d'avoir accepté d'encadrer ce travail.
Je remercie les membres de jury, René Garcia et Benoit Perron d'avoir accepté de lire ce mémoire.
Je remercie également Raphael Godefroy pour tous ses nombreux conseils, sans oublier Mathieu Marcoux.
J'exprime enfin ma reconnaissance à tous mes amis en recherche dans le département de sciences économiques de l'Université de Montréal.

Nonparametric Estimation of Risk Neutral Density

By

Marcel Djossaba*

Under the guidance of

Marine Carrasco†

November 3, 2023

---

*Student, Department of Economics, Université de Montréal. Email: adjimon.marcel.djossaba@umontreal.ca

†Professor, Department of Economics, Université de Montréal. Email: marine.carrasco@umontreal.ca

1 Introduction

The estimation of the risk neutral density (RND) remains an important concern in finance. Indeed, the risk-neutral density is a density function that contains rich information about market expectations regarding the price process of the underlying assets and is used to value illiquid derivatives; assess market beliefs; estimate investors' risk preferences; manage risks; market timing; etc. On a theoretical level, it is well known that in the risk-neutral valuation framework, any European-style derivative can be valued by taking the expectation of its payoff function with respect to the risk-neutral distribution and then discounted at the risk-free rate over the life of the contract (see Cox and Ross (1976)). For example, the mathematical expression for the price of a European put option is: $P_t = e^{-r_{t,T}(T-t)} \int_0^\infty max(\kappa - S_T, 0) f(S_T|(T-t))\, dS_T$; where S_T is the stock price at maturity, κ is the strike price, $f(S_T|(T-t))$ is the unobserved risk-free density (RND) of S_T conditional to the maturity T, $r_{t,T}$ is the riskless interest rate between date t and T.

The main challenge in this asset pricing problem is to be able to recover this unobserved function $f(S_T|(T-t))$. Indeed, there is a large literature on how to estimate this risk neutral density (RND). Two categories can be distinguished: parametric methods and non-parametric methods (including semi-parametric methods). Parametric methods assume that the RND function can be approximated by a fixed distribution and estimate the parameters of this distribution by minimizing prediction errors (see Rompolis and Tzavalis (2008), Corrado and Su (1996) who consider a log-normal distribution, Bahra (1997) who uses a mixture of log-normal distribution). However, this class of methods has demonstrated to be insufficiently flexible to adapt to real-world option prices. Indeed, as proven by Ait-Sahalia and Lo (2000), parametric methods have the advantage of involving only a few parameters, so that the calculation process is not cumbersome; but, they have disadvantages if an inappropriate process is assumed, or if a distribution is chosen that is not flexible enough to adapt to the data. Thus non-parametric methods (including semi-parametric methods), for their part, rather than assuming a probability distribution, opt for more general functions in order to obtain greater flexibility in the adjustment of option prices based on certain criteria. We can cite: Garcia and Gençay (2000) for the neural network method; Bondarenko (2003) for the positive convolution method; Carrasco and Tsafack (2023) for the functional analysis technique; Lu and Qu (2021) for a sieve estimator; Cui and Xu (2022) for the risk-neutral density established using exact representations of the Dirac delta function.

What all these papers have in common is the consideration of the explanatory variable, the strike price κ, as an exogenous variable. However, this exogeneity assumption may not hold as the determination of the strike price depends on numerous factors in the stock market; and the correlation between these latter factors (for example, the volatility or return of an asset) and the strike price cannot be overlooked. This is because the strike price is a fixed price at which a derivative contract can be bought or sold when it is exercised; and like all commodities, an imbalance between supply and demand will raise and lower the stock price. Thus, any variable correlated to the demand and supply of a stock would explain the strike price. This may be the current price (price on date t) of the stock, its volatility, its transaction cost, etc.

Based on these observations, the main idea of this work is to develop an estimator of RND that allows endogeneity in the model and to study its properties. Concretely, the paper proposes to address this problem of the endogeneity using instrumental variables in a nonparametric approach that is functional regression. The proposed estimator is indeed the solution of a Fredholm equation of the first kind which is an ill-posed problem because of the lack of stability of the solution. The Tikhonov regularization is used to avoid this problem of the instability. We analyse the performance of the proposed estimator using some Monte-Carlo simulations. For that, we make the assumption of lognormal mixture distribution for the true density as in Carrasco and Tsafack (2023). The results of these simulations show the importance of considering the endogeneity, but the performance is poor. We deal with this issue by considering Black-Scholes smile distribution as in Lu and Qu (2021). We therefore modify the way they generate data to account for endogeneity. We find a good performance suggesting that the performance depends on the true distribution considered.

The remainder of the paper is organized as follows. Section 2 describes the estimation method and suggests a data-driven method to select the optimal hyper parameter. Section 3 presents on overview of the related literature. Section 4 explains how to compute the estimator. It also presents the Monte-Carlo simulations results. Section 5 focuses on the impact of the distribution of the true density considered on the results of prediction, using Black-Scholes smile and lognormal mixture distributions. Section 6 presents an application case and section 7 concludes.

2 Definition and estimation of RND

This section shows how to obtain a consistent risk-neutral density estimator in the presence of endogeneity using the Tikhonov regularization technique.

2.1 Definition

In Cox and Ross (1976), the price of a European Put option P_t of a stock is expressed as follows:

$$P_t = e^{-r_{t,T}(T-t)} \int_0^{\infty} max(\kappa - S_T, 0) f(S_T|(T-t))\, dS_T \tag{1}$$

where S_T is the stock price at maturity, κ is the strike price, $f(S_T|(T-t))$ is the unobserved risk-free density (RND) of S_T conditional on the maturity T, $r_{t,T}$ is the riskless interest rate between date t and T.

Equation (1) means that, in the context of a complete market, the price of a European put option P_t with a maturity T, with an underlying price at maturity S_T and a strike price κ, is equal to the expected pay-offs $Z(S_T) = max(\kappa - S_T, 0)$ discounted back to the present.

In reality, markets are incomplete, meaning that some payoffs cannot be replicated by trading in marketed securities (Staum (2007)). Hence, the previous equation must be rewritten to take into account an error term. So, let's consider put options indexed by i P_i contracted at the same fixed date t with different corresponding strike price κ_i and of same end of the contract T. Then,

$$Y_i = \int_0^{\infty} Z_i(s) f(s|\tau)\, ds + \epsilon_i \tag{2}$$

where $Y_i = P_i e^{r_{t,T}(T-t)}$, $Z_i(s) = max(\kappa_i - s, 0)$, $\tau = T - t$.

In Carrasco and Tsafack (2023), the authors estimated directly this equation by assuming that the explanatory variable, κ_i is exogenous and the error term is conditionally zero-mean and homoskedastic.

This exogeneity hypothesis is not obvious and may not hold because the determination of the strike price depends on some factors in the stock market; and the correlation between these factors and the strike price cannot be overlooked (see Roberts and Whited (2012) for sources of endogeneity in Empirical Corporate Finance). Indeed, the strike price is a set price at which a derivative contract can be bought or sold when it is exercised; and like all commodities, an imbalance between supply and demand will raise and lower the price of stock. Thereby, any variable correlated with demand and supply of a stock would explain the strike price.

For instance, considering the S&P 500 index (SPX) data used for the real application in Carrasco and Tsafack (2023), we observe a strong negative correlation (around -0.90) between strike price and transaction cost (bid-ask spread) as shown in figure 1

One can also see the strike price as the future price of a stock from its definition. And in

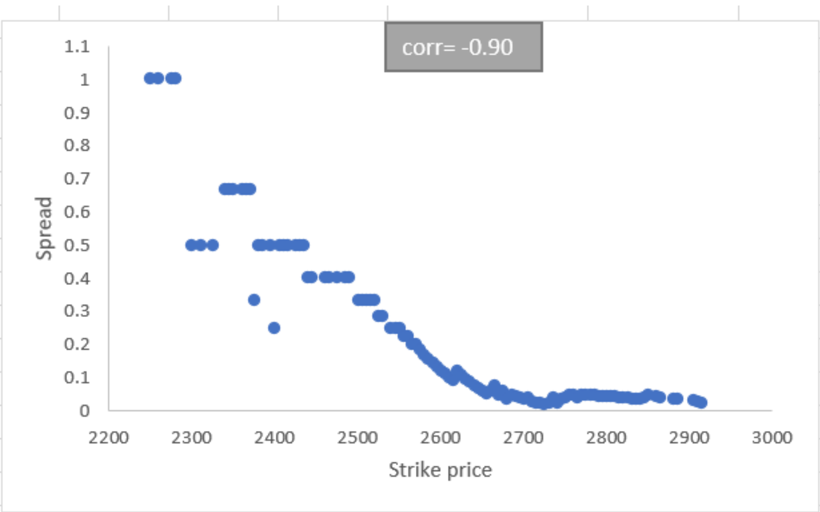

Figure 1: Strike price and spread
(Constructed from S&P 500 index (SPX) data)

this case, the strike price depends on the Volatility of its past prices of security. Indeed, the volatility is the rate at which the price of a stock fluctuates over a particular period. Higher stock price volatility often means higher risk and helps an investor to evaluate the uncertainty that may happen in the future.

These different variables could be very important in estimating of the RND and not considering them would bias the estimate.

For this reason, in this paper, the explanatory variable κ_i is considered as an endogenous variable and the estimation of the RND density will be done with instrumental variable.

2.2 Estimator of RND $\hat{f}$ using IV

Let $\mathcal{H} = L^2([0,\infty))$ be the space of square integrable functions mapping from the interval $[0,\infty)$ to the set of real numbers $\mathcal{R}$. $\mathcal{H}$ is a Hilbert-space endowed with an inner product $< .,. >$ and a norm $\|.\|$, which are respectively defined as follows : $< f, g >= \int_0^\infty f(t)g(t)dt$ and $\|f\| = \int_0^\infty (f^2)^{1/2}$.
Let us consider the sample $(Y_1,\kappa_1,W_1),(Y_2,\kappa_2,W_2),...,(Y_n,\kappa_n,W_n)$ of independent variables following the same distribution as the population version (Y,κ,W). Let $Z_i(s) = max(\kappa_i - s, 0)$ the pay-off at maturity of the put option with the strike price κ_i ($Z_i(s)$, for each s, is random only through κ_i). Z_i is a continuous time process indexed by by s and is an element of $\mathcal{H}$
We consider the following functional linear model :

$$Y_i = \int_0^\infty Z_i(s)f(s)\,ds + \epsilon_i \tag{3}$$

where Y is the scalar dependent variable. A possible instrument is W.
The identification in the linear IV regression relies on the uncorrelatedness between the instrumental variable W and the unobservables ϵ, and the rank condition. One of the requirements of the rank condition is that the number of instrumental variables corresponds to the dimension of the endogenous covariate. In our settings, the endogenous covariate Z is a function (high-dimensional realization), hence, one would need a high-dimensional instrumental variable W to identify f. This condition may or may not be satisfied. We therefore distinguish between two cases.

Case1: Condition of dimension of W is satisfied. That means W is a function like Z. In this case, we assume that W is an element of $\mathcal{H}$ such that: $E(W\epsilon) = 0$.
Then, by premultiplying both sides of Equation (3) by $W_i(u)$ and taking the expectation, we obtain:

$$E(W_i(u)Y_i) = \int_0^\infty E(W_i(u)Z_i(s))f(s)\,ds + E(W_i(u)\epsilon_i)$$

$$\text{Then} \quad E(W_i(u)Y_i) = \int_0^\infty E(W_i(u)Z_i(s))f(s)\,ds, \text{ since } E(W_i(u)\epsilon_i) = 0$$

We can write the compact form of this equation as following:

$$C_{WY} = Kf \tag{4}$$

where $C_{WY}(u) = E(W_i(u)Y_i)$ is the cross-covariance function between the instrumental variable W and the response variable Y and K is the covariance operator from $\mathcal{H}$ to $\mathcal{H}$ defined as:

$$Kf = \int_0^\infty E(W_i(u)Z_i(s))f(s)\,ds$$

Case2: Condition of dimension of W is not satisfied.
We consider the case where W_i is a finite dimensional random variable and $E(\epsilon|W) = 0$. Babii

(2021) called this type of regression a high-dimensional mixed-frequency IV Regression.
Under this assumption, we have:

$$E(Y_i|W=w) = \int_0^\infty E(Z_i(s)|W=w)f(s)\,ds$$

Estimating directly this equation is not convenient. Thus, Babii (2021) shows that it is possible to estimate $f(s)$ under the condition of the linear completeness property of the distribution of $(Z;W)$[1]. He proposes to convert the conditional moment condition into unconditional moment condition using a square-integrable function of the instrumental variable, denoted Ψ. In his paper, he recommended the logistic CDF: $\Psi(u,W) = \frac{1}{1+exp(-u^TW^1))}$; where $W^1 = (1, W^T)^T$, $u \in V \subset \mathcal{R}^{1+q}$, with V is a compact set with nonempty interior.
With this transformation, $\Psi_i(u,W)$ plays exactly the role of $W_i(u)$ in the previous case1. So, we have:

$$E(\Psi_i(u,W)Y_i) = \int_0^\infty E(\Psi_i(u,W)Z_i(s))f(s)\,ds$$

Then, we have the similar compact form:

$$C_{\Psi Y} = K_\Psi f \tag{5}$$

Our goal is to estimate the density function f solution of equation (4) or equation (5) and verify the validity of the instrument used. Those equations are indeed integral equations, also called Fredholm equations of the first kind (see Carrasco et al. (2007)). Such equations in f are ill-posed problems because of the lack of stability of the solution. To overcome this problem, we will propose a regularized solution of f. The Tikhonov regularization technique is proposed to stabilize the solution.

2.3 Estimation using Tikhonov method

We want to estimate f solution of equation (4): $C_{WY} = Kf$ (it is the same technique for the equation 5: $C_{W_\Psi Y} = K_\Psi f$.). Let K^* be the adjoint operator of K, this operator is such that: $< Kf, g >=< f, K^*g >$, for all $f \in H$ and $g \in H$.
We multiply both sides of equation (4) by K^*.

$$K^*Kf = K^*C_{WY} \tag{6}$$

The Tikhonov regularization solution is characterized by:

$$\hat{f}^\alpha = (\alpha I + K^*K)^{-1}K^*C_{WY} \tag{7}$$

where α (with $\alpha > 0$) is a regularization parameter which will be allowed to converge to zero as n goes to infinity. This estimator can be viewed as a generalization of ordinary least-squares estimator (with Ridge regularization).

[1]We say that the stochastic process $Z \in \mathcal{H}$ is linearly complete for W if for all $b \in \mathcal{H}$ with $E| < Z, b > | < \infty$, we have: $E(< Z, b > /W) = 0 \implies b = 0$. The linear completeness is a generalization of the rank condition imposed in the finite-dimensional linear IV regression.

2.4 Choice of instrument W and form of Ψ

This subsection gives some ideas about relevant instrumental variables that could be used. The instrument must be correlated with the strike price κ (because the endogenous variable Z is random only through κ) and must be exogenous. Since, the strike price is a set price at which a security can be bought or sold when the option is exercised; and like all commodities, an imbalance between supply and demand will raise and lower the price of stock. Thereby, any variable correlated with demand and supply of a stock would explain the strike price. Thus, the strike price of the security at time t may depend on its past price at the previous time $t-1$. The volatility of the security (using its past prices) or the performance of a share on the stock market influence the fixation of the strike price.

Basically, the past volatility or the past return of the asset could be potential instrumental variables, because these variables are correlated with strike price. Besides, asset returns are only slightly self-correlated (see Cont (2001)).
As we work in the case of European options, i.e. we consider options contracted on the same day and which mature all on the same day, the past price would be nothing other than the price just before the signature of the contract (just before the sale or purchase of the option) and as for the volatility or the return, it would take into account all the prices just before the signing of the contract. For example, if we consider the $S\&P500$ index, we will need to know its price at every moment of the day, when each European option based on this index is contracted.

As can be seen, the instrumental variable will likely be a scalar. Thus, the use of a square-integrable function is necessary to convert a conditional moment condition into an unconditional moment condition. The following two functions are considered in this paper:

- $\Psi(u, W) = \frac{1}{1+exp(-u*W)}$, for $u \in [0, 1]$
- $\Psi(u, W) = max(0, W - u)$, for $u \in [min(W), max(W)]$

Both provide fairly close results. Therefore, either can be used.

2.5 Validation of the instrument

It would be important to verify if the instrument used is effectively a valid instrument. Indeed, the instrument W must verify two important conditions:

1. Correlation with the explanatory variable Z: the covariance function between Z and W (or Ψ) is not null.
2. Exogeneity with the error term: $E(W\epsilon) = 0$ or $E(\epsilon|W) = 0$

For this purpose, an improvement of Hansen (1982)'s J-test proposed by Amengual et al. (2020) in the case of a continuum of moment conditions would be used.

2.6 Correction of $\hat{f}^\alpha$

The estimated Risk neutral density $\hat{f}^\alpha$, like any density, must keep the properties of positivity and integrability to one. However, the estimation technique used does not guarantee these properties. For that reason, one need to correct the estimator. We propose to correct $\hat{f}^\alpha$ into a density noted $\tilde{f}^\alpha$ using the following simple correction:

$\tilde{f}^\alpha(s) = \frac{max(0,\hat{f}^\alpha(s))}{\int_0^\infty max(0,\hat{f}^\alpha(s))ds}$.
$\tilde{f}^\alpha$ is positive and its integral is equal to one.

2.7 Selection of the hyperparameter α

In order to have a stable solution, the estimated RND depends on the hyperparameter α which must be obtained optimally. Since the main goal is to estimate the RND and therefore predict the put prices, we define a prediction criterion to select the optimal hyperparameter α. That means we choose the regularization parameter in such a way that the mean squared prediction error (MSPE) is minimized. We use the K-fold cross-validation for the selection procedure. Indeed, we split the initial sample in K subsamples denoted I_1,...,I_K ($K = 10$ in this work by following numerous articles in the literature such as Carrasco and Tsafack (2023) or Lu and Qu (2021)). To do this:

- We divide the number of observations n by K, denoted nb, i.e, $nb = n/K$;
- Then, th first K-1 groups are made up of observations equal to the integer part of nb, and the remaining observations form the last group.

$\alpha_{op} = \underset{\alpha \in I_l}{argmin} \frac{1}{K} \sum_{l=1}^{K} \frac{1}{card(I_l)} \sum_{j \in I_l} (\hat{Y}_j - Y_j)^2$
For $l \in \{1, 2, ..., K\}$, we estimate f in sample I_{-l} representing all observations except those in I_l. With this $\hat{f}$, we predict the response variable in I_l: $\hat{Y}_j$ is the prediction of the j^{th} observation in I_l. The MSPE is calculated for each value of α in I_α, where I_α is the set of candidate α.

3 Related literature

In this section, we will begin by examining several methodologies used to estimate the risk-neutral density. Next, we will reference papers that are related to the proposed technique of functional regression. Lastly, we will provide an overview of the IV method.

3.1 Literature on RND

Several papers have been written on the estimation of the Risk-Neutral density (RND), with different methods of estimation, which preserve the properties of density (positiveness and integration to one). These methods can be categorized as: Parametric methods and semi-parametric or non-parametric methods. Considering the parametric techniques, they start with a distribution involving a set of parameters and make adjustments to the assumed distribution.

Then based on the distribution one can price the options and determine the parameters of the distribution by minimizing the pricing error. The most classical approach in this category is modeling the RND as a log-normal distribution (see Rompolis and Tzavalis (2008), Corrado and Su (1996) and Jarrow and Rudd (1982)) and a log-normal mixture distribution (see Bahra (1997)). One can also see the paper of Fabozzi et al. (2009) which extracts the RND with a three-parameter distribution, such as the generalized gamma distribution, that falls between the Weibull and the generalized beta distributions.

The second category is non-parametric (including semi-parametric) methods. These methods, rather than assuming a probability distribution, use more general functions to achieve greater flexibility in fitting option prices using certain criteria.

One well-known approach within this category is the smoothed implied volatility smile method (SML). SML utilizes the fact that the second derivative of the option price function is proportional to the RND; that means, to recover the RND, one needs to estimate the second derivative of the option (put or call) price.

Approaches within the semi and non-parametric methods category are diverse. Ait-Sahalia and Lo (2000) suggest to estimate the function option price by nonparametric kernel regression and then differentiate it twice to recover the RND. However, the resulting estimator is very volatile and the authors use a semiparametric approach based on Black and Scholes formula and realized volatility. To be fully non-parametric, Ait-Sahalia and Duarte (2003) propose to use a constrained locally polynomial kernel smoothing (see also Grith et al. (2012)) which has the advantage to reduce the variance.

Methods to fit the option pricing function using neural network are proposed by Garcia and Gençay (2000), Ludwig (2015). An extension of the concept of mixture density networks method was proposed by Schittenkopf and Dorffner (2001). Positive convolution methods is proposed by Bondarenko (2003). Estimating the risk-neutral density based on support vector regression is proposed by Feng and Dang (2016).

Some papers use flexible functions such as polynomials or splines to fit the option prices or implied volatilities and then convert them to the RND or fit the RND directly. One can cite : Monteiro et al. (2008), Yijun et al. (2012), Barletta et al. (2019), Monteiro and Santos (2022). Recently, Lu and Qu (2021) propose a sieve estimator for RND by using Hermite approximation; Jiang et al. (2021) estimate the RND by Piecewise Constant Nonparametric Approach. They propose per intervalle of strike price, a constant that is estimating by cross-validation. Carrasco and Tsafack (2023) use the functional analysis technique to estimate RND. They are the first authors who use this technique (to our knowledge). Cui and Xu (2022) propose a novel representation of the risk-neutral density established using exact representations of the Dirac Delta function.

3.2 Literature on estimation with functional regression

This paper is related to the literature on functional data analysis. For a general treatment, see the books by Bosq (2000), Ferraty and Vieu (2000). Theoretical results on the model where the predictor is a function and the response is a scalar are developed by Cai and Hall (2006), Delaigle and Hall (2012), Tsafack (2020), etc.. The more general case where both predictor and response variables are functions is discussed by Benatia et al. (2017) (the authors studied in this paper the theorical case of endogeneity of explanatory variables), Kargin and Onatski (2008), Chang et al. (2016), etc..

3.3 Literature on IV

In this paper, we propose to use IV and verify its validity because of endogeneity. The following papers talk about endogeneity analysis. Roberts and Whited (2012) presented the different sources of endogeneity and discuss on some econometric techniques to address this issue (see also Hill et al. (2021)). Wang (2010) propose to use instrumental variables approach to correct the problem of endogeneity in finance and Andrews et al. (2019) treats the case of weak instrumental variables. For the instrumental variables treatment in non-parametric models, see Horowitz (2011) and Amengual et al. (2020). For a general discussion in the case of non (semi) parametric, see the book Dewatripont et al. (1997) (chapter8). The most recent paper in this field and most related to our work is that of Babii (2021) which treats the case of low-frequency IV and proposes the solution (a continuous function) to overcome this issue.

3.4 Contribution

Based on these previous findings, this work focuses on the introduction of instrumental variables to estimate RND, using functional analysis technique. This case has not previously been treated in the literature. This paper can be seen as the continuity of Carrasco and Tsafack (2023).

4 Computation and simulation of the estimator

In this section, we show how to compute the estimator in both exogenous and endogenous cases, focusing on the various integrals without the need for discretization, thus preserving crucial information. Indeed, we can use the trapezoidal rule to approximate all integrals as in Carrasco and Tsafack (2023). However, we observed that the results are not satisfactory, in particular excessive noise in the density tails. Consequently, we use the following method to avoid discretizing certain integrals. We use this computational technique to run simulations in each case.

Remember: The goal is to estimate f such that:

$$Y_i = \int_0^\infty Z_i(s) f(s|\tau)\, ds + \epsilon_i \tag{8}$$

where $Y_i = P_i e^{r_{t,T}(T-t)}$, $Z_i(s) = max(\kappa_i - s, 0)$, $\tau = T - t$
N.B: In practice, s takes values between a and b, finite values. That means we solve:
$Y_i = \int_a^b Z_i(s) f(s|\tau)\, ds + \epsilon_i$.

4.1 Exogenous case

4.1.1 Computation of the estimator

We have: $E(Z_i(l)Y_i) = \int_0^\infty E(Z_i(l)Z_i(s))f(s|\tau)\,ds + E(Z_i(l)\epsilon_i)$

$E(Z_i(l)Y_i) = \int_0^\infty E(Z_i(l)Z_i(s))f(s|\tau)\,ds$ because $E(Z_i(l)\epsilon_i) = 0$

The compact form of this equation is: $C_{ZY} = Kf$
where $C_{ZY} = E(Z_i(l)Y_i)$ is a cross-covariance function between Z and Y.
K is a covariance operator from $\mathcal{H}$ to $\mathcal{H}$.

The regularized estimation of f based on Tikhonov regularization is: $\hat{f}_\alpha = (\hat{K} + \alpha I)^{-1}\hat{C}_{ZY}$, where $\hat{K}$ and $\hat{C}_{ZY}$ are the sample counterparts of K and C_{ZY}.

Let define $\mathcal{R}^n$ the space of $n \times 1$ vectors endowed with inner product $v'w/n$ for any pair of $n \times 1$ vectors v and w and T the operator from $\mathcal{H}$ to $\mathcal{R}^n$ such that:

$$Tf = \begin{bmatrix} < Z_1, f > \\ . \\ . \\ < Z_n, f > \end{bmatrix} \qquad \text{for any function } f \text{ in } \mathcal{H}$$

Let T^* be the adjoint operator of T such that: $\qquad (T^*v)(s) = \frac{1}{n}\sum_{i=1}^n Z_i(s)v_i$

With these notations, we can remark that: $\hat{C}_{ZY} = T^*Y$ and $\hat{K} = T^*T$

Then the Tikhonov estimator is: $\hat{f}_\alpha = (T^*T + \alpha I)^{-1}T^*Y$

This equation is equivalent to: $\hat{f}_\alpha = T^*(TT^* + \alpha I)^{-1}Y = T^*(C + \alpha I)^{-1}Y$, with $C = TT^*$. Indeed, the equivalence comes from Engl et al. (1999).

- $C = TT^*$ is such that: for a vector g

$$(Cg) = \begin{bmatrix} < Z_1, (T^*g) > \\ . \\ . \\ < Z_n, (T^*g) > \end{bmatrix} = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n < Z_1, Z_i > g_i \\ . \\ . \\ \frac{1}{n}\sum_{i=1}^n < Z_n, Z_i > g_i \end{bmatrix}$$

Then, $C = TT^* = \frac{1}{n}\begin{bmatrix} < Z_1, Z_1 > & < Z_1, Z_2 > & . & . & < Z_1, Z_n > \\ < Z_2, Z_1 > & < Z_2, Z_2 > & . & . & < Z_2, Z_n > \\ . & . & . & . & . \\ . & . & . & . & . \\ < Z_n, Z_1 > & < Z_n, Z_2 > & . & . & < Z_n, Z_n > \end{bmatrix}$, $\quad C$ is $n \times n$ matrix.

- To compute T^*, we need to consider a grid of values of $s = s_1, s_2, ..., s_m$ in order to have an approximation, denoted $\underline{T^*}$. We therefore estimate the vector: $\underline{\hat{f}}_\alpha = \begin{bmatrix} \hat{f}_\alpha(s_1) \\ \hat{f}_\alpha(s_2) \\ . \\ . \\ . \\ \hat{f}_\alpha(s_m) \end{bmatrix}$ an approximation of $\hat{f}_\alpha$. $\underline{T^*}$ is $m \times n$ matrix such that:

$$\underline{T^*} = \frac{1}{n} \begin{bmatrix} Z_1(s_1) & Z_2(s_1) & . & . & Z_n(s_1) \\ Z_1(s_2) & Z_2(s_2) & . & . & Z_n(s_2) \\ . & . & . & . & . \\ . & . & . & . & . \\ Z_1(s_m) & Z_2(s_m) & . & . & Z_n(s_m) \end{bmatrix}$$

- $< Z_i, Z_j >$ is computed as following:

$< Z_i, Z_j >= \int_a^b Z_i(s)Z_j(s)ds = \int_a^b max(0, \kappa_i - s)max(0, \kappa_j - s)ds$

-if $\kappa_i \leq \kappa_j$, then we have:

$$\begin{aligned} < Z_i, Z_j > &= \int_a^{\kappa_i} (\kappa_i - s)(\kappa_j - s)ds && (9) \\ &= \int_a^{\kappa_i} (\kappa_i * \kappa_j - (\kappa_i + \kappa_j)s + s^2)ds && (10) \\ &= (\kappa_i - a)\kappa_i\kappa_j - \frac{1}{2}(\kappa_i + \kappa_j)(\kappa_i^2 - a^2) + \frac{1}{3}(\kappa_i^3 - a^3) && (11) \end{aligned}$$

-if $\kappa_i > \kappa_j$, then we have:

$$\begin{aligned} < Z_i, Z_j > &=< Z_j, Z_i > && (12) \\ &= (\kappa_j - a)\kappa_j\kappa_i - \frac{1}{2}(\kappa_j + \kappa_i)(\kappa_j^2 - a^2) + \frac{1}{3}(\kappa_j^3 - a^3) && (13) \end{aligned}$$

Now, the estimator can be computed in the exogenous case. It is evident that the inverse matrix is independent of the discretization value of the integral.

4.1.2 Simulation results

The Monte-Carlo simulations are used to show how the proposed computation method works. For that, we generate the data for different values of n, representing the number of observations. This simulation follows the same idea as the one by Carrasco and Tsafack (2023) with a change on how to compute the estimator.

The simulation procedure is as follows.

1. Generate $\epsilon \sim \mathcal{U}(-\sqrt{3}, +\sqrt{3})$ (n observations) to reflect the variation of errors in reality in a similar way to Carrasco and Tsafack (2023);

2. Generate κ a grid of n observations between 1500 and 3000, where the interval [1500, 3000] corresponds to the range of κ values observed in the application done by Carrasco and Tsafack (2023);

3. Define for s, a grid of m values between $min(\kappa)-1$ and $max(\kappa)+1$ (In practice, $m = 1200$ represents a large dimension of covariates which are initially a function, see Carrasco and Tsafack (2023).);

4. Define a matrix of pay-offs Z such that $Z[i,j] = max(\kappa_i - s_j, 0)$ where $i = 1, 2, ..., n$ and $j = 1, 2, ..., m$

5. Compute the true density of f as a lognormal mixture distribution : $f(s) = \pi_1 \mathcal{LN}(s|\eta_1, \sigma_1) + \pi_2 \mathcal{LN}(s|\eta_2, \sigma_2) + \pi_3 \mathcal{LN}(s|\eta_3, \sigma_3)$, with $\pi_1 + \pi_2 + \pi_3 = 1$ ($\pi_1 = 0.0812, \pi_2 = 0.0914, \pi_3 = 0.8274$) ;

6. Compute $y = Zf + \varepsilon$ as following: $y_i = \sum_{j=1}^{m} Z[i,j] * f(s_{j+1})(s_{j+1} - s_j) + \varepsilon_i$;

7. Run the 10-fold cross-validation procedure in order to select the optimal hyperparameter. The whole sample is randomly divided into 10 groups. 9 groups are used as the training sample and the last one is used as the the validation sample.The validation sample is defined to select the optimal hyperparameter α. This operation is repeated 10 times as we have 10 different groups in the cross-validation procedure.

8. Compute an estimation of f denoted $\hat{f}$ using the selected optimal hyperparameter. We call this estimator, Tikhonov estimator;

9. Calculate $\tilde{f}$, corrected Tikhonov estimator ($\tilde{f}(s) = \frac{max(0,\hat{f}(s))}{\int max(0,\hat{f})ds}$);

10. Calculate the root mean-square prediction error (RMSPE) by 10-fold cross-validation;

11. Repeat the steps 1 to 10 with 200 iterations and calculate the average RMSPE, average Tikhonov estimator and average corrected Tikhonov estimator over the simulations.

N.B: f is such that $\mathcal{LN}(s|\eta_j, \sigma_j) =$ with $\eta_1 = 7.8020, \eta_2 = 7.7023, \eta_3 = 7.8052$ and $\sigma_1 = 0.0285, \sigma_2 = 0.0987, \sigma_3 = 0.0245$.
The following three figures (2, 3, 4) represent the average estimators computed with three different numbers of observations (1000, 500 and 200) and 200 iterations. We can see that the results are very good regardless of the values of n. And also, the corrected estimate almost coincides with the uncorrected one. We can therefore say that the proposed estimator performs well in the exogenous case.

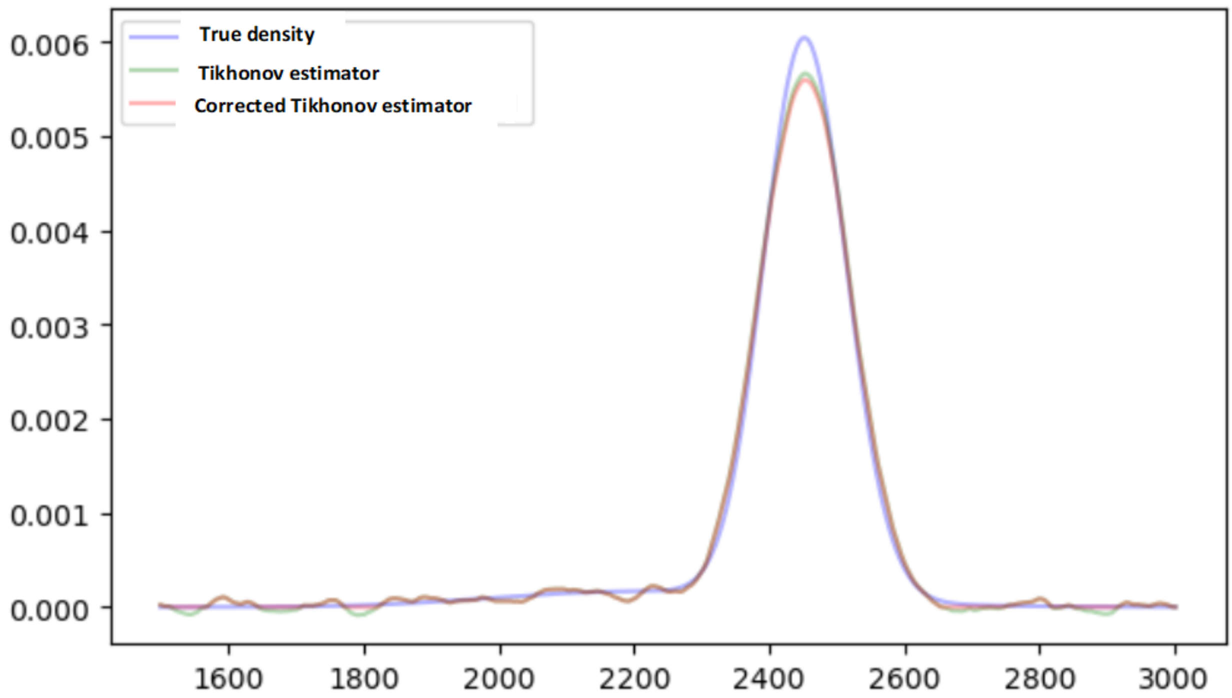

Figure 2: Data generated with exogenous error term. f average of 200 replications. Number of observations n=1000.
Blue curve is true density, green curve is estimated density and red curve is estimated density corrected

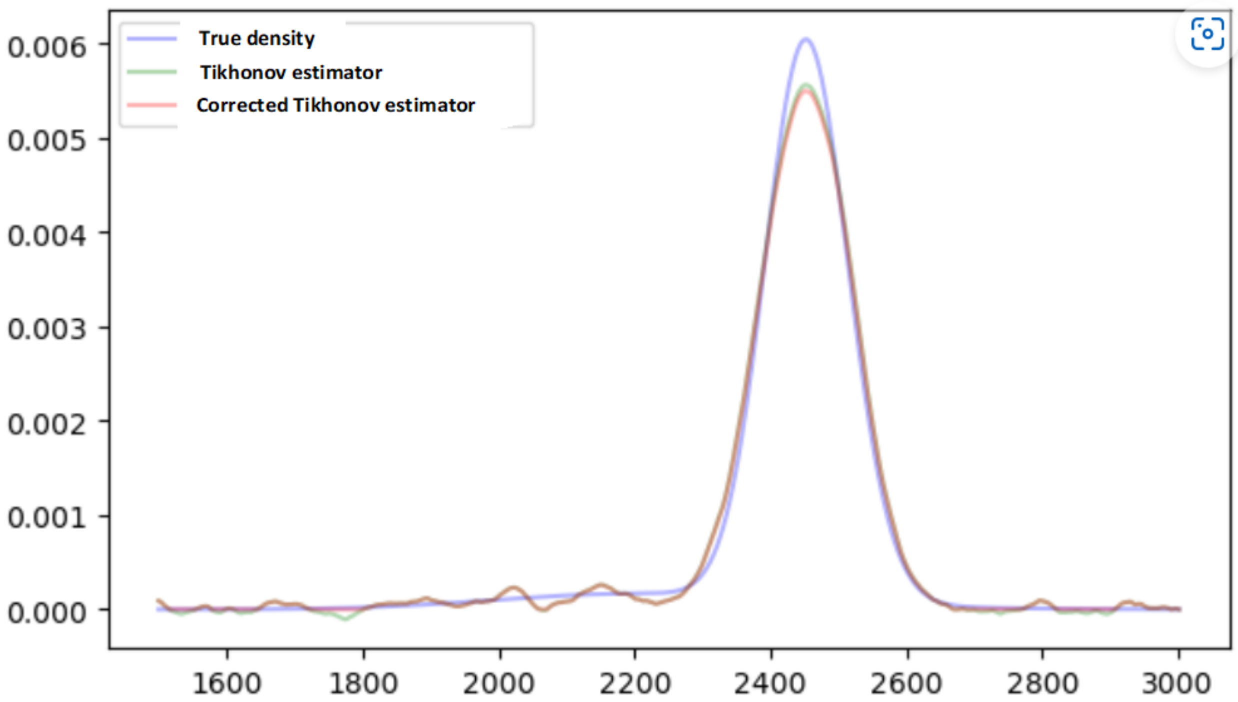

Figure 3: Data generated with exogenous error term. f average of 200 replications. Number of observations n=500.
Blue curve is true density, green curve is estimated density and red curve is estimated density corrected

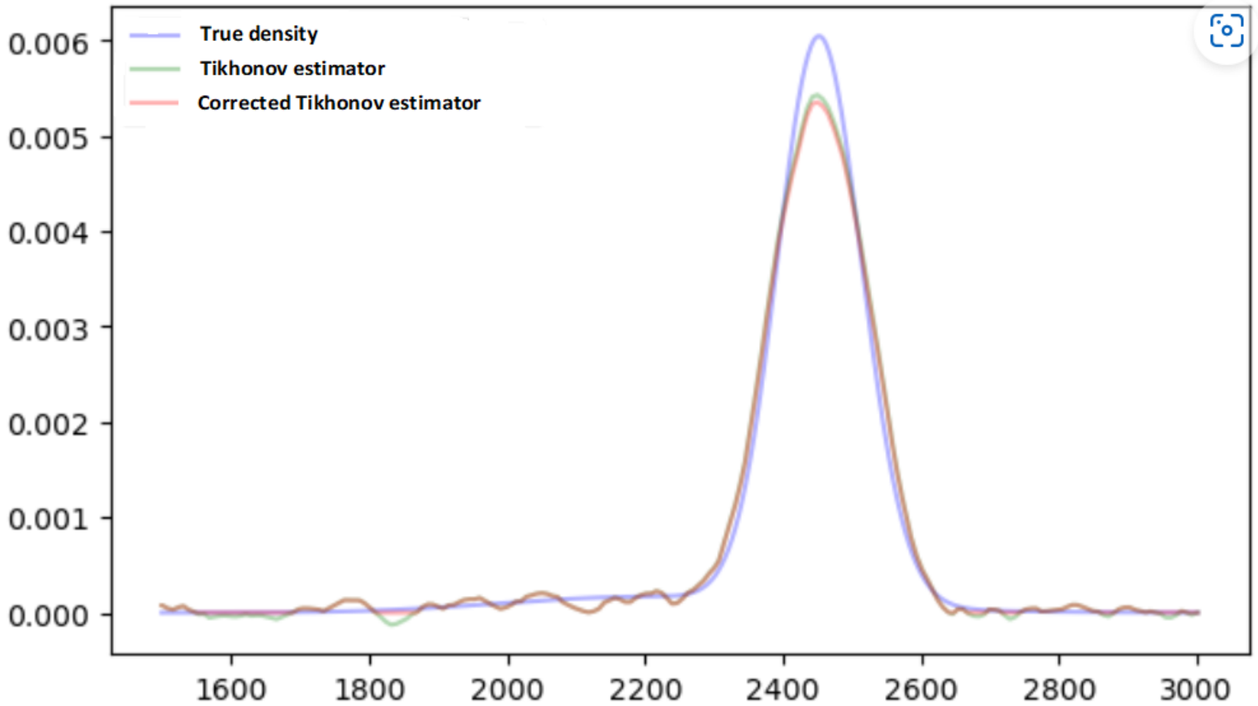

Figure 4: Data generated with exogenous error term. f average of 200 replications. Number of observations n=200.
Blue curve is true density, green curve is estimated density and red curve is estimated density corrected

4.2 Endogenous case

4.2.1 Computation

Now, we will demonstrate how to calculate the estimator, focusing on directly solving the involved integrals, when utilizing the instrumental variable noted as W to tackle the issue of endogeneity, which is the primary objective of this research.
For that, we have: $E(\Psi_i(l)Y_i) = \int_a^b E(\Psi_i(l)Z_i(s))f(s|\tau)\,ds + E(\Psi_i(l)\epsilon_i)$, where Ψ_i is a function of W_i and l.

$E(\Psi_i(l)Y_i) = \int_a^b E(\Psi_i(l)Z_i(s))f(s|\tau)\,ds$, because $E(\Psi_i(l)\epsilon_i) = 0$

The compact form is: $C_{\Psi Y} = C_{\Psi Z} f$

Then, $C_{Z\Psi}C_{\Psi Y} = C_{Z\Psi}C_{\Psi Z}f$, because the adjoint operator of $C_{\Psi Z}$ is $C_{Z\Psi}$

The Tikhonov estimator is: $\hat{f}^\alpha = (C_{Z\Psi}C_{\Psi Z} + \alpha I)^{-1}C_{Z\Psi}C_{\Psi Y}$

This solution is equivalent to: $\hat{f}^\alpha = C_{Z\Psi}(C_{\Psi Z}C_{Z\Psi} + \alpha I)^{-1}C_{\Psi Y}$

Now, we explain how to compute $C_{\Psi Z}C_{Z\Psi}$. Note that:

$C_{\Psi Z} C_{Z\Psi}\omega = C_{\Psi Z} g, \quad \text{where} \quad g = C_{Z\Psi}\omega$

$$C_{\Psi Z} C_{Z\Psi}\omega(t) = \int \frac{1}{n}\sum_i \Psi_i(t) Z_i(s) g(s) ds, \ for \ t \in [a,b] \tag{14}$$

$$= \frac{1}{n}\sum_i \Psi_i(t) \int \frac{1}{n} Z_i(s) g(s) ds, \ for \ t \in [a,b] \tag{15}$$

$$= \frac{1}{n}\sum_i \Psi_i(t) \int Z_i(s)(\frac{1}{n}\sum_j Z_j(s) \int \Psi_j(t)\omega(t)dt) ds, \ for \ t \in [a,b] \tag{16}$$

$$= \frac{1}{n}\sum_i \Psi_i(t) \frac{1}{n}\sum_j \int Z_i(s) Z_j(s) ds \int \Psi_i(t)\omega(t)dt, \ for \ t \in [a,b] \tag{17}$$

$$= \frac{1}{n^2}\sum_i \sum_j \Psi_i(t)(\int Z_i(s) Z_j(s) ds) \int \Psi_i(t)\omega(t)dt, \ for \ t \in [a,b] \tag{18}$$

So, $C_{\Psi Z} C_{Z\Psi}\underline{\omega} = \frac{1}{n}\Phi' C \Phi \underline{\omega}$

where: $\underline{\omega} = \begin{bmatrix} \omega(t_1) \\ \omega(t_2) \\ . \\ . \\ \omega(t_k) \end{bmatrix}$

-$C = \frac{1}{n}\begin{bmatrix} < Z_1, Z_1 > & < Z_1, Z_2 > & . & . & < Z_1, Z_n > \\ < Z_2, Z_1 > & < Z_2, Z_2 > & . & . & < Z_2, Z_n > \\ . & . & . & . & . \\ . & . & . & . & . \\ < Z_n, Z_1 > & < Z_n, Z_2 > & . & . & < Z_n, Z_n > \end{bmatrix}$ is $n \times n$ matrix, the matrix computed in exogenous case.

An approximation of Φ, denoted $\underline{\Phi}$, is given by considering a grid of values of $t = t_1, t_2, ..., t_k$. $\underline{\Phi} = \begin{bmatrix} \Psi_1(t_1) & \Psi_1(t_2) & . & . & \Psi_1(t_k) \\ \Psi_2(t_1) & \Psi_2(t_2) & . & . & \Psi_2(t_k) \\ . & . & . & . & . \\ . & . & . & . & . \\ \Psi_n(t_1) & \Psi_n(t_2) & . & . & \Psi_n(t_k) \end{bmatrix}$ is $n \times k$ matrix with $\Psi_i(t_j) = max(0, w_i - t_k)$

Thus, an approximation of $\hat{f}_\alpha$ is: $\underline{\hat{f}^\alpha} = \underline{C_{Z\Psi}}(C_{\Psi Z} C_{Z\Psi} + \alpha I)^{-1}\underline{C_{\Psi Y}}$ with $\underline{C_{Z\Psi}} = \underline{T}^*\underline{\Phi}$ and $\underline{C_{\Psi Y}} = \frac{1}{n}\underline{\Phi}' y$

Finally, the Tikhonov estimator can be computed in the endogenous case.

4.2.2 Simulations in the endogenous case

For this simulation part, we modify just a few steps of the exogenous case to allow endogeneity. Simulation procedure is as follows. Data are generated (in particular steps 1 to 4) to preserve the range of κ values as in the exogenous case (see previous simulation), and to check IV conditions, such as the correlation between W and κ, and the absence of correlation between W and ϵ.

1. Generate $v \sim \mathcal{U}(-1/2, +1/2)$ ($n = 1000$ observations);
2. Generate the instrument variable $W \sim \mathcal{U}(1650, 2850)$ ($n = 1000$ observations);
3. Compute $\kappa = w + 300 * v$;
4. Compute error terms $\epsilon = 100 * v$;
5. Define s, as a grid of $m = 1200$ values between 1499 and 3001;
6. Follow Steps 5 to 11 of the exogenous simulation part. The difference here is that we determine two estimators and calculate RMSPE and RISE:
   - Tikhonov estimation of f without considering endogeneity, that means we estimate f with the method described in the exogenous case;
   - Estimation of f using instrument, so by considering the presence of endogeneity.

Indeed, we use two criteria to compare the results:

- The Root Mean Squared Prediction Error (RMSPE) between the estimated put prices $\hat{Y}$ and the theoretical one Y. It is the root of MSPE corresponding to the optimal hyperparameter presenting in 2.7:
  $RMSPE = \sqrt{\frac{1}{K}\sum_{l=1}^{K} \frac{1}{card(I_l)} \sum\limits_{j \in I_l} (\hat{Y}_j - Y_j)^2}$.
- The Root Integrated Squared Error (RISE) between the estimated density $\hat{f}$ and the theoretical one f:
  $RISE = \sqrt{\int_0^\infty (\hat{f}_\alpha(s) - f(s))^2 ds}$, with $\hat{f}_\alpha$ is the RND corresponding to the optimal hyperparameter α. In practice, we calculate an approximation of this error as follows:
  $\widehat{RISE} = \sqrt{\sum_{j=1}^{m} (\hat{f}_\alpha(s_j) - f(s_j))^2 \Delta s_j}$, where $\Delta s_j = s_j - s_{j-1}$, for $j = 2, 3, ..., m$ and $\Delta s_1 = s_2 - s_1$.
  N.B: $m = 1200$ is the number of points considered to calculate f.

The following figure 5 presents the results of the simulations: blue curve is true density, green curve and red curve are respectively density estimated without and with considering endogeneity. Using the RMSPE and RISE comparison criteria (see table 1), the estimator with IV outperforms the estimator without IV. Indeed, the RMSPE is equal to 0.001098 for the estimator when the IV is used, versus 0.001375 when the estimator is obtained by neglecting endogeneity. We can therefore say that the IV helps to estimate the RND.

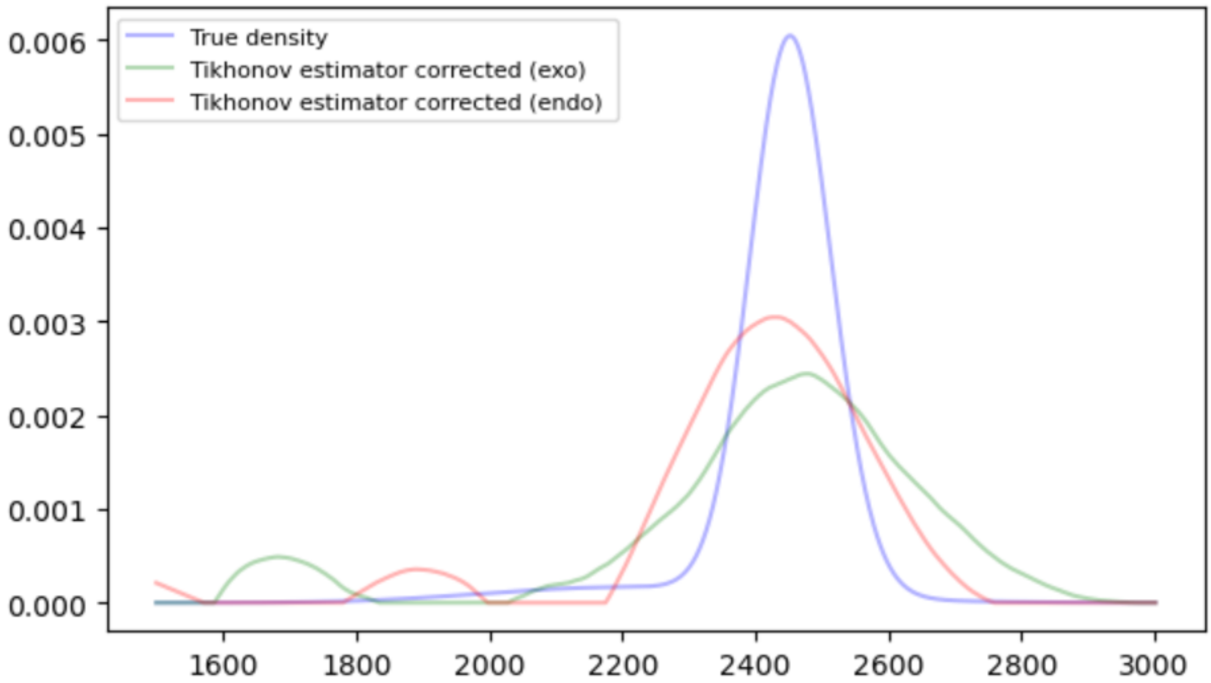

Figure 5: Data generated by introducing endogeneity and f estimated with and without considering this endogeneity.
Blue curve is true density, green curve and red curve density estimated without and with considering endogeneity.

	$RMSPE$	$RISE$
$\hat{f}_{IV}$	0.001098	0.00082
$\hat{f}$	0.001375	0.001104

Table 1: Comparison of the estimation results

4.3 Discussion

As we can see, we estimate and compare the estimator obtained by using IV with that obtained by neglecting the presence of endogeneity. We find that the estimator that takes endogeneity into account is more efficient than the one that neglects it.
Moreover, neither estimator performs well. That is, neither predicts the true density well. This raises the question: (i) Is it related to the way the data are generated? or (ii) Is it related to the assumption about the true density?
In the next section we will focus on the second question, because in the literature authors consider different distributions of the true density.

5 Comparison with sieve estimator

In this section, we consider the approach of Lu and Qu (2021). We first summarize their work and introduce the endogeneity within their estimation technique; and at the end we bring the distribution of the true density they use into our proposed estimation technique. .

5.1 Sieve estimator of RND

To estimate RND, Lu and Qu (2021) use a sieve estimator. Also, they didn't estimate the density of an underlying price at expiration S_T, but the density of their return $log(S_T/S_0)$. The problem is still ill-posed, but they used the Hermite approximation of the density, used put and call options together, considered a distribution that is not exactly normal but close to it as the true density, and transformed the initial data.
We summarize their idea and generate data like they did, while introducing endogeneity.

5.1.1 Description of the Sieve estimator

Lu and Qu (2021) propose a nonparametric estimator of the RND implied by a single cross-section of European options with different strikes and the same maturity.
They consider both put and call options:

$$C_t(\kappa) = e^{-r_t\tau} \int_0^{\infty} (S_T - \kappa)^+ f_t^*(S_T) dS_T \tag{19}$$

and

$$P_t(\kappa) = e^{-r_t\tau} \int_0^{\infty} (\kappa - S_T)^+ f_t^*(S_T) dS_T \tag{20}$$

where κ is the strike price, $\tau = T - t$, $(S_T - \kappa)^+ = max(S_T - \kappa, 0)$, r_t is the risk-free rate at time t, and S_T denotes the potential price of the underlying asset at time T.

There are two issues that need to be addressed for a sieve estimator:

1. The first concerns the fact that the RND is a dynamic object, so they standardize it by the following change of variables:

$$X = \frac{log(S_T/S) - r\tau}{\sigma\sqrt{\tau}} \tag{21}$$

and

$$\tilde{Z} = \frac{log(\kappa/S) - r\tau}{\sigma\sqrt{\tau}} \tag{22}$$

where $S = S_t$ is the spot price of the underlying asset at time t and σ is equal to the Black–Scholes implied volatility for at-the-money call option of the same underlying asset and maturity.
With these transformations, the rewriting of (19) and (20) give:

$$C(\tilde{Z}) = \int_{-\infty}^{\infty} S(e^{\sqrt{\tau}\sigma X} - e^{\sqrt{\tau}\sigma\tilde{Z}})^+ f(X) dX \tag{23}$$

and

$$P(\tilde{Z}) = \int_{-\infty}^{\infty} S(e^{\sqrt{\tau}\sigma\tilde{Z}} - e^{\sqrt{\tau}\sigma X})^+ f(X) dX \tag{24}$$

where $f(.)$ denotes density function after the change of variables.

Then, they recover f^* from f through

$$f^*(S_T) = \frac{1}{\sigma\sqrt{\tau}S_T} f(\frac{log(S_T/S) - r\tau}{\sigma\sqrt{\tau}}) \tag{25}$$

.

Also, the density of the log-return $R_T = log(S_T/S)$ is recovered : $f^*(R_T) = \frac{1}{\sigma\sqrt{\tau}} f(\frac{R_T - r\tau}{\sigma\sqrt{\tau}})$ (N.B: In the following, the paper focuses on the density of the log-return).

2. The second concerns the choice of basis functions for the approximation (because the sample size is usually small). Hermite functions are suitable basis functions for this approximation. This choice is based on three features of f:

    - the support of f has no fixed boundary,
    - f is closely related to the normal density,
    - f can have thicker tails than the normal distribution.

    So,

$$f(X) \sim \sum_{j=0}^{\infty} \beta_j h_j(X) \tag{26}$$

where $h_j(X) = \frac{H_j(X)}{(2^j j! \pi^{1/2})^{1/2}} e^{-X^2/2}$ with $H_j(X) = (-1)^j e^{X^2} \frac{d^j}{dX^j} e^{-X^2}$, $j = 0, 1, 2, 3, ...$
Then $\beta_j = \int_{-\infty}^{\infty} f(X) h_j(X) dX$

Practical implementation of sieve estimator.
Let $\{y_i, Z_i\}(i = 1, ..., n)$ denote the sample of observed option prices and transformed strike prices at time t. They assume that the first n_c observations are call options and the remaining $n - n_c$ observations are put options. Thus, they obtain the following system:

$$y_i = \begin{cases} \int_{-\infty}^{\infty} S(e^{\sqrt{\tau}\sigma X} - e^{\sqrt{\tau}\sigma \tilde{Z}_i})^+ f(X) dX + \varepsilon_i, & for\ i = 1, 2, ..., n_c \\ \int_{-\infty}^{\infty} S(e^{\sqrt{\tau}\sigma \tilde{Z}_i} - e^{\sqrt{\tau}\sigma X})^+ f(X) dX + \varepsilon_i, & for\ i = n_c + 1, ..., n \end{cases} \tag{27}$$

The approximation of f is: $f(X) \approx \sum_{j=0}^{J} \beta_j h_j(X)$, J is the truncation order and they suggest to choose $J = ceiling(2(n/logn)^{1/5})$.
To compute this f, they follow three steps:

1. Step1: For $j = 0, 1, ..., J$, compute

$$X_{i,j} = \begin{cases} \int_{-m}^{m} S(e^{\sqrt{\tau}\sigma X} - e^{\sqrt{\tau}\sigma Z})^+ h_j(X) dX, & for\ i = 1, 2, ..., n_c \\ \int_{-m}^{m} S(e^{\sqrt{\tau}\sigma Z} - e^{\sqrt{\tau}\sigma X})^+ h_j(X) dX, & for\ i = n_c + 1, ..., n \end{cases} \tag{28}$$

    where $m = max\{10, logn\}$.

2. Step2: Solve the optimization problem:
    $\min_{\beta \in \mathcal{H}_J} \sum_{i=1}^{n} (y_i - X_i^{'}\beta)^2 + \beta^{'} Q_\alpha \beta$
    Where:

- $X_i = (X_{i,0}, X_{i,1}, ..., X_{i,J})'$
- $\beta = (\beta_0, \beta_1, ..., \beta_J)'$ and $\mathcal{H}_J = \{\beta \in \mathcal{R}^{J+1} : \inf_{X \in \mathcal{R}} \sum_{j=0}^{J} \beta_j h_j(X) \geq \eta\}$. with η choosing between -10^{-4} and -10^{-3}
- Q_α is a $(J+1)$-dimensional regularization matrix. They use two specifications.
  - The first is $Q_\alpha = \alpha I$, I is an identity matrix.
  - The second is $Q_\alpha = VD_\alpha V'$ with $D_\alpha = diag[max\{\alpha - n\lambda_1, 0\}, max\{\alpha - n\lambda_2, 0\}, ..., max\{\alpha - n\lambda_{J+1}, 0\}]$, where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_{J+1}$ are the ordered eigenvalues of $n^{-1} \sum_{i=1}^{n} X_i X_i'$, and V contains the corresponding orthonormal eigenvectors as columns.
  - α is chosen by K-fold cross validation (here they choose $K = 10$).

The solution of this optimization leads to: $\hat{f}(X) = \sum_{j=0}^{J} \hat{\beta}_j h_j(X)$.

3. Step3 (This step is optional): correction of $\hat{f}$:
   $\tilde{f}(X) = max(0, \hat{f}(X) - c)$, where c is a constant is such that $\int_{-m}^{m} \tilde{f}(X)dX = 1$.
   For the authors, this step is optional, because the differences between $\tilde{f}(X)$ and $\hat{f}(X)$ are small.

5.1.2 Simulation of Sieve estimator: exogenous case

They consider an equal number of call n_c and put n_p options, $n_c = n_p = 25$ (and also $n_c = n_p = 50$) with strike prices which take values between 1000 and 1700; and have the same time to maturity $T = 30$. The spot price, interest rate, and dividend yield are equal to $S_t = 1365$, $r_t = 4.5\%$, and $\delta_t = 2.5\%$, respectively.

1. Generate κ, a grid of $n = 25$ values equally spaced between 1000 and 1700

2. Compute smile volatility which is a linear function of the strike price:
   $smile = \sigma(\kappa) = c + b * \kappa + a * \kappa^2$,
   where $a = 0$,
   $b = (\sigma(\kappa_{min}) - \sigma(\kappa_{max}))/(\kappa_{min} - \kappa_{max})$,
   and $c = \sigma(\kappa_{min}) - b * \kappa_{min}$;
   ($\sigma(\kappa_{min}) = 0.4 \quad \sigma(\kappa_{max}) = 0.2$)

3. Compute call price and put price using the Black-Scholes formula:

$$call_i = call(S_t, \kappa_i, r_t, \delta_t, T, \sigma(\kappa_i)) = S_t * exp(-\delta_t * T/365) * \Phi(d1) - exp(-r_t * T/365) * \kappa_i * \Phi(d2) \tag{29}$$

$$put_i = put(S_t, \kappa_i, r_t, \delta_t, T, \sigma(\kappa_i)) = exp(-r_t * T/365) * \kappa_i * \Phi(-d2) - S_t * exp(-\delta_t * T/365) * \Phi(-d1) \tag{30}$$

   where Φ denotes the standard normal cumulative distribution function;
   $d1 = log(\frac{S_t * exp((r_t - \delta_t) * T/365)}{\kappa}) + \frac{(\sigma(\kappa)^2/2) * T/252)}{\sigma(\kappa) * \sqrt{T/252}}$;
   $d2 = d1 - \sigma(\kappa) * \sqrt{T/252}$

4. After that, they construct the error terms for each price options as following:

   - $error1 = 15/70000 * \kappa - 0.1842857$ (it is a vector)
   - $call_{error} = 2\sqrt{3} * error1(\mathcal{U}(0,1) - 0.5)$
   - $put_{error} = 2\sqrt{3} * rev(error1)(\mathcal{U}(0,1) - 0.5)$

   $u.v$ denotes product element by element;
   $\mathcal{U}$ uniform distribution ($n = 25$ draw)
   rev denotes the reverse of the vector (Example: $rev(1\ 2\ 3) = (3\ 2\ 1)$)

5. Finally the vector of put price and call price are:
   $call = call * (1 + call_{error})$
   and $put = put * (1 + put_{error})$;

6. They define s, stock price at maturity as a grid of 100 values equally spaced between 1000 and 1700

7. They develop Black-Scholes smile density in following steps : (Indeed, f the second derivative of put or call (see 29 or 30) with respect to a κ_i. We can demonstrate using $call_i = \int_0^\infty max(k_i - s, 0) f(s) ds$)

   (a) $f_1(s) = exp(-rT/365) * exp(-d2^2/2) * 1/\sqrt{2 * \pi}$
   (b) $vega = S * \sqrt{T/252} * exp(-\delta * T/365) * exp(-d1^2/2) * 1/\sqrt{2 * \pi}$
   (c) $dvega = vega * d1 * d2/smile$
   (d) $dvegas = vega * d1 * 1/(s * smile * \sqrt{T/252})$
   (e) $f_2(s) = f_1(s) + dvega * (b + 2 * a * s)^2$
   (f) $f_3(s) = f_2(s) + 2 * dvegak * (b + 2 * a * s)$
   (g) To finish, $f(s) = f_3(x) * exp(rT/365)$.

8. Compute $Sigma$ to get the transformation in equation (21) and (22) as following:
   $Sigma = 1/2 * (\sigma_{put} + \sigma_{call}) * \sqrt{1/365}$, where:

   - σ_{put} is implied volatility of a put price and strike price κ corresponding to $min(\kappa - S)$,
   - σ_{call} is implied volatility of a call price and strike price κ corresponding to $min(\kappa - S)$.

9. Then, they compute the transformation in equation (22).

10. They define for X in equation (21), a grid of $m = 1020$ points on $[-10, 10]$.

11. With $n = n_{call} + n_{put} = 25 + 25 = 50$, the truncation order is $J = 4$ by applying the formula in equation (27).

12. They compute equation (28) by using Riemann sum.

13. They solve the optimization problem of step2: α is chosen by 10-fold cross-validation.

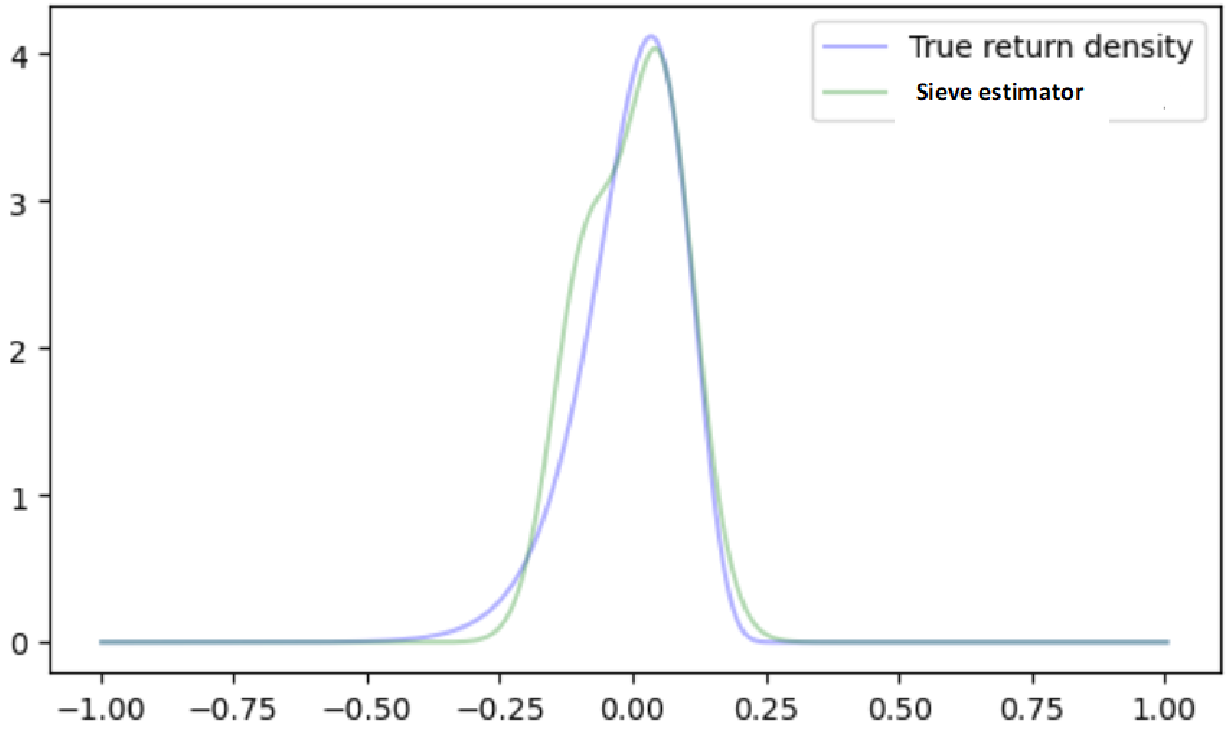

Figure 6: Data generated as in Lu and Qu (ncall=nput=25); f estimated following Lu and Qu. The Blue curve is true return density and green curve, estimated density

14. They compute $\hat{f}(X)$ from Hermite polynomials.

15. They run 5000 replications (Here, we run 200 replications).

16. They compute the density of return $\hat{f}^*(R) = 1/(Sigma * \sqrt{\tau} * \hat{f}(X))$ and determine the grid of R associated with the grid of X by applying the formula in equation (21).

17. With the grid of R, they get a grid of s which is $s = S * exp(R)$.

18. With this new grid of s, they compute the Black-Scholes smile density as in point 7.

19. To finish, they represent the graph of true return density and estimated return density (mean over the 5000 replications, median, quantile 0.25 and quantile 0.9725. Here, we only represent the mean over the 200 replications considered.)

The results show that the sieve estimator performs well in the exogenous case, as in their paper (see figure 6).

5.1.3 Simulation of Sieve estimator: endogenous case

Here, we generate data following the idea of Lu and Qu while introducing endogeneity. We then propose a modification of the Lu and Qu estimation method to account for endogeneity. In fact, data are generated to preserve the range of κ values as in the paper of Lu and Qu (2021), and to check IV conditions such as the correlation between W and κ, and the absence of correlation between W and ϵ.

- Generate $v \sim \mathcal{U}(-1, +1)$ (ncall=nput=200, so 200 observations).
- Generate $W \sim \mathcal{U}(1150, 1550)$ (200 observations).
- Compute $\kappa = W + 150 * v$ and $\varepsilon = 10v$.
- Define a grid of m discrete points for s between 1000 and 1700 (s is used to compute the density function $f(s)$);
- I add to y_{put} and y_{call} the same error term.
- I solve the problem by two ways:
  - Benchmark: I solve the problem as exogenous (previous situation presented).
  - I solve it as endogenous: for that, the problem becomes $E(\Psi_{W_i} y_i) = \int E(\Psi_{W_i} Z_i(s) f(s) ds$. Then, we construct the sieve estimator in endogenous case as follows:
    * We premultiply the equation (27) by the instrument $\psi_i(l)$:

$$\psi_i(l) y_i = \begin{cases} \int_{-\infty}^{\infty} \psi_i(l) S(e^{\sqrt{\tau}\sigma X} - e^{\sqrt{\tau}\sigma \tilde{Z}_i})^+ f(X) dX + \psi_i(l)\varepsilon_i, & for\ i = 1, 2, ..., n_c \\ \int_{-\infty}^{\infty} \psi_i(l) S(e^{\sqrt{\tau}\sigma \tilde{Z}_i} - e^{\sqrt{\tau}\sigma X})^+ f(X) dX + \psi_i(l)\varepsilon_i, & for\ i = n_c + 1, ..., n \end{cases} \tag{31}$$

    * By following the same computation technique (see equation (28)), the sieve estimator in endogenous case is the solution of:
      $\min_{\beta \in \mathcal{H}_J} \sum_{l=1}^{L} \sum_{i=1}^{n} (\Psi_i(l) y_i - \Psi_i(l) X_i^{'} \beta)^2 + \beta^{'} Q_\alpha \beta$
      where $\Psi_i(w, l) = 1/(1 + exp(-wl))$ or $\Psi_i(w, l) = max(0, w - l)$
- we run 200 replications.

The following graph (Figure 7) compares two Sieve estimators of the RND of return (the red curve is the density estimated neglecting endogeneity and the green curve is the density estimated considering endogeneity) to the true return density (blue curve). It shows that both estimators perform well overall, with a small gain when we use instruments to solve the problem, considering the RMSPE and RISE criteria (see table 2).

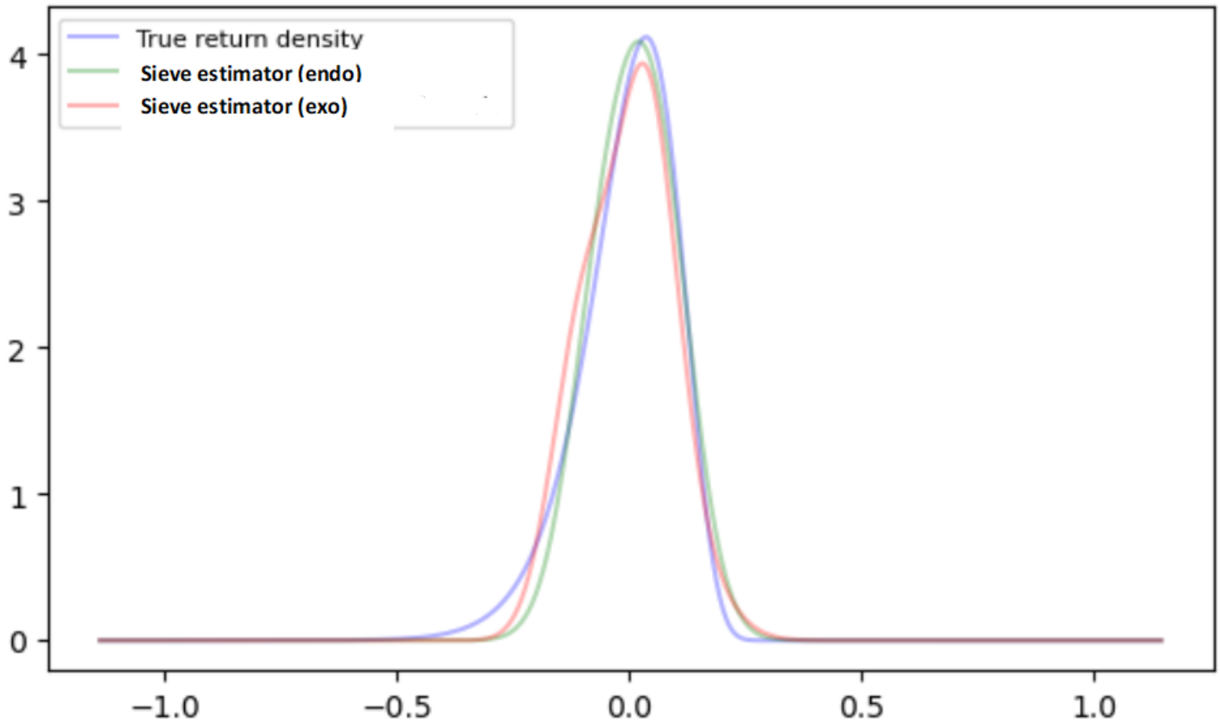

Figure 7: Data generated similary to Lu and Qu, plus endogenous; f estimated following Lu and Qu; ncall=nput=200.
Blue curve is true density, red curve is density estimated neglecting endogenous and green curve is density estimated considering endogeneity

	$RMSPE$	$RISE$
$\hat{f}_{sieve_{IV}}$	$4.213e^{-5}$	$1.954e^{-5}$
$\hat{f}_{sieve}$	$5.651e^{-5}$	$3.232e^{-5}$

Table 2: Comparison of the Sieve estimation results

In what follows, we will focus on how Lu and Qu (2021) generate data. We will make a few modifications in order to study the effect of the generated true density.

5.2 Sieve and Tikhonov estimators with Black-scholes smile distribution

In this subsection, we present the Sieve and Tikhonov estimators in two cases: exogenous and endogenous, without the use of instruments. To achieve this, we generate data for put options (n=500 observations) using the technique presented in the previous section (from Lu and Qu (2021)). We compute three estimators and display the results for the underlying price directly, rather than its log-return.

- First estimator: Sieve estimator. Data are standardized and we use Hermite polynomials to approximate f .
- Second estimator: Tikhonov estimator, we estimate directly f as presented in Section 2 (the data have not been processed here).

- Last estimator: Tikhonov estimator (Standard). Here we have just standardized data and estimated directly f using the technique proposed in Section 2.

Case1: Exogenous case

Figure 8 shows the three estimators and the true density in the exogenous case. The data was generated as described in Lu and Qu (2021). The results indicate good performance overall for all three estimators. Specifically, the Tikhonov estimator performs the best with a RISE of $8.129e^{-6}$ compared to the Tikhonov estimator (Standard) and Sieve estimator with RISE of $1.032e^{-5}$ and $3.532e^{-5}$, respectively. Therefore, it can be inferred that the Tikhonov estimator yields strong results when implemented on exogenously created data, as seen in Carrasco and Tsafack (2023) and Lu and Qu (2021). Additionally, it surpasses the Sieve estimator.

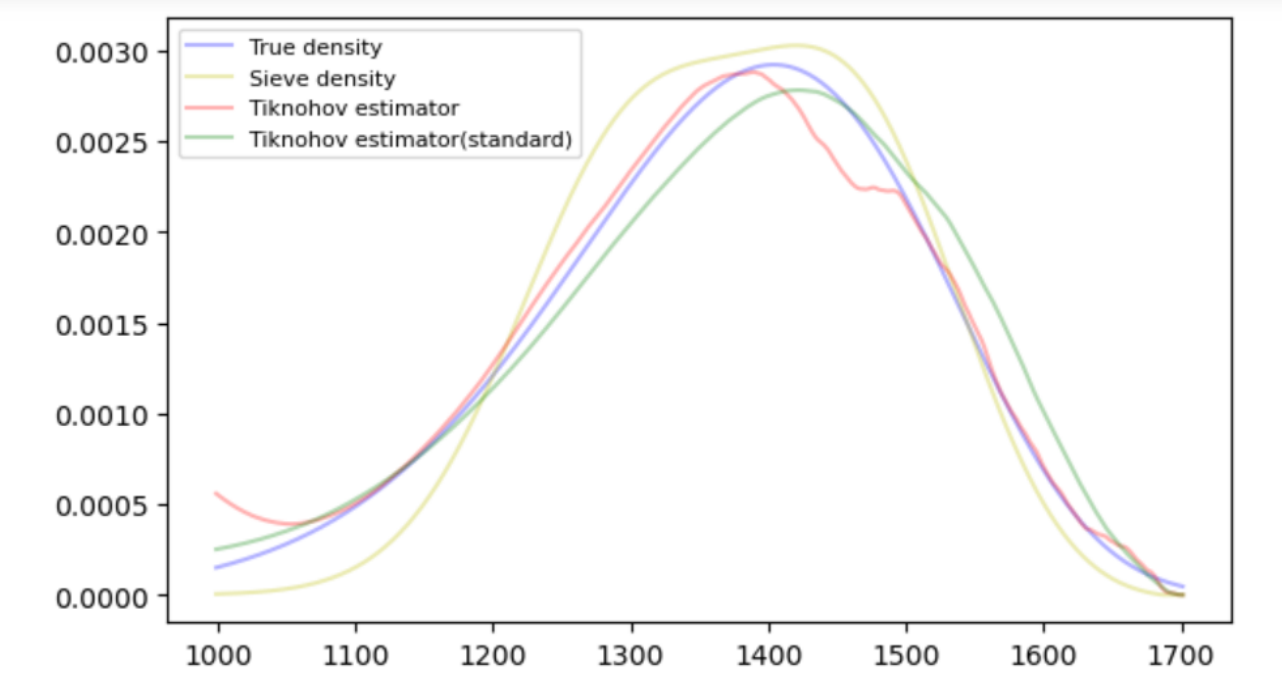

Figure 8: Data generated similarly to Lu and Qu.
Blue curve is true density, red curve is Tikhonov estimator, green curve is Tikhonov estimator(standard) and yellow curse is Sieve estimator.

	$RMSPE$	$RISE$
$\hat{f}_T$	$2.231e^{-5}$	$8.129e^{-6}$
$\hat{f}_{TS}$	$3.523e^{-5}$	$1.032e^{-5}$
$\hat{f}_{sieve}$	$5.514e^{-5}$	$3.532e^{-5}$

Table 3: Comparison of the estimation results, exogenous case
$\hat{f}_T$: Tikhonov estimator; $\hat{f}_{TS}$: Tikhonov estimator(standard); $\hat{f}_{sieve}$: Sieve estimator

Case2: Endogenous case without the use of instruments

Data are generated to take account of endogeneity and the estimators are computed by neglecting this endogeneity. The figure 9 illustrates that all of the estimators perform well. Upon comparison, it is observed that the (standard) Tikhonov estimator performs better than the Tikhonov estimator, but the difference between them is minimal. Additionally, the Tikhonov estimator outperforms the Sieve estimator. The Table 4 depicts the values of the different comparison criteria. Indeed, the RISE value is $2.409e^{-}5$ for the Tikhonov estimator, $2.088e^{-}5$ for the Tikhonov estimator (Standard), and $5.002e^{-}5$ for the Sieve estimator. The study indicates that the Tikhonov estimator achieves better performance with the Black-Scholes smile distribution compared to the Tikhonov estimator using the lognormal mixture distribution (see Figure 5). This implies that the true density may significantly affect the estimator's accuracy.

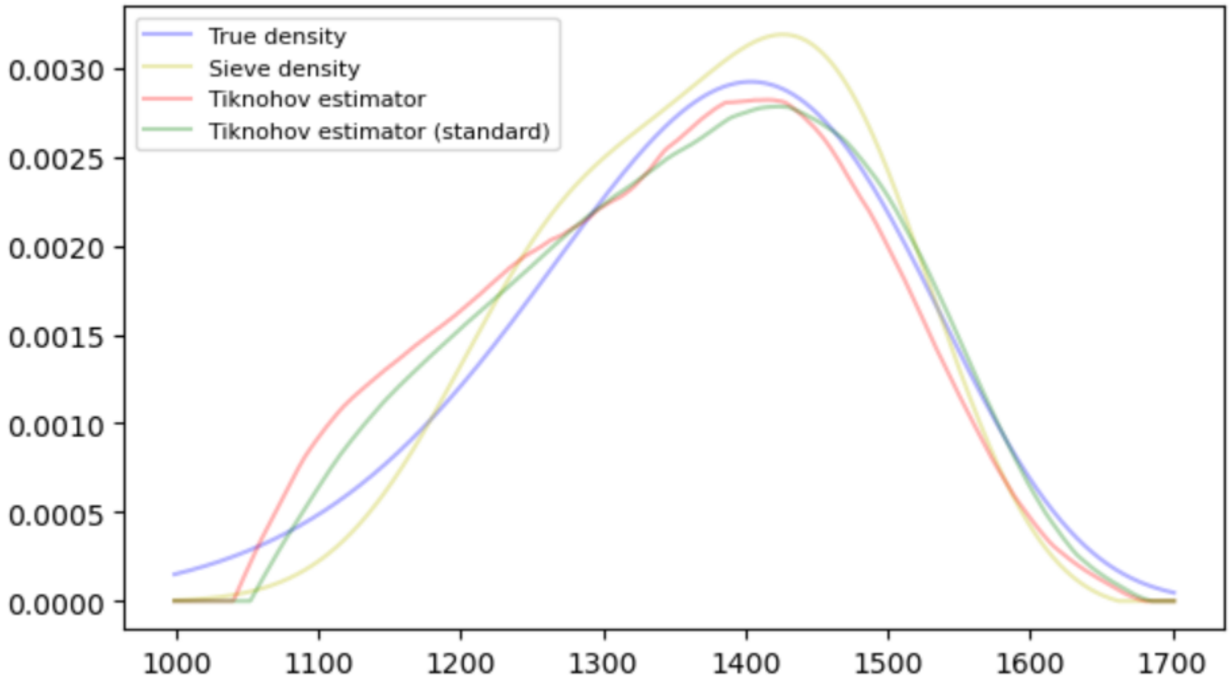

Figure 9: Data generated similary to Lu and Qu, plus endogeneity, solve as exogenous. Blue curve is true density, red curve is Tikhonov estimator, green curve is Tikhonov estimator(standard) and yellow curse is Sieve estimator.

	$RMSPE$	$RISE$
$\hat{f}_T$	$4.739e^{-5}$	$2.409e^{-5}$
$\hat{f}_{TS}$	$4.061e^{-5}$	$2.088e^{-5}$
$\hat{f}_{sieve}$	$6.121e^{-5}$	$5.002e^{-5}$

Table 4: Comparison of the estimation results, endogenous case and solve as exogenous $\hat{f}_T$: Tikhonov estimator; $\hat{f}_{TS}$: Tikhonov estimator(standard); $\hat{f}_{sieve}$: Sieve estimator

We now present our proposed Tikhonov estimator with instruments, utilizing the Black-Scholes smile distribution.

5.3 Tikhonov estimator with Black-scholes smile distribution

In this subsection, we want to investigate the impact of the generated true density on our proposed Tikhonov estimator using simulated instruments (see 5.1.3 for simulation procedure). Figure 10 displays a Tikhonov estimator with instrumental variables (IV) and a Tikhonov estimator without IV for a Black-Scholes smile distribution. Based on comparison criteria of RMSPE and RISE, the IV estimator outperforms the estimator without IV, which neglects endogeneity (refer to table 5). The Figure 10 also demonstrates that our proposed estimation technique performs better when using the Black-Scholes distribution, as considered by Lu and Qu (2021), than the log-normal mixture distribution, as considered by Carrasco and Tsafack (2023). For further details, refer to Figure 5, which shows the true density as a log-normal mixture distribution. Consequently, our suspicions were confirmed that the estimator's performance may be influenced by the distribution of the true density.

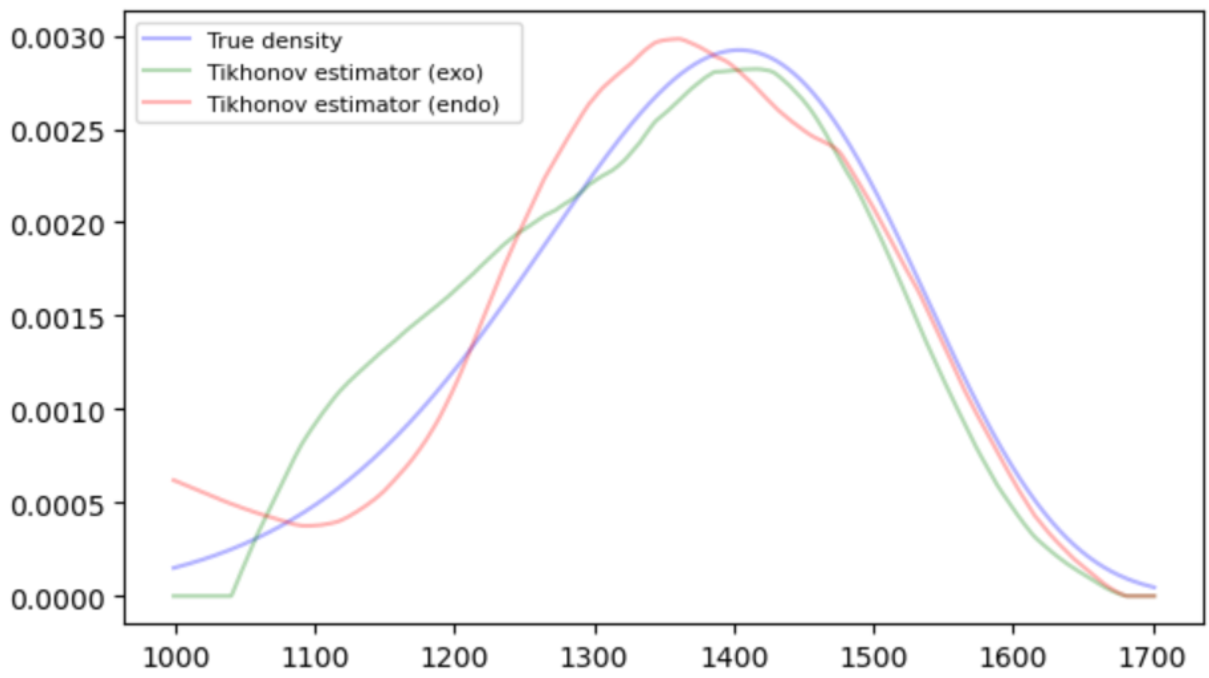

Figure 10: Data generated similarly to Lu and Qu, plus endogeneity.
Blue curve is true density, green curve and red curve are density estimated without and with considering endogeneity respectively.

	$RMSPE$	$RISE$
$\hat{f}_{IV}$	$2.976e^{-5}$	$1.611e^{-5}$
$\hat{f}$	$4.739e^{-5}$	$2.409e^{5}$

Table 5: Comparison of the Tikhonov estimation results, with IV and without IV.

6 Application

This section focuses on a real data example. We consider the data employed by Carrasco and Tsafack (2023). The dataset utilizes the $S\&P$ 500 index (SPX) to derive the underlying and the strikes. Specifically, the database archives the optimal bid and ask prices for all SP 500 European options that are collected on June 3, 2019, with a maturity period of 25 days. The price of the index at the end of the same day is also collected: it was evaluated at 2744.45 dollars. The expiration price ranges from 1500 to 3999.50 dollars. This study solely examines put options, while Carrasco and Tsafack (2023) investigate both puts and calls. The sample size is $n = 108$. We aim to utilize instrumental variables, comparing outcomes in the exogenous case to those obtained in the endogenous case. Unfortunately, we could not build the instruments as outlined in section 2.4. Consequently, we present the findings in the exogenous scenario utilizing two estimators: the Tikhonov and Sieve estimators.

The root mean square prediction error (RMSPE) is used as the criterion to select the optimal hyperparameter and estimate RND, following 10-fold cross-validation. The results of Tikhonov and Sieve estimators are presented in figure 11, with Tikhonov's RMSPE at 0.100 and Sieve's at 2.061. Thus, our findings suggest that, in the exogenous case, the Tikhonov estimator outperforms the Sieve estimator, consistent with the simulation section. The difference in the RMSPE may be attributed to the greater flexibility of the Tikhonov estimator compared to the Sieve estimator. The Sieve estimator is less flexible due to its use of Hermite polynomials in estimation, resulting in a reduction of parameters considered.

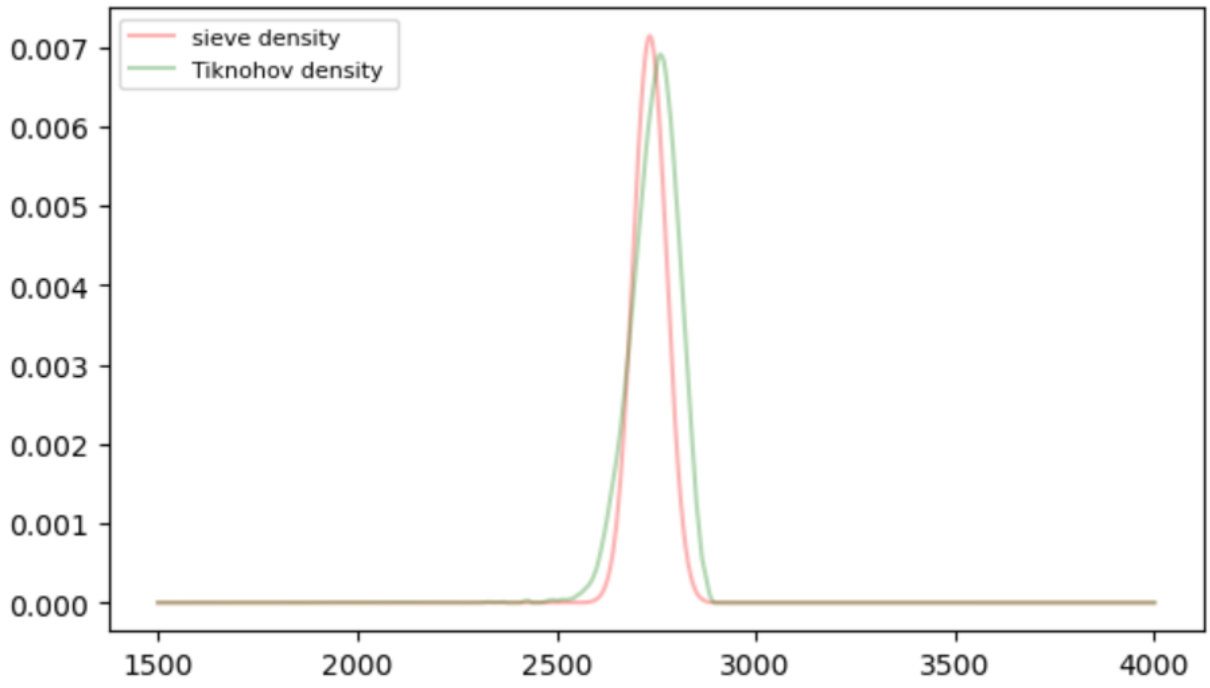

Figure 11: Risk neutral density with real data.
Green curve is Tikhonov estimator and red curve is Sieve estimator.

7 Conclusion

We propose in this paper, a non-parametric estimation for the Risk Neutral Density (RND) in European option context by using IV method. The non-parametric model examined is the functional linear model. One of the main issue of this model is the high dimensionality problem as the inverse of the covariance operator of the predictor variable is not continuous. This problem leads to an unstable estimated function. To overcome this issue, we propose to use the Tikhonov regularization technique. We control for the positivity and the integration to one of the density by applying a density correction. We also address the discretization of integrals by proposing a way to calculate our estimator. The simulation results effectively show that in the presence of endogeneity, the estimator without considering this endogeneity underperforms. And, when, we use the IV, the result is better (in term of RISE or RMSPE). We also note that the overall performance can depend on the true distribution considered. Indeed, in this work, we consider two types of true density: the lognormal mixture distribution and the Black-Scholes smile distribution. The results are better with the last density than with the first one. It might be interesting to investigate what the true risk-neutral density would be, as an extension of this work. At the end, we present an application case with real data on $S\&P$ 500 options in the exogenous case because we were unable to find instrumental variables in the European options case as described in subsection 2.4. The results show that the Tikhonov estimator outperforms the Sieve estimator. Consequently, the Tikhonov estimator method proposed in this work, in particular the computational technique, can be seen as a promising alternative to existing methods.

References

Ait-Sahalia, Y. and Duarte, J. (2003). Nonparametric option pricing under shape restrictions. *Journal of Econometrics*, 116(1-2):9–47.

Ait-Sahalia, Y. and Lo, A. (2000). Nonparametric risk management and implied risk aversion. *Journal of econometrics*, 94(1-2):9–51.

Amengual, D., Carrasco, M., and Sentana, E. (2020). Testing distributional assumptions using a continuum of moments. *Journal of Econometrics*, 218:655–689.

Andrews, I., Stock, J. H., and Sun, L. (2019). Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics*, 11(1):727–753.

Babii, A. (2021). High-dimensional mixed-frequency iv regression. *Journal of Business & Economic Statistics*, 0(0):1–14.

Bahra, B. (1997). Implied risk-neutral probability density functions from option prices: theory and application.

Barletta, A., Santucci de Magistris, P., and Violante, F. (2019). A non-structural investigation of vix risk neutral density. *Journal of Banking Finance*, 99:1–20.

Benatia, D., Carrasco, M., and Florens, J.-P. (2017). Functional linear regression with functional response. *Journal of Econometrics*, 201(2):269–291.

Bondarenko, O. (2003). Estimation of risk-neutral densities using positive convolution approximation. *Journal of Econometrics*, 116(1-2):85–112.

Bosq, D. (2000). Linear processes in function spaces: Theory and applications. *volume 149 of lecture notes in statistics. Springer-Verlag New York Inc.*

Cai, T. T. and Hall, P. (2006). Prediction in functional linear regression. *The Annals of Statistics*, 34(5):2159 – 2179.

Carrasco, M., Florens, J.-P., and Renault, E. (2007). Chapter 77: Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. *Handbook of Econometrics*, 6:5633–5751.

Carrasco, M. and Tsafack, I. (2023). Risk neutral density estimation with a functional linear model. *in Essays in honor of Joon Y. Park, Advances in Econometrics*, Vol.45B:133–157.

Chang, Y., Kim, C., and Park, J. (2016). Nonstationarity in time series of state densities. *Journal of Econometrics*, 192(1):152–167.

Cont, R. (2001). Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative Finance*, 1:223–236.

Corrado, C. J. and Su, T. (1996). Skewness and kurtosis in s&p 500 index returns implied by option prices. *Journal of Financial research*, 9(2):175–192.

Cox, J. and Ross, S. (1976). The valuation of options for alternative stochastic processes. *Journal of financial Economics*, 3(1):145–166.

Cui, Z. and Xu, Y. (2022). A new representation of the riskneutral distribution and its applications. *Quantitative Finance*, 22(5):817–834.

Delaigle, A. and Hall, P. (2012). Methodology and theory for partial least squares applied to functional data. *The Annals of Statistics*, 40(1):322–352.

Dewatripont, M., Hansen, L. P., and Turnovsky, S. J. (1997). *Advances in Economics and Econometrics : Theory and Applications : Seventh World Congress*, volume 1 of *Econometric Society Monographs*. Cambridge University Press.

Engl, H. W., Hanke, M., and Newbauer, A. (1999). Regularization of inverse problems (mathematics and its applications, 375). *Springer; 1996th edition (July 31, 1996).*

Fabozzi, F. J., Tunaru, R., and Albota, G. (2009). Estimating risk-neutral density with parametric models in interest rate markets. *Quantitative Finance*, 9(1):55–70.

Feng, P. and Dang, C. (2016). Shape constrained risk-neutral density estimation by support vector regression. *Information Sciences*, 333:1–9.

Ferraty, F. and Vieu, P. (2000). Nonparametric functional data analysis: theory and practice. *Springer Science Business Media.*

Garcia, R. and Gençay, R. (2000). Pricing and hedging derivative securities with neural networks and a homogeneity hint. *Journal of Econometrics*, 94(1-2):93–115.

Grith, M., Hrdle, W., and Schienle, M. (2012). Nonparametric estimation of risk-neutral densities. *Handbook of Computational Finance*, pages 277–304.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054.

Hill, A. D., Johnson, S. G., Greco, L. M., and Greco, L. M. (2021). Endogeneity: A review and agenda for the methodology-practice divide affecting micro and macro research. *Journal of Management*, 47(1):105– 143.

Horowitz, J. L. (2011). Applied nonparametric instrumentalvariables estimation. *Econometrica*, 79(2):347–394.

Jarrow, R. and Rudd, A. (1982). Approximate option valuation for arbitrary stochastic processes. *Journal of financial Economics*, 10(3):347–369.

Jiang, L., Zhou, S., Li, K., Wang, F., and Yang, J. (2021). A new nonparametric estimate of the risk-neutral density with applications to variance swaps. *Frontiers in Applied Mathematics and Statistics*, 6.

Kargin, V. and Onatski, A. (2008). Curve forecasting by functional autoregression. *Journal of Multivariate Analysis*, 99(10):2508–2526.

Lu, J. and Qu, Z. (2021). Sieve estimation of option-implied state price density. *Journal of Econometrics*, 224(1):88–112. Annals Issue: PI Day.

Ludwig, M. (2015). Robust estimation of shape-constrained state price density surfaces. *The Journal of Derivatives*, 22(3):56–72.

Monteiro, A. M. and Santos, A. A. F. (2022). Option prices for risk-neutral density estimation using nonparametric methods through big data and large-scale problems. *The journal of futures markets*, 42:152–171.

Monteiro, A. M., Tütüncü, R., and Vicente, L. (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.

Roberts, M. R. and Whited, T. M. ((2012). Endogeneity in empirical corporate finance. *Simon School Working Paper No. FR 11-29*.

Rompolis, L. S. and Tzavalis, E. (2008). Recovering risk neutral densities from option prices: A new approach. *Journal of Financial and Quantitative Analysis*, 43(04):1037–1053.

Schittenkopf, C. and Dorffner, G. (2001). Risk-neutral density extraction from option prices: improved pricing with mixture densitynetwork s. *in IEEE Transactions on Neural Networks*, 12(4):716–725.

Staum, J. (2007). Chapter 12 incomplete markets. In Birge, J. R. and Linetsky, V., editors, *Financial Engineering*, volume 15 of *Handbooks in Operations Research and Management Science*, pages 511–563. Elsevier.

Tsafack, I. (2020). Intraday stock market forecasting via functional time series. *Working paper*.

Wang, C.-J. (2010). *The Instrument Variable Approach to Correct for Endogeneity in Finance*, pages 1357–1369. Springer US, Boston, MA.

Yijun, D., Chen, W., and Yibing, D. (2012). Inversion of option prices for implied risk-neutral probability density functions: general theory and its applications to the natural gas market. *Quantitative Finance*, 12(12):1877–1891.