A SEMIPARAMETRIC ESTIMATION OF LIQUIDITY EFFECTS ON OPTION PRICING

Eva Ferreira  
(Universidad del País Vasco)  

Mónica Gago  
(Universidad del País Vasco)  

and  

Gonzalo Rubio  
(Universidad del País Vasco)  

First draft: July 1999  
This draft: March 2000  

Keywords: Multivariate Kernel and SNN Regressions, Volatility Smile, Option Pricing  
JEL classification: G12, G13  

Corresponding author: Gonzalo Rubio, Dpto. Fundamentos del Análisis Económico, Facultad de Ciencias Económicas, Universidad del País Vasco, Avda. L. Aguirre 83, 48015 Bilbao, Spain; e-mail: jepruig@bs.ehu.es  

Eva Ferreira and Gonzalo Rubio acknowledge the financial support provided by Dirección Interministerial Científica y Técnica (DGICYT) grants PB98-0149 and PB97-0621 respectively. The three authors acknowledge the financial support provided by Universidad del País Vasco (UPV/EHU) grant UPV 038.321-HA129/99, and the BSI Gamma Foundation. We appreciate the helpful comments of Ángel León and Javier Fernández Navas, seminar participants at the Bank of Spain, and the computational assistance of Gregorio Serna. We thank Juan Ayuso and MEFF for providing the data used in this article. The contents of this paper are the sole responsibility of the authors.
Abstract

This paper proposes a semiparametric option pricing model with liquidity, as proxied by the relative bid-ask spread. The nonparametric volatility function with liquidity costs as an explanatory variable is estimated using the Symmetrized Nearest Neighbors (SNN) estimator rather than the traditional kernel estimator. It is argued that the SNN estimator is particularly suitable for option data. Moreover, special care is taken in obtaining the smoothing parameter. A statistical design to test competing option pricing models which takes into account the lack of independence between them is also presented. The in-sample performance of the model turns out to be statistically favorable relative to a competing model without liquidity. However, the out-of-sample performance is quite disappointing independently of the option pricing model employed in the estimation.
1. Introduction

It is well understood that the central point for the empirical testing of option pricing models is whether the actual distribution of the underlying asset implied by the option market data is consistent with the distribution assumed by the theoretical option pricing model.

Given the Black-Scholes (BS) (1973) assumptions, all option prices on the same underlying security with the same expiration date but with different exercise prices should have the same implied volatility. However, the well known volatility smile pattern suggests that the BS formula tends to misprice deep in-the-money and deep out-of-the-money options. There have been various attempts to deal with this apparent failure of the BS valuation model. In principle, as explained by Das and Sundaram (1999) and others, the existence of the smile may be attributed to the well known presence of excess kurtosis in the conditional return distributions of the underlying assets. It is clear that excess kurtosis makes extreme observations more likely than in the BS case. This increases the value of out-of-the-money and in-the-money options relative to at-the-money options, creating the smile. However, at least in equity markets, the pattern shown by data tends to contain an asymmetry in the shape of the smile. This may be due to the presence of skewness in the distribution which has the effect of accentuating just one side of the smile.

Given this evidence, extensions to the BS model that exhibit excess kurtosis and skewness have been proposed in recent years along two lines of research: Jump-diffusion models with a Poisson-driven jump process, and the stochastic volatility framework are the two key developments in theoretical option pricing literature.

---

1 After the October 1987 crash, the implied volatility computed from options on stock indexes in the US market inferred from the BS formula appears to be different across exercise prices. This is the so-called “volatility smile”. In fact, as pointed out by Rubinstein (1994), Aït-Sahalia and Lo (1998) and Dumas, Fleming and Whaley (1998), implied volatilities of the S&P 500 options decrease monotonically as the exercise price becomes higher relative to the current level of the underlying asset. On the other hand, Taylor and Xu (1994) show that currency options tend to present a much more pronounced smile. Similar patterns of implied volatilities across exercise prices are found by Peña, Rubio and Serna (1999) in the Spanish options market. Moreover, Bakshi, Cao and Chen (1997), Chernov and Ghysels (1999) and Pan(1999) report smile shapes for the (implied) instantaneous volatility under stochastic volatility and jump-diffusion option pricing models. León and Rubio (2000) theoretically study the behavior of the implied volatility function (smile) when the true distribution of the underlying asset is consistent with the stochastic volatility model proposed by Heston (1993).
Unfortunately, however, the empirical evidence regarding these new models is quite disappointing. Bates (1996), and Bakshi, Cao and Chen (1997) reject the jump-diffusion option pricing model on currency options and equity options respectively. The stochastic volatility model proposed by Heston (1993) is rejected by Bakshi, Cao and Chen (1997), Chernov and Ghysels (1999), and Pan (1999) for options written on the S&P 500 index.2 Fiorentini, León and Rubio (1999) reject the same model for equity options on the Spanish IBEX-35 index.

These latter authors argue that the ultimate reasons behind the performance failure of Heston’s model are closely related to the time-varying skewness and kurtosis found in the data. In particular, they suggest that the assumption of a constant correlation coefficient between returns and stochastic volatility should be relaxed if we really want to have a richer model. Unfortunately, the complexities needed to price options seem to increase without bounds. It may be the case that simple nonparametric (semiparametric) methodologies are able to incorporate the missing (realistic) factors in our option pricing models.

Along these lines, it should be pointed out that all previous models have been developed in a competitive, frictionless framework. It may the case that liquidity costs, as represented by the percentage bid-ask spread, account for some of the differences observed between market prices and theoretical prices. In fact, Lonstaff (1995) shows that the relative bid-ask spread explains a statistically significant percentage of the BS pricing errors. Interestingly, Peña, Rubio and Serna (1999) show that liquidity costs significantly cause the magnitude of the smile in equity options written on the Spanish IBEX-35 index. After the October 1987 crash, portfolio insurers began buying index options to implement their insurance strategies. In particular, institutional investors are interested in buying out-of-the-money put options as an insurance mechanism. This institutional buying pressure on out-of-the-money put options will increase put prices to a level where market makers are induced to accept the bet that the index level will not fall below the exercise price before the option’s expiration. Therefore, independently of the distributional characteristics of the underlying

---

2 Pan (1999) shows that jump-risk premia play a key role in reconciling the dynamics implied by S&P 500
asset, liquidity costs, as proxied by the relative bid-ask spread, may induce patterns in implied volatilities.

A potentially relevant area of research might be related to endogenously incorporating liquidity costs in option pricing models with either stochastic volatility, stochastic jumps or both. A much simpler but, at the same time, more effective approach would be based on the estimation of the implied volatility function with semiparametric methodologies, where the Black-Scholes implied volatility is replaced by a nonparametric function which depend upon a vector of explanatory variables. This is the multivariate kernel regression approach which has been recently followed by Aït-Sahalia and Lo (1998). However, they ignore the potential effects of market frictions on the nonparametric volatility function. The objective of our paper is to fill up this gap by incorporating the percentage bid-ask spread as an additional explanatory variable on the nonparametric volatility function. Thus, we construct the corresponding call pricing function under liquidity costs, and compare its performance relative to more traditional option pricing models. Hence, our nonparametric volatility function depends on moneyness, time-to-expiration and the percentage or relative bid-ask spread. In this sense, we are dealing with a multivariate nonparametric estimation.

Moreover, we also estimate the state price density (SPD) or the so called (under non-arbitrage models) risk-neutral density, with the added potential effects of market frictions as proxied by the bid-ask spread. This is a key contribution of this paper to literature on option pricing. At the same time, from a statistical point of view, our work improves the technique used by Aït-Sahalia and Lo (1998) in, at least, two important ways: (i) The use of a multivariate kernel based on a global smoothing parameter may lead to estimation problems when obtaining the volatility nonparametric function in moneyness intervals for which the amount of data is relatively small. These intervals coincide with extreme out-of-the-money and in-the-money options and, of course, these are precisely the sections of the smile in which we are particularly interested. Given these arguments, we are planning to

---

3 Using a linear and a quadratic parametric approach, Peña, Rubio and Serna (2000) solve numerically a forward partial differential equation with liquidity costs also proxied by the relative bid-ask spread. Independently of the parametric specification, they find that these models seem to perform poorly relative to Black-Scholes.
use the so called Symmetrized Nearest Neighbors (SNN) estimation instead of the more traditional kernel approach. (ii) Despite the fact that they have a three-dimensional kernel estimator, Aït-Sahalia and Lo employ a univariate smoothing parameter criterion. Moreover, they simplify the problem by eliminating the bias term in choosing the necessary bandwidth to estimate their nonparametric volatility function. We employ several criteria in order to calculate the bandwidth used in our estimations. In particular, we employ a plug-in criterion and multivariate approaches under two alternative specifications. Robusteness relative to the bandwidth parameter is an important issue in nonparametric statistics.

A statistical design is proposed to compare competing option pricing models which are not independent. In-sample and out-of-sample option pricing empirical results are reported where nonparametric multivariate volatility function and the corresponding option prices are estimated quarterly. Independently of the bandwidth criteria used, the in-sample results show an important improvement whenever we incorporate liquidity effects on the estimation. This result may have serious implications for option pricing research. However, the out-of-sample results are generally quite poor. All pricing models have a significant degree of mispricing. This should come as no surprise taking into account that all models are estimated using quarterly data. This introduces a demanding requirement of stability to our estimated nonparametric functions, even though it should be noticed that, using a randomization test, we are not able to reject the stability of risk-neutral densities between quarters.

The paper is organized as follows. Section 2 introduces our nonparametric estimation. In Section 3 we present the data available for our research. Section 4 discusses the nonparametric estimation of both the volatility smile and the risk-neutral density. The main empirical results are reported in Section 5. We discuss both the in-sample results and the out-of-sample performance of alternative pricing models. We conclude in Section 6 with a summary of results and a brief discussion of possible future work.
2. Nonparametric Estimation of Risk-Neutral Densities

2.1 Non-arbitrage Pricing

It is well known that under risk-neutrality the price of any financial asset can be expressed as the expected present value of its future payoffs, where the present value is obtained relative to the riskless rate and the expectation is taken relative to the risk-neutral density function of the payoffs. When the asset is a European call option with expiration at $T$ and exercise price $X$, its price is given by:

$$c = e^{-r(T-t)} \int_0^\infty \max[S_T - X] f_t^*(S_T) dS_T$$

(1)

where $r$ is the constant riskless rate between $t$ and $T$, and $f_t^*(S_T)$ is the date-$t$ risk-neutral density of the stock price at future date $T$.

Finally,

$$f_t^*(S_T) = e^{r(T-t)} \frac{\partial^2 c}{\partial X^2} \bigg|_{X=S_T}$$

(2)

so that, the risk-neutral density is proportional to the second derivative of the option price function with respect to the exercise price.

2.2 Nonparametric Estimation

As discussed in the introduction, the idea of the paper is to estimate the risk-neutral density nonparametrically, and to be able to price options. Our procedure is based on the following sequence of estimations: We employ option market prices to estimate a nonparametric volatility function which depends upon the degree of moneyness, time-to-expiration and liquidity, proxied by the relative bid-ask spread. Then, given this function in which volatility is allowed to vary with moneyness, time-to-expiration and the bid-ask spread, the Black-Scholes formula can be used to obtain, option prices semiparametrically\(^4\). In the last

\(^4\) This semiparametric estimation of option prices considerably reduces the dimensionality of the problem. As pointed out by Aït-Sahalia and Lo (1998), the sample size required to achieve the same degree of accuracy as in the full nonparametric estimation may be much smaller.
step, we differentiate this option estimator twice with respect to the exercise price to obtain (2), given the appropriate interest rate. The issue, of course, is how to estimate the multivariate volatility function nonparametrically.

It is important to realize that there is no obvious way to model the influence of moneyness, time-to-expiration and liquidity on the volatility function. It is precisely in this sense that the nonparametric framework provides a very flexible approach.

Let us consider a multivariate kernel estimator of the volatility function, with possibly different smoothing parameters for the covariates:

\[
\tilde{\sigma}(\xi, \text{SP}, \tau) = \frac{1}{nh_{\xi} h_{\text{SP}} h_{\tau}} \sum_{i=1}^{n} K \left( \frac{\xi - \xi_{i}}{h_{\xi}} \right) K \left( \frac{\text{SP} - \text{SP}_{i}}{h_{\text{SP}}} \right) K \left( \frac{\tau - \tau_{i}}{h_{\tau}} \right) \sigma_{i}
\]

where \( \xi \equiv X/F \) is the degree of moneyness, where \( X \) is the exercise price and \( F \) is the futures price (underlying asset)\(^5\), \( \text{SP} \) is the relative bid-ask spread, \( \tau \) is the time-to-expiration, \( \sigma_{i} \) is the volatility implied by the option price \( c_{i} \), \( h_{j} \) is the bandwidth or smoothing parameter for each covariate \( j = \xi, \text{SP}, \tau \), and \( \tilde{\sigma}(\xi, \text{SP}, \tau) \) is the three-dimensional nonparametric volatility function to be estimated.

It is important to point out that a kernel estimator based on a global smoothing parameter may lead to poor estimation results basically in those zones where we have a relatively small amount of data. Translating these effects to our case, it suggests that we may obtain poor estimations for the volatility function for (deep) out-of-the-money and (deep) in-the-money options. Of course, from a financial point of view, these are precisely the options (and the sections of the volatility smile) in which we are particularly interested.
Given this fact, in this paper we employ the Symmetrized Nearest Neighbors (SNN) estimation as an alternative to the classical kernel estimator. This kind of estimator was proposed by Yang (1981), and studied in detail by Stute (1984). The idea behind them is very simple. When estimating in one point we calculate the weight for the rest of observations looking at the distance between the values of the empirical distribution at each point rather than the distance between the points themselves. Hence, the estimator is defined as:

\[
\hat{\sigma}(\xi, \text{SP}, \tau) = \frac{1}{n\xi h_{\text{SP}}h_{\tau}} \sum_{i=1}^{n} \left( \frac{\hat{F}_n(\xi) - F_n(\xi)}{h_{\xi}} \right) \left( \frac{F_n(\text{SP}) - F_n(\text{SP})}{h_{\text{SP}}} \right) \left( \frac{F_n(\tau) - F_n(\tau)}{h_{\tau}} \right) \sigma_i.
\]

where \(F_n(.)\) denotes the empirical distribution of the corresponding variable, \(\xi\), SP or \(\tau\).

Roughly speaking, the empirical distribution changes the random design to a uniform design with the knots uniformly spaced between zero and one. In practice, using SNN estimators is basically the same as employing kernel estimators for \(F_n(X_i)\) instead of \(X_i\). A detailed discussion on the differences between these estimators is contained in Appendix A, where we present and compare the minimum asymptotic mean square error (MSE) for both kernels and SNN. Moreover, to provide some intuition related to our particular case, we discuss an example that illustrates the behavior of our dataset.

As shown in Appendix A, both estimators have the same MSE under a uniform design. On the other hand, if we assume the bias to be negligible with respect to variance, it is easy to show that using the SNN estimator with bandwidth \(h\) is equivalent to employ a kernel estimator with variable bandwidth equal to \(hf(x)\). Also, the discussion provided in Appendix A allows us to argue that, in the tails, a smaller MSE is obtained for the SNN.

Note that the underlying asset in the Spanish market for which we have data is the futures price on the stock exchange index.
estimator. It should be recalled that we are particularly concerned with the tails of the
distribution given, of course, that extreme degrees of moneyness are a key issue in terms of
both the volatility smile and pricing.

Once we have estimated the volatility function given by (4), we have to estimate the call-
pricing function. This function is evaluated as,

\[ \hat{c}(\xi, r, \tau, \hat{\sigma}(\xi, SP, \tau)) \]

where the function \( \hat{c}(.) \) is the same as in the Black-Scholes expression with the nonparametrically estimated
volatility. This is to say,

\[ c(\xi, \tau, r, \sigma(\xi, SP, \tau)) = c_{BS}(\xi, \tau, r, \sigma(\xi, SP, \tau)) \] \hspace{1cm} (5)

The risk-neutral density estimator follows by taking the appropriate partial second
derivative of \( \hat{c}(.) \) with respect to the exercise price. The detailed derivation of this second
derivative is reported in Appendix B:

\[ \hat{f}_T^r(S_T) = e^{\gamma T} \left[ \frac{\partial^2 c(\xi, \tau, r, \sigma(\xi, SP, \tau))}{\partial X^2} \right]_{X=S_T} \] \hspace{1cm} (6)

In practice, the last and probably most important problem faced by any researcher is the
selection of the smoothing parameters, \((h_\xi, h_{SP}, h_r)\). It is interesting to point out that there
is a tremendous amount of literature developed for the univariate case. See Härdle (1990)
for a general presentation of this literature. Unfortunately, bandwidth selection becomes
much more complicated in the multivariate context.

Generally speaking, there are two groups of methods for selecting bandwidth: plug-in
methods and methods based on the minimization of some penalized least square error
measure. Deciding the particular selection method to be employed is no trivial task. Even in
the simpler case, in which we have a fixed design and a univariate estimator, different
asymptotically optimal methods may lead to different smoothing parameters. In our case,
we have not only a random design but also a multivariate context.
For a similar context to ours, Aït-Sahalia and Lo (1998) use a global univariate selection criterion. However, this criterion does not take into account the multivariate character of the estimator (the optimum univariate bandwidths might be different from the optimal bandwidths in the multivariate context), and, moreover, they do not offer any criteria for choosing the constant involved in the estimation.

It seems clear to us that the bandwidth parameter is the most important quantity to be selected for any nonparametric estimation; it must definitely be carefully selected. It seems, therefore, convenient to analyze the stability of the smoothing estimators, and the robustness of results, relative to alternative selection methodologies. Appendix C contains a detailed discussion of the alternative techniques employed in this paper to calculate bandwidths.

With these considerations in mind, and being concerned with computation costs whenever a very sophisticated method is used, we propose the following methodology. We first compute the univariate pilot bandwidths for our three explanatory variables by using a plug-in method, where the constants are selected with an iterative method due to Gasser, Kneip and Köhler (1991) and are discussed in Appendix C. Since the asymptotical rates of convergence suggest that the bandwidth is influenced by the dimensionality of the problem, we check the validation of these pilot bandwidths by using two alternative multivariate cross-validation criteria. In particular, we employ the natural extensions of the Generalized Cross-Validation (GCV) method and Rice’s bandwidth selectors to the multivariate case. All of these are presented in Appendix C. They are evaluated in a grid of bandwidths around the pilot smoothing parameters.

The need of a multivariate criterion may be easily justified. The rate at which h must go to zero is of order $n^{-1/(4+d)}$, where d denotes the dimension of the covariate vector. Therefore, we may expect a large h whenever we employ a multivariate criterion. In practice, however, there are no substantial differences among the alternative criteria.
3. The Data: The Spanish IBEX-35 and the Option Contract

The Spanish IBEX-35 index is a value-weighted index comprising the 35 most liquid Spanish stocks traded in the continuous auction market system. The official derivative market for risky assets, which is known as MEFF, trades a futures contract on the IBEX-35, the corresponding option on the IBEX-35 futures contracts for calls and puts, and individual option contracts for blue-chip stocks.

The Spanish option contract on the IBEX-35 futures is a cash settled European option with trading during the three nearest consecutive months and the other three months of the March-June-September-December cycle. The expiration day is the third Friday of the contract month. During the sample period covered by this research, the multiplier changed from 100 Spanish pesetas times the IBEX-35 index at the beginning of the sample period to 1000 pesetas during 1998. Prices are quoted in full points, with a minimum price change of one index point. The exercise prices are given by 50 index point intervals.

Our database is comprised of all call and put options on the IBEX-35 index futures traded daily on MEFF during the period January 1996 through November 1998. Liquidity is concentrated on the nearest expiration contract. In fact, during the sample period almost 90% of crossing transactions occurred in this type of contracts. Given the concentration in liquidity, our daily set of observations includes only calls and puts with two possible expiration dates. We only include options which expire at between five and forty days. That is, we eliminate all transactions taking place during the last five days before expiration, and transactions which will expire in more than forty days.

As usual in this type of research, our primary concern is the use of simultaneous prices for the options and the underlying security. The data, which are based on all reported transactions during each day throughout the sample period, do not allow us to observe simultaneously enough options with the same time-to-expiration on exactly the same underlying security price but with different exercise prices. In order to avoid large

---

6 Starting in January 1999, it has been changed to 10 euros.
variations in the underlying security price, we restrict our attention to the 45-minute window from 16:00 to 16:45. It turns out that almost 25% of crossing transactions occur during this interval. Moreover, care was also taken to eliminate the potential problems with artificial trading that are most likely to occur at the end of the day. Thus, all trades after 16:45 were eliminated so that we avoid data which may reflect trades to influence market maker margin requirements. Finally, we eliminate all call and put prices that violate the well known arbitrage bounds. The number of observations within a day may vary according to the number of crossing transactions associated with different exercise prices available each day.

These exclusionary criteria yield a final daily sample of 8321 observations (4798 calls and 3523 puts). The implied volatility for each of our 8321 options is estimated next. To do this, we take as the underlying asset the average of the bid and ask price quotation given for each futures contract associated with each option during the 45-minute interval. To proxy for riskless interest rates, we use the daily series of annualized repo T-bill rates with either one week, two weeks or three weeks to maturity. One of these three interest rates will be employed depending upon how close the option is to the expiration day.

Table 1 describes the sample properties of the call and put option prices employed in this work. Average prices, average relative bid-ask spread and the average number of contracts per day are reported for each moneyness category. Moneyness is defined as the ratio of the exercise price to futures price. A call (put) option is said to be deep out-of-the-money (deep in-the-money) if the ratio X/F is >1.03; out-of-the-money (in-the-money) if 1.03 ≥ X/F > 1.01; at-the-money when 1.01 ≥ X/F > 0.99; in-the-money (out-of-the-money) when 0.99 ≥ X/F > 0.97; and deep-in-the-money (deep out-of-the-money) if 0.97 > X/F. As we already discussed, there are 4798 call option observations (3523 puts), with OTM, ATM and ITM call (put) options respectively representing 61% (68), 30% (25) and 9% (7). The average call (put) price ranges from 61.6 (64.1) pesetas for deep OTM call (put) options to 381.2 (461.7) pesetas for deep ITM call (put) options. The average relative bid-ask spread tends

---

7 Note that the window employed in this research is characterized as being a highly liquid interval. This is important since it works against finding liquidity effects.
to move in the opposite direction to the average price. In particular, it ranges from 0.38 (0.33) for deep OTM call (put) options to 0.12 (0.20) for deep ITM calls (ITM puts).

4. Smiles and Risk-neutral Densities

It is well recognized that option prices provide market participants with a tremendous amount of information. In particular, we have already discussed how to infer the risk-neutral density or, alternatively, the Arrow-Debreu prices from trading options on the market portfolio. This section presents our nonparametric estimation of both the risk-neutral density and the volatility smile for 1998.

We first discuss the nonparametric estimation of the univariate (traditional) volatility smile. As before, we employ the SNN estimator instead of the classical kernel estimator, and the plug-in bandwidth selection method:

\[ \hat{\sigma}(\xi) = \frac{1}{nh_\xi} \sum_{i=1}^{n} K \left( \frac{F_n(\xi_i) - F_n(\xi)}{h_\xi} \right) \sigma_i \]

Expression (7) is firstly estimated using only call options transacted during 1998. It is interesting to note the major differences obtained when we employ the SNN procedure rather than the traditional kernel estimator. Figure 1 contains the nonparametric volatility smile estimated with both methods. It turns out that the optimal bandwidth under the SNN estimator is 0.093 and, as we can see in Figure 1, the smile is optimally smoothed. A very different pictures emerges when we employ the traditional kernel estimator given by:

---

8 During 1998, financial markets experienced great volatility. Our presentation for 1998 should be taken as just a working example. In any case, the main implications of our analysis could have been obtained with any other year of the sample period.
The effect of undersmoothing is clearly reflected in Figure 1. The estimator is highly variable because the optimal bandwidth selected according to the plug-in method tends to zero. We argue that, given the specific data we usually have when doing research in option pricing, where a lot of observations are centered around the at-the-money options and relatively few observations are available in the extremes, the SNN estimator is more appropriate.

It is also well known that, given equation (6), the pattern of implied volatilities for alternative exercise prices (the smile) gives us direct evidence of the risk-neutral density function actually embedded in option price data. Using call options transacted during 1998, as an example, we next estimate the implied risk-neutral distribution recognizing the potential effects of not only moneyness but also time-to-expiration and liquidity.

As already mentioned in Section 2, we first estimate the nonparametric function, \( \hat{\sigma}(\xi, SP, \tau) \) where \( \xi \) is the degree of moneyness, SP is the relative bid-ask spread, and \( \tau \) is time-to-expiration. We employ the SNN estimator given by expression (4). It should be pointed out that we also estimate the nonparametric volatility function without the liquidity variable. This is an important issue in this paper. These estimations will allow us to compare the implied risk-neutral densities with and without liquidity effects.

Table 2 contains the optimal bandwidths given by the plug-in criterion for our three explanatory variables and for the three years in our sample. Given the high degree of uncertainty experienced by the market during 1998, it should come as no surprise to observe that the bandwidths become larger for all variables during 1998. Otherwise, the results seem to be reasonable.
Once we have the nonparametric volatility functions with and without liquidity, we employ equation (6) to estimate the risk-neutral density function for 1998. In order to do so, we first observe all exercise prices available during 1998. The rest of the explanatory variables are assumed to be constant in their means for that year, so that the future price, the relative spread and time-to-expiration remain constant in their means, $\bar{F}, \bar{S}, \bar{\tau}$. Note that the only variable allowed to vary is the exercise price. We now estimate the nonparametric volatility function given by (4) in the new knots $(X_i/F, S, \tau)$ where $X_i, i = 1, \ldots, n$, are the number of exercise prices observed for that particular period. The implied risk-neutral distribution is then given by\textsuperscript{9}:

$$f_t^*(S_T) = e^{\tau} \left[ \frac{\partial^2 c(X_i/F, \tau, r, \sigma(X_i/F, S, \tau))}{\partial X^2} \right]_{X_i = S_T}$$ (9)

Since we are interested in the effects of liquidity on option pricing, the same procedure is repeated without taking into account the bid-ask spread variable. Thus, we have two risk-neutral densities, in the first of which the potential effects of market frictions are explicitly considered. As a reference, we also present the Black-Scholes distribution by employing the at-the-money mean implied volatility during 1998 as an input in the expression of the lognormal density. Note that this (mean) implied volatility is used as a constant volatility in all knots where the density is estimated.

Figure 2 contains the estimated risk-neutral density for 1998 in terms of returns rather than levels of exercise prices. This figure is obtained using all calls available in the sample. The returns are simply calculated as $R_i = \ln(X_i/F)$. The results imply that the density estimated with liquidity presents a (slightly) thinner left tail and a fatter right tail relative to BS. Moreover, it also presents a (slightly) thinner left tail with respect to the non-liquidity case. Finally, the estimation incorporating liquidity costs is clearly more leptokurtic than the other two cases. All of these differences suggest that there might be important

\textsuperscript{9} In fact, the estimation of the implied risk-neutral density is easily simplified by noting that the only relevant term of equation (B.2), in terms of magnitude, is the first component of the expression. The last two terms are
differences in pricing options once liquidity costs are taken into account. They will be further investigated in the next section.

5. Empirical Results

5.1 In-sample pricing performance

This section evaluates the performance of the alternative semiparametric option pricing models described above.

Two theoretical option prices are calculated using the nonparametric volatility functions estimated according to our two versions of equation (4) with and without liquidity. Once we have these functions, we employ Black’s (1976) model to obtain our theoretical semiparametric option prices for each call in the sample. In particular, we calculate the following call prices by:

\[
\tilde{c}_i \left( \xi_i, \tau_i, r_i, \sigma_i \right) = e^{-r_i \tau_i} \left[ F_i N(d_{li}) - X_i N(d_{2i}) \right] 
\]

(10)

where,

\[
d_{li} = \frac{\ln(F_i/X_i) + (\hat{\sigma}_i/2)\tau_i}{\hat{\sigma}_i \sqrt{\tau_i}}, \quad d_{2i} = d_{li} - \frac{\sigma_i \sqrt{\tau_i}}{2} 
\]

(11)

where each option i available in the sample is characterized by a futures price, \( F_i \), an exercise price, \( X_i \), a time-to-expiration, \( \tau_i \) and, given the days to maturity, the corresponding repo rate with similar maturity, \( r_i \). As the input for volatility, \( \hat{\sigma}_i \), we introduce the (nonparametrically) estimated volatility function, which is obtained by either of the following two SNN estimators:

---

very small since they are multiplied by 1/X and 1/F² respectively.
\[ \hat{\sigma}_{WL} = \frac{1}{nh_\tau h_{SP}} \sum_{i=1}^{n} K \left( \frac{F_n(\xi_i) - F_n(\xi_j)}{h_\xi} \right) K \left( \frac{F_n(SP) - F_n(SP)}{h_{SP}} \right) K \left( \frac{F_n(\tau) - F_n(\tau_j)}{h_\tau} \right) \]  

\[ \hat{\sigma}_{WOL} = \frac{1}{nh_\tau h_{SP}} \sum_{i=1}^{n} K \left( \frac{F_n(\xi_i) - F_n(\xi_j)}{h_\xi} \right) K \left( \frac{F_n(\tau) - F_n(\tau_j)}{h_\tau} \right) \]  

It should be pointed out that these two equations are estimated quarterly over the whole sample period from January 1996 to November 1998. The plug-in criterion is employed to calculate the optimal bandwidth for each explanatory variable and for each quarter. In this way, we have 4798 pricing errors for calls from January 2, 1996 to November 10, 1998, for each of the models analyzed.

The statistical significance of performance is assessed by analyzing the proportion of theoretical prices lying outside their corresponding bid-ask spread boundaries. This will allow us to test whether or not the differences between the pricing performance of our two competing models are statistically different from zero.

For each model and each moneyness category, we compute the proportion of options such that the estimated theoretical price falls outside the bid-ask boundary. Let us denote by:

- \( p^1 \) the proportion of calls whose theoretical price lies outside the bid-ask spread, when we price with liquidity costs.
- \( p^2 \) the proportion of calls whose theoretical price lies outside the bid-ask spread, without liquidity costs.
For each call in the sample:

- Let $Z_i$ be 1 if theoretical price is outside the spread and 0 otherwise (with liquidity)
- Let $Z_j$ be 1 if theoretical price is outside the spread and 0 otherwise (without liquidity).

We want to test whether $p^1 < p^2$. If this is the case, we may argue that liquidity is a relevant variable in pricing call options. One should be careful in defining this statistic. It should be noticed that these competing models are not independent. Hence,

$$Z_i - Z_j = \begin{cases} 
-1 & \text{with } p_i \\
0 & \text{with } p_2 \\
1 & \text{with } 1 - p_1 - p_2 
\end{cases}$$

Under the null hypothesis of equal proportions we have that

$$E(Z_i - Z_j) = 0 \Rightarrow p_2 = 1 - 2p_1$$
$$\text{var}(Z_i - Z_j) = 2p_1$$

Note that,

$$\frac{1}{n} \sum_{i=1}^n Z_i = p^1 \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^n Z_j = p^2$$

We test $H_0 : p^1 \geq p^2$ against $H_a : p^1 < p^2$. In this context, the Z-statistic is given by:

$$Z = \frac{1}{n} \left[ (Z_i^1 - Z_i^2) + ... + (Z_j^1 - Z_j^2) \right]$$

(14)
where \( n \) is the sample size. This statistic is normally distributed with mean 0 and variance 
\[
\frac{2p_1}{n}
\]. The maximum likelihood estimator of \( p_1 \) is:
\[
\hat{p}_1 = \frac{\#\{1\} + \#\{-1\}}{2n}
\] (15)

Therefore, the Z-statistic proposed in this paper to test the differences in proportions is given by:
\[
Z = p_1 - p^2 
\approx N\left(0, \frac{2\hat{p}_1}{n}\right)
\] (16)

We might be also interested in knowing whether a given theoretical pricing model undervalues or overvalues market prices. Thus, the Z-statistic is also calculated to obtain the proportion for which the theoretical model yields a price below the bid quote, and the proportion for which the model gives a price above the ask quote. If a theoretical model tends to undervalue market prices, it will yield a higher proportion of prices below the bid quote. If, on the other hand, the model tends to overvalue market prices, it will have a higher proportion of prices above the ask quote.

Table 3 reports the in-sample results for each year and for each moneyness category. The empirical evidence contained in the table are quite impressive. For OTM and ATM calls the pricing performance of the semiparametric option pricing model with liquidity is statistically and systematically superior to the model without liquidity. In-sample pricing performance is clearly improved by recognizing that liquidity effects are present in the pricing of options. This is also the case when we consider all calls together. The evidence is slightly less clear for ITM calls, particularly for 1998. Liquidity, as proxied by the relative bid-ask spread, seems to be an important variable in pricing call options in the Spanish market.
In order to validate our previous results we carried out a slightly different in-sample test. The sample available in each each quarter is separated into two subsamples. The first subsample contains 10% of calls for each moneyness degree (subsample 10). The other subsample (subsample 90) contains the rest of the options. The price of each call in subsample 10 is estimated with the information contained in subsample 90. The bandwidths are those selected by the plug-in method with all available calls in the corresponding quarter. Note that this is not the usual in-sample pricing because the calls in subsample 10 are not included in their estimation. In Table 4, as in the previous case, we again analyze the proportion of theoretical prices lying outside their corresponding bid-ask spread boundaries. The results are reported for the case in which we consider all calls together. Again, the option pricing model with liquidity is superior to the model without liquidity, although in some cases we are not able to reject the null hypothesis. The same results hold when we separate the subsamples according to their degree of moneyness\(^{10}\). We may conclude that liquidity is a relevant factor in determining market prices for call options.

### 5.2 In-sample pricing performance with alternative smoothing parameters

As mentioned in Section 2.2, we check the appropriateness of the plug-in smoothing parameter estimators by calculating three different bandwidths obtained from the three alternative multivariate criteria described in Appendix C. They are given by expressions (C.4) and (C.5).

It should be recognized that our nonparametric problem is inherently a multivariate estimation problem. These alternative criteria -the GCV and the Rice’s estimators- have been used in univariate contexts. We employ their extensions to the multivariate cases. Although these extensions are quite natural, to the best of our knowledge, they have not been used before in empirical work.

Table 5 reports the bandwidths obtained for each criteria and for each year during the sample period, and Table 6 compares the effects of the GCV selection method with the

\(^{10}\) Results are available upon request.
plug-in method on option pricing. As we can observe from the table, when we take out 10% of the sample, we cannot statistically reject the equality of proportions between the two selection models. This result suggests that the plug-in method produces quite a good estimation of the smoothing parameter. In any case, although not reported, the results in terms of lower proportions of prices lying outside the bid-ask spread boundaries in the model with liquidity remain the same. Further research specifically directed toward this issue is clearly justified\(^\text{11}\).

### 5.3 The out-of-sample performance of competing models

Before reporting the out-of-sample results, it should be mentioned that we test the stability of risk-neutral densities from one quarter to another. Practical reasons have led us to use a simple and easy-to-implement test based on the so called *randomization tests* to detect whether the differences between the risk-neutral densities are statistically significant. This test lies on the permutation of options in our dataset. More specifically, we assume that options may be randomly assigned to different testing periods. Data are permuted repeatedly and a test statistic is computed for each of the resulting data permutations. These data permutations, including the one representing the results obtained, constitute the reference set for determining the level of significance. The proportion of data permutations in the reference set that have a value of the test statistic greater than or equal to the value of the experimentally obtained results is the p-value. It should be noted that, in the proposed test, the basis for permuting the data is random assignment. This is why it is known as a randomization test. See Good (1994) for details on randomization tests. Although the results are not reported for space reasons, it is interesting to point out that for all cases the null hypothesis of no significant differences between quarters cannot be rejected\(^\text{12}\).

This surprising result seems to be useful in justifying the use of a given risk-neutral density estimated for a given quarter as the basis for pricing options during the following quarter. Having this in mind, the out-of-sample performance is investigated next. We use the

---

\(^\text{11}\) Note that Table 6 is different from Table 4. This is because in Table 4 we price quarterly, with the bandwidths selected for each quarter, and in Table 6 we price with the bandwidths selected yearly.

\(^\text{12}\) Results available upon request.
A nonparametric volatility function estimated for a given quarter \( q \) as the true volatility function for the following quarter \( q+1 \).

We again calculate the theoretical option prices of our two competing models using expression (10) by employing the nonparametric volatility function estimated over the previous quarter. As before, the Z-statistic given in equation (16) is used.

The results are contained in Table 7. Unfortunately, the empirical performance of both models is really disappointing. Independently of the degree of moneyness considered, we are not able to reject equality of performance between the two models. Moreover, the proportion of theoretical prices lying outside the bid-ask spread is considerably higher than in the in-sample case. The conclusion seems to be clear. The out-of-sample performance of our semiparametric models is quite poor both with liquidity and without liquidity.

It should be noted that the results of Table 7 are based on the comparison between the second and third quarter of each year during the sample period. We report this specific out-of-sample quarter performance as a representative case of what is found in all quarters. In fact, as a way of aggregating quarter Z statistics over a single year, the quarter Z´s were added together and divided by the square root of the number of quarters where we do the pricing (\( \sqrt{4} \) in 1997, \( \sqrt{3} \) in 1996 and 1998), to obtain an aggregate N(0,1) statistic from which an aggregate p-value is obtained. Although not reported, these aggregate p-values suggest that we are never able to reject equality of proportions between the model with liquidity and the model without liquidity. The out-of-sample performance is always poor regardless of the model.

This is an interesting result. In spite of the fact that we are not able to reject the stability of the risk-neutral densities over time, the pricing results are very disappointing. This should be understood as a warning about the stability test as a sufficient condition for out-of-sample pricing.
6. Conclusions

This paper has investigated the effects of liquidity, as proxied by the relative bid-ask spread, on the pricing of options. Given the evidence contained in Lonstaff (1995) and Peña, Rubio and Serna (1999), where linear (and non-linear) causality tests between the shape of the volatility smile and the bid-ask spread show a bidirectional Granger causality, this paper estimates a nonparametric volatility function in which liquidity is incorporated as a key explanatory variable. Given the structure of our dataset across the degree of moneyness, the paper employs the Symmetrized Nearest Neighbors (SNN) nonparametric estimator rather than the traditional kernel estimator. Moreover, special care is taken in the estimation of the smoothing parameter. The results show that in-sample performance is clearly better in a semiparametric model with liquidity than in a similar model estimated without liquidity. In fact, liquidity is very useful if we employ options close together in time. This is supported by the in-sample pricing when we take off 10% of data.

Unfortunately, however, the out-of-sample performance results are quite disappointing. We employ the nonparametric volatility function estimated for a given quarter q as the true volatility function for the following quarter q+1. Our two semiparametric competing option pricing models present high proportions of theoretical prices lying outside the bid-ask spread boundaries. At the same time, we are not able to reject the equality of these proportions across both models. This result is found despite the fact that we do not reject the stability of risk-neutral densities over the quarters covered by the research. In any case, it just might be that the power of this test is low, although it is also true that the chosen length to carry out the out-of-sample test is an issue in itself.

Given our in-sample result, we are planning to analyze the pricing performance of our semiparametric option pricing model with liquidity on a daily basis. That is to say, we will estimate the nonparametric volatility function daily with the SNN estimator, and this function will be used as an input in the option pricing function for the following day. This procedure is closer in spirit to other empirical related papers in option pricing.
APPENDIX A
SYMMETRIZED NEAREST NEIGHBORS (SNN) METHODOLOGY

For the sake of simplicity, we make a comparison between kernel and SNN estimators in the univariate context. Let us assume the following data generating process,

\[ Y_i = m(X_i) + \varepsilon_i ; \quad i = 1, \ldots, n \]  

(A.1)

where \( m(.) \) is the unknown function, and \( X_i \) are i.i.d. random variables having a density function \( f(x) \). The disturbance terms are also assumed to be i.i.d., and a sample of size \( n \) is taken. The main advantage of the nonparametric estimation is that we do not have to assume anything about the functional form of the regression function. We only assume that this function is smooth and (at least) twice differentiable. The kernel estimator is:

\[
\hat{m}_K(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) Y_i
\]

(A.2)

where \( h \) is the smoothing or bandwidth parameter to be selected.

The kernel or weight function \( K \) has the following properties:

\[
\int K(u) du = 1
\]
\[
\int K(u)udu = 0
\]
\[
\int K(u)^2 du = d_k \quad \text{(finite)}
\]
\[
\int K(u)^2 du = c_k \quad \text{(finite)}
\]

In practice, we employ the Gaussian kernel which has these properties and is given by,
The estimation proposed by Yang (1981) and studied by Stute (1984) is:

\[
\hat{m}_S(x) = \frac{1}{nh} \sum_{i=1}^{n} K \left( \frac{F_n(x) - F_n(X_i)}{h} \right) Y_i
\]

(A.3)

The known \textit{k-nearest neighbors} estimate defines neighbors in terms of the k-nearest points of \( x \). On the other hand, with the SNN methodology, neighbors are defined in terms of the distance between the values of the empirical distribution function at each point. Since expression (A.3) picks up its neighbors symmetrically, it is known as the \textit{Symmetrized Nearest Neighbors} (SNN).

Under the usual assumptions in kernel estimation, it is well known that the leading term in the mean square error is given by \(^{13}\):

\[
\text{MSE} \left( \hat{m}_K(x) \right) = \frac{d_K^2}{4} h^4 \frac{(m^2 f + 2m' f')^2(x)}{f^2(x)} + \frac{s^2}{nh} c_K
\]

(A.4)

where \( d_K \) and \( c_K \) are constants (already defined above) that depend on the kernel chosen, and \( s^2 \) is the variance of the disturbance term.

The MSE for the SNN estimator is:

\[
\text{MSE} \left( \hat{m}_S(x) \right) = \frac{d_K^2}{4} h^4 \left( \frac{(m^2 f - m' f')(x)}{f^3(x)} \right)^2 + \frac{s^2}{nh} c_K
\]

(A.5)
By minimizing the MSE for each case, the optimal bandwidth $h$ is derived. This is substituted into (A.4) and (A.5) to get the minimum MSE denoted by $\text{MSE}^*$. If we employ the kernel estimator:

$$\text{MSE}^*(\hat{m}_K(x)) = \frac{5}{4} \left( \frac{d_K^2}{2} \right)^{1/5} \left( \frac{s^2}{2} \right)^{4/5} c_K \frac{H(x)^{1/5}}{n^{4/5} f(x)^{4/5}}$$

(A.6)

where,

$$H(x) = \frac{(m''f + 2m'f')^2(x)}{f^2(x)}$$

On the other hand, if we use an SNN estimator there is an equivalent expression which asymptotically is equal to:

$$\text{MSE}^*(\hat{m}_S(x)) = \frac{5}{4} \left( \frac{d_K^2}{2} \right)^{1/5} \left( \frac{s^2}{2} \right)^{4/5} c_K \frac{S(x)^{1/5}}{n^{4/5}}$$

(A.7)

where,

$$S(x) = \frac{(m''f - m'f')^2(x)}{f^6(x)}$$

Thus, we can conclude that it is advisable to employ the SNN estimator instead of the kernel estimator when $\text{MSE}^*(\hat{m}_S(x)) < \text{MSE}^*(\hat{m}_K(x))$. This will be the case for any $x$ such that:

$$\frac{H(x)}{f^4(x)} > S(x)$$

(A.8)

Alternatively, condition (A.8) is equivalent to:

$^{13}$ MSE = (Bias)$^2$ + Variance.
\[(m''f + 2m'f')^2 > (m''f - m'f')^2 \iff m'f'(m'f' + 2mf) > 0 \quad (A.9)\]

As we have already mentioned, depending on the specific functional form of the unknown function, \(m(x)\), and the density function, \(f(x)\), we would prefer one or the other. We illustrate this behavior with a simple example. Suppose that we have the following parametric model:

\[y_i = m(x_i) + \varepsilon_i\]

where

\[m(x_i) = ax_i^2\]

Let us also assume that the errors are i.i.d. and that \(x\) have a normal density with mean 0 and variance \(s^2\):

\[f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/s^2}\]

By taking the appropriate derivatives for \(m(x)\) and \(f(x)\) and substituting them into expression (A.9), we obtain that the MSE under the kernel estimation is higher than in the SNN estimation as long as \(|x| > s\). In words, when the points are at a distance from the mean (in this example zero, but it may be generalized) greater than the standard deviation, we obtain better results using the SNN estimation. Note the choice of the particular form of \(m(.)\) and \(f(.)\). A quadratic form for \(m(.)\) could be similar to the pattern of the volatility smile. On the other hand, the normality of \(x\) tries to show the behavior where the design has a few amount of data in the tails relative to the center. Of course, this is similar to the moneyness distribution we observe in practice.

This example suggests that whenever we are interested in estimating in those places where the density is very low (away from the mean), the SNN may yield better results than the
traditional kernel estimators. Given the empirical distribution of moneyness, this suggests that, in general, the $\text{MSE}\ast\left(\hat{m}_S(x)\right)$ might be lower in those sections of the smile where the degree of moneyness is far away from at-the-money options. Of course, as we said before, whether in general $\text{MSE}\ast\left(\hat{m}_S(x)\right) < (>) \text{MSE}\ast\left(\hat{m}_K(x)\right)$ depends on the particular functional form of the unknown function $m(.)$ and the density $f(x)$. We suspect that in most cases the SNN estimator would present better results when estimating in places where the density is low.
APPENDIX B
APPROPRIATE DERIVATIVES FOR THE CALCULATION
OF THE RISK-NEUTRAL DENSITY

The Black-Scholes formula for a European futures call option is\(^{14}\):

\[
c = e^{-r\tau}[ FN(d_1) - XN(d_2) ]
\]

where,

\[
d_1 = \frac{\ln(F/X) + \left(\sigma^2/2\right)\tau}{\sigma\sqrt{\tau}}
\]

\[
d_2 = \frac{\ln(F/X) - \left(\sigma^2/2\right)\tau}{\sigma\sqrt{\tau}} = d_1 - \sigma\sqrt{\tau}
\]

\(N(d_1)\) is the value for the accumulative probability distribution of a normal variable with mean 0 and variance 1. Its derivative is, therefore, the normal density function:

\[
N'(d_1) = \frac{1}{\sqrt{2\pi}} e^{-d_1^2/2}
\]

It is easy to show that,

\[
N'(d_2) = \frac{1}{\sqrt{2\pi}} e^{-d_2^2/2} = N'(d_1) \frac{F}{X}
\]

Our objective is to find the second derivative of the call price with respect to the exercise price. It should be noted that, in our case, the volatility is also a function of the exercise price:

\(^{14}\) In fact, this is Black (1976).
\[
\frac{\partial^2 \hat{c}}{\partial X^2} = \frac{\partial^2 c}{\partial X^2} + 2 \frac{\partial^2 c}{\partial X \partial \sigma} \frac{\partial \hat{c}}{\partial X} + \left[ \frac{\partial^2 c}{\partial \sigma^2} + \frac{\partial c}{\partial \sigma} \right] \frac{\partial^2 \hat{c}}{\partial X^2}
\]

(B.1)

We need the following results,

\[
\frac{\partial c}{\partial X} = -e^{-\tau} N(d_2)
\]

\[
\frac{\partial^2 c}{\partial X^2} = -\tau e^{-\tau} \frac{1}{X \sqrt{2\pi \sigma^2 \tau}} \exp \left\{ \frac{\ln(F/X) - \left(\sigma^2/2\right) \tau}{2\sigma^2 \tau} \right\}
\]

\[
\frac{\partial d_1}{\partial X} = -\frac{1}{X \sigma \sqrt{\tau}}
\]

\[
\frac{\partial c}{\partial \sigma} = -\tau e^{-\tau} N'(d_1) F \sqrt{\tau}
\]

\[
\frac{\partial c}{\partial \sigma \partial X} = -\tau e^{-\tau} \frac{F}{X} N'(d_1) \frac{1}{\sigma}
\]

\[
\frac{\partial^2 c}{\partial \sigma^2} = -\tau e^{-\tau} F \frac{d_2}{\sigma} \sqrt{\tau} N'(d_1)
\]

We need to calculate the derivatives of our volatility function estimation with respect to the exercise price, \( \partial \hat{\sigma} / \partial X \) and \( \partial^2 \hat{\sigma} / \partial X^2 \). To obtain them, a kernel estimation is used. But recall that our nonparametric estimation of volatility depends on moneyness, and not on the exercise price. Hence, the appropriate derivatives for the volatility with respect to the exercise price are:
\[ \frac{\partial^m \hat{\sigma}}{\partial X^m} = \frac{\partial^m \hat{\sigma}}{\partial (X/F)^m} \frac{1}{F^m} \]

Therefore,

\[ \frac{\partial^2 \hat{c}}{\partial X^2} = e^{-r\tau} \frac{1}{X \sqrt{2\pi \sigma^2 \tau}} \exp \left\{ -\left[ \ln \left( \frac{F}{X} \right) - \left( \frac{\sigma^2}{2} \right) \frac{1}{\tau} \right]^2 \frac{1}{2\sigma^2 \tau} \right\} \]

\[ -2e^{-r\tau} \frac{F}{X} N^*(d_1) \frac{1}{\sigma} \frac{\partial \sigma}{\partial (X/F)} \frac{1}{F} \]

\[ + \left[ -e^{-r\tau} \frac{F}{\sigma} \frac{d_2}{x} \sqrt{\tau} N^*(d_1) + e^{-r\tau} N'(d_1) F \sqrt{\tau} \right] \frac{\partial^2 \sigma}{\partial (X/F)^2} \frac{1}{F^2} \]

(B.2)
APPENDIX C
SELECTING THE SMOOTHING PARAMETER

We know that the accuracy of kernel smoothers as estimators of any function \(m(.)\), as in expression (A.1) is a function of the kernel and the bandwidth \(h\). In practice, it is well accepted that accuracy depends mainly on the chosen smoothing parameter \(h\). In this appendix we first discuss the plug-in method as applied to our multivariate particular case. Secondly, we present the multivariate cross-validation criteria also employed in our estimations.

Let us briefly describe the appropriate algorithm for the simple kernel univariate case. We know that the asymptotic mean squared error for the usual kernel estimator is given by the following equation:

\[
\text{AMSE} = \frac{1}{nh} s^2 c_K + \frac{h^4 d_K^2 [m''(x)]^2}{4}
\]  

(C.1)

and the asymptotic integrated mean squared error is:

\[
\text{AMISE} = \frac{1}{nh} s^2 c_K + h^4 d_K^2 \int [m''(x)]^2 dx \left[\frac{1}{4}\right]
\]

(C.2)

\[
\frac{\partial \text{AMISE}}{\partial h} = 0 \Rightarrow h = \left(\frac{1}{n}\right)^{\frac{1}{5}} \left(c_K \left(\frac{s^2}{d_K} \int [m''(x)]^2 dx\right)\right)^{\frac{1}{5}}
\]

where the problem is that \(m''(x)\) is unknown and must be estimated. Note that this is the case given that the constants \(c_k\) and \(d_k\) depend on the kernel function assumed, \(n\) is the sample size, and the variance of the response variable may easily be estimated as,

\[
s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2
\]
Let $\hat{m}_2(x, h_2)$ be the estimator of $m''(x)$ where $h_2$ refers to the second derivative. It turns out that this can be estimated using a kernel estimator again.

$$\hat{m}_2(x, h_2) = \frac{1}{nh_2^3} \sum_{i=1}^{n} K_h \left( \frac{x - X_i}{h_2} \right) Y_i$$

where the specific kernel assumed must have the following properties:

- $\int K''(u) du = 0$
- $\int K''(u) u du = 0$
- $\int K''(u) u^2 du = 2$
- $\int K''(u) u^3 du = 0$
- $\int K''(u) u^4 du = d_k$ (finite)
- $\int K''(u)^2 du = c_k$ (finite)

In particular, the kernel function $K''(u) = \frac{105}{16} (-5u^4 + 6u^2 - 1)$ has these properties and has been used in this estimation.

The bandwidth $h_2$ is estimated directly according to the algorithm proposed by Gasser, Kneip, and Köhler (1991) for the univariate case. The algorithm is performed by the following steps:

(i) $\hat{h}_0 = \frac{1}{n}$

(ii) $\hat{h}_j = \left( \frac{c_k s^2}{nd_k^2 \int \hat{m}_2(x, h_{j-1}^{1/10})^2 dx} \right)^{1/5}$

(iii) stop when it converges
In words,

- Obtain the expression for the mean integrated squared error for the nonparametric regression (expression C.2)
- Estimate the partial derivatives involved using pilot bandwidths. In this case we need the second derivative of the m(x) function, where the pilot bandwidth is $h_2$.
- Get the initial bandwidth and transform it to obtain other pilot bandwidths.
- Iterate until convergence.

The extension of this univariate plug-in method to the multivariate case is not trivial. For this reason, in order to check the robustness of our results, we employ the natural extensions of the univariate cross-validation criteria given in Härdle (1990). In particular, the criteria employed in this research and their penalizing functions are the following:

(i) The Generalized Cross-Validation (GCV), where we have to minimize:

$$GCV(h_1, h_2, h_3) = \frac{\text{RSS}(h_1, h_2, h_3)}{1 - \frac{K(0)^3}{n} \sum_{j=1}^{n} \left( \frac{\xi_j - \hat{\xi}_i}{h_1} \right) \left( SP_j - SP_i \right) \left( \frac{\tau_j - \tau_i}{h_3} \right)}$$

(ii) Rice’s bandwidth selector, where we have to minimize:

$$\text{RICE}_1(h_1, h_2, h_3) = \frac{\text{RSS}(h_1, h_2, h_3)}{1 - \frac{2K(0)^3}{n} \sum_{j=1}^{n} \left( \frac{\xi_j - \hat{\xi}_i}{h_1} \right) \left( SP_j - SP_i \right) \left( \frac{\tau_j - \tau_i}{h_3} \right)}$$
where \(\text{RSS}(.)\) is the sum of squared residuals; this is to say, the residuals we obtain in the nonparametric estimation with the kernel. Finally, \(K(0)\) is the kernel evaluated at 0. It is important to point out that (C.4) or (C.5) are truly multivariate bandwidth selection methods.
REFERENCES

Aït-Sahalia, Y., and A. Lo (1998). “Nonparametric estimation of state-price densities implicit in financial asset prices”, Journal of Finance 53, pp.499-547.

Bakshi, G., Cao, C., and Z. Chen (1997). “Empirical performance of alternative option pricing models”, Journal of Finance 52, pp. 2003-2049.

Bates, D. (1996). “Jumps and stochastic volatility: exchange rate processes implicit in Deutsche mark options”, Review of Financial Studies 9, pp. 69-107.

Black, F. (1976). “The pricing of commodity contracts”, Journal of Financial Economics 3, pp. 167-179.

Black, F. and M. Scholes (1973). “The pricing of options and corporate liabilities”, Journal of Political Economy 81, pp. 637-659.

Chernov, M. and E. Ghysels (1999). “A study towards a unified approach to the joint estimation of objective and risk neutral measures for the purpose of option valuation”, unpublished manuscript, Pennsylvania State University.

Das, S., and R. Sundaram (1999). “Of smiles and smirks: A term-structure perspective”, Journal of Financial and Quantitative Analysis, 34, pp. 211-239.

Dumas, B., J. Fleming and R. Whaley (1998). “Implied volatility functions: empirical tests”, Journal of Finance, 53, pp. 2059-2106.

Fiorentini, G., León A. and G. Rubio (1999). “Short-term options with stochastic volatility: Estimation and empirical performance”, Studies on the Spanish Economy, Fedea 02, Madrid.
Gasser, T., Kneip, A. and W. Köhler (1991). “A flexible and fast method for automatic smoothing”, Journal of the American Statistical Association, 86, pp. 643-652.

Good, P. (1994). “Permutation Tests. A practical guide to resampling methods for testing hypotheses”, Springer-Verlag, New York.

Härdle, W. (1990). “Applied nonparametric regression”, Cambridge University Press, Cambridge.

Heston, S. (1993). “A closed-form solution for options with stochastic volatility with applications to bond and currency options”, Review of Financial Studies 6, pp. 327-344.

León, A. and G. Rubio (2000). “Smiling under stochastic volatility”, forthcoming in the Journal of Financial and Quantitative Analysis.

Longstaff, F. (1995). “Option pricing and the martingale restriction”, Review of Financial Studies 8, pp. 1091-1124.

Pan, J, (1999). “Integrated time-series analysis of spot and option prices”, unpublished manuscript, Stanford University.

Peña, I., Rubio, G. and G. Serna (1999). “Why do we smile? On the determinants of the implied volatility function”, Journal of Banking and Finance 23, pp. 1151-1179.

Peña, I., Rubio, G. and G. Serna (2000). “Smiles, bid-ask spreads and option pricing”, forthcoming in the European Financial Management.

Rubinstein, M. (1994). “Implied binomial trees”, Journal of Finance 49, pp. 771-818.

Stute, W. (1984). “Asymptotic normality of nearest neighbor regression function estimates”, Annals of Statistics 12. pp. 917-926.
Taylor, S. J. and X. Xu (1994). “The magnitude of implied volatility smiles: theory and empirical evidence for exchange rates”, The Review of Futures Markets 13, pp. 355-380.

Yang, S. (1981). “Linear functions of concomitants of order statistics with applications to nonparametric estimation of a regression function”, Journal of the American Statistical Association 76, pp. 658-662.
TABLE 1
SAMPLE CHARACTERISTICS OF IBEX-35 FUTURES OPTIONS

Average prices, average relative bid-ask spread and the average number of contracts per day are reported for each moneyness category. All call options transacted over the 45 minute interval from 16:00 to 16:45 are employed from January 2, 1996 to November 10, 1998. X is the exercise price and F denotes the futures price of the IBEX-35 index. Moneyness is defined as the ratio of the exercise price to the futures price. OTM, ATM, and ITM are out-of-the-money, at-the-money, and in-the-money options respectively.

| Moneyness | Average Price | Average Bid-Ask Spread | Average Number of Contrats per Day |
|-----------|---------------|------------------------|-----------------------------------|
| OTM: 1.01-1.03 | 81.7 | 0.237 | 404 |
| ATM: 0.99-1.01 | 120.4 | 0.174 | 340 |
| ITM: 0.97-0.99 | 180.0 | 0.122 | 271 |

CALLS

| Moneyness | Average Price | Average Bid-Ask Spread | Average Number of Contrats per Day |
|-----------|---------------|------------------------|-----------------------------------|
| DEEP OTM: > 1.03 | 61.6 | 0.378 | 154 |
| OTM: 1.01-1.03 | 81.7 | 0.237 | 404 |
| ATM: 0.99-1.01 | 120.4 | 0.174 | 340 |
| ITM: 0.97-0.99 | 180.0 | 0.122 | 271 |
| DEEP ITM: < 0.97 | 381.2 | 0.124 | 212 |
| ALL CALLS: - | 99.6 | 0.251 | 295 |

PUTS

| Moneyness | Average Price | Average Bid-Ask Spread | Average Number of Contrats per Day |
|-----------|---------------|------------------------|-----------------------------------|
| DEEP ITM: > 1.03 | 464.6 | 0.194 | 277 |
| ITM: 1.01-1.03 | 165.2 | 0.119 | 276 |
| ATM: 0.99-1.01 | 128.1 | 0.152 | 487 |
| OTM: 0.97-0.99 | 91.9 | 0.196 | 576 |
| DEEP OTM: < 0.97 | 65.1 | 0.325 | 243 |
| ALL PUTS: - | 99.3 | 0.233 | 399 |
TABLE 2

SMOOTHING PARAMETERS CALCULATED BY THE ITERATIVE PLUG-IN METHOD:
THE NONPARAMETRIC VOLATILITY FUNCTION

All call option data for each year separately are used in calculating the optimal bandwidth parameters for each of our three explanatory variables when estimating the nonparametric volatility function, $\hat{\sigma}(\xi, SP, \tau)$. We follow the univariate iterative procedure suggested by Gasser, Kneip, and Köhler (1991) in which the optimal bandwidth parameter, $h_j, (j = \xi, SP, \tau)$ is obtained by minimizing the mean integrated squared error.

| BANDWIDTH PARAMETERS | 1996 | 1997 | 1998 |
|-----------------------|------|------|------|
| $h_\xi$ (moneyness)   | 0.0537 | 0.0614 | 0.0931 |
| $h_{SP}$ (bid-ask spread) | 0.0383 | 0.0434 | 0.0721 |
| $h_\tau$ (time-to-expiration) | 0.0397 | 0.0502 | 0.0941 |
### TABLE 3
**IN-SAMPLE STATISTICAL SIGNIFICANCE OF PERFORMANCE FOR ALTERNATIVE SEMIPARAMETRIC OPTION PRICING MODELS: THE LIQUIDITY EFFECTS**

The nonparametric volatility function is estimated for each quarter from January 2, 1996 to November 10, 1998 using all available call options in each quarter. The corresponding call price is calculated by using the Black-Scholes pricing function evaluated at the previously (nonparametrically) estimated volatility. The bandwidth parameters are obtained by the plug-in method. The statistical performance for pricing errors is assessed by analyzing the proportion of theoretical prices lying outside their corresponding bid-ask spread boundaries. The Z-statistic for testing the differences between two proportions is employed. We report the statistical significance of pricing errors between the semiparametric option pricing model with liquidity (WL) and the semiparametric option pricing model without liquidity (WOL). For presentation reasons, the empirical results are aggregated for years. Moneyness is defined as the ratio of the exercise price to the futures price. OTM, ATM, and ITM are out-of-the-money, at-the-money, and in-the-money options respectively.

| CATEGORIES     | 1996  | 1997  | 1998  |
|----------------|-------|-------|-------|
| OTM CALLS:     |       |       |       |
| p(Bid>c>Ask)   | 0.110 | 0.171 | 0.306 |
|                | (0.000) | (0.000) | (0.000) |
| p(c > Ask)     | 0.064 | 0.086 | 0.156 |
|                | (0.000) | (0.000) | (0.000) |
| p(c < Bid)     | 0.046 | 0.084 | 0.149 |
|                | (0.000) | (0.000) | (0.000) |
| ATM CALLS:     |       |       |       |
| p(Bid>c>Ask)   | 0.100 | 0.180 | 0.245 |
|                | (0.000) | (0.000) | (0.000) |
| p(c > Ask)     | 0.050 | 0.117 | 0.141 |
|                | (0.000) | (0.000) | (0.000) |
| p(c < Bid)     | 0.050 | 0.101 | 0.103 |
|                | (0.000) | (0.000) | (0.000) |
| ITM CALLS:     |       |       |       |
| p(Bid>c>Ask)   | 0.139 | 0.137 | 0.302 |
|                | (0.000) | (0.000) | (0.000) |
| p(c > Ask)     | 0.054 | 0.166 | 0.151 |
|                | (0.003) | (0.000) | (0.000) |
| p(c < Bid)     | 0.084 | 0.101 | 0.151 |
|                | (0.024) | (0.000) | (0.000) |
| ALL CALLS:     |       |       |       |
| p(Bid>c>Ask)   | 0.109 | 0.170 | 0.290 |
|                | (0.000) | (0.000) | (0.000) |
| p(c > Ask)     | 0.057 | 0.228 | 0.152 |
|                | (0.000) | (0.000) | (0.000) |
| p(c < Bid)     | 0.052 | 0.172 | 0.138 |
|                | (0.000) | (0.000) | (0.000) |
TABLE 4
THE 10% IN-SAMPLE PRICING ERRORS FOR ALTERNATIVE SEMIPARAMETRIC OPTION PRICING MODELS: THE LIQUIDITY EFFECTS

The nonparametric volatility function is estimated from January 2, 1996 to November 10, 1998. The price of each call is calculated in a subsample containing 10% of all available calls using the nonparametric volatility function estimated with the remaining 90% of calls. The corresponding call price is obtained by using the Black-Scholes pricing function evaluated at the previously (nonparametrically) estimated volatility. The bandwidth parameters are estimated by the plug-in method. The statistical performance for pricing errors is assessed by analyzing the proportion of theoretical prices lying outside their corresponding bid-ask spread boundaries. The Z-statistic for testing the differences between two proportions is employed. The empirical results are aggregated for years. Moneyness is defined as the ratio of the exercise price to the futures price. We compare the pricing errors between the semiparametric option pricing model with liquidity (WL) and the semiparametric option pricing model without liquidity (WOL).

| CATEGORIES | 1996 | 1997 | 1998 |
|------------|------|------|------|
|            | WL   | WOL  | Z-STAT | WL  | WOL  | Z-STAT | WL  | WOL  | Z-STAT |
| ALL CALLS: |      |      |         |      |      |         |      |      |         |
| p(Bid > c > Ask) | 0.385 | 0.452 | -1.62 | 0.614 | 0.648 | -2.40 | 0.520 | 0.628 | -2.60 |
|              | (0.052) | (0.008) | (0.004) |      |      |         |      |      |         |
| p(c > Ask)  | 0.236 | 0.256 | -0.68 | 0.319 | 0.361 | -1.52 | 0.322 | 0.371 | -1.73 |
|              | (0.248) | (0.064) | (0.041) |      |      |         |      |      |         |
| p(c < Bid)  | 0.148 | 0.195 | -1.45 | 0.295 | 0.337 | -1.45 | 0.198 | 0.256 | -1.80 |
|              | (0.073) | (0.073) | (0.035) |      |      |         |      |      |         |
**TABLE 5**

**A COMPARISON OF ALTERNATIVE SMOOTHING PARAMETERS: THE NONPARAMETRIC VOLATILITY FUNCTION**

All call option data for each year separately are used in calculating the optimal bandwidth parameters for each of our three explanatory variables when estimating the nonparametric volatility function, $\hat{\sigma}(\xi, SP, \tau)$. We compare the univariate iterative procedure suggested by Gasser, Kneip, and Köhler (1991) in which the optimal bandwidth parameter is obtained by minimizing the mean integrated squared error, the multivariate the Generalized Cross-Validation (GCV), and the multivariate Rice’s bandwidth selector.

1996

| BANDWIDTH PARAMETERS | PLUG-IN | GCV   | RICE  |
|----------------------|---------|-------|-------|
| $h_{\xi}$ (moneyness) | 0.0537  | 0.0268| 0.0537|
| $h_{SP}$ (bid-ask spread) | 0.0383  | 0.0766| 0.0766|
| $h_{\tau}$ (time-to-expiration) | 0.0397  | 0.0198| 0.0635|

1997

| BANDWIDTH PARAMETERS | PLUG-IN | GCV   | RICE  |
|----------------------|---------|-------|-------|
| $h_{\xi}$ (moneyness) | 0.0614  | 0.0307| 0.0307|
| $h_{SP}$ (bid-ask spread) | 0.0434  | 0.0434| 0.0434|
| $h_{\tau}$ (time-to-expiration) | 0.0502  | 0.0251| 0.0251|

1998

| BANDWIDTH PARAMETERS | PLUG-IN | GCV   | RICE  |
|----------------------|---------|-------|-------|
| $h_{\xi}$ (moneyness) | 0.0931  | 0.0465| 0.0465|
| $h_{SP}$ (bid-ask spread) | 0.0721  | 0.0721| 0.1154|
| $h_{\tau}$ (time-to-expiration) | 0.0941  | 0.0470| 0.0470|
TABLE 6

THE 10% IN-SAMPLE PRICING PERFORMANCE FOR THE SEMIPARAMETRIC OPTION PRICING MODEL WITH LIQUIDITY FOR ALTERNATIVE SMOOTHING PARAMETERS:

The nonparametric volatility function is estimated from January 2, 1996 to November 10, 1998. The price of each call is calculated in a subsample containing 10% of all available calls using the nonparametric volatility function estimated with the remaining 90% of calls. The corresponding call price is obtained by using the Black-Scholes pricing function evaluated at the previously (nonparametrically). We compare the statistical significance of the in-sample pricing performance when the following smoothing parameter selection procedures are employed: the univariate iterative procedure suggested by Gasser, Kneip, and Köhler (1991) in which the optimal bandwidth parameter is obtained by minimizing the mean integrated squared error, and the multivariate the Generalized Cross-Validation (GCV). The statistical performance is assessed by analyzing the proportion of theoretical prices lying outside their corresponding bid-ask spread boundaries. The Z-statistic for testing the differences between two proportions is employed.

|                | 1996  | 1997  | 1998  |
|----------------|-------|-------|-------|
|                | Plug-in GCV Z-STAT | Plug-in GCV Z-STAT | Plug-in GCV Z-STAT |
| ALL CALLS:     |       |       |       |
| p(Bid>c>Ask)   | 0.542 | 0.483 | 1.87  | 0.744 | 0.651 | 2.46  | 0.817 | 0.790 | 0.94  | 0.969 | 0.993 | 0.827 |
| p(c > Ask)     | 0.281 | 0.254 | 1.26  | 0.465 | 0.401 | 2.11  | 0.527 | 0.506 | 0.77  | 0.897 | 0.982 | 0.780 |
| p(c < Bid)     | 0.261 | 0.228 | 1.21  | 0.279 | 0.250 | 1.09  | 0.290 | 0.283 | 0.27  | 0.887 | 0.862 | 0.609 |
TABLE 7
OUT-OF-SAMPLE STATISTICAL SIGNIFICANCE OF PERFORMANCE FOR ALTERNATIVE SEMIPARAMETRIC OPTION PRICING MODELS: THE LIQUIDITY EFFECTS

The nonparametric volatility function is estimated for each quarter from January 2, 1996 to November 10, 1998 using all available call options in each quarter. The corresponding call price is calculated by using the Black-Scholes pricing function evaluated at the previously (nonparametrically) estimated volatility. The bandwidth parameters are obtained by the plug-in method. The statistical performance for pricing errors is assessed by analyzing the proportion of theoretical prices lying outside their corresponding bid-ask spread boundaries. The Z-statistic for testing the differences between two proportions in each quarter is employed. We report the statistical significance of pricing errors between the semiparametric option pricing model with liquidity (WL) and the semiparametric option pricing model without liquidity (WOL), using data and the estimated nonparametric volatility function from one quarter to price options in the following quarter. For presentation reasons, the empirical results are reported only for the 2nd vs. the 3rd. quarter of each year. Moneyness is defined as the ratio of the exercise price to the futures price. OTM, ATM, and ITM are out-of-the-money, at-the-money, and in-the-money options respectively.

| CATEGORIES | 1996 (2nd vs. 3rd) | 1997 (2nd vs. 3rd) | 1998 (2nd vs. 3rd) |
|------------|---------------------|---------------------|---------------------|
|            | WL      | WOL     | Z-STAT | WL      | WOL     | Z-STAT | WL     | WOL     | Z-STAT |
| OTM CALLS: |         |         |         |         |         |         |         |         |         |
| p(Bid>c>Ask) | 0.670   | 0.592   | 2.16 (0.984) | 0.789   | 0.778   | 0.56 (0.712) | 0.915   | 0.907   | 0.60 (0.725) |
| p(c > Ask) | 0.111   | 0.139   | -1.88 (0.030) | 0.282   | 0.277   | 0.40 (0.655) | 0.384   | 0.382   | 0.22 (0.587) |
| p(c < Bid) | 0.558   | 0.452   | 3.21 (0.999) | 0.506   | 0.501   | 0.32 (0.625) | 0.530   | 0.525   | 0.81 (0.791) |
| ATM CALLS: |         |         |         |         |         |         |         |         |         |
| p(Bid>c>Ask) | 0.569   | 0.512   | 1.25 (0.894) | 0.771   | 0.638   | 2.74 (0.996) | 0.838   | 0.838   | 0.00 (0.500) |
| p(c > Ask) | 0.227   | 0.138   | 2.84 (0.997) | 0.180   | 0.085   | 2.50 (0.993) | 0.462   | 0.451   | 0.37 (0.644) |
| p(c < Bid) | 0.341   | 0.373   | -0.89 (0.186) | 0.590   | 0.552   | 1.00 (0.841) | 0.376   | 0.387   | -1.00 (0.158) |
| ITM CALLS: |         |         |         |         |         |         |         |         |         |
| p(Bid>c>Ask) | 0.703   | 0.741   | -0.57 (0.284) | 0.400   | 0.600   | -1.89 (0.029) | 0.937   | 0.937   | 0.00 (0.500) |
| p(c > Ask) | 0.148   | 0.148   | 0.00 (0.500) | 0.033   | 0.200   | -1.88 (0.030) | 0.937   | 0.937   | 0.00 (0.500) |
| p(c < Bid) | 0.555   | 0.592   | -0.57 (0.284) | 0.366   | 0.400   | -0.57 (0.384) | 0.000   | 0.000   | 0.00 (0.500) |
| ALL CALLS: |         |         |         |         |         |         |         |         |         |
| p(Bid>c>Ask) | 0.635   | 0.574   | 2.29 (0.988) | 0.762   | 0.737   | 1.29 (0.901) | 0.901   | 0.894   | 0.50 (0.691) |
| p(c > Ask) | 0.158   | 0.139   | 1.27 (0.897) | 0.245   | 0.231   | 1.02 (0.848) | 0.419   | 0.415   | 0.36 (0.640) |
| p(c < Bid) | 0.477   | 0.434   | 1.83 (0.966) | 0.516   | 0.506   | 0.66 (0.745) | 0.481   | 0.479   | 0.37 (0.644) |
Figure 1: Nonparametric Volatility Smiles in 1998: Kernel Estimator versus the SNN estimator.
Figure 2: Risk-Neutral Densities for Calls in 1998

Legend:
- Black-Scholes
- Nonparametric with liquidity
- Nonparametric without liquidity