Stat Comput (2010) 20: 421–434
DOI 10.1007/s11222-009-9134-y

Semi-parametric forecasts of the implied volatility surface using regression trees

Francesco Audrino · Dominik Colangelo

Received: 13 May 2008 / Accepted: 19 May 2009 / Published online: 6 June 2009

Abstract We present a new semi-parametric model for the prediction of implied volatility surfaces that can be estimated using machine learning algorithms. Given a reasonable starting model, a boosting algorithm based on regression trees sequentially minimizes generalized residuals computed as differences between observed and estimated implied volatilities. To overcome the poor predictive power of existing models, we include a grid in the region of interest, and implement a cross-validation strategy to find an optimal stopping value for the boosting procedure. Back testing the out-of-sample performance on a large data set of implied volatilities from S&P 500 options, we provide empirical evidence of the strong predictive power of our model.

Keywords Implied volatility · Implied volatility surface · Option pricing · Forecasting · Tree boosting · Regression tree · Functional gradient descent

1 Introduction

Despite the discrepancy between the Black and Scholes (BS) theory and reality, the concept of implied volatility surfaces (IVS) is still very popular. The mapping from observed market prices to *implied volatilities* (IV) is used as a way to compare options with different characteristics, i.e. different strike, expiry date, underlying,.... Implied volatilities of options are calculated and stored in financial databases. Market makers, traders, and risk managers rely on IV to calibrate their pricing models.

Classical approaches assume volatility to be a deterministic function of spot price and time. Dumas et al. (1998) find that estimated parameters are highly unstable over time, allowing only for short time predictions. As Poon and Granger (2003) point out, classical time series-based methods do not perform well in predicting volatility. In addition, they report that forecasting the volatility based on option-implied standard deviation provides superior performance across different assets and over very long forecast horizons (up to three years). This approach requires employing a strategy to decide which point on the IVS or which weighting scheme will be used to obtain a forecast of the volatility. Implied volatility function (IVF) models allow for dependency on state variables other than spot price and time. For example, Shimko (1993) models the observed IV as a function of the strike price and recovers the risk-neutral probability density function. Rosenberg (2000) analyzes options on S&P 500 futures with a dynamic IVF model.

More recently, Gonçalves and Guidolin (2006) combine a cross-sectional approach similar to that of Dumas et al. (1998) with vector autoregressive models in order to model the IVS dynamics. They try and partially succeeded (depending on the transaction costs) in exploiting the one-step and multi-step ahead volatility predictions produced by their model to form profitable volatility-based trading strategies. Skiadopoulos et al. (1999) popularize principal components analysis (PCA) in the IVS literature. They apply PCA on a multivariate time series of IV differences for a given moneyness level and within a certain expiry range. For a surface analysis, they only use three "expiry buckets" with 10 to 90,

F. Audrino (✉)
Fachbereich für Mathematik und Statistik, University of St. Gallen, Bodanstrasse 6, 9000 St. Gallen, Switzerland
e-mail: francesco.audrino@unisg.ch

D. Colangelo
Swiss Finance Institute, Università della Svizzera italiana, USI, Via G. Buffi 13, 6904 Lugano, Switzerland
e-mail: dominik.colangelo@lu.unisi.ch
url: http://www.people.lu.unisi.ch/colanged

90 to 180, and 180 to 270 days to expiry. With the same goal, Cont and da Fonseca (2002) apply the Karhunen-Loève decomposition, a PCA method for random surfaces. Fengler et al. (2007) combine methods from functional PCA and backfitting techniques for additive models in their dynamic semiparametric factor model (DSFM). By taking the degenerated option data structure explicitly into account, i.e. the fact that there is only a discrete set of strikes with a very small number of maturities available at each moment in time, they overcome some of the difficulties that the above-mentioned models based on PCA have encountered. They fit their functional model directly on the aggregated data, without the need to estimate IV with a non-parametric smoothing estimator on a fixed grid or to sort IV into moneyness/time to expiry buckets in order to obtain a high dimensional time series of IV classes as an approximation of the IVS. In a comparison of the one-day out-of-sample prediction error, the DSFM performs just 10 percent better on DAX option data than a simple sticky-moneyness model, where IV is taken to be constant over time at a fixed moneyness.

In contrast with the studies mentioned above, we propose a new semi-parametric model based on an additive expansion of simple fitted regression trees that does not resort to variance reduction techniques like factor analysis or PCA to forecast the dynamics of the IVS. Similarly to Fengler et al. (2007), our model takes into account the degenerated option data structure and can be easily estimated using standard boosting techniques. Moreover, in contrast with most of the previous studies, we are able to handle all options traded on the market without applying preliminary filtering or discarding any information. Any starting model can be enhanced with the help of our framework by including exogenous factors. The most relevant ones are chosen automatically by the regression trees used in our model to minimize the difference between observed and estimated implied volatilities. In this way, our methodology aims to improve the out-of-sample prediction of the IVS and is able to handle very high values of IV for both in- and out-of-the-money options.

We fit our IVS model to S&P 500 option data using different starting model specifications in the estimation procedure. The results are very promising. Starting from the model proposed by Dumas et al. (1998), our boosting algorithm reduces the out-of-sample mean square prediction error by 31 percent. When the starting model is taken to be the DSFM model, improvements range up to 3 percent. Furthermore, using a very simple regression tree as a starting model, we get a model that outperforms all alternative competitors, and consistently predicts implied volatilities up to 60 days out-of-sample with a daily averaged mean square prediction error of not more than 0.0225.

The paper is organized as follows. Section 2 presents our model and the estimation procedure used to estimate it in detail. Section 3 contains the empirical part. We discuss the included factors, the starting models, and compare the performance of the alternative models in terms of IV forecast accuracy. Section 4 concludes.

2 Model and estimation procedure

In this section we propose a new semi-parametric model for the analysis and prediction of the time-varying dynamics of the implied volatility surface. Our model can be easily estimated using a classical boosting algorithm based on regression trees.

2.1 Implied volatility surface (IVS)

The multivariate time series of interest is the time-varying implied volatility surface. Implied volatility (IV) should be considered as a mapping from time t, option's strike price K, and expiry date T. The mapping

$$\tilde{\sigma}_t^{\mathrm{IV}}: \quad (K,T) \longmapsto \tilde{\sigma}_t^{\mathrm{IV}}(K,T) \tag{1}$$

is called the implied volatility surface (IVS). Plugging the price of the underlying stock S_t, K, the risk-free interest rate r, T, and $\tilde{\sigma}_t^{\mathrm{IV}}(K,T)$ back into the well-known Black and Scholes (BS) formula leads (by definition of implied volatility) to the observed market price.[1] Thus, knowing the price of the underlying stock, the risk-free interest rate, and the IVS at time t is equivalent to knowing the market price of any option with any given contract characteristic.[2] The mapping allows us to compare two options, even ones with different characteristics. The one with the higher IV is priced relatively higher compared to the other one with lower IV.

As is usually done in the IV literature, we also describe the IVS in relative coordinates. Let $m = K/S_t$ denote the moneyness and $\tau = T - t$ the time to maturity. The IVS is then given by

$$\sigma_t^{\mathrm{IV}}: \quad (m,\tau) \longmapsto \sigma_t^{\mathrm{IV}}(m,\tau) = \tilde{\sigma}_t^{\mathrm{IV}}(m \cdot S_t, t+\tau). \tag{2}$$

2.2 Inspiration

As a starting point, we consider a non-parametric kernel-smoothing estimator for the IVS introduced by Gouriéroux et al. (1995) and Ait-Sahalia and Lo (1998). In particular, the

[1] According to the BS assumptions, this implicitly calculated volatility should be constant. It can be easily shown empirically that a cross-section of options with different strikes and expiry dates have different IV (volatility smiles/smirks, term structure of IV). Even worse, IV also changes over time.

[2] Type of option, strike, and expiry date.

least square kernel (LSK) smoothing estimator introduced by Gouriéroux et al. (1995) is defined by

$$\hat{\sigma}^{\mathrm{IV}}(m,\tau) = \arg\min_{\tilde{\sigma}} \sum_{i=1}^{n} \bigl(c_{t_i} - c^{\mathrm{BS}}(\cdot,\tilde{\sigma})\bigr)^2 \omega(m_{t_i}) \times K_1\left(\frac{m-m_{t_i}}{h_1}\right) K_2\left(\frac{\tau-\tau_{t_i}}{h_2}\right). \tag{3}$$

The observed call prices are normalized by the price of the underlying stock, $c_t = C_t/S_t$, and c^{BS} denotes the normalized price obtained using the BS formula in terms of moneyness. No inversion of the BS formula is needed, because observed market prices act as inputs. The estimate for a particular point on the IVS is given by the minimum of the loss function, which in this case is the weighted sum of least squares. K_1 and K_2 are univariate kernel functions with bandwidths of h_1 and h_2, respectively. $\omega(m)$ denotes a uniformly continuous and bounded weight function, depending on m. Fengler (2005) presents a summary of possible weight functions from the early literature on IV. Gouriéroux et al. (1994) prove that under some weak conditions, the LSK estimator $\hat{\sigma}^{\mathrm{IV}}(m_t,\tau)$ converges in probability to the true volatility of the underlying asset price process and that it belongs to the class of kernel M estimators. It is therefore shown to be both consistent and asymptotically normal.

Out-of-sample prediction is the greatest disadvantage of smoothing techniques, because kernel functions explicitly depend on observed data. To be able to obtain accurate forecasts, we suggest modifying the LSK estimator along the lines that will be discussed in the next section.

2.3 The model

In a general nonparametric model, IV is regressed on a vector of predictors $\mathbf{x}^{\mathrm{pred}}$ through unspecified functions $f_{m,\tau}$ such that

$$\sigma^{\mathrm{IV}}_{m,\tau} = f_{m,\tau}(\mathbf{x}^{\mathrm{pred}}) + \varepsilon_{m,\tau} \tag{4}$$

with $E[\varepsilon_{m,\tau}] = 0$ and $E[\varepsilon^2_{m,\tau}] < \infty$ for each $m, \tau > 0$.

The regression functions $f_{m,\tau}(\cdot)$ are implicitly defined in such a way that the expectation of a given loss function λ (which is known as *risk* in supervised learning),

$$E\bigl[\lambda(\sigma^{\mathrm{IV}}_{m,\tau}, f_{m,\tau}(\mathbf{x}^{\mathrm{pred}}))\bigr]$$

is minimized for each $m, \tau > 0$.

According to (2), the IVS changes with m, τ and also over time t, but one might include other factors and allow for put or call dependency.[3] Specifically, the predictor space is assumed to be $\mathbf{x}^{\mathrm{pred}} = (m, \tau, cp\,flag, factors)$ where *cp flag* indicates the type of the option (call or put) and the *factors* are time dependent, either directly or indirectly, through time-lagged and *forecasted* time-leading versions of themselves.[4]

The model we propose is a semi-parametric one, based on a given (parametric or nonparametric) starting model $F_0(\mathbf{x}^{\mathrm{pred}})$ that might fit the IVS quite well in certain (m,τ) areas but not necessarily everywhere. To be able to estimate the model using classical boosting algorithms, for each m, $\tau > 0$ we restrict the regression function $f_{m,\tau}$ to be a linear additive expansion of the form

$$f_{m,\tau}(\mathbf{x}^{\mathrm{pred}}) = F_0(\mathbf{x}^{\mathrm{pred}}) + \sum_{j=1}^{M} B_j(\mathbf{x}^{\mathrm{pred}}) \tag{5}$$

where each B_j denotes a general, arbitrary statistical procedure (function) called *base learner* in the machine-learning context. The possible choices of the functions B_j are restricted in the following way: B_j is assumed to belong to a pre-defined class of statistical procedures that must be *weak*, in the sense that they avoid overfitting by limiting the number of parameters involved in the estimation. Typical examples of base learners are regression trees and projection pursuit regressors. When choosing F_0 and the base learners, one needs to ensure that the IVS remains positive.

In our study, we decided to use regression trees as base learners for several reasons. Classification and Regression trees (CART) were popularized by Breiman et al. (1984), and later their appeal spread because of their simplicity and interpretability. A regression tree is a set of logical if/then conditions that creates a binary partition of the predictor space and models the response as a constant for each region. Its ability to chose automatically $L-1$ split variables (i.e. predictor variables on which the conditions are stated) in order to construct a regression tree with L end-nodes is of great importance. In our case, a very simple example of a regression tree with three end-nodes where the predictor variables are moneyness m and time to maturity τ may be given by:

$$B_j(\mathbf{x}^{\mathrm{pred}}) = \begin{cases} c_1, & \text{if } m \le d_1, \\ c_2, & \text{if } m > d_1 \text{ and } \tau \le d_2, \\ c_3, & \text{if } m > d_1 \text{ and } \tau > d_2, \end{cases}$$

where c_i, $i = 1,2,3$, are constant parameters, and d_i, $i = 1,2$, are constant thresholds for the relevant predictor variables that must be estimated from the data.

[3] According to Noh et al. (1994), there are considerable advantages in separately modeling the IVS for call and put options.

[4] Often, IV is interpreted as the market's expectation of average volatility through the life time of the option. As a consequence IV at time t is a forward-looking measure, depending on S_t and possibly other factors at time t for $t \in [t, T]$.

Previous studies showed that accurate out-of-sample predictions can be obtained using regression trees as base learners, in particular when the number of end-nodes L is kept small, i.e. $L \leq 5$. The lack of smoothness of the prediction surface obtained using regression trees is not a disadvantage: often m or τ are chosen as split variables when fitting the jth regression tree, and plotting the contribution of B_j to $f_{m,\tau}$ mainly shows that IV residuals for small τs are improved. This is in line with results from the stochastic volatility literature, where the shape of the IVS for small τs is better fitted when jumps are introduced in the dynamics of the underlying.

2.4 Estimation

We propose to use an optimization technique in function space called functional gradient descent (FGD).[5] This machine learning technique has shown its power in improving volatility forecasts in high-dimensional GARCH models (see Audrino and Bühlmann 2003). In another case, FGD-based filtered historical simulation was conducted to compute reliable out-of-sample yield curve scenarios and confidence intervals (see Audrino and Trojani 2007). Since only a finite sample of observed IV is available, the FGD estimate of $f_{m,\tau}(\cdot)$ is constructed from a constrained minimization of the average observed loss (*empirical risk*). The constraints require that $\hat{f}_{m,\tau}(\cdot)$ is an additive expansion of base learner functions as in (5). Boosting based on regression trees is a simple version of FGD, using regression trees as base learners and a quadratic loss function.

2.4.1 Empirical local criterion

Let $(m_{ti}, \tau_{ti}, \sigma_{ti}^{\mathrm{IV}})$, $i \in \{1, \ldots, L_t\}$ denote the observations of moneyness, time to maturity, and IV at day t. The daily number of observations L_t varies over time. As Fengler (2005) points out, the biggest problem in estimating the IVS comes from the degenerated design of the data. There is only a discrete set of strikes with a very small number of maturities available at each moment in time. This string structure calls for aggregation over time. This is necessary in order to obtain a region where observed location parameters form quasi a continuum, but this requires controlling for the time to expiry. Long dated options can appear every day over the aggregated sample period, whereas short dated ones soon expire and are replaced by others. In order to improve the implied volatility estimates in each iteration step, reducing the errors of short expiring options needs to be the aim. We propose an empirical local criterion that allows for learning from in-the-money (ITM), at-the money (ATM), and out-of-the-money (OTM) as well as from very short- to long-term options and that is able to control for over-fitting.

[5] A short introduction to FGD is provided in Appendix.

Our approach relies on a fixed grid in the (m, τ) domain, which should be laid over the region where forecasts of the IVS are to be calculated. Using the following indexing

$$
\begin{aligned}
&[1] := (m_{(1)}, \tau_{(1)}), \qquad [2] := (m_{(2)}, \tau_{(1)}), \ldots,\\
&[N_m] := (m_{(N_m)}, \tau_{(1)}),\\
&[N_m + 1] := (m_{(1)}, \tau_{(2)}), \qquad [N_m + 2] := (m_{(2)}, \tau_{(2)}), \ldots,\\
&[2 \cdot N_m] := (m_{(N_m)}, \tau_{(2)}),\\
&\vdots\\
&[(y-1) \cdot N_m + x] := (m_{(x)}, \tau_{(y)}),\\
&x \in \{1, \ldots, N_m\}, \quad y \in \{1, \ldots, N_\tau\}\\
&(m_{(1)} < m_{(2)} < \cdots < m_{(N_m)},\ \tau_{(1)} < \tau_{(2)} < \cdots < \tau_{(N_\tau)})
\end{aligned}
$$

a grid with grid points $\mathrm{GP} = \{[1], [2], \ldots, [N_m \cdot N_\tau]\}$ is obtained. This helps to get reasonable estimates (and forecasts), because our model is fitted via kernel weighting in such a way that the focus is set to the region of the grid.

Starting from an initial model, additive expansions in the form of regression trees with a small number of end-nodes (typically $L = 2, 3$ or 5) are used to fit the data. Similar to (3), we focus on a quadratic loss function which depends directly on implied volatilities:

$$\lambda(\sigma_t^{\mathrm{IV}}, \hat{\sigma}_t^{\mathrm{IV}}) = \left(\sigma^{\mathrm{IV}} - \hat{\sigma}^{\mathrm{IV}}\right)^2.$$

The empirical local criterion to minimize over the grid specified above is then defined by:

$$\Lambda_{\mathrm{grid}} = \sum_{t=1}^{N} \sum_{i=1}^{L_t} \sum_{[g] \in \mathrm{GP}} \left(\sigma_{ti}^{\mathrm{IV}} - \hat{\sigma}_{ti}^{\mathrm{IV}}\right)^2 w_t(i, [g]), \tag{6}$$

with weights specified by

$$
\begin{aligned}
w_t(i, [g]) &= \omega_1(m_{ti}) \cdot \omega_2(\tau_{ti})\\
&\quad \times K\left(\frac{m_{(x)} - m_{ti}}{h_1}, \frac{\tau_{(y)} - \tau_{ti}}{h_2}\right).
\end{aligned} \tag{7}
$$

In the above equation, the different quantities are defined as

$$
\begin{aligned}
&[g] = (m_{(x)}, \tau_{(y)}) \in \mathrm{GP}\\
&\quad x \in \{1, \ldots, N_m\},\ y \in \{1, \ldots, N_\tau\},\\
&K(u, v) = \frac{1}{2\pi} \cdot e^{-\frac{1}{2}(u^2 + v^2)},\\
&\omega_1(m_{ti}) = \begin{cases} 1/\pi \cdot \arctan(\alpha_1(m_{ti} - 1)) + 1/2, \\ \quad \text{if option } i \text{ is a call}, \\ 1/\pi \cdot \arctan(\alpha_1(1 - m_{ti})) + 1/2, \\ \quad \text{if option } i \text{ is a put}, \end{cases}\\
&\omega_2(\tau_{ti}) = 1/\pi \cdot \arctan(\alpha_2(1 - \tau_{ti})) + 1/2.
\end{aligned}
$$

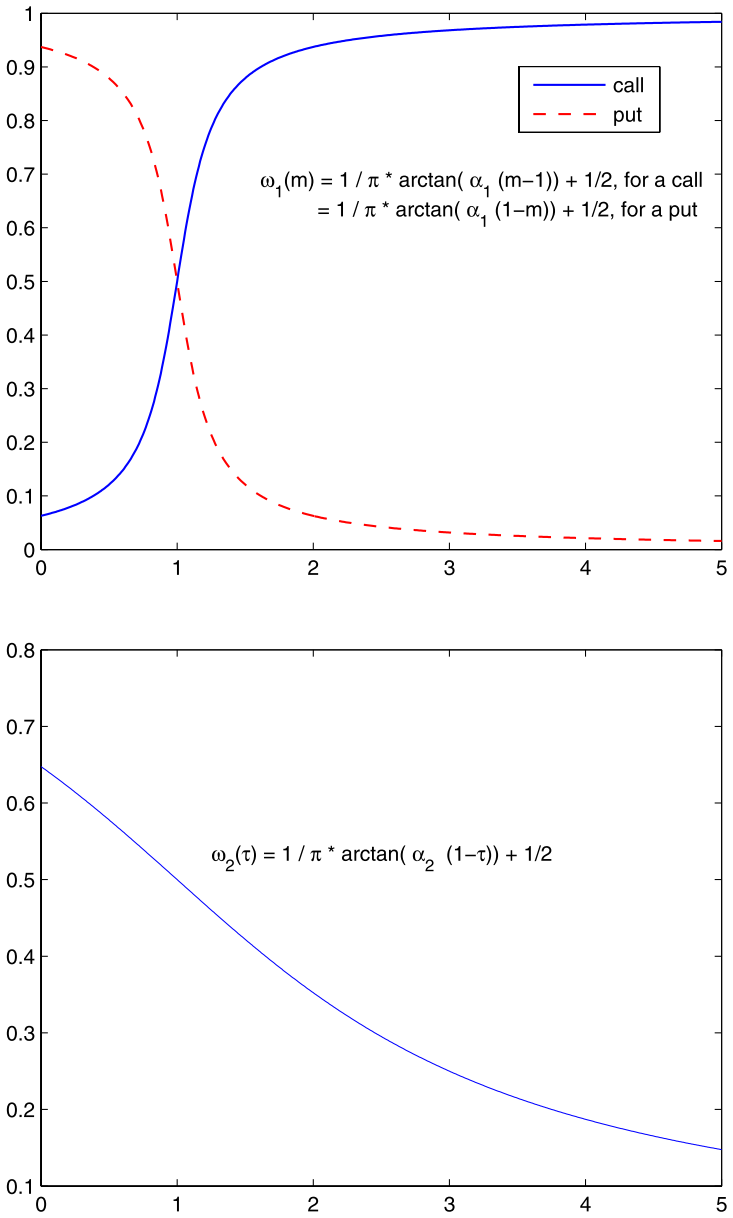

Fig. 1 Plot of the function $\omega_1(m)$ with $\alpha_1 = 5$ *above* and $\omega_2(\tau)$ with $\alpha_2 = 0.5$ *below*

The weight function (7) consists of three factors. The first is ω_1, taken from Fengler et al. (2007), with slight corrections such that OTM options have more influence than ITM options. The second one is ω_2, and depends on the time to maturity. It was chosen to reduce the influence of options which expire far in the future and to increase the importance of options that are soon due. Figure 1 shows a plot of ω_1 and ω_2.[6]

[6]From a numerical point of view, it is convenient to normalize the weight function in such a way that:

$$\sum_{i=1}^{L_t} w_t(i,[g]) = 100$$

for every $[g] \in \mathrm{GP}$ and t, because the product of three small factors can become very small.

2.4.2 *A feasible algorithm*

A cross-validation scheme is needed to avoid overfitting the model. The first 70 percent of the days in the training data are considered to be a learning sample and the remaining 30 percent a validation sample. The model is fitted on the aggregated IV observations in the learning sample only. The more additive components in the expansion there are, the smaller the error in the learning sample becomes. It tends to zero as the number of iterations increases, but this generally goes along with a worsening predictive power. The empir-

ical local criterion (6) was tailored to highlight the importance of prediction errors in the grid region. The addition of expansions is stopped when the empirical criterion takes on its minimum in the validation sample. Slow convergence is desired in order to find the optimal number of iterations M, but using the negative gradient of the loss function in (6) leads to

$$\sum_{[g]\in GP} 2(\sigma_{ti}^{\mathrm{IV}} - \hat{\sigma}_{ti}^{\mathrm{IV}}) w_t(i,[g])$$

and requires too many iterations (>500). To make computation feasible, we use the unweighted residuals in growing the additive expansion (see step 2 of Algorithm below). The learning rate can be further controlled by introducing a shrinkage factor $0 < \nu \leq 1$.

Taking all the above considerations into account, the following algorithm is proposed for estimating the IVS.

Algorithm: Tree-boosting for Implied Volatility Surfaces (treefgd)

1. Fit initial model $F_0(m,\tau,cp\,flag,factors)$ to the data $t \in \{1,\ldots,N\}$, $i \in \{1,\ldots,L_t\}$ (separate for call (c) and put (p) options):

$$\begin{aligned}\hat{\sigma}_{ti}^{\mathrm{IV},0} &= \hat{F}_0(m_{ti},\tau_{ti},cp\,flag_{ti},factors_t)\\ &= \hat{F}_0^c(m_{ti},\tau_{ti},factors_t)\mathbb{1}_{\{cp\,flag_{ti}=\text{call}\}}\\ &\quad + \hat{F}_0^p(m_{ti},\tau_{ti},factors_t)\mathbb{1}_{\{cp\,flag_{ti}=\text{put}\}},\end{aligned} \tag{8}$$

   where *cp flag* indicates the type of the option (call or put), and *factors* indicates that relevant exogenous factors may be included as additional predictor variables fitting the model.
2. For $j = 1,\ldots,M$:
   (a) For $t = 1,\ldots,N, i = 1,\ldots,L_t$ compute:

$$\begin{aligned}\text{residual}_{ti} &= \sigma_{ti}^{\mathrm{IV}} - \hat{\sigma}_{ti}^{\mathrm{IV},j-1}\\ &= \sigma^{\mathrm{IV}}(m_{ti},\tau_{ti},cp\,flag_{ti})\\ &\quad - \hat{F}_{j-1}(m_{ti},\tau_{ti},cp\,flag_{ti},factors_t).\end{aligned}$$

   (b) Fit a regression tree with L end-nodes to the residuals (separate for call (c) and put (p) options):

$$\begin{aligned}&\widehat{\text{tree}}(m,\tau,cp\,flag,factors)\\ &\quad = \widehat{\text{tree}^c}(m,\tau,factors)\mathbb{1}_{\{cp\,flag=\text{call}\}}\\ &\qquad + \widehat{\text{tree}^p}(m,\tau,factors)\mathbb{1}_{\{cp\,flag=\text{put}\}}.\end{aligned}$$

   (c) Update:

$$\begin{aligned}&\hat{F}_j(m,\tau,cp\,flag,factors)\\ &\quad = \hat{F}_{j-1}(m,\tau,cp\,flag,factors)\\ &\qquad + \nu\cdot\widehat{\text{tree}}(m,\tau,cp\,flag,factors),\end{aligned}$$

   with shrinkage factor $0 < \nu \leq 1$ small.
3. Choose $\hat{M}$ such that $\Lambda_{\text{grid}}(\hat{F}_{\hat{M}})$ is minimal over the validation sample. Use only the learning sample for fitting the models in Steps 1 and 2.

Steps 1 and 2 of the algorithm are independent of the chosen grid points. This makes estimation faster. On a standard PC, 250 iterations of step 1 and 2 are calculated within 10 minutes for a learning sample of 175 days with about 70,000 observed IV and a 14 dimensional $\mathbf{x}^{\text{pred}}$. The same calculations for a 278 dimensional $\mathbf{x}^{\text{pred}}$ require computation over night.

Evaluating $\hat{F}_j$ at $(m_{ti},\tau_{ti},cp\,flag_{ti},factors_t)$ for $t \in$ validation sample, $i \in \{1,\ldots,L_t\}$, $j \in \{1,\ldots,250\}$ and calculating $\Lambda_{\text{grid}}(\hat{F}_j)$ as in (6) is straightforward. The time needed for the cross-validation scheme depends on the size of the chosen grid and should not be underestimated. Once the optimal stopping value $\hat{M}$ has been found, $F_{\hat{M}}$ has to be estimated again, now using the whole sample and not only the first 70% of the data.

The distributed computing capability of today's standard software makes it possible to optimize parameters like ν, L and the number of time-lagged or leading factors by a simple grid search. The weight function (7) is trickier to handle: it depends on the chosen grid, kernel, bandwidths and α_1,α_2. The choice of the grid is the most important. To avoid complex adjustment procedures, we fit our model with the following default values: $\nu = 0.5$, $L = 5$, 5 time-lagged and forecasted time-leading factors, $h_1 = h_2 = 0.5$, $\alpha_1 = 5$ and $\alpha_2 = 0.5$. Even with these fixed settings, the tree-boosting algorithm is able to improve any starting model.

2.5 Keeping extremal IV in the sample

In-the-money options (ITM, $m < 1$ for call, $m > 1$ for put) are often excluded in the IVS literature. ITM options contain a liquidity premium: they have an intrinsic value which increases their costs and leaves less leverage for speculation. The costs in portfolio hedging are higher with those options. Hence, they are traded less frequently. Cont and da Fonseca (2002) claim that out-of-the-money (OTM) options contain the most information about the IVS.

Gonçalves and Guidolin (2006) apply five exclusionary criteria to filter their IVS data. They exclude thinly traded options, options that violate at least one basic no-arbitrage condition, options with fewer than six trading days to maturity or more than one year, options with moneyness smaller than 0.9 and larger than 1.1, and finally contracts with a price lower than three-eighths of a dollar. Cassesse and Guidolin (2006) investigate the pricing efficiency in a bid-ask spread and transaction cost framework. They find a frictionless data

set by dropping 51 percent of the original observations. Skiadopoulos et al. (1999) also screen the raw data. They eliminate data where the option price is less or equal to its intrinsic value, where prices are less than 10 cents and where $\tau < 10$ days. They construct smiles using OTM puts for low strikes and OTM calls for high strikes only, relying on the put-call parity. They also set a vega cutoff: options with vega less than 8 are dropped from the sample. In this way, only 40 percent of the observations for calls and 70 percent for puts are retained in the sample.

It is known that ITM calls and OTM puts are traded at higher prices compared with corresponding at-the-money (ATM) options in general. Especially when the expiry date nears, observed prices and IV react violently: see Hentschel (2003). Options with expiry further in the future have more vega and less gamma than those expiring sooner. Low IV precision close to expiration is inherent to options. These options have very little vega: thus, inverting the pricing formula gives a big change of volatility for a tiny price change. This is usually amplified by the wider bid-ask spreads for ITM options close to maturity. The usual trick is to focus on ATM and OTM options. Excluding the strangely behaving options from the sample helps any model to perform better, but this still completely neglects the reality of having higher IV values. Regardless of what causes very high IV, removing them leads to a loss of information that may be important for prediction.

The model proposed in this study should be able to handle all options, including ITM, ATM and OTM as well as options shortly before the expiry date. A weight function guarantees that extremal IV can be kept in the sample while controlling their influence.

3 Application

3.1 Data

We use Option Metrics' Ivy database and analyze implied volatilities of call and put options with different strikes and maturities on the S&P 500 index for this project. The data range from 4 January 1996 to 29 August 2003 and consist of 777,887 observations on 1,928 days. There are approximately 400 observations of IV per day on average over the whole sample. They appear in a so-called string structure, which means that usually only options with a few standardized times to maturity are traded, but for each τ there are many different strikes. The observations of the next day contain yesterday's options where the times to maturity are net of 1/365 and the moneyness of each option has changed because the underlying has as well. Table 1 shows summarized statistics of the option data under investigation.

We fit our model to five different sub-samples 250 days in length. They are chosen to occur before five special days of interest, where—from today's perspective—a more or less heavy structural break is expected to happen. On 7 August 1998 two bomb attacks on US embassies in Africa occurred. The impeachment trial of President Clinton was opened in the Senate on 7 January 1999. The first date for which we have IVS observations in our sample after President Bush's oath of office and the disruption caused by the unclear outcome in the 2000 presidential election is 22 January 2001. 17 September 2001 is six days after the 11 September terrorist attacks, and 20 March 2003 marks the official beginning of the military campaign against Iraq. Table 2 summarizes the explicit sub-samples.

The sub-samples end 4 to 25 days before the special days of interest. Our goal is to attain 60 days out-of-sample predictions of the IVS, such that we can compare the observed IV with the predicted one, before and after a supposed structural break. The accuracy is measured by evaluating our fitted model at exactly the same (m, τ) locations as observed, and finally by calculating the sum of squared residuals (SSR) and the value of the empirical criterion. We use a linearly spaced 15×15 grid with values from $m = 0.2$ to 2 and from $\tau = \frac{1}{365}$ to 3 for the empirical local criterion defined in Sect. 2.4.1.

3.2 Forecasting accuracy: the setting

We specify in this section the general setting we use for deriving the out-of-sample results of the next section. We introduce performance measures, predictor variables used in the estimation and for forecasting, and the competing approaches under investigation.

3.2.1 Out-of-sample (OS) performance measures

We compare the OS performance by evaluating the IVS models at the same $(m, \tau, cp\,flag)$ locations as the ones of recorded OS IV entries in our database. We measure goodness-of-fit of the different competitors with respect to the daily and overall averaged mean square forecast errors:

$$\text{daily SSR}_t = \frac{1}{L_t} \sum_{i=1}^{L_t} (\sigma_{ti}^{\text{IV}} - \hat{\sigma}^{\text{IV}})^2, \tag{9}$$

$$\text{overall SSR} = \frac{1}{N} \sum_{t=1}^{N} \text{daily SSR}_t. \tag{10}$$

We also consider as additional performance measures the daily and the overall averaged empirical criteria, daily EC_t and overall EC, defined as

$$\text{daily EC}_t = \frac{1}{L_t} \sum_{i=1}^{L_t} \sum_{[g] \in \text{GP}} \left(\sigma_{ti}^{\text{IV}} - \hat{\sigma}_{ti}^{\text{IV}}\right)^2 w_t(i, [g]), \tag{11}$$

Table 1 Descriptive statistics of implied volatilities of options on the S&P 500 index, from 4 January 1996 to 29 August 2003, 777,887 observations on 1,928 days. Mean and standard deviation (Std) are in percentage. Moneyness m is defined as strike price divided by the closing price of the underlying asset. Maturity is measured in calendar days. Moneyness categories for call options are defined in the following way: deep in-the-money (DITM) equals $m \leq 0.8$, in-the-money (ITM) $0.8 < m \leq 0.94$, at-the-money (ATM) $0.94 < m \leq 1.04$, out-of-the-money (OTM) $1.04 < m \leq 1.2$, deep out-of-the-money (DOTM) $1.2 < m$. For put options, the reverse order has to be considered

		Maturity in days					
		Less than 90		90 to 240		More than 240	
		Call	Put	Call	Put	Call	Put
DITM	Mean	76.75	71.96	35.33	24.49	29.92	19.73
	Std	61.99	62.21	9.34	7.77	6.57	3.13
	Observations	15,176	4,667	16,302	2,510	25,227	4,694
ITM	Mean	35.50	27.53	25.56	20.12	24.79	21.02
	Std	21.75	18.87	5.08	3.53	4.59	3.33
	Observations	41,347	20,673	22,489	12,782	30,309	22,011
ATM	Mean	22.10	22.15	21.06	21.17	22.14	22.36
	Std	7.09	6.34	4.33	4.36	4.33	3.97
	Observations	61,768	61,707	19,771	19,910	25,368	25,443
OTM	Mean	21.83	32.06	18.64	25.64	19.86	25.13
	Std	10.27	10.97	3.50	4.96	3.65	4.33
	Observations	44,562	49,061	21,285	23,037	31,920	30,544
DOTM	Mean	43.07	49.25	20.65	34.95	18.38	30.25
	Std	34.72	19.41	5.67	7.33	2.90	5.65
	Observations	19,642	24,037	19,521	22,724	29,906	29,494

Table 2 In-sample (IS) periods of length 250 days and out-of-sample (OS) periods of length 60 days. The special days of interest in the OS periods for the 5 different sub-samples under investigation

Sub-sample	Training data In-sample period	Forecasting Out-of-sample period	Special day of interest	
1	21 Jul. 1997–16 Jul. 1998	17 Jul. 1998–09 Oct. 1998	07 Aug. 1998	16th OS day
2	06 Jan. 1998–31 Dec. 1998	04 Jan. 1999–30 Mar. 1999	07 Jan. 1999	4th OS day
3	20 Dec. 1999–13 Dec. 2000	14 Dec. 2000–13 Mar. 2001	22 Jan. 2001	25th OS day
4	15 Aug. 2000–10 Aug. 2001	13 Aug. 2001–09 Nov. 2001	17 Sep. 2001	21st OS day
5	18 Mar. 2002–13 Mar. 2003	14 Mar. 2003–09 Jun. 2003	20 Mar. 2003	5th OS day

$$\text{overall EC} = \frac{1}{N}\sum_{t=1}^{N} \text{daily EC}_t. \tag{12}$$

3.2.2 Predictor variables and factors

Unlike in the theoretical Black and Scholes (BS) framework, option prices might depend on more than their contract characteristics, the price of the underlying and the risk-free interest rate. It is not appropriate to use the 3-month US Treasury-bill rate, fixed at the day the option is issued, as the constant r in the BS formula. Allowing for stochastic interest rates or including proxies for the term structure of interest rates is common nowadays.

Our model allows for an arbitrary number of exogenous factors, directly or indirectly time-dependent. IV can be seen as a predictor of future volatility, so time-lagged as well as *forecasted* time-leading factors should be included in the predictors.[7] Time-leading factors are predicted *without using any future information* in the following way: For simplicity, we model the log returns of each factor time series in the case of an asset price, the first differences in the case

[7]Note that including forecasts for time-leading factors in the predictor space of a regression tree does not cause problems. Regression trees can handle missing values and a huge number of predictors. Only the most relevant ones are automatically chosen as split variables.

of interest rates, as a univariate ARMA(1,1)–GARCH(1,1) process and use its mean forecast as a prediction of the unknown, future time-leading predictor variable in the regression trees. We use standard filtered historical simulation[8] to reduce the forecast errors. Parameters of the ARMA(1,1)–GARCH(1,1) process are estimated on a (rolling) time-window of the past 500 observations. Then the predictions for future returns (or differences) are based on the estimated parameters.

Possible exogenous factors to consider are implied asset prices (Garcia et al. 2003), the bid-ask spread, net buying pressure (Bollen and Whaley 2004), trading volume, other stock or index returns, and interest rates. Another strategy would be to include IV time series for options with fixed specifications as predictors. If the goal is for example to model the IVS in the neighborhood of 30 days ATM, one could make use of the explanatory capabilities of competing option pricing methods by calculating IV time series for a call and a put with this specification.

In this study, we limit ourselves to three different sets of predictor variables (pv set). We define them such that

$$\text{pv set } 1 \subset \text{pv set } 2 \subset \text{pv set } 3.$$

In addition to the original factors, we include five time-lagged and five forecasted time-leading versions of each.

pv set 1 consists of the closing prices of the underlying.

pv set 2 additionally includes the {3,6}-month and {1, 3, 5, 10}-year Treasury Constant Maturity Rates as representatives of the interest rates' term structure. The time series are available on the St. Lous Fed Homepage in the FRED database and their series IDs are labeled DGS.

pv set 3 comprises another 18 factors. The factors are actually option prices for different characteristics (call and put, $m \in \{0.8, 1, 1.2\}$), and $\tau \in \{30, 60, 90 \text{ days}\}$, obtained with the Heston Nandi GARCH (HNG) model. We choose it because HNG is a discrete time series model based on an asymmetric GARCH process for the spot asset price with a closed-form solution for option prices (Heston and Nandi 2000). Thus, it is simple and can be quickly implemented computationally. In the continuous-time limit, it contains Heston (1993) stochastic model which is still very popular in practice. Including the 18 HNG factors acts as guidance for our model. Intuitively, if a regression tree chooses a HNG factor as split variable, it implies realigning the minimization of the sum of squared differences between observed and estimated IV to option prices calculated with the analytical Heston-Nandi pricing formula.[9]

3.2.3 Competitors and starting models

We compare the accuracy of the IVS predictions from our model with several alternative approaches. Each alternative competitor is also used as starting model in the boosting procedure explained in Sect. 2.4.2. In particular, we compare results from our model with those from the following competitors.

Regression tree (regtree) We fit a regression tree with 10 end-nodes on all observed in-sample calls and puts separately. We use Gini's diversity index as a criterion for choosing a split. Positivity of the IVS is guaranteed since the model depends on the aggregated observed positive IV.

Ad hoc BS model (adhocbs) Dumas et al. (1998) perform a goodness-of-fit test for several functions of quadratic form in a deterministic volatility framework. They find that the best parametrization is given by

$$\sigma(K, \tau) = \max(0.01, a_0 + a_1 K + a_2 K^2 + a_3 \tau + a_4 K \tau).$$

Since we use relative coordinates $m = K/S_t$ and $\tau = T - t$, we fit

$$\sigma_{ti}^{\text{IV}} = a_{t0} + a_{t1} m_{ti} S_t + a_{t2} (m_{ti} S_t)^2 + a_{t3} \tau_{ti} + a_{t4} m_{ti} S_t \tau_{ti} + \epsilon_{ti} \tag{13}$$

by least square, using observations on day t. In case of a negatively estimated IV, values are also set to 0.01. The coefficients estimated on a reference day $\tilde{t} < t_f$ are used to obtain the IVS on a future date t_f.

Sticky moneyness model (stickym) The sticky moneyness model is a 'naïve trader model.' It assumes that IV is constant at fixed moneyness. The term structure of the IVS at a reference day $\tilde{t} < t_f$ is used to interpolate the IV on a future date t_f.

Bayesian vector autoregression (bvar) We evaluate the IVS on a linearly spaced 10×10 grid with values from $m = 0.2$ to 2 and from $\tau = 1/365$ to 3 for each day in the sample using a Nadaraya-Watson estimator with a Gaussian product kernel and step width chosen to minimize the mean squared error at each grid point. We obtain a 100 dimensional time series over the whole in-sample period on which, for simplicity, we fit a Bayesian vector autoregression of order 2. The Econometrics Toolbox by James P. LeSage in Matlab contains functions to estimate, evaluate, and forecast Bayesian vector autoregressive models.

[8] Filtered historical simulation is a particular technique based on the bootstrap of the estimated residuals. See Barone-Adesi and Giannopoulos (1998) and Barone-Adesi and Vosper (1999) for a detailed description of filtered historical simulation.

[9] The choice of the HNG model is not restrictive: other models can be used to obtain reliable and informative factors.

Table 3 60 days out-of-sample performance in terms of overall averaged mean square forecast errors (10) for different predictor variable sets (Sect. 3.2.2) and sub-samples (Table 2). The alternative models considered are a regression tree (regtree), the ad-hoc Black-Scholes model (adhocbs), and the dynamic semiparametric factor model (dsfm) proposed by Fengler et al. (2007). The models are considered alone (first three columns) and in connection with our boosting procedure based on regression trees (treefgd, second three columns)

	Sub-sample	Only starting model			Improved with treefgd		
		regtree	adhocbs	dsfm	regtree	adhocbs	dsfm
pv set 1	1	0.0188	0.0223	0.0280	0.0166	0.0267	0.0260
	2	0.0118	0.0661	0.4993	0.0032	0.0191	0.4979
	3	0.0112	0.0264	0.1325	0.0022	0.0115	0.1304
	4	0.0225	0.0755	15.4498	0.0061	0.0252	15.4485
	5	0.0117	0.1362	0.0267	0.0066	0.0910	0.0240
pv set 2	1	0.0188	0.0223	0.0280	0.0149	0.0234	0.0271
	2	0.0120	0.0661	0.4993	0.0027	0.0173	0.4987
	3	0.0112	0.0264	0.1325	0.0026	0.0378	0.1314
	4	0.0225	0.0755	15.4498	0.0066	0.0259	15.4492
	5	0.0117	0.1362	0.0267	0.0055	0.0968	0.0243
pv set 3	1	0.0188	0.0223	0.0280	0.0149	0.0279	0.0275
	2	0.0120	0.0661	0.4993	0.0028	0.0260	0.4987
	3	0.0112	0.0264	0.1325	0.0028	0.0213	0.1314
	4	0.0225	0.0755	15.4498	0.0101	0.0376	15.4492
	5	0.0117	0.1362	0.0267	0.0045	0.0968	0.0244

Dynamic semiparametric factor model (DSFM) In the DSFM of Fengler et al. (2007), the smooth functions are multiplied by latent factor loadings. It is accordingly more difficult to fit a DSFM to data. We choose 4 smooth basis functions, each a linear combination of cubic B-splines on a uniformly spaced knot sequence of length 6 between the minimum and maximum values of time-aggregated observation in both dimensions of the (m, τ) plane. We apply the DSFM algorithm directly on IV data instead of on the log IV. We find that this improves the out-of-sample predictions; otherwise errors in predicting the latent factor loadings are raised to a higher power. We set eventual negative IV to 0.01.

Our choice of the tuning parameters in the different alternative models described above is driven by a trade-off between computational feasibility and optimization. Given some restriction on the maximal number of parameters, we derived the optimal tuning parameters by minimizing the expected square prediction error approximated by using the same cross-validation scheme adopted in the estimation procedure introduced in Sect. 2.4.2.[10]

[10]Nevertheless, the different competing approaches are also used as starting models in our estimation procedure and therefore only need to provide a first rough approximation of the true IVS dynamics.

3.3 Out-of-sample results

Result of our out-of-sample investigations are summarized in Table 3. It shows the OS averaged mean square forecast error introduced in (10) over the whole OS period of 60 days for the different approaches under investigation alone and taken as starting models in our estimation procedure. Two models, stickym and bvar, are not listed. The former is omitted because the OS errors are still quite large, even after using a filtered sample to fit the model. The latter is excluded because of instability, since the forecast for the Bayesian vector autoregression yields unexpectedly high predicted values after maximal 10 OS days.

A regression tree as starting model (regtree) in combination with our boosting algorithm based on regression trees (treefgd) beats all other models that we have considered. This starting model partitions the $(m, \tau, cp\ flag)$ domain into regions with 10 different IV levels each for calls and puts. We obtain two piecewise constant IVS, one for calls and one for puts. Each of them captures to a certain extent an average IVS over time. The additive expansions in the boosting algorithm are regression trees as well, but with larger predictor spaces, including time series of factors. Our suggested cross-validation strategy works very well with the simple starting model regtree. We obtain a precise dynamical IVS model that does not overfit the data.

As an illustration, let us examine the out-of-sample predictions from that model for the 23 February 1999, the 35th

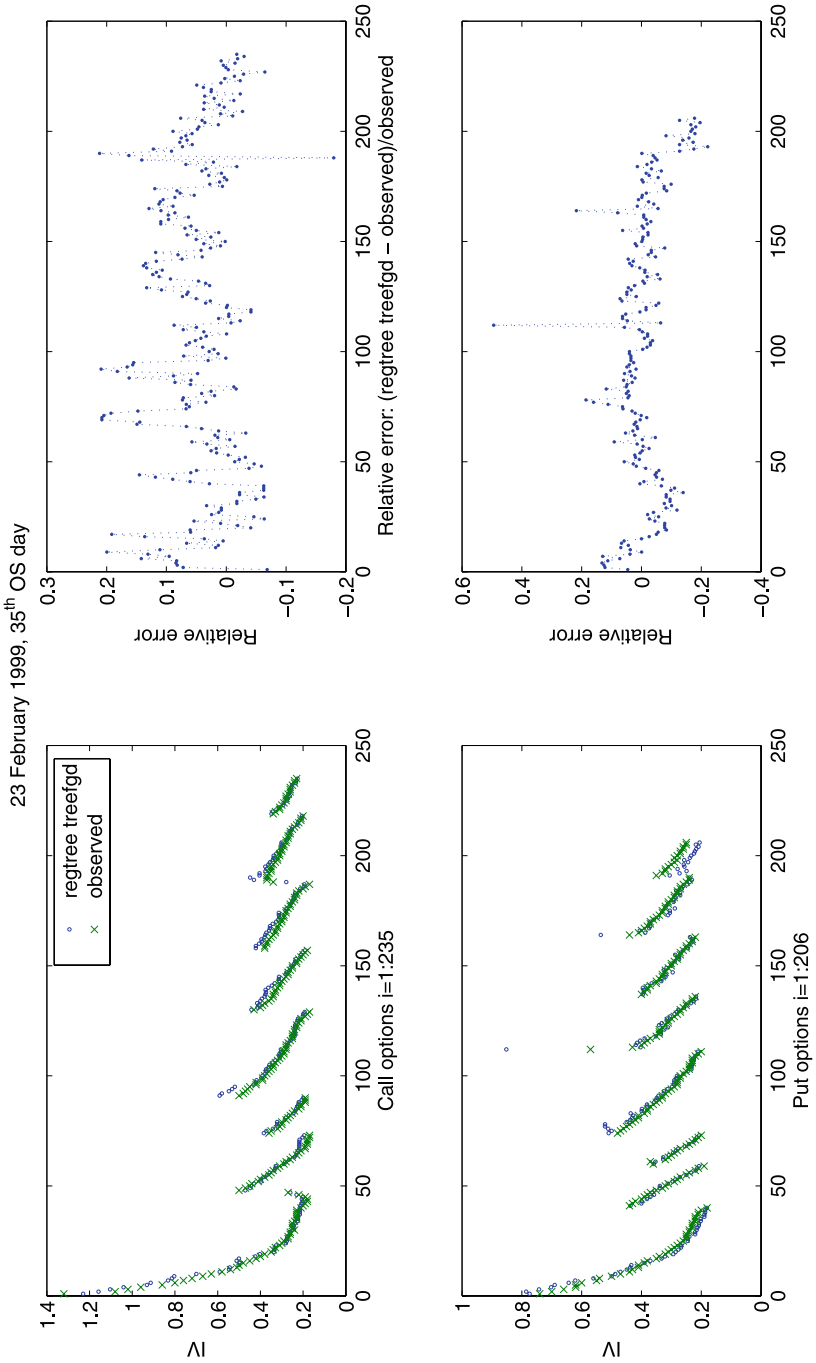

Fig. 2 35th Out-of-sample day predictions from our model with a regression tree as starting model, improved with the boosting algorithm described in Sect. 2.4.2. The predicted values are compared to the effectively observed IV on that day. Observations are sorted first by τ, then by m in ascending order. The shape of each IV string is retained

Table 4 Summary of automatically chosen split variables. Each row shows the composition of chosen split variables when applying the boosting procedure based on the regression tree algorithm proposed in Sect. 2.4.2 on all five sub-samples. We vary starting models (Sect. 3.2.3) and predictor variable sets (Sect. 3.2.2). m and τ are location parameters, close is the option's underlying closing price. DGS are treasury-constant maturity rates with different maturities. HNG stands for option prices calculated according to the Heston Nandi GARCH model. Time-lagged and *forecasted* time-leading versions of close, DGS, and HNG are included as predictor variables; past = $\{t-5,\ldots,t-1\}$, contemporaneous = t, and future = $\{t+1,\ldots,t+5\}$

		Split variables									
		$\hat{M}_{\text{total}}$	# splits	m	τ	close	DGS	HNG	past	contemp	future
pv set 1	regtree	1,066	8,528	39.8%	29.6%	30.6%			1,146	269	1,191
	adhocbs	228	1,824	46.4%	22.8%	30.8%			262	49	251
	stickym	10	80	36.3%	53.7%	10.0%			2	4	2
	bvar	430	3,440	35.6%	29.5%	34.9%			496	138	568
	dsfm	282	2,256	20.8%	35.6%	43.6%			459	87	437
pv set 2	regtree	1,185	9,480	42.0%	30.7%	9.2%	18.1%		1,173	289	1,124
	adhocbs	152	1,216	43.9%	22.3%	7.9%	25.9%		199	46	166
	stickym	21	168	29.8%	37.4%	1.8%	31.0%		7	3	45
	bvar	446	3,568	36.2%	28.4%	12.2%	23.2%		614	105	544
	dsfm	11	88	36.4%	43.2%	3.4%	17.0%		10	1	7
pv set 3	regtree	890	7,120	42.3%	31.1%	0.8%	6.6%	19.1%	841	243	806
	adhocbs	182	1,456	43.7%	19.3%	0.8%	9.3%	26.9%	235	53	250
	stickym	13	104	35.6%	37.5%	1.0%	1.9%	24.0%	2	20	6
	bvar	497	3,976	38.8%	25.9%	1.0%	9.6%	24.7%	658	127	619
	dsfm	11	88	31.8%	47.7%	1.2%	10.2%	9.1%	10	2	6
	Total	5,424	43,392	39.4%	29.3%	15.9%	9.1%	6.4%	6,114	1,436	6,022

OS day of sub-sample 2. There are 235 calls and 206 puts in our database for that day. Figure 2 shows how close the IV predictions are to the effectively observed IV values. Put in numbers, daily SSR equals 0.0010 and daily EC 0.0002.

Our estimation procedure is able in all cases to improve the predictive accuracy of the starting model considered. In particular, the boosting algorithm reduces the overall averaged SSR by 58, 31 and 3 percent for the starting models regtree, adhocbs and dsfm on average over 60 OS days and all combinations of sub-samples and predictor variable sets in Table 3.

Let us finish the discussion of the OS results with some other comments on the alternative approaches considered. First, the ad hoc BS model (adhocbs) works most of the time quite well when its parameters are estimated on a filtered sample.[11] By contrast, it produces IV estimates and forecasts that are not accurate when considering the sample data on its entirety. The relative high OS prediction errors can be traced back to a substantial number of outliers.

Second, the sticky moneyness model (stickym) is a simple interpolation scheme that produces good IS results, but it is inappropriate for OS predictions. Our cross-validation scheme suggests an optimal value of only three linear additive expansions on average. At best, our tree-boosting algorithm improves OS predictions slightly for the first 30 OS days. Over the whole period of 60 OS days, the OS overall averaged sum of squared residuals remains so high that it does not make worthwhile to report the results in Table 3.

Third, the Bayesian vector autoregressive model (bvar) catches the IVS dynamic reasonably well in-sample. Unfortunately, bvar does not produce stable OS predictions. After maximum 10 OS days, OS prediction errors increase exorbitantly. Our tree-boosting algorithm is not able to stabilize the function out-of-sample. Fourth, the dynamic semiparametric factor model (dsfm) fits the IVS very well in-sample. Predicting the latent factors over the whole OS period of 60 days is very inaccurate, especially in the presence of a possible structural break.

3.4 Relative importance of predictor variables

In the same fashion as our interpretation of boosting algorithms, here we address the question of the relevance of the different predictors in our real data application. Performing cross-validation on all 5 sub-samples leads to a total number

[11] When filtering the data in the spirit of Gonçalves and Guidolin (2006), 60 to 70 percent of the IV data in the sub-samples are excluded.

of optimal stopping values $\hat{M}_{\text{total}} = \sum_{i=1}^{5} \hat{M}_i$. Each additive expansion consists of two regression trees with $L = 5$ end-nodes, one for the calls and one for the puts. Hence, the number of total split variables is $\hat{M}_{\text{total}} \cdot (5-1) \cdot 2$. Table 4 summarizes all of the information about the predictor variables that are chosen in the boosting procedure.

Regardless which starting model or predictor variable set is used, location parameters m and τ are chosen about 70 percent of the time. Knowing the cut values[12] enables us to see on which regions the base learners improve the starting model. The cut values for m lie between 0.4 and 1.5, uniformly distributed over this interval. Only 5 percent of the cut values are greater than 1.5, the mean is 0.9067, and the maximum 2.4007. The distribution of cut values for split variable τ is concentrated around small values. 25 percent are smaller than 0.0164 (6 days), 50 percent are smaller than 0.0466 (17 days), and the average is 0.2112 (77 days).

In only 30 percent of the time, regression trees select time-lagged/forecasted time-leading factors as split variables. Including forecasts of time-leading factors in the predictor variable sets turns out to be as important as including time-lagged factors: both are chosen about the same number of times. Adding ever more exogenous factors reduces the OS errors, but it is computationally expensive and the gain in precision is small.

3.5 A robustness check

Recently, Battalio and Schultz (2006) discussed problems related to the use of the Option Metric's Ivy database for academic studies when arbitrage violations must be taken into account. The problems are mainly due to the non-synchronicity of the prices stored in the database: in many cases, time stamps of the options differ from time stamps of the underlying. To verify that the forecasting results discussed in the previous sections are not a consequence of this non-synchronicity issue, we perform a small robustness check. We consider the first sub-sample, explicitly given in Table 2. We construct a new data set of option prices. The underlying S&P 500 index levels are implicitly obtained from the original reported option prices, similar to the procedure proposed by Manaster and Rendleman (1982). We use non-linear least-square estimates of index levels and dividend yields that minimize the sum of the squared differences between the Black and Scholes (BS) option prices and the option prices reported in the Option Metrics database over each day. Then, we recompute BS implied volatilities and moneyness, using the implied S&P 500 levels and implied dividend yields. This new data set of option IVs is not dependent on underlying prices that are asynchronous with respect to the reported closing option prices.

There is no qualitative difference in the results. Comparing the accuracy of the OS predictions obtained using the new database with those obtained using the Option Metrics database, we observe only small changes. In particular, our boosting procedure of using a regression tree as starting model outperforms all other competitors. The overall averaged mean square forecast error is 0.0101. Using the Option Metrics database yielded 0.0149. Stickym and adhocbs produce slightly better OS results with the filtered,[13] implied database than the filtered Option Metrics database. The overall averaged mean square forecast errors are now 0.0502 and 0.0170 for stickym and for adhocbs, respectively. Previous results were 0.0608 and 0.0223. Qualitatively identical results are found when using the averaged overall EC. Results for the other alternative approaches do not change significantly.

4 Conclusions

We proposed a new model to estimate and, in particular, predict implied volatility surfaces. Our approach relies on a starting model that is improved by semi-parametric additive expansion of simple fitted functions using regression trees for the dynamics of implied volatilities. A modified version of classical boosting procedures can handle very high dimensional predictor variable sets. Consequently, there is no need for variance reduction or other excluding data techniques to fit the model to real data, avoiding the possibility of a dangerous information loss.

We tested the predictive potential of various IVS models on a huge data set of S&P 500, options collecting strong empirical evidence that our method improves the performance of any reasonable starting model in forecasting short- and middle-term future implied volatilities (up to 60 days), and even under possible structural breaks in the time series. Similar results are also obtained when fitting the models to the data of a volatile stock option. The model completely based on regression trees (i.e. regression tree as starting model and regression trees as base learners) turns out to be the best performing model and proves to be a powerful tool in forecasting IVS dynamics.

Acknowledgements We would like to thank Loriano Mancini, Wolfgang Härdle, Szymon Borak, Julien Gosme, and two anonymous referees for their useful comments and discussions.

[12] The CART algorithm of Breiman et al. (1984) provides us with four split variables and cut values in order to obtain a base learner that partitions the predictor space into five areas.

[13] The filtered option data contain only those options from the new data set with moneyness between 0.9 and 1.1 and time to maturity between six days and one year.

Appendix: Functional gradient descent method

Gradient descent methods are an iterative way of finding a minimum of a function f of several real-valued variables. The negative gradient $g_i = -\nabla f(P_i)$ is the direction of the steepest descent at the point P_i. In the line search step, we find $\lambda_i \in \mathbb{R}$, such that $P_{i+1} = P_i + \lambda_i g_i$ is the lowest point along this path. Iterating those two steps leads to a sequence of points which converges to the minimum of f. The drawback of this method is that it converges slowly for functions that have a long, narrow valley. A better choice for the direction in this case would be the conjugate gradient.

Applying the steepest descent method in a function space $\mathcal{F} = \{f | f : \mathbb{R}^d \to \mathbb{R}\}$ leads, as the name indicates, to the FGD technique. Based on data (Y_i, X_i), $i = 1, \ldots, n$, an estimation of a function $F \in \mathcal{F}$ which minimizes an expected loss function $\mathbb{E}[\lambda(Y, F(X))]$, where $\lambda : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ is developed.

The FGD estimate of $F(\cdot)$ is found by minimizing Λ, the empirical risk, defined as:

$$\Lambda(F)(x_1, \ldots, x_n, y_1, \ldots, y_n) = \frac{1}{n} \sum_{i=1}^{n} \lambda(y_i, F(x_i)). \tag{14}$$

Starting from an initial function $\hat{F}$, the steepest descent direction would be given by the negative functional derivative $-d\Lambda(\hat{F})$. Due to smoothness and regularization constraints on the minimizer of $\Lambda(\hat{F})$, we must restrict the search to finding a function $\hat{f}$ which is in the linear span of a class of simple base learners $\mathcal{S}$ and close to $-d\Lambda(\hat{F})$ in the sense of a functional metric. This is equivalent to fitting the base learner $h(x, \theta) \in \mathcal{S}$ to the negative gradient vectors:

$$U_i = -\left.\frac{\partial \lambda(Y_i, Z)}{\partial Z}\right|_{Z=\hat{F}(X_i)}, \quad i = 1, \ldots, n. \tag{15}$$

The minimal function $F \in \mathcal{F}$ is approximated in an additive way with simple functions $\hat{f}_m(\cdot) = h(\cdot, \hat{\theta}_{U,X}) \in \mathcal{S}$:

$$\hat{F}_M(\cdot) = \sum_{m=0}^{M} \hat{w}_m \hat{f}_m(\cdot), \tag{16}$$

where the $\hat{w}_m$s are obtained in a line search step as in the previous procedure.

FGD is a derivative of boosting and bagging (cf. Friedman et al. 2000; Friedman 2001). Audrino and Bühlmann (2003) already successfully applied FGD to estimate volatility in high-dimensional GARCH models, and Audrino et al. (2005) used it to model interest rates.

References

Ait-Sahalia, Y., Lo, A.: Nonparametric estimation of state-price densities implicit in financial asset prices. J. Financ. **53**(2), 499–547 (1998)

Audrino, F., Bühlmann, P.: Volatility estimation with functional gradient descent for very high-dimensional financial time series. J. Comput. Financ. **6**(3), 1–26 (2003)

Audrino, F., Trojani, F.: Accurate short-term yield curve forecasting using functional gradient descent. J. Financ. Econom. **5**(4), 591–623 (2007)

Audrino, F., Barone-Adesi, G., Mira, A.: The stability of factor models of interest rates. J. Financ. Econom. **3**(3), 422–441 (2005)

Barone-Adesi, G.B.F., Giannopoulos, K.: Don't look back. Risk **11**, 100–104 (1998)

Barone-Adesi, G.G.K., Vosper, L.: Var without correlations for portfolio of derivative securities. J. Futures Mark. **19**, 583–602 (1999)

Battalio, R., Schultz, P.: Options and the bubble. J. Financ. **61**(5), 2071–2102 (2006)

Bollen, N.P.B., Whaley, R.E.: Does net buying pressure affect the shape of implied volatility functions? J. Financ. **59**(2), 711–753 (2004)

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A.: Classification and Regression Trees. Chapman & Hall/CRC Press, London/Boca Raton (1984)

Cassesse, G., Guidolin, M.: Modelling the MIB30 implied volatility surface. Does market efficiency matter? Int. Rev. Financ. Analy. **15**(2), 145–178 (2006)

Cont, R., da Fonseca, J.: Dynamics of implied volatility surfaces. Quantitative Financ. **2**(1), 45–60 (2002)

Dumas, B., Fleming, J., Whaley, R.E.: Implied volatility functions: empirical tests. J. Financ. **53**(6), 2059–2106 (1998)

Fengler, M.R.: Semiparametric Modeling of Implied Volatility. Springer, Berlin (2005)

Fengler, M.R., Härdle, W.K., Mammen, E.: A semiparametric factor model for implied volatility surface dynamics. J. Financ. Econom. **5**(2), 189–218 (2007)

Friedman, J.: Greedy function approximation: a gradient boosting machine. Ann. Stat. **29**(5), 1189–1232 (2001)

Friedman, J., Hastie, T., Tibshirani, R.: Additive logistic regression: a statistical view of boosting. Ann. Stat. **28**(2), 337–407 (2000)

Garcia, R., Luger, R., Renault, E.: Pricing and hedging options with implied asset prices and volatilities. Working Paper, CIRANO, CIREQ and Université de Montréal (2003)

Gonçalves, S., Guidolin, M.: Predictable dynamics in the S&P 500 index options implied volatility surface. J. Bus. **79**(3), 1591–1635 (2006)

Gouriéroux, C., Monfort, A., Tenreiro, C.: Nonparametric diagnostics for structural models. Document de travail 9405, CREST, Paris (1994)

Gouriéroux, C., Monfort, A., Tenreiro, C.: Kernel M-estimators and functional residual plots. Document de travail 9546, CREST, Paris (1995)

Hentschel, L.: Errors in implied volatility estimation. J. Financ. Quant. Anal. **38**(4), 779–810 (2003)

Heston, S.L.: A closed-form solution for options with stochastic volatility with applications to bond and currency options. Rev. Financ. Stud. **6**(2), 327–343 (1993)

Heston, S.L., Nandi, S.: A closed-form GARCH option valuation model. Rev. Financ. Stud. **13**(3), 585–625 (2000)

Manaster, S., Rendleman, R.: Option prices as predictors of equilibrium stock prices. J. Financ. **37**, 1043–1057 (1982)

Noh, J., Engle, R.F., Kane, A.: Forecasting volatility and option prices of the S&P 500 index. J. Deriv. **2**, 17–30 (1994)

Poon, S.-H., Granger, C.W.J.: Forecasting volatility in financial markets: a review. J. Econ. Lit. **41**(2), 478–539 (2003)

Rosenberg, J.: Implied volatility functions: a reprise. J. Deriv. **7**, 51–64 (2000)

Shimko, D.: Bounds of probability. Risk **6**(4), 33–37 (1993)

Skiadopoulos, G.S., Hodges, S.D., Clewlow, L.: The dynamics of the S&P 500 implied volatility surface. Rev. Deriv. Res. **3**(3), 263–282 (1999)