Neural Tangent Kernel in Implied Volatility Forecasting: A Nonlinear Functional Autoregression Approach

Chen, Ying and Grith, Maria and Lai, Hannah L. H.

National University of Singapore, Erasmus University Rotterdam, National University of Singapore

24 October 2023

Neural Tangent Kernel in Implied Volatility Forecasting: A Nonlinear Functional Autoregression Approach

Ying Chen* Maria Grith† Hannah L. H. Lai‡

Abstract

Implied volatility (IV) forecasting is inherently challenging due to its high dimensionality across various moneyness and maturity, and nonlinearity in both spatial and temporal aspects. We utilize implied volatility surfaces (IVS) to represent comprehensive spatial dependence and model the nonlinear temporal dependencies within a series of IVS. Leveraging advanced kernel-based machine learning techniques, we introduce the functional Neural Tangent Kernel (fNTK) estimator within the Nonlinear Functional Autoregression framework, specifically tailored to capture intricate relationships within implied volatilities. We establish the connection between fNTK and kernel regression, emphasizing its role in contemporary nonparametric statistical modeling. Empirically, we analyze S&P 500 Index options from January 2009 to December 2021, encompassing more than 6 million European calls and puts, thereby showcasing the superior forecast accuracy of fNTK. We demonstrate the significant economic value of having an accurate implied volatility forecaster within trading strategies. Notably, short delta-neutral straddle trading, supported by fNTK, achieves a Sharpe ratio ranging from 1.45 to 2.02, resulting in a relative enhancement in trading outcomes ranging from 77% to 583%.

Keywords: Implied Volatility Surfaces; Neural Networks; Neural Tangent Kernel; Implied Volatility Forecasting; Nonlinear Functional Autoregression; Option Trading Strategies.

JEL classification: C14, C45, C58, G11, G13, G17

*E-mail: matcheny@nus.edu.sg. Department of Mathematics, Asian Institute of Digital Finance, Risk Management Institute, National University of Singapore

†E-mail: grith@ese.eur.nl. Erasmus School of Economics, Erasmus University Rotterdam

‡E-mail: hlhlai@u.nus.edu. Department of Mathematics, Integrative Sciences and Engineering Programme, National University of Singapore

1 Introduction

Implied Volatility (IV) forecasting represents a crucial frontier in financial econometrics, carrying profound implications for financial markets. Serving as a prominent gauge and an 'observable' measure of volatility, IV intricately weaves through investment landscapes, risk management tactics, and strategic financial blueprints. It offers valuable insights into market expectations, influences option pricing models, and impacts the theoretical value of options. Its dynamism captures current market sentiment and projects forward-looking market expectations. Within this omnipresence of IV lies the pressing challenge of precise forecasting—a conduit to maximize its potential in shaping well-informed financial strategies. This underscores the critical need for innovation in IV forecasting methods.

Numerous studies have dedicated considerable effort to forecasting IV, recognizing its significance within the realm of financial markets Audrino and Colangelo (2010), Bernales and Guidolin (2014, 2015), and Kearney, Shang, and Sheenan (2019). Achieving accurate forecasts in this context demands that modeling approaches align with the actual characteristics and dynamics of IVs found in real-world data. Among these characteristics, two distinct types of dependencies emerge. Spatial dependence arises from the reliance of IVs on various moneyness levels and maturities. Concurrently, temporal dependencies come into play as IVs dynamically evolve in response to shifts in the financial market. Both of these dependencies are characterized by intricate, nonlinear patterns Almeida et al. (2022) and Zhang, Li, and Zhang (2023).

Modeling the dynamics of IVs gains a comprehensive perspective when viewed through the Implied Volatility Surface (IVS) prism. The IVS offers a holistic characterization of IV, encapsulating its systematic variations in spatial aspects with respect to both strike price and expiration date. By modeling the entire IV surface, we can comprehensively represent the spatial dependence structures as a function, and the focus shifts toward understanding the temporal dependence within the series of functions. There is an active area of research for embedding features of IVS in the estimation of parametric models (see Carr and Wu, 2016; Aït-Sahalia, Li, and Li, 2021a; Aït-Sahalia,

Li, and Li, 2021b for some recent progress). While less commonly employed in IV forecasting, functional time series analysis has gained prominence as a lively area of research.[1] This includes modeling the dynamics of IVS, albeit typically through linear modeling approaches, see e.g., Cont and Da Fonseca (2002), Fengler, Härdle, and Mammen (2007), Park et al. (2009), Fengler and Hin (2015), Ulrich and Walther (2020). However, it becomes evident that a model aiming to elucidate the IVS dynamics should be flexible to encapsulate nonlinear nature.

To achieve our goal, we introduce the functional Neural Tangent Kernel (fNTK) estimator within the Nonlinear Functional Autoregression (NFAR) modeling framework. In this framework, future implied volatility surfaces depend on past implied volatility surfaces through a nonlinear function-on-function regression framework based on a linear operator within the Reproducing Kernel Hilbert Space (RKHS), see Kadri et al. (2010) and Sang and Li (2022). By leveraging advanced kernel-based machine learning techniques, fNTK is specifically designed to capture intricate complex relationships within implied volatilities. We show the connection between fNTK and kernel regression, highlighting its role in contemporary nonparametric statistical modeling. Our innovation extends the NTK approach proposed by Arora et al. (2019), Jacot, Gabriel, and Hongler (2018) and Domingos (2020) into the functional setup. Our contributions are multifaceted: we establish a comprehensive methodological framework for modeling complex nonlinearity, elucidate the relationship between NTK and kernel regression in the functional setup, and emphasize the significance of fNTK in the modern analytical landscape.

Transitioning from theoretical concepts to empirical exploration, our analysis of the S&P 500 index Implied Volatility Surface (IVS) spanning from January 2009 to December 2021 unfolds a compelling narrative. The fNTK estimator demonstrates a noteworthy improvement, averaging between 4.54% to 39.44% in RMSE forecast accuracy relative to a linear benchmark and a range of alternative models, across 5 to 20-day-ahead predictions. Furthermore, we delve into an important economic question concerning the added value of accurate IV forecast accuracy within various

[1] Examples include forecasting natural gas transmission network (Chen et al., 2021; Xu et al., 2022), yield curves (Xu, Li, and Chen, 2017), probability density functions (Petersen and Müller, 2016; Grith et al., 2018), Value-at-Risk (Cai et al., 2019), limit order books (Chen, Chua, and Härdle, 2019), medical and genetic data (Li, Huang, and Härdle, 2019), brain networks (Solea and Li, 2022).

trading strategies. Remarkably, short delta-neutral straddle trading, empowered by fNTK, yields a Sharpe ratio between 1.45 and 2.02, resulting in a substantial enhancement in trading outcomes, ranging from 77% to 583%. This convergence of methodological rigor and empirical results not only advances statistical understanding but also serves as a guiding light for practitioners in the options market. Our work underscores the presence of nonlinear patterns within the IVS series and the advantages of acknowledging nonlinear dependencies in both modeling and forecasting. Additionally, our study underscores the significant impact that accurate IV forecasting can have on investment decision-making.

The paper is organized as follows. Section 2 gives an overall data description. Section 3 presents the NFAR model, the estimation procedure, and the asymptotic properties of the estimator. Section 4 reports the implementation of the model on the data. Section 5 shows the economic values of the predictability of the NFAR model. Section 6 concludes the paper.

2 Data

We consider daily options on the S&P 500 Index obtained from IvyDB OptionMetrics for the period spanning from January 1, 2009, through to December 31, 2021, encompassing about 6.4 million European calls and puts. This repository facilitates a comprehensive insight into option contract specifics: best bid and ask quotes, expiration date, strike, implied volatility,[2] open interest and volume. We also collect data on the closing value of the S&P 500 index, dividend, forward prices, and the yield curve of riskless interest rates for constructing implied volatility surfaces and further analysis.[3]

In the realm of implied volatility forecasting and trading strategy, options on the S&P 500 Index

[2]For a given option, IvyDB computes the Black-Scholes implied volatility from to the midpoint of the best closing bid price and best ask of the option. The appropriate interest rate input corresponds to the zero-coupon rate that has a maturity equal to the option's expiration and is obtained by linearly interpolating between the two closest zero-coupon rates on the zero curve. The daily dividend rate is assumed constant and is computed by IvyDB from the put-call parity under a "constant dividend yield" assumption using the call's bid price with the offer price of the put and vice versa.

[3]The zero-coupon interest rates curve used by IvyDB is derived from ICE IBA LIBOR rates and settlement prices of CME Eurodollar futures.

(SPX) stand as a paramount dataset. The SPX's encompassing representation of the most prominent U.S.-based publicly traded companies ensures that its options' implied volatility serves as a reliable barometer for broad market volatility expectations. The prodigious liquidity associated with SPX options, attributable to their rank among the most frequently traded options globally, ensures that the derived volatility signals are less tainted by liquidity-induced noise. The diverse expiration cycles and an array of strike prices in these options allow for a meticulous investigation into the intricacies of the implied volatility surfaces. Such data granularity and the economic significance of the SPX movements resonate not just with traders but extend to economists, policymakers, and risk management professionals. The cash-settled nature of SPX options further simplifies the process for those focusing on volatility without the encumbrances of directional biases or the complexities of physical settlement. Furthermore, while individual stocks are often subject to sharp volatility swings due to idiosyncratic events, the aggregated nature of the SPX provides a more tempered and consistent snapshot of market volatility. As such, SPX options have been popular tools for traders and researchers alike when studying implied volatility.

Drawing inspiration from the methodology outlined by Büchner and Kelly (2022), we meticulously apply a set of filters to our dataset to eliminate entries that may arise from recording discrepancies or erroneous inputs. Specifically, we discard all options in which i) the bid price is negative or zero, ii) the bid exceeds the ask, iii) no-arbitrage conditions are violated,[4] iv) implied volatility is missing or non-positive, and (v) the open interest is negative.

For each option contract, we define time-to-maturity τ measured in years as the number of trading days to expiration divided by 252. To measure the moneyness of a contract, we follow Andersen, Fusari, and Todorov (2017) and define it as: $m = \frac{\ln(K/F_\tau)}{\sqrt{\tau} IV_{atm,\tau}}$, where K is the strike price, F_τ denotes the forward price for transactions τ years into the future, while $IV_{atm,\tau}$ denotes the annualized implied volatility of the option with the strike price closest to F_τ.[5] This definition of moneyness has two advantages: (i) an exactly at-the-money option (i.e., $K = F_\tau$) attains a delta of

[4]For example, we ensure the monotonicity of option prices with respect to the strike price.

[5]If a forward contract for a desired time-to-maturity τ is not available, we apply linear interpolation between the two closest forward prices.

roughly 0.5 in absolute value, and (ii) it ensures that the implied volatility surfaces are comparable across different maturities and volatility regimes.

Our study models the implied volatility surface (IVS). The raw data are discretely observed at irregular points (along with moneyness and maturity). We process a pre-smoothing to transform the discrete observations on the same day into a daily continuous function. Specifically, we employ a two-dimensional B-splines expansion[6]. Compared to discrete implied volatilities, IVS holistically offers a comprehensive perspective, capturing the intricate dynamics across different strike prices and expiration times. By encapsulating rich volatility information across the entire options chain in a unified view, the inherent spatial relationships between implied volatilities at varying maturities and moneyness are seamlessly integrated, eliminating the need for separate spatial models and reducing the modeling complexity. Even for SPX, specific moneyness and maturities might be sparse. Thus, we construct IV surfaces with tenors between 5 and 252 trading days, with moneyness ranging from $-2 \leq m \leq 2$. To construct smooth implied volatility surfaces using daily implied volatilities, we encompass options with moneyness stretching from -2.5 to a cap of 2.5, and a time-to-maturity τ of up to 280 days. Our consolidated dataset spans 3273 days, averaging 908 call options and 875 put options daily[7].

The implied volatilities of put and call options with the same strike and maturity provided by OptionMetrics deviate from one another, thereby infringing upon the principles of put-call parity[8]. To address this issue, we construct the implied volatility surfaces of call and put options separately. By independently tailoring the IVS for puts and calls, we accommodate each option type's disparate dynamics and trading patterns. In Figure 1, we provide an illustration that showcases

[6]Specifically, we utilize two-dimensional cubic B-splines with intercepts and without interior knots. This approach results in 16 basis functions. We experimented with different numbers of equidistant knots, such as a single knot at $1/2$ and two knots at $\{1/3, 2/3\}$. However, the RMSE for the smoothed surfaces remained largely consistent across these variations.

[7]In Appendix, Table A.1 summarizes the composition of our consolidated sample. Figure A.1 illustrates histograms of moneyness m and K/F_τ ratios for both option types. Notably, put options manifest a pronounced inclination towards being out of the money, reflected by condition $m < 0$. In contrast, the moneyness distribution for call options seems more equitably spread.

[8]This inconsistency can be attributed to varied factors. For instance, potential discrepancies in IvyDB's dividend computation methods have been alluded to (as elaborated in footnote 12 of in Bernales and Guidolin (2014)). Meanwhile, Büchner and Kelly (2022) hypothesizes that price pressure disparities may be the culprit.

four IV surfaces for both call and put options: one from a calm period in 2019 and another from a volatile period in 2020. This figure also includes snapshots of the discrete implied volatilities of traded options for these periods. An IV surface clearly provides a smoother and more consistent representation of volatility across diverse moneyness and maturity combinations. On the other hand, a focus solely on discrete volatilities can introduce inconsistencies or abrupt transitions that might not truly capture the market's underlying volatility structure.

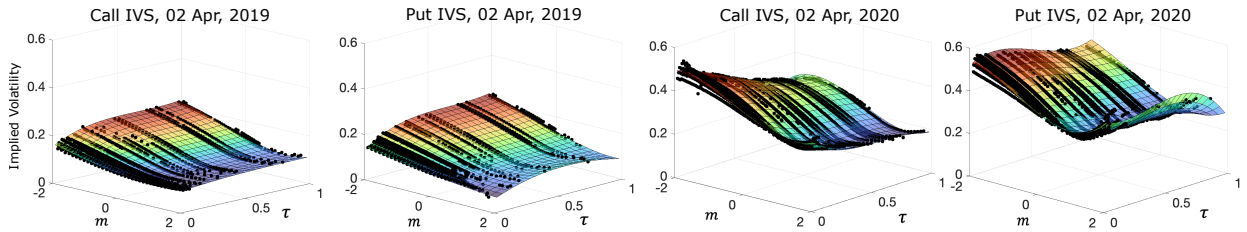
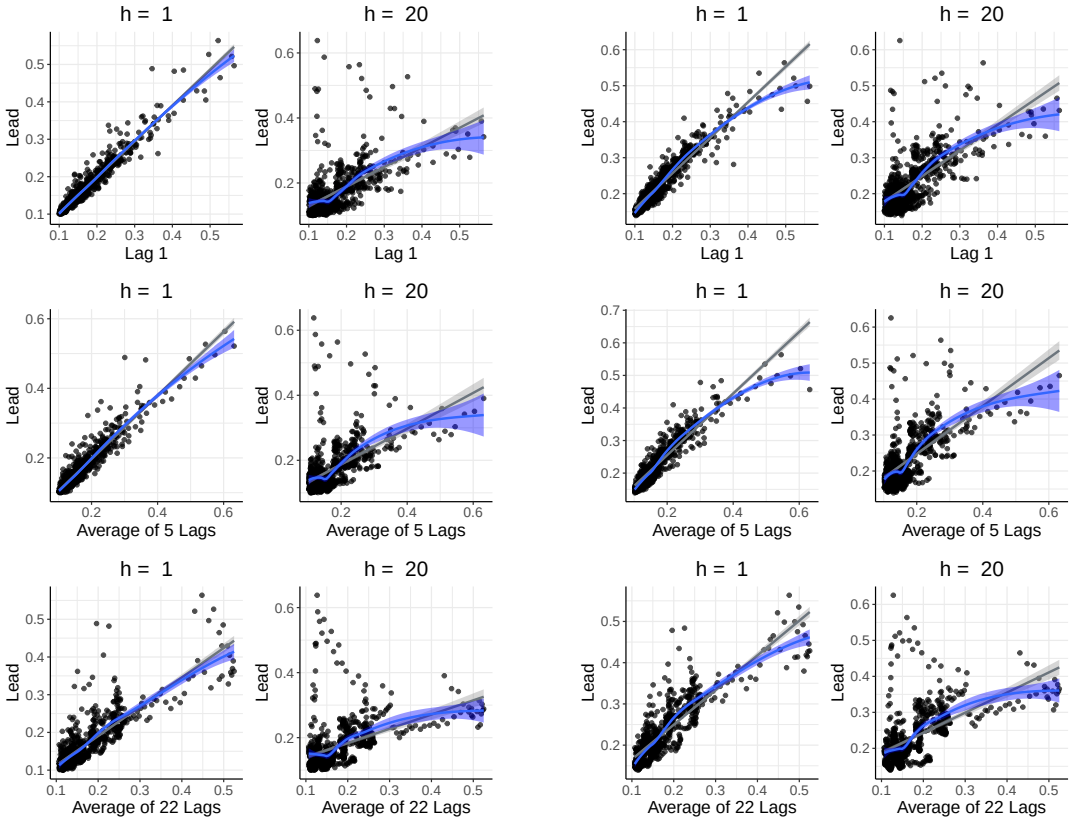


Figure 1: IVS of S&P500 calls and puts options on two different days with low and high volatility. The observed IV values are black dots, and the surfaces are smoothed with two-dimensional B-splines.

Empirical studies unambiguously indicate a pronounced temporal dependence within the Implied Volatility Surface (IVS), with shocks and co-movements manifesting pronounced correlations across diverse time-to-maturities and moneynesses, see Fengler (2012). Figure 2 display the lead-lag relationships and cross-lead-lag regressions for the average of call options, accentuating the interconnectedness of the IV values from various days and maturities. Yet, this temporal relationship is not linear. This is evident in Figure 2, where the nonparametric fit signifies the existence of a nonlinear pattern in the dynamics of IV, especially in the longer forecasting horizons. While the temporal nuances necessitate the application of autoregressive models, the intricate nonlinearities observed command a more sophisticated nonlinear modeling approach.

3 Methodology

In this section, we introduce the nonlinear functional autoregressive (NFAR) model to capture the intricacies of implied volatility surfaces using lagged surfaces. This is achieved through a nonlinear function-on-function regression framework, facilitated by a linear operator in the Reproducing



(a) Lead-lag regression

(b) Cross lead-lag regression

Figure 2: Temporal dependence in the average IVS of call and put options. We plot the lead-lag (both IV on day t and on day $t+h$ at 30 day-to-maturity and moneyness $m=0$) as well as the cross-lead-lag regression (IV on day t at 30 day-to-maturity and $m=0$ while IV on day $t+h$ at 60 day-to-maturity and $m=-0.75$) at two horizons, $h=1$ and 20 and different lags (the last one day, average of the last five days, and average of the last 22 days), using data of the prediction period, from Jan 05, 2019 to Dec 31, 2021.

Kernel Hilbert Space (RKHS) - a functional space where regularized models can be defined from data with the representer theorem. It enables the representation of high-dimensional objects and their dependencies in a more feasible manner. We show that our model demonstrates adaptability to a wide range of kernels. Leveraging advanced kernel-based machine learning techniques, we utilize the Neural Tangent Kernel (NTK), a versatile kernel capable of capturing the complex nonlinear relationship in the feature space of the original curves through convolutional neural networks. We rigorously draw the connection between NTK and kernel regression through derivation, highlighting the NTK's position as a contemporary nonparametric statistical model.

3.1 Nonlinear functional autoregression

We denote by $\mathscr{H} = L^2(\mathscr{I})$ the Hilbert space consisting of all square-integrable surfaces defined on a compact set $\mathscr{I} \subset \mathbb{R}^q$ and equipped with the inner product $\langle f, g \rangle_{\mathscr{H}} = \int_{\mathscr{I}} f(u)g(u)\,du$, for any $f, g \in L^2(\mathscr{I})$. Define the squared L^2 norm of a function by $\|f\|_{\mathscr{H}} = \langle f, f \rangle_{\mathscr{H}}$.

Let $\{Y_i\}_{i=1}^n$ be a series of n random surfaces that take values on $\mathscr{H}_Y = L^2(\mathscr{I}_Y)$. Associated with each Y_i, there is a regressor surface $X_i \in \mathscr{H}_X = L^2(\mathscr{I}_X)$. We consider functions with finite second moment, i.e., $\mathbb{E}\left[||Y_i||^2_{\mathscr{H}_Y}\right] < \infty$ and $\mathbb{E}\left[\|X_i\|^2_{\mathscr{H}_X}\right] < \infty$. For simplicity, we assume that Y_i and X_i are centered functions, i.e., $\mu_X(v) = \mathbb{E}[X_i(v)] = 0, \forall v \in \mathscr{I}_X$ and $\mu_Y(u) = \mathbb{E}[Y_i(u)] = 0, \forall u \in \mathscr{I}_Y$. Let P_X and P_Y denote the distributions of X and Y, and $P_{Y|X} : \mathscr{H}_X \times \mathscr{H}_Y \to \mathbb{R}$ the conditional distribution of Y given X. If $L_2(P_X)$ represents the class of all measurable functions of X with $\mathbb{E}[f^2(X)] < \infty$ under P_X, then $L_2(P_Y)$ is similarly defined for Y. Our goal is to capture the dependence between Y_i and X_i through a function $f : \mathscr{H}_X \to \mathscr{H}_Y$ such that

$$Y_i = f\big(X_i\big) + \varepsilon_i, \tag{1}$$

where the innovation ε_i is a noise function with $\mathbb{E}[\varepsilon_i(u)] = 0, \forall u \in \mathscr{I}_Y$ and $\mathbb{E}\left[\|\varepsilon_i\|^2_{\mathscr{H}_Y}\right] < \infty$. In our study, X_i can be a lagged surface Y_{i-1}, multiple lagged surfaces $Y_{i-1}, \ldots, Y_{i-p}$, or linear combinations of the lags. Hence, model (1) is a nonlinear functional autoregression model (NFAR).

To elucidate the nonlinear relation between X_i and Y_i, we introduce another Hilbert space of functions generated by a positive-definite kernel $K : \mathscr{H}_X \times \mathscr{H}_X \to \mathbb{R}$ defined on the inner product of $\mathscr{H}_X$ through a function $\rho : \mathbb{R}^3 \to \mathbb{R}^+$, such that

$$K(X_i, X_j) = \rho(\langle X_i, X_i \rangle_{\mathscr{H}_X}, \langle X_i, X_j \rangle_{\mathscr{H}_X}, \langle X_j, X_j \rangle_{\mathscr{H}_X}), \tag{2}$$

for any $X_i, X_j \in \mathscr{H}_X$. The kernel K satisfies the kernel property $K(X_i, X_j) = \langle K(.,X_i), K(.,X_j) \rangle_{\mathfrak{M}_X}$.[9]

The space induced by K, denoted by $\mathfrak{M}_X$, is a nested space of $\mathscr{H}_X$ via ρ, see Li and Song

[9] Intuitively, $K(.,X_i) : \mathscr{H}_X \to \mathfrak{M}_X$ can be thought of as a feature map defined by the kernel, and $K(X_i,X_j)$ as a measure of similarity between any two curves X_i, X_j in the $\mathfrak{M}_X$ space.

(2017). For ease of notation, we assume that kernels are demeaned, i.e., for any $X_i \in \mathscr{H}_X$, $\mu_X(X_i) = \langle \mu_X, K(.,X_i)\rangle_{\mathfrak{M}_X} = \mathbb{E}[K(X,X_i)] = 0$. The space $\mathfrak{M}_X$ is called RKHS since $\mathfrak{M}_X = \text{span}\{K(\cdot,X_i) : X_i \in \mathscr{H}_X\}$ and K has the reproducing property, i.e., for any function $g \in \mathfrak{M}_X$, $g(X_i) = \langle g, K(\cdot,X_i)\rangle_{\mathfrak{M}_X}$.

The introduction of the kernel K is crucial in capturing the essence of the underlying nonlinear relationship. With the RKHS $\mathfrak{M}_X$ generated by K, the nonlinear function in the model is represented by a linear expansion of functions in the nested space. Let $\mathscr{B}(\mathscr{H}_1, \mathscr{H}_2)$ be the class of bounded linear operators mapping a Hilbert space $\mathscr{H}_1$ to another Hilbert space $\mathscr{H}_2$. Then $Bg(X_i) = \langle Bg, K(\cdot,X_i)\rangle_{\mathfrak{M}_X} = \langle g, B^*K(\cdot,X_i)\rangle_{\mathscr{H}_Y}$, for $g \in \mathscr{H}_Y, B \in \mathscr{B}(\mathscr{H}_Y, \mathfrak{M}_X)$ and $B^* \in \mathscr{B}(\mathfrak{M}_X, \mathscr{H}_Y)$ the adjoint operator of B. This means that we can represent functions in $\mathscr{H}_Y$ by means of the kernel.

The function-on-function regression problem in Equation (1) can be reformulated as a functional kernel regression, in which the task is to find $B_0 \in \mathscr{B}(\mathscr{H}_Y, \mathfrak{M}_X)$ such that

$$B_0 = \underset{B \in \mathscr{B}(\mathscr{H}_Y, \mathfrak{M}_X)}{\arg\min} \; \mathbb{E}\left[\|Y_i - B^*K(.,X_i)\|^2_{\mathscr{H}_Y}\right]. \tag{3}$$

This model can be viewed as an extension of the traditional multivariate linear model that associates vector responses with vector covariates. The functional normal equation of the least squares regression from the RKHS to $\mathscr{H}_Y$ takes the form $\Sigma_{XY} = \Sigma_{XX}B_0$, with the (cross-)covariance operators $\Sigma_{XX} \in \mathscr{B}(\mathfrak{M}_X, \mathfrak{M}_X)$, $\Sigma_{XY} \in \mathscr{B}(\mathscr{H}_Y, \mathfrak{M}_X)$ such that $\Sigma_{XX} = \mathbb{E}[K(.,X) \otimes K(.,X)]$ and $\Sigma_{XY} = \mathbb{E}[K(.,X) \otimes Y]$. Since Σ_{XX} is a compact operator in L^2, its inverse is not bounded, leading to an ill-posed problem. To address this issue, we define $\Sigma^{\dagger}_{XX}$ to be the Moore-Penrose inverse of Σ_{XX}. Theorem 2.1. in Sang and Li (2022) states that a solution to the regression (3) is given by

$$B_0 = \Sigma^{\dagger}_{XX}\Sigma_{XY}. \tag{4}$$

The solution B_0 is well-defined under the following additional assumptions:

Assumption 1. *$\mathfrak{M}_X$ is a dense subset of $L_2(P_X)$ and $\mathbb{E}[\|Y\|^2_{\mathscr{H}_Y}] < \infty$;*

Assumption 2. *There exists a constant $C > 0$ so that for any $f \in \mathfrak{M}_X$, $\mathbb{E}[f^2(X)] \leq C\|f\|^2_{\mathfrak{M}_X}$;*

Assumption 3. *$ran(\Sigma_{XY}) \subseteq ran(\Sigma_{XX})$ and $\Sigma_{XX}^{\dagger}\Sigma_{XY}$ is a bounded operator.*

Proposition 2.2. and Proposition 2.3. of Sang and Li (2022) lead to the following relationship between B_0 and the predicted value of Y_i for a given $X_i \in \mathscr{H}_X$ and the kernel K

$$\begin{aligned}\mathbb{E}[Y_i|X = X_i] &= B_0^* K(.,X_i) \\ &= \Sigma_{YX}\Sigma_{XX}^{\dagger}K(.,X_i) \\ &= \mathbb{E}\left[\{(\Sigma_{XX}^{\dagger}K(.,X_i))(X)\}Y\right].\end{aligned} \tag{5}$$

where $B_0^* = \Sigma_{YX}\Sigma_{XX}^{\dagger} \in \mathscr{B}(\mathfrak{M}_X, \mathscr{H}_Y)$ is the adjoint operator of B_0, and $\Sigma_{YX} = \Sigma_{XY}^* = \mathbb{E}[Y \otimes K(.,X)] \in \mathscr{B}(\mathfrak{M}_X, \mathscr{H}_Y)$. The last equality is an expectation of weighted function Y, where the random weights are defined as $W : \mathscr{H}_X \times \mathscr{H}_X \to \mathbb{R}$, $W(X_i,X) := (\Sigma_{XX}^{\dagger}K(.,X_i))(X) = \langle \Sigma_{XX}^{\dagger}K(.,X_i), K(.,X)\rangle_{\mathfrak{M}_X}$. Note that the conditional expectation in Equation (5) essentially depends on the kernel K. In what follows, we transform the function-to-function regression into a vector-to-vector regression, resulting, under suitable assumption, in a reduced dimensional space representation of the curves.

3.1.1 From RKHS of functions to RKHS of vectors

For $\mathbb{E}[\|X_i\|^2_{\mathscr{H}_X}] < \infty$, we apply the Karhunen–Loève representation of X_i in terms of eigenfunctions $(\psi_1, \psi_2, \ldots)$ of the covariance operator $C_X = \mathbb{E}[X \otimes X]$, with $\psi_v \in \mathscr{H}_X$, and projection coefficients $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots)^T \in \mathscr{H}_{\boldsymbol{x}} \subseteq \mathbb{R}^{\infty}$. Hence,

$$X_i = \sum_{v=1}^{\infty} x_{iv}\psi_v, \quad \text{with } x_{iv} = \langle X_i, \psi_v\rangle_{\mathscr{H}_X}. \tag{6}$$

Then, we can define a new kernel $k : \mathscr{H}_{\boldsymbol{x}} \times \mathscr{H}_{\boldsymbol{x}} \to \mathbb{R}$ such that for any $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathscr{H}_{\boldsymbol{x}}$

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \rho(\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle, \langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle, \langle \boldsymbol{x}_j, \boldsymbol{x}_j\rangle). \tag{7}$$

Denote $\mathfrak{M}_{\boldsymbol{x}}$ the RKHS induced by k and determined by the inner product of $\mathscr{H}_{\boldsymbol{x}}$. The following theorem states the isomorphism between the two Reproducing Kernel Hilbert Spaces $\mathfrak{M}_X$ and $\mathfrak{M}_{\boldsymbol{x}}$.

Theorem 1 (**Isomorphism between Reproducing Kernel Hilbert Spaces**). *Under Equations* (6) *and* (7)*, it holds that*

$$\begin{aligned} k(\boldsymbol{x}_i, \boldsymbol{x}_j) &= \langle k(.,\boldsymbol{x}_i), k(.,\boldsymbol{x}_j)\rangle \\ &= \langle K(.,X_i), K(.,X_j)\rangle_{\mathfrak{M}_X} = K(X_i,X_j). \end{aligned} \tag{8}$$

Then the RKHS $\mathfrak{M}_X$ nested on $\mathscr{H}_X$ is isometrically isomorphic to the RKHS $\mathfrak{M}_{\boldsymbol{x}}$ nested on $\mathscr{H}_{\boldsymbol{x}}$.

3.1.2 From function-to-function to vector-to-vector regression.

Functions contain abundant spatial dependence information but are defined in infinite regions, leading to numerical challenges in analysis and estimation. Therefore, there is a motivation to shift from the functional domain to the vector domain, where the vectors represent extracted factors from the infinite dimensionality while encapsulating the crucial features of spatial dependence.

For $\mathbb{E}\left[||Y_i||^2_{\mathscr{H}_Y}\right] < \infty$, we apply the spectral decomposition of the covariance operator $C_Y = \mathbb{E}[Y \otimes Y]$ to project the Y_i onto the orthonormal eigenfunctions $\boldsymbol{\varphi} = (\varphi_1, \varphi_2 \ldots)^T$ where $\varphi_j \in \mathscr{H}_Y$

$$Y_i = \sum_{j=1}^{\infty} y_{ij}\varphi_j, \quad \text{with } y_{ij} = \langle Y_i, \varphi_j\rangle_{\mathscr{H}_Y}, \tag{9}$$

satisfying $\mathbb{E}\left[y_{ij}y_{rv}\right] = 0$ for $j \neq v$ and any i, r. Let $\Sigma_{\boldsymbol{xx}} = \mathbb{E}\left[k(.,\boldsymbol{x}) \otimes k(.,\boldsymbol{x})\right]$ be the covariance matrix of $k(.,\boldsymbol{x})$ and $\Sigma^{\dagger}_{\boldsymbol{xx}}$ its Moore-Penrose inverse. Further define $\boldsymbol{y} = (y_{i1}, y_{i2}, \ldots)^T \in \mathscr{H}_{\boldsymbol{y}} \subseteq \mathbb{R}^{\infty}$, and $\Sigma_{\boldsymbol{yx}} = \mathbb{E}[\boldsymbol{y} \otimes k(.,\boldsymbol{x})]$. Now, we can establish the link between function-to-function regression to vector-to-vector regression.

Theorem 2 (**Vector-to-vector regression**). *Given the decomposition of X_i in Equations* (6) *and Y_i in Equation* (9)*, under Assumptions* (1) - (3) *and Theorem 1, for a positive definite kernel k defined by Equation* (7)*, if there is a covariance matrix $\Sigma_{\boldsymbol{xx}}$ of $k(.,\boldsymbol{x})$ that is diagonal, then the function-to-function regression model in Equation* (3) *may be represented equivalently by*

$$\beta_0 = \underset{\beta \in \mathscr{B}(\mathscr{H}_{\boldsymbol{y}}, \mathfrak{M}_{\boldsymbol{x}})}{\arg\min} \; \mathbb{E}\left[\|\boldsymbol{y}_i - \beta^* k(.,\boldsymbol{x}_i)\|^2_2\right], \tag{10}$$

with solution $\beta_0 = \Sigma^{\dagger}_{\boldsymbol{xx}}\Sigma_{\boldsymbol{xy}}$. *This leads to*

$$\begin{aligned}\mathbb{E}[\boldsymbol{y}_i|\boldsymbol{x}=\boldsymbol{x}_i] &= \beta_0^* k(.,\boldsymbol{x}_i) \\ &= \Sigma_{\boldsymbol{yx}}\Sigma^{\dagger}_{\boldsymbol{xx}} k(.,\boldsymbol{x}_i) \\ &= \mathbb{E}\left[\{(\Sigma^{\dagger}_{\boldsymbol{xx}} k(.,\boldsymbol{x}_i))(\boldsymbol{x})\}\boldsymbol{y}\right].\end{aligned} \tag{11}$$

Although vectors offer a more compact representation of functions, they still exist within an infinite-dimensional framework. This inherent complexity makes practical computations of $\mathbb{E}[\boldsymbol{y}_i|\boldsymbol{x}=\boldsymbol{x}_i]$ in Equation (11) challenging when working with finite sample sizes. To address this issue, we employ classical sieve methods, as explained in detail in Appendix A.3. These methods involve truncating the regression for the original functions while striving to minimize any loss of information. More specifically, we adopt truncated Karhunen-Loéve representations to capture the majority of the variance within the processes, as outlined in Appendix A.4.1.

3.2 Neural Tangent Kernel

Kernel functions are associated with particular feature spaces that may possess either finite or infinite dimensions. Generally, these functions have simple parameterizations through the imposition of regularity constraints on the feature space that facilitates the transformation of a regression problem with numerous features into a weighted average of observations, wherein the kernel utilized plays a straightforward role in determining the weights assigned. While nonparametric kernel regressions depend on the choice of kernel, parametric kernels can be limited in their ability to describe nonlinear dependence. We propose a flexible class of kernels, the Neural Tangent Kernel (NTK), which leverages neural networks to effectively capture complex nonlinear dependencies in the data. Effectively, the NTK characterizes the behavior of neural networks during training with first-order gradient descent and is specific to a given neural network architecture. It is computed as the inner product of the gradients of the neural network with respect to the weights in the network. Recent studies have shown that for infinite widths and specific parameterization, neural

networks are equivalent to kernel regressions that depend on a deterministic limiting kernel determined solely by the network architecture, see Jacot, Gabriel, and Hongler (2018), Arora, Du, Hu, Li, Salakhutdinov, and Wang (2019) and Chen et al. (2020). Our paper employs the NTK of shallow, fully connected neural networks with wide widths. These are shown to give good results in finite samples, especially with weight decay, see Lee et al. (2020).

Suppose we have a NN of depth L, $f(.;\theta) : \mathbb{R}^{n_0} \to \mathbb{R}^{n_L}$ with parameters θ, where layers are indexed from 0 (input $\boldsymbol{x}$) to L (output $\boldsymbol{y}$), each layer containing $n_0, n_1, ..., n_L$ neurons. We use the NTK parameterization of Jacot, Gabriel, and Hongler (2018) for the NN

$$\begin{aligned}
\text{input layer} &: \alpha^{(0)}(\boldsymbol{x};\theta) = \boldsymbol{x} \in \mathbb{R}^{n_0}, \\
\text{preactivation} &: \tilde{\alpha}^{(\ell+1)}(\boldsymbol{x};\theta) = \frac{1}{\sqrt{n_\ell}} W^{(\ell)} \alpha^{(\ell)}(\boldsymbol{x};\theta) + \eta b^{(\ell)}, \\
\text{activation} &: \alpha^{(\ell+1)}(\boldsymbol{x};\theta) = \sigma(\tilde{\alpha}^{(\ell+1)}(\boldsymbol{x};\theta)), \\
\text{output layer} &: f(\boldsymbol{x};\theta) = \tilde{\alpha}^{(L)}(\boldsymbol{x};\theta) \in \mathbb{R}^{n_L},
\end{aligned}$$

where parameters θ consist of the connection matrices $W^{(\ell)} \in \mathbb{R}^{n_{\ell+1} \times n_\ell}$, and bias vectors $b^{(\ell)} \in \mathbb{R}^{n_{\ell+1}}$. All parameters are initialized as i.i.d. Gaussians $\mathcal{N}(0,1)$; the constant $\eta > 0$ controls the influence of the bias on the training; and the nonlinear activation function $\sigma(.)$ is applied element-wise, to each element of $\tilde{\alpha}^{(\ell+1)}(\boldsymbol{x};\theta)$. Note that NTK models are parameterized differently from the standard NNs, commonly used in previous popular studies in finance, such as Gu, Kelly, and Xiu (2020) and Almeida et al. (2022), which makes them more suitable in a kernel regression setting[10]. Given a training dataset $\{(\boldsymbol{x}_i, \boldsymbol{y}_i) : i = 1, ..., n\}$, the least-squares loss is defined as

$$\mathscr{L}(\theta) = \sum_{i=1}^{n} \|\boldsymbol{y}_i - f(\boldsymbol{x}_i;\theta)\|^2. \tag{12}$$

The parameters of the NN model are updated and learned by minimizing a least-squares loss

[10]Usually, the standard NNs do not have the factors $\frac{1}{\sqrt{n_\ell}}$, and their parameters are initialized using LeCun initialization, with $W_{ij}^{\ell} \sim \mathcal{N}(0, 1/n_\ell)$ and $b_j^{l} \sim \mathcal{N}(0,1)$. The factors $\frac{1}{\sqrt{n_\ell}}$ are essential for obtaining a consistent asymptotic behavior of the NNs as the number of neurons $n_1, n_2, ..., n_{L-1}$ go to infinity; while the factor η is introduced to balance the influence of the bias and the weights.

function $\mathscr{L}(\theta)$ via the back-propagation algorithm and the Gradient Descent (GD) method with learning rate ζ. At each updating step s, the parameters θ_s are updated to θ_{s+1} by

$$\theta_{s+1} = \theta_s - \zeta \nabla_\theta \mathscr{L}(\theta_s), \tag{13}$$

with $\nabla_\theta \mathscr{L}(\theta_s)$ the gradient of the loss function with respect to θ at step s. Using GD, the weights and biases in all layers are simultaneously updated towards the parameters that minimize the loss function, effectively training the neural network to make accurate predictions. Let the prediction vector of the NN be $f(\boldsymbol{x};\theta) = (f_1(\boldsymbol{x};\theta),...,f_{n_L}(\boldsymbol{x};\theta))^T$, with $f_j(\boldsymbol{x};\theta)$ denoting the j-th output of the NN. Daniely, Frostig, and Singer (2016) have shown that at initialization, as the number of neurons in each layer goes to infinite, the outputs $f_j(\boldsymbol{x};\theta)$ for $j = 1,...,n_L$ tend to i.i.d. centered Gaussian processes (GP) with scalar covariance $\Sigma^{(L)}$ defined recursively by

$$\begin{aligned} \Sigma^{(1)}(\boldsymbol{x},\boldsymbol{x}') &= \frac{1}{n_0}\boldsymbol{x}^T\boldsymbol{x}' + \eta^2 \\ \Sigma^{(\ell+1)}(\boldsymbol{x},\boldsymbol{x}') &= \mathbb{E}_\theta\Big[\sigma(f(\boldsymbol{x}))\sigma(f(\boldsymbol{x}'))\Big] + \eta^2, \end{aligned} \tag{14}$$

where $\Sigma^{(1)}(\boldsymbol{x},\boldsymbol{x}')$ can be computed from the network architecture. Suppose the NN is trained with gradient descent over a number of updating steps indexed by s. Thus, $f(\boldsymbol{x};\theta_s)$ is the output of the NN using $\theta = \theta_s$, and $\nabla_\theta f(\boldsymbol{x};\theta_s)$ is the rate of change in output $f(\boldsymbol{x};\theta_s)$ with respect to parameter θ at step s. The empirical neural tangent kernel matrix in $\mathbb{R}^{n_L \times n_L}$ for the depth L NN is defined as

$$\tilde{k}_s^{(L)}(\boldsymbol{x},\boldsymbol{x}') = \nabla_\theta^T f(\boldsymbol{x};\theta_s)\nabla_\theta f(\boldsymbol{x}';\theta_s). \tag{15}$$

By Theorem 1 and Theorem 2 in Jacot, Gabriel, and Hongler (2018), for $\sigma(.)$ being Lipschitz, as $n_1, n_2, ..., n_{L-1} \to \infty$, the empirical NTK $\tilde{k}_s^{(L)}(\boldsymbol{x},\boldsymbol{x}')$ converges in probability to a deterministic limiting kernel matrix $k_\infty^{(L)}(\boldsymbol{x},\boldsymbol{x}') \otimes I_{n_L}$ for all $\boldsymbol{x},\boldsymbol{x}' \in \mathbb{R}^{n_0}$ and $s \geq 0$, where I_{n_L} is an identity matrix in $\mathbb{R}^{n_L \times n_L}$. This implies that an NN with n_L outputs behaves asymptotically like n_L NNs with scalar

outputs trained independently. The scalar kernel $k_\infty^{(L)}(\boldsymbol{x},\boldsymbol{x}') : \mathbb{R}^{n_0} \times \mathbb{R}^{n_0} \to \mathbb{R}$ is defined recursively

$$\begin{aligned} k_\infty^{(1)}(\boldsymbol{x},\boldsymbol{x}') &= \Sigma^{(1)}(\boldsymbol{x},\boldsymbol{x}') \\ k_\infty^{(\ell+1)}(\boldsymbol{x},\boldsymbol{x}') &= k_\infty^{(\ell)}(\boldsymbol{x},\boldsymbol{x}')\dot{\Sigma}^{(\ell+1)}(\boldsymbol{x},\boldsymbol{x}') + \Sigma^{(\ell+1)}(\boldsymbol{x},\boldsymbol{x}'), \end{aligned} \tag{16}$$

where $\dot{\Sigma}^{(\ell+1)}(\boldsymbol{x},\boldsymbol{x}') = \mathbb{E}_\theta\left[\dot{\sigma}(f(\boldsymbol{x}))\dot{\sigma}(f(\boldsymbol{x}'))\right]$ and $\dot{\sigma}$ is the derivative of σ with respect to θ. This means that the empirical NTK is independent of the initialization value and is solely determined by the NN architecture. Training the NN under least-square loss $\mathscr{L}(\theta)$ by gradient descent is equivalent to a kernel regression using the NTK. Although the limiting NTK $k_\infty^{(L)}$ has a different form compared to the standard kernels, it is suitable for our modeling framework. It is defined recursively via the covariance of the GPs, $\Sigma^{(\ell)}$ in (14) and (16) and utilizes the inner product of $\boldsymbol{x}$ and $\boldsymbol{x}'$ in each iterative steps. Hence, it satisfies Equation (7) and can be used in Equation (11).[11].

3.3 fNTK algorithm

We develop an fNTK algorithm for the NFAR model, see Figure 3. In estimation, we project an infinite-dimensional process onto a finite parameter space for computational feasibility. Employing the sieve method, the infinite coefficients are estimated within a finite sample space based on a specific parameter space derived from a dataset with n observations. Functional Principal Component Analysis (FPCA) identifies orthonormal bases for distinct functional spaces using empirical mean and variance operators. We provide detailed information about the implementation in Appendix A.4. We also conduct a simulation study in Appendix A.5, showing that the proposed estimation is numerically consistent.

[11] In practice, it is infeasible to implement an NN with infinite widths. However, for large enough widths of the hidden layers, the NN provides a fair approximation of $k_\infty^{(L)}$.

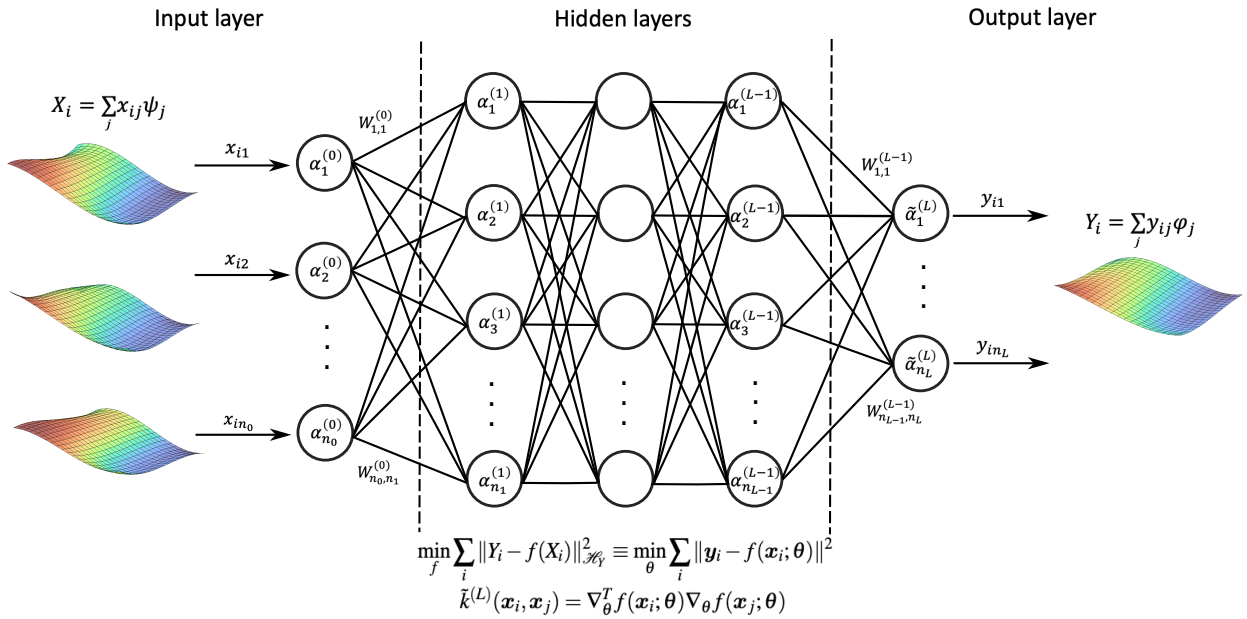


Figure 3: Training an fNTK parametrized network. Parameters $\theta = \left\{W^{(\ell)}, b^{(\ell)} : \ell = 1, L-1\right\}$ are initialized as i.i.d. $\mathcal{N}(0,1)$ and updated simultaneously in each gradient descent step. The pre-activation is $\tilde{\alpha}^{(\ell+1)}(x_i;\theta) = \frac{1}{\sqrt{n_\ell}} W^{(\ell)}\alpha^{(\ell)}(x_i;\theta) + \eta b^{(\ell)}$, and the activation is $\alpha^{(\ell+1)}(x_i;\theta) = \sigma(\tilde{\alpha}^{(\ell+1)}(x_i;\theta))$. For infinitely wide networks, the empirical NTK $\tilde{k}^{(L)}(x_i, x_j) \xrightarrow{p} k_\infty^{(L)}(x_i, x_j) \otimes I_{n_L}$ for all $x_i, x_j \in \mathbb{R}^{n_0}$.

4 Forecasting Implied Volatility Surfaces

This section aims to investigate the effectiveness of the proposed neural-based machine learning estimation approach, fNTK, for nonlinear functional autoregressive models in forecasting implied volatility surfaces (IVS). We compare the fNTK approach with classical IV forecasting alternatives and nonparametric functional estimation using alternative kernels to assess their predictive performance.

4.1 Forecasting Framework

To forecast IVS, incorporate different lags of implied volatility surfaces as the explanatory function X_t. The response function Y_t represents the future implied volatility surface $IV_{t+h}(u)$ of day $t+h$, with a forecasting horizon of h. Following previous studies by Bernales and Guidolin (2014) and Goncalves and Guidolin (2006), we consider IVS defined on two features representing continuous

domains in maturity and moneyness, denoted as $u = (\tau, m)$.

$$IV_t^{(1)}(u) = IV_t(u), \quad IV_t^{(5)}(u) = \frac{1}{5}\sum_{k=t-4}^{t} IV_k(u), \quad IV_t^{(22)}(u) = \frac{1}{22}\sum_{k=t-21}^{t} IV_k(u).$$

The explanatory variables capture the daily, weekly, and monthly features of the IVS, respectively. This approach effectively captures the three primary volatility components associated with different types of traders: short-term, medium-term, and long-term. We standardize the three predictor surfaces, $IV_t^{(1)}(u)$, $IV_t^{(5)}(u)$, and $IV_t^{(22)}(u)$, and combine them into $X_t(u)$. It's important to note that standardization is based solely on the training data to prevent information leakage from the test data.

The Nonlinear Functional Autoregressive (NFAR) model is trained using a rolling window spanning 2500 days, ensuring that the model remains adaptable to the most recent market conditions. This training practice results in a roughly 80%/20% train/test split ratio. Model hyperparameter tuning is conducted at six-month intervals, with detailed values provided in Appendix A.7. During the hyperparameter tuning process, the training data is partitioned into a 2000-day training set and a 500-day validation set. The optimal hyperparameters derived from this tuning process are then retained until the subsequent update. The validation set is used for model and hyperparameter selection, while the test set assesses how well the models generalize to new, unseen data, where no estimation or tuning is conducted [12]. Additionally, we perform a robustness check by training models using rolling windows of 1000 and 2000 days, further partitioned into 500/500 and 1500/500 days for training and validation, respectively. The results closely align with the reported findings and are accessible for reference in Appendix A.9.

4.2 Alternative Models

In addition to fNTK, we explore alternative nonparametric functional estimation approaches. We delve into a kernel regression approach, using linear, Gaussian, and Laplacian kernels within the

[12] Standard k-fold cross-validation is not considered since it violates the time series structure.

proposed NFAR framework. The linear kernel (LinK) is defined by $k(\boldsymbol{x},\boldsymbol{x}') = \boldsymbol{x}^T\boldsymbol{x}'$. In this case, the new RKHS is the same as the original Hilbert input variable space, and the kernel regression is equivalent to a linear regression. In other words, it becomes functional autoregression. We also utilize two popular nonlinear parametric kernels: the Gaussian kernel (GauK), also known as the radial basis function kernel ($k(\boldsymbol{x},\boldsymbol{x}') = \exp(-\gamma\|\boldsymbol{x}-\boldsymbol{x}'\|^2)$) and the Laplacian kernel (LapK) ($k(\boldsymbol{x},\boldsymbol{x}') = \exp(-\gamma\|\boldsymbol{x}-\boldsymbol{x}'\|_1)$). This also equates to an infinitely dimensional feature space in the context of kernels. γ is a constant controlling the roughness of the kernel.[13] For the parametric kernels, exact solutions are obtained in a kernel-ridge regression as explained in Appendix A.6.

We also consider several classical (non-functional) models for implied volatility forecasting. In these classical models, the discrete implied volatilities are forecasted individually. The objective is to scrutinize the importance of nonparametric modeling in capturing nonlinear dynamics. The classical models encompass the renowned Carr and Wu model by Carr and Wu (2016) and the Ad-Hoc Black-Scholes model by Dumas, Fleming, and Whaley (1998). These models have demonstrated effectiveness for S&P 500 IVS and were previously employed in studies by Goncalves and Guidolin (2006), Bernales and Guidolin (2014), and Almeida et al. (2022). Additionally, we adapt the random walk model to our functional framework.

Carr and Wu (CW) Model. The approach introduced by Carr and Wu (2016) offers an option pricing framework that characterizes implied volatility dynamics across various strikes and maturities. For an option with strike and time to maturity, the risk-neutral measures encapsulate the dynamics of the underlying spot price and the option implied volatility. The model involves parameters like the instantaneous variance rate, the average implied volatility drift and the exponential dampening parameter. We employ the relative strike and formulate a quadratic Equation for the square implied volatility. Parameter estimation involves minimizing a nonlinear least squares problem, and predictions are derived by solving the quadratic Equation using the estimated parameters. We refer to Appendix A.8.1 for detailed formulation.

[13] We experimented with polynomial kernels of different degrees of 2, 3, 4, and 5. However, their performance is significantly worse than all other kernels; hence, our methodology and results do not report it.

Ad-Hoc Black–Scholes (AHBS) Model. The approach proposed by Goncalves and Guidolin (2006) extracts features of a cross-section of options on moneyness and maturities based on the ad-hoc Black-Scholes (AHBS) model by Dumas, Fleming, and Whaley (1998) and utilizes a vector autoregressive model to predict the features, enabling the forecasting of implied volatilities. We refer to Appendix A.8.2 for detailed formulation.

Functional Random Walk (fRW) Model. The functional random walk model, adapted from Bernales and Guidolin (2014), predicts the implied volatility surface of a future day to be the same as the current day's. This straightforward approach is a comparative baseline to gauge the significance of incorporating implied volatility surface dynamics in enhancing predictions.

4.3 Forecasting performance measures

We assess prediction accuracy and goodness of fit derived from observed test data. Specifically, root mean square error (RMSE) captures prediction accuracy, and out-of-sample R^2 (OoR^2) measures the goodness of fit, depicting the proportion of variance the models explain.[14]

$$\mathrm{RMSE}_h = \sqrt{\frac{1}{\sum_{t=t_0}^{T-h} N_t} \sum_{t=t_0}^{T-h} \sum_{j=1}^{N_t} (Y_{t+h}(u_j) - \hat{Y}_{t+h}(u_j))^2},$$

$$\mathrm{Oo}R^2_h = 1 - \frac{\sum_{t=t_0}^{T-h} \sum_{j=1}^{N_t} (Y_{t+h}(u_j) - \hat{Y}_{t+h}(u_j))^2}{\sum_{t=t_0}^{T-h} \sum_{j=1}^{N_t} (Y_{t+h}(u_j) - \bar{Y}_{t+h})^2}.$$

4.4 Results

Table 1 presents an overview of model performance across different forecasting horizons ($h = 1, 5, 10, 15$, and 20 days ahead).[15] This analysis underscores the exceptional performance of the fLink model for one-step-ahead predictions. This can be attributed to the inherently linear nature of implied volatility surfaces for short lags, where linear predictors effectively capture the high

[14] In addition, we report mean absolute percentage error (MAPE) and mean correct prediction of direction of change (MCPDC) in the Appendix A.9.

[15] For clarity, it is worth noting that the results reported represent the average performance measures for both put and call options. Detailed results for put and call options are available upon request.

persistence pattern. This observation aligns with analogous findings in prior financial time series research, such as Christensen, Siggaard, and Veliyev (2022) in the context of forecasting realized variance and Nunes et al. (2019) for yield curve prediction. However, this dynamic changes as the forecasting horizon extends. These prior studies generally show that nonlinear machine learning models tend to outperform their linear counterparts, particularly when dealing with medium to long forecasting horizons.

Notably, surpassing all other models, the fNTK model exhibits remarkable performance gains for forecasting horizons beginning at $h = 5$, especially for extended prediction horizons. For instance, in the case of a 20-day ahead forecast, fNTK outperforms both fLinK and classical models CW and AHBS by approximately 45%. Additionally, fNTK surpasses the two other nonlinear kernel models, fGauK and fLapK, by around 20% in terms of RMSE. The out-of-sample R^2 (OoR^2) of fNTK consistently outperforms all other models, firmly establishing its superiority. The good performance of fNTK is further demonstrated in Figure 4, where we visualize RMSE and OoR^2 across various forecasting horizons, accompanied by corresponding 95% confidence intervals derived using block bootstrapping.[16] In addition to RMSE and OoR^2, we provide insights into the model performance through mean absolute percentage error (MAPE) and mean correct prediction of the direction of change (MCPDC) in Table A.3. These metrics exhibit similar patterns to RMSE, with fNTK consistently outperforming parametric and classical models, except for fGauK.[17] These findings remain robust when varying the training sample size to 1000 or 2000 days (Table A.4).

To ensure robustness, we also conduct an analysis of model performance on an annual basis, as presented in Table 1. The consistent superiority of nonlinear models for extended forecasting horizons is evident across three distinct periods: 2019, the year prior to the Covid-19 pandemic; 2020, marked by pandemic-induced market volatility; and 2021, a year of market recovery. Remarkably, nonlinear models, with fNTK leading the way, exhibit enhanced performance relative to fLinK, AHBS, CW, and fRW, especially during the turbulent year of 2020. The performance

[16] The 95% confidence intervals are obtained using the non-overlapping block bootstrap method with a bootstrap length of 20 days.

[17] Diebold-Mariano tests validate the statistical significance of fNTK's superior performance in terms of both MAPE and RMSE for extended forecasting horizons (Tables A.5 and A.6).

	RMSE					OoR^2				
	$h=1$	$h=5$	$h=10$	$h=15$	$h=20$	$h=1$	$h=5$	$h=10$	$h=15$	$h=20$
Overall (from Jan 9, 2019 to Dec 31, 2021)										
CW	3.57	5.45	7.23	8.87	9.91	88.74	74.18	54.50	32.01	15.21
AHBS	3.72	5.53	7.37	8.69	9.55	87.79	73.35	52.87	34.83	21.33
fRW	3.65	5.72	7.66	9.27	10.57	88.28	71.59	48.97	25.69	3.49
fLinK	**3.48**	5.31	7.29	8.41	9.40	**89.19**	75.46	53.78	38.76	23.59
fGauK	9.23	9.38	9.38	8.95	7.67	25.54	23.32	23.67	30.88	49.32
fLapK	7.51	8.32	8.05	7.99	7.49	50.70	39.71	43.74	44.90	51.63
fNTK	4.11	**5.07**	**6.03**	**6.32**	**5.73**	85.24	**77.24**	**68.17**	**65.38**	**71.67**
From Jan 9, 2019 to Dec 31, 2019										
CW	1.53	2.23	2.65	2.91	3.19	86.13	70.52	58.33	50.05	40.40
AHBS	1.62	2.28	2.66	2.91	3.04	84.54	69.35	58.10	49.89	45.86
fRW	1.51	2.39	2.89	3.19	3.51	86.62	66.28	50.48	39.74	27.68
fLinK	**1.41**	2.15	2.48	2.75	2.86	**88.28**	72.65	63.63	55.34	51.97
fGauK	1.56	**1.88**	**1.82**	**1.92**	**1.91**	85.67	**79.15**	**80.43**	**78.16**	**78.48**
fLapK	1.61	1.92	2.01	2.07	2.10	84.65	78.26	75.97	74.70	74.01
fNTK	1.66	1.95	1.97	1.96	1.96	83.73	77.43	76.99	77.15	77.43
From Jan 1, 2020 to Dec 31, 2020										
CW	5.29	8.11	10.95	13.57	15.19	82.92	59.91	24.89	-16.35	-48.49
AHBS	5.46	8.17	11.11	13.20	14.54	81.81	59.28	23.18	-10.04	-35.87
fRW	5.43	8.43	11.53	14.10	16.15	82.00	56.78	17.11	-25.61	-67.87
fLinK	**5.21**	7.86	11.02	12.78	14.38	**83.14**	62.38	23.93	-3.81	-33.68
fGauK	14.74	14.87	14.80	13.98	11.85	-31.32	-35.54	-36.20	-23.39	9.91
fLapK	11.92	13.11	12.55	12.33	11.50	14.09	-5.34	1.89	3.99	15.23
fNTK	6.16	**7.65**	**9.23**	**9.68**	**8.68**	77.18	**63.24**	**46.11**	**40.48**	**51.41**
From Jan 1, 2021 to Dec 31, 2021										
CW	1.97	2.81	3.22	3.40	3.49	89.69	78.95	72.39	69.14	66.14
AHBS	2.18	3.06	3.47	3.66	3.80	87.42	75.14	67.94	64.42	59.98
fRW	1.96	3.14	3.63	3.83	3.95	89.79	73.92	65.01	61.03	56.83
fLinK	**1.83**	2.91	3.44	3.63	3.61	**91.08**	77.54	68.52	65.00	63.98
fGauK	1.96	**2.29**	**2.35**	2.59	2.56	89.79	**86.17**	**85.28**	82.16	81.80
fLapK	1.89	2.37	2.63	2.90	2.76	90.54	85.16	81.59	77.71	78.86
fNTK	2.06	2.47	2.50	**2.48**	**2.39**	88.72	83.81	83.38	**83.61**	**84.21**

Table 1: Prediction accuracy for all models in different forecasting horizons. Results are for four prediction periods: overall (from Jan 09, 2019 to Dec 31, 2021), from Jan 09, 2019 to Dec 31, 2019, from Jan 01, 2020 to Dec 31, 2020, and from Jan 01, 2020 to Dec 31, 2020. Bold numbers indicate the best-performing model (or models) in a given column.

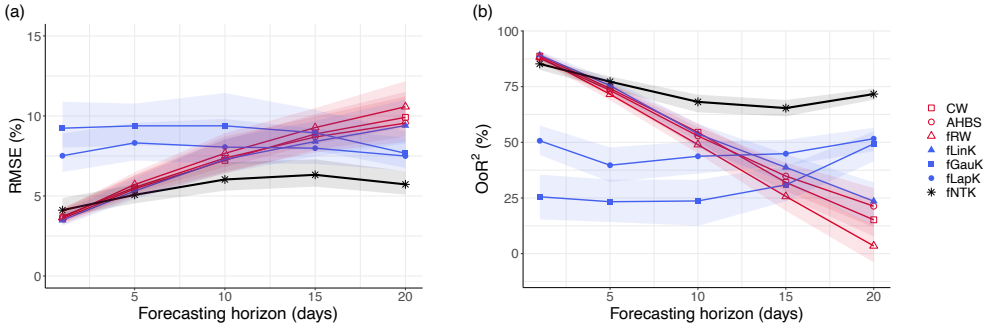


Figure 4: Prediction measurement in terms of (a) RMSE, (b) OoR^2 of all models for forecasting horizons $h = 1, 5, 15, 10$ and 20. The prediction period is from Jan 09, 2019 to Dec 31, 2021. Shaded areas are the 95% confidence intervals using block bootstrapping.

gap widens during periods of heightened market volatility, as exemplified in 2020. For instance, in the 20-day ahead prediction scenario in 2020, fNTK demonstrates an OoR^2 that is approximately 170% to 240% higher than that of linear and classical models, underscoring its dominant performance during times of increased uncertainty. We further investigate the robustness of these models with respect to hyperparameters, moneyness, and maturity groups, as detailed in Appendix A.9.1. These findings collectively emphasize the robustness of nonlinear models, particularly fNTK, in capturing the intricate dynamics of implied volatility surfaces across diverse market conditions. They also provide valuable insights into the interplay between model complexity and performance across various prediction horizons. In essence, our results highlight the capability of fNTK to capture the intricate dynamics of implied volatility surfaces across a range of forecasting horizons and performance metrics. Its capacity to outperform alternative models underscores its effectiveness in handling the complex nonlinearities inherent in financial data.

5 Economic Value of Predictability

This section investigates the economic value of an accurate IV forecaster. By leveraging the enhanced predictability offered by the fNTK model, we demonstrate its potential benefits for investors in the options market. Our evaluation encompasses various trading strategies: call and put delta-hedging and delta-neutral straddle strategies, both short and long, that exploit the h-step ahead forecasts of IVS. We also conduct robustness checks under different test periods, transaction costs, and varying filtering thresholds in the trading.

5.1 Trading strategies

The trading strategies we consider enable option traders to focus on fluctuations in volatility. The delta hedging strategy is specifically designed to mitigate the risk associated with underlying asset price movements by establishing a corresponding position in an option and the asset. Conversely, the delta-neutral straddle strategy relies on option deltas and attains directional neutrality by bal-

ancing opposite exposures in the underlying and simultaneously investing in a weighted combination of call and put options with the same characteristics, see Gao, Xing, and Zhang (2018). An investor who holds a short position is essentially betting against volatility, while someone with a long position is betting that volatility will increase (see Coval and Shumway, 2001; Bakshi and Kapadia, 2003; Driessen and Maenhout, 2007). Our setup relies on trading signals extracted from the predicted IVS and includes only non-zero volume options at time t. An increase (decrease) in the implied volatility is equivalent to an increase (decrease) in the options prices. For brevity, we describe the straddle strategies below and delegate the details of delta-hedging strategies to Appendix A.10.

Delta-neutral straddles. For the short delta-neutral straddle strategy, on each day t, we short w_t units of call option and $(1-w_t)$ units of put option of the same strike and expiration date, where the weights $w_t = -\Delta_t^{sp}/(\Delta_t^{sc} - \Delta_t^{sp})$ and $1-w_t$ are used to ensure the straddle delta equal 0. We trade a pair of put and call options if the IV of both options is predicted to decrease on the day $t+h$, see Gao, Xing, and Zhang (2018). Recall that put and call IVS are modeled and predicted separately. Define Q_t^s as the set of put and call option pairs to be traded. On day t, we sell all the pairs of options in Q_t^s and gain a cash inflow of $\sum_{i\in Q_t^s}(w_{i,t}C_{i,t} + (1-w_{i,t})P_{i,t})$, and close off the positions by paying a cost of $\sum_{i\in Q_t^s}\left(w_{i,t}C_{i,t+h} + (1-w_{i,t})P_{i,t+h}\right)$ on day $t+h$. Similarly, in the long delta-neutral straddle strategy, we only trade a pair of put and call options if the IV of both options is predicted to increase on the day $t+h$. The returns R_t^s of the short and R_t^l of the long delta-neutral straddle portfolios are

$$R_t^s = 1 - \frac{\sum_{i\in Q_t^s}(w_{i,t}C_{i,t+h} + (1-w_{i,t})P_{i,t+h})}{\sum_{i\in Q_t^s}(w_{i,t}C_{i,t} + (1-w_{i,t})P_{i,t})} \quad \text{and} \quad R_t^\ell = \frac{\sum_{i\in Q_t^\ell}(w_{i,t}C_{i,t+h} + (1-w_{i,t})P_{i,t+h})}{\sum_{i\in Q_t^\ell}(w_{i,t}C_{i,t} + (1-w_{i,t})P_{i,t})} - 1.$$

We use a filtering threshold of 0.5% deviation in implied volatility for each trading strategy. Specifically, on a given day t, if the implied volatility of an option is predicted to increase (or decrease) by at least 0.5% on day $t+h$, we buy (or sell) the option. The practice of using filtering thresholds to avoid noisy signals is also used in Goncalves and Guidolin (2006) and Liu et al.

(2021). In the robustness check, we also investigate a threshold of 1% and 5%. Additionally, to assess whether trading strategies remain profitable after accounting for transaction costs, we measure these costs using effective bid-ask spreads when selling and buying put and call options, as in Zhan et al. (2022). Given the absence of transaction data, we will recompute option returns using assumed effective option spreads equal to 50%, 75%, and 100% of the quoted spread.

5.2 Performance of Trading Strategies

In evaluating the performance of our trading strategies, we report two key metrics: mean returns (MR) as a percentage and the annualized Sharpe ratio (SR).[18] We present the detailed performance of the short delta-neutral straddle strategy exclusively in the main text due to space constraints. However, we will provide a general overview of the performance of the remaining strategies. More detailed performance metrics for these additional strategies can be found in Appendix A.11.

The upper section of Table 2 and Figure 5 present the results for the short delta-neutral straddles. It is evident that the performance of fNTK stands out prominently, surpassing most benchmark strategies across various forecasting horizons. It exhibits remarkable mean returns and Sharpe ratios, with a weekly mean return of 4.68% and a monthly return of 11.91%. This represents an improvement ranging from 38% to 173% in returns and a Sharpe ratio ranging between 1.45 and 2.02. These results translate to a substantial 77% to 583% relative enhancement in trading outcomes compared to a random walk benchmark. While other nonlinear kernel models also produce impressive results, fNTK remains the clear leader in terms of overall performance.

Further analysis of performance over multiple years, as shown in Table 2, reveals that during periods of high market volatility, most benchmarks experience a decrease in performance, resulting in negative returns and Sharpe ratios, particularly in 2020. In contrast, fNTK consistently maintains a positive Sharpe ratio, ranging between 0.37 and 0.61 across all trading horizons. All

[18] We base our analysis on simple returns R_t. The excess return is calculated as $ER_t = R_t - \left(\exp\left(\frac{h \cdot r_{t,h}}{252}\right) - 1\right)$, where $r_{t,h}$ represents the annualized riskless interest rate with a time-to-maturity of h days. The Sharpe ratio is defined as $\frac{\mu_{ER}}{\sigma_{ER}}$, where $\mu_{ER} = \frac{1}{N_{traded}} \sum_{t=1}^{N_{traded}} ER_t$ and $\sigma_{ER} = \sqrt{\frac{1}{N_{traded}-1} \sum_{t=1}^{N_{traded}} (ER_t - \mu_{ER})^2}$ denote the mean and standard deviation of excess returns, and N_{traded} is the number of traded days.

	Mean return (%)					Sharpe ratio				
	$h=1$	$h=5$	$h=10$	$h=15$	$h=20$	$h=1$	$h=5$	$h=10$	$h=15$	$h=20$
Overall (from Jan 9, 2019 to Dec 31, 2021)										
CW	0.21	0.36	-1.10	-2.44	-2.34	0.57	0.14	-0.21	-0.28	-0.19
AHBS	-0.28	-0.32	-1.22	-3.13	-3.84	-0.80	-0.13	-0.20	-0.30	-0.23
fRW	**0.45**	3.37	6.16	5.62	4.25	**1.09**	1.14	0.93	0.50	0.24
fLinK	0.28	0.60	-0.14	-2.05	-2.54	0.66	0.20	-0.03	-0.18	-0.15
fGauK	-0.01	2.68	5.66	7.98	8.65	-0.04	1.12	1.11	1.09	1.13
fLapK	0.01	1.62	5.52	4.09	7.38	0.01	0.60	0.99	0.40	0.68
fNTK	-0.03	**4.68**	**8.23**	**9.94**	**11.91**	-0.07	**2.02**	**1.78**	**1.45**	**1.64**
From Jan 9, 2019 to Dec 31, 2019										
CW	0.33	0.81	1.21	1.86	3.72	1.13	0.51	0.51	0.64	0.89
AHBS	0.03	1.26	4.74	5.84	9.47	0.06	0.60	1.42	1.37	1.65
fRW	**0.75**	4.23	**9.72**	10.45	11.10	**2.12**	2.06	2.68	2.36	2.03
fLinK	0.49	1.43	5.32	8.69	9.44	1.34	0.68	1.67	2.25	2.04
fGauK	0.05	4.27	7.80	**10.92**	10.80	0.10	1.98	2.44	**2.59**	2.10
fLapK	0.42	2.54	7.77	10.69	11.54	1.07	1.24	2.58	2.52	2.25
fNTK	0.57	**6.42**	8.22	8.96	**11.67**	1.36	**2.99**	**3.07**	2.27	**2.38**
From Jan 1, 2020 to Dec 31, 2020										
CW	**-0.03**	-1.89	-9.31	-17.49	-23.15	**-0.06**	-0.49	-1.01	-1.16	-1.08
AHBS	-0.62	-2.77	-8.07	-12.52	-16.38	-1.36	-0.74	-0.91	-0.85	-0.69
fRW	-0.30	-1.96	-6.41	-15.77	-26.09	-0.63	-0.46	-0.63	-0.90	-0.93
fLinK	-0.34	-2.85	-9.27	-19.00	-24.42	-0.69	-0.68	-0.96	-1.09	-0.92
fGauK	-0.51	-0.81	-1.48	-1.37	0.33	-1.04	-0.27	-0.22	-0.14	0.03
fLapK	-0.74	-1.44	-1.29	-9.61	-3.64	-1.59	-0.40	-0.17	-0.68	-0.25
fNTK	-0.60	**1.81**	**3.26**	**3.67**	**5.49**	-1.21	**0.61**	**0.50**	**0.37**	**0.54**
From Jan 1, 2021 to Dec 31, 2021										
CW	0.33	2.01	3.53	5.75	8.79	1.27	2.18	2.50	2.74	2.68
AHBS	-0.23	0.99	2.69	4.50	6.31	-0.82	0.98	1.79	2.07	2.88
fRW	**0.97**	**7.74**	**14.79**	**19.51**	**23.64**	**2.52**	**4.72**	**5.75**	4.96	**5.40**
fLinK	0.76	4.11	7.40	11.93	20.60	2.04	2.88	2.90	3.33	4.15
fGauK	0.52	5.54	13.14	17.75	19.08	1.37	3.93	4.94	4.66	4.94
fLapK	0.56	4.77	13.04	18.14	20.43	1.48	3.50	5.12	**5.29**	5.24
fNTK	0.05	6.38	14.30	18.16	19.07	0.13	4.60	5.37	4.83	4.41

Table 2: Mean simple returns (%) and annualized Sharpe ratio of short delta-neutral straddles over the whole test period, from Jan 9, 2019 to Dec 31, 2021, and over each year (2019, 2020, and 2021) of the test period. Bold numbers indicate the best-performing model (or models) in a given column.

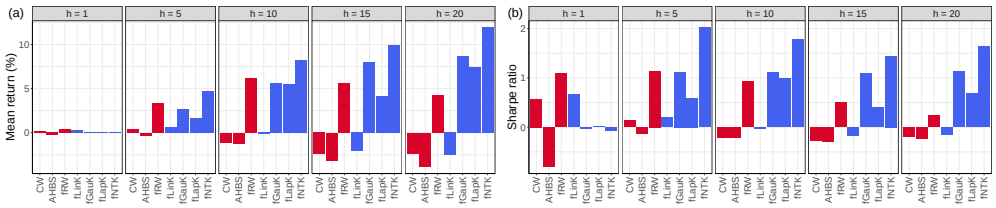


Figure 5: Mean simple returns (%) and annualized Sharpe ratio of short delta-neutral straddle strategy. The prediction period is from Jan 09, 2019 to Dec 31, 2021. The blue color is for functional kernel models, while the red color is to indicate classical models.

nonlinear kernel models yield profitable strategies in 2019 and 2021, achieving the best Sharpe ratios ranging between 3.07 and 5.37. These compelling findings underscore the critical role of

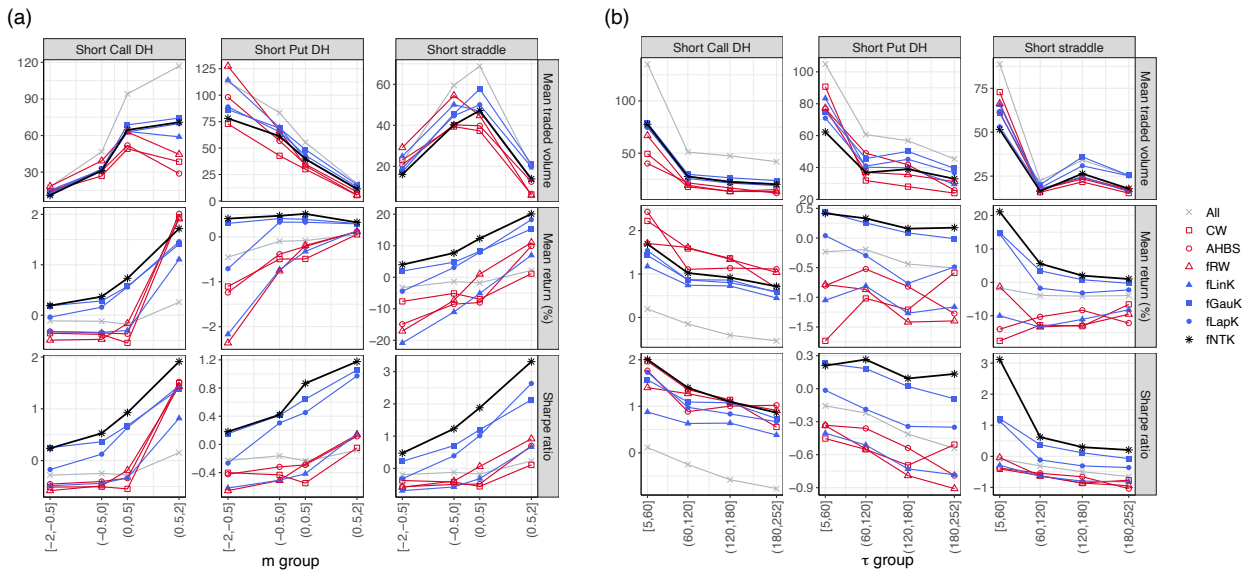


Figure 6: Mean traded volume, mean simple returns (MR) in percentage, Sharpe ratio (SR) of the models for all the short trading strategies at the forecasting horizon $h = 20$, across different moneyness m groups $[-2,-0.5],(-0.5,0],(0,0.5]$, and $(0.5,2]$, and time-to-maturity τ groups $[5,60],(60,120],(120,180]$, and $(180,252]$. The prediction period is from Jan 09, 2019 to Dec 31, 2021.

nonlinearities in modeling implied volatility surface (IVS) dynamics and extracting valuable trading signals. Moreover, they suggest that integrating fNTK into trading strategies is particularly rewarding during turbulent market conditions.

When evaluating the performance of a trading strategy, it is essential to go beyond mere aggregate results. Different options, distinguished by their features, carry varying degrees of risk. Therefore, it is crucial to investigate how their predictability impacts strategy performance across different moneyness and maturity groups for a more comprehensive understanding of the strategy's effectiveness. In Figure 6, we present data on the mean daily traded volume, mean returns (MR), and annualized Sharpe ratio for short call delta-hedging, short put delta-hedging, and short delta-neutral straddle strategies, all at the 20-step ahead forecasting horizon, representative of longer forecasting horizons. The results reveal a clear trend: trading options with higher moneyness and shorter time-to-maturity consistently leads to superior trading strategy performance. Notably, the kernel-based models, with fNTK in the lead, exhibit substantial differences from the benchmark strategies. We also show trading returns of all available options on both day t and day $t+h$. The benefits of employing trading signals, particularly those generated by nonlinear kernel models, are

striking across all moneyness and time-to-maturity groups.[19]

Our analysis also considers the robustness of trading strategies regarding transaction costs. The results for short delta-neutral straddles, as presented in Table 3, highlight that fNTK maintains its leading position when transaction costs are introduced. While there is a decrease in performance, it is not substantial, with a roughly 12% reduction in mean returns and a 14% reduction in Sharpe ratio for portfolios with a 20-day horizon and effective spread of 100%. Furthermore, nonlinear models consistently outperform linear and classical models across various spread values, especially for longer forecasting horizons. In the appendix, additional details on this matter are available in Figure A.12. Overall, our results suggest that the trading strategy remains robust in the presence of transaction costs, with a moderate decline in returns and Sharpe ratios. The superiority of nonlinear models persists across different effective spreads and forecasting horizons.

Another critical variable affecting trading returns is the filtering threshold. Higher threshold values require a higher deviation from option implied volatility (IV) on day t for the option to be traded, resulting in fewer trades. We repeat the trading exercises with two additional threshold values: 5% and 10%. The results in Table 4 demonstrate that as threshold values increase, mean returns and Sharpe ratios for nonlinear models generally improve for both short and long delta-neutral straddle strategies. For instance, in the short delta-neutral straddle strategy with a 20-step ahead prediction, the fNTK model exhibits an increase in mean return from 11.91% to 24.02% and an increase in Sharpe ratio from 1.64 to 2.57. In contrast, the fRW model's mean return decreases from 4.25% to -1.85%, and the corresponding Sharpe ratio decreases from 0.24 to -0.08. [20]

In summary, the fNTK model consistently demonstrates strong performance across various trading strategies and market conditions. It outperforms both functional alternatives and classical models in terms of mean simple returns and annualized Sharpe ratios, particularly for strategies involving delta-neutral straddle options. This robustness is evident across different prediction peri-

[19] Detailed results for the long strategies at the 20-step ahead forecasting horizon are reported in Figure A.23.

[20] This trend can potentially be attributed to the fact that nonlinear models exhibit statistically superior MAPE and MCPDC measurements compared to fRW, fLinK, and classical models, particularly for longer horizons, as documented in Appendix A.9. Consequently, when the filtering threshold is raised, nonlinear models tend to capture options with more accurate high deviations, whereas benchmark models may select options with higher errors on day $t+h$.

	Mean return (%)					Sharpe ratio				
	$h=1$	$h=5$	$h=10$	$h=15$	$h=20$	$h=1$	$h=5$	$h=10$	$h=15$	$h=20$
EM = 50%										
CW	-0.55	-0.44	-1.91	-3.27	-3.18	-1.50	-0.19	-0.35	-0.37	-0.25
AHBS	-0.93	-1.06	-2.06	-3.97	-4.67	-2.56	-0.40	-0.34	-0.37	-0.27
fRW	**-0.41**	2.46	5.24	4.75	3.40	**-0.99**	0.82	0.78	0.42	0.19
fLinK	-0.53	-0.35	-1.05	-2.99	-3.48	-1.26	-0.12	-0.16	-0.25	-0.20
fGauK	-0.91	1.77	4.81	7.16	7.87	-2.08	0.73	0.92	0.96	1.01
fLapK	-0.85	0.71	4.62	3.15	6.55	-2.01	0.25	0.81	0.30	0.60
fNTK	-0.93	**3.75**	**7.40**	**9.14**	**11.16**	-2.14	**1.59**	**1.58**	**1.31**	**1.52**
EM = 75%										
CW	-0.93	-0.85	-2.32	-3.68	-3.60	-2.51	-0.35	-0.43	-0.42	-0.28
AHBS	-1.25	-1.44	-2.49	-4.39	-5.09	-3.41	-0.54	-0.40	-0.41	-0.29
fRW	**-0.85**	2.00	4.77	4.30	2.97	**-2.00**	0.66	0.70	0.38	0.16
fLinK	-0.93	-0.83	-1.51	-3.47	-3.95	-2.22	-0.28	-0.22	-0.29	-0.22
fGauK	-1.36	1.31	4.37	6.75	7.48	-3.06	0.53	0.83	0.90	0.96
fLapK	-1.28	0.24	4.17	2.66	6.13	-3.01	0.08	0.73	0.25	0.56
fNTK	-1.39	**3.27**	**6.98**	**8.73**	**10.78**	-3.17	**1.38**	**1.48**	**1.24**	**1.46**
EM = 100%										
CW	-1.31	-1.25	-2.74	-4.10	-4.03	-3.51	-0.51	-0.50	-0.46	-0.32
AHBS	-1.58	-1.82	-2.91	-4.82	-5.51	-4.24	-0.68	-0.46	-0.45	-0.32
fRW	**-1.29**	1.52	4.30	3.85	2.53	**-2.98**	0.49	0.62	0.33	0.14
fLinK	-1.34	-1.31	-1.97	-3.95	-4.43	-3.17	-0.43	-0.29	-0.33	-0.25
fGauK	-1.82	0.85	3.94	6.33	7.09	-3.99	0.34	0.74	0.84	0.90
fLapK	-1.72	-0.23	3.71	2.17	5.71	-4.00	-0.09	0.64	0.20	0.51
fNTK	-1.85	**2.80**	**6.55**	**8.31**	**10.40**	-4.20	**1.17**	**1.38**	**1.17**	**1.40**

Table 3: Mean returns (%) and annualized Sharpe ratio of short delta-neutral straddles with three levels of effective measurement (50%, 75%, and 100%) of all models over the whole test period, from Jan 9, 2019 to Dec 31, 2021. Bold numbers indicate the best-performing model (or models) in a given column.

	Mean return (%)					Sharpe ratio				
	$h=1$	$h=5$	$h=10$	$h=15$	$h=20$	$h=1$	$h=5$	$h=10$	$h=15$	$h=20$
Threshold = 5%										
CW	**0.93**	0.98	-2.07	-5.81	-7.48	**2.13**	0.31	-0.29	-0.49	-0.45
AHBS	-0.19	-1.27	-4.02	-6.28	-9.51	-0.44	-0.41	-0.54	-0.50	-0.46
fRW	0.49	2.99	5.61	4.83	4.14	0.82	0.79	0.69	0.36	0.20
fLinK	-0.61	-0.04	-1.69	-2.76	-3.22	-0.98	-0.01	-0.20	-0.20	-0.16
fGauK	-0.05	3.77	9.03	11.73	11.84	-0.09	1.29	1.49	1.43	1.04
fLapK	-0.57	1.61	7.23	10.80	13.91	-0.95	0.44	1.00	1.08	1.44
fNTK	-0.58	**6.98**	**12.78**	**13.65**	**17.82**	-0.92	**2.40**	**2.46**	**1.50**	**2.26**
Threshold = 10%										
CW	**-0.30**	-1.62	-9.03	-12.35	-14.52	**-0.45**	-0.38	-0.97	-0.87	-0.79
AHBS	-2.03	-5.64	-8.40	-12.26	-16.62	-2.98	-1.33	-0.86	-0.77	-0.66
fRW	-5.45	-4.73	-4.55	-3.56	-1.85	-7.14	-1.05	-0.48	-0.24	-0.08
fLinK	-5.98	-5.49	-7.44	-9.28	-10.55	-6.90	-1.18	-0.76	-0.60	-0.45
fGauK	-3.20	1.05	9.43	10.94	15.92	-4.12	0.27	1.33	1.06	1.73
fLapK	-4.93	-3.80	4.05	7.72	14.36	-5.89	-0.88	0.46	0.63	1.23
fNTK	-3.42	**7.03**	**16.66**	**14.99**	**24.02**	-3.81	**1.84**	**2.58**	**1.31**	**2.57**

Table 4: Mean simple returns (%) and annualized Sharpe ratio of all models for short delta-neutral straddles, using two different filtering thresholds of 5% and 10% over the whole test period (Jan 09, 2019, to Dec 31, 2021). Bold numbers indicate the best-performing model (or models) in a given column.

ods, even during turbulent market conditions such as the Covid-19 pandemic in 2020. Furthermore, fNTK's superiority holds when accounting for transaction costs, with its positive returns persisting and remaining relatively unaffected. The model's resilience is further underscored by its ability to maintain its advantage as filtering thresholds increase. Overall, the fNTK model's remarkable performance and consistency highlight its potential as a valuable tool for investors in navigating option markets and optimizing trading strategies.

6 Conclusion

This study introduces a novel fNTK (Functional Neural Tangent Kernel) estimator for the Nonlinear Functional AutoRegressive (NFAR) model, focusing on its applicability to the financial options market. Our research uncovers the intricate cross-dependent dynamics between functional responses and multiple lagged functional predictors in this complex domain. Leveraging state-of-the-art kernel-based machine learning techniques, fNTK facilitates the application of NN to functional data. We rigorously draw the connection between NTK and kernel regression through derivation, emphasizing the fNTK's position as a contemporary nonparametric statistical model.

Our empirical findings consistently highlight the superior predictive power of the fNTK-based NFAR model, especially for longer forecasting horizons. These results are not only statistically significant but also carry substantial economic implications. By conducting extensive trading strategy simulations, we demonstrate the real-world value of the fNTK-based NFAR model. It consistently outperforms alternative models, generating superior trading returns and Sharpe ratios across various market conditions. Moreover, its robustness maintains its competitive edge even when factoring in practical transaction costs or filtering thresholds, making it a valuable tool for investors.

While rooted in IVS forecasting, our modeling framework has broader implications. Its generality makes it versatile for applications beyond IV forecasting, heralding a new era in econometric modeling capable of analyzing data with complex, nonlinear dynamics. This work serves as both a technical contribution and a clarion call for the broader analytical community to recognize the

vast potential of accurate forecasting in shaping outcomes across a spectrum of applications.

Some extensions will be considered in future research. First, there are more sophisticated procedures for dimensional reduction. For example, Li and Song (2017) extends the theory of nonlinear sufficient dimension reduction to find the relevant subspace where the relevant information about the relationship between the predictor and response functions lies. Liang, Sun, and Liang (2022) also offers an innovative method for nonlinear sufficient dimension reduction for large-scale data using stochastic neural networks. This inspires another approach in which we can incorporate dimensional reduction and nonlinear modeling into one step instead of using two separate steps. Second, there has been some work showing that there are equivalent NTKs for different architectures of NNs, such as deep Gaussian processes (see Finocchio and Schmidt-Hieber, 2023).

7 Acknowledgements

The authors would like to thank Cees Diks, Gianluca Finocchio, Gustavo Freire, and Eftychia Solea, and participants to the AAAI 2023 Inaugural Summer Symposium on Artificial Intelligence for FinTech (Singapore), the 25th International Conference on Computational Statistics (London), Statistics for Machine Learning Conference (Prague), ETH Forschungsinstitut für Mathematik, ETH Risk Centre, for their helpful feedback and support. The authors would also like to thank the Ministry of Education Singapore for their research funding under the project A-8000014-00-00.

References

Aït-Sahalia, Y., C. Li, and C. X. Li (2021a). "Closed-form implied volatility surfaces for stochastic volatility models with jumps". *Journal of Econometrics*, 222.1, 364–392.

Aït-Sahalia, Y., C. Li, and C. X. Li (2021b). "Implied stochastic volatility models". *The Review of Financial Studies*, 34.1, 394–450.

Almeida, C., J. Fan, G. Freire, and F. Tang (2022). "Can a Machine Correct Option Pricing Models?" *Journal of business & economic statistics: a publication of the American Statistical Association*, 1–14.

Andersen, T. G., N. Fusari, and V. Todorov (2017). "Short-term market risks implied by weekly options". *The Journal of Finance*, 72.3, 1335–1386.

Arora, S., S. S. Du, W. Hu, Z. Li, R. R. Salakhutdinov, and R. Wang (2019). "On exact computation with an infinitely wide neural net". *Advances in neural information processing systems*, 32.

Audrino, F. and D. Colangelo (2010). "Semi-parametric forecasts of the implied volatility surface using regression trees". *Statistics and Computing*, 20.4, 421–434.

Bakshi, G. and N. Kapadia (2003). "Delta-Hedged Gains and the Negative Market Volatility Risk Premium". *The review of financial studies*, 16.2, 527–566.

Bernales, A. and M. Guidolin (2014). "Can we forecast the implied volatility surface dynamics of equity options? Predictability and economic value tests". *Journal of Banking & Finance*, 46, 326–342.

Bernales, A. and M. Guidolin (2015). "Learning to smile: Can rational learning explain predictable dynamics in the implied volatility surface?" *Journal of Financial Markets*, 26, 1–37.

Büchner, M. and B. Kelly (2022). "A factor model for option returns". *Journal of Financial Economics*, 143.3, 1140–1161.

Cai, C. X., M. Kim, Y. Shin, and Q. Zhang (2019). "FARVaR: functional autoregressive value-at-risk". *Journal of Financial Econometrics*, 17.2, 284–337.

Carr, P. and L. Wu (2016). "Analyzing volatility risk and risk premium in option contracts: A new theory". *Journal of Financial Economics*, 120.1, 1–20.

Chen, Y., W. S. Chua, and W. K. Härdle (2019). "Forecasting limit order book liquidity supply–demand curves with functional autoregressive dynamics". *Quantitative Finance*, 19.9, 1473–1489.

Chen, Y., T. Koch, K. G. Lim, X. Xu, and N. Zakiyeva (2021). "A review study of functional autoregressive models with application to energy forecasting". *Wiley Interdisciplinary Reviews: Computational Statistics*, 13.3, e1525.

Chen, Z., Y. Cao, Q. Gu, and T. Zhang (2020). "A generalized neural tangent kernel analysis for two-layer neural networks". *Advances in Neural Information Processing Systems*, 33, 13363–13373.

Christensen, K., M. Siggaard, and B. Veliyev (2022). "A Machine Learning Approach to Volatility Forecasting". *Journal of Financial Econometrics*.

Cont, R. and J. Da Fonseca (2002). "Dynamics of implied volatility surfaces". *Quantitative finance*, 2.1, 45.

Coval, J. D. and T. Shumway (2001). "Expected option returns". *The Journal of finance*, 56.3, 983–1009.

Daniely, A., R. Frostig, and Y. Singer (2016). "Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity". *Advances in neural information processing systems*, 29.

Domingos, P. (Nov. 2020). "Every Model Learned by Gradient Descent Is Approximately a Kernel Machine". *arXiv*.

Driessen, J. and P. Maenhout (2007). "An Empirical Portfolio Perspective on Option Pricing Anomalies". *Review of Finance*, 11.4, 561–603.

Dumas, B., J. Fleming, and R. E. Whaley (1998). "Implied volatility functions: Empirical tests". *The Journal of Finance*, 53.6, 2059–2106.

Fengler, M. R. (2012). *Option data and modeling BSM implied volatility*. Springer.

Fengler, M. R., W. K. Härdle, and E. Mammen (2007). "A semiparametric factor model for implied volatility surface dynamics". *Journal of Financial Econometrics*, 5.2, 189–218.

Fengler, M. R. and L.-Y. Hin (2015). "Semi-nonparametric estimation of the call-option price surface under strike and time-to-expiry no-arbitrage constraints". *Journal of Econometrics*, 184.2, 242–261.

Finocchio, G. and J. Schmidt-Hieber (2023). "Posterior contraction for deep Gaussian process priors". *Journal of Machine Learning Research*, 24.66, 1–49.

Gao, C., Y. Xing, and X. Zhang (2018). "Anticipating uncertainty: straddles around earnings announcements". *Journal of Financial and Quantitative Analysis*, 53.6, 2587–2617.

Goncalves, S. and M. Guidolin (2006). "Predictable dynamics in the S&P 500 index options implied volatility surface". *The Journal of Business*, 79.3, 1591–1635.

Grith, M., H. Wagner, W. K. Härdle, and A. Kneip (2018). "Functional principal component analysis for derivatives of multivariate curves". *Statistica Sinica*, 28.4, 2469–2496.

Gu, S., B. Kelly, and D. Xiu (2020). "Empirical asset pricing via machine learning". *The Review of Financial Studies*, 33.5, 2223–2273.

Jacot, A., F. Gabriel, and C. Hongler (2018). "Neural tangent kernel: Convergence and generalization in neural networks". *Advances in neural information processing systems*, 31.

Kadri, H., E. Duflos, P. Preux, S. Canu, and M. Davy (2010). "Nonlinear functional regression: a functional RKHS approach". *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings, 374–380.

Kearney, F., H. L. Shang, and L. Sheenan (2019). "Implied volatility surface predictability: The case of commodity markets". *Journal of Banking & Finance*, 108, 105657.

Lee, J., S. Schoenholz, J. Pennington, B. Adlam, L. Xiao, R. Novak, and J. Sohl-Dickstein (2020). "Finite versus infinite neural networks: an empirical study". *Advances in Neural Information Processing Systems*, 33, 15156–15172.

Li, B. and J. Song (June 2017). "Nonlinear sufficient dimension reduction for functional data". *The Annals of Statistics*, 45.3, 1059–1095.

Li, Y., C. Huang, and W. K. Härdle (2019). "Spatial functional principal component analysis with applications to brain image data". *Journal of Multivariate Analysis*, 170, 263–274.

Liang, S., Y. Sun, and F. Liang (2022). "Nonlinear Sufficient Dimension Reduction with a Stochastic Neural Network". *Advances in Neural Information Processing Systems*, 35, 27360–27373.

Liu, D., Y. Liang, L. Zhang, P. Lung, and R. Ullah (2021). "Implied volatility forecast and option trading strategy". *International Review of Economics & Finance*, 71, 943–954.

Nunes, M., E. Gerding, F. McGroarty, and M. Niranjan (2019). "A comparison of multitask and single task learning with artificial neural networks for yield curve forecasting". *Expert Systems with Applications*, 119, 362–375.

Park, B. U., E. Mammen, W. Härdle, and S. Borak (2009). "Time series modelling with semiparametric factor dynamics". *Journal of the American Statistical Association*, 485, 284–298.

Petersen, A. and H.-G. Müller (2016). "Functional data analysis for density functions by transformation to a Hilbert space". *Annals of statistics*, 44.1, 183–218.

Sang, P. and B. Li (2022). "Nonlinear function-on-function regression by RKHS". *arXiv preprint arXiv:2207.08211*.

Solea, E. and B. Li (2022). "Copula Gaussian graphical models for functional data". *Journal of the American Statistical Association*, 117.538, 781–793.

Ulrich, M. and S. Walther (2020). "Option-implied information: What's the vol surface got to do with it?" *Review of Derivatives Research*, 23.3, 323–355.

Xu, M., J. Li, and Y. Chen (2017). "Varying coefficient functional autoregressive model with application to the US treasuries". *Journal of Multivariate Analysis*, 159, 168–183.

Xu, X., Y. Chen, G. Zhang, and T. Koch (2022). "Modeling functional time series and mixed-type predictors with partially functional autoregressions". *Journal of Business & Economic Statistics*, 1–18.

Zhan, X., B. Han, J. Cao, and Q. Tong (2022). "Option return predictability". *The Review of Financial Studies*, 35.3, 1394–1442.

Zhang, W., L. Li, and G. Zhang (2023). "A two-step framework for arbitrage-free prediction of the implied volatility surface". *Quantitative Finance*, 23.1, 21–34.

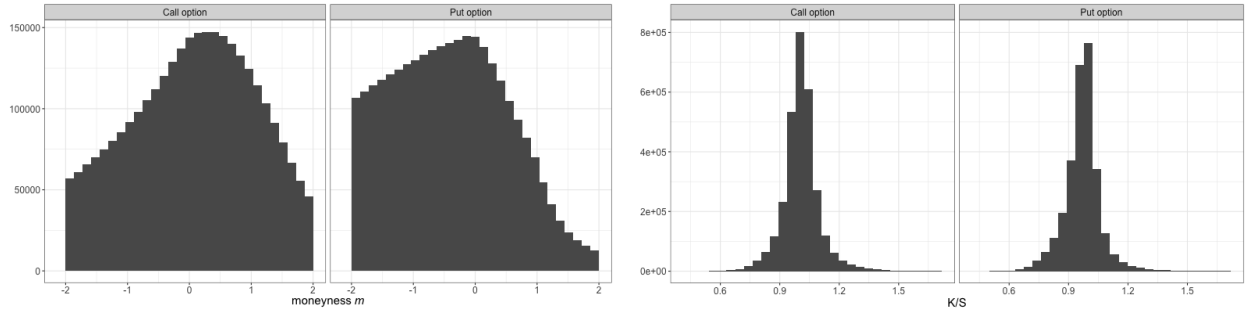
Appendix

A Appendix

A.1 Summary statistics

Panel A: Call options					
	$5 \leq \tau \leq 60$	$60 < \tau \leq 120$	$120 < \tau \leq 180$	$180 < \tau$	Total
$m < -0.5$					
Contract (%)	11.46	3.04	1.55	1.33	17.38
Average IV	0.22	0.26	0.27	0.27	0.24
$\|m\| \leq 0.5$					
Contract (%)	19.87	7.09	3.09	3.72	33.77
Average IV	0.17	0.19	0.19	0.19	0.18
$0.5 < m$					
Contract (%)	32.59	8.12	4.03	4.11	48.85
Average IV	0.15	0.17	0.17	0.17	0.16
Total					
Contract (%)	63.92	18.25	8.68	9.15	100.00
Average IV	0.17	0.19	0.20	0.19	0.18
Panel B: Put options					
	$5 \leq \tau \leq 60$	$60 < \tau \leq 120$	$120 < \tau \leq 180$	$180 < \tau$	Total
$m < -0.5$					
Contract (%)	18.75	5.84	2.49	2.75	29.83
Average IV	0.23	0.26	0.28	0.28	0.24
$\|m\| \leq 0.5$					
Contract (%)	20.51	7.21	3.09	3.63	34.45
Average IV	0.18	0.19	0.20	0.19	0.18
$0.5 < m$					
Contract (%)	23.03	6.56	3.07	3.07	35.72
Average IV	0.19	0.20	0.20	0.20	0.19
Total					
Contract (%)	62.29	19.62	8.65	9.45	100.00
Average IV	0.20	0.22	0.22	0.22	0.20

Table A.1: Summary Statistics for Implied Volatilities by maturity and moneyness. Percentage of contracts and mean of IV of call options (Panel A) and put options (Panel B), over different combinations of time-to-maturity (τ, in days) and moneyness (m) between January 1, 2009, and December 31, 2021.



(a) Moneyness m **(b)** Strike over spot ratio K/S

Figure A.1: Histograms of moneyness m and strike over spot ratio K/S for call and put options.

A.2 Technical proofs

Theorem 1 (Isomorphism between Reproducing Kernel Hilbert Spaces). *Under Equations (6) and (7), it holds that*

$$\begin{aligned} k(\boldsymbol{x}_i, \boldsymbol{x}_j) &= \langle k(.,\boldsymbol{x}_i), k(.,\boldsymbol{x}_j)\rangle \\ &= \langle K(.,X_i), K(.,X_j)\rangle_{\mathfrak{M}_X} = K(X_i,X_j). \end{aligned} \tag{1}$$

Then the RKHS $\mathfrak{M}_X$ nested on $\mathscr{H}_X$ is isometrically isomorphic to the RKHS $\mathfrak{M}_{\boldsymbol{x}}$ nested on $\mathscr{H}_{\boldsymbol{x}}$.

Proof. Given that $\langle \psi_\ell, \psi_v\rangle = \delta_{\ell v}$, the Kronecker delta, it follows that

$$\begin{aligned} \langle X_i, X_j\rangle_{\mathscr{H}_X} &= \langle \sum_{\ell=1}^{\infty} x_{i\ell}\psi_\ell, \sum_{v=1}^{\infty} x_{jv}\psi_k\rangle_{\mathscr{H}_X} \\ &= \sum_{\ell=1}^{\infty}\sum_{v=1}^{\infty} x_{i\ell}x_{jv}\langle \psi_\ell, \psi_v\rangle_{\mathscr{H}_X} \\ &= \sum_{v=1}^{\infty} x_{iv}x_{jv} \\ &= \langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle. \end{aligned} \tag{2}$$

For $k : \mathscr{H}_{\boldsymbol{x}} \times \mathscr{H}_{\boldsymbol{x}} \to \mathbb{R}$ that satisfies (7) and by Equation (2)

$$\begin{aligned} k(\boldsymbol{x}_i, \boldsymbol{x}_j) &= \rho(\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle, \langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle, \langle \boldsymbol{x}_j, \boldsymbol{x}_j\rangle) \\ &= \rho(\langle X_i, X_i\rangle_{\mathscr{H}_X}, \langle X_i, X_j\rangle_{\mathscr{H}_X}, \langle X_j, X_j\rangle_{\mathscr{H}_X}) \\ &= K(X_i, X_j). \end{aligned}$$

Let $\mathfrak{M}_{\boldsymbol{x}}$ be the RKHS generated by $k(.,\boldsymbol{x})$, then we also have

$$\begin{aligned} K(X_i,X_j) &= \langle K(.,X_i), K(.,X_j)\rangle_{\mathfrak{M}_X} \\ &= \langle k(.,\boldsymbol{x}_i), k(.,\boldsymbol{x}_j)\rangle = k(\boldsymbol{x}_i,\boldsymbol{x}_j), \end{aligned} \tag{3}$$

The map $T : \mathscr{H}_X \to \mathbb{R}^\infty$ defined as $TX = (\langle X_i, \psi_v\rangle_{\mathscr{H}_X})_{v=1,2,\ldots}$ is a bijective linear mapping,

and is also distance-preserving as shown in Equation (3). Thus, by extension of Lemma 4.2. of Klepsch and Klüppelberg (2017), the Hilbert space $\mathfrak{M}_X$ of $K(.,X)$ is isometrically isomorphic to the $\mathfrak{M}_{\boldsymbol{x}}$ space of $k(.,\boldsymbol{x})$. Moreover, the function process $\{X_i\}$ is isometrically isomorphic to the vector process $\{x_i\}$.

□

Theorem 2 (**Vector-to-vector regression**). *Given the decomposition of X_i in Equations (6) and Y_i in Equation (9), under Assumptions (1) - (3) and Theorem 1, for a positive definite kernel k defined by Equation (7), if there is a covariance matrix $\Sigma_{\boldsymbol{xx}}$ of $k(.,\boldsymbol{x})$ that is diagonal, then the function-to-function regression model in Equation (3) may be represented equivalently by*

$$\beta_0 = \underset{\beta \in \mathscr{B}(\mathscr{H}_{\boldsymbol{y}}, \mathfrak{M}_{\boldsymbol{x}})}{\arg\min} \; \mathbb{E}\left[\|\boldsymbol{y}_i - \beta^* k(.,x_i)\|_2^2\right], \tag{4}$$

with solution $\beta_0 = \Sigma^{\dagger}_{\boldsymbol{xx}}\Sigma_{\boldsymbol{xy}}$. This leads to

$$\begin{aligned} \mathbb{E}[\boldsymbol{y}_i|\boldsymbol{x}] &= \beta_0^* k(.,\boldsymbol{x}_i) \\ &= \Sigma_{\boldsymbol{yx}}\Sigma^{\dagger}_{\boldsymbol{xx}} k(.,\boldsymbol{x}_i) \\ &= \mathbb{E}\left[\{(\Sigma^{\dagger}_{\boldsymbol{xx}} k(.,\boldsymbol{x}_i))(\boldsymbol{x})\}\boldsymbol{y}\right]. \end{aligned} \tag{5}$$

Proof. We first show that we can move from function-to-function to vector-to-function regression. By projecting function Y_i onto the set of orthonormal eigenfunctions $\varphi = (\varphi_1, \varphi_2 \ldots)^T$ where $\varphi_j \in \mathscr{H}_Y$

$$Y_i = \sum_{j=1}^{\infty} y_{ij}\varphi_j, \quad \text{with } y_{ij} = \langle Y_i, \varphi_j \rangle_{\mathscr{H}_Y},$$

and let the number of bases j go to infinity, there is no information loss in the expansion. Given the fixed form of φ, the coefficients y_{ij}-s explain the functional variables uniformly. Since Y_i is centered, its basis coefficient vector $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots)^T$ has zero mean. As $B_0^* K(.,X_i) \in \mathscr{H}_Y$, we can

express it in terms of $\{\varphi_j\}$

$$\begin{aligned}\mathbb{E}[Y_i|X=X_i] = B_0^* K(.,X_i) &= \sum_{j=1}^{\infty} \langle \varphi_j, B_0^* K(.,X_i)\rangle_{\mathscr{H}_Y} \varphi_j \\ &= \sum_{j=1}^{\infty} \langle B_0\varphi_j, K(.,X_i)\rangle_{\mathfrak{M}_X} \varphi_j \\ &= \sum_{j=1}^{\infty} \langle b_{0j}, K(.,X_i)\rangle_{\mathfrak{M}_X} \varphi_j \\ &= \sum_{j=1}^{\infty} b_{0j}(X_i)\varphi_j,\end{aligned} \tag{6}$$

where $b_{0j} = B_0\varphi_j \in \mathfrak{M}_X$ and $b_{0j}(X_i) \in \mathbb{R}$. On the other hand, taking conditional expectation in $Y_i = \sum_{j=1}^{\infty} y_{ij}\varphi_j$ leads to

$$\mathbb{E}[Y_i|X=X_i] = \sum_{j=1}^{\infty} \mathbb{E}[y_{ij}|X=X_i]\varphi_j. \tag{7}$$

Hence, predicting Y_i is equivalent to predicting y_{ij}-s given $X=X_i$ and kernel K. Furthermore, by (6), (7), and the orthonormality of $\{\varphi_j\}$, the predicted value of y_{ij} given $X=X_i \in \mathscr{H}_X$ and kernel K is

$$\mathbb{E}[y_{ij}|X=X_i] = b_{0j}(X_i) = \langle B_0\varphi_j, K(.,X_i)\rangle_{\mathfrak{M}_X}. \tag{8}$$

This means that the original function-to-function regression is equivalent to vector-to-function regression, where we predict the basis coefficient vector $\boldsymbol{y}_i$ of Y_i given the functional input $X=X_i$. Next, we show that we can move from vector-to-function to vector-to-vector regression. Next, we show that we can move from vector-to-function to vector-to-vector regression. For this, we can use different orthogonal projection of $K(.,X_i)$. Although we illustrate the idea for the spectral representation of $K(.,X_i)$, the approach is generalisable to other basis functions in $\mathfrak{M}_X$.

If $\mathbb{E}(\|K(.,X_i)\|^2_{\mathfrak{M}_X}) < \infty$, then $K(.,X_i)$ admits the Karhunen-Loève decomposition

$$K(.,X_i) = \sum_{\ell=1}^{\infty} \zeta_{i\ell}\omega_\ell, \quad \text{with } \zeta_{i\ell} = \langle K(.,X_i), \omega_\ell\rangle_{\mathfrak{M}_X}, \tag{9}$$

with $\{\omega_\ell\}$ the eigenfunctions and $\zeta_{i\ell}$ the $\ell-$th basis coefficient of $K(.,X_i)$.

Assume that $k(.,\boldsymbol{x}_i) = (\zeta_{i1}, \zeta_{i2}, \ldots)^T$ and denote $\boldsymbol{\omega} = (\omega_1, \omega_2, ...)$ the vector of basis functions of $\mathfrak{M}_X$. Then, the covariance operator Σ_{XX} can be decomposed into

$$\begin{aligned}
\Sigma_{XX} &= \mathbb{E}[K(.,X) \otimes K(.,X)] \\
&= \mathbb{E}\Big[\Big(\sum_{\ell=1}^{\infty} \zeta_\ell \omega_\ell\Big) \otimes \Big(\sum_{j=1}^{\infty} \zeta_j \omega_j\Big)\Big] \\
&= \mathbb{E}\Big[\sum_{\ell=1}^{\infty}\sum_{j=1}^{\infty} \zeta_\ell \zeta_j \big(\omega_\ell \otimes \omega_j\big)\Big] \\
&= \boldsymbol{\omega} \Sigma_{\boldsymbol{xx}} \boldsymbol{\omega}^T,
\end{aligned} \tag{10}$$

where $\Sigma_{\boldsymbol{xx}} = \mathbb{E}\big(k(.,\boldsymbol{x}) \otimes k(.,\boldsymbol{x})\big) = \text{diag}\big(\mathbb{E}[\zeta_1^2], \mathbb{E}[\zeta_2^2], ...\big)$ is the covariance matrix of $k(.,\boldsymbol{x})$ and $\Sigma^\dagger_{\boldsymbol{xx}} = \text{diag}\big(\mathbb{E}[\zeta_1^2]^{-1}\mathbb{1}_{\mathbb{E}[\zeta_1^2]>0}, \mathbb{E}[\zeta_2^2]^{-1}\mathbb{1}_{\mathbb{E}[\zeta_2^2]>0}, ...\big)$ is its Moore-Penrose inverse. We can also express $\Sigma^\dagger_{XX}$ in terms of $\Sigma^\dagger_{\boldsymbol{xx}}$ and $\boldsymbol{\omega}$

$$\Sigma^\dagger_{XX} = \boldsymbol{\omega} \Sigma^\dagger_{\boldsymbol{xx}} \boldsymbol{\omega}^T. \tag{11}$$

Multiply both side by ω_ℓ, we obtain

$$\begin{aligned}
\Sigma^\dagger_{XX} \omega_\ell &= \boldsymbol{\omega} \Sigma^\dagger_{\boldsymbol{xx}} (\langle \omega_1, \omega_\ell \rangle, \langle \omega_2, \omega_\ell \rangle, ...)^T \\
&= (\omega_1 \Sigma^\dagger_{\boldsymbol{xx},1}, \omega_2 \Sigma^\dagger_{\boldsymbol{xx},2}, ...) e_\ell \\
&= \omega_\ell \Sigma^\dagger_{\boldsymbol{xx},\ell}.
\end{aligned} \tag{12}$$

where $e_\ell = (0,0,...,0,1,0,...)^T$ is a vector with all elements equal 0 except for the $\ell-$th place,

which is equal to 1, and $\Sigma^{\dagger}_{\boldsymbol{xx},\ell}$ is the $\ell-$th diagonal term of $\Sigma^{\dagger}_{\boldsymbol{xx}}$.

$$\begin{aligned}
\langle \varphi_j, \Sigma_{YX}\omega_\ell\rangle_{\mathscr{H}_Y} &= \langle \varphi_j, \mathbb{E}[Y\otimes(K(.,X))]\omega_\ell\rangle_{\mathscr{H}_Y} \\
&= \langle \varphi_j, \mathbb{E}[Y\langle(K(.,X),\omega_\ell\rangle_{\mathfrak{M}_X}]\rangle_{\mathscr{H}_Y} \\
&= \langle \varphi_j, \mathbb{E}[Y\zeta_\ell]\rangle_{\mathscr{H}_Y} \\
&= \mathbb{E}[\langle \varphi_j, Y\rangle_{\mathscr{H}_Y}\zeta_\ell] \\
&= \mathbb{E}[y_j\zeta_\ell].
\end{aligned} \tag{13}$$

For $k(.,\boldsymbol{x}_i) = (\zeta_{i1}, \zeta_{i2}, \ldots)^T$ the isomorphism also holds

$$\begin{aligned}
K(X_i,X_j) &= \langle K(.,X_i), K(.,X_j)\rangle_{\mathfrak{M}_X} \\
&= \langle \sum_{\ell=1}^{\infty}\zeta_{i\ell}\omega_\ell, \sum_{v=1}^{\infty}\zeta_{iv}\omega_v\rangle_{\mathfrak{M}_X} \\
&= \sum_{\ell=1}^{\infty}\sum_{v=1}^{\infty}\zeta_{i\ell}\zeta_{iv}\langle\omega_\ell,\omega_v\rangle_{\mathfrak{M}_X} \\
&= \sum_{\ell=1}^{\infty}\zeta_{i\ell}\zeta_{l\ell} \\
&= \langle k(.,\boldsymbol{x}_i), k(.,\boldsymbol{x}_j)\rangle \\
&= k(\boldsymbol{x}_i,\boldsymbol{x}_j).
\end{aligned} \tag{14}$$

Now we can further rewrite $b_{0j}(X_i)$ of Equation (8) as

$$
\begin{aligned}
b_{0j}(X_i) &= \langle B_0\varphi_j, K(.,X_i)\rangle_{\mathfrak{M}_X} \\
&= \langle B_0\varphi_j, \sum_{\ell=1}^{\infty} \zeta_{i\ell}\omega_\ell\rangle_{\mathfrak{M}_X} \\
&= \sum_{\ell=1}^{\infty} \langle B_0\varphi_j, \omega_\ell\rangle_{\mathfrak{M}_X}\ \zeta_{i\ell} \\
&= \sum_{\ell=1}^{\infty} \langle \varphi_j, B_0^*\omega_\ell\rangle_{\mathscr{H}_Y}\ \zeta_{i\ell} \\
&= \sum_{\ell=1}^{\infty} \langle \varphi_j, \Sigma_{YX}\Sigma_{XX}^{\dagger}\omega_\ell\rangle_{\mathscr{H}_Y}\ \zeta_{i\ell} \\
&= \sum_{\ell=1}^{\infty} \langle \varphi_j, \Sigma_{YX}\omega_\ell\Sigma_{\boldsymbol{xx},\ell}^{\dagger}\rangle_{\mathscr{H}_Y}\ \zeta_{i\ell} \quad \text{by Equation (12)} \\
&= \sum_{\ell=1}^{\infty} \langle \varphi_j, \Sigma_{YX}\omega_\ell\rangle_{\mathscr{H}_Y}\ \Sigma_{\boldsymbol{xx},\ell}^{\dagger}\ \zeta_{i\ell} \\
&= \sum_{\ell=1}^{\infty} \mathbb{E}[y_j\omega_\ell(\boldsymbol{x})]\ \Sigma_{\boldsymbol{xx},\ell}^{\dagger}\ \zeta_{i\ell} \quad \text{by Equation (13)} \\
&= \mathbb{E}[y_j k(.,\boldsymbol{x})]\ \Sigma_{\boldsymbol{xx}}^{\dagger}\ k(.,\boldsymbol{x}_i) \quad \text{by diagonality of } \Sigma_{\boldsymbol{xx}}^{\dagger} \\
&= \Sigma_{y_j\boldsymbol{x}}\Sigma_{\boldsymbol{xx}}^{\dagger}k(.,\boldsymbol{x}_i),
\end{aligned}
\tag{15}
$$

where $\Sigma_{y_j\boldsymbol{x}} = \mathbb{E}[y_j k(.,\boldsymbol{x})]$. The predicted value of the $j-$th basis coefficient y_{ij} of Y_i given the vector of basis coefficients $\boldsymbol{x}_i$ of X_i is

$$
\begin{aligned}
\mathbb{E}[y_{ij}|\boldsymbol{x}=\boldsymbol{x}_i] &= b_{0j}(X_i) \\
&= \Sigma_{y_j\boldsymbol{x}}\Sigma_{\boldsymbol{xx}}^{\dagger}k(.,\boldsymbol{x}_i) \\
&= \mathbb{E}\left[\{(\Sigma_{\boldsymbol{xx}}^{\dagger}k(.,\boldsymbol{x}_i))(\boldsymbol{x})\}y_j\right].
\end{aligned}
\tag{16}
$$

Thus, the function-to-function regression model can be represented equivalently by the vector-to-vector regression. We predict each element y_{ij} of the basis coefficient vector $\boldsymbol{y}_i$ of Y_i using the basis coefficient vector $\boldsymbol{x}_i$ of X_i and kernel k.

Denote $\beta_0 = \Sigma_{\boldsymbol{xx}}^{\dagger}\Sigma_{\boldsymbol{xy}}$, with the adjoint operator $\beta_0^* = \Sigma_{\boldsymbol{yx}}\Sigma_{\boldsymbol{xx}}^{\dagger}$. Equation (16) can be further

written for predicting vector $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots)^T$

$$\begin{aligned}\mathbb{E}[\boldsymbol{y}_i|\boldsymbol{x} = \boldsymbol{x}_i] &= \mathbb{E}\left[\{(\Sigma^{\dagger}_{\boldsymbol{x}\boldsymbol{x}} k(.,\boldsymbol{x}_i))(\boldsymbol{x})\}\boldsymbol{y}\right] \\ &= \Sigma_{\boldsymbol{y}\boldsymbol{x}}\Sigma^{\dagger}_{\boldsymbol{x}\boldsymbol{x}} k(.,\boldsymbol{x}_i) \\ &= \beta_0^* k(.,\boldsymbol{x}_i).\end{aligned} \tag{17}$$

□

A.3 Dimension reduction via Sieve

Estimating an infinite number of coefficients using a finite sample is computationally infeasible. The sieves method (Chen, 2007) projects the infinite-dimensional process onto a finite parameter space, minimizing information loss. Specifically, we construct sieves, a sequence of subspaces $\{\Theta_s\}$ from the original infinite-dimensional space Θ, which is compact and non-decreasing with each subspace satisfying the condition $\Theta_s \subseteq \Theta_{s+1} \subseteq \cdots \subseteq \Theta$ and the union of these subspaces, $\bigcup_s \Theta_s$, is dense in Θ.

Given a dataset with n observations, the strategy is to select Θ_{K_n}, a parameter space of degree K_n, such that the loss function is well-defined in the finite-dimensional linear space:

$$\Theta_{K_n} = \left\{ f(\tau) \in L^2(\mathscr{C}) \mid f(\tau) = \sum_{k=1}^{K_n} \theta_k \phi_k(\tau), \sum_{k=1}^{K_n} k^2 \theta_k^2 \leq cK_n, \theta_k \in \mathbb{R}, \tau \in \mathscr{C} \right\},$$

where $\{\theta_k\}$ denotes the expansion coefficients for functional terms, and c is a positive constant that controls the growth rate of K_n. We consider K_n as a hyperparameter in the sieve approach and will address its selection in Section A.4. Under the sieves with degree K_n, the (approximated) projection operates within a finite parameter space for $k = 1, \ldots, K_n$.

A.4 Implementation

A.4.1 Functional Principal Component Analysis

Given the smooth curves[1] $\{Y_i, X_i\}$, $i = 1, \ldots, n$, we can estimate the orthonormal bases of $\mathscr{H}_X$ and $\mathscr{H}_Y$ using the empirical mean and covariance operators

$$\begin{aligned} \hat{\mu}_Y(u) &= \frac{1}{n}\sum_{i=1}^{n} Y_i(u), \quad \hat{\mu}_X(v) = \frac{1}{n}\sum_{i=1}^{n} X_i(v), \\ \hat{C}_X(v,v') &= \frac{1}{n}\sum_{i=1}^{n}\{X_i(v) - \hat{\mu}_X(v)\}\{X_i(v') - \hat{\mu}_X(v')\}, \\ \hat{C}_Y(u,u') &= \frac{1}{n}\sum_{i=1}^{n}\{Y_i(u) - \hat{\mu}_Y(u)\}\{Y_i(u') - \hat{\mu}_Y(u')\}. \end{aligned} \tag{18}$$

Throughout the paper, we assume that the eigenfunctions and projection coefficients correspond to decreasing eigenvalues of the covariance operator. By performing functional principal component analysis (FPCA) on $\hat{C}_X$ and $\hat{C}_Y$, we obtain a set of orthonormal basis $\{\hat{\psi}_j(.)\}$ for $\mathscr{H}_X$ and a set of orthonormal basis $\{\hat{\varphi}_j(.)\}$ for $\mathscr{H}_Y$, such that we may write

$$X_i(v) - \hat{\mu}_X(v) \approx \sum_{j=1}^{d_x} \hat{x}_{ij}\hat{\psi}_j(v), \quad Y_i(u) - \hat{\mu}_Y(u) \approx \sum_{j=1}^{d_y} \hat{y}_{ij}\hat{\varphi}_j(u), \tag{19}$$

where the loadings $\hat{x}_{ij} = \int_{\mathscr{I}_x}(X_i(v) - \hat{\mu}_X(v))\hat{\psi}_j(v)dv$ and $\hat{y}_{ij} = \int_{\mathscr{I}_y}(Y_i(u) - \hat{\mu}_Y(u))\hat{\varphi}_j(u)du$;[2] d_x and d_y are the truncated dimensions of the basis coefficient vectors $\hat{\boldsymbol{x}}_i$ and $\hat{\boldsymbol{y}}_i$, respectively, using the sieve method in Section A.3.

A.4.2 fNTK algorithm

We introduce the algorithm for the proposed functional Neural Tangent Kernel (fNTK) approach within the context of the nonlinear functional autoregressive model. This algorithm aims to lever-

[1] Otto and Salish (2022) note that if the discrete observations are dense enough, the eigenfunctions and projection coefficients can be estimated at the same $\sqrt{n}$ rate as if the curves were fully observed.

[2] Since the B-splines are not orthogonal, for the implementation we follow Ramsay and Silverman (2005) and discretize the observed functions X_i to a fine grid of n_x equally spaced values $v_1, \ldots, v_{n_x}$ that span the interval $\mathscr{I}_v$. Similarly for Y_i, we have a fine grid of n_y equally spaced values $u_1, \ldots, u_{n_y}$ that span the interval $\mathscr{I}_u$.

age the power of neural networks to capture intricate relationships within the implied volatility surfaces, aiming for enhanced forecasting accuracy. An illustration of Algorithm 1 is shown in Figure 3. We conduct a simulation study in Appendix A.5, and it shows that the proposed estimation is numerically consistent.

Algorithm 1 fNTK regression

Given training data $\{(Y_i, X_i)\}_{i=1}^n$, predict for (Y_{test}, X_{test}).

1. Perform FPCA on $\hat{C}_Y$ and $\hat{C}_X$ of smoothed surfaces $\{Y_i\}_{i=1}^n$ and $\{X_i\}_{i=1}^n$
2. Truncate the representation to account for 99.99% of variance explained. Obtain set of orthonormal basis $\{\hat{\varphi}_{train,j}(.)\}_{j=1}^{d_y}$ and $\{\hat{\psi}_{train,j}(.)\}_{j=1}^{d_x}$, and set of scores $\hat{\boldsymbol{y}}_{train,i} \in \mathbb{R}^{d_y}$ and $\hat{\boldsymbol{x}}_{train,i} \in \mathbb{R}^{d_x}$, respectively.
3. Train a NTK parameterized neural network with $\{(\hat{\boldsymbol{y}}_{train,i}, \hat{\boldsymbol{x}}_{train,i})\}_{i=1}^n$ to minimize a least-square loss function
4. Use estimated basis functions from training data to get scores $\hat{\boldsymbol{x}}_{test}$ of test data.
5. Use trained NN and $\hat{\boldsymbol{x}}_{test}$ to predict scores of Y_{test}.
6. Using predicted score $\tilde{\boldsymbol{y}}_{test}$, obtain predicted surface $\hat{Y}_{test}(u) = \sum_{j=1}^{d_y} \tilde{y}_{test,j} \hat{\varphi}_{train,j}(u)$, and get predictions for observed values of Y_{test}.

A.5 Simulation study

In this section, we study the finite-sample performance of the proposed estimation in the NFAR modeling framework. We consider settings with different linear and nonlinear dynamics, generated with the parameters of the Ad-hoc Black–Scholes (AHBS) model.

For each day t, for a given moneyness and time-to-maturity (TTM), the implied volatility is simulated with

$$IV_{i,t}(m_{i,t}, \tau_{i,t}) = \alpha_{0,t} + \alpha_{1,t} m_{i,t} + \alpha_{2,t} m_{i,t}^2 + \alpha_{3,t} \tau_{i,t} + \alpha_{4,t} m_{i,t} \tau_{i,t} + \varepsilon_{i,t} \quad \text{for } i = 1, 2, ..., N_t,$$

where $IV_{i,t}(m_{i,t}, \tau_{i,t})$, $m_{i,t}$ and $\tau_{i,t}$ are the simulated implied volatilities, moneyness, and TTM (in years) of the option i on day t, respectively; $\varepsilon_{i,t}$ is the normally distributed random error term

with mean 0 and standard deviation 0.01, and N_t is the number of options available for day t. For simplicity, we assume that the IVS are observed at regularly spaced points on each day t. The moneyness m values are set at 50 equidistant points in the range of -2.5 to 2.5, while for the time-to-maturity τ, we use the sequence of values from 0.02 to 1 by step of 0.05. This mimics our real data setup where we focus on options with $m \in [-2.5, 2.5]$ and time-to-maturity τ of at least 5 trading days and at most one year.

We set $T = 2000$, that is, the sample size, which is split into $T_1 = [1, 1200]$ as the training set to perform in-sample estimation, $T_2 = [1201, 1600]$ as the validation set for tuning hyperparameters, and $T_3 = [1601, 2000]$ as the test set to conduct the out-of-sample prediction. The temporal dependence between IVS is captured by the simulated dynamics of $\alpha = (\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4)$. We consider the two experiments described below. The generation was repeated 100 times in each experiment.

In the first experiment, which is referred to as the Linear experiment, we aim to investigate the performance of functional models under simple linear dynamics. To reflect the values of observed implied volatility of S&P 500 options, each $\alpha_j, i = 0, 1, ..., 4$ is simulated with an autoregressive (AR) model

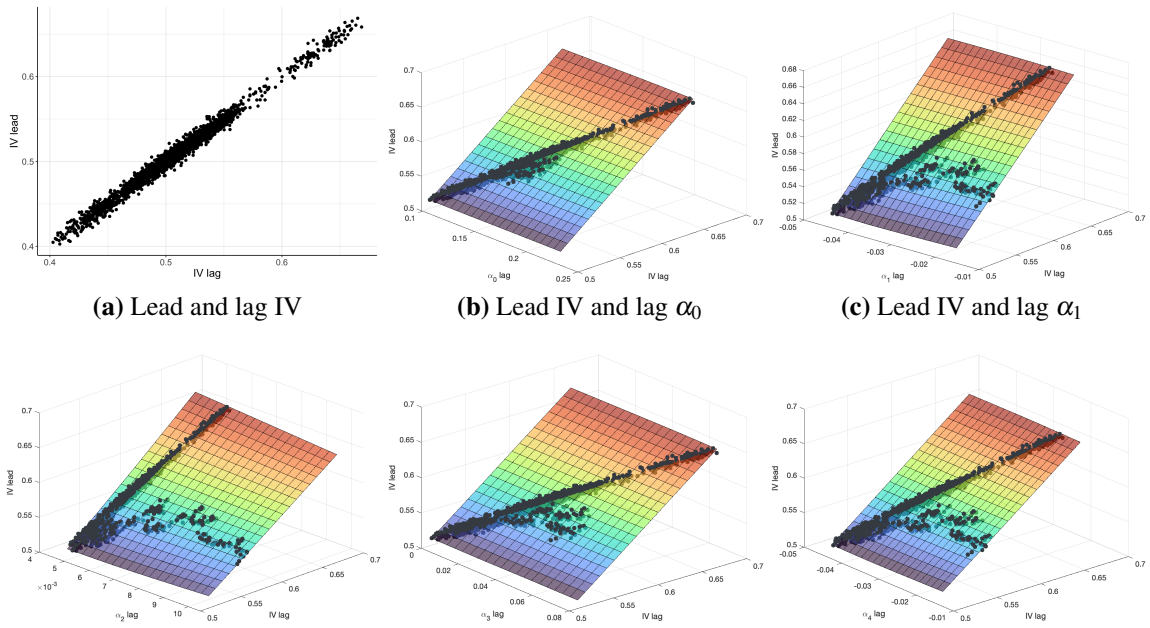
$$\alpha_{j,t} = a_j \alpha_{j,t-1} + e_{j,t},$$

with unconditional mean and variance that match the mean and variance of α_j estimated from our S&P 500 data in the sample period 2009 to 2021. The error term $e_{j,t}$ is sampled from $\mathcal{N}(0, \sigma_j^2)$ where σ_j^2 is set to be 2% of the variance of α_j estimated from real data.

Next, we have the Nonlinear experiment, in which we incorporate the nonlinearity of temporal dependence of IVS by using a nonlinear model to simulate the dynamics of each α_j

$$\alpha_{j,t} = 2\sin(\alpha_{j,t-1}) + 4\cos(\alpha_{j,t-1}) + u_{j,t}.$$

The simulated values of α_j are rescaled to be in the range of α_j estimated from real data, and the error term $u_{j,t}$ is sampled from $\mathcal{N}(0, \sigma_j^2)$ where σ_j^2 is set to be 2% of the variance of α_j estimated from real data.



(a) Lead and lag IV **(b)** Lead IV and lag α_0 **(c)** Lead IV and lag α_1

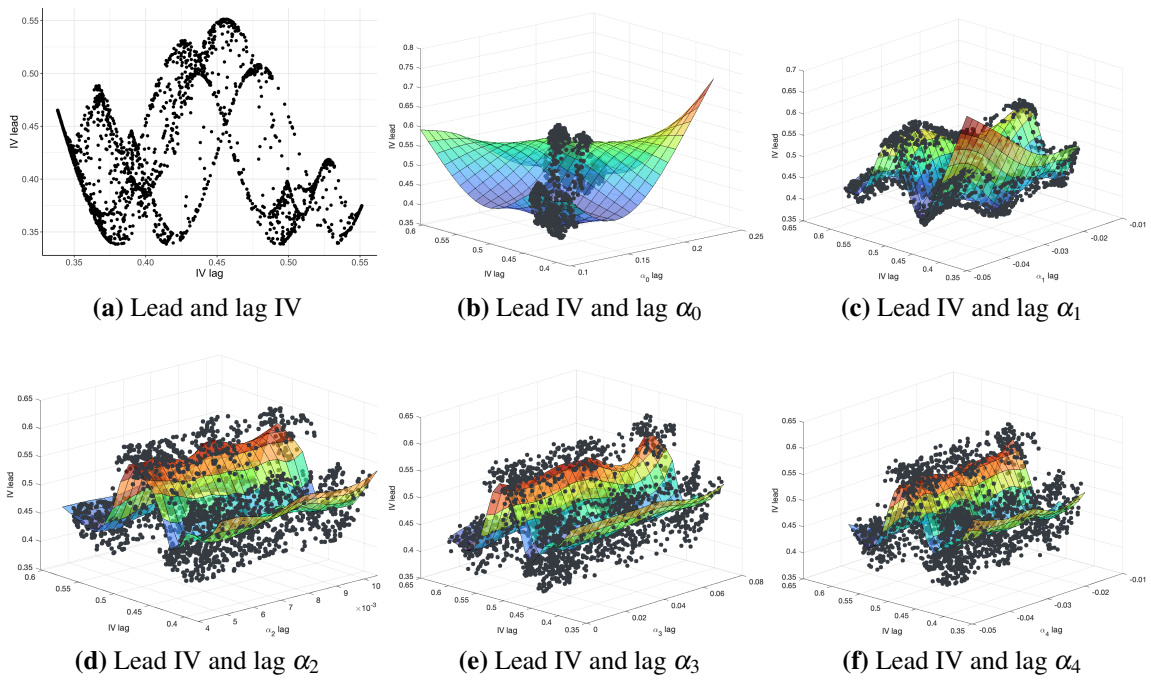
(d) Lead IV and lag α_2 **(e)** Lead IV and lag α_3 **(f)** Lead IV and lag α_4

Figure A.2: Lead values of $IV(0,0.5)$ on day $t+1$ versus lag values of $IV(0,0.5)$ in (a) and each $\alpha_i, i = 0,...,4$ on day t in (b) to (f), in one of the simulations for the Linear experiment.

Figure A.2 shows the linear relationships between lead and lag of $IV(0,0.5)$, i.e., IV at moneyness $m = 0$ and time-to-maturity of half a year, time series, as well as the 3D plots when we incorporate in the lag of each parameter $\alpha_i, i = 0,...,4$. In Figure A.3, we illustrate IV of day $t+1$ as a nonlinear function of the parameters $\alpha_i, i = 0,...,4$. The sine-cosine dynamics $\alpha_{j,t} = a_j \alpha_{j,t-1} + e_{j,t}$ is well reflected in the lead-lag IV relationship, and also in the 3D plots with the lag of each parameter α_i.

	Linear			Nonlinear		
	RMSE	MAPE	OoR^2	RMSE	MAPE	OoR^2
fRW	0.83 (0.12)	**1.37 (0.20)**	94.47 (2.97)	9.16 (0.28)	18.60 (0.69)	-107.88 (4.22)
fLinK	**0.83 (0.11)**	**1.37 (0.20)**	**94.49 (2.96)**	6.38 (0.18)	12.91 (0.42)	-0.87 (1.40)
fGauK	0.85 (0.13)	1.40 (0.26)	94.30 (3.01)	1.52 (0.06)	2.98 (0.13)	94.27 (0.59)
fLapK	0.96 (0.39)	1.58 (0.76)	92.40 (7.79)	1.37 (0.13)	2.52 (0.24)	95.27 (0.90)
fNTK1	**0.83 (0.11)**	1.38 (0.20)	94.45 (2.97)	1.57 (0.07)	3.04 (0.13)	93.82 (0.57)
fNTK3	0.84 (0.11)	1.39 (0.20)	94.39 (2.98)	**0.97 (0.15)**	**1.93 (0.35)**	**97.58 (0.85)**
fNTK5	0.85 (0.12)	1.40 (0.21)	94.28 (3.04)	1.08 (0.38)	2.19 (0.82)	96.69 (2.89)

Table A.2: Mean and standard deviation (in brackets) of prediction accuracy in terms of RMSE, MAPE and OoR^2 of the functional models over the test set T_3, under Linear and Nonlinear experiments.



(a) Lead and lag IV **(b)** Lead IV and lag α_0 **(c)** Lead IV and lag α_1

(d) Lead IV and lag α_2 **(e)** Lead IV and lag α_3 **(f)** Lead IV and lag α_4

Figure A.3: Lead values of $IV(0,0.5)$ on day $t+1$ versus lag values of $IV(0,0.5)$ in (a) and each $\alpha_i, i = 0,...,4$ on day t in (b) to (f), in one of the simulations for the Nonlinear experiment.

Table A.2 shows the mean and standard deviation of prediction performances in terms of RMSE, MAPE, and OoR^2 for the test set T_3 in the Linear and Nonlinear experiments. As expected, we observe that for the Linear experiment, the linear model performs the best, with 0.83% RMSE and 94.49% OoR^2, followed closely by fNTK1 with 0.83% RMSE and 94.45% OoR^2. The random walk model, fRW also performs very well. After introducing nonlinearity, the nonlinear models start to outperform fLinK and fRW significantly. In the Nonlinear experiment, fNTK3 achieves the best performance of 0.97% RMSE, compared to 6.38% for fLinK and 9.16% for fRW. The outperformance is even more profound when we look at OoR^2 and MAPE, with fNTK achieving the best OoR^2 of 97.58%, while fRW and fLinK both have negative OoR^2. The other nonlinear models also perform well under the nonlinear setup, for example, RMSE is 1.37% for fLapK and 1.52% for fGauK.

A.6 Estimation with parametric kernels

Each task j can be reformulated as a regression that can be solved by estimating a kernel-ridge model. Suppose the reproducing kernel k has an eigen-expansion

$$k(\hat{\boldsymbol{x}}, \hat{\boldsymbol{x}}^{'}) = \sum_{i=1}^{\infty} \nu_i \phi_i(\hat{\boldsymbol{x}}) \phi_i(\hat{\boldsymbol{x}}^{'}) \tag{20}$$

with $\nu_i \geq 0, \sum_{i=1}^{\infty} \nu_i^2 < \infty$. We can express each function $f_j \in \mathfrak{M}_{\boldsymbol{x}}$ in terms of these eigenfunctions

$$f_j(\hat{\boldsymbol{x}}) = \sum_{\ell=1}^{\infty} c_{j\ell} \phi_\ell(\hat{\boldsymbol{x}}) \tag{21}$$

with a generalized ridge penalty

$$J(f_j) = \|f_j\|_{\mathfrak{M}_{\boldsymbol{x}}}^2 \stackrel{\text{def}}{=} \sum_{\ell=1}^{\infty} c_{j\ell}^2 / \nu_\ell < \infty$$

where $\|f_j\|_{\mathfrak{M}_{\boldsymbol{x}}}^2$ is the norm induced by k. The function $f_j \in \mathfrak{M}_{\boldsymbol{x}}$ can be found by the minimization problem

$$\min_{f_j \in \mathfrak{M}_{\boldsymbol{x}}} \left[\sum_{i=1}^{n} \left(\hat{y}_{ij} - f_j(\hat{\boldsymbol{x}}_i) \right)^2 + \lambda \|f_j\|_{\mathfrak{M}_{\boldsymbol{x}}}^2 \right] \tag{22}$$

By the representer theorem from Schölkopf, Herbrich, and Smola (2001), the solution of (22) is finite-dimensional and has the form

$$f_j(\hat{\boldsymbol{x}}) = \sum_{i=1}^{n} \alpha_i k(\hat{\boldsymbol{x}}, \hat{\boldsymbol{x}}_i) \tag{23}$$

Using the reproducing properties of kernel k, we have $\langle k(., \hat{\boldsymbol{x}}_i), h \rangle_{\mathfrak{M}_{\boldsymbol{x}}} = h(\hat{\boldsymbol{x}}_i)$ and $\langle k(., \hat{\boldsymbol{x}}_i), k(., \hat{\boldsymbol{x}}_j) \rangle_{\mathfrak{M}_{\boldsymbol{x}}} = k(\hat{\boldsymbol{x}}_i, \hat{\boldsymbol{x}}_j)$, and the penalty term $J(f_j)$ can be further expressed as

$$J(f_j) = \sum_{i=1}^{n} \sum_{\ell=1}^{n} k(\hat{\boldsymbol{x}}_i, \hat{\boldsymbol{x}}_\ell) \alpha_i \alpha_\ell \tag{24}$$

Let $\hat{\boldsymbol{y}}_j = (\hat{y}_{1j}, \ldots, \hat{y}_{nj})^T$, $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_n)^T$, and the gram matrix G where $G_{ij} = k(\hat{\boldsymbol{x}}_i, \hat{\boldsymbol{x}}_j)$. From (22), (23) and (24), we can rewrite the minimization problem in terms of $\boldsymbol{\alpha}$ and G

$$\min_{\boldsymbol{\alpha}} \quad (\hat{\boldsymbol{y}}_j - G\boldsymbol{\alpha})^T(\hat{\boldsymbol{y}}_j - G\boldsymbol{\alpha}) + \lambda \boldsymbol{\alpha}^T G\boldsymbol{\alpha}. \tag{25}$$

Thus, the solution for $\boldsymbol{\alpha}$ is simply $\widehat{\boldsymbol{\alpha}} = (G + \lambda \boldsymbol{I})^{-1}\hat{\boldsymbol{y}}_j$. The solution for the kernel ridge regression is therefore $\hat{h}_j(\hat{\boldsymbol{x}}) = \sum_{i=1}^n \widehat{\alpha}_i k(\hat{\boldsymbol{x}}_i, \hat{\boldsymbol{x}})$. The minimization problem in (25) is solved in one step for known kernels.

A.7 Hyperparameters

There are several hyperparameters involved in the estimation process. The sieve hyperparameter K_n dictates the dimensions of the reduced parameter space. Tuning hyperparameters γ and λ control the bandwidth and sparsity of curve predictors in regressions with parametric kernels. Additionally, for the NTK, unique hyperparameters pertain to the NN architecture.

Sieve approximation hyperparameters. The number of sieves for predictor and response curves is determined using the explained variance principle. We truncate the number of eigenfunctions to account for 99.99% of the variance explained. Consequently, the final number of eigenvectors retained is 15 for Y and 46 for X.

Ridge regularization and bandwidth hyperparameters. Kernel ridge regression is implemented with a ridge regularization strength $\lambda > 0$ for each kernel. The hyperparameter λ modulates the balance between fitting the training data and preventing overfitting, dictating the extent to which the regularization term influences the final solution. The hyperparameters γ and λ are fine-tuned across a value grid, ranging from 0.005 to 0.025 for γ and from 10^{-5} to 10^{-1} for λ, at the start of every six months using cross-validation (see, e.g., Yao, Müller, and Wang (2005)).

NN hyperparameters. Implementation of an NN model requires the specification of several hyperparameters, including the number of hidden layers, choice of activation functions, and regularization strengths. However, an exhaustive search for the optimal architecture by evaluating

infinite hyperparameter combinations is typically impractical. In this study, we employ an NN architecture featuring large widths of 500 neurons in each hidden layer. This design ensures that the empirical NTK of the NN closely approximates its limiting NTK. We opt for three hidden layers and deploy the ReLU activation function across all of them, primarily since most theoretical results for the NTK align with ReLU. We also conduct a robustness test using NNs with one and five hidden layers, with findings presented in Appendix A.9.

The NN parameters are refined and learned by minimizing $\mathscr{L}(\Omega_{MLP^{(L)}})$ through the gradient descent optimizer with a learning rate of 0.05. Every six months, we adjust the weight decay rate over a set of five possible values, spanning from 10^{-5} to 10^{-1}. Weight decay, also recognized as L2 regularization, offers multiple benefits: it counteracts overfitting, enhances generalization to unseen datasets, regulates model complexity, lowers sensitivity to noise, and stabilizes the training process. Through penalization of large weight values, weight decay promotes the development of simpler models, reducing susceptibility to overfitting and ensuring better generalization to novel data.

A.8 Alternative models

A.8.1 Car and Wu model

Carr and Wu (2016) proposes an option pricing framework that models the near-term dynamics of the implied volatility across different strikes and expiries. For an option with strike K and time to maturity τ, the dynamics of the underlying spot price S_t and the option implied volatility $IV_t(K,\tau)$ under the risk-neutral measures are captured as

$$\begin{aligned} dS_t/S_t &= \sqrt{v_t}dW_t, \\ dIV_t(K,\tau)/IV_t(K,\tau) &= e^{-\eta_t\tau}(a_t dt + w_t dZ_t), \end{aligned} \tag{26}$$

where v_t denotes the time-t instantaneous variance rate of the underlying asset log-returns, a_t is the average drift of the implied volatility, and the exponential dampening parameter $e^{-\eta_t\tau}$ accom-

modates the empirical observation that implied volatilities of long-dated options tend to move less. W_t and Z_t are the Wiener processes with correlation process $\rho_t \in [-1,1]$. Additionally, a_t, w_t and η_t are stochastic processes independent of K, τ and $IV_t(K,\tau)$.

Denote $k = log(KS_t)$ the relative strike. It is shown by Carr and Wu (2016) that the square implied volatility $IV_t^2(k,\tau)$ satisfies the quadratic equation

$$\frac{1}{4}e^{-2\eta_t\tau}w_t^2\tau^2 IV_t^4 + (1-2e^{-\eta_t\tau}a_t\tau - e^{-\eta_t\tau}w_t\rho_t\sqrt{v_t}\tau)IV_t^2 - (v_t + 2e^{-\eta_t\tau}w_t\rho_t\sqrt{v_t}k + e^{-2\eta_t\tau}w_t^2k^2) = 0. \tag{27}$$

The implied volatility surface on day t is fitted with the values of parameters $\boldsymbol{\theta}_t = (v_t, a_t, w_t, \eta_t, \rho_t)$ at t. Given the set of options on day t with implied volatility $IV_{i,t}$, relative strike $k_{i,t}$ and maturity $\tau_{i,t}$, the parameters $\boldsymbol{\theta}_t$ are estimated by minimizing the nonlinear least squares

$$\begin{aligned}\hat{\boldsymbol{\theta}}_t = \arg\min_{\boldsymbol{\theta}_t} \sum_{i=1}^{N_t} \Big[&\frac{1}{4}e^{-2\eta_t\tau_{i,t}}w_t^2\tau_{i,t}^2 IV_{i,t}^4 + (1-2e^{-\eta_t\tau_{i,t}}a_t\tau_{i,t} - e^{-\eta_t\tau_{i,t}}w_t\rho_t\sqrt{v_t}\tau_{i,t})IV_{i,t}^2 \\ &- (v_t + 2e^{-\eta_t\tau_{i,t}}w_t\rho_t\sqrt{v_t}k_{i,t} + e^{-2\eta_t\tau_{i,t}}w_t^2k_{i,t}^2)\Big].\end{aligned} \tag{28}$$

With the estimated parameters $\hat{\boldsymbol{\theta}}_t$, the parameters on day $t+h$ is predicted to be $\hat{\boldsymbol{\theta}}_{t+h} = \hat{\boldsymbol{\theta}}_t$ and hence the implied volatility $IV_{i,t+h}$ of day $t+h$ predicted by the Carr and Wu (CW) model is obtained by solving equation (27) using $\hat{\boldsymbol{\theta}}_{t+h}$ as inputs as well as the option relative strike and time to maturity.[3]

A.8.2 Ad-hoc Black–Scholes model

Goncalves and Guidolin (2006) propose a vector autoregressive approach to model the dynamics of the parameters of the Ad-hoc Black–Scholes (AHBS) model of Dumas, Fleming, and Whaley (1998) and predict their values in the future. For each day t, the AHBS model is estimated with a

[3] We are grateful to Gustavo Freire for sharing his codes implementing the Carr and Wu model.

cross-section of $i = 1, ..., N_t$ options using the following regression

$$IV_{i,t} = \alpha_{0,t} + \alpha_{1,t} m_{i,t} + \alpha_{2,t} m_{i,t}^2 + \alpha_{3,t} \tau_{i,t} + \alpha_{4,t} m_{i,t} \tau_{i,t} + \varepsilon_{i,t}, \tag{29}$$

where $IV_{i,t}$, $m_{i,t}$ and $\tau_{i,t}$ are the observed implied volatilities, moneyness, and time to maturity (in years) of the option i on day t, respectively; $\varepsilon_{i,t}$ is the random error term, and N_t is the number of options available for day t. A VAR(p) model is fitted to capture the dynamics of $\boldsymbol{\alpha}_t = (\alpha_{0,t}, \alpha_{1,t}, \alpha_{2,t}, \alpha_{3,t}, \alpha_{4,t})$:

$$\boldsymbol{\alpha}_{t+h} = \boldsymbol{\mu} + \boldsymbol{\Phi}_1 \boldsymbol{\alpha}_t + \frac{1}{5} \boldsymbol{\Phi}_5 \sum_{j=t}^{t-4} \boldsymbol{\alpha}_j + \frac{1}{22} \boldsymbol{\Phi}_{22} \sum_{j=t}^{t-21} \boldsymbol{\alpha}_j + \boldsymbol{\varepsilon}_{t+1}, \tag{30}$$

where $\boldsymbol{\varepsilon}_{t+1} \overset{\text{i.i.d}}{\sim} N(0, \Xi)$. After estimating $\hat{\boldsymbol{\alpha}}_{t+h}$ with equation (30), the implied volatility predicted by the AHBS model is attained by using $\hat{\boldsymbol{\alpha}}_{t+h}$ and equation (29). Comparing the NFAR models with the CW and AHBS models allows us to see whether using a nonparametric nonlinear model can help improve prediction accuracy, as opposed to using a traditional parametric model for predicting IV.

A.9 Additional results on statistical performance

On top of RMSE and OoR^2 reported in the main text, we assess prediction accuracy from two more angles, error magnitude, and directional changes, both derived from observed test data. Mean absolute percentage error (MAPE) captures prediction accuracy and error magnitude, while mean correct prediction of direction of change (MCPDC) evaluates the ability of the models to anticipate the direction of price movements. These two metrics contribute another unique insight into the

performances of the models.

$$\text{MAPE}_h = \frac{1}{\sum_{t=t_0}^{T-h} N_t} \sum_{t=t_0}^{T-h} \sum_{j=1}^{N_t} \left| \frac{Y_{t+h}(u_j) - \hat{Y}_{t+h}(u_j)}{Y_{t+h}(u_j)} \right|,$$

$$\text{MCPDC}_h = \frac{1}{\sum_{t=t_0}^{T-h} N_t} \sum_{t=t_0}^{T-h} \sum_{j=1}^{N_t} \times \mathbb{1}_{(Y_{t+h}(u_j) - Y_t(u_j))(\hat{Y}_{t+h}(u_j) - Y_t(u_j)) > 0}.$$

Here, $t_0 = 2523$ and $T = 3273$ mark the start and end of the testing period. The MCPDC_h is computed solely for options traded on both day t and day $t+h$.

A.9.1 Robustness of statistical performance

The outcomes documented in Table A.3 underscore an intriguing observation: while utilizing a Neural Tangent Kernel (NTK) framework with three hidden layers contributes to improved prediction accuracy, the benefits of incorporating a higher number of hidden layers are not uniformly positive. For example, when focusing on the forecasting horizon of $h = 20$, the Mean Absolute Percentage Error (MAPE) diminishes from 14.85% to 10.83% with the adoption of fNTK with three hidden layers instead of one, whereas employing five layers slightly exacerbates MAPE to 10.94%.

Furthermore, we categorize options into distinct groups based on their moneyness (m) and time-to-maturity (τ), subsequently examining the statistical performance of the models within each group. With respect to moneyness, the options are stratified into four groups: $[-2,-0.5]$, $(-0.5,0]$, $(0,0.5]$, and $(0.5,2]$, while maturity is divided into four intervals based on days to maturity: $[5,60]$, $(60,120]$, $(120,180]$, and $(180,252]$ days. The performance of the models across moneyness and time-to-maturity groups, as measured by the Root Mean Square Error (RMSE), is illustrated in Figure A.4.[4]

The analysis reveals that, on the whole, prediction errors, as quantified by RMSE, tend to be higher for options with larger moneyness (m) and shorter maturity (τ). Remarkably, the consistent

[4]We tried the put and call options separately and the same conclusions hold. Hence, it does not seem to be a matter of liquidity. Additional results for put and calls separately aare available upon request.

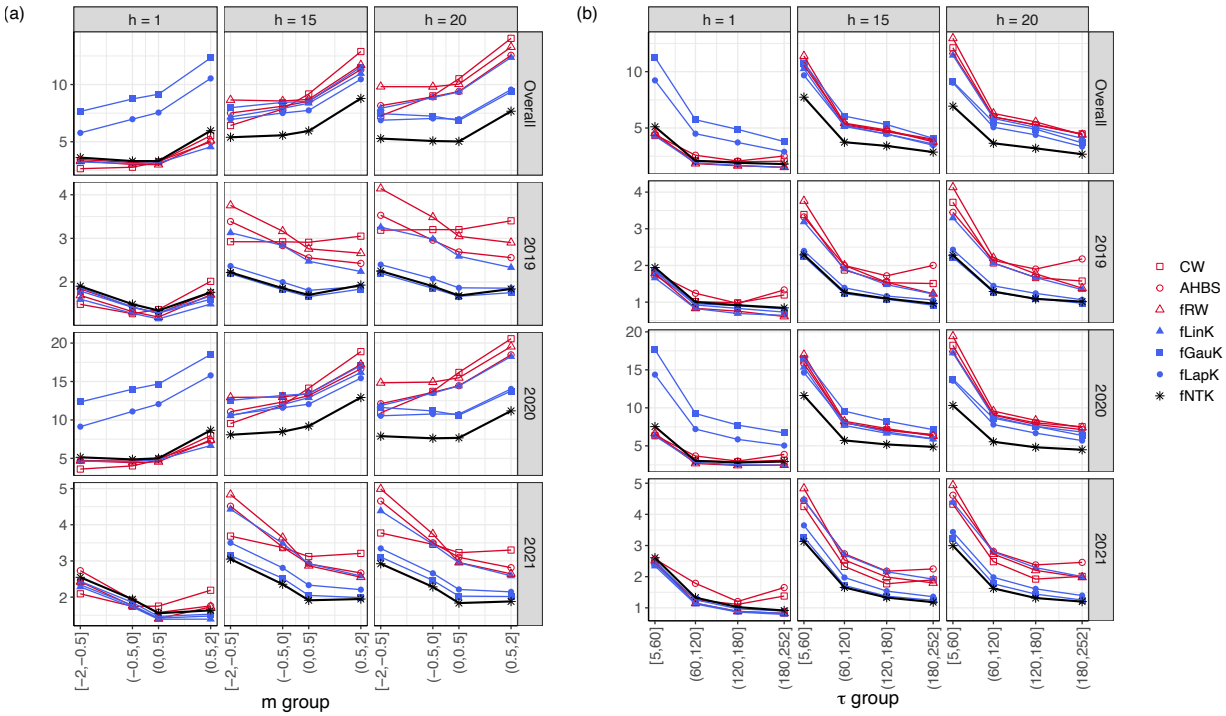


Figure A.4: Prediction accuracy in terms of RMSE of all the models across forecasting horizons $h = 1, 15$ and 20, in (a) four different moneyness m values: $[-2,-0.5], (-0.5,0], (0,0.5]$, and $(0.5,2]$ and (b) time-to-maturity τ values: $[5,60], (60,120], (120,180]$, and $(180,252]$ days. The performance period is split into overall (from Jan 09, 2019, to Dec 31, 2021), year 2019, year 2020, and year 2021.

pattern across all moneyness and time-to-maturity groups, as well as prediction periods, is the superior performance of nonlinear functional models, notably exemplified by fNTK, particularly in longer forecasting horizons. It is noteworthy that across the entirety of the analysis, all models exhibit diminished forecasting accuracy in the year 2020 when compared to both 2019 and 2021.

These findings collectively underscore the robustness of nonlinear models across various market conditions and reveal nuanced insights into the interplay between model complexity and performance across distinct prediction horizons.

	RMSE					MAPE					OoR^2					MCPDC				
	$h=1$	$h=5$	$h=10$	$h=15$	$h=20$	$h=1$	$h=5$	$h=10$	$h=15$	$h=20$	$h=1$	$h=5$	$h=10$	$h=15$	$h=20$	$h=1$	$h=5$	$h=10$	$h=15$	$h=20$
Overall (from Jan 9, 2019 to Dec 31, 2021)																				
CW	3.57	5.45	7.23	8.87	9.91	8.46	12.49	16.05	18.36	20.23	88.74	74.18	54.50	32.01	15.21	41.80	43.40	45.35	49.11	51.85
AHBS	3.72	5.53	7.37	8.69	9.55	8.92	12.78	15.73	17.72	19.02	87.79	73.35	52.87	34.83	21.33	45.05	44.61	48.28	51.70	54.86
fRW	3.65	5.72	7.66	9.27	10.57	7.21	11.89	15.80	18.36	20.45	88.28	71.59	48.97	25.69	3.49	39.92	44.64	46.23	49.61	51.41
fLinK	**3.48**	5.31	7.29	8.41	9.40	**6.95**	11.49	14.61	16.62	17.88	**89.19**	75.46	53.78	38.76	23.59	43.90	46.68	51.40	55.02	57.29
fGauK	9.23	9.38	9.38	8.95	7.67	11.47	13.00	13.57	14.01	13.97	25.54	23.32	23.67	30.88	49.32	47.99	58.50	66.57	69.83	72.55
fLapK	7.51	8.32	8.05	7.99	7.49	9.21	11.85	12.72	13.53	13.65	50.70	39.71	43.74	44.90	51.63	45.92	55.78	63.63	66.69	69.50
fNTK1	4.44	5.36	6.95	7.28	7.35	8.34	11.31	13.03	14.01	14.85	82.81	74.42	56.93	54.15	53.19	45.48	53.71	60.13	65.09	68.76
fNTK3	4.11	**5.07**	**6.03**	**6.32**	**5.73**	8.10	**9.80**	10.85	10.85	**10.83**	85.24	**77.24**	**68.17**	**65.38**	**71.67**	46.40	58.97	65.83	71.55	74.59
fNTK5	4.08	5.08	6.49	6.46	6.06	8.44	9.95	**10.48**	**10.51**	10.94	85.49	77.23	63.18	63.77	68.27	**48.08**	**59.27**	**67.21**	**72.18**	**74.66**
From Jan 9, 2019 to Dec 31, 2019																				
CW	1.53	2.23	2.65	2.91	3.19	7.57	11.24	13.90	15.34	17.44	86.13	70.52	58.33	50.05	40.40	40.90	40.98	44.85	50.52	53.52
AHBS	1.62	2.28	2.66	2.91	3.04	7.87	11.30	13.70	15.46	16.52	84.54	69.35	58.10	49.89	45.86	44.94	45.84	49.23	53.31	56.45
fRW	1.51	2.39	2.89	3.19	3.51	6.38	10.73	13.83	15.38	17.67	86.62	66.28	50.48	39.74	27.68	39.20	41.97	45.52	51.87	53.55
fLinK	**1.41**	2.15	2.48	2.75	2.86	**6.15**	9.71	11.42	12.39	13.19	**88.28**	72.65	63.63	55.34	51.97	**43.86**	47.15	52.27	55.00	54.82
fGauK	1.56	**1.88**	1.82	**1.92**	**1.91**	7.03	**8.75**	8.69	9.44	**9.48**	85.67	**79.15**	80.43	**78.16**	78.48	43.83	57.43	66.50	69.21	73.42
fLapK	1.61	1.92	2.01	2.07	2.10	7.42	9.16	9.75	10.36	10.25	84.65	78.26	75.97	74.70	74.01	41.95	54.35	62.85	65.27	69.68
fNTK1	1.71	2.24	2.32	2.49	2.37	7.82	10.95	11.45	12.25	11.96	82.65	70.25	68.23	63.21	66.93	44.15	53.35	61.33	64.54	71.40
fNTK3	1.66	1.95	1.97	1.96	1.96	7.49	8.82	9.22	**9.17**	9.53	83.73	77.43	76.99	77.15	**77.43**	43.72	**59.57**	65.25	**70.60**	**74.58**
fNTK5	1.74	1.92	**1.81**	2.03	1.99	8.08	8.78	**8.41**	9.48	9.68	82.13	78.20	**80.65**	75.38	76.72	44.69	58.42	**67.08**	67.96	74.04
From Jan 1, 2020 to Dec 31, 2020																				
CW	5.29	8.11	10.95	13.57	15.19	9.35	14.52	20.28	24.00	26.71	82.92	59.91	24.89	-16.35	-48.49	42.92	45.15	44.40	46.96	47.14
AHBS	5.46	8.17	11.11	13.20	14.54	9.88	14.63	18.67	21.02	22.37	81.81	59.28	23.18	-10.04	-35.87	46.49	44.59	48.53	51.22	52.79
fRW	5.43	8.43	11.53	14.10	16.15	8.38	14.00	20.04	24.14	27.26	82.00	56.78	17.11	-25.61	-67.87	41.99	46.81	44.82	46.59	45.60
fLinK	**5.21**	7.86	11.02	12.78	14.38	**8.11**	13.42	17.76	20.66	22.73	**83.14**	62.38	23.93	-3.81	-33.68	**45.33**	47.05	52.16	54.77	55.42
fGauK	14.74	14.87	14.80	13.98	11.85	19.03	20.28	21.47	21.12	20.28	-31.32	-35.54	-36.20	-23.39	9.91	51.18	58.17	65.98	69.42	70.48
fLapK	11.92	13.11	12.55	12.33	11.50	13.00	16.81	17.61	17.76	17.98	14.09	-5.34	1.89	3.99	15.23	49.68	56.74	65.76	70.20	70.39
fNTK1	6.71	8.06	10.65	11.04	11.20	9.60	13.25	16.15	16.88	19.13	72.88	58.75	26.30	22.69	18.47	47.96	55.04	62.53	69.46	67.95
fNTK3	6.16	**7.65**	**9.23**	**9.68**	**8.68**	9.39	**11.41**	**13.52**	13.63	**13.34**	77.18	63.24	**46.11**	**40.48**	**51.41**	46.74	59.95	**68.29**	**73.99**	**75.31**
fNTK5	6.07	7.66	10.04	9.96	9.27	9.51	11.51	13.57	**13.51**	13.91	77.74	**63.40**	36.38	36.85	44.68	48.56	**61.10**	67.81	73.70	74.82
From Jan 1, 2021 to Dec 31, 2021																				
CW	1.97	2.81	3.22	3.40	3.49	8.12	11.21	13.03	14.35	15.05	89.69	78.95	72.39	69.14	66.14	41.22	43.15	46.61	50.49	55.60
AHBS	2.18	3.06	3.47	3.66	3.80	8.61	11.82	13.96	15.67	17.00	87.42	75.14	67.94	64.42	59.98	43.69	43.85	47.44	51.24	56.02
fRW	1.96	3.14	3.63	3.83	3.95	6.54	10.46	12.65	14.16	14.90	89.79	73.92	65.01	61.03	56.83	38.30	44.15	48.07	51.30	55.96
fLinK	**1.83**	2.91	3.44	3.63	3.61	**6.29**	10.63	13.37	15.08	15.75	**91.08**	77.54	68.52	65.00	63.98	42.51	46.04	50.09	55.21	60.55
fGauK	1.96	**2.29**	**2.35**	2.59	2.56	6.65	**8.27**	**8.46**	9.44	10.08	89.79	**86.17**	**85.28**	82.16	81.80	47.35	**59.47**	**67.23**	70.65	74.10
fLapK	1.89	2.37	2.63	2.90	2.76	6.54	8.50	9.55	11.10	11.20	90.54	85.16	81.59	77.71	78.86	44.60	55.75	62.06	64.11	68.52
fNTK1	2.08	2.59	2.85	3.14	2.92	7.40	9.59	10.83	12.10	12.12	88.48	82.21	78.37	73.81	76.41	43.84	52.60	57.09	61.06	68.01
fNTK3	2.06	2.47	2.50	2.48	2.39	7.20	8.77	9.12	8.99	8.97	88.72	83.81	83.38	83.61	84.21	47.67	57.62	63.82	69.75	73.87
fNTK5	2.10	2.52	2.42	**2.27**	**2.25**	7.59	9.10	8.59	**8.01**	**8.56**	88.27	83.19	84.45	**86.33**	**86.01**	**49.64**	57.97	66.75	**73.19**	**74.89**

Table A.3: Prediction accuracy RMSE, MAPE, OoR^2 and MCPDC of all models over the whole test period (Jan 09, 2019 to Dec 31, 2021) and in each of three years 2019 (before Covid), 2020 (Covid year) and 2021 (recovery year). In this table, fNTKℓ refers to the fNTK model with ℓ hidden layers, and fNTK3 is the fNTK in the main text. Bold numbers indicate the best-performing model (or models) in a given column.

Panel A: Training size of 1000 days

	RMSE					MAPE				
	$h=1$	$h=5$	$h=10$	$h=15$	$h=20$	$h=1$	$h=5$	$h=10$	$h=15$	$h=20$
CW	**3.57**	**5.45**	7.23	8.87	9.91	8.46	12.49	16.05	18.36	20.23
AHBS	3.82	5.57	7.60	8.85	9.68	9.12	13.17	16.47	18.51	19.86
fRW	3.65	5.72	7.66	9.27	10.57	**7.21**	11.89	15.80	18.36	20.45
fLinK	3.94	5.74	7.56	8.24	9.01	7.37	12.32	16.47	18.09	19.05
fGauK	10.13	10.31	10.32	9.87	8.45	13.16	14.30	14.64	14.72	14.54
fLapK	8.64	8.88	9.30	9.08	8.21	10.49	12.36	13.84	14.68	14.95
fNTK	5.67	5.88	**6.73**	**7.19**	**6.14**	9.34	**10.57**	**11.13**	**11.51**	**11.07**
	OoR^2					MCPDC				
	$h=1$	$h=5$	$h=10$	$h=15$	$h=20$	$h=1$	$h=5$	$h=10$	$h=15$	$h=20$
CW	**88.74**	**74.18**	54.50	32.01	15.21	41.80	43.40	45.35	49.11	51.85
AHBS	87.11	72.99	49.89	32.23	19.17	45.37	44.50	47.65	49.95	53.54
fRW	88.28	71.59	48.97	25.69	3.49	39.92	44.64	46.23	49.61	51.41
fLinK	86.28	71.25	49.92	41.40	30.08	44.60	48.09	52.09	56.45	59.93
fGauK	10.33	7.48	7.62	15.75	38.35	**48.48**	58.84	65.65	69.58	72.11
fLapK	34.81	30.99	24.87	28.69	41.91	46.57	55.59	63.00	66.12	68.61
fNTK	71.91	69.93	**60.70**	**55.14**	**67.48**	45.62	**59.11**	**65.87**	**71.62**	**75.33**

Panel B: Training size of 2000 days

	RMSE					MAPE				
	$h=1$	$h=5$	$h=10$	$h=15$	$h=20$	$h=1$	$h=5$	$h=10$	$h=15$	$h=20$
CW	**3.57**	5.45	7.23	8.87	9.91	8.46	12.49	16.05	18.36	20.23
AHBS	3.76	5.53	7.47	8.75	9.59	8.97	12.77	15.70	17.55	18.72
fRW	3.65	5.72	7.66	9.27	10.57	7.21	11.89	15.80	18.36	20.45
fLinK	3.70	5.38	7.38	8.30	9.06	**7.03**	11.49	14.72	16.50	17.74
fGauK	9.82	9.90	10.05	9.61	8.21	12.51	13.64	14.26	14.59	14.32
fLapK	8.40	8.38	8.91	8.73	8.01	10.02	11.77	13.38	14.09	14.07
fNTK	4.39	**4.96**	**6.50**	**6.64**	**6.17**	8.39	**9.57**	**10.89**	**10.85**	**11.12**
	OoR^2					MCPDC				
	$h=1$	$h=5$	$h=10$	$h=15$	$h=20$	$h=1$	$h=5$	$h=10$	$h=15$	$h=20$
CW	**88.74**	74.18	54.50	32.01	15.21	41.80	43.40	45.35	49.11	51.85
AHBS	87.52	73.35	51.52	33.75	20.73	45.07	44.44	47.77	50.86	54.38
fRW	88.28	71.59	48.97	25.69	3.49	39.92	44.64	46.23	49.61	51.41
fLinK	87.87	74.81	52.68	40.44	29.14	44.10	48.07	52.68	56.69	59.37
fGauK	15.75	14.71	12.44	20.11	41.84	**48.25**	58.57	66.43	69.42	72.47
fLapK	38.40	38.75	31.14	34.17	44.70	47.52	56.60	63.78	66.38	70.04
fNTK	83.18	**78.48**	**62.34**	**61.87**	**66.95**	46.74	**59.86**	**67.04**	**72.02**	**73.58**

Table A.4: Prediction accuracy of all models in terms of RMSE, MAPE, OoR^2, and MCPDC. The models are trained with two different training sizes: 1000 days in Panel A and 2000 days in Panel B. The prediction period is from Jan 09, 2019 to Dec 31, 2021. Bold numbers indicate the best-performing model (or models) in a given column.

A.9.2 Diebold-Mariano tests

We use the Diebold-Mariano (DM) test to determine whether the two prediction performances are significantly different, see Diebold and Mariano (2002). Let $e_{h,1t}$ and $e_{h,2t}$ denote the $h-$day forecasting error of model 1 and model 2 at time $t \in [t_0, T-h]$. Since we have multiple observations and forecasted values on each test day t, we define $e_{h,t} = \frac{1}{N_t}\sum_{j=1}^{N_t}\left|\frac{Y_{t+h}(u_j)-\hat{Y}_{t+h}(u_j)}{Y_{t+h}(u_j)}\right|$ for MAPE, and $e_{h,t} = \sqrt{\frac{1}{N_t}\sum_{j=1}^{N_t}(Y_{t+h}(u_j)-\hat{Y}_{t+h}(u_j))^2}$ for RMSE. The differential loss is computed by $d_{h,t} = e_{h,1t} - e_{h,2t}$, and the DM test statistic is computed as follows

$$\mathrm{DM}_h = \frac{\frac{1}{T-h-t_0+1}\sum_{t=t_0}^{T-h} d_{h,t}}{\sqrt{(\hat{\sigma}_0 + 2\sum_{k=1}^{h}\hat{\sigma}_k)/(T-h-t_0-22+1)}},$$

where $\hat{\sigma}_0$ is the sample standard deviation, and $\hat{\sigma}_k$ is the autocovariance at lag $k \geq 1$ of the series d_t^h. The null hypothesis $H_0 : \mathrm{DM}_h = 0$ is that there is no significant difference in the accuracy of the two models.

	CW	AHBS	fRW	fLinK	fGauK	fLapK	fNTK
$\underline{h=1}$							
CW	–	0.000	1.000	1.000	0.000	0.013	0.766
AHBS	1.000	–	1.000	1.000	0.000	0.295	1.000
fRW	0.000	0.000	–	1.000	0.000	0.000	0.000
fLinK	0.000	0.000	0.053	–	0.000	0.000	0.000
fGauK	1.000	1.000	1.000	1.000	–	1.000	1.000
fLapK	0.998	0.881	1.000	1.000	0.000	–	0.996
fNTK	0.955	0.000	1.000	1.000	0.000	0.006	–
$\underline{h=5}$							
CW	–	0.002	0.732	0.993	0.248	0.517	1.000
AHBS	1.000	–	0.987	1.000	0.398	0.701	1.000
fRW	0.999	0.039	–	0.994	0.228	0.517	1.000
fLinK	0.172	0.000	0.033	–	0.142	0.313	1.000
fGauK	0.872	0.678	0.799	0.890	–	0.995	0.983
fLapK	0.657	0.349	0.517	0.707	0.016	–	0.956
fNTK	0.000	0.000	0.000	0.002	0.025	0.069	–
$\underline{h=10}$							
CW	–	0.592	0.063	0.993	0.898	0.998	1.000
AHBS	0.971	–	0.273	1.000	0.836	0.984	1.000
fRW	1.000	0.760	–	0.992	0.937	1.000	1.000
fLinK	0.222	0.006	0.032	–	0.649	0.891	1.000
fGauK	0.453	0.308	0.198	0.553	–	0.915	0.927
fLapK	0.095	0.039	0.003	0.256	0.110	–	0.920
fNTK	0.000	0.000	0.000	0.000	0.074	0.105	–
$\underline{h=15}$							
CW	–	0.748	0.007	0.940	1.000	1.000	1.000
AHBS	0.672	–	0.154	0.964	0.954	0.999	1.000
fRW	0.999	0.872	–	0.966	1.000	1.000	1.000
fLinK	0.192	0.104	0.058	–	0.878	0.980	1.000
fGauK	0.046	0.100	0.000	0.320	–	0.855	0.953
fLapK	0.000	0.003	0.000	0.113	0.201	–	0.966
fNTK	0.000	0.000	0.000	0.000	0.058	0.039	–
$\underline{h=20}$							
CW	–	0.818	0.008	0.946	1.000	1.000	1.000
AHBS	0.413	–	0.117	0.949	0.983	1.000	1.000
fRW	0.999	0.903	–	0.972	1.000	1.000	1.000
fLinK	0.161	0.180	0.057	–	0.956	0.994	0.998
fGauK	0.001	0.045	0.000	0.166	–	0.758	0.949
fLapK	0.000	0.001	0.000	0.028	0.295	–	0.952
fNTK	0.001	0.000	0.001	0.003	0.064	0.059	–

Table A.5: P-values of Diebold-Mariano (DM) tests using RMSE. The reported p-values are the maximum of the p-values of the DM tests using RMSE for put and call options separately. The hypothesis being tested is $H_0 : RMSE_i = RMSE_j$ against a one-sided alternative $H_1 : RMSE_i < RMSE_j$, where model i is the label of the selected row, whereas model j is the label of the selected column. Thus, for each forecasting horizon h, if the reported p-value at the i^{th} row and j^{th} column is less than 0.05, it means that the model i performs statistically better than model j at 5% level of significance.

	CW	AHBS	fRW	fLinK	fGauK	fLapK	fNTK
$h=1$							
CW	–	0.002	1.000	1.000	0.007	0.951	1.000
AHBS	1.000	–	1.000	1.000	0.039	0.997	1.000
fRW	0.000	0.000	–	1.000	0.000	0.000	0.000
fLinK	0.000	0.000	0.071	–	0.000	0.000	0.000
fGauK	1.000	1.000	1.000	1.000	–	1.000	1.000
fLapK	0.880	0.033	1.000	1.000	0.000	–	0.928
fNTK	0.555	0.000	1.000	1.000	0.000	0.133	–
$h=5$							
CW	–	0.583	1.000	1.000	0.930	0.997	1.000
AHBS	1.000	–	1.000	1.000	0.913	0.995	1.000
fRW	0.448	0.000	–	0.932	0.536	0.908	1.000
fLinK	0.039	0.000	0.098	–	0.398	0.746	1.000
fGauK	0.464	0.186	0.471	0.666	–	0.991	0.986
fLapK	0.089	0.011	0.167	0.314	0.062	–	0.972
fNTK	0.000	0.000	0.000	0.000	0.018	0.075	–
$h=10$							
CW	–	0.922	1.000	1.000	1.000	1.000	1.000
AHBS	0.847	–	0.692	0.999	0.998	1.000	1.000
fRW	0.998	0.505	–	0.993	0.999	1.000	1.000
fLinK	0.127	0.019	0.050	–	0.954	0.994	1.000
fGauK	0.019	0.019	0.007	0.132	–	0.731	0.924
fLapK	0.000	0.000	0.000	0.022	0.350	–	0.969
fNTK	0.000	0.000	0.000	0.000	0.096	0.087	–
$h=15$							
CW	–	0.856	1.000	0.996	1.000	1.000	1.000
AHBS	0.629	–	0.430	0.952	0.999	1.000	1.000
fRW	1.000	0.690	–	0.963	1.000	1.000	1.000
fLinK	0.181	0.070	0.081	–	0.987	0.999	1.000
fGauK	0.001	0.010	0.000	0.082	–	0.631	0.974
fLapK	0.000	0.000	0.000	0.022	0.437	–	0.995
fNTK	0.000	0.000	0.000	0.000	0.033	0.005	–
$h=20$							
CW	–	0.866	1.000	0.993	1.000	1.000	1.000
AHBS	0.392	–	0.260	0.924	0.999	1.000	1.000
fRW	1.000	0.828	–	0.976	1.000	1.000	1.000
fLinK	0.127	0.077	0.054	–	0.994	1.000	1.000
fGauK	0.000	0.008	0.000	0.054	–	0.644	0.976
fLapK	0.000	0.000	0.000	0.004	0.399	–	0.994
fNTK	0.000	0.000	0.000	0.000	0.044	0.022	–

Table A.6: P-values of Diebold-Mariano (DM) tests using MAPE. The reported p-values are the maximum of the p-values of the DM tests using MAPE for put and call options separately. The hypothesis being tested is $H_0 : MAPE_i = MAPE_j$ against a one-sided alternative $H_1 : MAPE_i < MAPE_j$, where model i is the label of the selected row, whereas model j is the label of the selected column. For each forecasting horizon h, if the reported p-value at the i^{th} row and j^{th} column is less than 0.05, it means that the model i performs statistically better than model j at 5% level of significance.

A.10 Delta-hedging of call and put options portfolios

Call delta-hedging. In a short call delta-hedging strategy, we sell one call option contract hedged by a long position in delta shares of S&P 500. Our setup relies on trading signals extracted from the predicted IVS and includes only non-zero volume call options at time t. Let Q_t^{sc} be the set of call options whose implied volatilities are predicted to decrease on day $t+h$, or in other words, the set of call options to be sold on day t. The initial investment cost for such a delta-hedged portfolio is $\sum_{i\in Q_t^{sc}}(\Delta_{i,t}^{sc}S_t - C_{i,t})$, where $C_{i,t}$ and $\Delta_{i,t}^{sc} \in [0,1]$ denote the price and the Black-Scholes delta of the call option $i \in Q_t^{sc}$, and S_t is the closing S&P 500 stock price on the day t. The capital required for the portfolio is always positive. To reduce transaction costs, we hold the position for an h-day period without rebalancing the delta-hedges, as described in Goyal and Saretto (2009). The payoff on day $t+h$ is $\sum_{i\in Q_t^{sc}}(\Delta_{i,t}^{sc}S_{t+h} - C_{i,t+h})$. Conversely, long-call delta-hedging involves buying Q_t^{lc} call options while selling shares with delta $\Delta_{i,t}^{lc}$, with $i \in Q_t^{lc}$. Thus, the return R_t^{sc} of the short call delta-hedging portfolio and return R_t^{lc} of a long call delta-hedging portfolio are

$$R_t^{sc} = \frac{\sum_{i\in Q_t^{sc}}(\Delta_{i,t}^{sc}S_{t+h} - C_{i,t+h})}{\sum_{i\in Q_t^{sc}}(\Delta_{i,t}^{sc}S_t - C_{i,t})} - 1 \quad \text{and} \quad R_t^{lc} = 1 - \frac{\sum_{i\in Q_t^{lc}}(\Delta_{i,t}^{lc}S_{t+h} - C_{i,t+h})}{\sum_{i\in Q_t^{lc}}(\Delta_{i,t}^{lc}S_t - C_{i,t})}.$$

Put delta-hedging. The short put delta-hedging strategy involves selling one contract of a put option and delta-hedging the position by shorting the S&P 500 index. Like the call delta-hedging, we use the predicted IVS to extract trading signals and discretely observed put options to build a portfolio. Denote by Q_t^{sp} the set of put options to be sold on the day t, i.e., those put options with IV predicted to decrease on the day $t+h$. On day t, we have a cash inflow of $\sum_{i\in Q_t^{sp}}(-\Delta_{i,t}^{sp}S_t + P_{i,t})$ where $P_{i,t}$ and $\Delta_{i,t}^{sp} \in [-1,0]$ denote the price and the Black-Scholes delta of the put option $i \in Q_t^{p}$. We hold the portfolio for h days without rebalancing, and on day $t+h$, we close the positions and pay a cost of $\sum_{i\in Q_t^{sp}}(-\Delta_{i,t}^{sp}S_{t+h} + P_{i,t+h})$. For long put delta-hedging, we buy one contract of a put option hedged by a long position in the corresponding number of shares of S&P 500. The returns

of the portfolios on the two strategies are, respectively

$$R_t^{sp} = 1 - \frac{\sum_{i \in Q_t^{sp}} (-\Delta_{i,t}^{sp} S_{t+h} + P_{i,t+h})}{\sum_{i \in Q_t^{sp}} (-\Delta_{i,t}^{sp} S_t + P_{i,t})} \quad \text{and} \quad R_t^{lp} = \frac{\sum_{i \in Q_t^{lp}} (-\Delta_{i,t}^{lp} S_{t+h} + P_{i,t+h})}{\sum_{i \in Q_t^{lp}} (-\Delta_{i,t}^{lp} S_t + P_{i,t})} - 1.$$

A.11 Additional results on economic performance

The performance of the remaining short and long strategies is detailed in this section. Overall, fNTK performs the best over different strategies and trading periods, transaction costs, and filtering thresholds. But some differences exist. We review the main findings below.

Compared to delta-hedging strategies, the straddle strategies show higher trading returns in terms of mean simple returns and Sharpe ratio. Additionally, it is worth noting that theoretically, delta-hedging (with put and call together), simple straddle (using one call and one put option) and delta-neutral straddle (described in the main text) strategies using at-the-money (ATM) options, defined by options with $|\Delta| = 0.5$, should yield the same returns. We verified this by using call and put ATM options with $0.48 \leq |\Delta| \leq 0.52$, and our results show that delta-hedging, simple straddle, and delta-neutral strategies with ATM options yield highly similar results , see Table A.9.

Short call delta-hedging strategies, Table A.7 and Figure A.5 is the only one that does not fully benefit from fNTK and is outperformed by the benchmark models. Its performance is surpassed in particular by fRW which records high Sharpe ratios between 4.74 and 5.28, particularly in 2021, while fNTK's records a Sharpe ratio 2.29 to 4.65 in the same period. However, this changes when we increase the filtering threshold to 10%; then fRW performs worse and fNTK performs the best in the longer horizons, e.g., for 20 steps ahead, fRW has a Sharpe ratio of 1.08 while fNTK has a Sharpe ratio of 2.66, as shown in Figure A.7. fNTK scores highest in short-put delta-hedging strategies, Table A.8 and Figure A.6, but the returns are close to zero and negative for most horizons.

For long strategies, the relative performance of the nonlinear kernels is evident in Figures A.14, A.15, A.16 and upper panels of tables A.10, A.11, A.12. fNTK scores highest in the overall per-

formance but the mean returns are relatively small compared to the short strategies. The returns of long delta-hedging strategies are around 0, with a more pronounced negative pattern for put-based strategies. Long delta-neutral straddles score a maximum return of 5.69% for 20-day portfolios, but their Sharpe Ratio is only 0.41. Compared to the performance of short delta-neutral straddles, this marks a sharp decrease in performance. However, this discrepancy in performance between long and short delta-neutral straddles aligns with the existing literature.

When looking at the trading returns across different years, we observe reasonable patterns for long and short strategies: for short strategies, the returns are generally better in the calmer periods of years 2019 and 2021; while for long strategies, most of the returns are from the highly volatile year 2021. Furthermore, across all trading strategies, it is also noteworthy that as we increase the filtering threshold, the nonlinear models, especially the fNTK model, tend to have better returns; but it is not always true for the classical methods, peculiarly the fRW model. With respect to transaction costs, as the effective spread measures, i.e., the trading costs become higher, the returns of the models tend to reduce, but not by significant amount.

A.11.1 Performance of short strategies

	Mean return (%)					Sharpe ratio				
	$h=1$	$h=5$	$h=10$	$h=15$	$h=20$	$h=1$	$h=5$	$h=10$	$h=15$	$h=20$
Overall (from Jan 9, 2019 to Dec 31, 2021)										
CW	0.00	0.26	0.59	1.06	1.56	-0.06	1.18	1.10	1.21	1.35
AHBS	0.01	0.19	0.57	0.96	1.48	0.32	0.97	1.24	1.23	1.30
fRW	0.06	**0.44**	**0.87**	**1.28**	**1.73**	1.63	**2.14**	**1.77**	**1.56**	**1.51**
fLinK	0.06	0.33	0.49	0.64	0.94	1.41	1.72	0.99	0.71	0.82
fGauK	0.05	0.20	0.40	0.63	1.03	1.22	1.03	0.86	0.86	1.29
fLapK	0.05	0.27	0.52	0.63	1.01	1.24	1.35	1.09	0.77	1.12
fNTK	**0.08**	0.27	0.54	0.74	1.16	**1.92**	1.47	1.24	1.08	1.39
From Jan 9, 2019 to Dec 31, 2019										
CW	0.01	0.22	0.53	0.86	1.45	0.53	2.10	2.40	2.33	2.64
AHBS	0.02	0.24	0.86	**1.54**	**2.11**	1.32	2.07	3.46	3.60	**3.69**
fRW	0.06	**0.46**	**0.94**	1.42	1.88	3.07	**4.01**	**4.23**	**3.99**	3.63
fLinK	0.06	0.38	0.61	0.73	0.84	3.33	3.48	3.21	2.11	2.21
fGauK	0.08	0.30	0.61	0.92	1.27	3.97	3.06	3.36	2.33	3.06
fLapK	**0.09**	0.38	0.71	0.99	1.23	**4.34**	3.34	3.74	2.59	2.73
fNTK	**0.09**	0.33	0.58	0.70	1.22	4.12	3.32	3.20	1.96	3.05
From Jan 1, 2020 to Dec 31, 2020										
CW	-0.03	**0.37**	**0.76**	**1.40**	**1.85**	-0.64	**1.06**	**0.90**	**1.02**	**1.03**
AHBS	-0.02	0.00	0.14	0.08	0.66	-0.36	-0.02	0.18	0.05	0.38
fRW	0.03	0.29	0.60	0.88	1.30	0.43	0.92	0.78	0.68	0.71
fLinK	0.03	0.18	0.06	-0.06	0.47	0.39	0.59	0.06	-0.07	0.26
fGauK	-0.03	-0.07	-0.15	-0.06	0.47	-0.55	-0.29	-0.24	-0.09	0.41
fLapK	-0.03	0.03	0.13	-0.18	0.37	-0.58	0.06	0.16	-0.17	0.26
fNTK	**0.06**	0.14	0.33	0.58	1.02	**0.90**	0.49	0.48	0.56	0.82
From Jan 1, 2021 to Dec 31, 2021										
CW	0.02	0.19	0.47	0.91	1.36	1.49	2.73	2.84	2.98	2.95
AHBS	0.04	0.33	0.72	1.31	1.69	2.22	2.95	3.34	3.37	3.26
fRW	**0.10**	**0.57**	**1.06**	**1.56**	**2.04**	**4.96**	**5.28**	**5.07**	**4.79**	**4.74**
fLinK	0.09	0.45	0.83	1.32	1.61	4.73	4.32	3.87	3.78	3.20
fGauK	0.09	0.37	0.78	1.04	1.35	4.68	4.60	4.11	2.71	2.25
fLapK	0.09	0.40	0.75	1.11	1.45	4.45	4.48	3.47	3.11	2.75
fNTK	0.10	0.34	0.73	0.96	1.25	4.65	3.52	3.69	2.60	2.29

Table A.7: Mean simple returns (MR) and annualized Sharpe ratio (SR) of short call delta-hedging over the whole test period, from Jan 9, 2019 to Dec 31, 2021, and in each year of the test period. Bold numbers indicate the best-performing model (or models) in a given column.

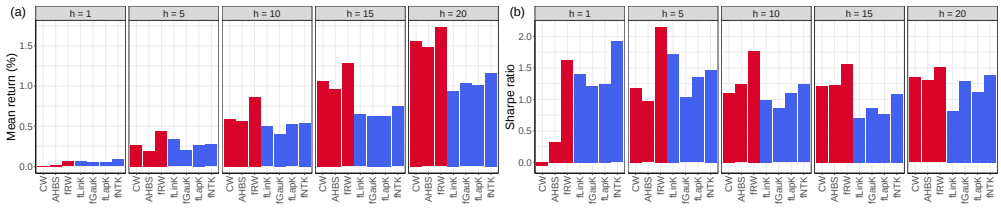


Figure A.5: Mean simple returns (MR) in percentage and Sharpe ratio (SR) of short call delta-hedging strategy. The prediction period is from Jan 09, 2019 to Dec 31, 2021.

	Mean return (%)					Sharpe ratio				
	$h=1$	$h=5$	$h=10$	$h=15$	$h=20$	$h=1$	$h=5$	$h=10$	$h=15$	$h=20$
Overall (from Jan 9, 2019 to Dec 31, 2021)										
CW	**0.03**	**0.07**	-0.03	-0.21	-0.50	**0.67**	**0.28**	-0.12	-0.27	-0.32
AHBS	0.00	0.02	-0.06	-0.20	-0.45	0.04	0.02	-0.19	-0.27	-0.31
fRW	0.00	-0.02	-0.06	-0.22	-0.44	-0.26	-0.20	-0.19	-0.30	-0.31
fLinK	0.00	-0.02	-0.12	-0.35	-0.62	-0.22	-0.19	-0.29	-0.39	-0.39
fGauK	0.00	0.02	0.05	-0.11	0.25	-0.29	0.03	0.04	**-0.12**	0.19
fLapK	0.00	0.01	-0.02	-0.49	-0.39	-0.27	-0.03	-0.11	-0.36	-0.21
fNTK	-0.01	0.05	**0.08**	**-0.09**	**0.30**	-0.42	0.17	**0.06**	**-0.12**	**0.22**
From Jan 9, 2019 to Dec 31, 2019										
CW	**0.03**	**0.13**	0.17	0.18	0.19	**1.51**	**1.12**	0.64	0.30	0.06
AHBS	0.02	0.07	0.14	0.22	**0.31**	0.54	0.33	0.38	0.51	0.57
fRW	0.01	0.06	0.12	0.17	0.23	-0.04	0.23	0.29	0.21	0.25
fLinK	0.01	0.07	**0.18**	**0.25**	**0.31**	0.16	0.28	0.69	**0.66**	**0.60**
fGauK	0.01	0.08	0.17	0.23	0.25	0.06	0.42	**0.73**	0.65	0.31
fLapK	0.01	0.06	0.17	0.22	0.22	-0.11	0.21	0.64	0.51	0.18
fNTK	0.00	0.08	0.17	0.23	0.29	-0.44	0.50	0.70	0.58	0.53
From Jan 1, 2020 to Dec 31, 2020										
CW	**-0.01**	-0.13	-0.62	-1.24	-2.24	**-0.13**	-0.42	-0.71	-0.77	-0.77
AHBS	-0.03	-0.14	-0.57	-1.05	-1.75	-0.77	-0.45	-0.70	-0.73	-0.70
fRW	-0.04	-0.27	-0.65	-1.32	-2.17	-0.93	-0.86	-0.79	-0.89	-0.82
fLinK	-0.04	-0.24	-0.76	-1.55	-2.31	-1.05	-0.76	-0.89	-0.94	-0.83
fGauK	-0.05	-0.19	-0.42	-1.11	-0.34	-1.22	-0.62	-0.57	**-0.51**	-0.26
fLapK	-0.04	-0.19	-0.57	-2.00	-1.75	-1.01	-0.61	-0.66	-0.83	-0.52
fNTK	-0.05	**-0.10**	**-0.31**	**-1.09**	**-0.21**	-1.16	**-0.32**	**-0.31**	-0.57	**-0.14**
From Jan 1, 2021 to Dec 31, 2021										
CW	**0.05**	**0.22**	0.37	0.46	0.62	**2.78**	3.05	2.81	2.21	2.08
AHBS	0.03	0.14	0.32	0.47	0.82	1.73	2.11	2.37	2.34	3.13
fRW	0.02	0.15	0.39	0.54	0.74	1.29	2.53	3.91	3.87	3.22
fLinK	0.03	0.14	0.34	0.56	0.83	1.46	2.01	2.51	2.63	3.65
fGauK	0.03	0.20	**0.50**	**0.77**	**1.05**	1.54	**3.29**	**4.82**	**5.17**	**5.20**
fLapK	0.02	0.18	0.43	0.70	0.93	1.29	2.59	3.90	4.84	4.41
fNTK	0.03	0.18	0.42	0.69	0.89	1.38	2.84	4.08	4.99	4.25

Table A.8: Mean simple returns in percentage and annualized Sharpe ratio of short put delta-hedging over the whole test period, from Jan 9, 2019 to Dec 31, 2021, and in each year of the test period. Bold numbers indicate the best-performing model (or models) in a given column.

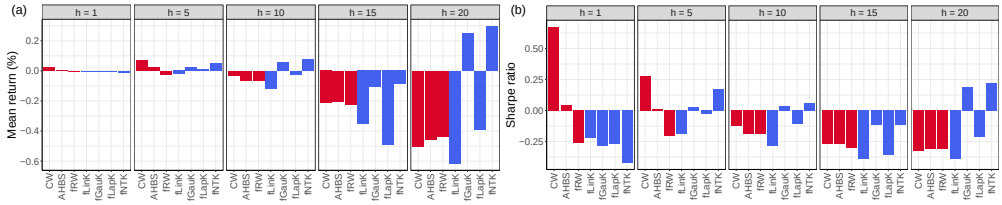


Figure A.6: Mean simple returns (MR) in percentage and annualized Sharpe ratio of short put delta-hedging strategy. The prediction period is from Jan 09, 2019 to Dec 31, 2021.

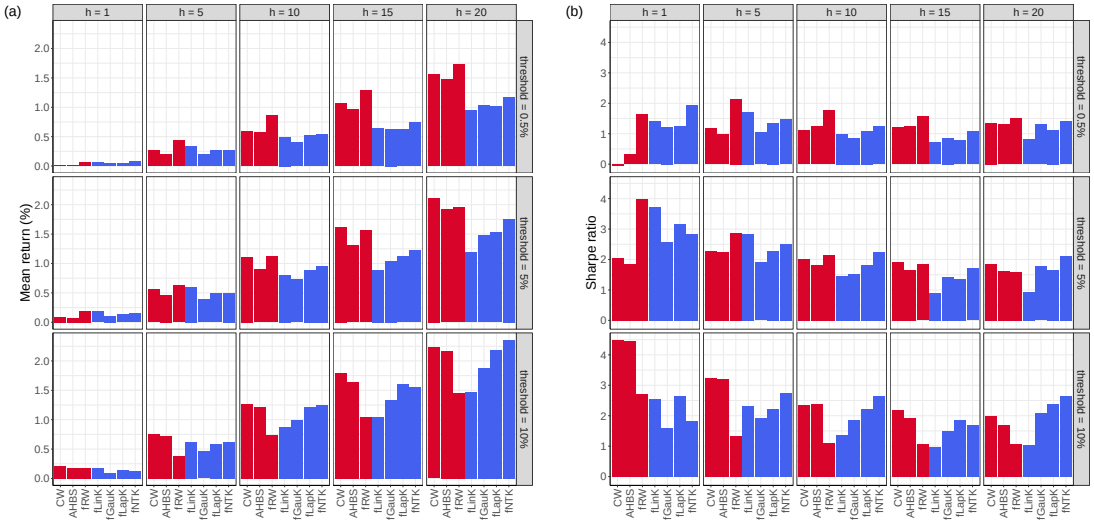


Figure A.7: Mean simple returns (MR) in percentage and annualized Sharpe ratio of short call delta-hedging strategy using three levels of filtering thresholds: 0.5%, 5% and 10%. The prediction period is from Jan 09, 2019 to Dec 31, 2021.

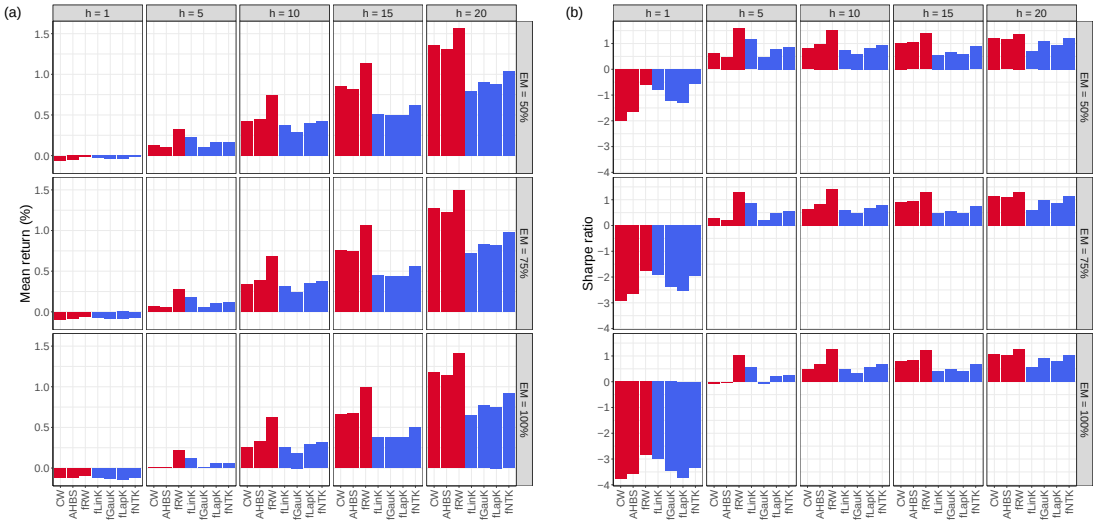


Figure A.8: Mean simple returns (MR) in percentage and annualized Sharpe ratio of short call delta-hedging strategy using three levels of effective measure (EM): 50%, 75% and 100%. The prediction period is from Jan 09, 2019 to Dec 31, 2021.

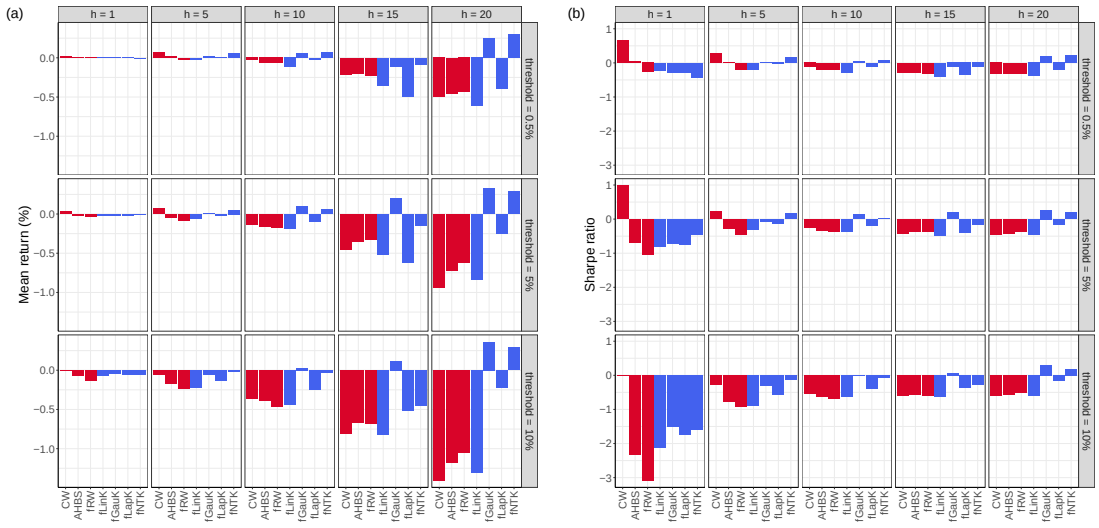


Figure A.9: Mean simple returns (%) and annualized Sharpe ratio of short put delta-hedging strategy using three levels of filtering thresholds: 0.5%, 5% and 10%. The prediction period is from Jan 09, 2019 to Dec 31, 2021.

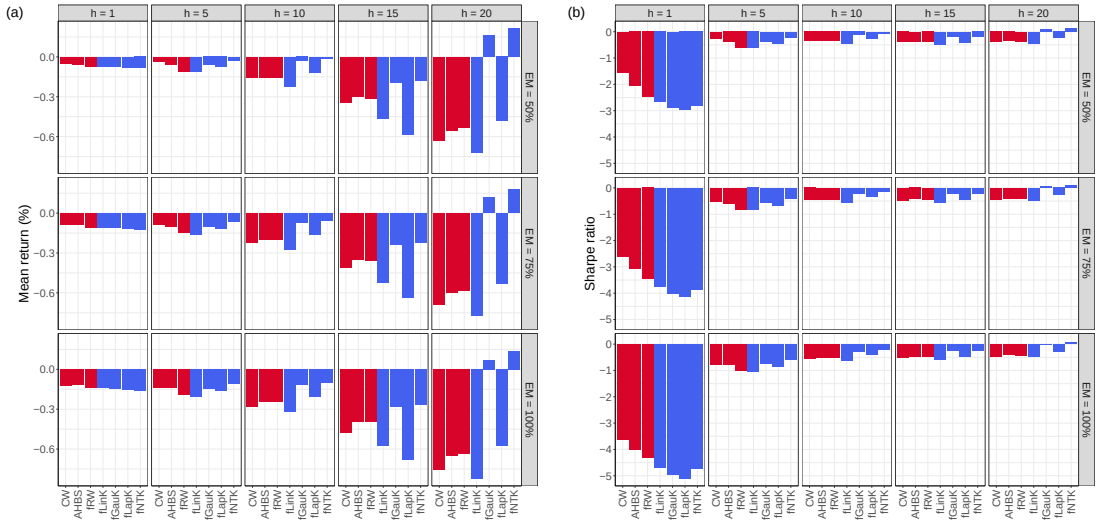


Figure A.10: Mean simple returns (%) and annualized Sharpe ratio of short put delta-hedging strategy using three levels of effective measure (EM): 50%, 75% and 100%. The prediction period is from Jan 09, 2019 to Dec 31, 2021.

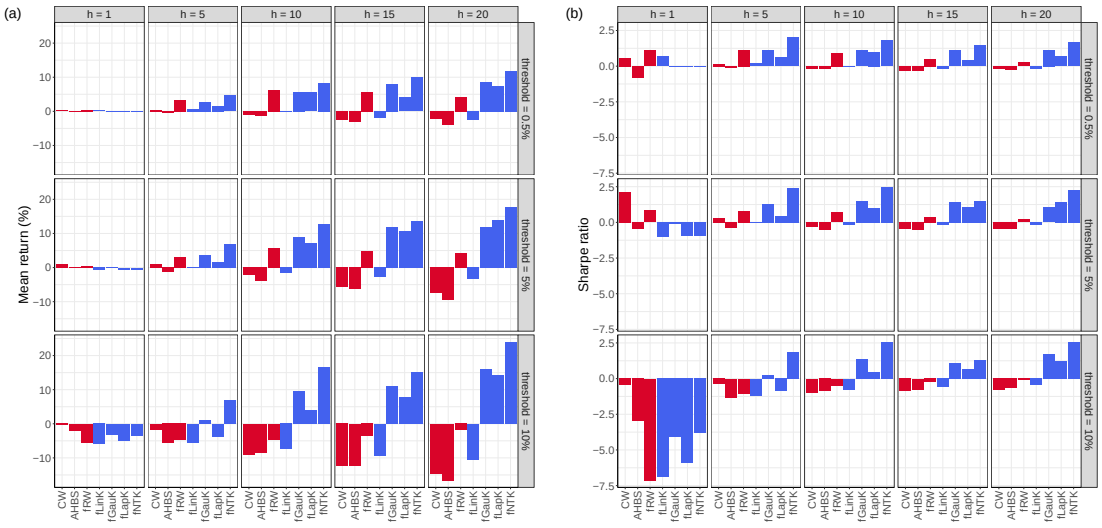


Figure A.11: Mean simple returns in percentage and Sharpe ratio of short delta-neutral straddle strategy using three levels of filtering thresholds: 0.5%, 5% and 10%. The prediction period is from Jan 09, 2019 to Dec 31, 2021.

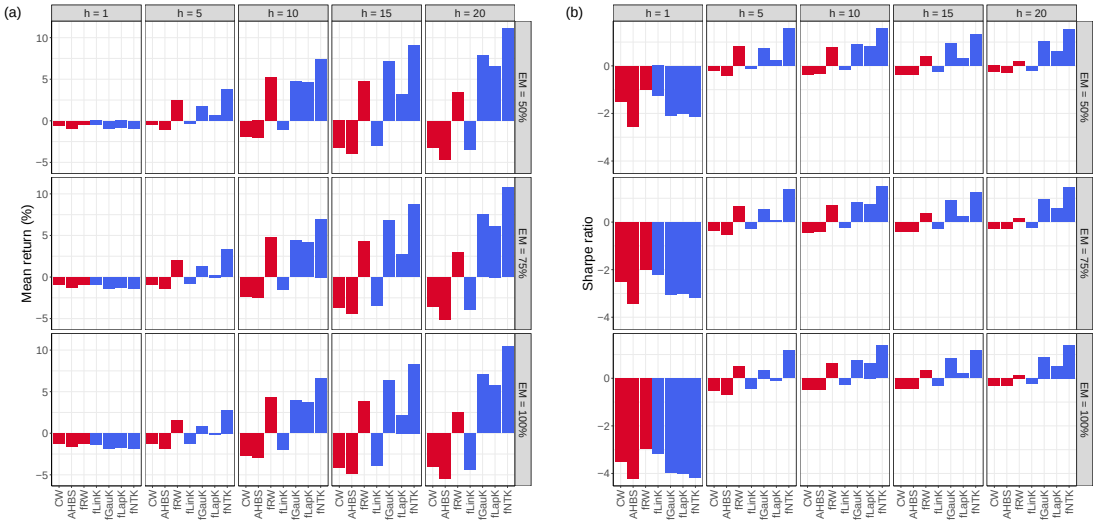


Figure A.12: Mean simple returns in percentage and Sharpe ratio of short delta-neutral straddle strategy using three levels of effective measure (EM): 50%, 75% and 100%. The prediction period is from Jan 09, 2019 to Dec 31, 2021.

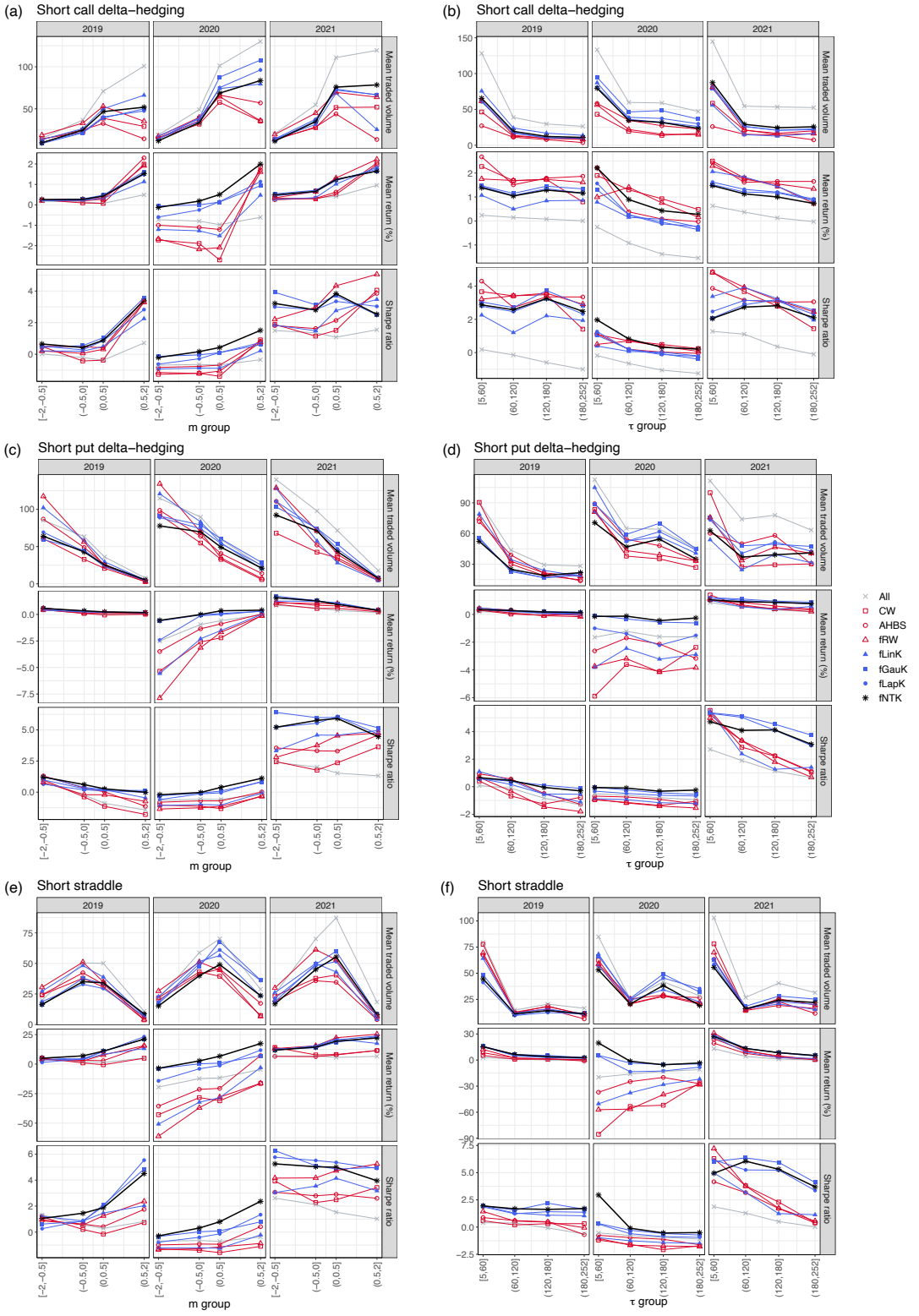


Figure A.13: Mean traded volume, mean simple returns in percentage, and annualized Sharpe ratio of the models for short call delta-hedging, short-put delta-hedging, and short delta-neutral straddle at the forecasting horizon $h = 20$, across different moneyness m groups $[-2,-0.5]$, $(-0.5,0]$, $(0,0.5]$, and $(0.5,2]$, and time-to-maturity τ groups $[5,60]$, $(60,120]$, $(120,180]$, and $(180,252]$, and in three prediction periods: 2019, 2020, and 2021.

	Mean return (%)					Sharpe ratio				
	$h=1$	$h=5$	$h=10$	$h=15$	$h=20$	$h=1$	$h=5$	$h=10$	$h=15$	$h=20$
Short delta-hedging										
CW	-0.43	-1.48	-6.22	-8.36	-8.78	-0.97	-0.27	-0.98	-0.74	-0.56
AHBS	-0.55	-0.70	-4.50	-7.26	-10.15	-1.37	-0.23	-0.66	-0.57	-0.52
fRW	**0.32**	1.90	-0.49	-4.21	-7.49	**0.70**	0.55	-0.08	-0.34	-0.32
fLinK	0.00	-1.07	-4.20	-7.86	-9.80	0.00	-0.34	-0.58	-0.60	-0.49
fGauK	0.04	2.02	3.65	3.87	6.54	0.07	0.65	0.74	0.53	0.85
fLapK	-0.31	1.36	1.84	2.89	5.33	-0.66	0.47	0.30	0.34	0.54
fNTK	-0.60	**2.95**	**6.23**	**7.47**	**10.08**	-1.26	**1.07**	**1.16**	**1.33**	**1.58**
Simple short straddle										
CW	-0.48	-2.76	-6.21	-7.93	-8.07	-1.25	-1.07	-1.00	-0.76	-0.58
AHBS	-0.42	-0.91	-4.35	-6.77	-9.44	-1.19	-0.36	-0.66	-0.58	-0.53
fRW	**0.22**	1.28	-0.25	-3.99	-5.04	**0.55**	0.46	-0.04	-0.35	-0.28
fLinK	-0.31	-0.85	-4.16	-7.23	-9.03	-0.76	-0.30	-0.59	-0.60	-0.50
fGauK	-0.42	1.65	3.28	3.60	6.77	-1.00	0.71	0.65	0.49	0.96
fLapK	-0.36	1.00	1.60	2.82	5.40	-0.85	0.40	0.27	0.34	0.63
fNTK	-0.65	**3.39**	**6.74**	**7.53**	**10.41**	-1.54	**1.47**	**1.61**	**1.39**	**1.95**
Delta-neutral short straddle										
CW	-0.49	-2.79	-6.23	-8.01	-8.17	-1.27	-1.08	-1.01	-0.77	-0.58
AHBS	-0.42	-0.93	-4.36	-6.80	-9.52	-1.19	-0.36	-0.66	-0.59	-0.53
fRW	**0.22**	1.25	-0.29	-4.06	-5.16	**0.54**	0.45	-0.05	-0.35	-0.29
fLinK	-0.32	-0.88	-4.19	-7.31	-9.10	-0.77	-0.31	-0.59	-0.60	-0.50
fGauK	-0.43	1.65	3.32	3.62	6.78	-1.02	0.71	0.66	0.49	0.96
fLapK	-0.37	0.98	1.61	2.80	5.37	-0.86	0.39	0.27	0.34	0.62
fNTK	-0.66	**3.34**	**6.74**	**7.51**	**10.37**	-1.55	**1.44**	**1.61**	**1.39**	**1.93**

Table A.9: Mean simple return (%) and annualized Sharpe ratio of short delta-hedging (using both call and put options), simple short straddle (consisting of one call and one put option), and short delta-neutral straddle for at-the-money options, defined by $|\Delta| \in [0.48, 5.02]$, where Δ is the delta of the options. Bold numbers indicate the best-performing model (or models) in a given column.

A.11.2 Performance of long strategies

	Mean return (%)					Sharpe ratio				
	$h=1$	$h=5$	$h=10$	$h=15$	$h=20$	$h=1$	$h=5$	$h=10$	$h=15$	$h=20$
Overall (from Jan 9, 2019 to Dec 31, 2021)										
CW	0.01	-0.04	-0.11	0.05	0.13	0.22	-0.34	-0.34	-0.01	0.06
AHBS	0.03	0.08	0.06	0.11	0.08	0.73	0.31	0.05	0.08	0.01
fRW	**0.06**	0.16	0.23	0.23	0.23	1.33	0.68	0.40	0.25	0.18
fLinK	**0.06**	0.18	0.25	0.34	0.40	1.38	0.77	0.47	0.39	0.34
fGauK	0.05	0.11	0.21	**0.47**	0.54	**1.54**	0.73	0.52	**0.59**	0.52
fLapK	0.05	0.13	0.25	0.40	0.54	1.32	0.67	0.50	0.52	0.50
fNTK	**0.06**	**0.19**	**0.31**	0.44	**0.56**	1.34	**0.84**	**0.63**	0.54	**0.54**
From Jan 9, 2019 to Dec 31, 2019										
CW	-0.02	-0.11	-0.24	-0.20	-0.18	-1.56	-1.95	-2.78	-1.46	-1.31
AHBS	0.00	-0.04	-0.09	-0.06	-0.05	-0.39	-0.94	-1.19	-0.80	-0.80
fRW	**0.02**	0.01	-0.01	0.01	0.01	**0.54**	-0.34	-0.56	-0.46	-0.56
fLinK	**0.02**	**0.02**	**0.00**	**0.04**	**0.17**	0.34	**-0.29**	**-0.50**	**-0.35**	**-0.03**
fGauK	0.01	0.01	-0.03	**0.04**	0.08	0.19	-0.38	-0.88	-0.40	-0.35
fLapK	0.01	0.00	-0.03	0.03	0.09	0.07	-0.46	-0.76	-0.42	-0.31
fNTK	0.01	0.01	**0.00**	0.01	0.11	-0.04	-0.45	-0.60	-0.55	-0.23
From Jan 1, 2020 to Dec 31, 2020										
CW	0.06	0.12	0.24	0.64	0.85	1.09	0.43	0.32	0.54	0.52
AHBS	0.11	0.39	0.54	0.72	0.69	1.64	1.17	0.69	0.67	0.54
fRW	0.15	0.47	0.76	0.73	0.72	2.20	1.37	0.95	0.65	0.51
fLinK	**0.16**	0.58	**0.99**	1.26	1.38	2.41	1.68	1.31	1.03	0.86
fGauK	0.14	0.41	0.91	**1.74**	**1.92**	**2.80**	**1.99**	**1.60**	**1.43**	**1.22**
fLapK	0.15	0.45	0.94	1.34	1.70	2.51	1.65	1.34	1.20	1.10
fNTK	**0.16**	**0.59**	0.98	1.36	1.49	2.36	1.74	1.33	1.14	1.01
From Jan 1, 2021 to Dec 31, 2021										
CW	-0.01	-0.13	-0.34	-0.35	-0.35	-0.72	-1.58	-2.03	-1.25	-0.96
AHBS	-0.01	-0.11	-0.28	-0.35	-0.41	-0.69	-1.40	-1.65	-1.25	-1.09
fRW	0.01	**0.00**	-0.08	-0.09	-0.07	0.43	**-0.02**	-0.45	-0.31	-0.19
fLinK	0.01	-0.04	-0.19	-0.24	-0.30	0.29	-0.43	-1.05	-0.82	-0.78
fGauK	**0.02**	-0.01	-0.08	-0.08	-0.07	**1.04**	-0.17	-0.53	-0.31	-0.22
fLapK	0.00	-0.03	-0.11	-0.08	-0.09	0.15	-0.33	-0.68	-0.32	-0.27
fNTK	0.01	-0.02	**-0.04**	**-0.04**	**0.03**	0.59	-0.32	**-0.26**	**-0.19**	**0.06**

Table A.10: Mean simple returns (%) and annualized Sharpe ratio of long call delta-hedging over the whole test period, from Jan 9, 2019 to Dec 31, 2021, and in each year of the test period. Bold numbers indicate the best-performing model (or models) in a given column.

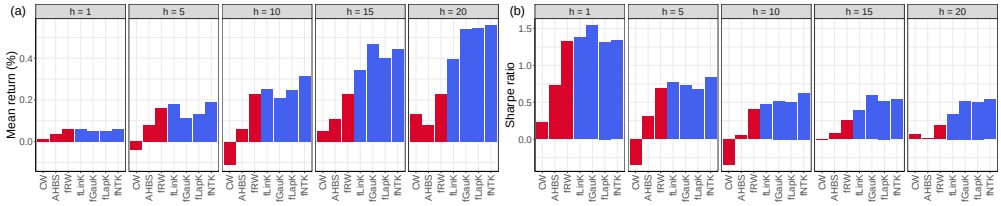


Figure A.14: Mean simple returns (%) and annualized Sharpe ratio of long call delta-hedging strategy. The prediction period is from Jan 09, 2019 to Dec 31, 2021.

	Mean return (%)					Sharpe ratio				
	$h=1$	$h=5$	$h=10$	$h=15$	$h=20$	$h=1$	$h=5$	$h=10$	$h=15$	$h=20$
Overall (from Jan 9, 2019 to Dec 31, 2021)										
CW	**-0.02**	-0.14	-0.26	-0.31	-0.21	**-1.60**	-1.77	-1.36	-0.96	-0.36
AHBS	-0.04	-0.20	-0.32	-0.34	-0.36	-2.57	-1.99	-1.80	-1.57	-1.27
fRW	-0.05	-0.21	-0.34	-0.32	-0.32	-3.01	-2.26	-1.62	-1.11	-0.80
fLinK	-0.05	-0.21	-0.23	-0.23	-0.23	-3.39	-2.37	-1.26	-0.95	-0.66
fGauK	-0.04	-0.10	-0.09	0.10	**0.35**	-4.22	-1.18	-0.47	0.05	**0.19**
fLapK	-0.05	-0.15	-0.13	0.02	0.26	-4.37	-1.95	-0.66	-0.05	0.14
fNTK	-0.04	**-0.04**	**-0.02**	**0.13**	0.30	-2.64	**-0.28**	**-0.15**	**0.10**	0.17
From Jan 9, 2019 to Dec 31, 2019										
CW	**-0.03**	-0.12	-0.22	-0.25	-0.27	**-3.73**	-3.50	-3.23	-2.48	-2.14
AHBS	-0.04	-0.12	-0.20	-0.21	-0.25	-4.67	-3.62	-3.26	-2.50	-1.96
fRW	-0.04	-0.16	-0.26	-0.26	-0.23	-5.60	-3.70	-3.02	-2.37	-1.85
fLinK	-0.05	-0.13	-0.24	-0.26	-0.27	-5.80	-3.20	-3.04	-2.47	-1.78
fGauK	-0.04	**-0.07**	**-0.11**	**-0.13**	-0.16	-5.41	**-1.78**	**-1.60**	**-1.44**	-1.55
fLapK	-0.04	-0.10	-0.12	-0.15	-0.16	-5.08	-2.26	-1.84	-1.64	-1.62
fNTK	**-0.03**	-0.08	**-0.11**	-0.16	**-0.12**	-4.31	-1.94	-1.68	-1.65	**-1.37**
From Jan 1, 2020 to Dec 31, 2020										
CW	**0.00**	-0.14	-0.26	-0.26	0.18	**-0.25**	-1.21	-0.82	-0.48	0.11
AHBS	-0.03	-0.24	-0.30	-0.27	-0.18	-1.29	-1.46	-1.04	-0.83	-0.49
fRW	-0.04	-0.20	-0.30	-0.21	-0.13	-1.59	-1.40	-0.91	-0.47	-0.22
fLinK	-0.04	-0.23	-0.05	0.07	0.21	-1.74	-1.69	-0.20	0.09	0.24
fGauK	-0.03	0.00	0.28	**1.09**	**2.20**	-1.93	-0.05	0.56	0.84	**0.82**
fLapK	-0.06	-0.14	0.15	0.73	1.59	-3.38	-1.14	0.33	0.71	0.68
fNTK	-0.03	**0.16**	**0.41**	1.08	1.55	-1.25	**0.45**	**0.68**	**0.85**	0.67
From Jan 1, 2021 to Dec 31, 2021										
CW	**-0.04**	**-0.15**	**-0.31**	-0.42	-0.55	**-3.24**	**-1.98**	**-1.93**	**-1.68**	**-1.60**
AHBS	-0.05	-0.24	-0.47	-0.55	-0.66	-4.46	-3.69	-3.62	-2.70	-2.16
fRW	-0.06	-0.26	-0.45	-0.50	-0.62	-4.77	-4.13	-3.04	-2.21	-1.88
fLinK	-0.07	-0.25	-0.41	-0.46	-0.59	-6.82	-4.42	-3.25	-2.32	-2.02
fGauK	-0.06	-0.20	-0.36	**-0.42**	**-0.53**	-6.28	-3.02	-2.52	-1.93	-1.62
fLapK	-0.06	-0.23	-0.42	-0.44	-0.57	-5.92	-3.47	-2.99	-2.13	-1.78
fNTK	-0.07	-0.20	-0.36	-0.50	-0.58	-5.49	-3.15	-2.71	-2.22	-1.79

Table A.11: Mean simple returns (%) and annualized Sharpe ratio of long put delta-hedging over the whole test period, from Jan 9, 2019 to Dec 31, 2021, and in each year of the test period. Bold numbers indicate the best-performing model (or models) in a given column.

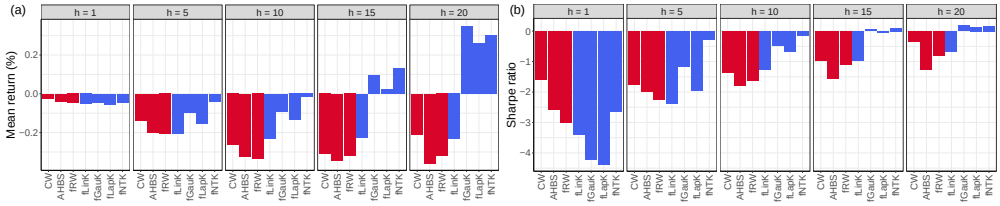


Figure A.15: Mean simple returns (%) and annualized Sharpe ratio of long put delta-hedging strategy. The prediction period is from Jan 09, 2019 to Dec 31, 2021.

	Mean return (%)					Sharpe ratio				
	$h=1$	$h=5$	$h=10$	$h=15$	$h=20$	$h=1$	$h=5$	$h=10$	$h=15$	$h=20$
Overall (from Jan 9, 2019 to Dec 31, 2021)										
CW	**-0.40**	-4.41	-5.48	-2.79	2.64	**-0.90**	-3.08	-1.52	-0.36	0.15
AHBS	-0.76	-3.29	-4.48	-3.52	-3.60	-2.34	-2.19	-1.49	-0.90	-0.80
fRW	-0.83	-1.86	-1.84	-2.37	-2.57	-2.15	-0.77	-0.39	-0.53	-0.60
fLinK	-0.73	-2.91	-2.23	-2.27	-3.30	-1.89	-1.34	-0.42	-0.34	-0.71
fGauK	-0.91	-0.11	0.58	3.57	5.57	-2.78	-0.04	0.10	0.38	0.39
fLapK	-1.04	-1.85	1.22	3.51	5.50	-2.91	-0.80	0.19	0.37	0.39
fNTK	-0.90	**0.27**	**1.90**	**4.09**	**5.69**	-2.60	**0.09**	**0.31**	**0.42**	**0.41**
From Jan 9, 2019 to Dec 31, 2019										
CW	-0.93	-4.46	-5.66	-3.77	-3.10	-2.70	-3.54	-2.94	-1.39	-1.06
AHBS	-0.92	-3.13	-4.52	-3.71	-3.49	-3.39	-2.74	-2.38	-1.27	-1.01
fRW	-0.99	-2.39	-2.98	-3.42	-2.86	-2.84	-1.28	-1.12	-1.16	-0.83
fLinK	**-0.61**	-2.75	-4.20	-5.10	-2.42	**-1.46**	-1.63	-1.26	-1.50	-0.75
fGauK	**-0.61**	**0.11**	-0.15	**-0.53**	**0.29**	-1.66	**0.03**	-0.07	**-0.18**	**0.03**
fLapK	-0.75	-1.52	**0.01**	-1.33	-0.47	-2.13	-0.91	**-0.02**	-0.42	-0.17
fNTK	-0.96	0.10	-0.48	-1.67	-0.06	-2.64	0.02	-0.16	-0.54	-0.06
From Jan 1, 2020 to Dec 31, 2020										
CW	-0.15	-3.27	-2.03	1.54	14.99	-0.33	-2.26	-0.39	0.12	0.51
AHBS	**-0.14**	-1.53	-1.68	0.00	-0.16	**-0.32**	-0.78	-0.40	0.00	-0.03
fRW	-0.22	0.65	2.32	0.27	-1.26	-0.47	0.19	0.32	0.04	-0.23
fLinK	-0.27	-0.89	4.91	5.29	0.35	-0.61	-0.29	0.57	0.47	0.05
fGauK	-0.91	3.86	9.85	**20.19**	**24.81**	-2.48	0.78	1.05	**1.23**	**0.96**
fLapK	-1.00	-0.07	9.24	17.66	22.63	-2.54	-0.02	0.92	1.13	0.94
fNTK	-0.60	**4.39**	**11.15**	18.71	20.49	-1.58	**1.04**	**1.20**	1.20	0.91
From Jan 1, 2021 to Dec 31, 2021										
CW	**-0.16**	-5.51	-8.81	-6.41	-4.73	**-0.29**	-3.52	-3.48	-2.02	-1.32
AHBS	-1.24	-5.21	-7.25	-6.70	-6.97	-4.76	-4.26	-3.29	-2.23	-1.74
fRW	-1.31	-4.08	-5.21	-4.32	-3.71	-3.96	-2.79	-2.28	-1.53	-1.07
fLinK	-1.33	-4.91	-6.57	-5.89	-6.74	-4.83	-3.46	-2.89	-1.95	-1.70
fGauK	-1.24	**-3.22**	-5.12	**-4.08**	-3.65	-5.05	**-2.15**	-2.01	**-1.26**	-0.89
fLapK	-1.40	-3.80	-5.28	-4.23	-4.01	-4.24	-2.64	-2.23	-1.35	-1.01
fNTK	-1.14	-3.46	**-4.56**	-4.20	**-3.03**	-3.80	-2.49	**-1.77**	-1.36	**-0.76**

Table A.12: Mean simple returns (%) and annualized Sharpe ratio of long delta-neutral straddles over the whole test period, from Jan 9, 2019 to Dec 31, 2021, and in each year of the test period. Bold numbers indicate the best-performing model (or models) in a given column.

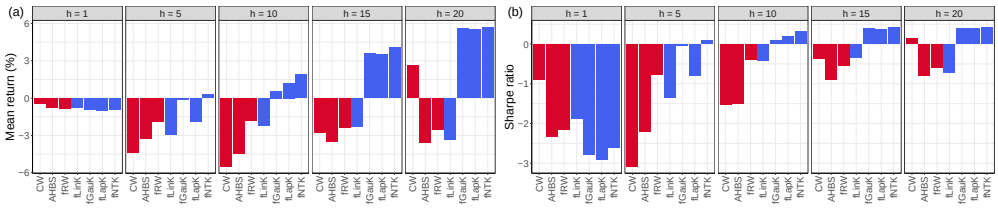


Figure A.16: Mean simple returns (%) and annualized Sharpe ratio of long delta-neutral straddle strategy. The prediction period is from Jan 09, 2019 to Dec 31, 2021.

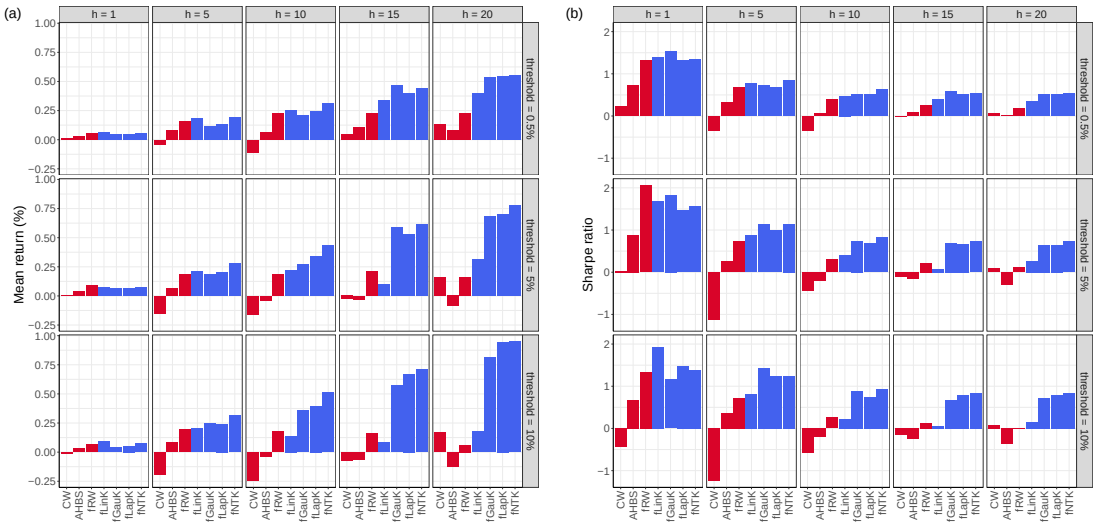


Figure A.17: Mean simple returns (%) and annualized Sharpe ratio of long call delta-hedging strategy using three levels of filtering thresholds: 0.5%, 5% and 10%. The prediction period is from Jan 09, 2019 to Dec 31, 2021.

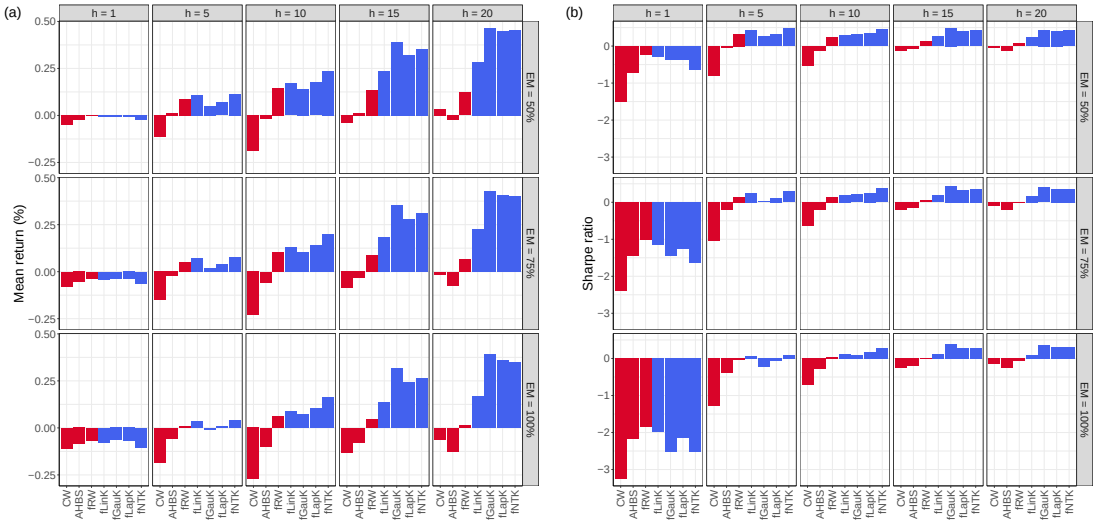


Figure A.18: Mean simple returns (%) and annualized Sharpe ratio of long call delta-hedging strategy using three levels of effective measure (EM): 50%, 75% and 100%. The prediction period is from Jan 09, 2019 to Dec 31, 2021.

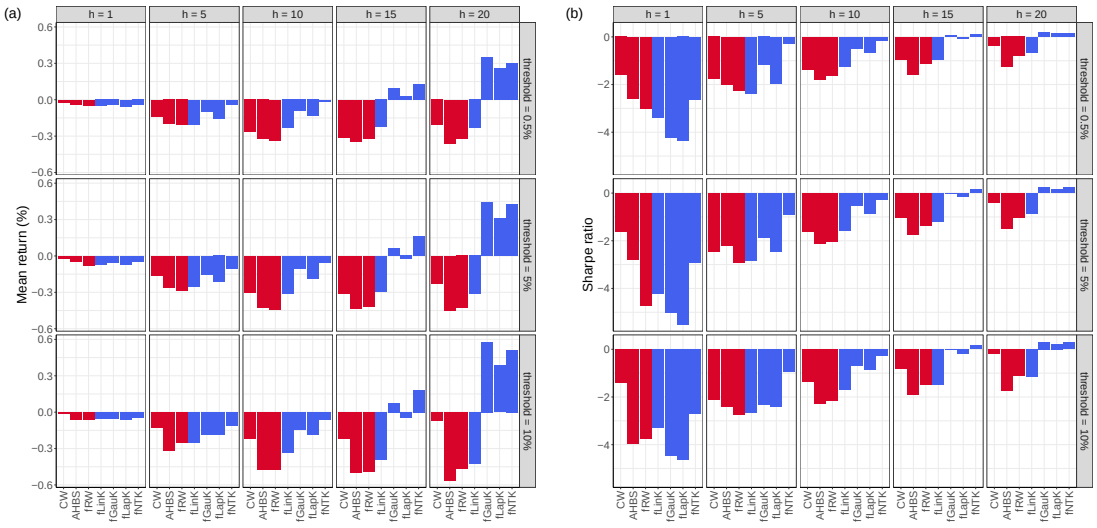


Figure A.19: Mean simple returns (%) and annualized Sharpe ratio of long put delta-hedging strategy using three levels of filtering thresholds: 0.5%, 5% and 10%. The prediction period is from Jan 09, 2019 to Dec 31, 2021.

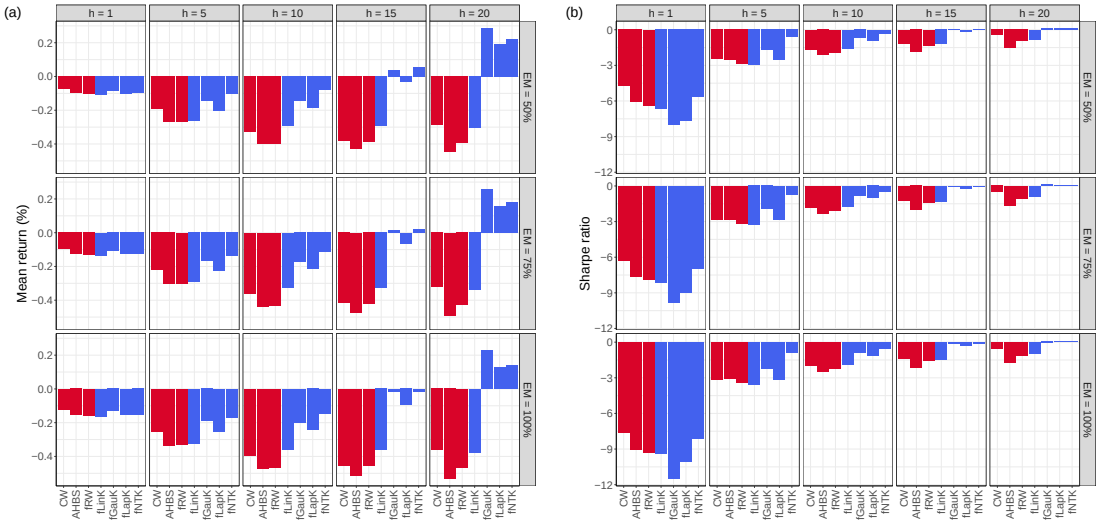


Figure A.20: Mean simple returns (%) and annualized Sharpe ratio of long put delta-hedging strategy using three levels of effective measure (EM): 50%, 75% and 100%. The prediction period is from Jan 09, 2019 to Dec 31, 2021.

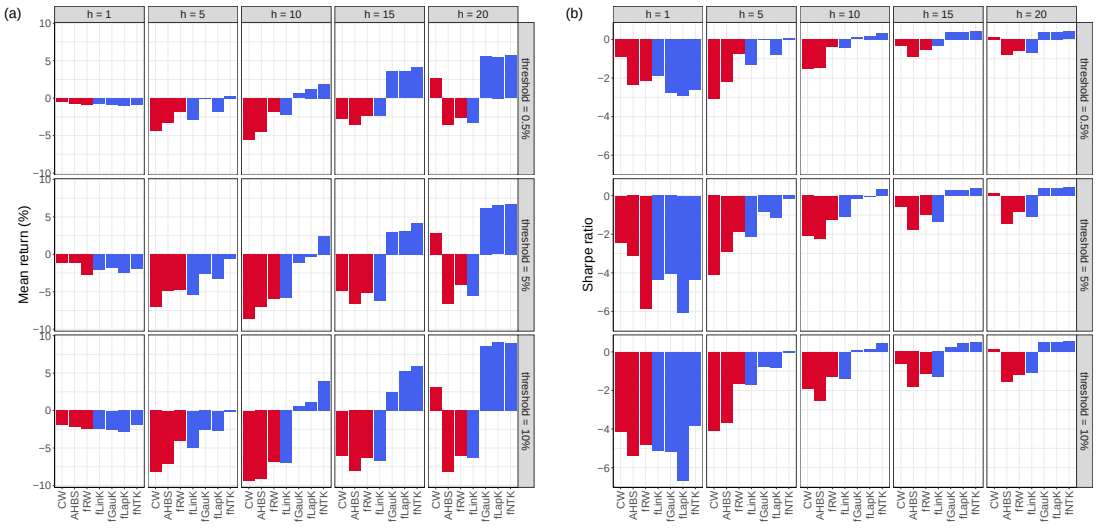


Figure A.21: Mean simple returns (%) and annualized Sharpe ratio of long delta-neutral straddle strategy using three levels of filtering thresholds: 0.5%, 5% and 10%. The prediction period is from Jan 09, 2019 to Dec 31, 2021.

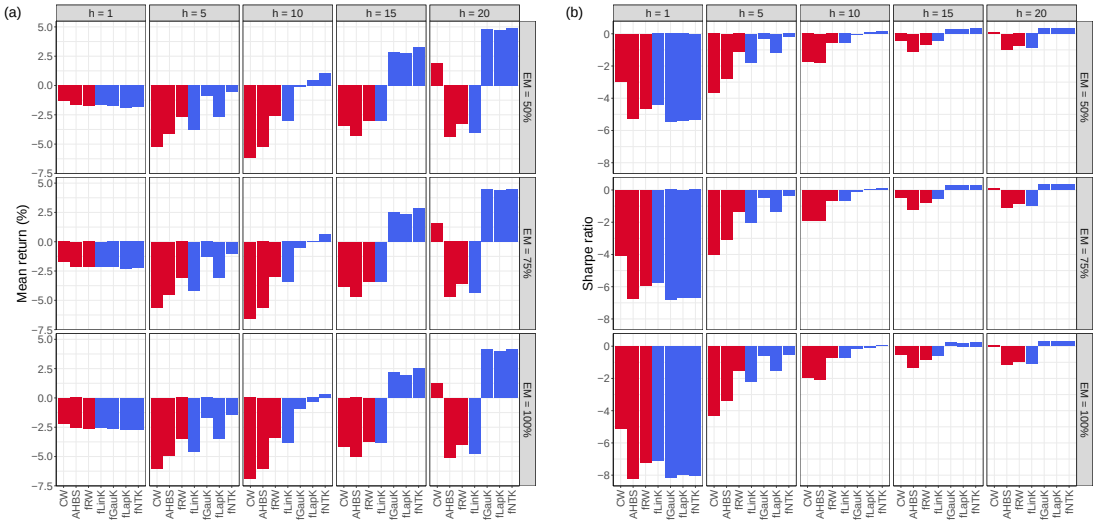


Figure A.22: Mean simple returns (%) and annualized Sharpe ratio of long delta-neutral straddle strategy using three levels of effective measure (EM): 50%, 75% and 100%. The prediction period is from Jan 09, 2019 to Dec 31, 2021.

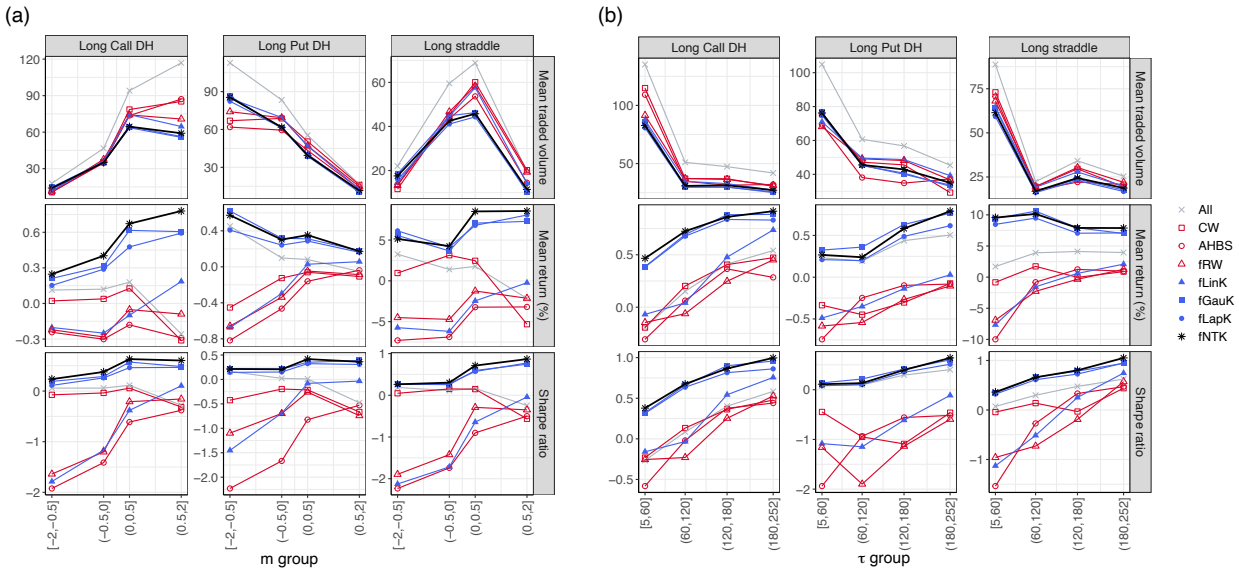


Figure A.23: Mean traded volume, level mean simple returns (MR) in percentage, level Sharpe ratio (SR) of the models for all the long trading strategies at the forecasting horizon $h = 20$, across different moneyness m groups $[-2,-0.5]$, $(-0.5,0]$, $(0,0.5]$, and $(0.5,2]$, and time-to-maturity τ groups $[5,60]$, $(60,120]$, $(120,180]$, and $(180,252]$. The prediction period is from Jan 09, 2019 to Dec 31, 2021.

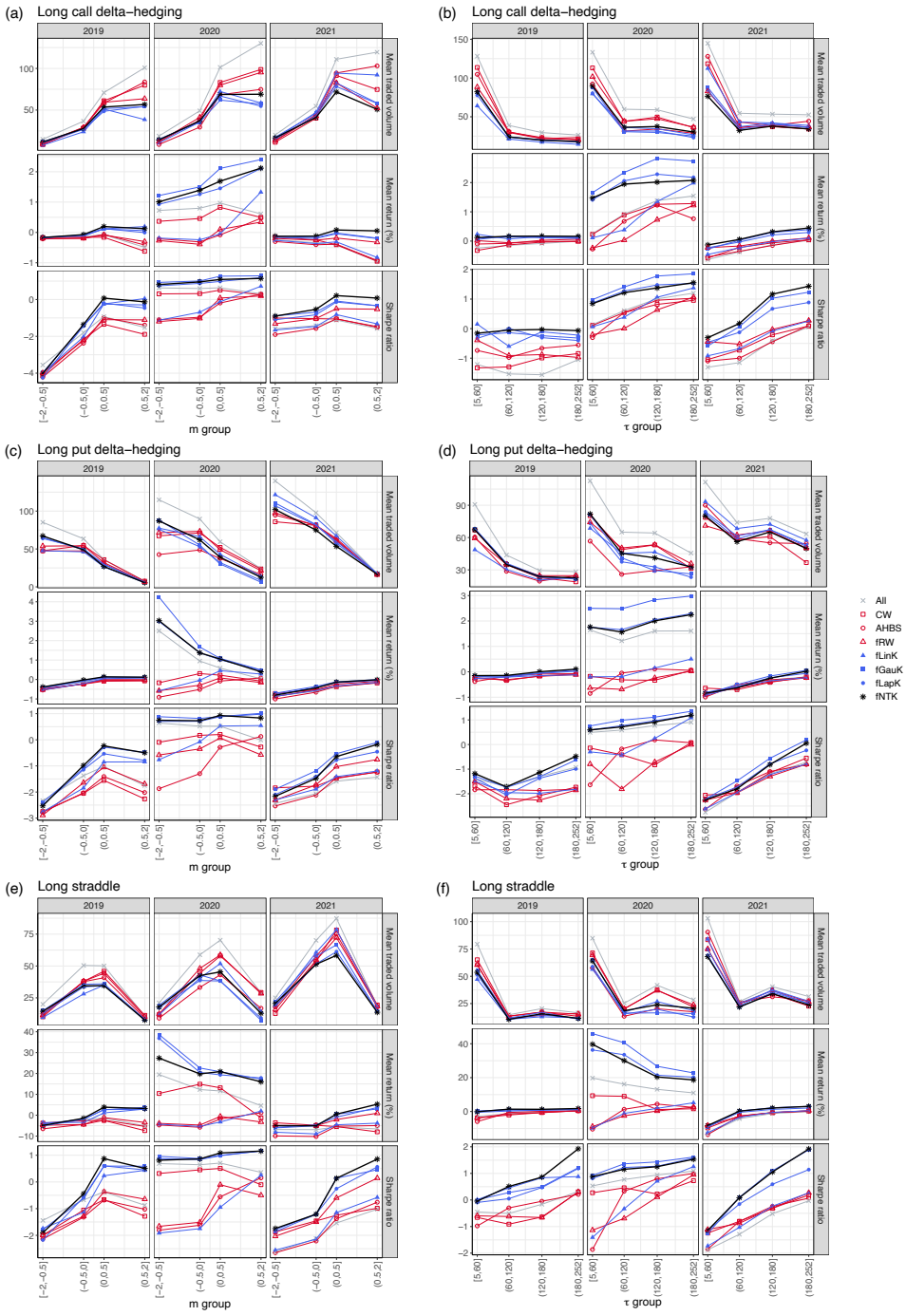


Figure A.24: Mean traded volume, mean simple returns in percentage, and annualized Sharpe ratio of the models for long call delta-hedging, long put delta-hedging, and long delta-neutral straddle at the forecasting horizon $h = 20$, across different moneyness m groups $[-2,-0.5]$, $(-0.5,0]$, $(0,0.5]$, and $(0.5,2]$, and time-to-maturity τ groups $[5,60]$, $(60,120]$, $(120,180]$, and $(180,252]$, and in three prediction periods: 2019, 2020, and 2021.

References

Carr, P. and L. Wu (2016). "Analyzing volatility risk and risk premium in option contracts: A new theory". *Journal of Financial Economics*, 120.1, 1–20.

Chen, X. (2007). "Large sample sieve estimation of semi-nonparametric models". *Handbook of econometrics*, 6, 5549–5632.

Diebold, F. X. and R. S. Mariano (2002). "Comparing predictive accuracy". *Journal of Business & economic statistics*, 20.1, 134–144.

Dumas, B., J. Fleming, and R. E. Whaley (1998). "Implied volatility functions: Empirical tests". *The Journal of Finance*, 53.6, 2059–2106.

Goncalves, S. and M. Guidolin (2006). "Predictable dynamics in the S&P 500 index options implied volatility surface". *The Journal of Business*, 79.3, 1591–1635.

Goyal, A. and A. Saretto (2009). "Cross-section of option returns and volatility". *Journal of Financial Economics*, 94.2, 310–326.

Klepsch, J. and C. Klüppelberg (2017). "An innovations algorithm for the prediction of functional linear processes". *Journal of Multivariate Analysis*, 155, 252–271.

Otto, S. and N. Salish (2022). "Approximate Factor Models for Functional Time Series". *arXiv*.

Ramsay, J. and B. Silverman (2005). *Functional Data Analysis*. Springer Series in Statistics. Springer. ISBN: 9780387400808.

Schölkopf, B., R. Herbrich, and A. J. Smola (2001). "A generalized representer theorem". *International conference on computational learning theory*. Springer, 416–426.

Yao, F., H.-G. Müller, and J.-L. Wang (2005). "Functional Data Analysis for Sparse Longitudinal Data". *Journal of the American Statistical Association*, 100.470, 577–590.