# Closing the Gap between State-Space and Score-Driven Models: an Application to Modeling Implied Volatility Surface Dynamics

Xia Zou[1], Yicong Lin[1], and Andre Lucas[1]

[1]Vrije Universiteit Amsterdam and Tinbergen Institute

August 27, 2024

**Abstract**

We compare state-space and score-driven models for option implied volatility surface dynamics. Point forecasts of both models behave similarly, but density forecasts of plain-vanilla score-driven models are substantially worse. We show how a simple adjustment of the measurement density of the score-driven model can put both models back on an equal footing. The score-driven models can subsequently be further enhanced with non-Gaussian features without complicating parameter estimation in any way to better accommodate the data. We illustrate our findings using a dataset on S&P500 index options implied volatility surfaces.

*Keywords:* implied volatility surface dynamics; score-driven; state-space; dynamic factor model.

# 1 Introduction

Implied Volatilities (IV) based on the Black and Scholes (1973) option pricing model can be computed for every option maturity and strike price. Together, these IVs constitute the so-called implied volatility *surface*, which has important applications in pricing, hedging, forecasting, and risk management (see for instance Jorion, 1995). IV surfaces are often modeled using common factors, such that the dynamics of the entire surface are captured by a limited set of shared dynamic components. A typical approach for this builds on a standard state-space framework (see, e.g., Bedendo and Hodges, 2009; Koopman et al., 2010; Doz et al., 2012; Jungbacker et al., 2014; Wel et al., 2016; Wang et al., 2017). A natural alternative to the state-space approach would be a score-driven framework using the methodology proposed in Creal et al. (2013) and Harvey (2013). Score-driven factor models have for instance also been used for modeling term-structure dynamics (see, e.g., Creal et al., 2013; Quaedvlieg and Schotman, 2022; Koopman et al., 2017). Although an IV surface, unlike a term-structure, has two dimensions rather than one, the modeling principle remains the same. The advantage of score-driven models over their state-space counterparts in this context is that they are observation-driven, as classified by Cox (1981), and have an explicit expression for the likelihood function. This facilitates estimation and inference using standard maximum likelihood (ML) methods, even when accounting for non-Gaussian error processes. In contrast, state-space models that deviate from a linear Gaussian case quickly become more challenging to estimate, often requiring approximate estimation techniques such as the extended Kalman Filter, Bayesian methods, and simulated ML.

Despite their relative simplicity from a computational perspective, score-driven models perform remarkably well in terms of point forecast quality, even if the true data generating process is of state-space form. Koopman et al. (2016) compare a range of time-varying parameter models (volatility, duration, intensity, counts) for univariate time-series and show that point forecasts based on simple score-driven models perform similarly to those based on non-Gaussian state-space models estimated by more complex machinery. The paper is silent, however, about the quality of the density forecasts. Results in Koopman et al. (2017) suggest that, from a density forecasting perspective, score-driven models might underperform compared to their state-space counterparts. In particular, the typical assumption of an exact factor structure, where all contemporaneous correlations are captured by a single common factor, appears too regid for score-driven models. Koopman et al. (2017) solve this by endowing the error terms in the factor model to have an equicorrelation structure. The origins of the difference between the two model classes in terms of point versus density forecasts, however, remain underexplored.

This paper provides two main contributions. First, in a two-dimensional setting of forecasting IV surfaces, we show that score-driven models perform worse in terms of density forecasts than a standard linear Gaussian state-space model. We can also pinpoint how this performance gap can be attributed to an overly restrictive assumption on the covariance structure of the measurement noise in the score-driven model. Second, we show how a simple adaptation of the measurement equation of the score-driven model may bring its density forecast performance much more in line with that of a state-space modeling framework. The key is to match the covariance structure of the measurement noise more closely with that of the predictive density of the state-space model. Indeed, when implementing this adjustment, the state-space and score-driven approaches perform almost at par, not only in terms of point forecasts as in Koopman et al. (2016), but also in terms of density forecasts.

Once the density forecast performance of the score-driven and state-space approach is put on an equal footing in the Gaussian case, we can easily extend the score-driven model with non-Gaussian features without complicating the maximum likelihood estimation and inference procedures. We find that incorporating such non-Gaussian features indeed increases the density forecast quality of the score-driven model beyond that achieved by the linear Gaussian state-space model. While the latter could benefit from adding non-Gaussian features, it would entail a more challenging estimation procedure.

It is widely recognized that integrating non-Gaussian features into the model leads to a more robust filtering procedure for time-varying parameter paths (see Creal et al., 2013; Harvey and Luati, 2014; D'Innocenzo et al., 2023; Gasperoni et al., 2023). We show that for more models like the two-dimensional time-varying IV surfaces in this paper, adding non-Gaussian features to the model may also help to unmask persistent model mis-specification in specific data segments. This can lead to seemingly accurate density forecasts, but poor point forecast accuracy and correlated forecast errors. This finding can lead to further improvements in the model, such as adjusting the number of factors or refining the specification of factor loadings.

We study the dynamics of IV surfaces of S&P500 index options using data from January 2010 to December 2022. The factor model follows Goncalves and Guidolin (2006), and we include five factors: the overall level of the IV surface, moneyness slope and curvature, time-to-maturity slope, and the interaction of moneyness and time-to-maturity. We find that a linear Gaussian state-space model outperforms a plain-vanilla score-driven model by a large margin, both in terms of density fit and Value-at-Risk (VaR) violation rates. However, when we incorporate the adjusted covariance structure for the measurement errors into the score-driven model, as proposed in this paper, the score-driven model performs comparably to the state-space model. Adding Student's $t$ error terms to the model even increases the density fit of the score-driven model beyond that of its state-space counterpart. This allows us to obtain a clearer signal than in the Gaussian setting that the initial model specification needs further enhancement to better capture some of the edges and corners of the volatility surface dynamics (see Appendix A). We show that such changes to the model indeed further improve the in-sample and out-of-sample fit.

The remainder of this paper is structured as follows. Section 2 presents the different modeling frameworks and discusses how to close the gap in density forecasting performance between score-driven and state-space models. Section 3 describes the data. Section 4 provides the empirical results. Section 5 concludes. Additional empirical results and the derivation of the Student's $t$ information matrix for score-driven models are available in the appendix.

# 2 The modeling frameworks

We begin by modeling IV using the standard frameworks of state-space and score-driven models in Section 2.1. In Section 2.2, we examine the disparity between these approaches and propose a solution to reconcile them.

## 2.1 Standard state-space and score-driven models

We model log implied volatilities $\boldsymbol{IV}_t \in \mathbb{R}^{N_t}$ for $t = 1, \ldots, T$, over a possibly time-varying grid of moneyness values $\mathbb{M}_t \subset \mathbb{R}^{\kappa_t}$ and times-to-maturity $\mathbb{T}_t \subset \mathbb{R}^{\tau_t}_{\geq 0}$, where the IVs may not be observed at each grid point at each time. The total number of IVs observed at time $t$ is given by $N_t \leq \kappa_t \cdot \tau_t$. This set-up accommodates a time-varying number of option contracts and allows for changes in the type of option contracts over time, in line with typical options data characteristics. For instance, because the expiry date of an option contract is fixed, its time-to-maturity decreases as time progresses. Upon expiry, the option contract is completely removed from the dataset.

We assume that $\log \boldsymbol{IV}_t$ evolves as follows:

$$\log \boldsymbol{IV}_t = \boldsymbol{M}_t \boldsymbol{\beta}_t + \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t \sim h\left(\boldsymbol{\varepsilon}_t \mid \boldsymbol{H}_t; \boldsymbol{\vartheta}_\varepsilon\right), \tag{1}$$

$$\boldsymbol{\beta}_{t+1} = \left(\boldsymbol{I}_p - \boldsymbol{B}\right) \boldsymbol{\vartheta}_\beta + \boldsymbol{B}\, \boldsymbol{\beta}_t + \boldsymbol{\xi}_t. \tag{2}$$

The measurement equation in (1) consists of the matrix of factor loadings $\boldsymbol{M}_t \in \mathbb{R}^{N_t \times p}$, $p \in \mathbb{Z}^+$, a vector of factors $\boldsymbol{\beta}_t \in \mathbb{R}^{p \times 1}$, and an independent innovation term $\boldsymbol{\varepsilon}_t$ with distribution $h(\cdot \mid \boldsymbol{H}_t; \boldsymbol{\vartheta}_\varepsilon)$, where $h$ denotes the distribution with mean zero, covariance matrix $\boldsymbol{H}_t$, and shape parameter vector $\boldsymbol{\vartheta}_\varepsilon$. The state transition equation in (2) has an intercept vector $(\boldsymbol{I}_p - \boldsymbol{B})\boldsymbol{\vartheta}_\beta \in \mathbb{R}^p$ where $\boldsymbol{\vartheta}_\beta$ denotes the unconditional mean of $\boldsymbol{\beta}_t$, autoregressive matrix $\boldsymbol{B} \in \mathbb{R}^{p \times p}$ with $\|\boldsymbol{B}\|_2 < 1$, and 'state increment' vector $\boldsymbol{\xi}_t \in \mathbb{R}^p$. Here, $\boldsymbol{I}_p$ denotes an identity matrix of size $p$, and $\|\cdot\|_2$ represents the spectral norm. We gather all static parameters of the model, such as $\boldsymbol{\vartheta}_\varepsilon$, $\boldsymbol{\vartheta}_\beta$, $\boldsymbol{B}$, as well as any parameters describing the matrices $\boldsymbol{M}_t$ and $\boldsymbol{H}_t$, or defining the shape of the distribution or the specification of $\boldsymbol{\xi}_t$, into a vector $\boldsymbol{\psi}$ that requires estimation.

This set-up unifies both state-space and score-driven frameworks, depending on our choice of $\boldsymbol{\xi}_t$. For instance, if $\left\{(\boldsymbol{\varepsilon}_t^\top, \boldsymbol{\xi}_t^\top)^\top\right\}$ is an independently and identially distributed (iid) sequence of innovations with mutually independent components $\boldsymbol{\varepsilon}_t$ and $\boldsymbol{\xi}_t$, then Eqs. (1)–(2) collapse to a standard linear state-space set-up (see Durbin and Koopman, 2012). Conversely, if $\boldsymbol{\xi}_t$ is a measurable function that depends solely on $\boldsymbol{\beta}_t$ and $\boldsymbol{IV}_t$, the model becomes observation-driven. If, furthermore, $\boldsymbol{\xi}_t$ is chosen as the derivative (with respect to $\boldsymbol{\beta}_t$) of the log predictive density of $\boldsymbol{IV}_t$ given $\boldsymbol{\beta}_t$, we recover the score-driven framework of Creal et al. (2013).

Eq. (1) does not yet fully specify the distribution of the error term $\boldsymbol{\varepsilon}_t$, other than its mean and covariance matrix. For instance, if $(\boldsymbol{\varepsilon}_t^\top, \boldsymbol{\xi}_t^\top)^\top$ is normally distributed, we obtain the linear Gaussian state-space model as used in for instance Goncalves and Guidolin (2006) for IV surfaces. For such a state-space specification, we can then estimate the static parameter vector $\boldsymbol{\psi}$ by maximizing $\mathcal{L}(\boldsymbol{\psi})$, given by

$$\mathcal{L}(\boldsymbol{\psi}) = -\frac{1}{2}\sum_{t=1}^{T}\left(\log|2\pi \boldsymbol{F}_t| + \boldsymbol{v}_t^\top \boldsymbol{F}_t^{-1}\boldsymbol{v}_t\right), \qquad \boldsymbol{v}_t = \log \boldsymbol{IV}_t - \log \boldsymbol{IV}_{t|t-1}, \tag{3}$$

where the prediction errors $\boldsymbol{v}_t$ and their conditional covariance matrix $\boldsymbol{F}_t$ follow directly from the Kalman filter. For a non-Gaussian $\boldsymbol{\varepsilon}_t$, the standard Kalman filter recursions break down, or more precisely, only provide minimum mean-squared error forecasts of the states. Other estimation techniques such as simulated maximum likelihood based on importance sampling or particle filtering can be used in such non-Gaussian and/or non-linear (see, e.g., Durbin and Koopman, 2012, for an overview). Such techniques are typically more challenging and computationally intensive.

In a score-driven framework, the parameter vector $\boldsymbol{\psi}$ can be estimated by standard maximum likelihood (ML) techniques, whether $\boldsymbol{\varepsilon}_t$ is normally distributed or not. In an observation-driven framework, $\boldsymbol{\xi}_t$ is predetermined such that the likelihood is known in analytic form through a standard prediction error decomposition. That is, $\boldsymbol{\xi}_t$ is $\mathcal{F}_{t-1}$-measurable, where $\mathcal{F}_t = \sigma\big(\boldsymbol{IV}_s, s \le t\big)$, containing information up to time $t$. For convenience, let $\mathbb{E}_t(\cdot) = \mathbb{E}(\cdot \mid \mathcal{F}_t)$. To illustrate, consider a normal distribution with covariance matrix $\boldsymbol{H}_t$ for the density $h(\,\cdot \,|\, \boldsymbol{H}_t; \boldsymbol{\vartheta}_\varepsilon)$. Given the conditional normality of $\boldsymbol{\varepsilon}_t$, there is no additional shape parameter $\boldsymbol{\vartheta}_\varepsilon$. Defining the scaled score as $\boldsymbol{s}_t = \mathbb{E}_{t-1}\big[\big(\nabla_t\nabla_t^\top\big)^{-1}\nabla_t\big]$ **YC says: "Should it be $\boldsymbol{s}_t = \Big[\mathbb{E}_{t-1}\big(\nabla_t\nabla_t^\top\big)\Big]^{-1}\nabla_t$ ??"** with $\nabla_t = \partial \log h(\boldsymbol{IV}_t \,|\, \boldsymbol{\beta}_t; \boldsymbol{H}_t, \boldsymbol{\vartheta}_\varepsilon)/\partial\boldsymbol{\beta}$, and letting $\boldsymbol{\xi}_t = \boldsymbol{A}\boldsymbol{s}_t$ for a parameter matrix $\boldsymbol{A} \in \mathbb{R}^{p\times p}$, we obtain

$$\boldsymbol{\xi}_t = \boldsymbol{A}\,\mathbb{E}_{t-1}\Big[\big(\nabla_t\nabla_t^\top\big)^{-1}\nabla_t\Big] = \boldsymbol{A}\,(\boldsymbol{M}_t^\top\boldsymbol{H}_t^{-1}\boldsymbol{M}_t)^{-1}\boldsymbol{M}_t^\top\boldsymbol{H}_t^{-1}\boldsymbol{\varepsilon}_t, \tag{4}$$

$$\mathcal{L}(\boldsymbol{\psi}) = -\frac{1}{2}\sum_{t=1}^{T}\left(\log|2\pi\boldsymbol{H}_t| + \boldsymbol{\varepsilon}_t^\top\boldsymbol{H}_t^{-1}\boldsymbol{\varepsilon}_t\right), \tag{5}$$

where $\boldsymbol{\varepsilon}_t = \log\boldsymbol{IV}_t - \boldsymbol{M}_t\boldsymbol{\beta}_t$, and where we used inverse information matrix scaling of the score as defined in Creal et al. (2013). The scaled-score step in Eq. (4) has an intuitive interpretation by adjusting the time-varying regression parameter $\boldsymbol{\beta}_t$ using a GLS improvement step. Moreover, when the errors follow a Student's $t$ distribution with a degree of freedom $\nu > 2$, the expressions change to

$$\boldsymbol{\xi}_t = \frac{1 + (N_t + 1)/\nu}{1 + \boldsymbol{\varepsilon}_t^\top \boldsymbol{H}_t^{-1} \boldsymbol{\varepsilon}_t/(\nu - 2)} \boldsymbol{A} \, (\boldsymbol{M}_t^\top \boldsymbol{H}_t^{-1} \boldsymbol{M}_t)^{-1} \boldsymbol{M}_t^\top \boldsymbol{H}_t^{-1} \boldsymbol{\varepsilon}_t, \tag{6}$$

$$\mathcal{L}(\boldsymbol{\psi}) = -\frac{1}{2} \sum_{t=1}^{T} \Bigg[ \log |(\nu - 2)\pi \boldsymbol{H}_t| \\ + (\nu + N_t) \log \left(1 + \frac{\boldsymbol{\varepsilon}_t^\top \boldsymbol{H}_t^{-1} \boldsymbol{\varepsilon}_t}{\nu - 2}\right) + 2 \log \Gamma \left(\frac{\nu}{2}\right) - 2 \log \Gamma \left(\frac{\nu + N_t}{2}\right) \Bigg]; \tag{7}$$

see Appendix B for a derivation of the scaled score in (6). **YC says: "From the appendix, I obtain $(N_t + 2)/\nu$ in Eq. (6)."** Note that as $\nu \to \infty$, Eqs. (6)–(7) collapse to (4)–(5). If $\nu < \infty$, the score in (6) downweighs the GLS step for large incidental outliers via the factor $\boldsymbol{\varepsilon}_t^\top \boldsymbol{H}_t^{-1} \boldsymbol{\varepsilon}_t$ in the denominator and thus mitigates their effect on the dynamics of the time-varying parameter $\boldsymbol{\beta}_t$; see also, for instance, Harvey and Luati (2014) and Gasperoni et al. (2023) for the robustness features of score-driven filters based on fat-tailed observations.

## 2.2 Adjusted covariance structures for score-driven models

So far, the state-space and score-driven models appear quite similar. The main difference lies in their choice of the state increment vector $\boldsymbol{\xi}_t$: it is random for the state-space set-up and pre-determined for the score-driven model. This distinction leads to similar behavior in point forecasts for both models (Koopman et al., 2016). However, in terms of density forecasts, the two models behave in a markedly different way, with the state-space specification generally performing better.

To understand this phenomenon, consider a simple specification of (1) with a diagonal error covariance matrix $\boldsymbol{H}_t$. In terms of point forecasts, since both models assume independent measurement errors $\boldsymbol{\varepsilon}_t$, they yield similar results when the estimates from the observation-driven ($\boldsymbol{\beta}_t^{\mathrm{sd}}$) and state-space models ($\boldsymbol{\beta}_{t|t-1}^{\mathrm{ss}}$) are comparably accurate, where $\boldsymbol{\beta}_{t|t-1}^{\mathrm{ss}} := \mathbb{E}[\boldsymbol{\beta}_t \mid \mathcal{F}_{t-1}]$. However, when it comes to density forecasts that use the full information set $\mathcal{F}_{t-1}$, the two models produce very different results. Note that Model (1) can be equivalently expressed as

$$\log \boldsymbol{IV}_t = \boldsymbol{M}_t \boldsymbol{\beta}_{t|t-1}^{\mathrm{ss}} + \boldsymbol{M}_t \left(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t|t-1}^{\mathrm{ss}}\right) + \boldsymbol{\varepsilon}_t. \tag{8}$$

Conditional on $\mathcal{F}_{t-1}$, the first component on the right side of the equation is fixed and does not contribute to the conditional variance. Therefore, for the state-space specification, we obtain

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left[\log \boldsymbol{IV}_t \mid \mathcal{F}_{t-1}\right] &= \boldsymbol{H}_t + \boldsymbol{M}_t \,\mathbb{V}\mathrm{ar}\left[\boldsymbol{\beta}_t \mid \mathcal{F}_{t-1}\right] \boldsymbol{M}_t^\top \\ &= \boldsymbol{H}_t + \boldsymbol{M}_t \boldsymbol{Q}_\xi^{\mathrm{ss}} \boldsymbol{M}_t^\top + \boldsymbol{M}_t \boldsymbol{B}\,\mathbb{V}\mathrm{ar}\left[\boldsymbol{\beta}_{t-1} \mid \mathcal{F}_{t-1}\right] \boldsymbol{B}^\top \boldsymbol{M}_t^\top, \qquad (9)\end{aligned}$$

where $\boldsymbol{Q}_\xi^{\mathrm{ss}}$ is the contemporaneous covariance matrix of the state-innovation in (2) given the state-space specification of the model. Even if $\boldsymbol{H}_t$ is diagonal, the resulting state-space predictive density clearly exhibits a non-diagonal covariance structure. On the other hand, the predictive density of the score-driven model yields a diagonal covariance $\boldsymbol{H}_t$, as the dynamics of $\boldsymbol{\beta}_t^{\mathrm{sd}}$ is pre-determined conditional on $\mathcal{F}_{t-1}$. Thus, even if the score-driven forecast $\boldsymbol{\beta}_t^{\mathrm{sd}}$ and the state-space forecast $\boldsymbol{\beta}_{t|t-1}^{\mathrm{ss}}$ are close, their forecasting densities are very different.

The non-diagonal covariance specification in (9) potentially provides a better fit to real data compared to a diagonal $\boldsymbol{H}_t$. To understand the intuition behind this, consider a simple one-factor set-up (i.e., $p = 1$) to illustrate the core of the issue. Assume that $\boldsymbol{M}_t$ consists of a single column of ones and that $\boldsymbol{B} = 1$, which models the IV surface using a single (random walk) level factor. Both models assume that for a given $\boldsymbol{\beta}_t$ the prediction errors around this level are uncorrelated. However, as discussed, while the state-space approach assumes that the future value of $\boldsymbol{\beta}_{t+1}$ cannot be known with certainty today and is therefore subject to a prediction error, the score-driven set-up excludes such prediction error by assuming $\boldsymbol{\beta}_{t+1}$ is pre-determined given $\mathcal{F}_t$. Accordingly, the score-driven set-up maintains a diagonal structure for the prediction errors in the measurement equation, while in the state-space framework, prediction errors are correlated due to the common prediction error in $\boldsymbol{\beta}_{t+1}$ given $\mathcal{F}_t$. Although a score-driven filter can still provide an accurate filtered or predicted value for $\boldsymbol{\beta}_{t+1}$, the assumption that $\boldsymbol{\beta}_{t+1}$ is pre-determined in the data generating process (DGP) is typically untenable in empirical situations.

The solution is straightforward. We can slightly adjust the covariance structure in the measurement equation of the score-driven model to better reflect the correlation structure of the prediction errors. We propose to replace the measurement equation of the score-driven factor model in (1) with

$$\log \boldsymbol{IV}_t = \boldsymbol{M}_t \boldsymbol{\beta}_t + \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t \sim h\left(\boldsymbol{\varepsilon}_t \mid \boldsymbol{H}_t + \boldsymbol{M}_t \boldsymbol{C} \boldsymbol{M}_t^{\top}; \boldsymbol{\vartheta}_{\varepsilon}\right), \tag{10}$$

where $\boldsymbol{C} \in \mathbb{R}^{p \times p}$ is an additional static parameter matrix to be estimated. With this new correlation structure for the score-driven modeling framework, the predictive densities of the score-driven and state-space approaches resemble each other much more. In particular, if the matrices in the state-space model are time-invariant, this captures the steady state predictive density. In such a setting, we expect the adjustment in (10) to largely close the gap in density fit between the score-driven and state-space model.

The suggested adjustment in (10) also explains the improvements in density fit obtained by Koopman et al. (2017) when modeling international term structures and imposing an equicorrelation structure. The level factor is the most important ingredient in their factor model. As explained before, if the DGP is of state-space with time-invariant parameter matrices, the steady state predictive density has an equicorrelation structure. Imposing this structure on the score-driven measurement equation therefore leads to a substantial improvement in density fit.

It is worth noting that the adjusted covariance structure in (10) does not hinge on a Gaussian distribution. The adjustment is equally applicable for fat-tailed or skewed density functions $h$. Therefore, the score-driven model can easily incorporate non-Gaussian features by adjusting the score dynamics accordingly. This is a major advantage over the state-space model for the estimation procedure, as it remains feasible using standard maximum likelihood methods. In contrast, including such non-Gaussian features in the state-space setting, as previously discussed, would be more cumbersome and usually necessitates a different estimation paradigm to approximate filtering techniques. We investigate such extensions to the score-driven model in the empirical application in Section 3.

# 3 Empirical data and model specification

## 3.1 Descriptives

Our dataset comprises European call options on the S&P 500 index and encompasses all call and put options traded on the Chicago Board Options Exchange (CBOE). The dataset, retrieved from OptionDX, spans the period from January 1, 2010, to January 1, 2022. It includes the daily closing price of the index, as well as the strike prices, expiration dates, option deltas ($\Delta$), and implied volatilities of each option contract.

We apply the filtering procedures of Barone-Adesi et al. (2008) and Wel et al. (2016) to clean the data. Initially, we restrict our analysis to out-of-the-money options, defined by a $\Delta$ less than 0.5 in absolute value, because out-of-the-money options typically have higher trading activity than their in-the-money counterparts. Moreover, focusing solely on out-of-the-money options is conceptually equivalent to studying only in-the-money options under the assumption of put-call parity. For example, a call option with a $\Delta$ of 0.1 should possess the same implied volatility as an out-of-the-money put with a $\Delta$ of -0.9. Next, we exclude observations with more than 360 days or less than 7 days to expiration, as these options are typically characterized by lower liquidity levels. Additionally, we discard options with implied volatilities greater than 0.7 or less than 0.05 to mitigate the effect of potential data errors. The final dataset comprises a total of 7,739,265 observations, with an average of 2,722 observations per day.

Table 1 provides some summary statistics. Following Wel et al. (2016), we divide the sample into 24 distinct groups based on time-to-maturity and moneyness. Specifically, the maturity component is partitioned into four groups with breakpoints at 45, 90, and 180 days-to-maturity, while moneyness is split into six groups with breakpoints at $\Delta$ values of -0.375, -0.125, 0, 0.125, and 0.375.

We classify options with $\Delta$ values ranging from -0.125 to 0 as deep out-of-the-money puts (DOTM puts). Options with $\Delta$ from -0.375 to -0.125 are classified as out-of-the-money puts (OTM puts), and options with $\Delta$ values between -0.5 and -0.375 are classified as at-the-money puts (ATM puts). Calls are classified as deep OTM, OTM, and ATM, using the same cutoffs, but with positive $\Delta$ values.

For each of these 24 groups, we present the time-series average and standard deviation of the implied volatility, days-to-maturity, moneyness, $\Delta$, and trading volumes (Trading Vol) for each bucket. The percentage of trade volumes represents the average daily number of contracts traded within a group relative to the total trading volume across all contracts.

Table 1 highlights some of the stylized facts about the implied volatility surface. First, the implied volatilities decrease as moneyness increases for each of the four maturity groups, a phenomenon commonly known as the volatility smile or smirk. Second, the implied volatilities increase as the time-to-maturity increases, known as the volatility term structure. Finally, we see that shorter-term or deeper out-of-the-money options have higher trading volumes than longer-term or at-the-money products.

## 3.2 Restricted factor representation

A common approach when modeling the IV surface is to express it as a function of moneyness ($m_{i,t}$) and time-to-maturity ($\tau_{i,t}$) for option contract $i = 1, \ldots, N_t$ at time $t = 1, \ldots, T$. Goncalves and Guidolin (2006) compare various parametric specifications as proposed by Dumas et al. (1998) and Pena et al. (1999). They conclude that a simple model, which linearly combines polynomial terms and interactions of moneyness and time-to-maturity, achieves a good fit to the S&P 500 IV surface. We adopt their set-up to illustrate the effect of using the adjusted covariance structure in score-driven factor framework for modeling the IV surface, specifying the following five-factor specification (Goncalves and Guidolin, 2006):

$$\log IV_{i,t} = \beta_{1,t} + \beta_{2,t} m_{i,t} + \beta_{3,t} m_{i,t}^2 + \beta_{4,t} \tau_{i,t} + \beta_{5,t} m_{i,t} \tau_{i,t} + \varepsilon_{i,t} =: \boldsymbol{m}_{i,t}^\top \boldsymbol{\beta}_t + \varepsilon_{i,t}, \qquad (11)$$

where $\boldsymbol{m}_{i,t} = \left(1, m_{i,t}, m_{i,t}^2, \tau_{i,t}, m_{i,t}\tau_{i,t}\right)^\top$ and $\boldsymbol{\beta}_t = (\beta_{1,t}, \ldots, \beta_{5,t})^\top$. Here, $\beta_{1,t}$ represents the time-varying level of the log implied volatility; $\beta_{2,t}$ and $\beta_{3,t}$ capture the slope and curvature of log implied volatilities in the moneyness dimension (i.e., the volatility smile), respectively; $\beta_{4,t}$ reflects the slope of log implied volatility in the time-to-maturity dimension (i.e., the implied volatility term structure); and $\beta_{5,t}$ captures the interaction between moneyness and time-to-maturity. The model can be expressed in the form (1), with $\boldsymbol{M}_t = \left(\boldsymbol{m}_{1,t}, \ldots, \boldsymbol{m}_{N_t,t}\right)^\top$. Richer factor structures can easily be specified by adding more terms to the right-hand side of Eq. (11). Alternatively, the factor loadings could be estimated rather than pre-specified as in the version of $\boldsymbol{M}_t$ in (11). However, this does not alter the results for our main focus in this paper, namely how to close the gap in density performance between the state-space and score-driven approach. We therefore primarily stick to the specification in (11), and investigate an additional factor with estimated factor loadings in the robustness analysis in Section 4.2.

Table 1: Summary Statistics

| | | 7 – 45 days | | 45 – 90 days | | 90 – 180 days | | 180 – 360 days | |
|---|---|---|---|---|---|---|---|---|---|
| | | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| DOTM put | IV | 0.32 | 0.13 | 0.33 | 0.11 | 0.35 | 0.11 | 0.35 | 0.10 |
| | DTM | 24.51 | 10.68 | 64.36 | 12.25 | 125.59 | 26.40 | 265.68 | 51.93 |
| | Moneyness | 0.84 | 0.09 | 0.76 | 0.12 | 0.68 | 0.14 | 0.57 | 0.15 |
| | $\Delta$ | -0.03 | 0.03 | -0.04 | 0.03 | -0.04 | 0.04 | -0.04 | 0.04 |
| | Trading Vol (%) | 22.83 | | 11.52 | | 7.74 | | 5.62 | |
| OTM put | IV | 0.20 | 0.09 | 0.21 | 0.08 | 0.23 | 0.08 | 0.24 | 0.07 |
| | DTM | 26.39 | 10.21 | 65.31 | 12.44 | 129.17 | 26.85 | 270.38 | 52.54 |
| | Moneyness | 0.96 | 0.02 | 0.94 | 0.03 | 0.91 | 0.04 | 0.87 | 0.06 |
| | $\Delta$ | -0.23 | 0.07 | -0.23 | 0.07 | -0.23 | 0.07 | -0.23 | 0.07 |
| | Trading Vol (%) | 6.42 | | 4.47 | | 4.26 | | 2.60 | |
| ATM put | IV | 0.18 | 0.10 | 0.17 | 0.07 | 0.19 | 0.07 | 0.19 | 0.05 |
| | DTM | 26.29 | 10.18 | 65.48 | 12.50 | 130.22 | 27.14 | 270.39 | 52.98 |
| | Moneyness | 0.99 | 0.01 | 0.99 | 0.01 | 0.98 | 0.01 | 0.98 | 0.02 |
| | $\Delta$ | -0.44 | 0.04 | -0.44 | 0.04 | -0.43 | 0.04 | -0.44 | 0.04 |
| | Trading Vol (%) | 1.81 | | 1.24 | | 1.27 | | 0.76 | |
| ATM call | IV | 0.17 | 0.10 | 0.16 | 0.07 | 0.18 | 0.06 | 0.18 | 0.04 |
| | DTM | 26.21 | 10.22 | 65.39 | 12.48 | 129.92 | 27.05 | 271.89 | 52.81 |
| | Moneyness | 1.01 | 0.01 | 1.01 | 0.01 | 1.02 | 0.01 | 1.03 | 0.02 |
| | $\Delta$ | 0.44 | 0.04 | 0.44 | 0.04 | 0.44 | 0.04 | 0.44 | 0.04 |
| | Trading Vol (%) | 1.59 | | 1.08 | | 1.09 | | 0.66 | |
| OTM call | IV | 0.15 | 0.09 | 0.14 | 0.06 | 0.15 | 0.05 | 0.15 | 0.04 |
| | DTM | 25.96 | 10.35 | 65.23 | 12.42 | 127.47 | 26.77 | 271.51 | 52.34 |
| | Moneyness | 1.03 | 0.02 | 1.04 | 0.02 | 1.06 | 0.03 | 1.09 | 0.04 |
| | $\Delta$ | 0.24 | 0.07 | 0.24 | 0.07 | 0.25 | 0.07 | 0.25 | 0.07 |
| | Trading Vol (%) | 3.33 | | 2.19 | | 1.90 | | 1.31 | |
| DOTM call | IV | 0.16 | 0.09 | 0.14 | 0.06 | 0.15 | 0.05 | 0.15 | 0.04 |
| | DTM | 23.89 | 10.41 | 64.01 | 12.30 | 125.64 | 26.34 | 266.00 | 52.26 |
| | Moneyness | 1.10 | 0.08 | 1.13 | 0.10 | 1.20 | 0.13 | 1.30 | 0.16 |
| | $\Delta$ | 0.03 | 0.03 | 0.04 | 0.04 | 0.04 | 0.04 | 0.04 | 0.04 |
| | Trading Vol (%) | 8.80 | | 3.50 | | 2.28 | | 1.74 | |

*Note*: This table presents summary statistics for the option data, including the mean and standard deviation (SD) over time for implied volatility, days to maturity (DTM), moneyness (the strike price divided by the index), option $\Delta$, and trading frequency across four maturity groups and six moneyness groups. The maturity groups are 7-45, 45-90, 90-180, and 180-360 days. The six moneyness groups are defined as deep out-of-the-money put ($-0.125 < \Delta < 0$, DOTM put), out-of-the-money put ($-0.375 < \Delta < -0.125$, OTM put), at-the-money put ($-0.5 < \Delta < -0.375$, ATM put), and similarly for call options (with positive $\Delta$s). Each day, we identify all contracts that fall within each maturity-moneyness group, and the numbers represent averages over time and across contracts for each group.

# 4 Estimation results

We present the main empirical results in this section. All of our analyses are done out-of-sample. We use a rolling window of 500 observations (about 2 years) to forecast the next 250 observations (1 year). This gives us $T^\star = 2,588$ out-of-sample observations from January 1, 2012, to January 1, 2022. We focus on the one-step-ahead forecasts of the log implied volatilities. The benchmark results are presented in Section 4.1, followed by several robustness checks in Section 4.2.

In the benchmark analysis (Section 4.1), we compare the state-space (SS) model with four different score-driven (SD) models. The SD models employ either a normal or Student's $t$ specification, as described in Eqs. (4)–(5) and (6)–(7), respectively. Moreover, for the SD models, We consider versions both with and without the covariance adjustment for the measurement equation as proposed in Eq. (10). In this first analysis, we use the non-bucketed option dataset. Therefore, the number of option contracts, and thus the dimension $N_t$ of $\log \boldsymbol{IV}_t$, changes over time.

We evaluate the performance of the different models in both statistical and economic terms. First, for the statistical measures, we compute log-likelihoods, AIC criteria, mean squared error (MSE), and mean absolute error (MAE) criteria. These are defined as $\text{MSE} = (N_t T^\star)^{-1} \sum_{t=1}^{T^\star} \sum_{i=1}^{N_t} (IV_{i,t} - IV_{i,t|t-1})^2$ and $\text{MAE} = (N_t T^\star)^{-1} \sum_{t=1}^{T^\star} \sum_{i=1}^{N_t} \left| IV_{i,t} - IV_{i,t|t-1} \right|$, respectively, where $\log IV_{i,t|t-1}$ the one-step-ahead forecast for $\log IV_t$. Second, for the economic evaluation, we conduct an out-of-sample Value-at-Risk (VaR) analysis at a 99% confidence level ($1 - \alpha = 99\%$). We concentrate on a setting where the correlation structure is crucial for forecast performance. In particular, we consider the unweighted overall cross-sectional average of the log IVs, $P_t = N_t^{-1} \sum_{i=1}^{N_t} \log IV_{i,t}$, and consider the one-step-ahead risk quantiles of $P_t$. The risk quantiles or Value-at-Risk for the score-driven specifications are straightforward to compute due to their observation-driven nature. For a $(1 - \alpha)$ confidence level, they are given by

$$\widehat{VaR}_{t+1} = P_{t|t-1} + \frac{Q(\alpha)}{N_t} \sqrt{\boldsymbol{\imath}_{N_t}^\top \hat{\boldsymbol{F}}_t \boldsymbol{\imath}_{N_t}}, \tag{12}$$

where $Q(\alpha)$ is the $\alpha$-quantile of the normal or unit-variance Student's $t$ distribution, $P_{t|t-1} = N_t^{-1} \sum_{i=1}^{N_t} \log IV_{i,t|t-1}$, and $\hat{\boldsymbol{F}}_t = \boldsymbol{H}_t$ for the standard score-driven model, and $\hat{\boldsymbol{F}}_t = \boldsymbol{H}_t + \boldsymbol{M}_t \boldsymbol{C} \boldsymbol{M}_t^\top$ for the adjusted model. For the state-space specifications, the predictions $P_{t|t-1}$ and forecast error variances $\hat{\boldsymbol{F}}_t = \boldsymbol{F}_t$ follow directly from the Kalman Filter recursions.

## 4.1 The benchmark analysis

Table 2 presents the out-of-sample MSE, MAE, log-likelihood (labelled as "loglik"), and AIC, for both the state-space and score-driven models used to model the log implied volatilities, as specified in Eqs. (1)–(2). For MSE and MAE, the ratios of all models to the corresponding measures of the state-space model are provided. It also includes the metrics for the static factor model, which assumes that the factors described in Eq. (1) remain constant over time (i.e., $\boldsymbol{\beta}_t \equiv \boldsymbol{\beta}$). Results are shown for the entire sample period and for two sub-sample periods: the pre-COVID period (2012-2020) and the COVID period (2020-2022). The latter period, marked by the COVID-19 pandemic, exhibits significantly higher volatility compared to the former. To quantify the statistical significance of performance differences, we use the Diebold-Mariano (DM) test, with the state-space models serving as benchmarks (Diebold and Mariano, 2002).

Table 2 highlights three main findings. First, the log-likelihood values indicate that the linear Gaussian SS model significantly outperforms the SD model with the same distribution assumption and without covariance adjustment for the measurement equation. However, when the covariance adjustment is applied to the SD model, its log-likelihood closely aligns with that of the linear Gaussian SS model This pattern is also reflected in the AIC results.

Second, the relative ratios of MSE and MAE suggest that the two approaches, whether with or without covariance adjustment, perform similarly in terms of point forecasts. This supports the findings of Koopman et al. (2016). For both the full sample and the pre-COVID period, the SD model under a Gaussian distribution without covariance adjustment slightly outperforms the SS model, achieving lower MSE and MAE. Conversely, the SD model with covariance adjustment shows slightly higher values for these metrics compared to the SS model. In the COVID period, however, the pattern reverses: the SD model with covariance adjustment slightly outperforms the SS model, while the model without covariance adjustment exhibits slightly higher values compared to the SS framework.

Third, the results for SD models utilizing a Student's $t$ distribution indicate that

Table 2: Out-of-sample performance using *non-bucketed* data, evaluated by MSE, MAE, log-likelihood (loglik), AIC, and BIC, for both state-space (SS) and score-driven (SD) models for the log implied volatility model from Eqs. (1)–(2). The rows labeled “Static” assume $\boldsymbol{\beta}_t \equiv \boldsymbol{\beta}$.

| Model | Distr. | Adj. | MSE | MAE | loglik | AIC | #Par. |
|---|---|---|---|---|---|---|---|
| | | | | | $\times 10^{-3}$ | $\times 10^{-3}$ | |
| Full sample (2012-2022) | | | | | | | |
| SS | $\mathcal{N}$ | — | 1.00 | 1.00 | 1557.45 | -3114.86 | 16 |
| SD | $\mathcal{N}$ | — | 0.99 | $0.99^{***}$ | $1203.19^{***}$ | -2406.35 | 16 |
| SD | $\mathcal{N}$ | yes | $1.02^{**}$ | $1.02^{***}$ | $1555.56^{***}$ | -3111.08 | 21 |
| SD | $t$ | — | 1.01 | 1.00 | 1551.76 | -3103.49 | 17 |
| SD | $t$ | yes | $1.05^{***}$ | $1.04^{***}$ | $1926.21^{***}$ | -3852.38 | 22 |
| Static | $\mathcal{N}$ | — | $2.12^{***}$ | $1.61^{***}$ | | | |
| Pre-COVID period (2012-2020) | | | | | | | |
| SS | $\mathcal{N}$ | — | 1.00 | 1.00 | 1095.87 | -2191.72 | 16 |
| SD | $\mathcal{N}$ | — | 0.98 | 0.99 | $885.25^{***}$ | -1770.47 | 16 |
| SD | $\mathcal{N}$ | yes | $1.04^{***}$ | $1.04^{***}$ | $1094.19^{***}$ | -2188.33 | 21 |
| SD | $t$ | — | 0.99 | 1.00 | 1144.60 | -2289.17 | 17 |
| SD | $t$ | yes | $1.07^{***}$ | $1.06^{***}$ | $1354.35^{***}$ | -2708.65 | 22 |
| Static | $\mathcal{N}$ | — | $1.55^{***}$ | $1.44^{***}$ | | | |
| COVID period (2020-2022) | | | | | | | |
| SS | $\mathcal{N}$ | — | 1.00 | 1.00 | 461.57 | -923.11 | 16 |
| SD | $\mathcal{N}$ | — | $1.02^{**}$ | $0.98^{***}$ | $317.94^{***}$ | -635.85 | 16 |
| SD | $\mathcal{N}$ | yes | $0.97^{**}$ | $0.96^{***}$ | $461.37^{***}$ | -922.70 | 21 |
| SD | $t$ | — | 1.06 | 1.01 | 407.16 | -814.29 | 17 |
| SD | $t$ | yes | $0.98^{*}$ | $0.98^{***}$ | $571.87^{***}$ | -1143.69 | 22 |
| Static | $\mathcal{N}$ | — | $3.84^{***}$ | $2.26^{***}$ | | | |

*Note*: The distribution (Distr.) can be normal ($\mathcal{N}$) or Student’s $t$, and the covariance structure can be diagonal or adjusted as in Eq. (10), as indicated by the column Adj. The last column specifies the number of parameters in each model. The log implied volatilities are forecasted for each option contract. For the DM test (with the state-space models as benchmarks), $^{***}$, $^{**}$, and $^{*}$ denote significance at the 1%, 5%, and 10% level, respectively.

incorporating non-Gaussian features enhances the density forecast performance beyond that of the linear Gaussian SS models. Specifically, the log-likelihood and AIC of the SD models with a Student’s $t$ distribution surpass those of the SS models after covariance adjustment. Even without covariance adjustment, the SD model with a Student’s $t$ distribution achieves log-likelihood and AIC values comparable to those of the SS models.

Table 3: Out-of-sample log-likelihood and 99% Value at Risk backtesting outcomes using *non-bucketed* data, including violation ratio (Viol ratio) and tick loss, for both state-space (SS) and score-driven (SD) models applied to the log implied volatility model from Eqs. (1)–(2). See the note in Table 2 for additional details.

| Model | Distr. | Adj. | loglik $\times 10^{-3}$ | Viol ratio $\times 10^{3}$ | Tick loss $\times 10^{3}$ |
|---|---|---|---|---|---|
| | | | Full sample (2012-2022) | | |
| SS | $\mathcal{N}$ | — | 1557.45 | 0.04 | 3.68 |
| SD | $\mathcal{N}$ | — | 1203.19*** | 41.07 | 9.99*** |
| SD | $\mathcal{N}$ | yes | 1555.56*** | 0.04 | 5.16*** |
| SD | $t$ | — | 1551.76 | 50.88 | 14.60*** |
| SD | $t$ | yes | 1926.21*** | 0.00 | 6.14*** |
| | | | Pre-COVID (2012-2020) | | |
| SS | $\mathcal{N}$ | — | 1095.87 | 0.05 | 3.16 |
| SD | $\mathcal{N}$ | — | 885.25*** | 40.43 | 9.84*** |
| SD | $\mathcal{N}$ | yes | 1094.19*** | 0.05 | 4.74*** |
| SD | $t$ | — | 1144.60 | 52.65 | 15.56*** |
| SD | $t$ | yes | 1354.35*** | 0.00 | 5.45*** |
| | | | COVID period (2020-2022) | | |
| SS | $\mathcal{N}$ | — | 461.57 | 0.00 | 6.43 |
| SD | $\mathcal{N}$ | — | 317.94*** | 44.47 | 10.76*** |
| SD | $\mathcal{N}$ | yes | 461.37*** | 0.00 | 7.40*** |
| SD | $t$ | — | 407.16 | 41.51 | 9.49*** |
| SD | $t$ | yes | 571.87*** | 0.00 | 9.79*** |

This observation holds true for both sub-sample periods.

In Table 3, we present the results for the 99%-Value-at-Risk (VaR). In terms of violation rates, our findings show that SD models without covariance adjustment, under both Gaussian and Student's $t$ distributions, exhibit a significantly higher violation ratio compared to the SS models. In contrast, the violation ratios for SD models with covariance adjustment are much closer to those of the SS models, and all models, including the SS model and the SD models with covariance adjustment, maintain conservative violation rates relative to the nominal level of 1%. Furthermore, both SD models without covariance adjustment demonstrate substantially higher tick loss compared to those with covariance adjustment, a pattern consistent across the two sub-sample periods. These results underscore the importance of covariance adjustment in the SD model, as it is crucial for narrowing the performance gap between the SS and SD frameworks in terms of density forecasts and VaR violation rates.

## 4.2 Robustness checks with bucketed data

To verify the robustness of our previous findings, we also apply our analysis to bucketed data. Following Wel et al. (2016), Bollen and Whaley (2004), and Barone-Adesi et al. (2008), we divide the data into four maturity groups, separated by maturities of 45, 90, and 180 days, and six moneyness groups, separated by $\Delta$ of -0.375, -0.125, 0, 0.125, and 0.375, as shown in Table 1. For each maturity-moneyness group, we select the contract closest to the mid-point. Stacking the IV vector for different groups leads to a fixed-dimensional vector of observations with 24 dimensions..

Tables 4 - 5 reproduce Tables 2 -3 using the *bucketed* data. As before, Table 4 reveals that the plain-vanilla SD models still have a significantly lower log-likelihood than the SS models. However, the log-likelihood of the SD models with covariance adjustment is significantly higher than that of the SS models for both the whole sample period and the first sub-sample period, providing even stronger evidence than what is shown in Table 2. For the second sub-sample period during COVID, there is no statistically significant difference. Moreover, in terms of point forecasts, the SD models, whether with or without covariance adjustment, generally perform better than their SS counterparts. Specifically, for MSE, both SD models under the Gaussian distribution significantly outperform the SS models for the entire sample period and the pre-COVID sub-sample. However, only the SD model with covariance adjustment shows significant improvement over the SS models during the volatile COVID period. Regarding MAE, the plain-vanilla SD model performs significantly better than the SS models in both sub-samples. Additionally, the SD model with Gaussian distribution and covariance adjustment matches the performance of the SS models for the entire sample and the pre-COVID period, but it significantly outperforms the SS models during the COVID period.

Table 5 presents the 99% VaR estimation results with the bucketed data. The violation rates align with our main findings in Section 4. Additionally, the tick loss results indicate that the SD models with covariance adjustment, regardless of whether they use Gaussian or Student's $t$ distributions, produce significantly lower values compared to the SS models.

In Appendix A, we provide additional robustness checks using an alternative factor

Table 4: Out-of-sample performance using *bucketed* data, evaluated by MSE, MAE, log-likelihood (loglik), AIC, and BIC, for both state-space (SS) and score-driven (SD) models for the log implied volatility model from Eqs. (1)–(2). The rows labeled "Static" assume $\boldsymbol{\beta}_t \equiv \boldsymbol{\beta}$. See the note in Table 2 for additional details.

| Model | Distr. | Adj. | MSE | MAE | loglik $\times 10^{-3}$ | AIC $\times 10^{-3}$ | #Par. |
|---|---|---|---|---|---|---|---|
| | | | | Full sample (2012-2022) | | | |
| SS | $\mathcal{N}$ | — | 1.00 | 1.00 | 62.11 | -124.18 | 16 |
| SD | $\mathcal{N}$ | — | 0.93** | 0.97*** | 54.57*** | -109.11 | 16 |
| SD | $\mathcal{N}$ | yes | 0.92* | 0.99 | 63.71*** | -127.38 | 21 |
| SD | $t$ | — | 2.65 | 1.17 | 57.00*** | -113.96 | 17 |
| SD | $t$ | yes | 1.35 | 1.07** | 63.57*** | -127.10 | 22 |
| Static | $\mathcal{N}$ | — | 9.36*** | 3.41*** | | | |
| | | | | Pre-COVID (2012-2020) | | | |
| SS | $\mathcal{N}$ | — | 1.00 | 1.00 | 53.72 | -107.40 | 16 |
| SD | $\mathcal{N}$ | — | 0.94*** | 0.97*** | 48.25*** | -96.47 | 16 |
| SD | $\mathcal{N}$ | yes | 0.97** | 1.00 | 55.03*** | -110.02 | 21 |
| SD | $t$ | — | 1.16 | 1.04** | 50.82*** | -101.61 | 17 |
| SD | $t$ | yes | 1.20** | 1.07*** | 55.19*** | -110.34 | 22 |
| Static | $\mathcal{N}$ | — | 7.35*** | 3.19*** | | | |
| | | | | COVID period (2020-2022) | | | |
| SS | $\mathcal{N}$ | — | 1.00 | 1.00 | 8.39 | -16.75 | 16 |
| SD | $\mathcal{N}$ | — | 0.92 | 0.95** | 6.32 | -12.61 | 16 |
| SD | $\mathcal{N}$ | yes | 0.87** | 0.95** | 8.68 | -17.31 | 21 |
| SD | $t$ | — | 4.25 | 1.54 | 6.18* | -12.32 | 17 |
| SD | $t$ | yes | 1.51 | 1.08 | 8.38 | -16.72 | 22 |
| Static | $\mathcal{N}$ | — | 11.52 | 4.00 | | | |

representation, where dummy variables are employed to capture group-specific levels. These results are consistent with the main findings of the paper. Moreover, it shows that the current model specifications can be improved to better capture some of the edges and corners of the volatility surface dynamics.

## 5 Conclusions

In this paper, we investigated the substantial difference in density fit between state-space and score-driven factor models and proposed a solution to bridge the gap. In particular,

Table 5: Out-of-sample log-likelihood and 99% Value at Risk backtesting outcomes using *bucketed* data, including violation ratio (Viol ratio) and tick loss, for both state-space (SS) and score-driven (SD) models applied to the log implied volatility model from Eqs. (1)–(2) See the note in Table 2 for additional details.

| Model | Distr. | Adj. | loglik | Viol ratio | Tick loss |
|---|---|---|---|---|---|
| | | | $\times 10^{-3}$ | $\times 10^{3}$ | $\times 10^{3}$ |
| Full sample (2012-2022) | | | | | |
| SS | $\mathcal{N}$ | — | 62.11 | 0.17 | 1.83 |
| SD | $\mathcal{N}$ | — | 54.57*** | 17.34 | 4.92*** |
| SD | $\mathcal{N}$ | yes | 63.71*** | 0.51 | 1.70** |
| SD | $t$ | — | 57.00*** | 18.16 | 5.32*** |
| SD | $t$ | yes | 63.57*** | 0.47 | 1.60* |
| Pre-COVID (2012-2020) | | | | | |
| SS | $\mathcal{N}$ | — | 53.72 | 0.10 | 1.82 |
| SD | $\mathcal{N}$ | — | 48.25*** | 19.14 | 5.35*** |
| SD | $\mathcal{N}$ | yes | 55.03*** | 0.56 | 1.75 |
| SD | $t$ | — | 50.82*** | 19.86 | 5.78*** |
| SD | $t$ | yes | 55.19*** | 0.56 | 1.61 |
| COVID period (2020-2022) | | | | | |
| SS | $\mathcal{N}$ | — | 8.39 | 0.54 | 1.91 |
| SD | $\mathcal{N}$ | — | 6.32 | 7.82 | 2.65 |
| SD | $\mathcal{N}$ | yes | 8.68 | 0.27 | 1.47** |
| SD | $t$ | — | 6.18* | 9.16 | 2.84 |
| SD | $t$ | yes | 8.38 | 0.00 | 1.59*** |

we introduced a change in the covariance structure of the measurement equation to put the state-space and score-driven model more on an equal footing. This adjustment facilitates the use of standard estimation and inference methods for non-Gaussian distributions, avoiding the complex techniques typically required by the state-space framework.

We applied the approach to model the implied volatility surface for S&P500 index option data. We confirmed that a plain-vanilla score-driven model has a significantly lower density forecast performance than a state-space competitor. However, when the modified covariance structure is applied to the score-driven model, this performance gap disappears and can even reverse, especially when incorporating non-Gaussian features. The covariance matrix adjustment introduced in this paper can also be applied in high-dimensional cases due to its parsimonious nature, which is rooted in the underlying factor model structure. Our empirical results reveal that accounting for correlated and heteroskedastic innovations in the score-driven framework elevates its log-likelihood to match that of the state-space model. We demonstrate that this adjustment significantly affects the quality of Value-at-Risk estimates.

# References

Barone-Adesi, Giovanni, Robert F Engle, and Loriano Mancini (2008). “A GARCH option pricing model with filtered historical simulation”. *Review of Financial Studies* 21.3, pp. 1223–1258.

Bedendo, Mascia and Stewart D Hodges (2009). “The dynamics of the volatility skew: A Kalman filter approach”. *Journal of Banking & Finance* 33.6, pp. 1156–1165.

Black, Fischer and Myron Scholes (1973). “The pricing of options and corporate liabilities”. *Journal of Political Economy* 81.3, pp. 637–654.

Bollen, Nicolas PB and Robert E Whaley (2004). “Does net buying pressure affect the shape of implied volatility functions?” *Journal of Finance* 59.2, pp. 711–753.

Cox, David R. (1981). “Statistical analysis of time series: Some recent developments”. *Scandinavian Journal of Statistics* 8.2, pp. 93–115.

Creal, Drew, Siem Jan Koopman, and André Lucas (2013). “Generalized autoregressive score models with applications”. *Journal of Applied Econometrics* 28.5, pp. 777–795.

D’Innocenzo, Enzo, Alessandra Luati, and Mario Mazzocchi (2023). “A robust score-driven filter for multivariate time series”. *Econometric Reviews* 42.5, pp. 441–470.

Diebold, Francis X and Robert S Mariano (2002). “Comparing predictive accuracy”. *Journal of Business & Economic Statistics* 20.1, pp. 134–144.

Doz, Catherine, Domenico Giannone, and Lucrezia Reichlin (2012). “A quasi-maximum likelihood approach for large, approximate dynamic factor models”. *Review of Economics and Statistics* 94.4, pp. 1014–1024.

Dumas, Bernard, Jeff Fleming, and Robert E Whaley (1998). “Implied volatility functions: Empirical tests”. *Journal of Finance* 53.6, pp. 2059–2106.

Durbin, James and Siem Jan Koopman (2012). *Time series analysis by state space methods*. Vol. 38. OUP Oxford.

Gasperoni, Francesca, Alessandra Luati, Lucia Paci, and Enzo D'Innocenzo (2023). "Score-driven modeling of spatio-temporal data". *Journal of the American Statistical Association* 118.542, pp. 1066–1077.

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

Harvey, Andrew C (2013). *Dynamic models for volatility and heavy tails: with applications to financial and economic time series*. Vol. 52. Cambridge University Press.

Harvey, Andrew and Alessandra Luati (2014). "Filtering with heavy tails". *Journal of the American Statistical Association* 109.507, pp. 1112–1122.

Jorion, Philippe (1995). "Predicting volatility in the foreign exchange market". *Journal of Finance* 50.2, pp. 507–528.

Jungbacker, Borus, Siem Jan Koopman, and Michel Van Der Wel (2014). "Smooth dynamic factor analysis with application to the US term structure of interest rates". *Journal of Applied Econometrics* 29.1, pp. 65–90.

Koopman, Siem Jan, Andre Lucas, and Marcel Scharth (2016). "Predicting time-varying parameters with parameter-driven and observation-driven models". *Review of Economics and Statistics* 98.1, pp. 97–110.

Koopman, Siem Jan, André Lucas, and Marcin Zamojski (2017). "Dynamic term structure models with score-driven time-varying parameters: estimation and forecasting". *Narowdy Bank Polski, NBP Working Paper* No. 258.

Koopman, Siem Jan, Max IP Mallee, and Michel Van der Wel (2010). "Analyzing the term structure of interest rates using the dynamic Nelson–Siegel model with time-varying parameters". *Journal of Business & Economic Statistics* 28.3, pp. 329–343.

Pena, Ignacio, Gonzalo Rubio, and Gregorio Serna (1999). "Why do we smile? On the determinants of the implied volatility function". *Journal of Banking & Finance* 23.8, pp. 1151–1179.

Quaedvlieg, Rogier and Peter Schotman (2022). "Hedging long-term liabilities". *Journal of Financial Econometrics* 20.3, pp. 505–538.

Wang, Jinzhong, Shijiang Chen, Qizhi Tao, and Ting Zhang (2017). "Modelling the implied volatility surface based on Shanghai 50ETF options". *Economic Modelling* 64, pp. 295–301.

Wel, Michel van der, Sait R Ozturk, and Dick van Dijk (2016). “Dynamic factor models for the volatility surface”. *Dynamic Factor Models*. Vol. 35. Emerald Group Publishing Limited, pp. 127–174.

# A Further robustness check

We propose an alternative factor representation, referred to as the *four-factor representation*, by replacing the interaction term between moneyness and time-to-maturity, as well as the squared moneyness term introduced in Section 3.2, with some dummy variables. Specifically, we divide the data into 24 groups, separated by maturities of 45, 90, and 180 days and moneyness intervals of -0.375, -0.125, 0, 0.125, and 0.375, as done previously. For clarity, let $g_{i,t} = \pi(m_{i,t}, \tau_{i,t}) \in 1, \ldots, 24$ represent the group number corresponding to $(m_{i,t}, \tau_{i,t})$, where $\pi(\cdot\ \cdot)$ is a function that assigns the group number according to the specified rule. To enable each maturity-moneyness group to have its own level, we introduce the dummy variables as follows: For $i = 1, \ldots, N_t$, $t = 1, \ldots, T$,

$$\log IV_{i,t} = \beta_{1,t} + \beta_{2,t} m_{i,t} + \beta_{3,t} \tau_{i,t} + \sum_{g=1}^{24} \beta_{4,g_{i,t},t} \mathbb{1}\{g_{i,t} = g\} + \varepsilon_{i,t}, \tag{A.1}$$

where $\mathbb{1}\{\cdot\}$ is an indicator function. Now, take $\boldsymbol{m}_{i,t} = \left(1, m_{i,t}, \tau_{i,t}, 1_{1,i,t}, 1_{2,i,t}, \ldots, 1_{24,i,t}\right)^\top$ and $\boldsymbol{\beta}_t = \left(\beta_{1,t}, \beta_{2,t}, \beta_{3,t}, \beta_{4,1,t}, \beta_{4,2,t}, \ldots, \beta_{4,24,t}\right)^\top$ in Eqs. (1)–(2).

The corresponding results are provided in Tables A.1 - A.2. We observe that the four-factor representation improves the models presented in the main sections by better capturing the edges and corners of the surface of the bucketed data.

Table A.1: Out-of-sample performance using *non-bucketed* data and with a *four-factor representation*, evaluated by MSE, MAE, log-likelihood (loglik), AIC, and BIC, for both state-space (SS) and score-driven (SD) models for the log implied volatility model from Eqs. (1)–(2). The rows labeled "Static" assume $\boldsymbol{\beta}_t \equiv \boldsymbol{\beta}$.

| Model | Distr. | Adj. | MSE | MAE | loglik | AIC | #Par. |
|---|---|---|---|---|---|---|---|
| | | | | | $\times 10^{-3}$ | $\times 10^{-3}$ | |
| | | | Full sample (2012-2022) | | | | |
| SS | $\mathcal{N}$ | — | 1.00 | 1.00 | 2529.16 | -5058.25 | 37 |
| SD | $\mathcal{N}$ | — | 1.00 | 1.00 | 2087.72$^{***}$ | -4175.37 | 37 |
| SD | $\mathcal{N}$ | yes | 1.00 | 1.01$^{***}$ | 2526.46$^{***}$ | -5052.83 | 41 |
| SD | $t$ | — | 1.05$^{***}$ | 1.02$^{**}$ | 2336.27$^{**}$ | -4672.46 | 38 |
| SD | $t$ | yes | 1.13$^{**}$ | 1.06$^{***}$ | 2817.99$^{***}$ | -5635.89 | 42 |
| Static | $\mathcal{N}$ | — | 2.98$^{***}$ | 1.91$^{***}$ | | | |
| | | | Pre-COVID (2012-2020) | | | | |
| SS | $\mathcal{N}$ | — | 1.00 | 1.00 | 1930.77 | -3861.47 | 37 |
| SD | $\mathcal{N}$ | — | 1.00 | 1.00$^{**}$ | 1649.28$^{***}$ | -3298.50 | 37 |
| SD | $\mathcal{N}$ | yes | 1.01$^{***}$ | 1.02$^{***}$ | 1928.41$^{***}$ | -3856.73 | 41 |
| SD | $t$ | — | 1.02 | 1.01$^{***}$ | 1833.79$^{**}$ | -3667.51 | 38 |
| SD | $\mathcal{N}$ | yes | 1.11$^{***}$ | 1.07$^{***}$ | 2132.30$^{***}$ | -4264.51 | 42 |
| Static | $\mathcal{N}$ | — | 2.01$^{***}$ | 1.68$^{***}$ | | | |
| | | | COVID sample (2020-2022) | | | | |
| SS | $\mathcal{N}$ | — | 1.00 | 1.00 | 598.39 | -1196.70 | 37 |
| SD | $\mathcal{N}$ | — | 1.00 | 1.00 | 438.44$^{**}$ | -876.80 | 37 |
| SD | $\mathcal{N}$ | yes | 0.97$^{***}$ | 0.99$^{**}$ | 598.05$^{**}$ | -1196.02 | 41 |
| SD | $t$ | — | 1.14 | 1.05 | 502.47 | -1004.87 | 38 |
| SD | $t$ | yes | 1.19 | 1.05 | 685.69$^{***}$ | -1371.30 | 42 |
| Static | $\mathcal{N}$ | — | 6.02$^{***}$ | 2.86$^{***}$ | | | |

*Note*: The distribution (Distr.) can be normal ($\mathcal{N}$) or Student's $t$, and the covariance structure can be diagonal or adjusted as in Eq. (10), as indicated by the column Adj. The last column specifies the number of parameters in each model. The log implied volatilities are forecasted for each option contract. For the DM test (with the state-space models as benchmarks), $^{***}$, $^{**}$, and $^{*}$ denote significance at the 1%, 5%, and 10% level, respectively.

Table A.2: Out-of-sample log-likelihood and 99% Value at Risk backtesting outcomes using *bucketed* data and with a *four-factor representation*, including violation ratio (Viol ratio) and tick loss, for both state-space (SS) and score-driven (SD) models applied to the log implied volatility model from Eqs. (1)–(2) See the note in Table A.1 for additional details.

| Model | Distr. | Adj. | loglik<br>$\times 10^{-3}$ | Viol ratio<br>$\times 10^{3}$ | Tickloss<br>$\times 10^{3}$ |
|---|---|---|---|---|---|
| | | | Full sample (2012-2022) | | |
| SS | $\mathcal{N}$ | — | 2529.16 | 0.21 | 1.75 |
| SD | $\mathcal{N}$ | — | 2087.72*** | 52.59 | 14.73*** |
| SD | $\mathcal{N}$ | yes | 2526.46*** | 0.13 | 2.15*** |
| SD | $t$ | — | 2336.27** | 53.40 | 16.56*** |
| SD | $t$ | yes | 2817.99*** | 0.00 | 5.72*** |
| | | | Pre-COVID period (2012-2020) | | |
| SS | $\mathcal{N}$ | — | 1930.77 | 0.20 | 1.65 |
| SD | $\mathcal{N}$ | — | 1649.28*** | 53.82 | 15.17*** |
| SD | $\mathcal{N}$ | yes | 1928.41*** | 0.15 | 1.98*** |
| SD | $t$ | — | 1833.79** | 55.55 | 17.50*** |
| SD | $t$ | yes | 2132.30*** | 0.00 | 6.02*** |
| | | | COVID period (2020-2022) | | |
| SS | $\mathcal{N}$ | — | 598.39 | 0.27 | 2.29 |
| SD | $\mathcal{N}$ | — | 438.44** | 46.09 | 12.40*** |
| SD | $\mathcal{N}$ | yes | 598.05** | 0.00 | 3.07*** |
| SD | $t$ | — | 502.47 | 42.05 | 11.57*** |
| SD | $t$ | — | 685.69*** | 0.00 | 4.10*** |

# B Information matrix for Student's $t$ distribution

The score for the Student's $t$ case with $\nu$ degrees of freedom and covariance matrix $\boldsymbol{H}_t$ is given by:

$$\nabla_t = \frac{\nu + N_t}{\nu - 2}\,\frac{\boldsymbol{M}_t^\top \boldsymbol{H}_t^{-1}\boldsymbol{\varepsilon}_t}{1 + \dfrac{\boldsymbol{\varepsilon}_t^\top \boldsymbol{H}_t^{-1}\boldsymbol{\varepsilon}_t}{\nu - 2}} = \frac{\nu + N_t}{\nu - 2}\,\sqrt{\frac{\nu - 2}{\nu}}\,\boldsymbol{M}_t^\top \boldsymbol{H}_t^{-1/2}\,\frac{\tilde{\boldsymbol{\varepsilon}}_t}{1 + \tilde{\boldsymbol{\varepsilon}}^\top \tilde{\boldsymbol{\varepsilon}}_t/\nu}\,, \tag{B.2}$$

where $\tilde{\boldsymbol{\varepsilon}}_t = \nu^{1/2}(\nu-2)^{-1/2}\boldsymbol{H}_t^{-1/2}\boldsymbol{\varepsilon}_t$ such that $\tilde{\boldsymbol{\varepsilon}}_t \sim t(0, \boldsymbol{I}, \nu)$. Therefore,

$$\begin{aligned}
\mathbb{E}_{t-1}[\nabla_t \nabla_t^\top] &= \frac{(\nu + N_t)^2}{(\nu-2)^2}\,\frac{\nu - 2}{\nu}\,\boldsymbol{M}_t^\top \boldsymbol{H}_t^{-1}\boldsymbol{M}_t\,\frac{1}{N_t}\,\mathbb{E}\left[\frac{\tilde{\boldsymbol{\varepsilon}}_t^\top \tilde{\boldsymbol{\varepsilon}}_t}{(1 + \tilde{\boldsymbol{\varepsilon}}_t^\top \tilde{\boldsymbol{\varepsilon}}_t/\nu)^2}\right] \\
&= \frac{(\nu + N_t)^2}{(\nu-2)}\,\frac{1}{\nu}\,\boldsymbol{M}_t^\top \boldsymbol{H}_t^{-1}\boldsymbol{M}_t\,\frac{1}{N_t}\,\mathbb{E}\left[\nu\frac{1}{1 + \tilde{\boldsymbol{\varepsilon}}_t^\top \tilde{\boldsymbol{\varepsilon}}_t/\nu}\left(1 - \frac{1}{1 + \tilde{\boldsymbol{\varepsilon}}_t^\top \tilde{\boldsymbol{\varepsilon}}_t/\nu}\right)\right] \\
&=: \frac{(\nu + N_t)^2}{(\nu-2)}\,\boldsymbol{M}_t^\top \boldsymbol{H}_t^{-1}\boldsymbol{M}_t\,\frac{1}{N_t}\mathbb{E}\left[b_{\nu,N_t}\left(1 - b_{\nu,N_t}\right)\right],
\end{aligned}$$

where $b_{\nu,N_t} = \left(1 + \tilde{\boldsymbol{\varepsilon}}_t^\top \tilde{\boldsymbol{\varepsilon}}_t/\nu\right)^{-1} \sim \text{Beta}(\nu, N_t)$. Using the expressions for the mean $\nu/(\nu + N_t)$ and the second-order uncentered moment $\nu(\nu+1)/[(\nu+N_t)(\nu+N_t+1)]$ of a beta distributed random variable, we therefore obtain

$$\begin{aligned}
\mathbb{E}_{t-1}[\nabla_t \nabla_t^\top] &= \frac{(\nu + N_t)^2}{(\nu-2)}\,\frac{1}{N_t}\,\left(\frac{\nu}{\nu + N_t} - \frac{\nu(\nu+1)}{(\nu+N_t)(\nu+N_t+1)}\right)\boldsymbol{M}_t^\top \boldsymbol{H}_t^{-1}\boldsymbol{M}_t \\
&= \frac{\nu}{(\nu-2)}\,\frac{\nu + N_t}{\nu + N_t + 1}\boldsymbol{M}_t^\top \boldsymbol{H}_t^{-1}\boldsymbol{M}_t.
\end{aligned}$$

**YC says: "I replace $q_t$ by $N_t$ above if I'm not mistaken. Moreover, it seems $b_{\nu,N_t}$ follows a distribution of Beta$(\nu/2, N_t/2)$. This should lead to**

$$\mathbb{E}\left[b_{\nu,N_t}(1 - b_{\nu,N_t})\right] = \frac{N_t \nu}{(\nu + N_t)(\nu + N_t + 2)}.$$

**Therefore,**

$$\mathbb{E}_{t-1}[\nabla_t \nabla_t^\top] = \frac{\nu}{(\nu-2)}\,\frac{\nu + N_t}{\nu + N_t + 2}\boldsymbol{M}_t^\top \boldsymbol{H}_t^{-1}\boldsymbol{M}_t.$$

"