TI 2025-036/III
Tinbergen Institute Discussion Paper

Improving Score-Driven Density Forecasts with an Application to Implied Volatility Surface Dynamics

Xia Zou[1]
Yicong Lin[2]
Andre Lucas[3]

1 Vrije Universiteit Amsterdam

2 Vrije Universiteit Amsterdam

3 Vrije Universiteit Amsterdam, Tinbergen Institute

Improving Score-Driven Density Forecasts with an Application to Implied Volatility Surface Dynamics

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

[1]Vrije Universiteit Amsterdam and Tinbergen Institute

May 27, 2025

Abstract

Point forecasts of score-driven models have been shown to behave at par with those of state-space models under a variety of circumstances. We show, however, that *density* rather than point forecasts of plain-vanilla score-driven models substantially underperform their state-space counterparts in a factor model context. We uncover the origins of this phenomenon and show how a simple adjustment of the measurement density of the score-driven model can put score-driven and state-space models approximately back on an equal footing again. The score-driven models can subsequently easily be extended with non-Gaussian features to fit the data even better without complicating parameter estimation. We illustrate our findings using a factor model for the implied volatility surface of S&P500 index options data.

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

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; Van der 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 models with shared dynamic components have also been used successfully for modeling term-structure dynamics (see, e.g., Creal et al., 2013; Koopman et al., 2017; Quaedvlieg and Schotman, 2022) and mixed measurement non-Gaussian factor models (Creal et al., 2014). Although an IV surface, unlike a term-structure, has two dimensions rather than one, the modeling principle remains the same.

An 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 parameter 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 set-up quickly become more challenging to estimate, often requiring Bayesian or other sampling-based methods or approximate estimation techniques such as the extended Kalman Filter or simulated ML (see, e.g., Durbin and Koopman, 2012).

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 as well as those based on non-Gaussian state-space models estimated using more complex machinery. The paper is silent, however, about the quality of the density forecasts. Results in Koopman et al. (2017) suggest that score-driven models might underperform from a density forecasting perspective compared to their state-space counterparts. In particular, the typical assumption of an exact factor structure in score-driven models, where all contemporaneous correlations are captured by a few common factors, appears too rigid. The origins of the difference between the density

forecast performance of the two model classes, however, remain largely underexplored.

This paper provides two main contributions. First, we show that standard score-driven models are outperformed by standard linear Gaussian state-space models in terms of density forecasts. We further 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 again in line with that of a state-space model, thus largely eliminating the difference in density forecast performance between the two model classes. The key is to match the covariance structure of the measurement noise more closely with that of the *predictive* rather than the *measurement* density of the state-space model. 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 we have eliminated the difference in density forecast performance between the score-driven and state-space approach for the Gaussian case, we can easily extend the score-driven model with non-Gaussian features without complicating the maximum likelihood estimation and inference procedures. Such non-Gaussian features may further increase the density forecast quality of the score-driven model beyond that achieved by the linear Gaussian state-space model. In addition, such non-Gaussian features result in a filtering procedure for time-varying parameter paths that is more robust to outliers (see Creal et al., 2013; Harvey and Luati, 2014; D'Innocenzo et al., 2023; Gasperoni et al., 2023). While adding non-Gaussian features to the state-space model is of course also possible, it would entail more challenges for the estimation procedure.

To illustrate our results empirically, we study the dynamics of IV surfaces for S&P500 index options using daily data from January 2010 to December 2022. We use the factor model of Goncalves and Guidolin (2006) and include five factors based on moneyness and time-to-maturity combinations. 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, even though the point forecasts are quite similar. However, when we incorporate the adjusted covariance structure for the measurement errors into the score-driven model as proposed in this paper, the density forecast performance of the score-driven and state-space models becomes very similar. Adding Student's t error terms to the score-driven model further increases its density fit beyond that of its state-space counterpart and helps us detect directions in which the model can be further improved.

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 forecast performance between score-driven and state-space models. Section 3 provides simulation evidence of the adjusted model's performance. Section 4 describes the data, while Section 5 provides the empirical results. Section 6 concludes. Additional empirical and technical results are available in the appendix.

2 The model

We first introduce the standard state-space and score-driven model set-up for implied volatility (IV) surfaces in Section 2.1. In Section 2.2, we then investigate the origins of the difference in density forecast performance between the two model classes and propose a solution to substantially reduce this disparity.

2.1 Standard state-space and score-driven models

We model a vector of 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 $\mathbb{T}_t \subset \mathbb{R}_+^{\tau_t}$, 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 automatically decreases as time progresses, thus changing its position on the time-to-maturity grid. Upon expiry, the option contract is completely removed from the dataset.

We assume that $\log \boldsymbol{IV}_t$ obeys the following factor structure:

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

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

The measurement equation in (1) consists of the matrix of exogenous, observed factor loadings $\boldsymbol{M}_t \in \mathbb{R}^{N_t \times p}$, a vector of factors $\boldsymbol{\beta}_t \in \mathbb{R}^p$, and an independent innovation term $\boldsymbol{\varepsilon}_t$ with distribution $h(\cdot \mid \boldsymbol{H}_t; \boldsymbol{\nu})$, where h denotes a distribution with mean zero, covariance matrix $\boldsymbol{H}_t$, and shape parameter vector $\boldsymbol{\nu}$. The state transition equation in (2) has an intercept vector $(\boldsymbol{I}_p - \boldsymbol{B})\bar{\boldsymbol{\beta}} \in \mathbb{R}^p$ where $\bar{\boldsymbol{\beta}}$ denotes the unconditional mean of $\boldsymbol{\beta}_t$, autoregressive matrix $\boldsymbol{B} \in \mathbb{R}^{p \times p}$ with all eigenvalues inside the unit circle, and 'state increment' vector $\boldsymbol{\xi}_t \in \mathbb{R}^p$. Here, $\boldsymbol{I}_p$ denotes an identity matrix of size p. We gather all static parameters of the model,

such as $\boldsymbol{\nu}$, $\bar{\boldsymbol{\beta}}$, $\boldsymbol{B}$, 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 parameter vector $\boldsymbol{\psi}$ that requires estimation.

This set-up unifies both state-space and score-driven models, depending on our choice of $\boldsymbol{\xi}_t$. For instance, if $\left\{(\boldsymbol{\varepsilon}_t^\top, \boldsymbol{\xi}_t^\top)^\top\right\}_{t\in\mathbb{Z}}$ is an independently and identically 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 estimate the static parameter vector $\boldsymbol{\psi}$ by maximizing the log-likelihood function $\mathcal{L}(\boldsymbol{\psi})$, 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 cases (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 log-likelihood function is known in analytic form through a standard prediction error decomposition. That is, $\boldsymbol{\xi}_t$ is $\mathcal{F}_{t-1}$-measurable, where $\mathcal{F}_t = \{\boldsymbol{IV}_t, \boldsymbol{IV}_{t-1}, \ldots, \boldsymbol{IV}_1\}$. Define $\mathbb{E}_t[\,\cdot\,] = \mathbb{E}[\,\cdot \mid \mathcal{F}_t]$. To illustrate, consider a normal distribution with covariance matrix $\boldsymbol{H}_t$ for the density $h(\,\cdot \mid \boldsymbol{H}_t; \boldsymbol{\nu})$. Given the conditional normality of $\boldsymbol{\varepsilon}_t$, the shape parameter $\boldsymbol{\nu}$ can be omitted. Defining the scaled score as $\boldsymbol{s}_t = \left[\mathbb{E}_{t-1}\left(\nabla_t\nabla_t^\top\right)\right]^{-1}\cdot\nabla_t$ with $\nabla_t = \partial \log h(\boldsymbol{IV}_t \mid \boldsymbol{\beta}_t; \boldsymbol{H}_t, \boldsymbol{\nu})/\partial\boldsymbol{\beta}$, and letting

$\boldsymbol{\xi}_t = \boldsymbol{A}\boldsymbol{s}_t$ for a parameter matrix $\boldsymbol{A} \in \mathbb{R}^{p\times p}$, we obtain

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

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

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: it adjusts 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 parameter $\nu > 2$ such that $\boldsymbol{\nu} = \nu$, the expressions change to

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

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

see Appendix B for a derivation of the scaled score in (6). Note that as $\nu \to \infty$, Eqs. (6)–(7) collapse to (4)–(5). If $\nu < \infty$, the score in (6) downweights 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 $\boldsymbol{\xi}_t$, which is random for the state-space set-up and pre-determined for the score-driven model. This distinction leads to a similar point forecast quality for both models (Koopman et al., 2016). However, in terms of density forecasts, the two models behave in markedly different ways, with the state-space specification generally performing better for the current class of factor models.

To understand this phenomenon, consider a simple version of model (1) with a diagonal error covariance matrix $\boldsymbol{H}_t$. Also define the point forecasts for the score-driven ($\boldsymbol{\beta}_t^{\text{sd}}$) and state-space model ($\boldsymbol{\beta}_{t|t-1}^{\text{ss}}$), respectively, where $\boldsymbol{\beta}_{t|t-1}^{\text{ss}} := \mathbb{E}[\boldsymbol{\beta}_t \mid \mathcal{F}_{t-1}]$. Assume that the point forecast quality of both models is similar, such that $\boldsymbol{\beta}_t^{\text{sd}} \approx \boldsymbol{\beta}_{t|t-1}^{\text{ss}}$, and both $\boldsymbol{\beta}_t^{\text{sd}}$ and $\boldsymbol{\beta}_{t|t-1}^{\text{ss}}$ are

$\mathcal{F}_{t-1}$-measurable. Note that Model (1) can be equivalently expressed as

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

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

$$\mathbb{V}\text{ar}\left[\log \boldsymbol{IV}_t \mid \mathcal{F}_{t-1}\right] = \boldsymbol{M}_t \, \mathbb{V}\text{ar}\left[\boldsymbol{\beta}_t \mid \mathcal{F}_{t-1}\right] \boldsymbol{M}_t^\top + \boldsymbol{H}_t. \tag{9}$$

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 in its original specification has a diagonal covariance $\boldsymbol{H}_t$, as $\boldsymbol{\beta}^{\text{sd}}_t$ is pre-determined conditional on $\mathcal{F}_{t-1}$. Thus, even if the score-driven forecast $\boldsymbol{\beta}^{\text{sd}}_t$ and the state-space forecast $\boldsymbol{\beta}^{\text{ss}}_{t|t-1}$ are close, their forecasting or predictive densities are very different.

The non-diagonal covariance specification in the predictive density typically provides a better fit to real data compared to a diagonal specification. To understand the intuition behind this, consider a simple one-factor set-up (i.e., $p = 1$). 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 state-space and score-driven models assume that for a given $\boldsymbol{\beta}_t$ the prediction errors around this (common) level are uncorrelated. However, as indicated above, 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 a 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 covariance matrix of the prediction errors, i.e., for the predictive density, 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 and does not fit the empirical correlation structure of the predictive density in factor model settings.

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. More specifically, we propose to replace the measurement equation of

the score-driven factor model in (1) with

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

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 closely. In particular, if the conditional covariance matrix $\mathbb{V}\text{ar}\left[\boldsymbol{\beta}_t \mid \mathcal{F}_{t-1}\right]$ would have a stable limit, e.g., in the case of a state-space model with time-invariant parameters, the above adjustment of the score-driven model would capture the steady state predictive density of the state-space specification. 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 improvements in density fit obtained by Koopman et al. (2017) when modeling international term structures and ad-hoc imposing an equicorrelation matrix structure in the score-driven specification. In their setting, the level factor is the most important element in their term-structure 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 adjusted score-driven model can easily be extended further to incorporate non-Gaussian features by modifying the score dynamics accordingly. This can be achieved without in any way complicating the parameter estimation procedure, which remains fully feasible using standard maximum likelihood methods based on an analytic expression of the log-likelihood function. We investigate some non-Gaussian extensions to the score-driven model in the empirical application in Section 4. In contrast, including such non-Gaussian features in a state-space setting typically comes with more challenges and usually requires a more complex estimation methodology based on numerical approximations and simulation techniques.

3 Simulation study

In this section, we investigate the performance of the different models, adjusted and unadjusted, in a Monte Carlo study. We simulate a time series of vector observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T$ from a state-space data-generating processes (DGPs) and compare the predictive perfor-

mances of the state-space (SS) model with four different score-driven (SD) models. The score-driven models assume either a normal or Student's t distribution for the measurement noise $\boldsymbol{\varepsilon}_t$, and are implemented both with and without the covariance adjustment in the measurement equation. We evaluate the performance of these models by comparing log-likelihoods, mean squared error (MSE), and mean absolute error (MAE) criteria.

We simulate data from a Gaussian state-space model with a factor structure as described in Eqs. (1)–(2). For the factor loadings $\boldsymbol{M}$, we use the restricted three factor specification, where we only include a constant, moneyness, and time-to-maturity as loadings. In the empirical application in Section 5 we augment this with two further factors in line with Van der Wel et al. (2016). We use moneyness levels $\boldsymbol{m}_t$ equal to $(0.9, 0.98, 1.05, 1.15, 1.3, 1.5)$ and time-to-maturity $\boldsymbol{\tau}_i$ $(10, 50, 100, 180)/255$ to be in line with the empirical application. The resulting factor loading matrix has dimension 24×3. The innovation term is drawn either from a multivariate normal or a Student's $t(3)$ distribution, both with covariance matrix $\boldsymbol{H}_t = \sigma_\varepsilon^2 \boldsymbol{I}_p$ for a high signal-to-noise ratio ($\sigma_\varepsilon^2 = 0.05$) and a low one ($\sigma_\varepsilon^2 = 0.50$). The high signal-to-noise ratio is closest to the empirical setting. We also set the remaining parameters in line with the empirical estimation results. For the state equation in Eq. (2), we choose a diagonal matrix $\boldsymbol{B} = \text{diag}(0.98, 0.93, 0.90)$, such that the first factors are most persistent, and the later factors in the data generating process are somewhat less.

The values for $\bar{\boldsymbol{\beta}}$ are randomly drawn from a uniform distribution over the range $(0, 1)$. The state innovations are drawn from a multivariate normal distribution with a zero mean and a covariance matrix $\boldsymbol{C} = \text{diag}(0.001, 0.005, 0.004)$, which again reflects the empirical estimation results.

We generate 1000 time series of $2 \cdot T$ observations for $T = 200, 1000$. The first T observations are used to estimate the model, while the remaining T observations are used to compute the performance criteria. In this way we avoid any potential biases that could be due to overfitting.

The results are presented in Table 1. The table has six panels, each corresponding to a different DGP. The left panels correspond to Gaussian measurement errors, while the right-hand panels are for the Student's $t(3)$ case. Within each panel, we present the out-of-sample MSE, MAE, and the average log-likelihood values for all estimated models. All these numbers are computed at the MLE for the corresponding statistical model and the data $\boldsymbol{IV}_t$.

The left-hand panels for the Gaussian DGPs in Table 1 highlight three key findings. First, in line with Koopman et al. (2016), we find little difference in terms of MSEs and MAEs between the different models. All of them perform well and at a similar level. In particular, the state-space model only performs marginally better, despite it being the true

Table 1: Out-of-sample performance for simulated data

Model	Distr.	Adj.	Gaussian DGP			$t(3)$ DGP		
			MSE	MAE	loglik	MSE	MAE	loglik
			$T = 200$, $\sigma_\varepsilon^2 = 0.05$					
SS	$\mathcal{N}$	—	0.061	0.197	0.762	0.060	0.169	0.684
SD	$\mathcal{N}$	—	0.063	0.200	-0.889	0.062	0.172	-0.907
SD	$\mathcal{N}$	yes	0.064	0.202	0.601	0.064	0.176	0.516
SD	t	—	0.063	0.200	-0.859	0.065	0.178	3.694
SD	t	yes	0.065	0.203	0.593	0.065	0.178	7.543
			$T = 200$, $\sigma_\varepsilon^2 = 0.50$					
SS	$\mathcal{N}$	—	0.520	0.575	-26.103	0.521	0.469	-26.514
SD	$\mathcal{N}$	—	0.528	0.580	-26.404	0.535	0.478	-26.888
SD	$\mathcal{N}$	yes	0.532	0.582	-26.266	0.533	0.478	-26.680
SD	t	—	0.528	0.580	-26.393	0.526	0.472	-19.582
SD	t	yes	0.532	0.582	-26.272	0.526	0.473	-19.079
			$T = 1000$, $\sigma_\varepsilon^2 = 0.05$					
SS	$\mathcal{N}$	—	0.061	0.197	0.830	0.060	0.168	0.859
SD	$\mathcal{N}$	—	0.062	0.199	-0.764	0.062	0.172	-0.756
SD	$\mathcal{N}$	yes	0.063	0.201	0.663	0.063	0.174	0.703
SD	t	—	0.062	0.199	-0.728	0.063	0.174	4.109
SD	t	yes	0.063	0.201	0.666	0.063	0.175	7.604

Note: This table presents the MSE, MAE, the average of log-likelihood (loglik) for a state-space (SS) model and score-driven (SD) models. The model distribution (Distr.) of the measurement noise $\boldsymbol{\varepsilon}_t$ is either normal ($\mathcal{N}$) or Student's t (t). The covariance structure in the score-driven (SD) specifications can be either diagonal (no Adj.) or adjusted (Adj.) as in Eq. (10); see the Adj. column. The table is based on 1000 simulations, and the different panels are for the different DGPs (small versus large sample size ($T = 200, 1000$), high ($\sigma_\varepsilon^2 = 0.05$) versus low ($\sigma_\varepsilon^2 = 0.50$) signal to noise ratio , and normal versus $t(3)$ measurement errors (with variance 1) in the DGP).

DGP. Second, consistent with the arguments of Section 2, the log-likelihood of the Gaussian score-driven model without covariance adjustment is substantially lower than that of the linear Gaussian state-space model (0.762 versus -0.889 for $(T, \sigma_\varepsilon^2) = (200, 0.05)$). However, when the covariance matrix adjustment is introduced, the log-likelihood of the Gaussian score-driven model aligns much closer with that of the linear Gaussian state-space model. Similarly, for score-driven models with a Student's t distribution, the model incorporating the covariance matrix adjustment exhibits a log-likelihood (0.516) much more comparable to that of the linear Gaussian state-space model (0.684), whereas the version without adjustment performs worse (-0.907). The gap cannot be closed completely given that the linear Gaussian

state space model is the true DGP here. Still, the results point to the fact that the density forecasting gap between the state-space and score-driven models can be made much smaller by the simple covariance matrix adjustment for the measurement errors. It underscores one of the more implicit conclusions in the original paper of Koopman et al. (2016). Also in their setting equal point forecasting performance of the Gaussian state-space model and the score-driven alternatives was only attainable if the measurement density in the score-driven specification was adjusted to allow for fatter tails than the tails of the conditional observation density in the state-space DGP. In other words, also in Koopman et al. (2005) the conditional measurement density for the score-driven model needed to be more flexible than its counterpart in the state-space DGP. Our result in this paper generalizes that finding and relates it to the covariance structure. In particular, in our factor model setting the covariance structure in the score-driven model needs to allow for cross-sectional correlations between the measurement errors to behave much more in line with the state-space model, not only in terms of point forecasts, but also in terms of density forecasts. The simulations show that this objective can indeed be achieved.

A third finding from the left-hand panel in Table 1 is that if we decrease the signal-to-noise ratio by considering $\sigma_\varepsilon^2 = 0.50$, then the density forecasting performance (loglik) is largely similar for the adjusted and unadjusted models. This again makes sense. If the signal is weak compared to the noise in the state-space DGP, then there is less cross-sectional correlation between the forecast errors. This explains why the performance of all score-driven models is similar. Finally, we see that the sample size hardly affects either the point or density forecast quality: both for $T = 200$ and $T = 1000$ the performance of the point and density forecasts is very similar for the adjusted score-driven models and their state-space counterparts, even though the latter is correctly specified. The unadjusted models, however, have a clearly worse density forecast performance (for instance -0.889 or -0.859 log-likelihood versus 0.762 for the state-space case for $T = 200$, $\sigma_\varepsilon^2 = 0.05$). Also the Gaussian and Student's t score-driven models behave at par: if the true measurement errors are Gaussian, the degrees of freedom parameter is typically estimated at a high value, rendering the Student's t and the Gaussian score-driven models very similar.

If we move to the Student's $t(3)$ DGPs in the right-hand panels of Table 1, the results are even more pronounced. Again, the point forecast quality is roughly similar for all models in the experiment. The density forecast performance, however, differs considerably across the different score-driven specifications. The unadjusted Gaussian score-driven model clearly has worse density forecast results in terms of log-likelihood for $\sigma_\varepsilon^2 = 0.05$. Its Gaussian adjusted counterpart already closes much of the performance difference with the state-space

model. Once we allow for non-Gaussian error terms in the measurement equation, the score-driven density forecasts clearly outperform those of the state-space model. Even though non-Gaussian error terms could in principle also be included in a state-space set-up, this would come at a substantial increase in computational costs (see, e.g., Durbin and Koopman, 2012). By contrast, the computational load for the score-driven models is rather insensitive to the use of a normal versus a Student's t distribution.

Interestingly, the density forecast performance of the score-driven Student's t model is improved further by including the covariance adjustment. The improvement is non-negligible, for instance, from 3.694 to 7.543 if the signal-to-noise ratio is sufficiently high ($T = 200$, $\sigma_\varepsilon^2 = 0.05$). Given that the high signal-to-noise ratio is closest to the empirical setting, we expect similar gains to be possible in the empirical application if we use the covariance-adjusted non-Gaussian score-driven factor model specification.

4 Empirical data and model specification

4.1 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 (Δ), and implied volatilities of each option contract.

We apply the filtering procedures of Barone-Adesi et al. (2008) and Van der Wel et al. (2016) to clean the data. Initially, we restrict our analysis to out-of-the-money options, defined by a Δ 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 Δ of 0.1 should possess the same implied volatility as an out-of-the-money put with a Δ of -0.9. 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 2 provides some summary statistics. Following Van der 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 Δ values of -0.375, -0.125, 0, 0.125, and 0.375.

We classify options with Δ values ranging from -0.125 to 0 as deep out-of-the-money puts (DOTM puts). Options with Δ from -0.375 to -0.125 are classified as out-of-the-money puts (OTM puts), and options with Δ 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 Δ values.

For each of these 24 groups, we present the time series average and standard deviation of the implied volatility, days-to-maturity, moneyness (the strike price divided by the index), Δ, 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 2 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.

4.2 Restricted factor representation

A common approach when modeling the IV surface is to express it as a function of moneyness ($m_{i,t}$, defined as the ratio of the strike price to the underlying index level) and time-to-maturity ($\tau_{i,t}$, expressed in year) 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 our adjusted covariance structure in the score-driven factor model for the IV surface. We therefore specify the following five-factor specification (Goncalves and Guidolin, 2006):

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

where $\boldsymbol{m}_{i,t}^\top = \left(1, m_{i,t}, m_{i,t}^2, \tau_{i,t}, m_{i,t}\tau_{i,t}\right)$ 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

Table 2: 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
	Δ	-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
	Δ	-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
	Δ	-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
	Δ	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
	Δ	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
	Δ	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 Δ, 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 Δs). 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.

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 is done in for instance the unrestricted specification in Van der Wel et al. (2016). This, however, does not alter any of the results that are the main focus of this paper, namely how to close the gap in density forecast performance between the state-space and score-driven approaches using an adapted covariance structure. We thus stick to the specification in (11) in our baseline analysis and investigate an additional factor with estimated factor loadings in the robustness analysis in Section 5.2.

5 Empirical results

In this section, we present our main empirical results. All of our analyses are performed 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 5.1, followed by robustness checks in Section 5.2.

5.1 The benchmark analysis

In our benchmark analysis, we compare the state-space (SS) model with four different score-driven (SD) models. The score-driven models use either a normal or Student's t specification for the measurement noise $\boldsymbol{\varepsilon}_t$, as described in Eqs. (4)–(5) and (6)–(7), respectively. For both distributional assumptions, we consider a version of the score-driven model with and without the covariance adjustment of the measurement equation as proposed in Eq. (10). In our 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. For the statistical measures, we compute the usual log-likelihoods, AIC criteria, mean

squared error (MSE), and mean absolute error (MAE) criteria. The latter are defined as

$$\text{MSE} = \frac{1}{T^\star}\sum_{t=1}^{T^\star}\frac{1}{N_t}\sum_{i=1}^{N_t}(\log IV_{i,t} - \log IV_{i,t|t-1})^2,$$
$$\text{MAE} = \frac{1}{T^\star}\sum_{t=1}^{T^\star}\frac{1}{N_t}\sum_{i=1}^{N_t}\left|\log IV_{i,t} - \log IV_{i,t|t-1}\right|,$$

respectively, where $\log IV_{i,t|t-1}$ denotes the one-step-ahead forecast of $\log IV_{i,t}$. 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 not only the conditional mean of $\log \boldsymbol{IV}_t$ matters, but where the whole forecasting density inclusive of its correlation structure plays a role. For this, we consider the unweighted overall cross-sectional average $P_t = N_t^{-1}\sum_{i=1}^{N_t}\log IV_{i,t}$ of the $\log \boldsymbol{IV}_t$s and consider the one-step-ahead risk quantiles of P_t. This provides a true density forecast performance contest for the different methods in economic terms. The one-step-ahead risk quantiles or Value-at-Risk for the score-driven specifications are straightforward to compute due to the observation-driven nature of the score-driven model. For a $(1-\alpha)$ confidence level, the Value-at-Risk is given by

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

where $Q(\alpha)$ is the α-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.

Table 3 presents the out-of-sample MSE, MAE, log-likelihood, and AIC, for both the state-space and score-driven models based on Eqs. (1)–(2). The MSE and MAE numbers are presented as ratios vis-à-vis the MSE and MAE of the linear Gaussian state-space benchmark model. As a second (more naive) benchmark, we also implemented a static factor model with $\boldsymbol{\beta}_t \equiv \boldsymbol{\beta}$. This static model, however, significantly underperformed all of the other specifications in all settings and is therefore left out of the discussion. Results are shown for the entire sample period and for two sub-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 model each time serving as the benchmark (Diebold and Mariano, 2002).

Table 3 highlights three main findings. First, the log-likelihood values indicate that the linear Gaussian state-space model significantly outperforms the Gaussian score-driven

Table 3: Out-of-sample performance for *non-bucketed* data

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	1557.08	-3114.12	16
SD	$\mathcal{N}$	—	0.99***	0.99***	1273.70***	-2547.37	16
SD	$\mathcal{N}$	yes	1.00	1.01***	1554.78***	-3109.52	21
SD	t	—	1.02	1.01	1537.55	-3075.07	17
SD	t	yes	0.99	1.00	1927.21***	-3854.38	22
Pre-COVID period (2012-2020)							
SS	$\mathcal{N}$	—	1.00	1.00	1095.66	-2191.29	16
SD	$\mathcal{N}$	—	0.99***	0.99***	922.00***	-1843.96	16
SD	$\mathcal{N}$	yes	1.01***	1.01***	1093.44***	-2186.84	21
SD	t	—	1.00	1.00	1141.53	-2283.02	17
SD	t	yes	1.00	1.01	1355.34***	-2710.63	22
COVID period (2020-2022)							
SS	$\mathcal{N}$	—	1.00	1.00	461.41	-922.79	16
SD	$\mathcal{N}$	—	1.00	1.00	351.70***	-703.37	16
SD	$\mathcal{N}$	yes	0.96***	0.99***	461.34**	-922.63	21
SD	t	—	1.11	1.05	396.03	-792.02	17
SD	t	yes	0.98	0.99	571.88***	-1143.71	22

Note: This table presents the MSE, MAE, log-likelihood (loglik), AIC, and BIC, for a state-space (SS) model and a score-driven (SD) factor model for the log implied volatilities of S&P500 index options as given in Eqs. (1)–(2). The distribution (Distr.) of the measurement noise ε_t is either normal ($\mathcal{N}$) or Student's t (t). The covariance structure in the score-driven (SD) specifications can be either 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 out-of-sample period covers January 1, 2012, through January 1, 2022. The log implied volatilities are forecast 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.

model without covariance adjustment. The log-likelihood is about 1557k for the state-space specification, whereas it only reaches 1273k for the score-driven model. However, when the covariance adjustment is applied to the score-driven model as suggested in this paper, the log-likelihood increases to 1554k, closely aligning with that of the linear Gaussian state-space model. This same pattern is also reflected in the AIC results, which reveal that the density forecasting gap between the state-space and score-driven model is nearly eliminated after introducing the simple covariance adjustment of the measurement equation in the score-

driven model.

Second, the relative ratios of MSE and MAE suggest that the state-space and score-driven 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 Gaussian score-driven model without covariance adjustment slightly outperforms the state-space model in terms of point forecasts as measured by the lower MSE and MAE. The differences, however, are small, though sometimes statistically significant. Conversely, the score-driven model with covariance adjustment exhibits a slightly higher MAE compared to the state-space model, while the MSE values are nearly identical. Interestingly, in the COVID period the patterns reverse: the score-driven model with covariance adjustment now outperforms the state-space model, while the score-driven model without covariance adjustment yields MSE and MAE values that are nearly identical to those of the state-space specification. In all cases, however, the density forecast performance of the adjusted score-driven model outperforms that of the unadjusted score-driven model.

Third, the results for score-driven models that use the Student's t distribution indicate that incorporating non-Gaussian features further improves the density forecast performance beyond that of the linear Gaussian state-space models. Specifically, the full-sample log-likelihood and AIC of the Student's t score-driven model without covariance matrix adjustment already closes most of the gap vis-à-vis the Gaussian state-space model and performs similarly to the Gaussian score-driven model *with* covariance matrix adjustment. It thus appears that both features can substantially improve density forecast performance. This is confirmed if we consider the Student's t score-driven model *with* covariance matrix adjustment: the out-of-sample log-likelihood for this model (1927k) is substantially larger than that of its Gaussian state-space counterpart (1557k) as well as that of the other score-driven specifications (1554k and 1537k). This is true for both tranquil and turbulent sub-periods. In sum, the unadjusted Gaussian score-driven model performs badly for density forecasts (though not for point forecasts), the covariance matrix adjustment largely remedies this problem, and the subsequent non-Gaussian model enhancements and covariance matrix adjustments further boost the density forecast performance of the score-driven model specification.

In Table 4, we evaluate the different models in economic terms by presenting the results for the 99%-Value-at-Risk (VaR). In terms of violation rates, we find that the score-driven models without covariance adjustment perform very badly. The predicted densities lie far from the true densities, as signaled by the VaR violation percentages above 47% for a nominal level of 1%. This holds for both the Gaussian and Student's t distributions, and for both the full sample and both sub-samples. As we are considering the quantiles of a sum $P_{t|t-1}$ of log

Table 4: 99% Value-at-Risk backtesting outcomes using *non-bucketed* data

Model	Distr.	Adj.	loglik $\times 10^{-3}$	Viol rate $\times 10^{3}$	Tick loss $\times 10^{3}$
Full sample (2012-2022)					
SS	$\mathcal{N}$	—	1557.08	0.43	3.46
SD	$\mathcal{N}$	—	1273.70***	470.66	11.78***
SD	$\mathcal{N}$	yes	1554.78***	0.43	4.19***
SD	t	—	1537.55	492.93	14.07***
SD	t	yes	1927.21***	0.00	4.91***
Pre-COVID (2012-2020)					
SS	$\mathcal{N}$	—	1095.66	0.51	2.98
SD	$\mathcal{N}$	—	922.00***	485.23	12.18***
SD	$\mathcal{N}$	yes	1093.44***	0.51	3.56***
SD	t	—	1141.53	518.33	15.23***
SD	t	yes	1355.34***	0.00	4.23***
COVID period (2020-2022)					
SS	$\mathcal{N}$	—	461.41	0.00	6.04
SD	$\mathcal{N}$	—	351.70***	393.53	9.65***
SD	$\mathcal{N}$	yes	461.34**	0.00	7.55***
SD	t	—	396.03	358.49	7.94**
SD	t	yes	571.88***	0.00	8.55***

Note: This table presents the out-of-sample log-likelihood (loglik), including the violation ratio (Viol ratio) and tick loss, for both the state-space (SS) and score-driven (SD) model applied to the log implied volatility model from Eqs. (1)–(2). The distribution (Distr.) of the measurement noise $\boldsymbol{\varepsilon}_t$ is either normal ($\mathcal{N}$) or Student's t (t). The covariance structure in the score-driven (SD) specifications can be either diagonal or adjusted as in Eq. (10), as indicated by the column Adj. The out-of-sample period covers January 1, 2012, through January 1, 2022. The log implied volatilities are forecast 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.

IVs, the correlation structure of the data matters substantially. As explained in Section 2, the unadjusted score-driven model assumes all forecast errors to be uncorrelated. This results in a relatively small VaR, as the forecast errors are assumed to cancel against each other given the uncorrelatedness assumption. In reality, of course, it is much more likely that the model makes an error in forecasting the level component ($\beta_{1,t}$). If, for instance, the level of the IV surface is forecast too low, the forecasts of all IVs are too low on average and all

forecast errors are correlated. Due to this, the VaR should be much higher than assumed by the unadjusted model, which results in a high violation rate for the unadjusted score-driven model.

If we use the covariance matrix adjustment for the score-driven model, the VaR violation rates of the Gaussian score-driven model immediately behave at par with those of the state-space model, or are even somewhat more conservative if the Student's t distribution is used for the predictive density and the score. Also, the tick loss functions for the adjusted score-driven models are much closer to the state-space model, signaling that the adjustment succeeds in making the density forecast performance of the score-driven and the state-space approach more similar. It underlines again that the uncorrelatedness assumption that is typical in state-space factor model specifications cannot simply be imposed in a score-driven setting, and that the covariance matrix adjustment for the measurement equation of the score-driven model is indispensable to make the models more competitive, not only in terms of point forecasts, but also in terms of density forecasts.

5.2 Robustness check with bucketed data

To verify the robustness of our previous findings for individual options' data, we also apply our analysis to bucketed data. This follows the approach of for instance Van der Wel et al. (2016), Bollen and Whaley (2004), and Barone-Adesi et al. (2008). Like these previous papers, we divide the data into four maturity groups, separated by maturities of 45, 90, and 180 days, and six moneyness groups, separated by Δs of -0.375, -0.125, 0, 0.125, and 0.375, as shown in Table 2. For each maturity-moneyness group, we select the contract closest to the mid-point. Stacking the log IVs into a vector for the different groups leads to a 24-dimensional vector at all times.

Tables 5 and 6 present the results for the *bucketed* data. Table 5 shows that the plain-vanilla, unadjusted score-driven models again have a significantly lower log-likelihood than the state-space models. The log-likelihoods of the score-driven models with covariance matrix adjustment are again significantly higher than their unadjusted counterparts. For the bucketed data, the adjusted score-driven model even has a higher out-of-sample likelihood than the state-space model. This result is robust for both the full sample and the two sub-samples, providing even stronger evidence in favor of the adjustment than the unbucketed results in Table 3. For the second sub-sample, which includes the COVID period, the Gaussian score-driven model behaves significantly better than its state-space counterpart in terms of log-likelihood.

In terms of point forecasts, the Gaussian score-driven models perform better than or at

Table 5: Out-of-sample performance using *bucketed* data

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	62.18	-124.33	16
SD	$\mathcal{N}$	—	0.97*	0.98***	54.41***	-108.8	16
SD	$\mathcal{N}$	yes	0.93**	0.97***	64.01***	-127.97	21
SD	t	—	3.33	1.25	55.73***	-111.42	17
SD	t		1.32	1.03	66.34***	-132.63	22
Pre-COVID (2012-2020)							
SS	$\mathcal{N}$	—	1.00	1.00	53.62	-107.21	16
SD	$\mathcal{N}$	—	0.94***	0.98***	48.20***	-96.36	16
SD	$\mathcal{N}$	yes	0.96***	0.99**	55.18***	-110.31	21
SD	t	—	1.21*	1.04*	50.41***	-100.79	17
SD	t		1.06	1.00	57.27***	-114.49	22
COVID period (2020-2022)							
SS	$\mathcal{N}$	—	1.00	1.00	8.56	-17.08	16
SD	$\mathcal{N}$	—	1.00	0.98	6.22*	-12.40	16
SD	$\mathcal{N}$	yes	0.90***	0.94***	8.83**	-17.62	21
SD	t	—	5.67	1.81	5.32	-10.60	17
SD	t		1.61	1.11	9.07**	-18.10	22

Note: This table presents the MSE, MAE, log-likelihood (loglik), AIC, and BIC, for a state-space (SS) model and a score-driven (SD) model factor model for the log implied volatilities of S&P500 index options as given in Eqs. (1)–(2). Individual options data are grouped into 24 (time-to-maturity, moneyness) bins and represented by the log IV for the option closest to the midpoint of each bucket. The distribution (Distr.) of the measurement noise $\boldsymbol{\varepsilon}_t$ is either normal ($\mathcal{N}$) or Student's t (t). The covariance structure in the score-driven (SD) specifications can be either 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 out-of-sample period covers January 1, 2012, through January 1, 2022. The log implied volatilities are forecast 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.

par with the state-space specification, whether with or without the covariance adjustment. The result holds both in terms of MSE and MAE. In several cases, the improvement is even statistically significant.

For the bucketed data, the Student's t distribution appears to have two different effects. When we compare the out-of-sample log-likelihoods of the adjusted Gaussian score-driven

model with that of its Student's t counterpart, we see that the out-of-sample log-likelihood increases by about 2300 points. The MSE and MAE results for the Student's t specification, however, are worse than those of the Gaussian model. The differences appear largest during the COVID period (lower panel). To understand the latter, Appendix C gives some more background. Particularly the rapid upward level shifts in the implied volatility surface during the early stages of the COVID-19 period is picked up better (out-of-sample) in terms of MSE and MAE by the Gaussian than the Student's t model. The Student's t based score-driven model makes a trade-off for every new observation, whether to ascribe it to a change in the factor loadings $\boldsymbol{\beta}_t$ or to the fat-tailedness of the error process. This is done through the weights $(1 + \boldsymbol{\varepsilon}_t^\top \boldsymbol{H}_t^{-1} \boldsymbol{\varepsilon}_t/(\nu-2))^{-1}$ in (6). As a result, the Student's t based model requires some more observations to react to a real level shift. The consequence for the sample at hand is that the Student's t model fails to fully adapt to the steep level shifts during the COVID lockdowns, leading to a worse point forecast performance. Our main conclusion, however, remains: the adjustment of the measurement equation in score-driven factor models seems indispensable for a good empirical performance of these models and improves both the density and point forecast quality.

Table 6 presents the Value-at-Risk results for the bucketed data. We confirm our earlier results. The violation rates near 15% for the unadjusted models are much too high compared to the 1% nominal level. For the adjusted model specifications, by contrast, the rejection percentages of 0.5% are only slightly more conservative than the nominal level. We also see that the state-space and adjusted score-driven models behave similarly in terms of VaR violations. In addition, the tick-loss functions of the adjusted score-driven models—under both Gaussian and Student's t distributions—are better than those of the state-space model. We again conclude that the simple adjustment of the measurement equation in score-driven factor models substantially improves their density forecast quality, without affecting their already adequate point forecast accuracy compared to the benchmark models.

As a final robustness check, we investigate whether a less restrictive factor model specification alters the results. In our last specification, which we refer to as the *four-factor representation*, we replace the interaction term between moneyness and time-to-maturity, as well as the squared moneyness term introduced in Section 4.2 by a sequence of estimated factor loadings, thus allowing for much more flexibility. We divide the *non-bucketed* data into the same 24 groups as before. Define the group membership or bucket numbers $g_{i,t} = g(m_{i,t}, \tau_{i,t}) \in \{1, \ldots, 24\}$ if contract i at time t belongs to bucket $g_{i,t}$, based on its moneyness and time-to-maturity value at time t. The specification then becomes

$$\log IV_{i,t} = \beta_{1,t} + \beta_{2,t} m_{i,t} + \beta_{3,t} \tau_{i,t} + \beta_{4,t} \omega_{4,g_{i,t}} + \varepsilon_{i,t}, \tag{13}$$

Table 6: 99% Value at Risk performance for *bucketed* data

Model	Distr.	Adj.	loglik $\times 10^{-3}$	Viol rate $\times 10^{3}$	Tick loss $\times 10^{3}$
Full sample (2012-2022)					
SS	$\mathcal{N}$	—	62.18	1.71	1.81
SD	$\mathcal{N}$	—	54.41***	157.17	4.63***
SD	$\mathcal{N}$	yes	64.01***	5.14	1.69***
SD	t	—	55.73***	134.48	4.18***
SD	t		66.34***	0.43	1.54**
Pre-COVID (2012-2020)					
SS	$\mathcal{N}$	—	53.62	1.02	1.78
SD	$\mathcal{N}$	—	48.20***	173.12	5.09***
SD	$\mathcal{N}$	yes	55.18***	5.09	1.70***
SD	t	—	50.41***	149.19	4.52***
SD	t		57.27***	0.51	1.52**
COVID period (2020-2022)					
SS	$\mathcal{N}$	—	8.56	5.39	1.95
SD	$\mathcal{N}$	—	6.22***	72.78	2.22***
SD	$\mathcal{N}$	yes	8.83***	5.39	1.62***
SD	t	—	5.32***	56.6	2.42***
SD	t		9.07***	0.00	1.70**

Note: Log-likelihood (loglik), Value-at-Risk, and tick-loss results for a state-space (SS) model and a score-driven (SD). Individual options data are grouped into 24 (time-to-maturity, moneyness) bins and represented by log IV for the option closest to the midpoint of each bucket. The distribution (Distr.) of the measurement noise $\boldsymbol{\varepsilon}_t$ is either normal ($\mathcal{N}$) or Student's t (t). The covariance structure in the score-driven (SD) specifications can be either diagonal or adjusted as in Eq. (10), as indicated by the column Adj. The out-of-sample period covers January 1, 2012, through January 1, 2022. The log implied volatilities are forecast 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.

where $\omega_{4,g}$ for $g = 1, \ldots, 24$ are group-specific coefficients that need to be estimated. This specification adds quite some flexibility compared to model (11), as we replace the rigid specification of the loadings ($m_{i,t}^2$ and $m_{i,t} \cdot \tau_{i,t}$) for the last two factors ($\beta_{4,t}$ and $\beta_{5,t}$) by a single factor ($\beta_{4,t}$) with a more flexible, non-parametric specification; compare Van der Wel et al. (2016).

The results for this new specification are presented in Tables A.1 and A.2 in the appendix. The new specification with flexible dynamic levels for each of the buckets clearly provides a substantial increase in fit for the individual, non-bucketed data. For instance, the out-of-sample log likelihood increases from a level of around 1500k for the Gaussian state-space and adjusted score-driven models in Table 3, to a level of around 2500k in Table A.1. This is mainly due to the new model's much more flexible factor loadings for the last factor. The main conclusions of this paper regarding the density forecasting performance between the state-space and score-driven model specification, however, remain unaltered. Also for this flexible version of the model, the unadjusted score-driven model performs badly in terms of density forecasting performance compared to the state-space version of the model. By contrast, the adjusted score-driven models largely close this gap in terms of both point and density forecasts. We therefore conclude that the proposed adjustment is useful if not indispensable when building score-driven factor models for forecasting.

6 Conclusion

Score-driven and state-space models are known to produce similar one-step-ahead forecast quality. In this paper, we showed that in terms of density forecasts, however, score-driven models can perform significantly worse. We investigated the origins of this performance difference in a factor model setting and suggested a simple remedy by adjusting the covariance matrix structure of the score-driven model to be more in line with that of the *predictive* rather than the *measurement error* density of its state-space counterpart.

Using this adjustment, we showed that the original substantial difference in density forecast quality between the two classes of models can be decreased substantially, or even closed, and that both model classes are put on an equal footing again in terms of point *and* density forecast quality. The advantage of this adjustment is that it can seamlessly be incorporated into the (adjusted) score-driven modeling approach, which can subsequently be extended with non-Gaussian features to fit the data even better. Such extensions do not complicate the estimation approach of score-driven models in any way: estimation can still be based on an analytic expression of the log-likelihood through a standard prediction error decomposition. Estimation for non-Gaussian error terms in a state-space context, however, is typically more challenging, requiring numerical approximations and simulation-based methods.

We applied the new approach to model the implied volatility surface for S&P500 index options data. We confirmed that a Gaussian plain-vanilla score-driven model has a significantly worse density forecast performance than a linear Gaussian state-space model. The

performance difference is substantial, both in statistical and economic terms. However, after adjusting the score-driven model's covariance structure as proposed in this paper, this out-of-sample performance gap vanishes and can even reverse, especially when incorporating non-Gaussian features.

We therefore conclude that the proposed adjustment provides a simple fix for the out-of-sample density forecasting performance of score-driven factor models. The adjustment approach can also be applied in high-dimensional cases if the number of factors remains limited. This is due to its parsimonious nature, which is rooted in the underlying factor model structure. Empirically, we illustrated this in the context of modeling implied volatility surfaces for individual (rather than bucketed) option contracts, in which case one easily has more than hundreds of time series observations every single period.

It is interesting of course to further investigate whether similar density forecast performance differences exist for other model settings than the factor model context. If so, one might investigate whether the solution proposed in this paper can also bring the state-space and score-driven approaches closer together in such an alternative context. We leave this for future research.

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.

Creal, Drew, Bernd Schwaab, Siem Jan Koopman, and Andre Lucas (2014). "Observation-driven mixed-measurement dynamic factor models with an application to credit risk". *Review of Economics and Statistics* 96.5, pp. 898–915.

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, Borus Jungbacker, and Eugenie Hol (2005). "Forecasting daily variability of the S&P 100 stock index using historical, realised and implied volatility measurements". *Journal of Empirical Finance* 12.3, pp. 445–475.

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.

Van der Wel, Michel, 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.

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.

A Further robustness checks

Table A.1: Out-of-sample performance with a *four-factor representation* for *non-bucketed* data

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	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	1.01*	1.02***	2820.04***	-5640.00	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	t	yes	1.01***	1.02***	2133.74***	-4267.40	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	1.02	1.02	686.30***	-1372.51	42
Static	$\mathcal{N}$	—	6.02***	2.86***			

Note: This table presents the MSE, MAE, log-likelihood (loglik), AIC, and BIC, for a state-space (SS) model, a score-driven (SD) model, and a static ($\boldsymbol{\beta}_t \equiv \boldsymbol{\beta}$) factor model for the log implied volatilities of S&P500 index options as given in Eqs. (1)–(2). This four-factor representation includes a level factor, slopes in the moneyness and time-to-maturity dimensions, and a nonparametric factor. The distribution (Distr.) of the measurement noise $\boldsymbol{\varepsilon}_t$ is either normal ($\mathcal{N}$) or Student's t (t). The covariance structure in the score-driven (SD) specifications can be either 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 out-of-sample period covers January 1, 2012, through January 1, 2022. The log implied volatilities are forecast 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: 99% Value at Risk performance with a *four-factor representation* using *non-bucketed* data

Model	Distr.	Adj.	loglik $\times 10^{-3}$	Viol ratio $\times 10^{3}$	Tickloss $\times 10^{3}$
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	2820.04***	0.86	2.47***
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	2133.74***	0.51	2.41***
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	—	686.30***	2.70	2.80***

Note: This table presents the out-of-sample log-likelihood (loglik), including the violation ratio (Viol ratio) and tick loss, for both the state-space (SS) and score-driven (SD) model applied to the log implied volatility model from Eqs. (1)–(2). This four-factor representation includes a level factor, slopes in the moneyness and time-to-maturity dimensions, and a nonparametric factor. The distribution (Distr.) of the measurement noise $\boldsymbol{\varepsilon}_t$ is either normal ($\mathcal{N}$) or Student's t (t). The covariance structure in the score-driven (SD) specifications can be either diagonal or adjusted as in Eq. (10), as indicated by the column Adj. The out-of-sample period covers January 1, 2012, through January 1, 2022. The log implied volatilities are forecast 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.

B Information matrix for Student's t distribution

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

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

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 $b_{\nu,N_t} = \left(1 + \tilde{\boldsymbol{\varepsilon}}_t^\top \tilde{\boldsymbol{\varepsilon}}_t / \nu\right)^{-1} \sim \text{Beta}(\nu/2, N_t/2)$. Using the expressions for the mean $\nu/(\nu + N_t)$ and the second-order uncentered moment $\nu(\nu+2)/[(\nu+N_t)(\nu+N_t+2)]$ of a beta distributed random variable, we therefore obtain

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

C MSE of the non-adjusted Student's t model for bucketed data

In this section, we examine the high MSE of the non-adjusted Student's t score-driven model reported in Table 5. Figure C.1 presents the yearly MSE for both the state-space model (denoted as SS), the non-adjusted Student's t score-driven model (denoted as E_T) and the adjusted Student's t (denoted as EAT) over time. The results indicate that the substantial discrepancy between the state-space and the Student's t models is concentrated in the year 2020.

A closer inspection of the 2020 MSE values, shown in Figure C.2, reveals that the poor performance of the Student's t model is particularly pronounced during the March–April period, which coincides with the onset of COVID-19 lockdowns in most countries. Figure C.3, which plots the mean level of implied volatility in 2020, highlights several significant upward shifts in the volatility surface during this time. Due to its heavy-tailed nature, the Student's t model tends to interpret such abrupt changes as outliers and therefore responds more conservatively. As a result, the model fails to adjust sufficiently to structural level shifts—especially when these shifts exhibit momentum—leading to poor forecasting performance during this period. In addition, we observe that incorporating a covariance adjustment mitigates the issue to a considerable extent, indicating the importance of accommodating a correlated covariance structure.

Finally, Figure C.4 shows that the estimated degree-of-freedom parameter remains relatively stable over time, suggesting that the deterioration in performance is not due to changes in the tail behavior of the fitted distribution itself.

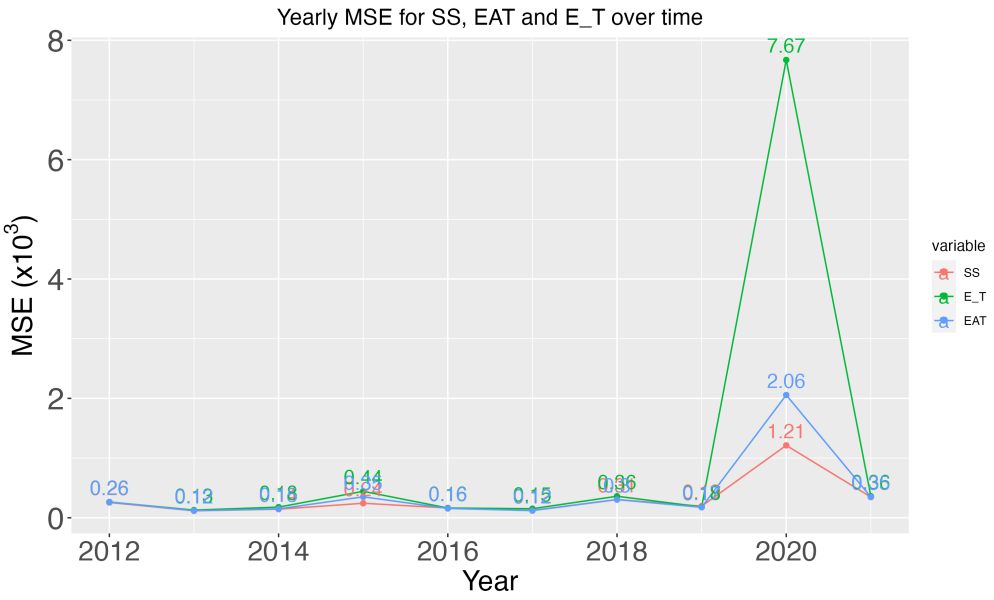

Figure C.1: Yearly MSE for the state space and non-adjusted Student's t model

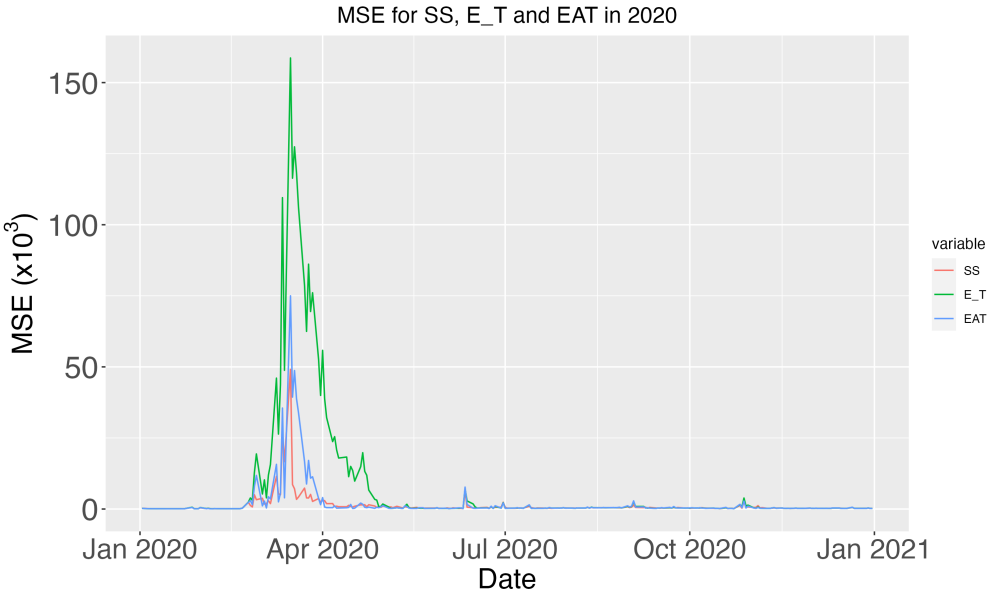

Figure C.2: MSE for state space and non-adjusted Student's t model in 2020

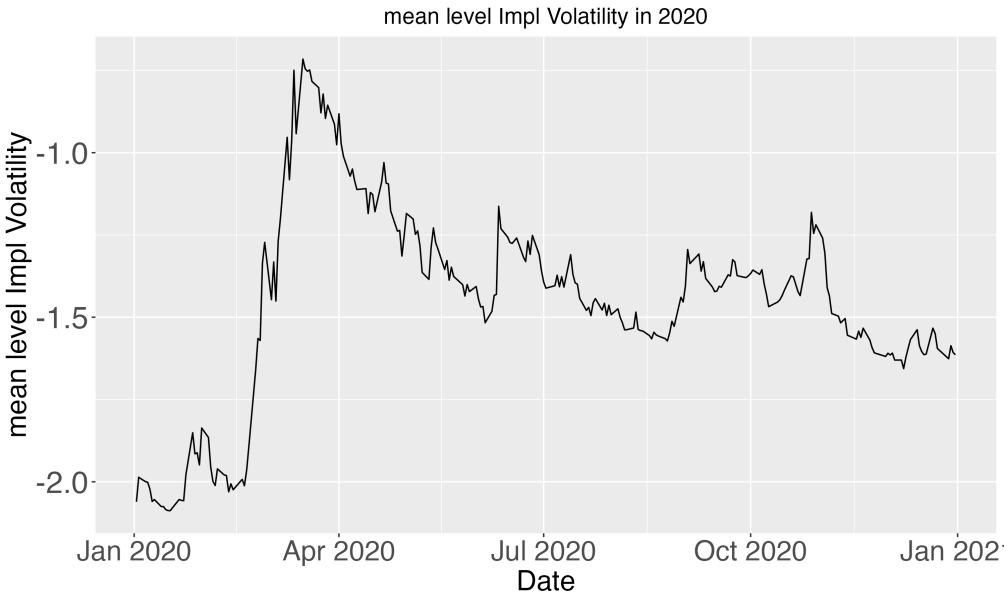

Figure C.3: Mean level of log implied volatility in 2020

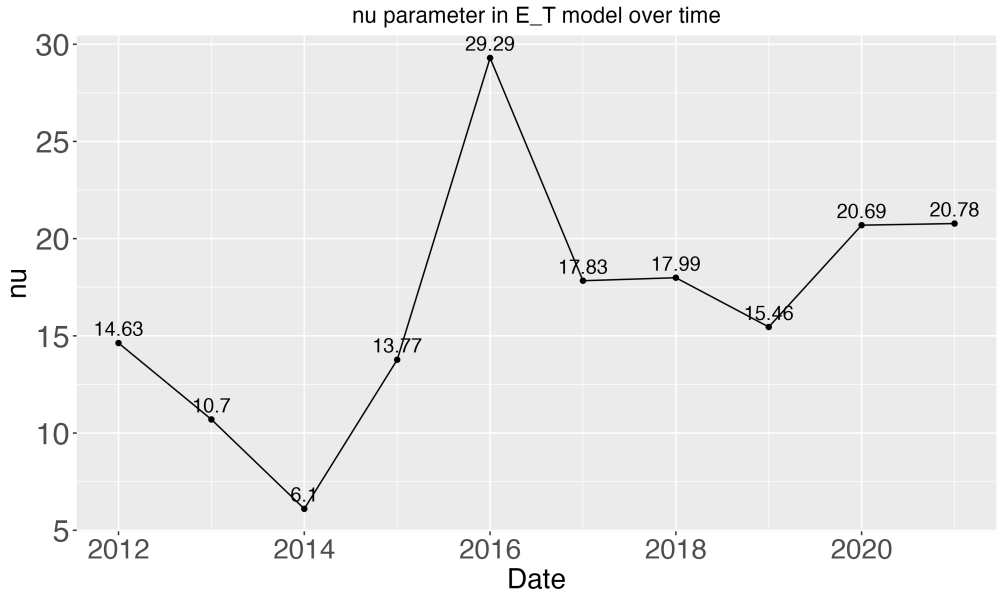

Figure C.4: Estimated degree of freedom in non-adjusted Student's t model for different prediction window