UNIVERSITY OF LIVERPOOL Management School

Working Paper in Economics

 # 20183

October 2018

A Multi-Factor Transformed Diffusion Model with Applications to VIX and VIX Futures

Ruijun Bu
Fredj Jawadi
Yuyi Li

https://www.liverpool.ac.uk/management/people/economics/

A Multi-Factor Transformed Diffusion Model with Applications to VIX and VIX Futures

Ruijun Bu* Fredj Jawadi† Yuyi Li‡

August 2018

Abstract

Transformed diffusions (TDs) have become increasingly popular in financial modelling for their model flexibility and tractability. While existing TD models are predominately one-factor models, empirical evidence often prefers models with multiple factors. We propose a novel distribution-driven nonlinear multi-factor TD model with latent components. Our model is a transformation of a underlying multivariate Ornstein Uhlenbeck (MVOU) process, where the transformation function is endogenously specified by a flexible parametric stationary distribution of the observed variable. Computationally efficient exact likelihood inference can be implemented for our model using a modified Kalman filter algorithm and the transformed affine structure also allows us to price derivatives in semi-closed form. We compare the proposed multi-factor model with existing TD models for modelling VIX and pricing VIX futures. Our results show that the proposed model outperforms all existing TD models both in the sample and out of the sample consistently across all categories and scenarios of our comparison.

JEL Classification: CC13, C32, G13, G15
Keywords: Transformation Model; Nonlinear Diffusion; Latent Factor; Kalman Filter; Volatility Index

*University of Liverpool, UK. Email: ruijunbu@liv.ac.uk.
†University of Evry, France. Email: fredj.jawadi@univ-evry.fr.
‡University of Liverpool, UK. Email: yuyili@liv.ac.uk.

1. Introduction

Since the seminal work of Merton (1973), continuous-time diffusion models have proved to be extremely useful in financial and economic modelling. In particular, they have been frequently applied to research in the dynamics of key economic variables such as short-term interest rates, exchange rates, and more recently volatility indices (c.f. Chan et al. 1992, Bu, Cheng and Hadri 2016, 2017, Bu, Jawadi and Li 2017).

A parametric univariate or one-factor continuous-time diffusion process, say $\{Y_t, t \geq 0\}$, is usually described by the following Stochastic Differential Equation (SDE):

$$dY_t = \mu_Y(y;\psi)\,dt + \sigma_Y(Y_t;\psi)\,dW_t \tag{1}$$

where $\mu_Y(y;\psi)$ and $\sigma_Y^2(y;\psi)$ are, respectively, the drift and diffusion functions with parameter ψ and $\{W_t, t \geq 0\}$ is a standard Brownian motion. Well known examples in finance include Merton (1973), Vasicek (1977), Cox et al. (1985), Duffie and Kan (1996), Aït-Sahalia (1996b), Conley et al. (1997), Ahn and Gao (1999), Detemple and Osakwe (2000), and more recently Bu et al. (2011), Eraker and Wang (2015), Bu, Jawadi and Li (2017), among others.

Maximum Likelihood (ML) is usually the preferred method of estimation. However, except for a few special cases such as the Geometric Brownian Motion (GBM) (c.f. Black and Scholes 1973), the Ornstein Uhlenbeck (OU) process (c.f. Vasicek 1977) and the square-root or CIR process (c.f. Cox et al. 1985), most continuous-time diffusion models do not possess closed-form transition densities. Nevertheless, nonlinearities beyond the assumptions of these models are often documented in the literature (c.f. Aït-Sahalia 1996b, Stanton 1997, Bu et al. 2011, Eraker and Wang 2015, and Bu, Cheng and Hadri 2017)[1]. Hence, one strand of literature focuses on developing density approximation techniques for nonlinear diffusions. Main examples include Lo (1988), Pedersen (1995), Brandt and Santa-Clara (2002), Shoji and Ozaki (1998), Kessler (1997), Elerian et al. (2001), and Aït-Sahalia (2002). Meanwhile, another strand of literature aims to find a balance between model flexibility and tractability with no recourse to density approximations. These studies advocate the use of the so-called transformed diffusion (TD) models. TDs are usually nonlinear transformations of tractable, typically affine, underlying diffusions (UDs). Hence, TDs are potentially flexible diffusion models capable of capturing nonlinear features in the data while at the same time possess some desirable analytical and statistical tractability inherited from the more tractable UDs. Primary examples of TDs, among others, include Ahn and Gao (1999), Bu et al. (2011), Goard and Mazur (2013), Forman and Sørensen (2014), Eraker and Wang (2015), and most recently Bu, Jawadi and Li (2017), Bu et al. (2018).

Compared with the density approximation methods for nonlinear diffusions, among which the closed-form expansion method of Aït-Sahalia (2002) is by far the most widely used in applications, the TD approach has several notable advantages. First, the theory of TDs is general enough to accommodate nonlinear UDs. Thus, the class of TDs is broad enough to encompass all parametric diffusions. For example, Bu et al. (2018) consider a nonparametric transformation of the Nonlinear

[1] Pritsker (1998) and Chapman and Pearson (2000) argued that nonparametric kernel-based specification tests used by Aït-Sahalia (1996b) and Stanton (1997) tend to suffer from severe size distortions. However, evidence from parametric specification tests often prefer nonlinear models. See for example Bu et al. (2011), Eraker and Wang (2015), and Bu, Cheng and Hadri (2017).

Drift Constant Elasticity Variance (NLDCEV) model of Aït-Sahalia (1996b)[2]. Second, simple transformations of affine models may actually perform better than popular general nonlinear diffusions. For example, Bu, Cheng and Hadri (2017) find that the transformed CIR model with the CEV diffusion function fits the VIX data significantly better than the NLDCEV model. Third, TDs inherit many important statistical and dynamic properties from their UDs. Forman and Sørensen (2014) show that these include stationarity, ergodicity, mixing, first passage time, and knowledge about the transition densities. For general nonlinear diffusions, however, these properties usually need to be determined on a case-by-case basis[3]. Fourth, TDs are particularly attractive for financial modelling if the UDs are affine. The affine class of models is by far the most widely adopted in asset pricing (c.f. Dai and Singleton 2000) because of their exceptional tractabilities. For transformed affine models, statistical inference (e.g. ML estimation and forecasting) and financial applications (e.g. derivative pricing, copula modelling) are often tractable too (c.f. Bu et al. 2011, Eraker and Wang 2015, Bu, Jawadi and Li 2017, Amengual and Xiu 2018). In contrast, financial modelling with general nonlinear diffusions can be significantly more difficult.

The issue of lack of tractability is most prominent for nonlinear multi-factor models with latent components. Aït-Sahalia (2008) developed a density approximation method for multivariate or multi-factor diffusions, which was then applied by Aït-Sahalia and Kimmel (2007) for estimating stochastic volatility models. Since the volatility process is not directly observed, model estimation requires the use of nonlinear filtering for general latent-factor diffusions, which can be computationally costly. They therefore used a proxy extracted from option data to replace the unobserved volatility[4]. However, in many other cases proxies are not always available. For instance, for the stochastic central tendency model of Mencía and Sentana (2013), no obvious proxy is available for the latent central tendency variable. Thus, most of the time, filtering directly applied to the observed series is inevitable. Therefore, there is a clear need for developing models that can circumvent these difficulties.

The most important property of TD models is their tractable transition densities, which facilitate statistical inferences and applications substantially. Ahn and Gao (1999) considered a transformed CIR model for modelling short-term interest rates and closed-form pricing of bonds. Detemple and Osakwe (2000) used a transformed OU model for modelling volatility index and closed-form pricing of VIX futures (VXFs) and options. Bu et al. (2011) also used TDs for modelling UK and US interest rates. Forman and Sørensen (2014) also considered a transformed OU model but study molecular dynamics data. Eraker and Wang (2015) considered a transformed CIR model with a cubic drift function for pricing VXFs. Most recently, Bu, Jawadi and Li (2017) proposed the distribution-driven transformed OU and CIR models with stationary distributions specified as the Skewed Student-t (SKST) distribution of Hansen (1994) for modelling VIX and

[2]The NLDCEV model is written as

$$dX_t = \left(a_{-1}X_t^{-1} + a_0 X_t + a_1 X_t + a_2 X_t^2\right) dt + \sigma X_t^{\gamma} dW_t$$

and is one of the most widely used general nonlinear diffusion models in finance. See for example Aït-Sahalia (1996b), Conley et al. (1997), Choi (2009).

[3]See Karlin and Taylor (1981) for more details.

[4]Estimation methods such as MCMC and particle filtering are computationally expensive and usually require the discretization of the models, which may lead to discretization errors. In addition, extracted proxies usually contain errors. The effect of these errors on parameter estimation can be complicated and is usually not properly investigated.

VXFs.

Although TDs have proved to be quite usefully in empirical applications by achieving a suitable balance between model flexibility and tractability, the literature on TDs has so far mainly focused on one-factor models, which inevitably suffer from some degrees of restrictions. For modelling the term structure of interest rates, for example, empirical evidence often finds one-factor models not fitting short term and long term yields satisfactorily at the same time. In addition, Bu et al. (2011) and Bu, Cheng and Hadri (2017) both find that TDs with time-varying transformations fit their data significantly better than TDs with constant transformations. Moreover, Forman and Sørensen (2014) find their distribution-driven TD model does not fit the autocorrelation structure of their protein unfolding data satisfactorily and subsequently estimate their model with an measurement error. All these evidence suggest that there is a clear and urgent need for introducing additional factors in order to make the TD framework empirically more flexible.

The main contributions of this paper are four-fold: First, we provide a review of existing parametric TD models, which is dominated by one-factor models. Second, we propose a novel distribution-driven nonlinear multi-factor TD model with latent components. While our approach is applicable to general multi-factor UDs, we propose more specifically a model that is the transformation of a multivariate Ornstein Uhlenbeck (MVOU) process with latent components, where the transformation function is endogenously determined by a flexible parametric specification of the stationary distribution of the observed variable. We show that exact ML inference for our model can be made efficiently by a modified Kalman filter algorithm. Third, exploiting the underlying affine structure, we derive a semi-closed form expression for the VXF price based on our latent-factor TD model. Finally, we examine the empirical performance of a two-factor specification of the proposed model in comparison with existing models for modelling the dynamics of VIX and for pricing VXF contracts. We base our comparison on both the in-sample model fitness criteria and the out-of-sample Root Mean Square Forecasting Error (RMSFE) for modelling VIX and on both in- and out-of-sample Root Mean Square Pricing Error (RMSPE) for pricing VXFs. Our results strongly favor our distribution-driven two-factor model, which outperforms all alternative TD models strongly and consistently across all the categories and scenarios of our comparison. Several specification tests applied to competing models also reveals the proposed two-factor model as the only model not rejected by the data.

The rest of the paper is organized as follows: In Section 2, we outline the TD framework and review existing one-factor TD models. In Section 3, we propose a multi-factor distribution-driven TD model with latent components and explains our modified Kalman filter procedure. Section 4 discusses the pricing of VXF contracts and our joint-measure estimation strategy. In Section 5, we empirically examine the performance of the proposed model in comparison with existing TDs for modelling VIX and pricing VXF contracts against several empirically relevant criteria. Some concluding remarks are included in Section 6.

2. Transformed Diffusion Models

2.1. The Framework

The TD approach assumes that the observed diffusion process Y is a strictly monotone and sufficiently smooth function[5] of some UD X. More specifically, it assumes that

$$\begin{aligned} Y_t &= V(X_t; \vartheta) & (2) \\ dX_t &= \mu_X(X_t; \omega)\, dt + \sigma_X(X_t; \omega)\, dW_t & (3) \end{aligned}$$

where $\mu_X(X_t; \omega)$ and $\sigma_X^2(X_t; \omega)$ are the drift and diffusion functions of X with parameter ω. $V(x; \vartheta)$ or equivalently its unique inverse $U(y; \vartheta) = V^{-1}(y; \vartheta)$ is known as the transformation function with parameter ϑ, satisfying $\partial V(x; \vartheta)/\partial x \neq 0$ for all x on its domain D_X.

Most importantly, both X and Y are assumed to satisfy the regularity conditions set out in, for example, Aït-Sahalia (1996b, Assumption A1)[6], so that they both admit a unique weak solution defined by their transition density p_X and p_Y, respectively. In addition, both X and Y are often assumed to be stationary with stationary distribution (cdf) F_X and F_Y and corresponding stationary density (pdf) f_X and f_Y, respectively. Under the assumption that $V(x; \vartheta)$ is strictly monotone and twice continuously differentiable, Ito's Lemma implies that the drift and diffusion functions of Y can be written as

$$\begin{aligned} \mu_Y(y; \psi) &= \frac{\mu_X(U(y; \vartheta); \omega)}{U'(y; \vartheta)} - \frac{\sigma_X^2(U(y; \vartheta); \omega)\, U''(y; \vartheta)}{2U'(y; \vartheta)^3} & (4) \\ \sigma_Y^2(y; \psi) &= \frac{\sigma_X^2(U(y; \vartheta); \omega)}{U'(y; \vartheta)^2} & (5) \end{aligned}$$

where $U'(y; \vartheta)$ and $U''(y; \vartheta)$ are its first two derivatives of $U(y; \vartheta)$. More specifically, let $p_X(x|x_0, \Delta; \omega)$ and $p_Y(y|y_0, \Delta; \psi)$ be the transition density function of X and Y, respectively, where Δ is the time interval. It follows immediately that

$$p_Y(y|y_0, \Delta; \psi) = |U'(y; \vartheta)|\, p_X(U(y; \vartheta)\, | U(y_0; \vartheta), \Delta; \omega)$$

Clearly, the specification of TD models has two components: the specification of the UD X in (3) and the specification of the transformation function V in (2). Since in practice X is often preferred to be tractable for inferential and derivative pricing purposes, the model in (3) is often, but not necessary, chosen from the affine class. The literature of TDs has so far mainly focused on one-factor models. Hence, unsurprisingly the CIR and the OU processes have been the most frequently used UDs. Specifically, the CIR process, which is usually defined on $D_X \in (0, \infty)$, is written as

$$dX_t = \kappa(\theta - X_t)dt + \sigma\sqrt{X_t}dW_t$$

satisfying the conditions $\kappa > 0$, $\theta > 0$ and $2\kappa\theta \geq \sigma^2 > 0$. The OU process defined on $D_X \in (-\infty, \infty)$ is given by

$$dX_t = \kappa(\theta - X_t)dt + \sigma dW_t$$

[5] The strict monotonicity condition ensures the invertibility of the transformation function and the smoothness condition ensures its differentiability so that necessary assumptions underlying the Ito's Lemma hold.

[6] See, for example, Theorem 5.15 in Chapter 5 in Karatzas and Shreve (1991) for more details.

satisfying the conditions $\kappa > 0$ and $\sigma^2 > 0$. For the CIR process, conditional on X_0, the random value of X after a time interval Δ has a non-central χ^2 distribution with fractional degrees of freedom, and in the limit as $\Delta \to \infty$, the stationary distribution of X is Gamma. For the OU process, both the conditional and stationary distributions are Gaussian.[7]

2.2. Specifications of Transformation Functions

For a given specification of the UD X, the specification of the TD Y will be completely determined by the specification of the transformation function V or equivalently but more conveniently U. Hence, most of the efforts in TD modelling have been focused on the specification of U. For one-factor TD models, Bu et al. (2011) show that for a given specification of X, the parametric form of either the drift μ_Y or the diffusion σ_Y^2 leads to the unique solution of U. In addition, Bu, Jawadi and Li (2017) show that under the stationarity assumption, U can also be uniquely determined from the stationary distributions F_X and F_Y when X is suitably normalized. Bu et al. (2018) formally discuss the identification of TD models. In what follows, we give a summary of existing one-factor TD models, focusing on their specification strategies.

2.2.1. TDs with Polynomial Drift Function

One class of TD models are specified to have a flexible polynomial drift function μ_Y. Bu, Jawadi and Li (2017) define this class of TD models as drift-driven. Note from (4) and (5) that for a given specification of X and that of μ_Y, the transformation function U is the solution to a system of 2nd-order Ordinary Differential Equations (ODEs), which depend on both μ_X and σ_X^2. For an arbitrary choice of μ_Y, closed-form solution for U is usually unavailable.

However, a special case arises when X is the CIR process and μ_Y is a cubic polynomial. Eraker and Wang (2015) propose a simple TD model, denoted as CIREW in this paper, which assumes that

$$U(y;\vartheta) = 1/(y-\delta) - \eta$$

It can be easily verified that

$$\begin{aligned}
\mu_Y(y;\psi) &= \kappa(y-\delta) + \left[\sigma^2 - \kappa(\theta+\eta)\right](y-\delta)^2 - \eta(y-\delta)^3 \\
\sigma_Y^2(y;\psi) &= \sigma^2 (y-\delta)^4 \left[(y-\delta)^{-1} - \eta\right]
\end{aligned}$$

where $\mu_Y(y;\psi)$ is a cubic polynomial if $\eta \neq 0$. Since $D_X = (0,\infty)$, the CIREW model has a bounded support on $D_Y = (\delta, 1/\eta + \delta)$. Note that the model of Ahn and Gao (1999) (CIRAG) for interest rates, also known as the 3/2 model by Goard and Mazur (2013) for VIX, arises as a special case of the CIREW model when $\delta = \eta = 0$, implying that $x = U(y) = 1/y$. Consequently, the CIRAG model has a quadratic drift function and a Constant Elasticity Variance (CEV) diffusion function $\sigma_Y^2(y;\psi) = \sigma^2 y^3$.

2.2.2. TDs with Constant Elasticity Variance

Another class of TD models begin with a desired diffusion function σ_Y^2. Bu, Jawadi and Li (2017) define this class of TD models as diffusion-driven. In this case, U is the solution to the 1st-order

[7] Grunbichler and Longstaff (1996) and Detemple and Osakwe (2000) consider the CIR model and the exponential transform of the OU model, denoted as the OUDO model in this paper, respectively, for pricing VIX derivatives.

ODE in (5). Solving (5) is relatively simple and in some cases closed-form solutions exist. An important specification frequently used in financial modelling is the following CEV diffusion

$$\sigma_Y^2(y;\psi) = \sigma^2 y^{2\gamma} \text{ for } \gamma \in [0,+\infty)$$

It was introduced by Chan et al. (1992) who considered a linear drift and subsequently studied by Aït-Sahalia (1996b) who promoted a nonlinear drift to improve the mean reversion. Bu et al. (2011) propose a TD model with CEV diffusion function where X is the CIR process, denoted as CIRCEV. They show that the required transformation is given by

$$U(y;\vartheta) = \begin{cases} \left[y^{1-\gamma}/(1-\gamma)\right]^2/4 & \text{for} \quad \gamma \in [0,1) \cup (1,+\infty) \\ (\log y)^2/4 & \text{for} \qquad \gamma = 1 \end{cases}$$

and the resulting drift function given by

$$\mu_Y(y;\psi) = \left[2\kappa\theta(1-\gamma) + \sigma^2(2\gamma-1)/2\right] y^{2\gamma-1} + \kappa y/(2\gamma-2)$$

In general, both the drift and the diffusion functions of the CIRCEV model are nonlinear. In addition to the CEV diffusion function, the drift function exhibits a much stronger pull at high levels of the state variable than the linear drift. Both properties are consistent with empirical findings about the two functions reported in, for example, Aït-Sahalia (1996a,b), Conley et al. (1997), Stanton (1997) and others. The CIRCEV model reduces to the CIR model when $\gamma = 0.5$ and to the CIRAG model when $\gamma = 1.5$. The CIRCEV model is clearly more general, since it can provide a varied degree of nonlinearity in both the drift and the diffusion functions by the data-driven choice of γ.[8] More importantly, in practice it is often necessary to test the validity of linear or affine constraints whenever a more general nonlinear alternative model can be specified. Although the CIREW model is the transformation of the CIR model, it does not nest the CIR model. Consequently, Eraker and Wang (2015) used a parametric bootstrap method to test the CIR model against the CIREW. In contrast, the CIRCEV model has the advantages that it strictly nests the CIR model as well as the CIRAG model. Therefore, testing the two models against the CIRCEV model only requires a standard procedure such as the nested Likelihood Ratio (LR) test. In addition, Bu et al. (2011) show that both the conditional and the unconditional moments implied by the CIRCEV model are available in closed form. This may be another important advantage over the CIREW model, since conditional mean forecasting and futures pricing based on the CIRCEV model are consequently extremely convenient.

2.2.3. TDs with Flexible Stationary Distribution

The third class of TD models are specified to have a flexible stationary distribution F_Y. Bu, Jawadi and Li (2017) define this class of TD models as distribution-driven. Under the stationarity assumption, the stationary distributions of both Y and X exist. Assuming that U is strictly increasing, we have

$$F_Y(y;\vartheta) = F_X\left[U(y;\vartheta);\omega\right]$$

[8] Bu, Cheng and Hadri (2016, 2017) extend the CIRCEV model by allowing γ and hence the transformation function to be time-varying.

which then implies that

$$U(y;\vartheta) = F_X^{-1}\left[F_Y(y;\vartheta);\omega\right] \tag{6}$$

The first two derivatives of $U(y;\vartheta)$ can be written as

$$U'(y;\vartheta) = \frac{f_Y(y;\vartheta)}{f_X\{U(y;\vartheta);\omega\}} \tag{7}$$

$$U''(y;\vartheta) = \frac{f'_Y(y;\vartheta)}{f_X\{U(y;\vartheta);\omega\}} - \frac{f'_X\{U(y;\vartheta);\omega\}}{f_Y(y;\vartheta)}U'(y;\vartheta)^3 \tag{8}$$

where $f_Y(y;\vartheta)$ and $f_X\{x;\omega\}$ are the stationary densities of Y and X, respectively, and $f'_Y(y;\vartheta) = \partial f_Y(y;\vartheta)/\partial y$ and $f'_X\{x;\omega\} = \partial f_X\{x;\omega\}/\partial x$. It follows that the transition density of Y is given by

$$p_Y(y|y_0,\Delta;\omega,\vartheta) = \frac{f_Y(y;\vartheta)}{f_X\{U(y;\vartheta);\omega\}}p_X(U(y;\vartheta)|U(y_0;\vartheta),\Delta;\omega) \tag{9}$$

and closed-form drift and diffusion functions can be obtained but plugging (6), (7) and (8) into (4) and (5).

It is important to point out that under the assumptions set out for diffusions X and Y in Section 2.1, the Jacobian of the the transformation (7) is continuous and non-negative on D_Y, which ensures the strict monotonicity of the transformation (6). Meanwhile, when the UD X satisfies the normalization conditions set out in Bu et al. (2018), the distribution-driven TD Y is uniquely identified.

This specification strategy was considered by Forman and Sørensen (2014) for modelling molecular dynamics, where they assume that X follows the OU process and the stationary distribution of Y is a mixture of two normal distributions. The motivation behind this specification is that the stationary distribution of their protein folding data exhibits bimodality, and existing models failed to model this feature adequately. Bu, Jawadi and Li (2017) considered transformed OU and CIR models with a SKST stationary distribution for modelling VIX and pricing VXFs.

3. A Multi-Factor Transformed Diffusion Model

Although TD models have proved to be quite useful for financial modelling, existing TDs are predominantly one-factor models. Meanwhile, plenty of empirical evidence finds one-factor models not flexible enough to model more complicated dynamics. Therefore, there is often a strong need for multi-factor models with latent components, particularly in areas such as term structure modelling and derivative pricing. For this reason, we consider extending the TD approach to the multi-factor case. We choose to present our modelling strategy in the context of a two-factor model. On one hand, this facilitates our exposition. On the other hand, both the VIX and VXF data in our empirical application indicate that a two-factor model is quite reasonable for our modelling objectives[9].

[9] In other financial contexts, such as modelling term structure of interest rates, it is often necessary to consider three-factor models. See for example Litterman and Scheinkaman (1991) and Balduzzi et al. (1996).

3.1. TDs with a Latent Factor

Suppose that we wish to model a diffusion process Y, assuming that $Y_t = V(X_t; \vartheta)$. Crucially, we now assume that the SDE of X can be written as

$$dX_t = \mu_X (X_t; \theta_t, \omega)\, dt + \sigma_X (X_t; \theta_t, \omega)\, dW_{X,t} \tag{10}$$
$$d\theta_t = \mu_\theta (\theta_t; X_t, \omega)\, dt + \sigma_\theta (\theta_t; X_t, \omega)\, dW_{\theta,t} \tag{11}$$

where θ_t is a latent process. The two-dimensional vector $Z_t = (X_t, \theta_t)^T$ follows a bivariate diffusion system with parameter ω. It is important to assume that the bivariate diffusion Z satisfy the regularity conditions set out in Aït-Sahalia (2008, Assumptions 1-4) which ensure that Z admits a unique weak solution in terms of the bivariate transition density $p_Z (z|z_0, \Delta; \omega)$. It then follows that continuous-time dynamics and the transition density of the transformed system $\tilde{Z}_t = (Y_t, \theta_t)^T = (V(X_t), \theta_t)^T$ can be obtained by the multivariate version of the Ito's Lemma and the usual Jacobian method, respectively.

3.2. Bivariate OU Process

While the dynamics of X written in (10) and (11) are general, the TD framework usually requires the UD process to be tractable. In the multi-factor setting, it is also preferable that the system of $Z_t = (X_t, \theta_t)^T$ is such that the marginal process X_t is tractable. For this reason, we consider the following Bivariate OU (BVOU) process as a preferred candidate in this paper. The continuous-time dynamics of the BVOU process can be written in a vector SDE as

$$dZ_t = \boldsymbol{\beta} (\alpha - Z_t)\, dt + \boldsymbol{\sigma} dW_t$$

where

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_{XX} & \beta_{X\theta} \\ \beta_{\theta X} & \beta_{\theta\theta} \end{pmatrix}, \alpha = \begin{pmatrix} \alpha_X \\ \alpha_\theta \end{pmatrix}, \boldsymbol{\sigma} = \begin{pmatrix} \sigma_X & 0 \\ \rho\sigma_\theta & \sqrt{1-\rho^2}\sigma_\theta \end{pmatrix}, dW_t = \begin{pmatrix} dW_{X,t} \\ dW_{\theta,t} \end{pmatrix}$$

and $W_{X,t}$ and $W_{\theta,t}$ are two independent Brownian motions. The parameters of this model can be summarized as $\omega = (\boldsymbol{\beta}, \alpha, \boldsymbol{\sigma})^T$.

The identification of the parameters for this particular model from discrete data is discussed in, for example, Philips (1973), Hansen and Sargent (1983) and Kessler and Rahbek (2004), and Aït-Sahalia (2008). In particular, Aït-Sahalia (2008) show that by imposing $\boldsymbol{\beta}$ to be triangular, the bivariate diffusion Z is stationary and the following matrix equation $\boldsymbol{\beta\lambda} + \boldsymbol{\lambda\beta}^T = \boldsymbol{\sigma\sigma}^T$ admits a unique solution for $\boldsymbol{\lambda}$, which is the 2×2 symmetric matrix $\boldsymbol{\lambda}$ given by

$$\boldsymbol{\lambda} = \frac{1}{2\mathrm{tr}[\boldsymbol{\beta}]\, \mathrm{Det}[\boldsymbol{\beta}]} \left(\mathrm{Det}[\boldsymbol{\beta}]\, \boldsymbol{\sigma\sigma}^T + (\boldsymbol{\beta} - \mathrm{tr}[\boldsymbol{\beta}]\, \mathbf{I})\, \boldsymbol{\sigma\sigma}^T (\boldsymbol{\beta} - \mathrm{tr}[\boldsymbol{\beta}]\, \mathbf{I})^T \right)$$

where $\mathbf{I}$ is the two-dimensional identity matrix. It then follows that the transition density of the BVOU system is bivariate normal with

$$\begin{aligned} p_{\mathbf{Z}} (z|z_0, \Delta; \omega) &= (2\pi)^{-1} \mathrm{Det}[\boldsymbol{\Omega}(\Delta)]^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} [z - m(\Delta, z_0)]^T \boldsymbol{\Omega}(\Delta)^{-1} [z - m(\Delta, z_0)] \right\} \\ \boldsymbol{\Omega}(\Delta; \omega) &= \boldsymbol{\lambda} - \exp\{-\boldsymbol{\beta}\Delta\}\, \boldsymbol{\lambda} \exp\{-\boldsymbol{\beta}^T \Delta\} \\ m(\Delta, z_0; \omega) &= \alpha + \exp\{-\boldsymbol{\beta}\Delta\} (z_0 - \alpha) \end{aligned}$$

where Δ is time interval between z_0 and z. The stationary bivariate distribution of Z can be obtained by taking the limit as $\Delta \to \infty$, which is given by

$$f_Z(z;\omega) = (2\pi)^{-1} \operatorname{Det}[\boldsymbol{\lambda}]^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}[z-\alpha]^T \boldsymbol{\lambda}^{-1}[z-\alpha]\right\}$$

Clearly, the stationary marginal distributions of X and θ are both normal.

3.3. Distribution-Driven Transformation

Bu et al. (2011) argue that the distribution-driven approach for specifying the transformation function U has a clear advantage over other approaches, since researchers can directly and purposefully specify a suitable parametric density $f_Y(y;\vartheta)$ to incorporate stationary (long-run) distributional information of the data. In addition, there is an enormous literature on density specification and estimation in statistics. Moreover, we can see from (7) and (9) that any closed-form $f_Y(y;\vartheta)$ can lead to a closed-form transformation function $U(y;\vartheta)$ and transition density $p_Y(y|y_0,\Delta;\omega,\vartheta)$.

For this reason, as in Section 2.2.3 we assume in this paper that the stationary distribution of the observed diffusion Y has a flexible parametric cdf $F_Y(y;\vartheta)$ and pdf $f_Y(y;\vartheta)$. Meanwhile, let $F_X(x;\omega)$ and $f_X(x;\omega)$ denote, respectively, the stationary marginal cdf and pdf of X in the bivariate system $Z=(X,\theta)^T$, and $f'_X(x;\omega)$ the derivative of $f_X(x;\omega)$. Under the maintained assumption $Y_t = V(X_t;\vartheta)$, we arrive at exactly the same expressions for the distribution-driven transformation $U(y;\vartheta)$ and its derivatives as in (6), (7) and (8), respectively. Therefore, similar to the one-factor case, the Jacobian of the transformation (7) is also guaranteed to be continuous and non-negative, ensuring one-to-one mapping between Y and X.

By the same argument in Bu et al. (2018), the bivariate system also needs to be suitably normalized to ensure the identification of ω. More specifically, in the BVOU case, the stationary marginal distribution of X is normal, which is completely specified by its mean and variance. Therefore, identical to the one-factor transformed OU model, the parameters representing the mean and variance of the stationary marginal distribution of X cannot be uniquely identified and thus must be fixed. Obviously, it is most convenient to normalize the system $Z=(X,\theta)^T$ so that the stationary marginal distribution of X is standard normal. Mathematically, this can be achieved by imposing

$$\alpha_X = 0 \text{ and } s^T \boldsymbol{\lambda} s = 1$$

where $s=(1,0)^T$ is the selection vector such that the up-left element of $s^T\boldsymbol{\lambda}s$ equals to one. Since the normalized X is symmetric around zero, there is no loss of generality to assume that the transformation is strictly increasing. Given the specification of $f_Y(y;\vartheta)$ and the normalized system $(X,\theta)^T$, the bivariate transition density for the system $\tilde{Z}=(Y,\theta)^T$ is given by

$$p_{\tilde{Z}}(y,\theta|y_0,\theta_0,\Delta;\omega;\vartheta) = U'(y;\vartheta)\, p_Z[U(y;\vartheta),\theta|U(y_0;\vartheta),\theta_0,\Delta;\omega]$$

Finally, since θ_t is a latent process, the identification of the parameters of the system follows from the fact that the corresponding discrete-time dynamics of Z follows the Gaussian Linear State-Space Model (LSSM). The conditions for the identification of LSSM systems are discussed in, for example, Hamilton (1994, Section 13.4) and references therein.

3.4. Specification of the Stationary Distribution

Our modelling strategy requires us to specify a parametric stationary density function f_Y of Y, which then implies $F_Y(y;\vartheta)$. Most financial data exhibit skewness, fat tails (c.f. Hansen 1994) and possibly multi-modality that linear diffusions are usually unable to produce (c.f. Aït-Sahalia 1996b, Fernandes 2006, Bu and Hadri 2007). The distribution-driven modelling strategy provides full flexibility in the specification of stationary distributions without sacrificing the dynamic structure. Forman and Sørensen (2014) used the mixture of two normal distributions for their protein unfolding data, and Bu, Jawadi and Li (2017) consider the SKST distribution of Hansen (1994) for the VIX data. Meanwhile, the formulation and subsequent estimation of distribution-driven TD models depend directly on the stationary pdf f_Y and cdf F_Y of Y, it is computationally advantageous to choose a distribution for which f_Y is in closed form and the cdf F_Y is easy to evaluate.

As in Forman and Sørensen (2014), we also promote the use of the mixture of distributions. Specifically, we propose to use a mixture of two lognormal (M2LN) distributions for Y on $D_Y \in (0,\infty)$. The pdf of a M2LN random variable Y can be written as

$$f_{Y,M2LN}(y|w,m_1,s_1,m_2,s_2) = wf_{Y,LN}(y|m_1,s_1) + (1-w)\,f_{Y,LN}(y|m_2,s_2)$$

where $f_{Y,LN}(\cdot)$ denotes the lognormal pdf, and (m_1,s_1) and (m_2,s_2) are the log means and log standard deviations of the two component distributions with $w \in [0,1]$ being the component weight. The corresponding cdf can be written as

$$F_{Y,M2LN}(y|w,m_1,s_1,m_2,s_2) = wF_{Y,LN}(y|m_1,s_1) + (1-w)\,F_{Y,LN}(y|m_2,s_2)$$

where $F_{Y,LN}(\cdot)$ denotes the lognormal cdf[10]. The M2LN distribution, with the above closed-form pdf and cdf, is also well known for its ability to generate large skewness, excess kurtosis, and as many as (but not necessarily) two modes, making it an attractive candidate distribution for modelling financial data with positive support such as interest rates, exchange rates, and volatility indices. In fact, the mixture of lognormal or normal distributions have been frequently used in statistical modelling. See for example Bahra (1997) and Melick and Thomas (1997), Söderlind and Svensson (1997), Brigo and Mercurio (2002), Forman and Sørensen (2014), and others.

That being said, distribution theories offer numerous potentially attractive candidate distributions that may also be considered for modelling financial variables. For example, in addition to the SKST distribution, the Generalized Exponential Distribution (GED) of Cobb et al. (1983) and Lye and Martin (1993) is an obvious alternative for empirical applications. The GED has the advantage of being functionally flexible, nesting many standard distributions such as the normal, the student-t, the lognormal, the Gamma, and so on, as special cases. It also has tractable recursive moment functions, making for example moment-based inference very convenient. Fernandes (2006) used a special case of GED, known as the Generalized Normal Distribution (GND) to model financial crashes. In fact, it is important to emphasize that the ability to allow for abundant choices of distribution functions for the stationary distribution of the data is the main

[10] In fact, modelling Y with a M2LN distribution is equivalent to modelling $\ln Y$ using a mixture of two normal distributions (M2N).

advantage of the distribution-driven TD modelling approach.[11]

When the underlying multi-factor diffusion process X is the (normalized) BVOU and the stationary distribution of Y is the M2LN, the resulting distribution-driven two-factor TD is denoted as the BVOUM2LN model in this paper. More specifically, for the BVOUM2LN model, we have

$$\begin{aligned} U(y;\vartheta) &= \Phi^{-1}\left[F_{Y,M2LN}(y;\vartheta)\right] \\ U'(y;\vartheta) &= \frac{f_Y(y;\vartheta)}{\phi\{U(y;\vartheta)\}} \end{aligned}$$

where $\Phi(\cdot)$ and $\phi\{\cdot\}$ are the standard normal cdf and pdf, respectively, and $\vartheta = (w, m_1, s_1, m_2, s_2)^T$. The complete specification of the continuous-time dynamics of the BVOUM2LN model is given Section 5.2.2.

3.5. Modified Kalman Filter Estimation

Since the newly proposed multi-factor model contains latent factors, certain filtering techniques must be used for estimating the model parameters. Instead of having to use sophisticated and computational expensive filtering algorithms such as the particle filter (e.g. Song and Xiu 2016) or the Bayesian MCMC method (e.g. Eraker 2001), the transformation structure and the tractability of the underlying MVOU system allows us to simply rely on the standard Kalman filter algorithm to evaluate the exact likelihood function of our model. Specifically, given observations of Y and the parametric specification of $F_Y(y;\vartheta)$, we can obtain parameter-dependent observations of X through the transformation as

$$X_t = U(Y_t;\vartheta) = \Phi^{-1}\left[F_{Y,M2LN}(Y_t;\vartheta)\right]$$

Since the discrete observations of X and θ form a bivariate linear Gaussian system, we can cast the discrete-time version of our BVOU system $(X,\theta)^T$ into the standard linear state-space form with parameter ω. Subsequently, using the standard Kalman filter algorithm, we can obtain the filtered density function for X as $p_X(x_t;\omega|I_t)$ for each observation of $X_t = U(Y_t;\vartheta)$ where I_t is the information set up to time t. The corresponding filtered density function for each observation of Y can then be obtained by the Jacobian method as

$$p_Y(y_t;\vartheta,\omega|I_t) = \frac{f_Y(y_t;\vartheta)}{\phi\{U(y_t;\vartheta)\}} p_X(U(y_t;\vartheta);\omega|I_t)$$

We refer this procedure as the our modified Kalman filter algorithm. Clearly, this procedure is computationally inexpensive compared to more sophisticated algorithms for general nonlinear state-space models. It can be easily verified that the ML likelihood estimator of the parameters of our multi-factor TD model with latent components obtained from our procedure is asymptotically normal with variance equal to the inverse of the information matrix.

[11] In our preliminary exercise, we also tried the GND for modelling the stationary distribution of the log of VIX. We found that the results are similar to those from using the M2LN distribution. However, the latter has closed-form pdf and cdf, whereas those of the former requires numerical integrations. We therefore choose to adopt the M2LN specification in our applications, as it significantly facilitates the implementation of our distribution-driven models, particularly in our bootstrap-based specification tests where 1000 replications are considered for each competing model.

4. VXF Pricing and Model Estimations

4.1. VXF Pricing

In the absence of arbitrage opportunities in a complete market, the price of a VXF contract is the conditional expectation of the value of VIX under the unique risk-neutral Martingale measure. This unique risk-neutral $\mathbb{Q}$-measure corresponding to the observed physical $\mathbb{P}$-measure can be established by applying Girsanov's theorem. Specifically, let $\Lambda\left(Y_t;\varphi\right)$ be a parametric function specifying the Market Price of Risk (MPR) with respect to the Brownian motion, where φ is the MPR parameter vector. For a diffusion process Y with $\mathbb{P}$-measure dynamics given by (1), the equivalent risk-neutral $\mathbb{Q}$-measure can be written as

$$dY_t = \left[\mu_Y\left(Y_t;\psi\right) - \Lambda\left(Y_t;\varphi\right)\sigma_Y\left(Y_t;\psi\right)\right]dt + \sigma_Y\left(Y_t;\psi\right)dW_t^{\mathbb{Q}} \tag{12}$$

Following the convention, the parametric specification of Y is assumed to be the same under both measures. Note that for TDs, the specification in (12) is determined jointly by the specification of X and $U\left(y;\vartheta\right)$. Thus, in order for $\sigma_Y^2\left(Y_t;\psi\right)$ to be identical under both measures, the parameters in the diffusion function of X and $U\left(y;\vartheta\right)$ must also be identical under both measures. Consequently, the diffusion function $\sigma_Y^2\left(Y_t;\psi\right)$ remains the same under both measures, but the drift parameters will differ under the two measures.

Define ω^Q as the parameter of X under the $\mathbb{Q}$-measure. Then, at time t the price of a VXF contract with time to maturity τ is simply the time-t conditional expectation of the value of VIX at maturity date $t+\tau$ under the $\mathbb{Q}$-measure, i.e.,

$$F\left(y_t,\theta_t,\tau;\omega^Q,\vartheta\right) = \mathbb{E}_t^Q\left[y|y_t,\theta_t,\tau;\omega^Q,\vartheta\right] = \int_0^\infty y p_Y\left(y|y_t,\theta_t,\tau;\omega^Q,\vartheta\right)dy \tag{13}$$

Since the risk-neutral conditional marginal density $p_Y\left(y|y_0,\theta_0,\Delta;\omega^Q,\vartheta\right)$ is in closed-form, the pricing formula in (13) for our multi-factor model only involves a 1-dimensional numerical integration. Meanwhile, it is important to note that θ_t is latent. Following the literature, we use the filtered value of θ_t, i.e., $\theta_{t|t} = E\left(\theta_t|I_t\right)$, in our application.

4.2. Model Estimations

Let $\{Y_{i\Delta}, i = 0,1,...,n_Y\}$ be a sample of VIX data, where Δ is the sampling interval. Define ω as the parameter of X under the $\mathbb{P}$-measure. Then, the log-likelihood (LL) function under the physical measure is given by

$$LL_Y\left(\omega,\vartheta\right) = \sum_{i=1}^{n_Y}\ln p_Y\left(Y_{i\Delta}|Y_{(i-1)\Delta},\Delta;\omega,\vartheta\right)$$

Meanwhile, let $\{F_j\left(\tau,Y_t\right), j = 1,2,...,n_F\}$ be a sample of VXF prices and assume that the VXF pricing error has the following distribution

$$\varepsilon_j(\omega^Q,\vartheta) = F_j\left(\tau,Y_t\right) - F_j\left(Y_t,\theta_{t|t},\tau;\omega^Q,\vartheta\right) \sim N\left(0,\sigma_F^2\right)$$

As such, we can profile out the parameter σ_{VXF} as

$$\hat{\sigma}_F = \sqrt{\frac{1}{n_F}\sum_{j=1}^{n_F}\left[F_j\left(Y_t,\tau\right) - F_j\left(Y_t,\theta_{t|t},\tau;\omega^Q,\vartheta\right)\right]^2}$$

which is actually the Root Mean Square Pricing Error (RMSPE) of our VXF pricing model. It follows that the profile LL for VXF data can be written as

$$LL_F\left(\omega^Q, \vartheta\right) = \sum_{j=1}^{n_F} \ln\left[\frac{1}{\hat{\sigma}_F}\phi\left(\frac{\varepsilon_j(\omega^Q, \vartheta)}{\hat{\sigma}_F}\right)\right]$$

where again $\phi(\cdot)$ is the standard normal pdf. Finally, the joint LL for a combination of VIX and VXF data can be written as the following sum

$$LL_{Total}\left(\omega, \omega^Q, \vartheta\right) = LL_Y\left(\omega, \vartheta\right) + LL_F\left(\omega^Q, \vartheta\right)$$

5. Empirical Comparison

5.1. The Data

We compare the empirical performance of the newly proposed distribution-driven multi-factor TD with latent component model with existing TD models for modelling the dynamics of VIX and pricing VXFs. Our data consist of daily VIX indices from January 2, 1990 to March 20, 2015 (6352 observations) and VXF closing prices from March 26, 2004 to February 17, 2015 (19215 observations). Following Eraker and Wang (2015), we construct seven series of daily constant maturity $(1, 2, 3, 4, 5, 6, 7$ month) VXF prices by linear interpolation, each containing 2742 observations. We use data up to December 31, 2013 for in-sample calibration and comparison and the remaining data for out-of-sample comparison.

We plot the time series of daily VIX and the term structure of constant maturity VXF prices in Figure 1 and 2, respectively, and some summary statistics are reported in Table 1. The evolution of VIX indicates that the mean reversion is weak when the level of VIX is low but much stronger when it is high. This suggests that a suitable diffusion model for VIX should have a drift function that is close to zero when VIX is low and strongly negative when VIX is high. Meanwhile, the local volatility of VIX is also low when VIX is low and substantially higher otherwise. This suggests that a suitable diffusion model should also have a diffusion function that increases rapidly in VIX. The mean of VIX is 20.61 and the standard deviation is 10.19. The large skewness 2.21 and kurtosis 9.25 suggest strong deviation from normality. Augmented Dickey-Fuller tests on these time series all rejected the unit root hypothesis with 4 lags at 5% significance level. Therefore, the use of stationary diffusion models is justified. More importantly, Mencía and Sentana (2013) show that the daily VIX series exhibits the ARMA(2,1) autocorrelation structure[12]. Thus, the use of our proposed two-factor TD model is justified, since it can be easily verified that it implies the ARMA(2,1) structure. Meanwhile, the term structure of VXFs is relatively flat and the evolutionary paths of the seven series are highly correlated. The first two eigenvalues of the correlation matrix dominate the others, explaining approximately 99.9% of the cross sectional variation in these series. This is further justification for the use of a two-factor models for pricing VXFs.

[Figure 1 and 2]
[Table 1]

[12] See Figure 2 of Mencía and Sentana (2013) for more details. Although our sample period is longer, we find the same ARMA(2,1) structure for our VIX data.

5.2. Competing Models

5.2.1. One-Factor Models

A total of eight models are considered in our empirical comparison, six of which are one-factor models and the other two are two-factor models. Four of the six one-factor models are transformed CIR models. They are: the benchmark CIR model, the drift-driven CIREW model, the diffusion-driven CIRCEV model, and finally the distribution-driven CIRM2LN model, which denotes the transformed CIR model with M2LN stationary distribution. Specifically, the CIRM2LN model can be written as

$$\begin{aligned} Y_t &= F_{Y,M2LN}^{-1}\left[\Gamma\left(X_t; 2\kappa\theta, 1/\left(2\kappa\right)\right); \vartheta\right] \\ dX_t &= \kappa\left(\theta - X_t\right)dt + \sqrt{X_t}dW_t \end{aligned}$$

where $\vartheta = (w, m_1, s_1, m_2, s_2)^T$ and $\Gamma\left(\cdot; 2\kappa\theta, 1/\left(2\kappa\right)\right)$ is the cdf of the Gamma distribution with shape parameter $2\kappa\theta$ and scale parameter $1/\left(2\kappa\right)$. The remaining two one-factor models are transformed OU models. One is the benchmark exponential-transformed OU model of Detemple and Osakwe (2000), denoted as the OUDO model, which can be written as

$$\begin{aligned} Y_t &= \exp\left(X_t\right) \\ dX_t &= \kappa\left(\theta - X_t\right)dt + \sigma dW_t \end{aligned}$$

The other is the distribution-driven OUM2LN model, which denotes the transformed OU model with M2LN stationary distribution. Specifically, the OUM2LN model can be written as

$$\begin{aligned} Y_t &= F_{Y,M2LN}^{-1}\left[\Phi\left(2\kappa X_t\right); \vartheta\right] \\ dX_t &= -\kappa X_t dt + dW_t \end{aligned}$$

The drift and diffusion functions of the CIRM2LN and the OUM2LN models can be obtained in closed form from (4), (5), (7), and (8), and their closed-form transition densities can be obtained from (9).

5.2.2. Two-Factor Models

We consider a couple of two-factor models in our empirical comparison. The first is the model considered by Mencía and Sentana (2013), which is included in our study as the two-factor benchmark model. Mencía and Sentana (2013) refers to this model as the Central Tendency OU model. This model effectively assumes that the transformation is exponential and can be written as

$$\begin{aligned} Y_t &= \exp\left(X_t\right) \\ dX_t &= \kappa\left(\theta_t - X_t\right)dt + \sigma dW_{X,t} \\ d\theta_t &= \kappa_\theta\left(\theta - \theta_t\right)dt + \sigma_\theta dW_{\theta,t} \end{aligned}$$

where $W_{X,t}$ and $W_{\theta,t}$ are independent. The purpose of the exponential transformation is to ensure $D_Y \in (0, \infty)$, but clearly it offers no additional degree-of-freedom. We refer to it as the BVOUMS model in this paper.

The main two-factor TD model to be investigated in our comparison is the newly proposed two-factor distribution-driven BVOUM2LN model. The BVOUM2LN model also assumes that X follows the above BVOU system, but crucially the transformation is parameter-dependent such that the stationary distribution of Y is the M2LN distribution, which allows us to model skewness, kurtosis and potential bimodality in the stationary distribution of our data with extra degrees of freedom. By construction, the domain of Y is on $D_Y \in (0, \infty)$, which is coherent for modelling and predicting volatility. This model can be written more specifically as

$$\begin{aligned} Y_t &= F^{-1}_{Y,M2LN}\left(\Phi\left(X_t\right);\vartheta\right) \\ dX_t &= \kappa\left(\theta_t - X_t\right)dt + \sigma dW_{X,t} \\ d\theta_t &= -\kappa_\theta \theta_t dt + \sigma_\theta dW_{\theta,t} \end{aligned}$$

where $W_{X,t}$ and $W_{\theta,t}$ are independent and the normalization constraint $\sigma_\theta = \sqrt{\left(2\kappa - \sigma^2\right)\left(\kappa + \kappa_\theta\right)\kappa_\theta/\kappa^2}$ is imposed. Moreover, we can write down the continuous-time dynamics in terms of SDE in closed form as

$$\begin{aligned} dY_t &= \left\{\frac{\kappa\left[\theta_t - U\left(Y_t;\vartheta\right)\right]}{U'\left(Y_t;\vartheta\right)} - \frac{\sigma^2 U''\left(Y_t;\vartheta\right)}{2U'\left(Y_t;\vartheta\right)^3}\right\}dt + \frac{\sigma}{U'\left(Y_t;\vartheta\right)}dW_{Y,t} \\ d\theta_t &= -\kappa_\theta\theta_t dt + \sqrt{\left(2\kappa - \sigma^2\right)\left(\kappa + \kappa_\theta\right)\kappa_\theta/\kappa^2}dW_{\theta,t} \end{aligned}$$

with

$$\begin{aligned} U\left(y;\vartheta\right) &= \Phi^{-1}\left[F_{Y,M2LN}\left(y;\vartheta\right)\right] \\ U'\left(y;\vartheta\right) &= \frac{f_{Y,M2LN}\left(y;\vartheta\right)}{\phi\left\{U\left(y;\vartheta\right)\right\}} \\ U''\left(y;\vartheta\right) &= \frac{f'_{Y,M2LN}\left(y;\vartheta\right)}{\phi\left\{U\left(y;\vartheta\right)\right\}} - \frac{\phi'\left\{U\left(y;\vartheta\right)\right\}}{f_{Y,M2LN}\left(y;\vartheta\right)}U'\left(y;\vartheta\right)^3 \end{aligned}$$

where $f'_{Y,M2LN}\left(y;\vartheta\right)$ and $\phi'\left(x\right)$ are the first derivatives of $f_{Y,M2LN}\left(y;\vartheta\right)$ and $\phi\left(x\right)$, respectively.

5.3. Analysis of Time Series of VIX

We first examine the performance of competing models for modelling the VIX time series. We investigate both the in-sample goodness-of-fit measure and out-of-sample forecasting accuracy as well as consider three specification tests. One of the main advantages of TDs is the availability of closed-form transition densities. Thus, ML is our preferred choice of estimation method. The ML estimates of the parameters of competing models are reported in the top panel Table 2. Our initial unconstrained estimation of the CIREW model resulted in a negative estimate of δ. Since δ is the lower bound of the support implied by the CIREW model, a negative δ is inconsistent with the nature of VIX. We then re-estimated the model by imposing $\delta = 0$. Meanwhile, η determines the upper bound of support. Therefore, no standard errors are reported for these two parameters. Moreover, when estimating the three distribution-driven models, we profiled out the parameters m_2 and s_2 of the M2LN distribution by matching the model-implied stationary mean and variance with the sample mean and variance of the VIX data[13]. Furthermore, for the BVOUM2LN model,

[13] This profiling method very effectively eliminates the possibility of the numerical optimization procedure converging to corner solutions without sacrificing any significant goodness-of-fit of the models to the data.

normalization requires the parameters θ and σ_θ to be constrained, and thus no standard errors are reported for the estimates of m_2, s_2, θ and σ_θ. For the same reason, no standard errors are reported for σ of the CIRM2LN model and θ and σ of the OUM2LN model either.

[Table 2]

5.3.1. In-Sample Performance

We first examine the general goodness-of-fit of each model to the VIX data in terms of LL, AIC and BIC measures reported in the middle panel of Table 2. The relative ranking of each model is the same in terms of any of the three measures. The worst performing model is the CIR model, followed closely by the OUDO model. This is expected, because the CIR model is a simple linear model and also the exponential transformation of the OUDO model offers no effective degrees of freedom to the simple linear underlying OU model. The remaining one-factor models all have parameter-dependent transformations. Consequently, the goodness-of-fit of these models are significantly better than the two benchmark models.

It is interesting to examine the relative performance of the three transformed CIR models, since they represent the drift-driven, the diffusion-driven and the distribution-driven models, respectively. The distribution-driven CIRM2LN model provided slightly better fit than the diffusion-driven CIRCEV, which slightly outperformed the drift-driven CIREW model. Compared to the CIREW model, the CIRCEV model has a well defined support on $D_Y = (0, \infty)$, making it a naturally coherent model for variables such as nominal interest rates and VIX. In addition, the CIRCEV model has a closed-form conditional mean and hence a closed-form pricing formula for the VXFs. These features make the CIRCEV model a very attractive alternative to the CIREW model in practice.

We plot in Figure 3 the estimated drift and diffusion functions of the one-factor models. We can see that when VIX is low, the estimated functions are relatively close among different models, but their differences increase quite dramatically as VIX increases. As we have seen from Figure 1, strong mean reversion and high volatility at high levels of VIX is a prominent feature of the VIX data. However, both functions of the CIR model are linear and very flat, unable to generate strong enough mean reversion or large enough volatility at high levels of VIX. All the other one-factor models have nonlinear drift and diffusion functions, but the distribution-driven CIRM2LN model has the strongest mean reversion and the largest volatility at high levels of VIX, unsurprisingly making it the best fitting one-factor model. Intuitively, the flexible M2LN distribution captures the information particularly in the right tail of the distribution much better than other models. This information is then suitably incorporated into the shapes of the drift and the diffusion functions to produce a better fit to the data. The estimated functions for the remaining one-factor models are relatively close.

[Figure 3]

We now turn our attention to the two-factor models. It is very interesting to note that despite the presence of a latent central tendency factor, the BVOUMS model only performed better than the benchmark CIR and OUDO models and was even outperformed by all other one-factor models. This is potentially an extremely important observation, as this suggests that at least for our data, the flexibility provided by the parameter-dependent nonlinear transformations play a more important role than the additional latent factor, if either but not both is included. Meanwhile, the BVOUM2LN model outperformed all other models by quite clear margins. This is expected, because the BVOUM2LN model contains not only a flexible parameter-dependent distribution-driven transformation function, designed to capture potentially crucial information in the stationary long-run behavior of VIX, but also a latent factor, which tracks the stochastic short-run central tendency of movement of VIX.

We plot in Figure 4 the estimated drift functions for the two two-factor models, conditional on the latent factors, taking several values between the 1st and the 99th quantile of their estimated stationary distributions. Note that the conditional drift functions for both models are nonlinear and relatively close when VIX is low, but their differences start to emerge as VIX goes up. Specifically, as VIX increases, the spread of the conditional drift functions across different values of the latent factor becomes wider for the BVOUM2LN model than for the BVOUMS model. Another striking difference is that the conditional drift functions of the BVOUMS model are globally concave in the level of VIX, but those of the BVOUM2LN model are not. We can see quite clearly that when VIX is in the middle range, the drift functions of the BVOUM2LN model have some degrees of convexities conditional on medium to low values of the latent factor. The wider spread and the higher degrees of nonlinearities of the conditional drift functions of the BVOUM2LN model are but potentially vital differences in explaining the dynamics of VIX. We can only attribute these to the flexible distribution-driven transformation. That is, the M2LN distribution can more flexibly capture the spread and variation in the density curve of the stationary distribution of the VIX data, and crucially such distributional features are then constructively translated into the variations in the estimated drift functions.

[Figure 4]

To further demonstrate the differences of the two models, we plot in the left panel of Figure 5 the estimated stationary densities of the two two-factor models together with that of the benchmark CIR model and the nonparametric kernel density[14]. As we can see, the implied stationary density by the BVOUM2LN model matches the kernel density very closely, incorporating most, if not all, key distributional features of the data. In contrast, a very large proportion in the middle of the stationary density implied by the BVOUMS model departed significantly from the kernel density, leading to significant differences in the estimated functions and goodness of fit to the data. Furthermore, we plot in the right panel of Figure 5 the estimated diffusion functions for the three models together with the nonparametric kernel diffusion estimate[15]. Compared to the flat linear diffusion function of the CIR model, that of the BVOUMS model is nonlinear and increases in VIX, but it is only to a limited extent. The BVOUM2LN model, however, shows much stronger nonlinearity and produces almost twice as much volatility for high levels of VIX, which is more consistent with our observation from the time series plot of the VIX. Most importantly, the estimated diffusion function of the BVOUM2LN model matches quite closely with the nonparametric estimate in terms of both the level and the slope. In clear contrast, however, that of the BVOUMS

[14] We use the bandwidth h_M which is justified in our discussion about the bandwidth selection in the specification test for diffusion models of Aït-Sahalia (1996b).

[15] We estimate the diffusion function by the Nadaraya-Watson estimator. The same estimator was considered by Jiang and Knight (1997), Stanton (1997). See Aït-Sahalia and Park (2016) for more details.

model deviates quite substantially from the nonparametric estimate, with no overlapping whatsoever except for very low levels of VIX. Above all, the superior suitability of the BVOUM2LN model over other models is quite clear.

[Figure 5]

To formally assess the specifications of competing models, we present the results from three specification tests in Table 3. First, we consider the kernel-based nonparametric stationary density test for parametric diffusion models proposed by Aït-Sahalia (1996b)[16]. This test assesses the suitability of a parametric diffusion model by evaluating the distance between its stationary density and the nonparametric kernel density. We denote the test statistic as $Q(h)$ where h is the bandwidth for the kernel density estimator. The test rejects the null hypothesis that a parametric model is correctly specified if $Q(h)$ exceeds certain critical value. We begin with two widely used data-driven bandwidths, namely, the Silverman's rule-of-thumb bandwidth h_S (c.f. Silverman 1986) and the K-fold blockwise log-likelihood cross validation (LLCV) bandwidth h_K with $K = 10$. In the upper panel of Figure 6, we plot the kernel densities based on h_S and h_K over the empirical support of VIX. It is important to note that the Silverman's bandwidth clearly undersmoothes the density, whereas the 10-fold LLCV bandwidth visibly oversmoothes slightly. Thus, we may argue that the most suitable bandwidth, which is difficult to pin down precisely, should lie somewhere between h_S and h_K. We therefore consider the third bandwidth $h_M = (h_S + h_K)/2$ and plot the resulting kernel density in the bottom panel of Figure 6. As we can see, the density with h_M has quite suitable smoothness and at the same time preserves distinctive features of the shape of the distribution[17].

[Table 3]
[Figure 6]

We report the results of the test based on h_S, h_M and h_K in the top panel of Table 3. Note that the stationary density of the CIR model is Gamma and those of the CIREW and the CIRCEV models are the transformed Gamma. The stationary density of the OUDO and the BVOUMS models are lognormal, and finally those of the CIRM2LN, the OUM2LN, and the BVOUM2LN models are the M2LN. Following Aït-Sahalia (1996b) and Fernandes (2006), we use the most favorable test statistic obtained by minimizing the measured distance with respect to the parameters of the stationary density of each model. Therefore, the test results for models with the same parametric stationary density are identical. As we can see, our results are consistent across all three bandwidths. The CIR, the CIREW, the CIRCEV, the OUDO and the BVOUMS models are all strongly rejected by the test, with p-values being practically zero, and all three models with the M2LN stationary distribution are not rejected. In particular, the p-value for the test with h_M, which has been shown in Figure 6 to be a very reasonable bandwidth, is fairly large. This means that the M2LN distribution fits very closely to the true but unknown stationary distribution of the VIX data. To further confirm the suitability of the M2LN specification, we plot the estimated

[16] Fernandes (2006) employ the same test in his application of diffusion models to forecasting financial crashes.

[17] Since the support of VIX is positive, we performed the test based on the distance between the kernel density of the log of VIX and the stationary density of the log of VIX implied by each parametric diffusion model. The values of the three bandwidths are $h_S = 0.064$, $h_M = 0.109$ and $h_K = 0.153$, respectively.

M2LN density on top of the kernel density with h_M in the bottom panel of Figure 6. As we can quite clearly see, the two densities are very close to each other over the empirical support of VIX. Therefore, we may argue that our choice of the M2LN distribution is quite reasonably justified.

Second, we employ a LR test to examine whether each of those relatively simple models is rejected in favor of the most general BVOUM2LN model. Since those models are not strictly nested in the BVOUM2LN model, following Eraker and Wang (2015) we perform the test using parametric bootstrap. Specifically, we simulate 1000 replications of artificial time series from each model under the null hypothesis using the ML estimates from the original data as the true parameters. For each replication, we estimate the model under the null and the BVOUM2LN model by ML. This leads to an empirical sample of 1000 bootstrap LR statistics, from which we then find the empirical p-value for the LR statistic obtained from the original data. We report the original LR statistics and their corresponding bootstrap p-values for each model in the middle panel of Table 3. As we can see, our bootstrap LR tests strongly reject all models in favor of the BVOUM2LN model. In particular, for every model considered as the null, none of the 1000 bootstrap LR statistics actually exceeds the original LR statistic. Therefore, the evidence in favor of our proposed distribution-driven two-factor model is very strong.

Finally, to examine the overall goodness-of-fit of each competing model to the data, we consider an information-theoretic approach for testing model specification. Specifically, we employ the Information Matrix (IM) test, originally proposed by White (1982) and then studied by Chesher (1983), Lancaster (1984), Orme (1990), Chesher and Spady (1991), Horowitz (1994), and others, for overall model specification[18]. Since the IM test tends to suffer quite severe size distortion if the asymptotic critical value is used, resulting in significant over-rejection in finite samples (c.f. Orme 1990, Chesher and Spady 1991 and Horowitz 1994), we adopt the parametric bootstrap approach suggested by Horowitz (1994). More specifically, we generate 1000 replications of artificial time series from each model and calculate the IM test statistic based on each of the 1000 replications. From the 1000 bootstrap IM test statistics, we obtain the empirical p-value for the IM statistic, denoted by D, obtained from the original data. The results are reported in the bottom panel of Table 3. Unsurprisingly, all one-factor models are strongly rejected, with empirical p-values equal to zero, meaning that literally none of the 1000 bootstrap samples can produce an IM test statistic larger than the statistic obtained from the original sample. For the BVOUMS model, however, the IM statistics from 7 out of the 1000 bootstrap samples turn out to be greater than the original statistic, resulting in an empirical p-value of 0.007 and a rejection at 1% significance level. These rejections are not surprising, because our first two tests both rejected these models. Finally, we find that, in contrast, the empirical p-value of the IM test for the BVOUM2LN model is 0.063, meaning that out of the 1000 replications we observe 63 replications for which the corresponding IM statistics actually exceed the original test statistic. Thus, the BVOUM2LN model is not rejected at 5% significance level. Although the non-rejection may seem to be a fairly small margin, considering the parsimony of the BVOUM2LN model and relatively large sample size, we may argue that the presented evidence of its goodness-of-fit to the data is fairly satisfactory.

[18] See, for example, Maasoumi and Racine (2002, 2008), Hall et al. (2015) for more examples of the information-theoretical approach for testing model specifications.

5.3.2. Out-of-Sample Performance

We now compare models in terms of their out-of-sample forecasting accuracy. For each model, we produce six series of rolling sample conditional mean forecasts for the VIX corresponding to forecasting horizons of 1 day (1D), 1 week (1W), and 1, 3, 5, 7 months (1M, 3M, 5M, 7M)[19]. For each horizon, we compute the RMSFE based on the observed out-of-sample series and its rolling sample forecasts produced by each model, and report the results in the bottom panel of Table 2.

For forecasting at 1D and 1W horizons, the CIR model is the best performing one-factor model (1.207 and 2.601) followed immediately by the OUDO model (1.208 and 2.618). At 1M horizon, however, the best performing one-factor model is the OUDO model (3.616) followed by the CIR model (3.626). The fact that they outperform the remaining one-factor models at these horizons suggests that at relatively short horizons, nonlinear parameter-dependent transformations may not significantly improve the ability of one-factor TD models to track the conditional mean. However, as the forecasting horizon increases to medium range (3M) and long range (5M and 7M), the performance of the CIR and OUDO models deteriorate significantly and are then exceeded by other one-factor TD models. This is not surprising, because at short forecasting horizons, the conditional distributions of all diffusion models are close to the normal distribution. Thus, the ability of more flexible models is minimized, but more parsimonious models usually have the advantage. At longer horizons, however, the conditional mean tends to depend more on the information in the stationary distribution (long-run behavior) implied by the forecasting model, for which more sophisticated models, particularly our distribution-driven models have the advantages. This explains why, among one-factor models, the two simplest models performed the best in short horizons and the worst in medium to long horizons forecasts.

The main advantage of the two two-factor models is that the additional central tendency variable can model the evolution of the conditional mean with more flexibility. Thus, we expect the two two-factor models to perform well at varied horizons. We also expect the newly proposed BVOUM2LN model to perform better than the BVOUMS model, since the distribution-driven transformation is expected to capture the nonlinear dynamics and particularly the information in the stationary distribution more effectively. Both of our expectations are confirmed by the forecasting results. Firstly, we find that both two-factor models outperformed all one-factor models at all forecasting horizons. In particular, the margins are more substantial for longer forecasting horizons than for shorter horizons. More importantly, both two-factor models outperformed one-factor models even at the shortest horizon. Comparing between the two two-factor models, the newly proposed BVOUM2LN model, which has additional degrees of freedom in the transformation function, outperformed the BVOUMS model by significant margins at all forecasting horizons. Most importantly, the advantage increases monotonically as forecasting horizon increases, confirming that the superiority of the BVOUM2LN model can indeed be attributed to its distribution-driven transformation design to incorporate information in the stationary distribution (long-run behavior) of the dynamics of VIX. In summary, the new BVOUM2LN model outperformed all competing models both in-sample and out-of-sample in every category and scenario of comparison that we considered.

[19] The shortest forecasting horizon of one day matches the frequency of the original VIX series, and the longest horizon of seven months matches the longest time-to-maturity of our interpolated constant maturity VIX futures.

5.4. Analysis of VXF Pricing

We now examine the performances of competing TD models in pricing VXFs. We first outline our assumptions on the risk-neutral $\mathbb{Q}$-measure dynamics of competing models. As explained in Section 4.1, the transformation function U must be the same under both $\mathbb{P}$- and $\mathbb{Q}$-measures, we therefore focus on the $\mathbb{Q}$-measure dynamics of the UD X for each model. For the CIR model and the three transformed CIR models, we assume that

$$dX_t = \kappa^{\mathbb{Q}} \left(\theta^{\mathbb{Q}} - X_t\right) dt + \sigma\sqrt{X_t} dW_t^{\mathbb{Q}}$$

satisfying $\kappa^{\mathbb{Q}} > 0$, $\theta^{\mathbb{Q}} > 0$ and $2\kappa^{\mathbb{Q}}\theta^{\mathbb{Q}} \geq \sigma^2 > 0$. For the two transformed OU models, we assume that

$$dX_t = \kappa^{\mathbb{Q}} \left(\theta^{\mathbb{Q}} - X_t\right) dt + \sigma dW_t^{\mathbb{Q}}$$

satisfying $\kappa^{\mathbb{Q}} > 0$. For the CIRM2LN and the OUM2LN models, we impose the additional normalization constraint $\sigma = 1$. For the two two-factor models, we follow Mencía and Sentana (2013) in assuming the following the $\mathbb{Q}$-measure dynamics for the underlying bivariate system:

$$\begin{aligned} dX_t &= \kappa \left(\theta_t^{\mathbb{Q}} - X_t\right) dt + \sigma dW_{X,t}^{\mathbb{Q}} \\ d\theta_t^{\mathbb{Q}} &= \kappa_\theta \left(\theta^{\mathbb{Q}} - \theta_t^{\mathbb{Q}}\right) dt + \sigma_\theta dW_{\theta,t}^{\mathbb{Q}} \end{aligned}$$

with $\theta_t^{\mathbb{Q}} = \theta_t - \sigma\varsigma_X/\kappa$ and $\theta^{\mathbb{Q}} = \theta - \sigma\varsigma_X/\kappa - \sigma_\theta\varsigma_\theta/\kappa_\theta$, where ς_X and ς_θ are the MPR for the two Brownian motions, respectively. For the BVOUM2LN model, we impose the additional normalization constraints $\theta = 0$ and $\sigma_\theta = \sqrt{\left(2\kappa - \sigma^2\right)\left(\kappa + \kappa_\theta\right)\kappa_\theta/\kappa^2}$.

Following Eraker and Wang (2015), we estimate model parameters under both $\mathbb{P}$- and $\mathbb{Q}$-measures jointly by ML using the combination of the VIX data and the $1, 3, 5$ and 7-month constant maturity VXF data. The parameter estimates are provided in the upper panel of Table 4. Note that the unconstrained estimate of δ in the CIREW model is positive from our joint measure estimation. Thus, no constrained estimation is required. For the same reason as in the VIX only case, no standard errors are reported for the set of parameters discussed in Section 5.3. The last two rows report the risk-neutral parameters $\kappa^{\mathbb{Q}}$ and $\theta^{\mathbb{Q}}$ for the six one-factor models and the MPR parameters ς_X and ς_θ for the two two-factor models. For all one-factor models, the difference between their parameter estimates under the two measures confirms the negative market price of volatility risk. For both two-factor models, this is also confirmed directly by the negative values of the MPR parameters.

[Table 4]

5.4.1. In-Sample Performance

The LL and the in-sample RMSPE for each model are reported in the middle panel of Table 4. Similar to the VIX only case, in terms of both LL and RMSPE, the CIR model is the worst performing model followed by the OUDO model, confirming again that models without effective transformations are too restrictive. The three transformed CIR models have slight advantages over the OUM2LN model, and the distribution-driven CIRM2LN model is again the best performing transformed CIR model.

We expect the two-factor models to perform better than the one-factor models, because pricing VXFs is effectively an exercise of conditional mean forecasting under the risk-neutral $\mathbb{Q}$-measure. In terms of both LL and RMSPE, again both two-factor models outperformed all one-factor models by clear margins. Specifically, the RMSPE from the BVOUMS model (2.162) is approximately 33% and 18% smaller than the worst performing one-factor CIR model (3.223), and the best performing one-factor CIRM2LN model (2.628), respectively. This confirms that the additional latent central tendency variable indeed helps capturing the conditional mean movement more effectively. Meanwhile, the performance of the BVOUM2LN model is much stronger than that of the BVOUMS model. The RMSPE from the BVOUM2LN model (1.694) is approximately 47% and 36% smaller than the CIR model and the CIRM2LN model, respectively. Between the two two-factor models, the RMSPE of the distribution-driven BVOUM2LN model is approximately 22% smaller than that of the BVOUMS model which has no effective transformation. Thus, as in the VIX only case, this confirms that the degree of freedom provided by the distribution-driven transformation significantly improves the flexibility of the model.

5.4.2. Out-of-Sample Performance

To examine out-of-sample performance in pricing VXFs, we calculate the RMSPE for pricing constant maturity VXFs in our forecasting sample. The RMSPE results obtained from VXFs with all the maturities and VXFs with only 1, 3, 5 and 7 month individual maturities are all presented in the bottom panel of Table 4. The CIR model and the OUDO model outperformed remaining one-factor models at 1M, 3M, 5M maturities and all maturities put together. However, as the maturity (forecasting horizon) increases to 7M, the advantage gradually disappears, and the two models are outperformed eventually by more flexibly single-factor models. This result is consistent with our findings in the VIX only case. That is, for the one-factor models, simple models tend to forecast well in short and at best at medium horizons, and TD models with more flexible transformations tend to forecast well in long horizons in addition to the fact that more flexible models almost always perform better in the sample.

We again expect our two-factor models to perform better in forecasting. In fact, our VXF forecasting results are even stronger and more convincing than the VIX forecasting results. Specifically, in terms of the RMSPE for VXFs with all maturities, the results from the BVOUMS model (1.220) is approximately 39% and 45% smaller than the best and the worst performing one-factor models, respectively. Moreover, the results from the BVOUM2LN model is even 54% and 58% smaller than the best and the worst performing one-factor models, respectively. Only at 1M maturity, which is the shortest maturity considered, the CIR model (0.978) performed slightly better than the BVOUMS model (1.108). However, even at this shortest horizon, the RMSPE from the newly proposed BVOUM2LN model (0.574) is only 57% of that of the CIR model. Thus, the advantage of our distribution-driven transformed two-factor model is substantial. Above all, in terms of the out-of-sample VXF pricing errors, the newly proposed BVOUM2LN model is always substantially better than any other models across all maturities considered.

6. Conclusion

We made an important contribution to the literature of TD modelling by proposing a novel framework for modelling multi-factor TD models with latent components. We advocated the use of the distribution-driven approach for the specification of the transformation functions and the use of the analytically tractable MVOU diffusion process as the underlying system. Our framework is intuitively constructive, statistically flexible, analytically tractable, and practically easy to implement. We examine the performance of the newly proposed latent-factor TD model in comparison with existing one-factor TDs in modelling the dynamics of VIX and pricing VXF contracts. The newly proposed distribution-driven two-factor BVOUM2LN model outperformed every competing model both in the sample and out of the sample across all forecasting horizons considered for both the VIX and VXFs.

References

Ahn, D. H., and Gao, B. (1999). A Parametric Non-linear Model of Term Structure Dynamics. Review of Financial Studies, 12, 721-762.

Aït-Sahalia, Y. (1996a). Nonparametric Pricing of Interest Rate Derivatives. Econometrica, 64, 527-560.

Aït-Sahalia, Y. (1996b). Testing Continuous-Time Models of the Spot Interest Rate. Review of Financial Studies, 9, 385-426.

Aït-Sahalia, Y. (2002). Maximum Likelihood Estimation of Discretely Sampled Diffusions: A Closed-form Approximation Approach. Econometrica, 70, 223-262.

Aït-Sahalia, Y. (2008). Closed-Form Likelihood Expansions for Multivariate Diffusions. Annals of Statistics, 36, 906-937.

Aït-Sahalia, Y., R. Kimmel., (2007). Maximum Likelihood Estimation of Stochastic Volatility Models. Journal of Financial Economics, 83, 413-452.

Aït-Sahalia, Y., and Park, J. (2016). Bandwidth Selection and Asymptotic Properties of Local Nonparametric Estimators in Possibly Nonstationary Continuous-Time Models, Journal of Econometrics, 192(1), 119-138.

Amengual, D., and Xiu, D. (2018). Resolution of Policy Uncertainty and Sudden Declines in Volatility, Journal of Econometrics, 203(2), 297-315.

Bahra, B. (1997). Implied Risk-Neutral Probability Density Functions from Option Prices: Theory and Application. Working Paper No 66, Bank of England.

Balduzzi, P., Das, S.R., Foresi, S., and Sundaram, R.K. (1996). A Simple Approach to Three-Factor Affine Term Structure Models. Journal of Fixed Income, 6(3), 43-53.

Black, F., and Scholes, M. (1973). The Pricing of Options and Corporate Liabilities, Journal of Political Economy, 81, 637-654.

Brandt, M., and Santa-Clara, P. (2002). Simulated Likelihood Estimation of Diffusions with an Application to Exchange Rate Dynamics in Incomplete Markets. Journal of Financial Econometrics, 63, 161–210.

Brigo, D., and Mercurio, F. (2002). Lognormal-Mixture Dynamics and Calibration to Market Volatility Smiles. International Journal of Theoretical and Applied Finance, 5(4), 427-446.

Bu, R., Cheng, J., and Hadri, K. (2016). Reducible Diffusions with Time-Varying Transformations with Application to Short-Term Interest Rates. Economic Modelling, 52, 266-277.

Bu, R., Cheng, J., and Hadri, K. (2017). Specification Analysis in Regime-Switching Continuous-Time Diffusion Models for Market Volatility. Studies in Nonlinear Dynamics and Econometrics, 21(1), 65-80.

Bu, R., Giet, L., Hadri, K., and Lubrano, M. (2011). Modelling Multivariate Interest Rates Using Time-Varying Copulas and Reducible Non-linear Stochastic Differential Equations. Journal of Financial Econometrics, 9(1), 198-236.

Bu, R., and Hadri, K. (2007). Estimating Option Implied Risk-Neutral Densities using Spline and Hypergeometric Functions. Econometrics Journal, 10, 216-244.

Bu, R., Hadri, K., and Kristensen, D. (2018). Diffusion Copulas: Identification and Estimation, Working Paper.

Bu, R., Jawadi, F., and Li, Y. (2017). An Empirical Comparison of Transformed Diffusion Models for VIX and VIX futures. Journal of International Financial Markets, Institutions & Money, 46, 116-127.

Chan, K. C., Karolyi, A., Longstaff, F., and Sanders, A. (1992). An Empirical Comparison of Alternative Models of the Short-Term Interest Rate. Journal of Finance, 47, 1209-1227.

Chapman, D.A, and Pearson, N.D. (2000). Is the Short Rate Drift Actually Nonlinear? Journal of Finance, 55(1), 355-388.

Chesher, A. (1983), The Information Matrix Test, Economics Letters, 13, 45-48.

Chesher, A., and Spady, R. (1991). Asymptotic Expansions of the Information Matrix Test Statistic, Econometrica, 59, 787-815.

Choi, S. (2009). Regime-Switching Univariate Diffusion Models of the Short-Term Interest Rate. Studies in Nonlinear Dynamics and Econometrics, 13(1): Article 4.

Cobb, L., Koppstein, P., and Chen. N.H. (1983) Estimation and Moment Recursion Relations for Multimodal Distributions of the Exponential Family, Journal of the American Statistical Association, 78(381), 124-130.

Conley, T. L., Hansen, L., Luttmer, E., and Scheinkman, J. (1997). Short Term Interest Rates as Subordinated Diffusion. Review of Financial Studies, 10, 525-578.

Cox, J., Ingersoll, J., and Ross, S. (1985). In Intertemporal General Equilibrium Model of Asset Prices. Econometrica, 53, 363-384.

Dai, Q., and Singleton, K.J. (2000). Specification Analysis of Affine Term Structure Models, Journal of Finance, 55, 1943-1978.

Detemple, J., and Osakwe, C. (2000). The Valuation of Volatility Option, European Finance Review, 4, 21-50.

Duffie, D., and Kan, R. (1996). A Yield Factor Model of Interest Rates. Mathematical Finance, 6, 379-406.

Elerian, O., Chib, S., and Shephard, N. (2001). Likelihood Inference for Discretely Observed Non-Linear Diffusions. Econometrica, 69, 959-993.

Eraker, B. (2001). MCMC Analysis of Diffusion Models with Application to Finance. Journal of Business and Economic Statistics, 19(2), 177-191.

Eraker, B., and Wang, J. (2015). A Non-linear Dynamic Model of the Variance Risk Premium. Journal of Econometrics, 187, 547-556.

Fernandes, M. (2006). Financial Crashes as Endogenous Jumps: Estimation, Testing and Forecasting, Journal of Economic Dynamics & Control, 30, 111-141.

Forman, J. L., and Sørensen, M., (2014). A Transformation Approach to Modelling Multi-Modal Diffusions. Journal of Statistical Planning and Inference, 146, 56-69.

Goard, J., and Mazur, M. (2013). Stochastic Volatility Models and the Pricing of VIX Options. Mathematical Finance, 23, 439-458.

Grünbichler, A., and Longstaff, F. (1996). Valuing Futures and Options on Volatility, Journal of Banking Finance, 20, 985-1001.

Hall, A.R., Li, Y., Orme, C.D., and Sinko, A. (2015). Testing for Structural Instability in Moment Restriction Models: An Info-Metric Approach. Econometric Reviews, 34(3), 286-327.

Hamilton, J. (1994). Time Series Analysis, Princeton University Press, Princeton.

Hansen, B. (1994). Autoregressive Conditional Density Estimation. International Economic Review, 35, 705-730.

Hansen, L.P. and Sargent, T.J. (1983). The Dimensionality of the Aliasing Problem in Models with Rational Spectral Densities. Econometrica, 51, 377-387.

Horowitz, J. (1994). Bootstrap-Based Critical Values for the Information Matrix Test, Journal of Econometrics, 61, 395-411.

Jiang, G., and J. Knight., (1997). A Nonparametric Approach to the Estimation of Diffusion Processes with an Application to a Short-Term Interest Rate Model. Econometric Theory, 13, 615–645.

Karlin, S., and Taylor, H.M. (1981). A Second Course in Stochastic Processes, Academic Press, New York.

Kessler, M. (1997). Estimation of an Ergodic Diffusion from Discrete Observations. Scandinavian Journal of Statistics, 24, 211-229.

Kessler, M., and Rahbek, A. (2004). Identification and Inference for Multivariate Cointegrated and Ergodic Gaussian Diffusions. Statistical Inference for Stochastic Process, 7, 137-151.

Karatzas, I., and Shreve, S.E. (1991). Brownian Motion and Stochastic Calculus, Second Edition, Springer-Verlag, New York.

Lancaster, T. (1984). The Covariance Matrix of the Information Matrix Test, Econometrica, 52(4), 1051-1053.

Litterman, R., and Scheinkman, J. (1991). Common Factors Affecting Bond Returns. Journal of Fixed Income, 1(1), 54-61.

Lo, A.W. (1988). Maximum Likelihood Estimation of Generalized Ito Processes with Discretely Sampled Data. Econometric Theory, 4, 231-247.

Lye, J., and Martin, V. (1993). Robust Estimation, Nonnormalities, and Generalized Exponential Distributions, Journal of the American Statistical Association, 88(421), 261-267.

Maasoumi, E., and Racine, J. (2002). Entropy and Predictability of Stock Market Returns. Journal of Econometrics, 1071(2), 291-312.

Maasoumi, E., and Racine, J. (2008). A Robust Entropy-Based Test for Asymmetry for Discrete and Continuous Processes. Econometric Reviews 28, 246-261.

Melick, W.R., and Thomas, C.P., (1997). Recovering an Asset's Implied PDF from Option Prices: An Application to Crude Oil during the Gulf Crisis. Journal of Financial and Quantitative Analysis, 32, 91-115.

Mencía, J., and Sentana, E. (2013). Valuation of VIX Derivatives. Journal of Financial Economics, 108, 367-391.

Merton, R. C. (1973). Theory of Rational Option Pricing. Bell Journal of Economics and Management Science, 4, 141-183.

Orme, C. (1990). Small Sample Performance of the Information-Matrix Test, Journal of Econometrics, 46, 309-331.

Pedersen, A.R. (1995). A New Approach to Maximum-likelihood Estimation for Stochastic Differential Equations Based on Discrete Observations. Scandinavian Journal of Statistics, 22, 55-71.

Philips, P.C.B. (1973). The Problem of Identification in Finite Parameter Continuous Time Models. Journal of Econometrics, 1, 351-362.

Pritsker, M. (1998). Nonparametric Density Estimation and Tests of Continuous Time Interest Rate Models. Review of Financial Studies, 11(3), 449-487.

Silverman, B.W. (1986). Density Estimation for Statistics and Data Analysis. Chapman and Hall, London.

Shoji, I., and Ozaki, T. (1998). Estimation for Non-linear Stochastic Differential Equations by a Local Linearization Method. Stochastic Analysis and Applications, 16, 733-752.

Silverman, B.W. (1986). Density Estimation for Statistics and Data Analysis. Chapman and Hall, London.

Söderlind, P., and Svensson, L. (1997). New Techniques to Extract Market Expectations from Financial Instruments, Journal of Monetary Economics, 40, 383-429.

Song, Z., and Xiu, D. (2016). A Tale of Two Option Markets: Pricing Kernels and Volatility Risk. Journal of Econometrics, 190, 176-196.

Stanton, R. (1997). A Nonparametric Model of Term Structure Dynamics and the Market Price of Interest Rate Risk. Journal of Finance, 52, 1973-2002.

Vasicek, O. (1977). An Equilibrium Characterization of the Term Structure, Journal of Financial Economics, 5, 177-188.

White, H. (1882). Maximum Likelihood Estimation of Misspecified Models, Econometrica, 50(1), 1-25.

Figure 1: Time Series of Daily VIX

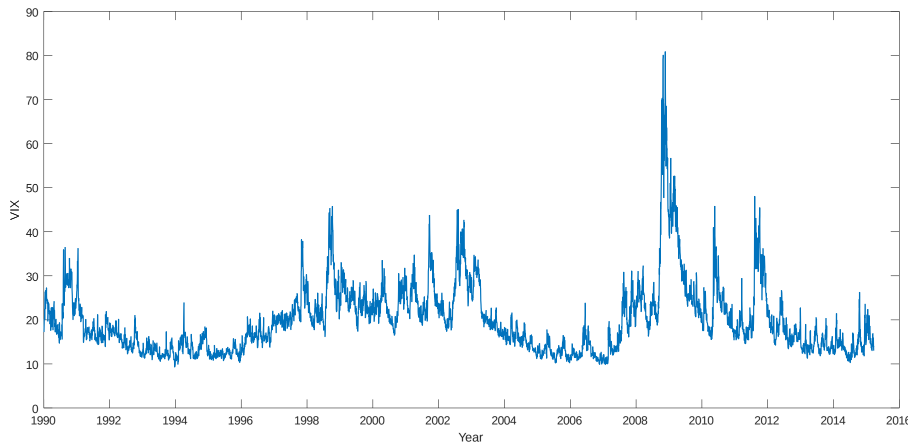


Figure 2: Term Structure of Constant Maturity VXFs

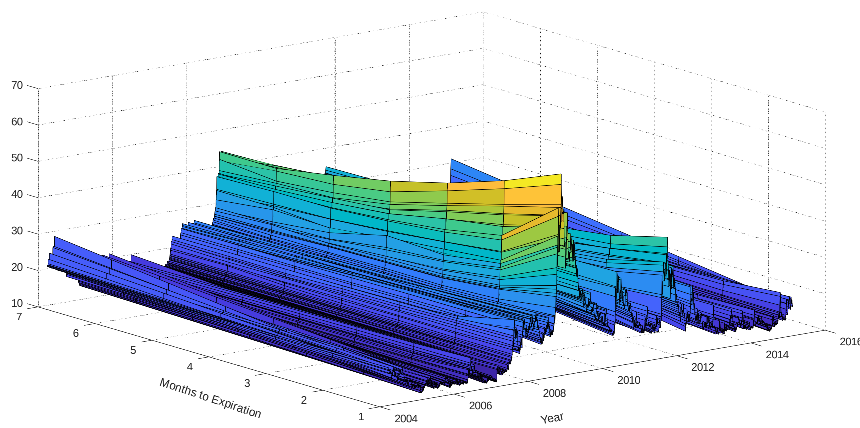


Figure 3: Estimated Drift and Diffusion Functions of One-Factor TD Models

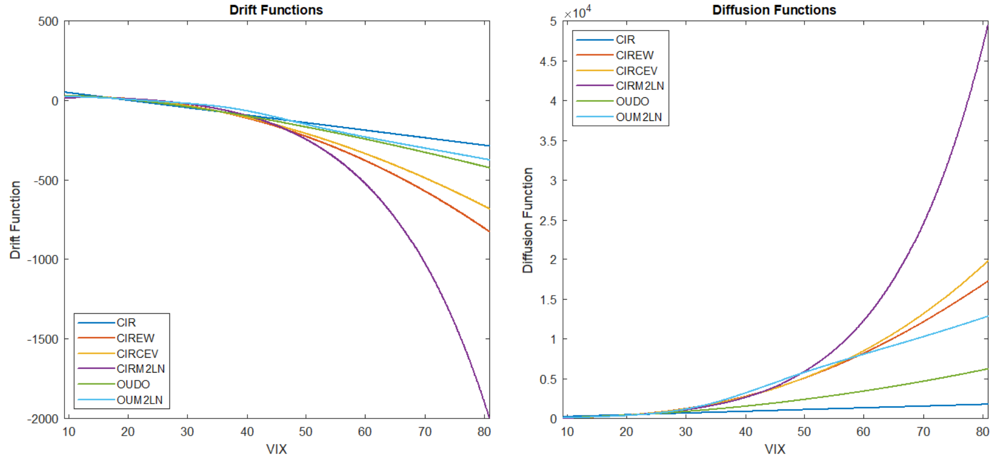


Figure 4: Estimated Conditional Drift Functions of Two-Factor TD Models

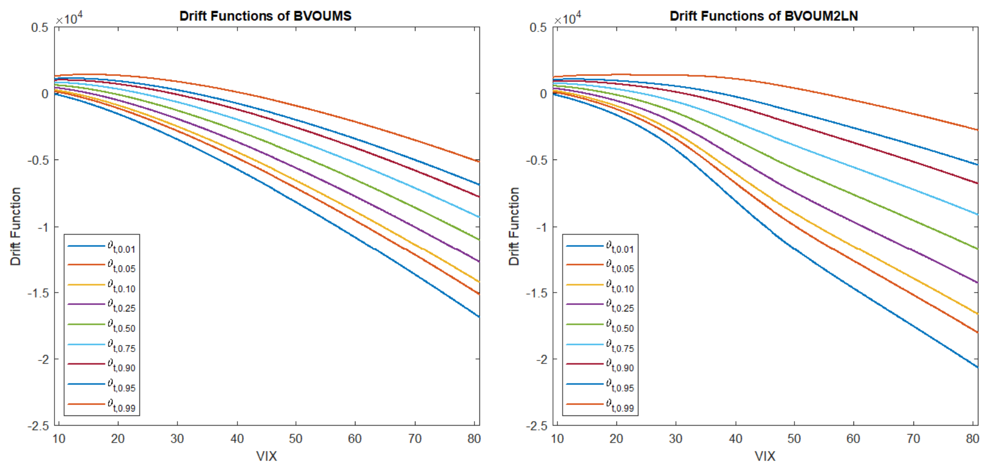


Figure 5: Estimated Stationary Densities and Diffusion Functions

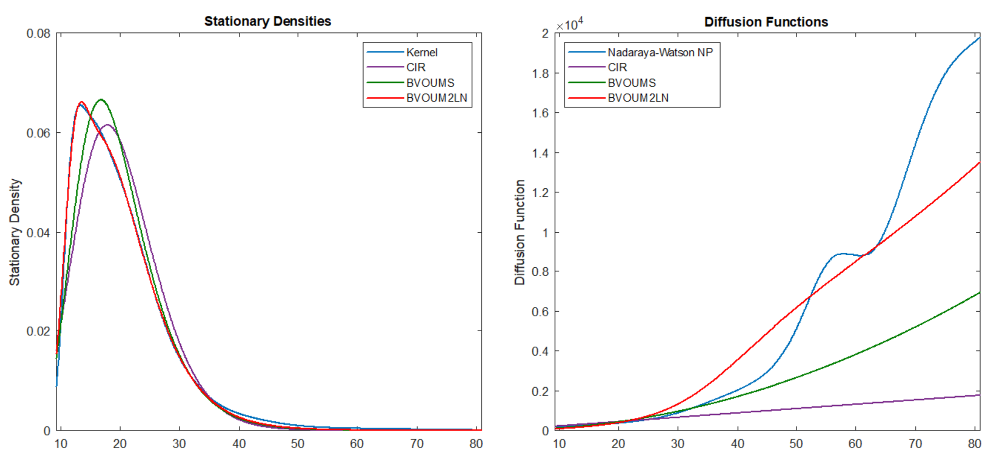
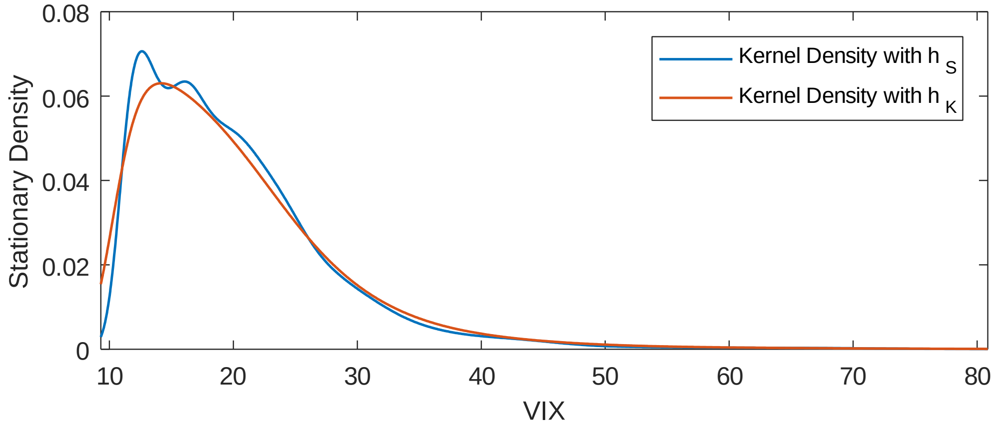


Figure 6: Stationary Densities of Daily VIX



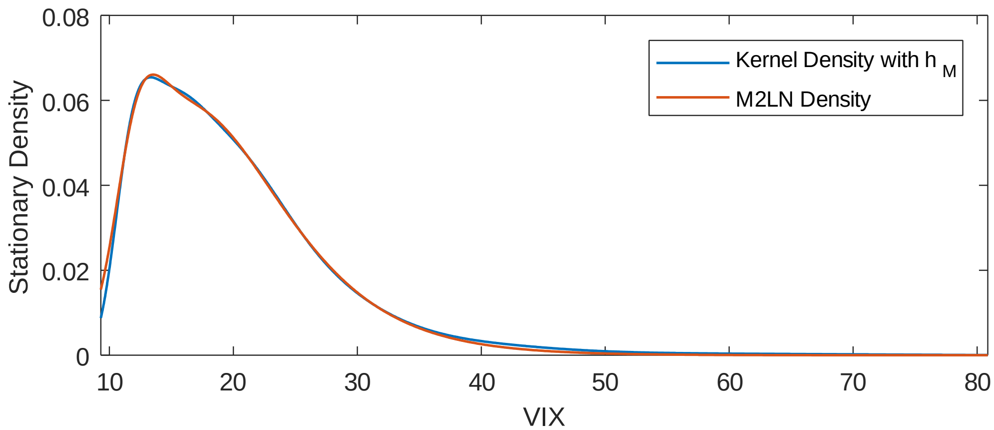


Table 1: **Summary of VIX and Constant Maturity VXFs**

	VXF							VIX
	1M	2M	3M	4M	5M	6M	7M	
Correlation	1.000	0.989	0.974	0.958	0.941	0.925	0.912	
		1.000	0.996	0.986	0.975	0.963	0.953	
			1.000	0.997	0.990	0.982	0.973	
				1.000	0.998	0.992	0.987	
					1.000	0.998	0.994	
						1.000	0.999	
							1.000	
Eigenvalue	6.852	0.136	0.008	0.002	0.001	0.000	0.000	
Maximum	65.462	59.004	53.782	49.727	46.707	44.635	44.085	80.860
Minimum	11.287	12.230	12.540	12.870	13.200	13.530	13.687	9.310
Mean	20.502	21.128	21.521	21.824	22.096	22.343	22.539	19.921
Median	18.180	19.312	19.944	20.506	20.972	21.272	21.534	18.130
Std. Dev.	8.351	7.672	7.153	6.806	6.566	6.384	6.231	7.982
Skewness	1.845	1.535	1.279	1.082	0.951	0.854	0.794	2.072
Kurtosis	7.070	5.815	4.717	3.936	3.508	3.232	3.085	10.466

Table 2: Estimation and Forecasting Results for VIX

	One-Factor Models						Two-Factor Models	
	Transformed CIR				Transformed OU		Transformed BVOU	
	CIR	CIREW	CIRCEV	CIRM2LN	OUDO	OUM2LN	BVOUMS	BVOUM2LN
κ	4.717	3.630	3.749	2.901	3.923	3.240	93.801	89.578
	(0.634)	(0.558)	(0.565)	(0.253)	(0.578)	(0.116)	(13.008)	(12.888)
θ	20.158	0.053	0.133	0.462	2.938	0	2.934	0
	(0.906)	(0.003)	(0.038)	(0.044)	(0.051)	-	(0.076)	-
σ	4.666	0.231	0.282	1	0.974	1	1.030	2.669
	(0.043)	(0.005)	(0.022)	-	(0.009)	-	(0.012)	(0.055)
κ_θ							1.588	1.157
							(0.372)	(0.140)
σ_θ							0.592	1.500
							(0.032)	-
γ			1.414					
			(0.027)					
δ		0						
		-						
η		0.005						
		-						
w				0.916		0.865		0.852
				(0.030)		(0.039)		(0.038)
θ_1				2.983		2.879		2.873
				(0.017)		(0.010)		(0.011)
s_1				0.365		0.331		0.330
				(0.005)		(0.006)		(0.006)
θ_2				2.483		3.336		3.333
				-		-		-
s_2				0.155		0.449		0.444
				-		-		-
LL $(\times 10^3)$	-10.006	-9.307	-9.305	-9.283	-9.436	-9.324	-9.364	**-9.251**
AIC $(\times 10^4)$	2.002	1.862	1.862	1.858	1.888	1.866	1.874	**1.852**
BIC $(\times 10^4)$	2.004	1.865	1.865	1.861	1.890	1.868	1.877	**1.856**
RMSFE_1D	1.207	1.211	1.210	1.211	1.208	1.210	1.206	**1.204**
RMSFE_1W	2.601	2.640	2.635	2.646	2.618	2.633	2.495	**2.484**
RMSFE_1M	3.626	3.697	3.674	3.727	3.616	3.639	3.402	**3.371**
RMSFE_3M	4.718	4.358	4.342	4.331	4.362	4.193	3.410	**3.244**
RMSFE_5M	5.474	5.185	5.164	5.160	5.162	5.093	4.066	**3.870**
RMSFE_7M	5.136	4.922	4.898	4.904	4.884	4.891	3.811	**3.630**

Table 3: Specification Diagnostics

	One-Factor Models						Two-Factor Models	
		Transformed CIR			Transformed OU		Transformed BVOU	
	CIR	CIREW	CIRCEV	CIRM2LN	OUDO	OUM2LN	BVOUMS	BVOUM2LN
Nonparametric Stationary Density Test (Aït-Sahalia 1996)								
$Q\left(h_S\right)$	2.868	2.983	2.703	0.319	2.629	0.319	2.629	0.319
p-value	0.000	0.000	0.000	**0.183**	0.000	**0.183**	0.000	**0.183**
$Q\left(h_M\right)$	2.414	2.278	1.965	0.071	1.909	0.071	1.909	0.071
p-value	0.000	0.000	0.000	**0.906**	0.000	**0.906**	0.000	**0.906**
$Q\left(h_K\right)$	1.867	1.329	1.055	0.023	1.070	0.023	1.070	0.023
p-value	0.000	0.000	0.000	**0.929**	0.000	**0.929**	0.000	**0.929**
Bootstrap Non-nested LR Test (Eraker and Wang 2015)								
LR	15.084	1.110	1.080	0.637	3.698	1.459	2.242	-
p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	-
Bootstrap Information Matrix Test (Horowitz 1994)								
D	9.550	1.820	2.877	1.544	39.887	1.444	5.029	2.452
p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.007	**0.063**

Table 4: Joint Estimation and Forecasting Results for VXFs

	One-Factor Models						Two-Factor Models	
	Transformed CIR				Transformed OU		Transformed BVOU	
	CIR	CIREW	CIRCEV	CIRM2LN	OUDO	OUM2LN	BVOUMS	BVOUM2LN
κ	4.717	3.047	3.396	1.917	3.925	2.560	74.990	7.966
	(0.634)	(0.547)	(0.537)	(0.077)	(0.578)	(0.061)	(0.543)	(0.187)
θ	20.158	0.052	0.048	0.526	2.938	0	2.831	0
	(0.906)	(0.005)	(0.005)	(0.012)	(0.051)	-	(0.042)	-
σ	4.666	0.291	0.199	1	0.975	1	1.138	2.349
	(0.043)	(0.003)	(0.005)	-	(0.009)	0	(0.011)	(0.022)
κ_θ							0.828	0.481
							(0.009)	(0.013)
σ_θ							0.178	0.817
							(0.003)	-
γ			1.517					
			(0.008)					
δ		1.575						
		-						
η		0.012						
		-						
w				0.799		0.842		0.688
				(0.014)		(0.018)		(0.013)
θ_1				2.799		2.816		3.077
				(0.011)		(0.009)		(0.008)
s_1				0.376		0.332		0.429
				(0.002)		(0.005)		(0.003)
θ_2				3.503		3.606		2.640
				-		-		-
s_2				0.231		0.428		0.167
				-		-		-
$\kappa^Q(\varsigma_X)$	1.324	0.692	1.274	0.574	1.342	1.421	-1.071	-1.025
	(0.016)	(0.020)	(0.013)	(0.025)	(0.013)	(0.018)	(0.269)	(0.021)
$\theta^Q(\varsigma_\theta)$	25.786	0.074	0.051	0.777	3.045	-0.030	-2.706	-0.180
	(0.099)	(0.002)	(0.004)	(0.018)	(0.006)	(0.009)	(0.193)	(0.020)
LL ($\times 10^4$)	-3.547	-3.293	-3.328	-3.286	-3.396	-3.335	-3.124	**-2.867**
RMSPE	3.223	2.665	2.763	2.628	2.929	2.764	2.162	**1.694**
RMSPE(ALL)	1.992	2.164	2.079	2.222	1.998	2.113	1.220	**0.925**
RMSPE(1M)	0.978	1.315	1.217	1.374	1.083	1.301	1.108	**0.574**
RMSPE(3M)	1.593	2.063	1.884	2.132	1.690	1.958	0.732	**0.713**
RMSPE(5M)	2.195	2.420	2.308	2.482	2.197	2.325	1.097	**0.957**
RMSPE(7M)	2.750	2.624	2.634	2.675	2.666	2.631	1.729	**1.293**