Analyzing volatility risk and risk premium in option contracts: A new theory*

Peter Carr[a], Liuren Wu[b†]

[a] *Courant Institute, New York University, 251 Mercer Street, New York, NY 10012, USA*

[b] *Baruch College, Zicklin School of Business, One Bernard Baruch Way, New York, NY 10010, USA*

Abstract

We develop a new option pricing framework that tightly integrates with how institutional investors manage options positions. The framework starts with the near-term dynamics of the implied volatility surface and derives no-arbitrage constraints on its current shape. Within this framework, we show that just like option implied volatilities, realized and expected volatilities can also be constructed specific to, and different across, option contracts. Applying the new theory to the S&P 500 index time series and options data, we extract volatility risk and risk premium from the volatility surfaces, and find that the extracted risk premium significantly predicts future stock returns.

JEL Classification: C13; C51; G12; G13

Keywords: Implied volatility surface; Option realized volatility; Expected volatility surface; Volatility risk premium; Vega-gamma-vanna-volga; Proportional variance dynamics

*We thank William Schwert (the editor), an anonymous referee, Torben Anderson, Hans Buehler, Bruno Dupire, Robert Engle, Travis Fisher, Jeremy Graveline, Rachid Lassoued, Alex Levin, Keith Lewis, Dilip Madan, Fabio Mercurio, Attilio Meucci, Alexey Polishchuk, Jason Roth, Angel Serrat, Mridul Tandon, Edward Tom, Ilya Ustilovsky, Arun Verma, Scott Weiner, and seminar participants at Baruch College, the Fields Institute, Northwestern University, Florida State University, Singapore Management University, the 2011 Western Finance Association meetings in Santa Fe, the 2011 China International Conference in Finance in Wuhan, China, and the 2013 Derivatives Conference at New York University for their comments and suggestions. Liuren Wu gratefully acknowledges the support by a grant from the City University of New York PSC-CUNY Research Award Program.

†Corresponding author. Tel: +1 646 312 3509; fax: +1 646 312 3451.
E-mail address: liuren.wu@baruch.cuny.edu (L. Wu).

1. Introduction

The option pricing literature has made great advances during the past decade; yet large gaps remain between theory and practice. First, traditional option pricing models specify the underlying price and variance rate dynamics and derive their implications on option prices; however, institutional investors manage their volatility views and exchange their quotes not through option prices, but through the option implied volatility computed from the Black-Merton-Scholes (BMS) model. This common practice does not mean that investors agree with the assumptions made by Black and Scholes (1973) and Merton (1973); rather, they use the BMS model as a transformation to enhance quote stability and to highlight the information in the option contract. Second, traditional option pricing theory requires the full specification of the instantaneous variance rate dynamics, not only about its current level, but also about its long run mean; yet in practice, investors do not observe the instantaneous variance rate, but instead observe many option implied volatilities across a wide spectrum of strikes and maturities. Furthermore, investors have much more confidence on how these implied volatilities move in the near term than in the very long run. The map between the implied volatility surface and the instantaneous variance rate dynamics is not always clear or well-determined, forcing modelers to frequently re-calibrate their models to match moving market conditions, with each re-calibration generating a new set of parameters that are supposed to be fixed over time. Such fudging practices create consistency concerns because the option pricing function would differ if one expects these parameters to be varying over time.

In this paper, we develop a new option pricing framework that tightly integrates with how institutional investors manage their option positions, thus closing the gap between theory and practice. Instead of modeling the full dynamics of an unobservable instantaneous variance rate and deriving the implication on option prices, the new framework models the near-term dynamics of the BMS implied volatility across different strikes and expires, and derives no-arbitrage constraints directly on the shape of the implied volatility surface. Under the assumed implied volatility dynamics, the shape of the whole implied volatility surface can be cast as the solution to a simple quadratic equation. The computational burden is dramatically reduced

compared to the standard option pricing literature. More importantly, by starting with the whole implied volatility surface instead of a single instantaneous variance rate, the new theory does not need to specify the full dynamics, but just the current levels of the drift and the diffusion processes. The current shape of the implied volatility surface only depends on the current levels of its drift and diffusion processes, but does not depend on how these processes will involve in the future. This "unspanned" nature allows the shape of the current implied volatility surface to be represented as a function of many state variables, but with no fixed model parameters. The high dimensionality renders the model flexible enough to fit the observed implied volatility surface well, whereas the absence of fixed model parameters dramatically simplifies model estimation, alleviates concerns on model stability over time, and allows continuous model recalibration to update the state variables without inducing any intertemporal inconsistency.

The fact that the new theory only specifies the near-term dynamics of the implied volatility surface while leaving its long-term variation unspecified highlights its "semi-parametric" flavor:[1] The theory specifies just enough dynamic structure to achieve a fully parametric characterization of the current implied volatility surface, while saying little about its long-run variation. Traditionally, one can either fit the surface parametrically or nonparametrically. Nonparametric fitting is easy to do, but with concerns that the nonparametrically smoothed implied volatility surface may not satisfy no-arbitrage conditions, may not be extrapolated with stability to regions where data are sparse or unavailable, and the method does not provide a mechanism to reduce the dimension of the surface to a few economically meaningful states. On the other hand, a fully specified parametric model can provide stable and arbitrage-free extrapolation, dimension reduction, and economic interpretation, but it has issues regarding its stability over time, its poor performance when the state dimension is low, and its numerical complexity and instability when the dimension is high. Our semi-parametric theory balances the two by providing a numerically simple approach to readily interpolate and extrapolate the surface while satisfying dynamic no-arbitrage constraints, and to reduce the dimension of the surface to a few economic states while leaving the state dynamics unspecified, thus avoiding introducing any fixed model parameters.

[1] We thank the referee for highlighting this feature.

The new theoretical framework does not replace the role played by fully parametric, equilibrium-based option pricing models; instead, it can provide a bridge between market observations and the fundamental valuations from these models. A well-specified parametric option pricing model may not fit the current market observations well, but its valuation can guide future market implied volatility movements. If one believes that option implied volatilities move toward their corresponding fundamental valuations from a parametric model, the new theory can readily embed the fundamental valuations from this model as the near-term targets of the implied volatility movements, and derives no-arbitrage constraints on the current shape of the implied volatility surface with the fundamental valuation as its reference point. To do so, the new theory only asks for the numerical valuation results from the parametric model, without needing to know its parametric model details.

Within the new theoretical framework, we propose a new concept that just like option implied volatilities, both realized and expected volatilities can be made specific to, and different across, option contracts. We define the *option realized volatility* (ORV) at each strike and expiry as the volatility level at which one achieves zero realized profit if one buys the option and performs daily delta-hedge based on the BMS model with this volatility input. Although this realized volatility can be estimated from the realized security price sample path, it is defined against a specific option contract and hence can differ across different strikes and expiries of the reference option contract. Since writing the option at this ORV level generates zero profit, the ex post premium from writing the option at its market price is directly given by the BMS value difference when evaluated at the option's implied volatility and its ORV level, respectively. This new option-specific volatility concept is tightly linked to the common practice of volatility investors, who usually take option positions and perform dynamic delta hedging to separate the volatility exposure from the directional price movement.[2] Taking an option position with delta hedge exposes the investor to variance risk during the life of the option, but the exposure to the different segments of the sample path differs for different option contracts. The ORV estimate for each option contract represents a weighted average of the variance risk

[2]Indeed, most institutional volatility investors and options market makers are required by their institutions to maintain delta neutrality.

over the sample path, with the weighting determined by the risk exposure of that contract.

To measure the ex ante volatility risk premium embedded in each option contract, we propose to estimate an *option expected volatility* (OEV) at each strike and expiry, defined as the volatility forecast that generates zero *expected* profit if one buys and delta-hedges the option at this volatility level. The difference between the OEV surface and the option implied volatility surface defines the volatility risk premium embedded in the option contracts across different strikes and maturities. Just as the current shape of the implied volatility surface is constrained by its near-term risk-neutral dynamics, the current shape of the OEV surface is analogously constrained by its near-term statistical dynamics. These constraints allow us to perform dimension reduction and extract meaningful economic states from the two surfaces.

We apply the new theoretical framework to the S&P 500 index time series and its options. Our data include nearly 18 years of over-the-counter SPX option implied volatility quotes from January 1997 to October 2014. At each date, the quotes are at a fixed grid of five relative strikes from 80% to 120% of the spot level and eight fixed time to maturities from one month to five years. Corresponding to these implied volatility quotes, we estimate the ORV at the corresponding relative strike and maturity levels based on the realized SPX sample paths, and we also propose a statistical procedure to estimate the corresponding OEV forecast based on exponential moving averages of BMS values of historical ORV estimates.

Given the unique feature of the new theoretical framework that the volatility surfaces are functions of several state variables but with no fixed model parameters, we propose a state-updating procedure based on an extended version of the classic Kalman (1960) filter. With this procedure, we can fit thousands of volatility surfaces and extract the corresponding state variables in a matter of seconds.

By fitting the statistical OEV dynamics to the current OEV surface shape and fitting the risk-neutral implied volatility dynamics to the current implied volatility surface shape, we obtain two sets of dynamics estimates that highlight how different economic states vary at different historical sample periods. The differences between the two sets of dynamics also highlight how the volatility risk premium varies over time. We project the volatility risk premium estimates to a return risk premium comopnent based on the

return-volatility correlation, and find that this projected return risk premium can predict future stock returns.

In a classic paper, Merton (1973) develops model-free bounds on option prices arising from no static arbitrage. These bounds can be classified into two types. Type I bounds are derived based on no-arbitrage arguments between European options of a fixed strike and maturity versus the underlying security and cash. Examples include: Call and put prices must not be smaller than their intrinsic value; call prices on a stock must not be larger than the dividend discounted stock price; put prices must not be larger than the present value of the strike price; and put-cal parity must hold. Type II bounds are derived based on no-arbitrage arguments between options of different strikes and maturities, such as the constraints that bull, bear, calendar, and butterfly spreads must be no less than zero. Hodges (1996) shows that by quoting an option in terms of a positive implied volatility, all Type I bounds are automatically guaranteed. This property makes it very attractive for market makers to quote and update implied volatilities based on options order flows while using an automated system to update the option prices whenever the underlying security price moves. Unfortunately, quoting positive implied volatilities does not exclude Type II arbitrages. Our new theory takes advantage of the BMS implied volatility transformation to exclude Type I arbitrages, and derives no-arbitrage constraints directly on the current shape of the implied volatility surface based on assumptions on its near-term movements.

In related work, Bakshi and Kapadia (2003a,b) articulate the idea that one can analyze the volatility risk premium by investigating the delta-hedged gains from each option contract. Our new theoretical framework formalizes their insights via the concept of option-specific expected and implied volatilities. To understand the risk profile of a portfolio of delta-hedged option positions, Engle and Figlewski (2015) propose a statistical model for the dynamics and correlations of implied volatilities across different individual stocks. Also related to our work is the growing literature on variance risk premium. Carr and Wu (2009) propose to use the difference between expected future realized variance and the variance swap rate to measure the variance risk premium. A growing list of studies build upon this variance risk premium measure, from developing theories explaining the large variance risk premium (Drechsler and Yaron (2011), Baele, Driessen,

Londono, and Spalt (2014)), modeling the variance swap term structure and developing variance swap allocation strategies (Egloff, Leippold, and Wu (2010)), documenting variance risk premium in other markets (Mueller, Vedolin, and Yen (2012)), to relating the equity variance risk premium to other financial markets (Bollerslev, Tauchen, and Zhou (2009) and Zhang, Zhou, and Zhu (2009)). Since over-the-counter variance swap rates are not readily available, most of these studies use vanilla options to form a replicating portfolio to approximate the variance swap rate (Carr and Wu (2006)) and Jiang and Tian (2005)). Our new framework provides a platform for analyzing volatility risk and volatility risk premium in each option contract, without resorting to option portfolio formulation.

Finally, there have been some largely unsuccessful attempts in the literature in directly modeling the implied volatility dynamics. Examples include Avellaneda and Zhu (1998), Ledoit and Santa-Clara (1998), Schonbucher (1999), Hafner (2004), Fengler (2005), and Daglish, Hull, and Suo (2007). These models are often called market models of implied volatility. These attempts have completely different starting point and ending objectives from our analysis. Instead of deriving no-arbitrage constraints on the implied volatility surface, these attempts take the observed implied volatility (on a single option, a curve, or over the whole surface) as given while specifying the continuous martingale component of the volatility surface. From these two inputs, they try to derive the no-arbitrage restrictions on the risk-neutral drift of the surface. The approach is analogous to the Heath, Jarrow, and Morton (1992) model on forward interest rates and can in principle be used for pricing derivatives written on the implied volatility surface. What this literature fails to recognize is that the knowledge of the current implied volatility surface places constraints on the specification of the continuous martingale component for its future dynamics. In this paper, rather than ignoring these constraints, we fully exploit them in building a simple, direct linkage between the current shape of the implied volatility surface and its near-term dynamics.

The remainder of the paper is organized as follows. Section 2 establishes the new theoretical framework by specifying near-term implied volatility dynamics and deriving the allowed shapes for the current implied volatility surfaces that exclude dynamic arbitrage. Section 3 defines the option realized and expected volatil-

ity surfaces across strikes and expiries based on the security price sample paths, and shows how the future statistical dynamics of the expected volatility surface determine the current shape of the surface. Section 4 documents the stylized evidence on the option implied and expected volatility surfaces for the S&P 500 index. Section 5 proposes a dynamic estimation procedure for extracting the economic states from the two volatility surfaces. Section 6 discusses the estimation results. Section 7 provides concluding remarks and directions for future research.

2. Implied volatility surface: From near-term dynamics to current shape

We consider a market with a riskfree bond, a risky asset, and a continuum of vanilla European options written on the risky asset.[3] For simplicity, we assume zero interest rates and zero carrying cost/benefit for the risky asset. In practical implementation, one can readily accommodate a deterministic term structure of financing rates by modeling the forward value of the underlying security and defining moneyness of the option against the forward. The risky asset can be any types of tradable securities, but we will refer to it as the stock for concreteness. We assume frictionless and continuous trading in the riskfree bond, the stock, and a basis option, and we assume no-arbitrage between the stock and the bond. As a result, there exists a risk-neutral probability measure $\mathbb{Q}$, equivalent to the statistical probability measure $\mathbb{P}$, such that the stock price S is a martingale.

We assume that the stock price S evolves in continuous time as a strictly positive and continuous semi-martingale. By the martingale representation theorem, there exists a standard Brownian motion W under $\mathbb{Q}$ such that S solves the following stochastic differential equation:

$$dS_t/S_t = \sqrt{v_t}dW_t, \tag{1}$$

[3] In the US, exchange-traded options on individual stocks are all American style. To apply our new theory to American options, a commonly used shortcut is to extract the BMS implied volatility from the price of an American option based on some tree/lattice method and use the implied volatility to compute a European option value for the same maturity date and strike. See Carr and Wu (2010) for a detailed discussion on data processing on individual stock options.

where v_t denotes the time-t instantaneous variance rate. We allow v to follow a positive, real-valued stochastic process such that there exists a unique solution to (1). However, in contrast to existing literature, we do not specify the risk-neutral dynamics of this process; instead, we will specify the risk-neutral dynamics of the BMS implied volatility for each vanilla option:

$$dI_t(K,T) = \mu_t dt + \omega_t dZ_t, \tag{2}$$

for all $K > 0$ and $T > t$. We refer to μ_t as the drift process and ω_t as the volatility of volatility process (henceforth "volvol" for short). Both processes can be stochastic and they can both depend on deterministic quantities such as calendar time t, strike price K, and maturity date T. In contrast to μ_t and ω_t, the standard Brownian motion Z_t is independent of the strike K and maturity T at all times. The two Brownian shocks on the stock price and the implied volatilities are allowed to be correlated,

$$\mathbb{E}_t\left[dW_t dZ_t\right] = \rho_t dt, \tag{3}$$

where ρ_t is a stochastic process taking values in the interval $[-1,1]$.

It is worth noting that equation (1) assumes a purely continuous security price dynamics, thus excluding discontinuous price movements from the security price specification, and equation (2) makes the strong assumption that instantaneously, the whole implied volatility surface is driven by one Brownian shock. On the other hand, we allow μ_t, ω_t, ρ_t to be stochastic processes.

The analysis is on European options. , but options on individual stocks in the US are American style. As is commonly done for option pricing on individual stock options,[4]one can perform de-Americanization before applying the model.

4

2.1. The fundamental PDE governing the implied volatility surface

For concreteness, let the basis option be a call with $C_t(K_0,T_0)$ denoting its value, and let all other options be puts, with $P_t(K,T)$ denoting the corresponding values. Let $B(S,\sigma,t;K,T):\mathbb{R}^+\times\mathbb{R}^+\times[0,T)\mapsto\mathbb{R}^+$ be the BMS model formula for a European put option:

$$B(S,\sigma,t;K,T)\equiv KN\left(\frac{\ln(K/S)}{\sigma\sqrt{T-t}}+\frac{\sigma\sqrt{T-t}}{2}\right)-SN\left(\frac{\ln(K/S)}{\sigma\sqrt{T-t}}-\frac{\sigma\sqrt{T-t}}{2}\right). \tag{4}$$

To reduce notation clutter, we henceforth suppress the notational dependence of B on contract characteristics K and T when no confusion shall occur.

By the definition of BMS implied volatility, we can write both the basis call option and the other put options in terms of the BMS put formula,

$$C_t(K_0,T_0)=B(S_t,I_t(K_0,T_0),t)+S_t-K_0,\quad P_t(K,T)=B(S_t,I_t(K,T),t), \tag{5}$$

for all $t\geq 0$, $K>0$, and $T>t$. It is well known that the function $B(S,\sigma,t)$ is $C^{2,2,1}$ on $\mathbb{R}^+\times\mathbb{R}^+\times[0,T)$, so Itô's formula can be used to relate increments of B to the increments of S, I, and t. To shorten the length of the following equations, we let subscripts of B denote partial derivatives and we suppress the arguments of B, which are always $(S_t,I_t(K,T),t)$.

Requiring the implied volatility for any option at (K,T) be positive, $I_t(K,T)>0$, guarantees no static arbitrage between the options at (K,T) and the underlying stock and cash (Hodges (1996)). We further require that no dynamic arbitrage be allowed on any option at (K,T) relative to the basis option at (K_0,T_0), the stock, and cash. This requirement for no dynamic arbitrage leads to a fundamental partial differential equation (PDE) between the functions $B(S,\sigma,t)$ and $I_t(K,T)$.

Proposition 1 *Under the stock price dynamics in (1), the implied volatility dynamics in (2), and the correlation specification in (3), the absence of dynamic arbitrage on an option contract $P_t(K,T)$ relative to*

the basis option at (K_0, T_0)*, the stock, and a riskfree bond dictates that the BMS option pricing function* $B(S,\sigma,t)$ *and the implied volatility function* $I_t(K,T)$ *for this option jointly solve the following fundamental PDE:*

$$-B_t = \mu_t B_\sigma + \frac{1}{2} v_t S_t^2 B_{SS} + \rho_t \omega_t \sqrt{v_t} S_t B_{S\sigma} + \frac{1}{2} \omega_t^2 B_{\sigma\sigma}. \tag{6}$$

Refer to Appendix A for the proof.

In the fundamental PDE in (6), the terms involving partial derivatives of B are called theta for B_t, vega for B_σ, dollar gamma for $S_t^2 B_{SS}$, dollar vanna for $S_t B_{S\sigma}$, and volga for $B_{\sigma\sigma}$. When μ_t and ω_t are independent of (K,T), equation (6) defines a linear relation between the BMS theta of the option and its vega, dollar gamma, dollar vanna, and volga. We christen the class of implied volatility surfaces defined by the fundamental PDE as the **Vega-Gamma-Vanna-Volga (VGVV)** model.

It is important to note that the PDE in equation (6) is not a PDE in the traditional sense. Traditionally, a PDE is specified to solve the value function. In our case, the value function $B(S_t, I_t, t)$ is simply the BMS put option formula in (4). Furthermore, the coefficients on traditional PDEs are deterministic, but they are stochastic in our PDE. Most important, our PDE is not derived to solve the value function, but rather to show that the various stochastic quantities have to satisfy this particular relation to exclude dynamic arbitrage.

Not only is the value function $B(S_t, I_t, t)$ well known, so are its various partial derivatives. Plugging these partial derivatives into the PDE in equation (6), we can reduce the PDE into an algebraic equation that links the implied volatility dynamics to the current shape of the implied volatility surface. This algebraic equation becomes particularly simple if we represent the current implied volatility surface as a function of the relative strike $k \equiv \ln(K/S)$ and time to maturity $\tau \equiv T - t$, i.e., $I_t(k,\tau)$.[5]

Proposition 2 *The fundamental PDE in (6) can be translated into a no dynamic arbitrage constraint on the current shape of the implied volatility surface* $I_t(k,\tau)$*, jointly determined by the current instantaneous variance rate level* v_t*, the current levels of the instantaneous drift* (μ_t) *and volatility* (ω_t) *of the implied volatility*

[5] To avoid introducing too many different notations, we use I_t to denote both the implied volatility value at time t and the various representations of the implied volatility function. The different representations are differentiated by the arguments that follow.

dynamics, and the current level of the instantaneous correlation process between return and implied volatility (ρ_t):

$$0 = \frac{1}{2}I_t^2 - \mu_t I_t \tau - \frac{1}{2}v_t - \rho_t \frac{\omega_t}{I_t}\sqrt{v_t}\left(k + \frac{I_t^2\tau}{2}\right) - \frac{1}{2}\frac{\omega_t^2}{I_t^2}\left(k^2 - \frac{1}{4}I_t^4\tau^2\right). \tag{7}$$

By specifying particular parametric functional forms for μ_t and ω_t, one can determine the algebraic nature of the manifold (7) in which the implied volatility function $I_t(k,\tau)$ resides. Just as integral transforms often convert partial differential equations into algebraic equations, the use of implied volatility has transformed the second order parabolic PDE in (6) into the simple algebraic relation in (7). Under our dynamic assumptions for the stock price and the implied volatilities, a necessary condition arising from no dynamic arbitrage is that the current implied volatility surface $I_t(k,\tau)$ resides in the manifold defined by equation (7).

The no-arbitrage constraint embedded in equation in (7) links the current shape of the implied volatility surface to the current levels of the drift process μ_t and the diffusion process ω_t for the implied volatility dynamics, as well as current levels of the correlation process ρ_t and the instantaneous variance rate process v_t; however, the no-arbitrage condition places no direct constraints on the exact dynamics specifications for these four processes $(\mu_t, \omega_t, \rho_t, v_t)$. Thus, the constraint on the current implied volatility surface shape only comes from the *near-term* dynamics of the implied volatility surface.

2.2. Proportional volatility dynamics

Different parameterizations for the drift μ_t and the diffusion ω_t of the implied volatility dynamics lead to different functional shapes for the implied volatility surface. As an illustrating example, we consider one particularly simple specification, where both the drift and the diffusion are proportional to the implied volatility level:

$$dI_t(K,T)/I_t(K,T) = e^{-\eta_t(T-t)}\left(m_t dt + w_t dZ_t\right), \quad w_t, \eta_t > 0, \tag{8}$$

where m_t, w_t, and η_t are stochastic processes that do not depend on K, T, or $I(K,T)$. We constrain w_t to be a strictly positive process with no loss of generality, and we use the exponential dampening $e^{-\eta_t(T-t)}$ to

accommodate the empirical observation that implied volatilities for long-dated options tend to move less.

Equation (8) represents a minimalist structure that captures the current levels of the drift and diffusion of the implied volatility surface. Full dynamic specifications for stochastic volatilities often accommodate mean reversion in the drift. Given our near-term focus, we use m_t to parsimoniously capture the direction and magnitude of the next expected move without delving into the particularly drivers of the move.

The literature often specifies the dynamics on the variance instead of volatility. Under the specification in (8), the implied variance dynamics also possess a proportional structure:

$$dI_t^2(K,T)/I_t^2(K,T) = 2e^{-\eta_t(T-t)}\left(\left(m_t + \frac{1}{2}e^{-\eta_t(T-t)}w_t^2\right)dt + w_t dZ_t\right). \tag{9}$$

It is worth noting that our implied variance diffusion specification deviates from the commonly used affine setting. Equation (9) dictates that in the limit, the diffusion for the instantaneous variance rate v_t is also proportional to the variance rate level. By contrast, the commonly used affine variance rate dynamics, e.g, Heston (1993), specifies the diffusion for the variance rate as proportional to the square root of the variance rate. The traditional literature relies heavily on the affine setting mainly for pricing tractability. Under our framework, the pricing is extremely tractable for a wide array of dynamics specifications. Our decision to deviate from the affine setting is motivated by better empirical performance.

Proposition 3 *Under the stock price dynamics in (1)-(3), when the implied variance follows the proportional dynamics in (8), no dynamic arbitrage requires that the implied variance surface as a function of relative strike k and time to maturity* τ*,* $I_t^2(k,\tau)$*, satisfy the following quadratic equation,*

$$0 = \frac{1}{4}e^{-2\eta_t\tau}w_t^2\tau^2 I_t^4 + \left(1 - 2e^{-\eta_t\tau}m_t\tau - e^{-\eta_t\tau}w_t\rho_t\sqrt{v_t}\tau\right)I_t^2 - \left(v_t + 2e^{-\eta_t\tau}w_t\rho_t\sqrt{v_t}k + e^{-2\eta_t\tau}w_t^2k^2\right). \tag{10}$$

Refer to Appendix C for the proof.

An interesting and important feature of equation (10) is that the no-arbitrage constraint depends on the

current levels of the five dynamic processes $(m_t, w_t, \eta_t, v_t, \rho_t)$, but it does not depend directly on the exact dynamics of these processes. Thus, the dynamics of the five state variables are left unspecified. As a result, fitting the relation to observed implied volatility surfaces only involves extracting the levels of the five dynamic states, but does not involve the estimation of any model parameters that govern the state dynamics.

At a fixed time to maturity, equation (10) describes a hyperbola in the variables I^2 and k,

$$I_t^2(k) = a_t + \frac{2}{\tau}\sqrt{(k-b_t)^2 + c_t}, \tag{11}$$

with

$$\begin{aligned} a_t &= \frac{-2\left(1-2e^{-\eta_t\tau}m_t\tau - e^{-\eta_t\tau}w_t\rho_t\sqrt{v_t}\tau\right)}{w_t^2\tau^2}, \\ b_t &= -\frac{\rho\sqrt{v_t}}{e^{-\eta_t\tau}w_t}, \\ c_t &= \frac{(1-\rho_t^2)v_t}{e^{-2\eta_t\tau}w_t^2} + \frac{\left(1-2e^{-\eta_t\tau}m_t\tau - e^{-\eta_t\tau}w_t\rho_t\sqrt{v_t}\tau\right)^2}{e^{-4\eta_t\tau}w_t^4\tau^2}. \end{aligned} \tag{12}$$

Equation (11) dictates that the implied variance smile $I_t^2(k)$ is convex in k. Furthermore, as $k \to \pm\infty$, the implied variance smile behaves linearly in k. Furthermore, equation (11) shows that the implied volatility smile at a fixed maturity is determined by three transformed covariates (a_t, b_t, c_t). The remaining degrees of freedom determine how the smiles vary across different maturities.

Gatheral (2006) proposes a similar, but more general form for the implied variance smile that involves five free covariates,

$$I_t^2(k) = a + b\left[\rho(k-m) + \sqrt{(k-m)^2 + \sigma^2}\right]. \tag{13}$$

Gatheral labels his specification as the SVI (stochastic volatility inspired) model. Although Gatheral makes some motivational linkages between these coefficients and stochastic volatility models, no stochastic volatility models have been proposed that lead exactly to this generalized SVI form. In particular, it is not clear how the five coefficients should vary across different time to maturities.

In the limit of $\tau = 0$, the implied variance is quadratic in the relative strike k,

$$I_t^2(k) = v_t + 2\rho_t \sqrt{v_t} w_t k + w_t^2 k^2, \tag{14}$$

with the volvol process w_t dictating the curvature and the correlation ρ_t dictating the asymmetry of the smile.

If we define "at-the-money" as the relative strike equal to the conditional mean of the log stock return $\ln(S_T/S_t)$ under the BMS model, $k = -\frac{1}{2} I_t^2(k,\tau)\tau$, the at-the-money implied variance takes the simple form,

$$A_t^2(\tau) = \frac{v_t}{1 - 2e^{-\eta_t \tau} m_t \tau}, \tag{15}$$

where the at-the-money implied variance term structure depends on the drift specification (μ_t), but is unaffected by the choice of the volvol process ω_t, nor by the choice of the correlation process ρ_t. The term structure starts at v_t at $\tau = 0$. As τ initially increases, the term structure can be either upward or downward sloping, depending on the drift of the implied volatility process m_t. It is upward sloping when the drift is positive ($m_t < 0$) and downward sloping when the drift is negative ($m_t < 0$).

2.3. A bridge between market observations and fundamental option valuations

Given the completely different starting points, one naturally wonders whether implied volatility dynamics specified under the new theoretical framework can be mapped tractably to corresponding dynamics for the instantaneous variance rate in the traditional modeling framework, and vice versa. While this endeavor can be an interesting direction for future research, making such a mapping is likely to be very difficult, except under certain special cases (e.g., Carr and Sun (2007)). What makes the mapping particularly difficult is the fact that under the new framework, the implied volatility dynamics are not fully specified. We specify merely the current drift and diffusion levels of the implied volatility dynamics, while saying nothing about their future movements. This partial specification gives us tremendous flexibility in fitting the implied volatility surface by updating the values of a set of state variables, but without the need to pin down the

dynamics of these state variables.

Given the partial specification, the new theory does not replace the roles played by fully parametric, equilibrium-based option pricing models; instead, it can provide a bridge between market observations and the fundamental valuations from these models. A well-specified parametric option pricing model may not fit the current market observations well, but its valuation can guide future market implied volatility movements. Specifically, if one believes that market-observed option implied volatilities tend to move toward their corresponding fundamental valuations from a parametric model, we can capture this reversion behavior via the following implied volatility dynamics specification,

$$dI_t^2(K,T) = \kappa_t \left(M_t(K,T) - I_t^2(K,T) \right) dt + 2e^{-\eta_t (T-t)} w_t I_t^2(K,T) dZ_t, \tag{16}$$

where the diffusion component remains the same as (9), but the size and direction of the drift is dictated by the deviation between the current market implied variance level and the model valuation $M_t(K,T)$, which denotes the value of the BMS implied variance generated from the particular parametric option pricing model. Equation (16) represents an analogous continuous-time specification to the error-correction specification of Engle and Granger (1987). Through the error-correcting drift specification, equation (16) drives the implied variance toward the model value $M_t(K,T)$, with κ_t controlling the error-correcting speed.

Given the specification in (9), one can analogously derive no-arbitrage constraints on the observed implied volatility surface by treating $M_t(K,T)$ as a numeric input for each contract.

$$\begin{aligned} 0 \;=\; & \tfrac{1}{4} e^{-2\eta_t \tau} w_t^2 \tau^2 I_t^4(k,\tau) + \left(1 + \kappa_t \tau + e^{-2\eta_t \tau} w_t^2 \tau - e^{-\eta_t \tau} w_t \rho_t \sqrt{v_t} \tau \right) I_t^2(k,\tau) \\ & - \left(v_t + \kappa_t M_t(k,\tau) \tau + 2 e^{-\eta_t \tau} w_t \rho_t \sqrt{v_t} k + e^{-2\eta_t \tau} w_t^2 k^2 \right). \end{aligned} \tag{17}$$

Equation (17) essentially contains results from two layers of dynamic modeling. The first layer is the traditional parametric option pricing model that generates the benchmark implied volatility surface valuation $M_t(K,T)$. The second layer is our near-term error-correction dynamics assumption that dictates how the observed implied volatility surface should vary around the parametric model valuation. The first layer

enters into the second layer only through the valuation at different strikes and maturities, with no direct reference to the particular model dynamics specification. This separation highlights the flexibility of the new theoretical framework, as it can build upon any fundamental model by capturing the market reversion to the fundamental valuation.

3. Option-contract specific realized and expected volatilities

Corresponding to the option implied volatility surface, we propose the new concept that realized and expected volatilities can also be defined in reference to a specific option contract.

3.1. Defining realized and expected volatility specific to an option contract

For each option contract, we define its **option realized volatility (ORV)** as the volatility input to the BMS model such that if one buys the option at this volatility level, with the invoice price generated from the BMS pricing formula, and performs daily BMS delta hedge on the option based on this volatility level through out the life of the option contract, the terminal profit and loss (PL) is zero. Formally, let $ORV(K,t,T)$ denote the option realized volatility corresponding to the reference option contract initiated at time t, struck at K, and expiring at T, and let $\{t_j\}_{j=1}^N$ denote the sequence of days during the life of the option with $t_0 = t$ and $t_N = T$, we can compute the ORV as

$$\begin{array}{rcl} ORV(K,t,T) & \equiv & x, \\ \text{s.t.} \qquad 0 & = & B(S_T,x,T) - B(S_t,x,t) - \sum_{j=1}^N B_S\left(S_{t_{j-1}},x,t_{j-1}\right)\left(S_{t_j} - S_{t_{j-1}}\right), \end{array} \tag{18}$$

where the second line defines the PL from buying the reference option contract and performing daily delta hedge, with x being the volatility input, $B(S_{t_j},x,t_j)$ denotes the BMS value of the reference option contract on date t_j, and $B_S(S_{t_j},x,t_j)$ denotes the corresponding BMS delta of the contract. The ORV is the BMS volatility level that makes the delta-hedged PL zero.

Although ORV is defined on a particular reference option contract, equation (18) highlights the fact that the value of ORV does not depend on the market price of that contract, but rather only on the sample path of the underlying security price. Its meaning becomes clearer if we further assume that the underlying price dynamics are purely continuous with stochastic volatility, in which case Carr and Madan (2002) show that if one buys the options at BMS volatility x and performs continuous delta hedge at this volatility, the delta-hedged PL can be written as

$$PL = \int_t^T \frac{1}{2} S_u^2 B_{SS}(S_u, x, u)(v_u - x^2) du. \tag{19}$$

Thus, setting this PL to zero to solve for x amounts to computing x^2 as a *weighted* average of the instantaneous variance rate with the weight given by the dollar gamma of the option at each point of the sample path. Therefore, based on the same price sample path, one can arrive at different variance estimates for different option contracts due to the different dollar gamma weighting for different option contracts. The traditionally defined realized variance can be regarded as a special case of this definition with equal weighting to each day's realization. Indeed, one can think of the traditionally defined realized variance with equal weighting as an ORV corresponding to the variance swap contract, which has constant dollar gamma.

Given an estimate of the ORV for a particular option contract, it becomes immediately clear that selling the contract makes money if its BMS implied volatility is higher than the ORV and loses money if its implied volatility is lower than the ORV. The delta-hedged PL from buying an option contract can be computed directly as the BMS value difference between using the ex post option realized volatility and the option's ex ante implied volatility as inputs, respectively. While the implied volatility $I_t(K,T)$ is known at time t, $ORV(K,t,T)$ is not fully realized until time T. The ex post realized dollar PL from buying the option contract and delta-hedging to expiration can be written as

$$PL(t,T) = B(S, ORV(K,t,T), t; K, T) - B(S, I_t(K,T), t; K, T). \tag{20}$$

Since the BMS option pricing formula is monotonic in its volatility input, the sign of the PL is determined by the sign of the difference between the ORV and the implied volatility.

Taking expectations on the ex post realized PL, one can obtain the ex ante expected volatility risk premium embedded in each option contract. To facilitate the ex ante volatility risk premium calculation, we propose a corresponding **option expected volatility (OEV)** surface, $V_t(K,T)$, defined as the time-t volatility forecast for each option contract at strike K and expiry T such that the *expected* PL is zero if one buys this option and delta-hedge to expiration at this volatility level:

$$B(S,V_t(K,T),t;K,T) = \mathbb{E}_t^{\mathbb{P}}\left[B(S,ORV(K,t,T),t;K,T)\right], \tag{21}$$

where $\mathbb{E}_t^{\mathbb{P}}[\cdot]$ denotes the expectation operator under the statistical measure $\mathbb{P}$ conditional on time-t filtration $\mathcal{F}_t$. According to this definition, if the current implied volatility level for the option contract is equal to the option's expected volatility, $I_t(K,T) = V_t(K,T)$, the expected delta-hedged PL from buying this option would be zero. On the other hand, if the implied volatility level differs from the expected volatility level, the expected delta-hedged PL, or the volatility risk premium (VRP) from buying this option can be computed simply as,

$$VRP_t(K,T) = B(S,V_t(K,T),t;K,T) - B(S,I_t(K,T),t;K,T). \tag{22}$$

Therefore, the difference between this option expected volatility surface and the implied volatility surface defines the surface of the volatility risk premium across different strikes and maturities.

3.2. No-arbitrage constraints on the option expected volatility surface

Analogous to the current shape of the implied volatility surface being determined by its future risk-neutral dynamics, the current shape of the expected volatility surface is constrained by its future statistical dynamics. In parallel to the risk-neutral proportional implied volatility dynamics in (8), we assume that the option expected volatility is proportional to its corresponding option implied volatility and that the two also follow

proportional dynamics under the statistical measure $\mathbb{P}$,

$$dV_t(K,T)/V_t(K,T) = e^{-\eta_t(T-t)}\left(m_t^{\mathbb{P}}dt + w_t dZ_t^{\mathbb{P}}\right), \tag{23}$$

where we use a different drift process $m_t^{\mathbb{P}}$ to capture the effects of market pricing of the volatility risk dZ_t. Under the continuous price dynamics assumption, both the implied volatility surface and the expected volatility surface converge to the same instantaneous volatility rate $\sqrt{v_t}$ as the time to maturity approaches zero. The two surfaces are also governed by the same return-volatility correlation level ρ_t. However, when the underlying security price can jump randomly, the instantaneous variance becomes an expectation of both diffusive movements and random jumps. When the jump risk is priced, the expectation can generate different values under the two measures $\mathbb{P}$ and $\mathbb{Q}$. The expected skewness of the return distribution can also differ under the two measures.[6] To partially accommodate the impacts of random price jumps, we relax the model assumption and use different variance rates ($v_t^{\mathbb{P}}$) and return-variance correlations ($\rho_t^{\mathbb{P}}$) under the statistical measure $\mathbb{P}_t$ to better match the expected volatility surface behaviors,

$$dS_t/S_t = \mu_t^{\mathbb{P}}dt + \sqrt{v_t^{\mathbb{P}}}dW_t, \quad \mathbb{E}_t^{\mathbb{P}}\left[dW_t^{\mathbb{P}}dZ_t^{\mathbb{P}}\right] = \rho_t^{\mathbb{P}}dt. \tag{24}$$

Proposition 4 *Under the stock price dynamics in (24) and the expected volatility dynamics in (23) under the statistical measure $\mathbb{P}$, the expected variance surface as a function of relative strike k and time to maturity τ, $V_t^2(k,\tau)$, satisfies the following quadratic equation,*

$$0 = \frac{1}{4}e^{-2\eta_t\tau}w_t^2\tau^2V_t^4 + \left(1 - 2e^{-\eta_t\tau}m_t^{\mathbb{P}}\tau - e^{-\eta_t\tau}w_t\rho_t^{\mathbb{P}}\sqrt{v_t^{\mathbb{P}}}\tau\right)V_t^2 - \left(v_t^{\mathbb{P}} + 2e^{-\eta_t\tau}w_t\rho_t^{\mathbb{P}}\sqrt{v_t^{\mathbb{P}}}k + e^{-2\eta_t\tau}w_t^2k^2\right). \tag{25}$$

Refer to Appendix D for the proof.

[6] See, for example, Polimenis (2006) for an illuminating discussion on how relative risk aversion interacts with the return cumulants under the two measures.

3.3. Linking variance risk premium to return risk premium

We assume that the presence of risk premium leads to a difference in the drift process of the implied volatility dynamics: m_t under the risk-neutral measure and $m_t^{\mathbb{P}}$ under the statistical measure. The difference between the two can be regarded as a measure of the instantaneous variance risk premium. Furthermore, since the return innovation dW_t and the variance innovation dZ_t is correlated, estimates on the variance risk premium have direct implications on the return risk premium.

Formally, we can perform a decompose on the stock return Brownian stock,

$$dW_t = \rho_t dZ_t + \sqrt{1-\rho_t^2}d\widetilde{W}_t,$$

where $d\widetilde{W}_t$ denotes the component of the return risk independent of the variance risk. With the decomposition, we assume the following pricing kernel dynamics,

$$dM_t/M_t = -\gamma_t\sqrt{v_t}dZ_t - \zeta_t\sqrt{v_t}d\widetilde{W}_t. \tag{26}$$

With this pricing kernel specification, the variance risk premium is given by

$$m_t - m_t^{\mathbb{P}} = -\gamma_t w_t \sqrt{v_t}. \tag{27}$$

The instantaneous return risk premium is given by $(\gamma_t\rho_t + \zeta_t\sqrt{1-\rho_t^2})v_t$. Without knowing ζ_t, we cannot fully identify the return risk premium, but we can estimate the contribution of the variance risk premium to the return risk premium (RRP) as

$$RRP = \gamma_t\rho_t v_t. \tag{28}$$

When the return-variance correlation is high in absolute magnitude, this RRP component becomes the major contribution of the total return risk premium.

4. Implied and expected volatility surfaces on S&P 500 Index options

We use options on S&P 500 index (SPX) to perform an empirical analysis on the new theory. We obtain matrix implied volatility quotes on SPX options from a major bank. The quotes are constructed to match the listed option prices at short maturities and to match the over-the-counter transactions at long maturities. The matrix quotes are on a grid of five fixed relative strikes from 80% to 120% of the spot level and eight fixed time to maturities from one month to five years. The data are available from January 8, 1997 to October 29, 2014. For our analysis, we sample the data weekly every Wednesday to avoid weekday effects. The weekly sampled data include 40 implied volatility series over 930 weeks, a total of 37,200 observations.

Corresponding to the implied volatility quotes, we obtain an extended sample of the SPX daily time series starting from January 8, 1982. At each date t and corresponding to each implied volatility quote $I_t(k,\tau)$, we compute a historical option realized volatility, $ORV(k,t-\tau,t)$, using the SPX time series data from time $(t-\tau)$ to time t. We start the historical ORV calculation ten years earlier than the implied volatility sample from January 8, 1987, which needs the time series data going back five additional years to January 8, 1982 for maturities up to five years. In estimating the conditional expectation of the BMS transformation at each time t and for each (k,τ) reference point, we apply exponential moving average to the BMS value of the historical ORV estimates, with an exponential decay speed of 0.03 per day. The ten-year additional history is for the exponential moving average estimates to stabilize. The moving average of the BMS value is then inverted back to obtain the OEV estimate $V_t(k,\tau)$.

4.1. The average behavior of volatility surfaces and volatility risk premiums

Table 1 reports the sample averages of the 40 implied volatility series in panel A. At each fixed time to maturity, the average implied volatility levels are higher for low strikes than for high strikes, forming the well-known negatively skewed implied volatility smirk pattern that is widespread across all global equity indexes (Foresi and Wu (2005)). At a fixed relative strike level, the average implied volatility declines with

maturity at low strikes but increases with maturity at high strikes. In particular, the at-the-money implied volatilities show an average upward sloping term structure.

[Table 1 about here.]

Panel B reports the sample averages of the corresponding option expected volatility (OEV) estimates at each relative strike and maturity. The negative skew along the strike dimension also shows up on the OEV surface, but the skew is not as monotone and becomes more of a smile, especially at short maturities. At a fixed relative strike level, the OEV term structure is downward sloping at low relative strikes, but mostly flat at other strikes, forming a contrast with the upward sloping term structures observed on at-the-money and high-strike implied volatilities. The fact that the implied volatility mean term structure is more upward sloping than the expected volatility mean term structure suggests that on average $m_t > m_t^{\mathbb{P}}$ and hence the market price of the variance risk γ_t, defined in the pricing kernel specification in (26), is negative.

The difference between the option expected volatility and the implied volatility defines the volatility risk premium on each option contract in volatility percentage points. A positive difference indicates positive expected PL from taking a long position in the option and delta-hedging until expiration, and hence a positive volatility risk premium. Panel C of Table 1 reports the average difference across different strikes and maturity. The average volatility risk premium is mostly negative except on high-strike, short-maturity options. The volatility risk premium is particularly negative for far out-of-the-money put options, where the average implied volatility can be over 10 volatility points higher than the corresponding average OEV.

To gauge the economic significance of the volatility risk premium, Panel D reports the annualized information ratio of a long option strategy: At each date t, we buy the option at (k,τ) and perform delta-hedge until expiration. The log expected return from this investment can be computed as $\ln B(V_t(k,\tau))/B(I_t(k,\tau))$, where $B(I_t(k,\tau))$ denotes the market cost of buying this option and $B(V_t(k,\tau))$ denotes the expected delta-hedged payoff from the long option position. We define the annualized information ratio as the ratio of the mean log return to the standard deviation of the log return, annualized by the square root of the time to

maturity of the option. The information ratio estimates are highly negative for all low-strike options and at-the-money options, but they become positive for some short-term high-strike options, as the OTC implied volatility quotes on these option are on average lower than the corresponding expected volatility estimates.

4.2. The time-series variation of implied and expected volatilities

Figure 1 compares the time-series variation of option implied and expected volatilities. Panels A and C plot the time series of the at-the-money implied and expected volatilities whereas Panels B and D plot the 90% strike minus 110% strike volatility differences as a measure of skewness on the return distribution. The three lines in each panel are for three selected maturities at one month (solid line), six months (dashed line), and 24 months (dash-dotted line), respectively.

[Fig. 1 about here.]

The time-series variations of at-the-money implied and expected volatilities show similar patterns. The volatility series show spikes during the 1998 Asian crises and the ensuing hedge fund crisis in 1999, during the mild recession in early 2000, and most prominently during the financial market melt down around 2008. The spike in 2012 corresponds to the European sovereign debt crisis.

The 90%-110% option implied volatility difference is uniformly positive over our whole sample period and across all option maturities, suggesting that the option implied SPX return distribution is persistently negatively skewed. By contrast, the corresponding option expected volatility shows much less negative skewness, and the estimates can turn positive from time to time. The large difference in skewness highlights the market's extreme aversion to stock market crashes (Wu (2006)).

4.3. Volatility of volatility dependence structure

Our dynamics specification assumes that the variance of the volatility changes is proportional to the variance level. This specification forms a contrast with the square-root instantaneous variance rate specification in the affine option pricing literature, e.g., Heston (1993), which implies that the variance of the instantaneous volatility changes is independent of the volatility level. To investigate how the variance of the volatility changes depends on the volatility level, we estimate a constant elasticity of variance (CEV) specification on the implied volatility time series,

$$\frac{1}{\Delta t} Var_t\left(\Delta I_t(k,\tau)\right) = C(\tau)\left(I_t^2(k,\tau)\right)^{\beta}, \tag{29}$$

where the free power coefficient β would be equal to one under our proportional volatility specification or a log normal stochastic volatility model (e.g., Hull and White (1987)), but equal to zero under the square-root variance specification (e.g. Heston (1993)).

To estimate this power coefficient β, we first estimate an exponentially weighted variance (EVI) on each implied volatility series,

$$EVI_t = \phi EVI_{t-1} + (1-\phi)\left[(\Delta I_t)^2/\Delta t\right], \tag{30}$$

where $\Delta t = 1/52$ denotes the weekly sampling frequency, ΔI_t denotes the weekly changes on an implied volatility series, and we set the exponential smoothing coefficient $\phi = 0.97$, corresponding to a half life of about half a year. Then, we perform the following linear regression to estimate the power coefficient,

$$\ln EVI_t(k,\tau) = \ln C(\tau) + \beta \ln I_t^2(k,\tau) + e_t. \tag{31}$$

Table 2 reports the slope estimates for each implied volatility time series. For all 40 series, the slope estimates are far away from the square-root hypothesis of $\beta = 0$, but is close to our proportional specification of $\beta = 1$, suggesting that the proportional volatility dynamics enjoys better empirical support than the square

root specification.

[Table 2 about here.]

5. Extracting economic states from implied and expected volatility surfaces

Under the proportional volatility dynamics specification, the time-t shape of the option implied volatility surface is governed by the time-t values of five covariates $(v_t, m_t, \rho_t, w_t, \eta_t)$, and we allow three additional covariates $(v_t^{\mathbb{P}}, m_t^{\mathbb{P}}, \rho_t^{\mathbb{P}})$ to capture the difference in the shape of the expected volatility surface. One particular feature of the model is that the shapes of the two volatility surfaces only depend on the levels of these state variables, but do not depend on the particular state dynamics specification. Therefore, the emphasis of the empirical analysis involves the extraction of the states from the two surfaces, without knowing the state dynamics. Based on this unique feature, we cast the model into a state-space form, where we treat the covariates as the hidden states and treat the observed option implied and expected volatility estimates as measurements with errors.

Among the eight covariates, four $(w_t, \eta_t, v_t, v_t^{\mathbb{P}})$ are constrained to be strictly positive, two (ρ_t and $\rho_t^{\mathbb{P}}$) are constrained to be between $(-1, 1)$. In defining the state vector X_t, we transform these covariates so that they can take values on the whole real line:

$$X_t \equiv \left[m_t, m_t^{\mathbb{P}}, \ln(w_t), \ln(\eta_t), \ln(v_t), \ln(v_t^{\mathbb{P}}), \ln\left(\frac{1+\rho_t}{1-\rho_t}\right), \ln\left(\frac{1+\rho_t^{\mathbb{P}}}{1-\rho_t^{\mathbb{P}}}\right) \right]^{\top}, \tag{32}$$

With the transformation, we assume that the state vector propagates as a random walk,

$$X_t = X_{t-1} + \sqrt{\Sigma_x}\varepsilon_t. \tag{33}$$

where the standardized error vector ε_t is normally distributed with zero mean and unit variance. We further assume that the covariance matrix is a diagonal matrix with distinct diagonal values so that the states can

have different degrees of variation but the movements are independent of each other.

In reality, the eight covariates represent eight different stochastic processes, which can follow much more complex dynamics than assumed in the state propagation equation (33). However, since their dynamics do not enter the pricing of the volatility surfaces, we leave them unspecified and use the simple random walk assumption to dictate the state propagation equation.

We define the measurement equations on the logarithm of the option implied and expected volatility estimates, assuming additive, normally distributed errors,

$$y_t = h(X_t) + \sqrt{\Sigma_y} e_t, \quad h(X_t) = \{\ln(I(X_t;k_j,\tau_j), \ln V(X_t;k_j,\tau_j)\}_{j=1}^{40} \tag{34}$$

where $y_t \in \mathbb{R}^{80}$ denotes the logarithm of the 40 implied volatility quotes and the 40 corresponding OEV estimates on date t, and $h(X_t)$ denotes the logarithm of the model value of the implied and expected volatility as a function of the states X_t, which can be solved from equations (10) and (25) in Propositions 3 and 4, respectively. By defining the measurement equations on the logarithms of the volatilities with additive, normally distributed errors, we guarantee the positivity of the volatilities. We assume that the additive pricing errors are iid normally distributed with error variance σ_I^2 for the 40 implied variance quotes and with error variance σ_V^2 for the 40 OEV series.

When the state-space model is Gaussian linear, the Kalman (1960) filter provides efficient forecasts and updates on the mean and covariance of the state and observations. Our state-propagation equations are constructed to be Gaussian and linear, but the measurement functions $h(X_t)$ are not linear in the state vector. We use the unscented Kalman filter (Wan and van der Merwe (2001)) to handle the nonlinearity.

The setup introduces ten auxiliary parameters that define the covariance matrices of the state propagation errors and the measurement errors. The relative magnitude of the state propagation error variance versus the measurement error variance controls the speed with which we update the states based on new observations. Intuitively, if the states vary a lot (large Σ_x) and the observations are accurate (small Σ_y), one would want

to update the states faster to better match the new observations. We choose these auxiliary parameters, and accordingly the optimal state updating speed, by minimizing the sum of squared forecasting errors in a quasi maximum likelihood setting.

6. Pricing performance and state dynamics analysis

We first examine the pricing performance of the model on the two volatility surfaces and then analyze the dynamic behaviors of the extracted states and their implications.

6.1. Pricing performance

Panel A of Table 3 reports the average pricing error on each volatility series. The pricing errors are defined as the difference between the observed volatility series and the model-generated values, in volatility percentage points. For the implied volatility surface, the most obvious average bias occurs at one-month maturity, where the observed implied volatilities are on average higher than the corresponding model values for far out-of-the-money options, but lower for at-the-money options. In essence, the model fails to fully capture the smile shape at short maturities. This deficiency comes mainly from the purely continuous price movement assumption. As shown in Carr and Wu (2003), continuous and discontinuous price dynamics generate very distinct behaviors for short-term out-of-the-money options. The data suggest that a jump component is needed to capture the short-term implied volatility smile. The average biases at longer maturities are less severe. At option maturities six months and longer, the model generates more negative skewness along the strike dimension than observed from the data. This bias is in part induced by the model's difficulty (and over compensation) in fitting the negative skewness at short maturities. The average biases on the expected volatility surface show less obvious structures.

[Table 3 about here.]

Panel B reports the explained variation on each series, defined as one minus the variance ratio of the model's pricing error to the original volatility series. The measure is analogous to the R-squared measure for a regression and captures the proportion of variation explained by the model. On average, the model explains 98 percent of the variation of the implied volatility series, and 80 percent of the variation of the expected volatility series. The lower explanatory power on the expected volatility series is expected as these series are noisy estimates of the true expected value based on historical movements. Across different strikes and maturities, the model fits the at-the-money implied volatility better than out-of-the-money implied volatilities, and fits the moderate-maturity volatility series better than series at very short or very long maturities. For reasons discussed above, the lowest explanatory power on implied volatilities are at the very short maturities.

In pricing the volatility surfaces, our model only depends on the current levels of the state variables, but does not depend on any fixed model parameters. The absence of fixed model parameters greatly simplifies model estimation and removes potential consistency issues encountered in model recalibration: A model with re-calibrated model parameters represents essentially a different model and thus generates different pricing and hedging implications from previous calibrations. Such consistency issues do not show up in our model as the pricing relation contains no fixed parameters. In our state-space approach to extract the state variables, we introduce ten auxiliary parameters to define the state propagation error variance and the measurement error variance. These variance estimates control the updating speed of the states based on new observations, and we use maximum likelihood estimation to determine the magnitudes of these parameters and accordingly the optimal updating speed. Altering the state propagation equation specification and/or the variance estimates does not induce consistency issues for the pricing relation, but can nevertheless change the state updates and thus change our valuation. In principle, the optimal updating speed can change with market conditions. For example, if the observations are becoming more accurate over time and/or the market starts to show larger variations, the optimal updating speed should become faster to put more weight on the most recent observation. To gauge how sensitive the model performance is to these auxiliary parameter estimates, we perform an out-of-sample analysis: We only use the first three years data (1997 to 1999) to

perform the maximum likelihood estimation and use the estimated parameters to filter the states over the whole sample period. We measure the correlation between the two sets of states based on the two sets of parameter estimates to determine how the auxiliary parameter estimates alter the values and movements of the extracted states. The correlation estimates are the highest at 99.6% for the two variance rates $(v_t, v_t^{\mathbb{P}})$, around 98% for $m_t^{\mathbb{P}}$ and $\rho_t^{\mathbb{P}}$, around 93% for m_t and w_t, and 86% for ρ_t. The lowest correlation is between the two sets of η_t estimates at 80%. The high correlations between the two sets of the extracted states, especially for the variance rates, suggest that the filtered states are not very sensitive to small variations in the auxiliary parameter estimates. The pricing performance is also similar under the two sets of auxiliary parameter estimates. The average explained variation on the implied volatility surface is around 98% based on both sets of parameters. The explained variation on the expected volatility surface experiences some deterioration from 80% based on the full-sample estimates to 67% based on the three-year sample estimates.

6.2. The time variation of short and long-term implied and expected volatilities

Figure 2 plots the time series of the instantaneous volatility $(\sqrt{v_t})$, with the solid line extracted from the options implied volatility surface and the dashed line from the expected volatility surface. The time series variation of the two instantaneous volatility series follows closely the time series variation of at-the-money implied and expected volatilities plotted in Figure 1. Due to the backward looking nature of the expected volatility estimation, the instantaneous volatility series extracted from the expected volatility surface seems to lag behind the solid line extracted from the implied volatility surface. Furthermore, during non-eventful time periods such as the bull market run from 2004 to 2007 and the most recent run since 2012, the solid line extracted from the implied volatility surface stays above the dashed line extracted from the expected volatility surface, but the two lines tend to cross in the aftermath of a volatility spike.

[Fig. 2 about here.]

Figure 3 plots the time series of the instantaneous drift processes under both the risk-neutral measure (m_t, solid line) and the statistical measure ($m_t^{\mathbb{P}}$, dashed line). The risk-neutral drift process dictates the term structure shape of the at-the-money implied volatility, whereas the statistical drift process governs the term structure shape of the expected volatility. The solid line stays above zero most of the time, except during the 2002 recession and the 2008 financial crisis. The on average positive risk-neutral drift suggests that the at-the-money implied volatility term structure is upward sloping most of the time. By contrast, the dashed line stays negative most of the time, suggesting that the expected volatility computed from the historical sample paths has a downward sloping term structure most of the time. The term structure difference reflects the volatility risk premium. The difference is particular large around the two financial crises (1998 and 2008) and during the 2003 recession.

[Fig. 3 about here.]

6.3. Stochastic variation of the return-volatility correlation and the volatility skew

Figure 4 plots the time series of the instantaneous correlation between the SPX index return and its volatility, again with the solid line extracted from the implied volatility surface and the dashed line from the expected volatility surface. The solid line stays strongly negative over the whole sample period, with a maximum of -0.47 and a minimum close to -0.98. These highly negative correlation estimates reflect the persistently negative skew observed from the implied volatility surface. By contrast, the dashed line varies much closer to zero and can switch signs, suggesting that the expected volatility surface is not always negatively skewed.

[Fig. 4 about here.]

Interestingly, the two financial crises during our sample period (the 1998 Asian crises and the 2008 financial meltdown) are both preceded by a divergence between the two correlation estimates, with the dashed line going above zero while the solid line reaching its most negative level. Before the financial crisis,

the options market becomes increasingly worried as shown by the extremely negative implied volatility skew. At the same time, the index return dynamics start to show abnormal behaviors as the return volatility starts increasing with rising index level, whereas at normal times return volatility tends to decline with rising index level. These behaviors, combined with the volatility spikes, seem to precede the upcoming of the financial crisis. By contrast, during the mild recession of 2003, although the volatility level also spiked up, the option implied volatility skew was not particularly negative, and the return-volatility correlation extracted from the expected volatility surface stayed negative. Thus, for future dynamic model designs, it is important to build different mechanisms for different types of volatility spikes.

6.4. The time series variation of volatility of volatilities

Figure 5 plots the extracted time series of the volvol process in panel A. The volvol estimates tend to be high when the volatility levels are high. A high volvol coefficient increases the convexity of the volatility smile along the strike dimension.

[Fig. 5 about here.]

Panel B of Figure 5 plots the time series of the maturity decay coefficient (η_t), which lowers the variation for long-term implied volatility series. The extracted series are stable except during the 2002 recession, when the estimates become much higher. This recession period seems to be unique in its behaviors, when the short-term volatility is high, the term structure for both implied and expected volatilities are downward sloping, and the short-term volatility varies much more than long-term volatilities. While the short-term volatility is high during both financial crises and during this recession, the long-dated implied volatilities do not go up as much during the recession, suggesting that investors are much less worried about this recession than about the financial crises.

The time-series variations of the different state variables depicted in Figures 2-5 provide guidance for future structural model designs. The variation of the instantaneous variance rate can be accommodated

by most stochastic volatility models. The return-variance correlation (hence volatility skew) variation can be accommodated by a two-factor volatility structure as in Carr and Wu (2007) on currency options and Christoffersen, Heston, and Jacobs (2009) on equity options. What is the most interesting and challenging is to come up with risk and risk preference specifications that can accommodate the risk premium variations as shown in the different term structure and skew variations in Figures 3 and 4 extracted from the two volatility surfaces.

6.5. Risk premiums and excess return predictions

Under the pricing kernel assumption in (26), we can identify the market pricing of the variance risk γ_t from the difference between the statistical and risk-neutral drift processes (m_t and $m_t^{\mathbb{P}}$), as shown in equation (27), $\gamma_t = (m_t^{\mathbb{P}} - m_t)/(w_t\sqrt{v_t})$. The market pricing of the variance risk contributes to the instantaneous return risk premium through the return-variance correlation by ρ_t as shown in equation (28). We label this component of the return risk premium as RRP. In this section, we analyze whether this return risk premium component has any actual predictive power of future excess returns on the SPX index. For comparison, we consider two benchmarks. One is the VIX index squared, which approximates the one-month variance swap rate of the S&P 500 index.[7] The VIX index is regarded as a fear gauge in the industry and has the potential to capture not only the risk level variation, but also risk preference changes over time. The second benchmark is the difference between VIX squared and the one-month realized variance ($VIX^2 - RV$), which is often labeled as the variance risk premium (VRP) and has been used to predict future stock returns by, among others, Bollerslev, Tauchen, and Zhou (2009).

To obtain an empirical estimate of the return risk premium, we regress future excess returns on the stock index on each of the three predictors,

$$ER_{t,j} = a_i + b_i x_t^i + e_{t,j}^i, \tag{35}$$

[7] See Carr and Wu (2006) for a detailed description of this index and its behaviors. We thank the anonymous referee for this suggestion.

where x_t^i denotes the time-t value of the ith predictor (VIX, VRP, or RRP) and $ER_{t,j}$ denotes the annualized future index excess return from time t to j days ahead,[8]

$$ER_{t,j} = \frac{365}{j}\sum_{s=1}^{j} r_{t+s} - R_t^f \tag{36}$$

with r_t denoting the daily return at time t, and R_t^f denotes the risk-free rate at time t, which we proxy with the US LIBOR rate of the corresponding horizon. The time series of the SPX, SPY, VIX, and the LIBOR rates are obtained from Bloomberg. The 30-day realized variance is computed from the historical return data on the SPX index. In computing the return risk premium (RRP) according to (28), we use the variance rate and return-variance correlation extracted from the expected volatility surface.

We perform an out-of-sample exercise based on the predictive regression in (35). Starting from January 2000 (three years from the starting date of the data sample), at each date t, we estimate each regression using data up to that point and make predictions for future excess returns from that point forward. To reduce the impact of data outliers on the forecasting results, we follow Campbell and Thompson (2008) and constrains all annualized excess return forecasts to be within $(0, 20\%)$. We compute the out-of-sample forecasting error as the difference between the future realized excess return and the forecasted excess return. As in Welch and Goyal (2008), the forecasting performance of each measure is compared with the historical average of the future excess return up to that point t,

$$\overline{ER}_{t,j} = \frac{1}{t-j}\sum_{s=j+1}^{t} ER_{s-j,j}. \tag{37}$$

The overall out-of-sample forecasting performance of each predictor is measured by the sum squared fore-

[8]Since the SPX index level does not adjust for dividend payments, log index level difference does not fully capture the returns from investing in the S&P 500 stocks. To obtain a return series that properly adjusts for dividend payments, we use the spider ETF (SPY) adjusted-price time series instead for the return calculation. The result difference from using SPX log index level difference is very small.

casting error (SSFE) over N out-of-sample observations,

$$SSFE_{i,j,N} = \sum_{t=1}^{N} \left(ER_{t,j} - \widehat{ER}^{i}_{t,j} \right)^2, \tag{38}$$

where $\widehat{ER}^{i}_{t,j}$ denotes the out-of-sample forecast from predictor i on excess return $ER_{t,j}$. Using the SSFE on the historical average as the benchmark,

$$SSFE_{0,j,N} = \sum_{t=1}^{N} \left(ER_{t,j} - \overline{ER}_{t,j} \right)^2, \tag{39}$$

we measure the relative performance of each predictor via an out-of-sample R-squared measure as in Rapach and Zhou (2013),

$$R^2_{i,j} = 1 - SSFE_{i,j,N} / SSFE_{0,j,N}. \tag{40}$$

A positive R-squared estimate indicates that the predictor outperforms the historical average benchmark.

To compute the return risk premium (RRP) from the VGVV model, we need to estimate the model to obtain the auxiliary parameters that control the state updates. For the out-of-sample exercise, we use the parameters estimated from the first three years of sample without further updating these estimates. As we have shown earlier, the extracted states are not particularly sensitive to the small variations in the auxiliary parameters. The extracted states are similar whether we re-estimate the model or not. In addition to performing forecasting regressions, given the structural meaning of RRP being the return risk premium, we can also directly use the RRP estimates as the forecast for future excess returns by setting the regression intercept to zero and the slope to one.

Table 4 reports the out-of-sample R-squared for the four sets of forecasts. The VIX itself can hardly outperform the historical average as the R-square estimates are close to zero at all forecasting horizons. The variance risk premium regression can outperform the historical average, with the best performance coming at quarterly forecasting horizon. The performance starts to deteriorate at longer forecasting horizons, potentially because the variance risk premium is constructed using only short-term option contracts. The

return risk premium regression underperforms the historical average at short forecasting horizons, but outperforms increasingly more as the forecasting horizon increases. The short-horizon underperformance is likely related to the model's difficulty in capturing the short-term behavior of the implied volatility surface. Its long-horizon outperformance, on the other hand, shows the benefit of extracting information from the two volatility surfaces.

More striking is the strong performance of directly applying the RRP as the excess return forecast. Since the RRP represents only part of the return risk premium, without accounting for the risk premium on the independent return risk, the RRP estimate can be regarded as a conservative estimate of the return risk premium. The forecasting regression can be used to adjust the scale, but it also brings in estimation error, especially out of sample. By discarding the regression and directly applying the structural implication of the model, one can avoid the noise introduced by the empirical fitting and generate much superior out-of-sample forecasting performance.

As in Rapach and Zhou (2013), Figure 6 plots the cumulative squared forecasting error difference,

$$CFED_{i,j,n} = SSFE_{0,j,n} - SSFE_{i,j,n}, \quad n = 1,2,\cdots,N. \tag{41}$$

The three lines in each panel denote three selected forecasting horizons at three months (solid line), six months (dashed line), and 12 months (dash-dotted line). Panel A plots the cumulative out-of-sample performance of the VIX squared regression. The performance is worse than the historical average at three-month forecasting horizon and only becomes slightly better at longer horizons. The drastic deterioration in 2009 is caused by the extreme spike in the implied volatility. Panel B shows that the VRP generates reasonably good out-of-sample forecasting performance at the quarterly forecasting horizon. The forecasting performance mainly come from the early 2000s and the 2008 financial crisis, but otherwise has little power during the long stretches between 2003-2008 and after 2010. At longer forecasting horizons, the VRP prediction deteriorates and can no longer outperform the historical average.

[Fig. 6 about here.]

Panel C plots the cumulative performance of the RRP regression. The regression is not stable at short forecasting horizons, but generates more consistent performance at longer horizons. In particular, the outperformance mainly comes from the early 2000 period and during the 2008 financial crises. The performance shows deterioration during the 2003-2008 stretch and after 2010. However, when we discard the regression and directly apply the RRP as the excess return forecast, the cumulative performance in panel D becomes much more uniform over different sample periods, especially at long forecasting horizons.

The out-of-sample predictive power of RRP highlights the information content of the two volatility surfaces in extracting risk dynamics and risk premiums. Several studies, e.g., Bollerslev, Tauchen, and Zhou (2009), Xing, Zhang, and Zhao (2010), Cremers and Weinbaum (2010), and Bakshi, Panayotov, and Skoulakis (2011) have found equity options to be informative of future stock returns. More recently in a seminal paper, Ross (2014) shows that under certain assumptions, one can identify both the risk-neutral and statistical dynamics, as well as the pricing kernel that links the two, using only information in the option implied volatility surface alone. Several researchers, e.g., Borovicka, Hansen, and Scheinkman (2014), Hansen and Scheinkman (2014), Walden (2014), Audrino, Huitema, and Ludwig (2014), Qin and Linetsky (2014), and Qin and Linetsky (2015), explore the implications of the underlying assumptions, potential extensions, and empirical performance. How to integrate the different perspectives to balance the need for structural assumptions and data constitutes a deeply interesting research direction.

7. Concluding remarks

Despite the fact that the BMS model assumptions are apparently violated, both practitioners and academics have accustomed to use the BMS implied volatility surface to represent the information in option contracts. Quoting a positive implied volatility for an option contract directly excludes arbitrage between this option and the underlying security, adding further attraction to the implied volatility quoting convention. Further-

more, delta hedging in practice is mostly based on the BMS delta at the implied volatility level. Despite much research, the literature has yet to propose an alternative delta ratio that outperforms the BMS delta in all practical situations.

Given this heavy reliance on the BMS implied volatility surface, it would be ideal if one can directly model the implied volatility dynamics and derive direct implications on the shape of the implied volatility surface. In this paper, we propose a new modeling framework that does just that. Given a one-factor pure diffusion dynamics on the implied volatility surface, we transform the dynamic no-arbitrage constraint between the underlying stock, a basis option, and any other option contract into a simple algebraic constraint on the shape of the current implied volatility surface. In particular, under a proportional volatility dynamics specification, the whole shape of the implied volatility surface becomes the solution to a simple quadratic equation. As a result, the numerical burden for option pricing and model estimation is dramatically reduced compared to the standard option pricing literature.

Corresponding to implied volatility surface, we also propose a new concept that both realized and statistically expected volatilities estimated from price sample paths can vary with the strike and maturity of a reference option contract. The idea is that although taking any option position with delta hedging exposes the investor to the variance risk during the life of the option, the weighting on the sample path differs for different option contracts. As such, one can estimate a realized and expected volatility corresponding to the risk exposure of each option contract. With this new concept, one can directly measure the volatility risk premium embedded in each option contract as the difference between this contract's implied and expected volatility. Furthermore, the current shape of the expected volatility surface is analogously governed by its future statistical dynamics.

A unique feature of our modeling framework is that by modeling the whole volatility surface, we only need to know the current level of the drift and diffusion of the volatility surface to determine the current shape of the surface. How the drift and diffusion processes vary in the future does not affect the current volatility shape. This unique feature allows us to specify a model that has many state variables but with no

fixed model parameters. The high dimensional state space allows the model to fit the observed volatility surface well without extra fudging, whereas the absence of fixed model parameters drastically simplify the model estimation process. This feature also makes the model as a perfect complement to traditional fully parametric option pricing models. In particular, volatility surface valuations from a chosen parametric model can be directly used as the starting point by assuming that market volatility observations converge to the model valuation via an error-correction specification.

Our new theoretical framework opens ample ground for future research. First, we use a simple proportional volatility dynamics for illustration. For future research, one can explore many different specifications, many of which can lead to extremely simple analytical solutions for the volatility surface. Second, the concept of option specific realized and expected volatility opens a whole new area of empirical research on option-specific volatility forecasting. Third, our current framework assumes diffusion return dynamics and a one-factor diffusion volatility surface dynamics, future research can investigate on how to accommodate discontinuous price and volatility movements and multiple volatility risk factors.

Appendix

A. Proof of Proposition 1

First, we form a portfolio between any put option at (K,T) and the basis call option at (K_0,T_0) to neutralize the exposure on the volatility risk dZ_t:

$$B_\sigma(S_t,I_t(K,T),t)\omega_t(K,T)-N_t^c B_\sigma(S_t,I_t(K_0,T_0),t)\omega_t(K_0,T_0)=0. \tag{42}$$

The two-option portfolio with no dZ_t exposure will in general be exposed to dW_t. As a result, a three-asset portfolio with N_t^S shares of the underlying stock is determined by requiring delta neutrality:

$$B_S(S_t,I_t(K,T),t)-N_t^c(1+B_S(S_t,I_t(K_0,T_0),t))-N_t^S=0. \tag{43}$$

Since shares have no vega, this three-asset portfolio retains zero exposure to dZ_t and by construction has zero exposure to dW_t.

By Ito's lemma, each option in this portfolio has risk-neutral drift given by:

$$B_t+\mu_t B_\sigma+\frac{1}{2}v_t S_t^2 B_{SS}+\rho_t\omega_t\sqrt{v_t}S_t B_{S\sigma}+\frac{\omega_t^2}{2}B_{\sigma\sigma}. \tag{44}$$

No dynamic arbitrage and no rates imply that both option drifts must vanish, leading to the fundamental partial differential equation (PDE):

$$-B_t=\mu_t B_\sigma+\frac{1}{2}v_t S_t^2 B_{SS}+\rho_t\omega_t\sqrt{v_t}S_t B_{S\sigma}+\frac{1}{2}\omega_t^2 B_{\sigma\sigma}. \tag{45}$$

This fundamental PDE applies to any option, as long as we require that this option allow no dynamic arbitrage relative to the basis option at (K_0,T_0), the stock, and cash.

B. Proof of Proposition 2

The BMS value function $B(S_t, I_t, t)$ is well known. So are its various partial derivatives:

$$\begin{aligned} B_t &= -\tfrac{\sigma^2}{2} S^2 B_{SS}, & B_\sigma &= \sigma(T-t) S^2 B_{SS}, \\ SB_{\sigma S} &= \left(\tfrac{\ln(K/S)}{\sigma\sqrt{T-t}} + \tfrac{\sigma\sqrt{T-t}}{2} \right) \sqrt{T-t} S^2 B_{SS}, & B_{\sigma\sigma} &= \left(\tfrac{\ln(K/S)^2}{\sigma^2(T-t)} - \tfrac{\sigma^2(T-t)}{4} \right) (T-t) S^2 B_{SS} \end{aligned} \tag{46}$$

where dollar gamma $S^2 B_{SS}$ is the common denominator of all the partial derivatives.

Evaluate these partial derivatives at $(S, \sigma, t) = (S_t, I_t(K,T), t)$, substitute them into the fundamental PDE in (6), and divide both sides of the equation by the dollar gamma $S_t^2 B_{SS}$ while noting that the dollar gamma is strictly positive at $T > t$, we transform the PDE into an algebraic equation,

$$\begin{aligned} 0 = \quad & \tfrac{1}{2} I_t^2(K,T) - \mu_t I_t(K,T)(T-t) - \tfrac{1}{2} v_t - \rho_t \omega_t \sqrt{v_t} \left(\tfrac{\ln(K/S)}{I_t(K,T)\sqrt{T-t}} + \tfrac{I_t(K,T)\sqrt{T-t}}{2} \right) \sqrt{T-t} \\ & - \tfrac{1}{2} \omega_t^2 \left[\left(\tfrac{\ln(K/S)^2}{I_t^2(K,T)(T-t)} - \tfrac{I_t^2(K,T)(T-t)}{4} \right) \right] (T-t). \end{aligned} \tag{47}$$

Re-define the implied volatility surface as a function of the relative strike $k \equiv \ln(K/S)$ and time to maturity $\tau \equiv T - t$, and re-arrange, we obtain the algebraic representation in (7).

C. Proof of Proposition 3

Under the proportional dynamics specification, the drift and volvol of the implied volatility process are

$$\mu_t = e^{-\eta_t (T-t)} m_t I_t(K,t), \quad \omega_t = e^{-\eta_t (T-t)} w_t I_t(K,T). \tag{48}$$

Substitute μ_t and ω_t in (48) into the no dynamic arbitrage restriction (7), fix the relative strike $k \equiv \ln(K/S)$ and time to maturity $\tau \equiv T - t$, and re-arrange terms, we can transform the no-arbitrage algebraic constraint into a quadratic function of $I_t^2(k, \tau)$ as shown in (10).

D. Proof of Proposition 4

Starting with the statistical dynamics in (23) and (24), if we assume that the variance risk is not priced, implied volatility would converge to the statistical expected volatility, $I_t(k,\tau) = V_t(k,\tau)$, and equation (23) would also become the risk-neutral dynamics for the implied volatility. In this hypothetical case, Proposition 3 shows that the shape of the implied variance surface as a function of relative strike and maturity is determined by the following quadratic equation,

$$\begin{aligned} 0 \quad = \quad & \tfrac{1}{4} e^{-2\eta_t \tau} w_t^2 \tau^2 I_t^4(k,\tau) + \left(1 + 2e^{-\eta_t \tau} m_t^{\mathbb{P}} \tau - e^{-\eta_t \tau} w_t \rho_t^{\mathbb{P}} \sqrt{v_t^{\mathbb{P}}} \tau\right) I_t^2(k,\tau) \\ & - \left(v_t^{\mathbb{P}} + 2e^{-\eta_t \tau} w_t \rho_t^{\mathbb{P}} \sqrt{v_t^{\mathbb{P}}} k + e^{-2\eta_t \tau} w_t^2 k^2\right). \end{aligned} \tag{49}$$

Furthermore, since $I_t(k,\tau) = V_t(k,\tau)$ under zero risk premium, the same quadratic equation in (49) also determines the shape of $V_t(k,\tau)$.

With non-zero risk premium, implied variance differs from the expected realized variance and the shape of the implied volatility is determined by a different quadratic equation in (10). In this case, the equation (49) only determines the shape of the expected volatility $V_t(k,\tau)$, with I_t being replaced by V_t in the equation.

References

Audrino, F., Huitema, R., Ludwig, M., 2014. An empirical analysis of the ross recovery theorem. Working paper. University of St. Gallen and University of Zurich.

Avellaneda, M., Zhu, Y., 1998. A risk-neutral stochastic volatility model. International Journal of Theoretical and Applied Finance 1, 289–310.

Baele, L., Driessen, J., Londono, J. M., Spalt, O. G., 2014. Cumulative prospect theory and the variance premium. Working paper. Tilburg University.

Bakshi, G., Kapadia, N., 2003a. Delta-hedged gains and the negative market volatility risk premium. Review of Financial Studies 16, 527–566.

Bakshi, G., Kapadia, N., 2003b. Volatility risk premium embedded in individual equity options: Some new insights. Journal of Derivatives 11, 45–54.

Bakshi, G., Panayotov, G., Skoulakis, G., 2011. Improving the predictability of real economic activity and asset returns with forward variances inferred from option portfolios. Journal of Financial Economics 100, 475–495.

Black, F., Scholes, M., 1973. The pricing of options and corporate liabilities. Journal of Political Economy 81, 637–654.

Bollerslev, T., Tauchen, G., Zhou, H., 2009. Expected stock returns and variance risk premia. Review of Financial Studies 22, 4463–4492.

Borovicka, J., Hansen, L. P., Scheinkman, J. A., 2014. Misspecified recovery. Working paper. New York University, University of Chicago, Columbia University.

Campbell, J. Y., Thompson, S. B., 2008. Predicting excess stock returns out of sample: Can anything beat the historical average?. Review of Financial Studies 21, 1509–1531.

Carr, P., Madan, D., 2002. Towards a theory of volatility trading. In: Jarrow, R. (Eds.), Risk Book on Volatility. Risk, New York.

Carr, P., Sun, J., 2007. A new approach for option pricing under stochastic volatility. Review of Derivatives Research 10, 87–150.

Carr, P., Wu, L., 2003. What type of process underlies options? A simple robust test. Journal of Finance 58, 2581–2610.

Carr, P., Wu, L., 2006. A tale of two indices. Journal of Derivatives 13, 13–29.

Carr, P., Wu, L., 2007. Stochastic skew in currency options. Journal of Financial Economics 86, 213–247.

Carr, P., Wu, L., 2009. Variance risk premiums. Review of Financial Studies 22, 1311–1341.

Carr, P., Wu, L., 2010. Stock options and credit default swaps: A joint framework for valuation and estimation. Journal of Financial Econometrics 8, 409–449.

Christoffersen, P. F., Heston, S. L., Jacobs, K., 2009. The shape and term structure of the index option smirk: Why multifactor stochastic volatility models work so well. Management Science 55, 1914–1932.

Cremers, M., Weinbaum, D., 2010. Deviations from put-call parity and stock return predictability. Journal of Financial and Quantitative Analysis 45, 335–367.

Daglish, T., Hull, J., Suo, W., 2007. Volatility surfaces: Theory, rules of thumb, and empirical evidence. Quantitative Finance 7, 507–524.

Drechsler, I., Yaron, A., 2011. What's vol got to do with it. Review of Financial Studies 24, 1–45.

Egloff, D., Leippold, M., Wu, L., 2010. The term structure of variance swap rates and optimal variance swap investments. Journal of Financial and Quantitative Analysis 45, 1279–1310.

Engle, R., Figlewski, S., 2015. Modeling the dynamics of correlations among implied volatilities. Review of Finance 19, 991–1018.

Engle, R. F., Granger, C. W., 1987. Cointegration and error correction: Representation, estimation and testing. Econometrica 55, 251–276.

Fengler, M. R., 2005. Semiparametric Modeling of Implied Volatility. Springer-Verlag, Berlin.

Foresi, S., Wu, L., 2005. Crash-o-phobia: A domestic fear or a worldwide concern?. Journal of Derivatives 13, 8–21.

Gatheral, J., 2006. The Volatility Surface: A Practitioner's Guide. John Wiley & Sons, New Jersey.

Hafner, R., 2004. Stochastic Implied Volatility: A Factor-Based Model. Springer-Verlag, Berlin.

Hansen, L. P., Scheinkman, J. A., 2014. Stochastic compounding and uncertain valuation. In: Weyl, E. G., Glaeser, E. L., Santos, T. (Eds.), Après le Déluge: Finance and the Common Good after the Crisis. University of Chicago Press.

Heath, D., Jarrow, R., Morton, A., 1992. Bond pricing and the term structure of interest rates: A new technology for contingent claims valuation. Econometrica 60, 77–105.

Heston, S. L., 1993. Closed-form solution for options with stochastic volatility, with application to bond and currency options. Review of Financial Studies 6, 327–343.

Hodges, H. M., 1996. Arbitrage bounds of the implied volatility strike and term structures of European-style options. Journal of Derivatives 3, 23–32.

Hull, J., White, A., 1987. The pricing of options on assets with stochastic volatilities. Journal of Finance 42, 281–300.

Jiang, G., Tian, Y., 2005. Model-free implied volatility and its information content. Review of Financial Studies 18, 1305–1342.

Kalman, R. E., 1960. A new approach to linear filtering and prediction problems. Transactions of the ASME–Journal of Basic Engineering 82, 35–45.

Ledoit, O., Santa-Clara, P., 1998. Relative pricing of options with stochastic volatility. manuscript. UCLA.

Merton, R. C., 1973. Theory of rational option pricing. Bell Journal of Economics and Management Science 4, 141–183.

Mueller, P., Vedolin, A., Yen, Y.-M., 2012. Bond variance risk premia. Working paper. London School of Economics.

Polimenis, V., 2006. Skewness correction for asset pricing. Working paper. University of California Riverside.

Qin, L., Linetsky, V., 2014. Long term risk: A martingale approach. Working paper. Northwestern University.

Qin, L., Linetsky, V., 2015. Positive eigenfunctions of markovian pricing operators: Hansen-Scheinkman factorization and Ross recovery and long-term pricing. Working paper. Northwestern University.

Rapach, D. E., Zhou, G., 2013. Forecasting stock returns. In: Elliott, G., Timmermann, A. (Eds.), Handbook of Economic Forecasting. Elsevier.

Ross, S., 2014. The recovery theorem. Journal of Finance forthcoming.

Schonbucher, P. J., 1999. A market model for stochastic implied volatility. Philosophical Transactions: Mathematical, Physical and Engineering Sciences 357, 2071–2092.

Walden, J., 2014. Recovery with unbounded diffusion processes. Working paper. University of California, Berkeley.

Wan, E. A., van der Merwe, R., 2001. The unscented Kalman filter. In: Haykin, S. (Eds.), Kalman Filtering and Neural Networks. Wiley & Sons Publishing, New York.

Welch, I., Goyal, A., 2008. A comprehensive look at the empirical performance of equity premium prediction. Review of Financial Studies 21, 1455–1508.

Wu, L., 2006. Dampened power law: Reconciling the tail behavior of financial asset returns. Journal of Business 79, 1445–1474.

Xing, Y., Zhang, X., Zhao, R., 2010. What does the individual option volatility smirk tell us about future equity returns?. Journal of Financial and Quantitative Analysis 45, 641–662.

Zhang, B. Y., Zhou, H., Zhu, H., 2009. Explaining credit default swap spreads with the equity volatility and jump risks of individual firms. Review of Financial Studies 22, 5099–5131.

Table 1
Average behavior of option volatilities and volatility risk premiums on the S&P 500 index
Entries report the sample average of the S&P 500 index option implied volatilities in panel A, the corresponding option expected volatilities (OEV) in panel B, the average expected-implied volatility difference in panel C, and the annualized information ratio from long the option contracts in panel D. The statistics are computed based on 40 over-the-counter implied volatility quotes series on the S&P 500 index options at a matrix grid of five relative strikes (K/S) and eight time to maturities (τ, in months). The corresponding OEV series are estimated based on the SPX index sample path over the same time period. The implied and expected volatility series are sampled weekly from January 8, 1997 to October 29, 2014, 930 weekly observations for each series.

K/S	0.8	0.9	1.0	1.1	1.2	0.8	0.9	1.0	1.1	1.2
Maturity	*A. Average option implied volatility*					*B. Average option expected volatility*				
1	34.88	26.81	19.19	15.47	15.31	20.48	20.08	17.56	17.57	18.56
3	30.02	24.79	19.87	16.25	14.92	20.04	19.35	17.59	15.98	16.20
6	27.85	24.00	20.37	17.35	15.55	20.09	19.08	17.72	16.41	15.59
12	26.34	23.55	20.90	18.54	16.74	19.62	18.92	18.20	17.28	16.35
24	25.56	23.54	21.63	19.88	18.37	18.92	18.37	17.96	17.61	17.21
36	25.49	23.84	22.29	20.85	19.56	18.85	18.19	17.75	17.45	17.27
48	25.66	24.24	22.92	21.69	20.56	18.68	17.88	17.43	17.22	17.12
60	25.94	24.69	23.52	22.44	21.43	18.42	17.73	17.08	16.82	16.71
	C. Average volatility risk premium					*D. Annualized information ratio*				
1	-14.40	-6.73	-1.63	2.10	3.25	-1.47	-2.57	-1.93	2.04	3.04
3	-9.98	-5.44	-2.28	-0.27	1.28	-1.62	-1.88	-1.52	-0.55	0.36
6	-7.75	-4.91	-2.65	-0.93	0.03	-1.22	-1.36	-1.14	-0.66	-0.45
12	-6.72	-4.63	-2.70	-1.26	-0.39	-0.97	-0.90	-0.73	-0.53	-0.31
24	-6.64	-5.18	-3.67	-2.27	-1.16	-0.76	-0.67	-0.54	-0.40	-0.28
36	-6.64	-5.65	-4.54	-3.41	-2.29	-0.59	-0.52	-0.45	-0.36	-0.28
48	-6.98	-6.36	-5.49	-4.47	-3.44	-0.49	-0.46	-0.40	-0.35	-0.29
60	-7.52	-6.97	-6.45	-5.62	-4.72	-0.45	-0.44	-0.40	-0.36	-0.31

Table 2
Constant elasticity of variance dependence of implied volatility dynamics
For each implied volatility series, we first estimate an exponentially weighted variance series on the weekly implied volatility changes, and then regress the logarithm of this variance estimator against the logarithm of the implied variance level. The regression slope captures the power dependence of the implied volatility variance on the implied variance level, β. Entries report the regression estimates (and standard errors in parentheses) for this power coefficient for each implied volatility series.

Maturity$\backslash(K/S)$	0.8	0.9	1.0	1.1	1.2
1	1.12 (0.03)	1.11 (0.03)	0.79 (0.02)	1.20 (0.03)	1.55 (0.04)
3	1.25 (0.03)	1.13 (0.03)	0.90 (0.02)	1.03 (0.03)	1.37 (0.04)
6	1.34 (0.03)	1.21 (0.03)	1.02 (0.03)	1.01 (0.03)	1.29 (0.03)
12	1.42 (0.03)	1.28 (0.03)	1.12 (0.03)	1.04 (0.03)	1.17 (0.03)
24	1.49 (0.03)	1.37 (0.03)	1.23 (0.03)	1.14 (0.03)	1.17 (0.03)
36	1.55 (0.03)	1.44 (0.03)	1.32 (0.03)	1.22 (0.03)	1.20 (0.03)
48	1.58 (0.04)	1.48 (0.04)	1.37 (0.04)	1.27 (0.04)	1.21 (0.04)
60	1.58 (0.04)	1.49 (0.04)	1.40 (0.04)	1.30 (0.04)	1.22 (0.04)

Table 3
Model pricing performance on SPX option implied and expected volatilities
Entries in panel A report the average pricing error of the model on each volatility series. The pricing error is defined as the difference between the observed volatility series and their corresponding model values, in volatility percentage points. Entries in panel B report the model's explained variation, defined as one minus to ratio of the pricing error variance to the variance of the regional volatility series. For each measure, the last row reports the grand average of the statistic for the 40 implied and 40 expected volatility series, respectively.

	Implied volatility surface					Expected volatility surface				
Maturity$\backslash(K/S)$	0.8	0.9	1.0	1.1	1.2	0.8	0.9	1.0	1.1	1.2
A. Average pricing error										
1	5.80	2.62	-0.97	-1.56	0.44	-1.92	0.83	-0.01	0.23	0.36
3	1.15	0.64	-0.40	-0.98	-0.18	-1.28	0.96	0.76	-0.64	-1.21
6	-0.76	-0.16	-0.12	-0.23	0.06	-0.28	1.37	1.42	0.33	-1.20
12	-1.88	-0.70	-0.04	0.25	0.44	0.12	1.70	2.19	1.49	0.02
24	-2.03	-0.89	-0.15	0.28	0.51	-0.16	0.98	1.50	1.39	0.69
36	-1.46	-0.62	-0.06	0.28	0.45	-0.14	0.44	0.71	0.65	0.34
48	-0.55	-0.02	0.31	0.50	0.57	0.02	0.12	0.19	0.20	0.08
60	0.53	0.79	0.91	0.95	0.91	0.33	0.27	0.01	-0.05	-0.12
Average:			**0.11**					**0.32**		
B. Explained variation										
1	0.90	0.95	0.96	0.97	0.90	0.77	0.71	0.73	0.69	0.78
3	0.97	0.99	0.99	0.99	0.99	0.88	0.90	0.91	0.88	0.80
6	0.97	0.99	0.99	0.99	0.99	0.87	0.90	0.88	0.80	0.71
12	0.98	0.99	0.99	0.98	0.98	0.75	0.77	0.73	0.67	0.56
24	0.97	0.99	0.99	0.99	0.98	0.80	0.81	0.78	0.76	0.68
36	0.96	0.99	0.99	0.99	0.98	0.85	0.89	0.88	0.84	0.77
48	0.95	0.98	0.99	0.99	0.99	0.86	0.89	0.91	0.88	0.80
60	0.93	0.96	0.98	0.98	0.97	0.74	0.78	0.78	0.76	0.72
Average:			**0.98**					**0.80**		

Table 4
Out-of-sample return forecasting R-squared
Entries report the out-of-sample return forecasting R-squared, defined as one minus the ratio of sum squared forecasting error of each method to that of the historical average benchmark.

Horizon	1	3	6	9	12
VIX regression	0.001	-0.017	0.008	0.007	0.010
VRP regression	0.010	0.040	0.028	0.004	0.002
RRP regression	-0.009	-0.015	0.054	0.071	0.086
RRP direct	0.017	0.061	0.159	0.223	0.260

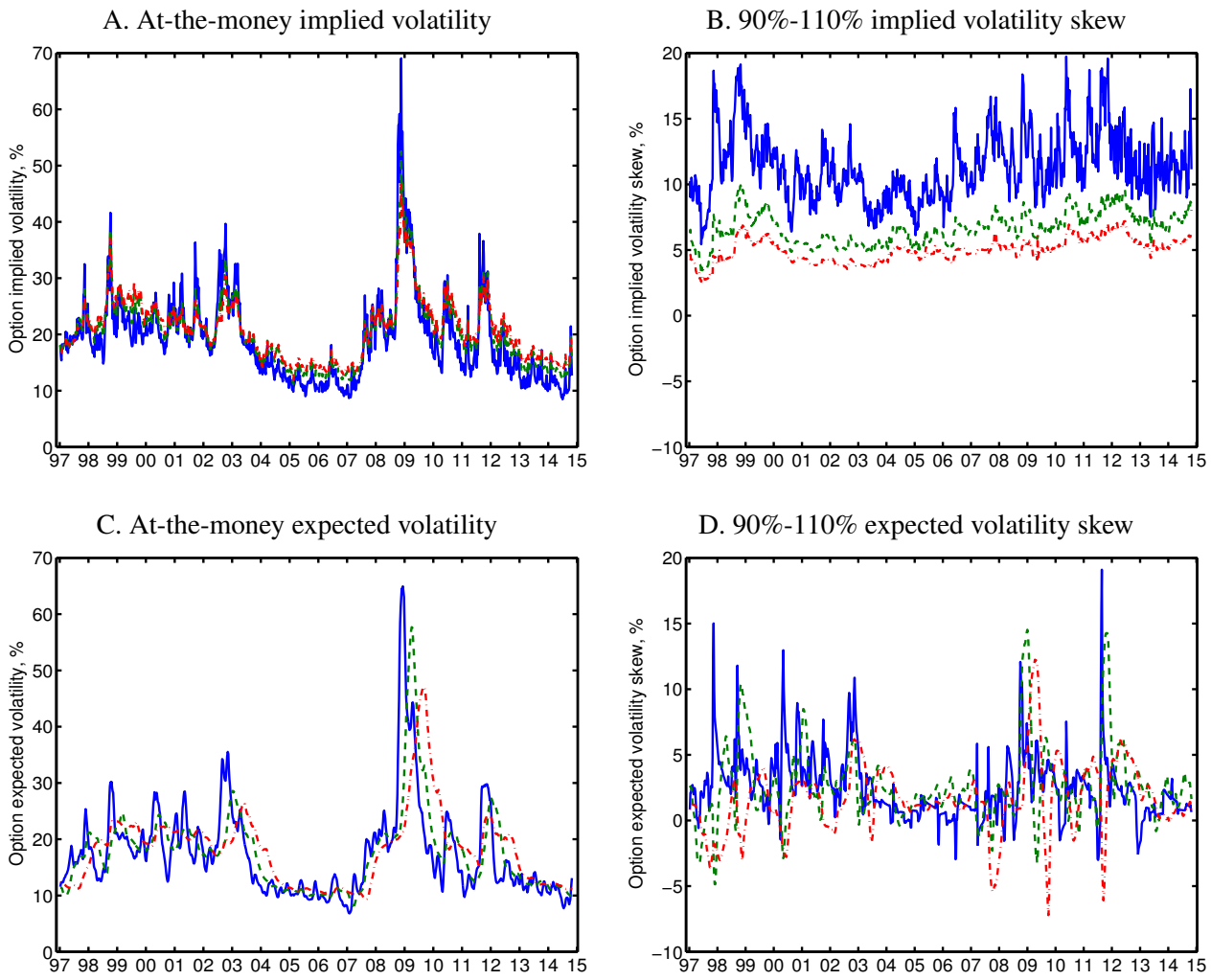

Fig. 1. Time-variation in option implied and expected volatilities and volatility skews. Lines plot the time series of option implied (Panels A & B) and expected volatilities (Panels C & D). Panels A & C are for at-the-money options whereas Panels B & D are for the volatility differences for 90% strike-110% strike risk reversals. The three lines in each panel are for three different time to maturities: one month (solid lines), six months (dashed lines), and the 24-months (dash-solid lines).

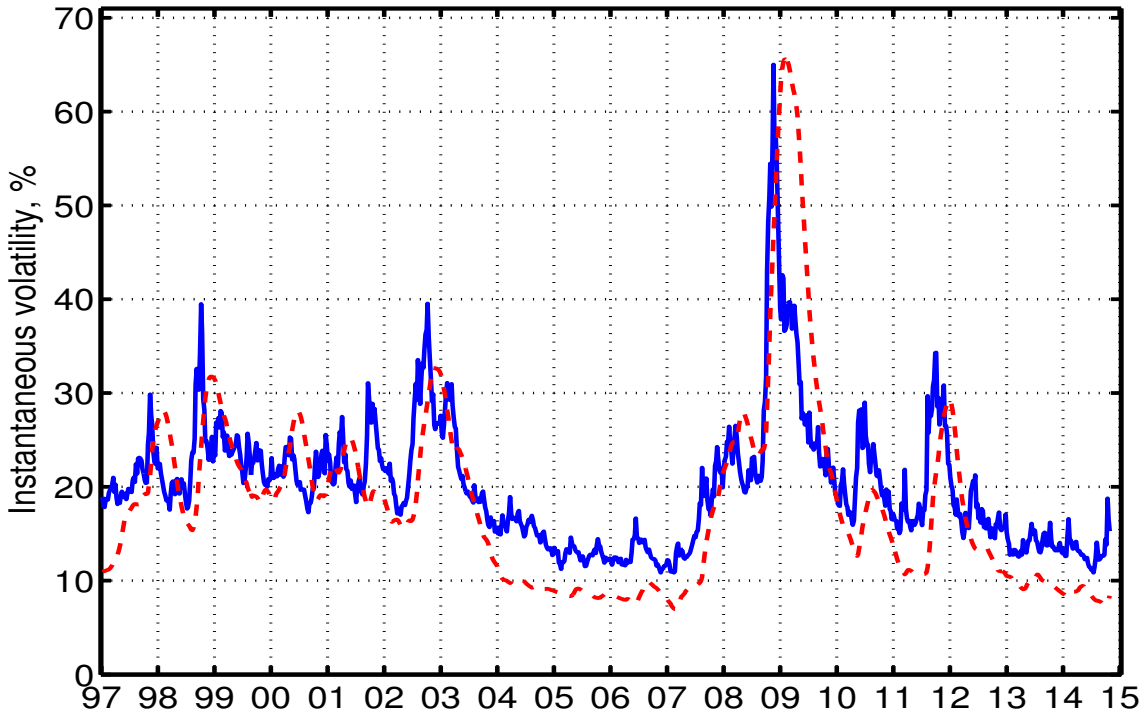

Fig. 2. The time-series variation of instantaneous volatilities. The solid line represents the time series of the instantaneous volatility ($\sqrt{v_t}$) extracted from the option implied volatility surface. The dashed line represents the corresponding instantaneous volatility extracted from the expected volatility surface.

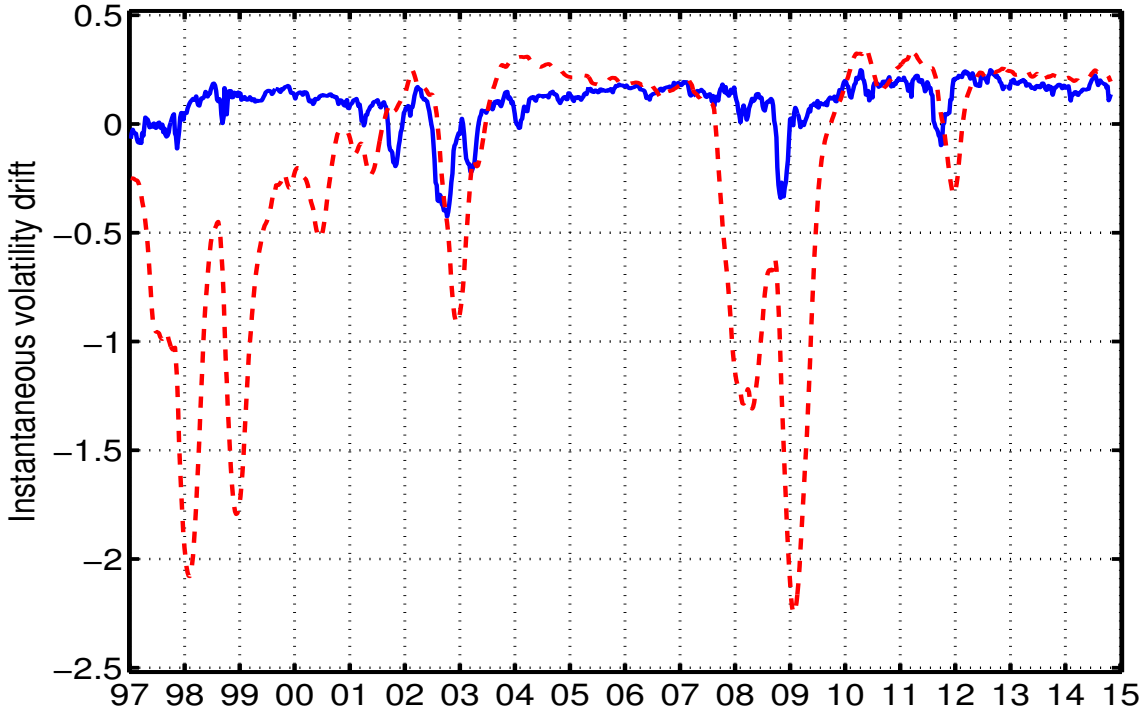

Fig. 3. The time-series variation of volatility drift processes. The solid line represents the time series of the risk-neutral volatility drift process (m_t), extracted from the option implied volatility surface. The dashed line represents the corresponding statistical volatility drift process ($m_t^{\mathbb{P}}$) extracted from the expected volatility surface.

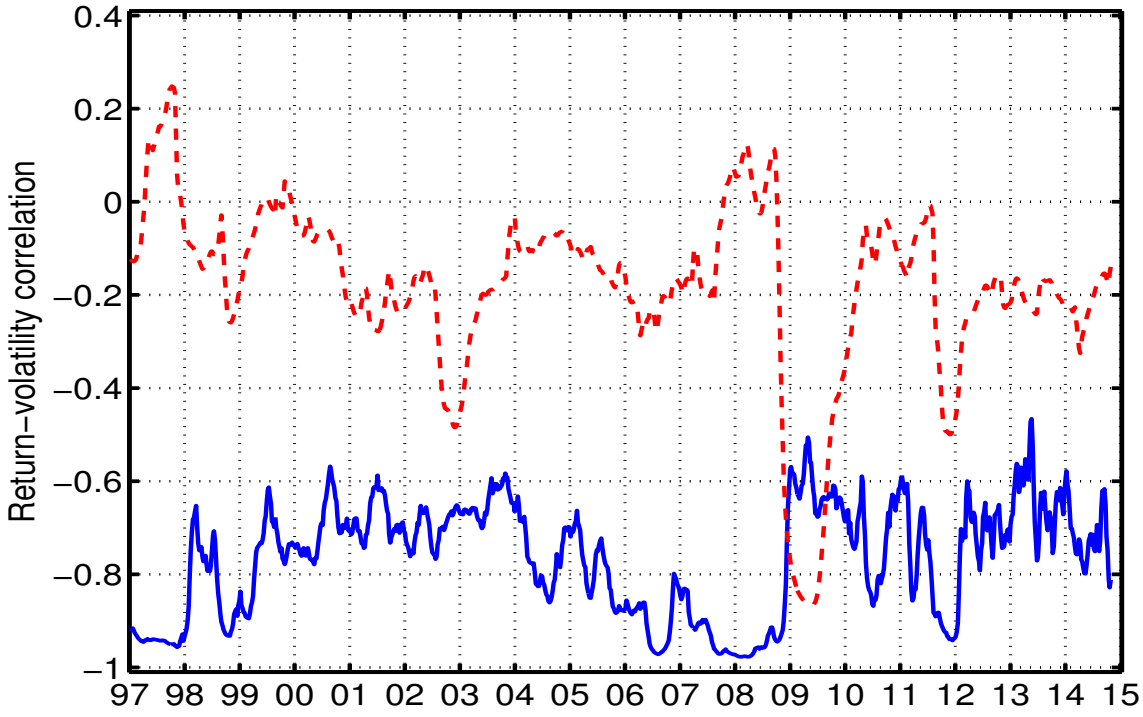

Fig. 4. The time-series variation of the return-volatility correlation. The solid line represents the time series of the instantaneous correlation between the index return and its volatility, extracted from the option implied volatility surface. The dashed line represents the corresponding correlation extracted from the expected volatility surface.

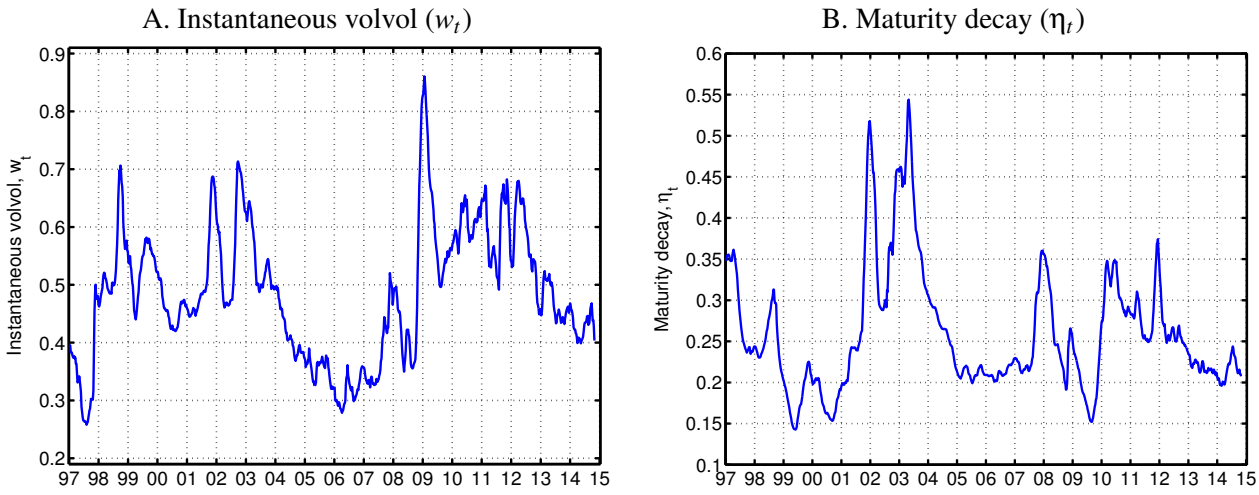

Fig. 5. The time series of the instantaneous volvol dyanmics w_t in panel A and the maturity decay process η_t in panel B.

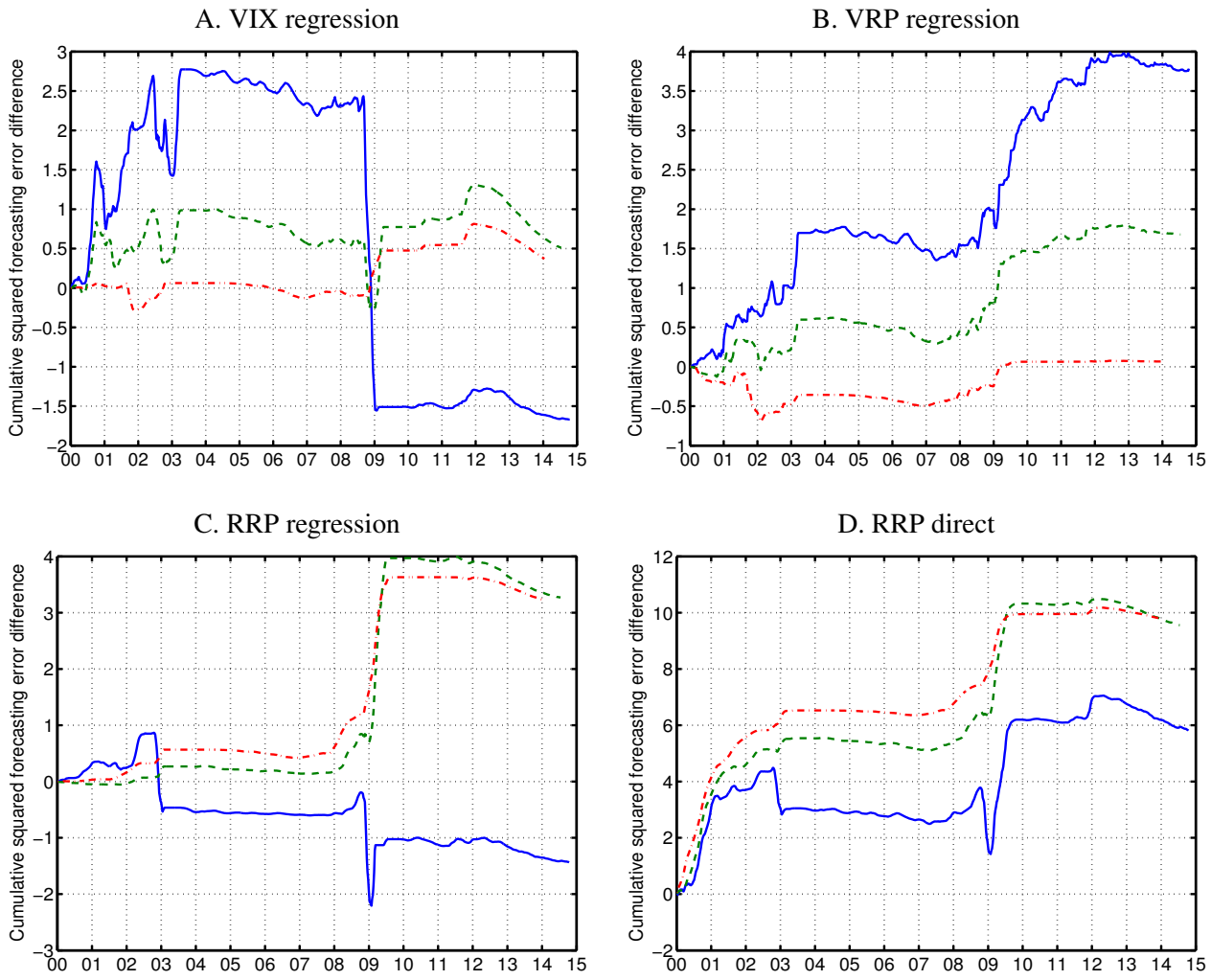

Fig. 6. Cumulative squared forecasting error difference. The three lines in each panel are the cumulative squared forecasting error difference between each method and the historical average benchmark, with each line representing one forecasting horizon: three months (solid line), six months (dashed line), and 12 months (dash-dotted line). The four panels are for four different forecasting methods.