A New Framework for Analyzing Volatility Risk and Volatility Risk Premium in Each Option Contract

Liuren Wu, Baruch College

Joint work with Peter Carr from Morgan Stanley

Florida State University
September 14, 2012
Traditional finance focuses on the trade off between return and risk.

- It has become apparent that the risk itself (volatility) is also risky.
  - The large literature on stochastic volatility, GARCH, HF volatility...

- It has also been found that the market heavily compensates investors for bearing volatility risk — the growing literature on variance risk premium.

- The cleanest way to gain variance risk exposure, without bearing return risk, seems to be through an OTC variance swap contract.
  - The P&L is determined by the difference between the realized variance and its risk-neutral expected value — the variance swap rate.
  - The academic literature on variance risk premium mostly focuses on this contract (natural or synthetic).

- The most available to investors are not the variance swap contracts, but plain vanilla options traded both on the exchanges and over the counter.

- We propose a new and simple framework for directly analyzing volatility risk and risk premium embedded in these vanilla option contracts.
Buy/sell an out-of-the-money option and *delta hedge*.

Daily delta hedge is a requirement for most institutional volatility investors and options market makers.

Call or put is irrelevant. What matter is whether one is long/short vega.

Delta hedged P&L is equal to dollar-gamma weighted variance difference:

$$PL_T = \int_0^T e^{r(T-t)}(\sigma_t^2 - IV_0^2) \frac{S_t^2}{2} \frac{n(d_1(S_t,t,IV_0))}{S_t IV_0 \sqrt{T-t}} dt.$$  

Views/quotes are expressed not in terms of dollar option prices, but rather in terms of *implied volatilities* (IV).

Implied volatilities are calculated from the Black-Merton-Scholes (BMS) model.

The fact that practitioners use the BMS model to quote options does not mean they agree with the BMS assumptions.

Rather, they use the BMS model as a way to transform/standardize the option price, for several practical benefits.
There are several practical benefits in transforming option prices into BMS implied volatilities.

1. **Information**: It is much easier to gauge/express views in terms of implied volatilities than in terms of option prices.

   - Option price behaviors all look alike under different dynamics: Option prices are monotone and convex in strike...
   - By contrast, how implied volatilities behavior against strikes reveals the shape of the underlying risk-neutral return distribution.
     - A flat implied volatility plot against strike serves as a benchmark for a normal return distribution.
     - Deviation from a flat line reveals deviation from return normality.
       \[ \Rightarrow \text{Implied volatility smile} — \text{leptokurtotic return distribution} \]
       \[ \Rightarrow \text{Implied volatility smirk} — \text{asymmetric return distribution} \]
**Why BMS implied volatility?**

2. **No arbitrage constraints:**
   - Merton (1973): model-free bounds based on no-arb. arguments:
     - Type I: No-arbitrage between European options of a fixed strike and maturity vs. the underlying and cash:
       - call/put prices $\geq$ intrinsic;
       - call prices $\leq$ (dividend discounted) stock price;
       - put prices $\leq$ (present value of the) strike price;
       - put-call parity.
     - Type II: No-arbitrage between options of different strikes and maturities: bull, bear, calendar, and butterfly spreads $\geq 0$.
   - Hodges (1996): These bounds can be expressed in implied volatilities.
     - Type I: *Implied volatility must be positive.*

$\Rightarrow$ *If market makers quote options in terms of a positive implied volatility surface, all Type I no-arbitrage conditions are automatically guaranteed.*

3. **Delta hedge:** The standard industry practice is to use the BMS model to calculate delta with the implied volatility as the input.
   *No delta modification consistently outperforms this simple practice in all practical situations.*
A new framework for analyzing volatility risk and premium

- The current literature:
  - Start with an instantaneous variance rate dynamics, derive no-arbitrage implications on option prices and then the implied volatility surface.
  - Volatility risk premium is defined on the instantaneous variance rate. Option value is a complicated function of dynamics and risk premium. \( \Rightarrow \) Fourier transforms are involved in the most tractable case. It is difficult to gauge the volatility risk premium embedded in an option without some complicated calculation.

- Our new framework is a lot simpler, much more direct, and much more in line with industry practice with vanilla options.
  - Start directly with implied volatility dynamics, derive no-arbitrage implications on the implied volatility surface. \( \Rightarrow \) Much simpler. The whole surface can be cast as solutions to a quadratic equation.
  - Define volatility risk premium on each option contract directly as the difference between the implied volatility and the expected value of a newly defined, contract specific option realized volatility measure.
The option realized volatility (ORV) surface

- Literature: To gauge the premium in a variance swap, one can directly compare the variance swap rate with a forecast of future realized variance.
  - The realized variance follows traditional definitions: sum of return squared, annualized with 252 over number of business days.

- Since our new framework directly models the BMS implied volatility surface, we propose a corresponding option realized volatility (ORV) surface:
  - The ORV for an option contract is the volatility that one uses in the BMS model to generate the option value and to perform daily delta hedge over the life of the option and leads to a zero terminal P&L.
  - Given the realized stock sample path, the ex post delta-hedged P&L of the option simply becomes $BMS_t(ORV_t(K, T)) - BMS_t(IV_t(K, T))$.
  - The time-$t$ forecast of ORV, $R_t(K, T)$, is defined as the contract-specific volatility forecast that generates zero expected P&L, $BMS_t(R_t(K, T)) = \mathbb{E}_t^P [BMS_t(ORV_t(K, T))]$.
  - Hence, $R_t(K, T) - IV_t(K, T)$ directly defines the “volatility risk premium” embedded in the option contract at strike $K$ and expiry $T$. 
**Implied volatility dynamics**

- Zero rates for notational clarity.

- Diffusion stock price dynamics: \( dS_t / S_t = \sqrt{v_t} dW_t \).

- The dynamics of the instantaneous variance rate \((v_t)\) is left **unspecified**.

- Instead, for each option struck at \( K \) and expiring at \( T \), we model its implied volatility \( I_t(K, T) \) dynamics under the risk-neutral \((\mathbb{Q})\) measure as,

\[
dI_t(K, T) = \mu_t dt + \omega_t dZ_t, \text{ for all } K > 0 \text{ and } T > t.
\]

  - \( \mu_t \) (drift) and \( \omega_t \) (volvol) can depend on \( K, T, \) and \( I(K, T) \).

  - One Brownian motion \( Z_t \) drives the whole implied volatility surface.

  - Correlation between implied volatility and return \( \rho_t dt = \mathbb{E}[dW_t dZ_t] \).

- \( I_t(K, T) > 0 \) guarantees no static arbitrage between any option \((K, T)\) and the underlying stock and cash.

- We further require that no dynamic arbitrage (NDA) be allowed between any option at \((K, T)\) and a basis option at \((K_0, T_0)\) and the stock.
No dynamic arbitrage

NDA: No dynamic arbitrage is allowed between any option at \((K, T)\) and a basis option at \((K_0, T_0)\) and the stock.

- 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 value.
- We can write both the basis call and other put options in terms of the BMS put formula:
  \[ P_t(K, T) = B(S_t, I_t(K, T), t), \quad C_t(K_0, T_0) = B(S_t, I_t(K_0, T_0), t) + S_t - K_0. \]
- We can form a portfolio between the two to neutralize the exposure on the volatility risk \(dZ\):
  \[ B_\sigma(S_t, I_t(K, T), t)\omega_t(K, T) - N_t^c B_\sigma(S_t, I_t(K_0, T_0), t)\omega_t(K_0, T_0) = 0. \]
- The 2-option portfolio with no \(dZ\) exposure can be exposed to \(dW\).
  We use \(N_t^S\) shares of the underlying stock to achieve delta neutrality:
  \[ B_S(S_t, I_t(K, T), t) - N_t^c(1 + B_S(S_t, I_t(K_0, T_0), t)) - N_t^S = 0. \]
- Since shares have no vega, this three-asset portfolio retains zero exposure to \(dZ\) and by construction has zero exposure to \(dW\).
The three-asset portfolio by design has no exposure to $dW$ or $dZ$.

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

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

No arbitrage and no rates imply that both option drifts must vanish, leading to the fundamental “PDE:”

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

When $\mu_t$ and $\omega_t$ are independent of $(K,T)$, the “PDE” defines a linear relation between the theta ($B_t$) of the option and its vega ($B_\sigma$), dollar gamma ($S_t^2 B_{SS}$), dollar vanna ($S_t B_{S\sigma}$), and volga ($B_{\sigma\sigma}$).

Hence, sometimes we call the class of implied volatility surfaces defined by our above fundamental PDE as the **Vega-Gamma-Vanna-Volga (VGVV)** model.
Our PDE is NOT a PDE in the traditional sense

\[-B_t = \mu_t B_\sigma + \frac{v_t}{2} S_t^2 B_{SS} + \rho_t \omega_t \sqrt{v_t} S_t B_{S\sigma} + \frac{\omega_t^2}{2} B_{\sigma\sigma}.\]

- Traditionally, PDE is specified to solve the value function. In our case, the value function $B(S_t, l_t, t)$ is well-known as it is simply the BMS formula.

- The coefficients on traditional PDEs are deterministic; they are stochastic in our “PDE.”

- Our “PDE” is not derived to solve the value function, but rather it is used to show that the various stochastic quantities have to satisfy this particular relation to exclude NDA.

$\Rightarrow$ Our “PDE” defines an NDA constraint on how the different quantities should relate to each other.
From the “PDE” to an algebraic restriction

\[-B_t = \mu_t B_\sigma + \frac{v_t}{2} S_t^2 B_{SS} + \rho_t \omega_t \sqrt{v_t} S_t B_{S\sigma} + \frac{\omega_t^2}{2} B_{\sigma\sigma}.\]

- The value function $B$ is well known, so are its various partial derivatives:
  
  \begin{align*}
  B_t &= -\frac{\sigma^2}{2} S_t^2 B_{SS}, \\
  B_\sigma &= \sigma \tau S_t^2 B_{SS}, \\
  S B_{\sigma S} &= -d_2 \sqrt{\tau} S_t^2 B_{SS}, \\
  B_{\sigma\sigma} &= d_1 d_2 \tau S_t^2 B_{SS},
  \end{align*}

where dollar gamma is the common denominator of all the partial derivatives, a result particular to the normal density function.

- The “PDE” constraint on $B$ is reduced to an algebraic restriction on the shape of the implied volatility surface $I_t(K, T)$,

\[
\frac{l_t^2}{2} - \mu_t l_t \tau - \left[ \frac{v_t}{2} - \rho_t \omega_t \sqrt{v_t} \tau d_2 + \frac{\omega_t^2}{2} d_1 d_2 \tau \right] = 0. \tag{1}
\]

- If $(\mu_t, \omega_t)$ do not depend on $I_t(K, T)$, we can solve the whole implied volatility surface as the solution to a quadratic equation.

- GVV (by Arslan, Eid, Khoury, and Roth from DB): $\mu_t = 0$, $\omega_t$ is independent. $\Rightarrow I_t^2$ is quadratic in $d_2$. 
Log-normal implied variance dynamics

- We focus on a log-normal implied variance (LNV) \( \mathbb{Q} \)-dynamics:
  \[
dl_t^2(K, T) = e^{-\eta_t(T-t)} \left( \kappa_t \left( \theta_t - l_t^2(K, T) \right) \right) dt + 2\omega_t l_t^2(K, T) dZ_t,
\]
  with \( \eta_t, \kappa_t, \theta_t, \omega_t > 0 \).

  - A log-normal specification has more empirical support than a square-root specification.
  - \( \kappa_t, \theta_t > 0 \) forces mean reversion into the process.
  - Exponential dampening makes long-term implied volatility less volatile and more persistent.
  - \( (\eta_t, \kappa_t, \theta_t, \omega_t) \) can all be stochastic processes, which are left unspecified, but they are not functions of \( K, T, \) or \( l(K, T) \).
  - The LNV is only a partial specification ...

- Implied volatility dynamics:
  \[
dl_t(K, T) = e^{-\eta_t(T-t)} \left( \frac{1}{2} \beta_t \left( \frac{\alpha_t}{l_t(K, t)} - I_k(K, T) \right) \right) + \omega_t l_t(K, T) dZ_t
\]
Under the LNV dynamics, we can re-cast the implied volatility surface in terms of log relative strike and time to maturity, \( I_t(k, \tau) \equiv I_t(K, T) \), with \( k = \ln K / S_t \) and \( \tau = T - t \).

The implied variance surface \( (I_t^2(k, \tau)) \) solves a quadratic equation:

\[
0 = \frac{1}{4} e^{-2\eta t \tau} w_t^2 \tau^2 I_t^4(k, \tau) + (1 + e^{-\eta t \tau} \beta_t \tau - e^{-\eta t \tau} w_t \rho_t \sqrt{v_t \tau}) I_t^2(k, \tau) - (v_t + e^{-\eta t \tau} \alpha_t \beta_t \tau + 2e^{-\eta t \tau} w_t \rho_t \sqrt{v_t} k + e^{-2\eta t \tau} w_t^2 k^2).
\]

Given the six covariates \((\rho_t, v_t, \theta_t, \kappa_t, \eta_t, w_t)\), the whole surface can be solved analytically from the quadratic equation (the positive solution).

Given the current levels of the six covariates, the current shape of the implied volatility surface must satisfy the above quadratic equation to exclude dynamic arbitrage.

However, the current shape of the surface does NOT depend on the exact dynamics of \((\rho_t, \theta_t, \kappa_t, \eta_t, w_t)\).

Their dynamics will only affect the future dynamics of the surface, but not its current shape.
Implied volatility smile at a fixed maturity

- At a fixed maturity, the implied variance variance smile can be solved as

\[ I^2(k, \tau) = a_t + \frac{2}{\tau} \sqrt{ \left( k + \frac{\rho \sqrt{v_t}}{e^{-\eta t \tau} w_t} \right)^2 + c_t}. \]

- When \( |k| \to \infty \), the asymptotic slope is \( s_\pm = 2 \).
- In the limit of \( \tau = 0 \), \( I^2_t(k, 0) = v_t + 2\rho_t \sqrt{v_t} w_k + w^2 k^2 \).
- Jim Gatheral’s SVI (“stochastic-volatility inspired”):

\[ I^2(k, \tau) = a + b \left[ \rho (k - m) + \sqrt{(k - m)^2 + \sigma^2} \right]. \]

- The asymptotes: \( s_+ = b\tau(1 + \rho), \quad s_- = b\tau(1 - \rho) \).
- A more flexible specification, but with no dynamics support or linkage across maturities.
The at-the-money implied variance term structure

- We define at-the-money as $K = \mathbb{E}_t^B[\ln S_T]$ or $d_2 = 0$.

- The at-the-money implied variance term structure is a weighted average of $\nu_t$ and its long-run value $\alpha_t$,

$$A^2_t(\tau) = \phi_t(\tau)\nu_t + (1 - \phi_t(\tau))\alpha_t, \quad \phi_t(\tau) = \frac{1}{1 + e^{-\eta_t\tau} \beta_t\tau}.$$  

- Only a function of implied volatility drift $(\alpha_t, \beta_t)$, does not depend on volvol.

- With strictly positive $\eta_t$, the long-run limit $\alpha_t = \frac{\kappa_t}{\kappa_t + e^{-\eta_t\tau} \omega_t^2} \theta_t$ is a function of time to maturity. It converges to $\theta_t$ as $\tau \to \infty$.

- The weight $\phi(\tau)$ is not monotone with maturity. It starts and ends with 1.

- $(\alpha_t - \nu_t)$ determines the initial slope of the term structure.
Let $\gamma_t$ denote the market price of Brownian risk on $dZ_t$ — It should not depend on $K$, $T$, or $I(K, T)$.

The statistical dynamics for the implied volatility becomes,

$$dl_t(K, T) = e^{-\eta_t(T-t)} \left( \frac{1}{2} \left( \frac{\alpha_t \beta_t}{I_t(K, t)} - \beta^P_t l_k(K, T) \right) + w_t l_t(K, T) dZ^P_t \right).$$

$$\beta^P_t = (\kappa_t - 2\gamma_t w_t) + e^{-\eta_t(T-t)} w_t^2 = \kappa^P_t + e^{-\eta_t(T-t)} w_t^2.$$

Let $R_t(K, T)$ denote the expected ORV that generates zero expected delta-hedged P&L: $BMS(R_t(K, T)) = \mathbb{E}^P_t [BMS_t(ORV_t(K, T))]$. In the absence of volatility risk premium ($\gamma_t = 0$), we have $R_t(K, T) = I_t(K, T)$. Both surfaces are determined by the same statistical dynamics:

$$0 = \frac{1}{4} e^{-2\eta_t \tau} w_t^2 \tau^2 R^4_t(k, \tau) + \left( 1 + e^{-\eta_t \tau} \beta^P_t \tau - e^{-\eta_t \tau} w_t \rho_t \sqrt{v_t \tau} \right) R^2_t(k, \tau) R^2_t(k, \tau) - \left( v_t + e^{-\eta_t \tau} \alpha_t \beta_t \tau + 2 e^{-\eta_t \tau} w_t \rho_t \sqrt{v_t} k + e^{-2\eta_t \tau} w_t^2 k^2 \right).$$

When $\gamma_t \neq 0$, this equation only determines the surface $R_t(k, \tau)$. 
Correcting for the effect of price jumps

- Our implied volatility surface constraint is derived under the assumption of diffusion return dynamics and a one-factor implied volatility structure.

- When price can jump, the instantaneous variance rate becomes an expectation and can differ under the two measures ($P$ and $Q$) if the jump risk is priced.

- We accommodate this potential difference by allowing $\nu_t^P$ different from $\nu_t$, with the difference measuring the risk premium induced by price jump risk.

- This rough adjustment only accounts for the short-term implied-realized volatility level difference, but does not account for the short-term implied volatility smiles/skews induced by price jumps.
  - Avoid using short-term options for estimating purely continuous price dynamics.
  - It would be interesting to see an extension explicitly allowing for jumps in the price dynamics.
The closest benchmark would be the pure diffusion, one-factor stochastic volatility model of Heston (1993).

If we allow the coefficients to vary, the extended Heston model would have five covariates to determine the two volatility surfaces \( (\kappa_t, \theta_t, \omega_t, \rho_t, \nu_t, \gamma_t) \).

- The stochastic nature of the covariates in the LNV model \( (\kappa_t, \theta_t, \omega_t, \rho_t, \nu_t, \gamma_t) \) does not affect the shape of the volatility surfaces.
- Extending the Heston parameters to be stochastic can have complicated effects on the surfaces.

The pricing performance of Heston is significantly worse than LNV, and the computational burden is at least 100 times larger.

- No comparable results yet...

A square-root implied variance dynamics (SRV) also generates a tractable implied volatility surface, but data support LNV dynamics better.
An empirical application to the SPX volatility surfaces

- SPX options are actively traded both on exchanges and over the counter.
- We use OTC implied volatility quotes that combine exchange transactions information at short maturity with OTC trades at longer term.
- Each date, implied volatility is quoted on a fixed grid of
  - 5 relative strike at 80, 90, 100, 110, 120% of the spot level.
  - 8 maturities from 1 month to 5 years.
- Data are available daily from January 1997 to March 2008. We sample the data weekly every Wednesday for 583 weeks.
- Corresponding to each implied volatility quote, $I_t(k, \tau)$, we also use the historical SPX price time series to compute
  - Historical contract-specific option realized volatility $ORV(k, t - \tau, t)$.
  - Forecasted option volatility surface $R_t(k, \tau)$ based on historical ORVs.
Negative skew (across strike) is observed for both implied and realized volatility, more for implied.

Average implied volatility level is higher than realized volatility level, more so at the short-term, low-strike region.
Variation of implied volatility changes

<table>
<thead>
<tr>
<th>$K/S$</th>
<th>0.8</th>
<th>0.9</th>
<th>1.0</th>
<th>1.1</th>
<th>1.2</th>
<th>0.8</th>
<th>0.9</th>
<th>1.0</th>
<th>1.1</th>
<th>1.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mat</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2.35</td>
<td>2.20</td>
<td>2.16</td>
<td>1.69</td>
<td>1.22</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
<td>0.96</td>
</tr>
<tr>
<td>3</td>
<td>1.56</td>
<td>1.51</td>
<td>1.48</td>
<td>1.37</td>
<td>1.06</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>0.97</td>
</tr>
<tr>
<td>6</td>
<td>1.19</td>
<td>1.16</td>
<td>1.15</td>
<td>1.11</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
</tr>
<tr>
<td>12</td>
<td>0.93</td>
<td>0.91</td>
<td>0.90</td>
<td>0.89</td>
<td>0.83</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td>24</td>
<td>0.75</td>
<td>0.74</td>
<td>0.74</td>
<td>0.73</td>
<td>0.71</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.98</td>
</tr>
<tr>
<td>36</td>
<td>0.67</td>
<td>0.66</td>
<td>0.66</td>
<td>0.65</td>
<td>0.65</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>48</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.59</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>60</td>
<td>0.57</td>
<td>0.56</td>
<td>0.56</td>
<td>0.56</td>
<td>0.56</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
</tbody>
</table>

- Both standard deviation and mean reversion declines as the implied volatility maturity increases.
- We apply exponential dampening $e^{-\eta (T-t)}$ on the drift and diffusion to capture this maturity pattern.
Assume implied volatility diffusion takes a CEV form, $dl = \mu dt + \omega l^{\beta} dZ_t$.

We estimate an exponentially weighted variance on weekly changes in implied volatility, $EVI$.

We then perform the following regression on each implied volatility series, $\ln EVI_t(k, \tau) = \text{intercept} + \beta \ln I^2(k, \tau) + e$.

$\beta = 1$ under log-normal dynamics, but $\beta = 0$ under square-root dynamics.

<table>
<thead>
<tr>
<th>$K/S$</th>
<th>0.8</th>
<th>0.9</th>
<th>1.0</th>
<th>1.1</th>
<th>1.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.02 ( 0.03 )</td>
<td>1.14 ( 0.03 )</td>
<td>0.88 ( 0.02 )</td>
<td>1.41 ( 0.03 )</td>
<td>1.62 ( 0.03 )</td>
</tr>
<tr>
<td>3</td>
<td>1.20 ( 0.03 )</td>
<td>1.16 ( 0.02 )</td>
<td>0.97 ( 0.02 )</td>
<td>1.24 ( 0.02 )</td>
<td>1.65 ( 0.03 )</td>
</tr>
<tr>
<td>6</td>
<td>1.28 ( 0.03 )</td>
<td>1.22 ( 0.03 )</td>
<td>1.07 ( 0.02 )</td>
<td>1.14 ( 0.02 )</td>
<td>1.54 ( 0.03 )</td>
</tr>
<tr>
<td>12</td>
<td>1.33 ( 0.03 )</td>
<td>1.23 ( 0.03 )</td>
<td>1.10 ( 0.03 )</td>
<td>1.07 ( 0.03 )</td>
<td>1.31 ( 0.03 )</td>
</tr>
<tr>
<td>24</td>
<td>1.37 ( 0.03 )</td>
<td>1.27 ( 0.03 )</td>
<td>1.15 ( 0.03 )</td>
<td>1.07 ( 0.03 )</td>
<td>1.15 ( 0.03 )</td>
</tr>
<tr>
<td>36</td>
<td>1.46 ( 0.04 )</td>
<td>1.36 ( 0.04 )</td>
<td>1.25 ( 0.04 )</td>
<td>1.14 ( 0.04 )</td>
<td>1.14 ( 0.03 )</td>
</tr>
<tr>
<td>48</td>
<td>1.47 ( 0.05 )</td>
<td>1.37 ( 0.04 )</td>
<td>1.25 ( 0.04 )</td>
<td>1.15 ( 0.04 )</td>
<td>1.09 ( 0.04 )</td>
</tr>
<tr>
<td>60</td>
<td>1.43 ( 0.05 )</td>
<td>1.34 ( 0.05 )</td>
<td>1.24 ( 0.05 )</td>
<td>1.14 ( 0.05 )</td>
<td>1.06 ( 0.05 )</td>
</tr>
</tbody>
</table>
Dynamic estimation of the volatility surfaces

- Treat the covariates as the hidden state vector $X_t$.

- Assume that the state vector propagates like a random walk:
  $$X_t = X_{t-1} + \sqrt{\Sigma_x} \varepsilon_t$$

  - One can assume more complicated dynamics (e.g., mean reversion).
  - The random walk assumption avoids estimating more parameters.

- The last fitted surface carries over until the arrival of new information.

- Transform the variates so that $X_t$ have the full support $(-\infty, +\infty)$.

- Assume diagonal matrix for $\Sigma_x$.

- Assume that the implied and realized volatility surfaces are observed with errors,
  $$y_t = h(X_t) + \sqrt{\Sigma_y} e_t.$$ 

  - $y_t$ includes 40 implied volatility series and 40 historical realized volatility series.
  - $h(\cdot)$ denote the model value (solutions to a quadratic equation)
  - Assume IID error for each of the two surfaces.

- The auxiliary parameters $(\Sigma_x, \sigma_e^2)$ control the relative update speed of the covariates $X_t$. 

Given the auxiliary parameters, the two volatility surfaces can be fitted quickly via unscented Kalman filter:

\[
\begin{align*}
\overline{X}_t &= \hat{X}_{t-1}, \quad \overline{V}_{x,t} = \hat{V}_{x,t-1} + \Sigma_x, \\
\chi_{t,0} &= \overline{X}_t, \quad \chi_{t,i} = \overline{X}_t \pm \sqrt{(k + \delta)(\overline{V}_{x,t})_j}, \\
\overline{y}_t &= \sum_{i=0}^{2^k} w_i \zeta_{t,i}, \quad \overline{V}_{y,t} = \sum_{i=0}^{2^k} w_i [\zeta_{t,i} - \overline{y}_t] [\zeta_{t,i} - \overline{y}_t]^\top + \Sigma_y, \\
\overline{V}_{xy,t} &= \sum_{i=0}^{2^k} w_i [\chi_{t,i} - \overline{X}_t] [\zeta_{t,i} - \overline{y}_t]^\top, \quad K_t = \overline{V}_{xy,t} (\overline{V}_{y,t})^{-1}, \\
\hat{X}_t &= \overline{X}_t + K_t (y_t - \overline{y}_t), \quad \hat{V}_{x,t} = \overline{V}_{x,t} - K_t \overline{V}_{y,t} K_t^\top.
\end{align*}
\]

Choose the auxiliary parameters to minimize the likelihood on forecasting errors.
Pricing performance comparison on SPX options

<table>
<thead>
<tr>
<th></th>
<th>LNV</th>
<th>SRV</th>
<th>Heston</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE on IV</td>
<td>0.74</td>
<td>1.75</td>
<td>–</td>
</tr>
<tr>
<td>RMSE on RV</td>
<td>2.94</td>
<td>3.82</td>
<td>–</td>
</tr>
<tr>
<td>Likelihood ($\times 10^5$)</td>
<td>1.64</td>
<td>1.47</td>
<td>–</td>
</tr>
</tbody>
</table>

- Log-normal dynamics fit the data better than square root dynamics.

- Heston model performance is a bit similar to SRV, but hundreds of times slower for estimation.

  - Given covariates, fitting all the surfaces over the 11-year period on my laptop takes about $\sim 1.4$ seconds for LNV/SRV, about 250 seconds for Heston.
The difference between the two short rates ($v_t, v_t^{IP}$) reflect risk premium induced by price jump risk.

The difference between the two long rates ($\alpha, \alpha^{IP}$) reflect the risk premium induced by the variance rate risk.

Traditional definition of volatility risk premium (implied minus realized) represents a combination of the two components.
Bollerslev, Tauchen, & Zhou (RFS, 2009) predict stock returns using $VIX^2 - RV^2$. The $R^2$ is 7% including the crisis period, and 4% during our sample period.

The two risk premium components extracted from the two surfaces both predict SPX returns, but with different strengths at different horizons.

A bivariate regression generates much higher forecasting R-squared.
Stochastic skew also predicts future stock returns (Xing, Zhang, Zhao (JFQA, 2010)), at least in the cross-section.

Can we use our extracted correlation to predict future SPX returns?

A. Stochastic return-volatility correlation

B. Predictability on index returns

The forecasting correlation between $\rho_t$ and future SPX return is negative.

⇒ The heavier the negative skew, the higher the return?
Concluding remarks

- Institutional option investors use BMS implied volatilities
  - to communicate their views and quotes,
  - to perform delta hedge, and
  - to gauge the delta-hedged gains in term of vol points.

- This paper provides a new, simple, transparent framework for analyzing volatility risk and volatility risk premium on each option contract, consistent with standard industry practice.
  - Model implied volatility dynamics and derive no-arbitrage constraints directly on the implied volatility surface — Extremely simple. The whole surface solves a quadratic equation.
  - Propose a new realized volatility measure that is specific to each strike and maturity — Realized volatility not only varies with term, but also with relative strike, similar to implied volatilities.
  - The volatility risk premium embedded in each contract becomes transparent — The difference between the implied volatility surface and the expected value of the ex post realized volatility surface defines the volatility risk premium for shorting each option contract.
Promise and future research

- The new framework has generated promising results when applied to the S&P 500 index.
- Despite its extreme simplicity, the proposed models fit the surface better than its counterpart in the standard option pricing literature.
- The extracted risk premiums and skews can predict future stock returns, with R-squared higher than those reported in the literature.

Many open questions remain.

- How to generate the ORV forecast?
- The PDE guarantees dynamic no-arbitrage between any option and a basis option under a single-factor continuous implied volatility dynamics. It remains open on how to guarantee (static) no-arbitrage among many options across different strikes and maturities.
- How to accommodate discontinuous price dynamics so that one can have a better handle of the short-term implied volatility smile?
- Is this framework limited to a one-factor volatility setting?