A New Framework for Analyzing Volatility Risk and Premium Across Option Strikes and Expiries

Liuren Wu, Baruch College

Joint work with Peter Carr from Morgan Stanley

Singapore Management University
July 4, 2013
Separating volatility risk from return risk

- 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 volatility risk exposure, without bearing return risk, seems to be through an OTC volatility swap contract.
  - The P&L is determined by the difference between the realized volatility and its risk-neutral expected value — the volatility swap rate.
  - The academic literature on variance risk premium mostly focuses on the variance swap contract (natural or synthetic).
- The most available to investors are not the volatility/variance swap contracts, but plain vanilla options.
  - We propose a new and simple framework for directly analyzing volatility risk and risk premium embedded in these vanilla option contracts.
How to gain volatility exposure via options in practice?

- 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 behave 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.
       - Implied volatility smile — leptokurtotic return distribution
       - Implied volatility smirk/skew — 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 at 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.*

⇒ *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. ⇒ Fourier transforms (numerical integration) are involved in the most tractable case.
    It is difficult to gauge the volatility risk premium embedded in an option contract 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 option implied volatility dynamics, derive no-arbitrage constraints on the implied volatility surface. ⇒ Much simpler. The whole surface solves a quadratic equation.
  - Define volatility risk premium on each option contract directly as the difference between the implied volatility and a newly defined, option-specific expected volatility measure, which can be estimated from the underlying price time series.
The option-specific volatility surface

- Literature: To gauge the premium in a volatility swap, one can directly compare the volatility swap rate with a forecast of future realized volatility.
  - Realized volatility follows traditional definitions: the square root of the sum of return squared, annualized.

- Since our new framework directly models the BMS implied volatility surface, we propose a corresponding option (strike-maturity)-specific realized and expected volatility surface:
  - The **option realized volatility** (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 **option expected volatility** (OEV), $V_t(K, T)$, is defined as the volatility forecast that generates zero expected delta-hedged P&L.
  - The difference $V_t(K, T) - IV_t(K, T)$ directly defines the “volatility risk premium” embedded in the option contract at strike $K$ and expiry $T$. 

Liuren Wu (Baruch)
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.
- For each option struck at \( K \) and expiring at \( T \), we model the near-future dynamics of its implied volatility \( I_t(K, T) \) under the risk-neutral \( (\mathbb{Q}) \) measure as,

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

- The drift \( (\mu_t) \) and volvol \( (\omega_t) \) processes 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 is \( \rho_t dt = \mathbb{E}[dW_t dZ_t] \).
- \( I_t(K, T) > 0 \) 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.
The three assets can be combined to neutralize exposure to $dW$ or $dZ$.

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

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

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

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

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 ($B_{\sigma\sigma}$).
Our PDE is NOT a PDE in the traditional sense.

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

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

- The coefficients on traditional PDEs are deterministic; they are stochastic processes 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 stochastic quantities should relate to each other.
From the “PDE” to an algebraic restriction

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

- The value function $B$ is well known, so are its various partial derivatives:

\[
\begin{align*}
B_t &= -\frac{\sigma^2}{2} S^2 B_{SS}, \\
B_\sigma &= \sigma \tau S^2 B_{SS}, \\
SB_{\sigma S} &= -d_2 \sqrt{\tau} S^2 B_{SS}, \\
B_{\sigma\sigma} &= d_1 d_2 \tau S^2 B_{SS},
\end{align*}
\]

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 $I_t(K, T)$,

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

- 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.
Proportional volatility dynamics, as an example

We consider a particularly simple and intuitive specification as an example:

\[
dl_t(K, T)/l_t(K, T) = e^{-\eta_t(T-t)}(m_t dt + w_t dZ_t),
\]

with \( \eta_t, w_t > 0 \).

- A proportional (log-normal) specification has more empirical support than a square-root variance specification.
- Exponential dampening makes long-term implied volatility less volatile and more persistent.
- \((\eta_t, m_t, w_t)\) are all stochastic processes, the dynamics of which are unspecified.
- Many other specifications can also work out, some with better pricing performance... but I like this extremely simple/parsimonious example.
Under the proportional volatility dynamics, we can re-cast the implied volatility surface in terms of log relative strike and time to maturity, 

\[ l_t(k, \tau) \equiv l_t(K, T), \] with \( k = \ln K/S_t \) and \( \tau = T - t \).

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

\[
0 = \frac{1}{4} e^{-2\eta_t \tau} w_t^2 \tau^2 l_t^4(k, \tau) + \left( 1 - 2 e^{-\eta_t \tau} m_t \tau - e^{-\eta_t \tau} w_t \rho_t \sqrt{\nu_t \tau} \right) l_t^2(k, \tau) \\
- \left( \nu_t + 2 e^{-\eta_t \tau} w_t \rho_t \sqrt{\nu_t} k + e^{-2\eta_t \tau} w_t^2 k^2 \right).
\]

Given the current levels of the five stochastic processes \((\rho_t, \nu_t, m_t, \eta_t, w_t)\), the whole implied volatility surface can be solved analytically from the quadratic equation (the positive solution).

- 5th grade math...
**Unspanned dynamics**

\[
0 = \frac{1}{4} e^{-\eta t \tau} w_t^2 \tau^2 l_t^4(k, \tau) + (1 - 2e^{-\eta t \tau} m_t \tau - e^{-\eta t \tau} w_t \rho_t \sqrt{v_t} \tau) l_t^2(k, \tau) \\
- \left(v_t + 2e^{-\eta t \tau} w_t \rho_t \sqrt{v_t} k + e^{-2\eta t \tau} w_t^2 k^2\right).
\]

- Given the current levels of the five stochastic processes \((\rho_t, v_t, m_t, \eta_t, w_t)\), the current shape of the implied volatility surface must satisfy the above quadratic equation to exclude dynamic arbitrage.

- The current shape of the surface does NOT depend on the exact dynamics of five stochastic processes.
  - The dynamics of the five processes are not spanned by the current shape of the implied volatility surface.
  - The current volatility surface shape is only linked to the *near-future* dynamics of the security return and implied volatility.

- The dynamics of the five processes will affect the future dynamics of the surface, but not its current shape.

- The current shape of the implied volatility surface is determined by *5 state variables*, but with *no parameters*!
OEV surface construction

- Forecasting option-specific expected volatility surface is analogous to the forecasting of the traditionally defined volatilities, except that we work with an OEV surface versus a traditional volatility term structure.

- We propose a simple procedure to get started, but look forward to econometric extensions:

  1. At each date $t$, we estimate a historical option realized volatility surface at fixed grids of relative strikes $k$ and time to maturity $\tau$ based on the historical sample path of the stock price, $ORV(k, t - \tau, t)$.

  2. We form a conditional expectation of the future based on historical ORV estimates,

     $$B(V_t(k, \tau)) \equiv \mathbb{E}_t^P [B(ORV(k, t, t + \tau))] \approx \sum_j w_j(t)B(ORV(k, t_j - \tau, t_j))$$

     with an exponential weighting $w_j(t) \propto e^{-\kappa|t - t_j|}, \forall t_j \leq t$.

  3. Note that the expectation is via the BMS model transformation as OEV is defined as the BMS volatility input that generates zero expected delta-hedged P&L.
No-arbitrage surface constrains on the OEV surface

- 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,
  \[ dl_t(K, T)/l_t(K, T) = e^{-\eta_t(T-t)} \left( m^P_t dt + w_t dZ^P_t \right), \]
  where the drift is adjusted as: $m^P_t = m_t + w_t \gamma_t$.

- Given our proportional implied volatility dynamics assumption, it is reasonable to assume that the expected volatility $V_t(K, T)$ is proportional to the corresponding implied volatility level $l_t(K, T)$ and is accordingly governed by the same proportional dynamics,
  \[ dV_t(K, T)/V_t(K, T) = e^{-\eta_t(T-t)} \left( m^P_t dt + w_t dZ^P_t \right). \]

- Analogously, the true OEV surface must satisfy the following equation,
  \[ 0 = \frac{1}{4} e^{-2\eta_t \tau} w^2 \tau^2 V^4_t(k, \tau) + \left( 1 - 2 e^{-\eta_t \tau} m^P_t \tau - e^{-\eta_t \tau} w_t \rho_t \sqrt{v_t} \tau \right) R^2_t(k, \tau) \]
  \[ - \left( v_t + 2 e^{-\eta_t \tau} w_t \rho_t \sqrt{v_t} k + e^{-2\eta_t \tau} w^2 \tau^2 k^2 \right). \]

Our OEV forecast estimators can be very noisy... This equation provides a filtering on the estimates.
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
  - Strike-expiry specific option historical realized volatility \( ORV(k, t - \tau, t) \).
  - An estimate of the option expected volatility surface \( V_t(k, \tau) \) based on historical ORV estimates.
The average implied/expected volatility surfaces

<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 A. Average implied volatility</td>
<td>33.75</td>
<td>25.81</td>
<td>18.71</td>
<td>15.06</td>
<td>14.26</td>
<td>20.18</td>
<td>19.03</td>
<td>16.59</td>
<td>16.43</td>
<td>18.87</td>
</tr>
<tr>
<td>3</td>
<td>28.86</td>
<td>23.88</td>
<td>19.31</td>
<td>15.97</td>
<td>14.56</td>
<td>18.85</td>
<td>18.13</td>
<td>16.44</td>
<td>14.95</td>
<td>15.09</td>
</tr>
<tr>
<td>6</td>
<td>26.61</td>
<td>23.04</td>
<td>19.68</td>
<td>16.90</td>
<td>15.21</td>
<td>18.51</td>
<td>17.57</td>
<td>16.44</td>
<td>15.31</td>
<td>14.53</td>
</tr>
<tr>
<td>12</td>
<td>25.12</td>
<td>22.57</td>
<td>20.14</td>
<td>17.97</td>
<td>16.32</td>
<td>17.81</td>
<td>17.22</td>
<td>16.78</td>
<td>16.13</td>
<td>15.49</td>
</tr>
<tr>
<td>36</td>
<td>24.41</td>
<td>22.93</td>
<td>21.50</td>
<td>20.16</td>
<td>18.93</td>
<td>16.35</td>
<td>15.96</td>
<td>15.82</td>
<td>15.81</td>
<td>16.01</td>
</tr>
<tr>
<td>48</td>
<td>24.52</td>
<td>23.25</td>
<td>22.04</td>
<td>20.90</td>
<td>19.83</td>
<td>15.85</td>
<td>15.61</td>
<td>15.45</td>
<td>15.46</td>
<td>15.66</td>
</tr>
<tr>
<td>60</td>
<td>24.72</td>
<td>23.60</td>
<td>22.55</td>
<td>21.55</td>
<td>20.60</td>
<td>15.15</td>
<td>14.99</td>
<td>15.05</td>
<td>15.15</td>
<td>15.33</td>
</tr>
<tr>
<td>Mat B. Average expected volatility</td>
<td>20.18</td>
<td>19.03</td>
<td>16.59</td>
<td>16.43</td>
<td>18.87</td>
<td>18.85</td>
<td>18.13</td>
<td>16.44</td>
<td>14.95</td>
<td>15.09</td>
</tr>
<tr>
<td>12</td>
<td>17.81</td>
<td>17.22</td>
<td>16.78</td>
<td>16.13</td>
<td>15.49</td>
<td>17.81</td>
<td>17.22</td>
<td>16.78</td>
<td>16.13</td>
<td>15.49</td>
</tr>
<tr>
<td>36</td>
<td>16.35</td>
<td>15.96</td>
<td>15.82</td>
<td>15.81</td>
<td>16.01</td>
<td>16.35</td>
<td>15.96</td>
<td>15.82</td>
<td>15.81</td>
<td>16.01</td>
</tr>
<tr>
<td>48</td>
<td>15.85</td>
<td>15.61</td>
<td>15.45</td>
<td>15.46</td>
<td>15.66</td>
<td>15.85</td>
<td>15.61</td>
<td>15.45</td>
<td>15.46</td>
<td>15.66</td>
</tr>
<tr>
<td>60</td>
<td>15.15</td>
<td>14.99</td>
<td>15.05</td>
<td>15.15</td>
<td>15.33</td>
<td>15.15</td>
<td>14.99</td>
<td>15.05</td>
<td>15.15</td>
<td>15.33</td>
</tr>
</tbody>
</table>

- 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>2.34</td>
<td>2.19</td>
<td>2.15</td>
<td>1.68</td>
<td>1.22</td>
<td>1.39</td>
<td>1.05</td>
<td>0.76</td>
<td>0.99</td>
<td>1.08</td>
</tr>
<tr>
<td>3</td>
<td>1.55</td>
<td>1.50</td>
<td>1.47</td>
<td>1.36</td>
<td>1.05</td>
<td>0.67</td>
<td>0.59</td>
<td>0.44</td>
<td>0.47</td>
<td>0.61</td>
</tr>
<tr>
<td>6</td>
<td>1.18</td>
<td>1.16</td>
<td>1.15</td>
<td>1.11</td>
<td>0.97</td>
<td>0.47</td>
<td>0.38</td>
<td>0.29</td>
<td>0.31</td>
<td>0.36</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>0.25</td>
<td>0.20</td>
<td>0.20</td>
<td>0.22</td>
<td>0.23</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>0.17</td>
<td>0.13</td>
<td>0.12</td>
<td>0.12</td>
<td>0.14</td>
</tr>
<tr>
<td>36</td>
<td>0.67</td>
<td>0.66</td>
<td>0.66</td>
<td>0.66</td>
<td>0.65</td>
<td>0.13</td>
<td>0.12</td>
<td>0.11</td>
<td>0.10</td>
<td>0.10</td>
</tr>
<tr>
<td>48</td>
<td>0.61</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.59</td>
<td>0.13</td>
<td>0.11</td>
<td>0.10</td>
<td>0.09</td>
<td>0.09</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>0.11</td>
<td>0.11</td>
<td>0.09</td>
<td>0.08</td>
<td>0.08</td>
</tr>
</tbody>
</table>

- Standard deviation declines as maturity increases.
- We apply exponential dampening $e^{-\eta(t)(T-t)}$ on the drift and diffusion to capture this maturity pattern.
**Constant elasticity of variance dependence**

- Assume implied volatility diffusion takes a CEV form, \( dl = \mu dt + wI^\beta dZ_t \).
- We estimate an exponentially weighted variance on weekly changes in implied volatility, \( EVI \).
- We then perform the following regression on each implied volatility series, \( \ln EVI_t(k, \tau) = \alpha + \beta \ln I^2(k, \tau) + e \).
- \( \beta = 1 \) under proportional 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>1.42 ( 0.03 )</td>
<td>1.62 ( 0.03 )</td>
</tr>
<tr>
<td>3</td>
<td>1.21 ( 0.03 )</td>
<td>1.16 ( 0.03 )</td>
<td>0.97 ( 0.02 )</td>
<td>1.24 ( 0.03 )</td>
<td>1.65 ( 0.03 )</td>
</tr>
<tr>
<td>6</td>
<td>1.29 ( 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>1.24 ( 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>1.28 ( 0.03 )</td>
<td>1.16 ( 0.03 )</td>
<td>1.07 ( 0.03 )</td>
<td>1.15 ( 0.03 )</td>
</tr>
<tr>
<td>36</td>
<td>1.47 ( 0.04 )</td>
<td>1.37 ( 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>1.48 ( 0.05 )</td>
<td>1.38 ( 0.04 )</td>
<td>1.26 ( 0.04 )</td>
<td>1.16 ( 0.04 )</td>
<td>1.10 ( 0.04 )</td>
</tr>
<tr>
<td>60</td>
<td>1.44 ( 0.05 )</td>
<td>1.35 ( 0.05 )</td>
<td>1.25 ( 0.05 )</td>
<td>1.15 ( 0.05 )</td>
<td>1.07 ( 0.05 )</td>
</tr>
</tbody>
</table>
Dynamic estimation of the volatility surfaces

- The fact that the model has six state variables \((5 + \gamma_t)\) with no parameters prompts us to think of model estimation a bit differently.

- Treat the six covariates as the hidden state vector \(X_t\) and assume that the state propagates like a random walk: \(X_t = X_{t-1} + \sqrt{\Sigma_x} \varepsilon_t\)
  - One can assume more complicated dynamics. 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 and 40 expected volatility series.
  - \(h(\cdot)\) denote the model value (solutions to two quadratic equations)
  - Assume IID error for each of the two surfaces.

- The auxiliary parameters \((\Sigma_x, \sigma^2_e)\) 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:

$$\bar{X}_t = \hat{X}_{t-1}, \quad \bar{V}_{x,t} = \hat{V}_{x,t-1} + \Sigma_x,$$

$$\chi_{t,0} = \bar{X}_t, \quad \chi_{t,i} = \bar{X}_t \pm \sqrt{(k + \delta)(\bar{V}_{x,t})_j},$$

$$\bar{y}_t = \sum_{i=0}^{2k} w_i \zeta_{t,i}, \quad \bar{V}_{y,t} = \sum_{i=0}^{2k} w_i [\zeta_{t,i} - \bar{y}_t] [\zeta_{t,i} - \bar{y}_t]^\top + \Sigma_y,$$

$$\bar{V}_{xy,t} = \sum_{i=0}^{2k} w_i [\chi_{t,i} - \bar{X}_t] [\zeta_{t,i} - \bar{y}_t]^\top, \quad K_t = \bar{V}_{xy,t} (\bar{V}_{y,t})^{-1},$$

$$\hat{X}_t = \bar{X}_t + K_t (y_t - \bar{y}_t), \quad \hat{V}_{x,t} = \bar{V}_{x,t} - K_t \bar{V}_{y,t} K_t^\top.$$
### Model explained variation

<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>A. Implied volatility surface</td>
<td>B. Expected volatility surface</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.89</td>
<td>0.93</td>
<td>0.95</td>
<td>0.95</td>
<td>0.94</td>
<td>0.65</td>
<td>0.71</td>
<td>0.75</td>
<td>0.67</td>
<td>0.59</td>
</tr>
<tr>
<td>3</td>
<td>0.97</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>0.99</td>
<td>0.59</td>
<td>0.63</td>
<td>0.69</td>
<td>0.69</td>
<td>0.63</td>
</tr>
<tr>
<td>6</td>
<td>0.98</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.52</td>
<td>0.60</td>
<td>0.65</td>
<td>0.66</td>
<td>0.64</td>
</tr>
<tr>
<td>12</td>
<td>0.94</td>
<td>0.98</td>
<td>0.99</td>
<td>0.98</td>
<td>0.98</td>
<td>0.50</td>
<td>0.59</td>
<td>0.58</td>
<td>0.54</td>
<td>0.54</td>
</tr>
<tr>
<td>24</td>
<td>0.92</td>
<td>0.98</td>
<td>0.99</td>
<td>0.99</td>
<td>0.98</td>
<td>0.79</td>
<td>0.77</td>
<td>0.73</td>
<td>0.68</td>
<td>0.57</td>
</tr>
<tr>
<td>36</td>
<td>0.93</td>
<td>0.97</td>
<td>0.99</td>
<td>0.98</td>
<td>0.96</td>
<td>0.86</td>
<td>0.86</td>
<td>0.85</td>
<td>0.82</td>
<td>0.76</td>
</tr>
<tr>
<td>48</td>
<td>0.90</td>
<td>0.93</td>
<td>0.94</td>
<td>0.94</td>
<td>0.93</td>
<td>0.61</td>
<td>0.73</td>
<td>0.78</td>
<td>0.78</td>
<td>0.73</td>
</tr>
<tr>
<td>60</td>
<td>0.80</td>
<td>0.83</td>
<td>0.84</td>
<td>0.85</td>
<td>0.85</td>
<td>0.32</td>
<td>0.48</td>
<td>0.55</td>
<td>0.55</td>
<td>0.52</td>
</tr>
</tbody>
</table>

**Average:** 0.95 0.65

- The model fits the implied volatility surface reasonably well.
- The worse performance on the OEV surface partially reflects the forecasting error of the OEV estimates.
A. Instantaneous volatility rate ($\sqrt{v_t}$)

B. Instantaneous volatility drifts ($m_t$, $m_t^\text{P}$)

- The instantaneous volatility time series look similar to the implied/expected volatility time series.
- The risk-neutral drift ($m_t$) dictates the term structure slope — The term structure is mostly upward sloping except during the 2002 recession.
- The drift difference ($m_t - m_t^\text{P}$) reflects the volatility risk premium and its variation.
The volvol is high during high vol periods (crises and recessions).
The correlation is more negative (implied vol smile is more skewed) during the financial crises but not during the recessions.
Volatility risk premiums

A. Market price of volatility risk

B. Instantaneous volatility risk premium

- Volatility risk premium is computed as $w_t \sqrt{v_t \gamma_t}$, risk times market price of risk.
- High during crises, low otherwise.
Variance and return risk premiums

C. Instantaneous variance risk premium

\[ w_t \gamma_t \]

D. Instantaneous return risk premium

\[ v_t \rho_t \gamma_t \]

- Variance risk premium is computed as \( 2w_t v_t \gamma_t \), variance risk times market price of risk.
- Return risk premium is computed as \( v_t \rho_t \gamma_t \), ignoring the risk premium contribution from the independent return movement. The average is 4.35%.
- Expected return is a lot harder to estimate than volatility — Identifying return risk premium via volatility risk premium can be a feasible route.
Bollerslev, Tauchen, & Zhou (RFS, 2009) predict stock returns using $VIX^2 - RV^2$. The $R^2$ is 4% during our sample period.

Our variance risk premium estimate predicts returns much stronger — the benefit of using more information from the two surfaces.

The inferred return risk premium works even better — The prediction is not due to higher-order dynamics, but simply due to its strong relation with return risk premium.
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 and simple 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/expected 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 our newly-defined expected volatility surface defines the volatility risk premium for each option contract.
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 can predict future stock returns, with R-squared much higher than those reported in the literature.

Many open questions remain:

- How to generate the “expected” volatility surface forecast based on historical sample paths?
- 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?

Beyond volatility: Link fundamental-based equilibrium models with supply-demand shocks...