Option Pricing & Implied Volatility based on Market ‘Big Data’

TEAM 4

Mentor: Greg Fish

Member: Zijie Chen, Jianfeng Liu, Mingjian Zhou, Zongtai Li
Our Plan

- Analyze market-quote for options on E-mini S&P 500 future
- Analyze ‘one-moment’ quote data, and calculate the implied volatility with respect of strikes
- Expand our analysis to other moments
- Do time-series and statistics analysis
- Scour the data to determine whether there were potential trading opportunities
Our Plan

- Research on the unknown
  - We were not sure if we could figure out some decent results
- Work with different tools and backgrounds
  - Combine Financial Mathematics and Statistics
  - With MATLAB, R, and Excel spreadsheet
- Parallel Research
  - We choose to go different ways to explore
  - Arbitrage/Prediction
Outline for Today

- Data Processing with MATLAB
- Time-series (Bollinger Band)
- Probability Distribution
- Curve Fitting & Volatility Surface with MATLAB
- Time-series (ARIMA-GARCH) with R
  - Model Calibration
  - Prediction
- Stochastic-Volatility with R
  - Model Calibration
  - Prediction
Data Processing

- Approx. 5 million records of data
  - In on day (Nov. 20\textsuperscript{th}. 2017)

- Computing Speed Improvement
  - Brent’s Method
  - Takes about 250s – 500s to compute 50K data

- Data Structure Improvement
  - MATLAB Table (Data Frame in MATLAB)
Brent’s Method*

Linear Interpolation (secant method)

\[ s = \begin{cases} b_k - \frac{b_k - b_{k-1}}{f(b_k) - f(b_{k-1})} f(b_k), & \text{if } f(b_k) \neq f(b_{k-1}) \\ m & \text{otherwise} \end{cases} \]

\[ m = \frac{a_k + b_k}{2}. \]

Bisection

\[ |\delta| < |b_k - b_{k-1}| \]

\[ |s - b_k| < \frac{1}{2} |b_k - b_{k-1}| \]

Interpolation

\[ |\delta| < |b_{k-1} - b_{k-2}| \]

\[ |s - b_k| < \frac{1}{2} |b_{k-1} - b_{k-2}| \]

True: Interpolation
False: Bisection

*From Wikipedia*
Data Structure in MATLAB ‘Table’

- Table Object
  - Pandas.DataFrame
  - Data.Frame in R
- More flexible than Pandas
  - Data objects in MATLAB are more compatible
  - (Numpy<->Pandas)
- SQL may be better
Bollinger Band

- We take ATM Option (Call/Put)
We know a butterfly strategy:
- Long 1 Call Option C1 with K1
- Short 2 Put Option C2 with K2
- Long 1 Call Option C3 with K3

Similar to Gamma, we have the probability:

\[
Prob = \frac{(C_3 - C_2) - (C_2 - C_1)}{\frac{1}{2} \times (K_3 - K_1)} = \frac{C_1 - 2C_2 + C_3}{K_2 - K_1}
\]

where the future settle on the K2
Probability Distribution

![Settlement Probability Graph](image)
The original data contains missing points and noise.

The best way to abstract the data is fit a curve and re-sample to reconstruct the data.
Curve Fitting (Polynomial)

- \( F(x) = C + a_1 x + a_2 x^2 \cdots a_n x^n \)

- Using R to fit the data and comparing the different models by F statistic and T statistic

- Find the best model and get the parameters
Order comparison

When the order is Two

When the order is Three

When the order is Four

When the order is Five

Red: Original Data    Black: Fitted Curve
3rd order polynomial to avid over fitted

After 3rd order The coefficients are too small and make no sense.
Volatility surface with MATLAB

Original Data

Fitted Curve
Time Series Analysis

- **ARIMA model**

\[ X_t - \alpha_1 X_{t-1} - \cdots - \alpha_p X_{t-p'} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}, \]

- **Differencing**

\[ y'_t = y_t - y_{t-1} \]

\[ y^*_t = y'_t - y'_{t-1} = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) = y_t - 2y_{t-1} + y_{t-2} \]
Time Series Analysis

- Sequence chart of ATM option volatility
Test the stationary of the sequence $X_t$

$H_0$: A unit root is presented

Result: $p = 0.5947 > 0.05$. Do not reject the null hypothesis.

Differencing: $Y_t = X_t - X_{t-1}$

Result: $p < 2.2e-16$. Reject the null hypothesis. We get the stationary sequence.
Time Series Analysis

- First differencing sequence chart of ATM option volatility
Time Series Analysis

- Preliminary define the model as: ARIMA(5, 1, 4)
Time Series Analysis

Residual Analysis

1. Ljung–Box test
   \[ H_0: \text{The data are independently distributed} \]
   Result: \( p = 0.94 > 0.05 \). Do not reject the null hypothesis.

2. McLeod Li test
   Result: There’s no garch effect.
Time Series Analysis

- Forecast

![Graph of Time Series Analysis]
Stochastic Volatility Analysis

- We try to use the Stochastic volatility model to get the estimation and prediction of implied volatility.

- The “stochvol” package in R can perform this quite well.

- The R package “stochvol” utilizes Markov Chain Monte Carlo (MCMC) sampler to conduct inference by obtain draws from the posterior distribution of parameters and latent variables which can be used for predicting future volatilities.
Set \( y = (y_1, y_2, \ldots, y_n) \) as a vector of returns with mean zero. Each observation \( y_t \) has its “own” contemporaneous variance \( e^{h_t} \).

The SV model can be conveniently expressed in such form:

\[
\begin{align*}
y_t | h_t & \sim N(0, \exp h_t) \\
h_t | h_{t-1}, \mu, \phi, \sigma_{\eta} & \sim N(\mu + \phi(h_{t-1} - \mu), \sigma^{2}_{\eta}) \\
h_0 | \mu, \phi, \sigma_{\eta} & \sim N(\mu, \frac{\sigma^{2}_{\eta}}{1 - \phi^{2}}) \\
\end{align*}
\]

\( h = (h_0, h_1, \ldots, h_n) \) is the latent time-varying volatility process (log-variance process).

\[
\begin{align*}
\mu & \sim N(b_\mu, B_\mu), (\phi + 1)/2 \sim B(a_0, b_0), \pm \sqrt{\sigma^{2}_{\eta}} \sim N(0, B_{\sigma_{\eta}})
\end{align*}
\]
Estimated Volatility

- Strike = 2600, Call option
- The latent volatilities in percent
- Priormu=N(0,100), Priorphi=\( B(10,1.5) \), Prior\( \sigma \)=0.1
Estimated volatilities in percent (1% / 10% / 50% / 90% / 99% posterior quantiles)
Posterior Draws and Distribution of Parameters

Posterior draws of parameters (thinning = 1):

<table>
<thead>
<tr>
<th>Parameter</th>
<th>mean</th>
<th>sd</th>
<th>5%</th>
<th>50%</th>
<th>95%</th>
<th>ESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>mu</td>
<td>-5.467</td>
<td>0.301</td>
<td>-5.893</td>
<td>-5.463</td>
<td>-5.038</td>
<td>2356</td>
</tr>
<tr>
<td>phi</td>
<td>0.802</td>
<td>0.161</td>
<td>0.483</td>
<td>0.845</td>
<td>0.978</td>
<td>149</td>
</tr>
<tr>
<td>sigma</td>
<td>0.364</td>
<td>0.171</td>
<td>0.128</td>
<td>0.340</td>
<td>0.677</td>
<td>151</td>
</tr>
<tr>
<td>exp(mu/2)</td>
<td>0.066</td>
<td>0.011</td>
<td>0.053</td>
<td>0.065</td>
<td>0.081</td>
<td>2356</td>
</tr>
<tr>
<td>sigma^2</td>
<td>0.162</td>
<td>0.146</td>
<td>0.016</td>
<td>0.116</td>
<td>0.459</td>
<td>151</td>
</tr>
</tbody>
</table>
Strike=2600, Put option
Residual

The dashed line indicates 2.5%/97.5% quantiles of the standard normal distribution.
The possible influence of the prior distributions of parameters

Analyzing three kinds of errors:
- homoscedastic
- SV
- GARCH(1,1)

Using empirical data to test the prediction results
Variance of prior $\mu=75$

Variance of prior $\mu=50$

Variance of prior $\mu=25$

Variance of prior $\mu=10$
Priorphi (5,1)

Priorphi (5,1.5)

Priorphi (10,1)

Priorphi (10,1.5)
Conclusion

THANKS