Estimation of Options Implied Volatility Curve

Seungki Min (mskyt@snu.ac.kr, www.mskyt.net)

• This document describes how I used mathematical methods for the estimation of implied volatility curve, a crucial engineering problem raised in our High Frequency Trading (HFT) strategy. I apologize for using many of jargons and unformatted figures. This is because the materials were not prepared for an academic purpose, but just for internal usage.
• All data used in this document are about KOSPI200 index futures and options.

1 Introduction

This estimation problem is to find the true values of implied volatility (IV) for every strike price over time. In the below figure, the green curve is such an estimated IV curve that nicely fits the market.

![Figure 1: Puts’ IV points (blue), calls’ IV points (purple), and estimated IV curve (green)](image)

For what? The IV curve is used to calculate the theoretical prices of options via Black-Scholes formula. The theoretical prices are unquestionably important in our strategy for detecting chances of arbitrage, and evaluating our positions.
Why implied volatility? In comparison to price, implied volatility is much more interpretable and time-invariant. Especially, the shape of IV curve, the so-called volatility smile, reflects the anticipation of market participants.

Our estimation process consists of two tasks as follows:

1. **IV curve fitting at a snapshot of the market**: Based on our own IV curve model, we have to find the optimal parameters that make the curve fit the market at the moment. Such parameters are measured every second.

2. **Curve parameter estimation over time**: We have to remove noise from the measurements of (1), then estimate the true values.

## 2 Curve Modeling

There are a variety of mathematical models for IV curve such as Corrado-Su or SABR, usually based on non-lognormal distribution of underlying asset price or non-constant volatility process. However, these models have significant discrepancy with the actual market. Rather, we use a geometric model to directly represent the shape of IV curve without any assumption.

![Figure 2: Our geometric modelling of implied volatility curve](image)

As shown above, we set several anchor points on the plane of volatility versus log-moneyness, and connect them smoothly via cubic spline interpolation. Moneyness is a ratio of a strike price to current underlying price, and its logarithmic value (log-moneyness) is the de facto standard for the horizontal axis of the IV curve, which we use too.

### 2.1 Parameters

In our model, a complete IV curve can be calculated if following parameters are given:
$F$: Current price of futures with the same underlying asset. Given by market.

$T$: Time to expiration. Given by market.

$r$: Interest rate. Given by market.

$b$: Basis between index futures and implied futures. For the quarterly options, both have the same expiry, thus $b = 0$. Almost given.

$\sigma_{50}$: Implied volatility of at-the-money (ATM) anchor point (0.5 delta)

$\sigma_{98}$: Implied volatility of d98 anchor point (0.98 call delta, -0.02 put delta)

$\sigma_{84}$: Implied volatility of d84 anchor point (0.84 call delta, -0.16 put delta)

$\sigma_{16}$: Implied volatility of d16 anchor point (0.16 call delta, -0.84 put delta)

$\sigma_{2}$: Implied volatility of d2 anchor point (0.02 call delta, -0.98 put delta)

downSlope: Slope of the leftmost segment of the curve.

upSlope: Slope of the rightmost segment of the curve.

**Why the futures price**, rather than the underlying asset price? Mainly because we cannot directly trade the underlying asset at that price. And partly because of data accessibility and data frequency. Thus, we use the futures price and Black model in our strategy.

**Why deltas of 0.98, 0.84, 0.16, 0.02?** The delta, one of the greeks, has an alternative meaning of ‘the likelihood of expiring in-the-money’. If a call option has 0.16 delta, then the probability of an event that the underlying asset expires above its strike price is 0.16, which means an ‘1-sigma’ event. Similarly, 0.98 or 0.02 delta means a ‘2-sigma’ event. This approach gives much more time-invariant parameters than when the anchor points have the fixed strike prices.

In detail, the x-values (log-moneyness) of ATM, d98, d84, d16, and d2 anchor points are calculated as follows:

\[
X_{98} = -2\sigma_{98}\sqrt{T} + \frac{1}{2}\sigma_{98}^2 T \\
X_{84} = -\sigma_{84}\sqrt{T} + \frac{1}{2}\sigma_{84}^2 T \\
X_{50} = \frac{1}{2}\sigma_{50}^2 T \\
X_{16} = +\sigma_{16}\sqrt{T} + \frac{1}{2}\sigma_{16}^2 T \\
X_{2} = +2\sigma_{2}\sqrt{T} + \frac{1}{2}\sigma_{2}^2 T 
\]

The **cubic spline interpolation** is used to get a smooth curve that passes through all five anchor points. For the strike prices inside the anchor points’ range, from 0.98 delta to 0.02 delta, the IV values can be calculated with this interpolation. For the strike prices out of this range, the IV values are calculated with a **linear extrapolation**, parameterized by downSlope and upSlope.
2.2 Model Validation

In our model described above, the shape of the IV curve is mainly determined by seven parameters ($\sigma_{50}$, $\sigma_{84}$, $\sigma_{16}$, $\sigma_{98}$, $\sigma_{2}$, downSlope, and upSlope), i.e., our model has a dimension of seven. It is notable that we can freely adjust its dimensionality by setting more or less anchor points.

It is valuable to check whether our model matches the dimensionality of the actual market data space, i.e., whether our IV curve model is flexible enough to describe the time-varying IV curve, but not over-fitted to the noise. We used Principal Component Analysis (PCA) to test this.

![Image 1](image-url)  
Figure 3: Implied volatility curve from 22-Sep-2014 to 2-Oct-2014

The actual IV curve, extracted from the live market, varies continuously over time, as shown in the above figure. We applied PCA for the three days of data, a $6105 \times 17$ matrix that consists of 6105 snapshots of IV curve for 17 strike prices. Each 17 column had been regularized by subtracting the mean of that column.

![Image 2](image-url)  
Figure 4: Results of PCA on the empirical data (left and middle), and the curve changes made by our modeling parameters (right)
Since we did not model the IV curve as a linear superposition of orthogonal components, the results of PCA would not be directly compared to our model. Nevertheless, as shown in the above figures, PCA’s most dominant components are similar to our model’s. And our model’s dimensionality, five\(^1\), seems to be reasonable when comparing the relative sizes of principal components.

\(^1\)The parameters \textit{downSlope} and \textit{upSlope} were not counted, because, in the data analyzed above, their scopes were out of the strike price range 240-280.
3 Curve Fitting

In our strategy, it is required to do curve fitting at a snapshot of the actual market every second, i.e., find a set of parameters that minimizes the discrepancy between our model and the market at the moment of fitting.

3.1 Object Function Design

The objective function, discrepancy or fitting cost that should be minimized, was carefully designed as a weighted sum of mis-pricing of individual options.

For a single option:

\[ f_i(P_{theo}^i(\Theta), P_{bid}^i, P_{ask}^i) = \log(1 + \exp(P_{bid}^i - P_{theo}^i)) + \log(1 + \exp(P_{theo}^i - P_{ask}^i)) \]

\[ P_{theo}^i(\Theta) : \text{Theoretical price of the } i\text{th option for the given parameter set } \Theta. \]
\[ P_{bid}^i, P_{ask}^i : \text{Market bid and ask price of the option at the moment of fitting.} \]

As a total:

\[ \epsilon(\Theta) = \frac{1}{N} \sum_{i} w_i \times f_i \]

\[ w_i : \text{Weight function that puts more weight on the out-of-the-money (OTM) options.} \]

Figure 5: Object function for a single option

The objective function for a single option looks like the above figure. It is designed to represent the amount of mis-pricing, which is approximately zero if the theoretical price is between bid price and ask price, and increases linearly as the theoretical price moves far outside of the bid-ask spread.

Note that the arguments are prices, not implied volatility. We do not directly fit our IV curve to the market’s IV curve. Rather, we convert our IVs to prices with Black model, then compare them with
the market prices. This saves lots of computation cost required to calculate IVs from prices, which is the inverse of Black model, usually implemented with bisection method.

3.2 Numerical Optimization

The objective function is almost convex over the entire parameter space. Especially near the optimal solution, it is much well convex as shown in the below figure.

![Figure 6: Objective function over ATM volatility $\sigma_{50}$ and basis $b$ domain (left), and $d98$ and $downSlope$ domain (right)](image)

Thanks to the convexity and continuity of this objective function, we can numerically find the local minima. For each parameter $\sigma_{50}$, basis $b$, $\sigma_{84}$, $\sigma_{16}$, $\sigma_{98}$, $\sigma_2$, $downSlope$, and $upSlope$, we try to change the parameter value by a small amount, then accept the change if the objective function is decreased. This iteration is repeated until any change cannot improve the objective function. Especially, to accelerate this convergence process, the size of change decreases over iterations, from a large step to a small step.

In the ordinary cases, it takes less than 100 iterations to find a new solution for a new snapshot, starting from the previous solution for the previous snapshot. Although the optimization process does not guarantee the global minima, the solution is highly reliable once we had reached the global minima, and it is indeed. In the extreme cases, such as when some parameters have marginal influence on the IV curve, or when the snapshot does not have enough information, the solution occasionally gets stuck at a non-global minima. In fact, we made great efforts to deal with these exceptions, but too detailed for this document.

4 Estimation

Through the curve fitting process, one set of IV curve parameters is measured from the market per one second. Such measurements, in the form of time series, contain lots of noise, which should be
removed to use. But the simple methods of noise filtering such as moving average inherently cause lagging. It is required to intelligently estimate the parameters by utilizing additional information: system dynamics and confidence of measurement.

4.1 Dynamics

![Figure 7: Movement of ATM implied volatility versus the futures price (the traces of ATM points are displayed with the color gradient from light gray to dark black)](image)

The IV curve parameters are not mutually independent. Especially they are strongly correlated with the futures price. As shown in the above figure, ATM implied volatility and the futures log-scaled price have a quite strong linear relationship. This linear relationship is hard to be modeled with explicit equations, because its slope varies unpredictably over time. We use simple linear regression for the recent traces of ATM point to measure the slope. The ‘recent’ points are carefully selected with a number of heuristic criteria. This dynamics, linearly simplified, is greatly helpful to follow the rapidly changing market, without lagging.

4.2 Bayesian Estimation

We use Bayesian estimator to obtain stable values from the noisy time series of parameters. Specifically, we interpret the value of the objective function as a measurement error of the parameters in curve fitting; the larger value, the larger mis-pricing the curve has, thus the parameters are likely to be erroneous. By putting more weights on the reliable measurement, which has the small discrepancy with the market, Bayesian estimator effectively filters the noise. In more detail, we slightly modified the estimator that its uncertainty of prior belief grows exponentially over time until a new measurement arrives, as follows:

\[
p(\theta | x_1, \cdots, x_t) \propto N(\hat{\theta}_t, \sigma_t^2)
\]

\[
p(\theta_{t+1} | x_1, \cdots, x_t) \propto N(\hat{\theta}_t, \sigma_t^2 + \sigma_s^2 \exp\left(\frac{\Delta t}{\tau_s}\right))
\]
\[
\frac{1}{\sigma_{t+1}^2} = \frac{1}{\sigma_t^2} + \frac{1}{\epsilon_{t+1}^2}
\]

\[
\hat{\theta}_{t+1} = \hat{\theta}_t - (\hat{\theta}_t - x_{t+1}) \left( \frac{\sigma_t^2 + \sigma_s^2 \exp(\triangle t/\tau_s)}{\sigma_t^2 + \sigma_s^2 \exp(\triangle t/\tau_s) + \epsilon_{t+1}^2} \right)
\]

Combined to the linear dynamics described above, the prior belief is shifted by the futures price movement, then updated with the new observation (measurement). This estimation process works in a similar manner to the Kalman filter.

Figure 8: Estimated basis \(b\) (green) from the noisy measurements (red)
5 Result

In fact, we did not measure the performance of this estimation process as a separated procedure. Rather, we verified the improvement of the entire strategy in terms of profit, via backtests or simulations with the empirical market data, then confirmed from the live market. With these mathematical investigations, for the other engineering problems as well, we have achieved very stable profits as shown in the below graph\(^2\). You can see that ‘intraday’ profit also had increased steadily. Especially, only one day of trading losses during the September we had.

![Figure 9: Accumulated profits of our High Frequency Trading strategy in an account in Korea](image)

\(^2\)The trading fees were not discounted in the plot. It takes approximately 30% of the profit as we trade a lot.