Volatility Surface Construction

December 16, 2015

Introduction

Binary.com uses a modified Stochastic alpha, beta, rho (SABR) model, a stochastic volatility (vol) model, to estimate the volatility smile in the derivatives market. For a more general overview of the general SABR model, please refer to (Clark, 2011).

An expression for the market-implied Black-Scholes-Merton volatility $\sigma_X(K,T)$ at strike K and expiry T, following Clark's (2011) terminology, can be written as

$$\sigma_X(K,T) = \frac{\alpha z}{\chi(z)}\left[1 + \left(\frac{\rho\nu\alpha}{4} + \frac{2-3\rho^2}{24}\nu^2\right)T\right] \tag{1}$$

where α is at-the-money (ATM) volatility, ν is volatility of volatility, ρ is correlation between spot and volatility, and we use

$$z(K,T) = \frac{\nu r}{\alpha}\tanh\left(\frac{\log\left(\frac{F_{0,T}}{K}\right)}{r}\right)\sqrt{\frac{t_0}{T}}, \tag{2}$$

where $F_{0,T}$ is ATM spot, $t_0 = 1$ yr,

$$\chi(z) = \log\left(\frac{\sqrt{1-2\rho z+z^2}+z-\rho}{1-\rho}\right).$$

The parameter r above is set as

$$r = \begin{cases}\log(s_+) & \text{if } F_{0,T} > K \\ \log(s_-) & \text{if } F_{0,T} \le K\end{cases},$$

where s_+, s_- are constants that represent maximum and minimum desired strike ratios. Also, we have used used the stochastic normal [Clark's (2011) $\beta = 1$] assumption as opposed to the stochastic lognormal ($\beta = 0$) or more general hybrid ($0 < \beta < 1$) implementations. In this respect, we could say we use the SAR (Stochastic alpha and rho) model. Note that the time t_0 in Eq. 2 is required for dimensional consistency and either should to be inserted in production code or removed everywhere with T always specified in years. In the standard

SAR model (Eq.1) the dependence on maturity, or term structure, comes from the factor in the square brackets with other factors remaining static. Our implementation, however, discards the square-bracket part and asserts maturity dependence with $\alpha(T)$ and $z(K,T)$. Binary.com's volatility surface is

$$\sigma(K,T) = \frac{\alpha(T)\, z(K,T)}{\chi(z)}$$

Definitions

- Moneyness is always calculated with respect to spot F_0.

$$\text{Moneyness} = K/F_0.$$

- Correlation denotes correlation between stock returns and implied volatility of the options. This is not simply the statistical correl function in excel. It is calculated using some other function based on kurtosis and skew. Discussed later.

- Skew is Implied Volatility of 90% Strike minus Implied Volatility of 100% Strike. This is scaled by square root of maturity.

Overview

Any distribution can be visibly perturbed using ATM Vol, Skew & Kurtosis. This model accepts these parameters and some other finer parameters to generate an extremely smooth surface.

When skew is very high it means OTM puts are priced at high premiums compared to ATM. Likewise if skew is very low OTM puts are priced at low premiums compared to ATM. When skew increases from low to high, what is the behavior of correlation between stock & volatility?

When the market is falling drastically and volatility is increasing this implies high Correlation. When the market is stable, and volatility is low implies low Correlation. Hence it can be said that the correlation between an index and volatility increases as negative skew increases. Kurtosis signifies that the volatility is clustered around the mean or that the distribution of ATM vol for different maturities will be similar. So high kurtosis means vols are clustered in the middle of the distribution and index movements will not cause much change in volatility. Hence lower correlation.

Similarly when low Kurtosis in volatility signifies that volatility is very dispersed or that small changes in the index might cause higher volatility changes. Hence for this case correlation between the index and volatility is high. So Kurtosis High implies Correlation Low, Kurtosis Low implies Correlation High.

Some quants might question this approach and say "Why not just use a simple correl function to calculate correlation between index returns and volatility?". In a perfect world this should be done but since the correlation value obtained using the correl function (assuming excel) is unstable, this is not a practical approach. This leads to unstable calibration parameters from one day to another.

Functional Forms

ATM Volatility and Skew

The calibration approach is based upon modeling the term structure of ATM Volatility and ATM Skew using exponential functions. It is widely observed that any ATM Vol term structure or skew term structure is convex (mostly) and rarely concave. An exponential function is in general a convex function.

We try to study this using the functional form:

$$w_1 e^{x_1} - w_2 e^{x_2}$$

The above form allows us to create both concave and convex functional forms or curves that are continuously differentiable. With appropriate scaling we can model the ATM term structure and skew based on the above equation. This will become our base function for calibrating ATM Vol and Skew term structure.

Volatility-Surface Model Parameters

We use the following parameters for our calibration:

Parameter	PERL Identifier	Dimensions
a_1	atmvolshort	$\mathrm{T}^{-1/2}$
a_2	atmvol1year	$\mathrm{T}^{-1/2}$
a_3	atmvolLong	$\mathrm{T}^{-1/2}$
a_4	atmWingL	T^{-1}
a_5	atmWingR	T^{-1}
a_6	skewshort	$\mathrm{T}^{-1/2}$
a_7	skew1year	$\mathrm{T}^{-1/2}$
a_8	skewlong	$\mathrm{T}^{-1/2}$
a_9	skewwingL	T^{-1}
a_{10}	skewwingR	T^{-1}
a_{11}	kurtosisshort	$\mathrm{T}^{-1/2}$
a_{12}	kurtosislong	$\mathrm{T}^{-1/2}$
a_{13}	kurtosisgrowth	T^{-1}
a_{14}	strikeDn	*
a_{15}	strikeUp	*

* dimensionless

For ATM vol we use

$$\alpha^2(T) = a_3^2 + \frac{\left(a_2^2 - a_3^2\right)\left(e^{-a_5 T} - e^{-a_4 T}\right) + \left(a_1^2 - a_3^2\right)\left(e^{-a_5 t_0 - a_4 T} - e^{-a_4 t_0 - a_5 T}\right)}{e^{-a_5 t_0} - e^{-a_4 t_0}}$$

Note that we are scaling variance with the exponential functions not the actual vols. This is as per the rationale that variance is scalable and can be linearly interpolated but volatility is the square root of variance so it is not easy to interpolate. Also variances are additive. The exponential function parameters a_4 =atmWingL and a_5 =atmWingR control the behavior of the vol term structure and have units of inverse time. If atmWL is greater than atmWR then the left side of the curve is lifted up, while the right side decreases. This effect is necessary for keeping the term structure smooth and preserving the shape of the smile on both sides of the ATM vol. We may also check that $\alpha(0) = a_1$, $\alpha(t_0) = a_2$, and $\lim_{T\to\infty} \alpha(T) = a_3$. Presumably this explains the naming of these parameters, but be cautious not to regard these as any meaningful volatility values.

For volatility skew (γ) we use

$$\gamma(T) = a_8 + \frac{(a_7 - a_8)\left(e^{-a_{10}T} - e^{-a_9 T}\right) + (a_6 - a_8)\left(e^{-a_{10}t_0 - a_9 T} - e^{-a_9 t_0 - a_{10}T}\right)}{e^{-a_{10}t_0} - e^{-a_9 t_0}}.$$

Here we find that $\gamma(0) = a_6$, $\gamma(t_0) = a_7$, and $\lim_{T\to\infty} \alpha(T) = a_8$.

For volatility of volatility, we use

$$\nu = \frac{2\gamma(T)}{\rho(T)},$$

where correlation ρ is described below.

$$\rho(T) = -\left[\frac{3}{2}\left(1 + \frac{k\alpha}{\gamma^2}\right)\right]^{-\frac{1}{2}}$$

and volatility kurtosis k is

$$k(T) = a_{12} + (a_{11} - a_{12})\, e^{-a_{13}T}.$$

As with α and γ, we can check for certain maturities, $k(0) = a_{11}$ and $\lim_{T\to\infty} k(T) = a_{12}$.

Example :
atmvolshort 0.13
atmvol1year 0.19
atmvolLong 0.12
atmWingL 1.27
atmWingR 1.56

In the example above these 5 points are mentioned in the same order. The two wings are basically the weights to stretch the ATM term structure on either end. As an input the exponential function doesn't take the volatilities but variances since variances can be linearly interpolated.

And similarly for skew exactly the same logic is applied for skew parameters.

1. Short term skew

2. 1 Year skew
3. Longest term skew

The ATM Vol term structure generated by the functional form will match the surface's ATM Vols approximately. The skew term structure similarly calculated from the functional form will match the skew calculated from the surface.

Skew Params & Values

Example:
skewshort -27%
skew1year -8%
skewlong -3%
skewwingL 2.01
skewwingR 2.03
All this generates the skew and atm vols only (not the surface).

Functional Relationships

Skew is asymmetric. When skew increases the left side shifts up and the right side shifts down. When skew decrease similarly the left side shifts down, and the right side shifts up. Kurtosis on the other hand provides a symmetric control over the wings of a surface. Kurtosis basically shifts the wings of the curve in a symetric way. We see that

$$k \text{ is large } \implies \rho \text{ is small}$$

and

$$k \text{ is small } \implies \rho \text{ is large.}$$

Similar logic applies to skew. When skew is higher it means OTM options are priced higher and might result in volatility increasing even for a slight fall in the market. So Spot movement is very much correlated with volatility when skew is high.

Sabr Tanh Parameters (Flattening):
strikeUp 120%
strikeDown 80%

Two extra params for control are strikeDn and strikeUp. These are the two strike limits beyond which the curvature effect is replaced by a flattening or flattening which occur beyond the designated strikes.

Volatility Calculation

As discussed previously, Eq.2 contains the tanh function. The limits of tanh are between -1 and +1, and since it is symmetric it is well suited for volatility surface modeling. After building the ATM and skew term structure the other strikes are weighted based on moneyness. The function below builds the surface.

Notation: Moneyness is calculated with respect to spot.

$$x = \frac{\log\left(\frac{K}{F_{0,T}}\right)}{\alpha}\sqrt{\frac{t_0}{T}}$$

So we basically scale moneyness with the tanh function. When strike is very near to ATM, the computation is numerically unstable, so we perform the following test:

```
    If Abs(x) < 0.0000000001 Then
        sabrVol2 = atmvol
        Exit Function
    End If
```

The tanh function is used to extend the curve between the ATM and the end points. The end point on one end is StrikeDn and for the other end is StrikeUp. It's basically a sort of trigonometric interpolation between two moneyness levels.

1. The ATM and the Upward extremum on one side
2. The ATM and the Downward extremum on other side

```
If (x > 0) Then
    x = xmax * tanh(x / xmax)
    Else
    x = xmin * tanh(x / xmin)
    End If
```

Since the functional form of the volatility as a whole is everywhere differentiable, it is continuous. Hence we can always get local volatility for any of the spot and time points. Second derivative problem never happens because the dependence on K is only through exponentials.

Optimization

We use a form of the Downhill Simplex Method or Nelder-Mead (available as the R function optim). This can also be coded in other languages. In the excel solver we can specify constraints using variable conditions. Here the function which is specified as parameter applies the constraints. For instance if I don't want ATM vol to go above 0.5, I will return a residual value of 1000 or any other large number. Thus, the optimization function that is calling my function will understand that ATM vol should not go above 0.5 (or 50 percent). It basically minimizes the function by assuming a simplex (or N vertices polygon). It is the shape by reflection, expansion, contraction and reduction operations on the geometrical figure. This results in an excellent optimization in few steps. A major problem with the above algorithm is that it will stop at the local minima. Given the large number of fitted parameters and nonlinear implied volatility function, using a global optimization technique such as simulated annealing or basin hopping should be used.

Validation

There are 8 checks we use to make sure we have a valid volatility surface.

1. Inputs
2. Identical Surface
3. Volatility Jump
4. Spot Reference
5. Age
6. Smile
7. Term Structure
8. Smile Admissibility

Checks 1 through 6 consist of a number of sanity verifications. Check 7 aims to eliminate arbitrage possibilities across tenor and check 8 prevents arbitrage across strikes. The following describes these final two checks.

Term Structure and Smile Admissibility

As necessary and sufficient conditions to prevent arbitrage possibilities (Roper, 2010), the implied European call price surface $C(K,T)$ must have the following properties.

1. $\frac{\partial C(K,T)}{\partial T} \geq 0$. We check that for any consecutive call prices ordered by maturity at a given strike, the second price is less than or equal to the first. (Calendar arbitrage check)
2. $\frac{\partial^2 C(K,T)}{\partial K^2} \geq 0$. We check that implied European call prices for any 3 consecutive strike points, the middle strike value is less than or equal to the average of the sum of the values of the end points. (Admissibility Check #2)
3. $C(K,T) \geq \max(S-K,0)$.
4. $C(K,0) = \max(S-K,0)$.
5. $\lim_{K\to\infty} C(K,T) = 0$. We check that the implied call price for large K is close to zero. (Admissibility Check #5)

Because we are interested in digital options rather than European-style options, we use the following checks in place of (3) and (4) in the list above.

- $\frac{\partial C(K,T)}{\partial K} \leq 0$. We check that for any consecutive strike points, the second price is less than or equal to the first. (Admissibility Check #1)
- $\lim_{K\to 0} C_{\text{dig}}(K,T) = 1$. We check that the implied digital call price for small K is close to unity. (Admissibility Check #4)