Binary.com’s volatility calibration model (a variant of SABR)

Introduction

The SABR (Stochastic alpha, beta, rho) model is a stochastic volatility model, which is used to estimate the volatility smile in the derivatives market. For a more general overview of the general SABR model, please refer https://en.wikipedia.org/wiki/SABR_volatility_model.

The implied vol approximation from the general SABR model is approximately given by

$$\sigma_{impl} = \alpha \frac{\log\left({}^{F_0}/_{K}\right)}{D(\zeta)} \left\{ 1 + \left[\frac{2\gamma_2 - \gamma_1^2 + {}^{1}/_{F_{mid}^2}}{24} \left(\frac{\sigma_0 C(F_{mid})}{\alpha} \right)^2 + \frac{\rho\gamma_1}{4} \frac{\sigma_0 C((F_{mid})}{\alpha} + \frac{2 - 3\rho^2}{24} \right] \varepsilon \right\}$$

where all the terms are as defined in the Wikipedia page mentioned above.

This term is basically expressing implied volatility as a function of some sort of moneyness function (the alpha Log(F0/K)/D(....) part. This term is being adjusted by some factor in square brackets and then added back. We are going to modify the term in square brackets in this model.

Definition

Moneyness is always calculated with respect to spot.

Moneyness = Strike/Spot.

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: 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 Correlation = High. When the market is stable, and volatility is low implies that Correlation = Low. 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 form day 1 to day 2 (any two consecutive days).

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 in the form:

$$w1 \times e^{x1} - w1 \times e^{x2}$$

The above form allows us to create both concave and convex functional forms or curves that are continuously differentiable in nature. 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.

Both variance and skew can be estimated by something like :

$$(var1year - varlong)\left(\frac{e^{-atmWR\times T} - e^{-atmWL\times T}}{e^{-atmWR} - e^{-atmWL}}\right) + (varshort - varlong)\left(\frac{e^{-atmWR-atmWL\times T} - e^{-atmWL-atmWR\times T}}{e^{-atmWR} - e^{-atmWL}}\right)$$

where WL and WR stand for Wing Left and Wing Right (direction as per the curve).

Note that we are scaling variance with 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 atmWL and atmWR control the behavior of vol term structure. 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.

SABR Tanh Model Parameters

Example :

atmShort	20%
atm1Year	24%
atmLongTerm	25%
atmWL	1.00
atmWR	1.02

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	-21%
skew1year	-14%
skewLongTerm	-10%
skewWL	1.00
skewWR	1.02

All this generates the skew and atm vols only (not the surface).

Kurtosis Functional Form

Lastly kurtosis can be manipulated using a simple growth rate function. In the below equation , K equals kurtosis

$$K = kshort + (klong - kshort)\times(1 - e^{-growth\times T})$$

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.

Correlation Functional Form

$$correl = \frac{1}{\sqrt{3\times\left(1 + kurtosis\times {}^{atmvol}/_{skew^2}\right)}}$$

It uses a standard function similar to $\frac{1}{1+x}$

It basically decreases to 0 as x approaches large values. We take the square root of this function to make sure that value of x is always positive. The x term in our case is

$$\frac{kurtosis\times atmvol}{skew^2}$$

This whole expression is scaled $\frac{1}{\sqrt{3}}$ for better adjustment.

We see that this is infact true for our functional form, from

Kurtosis = High, implies Correlation = Low.
Kurtosis = Low, implies Correlation = High

The value x is directly proportional to kurtosis. So when x is high we have low correlation or in other words high kurtosis = low correlation and vice versa.
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, the term in square brackets is modified with the tanh function.

$$\sigma_{impl} = \alpha \frac{\log\left({}^{F_0}\!/_{K}\right)}{D(\zeta)} \left\{ 1 + \left[\frac{2\gamma_2 - \gamma_1^2 + {}^{1}\!/_{F_{mid}^2}}{24} \left(\frac{\sigma_0 C(F_{mid})}{\alpha} \right)^2 + \frac{\rho\gamma_1}{4} \frac{\sigma_0 C((F_{mid})}{\alpha} + \frac{2 - 3\rho^2}{24} \right] \varepsilon \right\}$$

The limits of tanh are between -1 and +1, and since it is symmetric it is very much suitable for the 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.

$$x = log \frac{{}^{strike}\!/_{forward}}{atmvol \times \sqrt{tenor}}$$

So we basically scale moneyness with the tanh function.

$$xmin = log \frac{(strikeDn)}{atmvol \times \sqrt{tenor}}$$

$$xmax = log \frac{(strikeUp)}{atmvol \times \sqrt{tenor}}$$

Just a minor check if strike is very near to ATM :

```
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 trignometric 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

Z = -volvol * x

asinh = (Sqr((1 - 2 * corr * Z + Z * Z) - corr + Z)) / (1 - corr)

Final Vol = atmvol * Z / asinh
```

Advantages over other models

1. Direct input of ATM Vol points as parameters
2. Direct input of skew points as parameters

Since the functional form of the volatility as a whole is continuous function it is differentiable everywhere. Hence we can always get local volatility for any of the spot and time points. Second derivative problem never happens because of this continuously differentiable feature.

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 a perfect optimization in few steps. One of the problems with the above algorithm was that it used to stop at local minima or local maxima. Here we are only concerned with local minima. This is prevented by using a technique called as simulated annealing which uses random shocks in passed parameters few times to get better optimization.