SABR goes normal

The benchmark stochastic alpha beta rho model for interest rate derivatives was designed for an environment of 5% base rates, but its traditional implementation method based on a lognormal volatility expansion breaks down in today's low-rate and high-volatility environment, returning nonsensical negative probabilities and arbitrage. ***Philippe Balland*** *and* ***Quan Tran*** *present a new method based on a normal volatility expansion with absorption at zero, which calibrates while eliminating arbitrage in the lower strike wing*

The stochastic alpha beta rho (SABR) model is the industry standard for interest rate derivatives. However, it was designed at a time when most curves were at much higher levels than today's ultra-low-rate environment. Problems with its implementation, through the so-called Hagan expansion, such as the breakdown of the expansion for high volatility and the possibility of negative probabilities for very low strikes, did not matter at the time but now constitute a pressing problem for the swaps and rates options markets. This article presents a solution to these problems.

We consider a forward rate S. The SABR model is based on the following dynamics:

$$dS_t = \sigma_t \varphi(S_t) dW_t$$
$$d\sigma_t / \sigma_t = \gamma dB_t$$

where the Brownian motions W and B have correlation ρ and are driftless under a probability measure Q. The function φ satisfies the usual linear growth and Hölder continuity conditions to ensure that the above stochastic differential equation admits a unique solution when appropriate boundary conditions are specified.

In the original SABR dynamics, $\varphi(S) = S^\beta$ and the forward rate is assumed to be absorbed at zero. The constant elasticity of variance (CEV) β is typically greater than zero and smaller than one in interest rates applications. Negative rates can be accommodated by assuming that $S + \Delta$ follows SABR dynamics. The model is very popular among practitioners because it provides an intuitive parameterisation of volatility smiles.

Unfortunately, the asymptotic formula derived by Hagan *et al* (2002) loses accuracy for long-dated expiries, especially when the CEV exponent is close to zero or when the volatility-of-volatility is large. This loss of accuracy is problematic from a practical point of view because the density can become negative near the forward.

New techniques have recently been proposed to improve the accuracy in the original expansion of the implied volatility. When the correlation is zero, Antonov & Spector (2011) derived an exact expression for the price of a vanilla option based on a double integral. When the correlation is non-zero, the authors proposed using an approximately equivalent SABR model with zero correlation.

Andreasen & Huge (2013) proposed solving the one-factor partial differential equation (PDE) corresponding to the equivalent SABR local volatility using a single implicit time-step. The solution is obtained by solving a single ordinary differential equation and delivers arbitrage-free option prices. However, this method does not address the lack of accuracy in the expansion with which traders are familiar, but rather defines a new smile interpolation that corresponds to a SABR process running on an independent gamma clock.

Small CEV exponents are typically used to represent swaption and caplet smiles at the long end of the curve, where the asymptotic formula also breaks down. Based on this observation, we perform an asymptotic expansion of the implied volatility corresponding to normal SABR with absorption at zero, instead of Black-Scholes. We find that the resulting approximation is more accurate than the original SABR expansion and results in significant calculation time saving when compared with solving the one-factor equivalent local volatility PDE.

Equivalent SABR local volatility

As explained in Andreasen & Huge (2013), Balland (2010) and Doust (2010), we can obtain an accurate approximation of the local volatility equivalent to SABR. The local volatility $g(t, K)$ for SABR is given by the following expression:

$$g(t,K)^2 = \frac{\varphi(K)^2 E\left[\sigma_t^2 \delta(S_t - K)\right]}{E\left[\delta(S_t - K)\right]}$$

We denote the numerator of this expression (the so-called local time) by L, and the denominator (the process's probability density) by D.

In this section, we derive an approximation for $g(t, K)$ by simple applications of Itô's lemma and Girsanov's theorem. We have included this derivation as it will serve as the basis for our normal SABR expansion.

The SABR local time L is approximated by introducing the process:

$$J_t \equiv J(S_t, \sigma_t) \equiv \frac{1}{\sigma_t} \int_K^{S_t} \frac{du}{\varphi(u)}$$

and observing that:

$$L = \sigma_0 \varphi(K) E\left[e^{\gamma B_t - \frac{1}{2}\gamma^2 t} \delta(J_t) \right]$$

By applying Itô's lemma and performing the change of measure $d\hat{Q}/dQ = e^{\gamma B_t - \frac{1}{2}\gamma^2 t}$, we derive:

$$L = \sigma_0 \varphi(K) \hat{E}\left[\delta(J_t)\right]$$
$$dJ_t = q(J_t)^{\frac{1}{2}} d\hat{U}_t - \frac{1}{2}\dot{\varphi}(S_t)\sigma_t dt$$

where $q(J) = 1 - 2\rho\gamma J + \gamma^2 J^2$ and $\hat{U}$ is a Brownian motion under $\hat{Q}$.

The SABR density D is similarly approximated by performing the change of measure $d\tilde{Q}/dQ = e^{-\gamma B_t - \frac{1}{2}\gamma^2 t}$:

$$D = \frac{E\left[\delta(J_t)/\sigma_t\right]}{\varphi(K)} = \frac{\tilde{E}\left[\delta(J_t)\right]e^{\gamma^2 t}}{\sigma_0\varphi(K)}$$

$$dJ_t = q(J_t)^{\frac{1}{2}}\,d\tilde{U}_t + \dot{q}(J_t)dt - \frac{1}{2}\dot{\varphi}(S_t)\sigma_t dt$$

We define the martingale:

$$d\bar{\rho}_t / \bar{\rho}_t = \frac{\dot{q}(J_t)}{q(J_t)^{\frac{1}{2}}}d\tilde{U}_t$$

and the measure $d\hat{Q}/d\bar{Q} = \bar{\rho}_t$. We note that J has the same dynamics under $\bar{Q}$ and $\hat{Q}$ except for the drift of σ:

$$dJ_t = q(J_t)^{\frac{1}{2}}\,d\bar{U}_t - \frac{1}{2}\dot{\varphi}(S_t)\sigma_t dt$$

Now, consider the process $X_t = q(J_0)/q(J_t)e^{\gamma^2 t}$ and observe that:

$$X_t = \bar{\rho}_t \exp\left(\frac{1}{2}\int_0^t \dot{\varphi}(S_u)\sigma_u \frac{\dot{q}(J_u)}{q(J_u)}du\right)$$

It follows that the SABR density satisfies:

$$\begin{aligned} D &= \frac{\tilde{E}\left[\frac{q(J_0)}{q(J_t)}\delta(J_t)\right]e^{\gamma^2 t}}{q(J_0)\sigma_0\varphi(K)} \\ &= \frac{\tilde{E}\left[\exp\left(\frac{1}{2}\int_0^t \dot{\varphi}(S_u)\sigma_u \frac{\dot{q}(J_u)}{q(J_u)}du\right)\middle| J_t = 0\right]}{q(J_0)\sigma_0\varphi(K)} \times \bar{E}\left[\delta(J_t)\right] \end{aligned}$$

Since the volatility σ only appears in the drift expression of J, we conclude that $\hat{E}[\delta(J_t)]/\bar{E}[\delta(J_t)] = 1 + O(t^2)$. We consequently have:

$$g(t,K)^2 = q(J_0)\sigma_0^2\varphi(K)^2 e^{\frac{\sigma_0}{2}\left(\rho\gamma\dot{\varphi}(K) - \frac{1}{2}\dot{\varphi}(S_0)\frac{\dot{q}(J_0)}{q(J_0)}\right)t} + O\left(t^2\right)$$

We finally derive the following first-order approximation in time of the SABR local volatility (see also Andreasen & Huge, 2013, and Balland, 2010):

$$g(K) = \sigma_0\varphi(K)\sqrt{1 + 2\rho\gamma f(K) + \gamma^2 f(K)^2}$$

$$f(K) = \frac{1}{\sigma_0}\int_{S_0}^{K}\frac{du}{\varphi(u)}$$

Using this equivalent local volatility, we obtain Hagan's first-order approximation for the implied volatility under SABR using standard results for local volatility:

$$IV(K;\varphi,\gamma,\rho) = \frac{\ln(K/S_0)}{\int_{S_0}^{K}\frac{du}{g(u)}} = \frac{\ln(K/S_0)}{\int_0^{f(K;\varphi)}\frac{dv}{\sqrt{1+2\rho\gamma v+\gamma^2 v^2}}}$$

The SABR local volatility behaves like a CEV dynamic near zero. The absorption at zero is ignored in the above approximation because we are using Black-Scholes as the base model for our implied volatility calculation. Hence, we can expect to improve accuracy by choosing a base model with a dynamic absorbed at zero.

Asymptotic expansion with different base models

Suppose that we can accurately integrate the following instance of the SABR dynamics:

$$dS_t = \sigma_t\varphi^{base}(S_t)dW_t$$

$$d\sigma_t / \sigma_t = \gamma dB_t$$

$$\sigma_{t=0} = b_0$$

where γ, ρ are as in SABR. By matching the first-order implied volatility approximations, that is, $IV(K; \varphi_{base}, \gamma, \rho) = IV(K; \varphi, \gamma, \rho)$, we derive the base implied volatility b_0 so that the base model and SABR share the same implied volatility to first order in time:

$$b_0 = \frac{\sigma_0(1-\beta)\int_{S_0}^{K}\frac{du}{\varphi^{base}(u)}}{K^{1-\beta} - S_0^{1-\beta}}$$

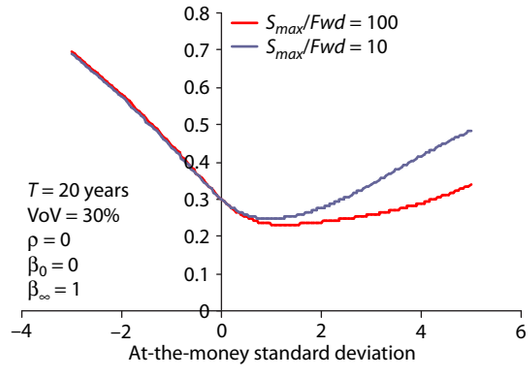
We consider two base candidates. Our first is SABR with a shifted lognormal backbone:

$$\varphi_1^{base}(S) = pS + (1-p)S_0$$

This base dynamics can be integrated by inverting a Laplace transform. Although tractable, this requires a double integration.

Our second candidate is normal SABR with absorption at zero:

1 Implied volatility for double-beta SABR



$$\varphi_2^{base}(S) = \lim_{\beta \downarrow 0} S^\beta = 1\{S > 0\}$$

We will see in the next paragraph that SABR with a normal backbone can be accurately approximated with limited calculation cost. We observe that the initial normal volatility to use when approximating SABR with this base model is as follows:

$$b_0 = \frac{\sigma_0 (1-\beta)(K - S_0)}{K^{1-\beta} - S_0^{1-\beta}}$$

In the case where we attempt to approximate SABR with an extended backbone $\varphi(S)$ instead of a CEV backbone, then our formula for b_0 is generalised as follows:

$$b_0 = \frac{\sigma_0 (K - S_0)}{\int_{S_0}^{K} \frac{du}{\varphi(u)}}$$

As observed in Andreasen & Huge (2013), the SABR dynamics calibrated to swaption smiles do not imply constant maturity swap levels consistent with the market. Various methods have been proposed to address this issue. These attempts to steepen the upper-strike wing while not affecting the liquid region and the lower wing too much. They are based on modifying either the density conditional of being in the upper-wing or directly the SABR dynamics.

We can gain control on the upper-wing steepness by assuming the following backbone:

$$\varphi(S) = S^{\beta(S)}$$
$$\beta(S) = \beta_0 + (\beta_\infty - \beta_0)\left(1 - e^{-S/S_{\max}}\right)$$

where S_{max} is typically much larger than the forward rate S_0 in order to localise the effect of double beta to the high-strike wing.

An alternative is to use the following double-beta backbone to control both lower and upper wings:

$$\varphi(S) = S^\beta \times \frac{(S/S_1)^{\beta_1} + 1}{(S/S_2)^{\beta_2} + 1}$$

This parameterisation allows us to account for the extra risk premium for high-strike volatilities and for the fact that traders typically increase β when interest rates become very low.

Approximation for normal SABR

We obtain the following formula for a call option under the normal SABR model by applying the Tanaka-Meyer formula to a call payout (see Benhamou & Croissant, 2007):

$$E\left[(S_T - K)^+\right] = (S_0 - K)^+ + \frac{1}{2}\int_0^T b_0^2 E\left[\alpha_t^2 \delta(S_t - K)\right] dt$$

where $\alpha_t = \sigma_t / b_0$ with $\sigma_0 = b_0$. We observe that:

$$E\left[\alpha_t^2 \delta(S_t - K)\right] = E\left[\alpha_t \delta(X_t)\right]$$
$$X_t = \frac{S_t - K}{\alpha_t}$$

Finally, we denote by P^α the probability measure associated with the Radon derivative $\alpha_t = \sigma_t / b_0$ and obtain the following formula for a call option under normal SABR:

$$E\left[(S_T - K)^+\right] = (S_0 - K)^+ + \frac{1}{2}\int_0^T b_0^2 E^\alpha\left[\delta(X_t)\right] dt$$

The process X satisfies:

$$dX_t = b_0 q(X_t)^{\frac{1}{2}} dW^\alpha_{t \wedge \tau}$$

where $q(X) = 1 - 2\rho\tilde{\gamma}X + \tilde{\gamma}^2 X^2$, $\tilde{\gamma} = \gamma / b_0$ and τ is the first time S hits zero.

The stopping of the diffusion is a consequence of using SABR with a vanishing CEV coefficient. As explained in Doust (2010), accounting for this stopping is important because the support of SABR is the positive half line and our base model must share with SABR the same behaviour at zero otherwise our lower-strike wing will be too steep. The importance of using interest rate models with absorbing and reflecting boundaries is discussed in Goldstein & Keirstead (1997).

Ignoring the volatility-of-volatility, we approximate τ as the first time X hits its expected barrier level under P^α, at which point $X_\tau = E^\alpha[-K/\alpha_\tau] = -K$. This approximation does not compromise the accuracy of our call price because this approximation only affects option prices with very low strikes. We can gain additional control on the lower-wing steepness by assuming that X is absorbed at the level $S_{min} - K$ where $S_{min} = (p-1)/p\, S_0$ is negative, that is, $0 < p \le 1$.

We define the following process:

$$I_t \equiv I(X_t) = \int_0^{X_t} \frac{du}{\sqrt{q(u)}} = \frac{1}{\tilde{\gamma}} \ln\left(\frac{\sqrt{q(X_t)} - \rho + \tilde{\gamma} X_t}{1 - \rho}\right)$$

In Appendix I, we derive an approximation for the density of X at zero using the reflection principle for Brownian motion:

$$E^\alpha\left[\delta(X_t)\right] = \frac{q(X_0)^{\frac{1}{4}}}{b_0 \sqrt{2\pi t}} \times \Lambda(t) \times \left[e^{-\frac{B^2}{b_0^2 t}} - e^{-\frac{C^2}{b_0^2 t}}\right]$$

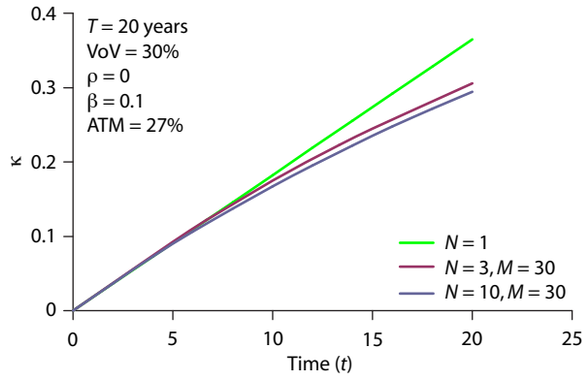
$$B = \frac{I(S_0 - K)}{\sqrt{2}}, \qquad C = \frac{2I(S_{\min} - K) - I(S_0 - K)}{\sqrt{2}}$$

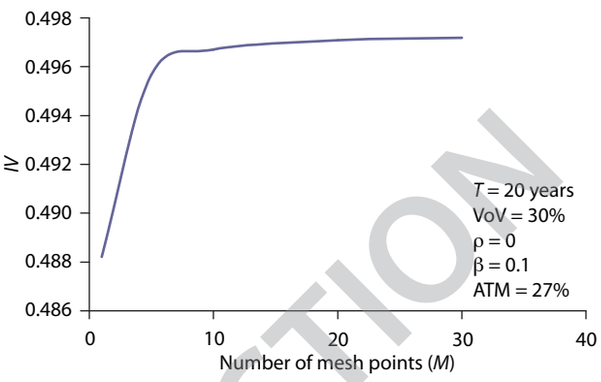
$$\Lambda(t) \approx e^{-\frac{1}{8}\gamma^2 t} \times \Phi\left(t, \frac{I_0}{b_0}\right)$$

$$\Phi(t,z) = E\left[\exp\left(\frac{3}{8}\gamma^2(1-\rho^2)\int_0^t \frac{1}{f(W_u)} du\right) \middle| W_t = z\right]$$

$$f(W) \equiv \frac{1}{4}\left((1+\rho)^2 e^{-2\gamma W} + (1-\rho)^2 e^{2\gamma W} + 2(1-\rho^2)\right)$$

2 Kappa for standard deviation = –0.5 (left); implied volatility (standard deviation = 2) (right)





Hence, we obtain the following approximation for call prices under SABR:

$$E\left[\left(S_T-K\right)^+\right]$$
$$=\left(S_0-K\right)^+ + \frac{q\left(S_0-K\right)^{\frac{1}{4}}}{2\sqrt{2\pi}} b_0 \int_0^T \frac{1}{\sqrt{t}} e^{\int_0^t \kappa\left(s,\frac{I_0}{b_0}\right)ds}\left(e^{-\frac{B^2}{b_0^2 t}} - e^{-\frac{C^2}{b_0^2 t}}\right) dt$$

$$b_0 = \frac{\sigma_0\left(K-S_0\right)}{\int_{S_0}^K \frac{du}{\varphi(u)}}$$

$$\kappa(t,z) \equiv -\frac{1}{8}\gamma^2 + \partial_t \ln \Phi(t,z)$$

The function $\kappa(t, z)$ is independent of K and only depends on the SABR parameters γ and ρ.

We have the following first-order approximation:

$$\kappa(t,z) = -\frac{1}{8}\gamma^2 + \frac{3}{16}\gamma^2\left(1-\rho^2\right)\left(\frac{1}{f(0)}+\frac{1}{f(z)}\right) + O(t)$$
$$\Phi(t,z) = \exp\left(-\frac{1}{8}\gamma^2 t + \frac{3}{16}\gamma^2\left(1-\rho^2\right)\left(\frac{1}{f(0)}+\frac{1}{f(z)}\right)t\right) + O\left(t^2\right)$$

We can estimate $\kappa(t, z)$, $\Phi(t, z)$ more accurately without any major increase in calculation time. Firstly, we pre-compute by forward induction $\Phi(T_i, \xi_j T_i^{1/2})$ on a fixed-time grid $\{T_i\}_{i<N}$ and an $N(0,1)$-mesh $\{\xi_j\}_{j<M}$ as explained in Appendix II. Finally, we approximate $\kappa(s, I_0/b_0)$ by a constant κ_i over each interval (T_{i-1}, T_i):

$$\kappa_i = -\frac{1}{8}\gamma^2 + \frac{\ln\Phi\left(T_i, \frac{I_0}{b_0}\right) - \ln\Phi\left(T_{i-1}, \frac{I_0}{b_0}\right)}{T_i - T_{i-1}}$$

where $\Phi(T_k, z)$ is obtained by cubic spline interpolation of $\{\Phi(T_k, \xi_j T_k^{1/2}) : j = 0, \ldots, M-1\}$.

Pricing formula with normal SABR as base

From our previous calculations, we derive the following approximation for the price of an option on a SABR underlying S using normal SABR as a base for our asymptotic expansion:

$$E\left[\left(S_T-K\right)^+\right]$$
$$=\left(S_0-K\right)^+ + \frac{q\left(S_0-K\right)^{\frac{1}{4}}}{\sqrt{2\pi}} b_0 \sum_{i=1}^N e^{-\left(\frac{1}{8}\gamma^2+\kappa_i\right)T_{i-1}} \Phi\left(T_{i-1}, \frac{I_0}{b_0}\right) \times J_i$$

$$J_i = \frac{1}{2}\int_{T_{i-1}}^{T_i} \frac{1}{\sqrt{t}} e^{\kappa_i t}\left(e^{-\frac{B^2}{b_0^2 t}} - e^{-\frac{C^2}{b_0^2 t}}\right) dt$$

The above integrals J_i are calculated using formula 7.4.33 in Abramowitz & Stegun (1972):

$$\frac{1}{2}\int_0^T \frac{1}{\sqrt{u}} e^{\kappa u - \frac{\lambda^2}{u}} du$$
$$= \frac{\sqrt{\pi}}{4\sqrt{-\kappa}}\left[e^{2|\lambda|\sqrt{-\kappa}}\left(erf\left(\sqrt{-\kappa T} + \frac{|\lambda|}{\sqrt{T}}\right) - 1\right)\right.$$
$$\left. + e^{-2|\lambda|\sqrt{-\kappa}}\left(erf\left(\sqrt{-\kappa T} - \frac{|\lambda|}{\sqrt{T}}\right) + 1\right)\right]$$

where:

$$erf(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2} dt = 2N\left(x\sqrt{2}\right) - 1$$

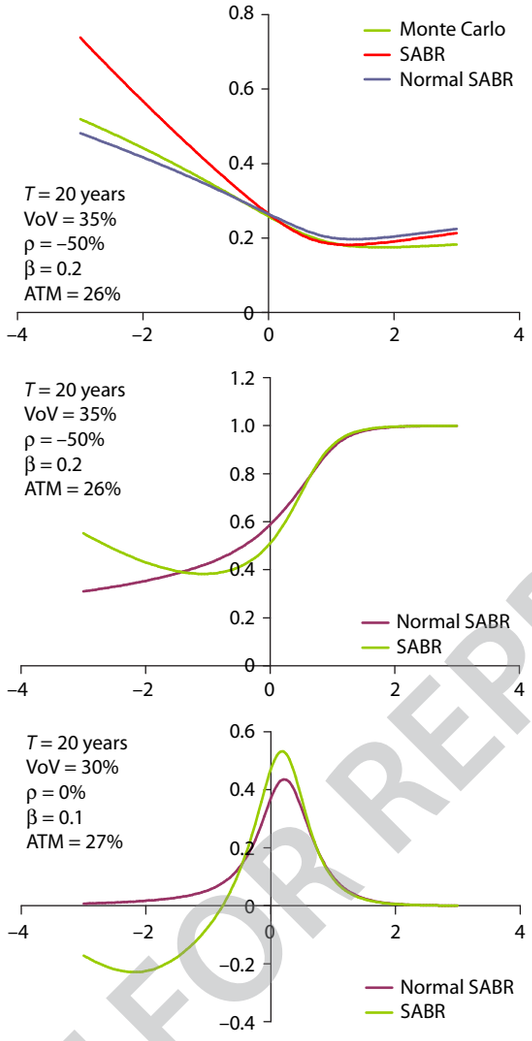
and $\sqrt{-\kappa}$ is either imaginary or real. The error function with complex argument can be estimated using the infinite series approximation of Abramowitz & Stegun (1972, see formula 7.1.29) as suggested in Benhamou & Croissant (2007):

$$erf(x+iy) = erf(x) + \frac{e^{-x^2}}{2\pi x}\left(1 - \cos 2xy + i\sin 2xy\right)$$
$$+ \frac{2}{\pi}e^{-x^2}\sum_{n=1}^{+\infty}\frac{e^{-\frac{n^2}{4}}}{n^2+4x^2}\left(f_n(x,y) + i g_n(x,y)\right)$$
$$f_n(x,y) = 2x - 2x\cosh ny\cos 2xy + n\sinh ny\sin 2xy$$
$$g_n(x,y) = 2x\cosh ny\sin 2xy + n\sinh ny\cos 2xy$$

In practical application, it is sufficient to include the first 10 terms to ensure a very good accuracy. From the above expression, we can calculate analytical expressions for the cumulative and density functions.

For moderate expiry and volatility-of-volatility, we can approximate $\kappa(t, I_0/b_0)$ using our first-order approximation, that is, $N = 1$. Otherwise, we approximate $\kappa(t, I_0/b_0)$ by a piecewise constant

3 Implied volatility SABR versus normal SABR (top); cumulative distribution (middle); density of ln*S* (bottom)



References

Abramowitz M and I Stegun, 1972
Handbook of mathematical functions
Dover Publications, New York

Andreasen J and B Huge, 2013
Expanded forward volatility
Risk January 2013, pages 101–107, available at *www.risk.net/2233952*

Antonov A and M Spector, 2011
Advanced analytics for the SABR model
SSRN paper

Balland P, 2010
Local volatility SABR
UBS Technical Note

Benhamou E and O Croissant, 2007
Local time for SABR model
SSRN paper

Doust P, 2010
No arbitrage SABR
SSRN paper

Goldstein R and W Keirstead, 1997
On the term structure of interest rates in the presence of reflecting and absorbing boundaries
SSRN paper

Hagan P, D Kumar, A Lesniewski and D Woodward, 2002
Managing smile risk
Wilmott Magazine 3, pages 84–108

function. For typical market data, we only need a limited number of grid and mesh points, that is, $N \sim 10$, $M \sim 30$, as illustrated in figure 2 representing $Kappa = \int_0^t \kappa(s, I_0/b_0)ds$ as a function of t and the implied volatility as a function of M.

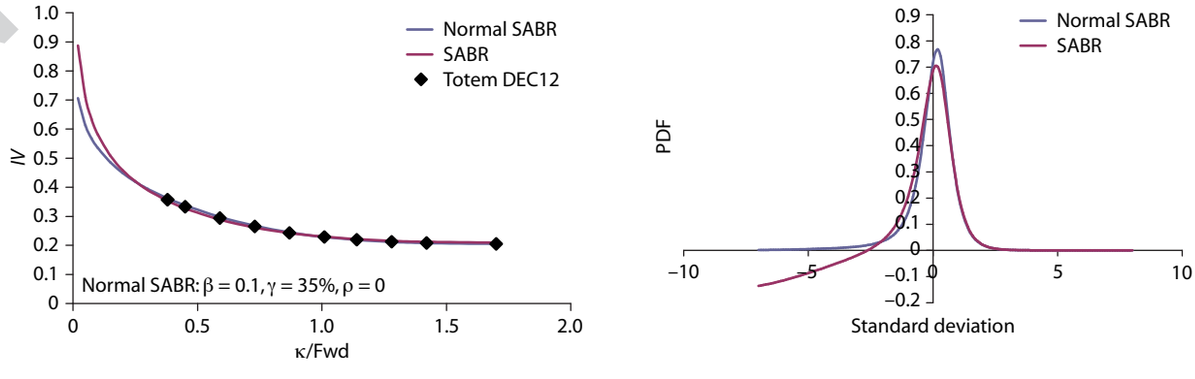
In figure 3, we show the accuracy of our implied volatility calculation comparing normal SABR and SABR using $S_{min} = 0$, $N = 10$ and $M = 30$. Our Monte Carlo results were obtained using 200 time-steps and 1*e*6 paths. The implied volatility is shown as a function of the at-the-money lognormal standard deviation.

In figure 4, we compare the smiles obtained using normal SABR approximation and Hagan with the market consensus for US dollar 10-year/10-year swaptions. Both smiles are calibrated to the same Totem data and so correspond to different SABR parameters. The calibration was performed by minimising the square of the calibration errors given a choice for the correlation.

Conclusion

We have proposed a simple approximation for call prices under the SABR dynamic based on an expansion of the normal SABR implied volatility. This approximation is exact when $\gamma = 0$ and $\beta = 0^+$ and remains accurate even with large volatility-of-volatility. It is well suited for interest rate smiles as these are typically associated with a small CEV exponent at the long end where the SABR formula breaks down. The approximation remains accurate and implies a positive density under extreme market data conditions.

4 US dollar 10-year/10-year swaption smile (left); density from US dollar 10-year/10-year swaption smile (right)



Appendix I: Density for normal SABR

We approximate the density of X at zero, that is, $E^{\alpha}[\delta(X_t)]$. As previously explained, the process X satisfies:

$$dX_t = b_0 q(X_t)^{\frac{1}{2}} dW^{\alpha}_{t\wedge\tau}$$
$$X_0 = S_0 - K$$

where $q(X) = 1 - 2\rho\tilde{\gamma}X + \tilde{\gamma}^2X^2$, $\tilde{\gamma} = \gamma/b_0$, W^{α} is a zero-drift Brownian motion under P^{α}, and τ is the first time X hits $-K$.

This is achieved by defining the following process:

$$I(X) = \int_0^X \frac{du}{\sqrt{q(u)}} = \frac{1}{\tilde{\gamma}} \ln\left(\frac{\sqrt{q(X)} - \rho + \tilde{\gamma}X}{1-\rho} \right)$$
$$q(X) = g(I) \equiv \frac{1}{4}\left((1+\rho)^2 e^{-2\tilde{\gamma}I} + (1-\rho)^2 e^{2\tilde{\gamma}I} + 2\left(1-\rho^2\right) \right)$$

The process $I_t \equiv I(X_t)$ admits the following dynamic:

$$dI_t = b_0 dW_{t\wedge\tau} - \frac{1}{2} \frac{\tilde{\gamma}^2 X_t - \rho\tilde{\gamma}}{\sqrt{1 + \tilde{\gamma}^2 X_t^2 - 2\rho\tilde{\gamma}X_t}} b_0^2 1\{t < \tau\} dt$$

We define the process $A_t = q(X_t)^{1/4}/q(X_0)^{1/4}$ and observe that:

$$d\ln A_t = d\ln \rho_t + \left(-\frac{1}{8} + \frac{3}{8}\frac{1-\rho^2}{q(X_t)} \right) \tilde{\gamma}^2 b_0^2 1\{t < \tau\} dt$$
$$d\rho_t / \rho_t = \frac{1}{2} \frac{\tilde{\gamma}^2 X_t - \rho\tilde{\gamma}}{\sqrt{1 + \tilde{\gamma}^2 X_t^2 - 2\rho\tilde{\gamma}X_t}} b_0 dW_{t\wedge\tau}$$

The martingale ρ defines a new measure Q and we have:

$$dI_t = b_0 dW^Q_{t\wedge\tau}$$

where W^Q is a Brownian motion under Q.

We observe that:

$$E^{\alpha}\left[\delta(X_t)\right] = q(X_0)^{\frac{1}{4}} E^{\alpha}\left[A_t \delta(X_t)\right] = q(X_0)^{\frac{1}{4}} \Lambda(t) E^Q\left[\delta(X_t)\right]$$
$$\Lambda(t) = E^Q\left[\exp\left(-\frac{1}{8}\tilde{\gamma}^2 b_0^2 t \wedge \tau + \frac{3}{8}\tilde{\gamma}^2 b_0^2 \left(1-\rho^2\right) \int_0^{t\wedge\tau} \frac{1}{g(I_u)} du \right) \middle| I_t = 0 \right]$$

Ignoring the stopping time in the above expression for $\Lambda(t)$ and using $\tilde{\gamma}b_0 = \gamma$, we derive:

$$\Lambda(t) \approx e^{-\frac{1}{8}\gamma^2 t} \times \Phi\left(t, \frac{I_0}{b_0}\right)$$
$$\Phi(t,z) = E^Q\left[\exp\left(\frac{3}{8}\gamma^2 \left(1-\rho^2\right) \int_0^t \frac{1}{f(W_u)} du \right) \middle| W_t = z \right]$$
$$f(W) \equiv \frac{1}{4}\left((1+\rho)^2 e^{-2\gamma W} + (1-\rho)^2 e^{2\gamma W} + 2\left(1-\rho^2\right) \right)$$

where W is a Q-Brownian motion with initial value zero.

Since $\Phi(t, z)$ depends exclusively on ρ, γ, this function can be pre-calculated or alternatively approximated as follows:

$$\Phi(t,z) = \exp\left(\frac{3}{16}\gamma^2 \left(1-\rho^2\right) \left(\frac{1}{f(0)} + \frac{1}{f(z)} \right) t \right) + O\left(t^2\right)$$

We define $\kappa(t, z) = -\frac{1}{8}\gamma^2 + \partial_t \ln \Phi(t, z)$:

$$\Lambda(t) = \exp\left(\int_0^t \kappa\left(s, \frac{I_0}{b_0}\right) ds \right)$$
$$\kappa(t,z) = -\frac{1}{8}\gamma^2 + \frac{3}{16}\gamma^2 \left(1-\rho^2\right) \left(\frac{1}{f(0)} + \frac{1}{f(z)} \right) + O(t)$$

Since $E^Q[\delta(I_t)] = E^Q[\delta(X_t)]$, we finally derive using the reflection principle for Brownian motions:

$$E^{\alpha}\left[\delta(X_t)\right] = \frac{q(X_0)^{\frac{1}{4}}}{b_0\sqrt{2\pi t}} \times e^{\int_0^t \kappa\left(s, \frac{I_0}{b_0}\right) ds} \times \left[e^{-\frac{B^2}{b_0^2 t}} - e^{-\frac{C^2}{b_0^2 t}} \right]$$
$$B = \frac{I(S_0 - K)}{\sqrt{2}}, \quad C = \frac{2I(S_{\min} - K) - I(S_0 - K)}{\sqrt{2}}$$

Appendix II: Calculation of functions Φ, κ

We propose a simple algorithm to calculate the functions Φ and κ. First, we fix a time grid $\{T_i\}_{i<N}$ and an N01-mesh $\{\xi_j\}_{j<M}$. (We can use the roots of the Hermite polynomial used in the Gauss-Hermite integration scheme.) Then, we evaluate $\Phi(T_i, \xi_j T_i^{1/2})$ by forward induction observing that:

$$\Phi(T_i, W_i) = E^Q\left[\Phi(T_{i-1}, W_{i-1}) \exp\left(\frac{\lambda}{2} \frac{\Delta T_i}{f(W_{i-1})} \right) \middle| W_i \right] \times \exp\left(\frac{\lambda}{2} \frac{\Delta T_i}{f(W_i)} \right)$$

where $\lambda = \frac{3}{8}\gamma^2(1 - \rho^2)$ and W_{i-1}, W_i have correlation $\rho_i = (T_{i-1}/T_i)^{1/2}$. We pre-calculate transition matrices $\{p_{kj}[i]\}_{i<N}$ depending only on the time grid $\{T_i\}_{i<N}$ so that we have for any natural cubic spline function F associated with nodes $\{\xi_j\}_{j<M}$:

$$E\left[F(\zeta_{i-1}) \middle| \zeta_i = \xi_k \right] = \sum_j p_{kj}[i] F(\xi_j)$$

where ζ_{i-1} and ζ_i are two normal random variables with correlation ρ_i and unit variance.

The calculation of the transition matrix is independent of the SABR parameters and can be performed analytically since the expectation $E[F(\rho_i\xi_k + (1 - \rho_i^2)^{1/2}\zeta)]$ with respect to ζ can be calculated analytically and written as a linear combination of $F(\xi_j)$.

We then derive:

$$\Phi\left(T_i, T_i^{\frac{1}{2}}\xi_k\right) = \sum_j p_{kj}[i] \Phi\left(T_{i-1}, T_{i-1}^{\frac{1}{2}}\xi_j\right) \exp\left(\frac{\lambda}{2} \left(\frac{1}{f\left(T_{i-1}^{\frac{1}{2}}\xi_j\right)} + \frac{1}{f\left(T_i^{\frac{1}{2}}\xi_k\right)} \right) \Delta T_1 \right)$$
$$\Phi\left(T_1, T_1^{\frac{1}{2}}\xi_k\right) = \exp\left(\frac{\lambda}{2} \left(\frac{1}{f(0)} + \frac{1}{f\left(T_1^{\frac{1}{2}}\xi_k\right)} \right) \Delta T_1 \right)$$

These equations allow us to construct N natural cubic spline functions $\{\Phi(T_i, z)\}_{i<N}$ depending on ρ, γ exclusively.

Finally, we approximate $\kappa(s, I_0/b_0)$ by a constant κ_i over each interval (T_{i-1}, T_i):

$$\kappa_i = -\frac{1}{8}\gamma^2 + \frac{\ln \Phi\left(T_i, \frac{I_0}{b_0}\right) - \ln \Phi\left(T_{i-1}, \frac{I_0}{b_0}\right)}{T_i - T_{i-1}}$$

The calculation of implied volatility is significantly faster when using this approximation than when solving the one-factor PDE based on the SABR local volatility. ■

Philippe Balland is global head of rates, currencies and credit analytics and Quan Tran is a rates quantitative analyst at UBS in London. Email: Philippe.Balland@ubs.com, Quan.Tran@ubs.com