Introducing the Implied Volatility Surface Parametrization (IVP): Application to the FX Market

Babak Mahdavi Damghani
e-mail: bmd@cantab.net

Abstract
The aim of this article[1] is to introduce a new parametrization of the implied volatility surface (IVP), which builds on the gSVI methodology recently introduced [4] but incorporates novel features like a bid–ask model and the methodology behind de-arbitraging a volatility surface and stressing it without re-adding arbitrages within the scope of the FX market – where the relationship between currencies is constrained by the triangle rule as well as the usual calendar and butterfly arbitrages.

Keywords
IVP, SVI, gSVI, SABR, arbitrage-free volatility surface, positive semi-definite implied correlation matrices, FX, Dupire local volatility, constraint optimization, butterfly spread, calendar spread.

1 Introduction

1.1 Scope

Whether in investment banks, hedge funds or clearing houses, risk managing at the portfolio level has become an active area of research for practitioners in quantitative analytics. Within the framework of options risk modeling, it is essential to define a volatility surface that works with a variety of pricing models. As it happens, most pricing systems used in practice are designed in such a way that they cannot accommodate volatility surfaces that would allow for arbitrage opportunities. In clearing, one needs to make sure that the scenarios used in the IM calculations are coherent, and hence having a stressed volatility surface with arbitrages on it violates this constraint. In order to address this issue we need to create a methodology that would first, test whether a volatility surface is arbitrage free and second, adjust the volatility surfaces that would prevent the presence of arbitrages. These steps have recently been introduced in [4], and the methodology has been proposed to work in the equities, commodities, and FX markets. Although the gSVI parametrization introduced in the same paper [4] happens to model the volatility surface geometrically in these specific markets, its de-arbitraging methodology is incomplete if applied to the FX market, specifically because of the "triangle rule" only relevant to the FX market. The objective of this article is to complete the de-arbitraging methodology suggested in [4], so that it works with the constraints induced by the triangle rule, as well as introducing an original liquidity model which enhances the parametrization as well as relaxing the de-arbitraging methodology a little.

[1]An alternative name for people already knowledgeable about parametrization could have been the Bid–Ask Wing-Adjusted De-arbed gSVI surface (BAWADgSVI). The renaming was implemented so as to avoid confusion with the generalized SVI model chosen by other authors subsequent to the publication of the gSVI presented in [4], and which may create confusion in the future over which model is being discussed.

1.2 Structure of the article

In Section 1.3 we discuss what people do in FX high-frequency trading (HFT) and how this impacts the measurement of risk, with our objectives in mind. In Section 2 we explore the conditions for an arbitrage-free volatility surface in the equities and commodities markets. In Section 3 we summarize the two key parametrizations that have led to the IVP model; that is, SVI and gSVI. Finally, in Section 4 we adjust the de-arbitraging methodology presented recently [4] to abide by the triangle and the implied correlation rules, in order to show an application of the model.

1.3 Understanding algorithmic HFT and the FX market

Let's call $S_{t,1}$ the exchange rate of the EUR/USD pair and σ_1 its implied volatility, $S_{t,2}$ the exchange rate of the USD/JPY pair and σ_2 its implied volatility, $S_{t,3}$ the exchange rate of the EUR/JPY pair and σ_3 its implied volatility. Now notice that $S_{t,3}$ must equal $S_{t,1} \times S_{t,2}$, else the arbitrage opportunity induced would be immediately taken advantage of by HFT. Therefore, $\ln(S_{t,3}) = \ln(S_{t,1} \times S_{t,2})$. Taking the variance on each side, we get the non-arbitrage condition on the volatility and the implied correlation given by equation (1):

$$\sigma_3^2 = \sigma_2^2 + \sigma_1^2 + 2\rho_{1,2}\sigma_1\sigma_2 \tag{1}$$

By rearranging, the implied correlation can be isolated and given by equation (2):

$$\rho_{1,2} = \frac{\sigma_3^2 - \sigma_2^2 - \sigma_1^2}{2\sigma_1\sigma_2} = \cos\phi_{1,2} \tag{2}$$

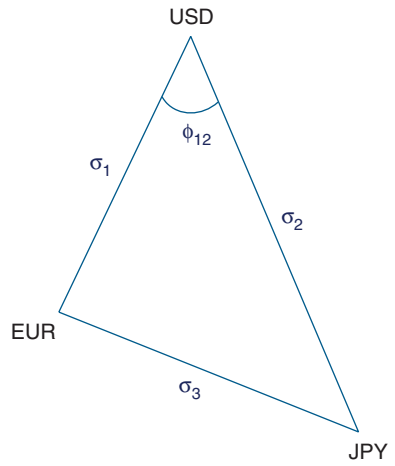
Figure 1 shows a visual representation of this non-arbitrage constraint, with $\phi_{1,2}$ representing the angle between σ_1 and σ_2.

The relationship between $\rho_{1,2}$ and $\phi_{1,2}$ is given by equation (3):

$$\arccos\rho_{1,2} = \phi_{1,2} \tag{3}$$

From a risk perspective – that is, on a 24-hr time scale – no matter what happens, this equality must hold true for any three pairs of FX available. It is not

Figure 1: FX triangle.



difficult to realize that in a fast-moving liquid market like FX, given the number of possible currency pairs, this equality may temporarily be violated. But these arbitrage situations are quickly absorbed by algorithmic traders. Given that the time scale of risk management is slower than HFT, we assume this equality always holds true. Note that we also assume equality (4) always holds true. This particular equality essentially says that in a triangle, the longest side must always be smaller than the sum of the remaining two sides. Although this is obvious intuitively, we will see in Section 4 that this constraint may be violated in the bumping phase and therefore needs to be enforced in the optimization by the constraint phase of the algorithm.

$$\sigma_1 + \sigma_2 + \sigma_3 > 2\max(\sigma_1 + \sigma_2 + \sigma_3) \tag{4}$$

2 Volatility arbitrage in the equities and commodities markets

The model setup is the usual. We have the probability space $(\Omega, (f)_{(t\geq 0)}, \mathbb{Q})$, with $(f)_{(t\geq 0)}$ generated by $(T+1)$-dimensional Brownian motion and $\mathbb{Q}$ the risk-neutral probability measure under which the discounted price of the underlier, rS, is a martingale. We also assume that the underlier can be represented as a stochastic volatility lognormal Brownian motion, as shown in equation (5):

$$dS_t = rS_t dt + \sigma_t S_t dW_t \tag{5}$$

In order to prevent arbitrages on the volatility surface, we start from basic principles and derive the constraints relevant to the strike and tenor.

2.1 Condition on the strike

2.1.1 Theoretical form

Using Dupire's work [5, 6], we can write the price of a call in the following way:

$$C(S_0, K, T) = e^{-rT}\mathbb{E}^{\mathbb{Q}}[S_T - K]^+ = e^{-rT}\int_K^{+\infty}(S_T - K)\phi(S_T, T)dS_T \tag{6}$$

with $\phi(S_T, T)$ being the final probability density of the call. Differentiating twice, we get equation (7):

$$\frac{\partial^2 C}{\partial K^2} = \phi(S_T, T) > 0 \tag{7}$$

Proof

$$C(S_0, K, T) = e^{-rT}\mathbb{E}^{\mathbb{Q}}[S_T - K]^+ = e^{-rT}\int_K^{+\infty}(S_T - K)\phi(S_T, T)dS_T$$

$$\frac{\partial C}{\partial K} = -e^{-rT}\int_K^{+\infty}\phi(S_T, T)dS_T$$

$$= -e^{-rT}\mathbb{E}(S_T > K)$$

We also know that $0 \leq -e^{-rT}\frac{\partial C}{\partial K} \leq 1$. Differentiating a second time and setting $r = 0$, we find $\phi(S_T, T) = \frac{\partial^2 C}{\partial K^2}$. □

Using numerical approximation we get equation (8), which is known in the industry as the arbitrage constraint of the positivity of the butterfly spread [19]:

$$\forall\Delta,\ C(K-\Delta) - 2C(K) + C(K+\Delta) > 0 \tag{8}$$

Proof. Given that the probability density must be positive, we have $\frac{\partial^2 C}{\partial K^2} \geq 0$. Using numerical approximation we get

$$\frac{\partial^2 C}{\partial K^2} = \lim_{\Delta\to 0}\frac{[C(K-\Delta) - C(K)] - [C(K) - C(K+\Delta)]}{\Delta^2}$$

$$= \lim_{\Delta\to 0}\frac{C(K-\Delta) - 2C(K) + C(K+\Delta)}{\Delta^2}$$

Therefore, $C(K-\Delta) - 2C(K) + C(K+\Delta) \geq 0$. □

Gatheral and Jacquier [11] proved that the positivity of the butterfly condition comes back to making sure that the function g(·) in equation (9) is strictly positive:

$$g(k) := \left(1 - \frac{Kw'(k)}{2w(k)}\right)^2 - \frac{w'(k)^2}{4}\left(\frac{1}{w(k)} + \frac{1}{4} + \frac{w''(k)}{2}\right) \tag{9}$$

Proof. We have shown in equation (7) that $\frac{\partial^2 C}{\partial K^2} = \varphi(\cdot)$. Applying this formula to the Black–Scholes equation leads to, for a given tenor:

$$\phi(k) = \frac{g(k)}{\sqrt{2\pi w(k)}}\exp\left(-\frac{d_2(k)^2}{2}\right) \tag{10}$$

where $w(k,t) = \sigma_{BS}^2(k,t)t$ is the implied volatility at strike K and $d_2(k) := \frac{-k}{\sqrt{w(k)}} - \sqrt{w(k)}$.

Function (9) yields a polynomial of second degree with negative highest order, which suggests that the function is inverse bell curve-like and potentially only positive given **two constraints** that may appear to contradict some of the initial slides Gatheral presented back in 2004. If g_1^e and g_2^e happen to be the exact roots of $g(k) = 0$, with $g_2^e \geq g_1^e$, then the volatility surface is arbitrage free with respect to the butterfly constraint if $w(k) \leq g_2^e$ and $w(k) \geq g_1^e$.

2.1.2 Practical form

There exists another version of this butterfly (equation (7)) condition that is a necessary but **not sufficient** condition to make a volatility surface arbitrage free, but remains useful when one has a more practical objective, which will be illustrated with an example in Section 3.3. This condition is given by equation (11):

$$\forall K, \forall T, |T\partial_K\sigma^2(K,T)| \leq 4 \tag{11}$$

Proof. The intuition behind the proof is taken from Rogers and Tehranchi [17], but is somewhat simplified for practitioners. Assuming $r = 0$, let us define the Black–Scholes call function $f : \mathbb{R}\times[0,\infty) \to [0,1)$ in terms of the tail of the standard Gaussian distribution $\Phi(x) = \frac{1}{\sqrt{2\Pi}}\int_x^{+\infty}\exp(\frac{-y^2}{2})dy$, given by

$$f(k,v)=\begin{cases}\Phi\left(\frac{k}{\sqrt{v}}-\frac{\sqrt{v}}{2}\right)-e^{k}\Phi\left(\frac{k}{\sqrt{v}}+\frac{\sqrt{v}}{2}\right) & \text{if} \quad v>0\\ (1+e^{k})^{+} & \text{if} \quad v=0\end{cases}$$

Let us call $V_t(k,\tau)$ the implied variance at time $t \geq 0$ for log-moneyness k and time to maturity $\tau \geq 0$. Let's now label our kappa and vega, with the convention that $\phi(x) = \frac{1}{\sqrt{2\Pi}}\exp(\frac{-x^2}{2})$:

$$f_k(k,v)=-e^{k}\Phi\left(\frac{k}{\sqrt{v}}+\frac{\sqrt{v}}{2}\right)$$

$$f_v(k,v)=\phi\left(\frac{k}{\sqrt{v}}+\frac{\sqrt{v}}{2}\right)/2\sqrt{v}$$

Now define the function $I:\{(k,c)\in\mathbb{R}\times[0,\infty):(1+e^{k})^{+}\leq c<1\}\rightarrow[0,1)$ implicitly by the formula

$$f(k,I(k,c))=c$$

Calculus gives $I_c=\frac{1}{f_v}$ and $I_k=-\frac{f_k}{f_v}$. Using the chain rule, and designating $\partial_{k+}V$ as the right derivative, we have

$$\partial_{k+}V=I_k+I_c\partial_k\mathbb{E}[(S_\tau-e^{k})^{+}]$$

$$\partial_{k+}V=-\frac{f_k}{f_v}-\frac{\mathbb{p}(S_\tau>e^{k})}{f_v}$$

$$<-\frac{f_k}{f_v}=2\sqrt{v}\frac{\Phi\left(\frac{k}{\sqrt{v}}+\frac{\sqrt{v}}{2}\right)}{\phi\left(\frac{k}{\sqrt{v}}+\frac{\sqrt{v}}{2}\right)}$$

Now, using the bounds of the Mills ratio $0\leq 1-\frac{x\Phi(x)}{\phi(x)}\equiv\epsilon(x)\leq\frac{1}{1+x^2}$, we have

$$\partial_{k+}V\leq\frac{4}{k/V+1}<4$$

Similarly we can show [17] that $\partial_{k-}V>-4$, therefore we have $|\partial_k V|<4$. □

One can think of the boundaries of the volatility surface, extrapolated by equation (11), as more relaxed boundaries (but still "close") in the strike space compared with the exact solution from equation (9) set to 0. These are both necessary and sufficient conditions for the volatility surface to be arbitrage free for the butterfly condition. Formally, if g_1^a and g_2^a happen to be the exact roots of $|T\partial_K\sigma^2(K,T)|-4=0$, with $g_2^a\geq g_1^a$, then we have $g_1^a\leq g_1^e\leq w(k)\leq g_2^e\leq g_2^a$. The reason why equation (11) is practical is because in de-arbitraging methodologies (as we will see in more detail in Section 3.3), there exists for the pricers a component of tolerance anyway (the pricers are stable if the volatility surface is slightly away from its arbitrage frontier). This suggests that finding a close-enough solution and building on top of that an iterative methodology to get closer and closer to the practical arbitrage frontier is almost equally fast, but with less computing trouble, than having the exact theoretical solution (and building an error tolerance finder on top). This is because there is less probability of making mistakes in typing the exact solution of equation (9) (or its numerical approximation), especially if your parametrized version of the volatility surface is complex, which is the case in most banks ($\{g_1^a, g_2^a\}$ is easier to find than $\{g_1^e, g_2^e\}$). Also, as we will see in Section 3.4.1, given that we would like a liquidity component around a mid price, having a simple "close-enough" constraint on the mid becomes very useful, especially if we are happy to allow the mid to have arbitrages on it, something which happens to be the case from time to time on the mid vol. in the markets anyway.

2.2 Condition on the tenor

2.2.1 Theoretical form

The condition on the tenor axis which insures the volatility surface will be arbitrage free is that the calendar spread should be positive:

$$C(K,T+\Delta)-C(Ke^{-r\Delta},T)\geq 0 \tag{12}$$

Proof. One application of Dupire's formula [5, 6] is that the pseudo-probability density must satisfy the Fokker–Planck equation [7, 16]. This proof is taken from El Karoui [14]. Let us apply Itô to the semi-martingale. This is done formally by introducing the local time Λ_T^K:

$$e^{-r(T+\epsilon)}\left(S_{T+\epsilon}-K\right)^{+}-e^{-r(T)}\left(S_T-K\right)^{+}$$
$$=\int_T^{T+\epsilon}re^{-ru}\left(S_u-K\right)^{+}du+\int_T^{T+\epsilon}e^{-ru}1_{\{S_u\geq K\}}dS_u$$
$$+\frac{1}{2}\int_T^{T+\epsilon}e^{-ru}d\Lambda_u^K$$

Local times are introduced in mathematics when the integrand is not smooth enough. Here the call price is not smooth enough around the strike level at expiry. Now we have $E\left(e^{-ru}1_{\{S_u\geq K\}}S_u\right)=C(u,K)+Ke^{-ru}P\left(S_u\geq K\right)=C(u,K)-K\frac{\partial C}{\partial K}(u,K)$. The term of the form $E\left(\int_T^{T+\epsilon}e^{-ru}d\Lambda_u^K\right)$ is found due to the formula for local times:

$$E\left(\int_T^{T+\epsilon}e^{-ru}d\Lambda_u^K\right)=\int_T^{T+\epsilon}e^{-ru}duE\left(\Lambda_u^K\right)=\int_T^{T+\epsilon}e^{-ru}du\sigma^2(u,K)K^2\phi(u,K)$$
$$=\int_T^{T+\epsilon}\sigma^2(u,K)K^2\frac{\partial^2C}{\partial K^2}(u,K)du$$

Plugging these results back into the first equation, we get

$$C(T+\epsilon,K)=C(T,K)-\int_T^{T+\epsilon}rC(u,K)du$$
$$+\left(r-q\right)\int_T^{T+\epsilon}\left(C(u,K)-K\frac{\partial C}{\partial K}(u,K)\right)du$$
$$+\frac{1}{2}\int_T^{T+\epsilon}\sigma^2(u,K)K^2\frac{\partial^2C}{\partial K^2}(u,K)du$$

If we want to give a PDE point of view of this problem, we can notice that $\phi(T,K)=e^{-rT}\frac{\partial^2C}{\partial K^2}(T,K)$ verifies the dual forward equation:

$$\phi_T'(T,K)=\frac{1}{2}\frac{\partial^2\left(\sigma^2(T,K)K^2\phi(T,K)\right)}{\partial K^2}-\frac{\partial^2\left(\left(r-q\right)K\phi(T,K)\right)}{\partial K}$$

Integrating twice by parts, we find

$$\frac{\partial e^{-rT}C(T,K)}{\partial T}=\frac{1}{2}\sigma^2(T,K)K^2e^{rT}\frac{\partial^2C(T,K)}{\partial K^2}$$
$$-\int_K^{+\infty}\left(r-q\right)Ke^{rT}\frac{\partial^2C(u,K)}{\partial K^2}\partial K(T,K)du$$

Now, integrating by parts again and setting dividends to 0, we find the generally admitted relationship

$$\frac{\partial C}{\partial t}=\frac{\sigma^2}{2}K^2\frac{\partial^2C}{\partial K^2}-rK\frac{\partial C}{\partial K}$$

and therefore we have

$$\sigma = \sqrt{\frac{2\frac{\partial C}{\partial t} + rK\frac{\partial C}{\partial K}}{K^2\frac{\partial^2 C}{\partial K}}}$$

From this formula and from the positivity constraint on equation (7), we find that

$$\frac{\partial C}{\partial t} + rK\frac{\partial C}{\partial K} \geq 0$$

Note that for very small Δ:

$$C(Ke^{-r\Delta}, T) \approx C(K - Kr\Delta, T)$$

Using the Taylor expansion:

$$C(K - Kr\Delta, T) = C(K, T) - Kr\Delta\frac{\partial C}{\partial K} + \ldots$$

Therefore

$$rK\frac{\partial C}{\partial K} \approx \frac{C(K, T) - C(Ke^{-r\Delta}, T)}{\Delta}$$

Using a forward difference approximation we also have

$$\frac{\partial C}{\partial K} = \frac{C(K, T + \Delta) - C(K, T)}{\Delta}$$

and from Fokker–Planck we have $\frac{\partial C}{\partial t} + rK\frac{\partial C}{\partial K} \geq 0$. Substituting, we obtain

$$\frac{C(K, T + \Delta) - C(K, T)}{\Delta} + \frac{C(K, T) - C(Ke^{-r\Delta}, T)}{\Delta} \geq 0$$

Simplifying, we find $C(K,T + \Delta) - C(Ke^{-r\Delta},T)$. □

2.2.2 Practical form

Similar to Section 2.1, there exists a more practical equivalent to the calendar spread criterion. This equivalent criterion is known as the falling variance criterion and states that

if S is a martingale under the risk neutral probability measure $\mathbb{Q}$,

$$\forall t > s, e^{-rt}\mathbb{E}^{\mathbb{Q}}(S_t - K)^+ \geq e^{-st}\mathbb{E}^{\mathbb{Q}}(S_s - K)^+ \tag{13}$$

Proof. $e^{-rt}\mathbb{E}^{\mathbb{Q}}(S_t - K)^+ \geq e^{-rs}\mathbb{E}^{\mathbb{Q}}(S_s - K)^+ \Rightarrow e^{-rt}\mathbb{E}^{\mathbb{Q}}(S_t - K)^+ - e^{-rs}\mathbb{E}^{\mathbb{Q}}(S_s - K)^+ \geq 0 \Rightarrow$ calendar spread $\geq 0 \Rightarrow C(K, T + \Delta) - C(Ke^{-r\Delta}, T) \geq 0$. □

2.2.3 Garman–Kohlhagen model

Another adjustment in the FX market necessary to compare with the gSVI model introduced in [4] is the use of the Garman–Kohlhagen model [8] instead of the Black–Scholes model to account for the presence of two interest rates relevant to pricing: r_d, the domestic risk-free simple interest rate and r_f, the foreign risk-free simple interest rate. Call and put pricing formulas adjusted to the FX market are summarized in equation (14), with the usual Black–Scholes naming conventions:

$$\begin{aligned} C &= S_0e^{-r_f T}(d_1) - Ke^{-r_d T}(d_2) \\ P &= Ke^{-r_d T}(-d_2) - S_0e^{-r_f T}(-d_1) \\ d_1 &= \frac{\ln(S_0/K) + (r_d - r_f + \sigma^2/2)T}{\sigma\sqrt{T}} \\ d_2 &= d_1 - \sigma\sqrt{T} \end{aligned} \tag{14}$$

3 Parametrizing the volatility surface via the IVP model

3.1 The stochastic volatility inspired (SVI) model

Like the SABR and Schonbucher models [12, 18], the advantage of the SVI is that it can be derived from Heston [10, 13], a model used by many financial institutions, and can therefore be taken to be legitimate. It is simple, yet came with linear wings (which yield a poor fit in the wings), no bid–ask liquidity model, and finally no non-arbitrage constraints. Further, its parameters are not as intuitive as they could be for traders. These are the main contributions of the SVI, developed by Gatheral [9] in 2005, and for which non-arbitrage-free constraints were clarified in 2012 [11]. For each time to expiry, he writes

$$\sigma_{BS}^2(k) = a + b[\rho(k - m) + \sqrt{(k - m)^2 + \sigma^2}] \tag{15}$$

- k is the log-moneyness ($\log\left(\frac{K}{F}\right)$, with S being the value of the forward);
- a adjusts the vertical displacement of the smile;
- b adjusts the angle between the left and right asymptotes;
- σ adjusts the smoothness of the vertex;
- ρ adjusts the orientation of the graph;
- m is the horizontal displacement of the smile.

The advantage of Gatheral's model was that it was a parametric model that was easy to use, yet had enough complexity to properly model the volatility surface and its dynamic[2] (or at least to the same extent that Schonbucher's model did). Note that Schonbucher's market model has one parameter less than the SVI: the m parameter, whose aim is to center the volatility surface around its minimum strike per tenor. Other than this, the two models are equivalent. At the same time, Gatheral's model was simple enough that a solution could be found using simple optimization by constraint algorithms. Figure 2 illustrates the change in the a parameter (general volatility level risk), Figure 3 illustrates the change in the b parameter (vol of vol risk), Figure 4 illustrates the change in the ρ parameter (skew risk), Figure 5 illustrates the change in the m parameter (horizontal displacement risk), and Figure 6 illustrates the change in the σ parameter (ATM volatility risk).

3.2 The SVI's constraints

The SVI has **three** necessary and sufficient conditions which make it arbitrage free. On top of these three constraints, the SVI has three other constraints that do not reduce the state space but decrease the probability of falling into a local minimum during the optimization process. We have seen in equation (11) the general condition that makes a volatility surface **"often"** arbitrage free along the strike axis. This condition translates into equation (16) for the SVI model:

$$b(1 + |\rho|) \leq \frac{4}{T} \tag{16}$$

Note we have mentioned that the volatility surface is often arbitrage free but not always. This is because equation (11) is a necessary but not sufficient condition for your volatility surface to be arbitrage free with respect to the butterfly condition. A counterexample was given by Axel Vogt on wilmott.com. He asserts that with the following SVI parameters: $(a, b, m, \rho, \sigma) = (-0.0410, 0.1331, 0.3586, 0.3060, 0.4153)$, one satisfies equation (16) yet violates the butterfly constraint. This counterexample has, in the past, put a negative hit on a useful constraint derived out of a necessary but not

[2] We will see its main limitation when we explore the gSVI.

Figure 2: Impact of a change in the *a* parameter in the SVI/gSVI/IVP model.

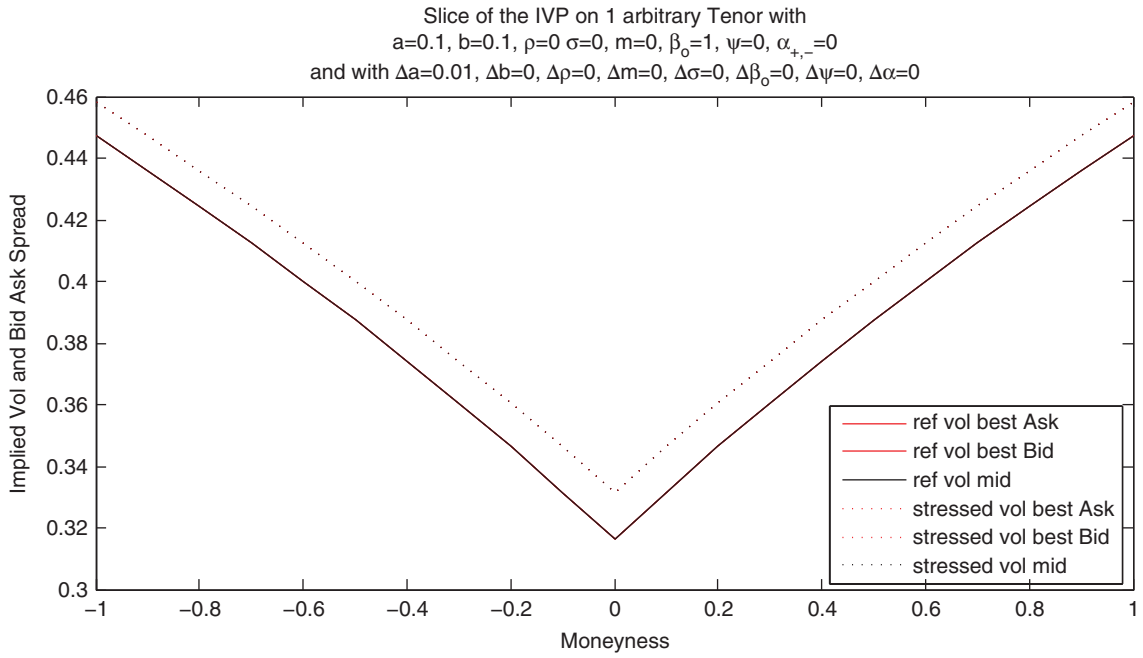
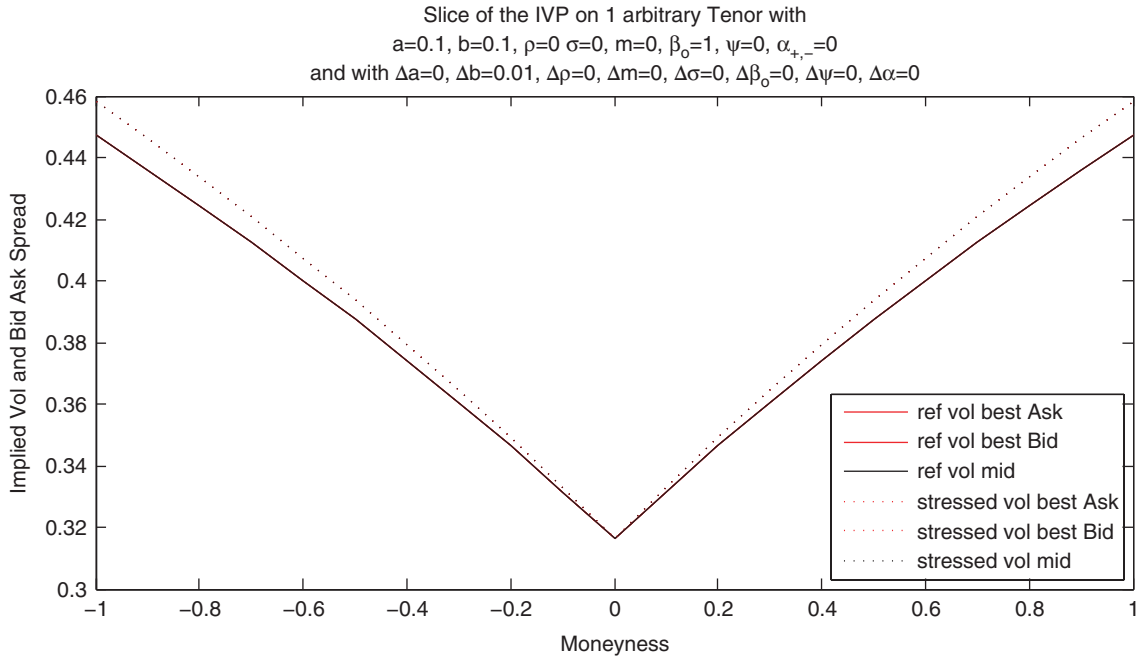


Figure 3: Impact of a change in the *b* parameter in the SVI/gSVI/IVP model.



sufficient condition [17] in a promising model [9]. We will see in Section 3.4 that this constraint becomes relevant again in the IVP model, because it is much more useful when associated with a bid–ask spread.

3.3 The generalized stochastic volatility inspired (gSVI) model

Gatheral developed the SVI model at Merrill Lynch in 1999 and implemented it in 2005. The SVI was subsequently decommissioned in 2010 because of its limitations in accurately pricing out-of-the-money (OTM) variance swaps (for example, short maturity Var swaps on the Eurostoxx are overpriced when using the SVI). This is because the wings of the SVI are linear and have a tendency to overestimate the OTM variance swaps. Benaim, Friz, and Lee [1] gave a mathematical justification for this market observation. Their paper suggests that the implied volatility cannot grow asymptotically faster than $\sqrt{k}$ but may grow slower than $\sqrt{k}$ when the distribution of the underlier does not have finite moments (e.g., has heavy tails). This suggest that the linear wings of the SVI model may overvalue really deeply OTM options, which

Figure 4: Impact of a change in the ρ parameter in the SVI/gSVI/IVP model.

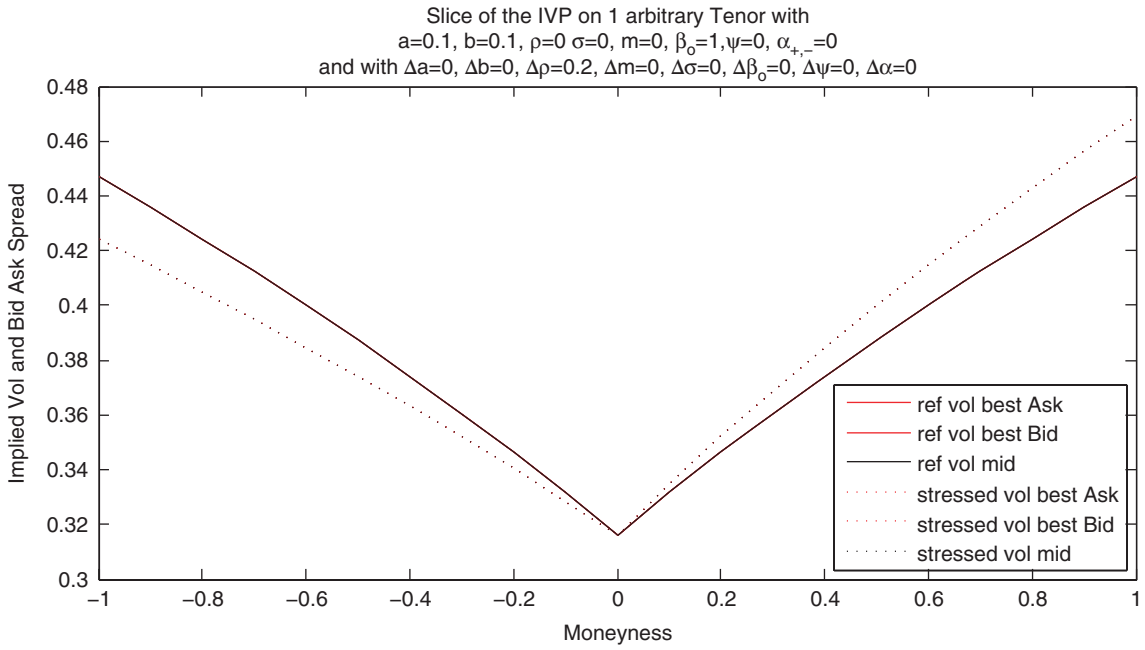
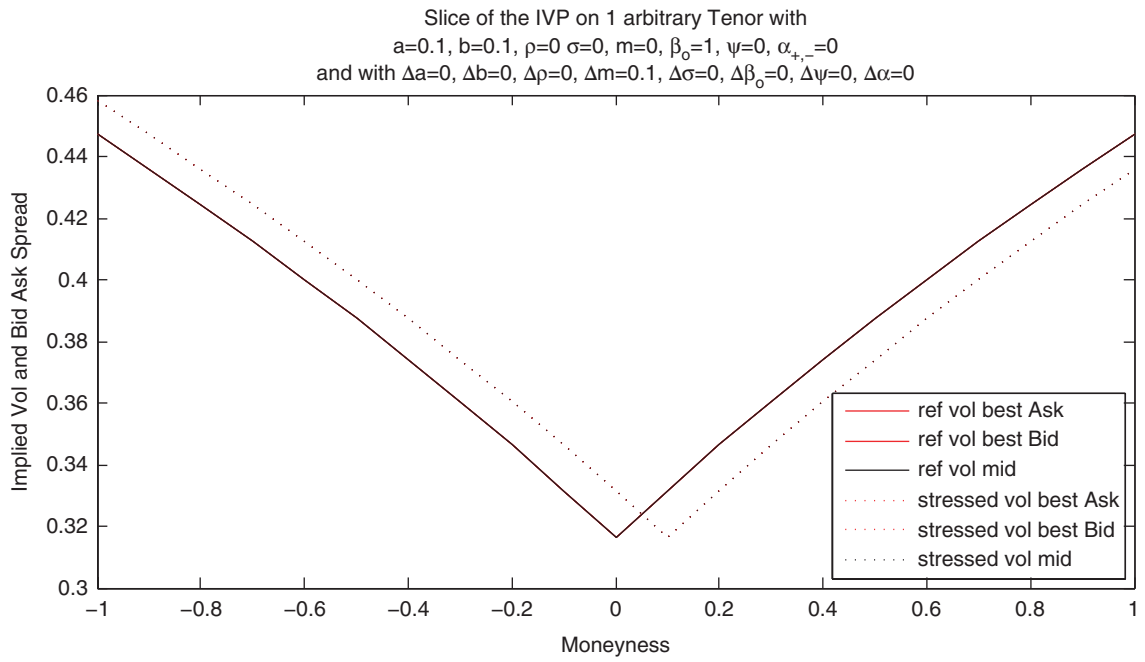


Figure 5: Impact of a change in the *m* parameter in the SVI/gSVI/IVP model.



is observable in the markets. In order to address the limitations of the SVI model in the wings, we propose a penalization of the wings function. The additional relevant parameter will be called β and aims to address this specific issue. The penalization will be symmetrical in the FX market, more significant on the left wing of the equities market, and more significant on the right wing of the commodities market (in general, e.g., excluding oil) due to the smile, skew, and inverse skew features observable on these different markets. The function needs to be increasing as it gets further away from m, majored by a linear function increasing in $[m;+\infty[$ and decreasing in $]-\infty;m]$, and increasing in concavity the further away it gets from the center. The real modeling contribution of the gSVI with respect to the SVI is this penalization change of variable and its corresponding constraint adjustments. Equation (17) summarizes the gSVI model. The penalization will be given by $z = \frac{k-m}{\beta^{|k-m|}}$, which is a strictly increasing function between log-moneyness 0 and 3 when $1 \leq \beta \leq 1.4$, similarly decreasing between –3 and 0. There are two main reasons why we have chosen the gSVI model. First, the more parameters a model has, the more flexibility it allows for in reproducing subtleties on the volatility surface. However, the more parameters

Figure 6: Impact of a change in the σ parameter in the SVI/gSVI/IVP model.

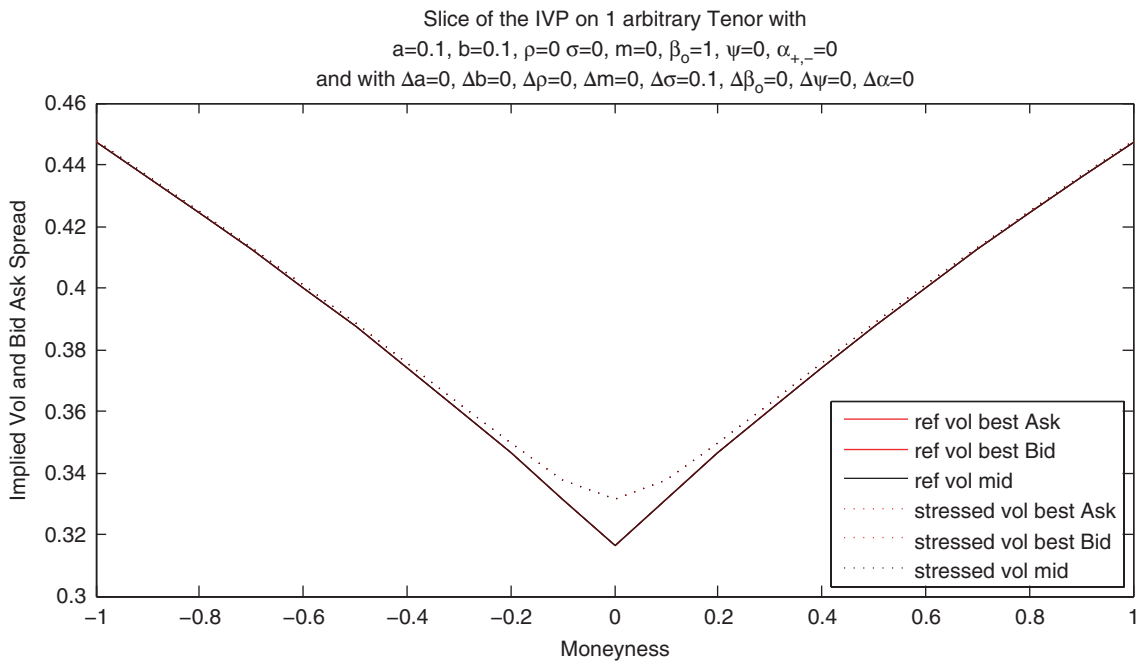
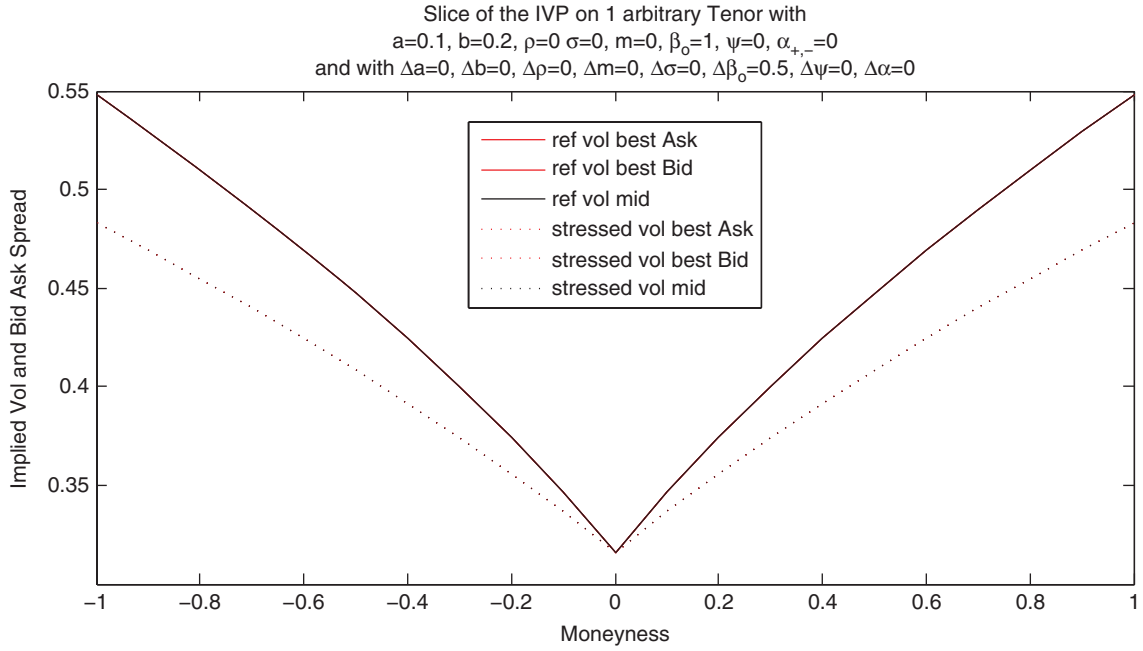


Figure 7: Impact of a change in the β parameter in the gSVI/IVP model.



a model has, the harder it is to calibrate it as the risk of falling into local minima increases. This means that the question of model selection is an optimization problem on its own. We believe that the gSVI has enough parameters to accurately model the volatility surface without the risk of falling into the traps of basic search algorithms. Also, the geometrical properties of the gSVI make it especially attractive when it comes to finding seed parameters for the optimization-by-constraints algorithm. We have already seen the changes in the a, b, ρ, m, and σ parameters in Figures 2–6, respectively. Figure 7 illustrates the change in the β parameter. The geometric properties of the gSVI, more specifically its ability to model the smile, skew, and inverse skew, while at the same time correcting the linear wings of the SVI, makes it applicable to the FX, commodities, and equities markets.

$$\sigma^2_{gSVI}(k) = a + b\left[\rho(z-m) + \sqrt{(z-m)^2+\sigma^2}\right]$$

$$z = \frac{k-m}{\beta^{|k-m|}}, 1 \le \beta \le 1.4 \qquad (17)$$

There exist two constraints that make the gSVI "often" (as explained in Section 3.2) arbitrage free. The first condition on the falling variance (equation (13)) does not change. However, we need to adjust for equation (16), which is replaced in the gSVI by equation (18):

$$\left| T\frac{1+|k-m|\ln\beta}{\beta^{|k-m|}}\left(b\rho+\frac{(\frac{k-m}{\beta^{|k-m|}}-m)}{\sqrt{(\frac{k-m}{\beta^{|k-m|}}-m)^2+\sigma^2}}\right)\right| \leq 4 \tag{18}$$

3.4 The implied volatility surface parametrization (IVP)

3.4.1 Expanding the idea of the wing adjustment to modeling bid–ask spread

The downside transform in the gSVI was given by $z=\frac{k-m}{\beta^{|k-m|}}$, $1\leq\beta\leq 1.4$. There are many ways of defining downside transforms. One general approach would be to define μ and η such that equation (19) defines the change of variable from strike space to modified strike space. The idea is that the further away you are from the at-the-money (ATM) position, the bigger the necessary adjustment on the wings.

$$z=\frac{k-m}{\beta^{\mu+\eta|k-m|}} \tag{19}$$

We can, for example, choose $\mu=1$ and $\eta=4$ and have the transformation in the form $z=\frac{k-m}{\beta^{1+4|k-m|}}$ because it yields better optimization results on the FX markets and also because it relaxes the constraint on β, since we incorporate a bid–ask layer. However, for the sake of making things simple we can use a linear change of variable in variance:

$$z_{\pm}=(1\pm\psi)\times z \tag{20}$$

3.4.2 Modeling the bid–ask wings curvature

One contribution of the gSVI [4] compared with the SVI [9] is the adjustment of the wings using a change of variable or downside transform (see equation (17)). Let's call the beta parameter whose aim is to adjust the wings of the mid, $\beta_{o,\tau}$. Someone wanting to sell an option would want to sell it at a higher price than the mid, so the dampening effect of the bid ($\beta_{+,\tau}$) should be smaller than that of the mid and therefore $\beta_{+,\tau}>\beta_{o,\tau}$. Using the same logic, the dampening of the ask price should be $\beta_{o,\tau}>\beta_{-,\tau}$. The constraints on the β's are given by equation (21):

$$\beta_{+,\tau}>\beta_{o,\tau}>\beta_{-,\tau} \tag{21}$$

In order to control the addition of new parameters, we set ψ as mentioned in equation (22) to account for the symmetry of this bid–ask adjustment. Figure 8 illustrates how the variable ψadjusts the bid–ask β's and hence the bid–ask curvatures.

$$\begin{aligned}\beta_{+,\tau}&=(1+\psi_\tau)\beta_{o,\tau}\\ \beta_{-,\tau}&=(1-\psi_\tau)\beta_{o,\tau}\end{aligned} \tag{22}$$

3.4.3 Modeling the bid–ask ATM spread

The curvature adjustment via the β parameters does in fact model the idea that the further away you are from the ATM, the bigger the bid–ask spread. However, this change of variable yields a bid–ask spread of 0 ATM. It is therefore necessary to adjust for this issue by adding an ATM bid–ask factor that will be a function $\min(\frac{a_\tau}{2},\alpha_\tau)$, where α_τ attempts an ATM bid–ask half spread, adjusted if its value is such that it will be higher than the lowest point of the implied vol. Figure 9 illustrates how the variable α adjusts the ATM bid–ask spread.

3.4.4 Liquidity factors as a function of position size

The current model we have does not take into account the position size or responsiveness of the market to liquidity. This subsection aims to address in spirit that particular issue through a couple of simple ideas that need further investigation. As we have seen in Sections 3.4.2 and 3.4.3, ψ_τ and α_τ are not functions of the market participant's position size. Even though the skeleton of the change of variable allows us to build abstraction in a useful manner, the current assumption around a fixed p is obviously not a very good one. Liquidity would be less favorable to

Figure 8: Impact of a change in the ψ parameter in the IVP model.

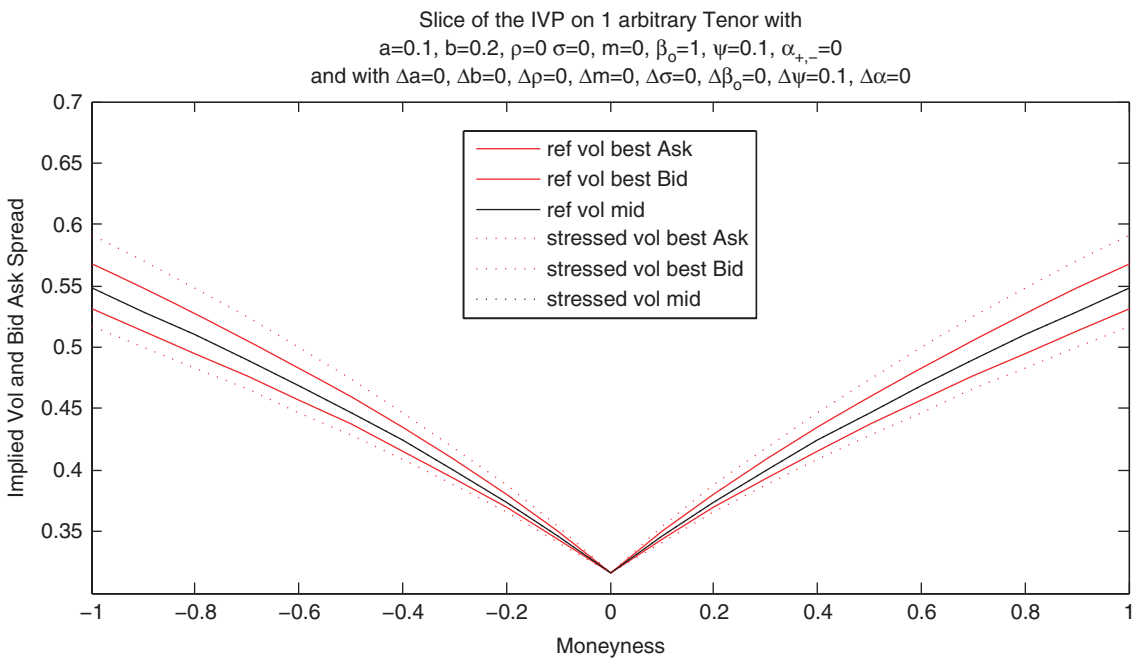
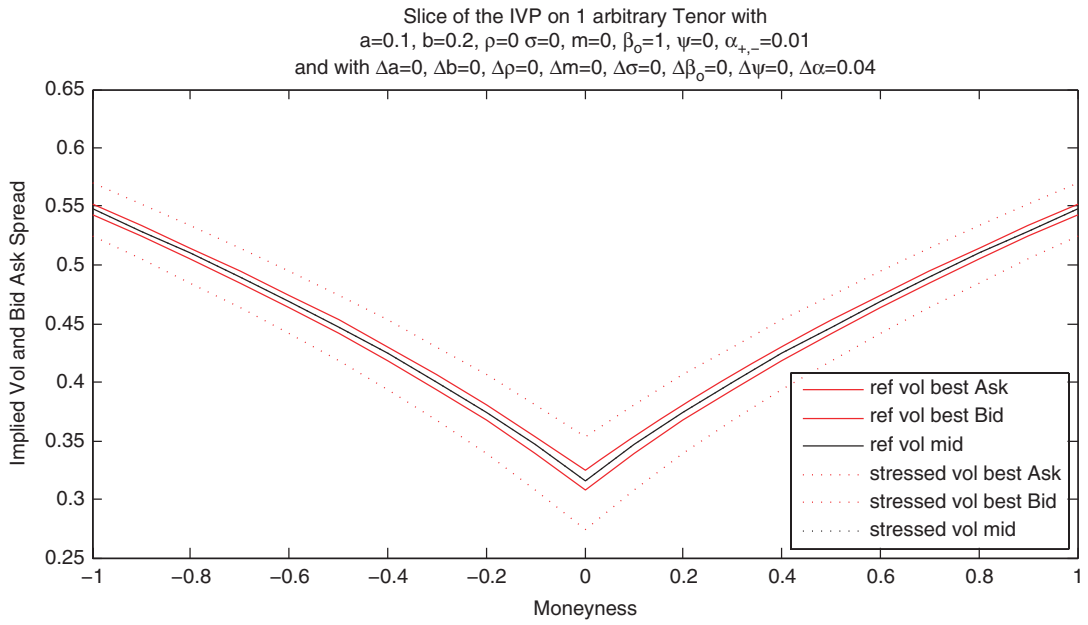


Figure 9: Impact of a change in the α parameter in the IVP model.



a market participant with a massive position compared with one with a smaller position. Suppose we call p our position size, with $\psi_\tau(p)$ and $\alpha_\tau(p)$ our liquidity functions expanded from Sections 3.4.2 and 3.4.3 to be functions as opposed to constants. We know that $p \in [0;\infty[$, $\psi_\tau(p) \in [0;1]$, and $\alpha_\tau(p) \in [0;+a[$, where a represents the ATM level. Finally, we also know that both functions should be increasing. Two simple functions which address this particular point are specified by equation (23):

$$\begin{aligned} \alpha_\tau(p) &= \alpha_0 + (a - \alpha_0)(1 - e^{-\eta_a p}) \\ \psi_\tau(p) &= \psi_0 + (1 - \psi_0)(1 - e^{-\eta_\psi p}) \end{aligned} \tag{23}$$

Here, η_a and η_w represent the liquidity elasticity of, respectively, the ATM and the wings, calibrated specifically for each product. The inspiration for these models has been taken from the inferred correlation formula [3]. A second desirable feature that the current model does not capture is the responsiveness of liquidity as it relates to various market conditions. It would be interesting to perhaps make η_a and η_w functions of market sentiment and/or functions of the rolling volatility of major macroeconomic indexes like the S&P.

3.4.5 Arbitrage constraint adjustment to liquidity factors

Note that once the bid–ask spread has been incorporated into the equation the arbitrage constraints are not assessed on the mid anymore but on the bid–ask. Equations (8) and (12) are adjusted to equations (24) and (25):

$$\forall \Delta,\ C(K - \Delta, \sigma_{IVP,+,t}(k)) - 2C(K, \sigma_{IVP,-,t}(k)) + C(K + \Delta, \sigma_{IVP,+,t}(k)) > 0 \tag{24}$$

$$C(K, T + \Delta, \sigma_{IVP,+,t}(k)) - C(Ke^{-r\Delta}, T, \sigma_{IVP,-,t}(k)) \geq 0 \tag{25}$$

3.4.6 The IVP equation

Incorporating the information on the gSVI, the ATM bid–ask spread, and the curvature adjustment of the wings, we get what we call the IVP in equation (26) with the constraint in equation (27):

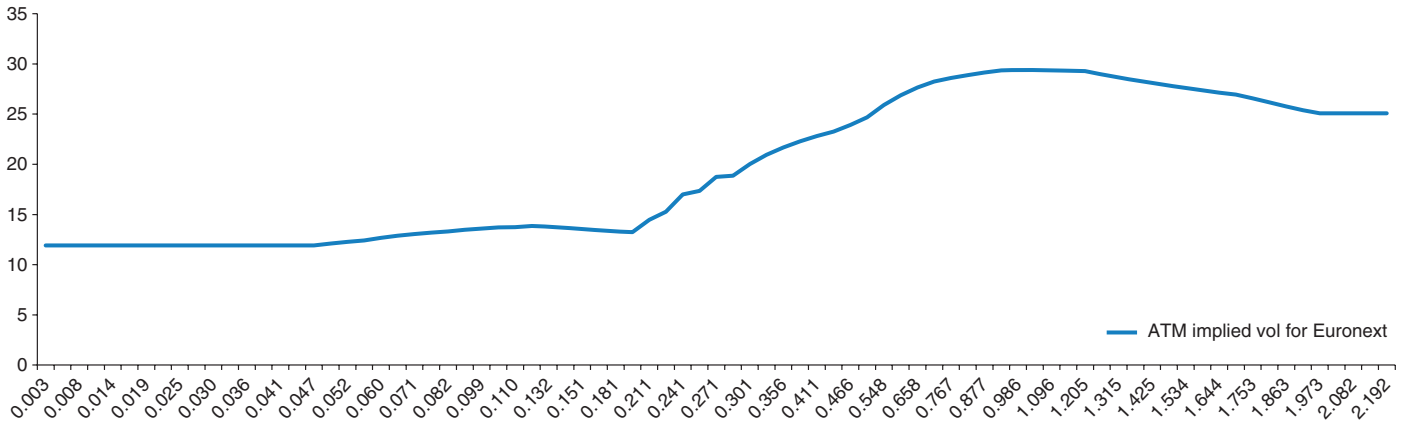
$$\begin{aligned}
\sigma^2_{IVP,o,\tau}(k) &= a_\tau + b_\tau \left[\rho_\tau (z_{o,\tau} - m_\tau) + \sqrt{(z_{o,\tau} - m_\tau)^2 + \sigma_\tau^2}\right] \\
z_{o,\tau} &= \frac{k}{\beta_{o,\tau}^{1+4|k-m|}} \\
\sigma^2_{IVP,+,\tau}(k,p) &= a_\tau + b_\tau \left[\rho_\tau (z_{+,\tau} - m_\tau) + \sqrt{(z_{+,\tau} - m_\tau)^2 + \sigma_\tau^2}\right] + \alpha_\tau(p) \\
z_{+,\tau} &= z_{o,\tau}[1 + \psi_\tau(p)] \\
\sigma^2_{IVP,-,p,\tau}(k,p) &= a_\tau + b_\tau \left[\rho_\tau (z_{-,\tau} - m_\tau) + \sqrt{(z_{-,\tau} - m_\tau)^2 + \sigma_\tau^2}\right] - \alpha_\tau(p) \\
z_{-,\tau} &= z_{o,\tau}[1 - \psi_\tau(p)] \\
\alpha_\tau(p) &= \alpha_0 + (a_\tau - \alpha_0)(1 - e^{-\eta_{\alpha_\tau} p}) \\
\psi_\tau(p) &= \psi_0 + (1 - \psi_0)(1 - e^{-\eta_{\psi_\tau} p})
\end{aligned} \tag{26}$$

$$\begin{aligned}
\beta_{+,\tau} &= [1 + \psi_\tau(p)]\beta_{o,\tau} \\
\beta_{-,\tau} &= [1 - \psi_\tau(p)]\beta_{o,\tau} \\
\tau &\in \{f, F\} \\
0 < \psi_\tau(p) &< 1 \\
0 < \alpha_\tau(p) & \\
\beta_{+,\tau}(p) > \beta_{o,\tau} &> \beta_{-,\tau}(p) \\
\eta_{\psi_\tau} &\in \mathbb{R}^+ \\
\eta_{\alpha_\tau} &\in \mathbb{R}^+
\end{aligned} \tag{27}$$

3.4.7 Modeling the decay factor

When creating a parsimonious volatility surface, in adding parameters (for example, here we have added the bid–ask spread) we need to take off some other parameters.

Figure 10: ATM implied volatility for Euronext 2004/01/02.



The IVP ignores tranches between the first and last tenor[3] and instead maps the different tenors through a decay factor λ, which we will see also becomes handy in the de-arbitraging optimization. The interpolation function is given by equation (28), with $\oplus \in \{+, -, o\}$ and $f < t < F$:

$$\sigma_{IVP,\oplus,t}(k) = \begin{cases} \sigma_{IVP,\oplus,f}(k) & \text{if} \quad t = f \\ \sigma_{IVP,\oplus,F}(k) & \text{if} \quad t = F \\ \sigma_{IVP,\oplus,f}(k) + [\sigma_{IVP,\oplus,F}(k) - \sigma_{IVP,\oplus,f}(k)] & \text{otherwise} \\ \quad \times \left[1 - \exp\left(-\lambda \frac{t-f}{F-f}\right)\right] & \end{cases} \tag{28}$$

The implied volatility surface may not always be strictly increasing or decreasing in the tenor space. Figure 10 exposes this specific observation, which would not yield a good fit with equation (28). Depending on whether or not we want to add a third tenor, we can adjust equation (28) by adding the tenor which yields the highest ATM vol – called the intermediate tenor I. The equation is adjusted to equation (29). An additional decay (λ_1 and λ_2 instead of λ) factor may be needed in order to account for this change of model.

$$\sigma_{IVP,\oplus,t}(k) = \begin{cases} \sigma_{IVP,\oplus,f}(k) & \text{if} \quad t = f \\ \sigma_{IVP,\oplus,f}(k) + [\sigma_{IVP,\oplus,I}(k) - \sigma_{IVP,\oplus,f}(k)] \times \left[1 - \exp\left(-\lambda_1 \frac{t-f}{I-f}\right)\right] & \text{if} \quad f < t < I \\ \sigma_{IVP,\oplus,I}(k) + [\sigma_{IVP,\oplus,F}(k) - \sigma_{IVP,\oplus,I}(k)] \times \left[1 - \exp\left(-\lambda_2 \frac{t-I}{F-I}\right)\right] & \text{if} \quad I < t < F \\ \sigma_{IVP,\oplus,F}(k) & \text{if} \quad t = F \end{cases} \tag{29}$$

3.4.8 Making the tradeoff between number of calibrated tenors and liquidity

The more parameters a model has, the better it can model the subtleties of the market. However, the more parameters it has, the harder it is to calibrate the model and the higher the risk of overfitting as well. Therefore, when one introduce a new model, one has to always examine the tradeoff between complexity and benefits. The IVP model makes the tradeoff between having fewer calibrated tenors compared with the gSVI [4] and accounting for more information per tenor fully calibrated thanks to a liquidity overlay model as well as a decay factor, which maps the first and last tenor and therefore enforces a full parametrization of the implied volatility surface, escaping the need for a cumbersome interpolation methodology which may reintroduce arbitrages. If we examine a typical situation, with 30 tenors to calibrate, the parametrization of the SVI model would need 150(5×30) parameters and the gSVI 180(6×30) parameters. None of these would be able to calibrate liquidity and both would suffer in the optimization by constraints, especially because of the calendar spread constraint that would require sequential optimization [4]. The IVP constraint on the contrary, in the worst case (containing implied vols that are not strictly increasing or decreasing in the tenor space like in Figure 10), would only require 24 (8×3) – that is, a decrease of 87% compared with the gSVI and 84% compared with the SVI.[4]

3.4.9 The IVP's constraints

There exist two constraints that make IVP "often" (as explained in Section 3.2) arbitrage free. The first condition on the falling variance (equation (13)) does not change. However, we need to adjust for equation (16), which is replaced in the IVP by equation (30):

$$\left| T \frac{1 + |k-m| \ln \beta_{o,T}}{\beta_{o,T}^{|k-m|}} \left[b \left(\rho + \frac{\left(\frac{k-m}{\beta_{o,T}^{|k-m|}} - m \right)}{\sqrt{\left(\frac{k-m}{\beta_{o,T}^{|k-m|}} - m \right)^2 + \sigma^2}} \right) \right] \right| \leq 4 \tag{30}$$

Proof. We have seen in equation (11) that $\forall K, \forall T, |T \partial_K \sigma^2(K, T)| \leq 4$. We know that $\partial_k \sigma_{IVP}^2(k) = \frac{\partial z}{\partial k} \times \frac{\partial \sigma}{\partial z}$. Calculus gives

$$\frac{\partial z}{\partial k} = \frac{1 + (k-m) \ln \beta_{o,T} (1_{k>m} - 1_{k<m})}{\beta_{o,T}^{|k-m|}} = \frac{1 + |k-m| \ln \beta_{o,T}}{\beta_{o,T}^{|k-m|}}$$

$$\frac{\partial \sigma}{\partial z} = b \left(\rho + \frac{2(z-m)}{2\sqrt{(z-m)^2 + \sigma^2}} \right) = b \left(\rho + \frac{\left(\frac{k-m}{\beta_{o,T}^{|k-m|}} - m \right)}{\sqrt{\left(\frac{k-m}{\beta_{o,T}^{|k-m|}} - m \right)^2 + \sigma^2}} \right)$$

Now, plugging in equation (11) the constraint yields equation (30). □

[3] Note that ignoring the in-between tranches is not obligatory. It is straightforward to add tranches if necessary, and fit each tenor with the same interpolation methodology.

[4] Note that these numbers would be even more substantial compared with a grid mode, where you typically have 30^2 parameters (30 tenors with 30 levels of moneyness).

4 Bumping and "de-arbitraging" the volatility surface

4.1 Bumping the volatility surface

There are many methodologies for bumping the volatility surface. One can either do scenario analysis, for example, if one wants to know what happens if the volatility of a particular point moves by x amount. One might like to know whether the induced volatility surface is arbitrage free or not and, if not, to what extent one can stress that particular point until an arbitrage has been reached. Similarly, one could be working within a risk department in an investment bank and be asked to investigate what the associated risk is for a certain product. More specifically, one might like to know the associated risk with respect to the change in volatility only. One could record the proportional historical move of the volatility surface on sticky[5] log-moneyness and apply these moves to today's volatility surface. One would want to make sure that the induced volatility surface is arbitrage free. Let us call these various bumped volatility surfaces a target volatility surface, σ_{target}, which may or may not have induced arbitrages as a result of being bumped.

4.2 De-arbitraging the implied vol in the equities and commodities markets

We would like to insure that σ_{target} is arbitrage free. The closest arbitrage-free volatility surface will be found by implementing the optimization problem specified in equation (31) subject to constraints (8) and (12). For the sake of clarity, we set $\Omega = \bigcup_{t_1 \le t \le T}(\rho_t, \sigma_t, a_t, b_t, m_t, \beta_t, \psi_t, \alpha_t, \lambda_t)$. Note that in the original paper [4], the authors added additional constraints for optimization purposes. We have taken these constraints off for clarity in this article. We refer to [4, 15] for tips on how to make the algorithm faster. Note that algorithm (31) differs from the one presented in [4] by taking into account the bid–ask spread, more specifically the fact that a mid implied vol can have arbitrage opportunities (which is perfectly fine as long as the bid–ask prevents the successful deployment of an arbitrage occurrence).

Solve:

$$\hat{\Omega} = \underbrace{\arg\min}_{\Omega} \sum_{t=t_1}^{T} \sum_{i=1}^{N} [\sigma_{IVP,t}(K_i) - \sigma_{target,t}(K_i)]^2$$

$$\forall \Delta, t, i \text{ subject to:}$$

$$C(K - \Delta, \sigma_{IVP,+,t}(k)) - 2C(K, \sigma_{IVP,-,t}(k)) + C(K + \Delta, \sigma_{IVP,+,t}(k)) > 0$$

$$C(K, T + \Delta, \sigma_{IVP,+,t}(k)) - C(Ke^{-r\Delta}, T, \sigma_{IVP,-,t}(k)) \ge 0 \qquad (31)$$

5 De-arbitraging methodology for the FX market

5.1 FX market conventions as they relate to the IVP parametrization

FX products are usually quoted in terms of "risk reversals" and "flies" in delta space, as opposed to moneyness space in equities.

5.1.1 Risk reversal as it relates to the ρ parameter

Risk reversals are defined in the FX market by equation (32), where RR_d refers to the risk reversal of delta d, $\sigma_{c,d}$ represents the call implied vol of delta d, and $\sigma_{p,d}$ represents its mirror put. A quick plot reveals an obvious connection to the ρ parameter (as plotted in Figure 4):

$$RR_d = \sigma_{c,d} - \sigma_{c,d} \qquad (32)$$

5.1.2 Fly_d as it relates to the b parameter

Flies are defined in the FX market by equation (33), where FLY_d refers to the fly of delta d and σ_{ATM} represents the ATM implied vol. A quick plot reveals an obvious connection to the b parameter (as plotted in Figure 3):

$$\text{FLY}_d = \frac{\sigma_{c,d} + \sigma_{p,d}}{2} - \sigma_{ATM} \qquad (33)$$

5.1.3 Fly_{25} and Fly_{10} as they relate to the β parameter

Figure 11 plots Fly25 and Fly10 for the CHF/JPY 1-month tenor on an arbitrary date. As one can see, the rate at which implied vol increases as a function of delta is negative as you move away from the ATM. This mirrors the concept of a downside transform.

5.2 Condition on the implied correlation

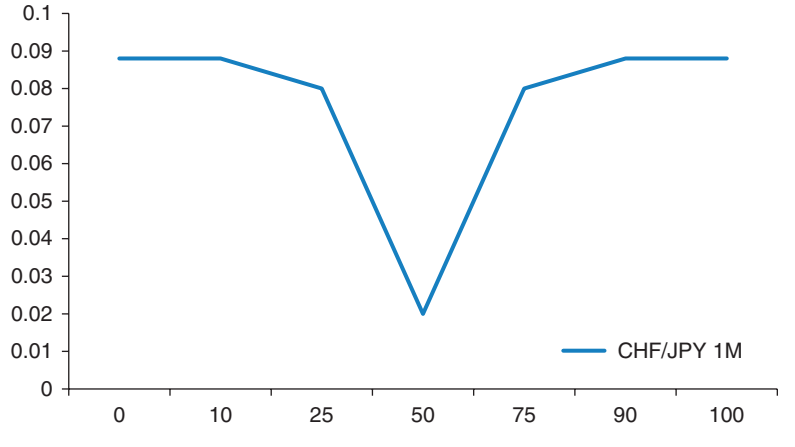
Let's call $\Omega_t = \{S_1, S_2, \ldots S_n\}$ the set of all n currencies with the same referential currency (say the US dollar) at time t. Let's also call $\Phi_t = \{\sigma_1, \sigma_2, \ldots, \sigma_n\}$ the set of all **direct** volatilities induced by the set Ω_t at time t. Let's call $\complement_t = \{\rho_{1,2}, \rho_{1,3}, \ldots, \rho_{2,3}, \ldots, \rho_{n-1,n}\}$ the set of implied correlations induced by, respectively, (σ_1, σ_2), (σ_1, σ_3),…, (σ_2, σ_3), …, (σ_{n-1}, σ_n), as seen in equation (34). Let's call $\Theta_t = \{\sigma_{1,2}, \sigma_{1,3}, \ldots, \sigma_{2,3}, \ldots, \sigma_{n-1,n}\}$ the set of **indirect** volatilities which are geometrically opposite to $\complement_t$ (see Figure 12). We also know that the correlation matrix $\complement_t$ (given by equation (34)) must, like every correlation matrix, be positive semi-definite.

$$\complement_t = \begin{pmatrix} 1 & \rho_{1,2} & \rho_{1,3} & \cdots & \rho_{1,n} \\ \rho_{2,1} & 1 & \rho_{2,3} & \cdots & \rho_{2,n} \\ \rho_{3,1} & \rho_{3,2} & 1 & \cdots & \rho_{3,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{n,1} & \rho_{n,2} & \cdots & \rho_{n,n-1} & 1 \end{pmatrix} \qquad (34)$$

5.3 Bumping the implied correlation coefficient

Within the scope of the full revaluation methodology, it may sound intuitive that one might also want to bump the implied correlation $\complement_t$ on top of the volatility surface,

Figure 11: Fly_{25} and Fly_{10} for the CHF/JPY 1-month tenor on an arbitrary date.



[5] The log-moneyness in the reference volatility surface does not change, so the strikes are adjusted with respect to the change in spot.

Figure 12: FX triangle for five currencies.

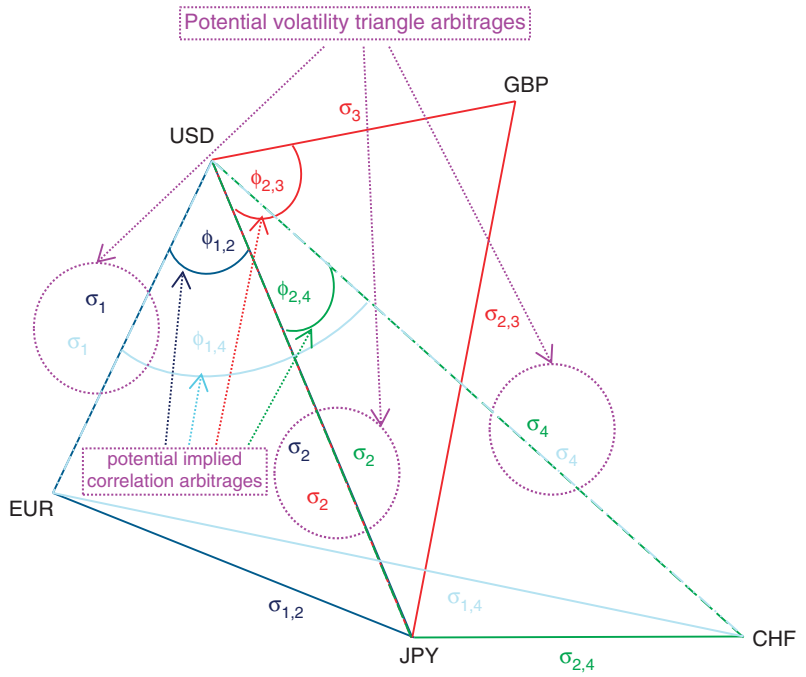
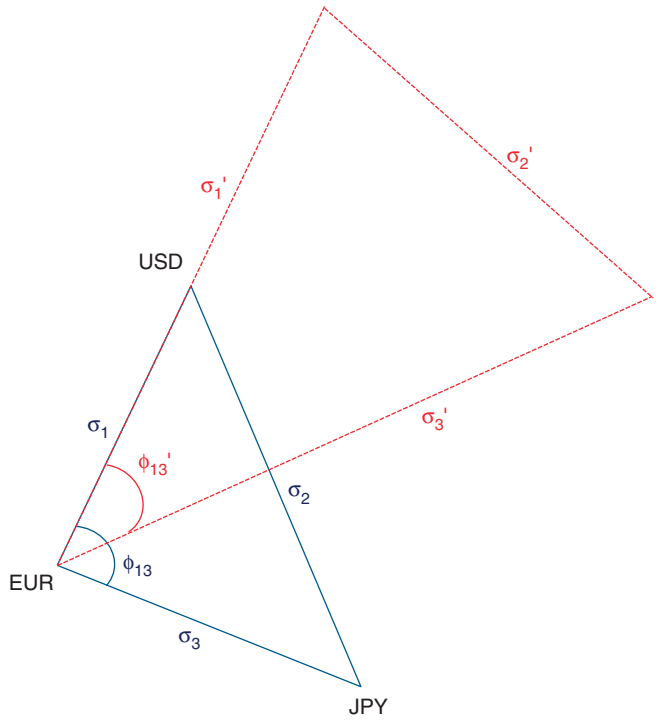


Figure 13: Example of bumped FX triangle.



the spots, the interest rates, etc. Figure 13 is a good example of why this is a bad idea. Suppose we take the same triangle as in Figure 1 and decide to bump σ_1 and σ_3 such that their new values, σ'_1 and σ'_3, are twice as big as in their non-bumped phase, and that we decide to keep σ'_2 unchanged. Figure 13 shows that the implied correlation ρ_{13} is automatically bumped to ρ_{13}. So, choosing to bump the implied correlation matrix on top of choosing to bump the respective volatilities increases spuriously the risk of our full revaluation system.

5.4 De-arbitraging the implied correlation matrix

If we call $\widetilde{\mathbb{C}}_{(1),t}$ the first upper left square matrix of $\widetilde{\mathbb{C}}_t$ (the de-arbitraged positive semi-definite correlation matrix), that is 1, $\widetilde{\mathbb{C}}_{(2),t}$ the second upper left square matrix of $\widetilde{\mathbb{C}}_t$, that is $\begin{pmatrix} 1 & \widetilde{\rho}_{1,2} \\ \widetilde{\rho}_{2,1} & 1 \end{pmatrix}$, and $\widetilde{\mathbb{C}}_{(n),t}$ the nth upper left square matrix of $\widetilde{\mathbb{C}}_t$, that is $\widetilde{\mathbb{C}}_t$ itself, then we know that for this correlation matrix to be positive semi-definite, the determinant of all these matrices should be greater than or equal to 0. That is, in order to adjust our de-arbitraging methodology introduced in [4], we need to do everything mentioned there and also perform the optimization problem introduced in equation (35):

$$
\begin{aligned}
&\text{Solve:}\\
&\widetilde{\mathbb{C}}_t = \underbrace{\arg\min}_{\widetilde{\rho}_{m,n}\forall m,n}\{(\mathbb{C}_t - \widetilde{\mathbb{C}}_t) \circledast (\mathbb{C}_t - \widetilde{\mathbb{C}}_t)\}\\
&\text{subject to:}\\
&\rho_{1,2} = \tfrac{\sigma_3^2-\sigma_2^2-\sigma_1^2}{2\sigma_1\sigma_2}\\
&\forall m \in [1, n], \det(\widetilde{\mathbb{C}}_{(m),t}) \geq 0\\
&\sigma_1 + \sigma_2 + \sigma_3 > 2\max(\sigma_1 + \sigma_2 + \sigma_3)\\
&\sigma_3^2 = \sigma_2^2 + \sigma_1^2 + 2\rho_{1,2}\sigma_1\sigma_2
\end{aligned}
\tag{35}
$$

Here, $\circledast$ represents the matrix element-by-element multiplication commonly represented by .* in many programming languages (e.g., Matlab). Incorporating this result into equation (31) and adjusting to the idea that there are as many implied correlation matrices as there are combinations of tenor and log-moneyness matrices, we get the optimization problem specified in equation (36):

$$
\begin{aligned}
&\text{Solve:}\\
&\{\hat{\Omega}, \widetilde{\mathbb{C}}\} = \underbrace{\arg\min}_{\Omega,\mathbb{C}} \sum_{t=t_1}^{T}\sum_{i=1}^{N}[\sigma_{IVP,t}(K_i) - \sigma_{target,t}(K_i)]^2 + (\mathbb{C}_{i,t} - \widetilde{\mathbb{C}}_{i,t}) \circledast (\mathbb{C}_{i,t} - \widetilde{\mathbb{C}}_{i,t})\\
&\text{subject to:}\\
&\forall \Delta,\ C(K-\Delta, \sigma_{IVP,+,t}(k)) - 2C(K, \sigma_{IVP,-,t}(k)) + C(K+\Delta, \sigma_{IVP,+,t}(k)) > 0\\
&C(K, T+\Delta, \sigma_{IVP,+,t}(k)) - C(Ke^{-r\Delta}, T, \sigma_{IVP,-,t}(k)) \geq 0\\
&\rho_{t,(i,k)} = \tfrac{\sigma^2_{t,(i,j)} - \sigma^2_{t,(j,k)} - \sigma^2_{t,(j,k)}}{2\sigma_{t,(i,j}\sigma_{t,(j,k)}}\\
&\forall m \in [1, n], \det(\widetilde{\mathbb{C}}_{(m),i,t}) \geq 0\\
&\forall i \neq j \neq k \in [1, n],\ \sigma^2_{t,(i,k)} = \sigma^2_{t,(j,k)} + \sigma^2_{t,(i,j)} + 2\rho_{t,(i,k)}\sigma_{t,(i,j)}\sigma_{t,(j,k)}\\
&\forall i \neq j \neq k \in [1, n],\ \sigma_{t,(i,k)} + \sigma_{t,(i,j)} + \sigma_{t,(j,k)} > 2\max(\sigma_{t,(i,k)} + \sigma_{t,(i,j)} + \sigma_{t,(j,k)})\\
&\forall i \neq j \neq k \neq l \in [1, n],\ S_{t,(i,k)} = S_{t,(i,j)} \times S_{t,(j,k)} = S_{t,(i,l)} \times S_{t,(l,k)}
\end{aligned}
\tag{36}
$$

6 Simplified version and closed-form optimization

6.1 Implied volatility from market prices

Algorithm 1 gives an example of how to fetch an implied volatility for a strike K and a tenor T out of the market observed price P through the bisection method. Note that there exist faster methods, like the Newton–Raphson method, which happen to be faster but have the undesirable property of sometimes not converging toward a solution. If both speed and accuracy are desirable, we recommend Brent's method [2], which is essentially a hybrid between the bisection method and Newton–Raphson.

Algorithm 1: Implied vol Fether finds the implied volatility given the options price

Input: Option's Price P, Option's Model M (Normal vs Log-Normal), Spot S_t, Strike K, Interest rate r, Dividend yield d, Expiry T
Output: Implied Vol σ

```
ε ← 0.01
N = 50
σ+ ← 3:0
σ− ← 0:01
for i ← 1 to N do
    σ ← (σ+ + σ−)/2
    if P > BS_M(S_t, K, σ, T, r, d) then
        σ+ ← σ
    else
        σ− ← σ
return σ
```

6.2 Simplifying the IVP equation

We present in this section a simplified version of the IVP (sIVP) which does not require any algorithmic optimization other than that associated with Algorithm 1. The rationale of this simplification is that the skeleton of the risk is given by the a, b, and ρ parameters. The downside transform for the mid will be set to 1 for this simplification. For the liquidity component, an α value of 2% in the first tenor and 1% in the last seems to be a simple hack for this simplification.[6] A wing liquidity ψ of 5% also seems a reasonable assumption.[7] Equation (37) summarizes these adjustments:

$$\begin{aligned}
\sigma^2_{M,+,\tau}(k) &= a_\tau + b_\tau(\rho_\tau[k(1+\psi_\tau)-m] + |k(1+\psi_\tau)-m|) + \alpha_\tau \\
\sigma^2_{M,o,\tau}(k) &= a_\tau + b_\tau(\rho_\tau(k-m) + |k-m|) \\
\sigma^2_{M,-,\tau}(k) &= a_\tau + b_\tau(\rho_\tau[k(1-\psi_\tau)-m] + |k(1-\psi_\tau)-m|) - \alpha_\tau
\end{aligned} \tag{37}$$

6.3 Calibrating the simplified IVP equation in closed form

Most quant models and systems should be well within IT capabilities. For this reason one may want to avoid difficult and time-consuming optimization algorithms at the cost of precision. Bearing this in mind, with the sgSVI we propose the closed-form calibration laid out in equation (38). The idea is to select six points in order to calibrate the skeleton of the first and last tenors. First, the general level of the volatility surface will be captured through the a parameter at the lowest point of the implied vol (assumed to be ATM). The vol of vol and skew will then be determined by solving a system of two equations in two unknowns, where x happens to be an arbitrary market observable moneyness (for example, 0.2). In equations (38), $\sigma_{M,o,\tau}(x)$ represents the mid implied vol observed in the market M at tenor τ for moneyness x.[8] Recall that o represents the "mid" and + the "asked" price. We have not performed the calculation on the "bid" because we assume that the bid–ask spread is symmetric, since we would like our vol to be as parsimonious as possible.

[6] The closed-form optimization still works with a downside transform bigger than one.
[7] Note that this combination solves most arbitrages on the implied vol.
[8] Note that in practice you may want to manually enforce $b > 0$ and $-1 < \rho < 1$ as market data is sometimes noisy and a function of asymmetric bid–ask spreads, and may corrupt the natural boundaries of these parameters.

$$\begin{aligned}
\hat{a}_\tau &= \sigma^2_{M,o,\tau}(m) \\
\hat{b}_\tau &= \frac{\sigma^2_{M,o,\tau}(m+k) + \sigma^2_{M,o,\tau}(m-k) - 2\hat{a}_\tau}{2|k|} \\
\hat{\rho}_\tau &= \frac{\sigma^2_{M,o,\tau}(m+k) - \sigma^2_{M,o,\tau}(m-k)}{2\hat{b}_\tau k} \\
\hat{\alpha}_\tau &= \sigma^2_{M,+,\tau}(0) - \hat{a}_\tau + m\hat{b}_\tau\hat{\rho}_\tau + |m|\hat{b}_\tau \\
\hat{\psi}_\tau &= \frac{\sigma^2_{M,+,\tau}(m) + \sigma^2_{M,-,\tau}(m) - 2\hat{a}_\tau}{2|m|\hat{b}_\tau}
\end{aligned} \tag{38}$$

Proof

$$\begin{aligned}
\sigma^2_{M,o,\tau}(m) &= a_\tau + b_\tau\left[\rho_\tau(m-m) + |(m-m)|\right] = a_\tau \\
\sigma^2_{M,o,\tau}(m-k) &= \sigma^2_{M,o,\tau}(m-k-m) \\
&\quad + b_\tau\left[\rho_\tau(m-k-m) + |m-k-m|\right] \\
\sigma^2_{M,o,\tau}(m+k) &= \sigma^2_{M,o,\tau}(m+k-m) \\
&\quad + b_\tau\left[\rho_\tau(m+k-m) + |m+k-m|\right]
\end{aligned}$$

$$\begin{aligned}
\sigma^2_{M,o,\tau}(m-k) - \sigma^2_{M,o,\tau}(m+k) &= 2 \times b_\tau\rho_\tau k \\
\sigma^2_{M,o,\tau}(m-k) + \sigma^2_{M,o,\tau}(m+k) &= 2 \times b_\tau|k| + a_\tau \\
\sigma^2_{M,+,\tau}(0) &= a_\tau + b_\tau\rho_\tau + b_\tau|-m| + \alpha_\tau \\
\sigma^2_{M,+,\tau}(m) &= a_\tau + b_\tau(\rho_\tau m\psi_\tau + |m\psi_\tau|) \\
\sigma^2_{M,-,\tau}(m) &= a_\tau + b_\tau(\rho_\tau(-m)\psi_\tau + |-m\psi_\tau|) \\
\sigma^2_{M,+,\tau}(m) + \sigma^2_{M,-,\tau}(m) &= 2a_\tau + 2b_\tau|m|\psi_\tau
\end{aligned}$$

□

7 Conclusion

We have shown that the positivity constraints on the butterfly and the calendar spread are the two necessary and sufficient conditions that make a volatility surface arbitrage free in the commodities and equities markets. We have also demonstrated that although the gSVI parametrization did model accurately the various behaviors of the FX market, its de-arbitraging methodology as presented in [4] failed to take into consideration the triangle constraint and the implied correlation constraints present in the FX market (which neither the equities nor the commodities markets have). We therefore laid out the specification to bump the various volatility surfaces in the FX market. More specifically, we showed that bumping each volatility surface individually prevented the need to also bump the implied correlation matrix. Rather, we correct it by enforcing that the determinant of each of the upper left-hand corners of the square matrices forming the implied correlation matrix be positive. We have also introduced the IVP model, which makes the tradeoff between having fewer calibrated tenors compared with the gSVI [4] but accounting for more information per tenor fully calibrated thanks to a liquidity overlay model as well as a decay factor, which maps the first and last tenor and therefore enforces a full parametrization of the implied volatility surface – escaping the need for a cumbersome interpolation methodology, which may reintroduce arbitrages.

Babak Mahdavi Damghani has been working in the financial industry within the Quantitative space (Exotics, High Frequency, Structuring, FO, and Risk Quant) through both the buy and the sell side. He did post-graduate studies in the applied mathematical and computational sciences at the University of Cambridge, University of Oxford, and Ecole Polytechnique.

REFERENCES

[1] Benaim, S., Friz, P., and Lee, R. 2008. On the Black–Scholes implied volatility at extreme strikes. In Cont, R. (ed.), *Frontiers in Quantitative Finance Volatility and Credit Modeling*. New York: Wiley Finance, page 2.
[2] Brent, R.P. 1973. *Algorithms for Minimization without Derivatives*. Englewood Cliffs, NJ: Prentice-Hall, chapter 4.
[3] Damghani, B.M. 2013. The non-misleading value of inferred correlation: An introduction to the cointelation model. *Wilmott*, 67, 50–61.
[4] Damghani, B.M. and Kos, A. 2013. De-arbitraging with a weak smile: application to skew risk. *Wilmott*, 64, 40–49.
[5] Dupire, B. 1994. Pricing with a smile. *Risk*, pp. 17–20.
[6] Dupire, B. 1997. Pricing and hedging with a smile. In Dempster, M.A.H. and Pliska, S.R. (eds), *Mathematics of Derivative Securities*. Cambridge: Cambridge University Press, pp. 103–111.
[7] Fokker, A. 1914. Die mittlere energie rotierender elektrischer dipole im strahlungsfeld. *Annalen der Physik*, 43, 810–820.
[8] Garman, M.B. and Kohlhagen, S.W. 1983. Foreign currency option values. *Journal of International Money and Finance*, 2(3), 231–237.
[9] Gatheral, J. 2006. *The Volatility Surface: A Practitioner's Guide*. New York: Wiley Finance.
[10] Gatheral, J. and Jacquier, A. 2011. Convergence of Heston to SVI. *Quantitative Finance*, 11, 1129–1132.
[11] Gatheral, J. and Jacquier, A. 2014. Arbitrage-free SVI volatility surfaces. *Quantitative Finance*, 14, 59–71.
[12] Hagan, P.S., Kumar, D., Lesniewski, A.S., and Woodward, D.E. 2002. Managing smile risk. *Wilmott*, Sept, 84–108.
[13] Heston, S.L. 1993. A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies*, 6, 327–343.
[14] Karoui, N. El. 2003. Couverture des risques dans les marchés financiers. Course at the *É*cole Polytechnique, Paris, p. 92. Available at: www.cmap.polytechnique.fr/elkaroui/masterfin034.pdf.
[15] Martini, C. 2009. Quasi-explicit calibration of Gatheral's SVI model. Available at: www.zeliade.com/whitepapers/zwp-0005.pdf.
[16] Planck, M. 1917. Sitz. ber. preu. Akad. p. 324.
[17] Rogers, C. and Tehranchi, M. 2009. The implied volatility surface does not move by parallel shifts. Available at: www.statslab.cam.ac.uk/chris/papers/iv.pdf.
[18] Schonbucher, P.J. 1999. A market model for stochastic implied volatility. SFB 303 Working Paper No. B-453.
[19] Tankov, P. and Touzi, N. 2010 Calcul stochastique en finance. Available at: http://www.cmap.polytechnique.fr/~touzi/Poly-MAP552.pdf.