Reconciling Caplets and Swaption Prices:

LMM-SABR and Model-Independent Results

R Rebonato

Head of Front-Office Market Risk and Quantitative Analytics, GBM, RBS

Visiting Lecturer, OCIAM, Oxford University

Adjunct Professor, Business School, Imperial College, London

1 What This Talk Is About

In this talk I ask whether the US$ swaption and caplet markets are internally consistent.

What do I mean by "internally consistent"?

For a given currency, the same quantities, ie, forward rates, constitute the common underliers for both markets.

By asking whether the caplet and swaption markets are internally consistent I therefore ask whether the same pricing model *with the same calibration* can simultaneously account for the observed market prices of swaptions and caplets.

The precise results will clearly depend on the chosen option model. However, that I am looking for almost model-independent indications as to whether the two markets may be congruent.

As the LMM-SABR model provides a set of no-arbitrage conditions within which different specific models can be naturally nested it constitutes a god starting point for my analysis.

Since the two markets are connected by the same underliers (forward rates), the question naturally arises as to how in an efficient market swaptions and caplet could fail to display coherence. A copious literature in the area of potential causes for market inefficiency (see, eg, Shleifer (2000)) points to the limits to and the riskiness of (pseudo)-arbitrage. I

therefore discuss in the concluding section to what extent a marked-to-market player could bring about the reconciliation of the two markets if they were indeed found to trade out of line with each other.

2 The LMM-SABR Pricing Framework

Both the LMM and its offspring, the LMM-SABR are usually referred to as 'models'. However, they are not truly models, but sets of no-arbitrage conditions.

For the sake of brevity we report only those features of the SABR and LMM-SABR model necessary to understand the analysis below. In the SABR model, the process for the forward (swap) rate, f_t^T, is of the stochastic-volatility CEV type:

$$df_t^T = \left(f_t^T\right)^{\beta^T} \sigma_t^T dz_t$$

$$\frac{d\sigma_t^T}{\sigma_t^T} = \nu^T dw_t$$

$$E\left[dz_t dw_t\right] = \rho^T dt$$

The superscript T indicates that the parameters of the SABR model (β^T, ρ^T and ν^T), as quoted in the market, depend on the expiry of each forward rate. On any given day, no single set of parameters describes either the whole European swaption *or* caplet market.

The situation is reminiscent of the Black implied volatilities that were quoted in the 1990s, and that also displayed a dependence on the expiry of the forward rate (the so-called 'term-structure of volatilities'). This expiry dependence suggests that the Black (and now the SABR) models are reduced-form models, that produce the correct market prices for *European* options, without having to specify a more complex latent structure that would only affect multi-forward options (such as swaptions).

In the LMM and in the LMM-SABR framework one then has to make assumptions about this latent structure. An infinity of specifications are consistent with a given set of European swaption (or caplet) prices. Each assumption will specify a particular 'model'.

Financial justification and statistical analysis must therefore be invoked to choose a desirable specifications for the volatility and correlation functions.

The link between these unobservable inputs and market observables is *via* the smile surface, because any choice for the volatility function will uniquely determine the current and future smile surface.

The future smile is, of course, unknown. However, historical observation of smile surfaces shows that their shapes remain remarkably stable as a function of the residual time to maturity of the forward rates.

The challenge for a dynamic model is therefore to explain the European prices for all strikes and expiries obtained by the reduced-form SABR parametrization in a way that is both parsimonious, financially justifiable and reflective of the observed regularities in the shape of the smile surface: **future smiles should 'look like' the smiles which have been observed in the past**. We shall show that some very simple and *prime facie* appealing choices for the volatility functions fail to pass this last test.

The LMM-SABR model can be madeto satisfy these requirements. It is characterized by the following constitutive equations*:

$$df_t^i = \left(f_t^i\right)^{\beta_i} s_t^{T_i} dz_t^i \tag{1}$$

$$s_t^{T_i} = k_t^{T_i} g_t^{T_i} \tag{2}$$

$$\frac{dk_t^{T_i}}{k_t^{T_i}} = \mu_k^i dt + h_t^{T_i} dw_t^i \tag{3}$$

$$E[dz_t^i dz_t^j] = \rho_{ij} dt \tag{4}$$

$$E[dw_t^i dw_t^j] = r_{ij} dt \tag{5}$$

*The equations are expressed in the measure under which the forward rate is a martingale. Expressions for the no-arbirage drfits of the forward rates and the volatilities are reported in Rebonato, McKay and White (200X). These drifts are not required in the present study.

$$E[dw_t^i dz_t^j] = R_{ij} dt \tag{6}$$

The function $s_t^{T_i}$ represents the instantaneous volatility of the forward rate of expiry T_i, and is made up of the product of a deterministic back-bone, the function $g()$, and a stochastic multiplicative scaling factor, $k_t^{T_i}$, that randomly changes the level of volatility through time.

We note that Rebonato and White (2010) have shown that very accurate analytic estimates of caplet and swaption prices can be obtained from the model parameters without having to resort to a Monte Carlo simulation. This feature greatly simplifies the task.

3 The $g()$ and $h()$ Functions

To ensure time homogeneity (that guarantees self-similarity of the future smile surface) we use $g_t^T = g(T-t)$ and $h_t^T = h(T-t)$.

The assumption that underlies this choice is that the only feature that differentiates two forward rates is their residual time to expiry.

Fitting market prices of caplets (both with deterministic or stochastic volatility) clearly indicates that the instantaneous volatility of a forward should not be constant throughout its life, with a maximum instantaneous volatility encountered on average when it is 6 to 18 months away from expiry.

This chimes well both with a financial account of how the actions of monetary authorities are likely to affect the volatility of forward rates (see,eg, Rebonato

(2004), Rebonato (2002) and Rebonato, McKay and White (2009)), and with direct econometric studies of the volatility of forward rates as a function of their residual time to expiry. There is a rich literature supporting humped volatility structures. See, eg, Amin and Morton (1994), Goncalves and Issler (1996), Ritchen and Chuang (1999), Brigo and Mercurio (2001), Rebonato, McKay and White (2009) among others. The time-homogeneity assumption is also common in the literature on interest-rate volatility: see, eg, Rebonato (2204), Amin and Morton (1994), Ritchen and Chuang (1999), Mercurio and Moraleda (2000).

We therefore choose for the forward-rate volatility function the following humped and time homogeneous functional form:

$$g(T_i, t) = g(T_i - t) = g(\tau_i) =$$

$$[a + b\,(\tau_i)]\,\mathsf{exp}\,(c\tau_i) + d$$

A typical shape for the $g()$ function is shown in Fig 1, and Fig 2 displays a typical shape of the $g()$ function obtained from fitting its parameters (a, b, c and d) to caplet prices as described in Rebonato, McKay and White (2009).

Note the goodness of the fit, and the natural recovery of a humped shape from market data.

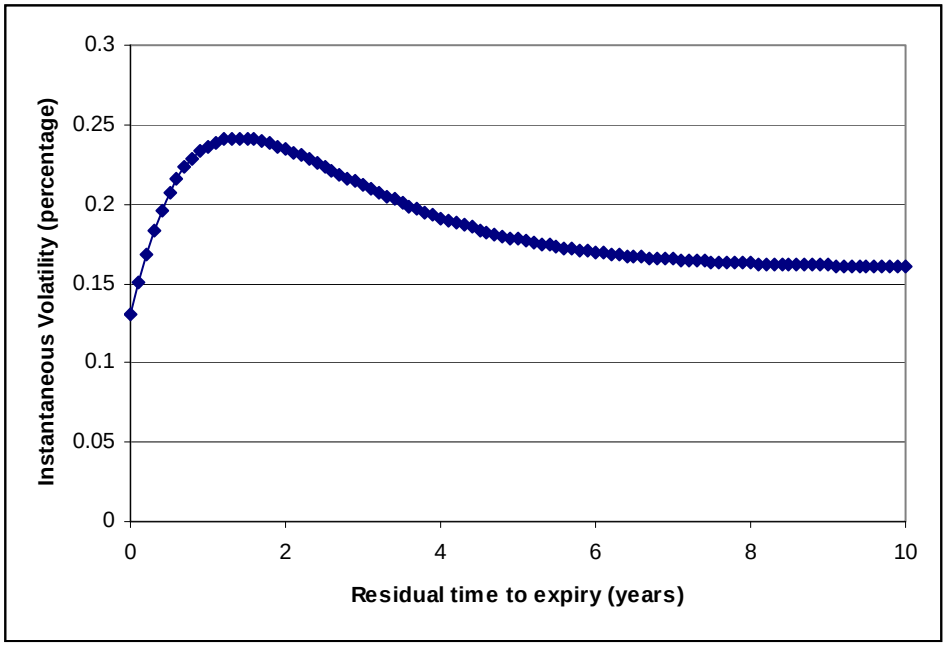

Figure 1: A typical shape for the time-homogenous deterministic backbone, $g()$, of the stochastic volatility, s.

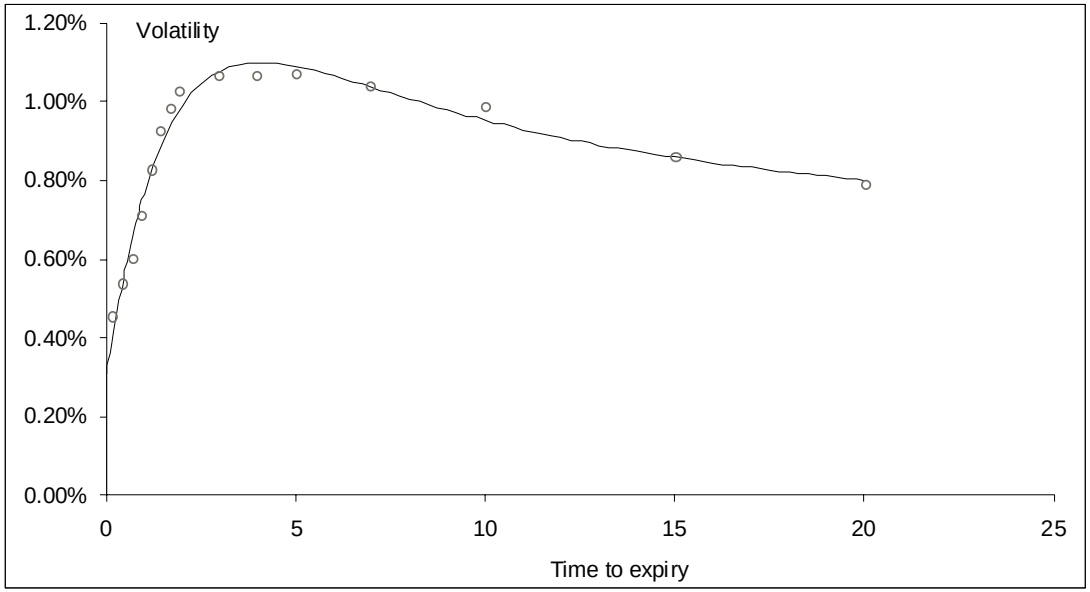

Figure 2: Calibration of $g()$ function to caplet prices, 5-July-2010, USD caplet market, RMS. $\beta = 0$ (normal case, 1% means 100 bps).

Finally, we note that on any given day the market implied volatilities of caplets of all expiries cannot in general be exactly recovered using the same parameters a, b, c and d of the $g()$ function. Perfect pricing is ensured by selecting slightly different values, $k_0^{T_i}$, for the initial values of the process $k_t^{T_i}$ in Equation (3).

Similarly, in order to fit match exactly the SABR volatility of volatility for all expiries exactly, constant, expiry-specific factors, $\xi_i^{T_i}$, pre-multiply the $h()$ function, so that the dynamics for the stochastic drivers, $k_t^{T_i}$, become:

$$\frac{dk_t^{T_i}}{k_t^{T_i}} = \mu_k^i dt + \xi_i^{T_i} h_t^{T_i} dw_t^i$$

Needless to say, doing so breaks exact time homogeneity. However, the corrections are in general very small, as can be appreciated by looking at Fig 2. Although a formal Bayesian estimation was not employed in the fitting, the adjustment factors $\left\{k_t^{T_i}\right\}$ and $\left\{\xi_i^{T_i}\right\}$ can be interpreted as the corrections most

compatible with the prior of a strictly time-homogeneous evolution of the smile surface, given the posited dynamics and the constraint of exact recovery of the market prices. As we show below, these adjustment factors also provide a transparent quantification of the quality of the fits, that conveys more information than the usual chi-squared statistic.

4 Gaining Intuition

The problem of reconciling the prices of caplets and swaptions is not trivial because of the imperfect correlation between the forward rates, $\{f_i\}$, that make up a given swap rate, SR, and of the time-dependence of the instantaneous volatility of the forward rates. It is therefore important to gain an intuitive understanding of how these two quantities affect the volatility of swap rates. To this effect, let's first define the *terminal* decorrelation from time t_0 to time T between forward rates j and k, by $\widehat{\rho}_{jk}(T)$:

$$\widehat{\rho}_{jk}(T) = \frac{\int_{t_0}^{T} \rho_{jk}(t)\sigma_j(t)\sigma_k(t)dt}{\sqrt{\int_{t_0}^{T} \sigma_j^2(t)dt \int_{t_0}^{T_{\text{exp}}} \sigma_k^2(t)dt}}$$

where $\sigma_j(t)$ is the instantaneous volatility at time t of forward rate j, and $\rho_{jk}(t)$ is the instantaneous correlation at time t between forward rates j and

k. Clearly, the terminal decorrelation will in general be lower than one even in the presence of perfect instantaneous correlation as long as the instantaneous volatility functions are not constants. We show below that terminal decorrelation is directly linked to the swap-rate volatility.

To see how this comes about, we begin by observing that a swap rate, SR_i, can always be expressed as a linear combination of forward rates:

$$SR_i = \sum_{k=1,n_i} w_{ik} f_k$$

where w_{ik} are suitable 'weights', defined, eg, in Rebonato (2004), and n_i signifies the number of forward rates in the ith swap rate, SR_i. Next, consider for simplicity the case of a lognormal process for the swap rate. (This assumption is purely made for ease of exposition.) Under lognormality, the Black formula (Jamshidian, Neubeger) provides the pricing for a European swaption. The volatility input for the Black formula is given by the 'implied' (in the lognormal case, root-mean-squared) volatility for the i-th swap rate:

$$[\sigma_{Black}^{SR_i}]^2 T_{exp} = \int_0^{T_{exp}} \sigma_{inst}^{SR_i}(u)^2 du \tag{7}$$

where T_{exp} is the expiry time of the European swaption in question, $\sigma_{inst}^{SR_i}(t)$ is the instantaneous volatility of the associated swap rate at time t, and $\sigma_{Black}^{SR_i}$ is the market-implied volatility for the swaption associated with swap rate SR_i.

If one makes a joint log-normal assumption for all the forward rates and for the swap rate (Rebonato (1999)), a straightforward application of Ito's lemma then links the instantaneous volatility of a given swap rate with the instantaneous volatilities, $\{\sigma_i(t)\}$, of all the underlying forward rates:

$$\sigma_{inst}^{SR_i}(t)^2 = \sum_{j,k=1,n_i} \zeta_j(t)\zeta_k(t)\rho_{jk}(t)\sigma_j(t)\sigma_k(t) \tag{8}$$

with

$$\zeta_j^i(t) = \frac{w_{ij}(t) + \sum_{k=1,n_i} f_k(t)\frac{\partial w_{ik}}{\partial f_j}}{\sum_{m=1,n_i} w_{im}(t) f_m(t)} \tag{9}$$

and

$$w_{ij}(t) = \frac{P_t^{T_j}}{\sum_{j=1,n_i} P_t^{T_j}\tau_j} \tag{10}$$

where $P_t^{T_j}$ is the time-t price of a discount bond expiring at time T_j and τ_j is the tenor of the jth forward rate.

Rebonato and Jaeckel (1993) explain why it is a good approximation to assume that in the expression above the new weights $\left\{\zeta_j^i(t)\right\}$ and the forward rates can be treated as deterministic with a value equal to their realization today, ie:

$$\sigma_{inst}^{SR_i}(t)^2 \simeq \sum_{j,k=1,n_i} \zeta_j^i(t_0)\zeta_k^i(t_0)\rho_{jk}(t)\sigma_j(t)\sigma_k(t) \tag{11}$$

Then the implied (root-mean-squared) swap-rate volatility, $\sigma_{Black}^{SR_i}$, to use in the Black formula to price the ith European swaption will be obtained using as:

$$[\sigma_{Black}^{SR_i}]^2 T = \int_0^{T_{\text{exp}}} \sigma_{inst}^{SR_i}(t)^2 dt = \sum_{j,k=1,n_i} \zeta_j(t_0)\zeta_k(t_0) \int_0^{T_{\text{exp}}} \rho_{jk}(t)\sigma_j(t)\sigma_k(t)dt$$

This clearly shows the importance of the covariance elements Cov_{jk}:

$$Cov_{jk} = \int_0^{T_{\text{exp}}} \rho_{jk}(t)\sigma_j(t)\sigma_k(t)dt \tag{12}$$

It is clear that the value of the covariance element will depend both on the correlation between the two forward rates and on the time dependence the the volatility functions. Indeed, Equation (12) shows the link with the terminal decorrelation defined above, and explains how the time dependence the the volatility functions affect the pricing by altering the terminal decorrelation. However, the time dependence of the instantaneous volatilities affects the pricing of swaptions in two distinct ways, as explained below.

The first effect is shown in Fig 3, that depicts the time dependence of the same (time homogeneous) volatility function in Fig 1 during the life of a 3 x 3-year semi-annual European swaption. The expiry of the swaption coincides with the expiry of the first forward rate only. However, by option expiry all the other forward rates will still be 'alive' and will have traversed different portions of their lives. Therefore the covariance elements between the various forward rates (that entail integrals out to time T_{exp}) will in general depend on the shape of the volatility function during the time to expiry (consider, for instance, $Cov_{1,5}$).

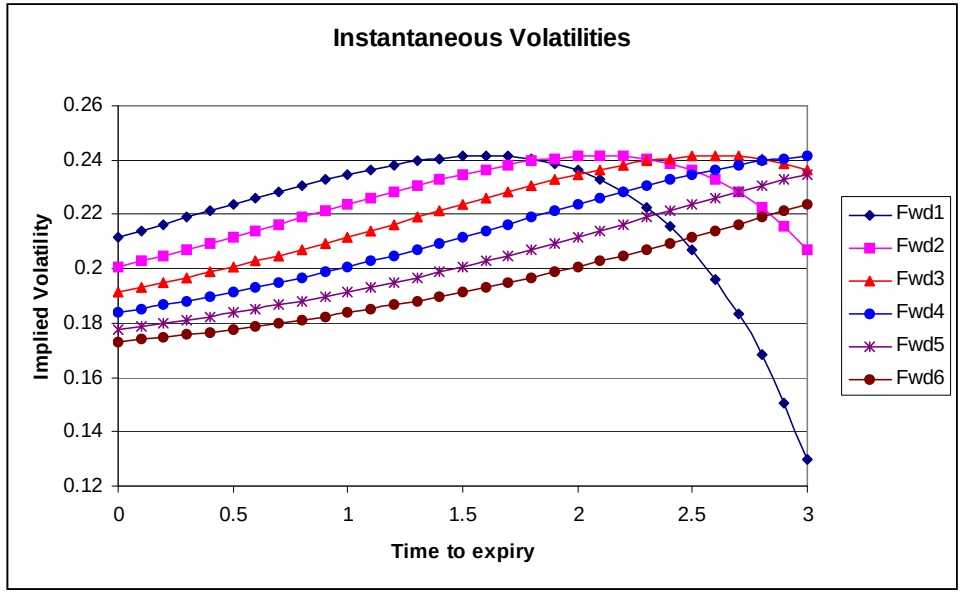

Figure 3: The case of a 3 x 3 year swaption

It is clear by a simple generalization of Schwartz's inequality, that, for any two portions of the instantaneous volatility functions with given root-mean-squared values over the life of the option, the covariance element (12) will attain its highest possible value when the two functions are constant *over the integration period*. Therefore, within the limits of the approximations and assumptions above, the root-mean-squared volatility of the swaption will almosty always be below the value, $\sigma^{SR_i}_{Black,\mathsf{max}}$, given by

$$[\sigma^{SR_i}_{Black,\mathsf{max}}]^2 = \sum_{j,k=1,n_i} \zeta^i_j(t_0)\zeta^i_k(t_0)\widehat{\sigma}^0_j\widehat{\sigma}^0_k \tag{13}$$

where the quantities $\widehat{\sigma}^0_j$ and $\widehat{\sigma}^0_k$ are the root-mean-square volatilities obtained when the instantaneous volatilities are constant over the integration period.†

†The root-mean-squared volatility of the swaption will almosty always be below the value, $\sigma^{SR_i}_{Black,\mathsf{max}}$, defined above for a short-dated option if the instantaneous volatility function were increasing.and the option expiry short.

In general, for fixed root-mean-squared caplet volatilities, the terminal decorrelation, and hence the covariance elements, and hence the swaption prices will therefore depend on the precise time dependence of the instantaneous volatilities. **This is the first way time dependent volatilities affect swaption prices**.

To understand the second effect alluded to above, consider (again in a Black world) a caplet with implied (root-mean-squared volatility), $\widehat{\sigma}_j\left(T_j\right)$, given by

$$\widehat{\sigma}_j\left(T_j\right) = \sqrt{\frac{1}{T_j}\int_0^{T_j}\sigma(u,T_j)^2 du} \tag{14}$$

Note that if the instantaneous volatility of the jth forward rate were constant, $\sigma(t,T_j)=\sigma_j^0$, then the root-mean-squared volatility would be independent of the upper integration limit in Equation (14):

$$\widehat{\sigma}_j\left(\tau\right) = \sqrt{\frac{1}{\tau}\int_0^{\tau}\left(\sigma_j^0\right)^2 du} = \widehat{\sigma}_j\left(T_j\right), \quad \forall\tau \;\; 0\leq\tau\leq T_j \tag{15}$$

Now, for a given swaption all but one of the underlying forward rates will still be 'alive' by swaption expiry.

Consider now the case of a short-expiry (3y) into a long-tail (7y) swaption. Fig 4 shows two possible instantaneous volatilities with the same root-mean-squared volatility over the full life of the last forward rate in the swap rate, ie, the 10-year forward rate.During the 3 years of the life of the swaption, however, the 10-year forward rate will experience a very different volatility in the two cases (around 8% in one case, and around 12% in the other), despite the fact that both functions are compatible with the market price of the 10-year caplet. This shows that the prices of a swaption obtained with two time-dependent volatility functions in Fig 4 (that price the caplet identically) can be very different.

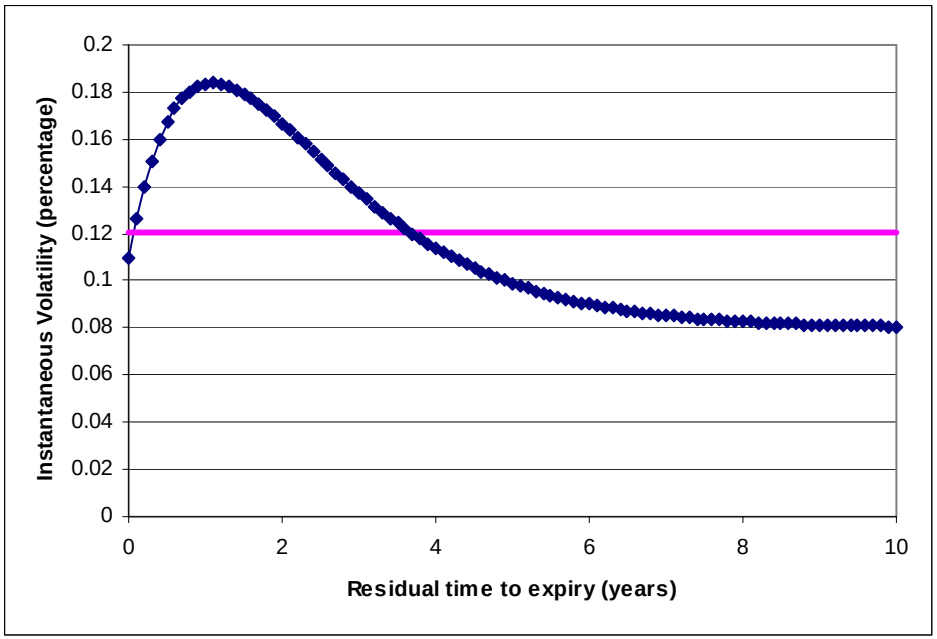

Figure 4: Two possible instantaneous volatilities with exactly the same root-mean-squared value out to caplet (but not swaption !) expiry.

So, **two instantaneous-volatility-related effects** (the 'Schwartz-inequality' effect for the covariance element, and the way a non-constant volatility is apportioned over the life of the option) **can affect the swaption price**.

Of course, the impact on the resulting swaption prices of these two effects will depend on the precise interplay between the time to swaption expiry, the length of the underlying swap rate and of the parameters (mainly d and c) of the $g()$ function.

It is for these combined reasons that commenting on the congruence of the caplet and swaption markets will in general depend on analyzing both the instantaneous volatilities of, and the correlation among, the underlying forward rates. In particular, without analyzing in detail the time dependence of the parametrized function $g()$, it is impossible to say *a priori* whether instantaneous correlation or time dependence of volatilities will have a bigger effect on the volatility of a swap rate.

The results above are obvious in the Black (log-normal) world discussed in this section. The intuition (although not the details) remain valid in the more complex case of the LMM-SABR model.

5 Quality of the Calibration to Market Caplet Prices

The LMM-SABR model was first calibrated to caplet prices for each trading day in the data set above using two different values of for the exponent β, $\beta = 0$ and $\beta = 0.5$. The results of the calibration are reported by showing the initial values, $k_0^{T_i}$, of the process $k_t^{T_i}$, (Figs 5 and 6) and the constants $\xi_i^{T_i}$ (Figs 7 and 8) for the two values of the exponent β. This shows both the goodness of the time-homogeneity assumption and the aptness of the chosen functional form. (Recall that,once the adjustment factors $\left\{k_t^{T_i}\right\}$ are applied, the market SABR caplet prices are recovered exactly by construction.)

We note that the vast majority of the $\left\{k_t^{T_i}\right\}$ and $\left\{\xi_t^{T_i}\right\}$ quantities are very close to 1. The only two forward rates for which there is an appreciable difference

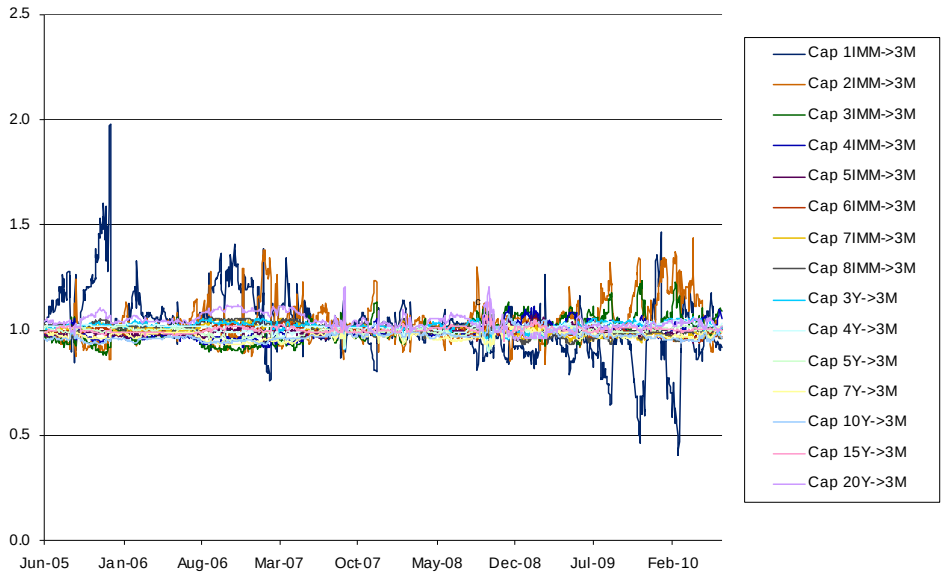

Figure 5: The initial value, $k_0^{T_i}$, of the process $k_t^{T_i}$ for the case of $\beta = 0$

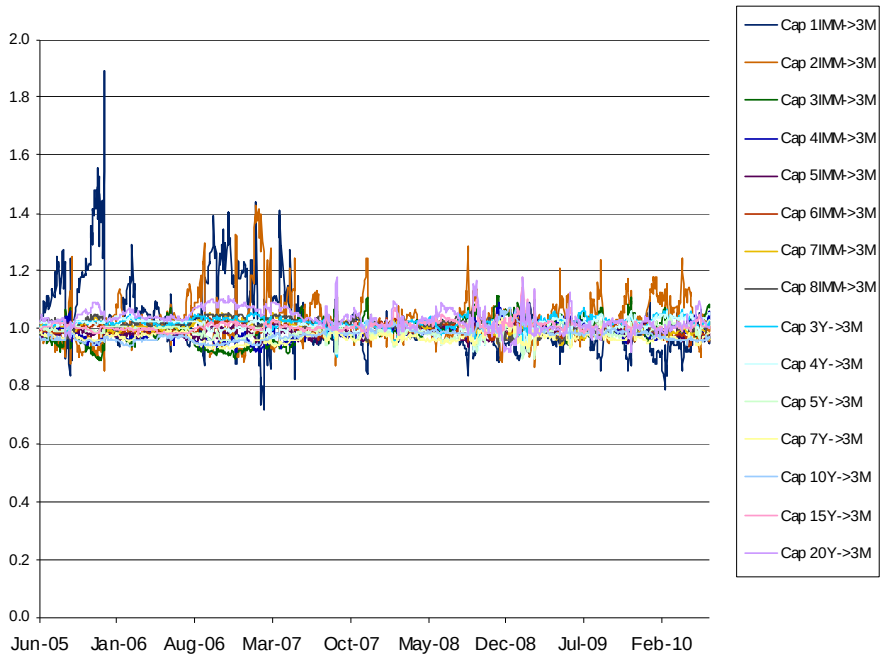

Figure 6: Same as Fig 4 for $\beta = 0.5$

from 1 are the first and, to a lesser extent, the second IMM (futures) contracts. As a result, on most trading days the market prices are approximately compatible with the time homogeneous assumption that underpins our analysis.

The largest discrepancies are encountered for the shortest-expiry caplets. This is not surprising, because in periods of market excitations the time homogeneity assumptions is not satisfied, and this affects most strongly the shortest-expiry forward rates, ie, the forward rates whose expiries fall during the short-lived period of market excitation.

The magnitude of the adjustment factors $\left\{\xi_t^{T_i}\right\}$ required to make the volatility of volatility match the shape of the market smile for all expiries is greater than for the factors $\left\{k_t^{T_i}\right\}$. However, for most days the deviations from 1 (the value that indicates perfect time homogeneity) are less than 20%.

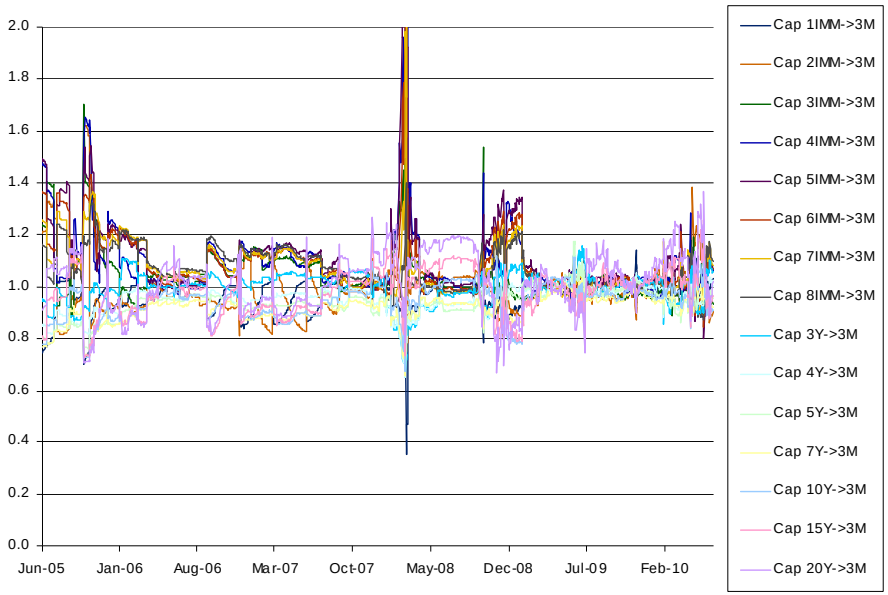

Figure 7: The factors $\xi_t^{T_i}$, for the case of $\beta = 0$

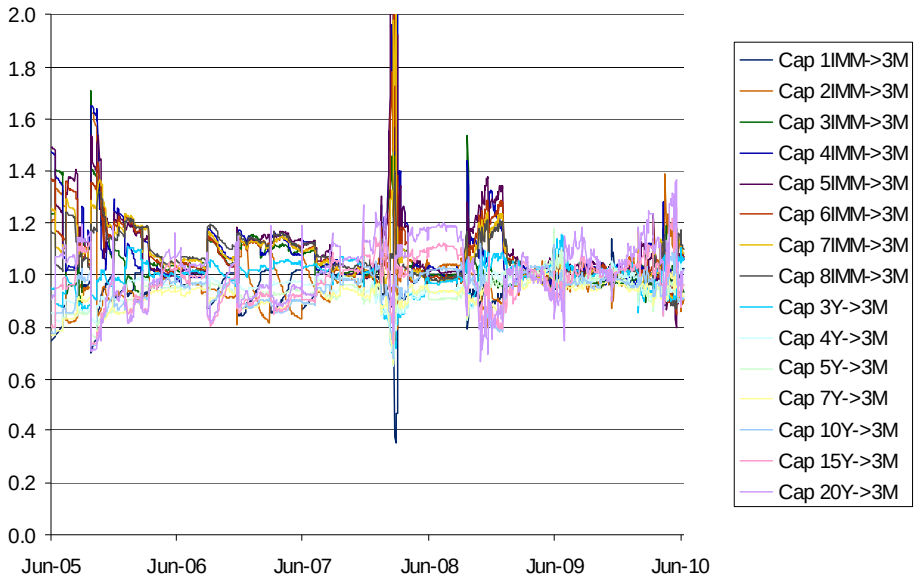

Figure 8: The factors $\xi_t^{T_i}$, for the case of $\beta = 0.5$

We therefore consider the calibration successfully carried out and, with the parameters obtained as described above, we finally are in a position to begin to explore the congruence of the caplet and swaption markets.

6 Results I: Time-Homogeneous $g()$ and $h()$ Functions

To calculate the model swaption prices we use the approximate but accurate formulae reported in Rebonato and White (2010), that allow the calculation of the prices of swaptions implied by a set of forward-rate parameters for the LMM-SABR model without having to use a Monte Carlo simulation. In this approach, the swaption price is produced by obtaining the SABR parameters for the swap rate process

$$dSW_t = (SW_t)^{\beta} \boldsymbol{\Sigma}_t dz_t \tag{16}$$

$$\frac{d\boldsymbol{\Sigma}_t}{\boldsymbol{\Sigma}_t} = V dw_t$$

given the parameters of the forward-rate LMM-SABR process. The LMM-SABR forward-rate parameters were chosen to match for each trading day the observed SABR market prices under the time-homogeneity *desideratum*, as described in the previous section. (We show in Appendix 1 the accuracy of these approximations.)

As the derivation of the expressions of the swaption SABR parameters, Σ_0 and V, is somewhat lengthy, we simply report the formulae employed for ease of reference:

$$\Sigma_0 = \sqrt{\frac{1}{T}\sum_{i,j}\left(\rho_{ij}W_i^0 W_j^0 k_0^i k_0^j \int_0^T g_t^i g_t^j dt\right)} \tag{17}$$

$$V = \frac{1}{\Sigma_0 T}\sqrt{2\sum_{i,j}\left(\rho_{ij} r_{ij} W_i^0 W_j^0 k_0^i k_0^j \int_0^T g_t^i g_t^j \widehat{h}_{ij}(t)^2 t dt\right)} \tag{18}$$

with

$$\widehat{h}_{ij}(t) = \sqrt{\frac{1}{t}\int_0^t h^i(s)h^j(s)ds} \tag{19}$$

and

$$W_k^t = w_k \frac{\left(f_t^k\right)^\beta}{(SR_t)^\beta} \tag{20}$$

and refer to Rebonato and White (2010) for a justification. The quantities ρ_{ij} and r_{ij} denote the correlations among the forward rates and among their volatilities, respectively.

We stress that, for reasons explained in the following, all these correlations were set to 1 in our study unless otherwise stated. As we shall see, doing so makes the results we obtain even stronger.

We have examined in our study a number of combinations of expiries and underlying swap lengths ("tails"). Without great loss of information we can limit the presentation of the results to the case of short and long expiries into short and long tails (where, in both cases, 'long' and 'short' indicate 10 and 2 years, respectively). We begin by looking at the values of the at-the-money (ATM) implied volatilities.

In the case of short tails, we find that the agreement is very good throughout the five years in our data, including the exceptionally turbulent period of September 2008 both for short (Figures 9 and 10) and long expiries (Figures 13 and 14). The results obtained with the exponent $\beta = 0$ perform somewhat better than $\beta = 0.5$.

In the case of long tails, the agreement is also very good up until the beginning of the recent financial crisis (Summer 2007) but then becomes very poor in September 2008. See Fig 11, 12, 15 and 16. After this date for long tails the model values remain consistently below the market prices. Note that we obtained the model values with a correlation of 1 both for ρ and for r. As we discussed below, a more realistic correlation structure would have lowered the model prices even more. We therefore consider our result as correlation-boundary cases.

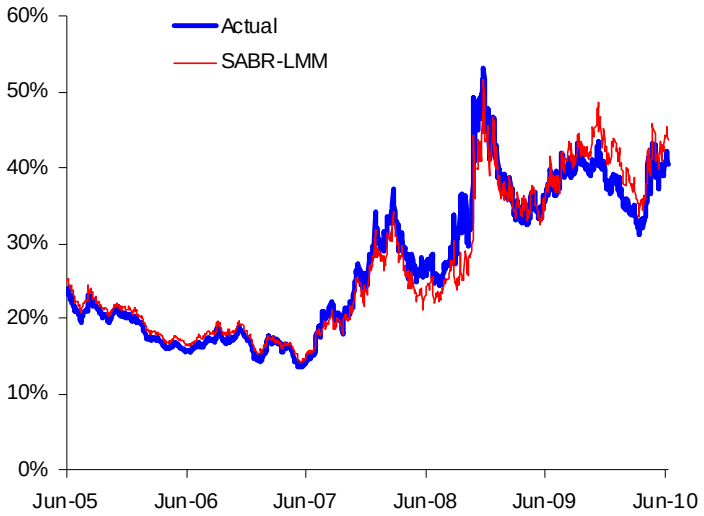

Figure 9: Market and model swaption ATM vol, 2Y x 2Y, $\beta = 0$

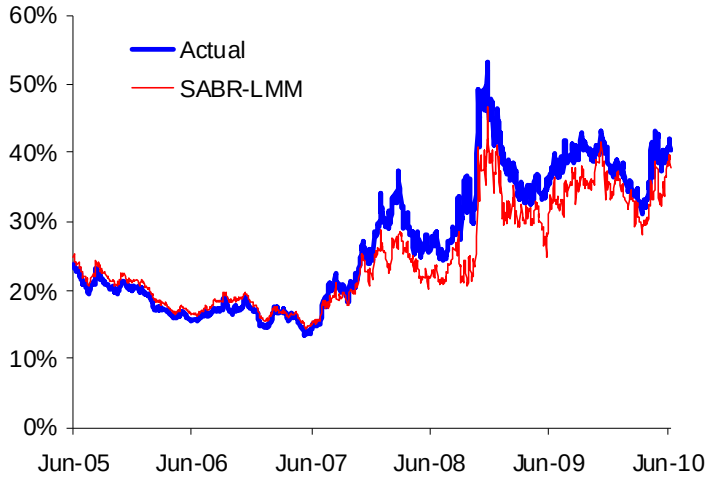

Figure 10: As above, ATM vol, 2Y x 2Y, $\beta = 0.5$

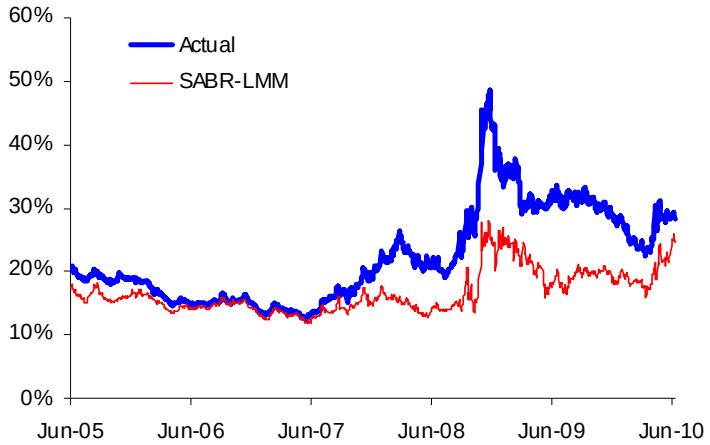

Figure 11: As above, ATM vol, 2Y10Y, $\beta = 0$

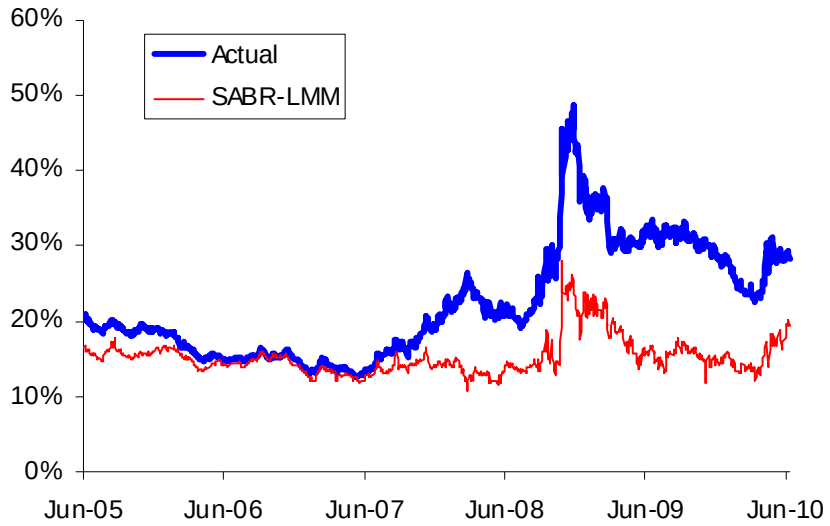

Figure 12: As abive, ATM vol, 2Y10Y, $\beta = 0.5$

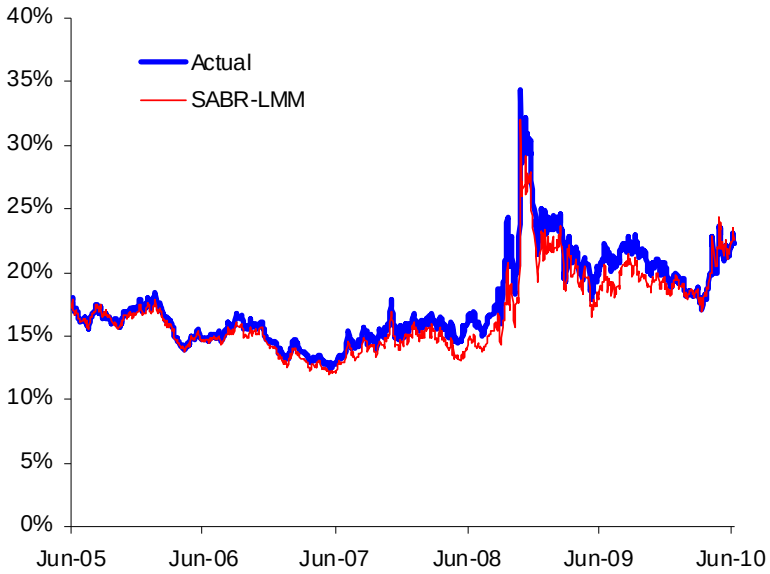

Figure 13: As above, ATM vol, 10Y x 2Y, $\beta = 0$

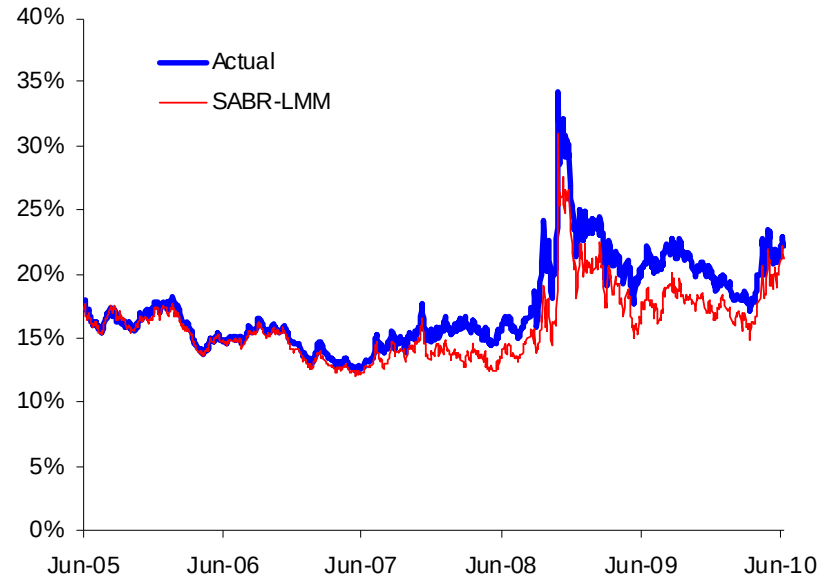

Figure 14: ATM vol, 10Y2Y, beta=0.5

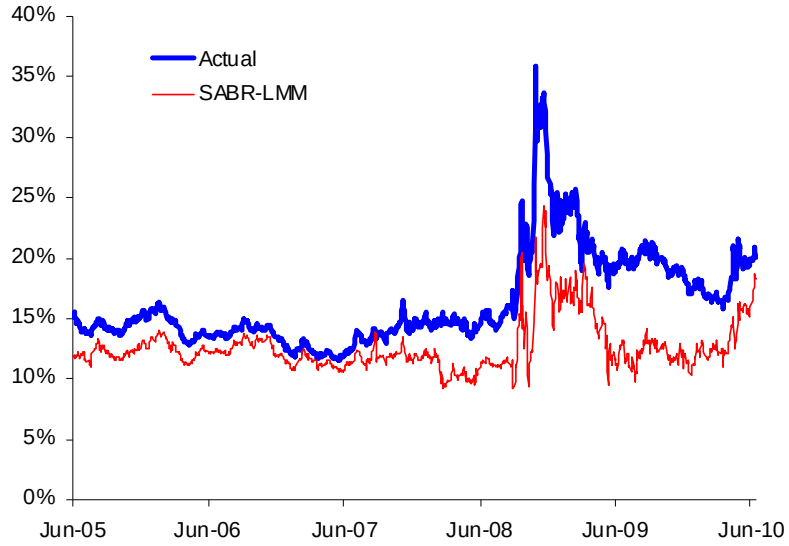

Figure 15: As above, ATM vol, 10Y x 10Y, $\beta = 0$

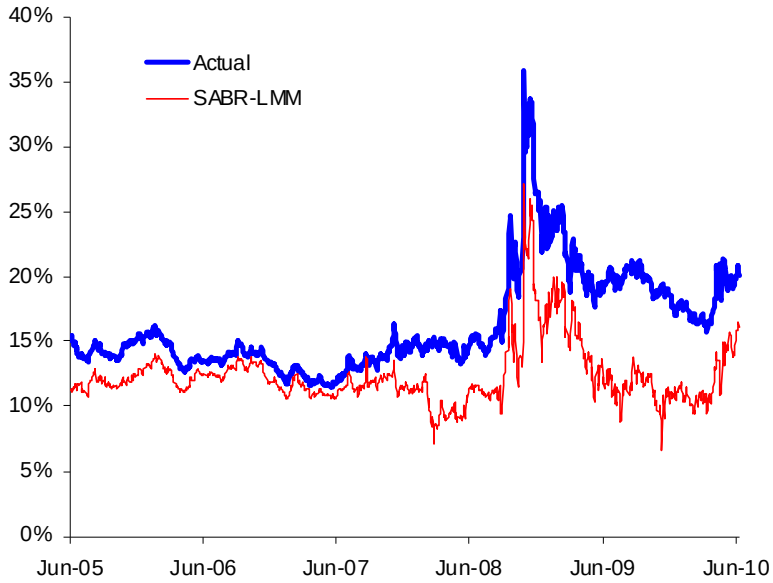

Figure 16: As above, ATM vol, 10Y x 10Y, $\beta = 0.5$

These results indicate

- that the agreement between model and market prices is always better for the exponent $\beta = 0$ than for $\beta = 0.5$;

- that the agreement is somewhat better for short expiries into short tails than or long expiries into long tails;

- that the agreement tends to break down during turbulent periods for long tails (and especially so for short expiries into long tails).

An explanation for these findings can be offered along the following lines.

To explain why we obtain better results with $\beta = 0$, we point to work by Rebonato (2003) that clearly shows that, for rate levels between approximately 2% and 6%, the dependence of swaption implied volatilities on the swap rate level is better explained by a normal (as opposed to lognormal) behaviour. This is also corroborated by we work by Rebonato, de Guillaume and Pogudin (2010) that explores the dependence of realized (as opposed to implied) volatility as a function of the level of rates. Also in this work for the level of swap rates encountered in the present study a normal behaviour (that corresponds to $\beta = 0$) accounts best for empirical data.

In order to understand why during excited periods long-tail swaptions may be underpriced under the time-homogeneous assumption that we have made so fat, consider Figure 17, which shows two fits to the $g()$ function, one obtained in normal market conditions and one during a period of turmoil.As one would expect, the instantaneous volatility function fitted during the excited period assumes much higher values for short times to expiries. What is more surprising is that it reaches much *lower* values for long times to expiry, as shown in Fig 17.

This is as an inescapable consequence of the time-homogeneity assumption. Consider in fact the pricing of a long-tail swaption during a period of market excitation. Assume that the market expects the turbulence to subside (and hence volatilities to decline) after a period of time of weeks or months[‡]. The

[‡]The study by White and Rebonato (2009) on the two-state Markov chain volatility model mentioned above indicates that the half-lives in the excited state typically range from a few weeks to several months.

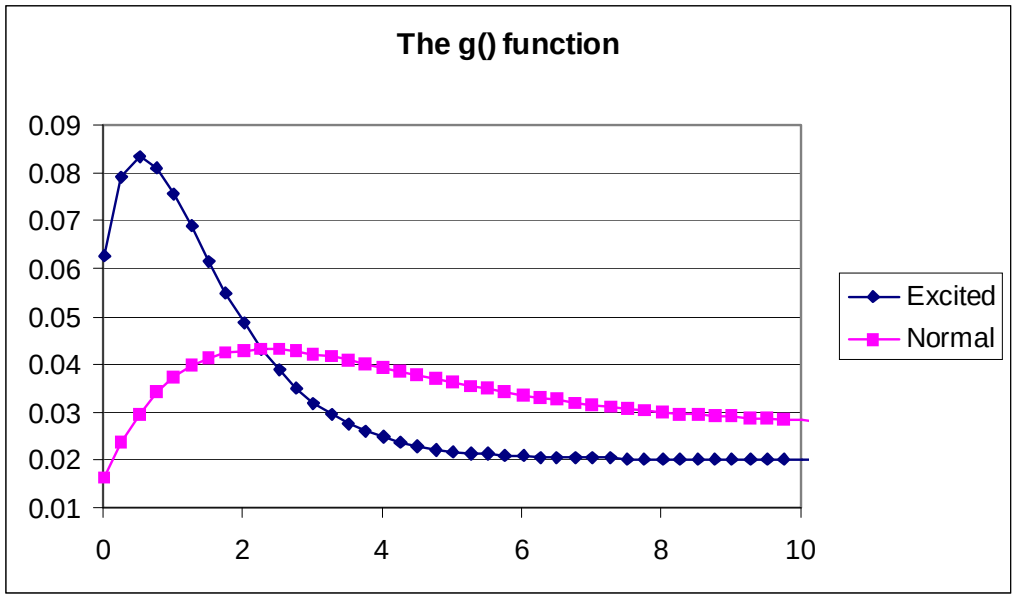

Figure 17: Normal and excited instantaneous volatility functions fitted on two different days.

time-homogeneity assumption, however, implies that the future smile surface will remain the same over time. In order to fit at the same time the observed short-dated (exceptionally high-volatility) option and the long-dated (almost-normal-volatility) options, a time-homogeneous model must therefore imply that all forward rates experience very high volatility when their expiries are short, and *extremely low volatility when they have a long time to expiry*.

The high instantaneous volatility at the short end is required to price correctly the short-dated and short-tail caplets. In order for long-expiry options to be priced correctly (at a close-to-normal implied volatility level), however, the instantaneous volatility has to decline very sharply for the correct root-mean-squared volatility to obtain.

It is now easy to see how this has a direct impact on long-tail options – the more so if the expiry is short – during excitation periods. Consider, for instance, the 2 x 10 year swaption. When the time homogeneous volatility function is fitted to the caplet market prices for an excited day, during the two years of the option life most of the underlying forward rates experience in the model the very low instantaneous volatility shown in Fig 17. This certainly contributes to the mispricing shown in Figs 11 an 12. Indeed, note that the discrepancy appears as soon as the market enters in a (transitory) phase of excitation (Summer of 2007), and reaches a maximum in the after-Lehman period. As markets then slowly return to normal, the discrepancy begins to narrow, and almost disappear.

The strength and the limitations of the time-homogeneity assumptions are clearly shown in Fig 18, which shows the ATM implied volatilities of caplets as a function of expiries for all the trading days in our data set from June 2005 to July 2010. Most of the times the time homogeneity assumption is well justified, as the term structure of ATM volatilities maintains over extended periods a remarkably self-similar shape.

However, the exceptional events of the 2007-2009 credit crisis show that the market expectation of where long-expiry volatilities will be relative to short-expiry level of volatility are not always the same.

7 Results II: Flat $g()$ and $h()$ Functions

The analysis presented above indicates that the inadequacy of the time-homogeneous assumption during periods of market excitation can account for some of the systematic underpricing of swaption implied by the LMM-SABR model calibrated to caplets with $g(t,T) = g(T-t)$.

However, it is apparent that *even during normal market conditions* the model prices for long-dated, long-tail swaptions tend to be lower than the corresponding market prices. Recall that this is true even for prices obtained with perfect instantaneous correlation. Any lower (and econometrically more realistic) value for the instantaneous correlation would produce an even greater discrepancy between model and market prices. See, for instance, the results presented in Fig 19 with the flat-volatility caplet calibration (which is, of course, independent of

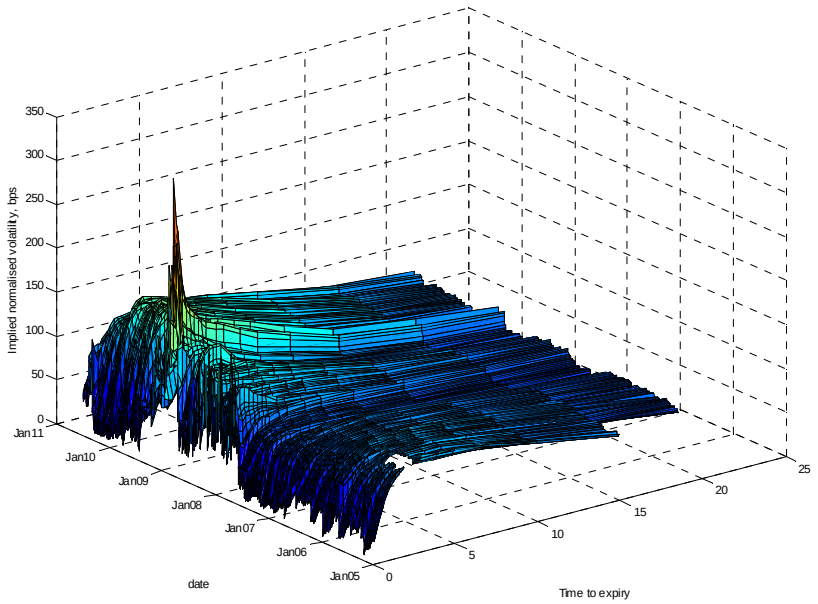

Figure 18: History of instantaneous volatilities fitted to caplet prices over the historical period explord in this study.

the correlation among forward rates and among volatilities) using either with perfect correlation or a 'realistic' correlation among forward rates (upper curve) for the 10y x 10y case. It is clear that the more realistic imperfect correlation calibration gives rise to an even greater degree of mispricing.

We therefore try to look more deeply at how the market may price swaptions relative to caplets.

Within the LMM-SABR set of no-arbitrage models, the natural way to increase the model prices of swaptions, while still recovering the prices of the market caplets and retaining absence of arbitrage, is to modify the shape of the instantaneous volatility functions in such a way as to keep their root-mean-squared unchanged. Recall from the discussion above that, for any given instantaneous correlation function, the maximum swaption volatility will be obtained with constant instantaneous volatilities (ie, when the $g()$ function assumes the functional form $g(t,T)=d$).

In order to see whether the swaption prices can be at all reconciled with the caplet prices, we therefore explore in this section the case of flat $g()$ and $h()$ functions, while retaining the perfect correlation assumption.

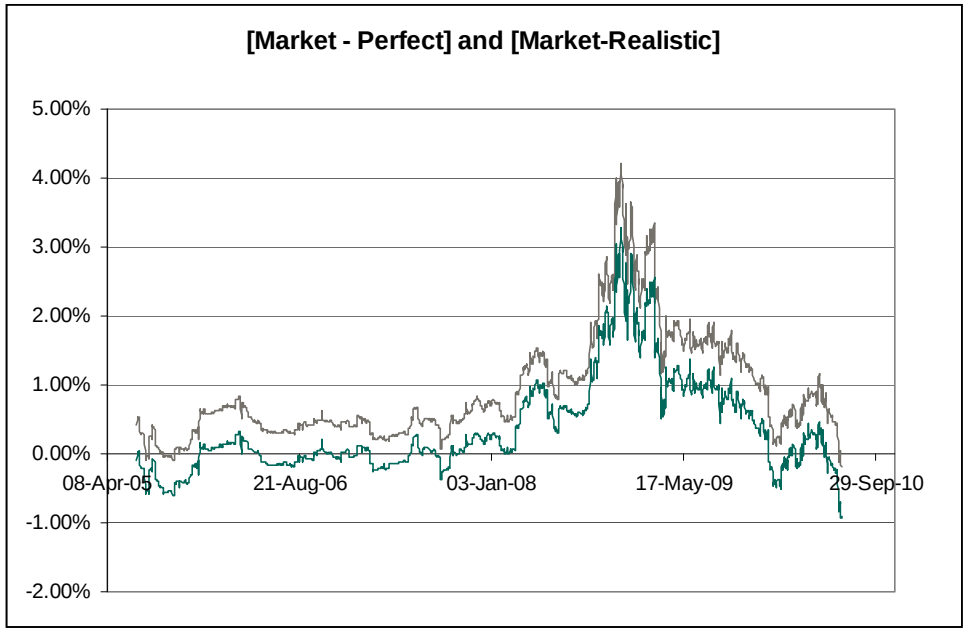

Figure 19: Difference between the market and model implied volatilities for the perfect correlation case, [Market - Perfect] (lower curve) and the realistic correlation case [Market - Realistic] for the 10y x 10 y swaption.

When these choice are made for the correlation and for the $g()$ and $h()$ functions, the pricing of swaptions can be profitably analyzed with reference to the pricing of **early-stopping caplets**, defined below.

Early-stopping caplets are the mirror image of forward-starting caplets and swaptions.

Consider a swaption with expiry time T_{exp} and strike K, and the forward rates that make up the reference swap rate of the swaption. The expiry of the first forward rate will coincide with the expiry of the swaptions, but all the other forward rates will expire at times $T_{exp} + i\tau$, $i, 1, 2, ..., n_i$, and will therefore still be 'alive' by swaption expiry. An early-stopping caplet settles for payment at time $t_{exp_i} < T_{exp} + i\tau$. Consider then the case when $t_{exp_i} = T_{exp}$, ie, when all the early-stopping caplets expire at the same time as the swaption. Then it is trivial to prove the following. (See Johnson and Nonas (2009) for a more general proof of the results below. The simple results presented in the following can be obtained by successive application of the triangular inequalities presented in Johnson and Nonas (2009)).

Proposition 1 *If the market price of a swaption expiring at time T_{exp} is greater than or equal to the price of a same-strike early-stopping cap made of the the underlying early-stopping caplets all with expiry T_{exp}, then there is a model-independent arbitrage profit that can be extracted with a static super-replication strategy. The strategy to realize the arbitrage profit is to sell the swaption and buy the early-stopping cap.*

The proof is straightforward. By the assumption, setting up the strategy either costs nothing or provides a positive cash flow at time t_0. Then note that at expiry time, T_{exp}, there is no time value left either in the swaption or in the early-stopping cap. We therefore only have to look at the intrinsic value.

Two situations can arise:

- either the yield curve is exactly flat, in which case the payoff from the swaption exactly equals the payoff from the early-stopping cap;

- or it is not flat, in which case the cap can only have the same or more value than the swaption (because of the the swap rate is a weighted average of forward rates, and therefore some caplets can be in the money even if the weighted average is not).

Note that, since only intrinsic values matter, the result is model-independent.

Proposition 2 *If*

i) *the instantaneous correlation between any two forward rates if equal to 1;*

ii) *the* $g()$ *function is flat,* $g(t,T) = d$*;*

iii) *the* $h()$ *function is flat:* $h(t,T) = \delta$*;*

iv) *the yield curve is flat; and*

v) *the term structure of volatilities is flat,*

then any model nested within the set of the LMM-SABR no-arbitrage conditions produces the same price for a swaption and an early-stopping cap.

The proof is also straightforward: if the yield curve is flat the swap rate is equal to any of the forward rates. Furthermore, with perfect instantaneous correlation, flat $g()$ and $h()$ functions and flat term structure of volatilities the volatility of the swap rate is identical to the volatility of the cap. The cap and the swaption must therefore have the same price.

Finally, we have the last result.

Proposition 3 *If any of the conditions in Proposition 2 are not met, the model-independent, arbitrage-free price of a swaption will always be lower than the price of the associated early-stopping cap.*

The result directly follows from Propositions 1 and 2 above[§].

The simple results above simply show that an early-stopping cap must be worth at least as much as or more than the associated swaption, but give no indication as to how much the cap will be more expensive when any of the conditions in Proposition 2 fail to be met. In practice we observe that reasonable variations

[§]More constructively, the proof also follows from recognizing i) that a cap is a portfolio of options and a swaption is an option on a protfolio of forward rates; ii) that with less-than-prefect instantaneous correlation or with time-dependent volatilities the swap rate volatility will be lower than the cap volatility; and iii) that if the term structure of volatilities is not flat the swap rate volatility will be lower than in the flat-term-structure-of-volatilities case. The price of the swaption must therefore be lower than the price of the early-stopping cap.

in the term structure of volatilities produce very small divergencies between the prices of the swaption and of the early-stopping cap. For similarly strong deviations from a flat yield curve the discrepancies are found to be greater.

Why does this matter for swaption pricing?

Let's look at long-expiry swaptions. We note that for these long expiries the forward portion of the term structure of rates (the portion that matters for the pricing of a swaption) can often be very flat for long-expiry swaptions. When this is true it is then reasonable to expect that the price of an early-stopping cap will be very close to the price of the associated swaption. This is indeed borne out in practice in Figs 20 and 21, that show the prices of a swaption obtained by calibrating the forward rates to caplets with flat volatilities and with perfect correlation, both for the 2y x 2y and for the 10y x 10y swaption. Note that for the second series (where indeed we expect the forward yield curve to be flatter) the two curves are most of the times virtually indistinguishable. The maximum difference, reached during the extreme post-Lehman turmoil, remains smaller than one vega.

With these results in mind, we can now turn to the market and model prices of swaptions obtained with perfect correlation and with flat $g()$ and the $h()$ functions that price all the caplets correctly. See Figures 23 and 24 for the two 'corner' cases, ie, for the 2 x 2 year and 10 x 10 year European swaptions. Similar results were obtained for other expiries and tails.

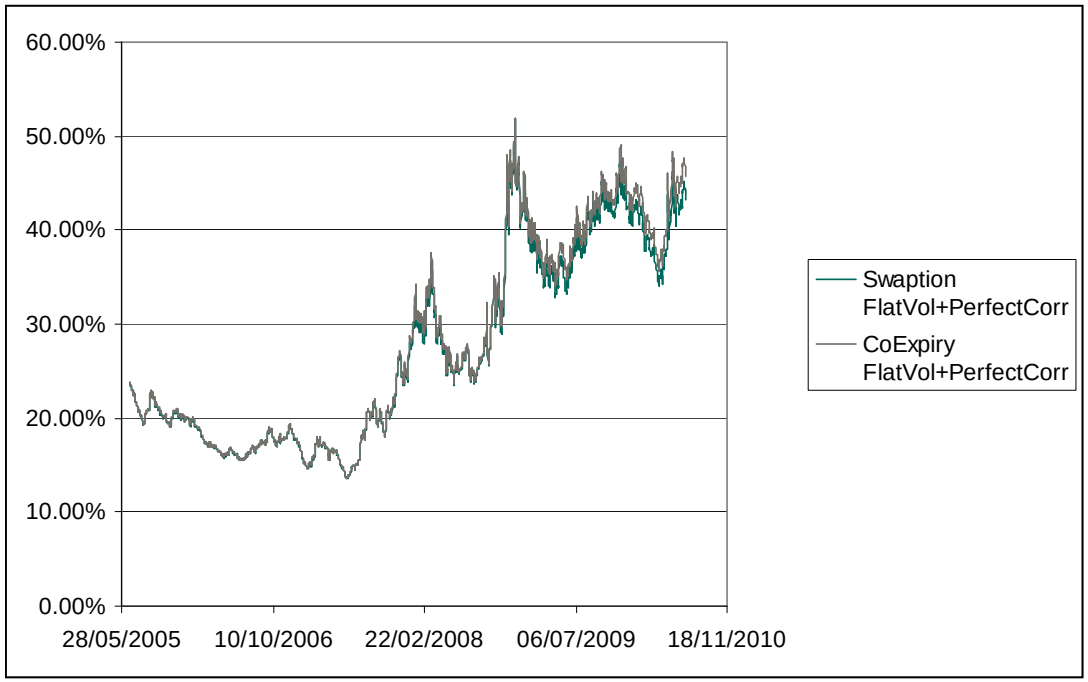

Figure 20: The prices of a 2 x 2 perfect-correlation swaption with flat volatilities and the price of an early-stopping (Co-Expiry) cap, also priced with flat volatilities.

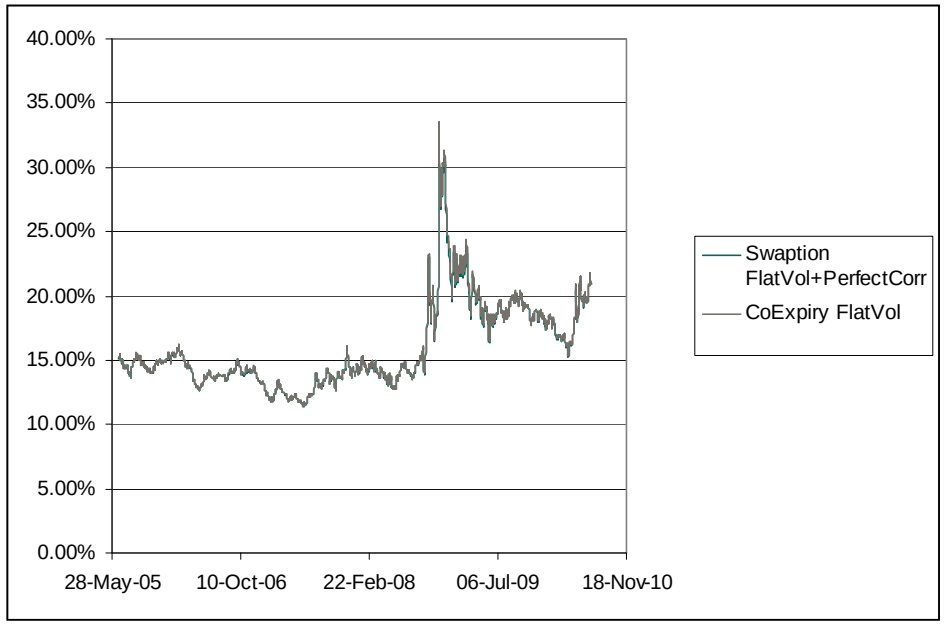

Figure 21: Same as above for the 10y x 10y swaptions.

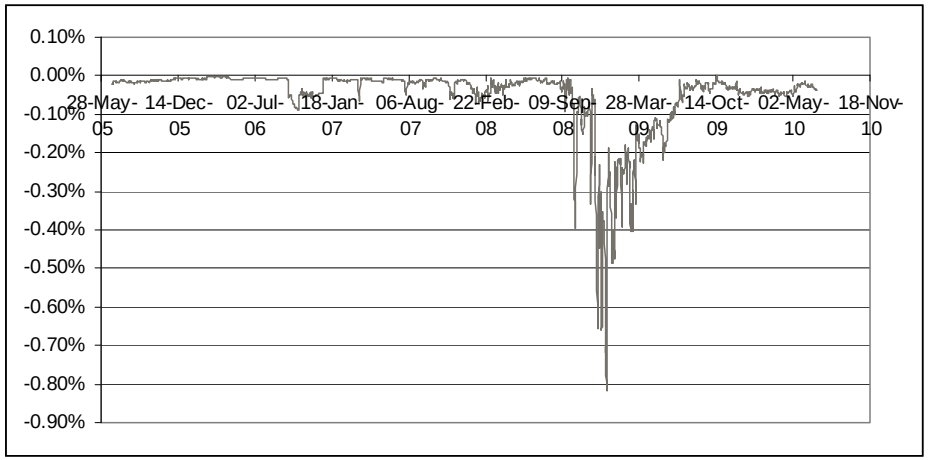

Figure 22: The difference between the time series in Fig 21

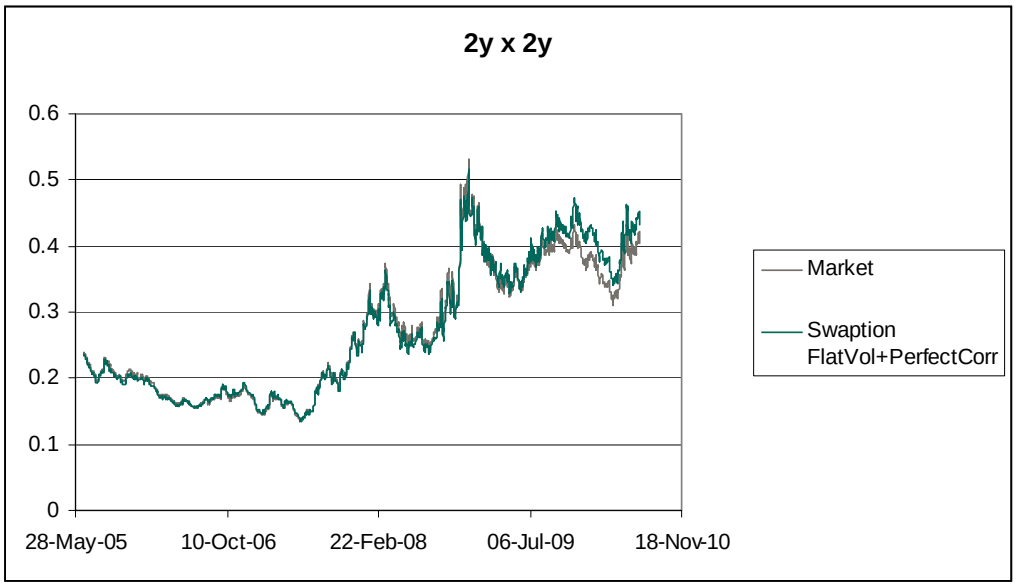

Figure 23: Market and model prices of 2y x 2y swaptions obtained with perfect correlation and flat volatilities.

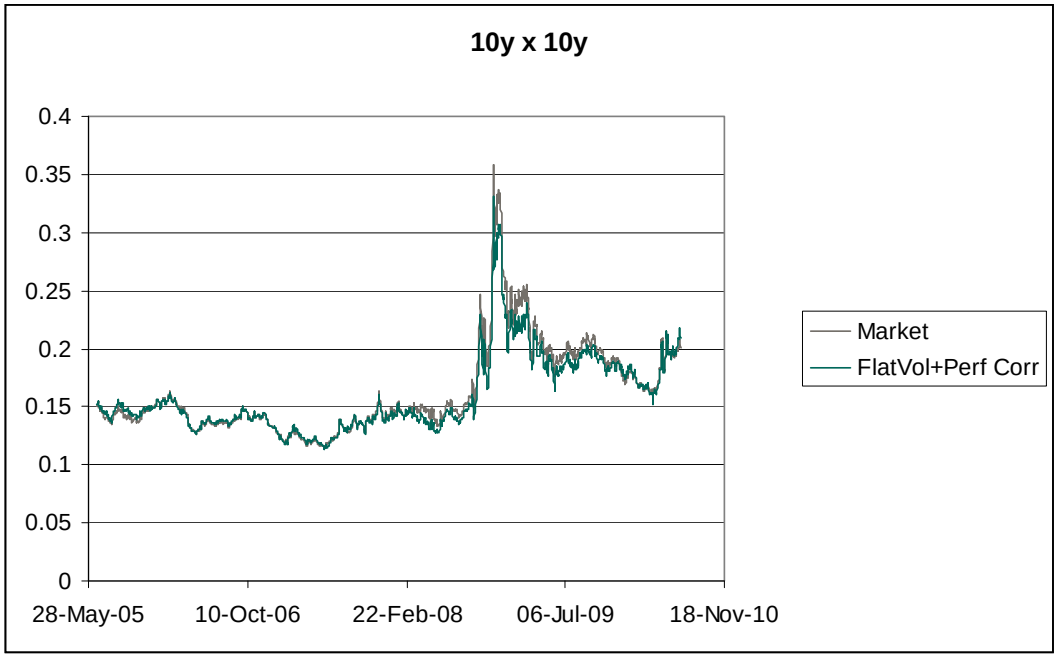

Figure 24: Same as above for the 10y x 10y swaptions

The first observation is that we now obtain extremely close agreement between model and market prices of swaptions during normal and excited periods, both for long and short tails and for long and short expiries.

Recall, however, that we also obtained extremely close agreement between the perfect-correlation, flat-volatility swaption prices and the price of an early-stopping cap (which constitutes a model-independent no-arbitrage boundary).

We therefore observe that, within the class of the LMM-SABR arbitrage conditions, swaptions trade in the market *very close to the model-independent no-arbitrage value of the associated super-replicating early-stopping cap*.

8 Implication of the Market Pricing

We have provided a convincing story of how the market may be pricing swaptions.

We have remarked that this pricing does not expose the trader to model-independent arbitrage.

However, pricing swaption *at* this boundary, as the market does, has some unpleasant financial implications.

This is most easily understood in terms of the specific parametrization of the LMM-SABR no-arbitrage conditions described above. Indeed, in order to obtain such a close agreement between the market and the model prices while pricing

caplets correctly, the correction factors, $\left\{k_t^{T_i}\right\}$, must now be very different from 1, indicating that time homogeneity is fully lost, both in normal and excited times.

The financial consequences of this loss of time homogeneity can be easily appreciated by looking at Figs 26 to 30, which display the future terms structure of volatilities implied by the model calibrated to caplets with flat instantaneous volatilities.

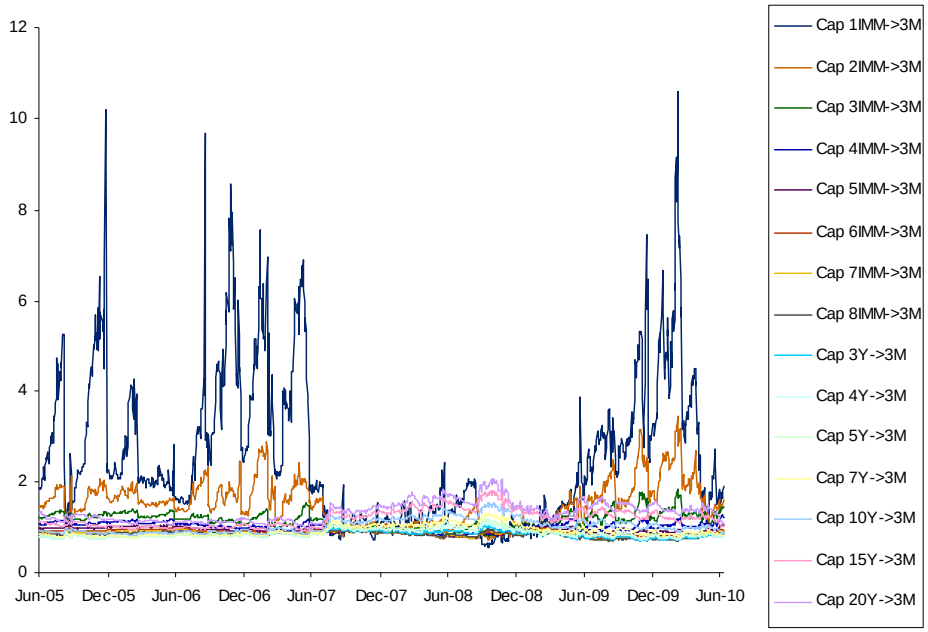

Figure 25: The adjustiments factors $\left\{k_t^{T_i}\right\}$ obtained with flat volatilities. This figures should be compared with Figure 5, that shows the same adjustment factors obtained with a time homogeneous $g()$ function. Note that with flat volatilities the adjustment factors $\left\{k_t^{T_i}\right\}$ are now one order of magnitude larger.

This calibration clearly implies a radical change of the term structure of volatilities towards shapes that have never been observed in the past, as can be seen by comparing the Figs 26 to 30 with Fig 18, which show a real-world time series for the same quantities.

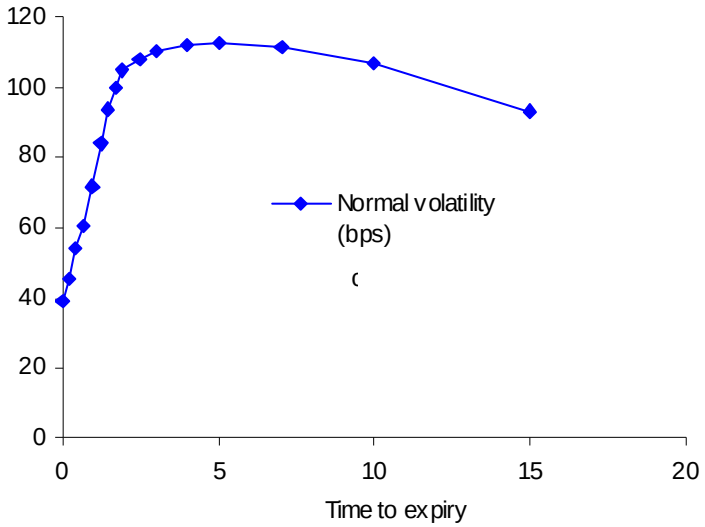

Figure 26: Time-0 term structure of ATM volatilities.

As long as today's prices are recovered, why should one worry about the realism

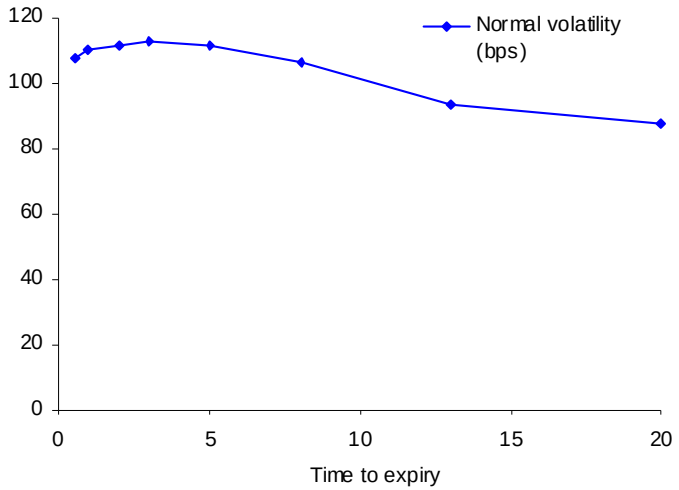

Figure 27: The same quantity in two years' time

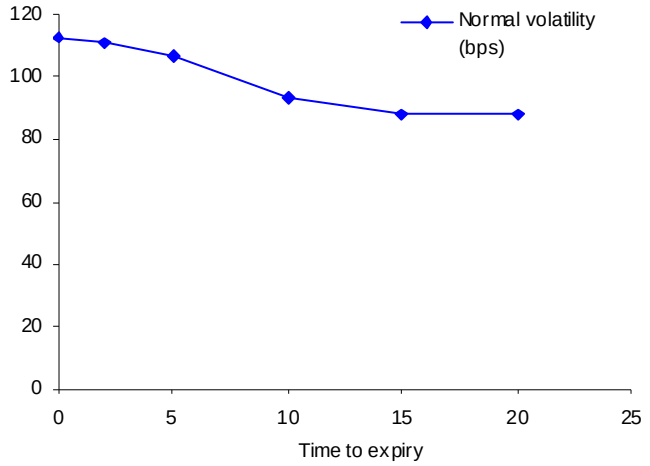

Figure 28: As above in 5 years

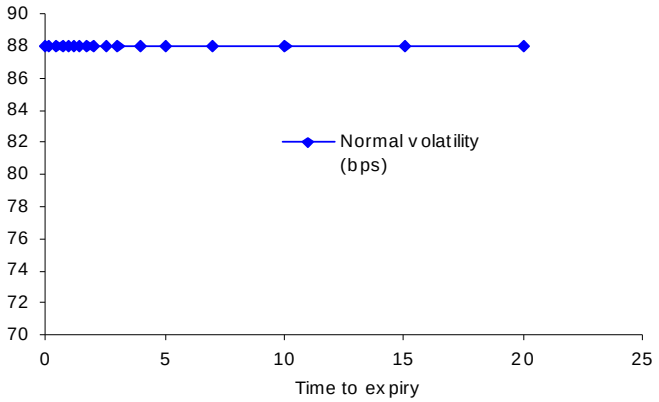

Figure 29: As above in 15 years

of the future term structure of volatilities (and, more generally, the future smile surface)? The reason is that, indeed, the future evolution of the smile surface does not affect today's prices, but it will fully determine future prices of plain-vanilla options, and will therefore affect future vega re-hedging costs. It is therefore important to choose a calibration that will produce future implied volatility surfaces as close as possible to what will be realized in the future, as failure to do so would predict today incorrect future vega re-hedging costs.

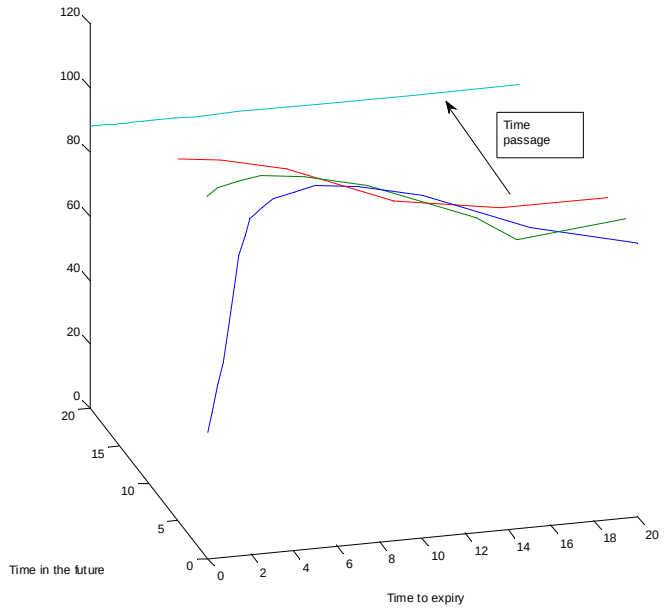

Figure 30: Evolution of the term structure of ATM volatilities for the case of falt instantaenous volatilities.

The effect is not small: a flat $g()$ function implies for the example shown in the figures above that in two years' time a trader will be able to sell a 1-month expiry caplet, that today trades at an implied volatility of 40 basis points, for an implied volatility of approximately 110 basis points. According to the model the same trader will also be able in fifteen years' time to buy for 88 basis point implied volatility a caplet that now trades close to 120 basis points.

In sum: our empirical analysis shows that the market prices of swaptions trade close to a boundary value that is attained only if very special modelling conditions (about the instantaneous correlation and the shape of the $g()$ and $h()$ functions) are met. However, these modelling requirements would imply an evolution of the term structure of volatilities that is totally ad odds with what has been observed over many years (see Fig 18), and that implies for a trader future re-hedging costs very different from what she is likely to encounter in reality.

Given this disconnect between statistically and financially justifiable prices on the one hand, and market prices on the other, why don't swaptions trade even higher?

We have pointed out that the flat-volatility swaption prices are often very close to the prices for an early stopping cap. This occurs when the forward part of the yield curve is close to flat, as it is often in practice the case for long-expiry swaptions.

We have also shown that in this case a model-independent static super-replicating strategy can in theory be put in place using early-stopping caplets. These are far from being liquid instruments, and swaption traders do not regularly use them in their hedging. However, traders are likely to keep these very easy-to-obtain boundary prices in the back of their minds as a 'sanity check', beyond which swaptions become 'too expensive'.

There is a market precedent for this: it is known, in fact, that in the case of other static replication strategies that *could* be put in place in practice (such as the replication using plain-vanilla calls of a double-barrier knock-out option), the strategy is rarely actually deployed in practice; however, the replication construction is extensively used as an alternative way to obtain a price for the option and to ensure that the trader may not expose herself to model-independent arbitrage.

Indeed, we have noted above that in the case of swaptions from time to time even the very special early-stopping-cap no-arbitrage boundary is exceeded, but that, when this happens, reversion to the boundary tends to be very swift.

9 Conclusions

We have looked at the differences between the market prices and the model prices for US$ European swaptions obtained using the LMM-SABR model. The model was calibrated exactly to caplets by using both realistic humped instantaneous volatility functions, and by using 'flat volatilities'. We also discussed and investigated the pricing impact of imposing a perfect or imperfect correlation among rates.

We found that, during normal market conditions, the agreement between market prices and model prices obtained using the realistic volatilities (and perfect instantaneous correlation) was fair, especially for short expiries. The discrepancy become larger, however, during periods of market excitation, and for larger expiries. We provided an explanation for this. We observed that introducing a mor realistic correlation structure among forward rates made the agreement between market and model prices worse.

Better, and indeed excellent, agreement for all expiries and tails and during both normal and excited periods was obtained using flat volatilities (and, again, perfect instantaneous correlation). We have shown that these model prices are very close to the model-independent no-arbitrage boundary (that we have expressed in terms of early-stopping caps). A first conclusion is therefore that the market appears to trade close to the no-arbitrage boundary.

This conclusion raises however some important questions. We have also shown in fact that the flat-volatility assumption required to obtain good swaption pricing produces very unrealistic evolutions for the future term structure of ATM volatilities.

As the volatility and correlation 'markets' are incomplete, this implies that the future re-hedging costs cannot be 'locked-in'. This raises the question as to why traders appear to endorse these unrealistic choices about instantaneous volatilities and correlations (and hence about future term structures of volatilities): to the extent that these implicit model assumptions about the future smile surface are not borne out in practice, a trader would in fact expose himself to the risk of hedging losses.

The reason is probably to be found in the difficulty in carrying out this arbitrage: for instance, even after vega hedging, caplets and swaptions have very different gamma profiles. Rebonato (2004) shows that, in order to mitigate these gamma mismatches, rather complex strangle-versus-straddle trades have to be carried out, which bring in a dependence on caplet and swaption smiles. And, of course, even if a trader were right, she would be exposed to the risk that, before option expiry, the market may move even more out of line with 'fundamentals, thereby causing a negative mark to market of her portfolio and forcing the trader to close her position. The persistence of this arbitrage would therefore be one instance of the limit-to-arbitrage phenomenon discussed in Shefrin (199X).

So, the data do not show any tendency for the market prices to converge to the prices that a financially appealing calibration (with imperfect correlation and non-flat volatilities) would suggest. On the contrary, market prices of swaptions at times even briefly exceed the price of a super-replicating portfolio of early-starting caplets.

On the other hand, when swaption market prices exceed the no-arbitrage boundary discussed above, we have shown that the reversion to the no-arbitrage boundary becomes swift.