Interest Rate Volatility Surface Diffusion Models and their Applications

Johann Power
Faculty of Economics, University of Cambridge
June, 2025

Abstract

This paper outlines how multi factor Linear Gaussian Models (MFLGMs) can be used to diffuse interest rate volatility surfaces and provides an in-depth theoretical derivation of how to calibrate such diffusion models using volatility ratios. This derivation is novel in that we have not seen it explicitly outlined in the literature and so had to reconstruct it however, calibrating to surface shapes is not a novel technique and calibrating to volatility ratios in particular is already used in industry. Although, this paper introduces some new additions to the calibration process including the construction of a moving-average reference volatility surface and weighting calibration points based on their importance score as defined by their squared reconstruction error from using a variational autoencoder (VAE). We then present an example calibrated volatility surface given an input volatility surface and methods to diffuse volatility surfaces over time. Finally we apply the MFLGM model to vega surface construction via a novel technique which includes the construction of a PCA score Jacobian matrix to back-out implied parameter shifts and thus their impact on a portfolio's P&L given a day's volatility surface change with ridge regularization added to deal with noisy market data. This gives us an interpretable explanation of how interest rate volatility surface movements impact a portfolio's P&L and where on the surface a portfolio's sensitivity to interest rate volatility lies.

1. Introduction: Swaptions and Implied Volatilities

We first briefly recap the financial products this paper works with. A swaption provides its holder with the option, without obligation, to enter into an interest rate swap at a predetermined strike rate K at some future date T. These are like equity options but with the underlying asset being an interest rate swap. Swaptions are categorized into payer and receiver swaptions, defined as follows:

- **Payer swaption:** right to pay fixed and receive floating.
- **Receiver swaption:** right to receive fixed and pay floating.

Under the Bachelier (normal) model, the price P_N of a payer swaption is (Henrard, 2014):

$$P_N = e^{-rT}\Big[(S_0 - K)\,\Phi(d) \;+\; \sigma_N\sqrt{T}\,\phi(d)\Big], \qquad d = \frac{S_0 - K}{\sigma_N\sqrt{T}}. \tag{1}$$

S_0 is the current forward interest rate swap rate, σ_N is the normal volatility, T is the time to expiration, r is the discount rate, and $\Phi(\cdot)$, $\phi(\cdot)$ denote the standard normal cumulative distribution and probability density functions respectively (proof in Appendix A.1).

1.1. Interest Rate Swaps

An interest rate swap (IRS) is an agreement between two counterparties to exchange interest rate payments over a predefined period, based on a notional amount. Typically, one party pays a fixed interest rate, while the other pays a floating interest rate indexed to a reference rate such as SOFR (for USD).

The valuation of an IRS at initiation (assuming no arbitrage and frictionless markets) is zero. The fixed rate is thus determined to equalize the present value of fixed payments against the present value of floating payments. This implies the fixed rate (swap rate) S_0 satisfies:

$$N\sum_{i=1}^{n} D(0,t_i)\,\tau_i\big(F(0;t_{i-1},t_i) - S_0\big) = 0 \tag{2}$$

Solving for S_0 gives:

$$S_0 = \frac{\sum_{i=1}^{n} D(0,t_i)\,\tau_i\,F\big(0;t_{i-1},t_i\big)}{\sum_{i=1}^{n} D(0,t_i)\,\tau_i} \tag{3}$$

where:

- $D(0,t_i) = e^{-\int_0^{t_i} r(s)ds}$ is the discount factor to time t_i,
- $\tau_i = t_i - t_{i-1}$ is the accrual period,
- $F(0;t_{i-1},t_i)$ is the forward rate observed at time 0 for the period $[t_{i-1},t_i]$.

At a later valuation date, the present value $V_{IRS}(0)$ of an existing swap with fixed rate K is the difference between the fixed and floating legs:

$$V_{IRS}(0) = N\Big(\underbrace{\sum_{i=1}^{n} D(0,t_i)\,\tau_i\,F(0;t_{i-1},t_i)}_{\text{Floating Leg PV}} - \underbrace{K\sum_{i=1}^{n} D(0,t_i)\,\tau_i}_{\text{Fixed Leg PV}}\Big). \tag{4}$$

Using the property of the floating leg under standard market conventions:

$$\sum_{i=1}^{n} D(0,t_i)\,\tau_i\,F\big(0;t_{i-1},t_i\big) = 1 - D(0,t_n) \tag{5}$$

we can simplify this formula to:

$$V_{IRS}(0) = N\big(1 - D(0,t_n) - K\sum_{i=1}^{n} D(0,t_i)\,\tau_i\big) \tag{6}$$

1.2. Implied Normal Volatility

To extract the implied normal volatility σ_N, given a market swaption price P_{mkt}, we solve (1) numerically (Brigo & Mercurio, 2006a). This inverse problem typically requires iterative numerical techniques such as the Newton-Raphson method:

$$\sigma_N^{(n+1)} = \sigma_N^{(n)} - \frac{f\big(\sigma_N^{(n)}\big)}{f'\big(\sigma_N^{(n)}\big)} \tag{7}$$

where:

$$\begin{aligned} f(\sigma_N) &= P_N(\sigma_N) - P_{\text{mkt}}, \\ f'(\sigma_N) &= \frac{\partial P_N}{\partial \sigma_N}. \end{aligned} \tag{8}$$

Explicitly, the derivative with respect to volatility σ_N is:

$$\frac{\partial P_N}{\partial \sigma_N} = e^{-rT}\,\sqrt{T}\,\phi(d) \tag{9}$$

1.3. Volatility Surfaces and Cubes

Market-implied volatilities for swaptions are typically presented in matrices or surfaces, characterized by different option expiries and underlying swap tenors. Such a structure is called a volatility surface. Extending this notion, when volatility also depends on strike rates, volatility is a function of three dimensions creating a 4D tensor, known as a volatility cube. Volatility cubes facilitate capturing the volatility smile or skew. Although, this paper will focus just on the at-the-money (ATM) volatility surface, i.e. only considering 0 moneyness points along the strike rate dimension of the volatility cube to form the 3D volatility surface.

2. The Multi-Factor Linear Gaussian Model of Volatility Surfaces

We now outline an industry-standard parametric model for interest rate volatility surface diffusion. The multi-factor Linear Gaussian Model (MFLGM) is a model of interest-rate processes (short rates, forward rates, swap rates). These interest rates tend to move together across maturities which MFLGM aims to capture in a lower-dimensional parsimonious manner. The implied-volatility surface is then a derived cross-sectional output of this model. The term-structure of interest rates is an *infinite-dimensional* object because, at each time t, the instantaneous forward rate $T \longmapsto f(t,T), \quad T \geq t$, must be specified for a continuum of maturities T. In practice we observe only a finite grid of tenors, but as the underlying curve lives in a function space it becomes “infinite-dimensional.” Attempting to model $f(t,T)$ for every T leads to a stochastic partial differential equation in an infinite-dimensional state space, which is intractable both analytically (no closed-form solutions) and numerically (requiring discretization of an unbounded system).

2.1. Model Definition

MFLGMs achieve tractability by projecting the curve onto a finite set of n factors. Concretely, one introduces an n-dimensional Markov state vector factor process $X(t) \in \mathbb{R}^n$ under the risk-neutral measure $\mathbb{Q}$, which drives all interest rates:

$$dX(t) = A\,X(t)\,dt + \Sigma\,dW^{\mathbb{Q}}(t), \tag{10}$$

where:

- $A \in \mathbb{R}^{n\times n}$ is the mean-reversion matrix,
- $\Sigma \in \mathbb{R}^{n\times n}$ is the volatility matrix,
- $W^{\mathbb{Q}}(t)$ is an n-dimensional Brownian motion under $\mathbb{Q}$ (Heath et al., 1992; Brace et al., 1997).

A risk-neutral measure is a probability measure under which all asset price processes when discounted by the appropriate numeraire become martingales. Under no-arbitrage, all zero-coupon bond prices $P(t,T)$ and thus forward rates are *affine* (linear + exponential) functionals of $X(t)$ (Brace et al., 1997; Brigo & Mercurio, 2006a).

The forward curve is then reconstructed by

$$f(t,T) = f(0,T) + \sum_{i=1}^{n} \phi_i(T)\, X_i(t) \tag{11}$$

where the $\phi_i(T)$ are deterministic loading functions.

2.2. PCA Interpretation of the Loading Functions $\phi_i(T)$

In practice one often constructs the LGM loading functions $\phi_i(T)$ via a principal-component analysis (PCA) of historical rate curves. The PCA analysis is done by:

- **Eigen-curves:** Collect a time series of discrete forward (or par-swap) curves $\mathbf{f}(t) = \big(f(t,T_1),\dots,f(t,T_K)\big)^\top$. Compute its sample covariance $\mathbf{C}$ and solve $\mathbf{C}\,\mathbf{v}_i = \lambda_i\,\mathbf{v}_i, \quad i = 1,\dots,K.$ Each eigenvector $\mathbf{v}_i \in \mathbb{R}^K$ is called an *eigen-curve*, and λ_i its eigenvalue (variance explained).
- **$\boldsymbol{\phi_i(T)}$**: The continuous-maturity loading function $\phi_i(T)$ is obtained by interpreting the entries of $\mathbf{v}_i$ at $T_1,\dots,T_K$ as samples of a smooth curve, then interpolating. Thus $\phi_i(T_j) = v_{i,j}$.
- **Eigenvalues vs. Loadings:** The *eigenvalues* λ_i are scalars measuring how much variance mode i explains. The *eigen-curves* $\phi_i(T)$ are the deterministic shapes (loadings) for factor i.
- **PCA Scores $\boldsymbol{X_i(t)}$:** Projecting each observed curve onto $\mathbf{v}_i$ yields the time-series $X_i(t)$, the PCA *scores*. In the LGM one then models

  $$dX_i(t) = -k_i\,X_i(t)\,dt + \sigma_i\,dW_i(t), \quad i = 1,2,3,$$

  and reconstructs (11).

This formulation provides several key benefits:

1. **Time-homogeneity**: Because K and Σ are constant, the law of $X(t)$ depends only on time increments, simplifying calibration and simulation in “rolling” frameworks.
2. **Factor stationarity**: Each component X_i is an Ornstein–Uhlenbeck process (or Gaussian), ensuring mean-reversion to zero and a well-defined long-run distribution, which matches empirical observations of yield-curve behavior (Litterman & Scheinkman, 1991).
3. **No-arbitrage**: building all bond and swap prices from the same factors enforces internal consistency across maturities.
4. **Closed-form bond prices**: Under the risk-neutral measure, zero-coupon bond prices:

   $$P(t,T) = \exp\Big(-\int_t^T f(t,u)\,\mathrm{d}u\Big) \tag{12}$$

   are analytic because $\int_t^T \phi_i(u)\,X_i(t)\,du$ is Gaussian (proof in Appendix A.2.).
5. **Closed-form factor distribution**: One can also show:

   $$X(t) \;\sim\; \mathcal{N}\Big(e^{-K\,t}\,X(0)\,,\;\int_0^t e^{-K\,s}\,\Sigma\Sigma^\top\,e^{-K^\top s}\,\mathrm{d}s\Big) \tag{13}$$

   allowing instantaneous evaluation of transition densities.
6. **Economic interpretability and computational efficiency.** The loading functions $\phi_i(T)$ typically correspond to “level,” “slope,” and “curvature” movements of the yield curve, which traders and risk managers recognize. Having only n factors dramatically reduces the dimensionality of Monte Carlo or PDE computations, enabling faster computation for valuing a wide range of interest-rate derivatives (Brigo & Mercurio, 2006b).

By capturing the essential dynamics across numerous individual interest rates with just a small number of Gaussian factors, MFLGMs strike a balance between realistic curve behavior and analytical/numerical tractability. For this reason they are a common model for interest rate dynamics modelling in industry. Once the rate-model factors are calibrated, the model implies a normal (or log-normal) swaption volatility surface via the factor loadings.

2.3. Swap Rates under the Annuity Measure

As introduced in Section 1, a par swap is an agreement to exchange a stream of floating payments tied to short rates which vary over time, for a stream of fixed payments at a constant rate S for notional N such that the swap's value at inception is zero. At $t = 0$, the present-value (PV) of each leg is:

where $P(0,u)$ is the zero-coupon bond discount factor to time u (which decrease with time), $\{t_i\}_{i=1}^N$ are the fixed-leg payment dates between T and $T+M$ and $\tau_i = t_i - t_{i-1}$ are accrual fractions. Equating these PVs (so the swap has zero net value) gives:

Leg	**PV at** $t = 0$
Floating (receive)	$N\big[P(0,T) - P(0,T+M)\big]$
Fixed (pay)	$N\,S(0;T,M)\sum_{i=1}^{N}\tau_i\,P(0,t_i)$

Table 1. Present values of the floating and fixed legs of a par swap with notional N, start T, tenor M at $t = 0$.

$$\begin{aligned} N\big[P(0,T) - P(0,T+M)\big] &= N\,S(0;T,M)\sum_{i=1}^{N}\tau_i\,P(0,t_i), \\ \implies S(0;T,M) &= \frac{P(0,T) - P(0,T+M)}{\sum_{i=1}^{N}\tau_i\,P(0,t_i)} \\ &= \frac{P(0,T) - P(0,T+M)}{A(0;T,M)}. \end{aligned} \tag{14}$$

In the LGM we treat the swap rate $S(t;T,M)$ as a stochastic process. If we choose the swap annuity (i.e. the PV of one unit paid at each fixed date):

$$A(t;T,M) \;=\; \sum_{i=1}^{N}\tau_i\,P(t,t_i) \tag{15}$$

as our numéraire (the *annuity measure* $\mathbb{Q}^A$), then the ratio

$$\frac{P(t,T) - P(t,T+M)}{A(t;T,M)} = S(t;T,M) \tag{16}$$

has zero drift and is therefore a martingale. The annuity measure effectively ensures that at any t, the swap is repriced to be ATM. A *martingale* is a stochastic process $\{X_t\}$ adapted to a filtration $\{\mathcal{F}_t\}$ such that, for all $s \le t$, $\mathbb{E}[X_t \mid \mathcal{F}_s] = X_s$ i.e. once you know everything up to time t, your best guess for time $t+1$ is simply the value today — there's no drift up or down. Any martingale $M(t)$ obeys a SDE of the form:

$$dM(t) = (\text{volatility}) \;\times\; dW(t), \quad \text{no } dt \text{ term.}$$

Hence,

$$\begin{aligned} dS(t;T,M) &= \big\langle \tilde{\gamma}(t;T,M),\, dW^A(t)\big\rangle, \\ &= \sum_{i=1}^{n}\gamma_i(t)\,dW_i(t). \end{aligned} \tag{17}$$

is a purely driftless normal diffusion under $\mathbb{Q}^A$. To prove this, first let's define the forward swap rate $F(t)$. Start from the floating-leg present value under the M-maturing bond numéraire $P(t,M)$:

$$\mathrm{PV}_{\text{float}} = P(t,T) - P(t,T+M) \tag{18}$$

T is when the forward swap starts and M is its duration. We introduce the forward swap rate by normalizing to $P(t,M)$:

$$F(t;T,M) = \frac{P(t,T) - P(t,T+M)}{P(t,M)} \tag{19}$$

We know that, under the M-forward measure $\mathbb{Q}^M$ (the probability measure from taking the zero-coupon bond maturing at time M, $P(t,M)$, as your numéraire), the forward swap rate is a martingale so its SDE has no drift:

$$dF(t;T,M) = \big\langle \gamma,\, dW^M\big\rangle = \sum_{i=1}^{n}\gamma_i(t;T,M)\,dW_i^M(t) \tag{20}$$

Using (15) so that:

$$\begin{aligned} S(t;T,M) &= \frac{P(t,T) - P(t,T+M)}{A(t;T,M)} \\ &= \frac{P(t,M)}{A(t;T,M)}\,\frac{P(t,T) - P(t,T+M)}{P(t,M)} \\ &= \frac{P(t,M)}{A(t;T,M)}\,F(t;T,M). \end{aligned} \tag{21}$$

Evaluating the prefactor at $t = 0$ gives the constant scaling:

$$S(t;T,M) = \frac{P(0,M)}{A(0;T,M)}\,F(t;T,M) \tag{22}$$

Differentiating and using the SDE for F gives:

$$\begin{aligned} dS(t;T,M) &= \frac{P(0,M)}{A(0;T,M)}\,dF(t;T,M) \\ &= \sum_{i=1}^{n}\Big[\tfrac{P(0,M)}{A(0;T,M)}\,\gamma_i(t;T,M)\Big]\,dW_i^M(t). \end{aligned} \tag{23}$$

Switching to the annuity measure $\mathbb{Q}^A$ renames $W^M \to W^A$ and we set

$$\tilde{\gamma}_i(t;T,M) = \frac{P(0,M)}{A(0;T,M)}\,\gamma_i(t;T,M) \tag{24}$$

which rescales the raw factor-loading $\gamma(t;T,M)$ so that it matches the swap-rate units and making

$$dS(t;T,M) = \sum_{i=1}^{n}\tilde{\gamma}_i(t;T,M)\,dW_i^A(t) \tag{25}$$

a purely driftless normal diffusion under $\mathbb{Q}^A$. Note, the volatility and drift affects of the stochastic prefactor which were normalized away at $t = 0$ in (22) have been absorbed into the definition of the annuity measure so that S is just a constant times F, and all of the remaining diffusion lives in F (scaled by that constant).

To recap, the n Gaussian factors (e.g. ”level”, ”slope”, ”curvature”) are sufficient to explain movements in the interest rate term structure. $\gamma(t; T, M)$ is a n-dimensional vector representing how strongly each factor shakes the forward swap rate $F(t; T, M)$ under the M-forward measure. $\tilde{\gamma}(t; T, M)$ is the same idea but a vector representing the factor impacts on the par swap rate, $S(t; T, M)$, rescaling γ so that you move from the forward-rate numéraire to the annuity numéraire. Switching and rescaling simplifies the swap-rate model to a zero-drift Gaussian diffusion so that ATM normal volatilities are just the integrated norm of $\tilde{\gamma}$. As under the bond numéraire, the forward rate, $F(t; T, M)$, has zero drift, but the par swap rate, $S(t; T, M)$, still carries a deterministic drift term. But by using the annuity itself as numéraire, you guarantee that $S(t; T, M)$ is a martingale.

2.4. Implied Volatility Surface as a Derived Output

Under the annuity measure $\mathbb{Q}^A$, the swap rate has the driftless SDE:

$$dS(t) = \big\langle \tilde{\gamma}(t),\, dW^A(t) \big\rangle \tag{26}$$

Because there is no dt term:

$$\mathbb{E}\big[dS(t)\big] = 0, \qquad \mathbb{E}\big[dS(t)^2\big] = \|\tilde{\gamma}(t)\|^2\, dt.$$

Integrating these infinitesimal variances over $[0, T]$ yields the total variance:

$$\operatorname{Var}\big(S(T) - S(0)\big) = \int_0^T \|\tilde{\gamma}(s)\|^2\, ds \tag{27}$$

For a normal (Gaussian) model the ATM normal implied volatility, $\sigma_N(T, M)$, for expiry T and tenor M, is:

$$\sigma_N(T, M) = \sqrt{\int_0^T \|\tilde{\gamma}(s; T, M)\|^2\, ds} \tag{28}$$

With this result one can write the implied volatility results of a caplet or approximate the implied volatility of a swaption. By varying (T, M) we generate the full surface $\sigma_N(T, M)$. Crucially, once we fix the factor loadings $\gamma(\cdot; T, M)$, the shape of $\sigma_N(T, M)$ is entirely determined — it is not an additional input but a direct output of the rate model. By varying T and M over a grid you fill out a swaption volatility surface.

2.5. Stochastic Evolution of the Surface

In reality the entire vol surface moves over time. A simple way to capture this is to add a single, common level-shift $v(t)$:

$$\sigma_N(t; T, M) = \sigma_N(0; T, M)\, \exp\big(v(t)\big) \tag{29}$$

We model $v(t)$ as a mean-reverting Ornstein–Uhlenbeck (Uhlenbeck & Ornstein, 1930) process:

$$dv(t) = -\lambda_v\, v(t)\, dt + \sigma_v\, dZ(t),$$

where:

- $\lambda_v > 0$ makes $v(t)$ pull back toward zero (today’s surface).
- σ_v controls the *volatility of vol*, i.e. the extent to which the surface can shift.
- $Z(t)$ may be correlated with the rate-factor Brownian motions $W^A(t)$.

This construction ensures the underlying interest-rate model remains arbitrage-free and the implied volatility surface exhibits realistic, stochastic up-and-down moves around its initial shape.

3. Calibration to Implied Normal-Volatility Ratios

Using the MFLGM theory for diffusing volatility surfaces outlined in Section 2 this section outlines how to calibrate the model’s parameters by calibrating to volatility ratios. This method greatly simplifies the math and computation for calibration. Calibrating to surface shapes is not a new addition to the literature (Ahdida et al., 2015) and calibrating specifically to volatility ratios is already done in industry. However, there is a lack of theoretical derivation for this technique which this section will explicitly outline. This paper also adds some additional novel improvements to the model by constructing a smoothed reference volatility surface, factor reparameterization and weighted calibration. Having derived in Section 2 that under the annuity measure $\mathbb{Q}^A$:

$$\begin{aligned} dS(t; T, M) &= \big\langle \tilde{\gamma}(t; T, M),\, dW^A(t) \big\rangle, \\ \sigma_N(T, M) &= \sqrt{\int_0^T \|\tilde{\gamma}(s; T, M)\|^2\, ds}\,. \end{aligned} \tag{30}$$

we now explain how to fit a single mean-reversion speed parameter k in $\gamma(t; T, M)$ so that the model’s normal-vol surface exactly matches market ATM quotes.

3.1. Why Use Volatility Ratios?

Directly matching each market-implied normal volatility $\sigma_N^{\mathrm{mkt}}(T, M)$ one-by-one would require calibrating a separate scaling for every (T, M) forcing us to fit multiple interdependent terms (annuity, discount differences, swap-rate levels). The naive objective would be:

$$\min_{k} \sum_{T,M} \Big(\sigma_N^{\mathrm{mdl}}(T, M; k) - \sigma_N^{\mathrm{mkt}}(T, M) \Big)^2 \tag{31}$$

where, under our normal-vol framework,

$$\begin{aligned} \sigma_N^{\mathrm{mdl}}(T, M; k) &= \sqrt{\int_0^T \|\tilde{\gamma}(s; T, M; k)\|^2 \, ds}\,, \\ \tilde{\gamma}_i(s; T, M; k) &= \frac{P(0, M)}{A(0; T, M)} \, \gamma_i(s; T, M; k)\,. \end{aligned} \tag{32}$$

Here:

- $\tilde{\gamma}(s; T, M; k)$ already includes the deterministic $\frac{P(0,M)}{A(0;T,M)}$ scaling,
- $\gamma_i(s; T, M; k)$ are the only unknown functions (or parameters) to fit.

Even in this form, σ_N^{mdl} mixes together the unknown γ with the known annuity $A(0; T, M)$ and discount $P(0, M)$. We would have to fit a separate scaling for each (T, M) to absorb noise in the input curves. The optimization would become high-dimensional with one parameter (or function) per grid point leading to instability and over-fitting while correlations across tenors (via shared annuities and discount differences) make the objective poorly conditioned. Forming ratios between adjacent tenors cancels all of the known terms, leaving a target that depends only on how $\tilde{\gamma}$ (and hence k) changes with tenor. This yields a parsimonious, well-conditioned one-dimensional least-squares problem per expiry.

3.2. Derivation of the Volatility Ratio

Under the annuity measure $\mathbb{Q}^A$, the model-implied normal volatility is:

$$\begin{aligned} \sigma_N^{\mathrm{mdl}}(T, M; k) &= \sqrt{\int_0^T \Big\|\tilde{\gamma}(s; T, M)\Big\|^2 \, ds}, \\ \tilde{\gamma}_i(s; T, M) &= \frac{P(0, M)}{A(0; T, M)} \, \gamma_i(s; T, M) \end{aligned} \tag{33}$$

Taking the ratio for adjacent tenors M_j and M_{j+1} gives:

$$\frac{\sigma_N^{\mathrm{mdl}}(T, M_{j+1})}{\sigma_N^{\mathrm{mdl}}(T, M_j)} = \frac{\dfrac{P(0, M_{j+1})}{A(0; T, M_{j+1})}}{\dfrac{P(0, M_j)}{A(0; T, M_j)}} \times \frac{\sqrt{\displaystyle\int_0^T \|\gamma(s; T, M_{j+1})\|^2 \, \mathrm{d}s}}{\sqrt{\displaystyle\int_0^T \|\gamma(s; T, M_j)\|^2 \, \mathrm{d}s}}\,. \tag{34}$$

Using

$$A(0; T, M) = \frac{P(0, T) - P(0, T + M)}{S(0; T, M)} \tag{35}$$

we rewrite the deterministic prefactor as

$$\frac{P(0, M)}{A(0; T, M)} = \frac{P(0, M)\, S(0; T, M)}{P(0, T) - P(0, T + M)} \tag{36}$$

3.2.1. 'Freeze' Approximation

In the exact ratio of deterministic prefactors for tenors M_j and M_{j+1}, one retains the term $\frac{P(0,M_{j+1})}{P(0,M_j)}$. However, because M_{j+1} and M_j are adjacent tenors (e.g. 1y and 2y), their zero-coupon bond prices at time 0 are almost identical: $P(0, M_{j+1}) \approx P(0, M_j)$. We therefore "freeze" this ratio to unity:

$$\frac{P(0, M_{j+1})}{P(0, M_j)} \;\approx\; 1 \tag{37}$$

Under this approximation, the $P(0, M)$ terms cancel out of the volatility-ratio, leaving only annuity and swap-rate quantities. This controlled simplification has negligible effect on adjacent-tenor ratios while greatly streamlining the algebra. Thus, under the "freeze" approximation, the $P(0, M)$ terms cancel in the ratio, leaving:

$$\frac{P(0, T) - P(0, T + M_{j+1})}{P(0, T) - P(0, T + M_j)} \times \frac{S(0; T, M_j)}{S(0; T, M_{j+1})}.$$

Putting this together defines the volatility ratio calibration target:

$$\begin{aligned} R_N^{\mathrm{mdl}}(T, M_j) = {} & \frac{\sigma_N^{\mathrm{mdl}}(T, M_{j+1})}{\sigma_N^{\mathrm{mdl}}(T, M_j)} \\ & \times \frac{P(0, T) - P(0, T + M_{j+1})}{P(0, T) - P(0, T + M_j)} \\ & \times \frac{S(0; T, M_j)}{S(0; T, M_{j+1})}\,. \end{aligned} \tag{38}$$

Likewise, the market volatility ratio which we calibrate our model to match is:

$$R_N^{\text{mkt}}(T, M_j) = \frac{\sigma_N^{\text{mkt}}(T, M_{j+1})}{\sigma_N^{\text{mkt}}(T, M_j)} \times \frac{P(0,T) - P(0, T+M_{j+1})}{P(0,T) - P(0, T+M_j)} \times \frac{S(0; T, M_j)}{S(0; T, M_{j+1})}. \tag{39}$$

By construction, at the calibrated mean-reversion speed k,

$$R_N^{\text{mdl}}(T, M_j; k) = R_N^{\text{mkt}}(T, M_j) \tag{40}$$

so to fit k we only need the ratio of the integrals $\int_0^T \|\gamma\|^2$. By forming the volatility ratio, we cancel known annuity and discount terms exactly leaving a target that depends only on how σ_N changes with tenor. This reduces each expiry T to a one-dimensional fit of the mean-reversion speed k, making the problem parsimonious, stable, and computationally efficient. Each fraction:

- $\frac{\sigma_N(T,M_{j+1})}{\sigma_N(T,M_j)}$ isolates how volatility changes with tenor,
- $\frac{P(0,T)-P(0,T+M_{j+1})}{P(0,T)-P(0,T+M_j)}$ removes discount-curve scaling,
- $\frac{S(0;T,M_j)}{S(0;T,M_{j+1})}$ aligns the swap-rate units.

Note, given J tenors $M_1 < \cdots < M_J$, the volatility ratio is only defined for $j = 1, \ldots, J-1$, because there is no M_{J+1}. Thus, we obtain exactly $J-1$ ratios from J tenors with the last tenor M_J not yielding its own ratio, but does appear as the numerator in the final ratio $R_N(T, M_{J-1})$.

3.3. Optimization of Mean-Reversion Speed

For a fixed expiry T, once the discount curve and annuities are known, the model implied normal volatility depends only on the single mean-reversion speed k. A small k means the factor-loading $\tilde{\gamma}(s; T, M)$ decays slowly in M, so long-tenor swap rates remain almost as volatile as short tenors — a flat vol-vs-tenor curve. Whereas, a large k makes $\tilde{\gamma}$ decay quickly in M, so the vol drops off sharply for longer tenors — a steep vol-vs-tenor curve. By adjusting the single parameter k we therefore steer how the entire curve $\sigma_N(T, M)$ "rolls off" as tenor length, M, increases. As ATM normal volatility typically starts higher at short tenors and then falls ("rolls off") with M. The calibration then solves:

$$k^*(T) = \arg\min_{k>0} \sum_j w_j \left[R_N^{\text{mdl}}(T, M_j; k) - R_N^{\text{mkt}}(T, M_j) \right]^2 \tag{41}$$

a one-dimensional optimization where w_j are node weights reflecting which surface points matter more. This is both computationally efficient (just one degree of freedom) and numerically stable (using ratios removes level/scale effects). The resulting formulation is parsimonious (only one parameter k per expiry), robust (ratios smooth out curve-building noise) and efficient (a small, well-behaved least-squares problem). We can use a bounded Levenberg–Marquardt solver (Nocedal & Wright, 2006) to numerically optimize for $k*$.

3.3.1. General Calibration of Mean-Reversion Speeds

In the most general setting with n Gaussian factors, one must calibrate the full vector of mean-reversion speeds: $\mathbf{k} = (k_1, k_2, \ldots, k_n)$ by minimizing the mismatch between model and market volatility ratios across all tenors as in (41).

3.3.2. Reduced Calibration to a Single Speed

Although, this study proposes a single mean reverting calibration as empirical studies (Litterman & Scheinkman, 1991; Diebold & Li, 2006) show that the "slope" and "curvature" factors mean-revert much faster and with less cross-expiry variation than the "level" factor. Numerical experiments also reveal that jointly fitting (k_1, k_2, k_3) via (41) is ill-conditioned and prone to overfitting as allowing all three k_i to vary makes the ratio-based calibration depend on a nontrivial combination of k_1, k_2, k_3, leading to multiple local minima which often chases noise in the ATM volatility quotes, especially on sparse tenor grids. Consequently, we fix $k_2 = 20\%$, $k_3 = 100\%$ and calibrate only the level speed k_1 with the other fixed k_i parameters acting as a form of regularization. The choices $k_2 = 20\%$ and $k_3 = 100\%$ are ofcourse arbitrary but based on standard historical anchor values for slope and curvature mean-reversion speeds with $k_2 = 0.20$ capturing medium-term slope dynamics and $k_3 = 1.00$ reflecting the very rapid curvature ('smile' type) adjustments. The one-dimensional calibration becomes:

$$k_1^*(T) = \arg\min_{0<k_1} \sum_j w_j \Big[R_N^{\text{mdl}}\big(T, M_j;\, k_1,\; k_2 = 0.20,\; k_3 = 1.00\big) - R_N^{\text{mkt}}(T, M_j) \Big]^2. \tag{42}$$

which under mild regularity the model-ratio $R_N^{\text{mdl}}(T, M_j; k_1)$ is continuous and strictly monotonic in k_1 and $\Phi(k_1)$ is unimodal and so has a unique global minimum. This means a standard 1D optimizer (e.g. Brent's method or bounded Levenberg–Marquardt) reliably converges to the global minimum from any starting guess. Numerical noise in market quotes can introduce tiny ripples, but these rarely generate additional minima of comparable depth. Thus, this reduction to a single free parameter ensures a robust, fast, and well-posed calibration for each expiry T.

Although, if one has very rich, high-quality data (dense tenor quotes, liquid mid-curve vols), it is possible to free k_2 and k_3 where one could then:

1. Introduce additional calibration moments sensitive to medium- and fast-decay (e.g. cross-expiry or higher-order ratios).
2. Solve a 3-dimensional least-squares or Levenberg–Marquardt problem for (k_1, k_2, k_3).
3. Impose regularization or Bayesian priors around the historical anchors (20%, 100%) to maintain stability.

In practice, however, the marginal improvement in fit unlikely justifies the added complexity, and so we retain the single-speed calibration for k_1.

The full exact formula for calibrating k_1 is thus:

$$
\begin{aligned}
k_1^*(T) = \arg\min_{0<k_1} \sum_{j=1}^{J-1} w_j \Bigg[& \frac{\sigma_N^{\mathrm{mdl}}(T, M_{j+1}; \mathbf{k})}{\sigma_N^{\mathrm{mdl}}(T, M_j; \mathbf{k})} \; \frac{P(0,T) - P(0, T + M_{j+1})}{P(0,T) - P(0, T + M_j)} \; \frac{S(0; T, M_j)}{S(0; T, M_{j+1})} \\
& - \frac{\sigma_N^{\mathrm{mkt}}(T, M_{j+1})}{\sigma_N^{\mathrm{mkt}}(T, M_j)} \; \frac{P(0,T) - P(0, T + M_{j+1})}{P(0,T) - P(0, T + M_j)} \; \frac{S(0; T, M_j)}{S(0; T, M_{j+1})} \Bigg]^2, \\
\sigma_N^{\mathrm{mdl}}(T, M; \mathbf{k}) = \sqrt{\int_0^T \sum_{i=1}^{3} \Big(k_i \, \frac{P(0,M)}{A(0; T, M)} \, \gamma_i(s; T, M) \Big)^2 \, \mathrm{d}s}. &
\end{aligned}
\tag{43}
$$

with $k_2 = 0.20$, $k_3 = 1.00$ held fixed.

3.4. Reparametrizing k for Sensible Values

When otpimizing over k_1, we want solutions to be bounded within 'sensible' economically meaningful values for k. Rather than optimize k directly (risking negative or runaway mean-reversion speeds), we introduce an unconstrained parameter ϕ and set

$$
k(\phi) = \frac{0.05}{1 + e^{-\phi}}, \quad \phi \in \mathbb{R} \tag{44}
$$

so that ϕ maps to a positive, bounded mean-reversion speed $0 < k < 5\%$. We then optimize ϕ via Levenberg–Marquardt, recovering $k^* = k(\phi^*)$, which guarantees a sensible k^* without manual bounds thus maintaining smooth differentiability for gradient-based solvers. We use this reparameterization henceforth in this paper for k_1 but alternatively if only a lower bound $k > 0$ with no upper limit is preferred, one can use the *softplus* transform:

$$
k(\phi) = \ln\big(1 + e^{\phi}\big) \tag{45}
$$

which yields $k > 0$ and allows the data to determine how large k can grow.

3.5. Calibration of Factor-Volatility Levels

With mean-reversion speeds $\{k_1^*,\, k_2,\, k_3\}$ fitted, we next calibrate the three factor-volatility parameters $\eta = (\eta_1, \eta_2, \eta_3)$ by minimizing the squared errors between model and market ATM normal vols. Calibrating all three η_i simultaneously captures both the overall level and cross-tenor shape of the volatility surface, while fixing $\{k_2, k_3\}$ preserves identifiability and robustness. The vol-of-vol calibration solves the nonlinear least-squares problem:

$$
\min_{\eta_i > 0} \sum_{p=1}^{P} \sum_{q=1}^{Q} w_{p,q} \left[\sigma_N^{\mathrm{mdl}}(T_p, M_q; \eta) \; - \; \sigma_N^{\mathrm{mkt}}(T_p, M_q) \right]^2, \tag{46}
$$

where

$$
\sigma_N^{\mathrm{mdl}}(T, M; \eta) = \sqrt{\int_0^T \sum_{i=1}^{3} \Big(\eta_i \, \tfrac{P(0,M)}{A(0;T,M)} \, \gamma_i(s; T, M) \Big)^2 \, ds} \, .
$$

Each η_i multiplies a different "mode" of the surface (level, slope and curvature). Those modes are orthogonal enough that the objective (46) behaves like a strictly convex function in the region $\eta_i > 0$. We again employ a bounded Levenberg–Marquardt (Trust-Region Reflective) algorithm. From any reasonable initial guess (e.g. historical vol-levels), it reliably converges in a handful of iterations to the unique best-fit (η_1, η_2, η_3). Unlike the mean-reversion speeds, where only the "level" speed k_1 is calibrated and the faster "slope" and "curvature" speeds k_2, k_3 are fixed, each factor's volatility parameter η_i has a distinct, non-redundant impact on the shape of the implied volatility surface:

- η_1 (*level* vol-of-vol) controls the overall height of the volatility surface, matching the average ATM normal volatility across tenors.
- η_2 (*slope* vol-of-vol) determines how sharply volatility tilts between short and medium tenors, anchoring the surface's roll-off.
- η_3 (*curvature* vol-of-vol) governs localized bumps or "smile" effects at very short or very long tenors.

If we calibrated only a single η, all three Gaussian factors would scale identically, making it impossible to simultaneously fit a high overall level (η_1) and a gentle slope (η_2) plus a pronounced curvature (η_3).

Therefore, we allow $\{\eta_1, \eta_2, \eta_3\}$ to float independently while keeping $\{k_2, k_3\}$ fixed so that the model can flexibly reproduce the full cross-sectional shape (level, tilt, and twist) of the market's ATM normal-vol surface. In contrast, a single vol-parameter would collapse all shape information into one degree of freedom and systematically mis-fit portions of the surface.

3.6. Reference Volatility Surface

We create *reference* implied volatilities for both calibration stages which in practice replace the market implied volatilities data in the calibration formulas outlined in previous sections. Market-quoted volatilities $\sigma_N^{\mathrm{mkt}}(T_i, M_j)$ can be noisy or sparse. To balance long-run values of data points against a noisy day's variation, which could cause trade pricing and P&L to erratically spike, we propose to compute all calibrations using a reference surface calculated as follows:

$$
\begin{aligned}
\sigma^{\mathrm{ref}}(T_i, M_j) = 0.50\, \overline{\sigma_N^{\mathrm{mkt}}}_{3\mathrm{y}}(T_i, M_j) + 0.25\, \overline{\sigma_N^{\mathrm{mkt}}}_{1\mathrm{y}}(T_i, M_j) \\
+ 0.25\, \sigma_N^{\mathrm{mkt}}(T_i, M_j),
\end{aligned}
\tag{47}
$$

where

$$
\begin{gathered}
\overline{\sigma_N^{\mathrm{mkt}}}_{\tau}(T_i, M_j) = \frac{1}{\tau} \int_{t-\tau}^{t} \sigma_N^{\mathrm{mkt}}(u; T_i, M_j) \, \mathrm{d}u, \\
\tau \in \{1, 3\}\, \mathrm{yr}.
\end{gathered}
\tag{48}
$$

is the τ-year moving average of ATM volatilities. We then replace σ_N^{mkt} by σ^{ref} in both calibration objectives (43, 46), ensuring our fitted parameters reflect a blend of long-run, medium-run and spot market conditions thus yielding stable, robust estimates even when some quotes are missing or erratic.

3.7. Volatility Surface Weighting

Our calibration methodology has thus far assumed equal weightings to each of the K grid-points (T_i, M_j) when calibrating the MFLGM parameters to the reference surface. However, some points may be less useful than others like illiquid nodes which have infrequent or noisy data. One potential solution is to train a variational autoencoder (VAE) (Kingma & Welling, 2013) to identify the nodes whose reconstruction error is largest, i.e. the points that carry the most "information" about surface shape. VAEs are a class of generative latent-variable models that combine probabilistic inference with deep learning. The *encoder* $q_\phi(z \mid x)$ maps an input datum x to a distribution over a lower d-dimensional latent $z \in^d$ (typically Gaussian). At the other end, a *decoder* $p_\theta(x \mid z)$, reconstructs x from a draw of z. This method could also be done via PCA but VAEs here have the advantage of non-linear latent dimensions which can capture more nuanced relationships in volatility surfaces over time. Concretely, let each observed surface $\sigma_{\text{ref}} \in^K$ be a column vector then the VAE is trained by maximizing the evidence lower bound (ELBO) over training surfaces $\{\sigma^{(n)}\}$:

$$\mathcal{L}(\phi, \theta) = \textstyle\sum_n \Big\{_{q_\phi(z|\sigma^{(n)})} [\log p_\theta(\sigma^{(n)} \mid z)] \; - \; [q_\phi(z \mid \sigma^{(n)}) \| p(z)] \Big\}. \quad (49)$$

Once trained, each node $k = 1, \ldots, K$ can be assigned an importance score:

$$I_k = \mathbb{E}_{\sigma \sim \text{data}} \big[(\sigma_k - \hat{\sigma}_k)^2 \big], \quad \hat{\sigma} = \mathbb{E}_{q_\phi(z|\sigma)} \big[p_\theta(\sigma \mid z) \big] \quad (50)$$

i.e. the expected squared reconstruction error at coordinate k. One then defines normalized weights:

$$w_k = \frac{I_k}{\sum_{\ell=1}^{K} I_\ell}, \qquad k = 1, \ldots, K \quad (51)$$

and incorporates them into the calibration objective by minimizing the weighted sum of squared differences:

$$\min_p \sum_{k=1}^{K} w_k \left[\sigma_{\text{model}}(p)_k - \sigma_{\text{ref},k} \right]^2 \quad (52)$$

This procedure focuses the fit on the most informative regions of the surface, reducing over-fitting to noisy or redundant nodes. As a sanity check, we should have at least some nodes (or sufficiently large enough weights) from across each region of the surface, so for instance, you avoid calibrating to just very short maturity points even though they may have a higher importance score due to them having more variation. This is an area which requires more research to reliably select the most important points and so our calibration results presented in this paper will maintain a uniform surface weighting.

3.8. Calibration Results

Using the reference EUR interest rate volatility surface for 10/06/2025, as displayed in Figure 1, our volatility ratio calibration process returns the following calibrated parameters:

The calibrated parameters in 2 generate the volatility surface in Figure 2 for that same day (10/06/2025).

Table 2. Calibrated MFLGM Parameters (10/06/2025)

Parameter	Calibrated Value
k_1	0.05
k_2	0.20
k_3	1.00
η_1	0.01454
η_2	0.00835
η_3	0.01129

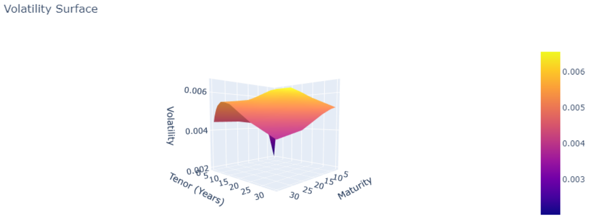

Figure 1. Reference Volatility Surface (30/01/2025).

We can then diffuse this surface using the *Closed-Form Expected Surface* method, which yields a single, analytic forecast of the mean surface, or use the *Monte Carlo Simulation* method, which produces an ensemble of possible surfaces, capturing the distributional uncertainty. Under MFLGM, each OU factor $X_i(t)$ and the log-level shift $v(t)$ of the vol-surface are Ornstein–Uhlenbeck processes:

$$\begin{aligned} dX_i(t) &= -k_i \, X_i(t) \, dt + \eta_i \, dW_i(t), \\ dv(t) &= -\lambda_v \, v(t) \, dt + \sigma_v \, dZ(t), \end{aligned} \quad (53)$$

for $i = 1, \ldots, n$, with calibrated parameters k_i, η_i, λ_v, σ_v and initial states $X_i(0)$, $v(0)$. The exact marginal expectations at for instance $\Delta t = 6$ months follow from standard OU theory:

$$\mathbb{E}[X_i(\Delta t)] = e^{-k_i \, \Delta t} \, X_i(0), \quad \text{Var}[X_i(\Delta t)] = \frac{\eta_i^2}{2 \, k_i} \big(1 - e^{-2k_i \, \Delta t} \big), \quad (54)$$

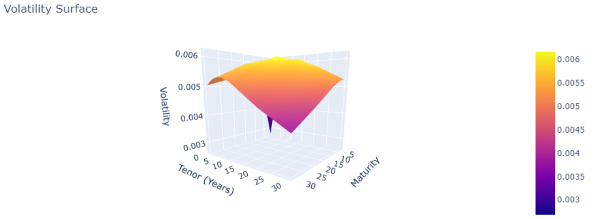

Figure 2. Calibrated Volatility Surface (30/01/2025).

$$\mathbb{E}[v(\Delta t)] = e^{-\lambda_v\,\Delta t}\,v(0). \tag{55}$$

By substituting these means into the continuous-maturity loadings $\phi_i(T)$ (cf. (Brigo & Mercurio, 2006a)) and the integrated diffusion formula

$$\sigma_N(T,M) = \sqrt{\int_0^T \left\|\tilde{\gamma}(s;T,M)\right\|^2 \mathrm{d}s \times \exp\!\big(\mathbb{E}[v(\Delta t)]\big)} \tag{56}$$

we obtain a single deterministic forecast $\sigma_N^{\exp}(T,M)$ at $t =$ 6 months. This approach is computationally trivial and fully analytic, but it omits any measure of variability around the mean.

Whereas Monte Carlo methods capture the full stochastic dispersion of future surfaces by simulating N independent joint paths of $\{X_i(t), v(t)\}$ over $t \in [0, \Delta t]$. For each path $n = 1, \ldots, N$:

1. Simulate $X_i^{(n)}(\Delta t)$ and $v^{(n)}(\Delta t)$ via an Euler-Maruyama or exact Gaussian integrator for OU processes (see (Uhlenbeck & Ornstein, 1930)).

2. Reconstruct the swap-rate loading $\tilde{\gamma}^{(n)}(s;T,M)$ and compute:

$$\sigma_N^{(n)}(T,M) = \sqrt{\int_0^T \left\|\tilde{\gamma}^{(n)}(s;T,M)\right\|^2 \mathrm{d}s \times \exp\!\big(v^{(n)}(\Delta t)\big)} \tag{57}$$

3. Store $\sigma_N^{(n)}(T,M)$ for all (T,M) on the desired grid.

Aggregating over the ensemble yields empirical estimates of the mean, median, confidence bands (e.g. 5th–95th percentiles) and the full distribution of $\sigma_N(T,M)$ at t months. While more computationally intensive, this method provides crucial insights into tail-risk and the volatility cone which is more useful for risk management and scenario analysis. We leave it to future research to test and compare the results of each diffusion method of our calibrated MFLGM model.

4. Application: Vega Surface Construction

4.1. Objective

We will now use our calibrated MFLGM for diffusing volatility surfaces, calibrated to 6 factors $(k_1, k_2, k_3, \eta_1, \eta_2, \eta_3)$ with k_2, k_3 fixed like previously, to quantify the sensitivity of a portfolio's P&L to local perturbations of swaptions across the volatility surface, i.e. construct a vega surface. Formally, let $\sigma(T_i, M_j)$ denote the ATM normal volatility at expiry T_i and tenor M_j. We aim to construct a *surface-vega grid*

$$K_{ij} = \frac{\partial\,\mathrm{PL}}{\partial\sigma(T_i, M_j)}, \quad i = 1, \ldots, n_T,\ j = 1, \ldots, n_M \tag{58}$$

where n_T and n_M are the numbers of expiries and tenors in our fixed grid.

4.2. Methodology

Key to our methodology is that rather than attempting to differentiate the model-implied volatilities at each grid node directly with respect to the model parameters, a route that leads to a high-dimensional, ill-conditioned system, we novely employ a two-stage projection through principal components followed by parameter-space sensitivities. This procedure ensures invertibility, numerical stability, and interpretability.

4.2.1. Historical Surface Differencing and Centering

Let $\sigma_t(T_i, M_j)$ be the raw ATM normal volatility observed on day t at expiry T_i and tenor M_j, for $t = 1, \ldots, N$, $i = 1, \ldots, n_T$, $j = 1, \ldots, n_M$, $K = n_T n_M$, between $01/01/2020 - 31/01/2025$. We daily difference the volatility surface:

$$\Delta\sigma_t = \sigma_t(T_i, M_j) - \sigma_{t-1}(T_i, M_j), \quad t = 2, \ldots, N \tag{59}$$

and flatten each day's volatility surface into a vector to then arrange into matrix form:

$$\Delta X = \begin{bmatrix} \Delta\sigma_2(T_1, M_1) & \cdots & \Delta\sigma_2(T_{n_T}, M_{n_M}) \\ \vdots & \ddots & \vdots \\ \Delta\sigma_N(T_1, M_1) & \cdots & \Delta\sigma_N(T_{n_T}, M_{n_M}) \end{bmatrix} \in \mathbb{R}^{(N-1)\times K} \tag{60}$$

Each column k of ΔX is then demeaned by its own historical mean

$$\begin{aligned} \Delta X_{\text{centered}} &= \Delta X - \mathbf{1}_{N-1}\,\bar{\Delta\sigma}^{\top}, \\ \bar{\Delta\sigma}_k &= \frac{1}{N-1}\sum_{t=2}^{N} \Delta X_{t-1,k}. \end{aligned} \tag{61}$$

This focuses the analysis on the stochastic fluctuations of the surface rather than on its absolute level. Demeaning is essential both to satisfy the zero-mean assumption of PCA and to remove any persistent drift in each node.

4.2.2. PCA Basis Extraction

Principal component analysis is then performed on $\Delta X_{\text{centered}}$ by solving:

$$\frac{1}{N-2}\,\Delta X_{\text{centered}}^{\top}\,\Delta X_{\text{centered}}\,v_\ell = \lambda_\ell\,v_\ell, \quad \ell = 1, \ldots, K \tag{62}$$

and retaining the first r eigenvectors $v_1, \ldots, v_r$ corresponding to the largest eigenvalues. These form the loading matrix

$$L = \begin{bmatrix} v_1 & \cdots & v_r \end{bmatrix} \in \mathbb{R}^{K\times r},$$

which furnishes a low-dimensional basis capturing the dominant modes of surface variation (commonly interpreted as level, slope, and curvature) so that any new daily shock $\Delta\sigma \in \mathbb{R}^K$ can be compressed into principal-component scores

$$s = L^{\top}\big(\Delta\sigma - \bar{\Delta\sigma}\big) \in \mathbb{R}^r \tag{63}$$

4.2.3. Reference Surface Calibration

Each trading day begins by constructing a smooth, *reference* volatility surface σ_{ref}, like explained in Section 3.6. We then calibrate our MFLGM model to σ_{ref} using the methodology outlined in Section 3, yielding the parameter vector $p_0 \in \mathbb{R}^P$. From p_0 one derives the model-implied base surface for that calibration day:

$$\sigma_0(T_i, M_j) = \sigma_{\text{model}}(T_i, M_j;\, p_0),$$

and its associated scores $s_0 = L^\top(\sigma_0 - \bar{\sigma})$.

4.2.4. PnL-per-Parameter Sensitivities

To quantify the P&L sensitivity with respect to each model parameter, we employ symmetric finite-differences. For $i = 1, \ldots, P$, define bumped parameter sets: $p_0^{(\pm,i)} = \big(p_{0,1}, \ldots, p_{0,i} \pm \delta p_i, \ldots, p_{0,P}\big)$, where δp_i is a small perturbation (e.g. 10bp or 0.001 in decimal terms). Re-pricing the entire book under $p_0^{(+i)}$ and $p_0^{(-i)}$ yields P&Ls: $\mathrm{PL}_i^{(+)}$ and $\mathrm{PL}_i^{(-)}$, from which we compute

$$\begin{aligned} S_{p_i} &= \frac{\mathrm{PL}_i^{(+)} - \mathrm{PL}_i^{(-)}}{2\,\delta p_i} && (\text{€ per unit } p_i), && (64)\\ S_{p_i}^{\text{bp}} &= S_{p_i} \times 10^{-4} && (\text{€ per bp}). && (65) \end{aligned}$$

We use central differencing to take into account any non-symmetric effects from positive compared to negative parameter bumps. In our code, we also convert the sensitivity with respect to a 1bp move.

4.2.5. Score-Jacobian Construction via PCA

In principle one could attempt to build the exact Jacobian $\partial\sigma/\partial p \in \mathbb{R}^{K\times P}$ by differentiating our model implied volatility formula with respect to each parameter and then propagate P&L sensitivities via the chain rule. However, (i) the tall $K \times P$ matrix is not directly invertible when $K > P$, and (ii) its inversion would amplify node-level noise. Instead, we introduce the Jacobian of principal-component scores with respect to parameters:

$$\begin{aligned} [J_{\text{score}}]_{j,i} &= \frac{L_{\cdot,j}^\top\big[\sigma\big(p_0^{(+,i)}\big) - \bar{\sigma}\big] - L_{\cdot,j}^\top\big[\sigma\big(p_0^{(-,i)}\big) - \bar{\sigma}\big]}{2\,\delta p_i}\\ &= \frac{s_{i,j}^{(+)} - s_{i,j}^{(-)}}{2\,\delta p_i}\,. \end{aligned} \tag{66}$$

where $L_{\cdot,j}$ denotes the jth column of L. This yields

$$J_{\text{score}} = \frac{\partial s}{\partial p} \;\in\; \mathbb{R}^{r\times P},$$

which is square and thus invertible (when $r = P$). Since our MFLGM methodology employs 4 free parameters to which we want to observe their sensitivity to the volatility surface, we perform PCA with 4 principal components which, whilst uncommon, is still a sensible number of components to describe the volatility surface (it is typically done using 3 components and Figure 3 shows its' use in explaining total surface variation observed).

4.2.6. Projection of Market-Observed Shocks

Let σ_{today} and $\sigma_{\text{yesterday}}$ be the ATM volatility surfaces from two consecutive trading days. Their difference $\Delta\sigma = \sigma_{\text{today}} - \sigma_{\text{yesterday}} \in \mathbb{R}^K$ is projected to PCA-space by

$$\Delta s = L^\top\,\Delta\sigma \;\in\; \mathbb{R}^r \tag{67}$$

4.2.7. Weighting PCA-Score Moves by Explained Variance

Although the raw principal-component scores $s = L^\top\,\Delta\sigma$ capture the projection of the surface shock onto each eigenmode, the absolute magnitude of each score is measured in arbitrary "units" of volatility change. In particular, higher-order components (e.g. PC4) may exhibit large loadings but explain only a tiny fraction of historical variation, while the first few components (level, slope, curvature) carry the bulk of the surface dynamics. To prevent less-relevant directions from disproportionately driving our parameter shifts and P&L attribution, we rescale each score by the proportion of total variance $\lambda_j / \sum_{\ell=1}^r \lambda_\ell$ that its eigenvalue λ_j represents. Concretely, if

$$\begin{aligned} \{\lambda_1, \ldots, \lambda_r\} &= \operatorname{eigvals}\big(\tfrac{1}{N-2}\,\Delta X_{\text{centered}}^\top \Delta X_{\text{centered}}\big),\\ \sum_{j=1}^r \lambda_j &= 1. \end{aligned} \tag{68}$$

then we define the *weighted* score vector

$$s^{(\text{w})} = \operatorname{diag}\big(\lambda_1, \ldots, \lambda_r\big)\, s \;=\; \operatorname{diag}\big(\lambda_1, \ldots, \lambda_r\big)\, L^\top\,\Delta\sigma, \tag{69}$$

so that each mode's contribution is naturally down-weighted in proportion to its statistical significance. Equivalently, one may use $\sqrt{\lambda_j}$ or any monotonic function of λ_j; the key principle is to suppress spurious noise from minor components.

4.2.8. Implied Parameter Shifts and P&L Computation

For an observed $\Delta s^{(\text{w})}$ we want to find the parameter shift implied by the Jacobian matrix which caused this move. Hence:

$$\begin{aligned} &\Delta s^{(\text{w})} \approx J^{\text{score}}\,\Delta p,\\ &\qquad \Longrightarrow\; \Delta p = \arg\min_{\Delta p} \big\| J^{\text{score}}\,\Delta p - \Delta s^{(\text{w})} \big\|^2,\\ &\qquad = \frac{(J^{\text{score}})^\top\,\Delta s^{(\text{w})}}{(J^{\text{score}})^\top\,J^{\text{score}}}\;(OLSsolution) \end{aligned} \tag{70}$$

The ensuing P&L change is then

$$\Delta\mathrm{PL} = \sum_{i=1}^P S_{p_i}^{\text{bp}}\,\Delta p_i \tag{71}$$

This explains portfolio P&L movements from day-to-day volatility surface changes. The contributions by each parameter are given by the individual products $S_{p_i}^{\text{bp}}\Delta p_i$ (both in bps).

4.2.9. Regularized Regression for Implied Parameter Shifts

In the context of mapping observed volatility-surface shocks into model-parameter moves, we solve a linear system: $J_{\text{score}}\,\Delta p \approx \Delta s$, where $J_{\text{score}} \in \mathbb{R}^{r\times P}$ is the PCA-score Jacobian, $\Delta s \in \mathbb{R}^r$ the vector of principal-component score shifts and $\Delta p \in \mathbb{R}^P$ the desired parameter perturbations. When J_{score} is square and full-rank, the ordinary least-squares (OLS) solution reduces to the direct inverse $\Delta p = J_{\text{score}}^{-1}\,\Delta s$. However, in practical applications especially when dealing with large P&L sensitivities and noisy surfaces:

- Some principal components explain very little historical variance, so the corresponding singular values of J_{score} are near zero.
- Direct inversion amplifies measurement noise or microstructure effects in the volatility surface, yielding spurious, excessively large Δp.

To stabilise the inversion and suppress over-sensitive directions, we introduce *ridge regularization* (Tikhonov smoothing), which adds a penalty on the norm of Δp. The ridge-regularized estimate is defined as the minimiser of a penalised least-squares objective:

$$\Delta p_\alpha = \arg\min_{\Delta p}\Big\{\big\|J_{\text{score}}\,\Delta p \;-\; \Delta s\big\|_2^2 \;+\; \alpha\,\big\|\Delta p\big\|_2^2\Big\}, \quad (72)$$

where $\alpha > 0$ is the *regularization parameter* with a closed-form solution:

$$\Delta p_\alpha = \big(J_{\text{score}}^\top J_{\text{score}} \;+\; \alpha\,I_P\big)^{-1}\,J_{\text{score}}^\top\,\Delta s. \quad (73)$$

The term $\alpha\,I_P$ augments the normal equations, lifting the smallest eigenvalues of $J_{\text{score}}^\top J_{\text{score}}$ away from zero. As $\alpha \to 0$, $\Delta p_\alpha \to \Delta p_{\text{OLS}}$. As $\alpha \to \infty$, $\Delta p_\alpha \to 0$, shrinking the solution towards the origin. Proper choice of α balances fidelity to the observed score change against suppression of spurious noise.

Ridge regularization ensures a stable mapping when faced with high-dimensional noisy market data where microstructure noise and low-variance components can otherwise dominate the solution. This thus mitigates against runaway parameter moves driven by near-singular inversion directions. The following methodology continues without assuming regularization for simplicity however the results in Section 6 illustrate outcomes under both scenarios.

4.2.10. Reconstruction of the Surface-Vega Grid

Finally, one can solve for the hypothetical \$/bp sensitivity at each node via:

$$K \;=\; L\,(\lambda_1,\dots,\lambda_r)\;J_{\text{score}}\,(J_{\text{score}}^\top J_{\text{score}})^{-1}\,S_{p_i}^{\text{bp}}, \quad (74)$$

We back-out this formula for K beginning with the P&L change in terms of parameter-space sensitivities $S_{p_i}^{\text{bp}} \in \mathbb{R}^P$:

$$\Delta\text{PL} = \big(S_{p_i}^{\text{bp}}\big)^\top\,\Delta p, \quad (75)$$

and then substitute Δp from (70) and $\Delta s^{(\text{w})}$ from (69) into (75) to get:

$$\Delta\text{PL} = \big(S_{p_i}^{\text{bp}}\big)^\top\Big[(J_{\text{score}}^\top J_{\text{score}})^{-1}J_{\text{score}}^\top\;\underbrace{(\lambda_1,\dots,\lambda_r)\,L^\top\,\Delta\sigma}_{\Delta s^{(w)}}\Big] \quad (76)$$

$$= \big[\,J_{\text{score}}\,(J_{\text{score}}^\top J_{\text{score}})^{-1}S_{p_i}^{\text{bp}}\big]^\top\,(\lambda_1,\dots,\lambda_r)\,L^\top\,\Delta\sigma$$

$$\big(\text{since } a^\top B\,c = c^\top B^\top a\big)$$

$$= \big[\,L\,(\lambda_1,\dots,\lambda_r)\,J_{\text{score}}\,(J_{\text{score}}^\top J_{\text{score}})^{-1}S_{p_i}^{\text{bp}}\big]^\top\,\Delta\sigma.$$

Thus, the K-vector of node-by-node vega sensitivities is:

$$K \;=\; L\,(\lambda_1,\dots,\lambda_r)\;J_{\text{score}}\,(J_{\text{score}}^\top J_{\text{score}})^{-1}\,S_{p_i}^{\text{bp}}, \quad (77)$$

so that $\Delta\text{PL} = K^\top\,\Delta\sigma$.

Reshaping the entries of $K \in \mathbb{R}^K$ back onto the $(n_T \times n_M)$ grid produces the surface-vega matrix K_{ij}. Since our volatilities are quoted in decimal units, each K_{ij} is in dollars per 1.0 move. To express the sensitivity per 1 bp instead, we simply divide by 10 000, yielding $\Delta\text{PL} = \sum_{i,j}\big(K_{ij}/10\,000\big)\,\Delta\sigma_{ij}^{\text{bp}}$.

4.3. Why PCA Rather Than an Analytical Jacobian?

A natural alternative is to derive the exact Jacobian

$$J_{\sigma,p} \;=\; \frac{\partial\,\sigma}{\partial p} \;\in\; \mathbb{R}^{K\times P}, \quad K = n_T\,n_M, \quad (78)$$

with $p = (p_1,\dots,p_P)$. However, two fundamental issues arise:

1. **Dimensionality and Invertibility.** Since $K \gg P$ in practice, $J_{\sigma,p}$ is non-square and cannot be directly inverted to obtain the parameter shifts Δp that reproduce an observed surface move $\Delta\sigma$. In contrast, by first projecting onto the leading r principal components ($r \le P$), one obtains

   $$J_{\text{score}} \;=\; \frac{\partial\,s}{\partial p} \;\in\; \mathbb{R}^{r\times P}, \quad (79)$$

   which is square when $r = P$ (or well-conditioned if $r < P$), and thus admits the unique least-squares solution

   $$\Delta p = \big(J_{\text{score}}^\top\,J_{\text{score}}\big)^{-1}\,J_{\text{score}}^\top\,\Delta s. \quad (80)$$

2. **Noise Reduction and Regularization.** Volatility surfaces contain microstructure noise from bid-ask bounces, missing quotes, and calendar effects especially at individual (T_i, M_j) nodes. Direct inversion of a high-dimensional, non-square Jacobian amplifies these spurious fluctuations and leads to unstable parameter estimates. By retaining only the statistically significant PCA modes (level, slope, curvature, etc.), one effectively filters out high-frequency noise and ensures a stable, interpretable mapping from surface moves to parameter adjustments.

5. Data

The inputs to our vega-surface construction are summarized as follows:

1. **Historical ATM volatility surfaces.** Daily EUR ATM normal-volatility surfaces were obtained from Fenics market data over the period $01/01/2020 - 31/01/2025$. Each surface is quoted in decimal normal-volatilities on a fixed grid of expiries T_i and tenors M_j.

2. **Bumped P&L results.** Finite-difference repricings (bumped by 0.001 or 10bps) of the four model parameters generated the following P&L outcomes in an example portfolio:

Table 3. Example Bumped P&L (€)

Bump Key	P&L
mrs_negative_bump	4419
mrs_positive_bump	-4412
vol1_negative_bump	-62519
vol1_positive_bump	57541
vol2_negative_bump	41961
vol2_positive_bump	-41961
vol3_negative_bump	-14158
vol3_positive_bump	14158

3. **Consecutive-day surfaces.** Raw ATM volatility surfaces for two adjacent trading days (denoted $t-1$ and t) are processed to obtain the reference volatility surface and then used to compute the realized surface shock $\Delta\sigma$. Here we use as example the volatility surfaces for 31/01/2025 and 30/01/2025.

6. Results

6.1. PCA Results on the EUR Swaption Surface

The first four principal components explain 96.88% of the variation observed in EUR swaption volatility surfaces with Figure 3 showing the cumulative explanation of each factor.

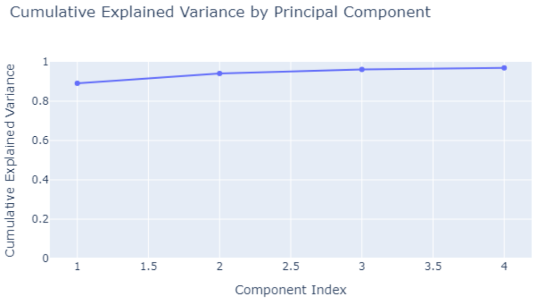

Figure 3. Cumulative Variation Explained by Principal Components.

The loading factors for each principal component are shown in Figures 4–7, respectively.

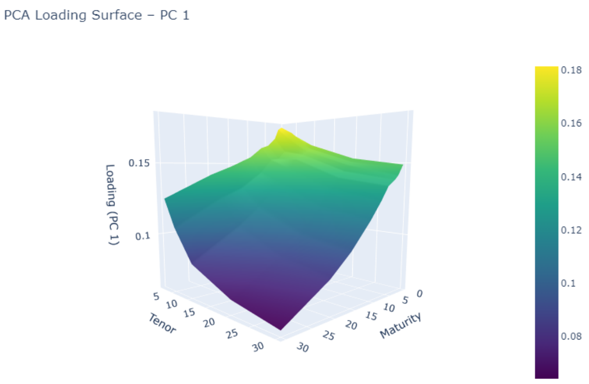

Figure 4. PCA Loading Surface: Principal Component 1 (Level).

Here PC1 displays the characteristic 'level' interpretation of surface movement as all loading factors are positive and are mainly parallel shocks except for the relatively stronger movement in shorter maturity swaptions.

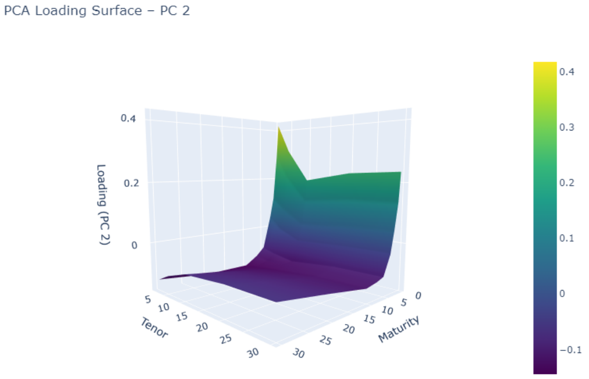

Figure 5. PCA Loading Surface: Principal Component 2 (Slope).

PC2 displays a 'slope' effect with loading factors moving from negative to positive representing opposite movements in shorter and longer dated maturity swaptions.

PC3 is typically meant to represent the 'curvature' effect of how the belly of the surface moves compared to its edges. Here this is less obvious but we do observe the factor loadings changing from positive to negative to positive again indicating some form of curvature effect.

PC4 accounts for very little of the variation observed and is usually harder to interpret but we include it in the analysis as by using 4 principal components we get the nice mathematical property of having a square, invertible Jacobian matrix.

These effects can be more clearly seen with a 2D plot of loading factors against maturity only (which show the average maturity loading factors across tenors for each maturity):

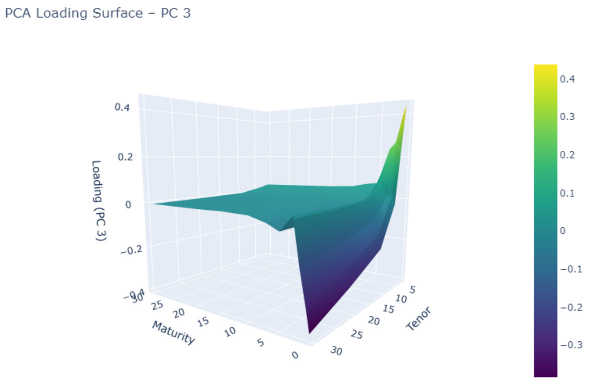

Figure 6. PCA Loading Surface: Principal Component 3 (Curvature).

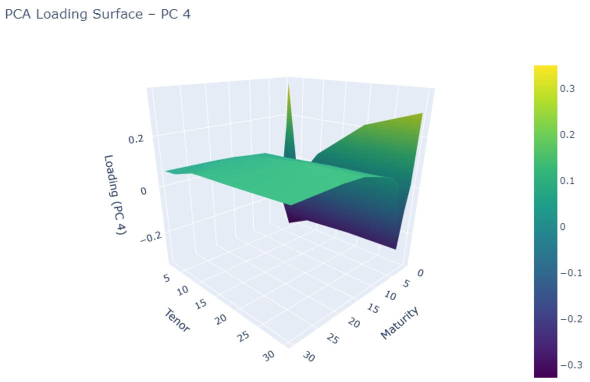

Figure 7. PCA Loading Surface: Principal Component 4 (Higher-Order).

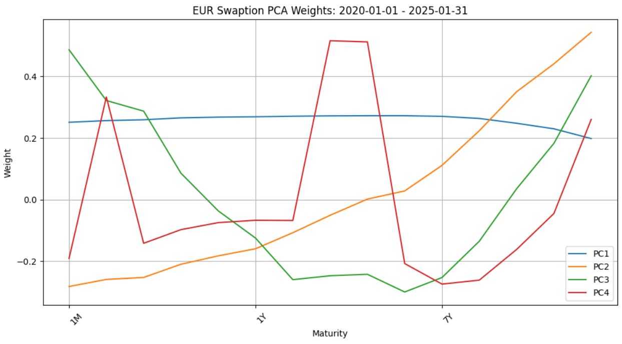

Figure 8. PCA Loading Factors 2D Plot Against Maturity.

6.2. Portfolio P&L Attribution

As an example, for the changes in the volatility surface over 30/01/2025 - 31/01/025 and our book's P&L parameter sensitivities, we observe a P&L impact of €206 452.84. Table 4 presents a summary of the results.

Table 4. Summary of Parameter- and Surface-Chain Sensitivities (in EUR)

Metric	**Value**
Parameter-chain $\Delta\mathrm{PL}$	€206 452.84
Surface-chain $\Delta\mathrm{PL}$	€206 452.84
€ per 1bp move in parameters $S_{\mathrm{per\_bp}}$	$[-441.55,\ 6003.00,\ -4196.10,\ 1415.80]$
J_{score} **(4×4)**	$\begin{bmatrix} -0.001245 & 0.002736 & 0.004303 & 0.001817 \\ -0.008853 & -0.001879 & -0.004833 & -0.005406 \\ 0.004714 & 0.002949 & 0.002618 & 0.007794 \\ 0.001570 & -0.001381 & 0.003037 & 0.007869 \end{bmatrix}$
Δs (PC-score shifts)	$[-0.000085,\ 0.000002,\ -0.000001,\ 0.000000]$
Δp (parameter shifts)	$[0.007396,\ -0.006858,\ -0.014521,\ 0.002931]$
Δp (in bp)	$[73.955992,\ -68.582902,\ -145.210325,\ 29.307965]$
PL contribution per parameter	$[-32655.27,\ -411703.16,\ 609317.05,\ 41494.22]$

6.3. Surface-Vega Map

The reconstructed dollar-per-1bp sensitivity matrix (K_{ij}) for each expiry–tenor node as backed out in (77) is presented in Figure 9 which shows the P&L move for a 1bp increase in each node.

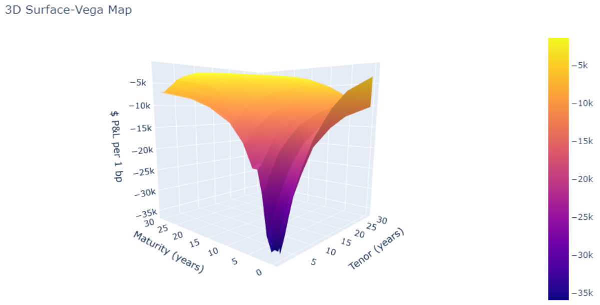

Figure 9. Surface-Vega Grid (€/bp).

This plot gives us an idea of which points on the surface we are most sensitive to movements in like shorter maturity swaptions and in particular the short maturity short tenor swaptions which have the largest P&L impact. Although, important to note is that the P&L impacts at each node are standalone moves when in fact they do not move independently but are correlated with each other (moving as described by the key factors we decomposed via our PCA analysis). So the main benefit of this plot is to visualize which regions on the surface we are generally most sensitive to.

6.4. Regularized Regression Results

When testing a series of random portfolios by re-pricing a portfolio before and after shocking each node on the surface by a uniform increase, e.g. 10bps, we find that setting $\alpha \approx 0.005$ produces similar historical P&L results between our method and direct surface shocks. This result also scales with different P&L parameter sensitivities and volatility surface shocks. Using the same P&L parameter sensitivities as in Table 4 and hyperparameter $\alpha = 0.005$ we construct the vega surface in Figure 10.

Figure 10 shows that whilst most the surface still displays negative vega, with the most negative point still short tenor short maturity points, there is a roll up for the shortest maturities across tenors $> 2Y$ to even having positive vega.

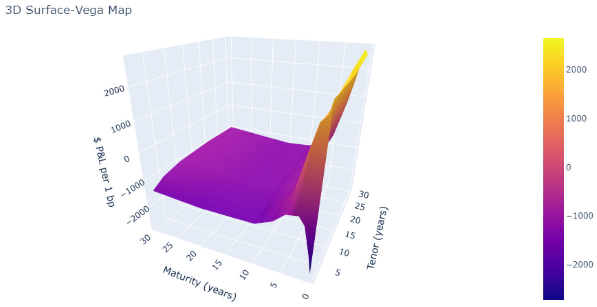

Figure 10. Surface-Vega Grid (€/bp).

In practice, when working with real noisy market data, we would recommend using the regularization approach. Whilst it loses some interpretability in the method it tends to produce more stable and, if well parameter-tuned, accurate P&L results.

7. Conclusion

We began by introducing swaptions and implied interest rate volatility surfaces in Section 1. We then outlined the theory of multi factor linear gaussian models (MFLGMs), an industry standard parametric model, to diffuse interest rate volatility surfaces. Section 3 then derived how to calibrate such models using the volatility ratio calibration approach. Some novel additions to this method include using a smoothed reference surface to avoid large day-to-day P&L fluctuations, factor reparameterization to ensure 'sensible' calibrated values and a node weighting procedure using VAEs. We then presented an example input volatility surface and its calibrated output. Alongside its primary role of diffusing volatility surfaces over time, we presented another useful application of the MFLGM calibrated model in Section 4 - to calculate a portfolio's sensitivity to the implied volatility surface, i.e. vega-surface construction. This novel method uses a score-Jacobian matrix calculated by finite differencing the projection of MFLGM parameter bumped surfaces onto the principal component dimensions to create an invertible square matrix we then use to back-out implied parameter moves from daily volatility surface changes and thus its associated impact on a portfolio's P&L given portfolio parameter sensitivities. When facing noisy market data we propose using ridge regularization to suppress over sensitive directions when backing-out implied parameter shifts. This process crucially gives us an interpretable explanation of P&L movements arising from daily volatility surface movements via their impact on the MFLGM parameters and hence P&L, as well as a vega-surface showing which regions a portfolio is most sensitive to - a crucial insight for risk management.

References

Ahdida, A., Alfonsi, A., and Palidda, E. Smile with the gaussian term structure model. 2015.

Brace, A., Gatarek, D., and Musiela, M. The market model of interest rate dynamics. *Mathematical Finance*, 7(2):127–154, 1997.

Brigo, D. and Mercurio, F. *Interest Rate Models—Theory and Practice*. Springer, 2nd edition, 2006a.

Brigo, D. and Mercurio, F. *Interest Rate Models—Theory and Practice*. Springer, Berlin, 2nd edition, 2006b.

Diebold, F. X. and Li, C. Forecasting the term structure of government bond yields. *Journal of Econometrics*, 130(2):337–364, 2006.

Heath, D., Jarrow, R., and Morton, A. Bond pricing and the term structure of interest rates: A new methodology. *Econometrica*, 60(1):77–105, 1992.

Henrard, M. *Interest Rate Modelling in the Multi-curve Framework*. Palgrave Macmillan, 2014.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Litterman, R. B. and Scheinkman, J. Common factors affecting bond returns. *Journal of Fixed Income*, 1(1):54–61, 1991.

Nocedal, J. and Wright, S. J. *Numerical Optimization*. Springer, 2nd edition, 2006.

Uhlenbeck, G. E. and Ornstein, L. S. On the theory of the brownian motion. *Physical Review*, 36(5):823–841, 1930.

A. Derivations

A.1. Derivation of the Bachelier Model Formula

The Bachelier model assumes that the underlying swap rate evolves according to a normal diffusion process:

$$dS_t = \sigma_N \, dW_t \tag{81}$$

where σ_N is the volatility and W_t is a standard Brownian motion. Integrating this stochastic differential equation from 0 to T gives:

$$S_T = S_0 + \sigma_N \sqrt{T} \, Z \tag{82}$$

with $Z \sim N(0, 1)$.

Under risk-neutral valuation, the value of a European payer swaption at expiration T with strike K is:

$$P_N(0, T) = e^{-rT} \, \mathbb{E}^Q \big[\max(S_T - K, \, 0)\big] \tag{83}$$

Using the normal distribution assumption:

$$\begin{aligned} P_N(0,T) &= e^{-rT} \int_{-\infty}^{\infty} \max\big(S_0 + \sigma_N \sqrt{T} \, z - K, \, 0\big) \, \phi(z) \, \mathrm{d}z, \\ &= e^{-rT} \big[(S_0 - K) \, \Phi(d) + \sigma_N \sqrt{T} \, \phi(d)\big]. \end{aligned} \tag{84}$$

where

$$d = \frac{S_0 - K}{\sigma_N \sqrt{T}},$$

and $\Phi(\cdot)$, $\phi(\cdot)$ are respectively the cumulative distribution and density functions of the standard normal distribution.

This provides the Bachelier formula commonly used to price swaptions under normal volatility assumptions.

A.2. Proof of the Zero-Coupon Bond Price Formula

Under the risk-neutral measure $\mathbb{Q}$, discounting by the money-market account

$$B(t) = \exp\bigg(\int_0^t r(u) \, \mathrm{d}u\bigg) \tag{85}$$

makes any asset price $S(t)$ satisfy

$$\frac{S(t)}{B(t)} \quad \text{is a } \mathbb{Q}\text{-martingale.} \tag{86}$$

In particular, the zero-coupon bond paying \$1 at T has

$$P(t, T) = \mathbb{E}^{\mathbb{Q}}\Big[e^{-\int_t^T r(u) \, \mathrm{d}u} \, \Big| \, \mathcal{F}_t\Big] \tag{87}$$

By definition of the instantaneous forward rate,

$$f(t, T) = -\frac{\partial}{\partial T} \ln P(t, T) \tag{88}$$

Integrate this in T from t to T:

$$\begin{aligned} \int_t^T f(t, u) \, du &= -\int_t^T \frac{\partial}{\partial u} \ln P(t, u) \, du \\ &= -\big[\ln P(t, T) - \ln P(t, t)\big] \\ &= -\ln P(t, T), \end{aligned} \tag{89}$$

since $P(t, t) = 1$. Exponentiating both sides gives the desired result:

$$P(t, T) = \exp\Big(-\int_t^T f(t, u) \, \mathrm{d}u\Big) \tag{90}$$