

Simulation of Arbitrage-Free Implied Volatility Surfaces

Rama Cont & Milena Vuletić

To cite this article: Rama Cont & Milena Vuletić (2023) Simulation of Arbitrage-Free Implied Volatility Surfaces, Applied Mathematical Finance, 30:2, 94-121, DOI: [10.1080/1350486X.2023.2277960](https://doi.org/10.1080/1350486X.2023.2277960)

To link to this article: <https://doi.org/10.1080/1350486X.2023.2277960>



© 2023 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



Published online: 22 Nov 2023.



Submit your article to this journal [↗](#)



Article views: 668



View related articles [↗](#)



View Crossmark data [↗](#)

Simulation of Arbitrage-Free Implied Volatility Surfaces

Rama Cont and Milena Vuletić

Mathematical Institute, University of Oxford, Oxford, England

ABSTRACT

We present a computationally tractable method for simulating arbitrage-free implied volatility surfaces. We illustrate how our method may be combined with a data-driven model based on historical SPX implied volatility data to generate dynamic scenarios for arbitrage-free implied volatility surfaces. Our approach conciliates static arbitrage constraints with a realistic representation of statistical properties of implied volatility co-movements.

ARTICLE HISTORY

Received 26 June 2023
Accepted 25 October 2023

KEYWORDS

Implied volatility; options markets; volatility index; simulation; arbitrage

1. Introduction

Market prices of options are quoted in terms of their Black-Scholes implied volatilities, obtained by inverting the Black Scholes formula given the market price of the option. It has been empirically documented across many options markets that the implied volatility $\Sigma_t(K, T)$ associated with a call option with exercise price K and maturity date T actually depends on (K, T) (Cont and da Fonseca 2002; Dumas, Fleming, and Whaley 1998; Dupire 1994; Gatheral 2011; Heynen 1994). The function $\Sigma_t : (K, T) \rightarrow \Sigma_t(K, T)$ which represents this dependence, called the *implied volatility surface* at date t , provides a snapshot of prices in the options market (Kamal and Gatheral 2010). An example is given in Figure 1 for SPX index options.

Two features of this surface have captured the attention of researchers in financial modeling. First, the non-flat instantaneous profile of the surface, whether it be a ‘smile’, ‘skew’ or the existence of a term structure, point out to the insufficiency of the Black Scholes model for matching a set of option prices at a given time instant and have led to various generalizations of the Black-Scholes model which aim at reproducing realistic instantaneous profiles for the surface $\Sigma_t(K, T)$. Second, the fact that the surface itself changes randomly with time as a result of supply and demand in the options market means that a good risk management model must not only fit the shape of the surface at a given date but also give realistic dynamics for co-movements of implied volatilities across strikes and maturities.

Market models of implied volatility (Babbar 2001; Carmona, Ma, and Nadtochiy 2017; Cohen, Reisinger, and Wang 2023; Cont and da Fonseca 2002; Cont, Fonseca, and Durrleman 2002; Gatheral and Jacquier 2014; Martini and Mingone 2022; Schönbucher 1999; Schweizer and Wissel 2008) attempt to directly model the cross-section and dynamics of implied volatilities. One of the challenges in modeling implied volatility surfaces is to

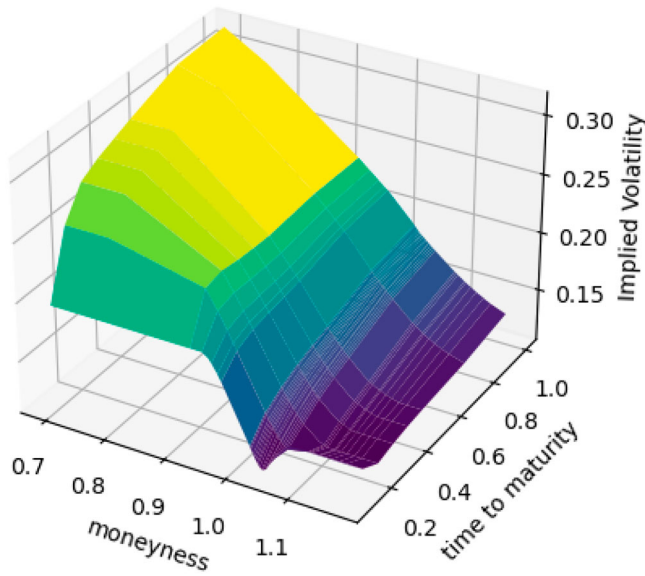


Figure 1. SPX implied volatility surface on 01/11/2021.

ensure that the absence of static arbitrage is satisfied. Indeed, the profile of the implied volatility surface cannot be arbitrary: static arbitrage constraints on the values of call and put options (Davis and Hobson 2007) put restrictions on the possible shape of the implied volatility surface. Analytical modeling has focused on obtaining parameterisations of implied volatility surfaces which guarantee that such arbitrage constraints are satisfied (Carmona, Ma, and Nadtochiy 2017; Cohen, Reisinger, and Wang 2023; Schweizer and Wissel 2008). Such models, however, are computationally challenging to implement, and even more challenging to calibrate to obtain realistic surface dynamics.

We present a computationally tractable method for simulating arbitrage-free implied volatility surfaces, which correctly captures the co-movements of implied volatility across a range of strikes and maturities. We first perform data analysis on the SPX implied volatility surface, and we then illustrate how our method may be combined with a factor model for the implied volatility surface to generate dynamic scenarios for arbitrage-free implied volatility surfaces. We give two examples: a stylized model using basis functions representing level, skew and curvature, and a data-driven example based on principal component analysis of daily changes in the logarithm of the SPX implied volatility surfaces. Our approach conciliates static arbitrage constraints with a realistic representation of statistical properties of implied volatility co-movements.

Outline Section 2 defines some notation for implied volatility surfaces and recalls some desired properties that market models of implied volatility have attempted to capture. In this section, we also perform data analysis of the SPX implied volatility and show that the co-movements can be captured by a small number of principal components. Section 3 recalls static arbitrage constraints on the implied volatility surface and introduces a penalty function for quantifying static arbitrage violations. We propose a Weighted Monte Carlo approach (Avellaneda et al. 2001) in Section 4 which prunes scenarios generated from a

base model using this penalty function. We illustrate in Section 5 how this approach may be applied to a factor model for the implied volatility surface (Cont and da Fonseca 2002).

2. Implied Volatility Surfaces

2.1. Properties of Implied Volatility Surfaces

Consider a market where (European) call and put options are traded on an underlying asset whose price we shall denote by S_t , across a range of strikes K and maturity dates T . The Black Scholes formula for the value of a call option with time to maturity $\tau = T - t$ and moneyness $m = K/S_t$ is:

$$C_{BS}(S_t, K, \tau, \sigma) = S_t N(d_1) - Ke^{-r\tau} N(d_2) \quad (1)$$

$$d_1 = \frac{-\ln m + \tau(r + \frac{\sigma^2}{2})}{\sigma\sqrt{\tau}} \quad d_2 = \frac{-\ln m + \tau(r - \frac{\sigma^2}{2})}{\sigma\sqrt{\tau}} \quad (2)$$

where $N(u) = (2\pi)^{-1/2} \int_{-\infty}^u \exp(-\frac{z^2}{2}) dz$.

Conversely, given the (observed) market price $C_t^*(K, T)$ of such a call option, the Black-Scholes implied volatility $\Sigma_t(K, T)$ is defined as the value of the volatility parameter which equates the market price with the Black-Scholes value:

$$\exists! \quad \Sigma_t(K, T) > 0, \quad C_{BS}(S, K, T - t, \Sigma_t(K, T)) = C_t^*(K, T) \quad (3)$$

From the implicit function theorem, one expects that in general Σ will depend on t, S, T, K (and of course on the randomness ω !). For fixed (K, T) , $\Sigma_t(K, T)$ is in general a stochastic process and, for fixed t , its value depends on the characteristics of the option: the maturity T and the strike level K . The function $\Sigma_t : (K, T) \rightarrow \Sigma_t(K, T)$ is called the *implied volatility surface* at date t . Using the *moneyness* $m = K/S_t$ of the option, one can also represent the implied volatility surface as a function of moneyness and maturity:

$$\sigma_t(m, \tau) = \Sigma_t(mS_t, t + \tau) \quad (4)$$

This representation is convenient since there is usually a range of moneyness around $m = 1$ for which the options are most liquid and therefore the empirical data is most readily available. The implied volatility surface today gives a snapshot of today's market prices of vanilla options: given the current term structure of interest rates and dividends, specifying the implied volatility surface is equivalent to specifying prices of all vanilla options quoted on the market.

A plethora of models have been proposed to model the instantaneous profile in (m, τ) of the implied volatility surface: local volatility models, jump-diffusion models and stochastic volatility models with or without jumps (Gatheral 2011). These 'smile' models are defined in terms of stochastic differential equations whose parameters describe the *infinitesimal* evolution of the asset price: since this evolution is not directly observed, *calibration* of model parameters to market prices of options turns out to be a non-trivial problem. However, even in cases where perfect calibration to today's option prices is achievable by a non-parametric model (for example, a local volatility model), getting a perfect fit of the implied volatility surfaces does not mean the model will generate realistic future scenarios.

This problem can be seen in the shape of the future smile (that is, the smile for forward-start options) generated by the model: many of these models, while giving good fits to today's implied volatility/ call prices generate unrealistic forms for future smiles, thus leading to a bias in prices of forward options.

Empirical studies of the behaviour of implied volatilities of exchange-traded options on various market indices (SP500, FTSE, DAX and others) point to many common statistical properties across markets (Avellaneda et al. 2020; Cont and da Fonseca 2002), which we summarize here and demonstrate on SPX data from 2000 to 2021 in the following subsection:

- (1) The implied volatility surface has a non-flat profile and exhibits both strike and term structure.
- (2) The shape of the implied volatility surface undergoes deformation in time.
- (3) Implied volatilities display high (positive) autocorrelation and mean-reverting behaviour.
- (4) The variance of the daily log-variations in implied volatility can be satisfactorily explained in terms of a small number of principal components.
- (5) The first principal component reflects an overall shift in the level of all implied volatilities.
- (6) The second principal component reflects opposite movements in (out of the money) call and put implied volatility.
- (7) The third and fourth principal components reflect the term structure and the changes in convexity of the implied volatility surface.
- (8) Global level shifts in implied volatility are negatively correlated with the returns of the underlying asset.
- (9) The projections of the surface on its principal components ('principal component processes') exhibit high (positive) autocorrelation and mean reversion. This autocorrelation structure is well represented by an AR(1)/ Ornstein Uhlenbeck process (Cont and da Fonseca 2002).

These dynamical properties of co-movements of implied volatilities and the underlying have important implications for hedging and should be reflected in any model used for risk management.

The possible shapes of implied volatility surfaces are limited by the arbitrage constraints on option prices. Call prices should be:

- increasing in time to maturity: $\partial_\tau C_{BS}(S_t, K, \tau, \sigma_t(m, \tau)) \geq 0$,
- decreasing in moneyness: $\partial_m C_{BS}(S_t, K, \tau, \sigma_t(m, \tau)) \leq 0$,
- convex in moneyness: $\partial_m^2 C_{BS}(S_t, K, \tau, \sigma_t(m, \tau)) \geq 0$.

These constraints translate to nonlinear inequalities involving σ_t , $\partial_m \sigma_t$, $\partial_m^2 \sigma_t$, $\partial_\tau \sigma_t$ (Cont, Fonseca, and Durrleman 2002). The resulting constraints on σ_t and the appropriate derivatives impose restrictions on the possible shapes.

Any option pricing model implies a dynamic model for the implied volatility surface. However, the corresponding shapes and dynamics are often intractable. 'Market models' of implied volatility aim to model implied volatility directly. The goal of such models has

been to correctly capture the co-movements of implied volatilities of options across different strikes and maturities while satisfying the no-arbitrage conditions, a challenging task, which has been capturing the attention of researchers for over 20 years. Statistical models of implied volatility dynamics (Avellaneda et al. 2020; Cont and da Fonseca 2002) have focused on correctly capturing the statistical properties of the market data and the co-movements across different strikes and maturities. These models are tractable and have been adopted for risk management applications, such as margin computations, but may lead to scenarios which are not compatible with arbitrage constraints. In parallel, analytical models have been developed with the goal of satisfying static (Gatheral and Jacquier 2014; Martini and Mingone 2022; Zhang, Li, and Zhang 2023), and dynamic arbitrage constraints (Carmona, Ma, and Nadtochiy 2017; Cohen, Reisinger, and Wang 2023; Schweizer and Wissel 2008; Wissel 2008). These models are computationally challenging to implement, simulate or estimate.

In the remainder of the paper, we describe an approach which aims to conciliate computational tractability, arbitrage constraints and realistic dynamics for the surface, and demonstrate its performance in two examples.

2.2. Case Study: Dynamics of the SPX Implied Volatility Surface

We consider a grid (m, τ) with 10 equispaced moneyness values between 0.6 and 1.4, and 8 time-to-maturity values of 30, 60, 91, 122, 152, 182, 273, 365 calendar days. We use daily time series of implied volatility for SPX options from the OptionMetrics SPX Implied Volatility Surface File for the period 2000-2021. Surfaces are interpolated linearly first in moneyness, and then in time to maturity to yield values on the grid (m, τ) . The average SPX implied volatility profile $\bar{\sigma}$ shown in Figure 2.

We note that the Implied Volatility Surface File is based on a previous interpolation of listed option prices so may not necessarily be arbitrage-free, as already noted in Cohen, Reisinger, and Wang (2023). We perform principal component analysis on the daily changes in the logarithm of the implied volatility surface

$$Y_t(m, \tau) = \log \sigma_t(m, \tau) \quad (5)$$

using a Karhunen-Loève decomposition (Cont and da Fonseca 2002). We denote by f_i the eigenvectors of the covariance operator of $\Delta Y_t = Y_{t+\Delta t} - Y_t$ ordered by decreasing eigenvalue. Each eigenvector may be represented as a function $(m, \tau) \mapsto f_i(m, \tau)$ of moneyness and time to maturity. We project $Y_t - \tilde{Y}$, where $\tilde{Y} = \log \bar{\sigma}(m, \tau)$, onto the eigenbasis:

$$Y_t(m, \tau) = \tilde{Y}(m, \tau) + \sum_{i=1}^k X_t^i f_i(m, \tau) + \epsilon_t(m, \tau), \quad (6)$$

where

$$X_t^i = \langle Y_t - \tilde{Y}, f_i \rangle = \sum_{(m, \tau) \in (m, \tau)} (Y_t(m, \tau) - \tilde{Y}(m, \tau)) f_i(m, \tau) \quad (7)$$

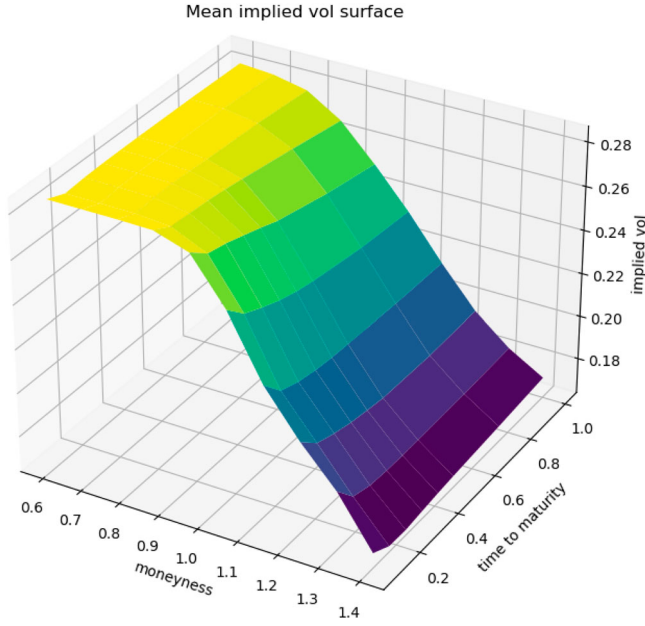


Figure 2. Average SPX implied volatility surface (2000–2021).

and $\epsilon_t(m, \tau)$ is the projection error. The rank- k approximation of implied volatility dynamics is given by

$$\sigma_t(m, \tau) \approx \bar{\sigma}(m, \tau) \exp \left[\sum_{i=1}^k X_t^i f_i(m, \tau) \right]. \quad (8)$$

2.2.1. Principal Component Analysis

To determine the number k of significant factors we consider the eigenvalues of the correlation matrix of the daily log-variations in the SPX implied volatility surface ΔY_t and compare them with the corresponding Marčenko-Pastur threshold λ_+ (Avellaneda et al. 2020; Dobi 2014):

$$\lambda_+ = \left(1 + \sqrt{\frac{N}{M}} \right)^2$$

where N is the number of points on the grid ($N = N_m N_\tau$), and M is the number of observations. We treat the eigenvalues below λ_+ as statistically insignificant.

As shown in Figure 3, there are $k = 4$ eigenvalues clearly above the Marčenko-Pastur threshold. The first four principal components of the daily changes in the log SPX implied volatility surface explain over 90% of the variance in the corresponding data (Table 1).

The first four eigenfunctions are shown in Figure 4. All four significant principal components f_1, f_2, f_3, f_4 (Equation (8)) have natural interpretations:

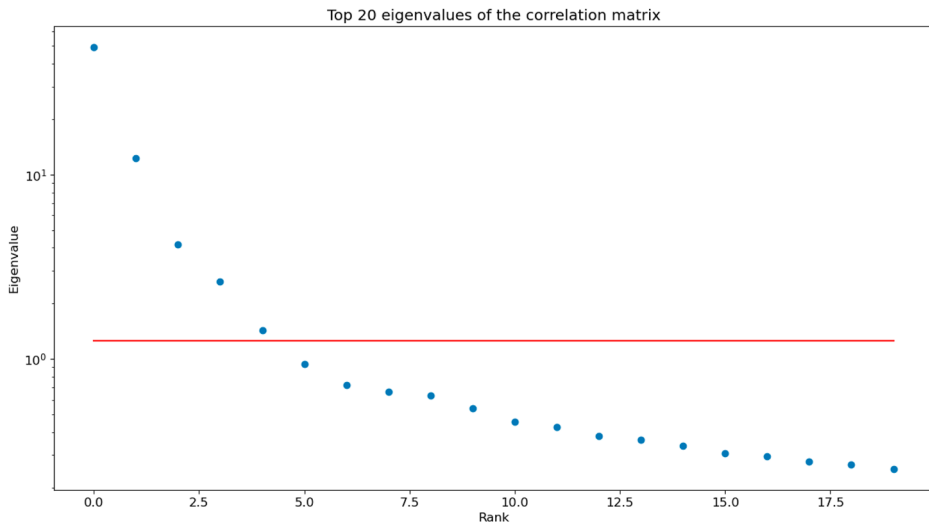


Figure 3. Eigenvalues of the correlation matrix of the daily changes in the log SPX implied volatility surface and the Marčenko-Pastur threshold λ_+ (in red).

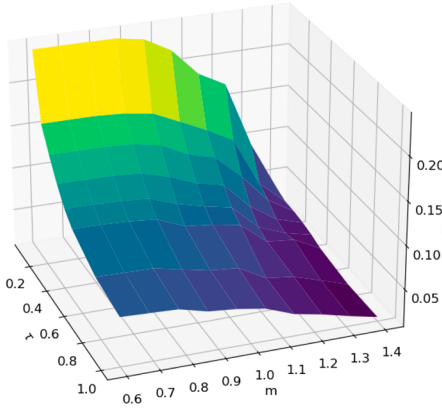
Table 1. Variance explained by the first five eigenvectors of the covariance and the correlation operator of the daily log returns of SPX implied volatilities.

	PC1	PC2	PC3	PC4	PC5
Variance explained (covariance)	68.88%	12.17%	5.67%	2.86%	1.41%
Cumulative variance explained (covariance)	68.88%	82.05%	87.72%	90.59%	92.00%
Cumulative variance explained (correlation)	61.39%	76.67%	81.88%	85.18%	86.96%

- The first principal component can be interpreted as the average *level* of implied volatilities. A positive shock along this mode would result in a global shift in implied volatility.
- The second principal component corresponds to the *skew*. Shocks along this direction result in opposite movements in out-of-the-money call and put implied volatilities.
- The third principal component reflects the *term structure* of implied volatilities.
- The fourth principal component can be interpreted as *curvature*. A positive shock along this mode would have an opposite effect on the implied volatilities close to the money and on the far out-of-the-money and in-the-money implied vols.

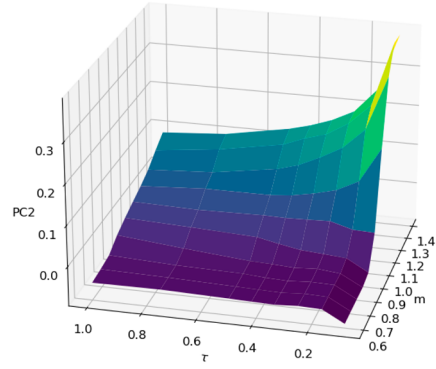
Projections of $Y_t(m, \tau) - \tilde{Y}_t(m, \tau)$ onto the first four principal components X_t^i , ($i = 1, \dots, 4$) (defined by Equation (7)) are shown in Figure 5. All processes exhibit high positive autocorrelation and mean-reverting behaviour with a period of mean reversion of several months. The ACF and PACF of X_t^1 are shown in Figure 6. As shown in Figure 6, the autocorrelation functions decay exponentially, suggesting that they can be modeled as Ornstein-Uhlenbeck/AR(1) processes, as already observed by Cont and da Fonseca (2002).

First principal component (log implied vol increments)



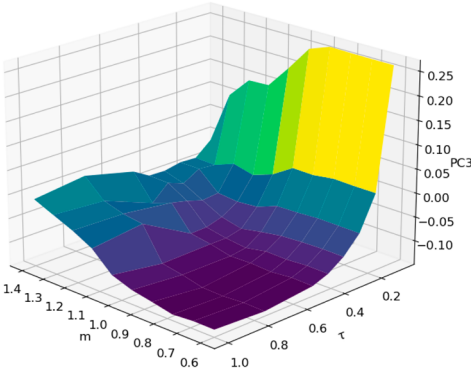
(a) Level

Second principal component (log implied vol increments)



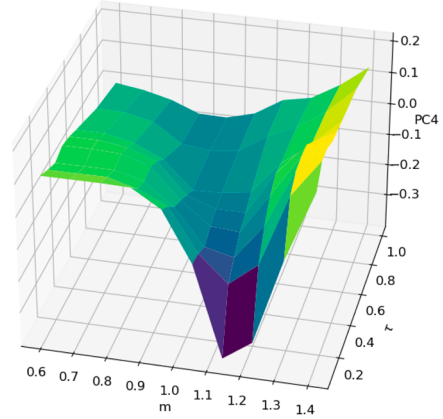
(b) Skew

Third principal component (log implied vol increments)



(c) Term structure

Fourth principal component (log implied vol increments)



(d) Curvature

Figure 4. The first four principal components of the daily changes in log SPX implied volatility surface. (a) Level (b) Skew (c) Term structure (d) Curvature.

Correlations of daily increments $\Delta X_t^i = X_{t+\Delta t}^i - X_t^i$ ($i = 1, \dots, 4$) and the log-returns of the underlying $R_t = \log S_{t+\Delta t} - \log S_t$ over a two-year rolling window are shown in Table 2. We note that the log-returns R_t are negatively correlated with $\Delta X_t^1, \Delta X_t^2, \Delta X_t^4$, while there is a positive correlation between the log-returns and the increments of the term-structure process ΔX^3 .

2.2.2. Relationship with the VIX

The CBOE Volatility Index (VIX) is constructed as a linear combination of out-of-the-money tradable calls and puts with one month to expiry (CBOE 2022). We investigate the

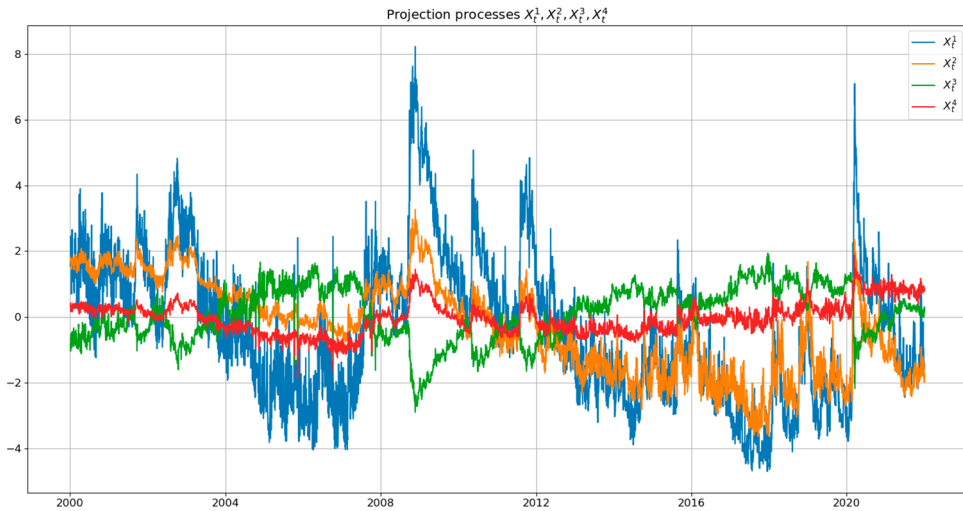


Figure 5. Principal component processes $X_t^1, X_t^2, X_t^3, X_t^4$.

Table 2. Long-term and 2-year rolling daily correlation between the log-increments of SPX and the increments of $X_t^1, X_t^2, X_t^3, X_t^4$ (Equation (7)).

Correlation between R_t and	ΔX_t^1	ΔX_t^2	ΔX_t^3	ΔX_t^4
Long-term	−38.21%	−25.64%	26.42%	−26.39%
Rolling: mean	−48.83%	−37.75%	31.60%	−32.63%

relationship between the VIX and different variables of interest by considering the historical closing VIX prices available on the CBOE website. Figure 7 displays the correlations between the log returns of VIX, one-month at-the-money SPX implied volatility, SPX and the increments of the level process X_t^1 over a rolling 2-year window. We note a high positive correlation between the level movements and the log-returns of VIX and ATM vol. Similarly, the returns of the underlying are negatively correlated with the increments of the level process (Table 2), log-returns of VIX and with the log-returns of the ATM vol. Correlations increasing in magnitude from 2006 onwards.

The one-month realized volatility $\hat{\sigma}_t$ is estimated as

$$\hat{\sigma}_t = \sqrt{\frac{21}{252} \sum_{i=0}^{20} R_{t-i\Delta t}^2}. \quad (9)$$

Figure 8 shows that the realized volatility is usually below the implied volatility. The average ratio of realized volatility to ATM volatility is 0.59, with a standard deviation of 0.209 (Figure 8).

3. Static Arbitrage Constraints

We now consider shape constraints on the implied volatility surfaces arising from static arbitrage inequalities (Davis and Hobson 2007). We are interested in a realistic setting

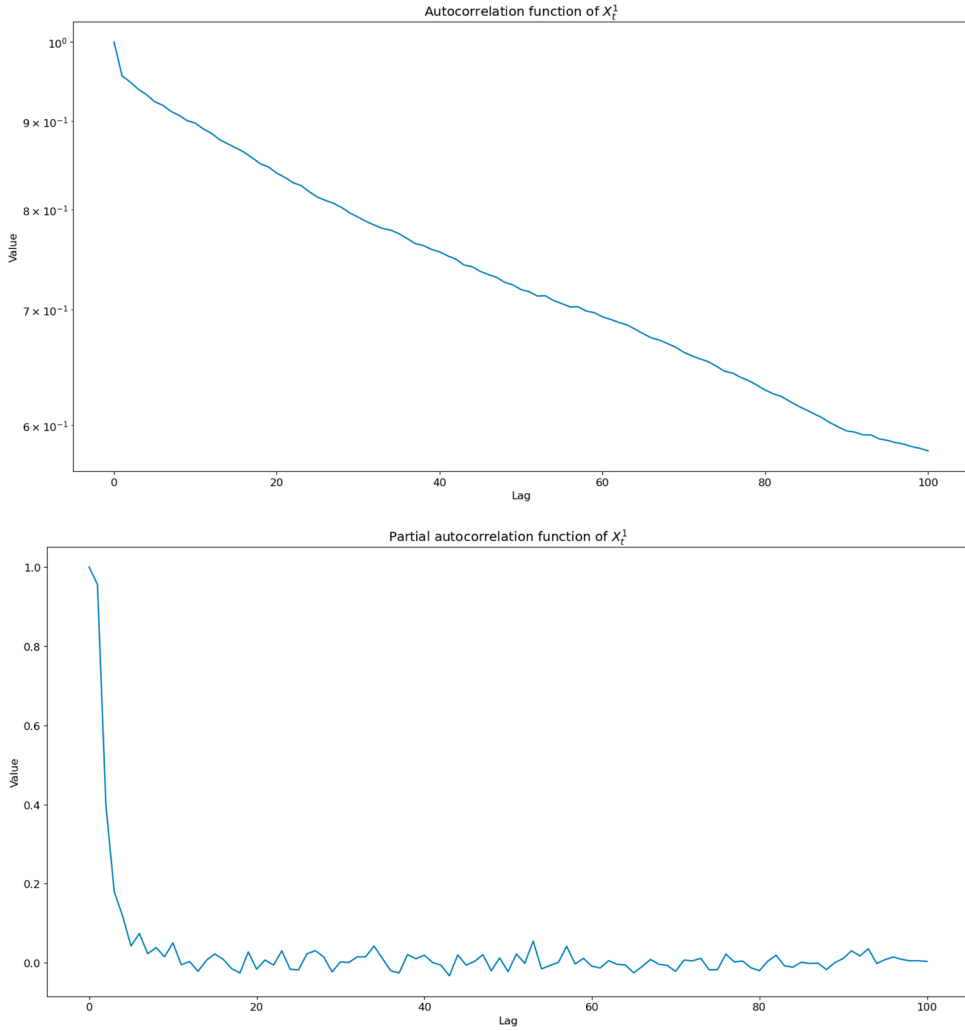


Figure 6. Autocorrelation and partial autocorrelation of the log implied volatility projection on the first principal component. The autocorrelation function (above) in logarithmic scale shows an exponential decay characteristic of OU processes.

where only a finite number of options are available. We fix a grid in moneyness and time to maturity $(\mathbf{m}, \boldsymbol{\tau}) = (m_i, \tau_j)_{i=1, \dots, N_m; j=1, \dots, N_\tau}$, with $m_i < m_{i+1}$ and $\tau_j < \tau_{j+1}$ for all i, j . Using the notation introduced in Section 2, denote by

$$c(m, \tau) := \frac{1}{S} C_{BS}(S, K, \tau, \sigma) = N(d_1) - me^{-r\tau} N(d_2)$$

the relative call price, which is a dimensionless quantity with $0 \leq c(m, \tau) \leq 1$.

3.1. Arbitrage Constraints and Arbitrage Penalty

As shown by Davis and Hobson (Corollaries 4.2 and 4.3 in Davis and Hobson 2007), absence of static arbitrage among options with strikes and maturities defined by $(\mathbf{m}, \boldsymbol{\tau})$ is equivalent to the following three conditions:

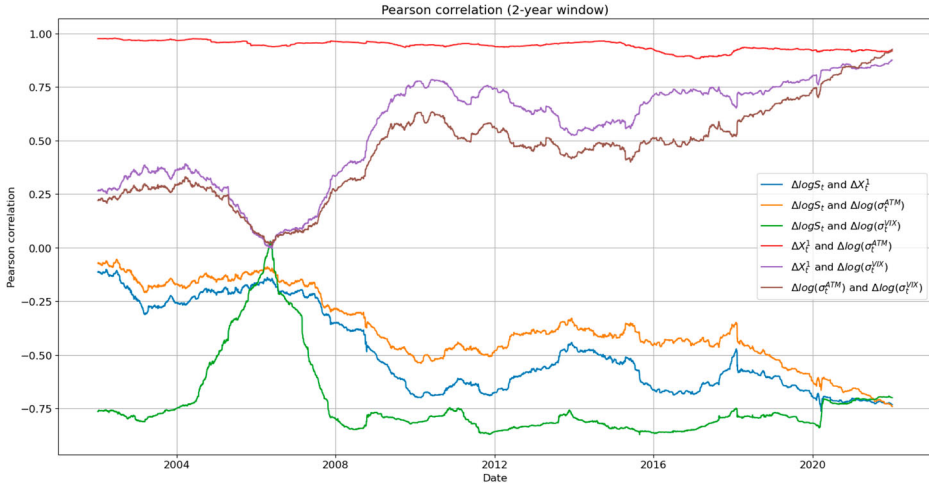


Figure 7. Correlation over a 2-year window between daily changes in the level process X_t^1 and daily log-returns of one-month ATM vol, VIX, and SPX.

(1) Absence of calendar spread arbitrage:

$$\tau_j \frac{c(m_i, \tau_j) - c(m_i, \tau_{j+1})}{\tau_{j+1} - \tau_j} \leq 0, \quad (10)$$

for $j = 1, \dots, N_\tau - 1$ and $i = 1, \dots, N_m$.

(2) Absence of call spread arbitrage:

$$\frac{c(m_{i+1}, \tau_j) - c(m_i, \tau_j)}{m_{i+1} - m_i} \leq 0 \quad (11)$$

for $i = 1, \dots, N_m - 1$ and $j = 1, \dots, N_\tau$.

(3) Absence of butterfly spread arbitrage:

$$\frac{c(m_i, \tau_j) - c(m_{i-1}, \tau_j)}{m_i - m_{i-1}} - \frac{c(m_{i+1}, \tau_j) - c(m_i, \tau_j)}{m_{i+1} - m_i} \leq 0 \quad (12)$$

for $i = 2, \dots, N_m - 1$ and $j = 1, \dots, N_\tau$.

Conversely, a non-zero positive part of the left-hand side of these inequalities indicates the presence of static arbitrage. We investigate whether an implied volatility surface $\sigma(\mathbf{m}, \boldsymbol{\tau})$ is arbitrage-free by considering the inequalities (10), (11) and (12).

Hence, for an implied volatility surface $\sigma(\mathbf{m}, \boldsymbol{\tau})$, we define the arbitrage penalty $\Phi(\sigma(\mathbf{m}, \boldsymbol{\tau}))$ as

$$\Phi(\sigma(\mathbf{m}, \boldsymbol{\tau})) = p_1(\sigma(\mathbf{m}, \boldsymbol{\tau})) + p_2(\sigma(\mathbf{m}, \boldsymbol{\tau})) + p_3(\sigma(\mathbf{m}, \boldsymbol{\tau})), \quad (13)$$

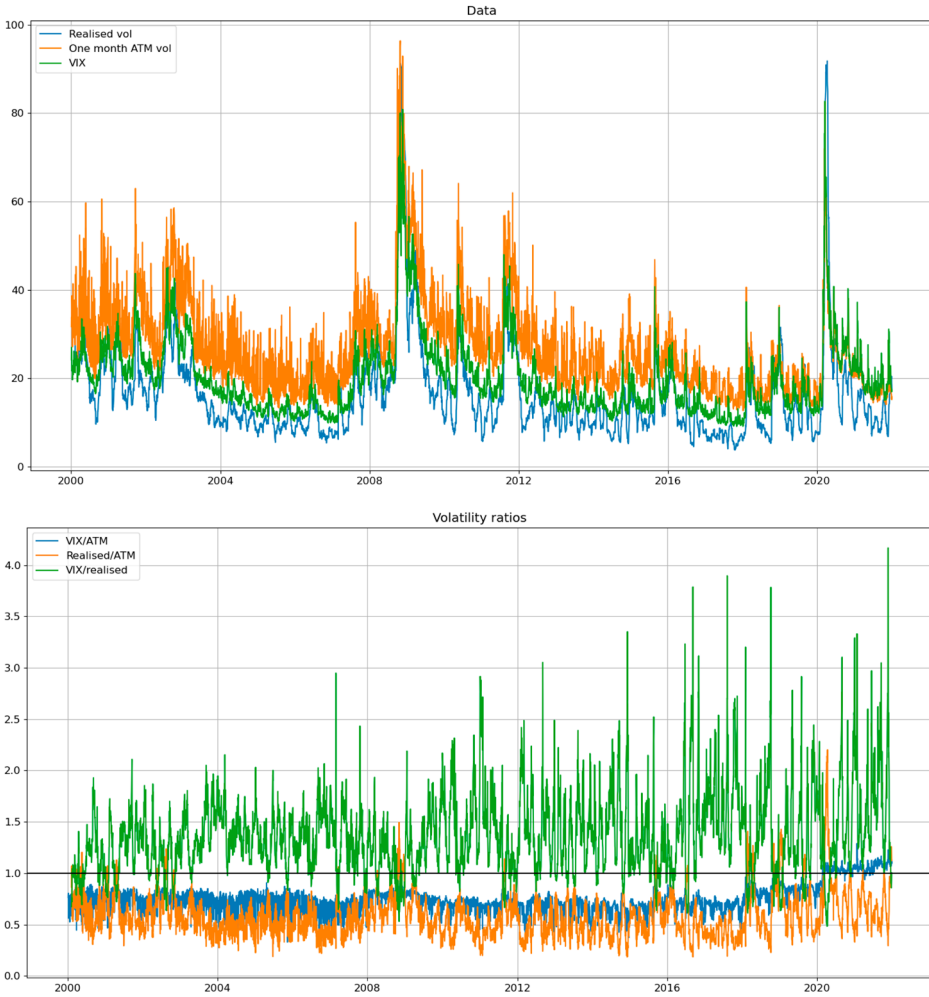


Figure 8. Above: SPX realized volatility (blue), one-month ATM volatility (orange) and VIX (green). Below: ratio of VIX to one-month ATM volatility (blue) and ratios of 21-day realized volatility to one-month ATM volatility, VIX to ATM volatility, and VIX to realized volatility.

where

$$p_1(\sigma(\mathbf{m}, \boldsymbol{\tau})) = \sum_{i=1}^{N_m} \sum_{j=1}^{N_\tau} \left(\tau_j \frac{c(m_i, \tau_j) - c(m_i, \tau_{j+1})}{\tau_{j+1} - \tau_j} \right)^+, \quad (14)$$

$$p_2(\sigma(\mathbf{m}, \boldsymbol{\tau})) = \sum_{i=1}^{N_m} \sum_{j=1}^{N_\tau} \left(\frac{c(m_{i+1}, \tau_j) - c(m_i, \tau_j)}{m_{i+1} - m_i} \right)^+, \quad (15)$$

$$p_3(\sigma(\mathbf{m}, \boldsymbol{\tau})) = \sum_{i=1}^{N_m} \sum_{j=1}^{N_\tau} \left(\frac{c(m_i, \tau_j) - c(m_{i-1}, \tau_j)}{m_i - m_{i-1}} - \frac{c(m_{i+1}, \tau_j) - c(m_i, \tau_j)}{m_{i+1} - m_i} \right)^+. \quad (16)$$

The quantities p_1, p_2, p_3 correspond to deviations from the calendar, call, and butterfly spread arbitrage constraints, respectively. They are the positive parts of the left-hand sides of the inequalities (10), (11), and (12). If p_1, p_2, p_3 are all equal to zero, there is no arbitrage. Conversely, if any of p_1, p_2, p_3 are non-zero, then there is arbitrage present in $\sigma(\mathbf{m}, \tau)$. Therefore,

$$\Phi(\sigma(\mathbf{m}, \tau)) = 0 \iff \sigma(\mathbf{m}, \tau) \text{ is arbitrage-free.}$$

We introduce the $N_m \cdot N_\tau$ penalty matrices P_1, P_2, P_3 defined as

$$(P_1)_{ij} = \left(\tau_j \frac{c(m_i, \tau_j) - c(m_i, \tau_{j+1})}{\tau_{j+1} - \tau_j} \right)^+, \quad (P_2)_{ij} = \left(\frac{c(m_{i+1}, \tau_j) - c(m_i, \tau_j)}{m_{i+1} - m_i} \right)^+ \\ (P_3)_{ij} = \left(\frac{c(m_i, \tau_j) - c(m_{i-1}, \tau_j)}{m_i - m_{i-1}} - \frac{c(m_{i+1}, \tau_j) - c(m_i, \tau_j)}{m_{i+1} - m_i} \right)^+,$$

with the appropriate endpoints being set to zero: $(P_1)_{i, N_\tau} = 0$ for $i = 1, \dots, N_m$, $(P_2)_{N_m, j} = 0$, $(P_3)_{N_m, j} = (P_3)_{0, j} = 0$ for $j = 1, \dots, N_\tau$. The arbitrage penalty may then be expressed as the 1-norm of the matrix $P_1 + P_2 + P_3$:

$$\Phi(\sigma(\mathbf{m}, \tau)) = \sum_{i=1}^{N_m} \sum_{j=1}^{N_\tau} (P_1 + P_2 + P_3)_{ij} = \|P_1 + P_2 + P_3\|_1.$$

Remark 3.1 (Extension to swaption implied volatility cube): When working with swaptions, for every tenor T^a there is an implied volatility surface $\sigma^a(\mathbf{m}, \tau)$. Hence, one could calculate the arbitrage penalty for each possible surface (across all of the available tenors) in order to reach an aggregated penalty for the swaption implied volatility cube. That is, suppose that we have available tenors T^1, \dots, T^A . Then we may define the arbitrage penalty for the swaption implied volatility cube $\{\sigma_t^a(\mathbf{m}, \tau)\}_{a=1 \dots A}$ by

$$\Phi^1(\{\sigma_t^a(\mathbf{m}, \tau)\}_{a=1 \dots A}) = \sum_{a=1}^A \Phi(\sigma_t^a(\mathbf{m}, \tau)), \quad (17)$$

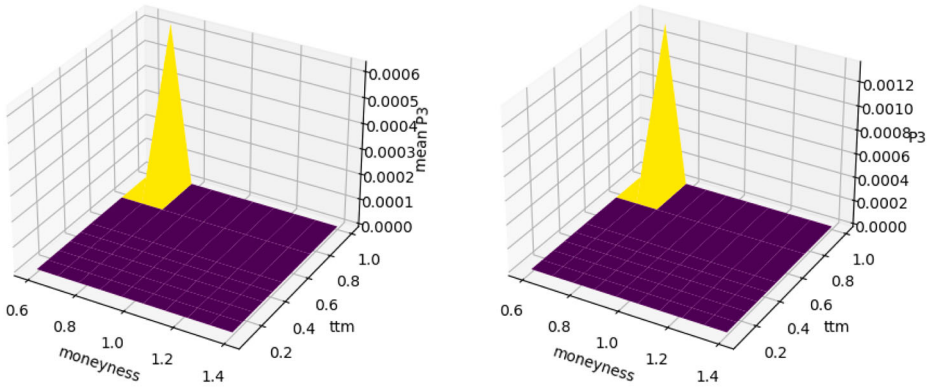
or by

$$\Phi^\infty(\{\sigma_t^a(\mathbf{m}, \tau)\}_{a=1 \dots A}) = \max_{a=1, \dots, A} \Phi(\sigma_t^a(\mathbf{m}, \tau)). \quad (18)$$

3.2. Behaviour of Arbitrage Penalty Under Perturbations

To gain intuition about the properties of arbitrage penalty (13) we investigate its behaviour under perturbations of an arbitrage-free implied volatility surface by IID noise and parallel shifts. In the numerical results below, the initial implied volatility surface is taken to be the SPX implied volatility surface on 31/12/2021.

Addition of IID noise We sample 10,000 implied volatility surfaces by adding IID noise (i.e. ,independent across strike and maturity) with a standard deviation of $\epsilon = 0.001$ to the initial arbitrage-free SPX implied volatility surface. We observe that 23% of generated surfaces exhibit butterfly spread arbitrage. The mean butterfly arbitrage penalty matrix P_3 is



(a) Mean effect of adding noise. (b) Sample parallel shift effect.

Figure 9. Butterfly penalty matrices (P_3) arising from noise and parallel shifts. (a) Mean effect of adding noise (b) Sample parallel shift effect.

displayed in Figure 9. We note that violations occur only for far from the money, long-dated options.

Parallel shifts Rogers and Tehranchi (2010) showed that moving implied volatility surfaces by parallel shifts will eventually result in configurations with static arbitrage. We explore this phenomenon quantitatively by adding a parallel shift to an initial arbitrage-free implied volatility surface (SPX implied volatility surface on 31/12/2021) and testing for static arbitrage. The absolute value of the largest negative shift is taken to be smaller than the smallest implied volatility value, guaranteeing non-negativity. The effect of parallel shifts on the arbitrage penalty are displayed in Figure 10: arbitrage constraints are violated for large enough positive shifts, and the arbitrage penalty grows linearly thereafter. For such large shifts, the constraint which is violated is convexity. A sample butterfly arbitrage penalty matrix P_3 is displayed in Figure 9(b). These results give a quantitative perspective on the results of Rogers and Tehranchi (2010).

3.3. Arbitrage Penalty in SPX Implied Volatility Data

Data sources on implied volatility, such as OptionMetrics, are often interpolated from actual market quotes, a procedure which may itself introduce static arbitrage. This has been previously noted by several studies, see e.g. Cohen, Reisinger, and Wang (2023). We study this phenomenon using daily SPX implied volatility surfaces from 2000 to end 2021. We observe non-zero arbitrage penalties, with a decomposition displayed in Figure 11 and Table 3. 90.5% of dates display no calendar arbitrage, 97.3% display no call spread arbitrage and 84.9% no butterfly arbitrage. Overall 80.2% of the observations correspond to arbitrage-free surfaces.

We observe a number of spikes in arbitrage penalties. The two largest spikes happen on 29/09/2008 (during the 2008 financial crisis) and on 13/03/2020 (the start of the Covid-19 pandemic). Figure 11 shows that the majority of arbitrage violations happen during the 2008 financial crisis and during the start of the Covid-19 pandemic. For comparisons, the

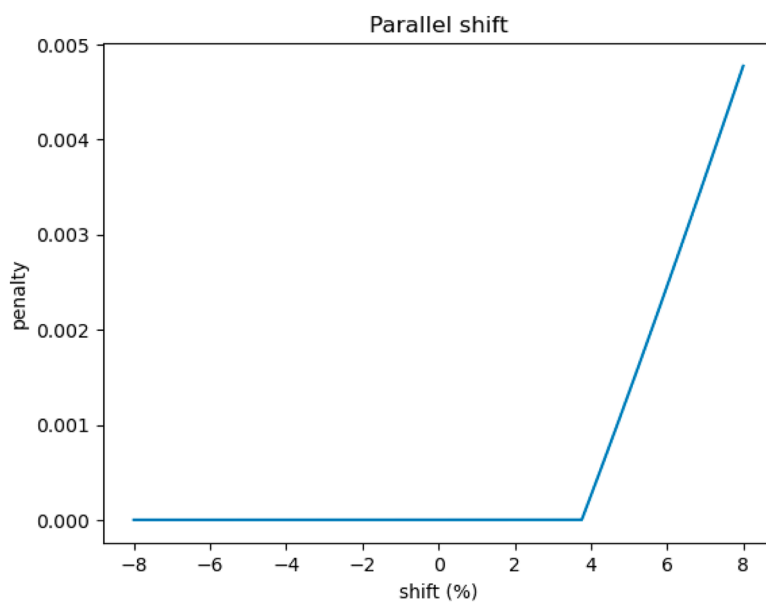


Figure 10. Arbitrage violations induced by parallel shifts on SPX implied volatility surface (31/12/2021).

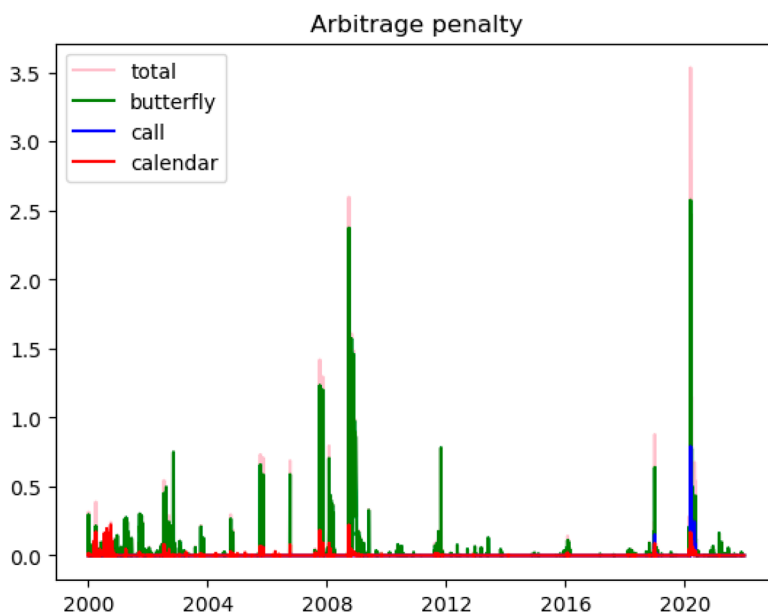
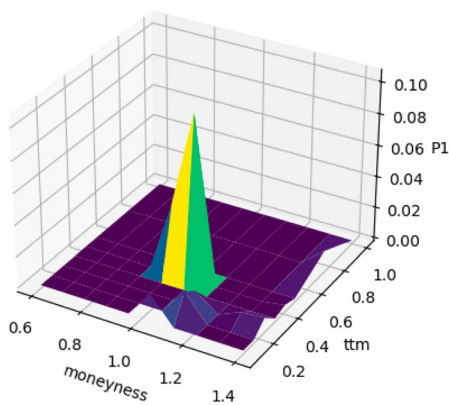


Figure 11. Arbitrage penalty decomposition for SPX options.

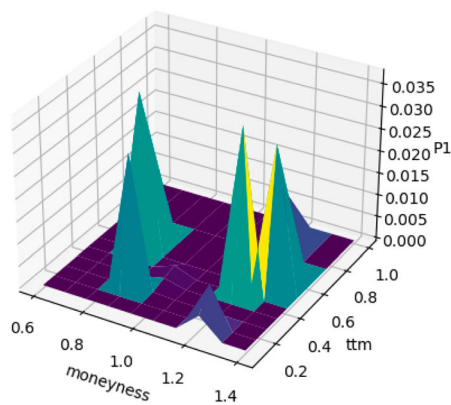
calendar, call, and butterfly arbitrage penalty matrices P_1, P_2, P_3 on dates 29/09/2008 and 13/03/2020 are displayed in Figures 12–14, respectively.

Table 3. Quantiles of arbitrage penalties for SPX options.

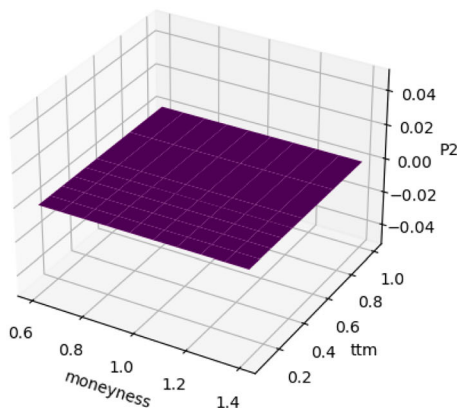
Penalty	Median	90th quantile	95th quantile	99th quantile
Total Φ	0	0.075	0.13	0.5
Calendar spread p_1	0	0	0.002	0.038
Call spread p_2	0	0	0	0.009
Butterfly spread p_3	0	0.01	0.06	0.458



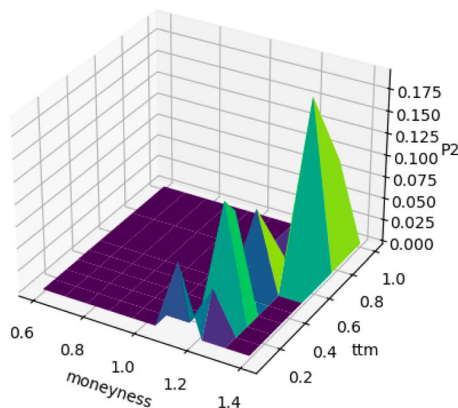
(a) 29.09.2008.



(b) 13.03.2020.

Figure 12. Calendar spread arbitrage (P1) for SPX options. (a) 29.09.2008 (b) 13.03.2020.

(a) 29.09.2008.



(b) 13.03.2020.

Figure 13. Call spread arbitrage (P2) for SPX options. (a) 29.09.2008 (b) 13.03.2020.

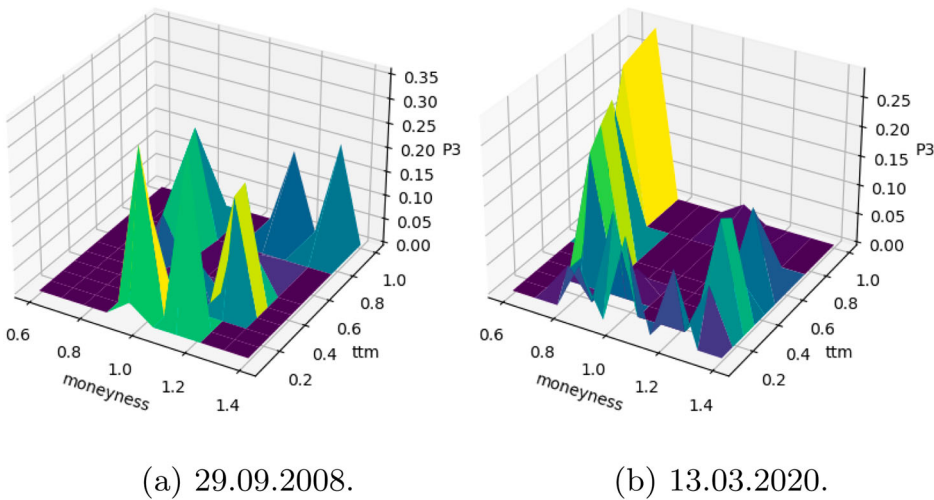


Figure 14. Butterfly spread arbitrage (P3) for SPX options. (a) 29.09.2008 (b) 13.03.2020.

4. Penalizing Static Arbitrage

4.1. Penalization Via Scenario Reweighting

Our starting point is a baseline model \mathbb{P}_0 for implied volatility surface dynamics, which correctly captures the co-movements and statistical properties of implied volatilities and the underlying asset, but may not necessarily be arbitrage-free. For example, this may be a factor model based on PCA, such as Avellaneda et al. (2020) and Cont and da Fonseca (2002).

We are interested in generating market scenarios over a time grid $\mathbb{T} = \{0, \dots, t_{\max}\}$. \mathbb{P}_0 may be a discrete-time or continuous-time model. For ease of notation, we will continue to denote by \mathbb{P}_0 the joint law of the variables $(S_t, \sigma_t(\mathbf{m}, \tau), t \in \mathbb{T})$ under \mathbb{P}_0 .

Our idea is to penalize arbitrage along the paths generated by \mathbb{P}_0 by ‘tilting’ the probabilities associated with such paths. We choose $\beta > 0$ and define a new probability \mathbb{P}_β on the space of market scenarios by

$$\frac{d\mathbb{P}_\beta}{d\mathbb{P}_0}(\omega) = \frac{\exp\left(-\beta \sum_{t \in \mathbb{T}} \Phi(\sigma_t(\mathbf{m}, \tau; \omega))\right)}{Z(\beta)} \quad (19)$$

where $Z(\beta)$ is a normalization factor:

$$Z(\beta) = \mathbb{E}^{\mathbb{P}_0} \left[\exp \left(-\beta \sum_{t \in \mathbb{T}} \Phi(\sigma_t(\mathbf{m}, \tau)) \right) \right]. \quad (20)$$

If the baseline model \mathbb{P}_0 is arbitrage-free then $\Phi(\sigma_t(\mathbf{m}, \tau)) = 0$ \mathbb{P}_0 -almost surely so $Z(\beta) = 1$ and $\mathbb{P}_\beta = \mathbb{P}_0$. If, however, \mathbb{P}_0 generates surfaces which violate static arbitrage constraints, then the change of measure (19) penalizes such scenarios, and may be thought of as an importance sampling method which penalizes static arbitrage violations. This penalization increases if we take large β and as $\beta \rightarrow \infty$ we reject all scenarios violating static arbitrage constraints. Thus one may think of $1/\beta$ as a ‘tolerance’ for static arbitrage.

Note that \mathbb{P}_β is absolutely continuous with respect to \mathbb{P}_0 so we are keeping the same paths but re-weighting them. In the case where the dynamics of variables are given by stochastic differential equations driven by Brownian motion, Girsanov's theorem implies that the re-weighting will impact the drift, but not the quadratic covariation of the variables. However, our approach does not assume that variables are driven by Brownian motion factors, and may be applied in a more general setting. Indeed, the whole procedure also makes sense for a discrete-time time series model.

4.2. A 'Weighted Monte Carlo' Approach

We now propose a method for sampling from \mathbb{P}_β , using a Weighted Monte Carlo approach (Avellaneda et al. 2001). We proceed as follows:

- Simulate N independent paths $(\omega_i, i = 1, \dots, N)$ from \mathbb{P}_0 . Each path corresponds to the joint evolution of the underlying asset and the implied volatility surface:

$$\omega_i = (S_t(\omega_i), \sigma_t(\mathbf{m}, \boldsymbol{\tau}; \omega_i); t \in \{0, \dots, t_{\max}\})$$

- Compute the arbitrage penalty $\varphi(\omega_i)$ along each path:

$$\varphi(\omega_i) = \sum_{t \in \{0, \dots, t_{\max}\}} \Phi(\sigma_t(\mathbf{m}, \boldsymbol{\tau}; \omega_i)). \quad (21)$$

- Associate a weight $w_i(\beta)$ with each path:

$$w_i(\beta) = \frac{\exp(-\beta\varphi(\omega_i))}{\sum_{j=1}^N \exp(-\beta\varphi(\omega_j))}. \quad (22)$$

- Sample from the **weighted** model \mathbb{P}_β^N defined as

$$\mathbb{P}_\beta^N(\omega_i) = w_i(\beta) \quad i = 1, \dots, N. \quad (23)$$

That is, instead of sampling each simulated path ω_i with probability $\frac{1}{N}$, sample it with probability $w_i(\beta)$.

This step-by-step procedure is summarized in Table 4.

Note that we keep the same paths but modify their weight. Thus, all quantities computed along the path, such as realized volatility and realized covariances will remain the same.

As $\beta \rightarrow \infty$, $w_i(\beta) \rightarrow 0$ as soon as $\varphi(\omega_i) > 0$ so only paths with arbitrage-free implied volatility surfaces survive for large β . Hence, $\frac{1}{\beta}$ can be viewed as an *arbitrage tolerance* parameter.

If the model \mathbb{P}_0 is arbitrage-free, then the re-weighting will have no influence as for every $\beta > 0$ we will have $w_i(\beta) = \frac{1}{N}$, implying that \mathbb{P}_β^N is simply the empirical distribution associated with the N simulated paths. In the general case, the relative entropy of \mathbb{P}_β^N with respect to this empirical distribution (i.e., the uniform distribution on $\{\omega_i, i = 1, \dots, N\}$) is

Table 4. Weighted Monte Carlo for implied volatility scenarios.**Ingredients**

- ‘Baseline model’ \mathbb{P}_0 for implied volatility surface dynamics.
- Time grid $\mathbb{T} = \{0 = t_0 < t_1 < \dots < t_{\max}\}$.
- Moneyness and time to maturity grid $(\mathbf{m}, \boldsymbol{\tau}) = (m_i, \tau_j)_{i=1, \dots, N_m; j=1, \dots, N_\tau}$.
- Number of paths N .
- Arbitrage penalty parameter $\beta > 0$.

Step 1: Simulate N independent scenarios $\omega_i, i = 1, \dots, N$ from the baseline model \mathbb{P}_0 . Each scenario ω_i represents a joint evolution of the underlying asset S_t and the implied volatility surface $\sigma_t(\mathbf{m}, \boldsymbol{\tau})$ for $t \in \{0, \dots, t_{\max}\}$:

$$\omega_i = (S_t(\omega_i), \sigma_t(\mathbf{m}, \boldsymbol{\tau}; \omega_i); t \in \{t_0, \dots, t_{\max}\}).$$

Step 2: For each simulated path $\sigma(\mathbf{m}, \boldsymbol{\tau})^i$, compute the arbitrage penalty

$$\varphi(\omega_i) = \sum_{t \in \mathbb{T}} \Phi(\sigma_t(\mathbf{m}, \boldsymbol{\tau}; \omega_i)).$$

Step 3: If $\varphi(\omega_i) = 0$ for all $i = 1 \dots N \rightarrow \text{STOP}$, else.

Step 4: Compute the weights

$$w_i(\beta) = \frac{\exp(-\beta \varphi(\omega_i))}{\sum_{j=1}^N \exp(-\beta \varphi(\omega_j))}.$$

Step 5: Compute the relative entropy $\mathcal{E}_N(\beta) = H(\mathbb{P}_\beta^N | \mathbb{P}_0^N)$.

Step 6: Sample the scenarios $\omega_i, i = 1, \dots, N$ with probability $w_i(\beta)$:

$$\mathbb{P}_\beta^N(\omega_i) = w_i(\beta).$$

an indicator of the ‘distance to no-arbitrage’:

$$\mathcal{E}_N(\beta) = H(\mathbb{P}_\beta^N | \mathbb{P}_0^N) = -N \ln N - N \sum_{i=1}^N w_i(\beta) \ln w_i(\beta). \quad (24)$$

When there is no static arbitrage in the scenarios ω_i generated by \mathbb{P}_0 , then relative entropy is zero: $\mathcal{E}_N(\beta) = 0$. On the other hand, the model \mathbb{P}_0 is far from being arbitrage-free, the arbitrage penalties $\varphi(\omega_i)$ are large and the relative entropy $\mathcal{E}_N(\beta)$ will be large.

Our approach is more efficient than rejection sampling, as we sample a fixed number of paths, regardless of the initial model \mathbb{P}_0 , so the complexity is of order $O(N)$. If the scenarios generated by \mathbb{P}_0 are likely to admit static arbitrage, even if the penalty is small and arises from interpolation, rejection sampling may result in an infinite loop.

The following result, which we state for completeness, clarifies the relation between the various probability measures involved. We use the notation of Section 4.1.

Proposition 4.1: (i) \mathbb{P}_β^N weakly converges to \mathbb{P}_β as the number of scenarios $N \rightarrow \infty$.
(ii) Let $U = \{\omega \in \Omega, \varphi(\omega) = 0\}$ be the set of scenarios free of static arbitrage. If $\mathbb{P}_0(U) > 0$ then, as $\beta \rightarrow \infty$, the support of \mathbb{P}_β concentrates on U :

$$\forall \varepsilon > 0, \quad \mathbb{P}_\beta(\{\varphi > \varepsilon\}) \xrightarrow{\beta \rightarrow \infty} 0. \quad (25)$$

Proof: (i) is a consequence of the weak law of large numbers. To show (ii), first note that if \mathbb{P}_0 is supported on arbitrage-free scenarios, then so is \mathbb{P}_β . Hence, suppose that \mathbb{P}_0 is not supported on the (closed) set $U = \{\omega \in \Omega, \varphi(\omega) = 0\}$. The arbitrage penalty

$$\varphi : \omega \in \Omega \mapsto \sum_{t \in \mathbb{T}} \Phi(\sigma_t; \omega)$$

defines a random variable on scenario space and $U = \{\varphi = 0\} \in \mathcal{F}_{t_{\max}}$. Note that since Φ defined by (13) is bounded, so is φ . Define

$$A_n = \{\omega \in \Omega, \varphi(\omega) > 1/n\} \in \mathcal{F}_{t_{\max}}.$$

Then $A = U^c = \{\varphi > 0\} = \cap_{n \geq 1} A_n$. If \mathbb{P}_0 is not supported on U , there exists $n \geq 1$ such that $\mathbb{P}_0(A_n) > 0$.

$$\mathbb{P}_\beta(A_n) = \frac{\int_{A_n} \exp(-\beta\varphi(\omega)) d\mathbb{P}_0(\omega)}{Z(\beta)}.$$

Since $\varphi = 0$ on $U \subset A_n^c$, we have

$$Z(\beta) = \int_{U^c} \exp(-\beta\varphi(\omega)) d\mathbb{P}_0(\omega) + \int_U d\mathbb{P}_0(\omega) \geq \mathbb{P}_0(U)$$

Also, since $\varphi > 1/n$ on A_n , we have

$$\int_{A_n} \exp(-\beta\varphi(\omega)) d\mathbb{P}_0(\omega) \leq \mathbb{P}_0(A_n) \exp(-\beta/n) \rightarrow 0 \quad \text{as } \beta \rightarrow \infty.$$

Hence,

$$\mathbb{P}_\beta(A_n) \leq \frac{\mathbb{P}_0(A_n) \exp(-\beta/n)}{\mathbb{P}_0(U)} \rightarrow 0 \quad \text{as } \beta \rightarrow \infty.$$

Taking $n > 1/\varepsilon$ yields the result. ■

5. Factor Models for Implied Volatility Dynamics

In order to illustrate our approach, we consider factor models for implied volatility dynamics as the baseline model \mathbb{P}_0 . We first simulate scenarios from the three-factor model introduced in Cont and da Fonseca (2002), and then from a four-factor model for the SPX implied volatility surface based on our previous analysis in Section 2.

5.1. Example: A Stylized Factor Model for Implied Volatility

We first consider a three-factor for implied volatility dynamics introduced in Cont and da Fonseca (2002), based on a Karhunen-Loeve decomposition of co-movements in implied volatilities. The evolution of implied volatility surface paths in this model is given by

$$\sigma_t(m, \tau) = \sigma_0(m, \tau) \exp(x_t^1 f_1(m, \tau) + x_t^2 f_2(m, \tau) + x_t^3 f_3(m, \tau)) \quad (26)$$

where the factors x_t^1, x_t^2, x_t^3 correspond to *level*, *skew* and *curvature*, as projections on principal components with the analogous representations (Cont and da Fonseca 2002). Their dynamics are modeled as independent Ornstein-Uhlenbeck processes:

$$dx_t^i = \lambda_i (\alpha_i - x_t^i) dt + \gamma_i dW_t^i, \quad (27)$$

where W_t^i are independent Brownian motions. The basis functions f_1, f_2, f_3 are the first three principal components of the log-implied volatility surface. The price of the underlying asset S is modeled as a diffusion with stochastic volatility $\sigma_t(1, 0)$, which corresponds

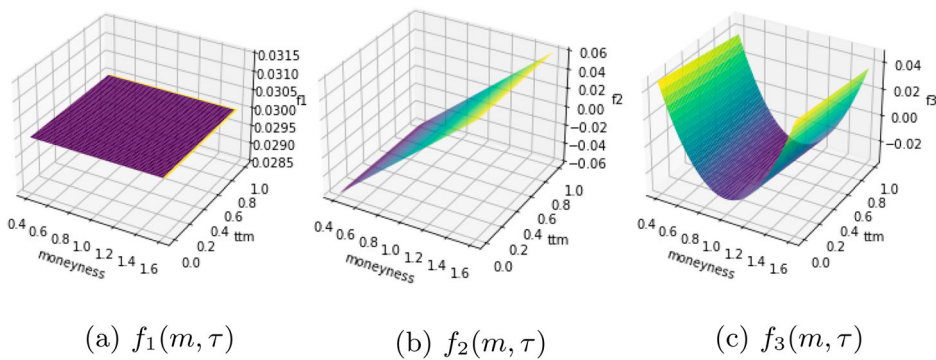


Figure 15. Basis functions f_1, f_2, f_3 corresponding to level, skew and curvature used to simulate scenarios from the factor model (26)–(27). (a) $f_1(m, \tau)$ (b) $f_2(m, \tau)$ (c) $f_3(m, \tau)$.

to the short-term at-the-money implied volatility:

$$dS_t = \sigma_t(1, 0)S_t dW_t^0, \quad W_t^0 = \rho W_t^1 + \sqrt{1 - \rho^2} B_t, \quad (28)$$

where $\rho < 0$ and B is a Brownian motion independent from $W^i, i = 1, 2, 3$. The increments of the first factor x_t^1 are negatively correlated with the returns of the underlying asset: $\text{cov}(W_t^0, W_t^1) = \rho t < 0$. We use $\rho = -0.5$ and $r = 0$ as an example.

Given that this model is based on a principal component analysis of market data, the simulated paths correctly capture the covariance structure of implied volatility comovements. Furthermore, the functional form (26) of the implied volatility surface guarantees smoothness of the surface and continuity of simulated paths.

In the first example, we suppose $\alpha_i = 0, \gamma_i = 1$ and use the coefficients λ_i given in Cont and da Fonseca (2002) for the SPX implied volatilities to drive the level, skew and curvature processes x_t^1, x_t^2, x_t^3 . The basis functions are based on the first three principal components of the market data and displayed in Figure 15. The initial surface σ_0 was taken to be the arbitrage-free SPX implied volatility surface on 31/12/2021.

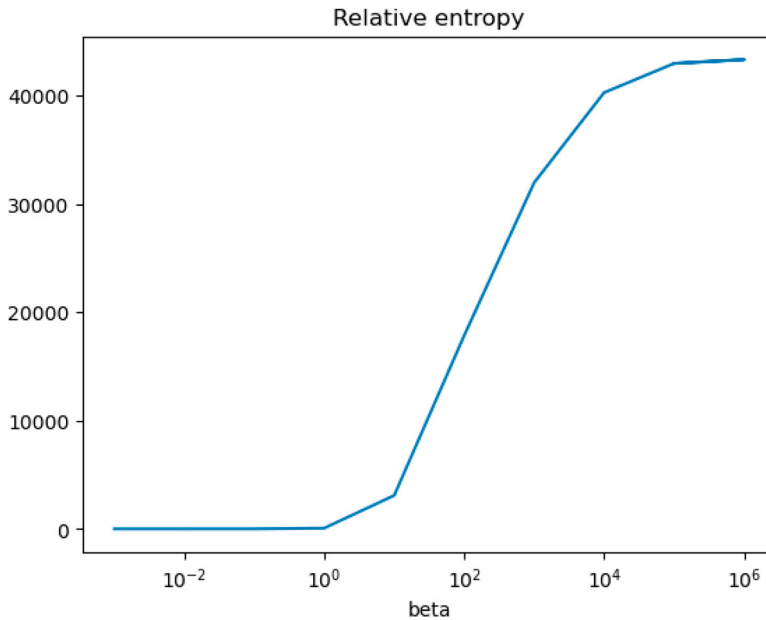
As shown by Rogers and Tehranchi (2010), an affine factor model such as (26)–(27) may violate static arbitrage constraints, so we apply the re-weighting procedure described in Table 4. We simulate $N = 100,000$ 3-month scenarios from the factor model (26)–(27).

Among these scenarios, 64.8% were arbitrage-free. However, even when the arbitrage penalty of a simulated path was non-zero, it was much lower than that of SPX implied vol data. Quantiles of arbitrage penalties across different \mathbb{P}_β are displayed in Table 5. When comparing the arbitrage penalty quantiles to those of SPX implied volatility displayed in Table 3, it is important to note that the factor model arbitrage penalties are calculated for paths, whereas the SPX market data arbitrage penalties are for individual surfaces only. In the scenarios generated by the factor model (26)–(27), only butterfly arbitrage was observed. All simulated implied volatility surfaces satisfied the absence of calendar and call spreads. The pattern of violations in the butterfly constraint resemble the violations induced by the addition of IID noise perturbations in Section 3.

Table 5 displays quantiles of the arbitrage penalty φ defined by (21) under \mathbb{P}_β^N . To compute the q th quantile of the arbitrage penalty under \mathbb{P}_β^N , we sort the scenarios in increasing order of arbitrage penalties $\varphi(\omega_{(1)}) \leq \varphi(\omega_{(2)}) \leq \dots \leq \varphi(\omega_{(N)})$. The q th quantile is then

Table 5. Quantiles of arbitrage penalty under \mathbb{P}_β^N in scenarios simulated from the factor model (26)–(27).

β	0	10^2	10^3	10^4	10^5
90th quantile	0.044	0.001	0	0	0
95th quantile	0.09	0.004	0.0001	0	0
99th quantile	0.206	0.014	0.001	0.0004	0


Figure 16. Relative entropy $H(\mathbb{P}_\beta^N | \mathbb{P}_0^N)$ in (26)–(27) as a function of β .

estimated as:

$$F_\varphi^{-1}(q) = \varphi(\omega_{(k)}), \quad k = \min\{j \in \{1, \dots, N\} : \sum_{i=1}^j w(\omega_{(i)}) \geq q\}. \quad (29)$$

As we increase β , we are less and less likely to sample a path with non-zero arbitrage penalty. Table 5 shows that with $\beta = 10^5$, 99% of scenarios are arbitrage-free. In this example, for $\beta = 10^{10}$, we are left with arbitrage-free paths only.

Figure 16 shows the relative entropy $\mathcal{E}(\beta) = H(\mathbb{P}_\beta^N | \mathbb{P}_0^N)$ as a function of β . We observe a sharp transition around $\beta = 100$, suggesting that for $\beta \gg 10^2$ the penalization eliminates scenarios with arbitrage.

The histogram of weights $w_i(10^2)$ (Figure 17) illustrates the clustering of weights into two groups: those corresponding to arbitrage-free scenarios, which are equally weighted, and those corresponding to scenarios with arbitrage penalty $\varphi(\omega_i) > 0$ whose weights are driven very close to zero. We note that even with $\beta = 10^2$, some of the weights are already very close to zero.

We conclude that for the factor model (26)–(27) the impact of the penalty step is small in terms of entropy distance.

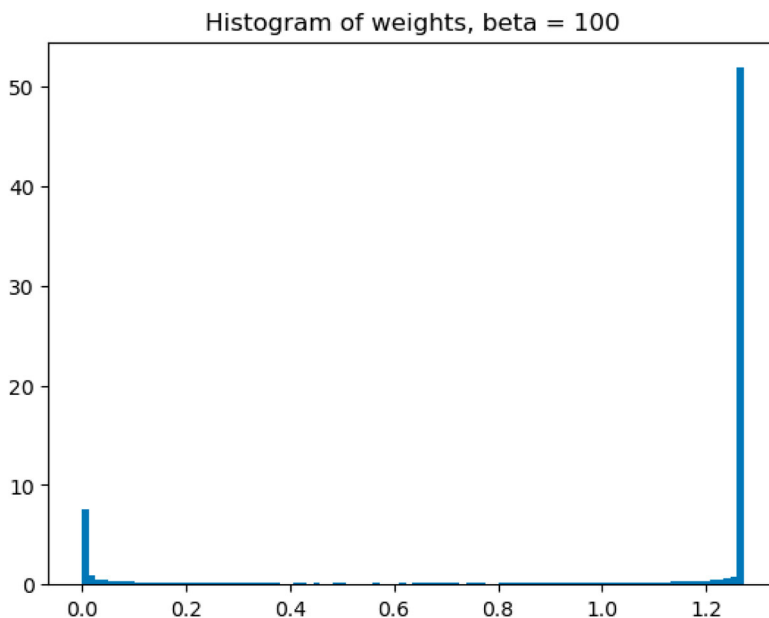


Figure 17. Histogram of $Nw(\beta)$ with $\beta = 10^2$ in (26)–(27).

5.2. Example: Factor Model for the SPX Implied Volatility Surface

We now give an example of a factor model for the SPX implied volatility surface based on the findings from Section 2. As in the factor model (26)–(27), we consider the following dynamics:

$$\sigma_t(m, \tau) = \bar{\sigma}(m, \tau) \exp \left(\sum_{i=1}^4 x_t^i f_i(m, \tau) \right), \quad (30)$$

where the factors $x_t^1, x_t^2, x_t^3, x_t^4$ in this case correspond to *level*, *skew*, *term structure* and *curvature*, as projections on principal components with the analogous representations. Their dynamics are once again modeled as independent Ornstein-Uhlenbeck processes:

$$dx_t^i = \lambda_i (\alpha_i - x_t^i) dt + \gamma_i dW_t^i, \quad i = 1, \dots, 4. \quad (31)$$

The underlying asset S is modeled as a process with stochastic volatility proportional to the one-month ATM volatility $\sigma_t(1, \frac{1}{12})$, as discussed in Section 2:

$$dS_t = \nu \sigma_t \left(1, \frac{1}{12} \right) S_t dW_t^0, \quad (32)$$

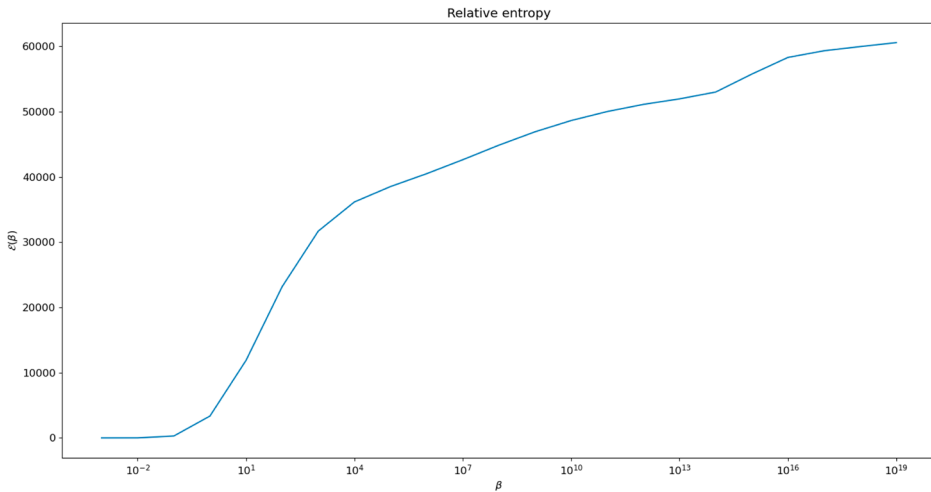
$$W_t^0 = \rho_1 W_t^1 + \rho_2 W_t^2 + \rho_3 W_t^3 + \rho_4 W_t^4 + \sqrt{1 - (\rho_1^2 + \rho_2^2 + \rho_3^2 + \rho_4^2)} B_t, \quad (33)$$

where $\nu > 0, \rho_1, \rho_2, \rho_4 < 0, \rho_3 > 0$ and $B, W^i, i = 1, \dots, 4$ are independent Brownian motions.

The factor model (30)–(31) may be adapted to any underlying asset. We demonstrate the approach for SPX options, by using as factors the first four principal components displayed

Table 6. Estimated OU parameters for SPX implied volatility factors.

	λ	α	γ
X_t^1	2.018	-0.422	4.414
X_t^2	0.986	-0.312	1.993
X_t^3	1.258	0.097	1.295
X_t^4	1.497	-0.021	0.824

**Figure 18.** Relative entropy $H(\mathbb{P}_\beta^N | \mathbb{P}_0^N)$ as a function of β : SPX factor model (30)–(31).

in Figure 4 for f_1, f_2, f_3, f_4 . We estimate $\alpha_i, \lambda_i, \gamma_i, i = 1, \dots, 4$ via a Generalized Method of Moments (GMM), using the first two moments and the autocorrelation function at various lags as moment conditions. Estimates are shown in Table 6. We set $\nu = \frac{1}{2}$ and the correlations $\rho_i, i = 1, \dots, 4$ to be the historical correlations from Table 2.

As in the previous example, we simulate 100,000 3-month paths from the model (30)–(31), using the above hyperparameters. The initial surface is the average SPX implied volatility $\bar{\sigma}$ (Figure 2), and the starting values for the level, skew, term structure and curvature processes are those observed on 31/12/2021. The percentage of paths and surfaces admitting a non-zero arbitrage penalty is shown in Table 7.

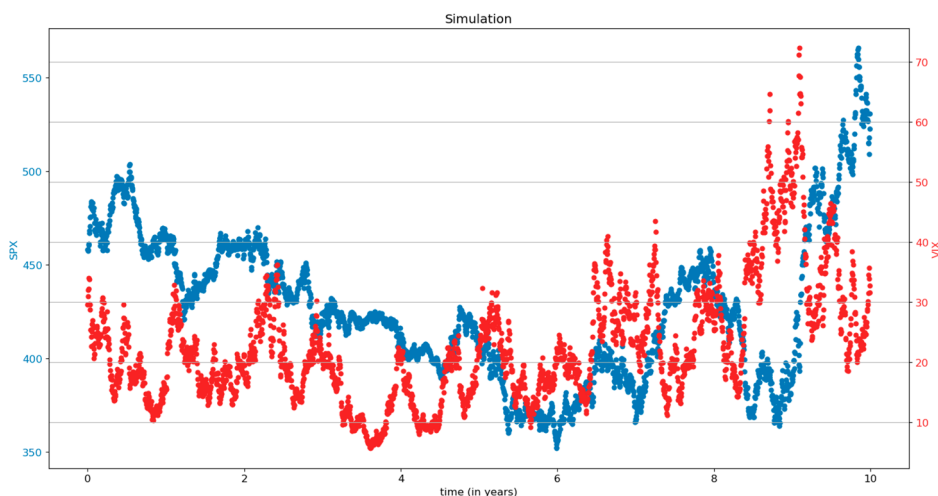
The relative entropy, shown in Figure 18, exhibits a sharp transition around $\beta = 10$ indicating that the penalization efficiently eliminates scenarios with arbitrage. The effect of β on arbitrage penalty is shown in Table 8. The arbitrage penalty in simulated scenarios is much lower than the historically observed penalties in the SPX implied volatilities (Table 3), considering that the penalties in Table 8 correspond to 3-month paths. It becomes negligible with $\beta = 10^2$, and arbitrage is effectively removed for $\beta > 10^4$ in this example. Furthermore, we note that the quantiles of the arbitrage penalty in the SPX model (30)–(31) (Table 8) decay faster with β compared to the corresponding quantiles in the factor model (26)–(27) (Table 5).

Table 7. Arbitrage presence in scenarios from the SPX factor model (30)–(31).

	Total	Calendar	Call	Butterfly
Paths	62.761%	21.245%	36.548%	36.548%
Surfaces	16.055%	4.004%	9.117%	7.073%

Table 8. Quantiles of arbitrage penalty under \mathbb{P}_β^N for SPX factor model (30)–(31).

β	0	10^2	10^4	10^6	10^{10}	10^{15}
90th quantile	0.20	$5 \cdot 10^{-5}$	$5 \cdot 10^{-10}$	$2 \cdot 10^{-13}$	0	0
95th quantile	0.77	0.002	$2 \cdot 10^{-7}$	$3 \cdot 10^{-9}$	$4 \cdot 10^{-15}$	0
99th quantile	4.49	0.01	$3 \cdot 10^{-7}$	$2 \cdot 10^{-9}$	$1 \cdot 10^{-11}$	$4 \cdot 10^{-16}$

**Figure 19.** Simulation of a 10-year scenario for VIX (red) and SPX (blue) using the SPX factor model (31)–(33).

Simulating the volatility index We simulate the VIX dynamics associated with the SPX four-factor model (30)–(33) using the methodology from CBOE (2022). We fix the mon-eyness grid by taking 100 equispaced values between 0.5 and 1.5. We simulate 10-year VIX and SPX paths using a frequency of one day $\Delta t = \frac{1}{252}$ with the average SPX implied volatility surface, and the SPX price on the 31st Dec 2021 as the starting point. The remaining hyperparameters are as in the previous simulations. Figure 19 displays simulated sample paths for the underlying and VIX. We note that the model (30)–(33) is able to produce high VIX values as historically observed during the 2008 financial crisis and during the Covid-19 pandemic. Figure 20 displays simulated paths for one-month realized vol, one-month ATM vol and VIX. We note that the VIX and the ATM vol are higher than the one-month forward realized vol. The ATM vol is usually below the VIX in the simulations, which is consistent with the post-pandemic dynamics (Figure 8).

We further investigate the relationship between the simulated values of VIX, ATM vol, SPX, and the level process. Pearson correlation between the simulated log-increments of SPX, ATM vol, VIX, and the increments of the level process is shown in Table 9. We note that the log-returns of the underlying are negatively correlated with the increments of the

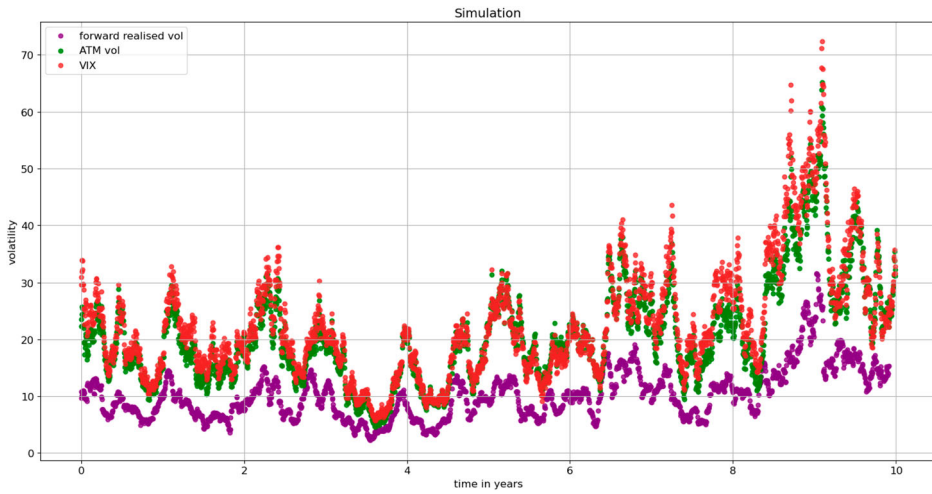


Figure 20. Simulation of VIX (red), ATM volatility (green), and the 30-day realized volatility (purple) using the SPX factor model (31)–(33).

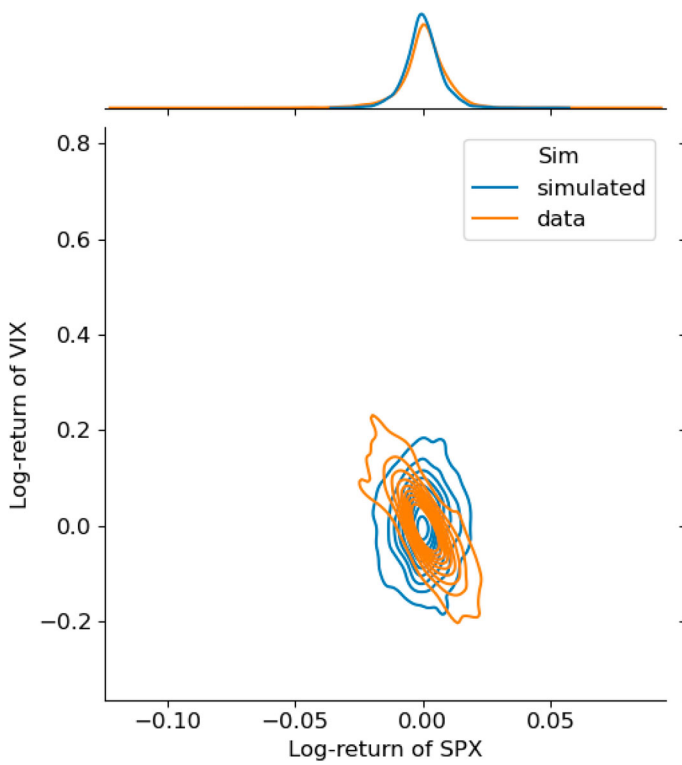


Figure 21. Joint distribution of log-returns of VIX and the log-returns of SPX in simulations via the SPX factor model (31)–(33) and in the historically observed data (2012–2021).

Table 9. Pearson correlation between simulated values of log-returns of SPX, returns of the level process, log-returns of the ATM vol, log-returns of VIX using the SPX factor model (31)–(33).

	$\Delta \log S_t$	ΔX_t^1	$\Delta \log \sigma_t^{ATM}$	$\Delta \log \sigma_t^{VIX}$
$\Delta \log S_t$	1.00	−0.45	−0.31	−0.30
ΔX_t^1	−0.45	1.00	0.96	0.94
$\Delta \log \sigma_t^{ATM}$	−0.31	0.96	1.00	0.99
$\Delta \log \sigma_t^{VIX}$	−0.30	0.94	0.99	1.00

level process, the log-returns of ATM vol, and the log-returns of VIX. There is a high positive correlation between the log increments of the ATM vol, VIX, and the increments of the level process. This is consistent with the historical correlations displayed in Figure 7.

In Figure 21 we compare the historical (2012–2021) and the simulated joint distribution of the log-returns of SPX and of VIX. The means of the two distributions align, and the corresponding marginal distributions of the simulated and historical values are close to each other. However, we note that as the historical correlation between the log-returns of SPX and of VIX is non-constant (Figure 7), the joint distribution changes through time as well. The correlation between the log-returns SPX and VIX being lower in the simulations than in the 2012–2021 historical data (Figure 21) can be contributed to the correlation ρ_1 used in simulations being lower than the average historical daily correlation of R_t and ΔX_t^1 for the time period 2012–2021 (Figure 7). Overall, we conclude that the four-factor model (31)–(33) is able to generate realistic scenarios for VIX, consistent with the historical observations.

6. Conclusion

We introduced a simple and computationally tractable method for simulating arbitrage-free implied volatility surfaces. Our approach offers flexibility with respect to the underlying model for implied volatility dynamics, whilst preserving the co-movements across strikes and maturities.

Our approach enables to combine a data-driven multifactor model with a Weighted Monte Carlo method in order to conciliate static arbitrage constraints with a statistically realistic representation of co-movements of implied volatilities.

Acknowledgments

We thank Katia Babbar, Andrey Chirikhin, Samuel N. Cohen, Bruno Dupire, Blanka Horvath, Terry Lyons, Fabio Mercurio, Christoph Reisinger, Justin Sirignano and seminar participants at QuantMinds 2022 for helpful comments and remarks.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

Milena Vuletić's research is supported by BNP Paribas through the EPSRC Centre for Doctoral Training in Mathematics of Random Systems: Analysis, Modelling and Simulation (EPSRC Grant EP/S023925/1).

References

- Avellaneda M., Buff R., Friedman C., Grandchamp N., Kruk L., and Newman J. 2001. "Weighted Monte Carlo: A New Technique for Calibrating Asset-Pricing Models." *International Journal of Theoretical and Applied Finance* 04 (01): 91–119. <https://doi.org/10.1142/S0219024901000882>.
- Avellaneda M., Healy B., Papanicolaou A., and Papanicolaou G. 2020. "PCA for Implied Volatility Surfaces." *The Journal of Financial Data Science* 2 (2): 85–109. <https://doi.org/10.3905/jfds.2020.1.032>.
- Babbar K. A. 2001. "Aspects of Stochastic Implied Volatility in Financial Markets." PhD diss., Imperial College London.
- Carmona R., Ma Y., and Nadtochiy S. 2017. "Simulation of Implied Volatility Surfaces Via Tangent Lévy Models." *SIAM Journal on Financial Mathematics* 8 (1): 171–213. <https://doi.org/10.1137/15M1015510>.
- CBOE. 2022. "Volatility Index Methodology: CBOE Volatility Index." Accessed May 08, 2023. https://cdn.cboe.com/api/global/us_indices/governance/VIX_Methodology.pdf.
- Cohen S. N., Reisinger C., and Wang S. 2020. "Detecting and Repairing Arbitrage in Traded Option Prices." *Applied Mathematical Finance* 27 (5): 345–373. <https://doi.org/10.1080/1350486X.2020.1846573>.
- Cohen S. N., Reisinger C., and Wang S. 2023. "Arbitrage-free Neural-SDE Market Models." *Applied Mathematical Finance* 30 (1): 1–46. <https://doi.org/10.1080/1350486X.2023.2257217>.
- Cont R., and da Fonseca J. 2002. "Dynamics of Implied Volatility Surfaces." *Quantitative Finance* 2 (1): 45–60. <https://doi.org/10.1088/1469-7688/2/1/304>.
- Cont R., Fonseca J. D., and Durrleman V. 2002. "Stochastic Models of Implied Volatility Surfaces." *Economic Notes* 31 (2): 361–377. <https://doi.org/10.1111/ecno.2002.31.issue-2>.
- Davis M. H., and Hobson D. G. 2007. "The Range of Traded Option Prices." *Mathematical Finance* 17 (1): 1–14. <https://doi.org/10.1111/mafi.2007.17.issue-1>.
- Dobi D. 2014. "Modeling Systemic Risk in the Options Market." PhD diss., Department of Mathematics, New York University.
- Dumas B., Fleming J., and Whaley R. E. 1998. "Implied Volatility Functions: Empirical Tests." *The Journal of Finance* 53 (6): 2059–2106. <https://doi.org/10.1111/jofi.1998.53.issue-6>.
- Dupire B. 1994. "Pricing with a Smile." *Risk* 7 (1): 18–20.
- Gatheral J. 2011. *The Volatility Surface: A Practitioner's Guide*. Chichester: John Wiley & Sons.
- Gatheral J., and Jacquier A. 2014. "Arbitrage-Free SVI Volatility Surfaces." *Quantitative Finance* 14 (1): 59–71. <https://doi.org/10.1080/14697688.2013.819986>.
- Heynen R. 1994. "An Empirical Investigation of Observed Smile Patterns." *Review of Futures Markets* 13:317–317.
- Kamal M., and Gatheral J. 2010. "Implied Volatility Surface." In *Encyclopedia of Quantitative Finance*, edited by R. Cont, Chichester: John Wiley & Sons, Ltd.
- Martini C., and Mingone A. 2022. "No Arbitrage SVI." *SIAM Journal on Financial Mathematics* 13 (1): 227–261. <https://doi.org/10.1137/20M1351060>.
- Rogers L. C. G., and Tehranchi M. R. 2010. "Can the Implied Volatility Surface Move by Parallel Shifts?." *Finance and Stochastics* 14 (2): 235–248. <https://doi.org/10.1007/s00780-008-0081-9>.
- Schönbucher P. J. 1999. "A Market Model for Stochastic Implied Volatility." *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* 357 (1758): 2071–2092. <https://doi.org/10.1098/rsta.1999.0418>.
- Schweizer M., and Wissel J. 2008. "Arbitrage-Free Market Models for Option Prices: The Multi-Strike Case." *Finance and Stochastics* 12 (4): 469–505. <https://doi.org/10.1007/s00780-008-0068-6>.
- Wissel J. S. 2008. "Arbitrage-Free Market Models for Liquid Options." PhD diss., ETH Zurich.
- Zhang W, Li L., and Zhang G. 2023. "A Two-Step Framework for Arbitrage-Free Prediction of the Implied Volatility Surface." *Quantitative Finance* 23 (1): 21–34. <https://doi.org/10.1080/14697688.2022.2135454>.