Behaviormetrika (2024) 51:291–317
https://doi.org/10.1007/s41237-023-00213-2

ORIGINAL PAPER

Interpolation of missing swaption volatility data using variational autoencoders

Ivo Richert[1] · **Robert Buch**[1]

Received: 31 October 2022 / Accepted: 30 October 2023 / Published online: 10 December 2023

Abstract

Albeit of crucial interest for financial researchers, market-implied volatility data of European swaptions often exhibit large portions of missing quotes due to illiquidity of the underlying swaption instruments. In this case, standard stochastic interpolation tools like the common SABR model cannot be calibrated to observed volatility smiles, due to data being only available for the at-the-money quote of the respective underlying swaption. Here, we propose to infer the geometry of the full unknown implied volatility cube by learning stochastic latent representations of implied volatility cubes via variational autoencoders, enabling inference about the missing volatility data conditional on the observed data by an approximate Gibbs sampling approach. Up to our knowledge, our studies constitute the first-ever completely non-parametric approach to modeling swaption volatility using unsupervised learning methods while simultaneously tackling the issue of missing data. Since training data for the employed variational autoencoder model is usually sparsely available, we propose a novel method to generate synthetic swaption volatility data for training and afterwards test the robustness of our approach on real market quotes. In particular, we show that SABR interpolated volatilities calibrated to reconstructed volatility cubes with artificially imputed missing values differ by not much more than two basis points compared to SABR fits calibrated to the complete cube. Moreover, we demonstrate how the imputation can be used to successfully set up delta-neutral portfolios for hedging purposes.

Keywords Swaption · Gibbs sampling · Variational autoencoder · Missing data imputation

Communicated by Alfonso Iodice D'Enza.

✉ Ivo Richert
ivo.richert@gmail.com

[1] Department of Financial Mathematics, Fraunhofer ITWM, Kaiserslautern, Germany

1 Introduction

A complete interest rate swaption volatility cube is of practical interest for both researchers and practitioners, enabling inference about the current state of the market by the former, while making both hedging and consistent valuation of swaptions and more exotic derivatives through all maturities and tenors possible for the latter (Dimitroff and de Kock 2011). The classic approach consists in fitting a given parametric model like the popular SABR model (Hagan et al. 2002) or the LIBOR market model to market-observed smiles, i.e., slices of the full volatility cube, and in interpolating the missing quotes by the calibrated models as well as extrapolating beyond them (see, e.g., Brigo and Mercurio (2007), for a variety of interest rate models).

Although the swaption market is approximately an order of magnitude larger than the next biggest interest rate derivative market being the cap/floor market, larger market volumes do not necessarily mean that volatility quotes are liquid in all parts of the swaption volatility cube. Indeed, one often observes that the at-the-money swaption market is very liquid; however, for various tenors and expiries, the away from-the-money quotes are missing or not at all reliable, especially when compared to corresponding cap/floor volatilities (Skantzos and Garston 2019, p. 2). When, however, for a given expiry and tenor, one only observes a single-quoted strike of the whole smile (typically the at-the-money point), the stochastic model-based approach cannot be applied directly. One then often needs to resort to simple interpolation schemes to fill the volatility cube and hence enable an ordinary calibration of a stochastic model. Given the sparsity of values in the away-from-the-money strike area, however, simple linear or even more advanced interpolation schemes often work exceptionally bad or even fail to work at all, due to no available simplex of data points surrounding a missing value on the cubic grid. Then, artificial extrapolation methods need to be considered for the boundaries of the cube. For a more detailed discussion of the limited usability of interpolation schemes for preprocessing volatility data with missing values, see Dimitroff and de Kock (2011, p. 8).

Alternatively, Hagan and Konikov (2004) suggest fitting the parameters of the SABR model both to the observed cap volatility surfaces as well as to quoted swaption volatilities, while Jäckel and Rebonato (2000) develop an explicit relationship between cap/floor and swaption volatilities by expressing the forward swap rate as a series of forward rates. These methods for preprocessing volatility cubes with missing data for calibration of a stochastic model are referred to as "lifting from caps" and will not be considered further here as they require additional sources of data in the form of cap quotes and explicitly build on a given parametric model like the SABR model. Instead, the aim of this paper is to intrinsically model the imputation of missing swaption quotes nonparametrically by a deep generative model.

Accordingly, we propose the idea of filling missing values in the volatility cube by a Gibbs sampling-inspired approach proposed by Rezende et al. (2014) that is able to asymptotically sample from a learned variational approximation

of the joint distribution of missing data and latent variables given the observed data of a volatility cube. Learning an approximative distribution driving the generation process of volatility cubes is carried out by the variational autoencoder (VAE) model of Kingma and Welling (2013) and Rezende et al. (2014). Using this approach, missing values on the cubic grid can simply be reconstructed from the existing ones and can afterwards be used as an input in the calibration procedure of a standard parametric model, like the SABR model. The motivation to use VAE as a special kind of an autoencoder model stems from two facts. First, as shown in Dai et al. (2018), there exists a connection of VAE to robust PCA, making this method a suitable candidate to model financial data, which are prone to different sources of noise. Second, given their stochastic nature, VAE forms a powerful generative model to holistically model high-dimensional data like implied swaption volatility cubes.

Since data for training the VAE model are sparsely available, and in particular not sufficiently available for training a machine learning model, we develop a novel method for generating synthetic swaption cubes from existing ones that can be used to train the VAE model, which has not been applied before in the context of financial swaption data. The robustness of this method is afterwards tested on real market-quoted volatility data and we show that SABR interpolated volatilities calibrated to reconstructed volatility cubes with artificially imputed missing values differ by not much more than two basis points compared to SABR fits calibrated to the true underlying complete swaption cube.

The imputation of missing data by deep generative models has become increasingly popular over the past years and has been studied by, e.g., Camino et al. (2019), Collier et al. (2020), Gondara and Wang (2018), Ipsen et al. (2021), Lewis et al. (2021), Ma et al. (2020), Ma and Zhang (2021), Mattei and Frellsen (2018), Nazabal et al. (2020), Qiu et al. (2023), or Roskams-Hieter et al. (2022). All of the aforementioned imputation studies are, however, not concerned with missing values inherent to financial data like swaption quotes. Past applications of supervised and unsupervised learning algorithms in the context of swaption data like, e.g., Kunsági-Máté et al. (2021), Thorin (2020) or Wang et al. (2018) are centered around learning the pricing or calibration routine of certain stochastic models like the SABR model. In contrast to that, our studies constitute the, up to our knowledge, first-ever completely nonparametric approach to modeling swaption volatility using unsupervised learning methods while simultaneously tackling the issue of sparsely available data which limits the applicability of the aforementioned pricing and calibration routines. Additionally, our synthetic data generation practice for training is novel in both the machine learning and financial literature.

In addition to data imputation by deep generative models described here, multiple other approaches for imputation of missing data were proposed by different authors. Sohl-Dickstein et al. (2015) and Goyal et al. (2019) study imputation of missing data with a Markov chain whose transition operator is learned by an appropriate statistical model and with unsupervised clustering algorithms. Moreover, Bachman and Precup (2015) suggest using LSTM networks to solve a sequential Markov decision problem arising in the context of data imputation, while Ivanov et al. (2019) propose

a variational autoencoder model that samples from a subset of missing features after being conditioned on arbitrary subsets of observed features.

Imputation of missing data using the Gibbs sampling-inspired approach of Rezende et al. (2014) was further studied in Mattei and Frellsen (2018) and Mattei and Frellsen (2019) who couple the algorithm described later with the importance-weighted autoencoder model from Burda et al. (2016) to handle cases where training data of the employed deep latent variable models contain missing values. Furthermore, Mattei and Frellsen (2018) propose to make use of a Metropolis-within-Gibbs extension of the algorithm presented here. However, we opt not to study this extension any further here, since the high dimensionality of the cubic swaption data makes calibrating a Metropolis proposal quite challenging.

The remaining paper is structured as follows. In the following section, we provide an overview over the SABR stochastic volatility model, before the third section deals with the variational inference paradigm utilized by the VAE model as well as with the Gibbs sampling-inspired approach of Rezende et al. (2014). The fourth section discusses means of synthetic data generation as it is necessary in the sparse swaption data environment. In the fifth section, we demonstrate the robustness of our approach and show the results of the Gibbs imputation on market-observed out-of-sample volatility cubes. Afterwards, we demonstrate how thereby obtained volatility values can be used to calibrate a SABR model to a volatility smile even when no quotes except from the at-the-money point are available and how one can successfully estimate the swaption's delta by the reconstructed volatility cubes. Finally, the sixth section concludes.

2 Swaption contracts and the SABR stochastic volatility model

A swaption is a financial instrument which gives its holder the right to enter into a financial swap with fixed rate K at some prespecified maturity T_0. If the holder of the contract decides to exercise the swaption at T_0, the holder and its counterparty enter into a series of exchanged payments at prespecified dates $T_1, \dots, T_M$, where one party pays a variable interest rate like the LIBOR rate on a fixed notional amount and the other party pays the fixed interest rate K on the same notional amount. Depending on whether the holder of the swaption pays or receives the fixed leg of the swap, the swaption is termed a payer swaption or a receiver swaption. The prespecified lifespan $T_M - T_0$ of the swap contract is usually termed the tenor of the swaption. From a theoretical point of view, a swaption constitutes a European option on the forward swap rate F_t with strike price K where the forward swap rate F_t is the fixed rate that would make the value of the swap expected at time t vanish at T_0. See Fig. 1 for an illustration of a typical swaption contract.

The SABR stochastic volatility model of Hagan et al. (2002) has become one of the effective standards in interest rate derivatives modeling which stems from its ability to accurately fit market implied volatility smiles with only four parameters

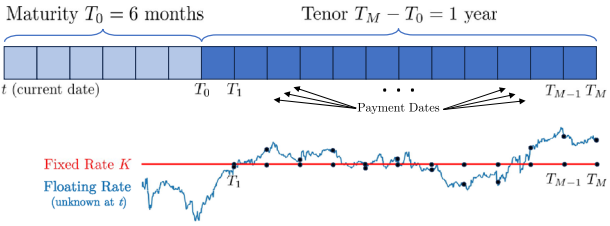

Fig. 1 Visualization of a typical swaption contract with a time to maturity of 6 months, a tenor of 1 year, and monthly payments. The highly nontrivial task of valuation of the swaption takes place at the current time t

α, ρ, ν, and β to fit. To account for the possibility of negative interest rates, a shifted version of the SABR model has become popular in the financial literature of the past years (see Antonov et al. 2019). The shifted SABR model describes the dynamics of the forward swap rate F for a given swaption maturity T_0 and tenor T_M under the so-called forward swap measure[1] Q^{T_0,T_M} with payment dates $T_1, \ldots, T_M$ by the two-factor CEV-type stochastic differential equations

$$\begin{aligned} \mathrm{d}F_t &= \sigma_t (F_t + b)^{\beta} \, \mathrm{d}W_t^{(1)} \\ \mathrm{d}\sigma_t &= \nu \sigma_t \, \mathrm{d}W_t^{(2)} \\ F_0 &= f \\ \sigma_0 &= \alpha, \end{aligned}$$

where f is the current forward swap rate, b is a displacement parameter allowing for negative rates, and where $W^{(1)}$ and $W^{(2)}$ are two standard Brownian motions with correlation $\rho \in [0, 1]$, controlling the instantaneous correlation between the forward swap rate and its volatility. The parameter $\nu \geq 0$ is often termed the volatility of volatility parameter and controls the curvature of the model-implied volatility smile, while the skewness parameter β controls the slope of the volatility smile and is usually fitted ex ante from historical time series analysis for the relevant market (see, e.g., West (2005)).

In practice, the value of a European swaption is generally quoted in terms of its Bachelier- or Black-model implied volatility. We refer to Bachelier-model implied volatility, which is typically quoted in basis points, as normal volatility or basis point volatility while we refer to Black-model implied volatility, which is typically quoted in percentage, as lognormal volatility. Given the low-interest regime that dominated global markets of the past years, we opt to represent swaption values in terms of normal volatility σ_N. In our studies, we fit the parameters of the shifted SABR model to the Bachelier-model implied normal volatility using the serial expansion formulas for the implied volatility from

[1] The forward swap measure Q^{T_0,T_M} is the equivalent martingale measure associated with the numeraire process $N_t := \sum_{i=1}^{m} \delta_i P(t, T_i)$, see the notation in Appendix A as well as Brigo and Mercurio (2007).

Hagan et al. (2002) with slight corrections from Oblój (2008) and using the explicit initial guesses derived in Le Floc'h and Kennedy (2014) for the iterative calibration procedure. The corresponding formulas for the implied normal and lognormal volatility can be found in Appendix A.

3 Variational inference and Gibbs-inspired sampling

The variational autoencoder model by Kingma and Welling (2013) aims to simultaneously train a generative model $p_\vartheta(x,z) = p_\vartheta(x|z)p(z)$ for the observable data $x \in \mathbb{R}^k$ given latent variables $z \in \mathbb{R}^d$ as well as an variational inference approximation $q_\theta(z|x)$ to the true posterior distribution $p(z|x)$. This is done by optimizing a variational lower bound to the logarithm of the intractable evidence $p_\vartheta(x) = \int p_\vartheta(x,z)\,\mathrm{d}z$

$$\log p_\vartheta(x) \geq \mathbb{E}_{q_\theta(z|x)}\left[\log \frac{p_\vartheta(x,z)}{q_\theta(z|x)}\right] = \mathbb{E}_{q_\theta(z|x)}[\log p_\vartheta(x|z)] - \mathrm{KL}(q_\theta(z|x)\|p(z)), \quad (1)$$

where KL denotes the Kullback–Leibler divergence of two distributions. While the latent prior distribution is commonly modeled by a standard normal distribution, the likelihood $p_\vartheta(x|z)$ and the variational posterior $q_\theta(z|x)$ are assumed to be Gaussians with diagonal covariance matrix parameterized by neural networks, called encoder and decoder, with parameters θ and ϑ for which maximizing (1) provides a tractable training criterion.

After training a variational autoencoder model on a complete data sample, we now assume that the remaining data x can be decomposed into an observed and into a missing component by $x = (x_{\text{obs}}, x_{\text{miss}})$. To infer the missing values inherent in the sample, the penultimate goal relies in sampling from the conditional distribution of x_{miss} given x_{obs}. We propose to utilize a Gibbs sampling-inspired approach to sample from an approximation of the conditional joint distribution of the random vector (x_{miss}, z) given x_{obs}. The details of this approach, which we will call Pseudo-Gibbs sampling, are confined in Appendix C. Samples from the required conditional distribution of x_{miss} given x_{obs} are then obtained via marginalization of the samples from the distribution of (x_{miss}, z) given x_{obs}, i.e., by discarding the latent code z from the joint samples of (x_{miss}, z). After obtaining samples $(x_{\text{miss}}^{(t)})_{0\leq t\leq T}$, we impute the missing values x_{miss} in the data by the sample average $\frac{1}{T}\sum_{t=0}^{T} x_{\text{miss}}^{(t)}$ as an estimator for the conditional expectation of the missing data given the observed. Due to its flexibility in sampling from an intractable conditional distribution, this Pseudo-Gibbs algorithm is frequently used in conjunction with deep latent variable models; see, e.g., Li et al. (2016), Mattei and Frellsen (2018), Rezende et al. (2016), or Du et al. (2018). The novelty of our proposal lies in applying the Pseudo-Gibbs methodology in the financial context and, in particular, in a financial environment which naturally exhibits large portions of missing data.

4 Employed data sets and synthetic data generation

To train the variational autoencoder model, daily (Bachelier model implied) swaption volatility cubes of European LIBOR swaptions were obtained from the FENICS market data provider between 21 August 2019 and 29 September 2021. Each volatility cube consists of $m = 21$ different maturities, $n = 14$ different tenors and 17 different strike prices per maturity–tenor combination. Multiple cubes within this 2 year period exhibit up to around 90% of their theoretical $21 \cdot 14 \cdot 17 = 4998$ entries missing. As there are only 163 fully observed swaption volatility cubes in our dataset, we need to apply synthetic data augmentation methods to provide enough training data coming from heterogeneous market states. Thereby, we enhance robustness and generalization capabilities while limiting overfitting in the VAE model. We decide to apply synthetic simulation in the parameter domain of the SABR stochastic volatility model. This is similar to an approach described in Andreichenko (2011).

To obtain realistic synthetic data, we fit the SABR stochastic volatility model with a fixed $\beta = 0.5$ to the 163 swaption cubes in the dataset because fixing β to a prespecified value like 0.5 is standard in practical applications of the SABR model (see, e.g., West (2005)). As it is the market-standard practice, we model the swap rate for each tenor T_M and maturity T_0 independently by a SABR model with parameters α_{T_0,T_M}, ν_{T_0,T_M} and ρ_{T_0,T_M}. This way, we obtain a dataset of $(m \times n)$-matrices of parameter values for each day and each parameter. Afterwards, we generate a synthetic α-parameter matrix in the following way:

(1) Generate the first row of the parameter matrix corresponding to the lowest swaption maturity: For each $j \in \{1, \dots, n\}$, fit a normal distribution to the increments $\alpha_{1,j} - \alpha_{1,j-1}$ between adjacent parameter values by matching means μ_j and variances σ_j^2 to the sample means and sample variances across the different days in the dataset.
(2) Choose $\kappa > 0$ as a scaling factor and sample these increments from the normal distribution $\mathcal{N}(\mu_j, \kappa \cdot \sigma_j^2)$. Apply the same generation procedure to simulate the first column of the parameter matrix.
(3) After obtaining the upper and left boundaries of the parameter matrix in this way, consecutively generate the values of parameters in the rest of the coordinates of the parameter matrix by the following method: The (i, j)th element $\alpha_{i,j}$ of the parameter matrix is obtained by fitting a normal distribution to the values $(\alpha_{i,j} - \alpha_{i-1,j})/(\alpha_{i-1,j} - \alpha_{i,j-1})$ in the dataset, which describe how the relative position $\alpha_{i,j}$ between the $(i-1, j)$th and $(i, j-1)$th value in the matrix is distributed, scaling the fitted variance by κ and afterwards sampling from this distribution.

This way, we maintain the inherent structure of the parameter values across the (T_0, T_M)-grid while simultaneously enriching the sparse dataset with new unseen parameter combinations. The hyperparameter $\kappa > 0$ controls the diversity of the generated parameter matrices. Choosing κ too small will result in parameter samples that differ only marginally from the average parameterization in the dataset, thereby limiting the generalization capabilities of the trained VAE model. Choosing κ too

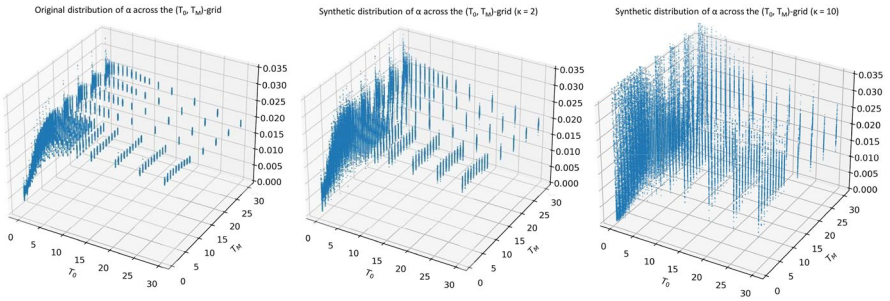

Fig. 2 Scatterplot of the original distribution of the α-parameter across the (T_0, T_M) grid on 20 October 2021 next to the distribution of synthetically sampled α-parameters for $\kappa = 2$ and for $\kappa = 10$. In both cases, the synthetic distribution displays relative overdispersion compared to the original distribution, i.e., for each tenor–maturity combination, the synthetic parameter distribution shows a higher variance than the distribution of the original parameter values for that tenor–maturity combination. This leads to a greater variety of volatility cubes for VAE training

large will feed the VAE with so much noise during training that it will be impossible for the neural network model to extract and learn the structural characteristics of the original volatility cubes.

We note here that, due to the random simulation practice, the obtained swaption cubes exhibit somewhat more roughness compared to the market-observed ones. We opt not to smooth out the synthetic data before training to make use of the denoising capabilities of VAE. It has been extensively studied (see, e.g., Vincent et al. (2008)) that inducing additional noise to the input improves generalization performances of deterministic autoencoder models, since it enhances robustness of adjacent data points in the latent manifold against the presence of small noise in the higher dimensional observation space. Moreover, Rezende et al. (2014) advocate that denoising enhances the generalization capabilities of probabilistic generative models as well, by noting that additional input noise is crucial for the recognition model to achieve the desired accuracy on unseen data. This motivates our choice to introduce the noise scaling hyperparameter κ, which should be optimized as an additional hyperparameter during training of the VAE model; see the following section.

A similar approach was used for generating synthetic matrices of the parameters ν and ρ. To make sure that the sampled α-, ν-, and ρ-values are positive and $[-1, 1]$-supported, respectively, we apply a log-transform to the α- and ν-parameter matrices and an artanh-transform (Fisher's z-transform) to the ρ-parameter matrices before sampling. Instead of the sampling approach described above, which yields sufficient results for our practices, other simulation methods for generating synthetic parameter matrices can also be used, e.g., fitting a matrix-variate normal distribution to given parameter matrices in the dataset (see, for example, Dutilleul 1999). A scatterplot of the original distribution of the α-parameter across the (T_0, T_M)-grid next to the distribution of synthetically sampled α-parameters in the cases $\kappa = 2$ and $\kappa = 10$ can be found in Fig. 2.

After obtaining synthetic realisations of SABR parameter matrices across the grid of swaption maturities and tenors, we transform them back into swaption volatility

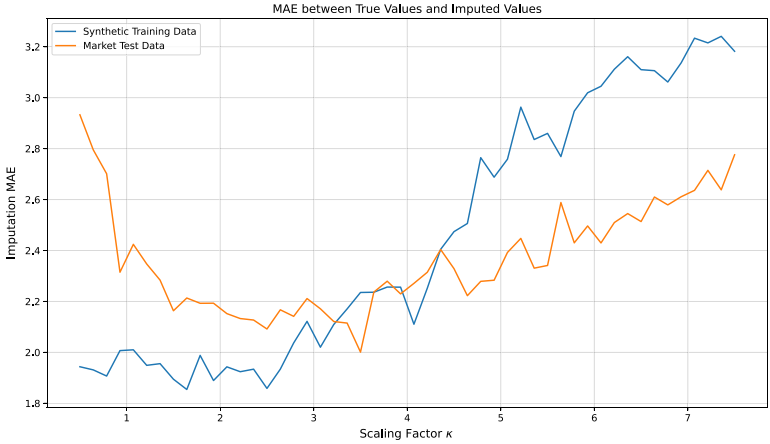

Fig. 3 Average imputation MAE in basis points across 10 000 synthetic data cubes and across 163 real market volatility cubes for different values of the noise scaling parameter κ between 0.5 and 7.5

cubes using the implied volatility expansion formulas from Hagan et al. (2002) (see formulas (2) and (3) in Appendix A).

5 Empirical results using VAE imputation on real and synthetic data

5.1 Assessment of imputation performance and hyperparameter optimization

After a VAE model has been trained on synthetic data, we can measure the accuracy of the Gibbs-inspired missing data imputation on a synthetic or real market swaption volatility cube in the following way: We set approximately 79.6% of the observations on the cube to zero and treat these points as missing data before the Pseudo-Gibbs imputation algorithm is applied to the cube. The locations of the missing points are chosen to match the locations of missing points of the volatility cube on 21 August 2019, of which only roughly over 20% of the values on the full three-dimensional grid were not missing. The algorithm is initialized at the missing points by an interpolation (resp. an extrapolation) of values at the points surrounding the missing values. Throughout, the pretrained inference model is a standard variational autoencoder with a ten-dimensional latent space which is trained for 50,000 epochs on 10,000 synthetically generated swaption cubes by the procedure described in Sect. 4. The missing values are imputed by the sample averages of the samples from a Gibbs Markov chain of length 2000 with a burn-in period of 100 (see Appendix C for further details). More details about the VAE model architecture and the original and reconstructed cube can be found in Appendix D. To obtain a global measure of the imputation accuracy, we compute the mean absolute error (MAE) between the true observed and the imputed values across the cube.

Our first assessment will consist of optimizing the noise scaling hyperparameter κ introduced in the synthetic data generation method from Sect. 4 with respect

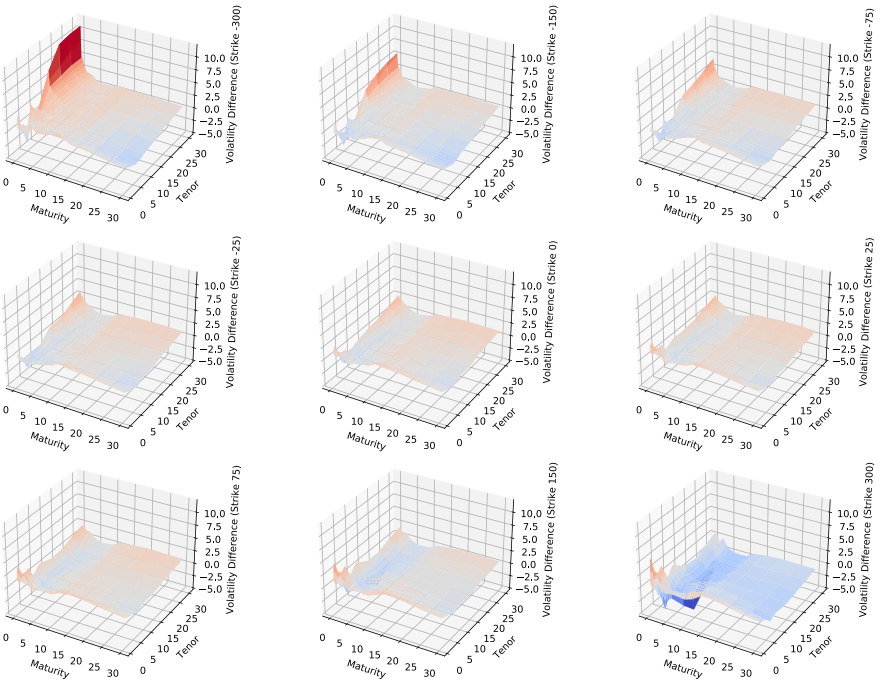

Fig. 4 Difference between the market-observed swaption volatility cube on 20 October 2021 and a VAE reconstructed volatility cube after 79.6% of the values were treated as missings

to this mean absolute deviation measure. Therefore, we generate synthetic data for 50 different values of κ between 0.5 and 7.5, train a VAE model for each value of κ, and compute the imputation MAE across all synthetic data samples and across all 163 fully observed real swaption cubes in our dataset. Figure 3 displays the average MAE in basis points across the different cubes and reveals that values of κ between 1.5 and 3.5 yield the best results on both the synthetic training set and the real test set. For values of κ below 1, the lack of sufficiently diverse training data severely hurts the imputation performance of the real market data, while of course the in-sample performance of the model on the rather homogeneous synthetic data is not hurt. On the other hand, values of κ above 4 introduce so much noise in the data set that the capabilities of the VAE to learn the true underlying data distribution are limited. Accordingly, we will use the value $\kappa = 2$ during synthetic training data generation for all further experiments.

Figure 4 shows the difference between the basis point volatility cube (Bachelier-model implied normal volatilities) of European LIBOR swaptions on 20 October 2021 and a reconstructed volatility cube using the VAE imputation. The mean absolute deviation of the imputed values from the true values is 1.9123 basis points, demonstrating the good out-of-sample generalization capabilities of the approach. The algorithm exhibits its greatest uncertainty of reconstruction in the low maturity-high tenor region which is also resembled later in Fig. 9.

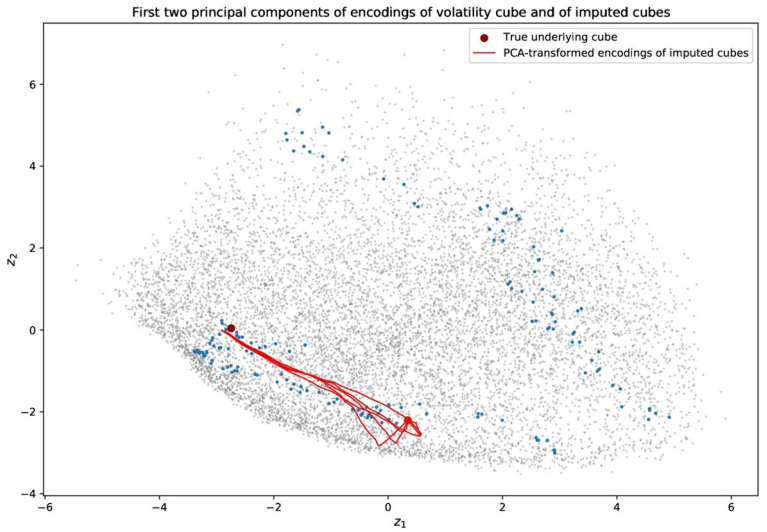

Fig. 5 First two principal components of ten-dimensional latent encodings of synthetic and market-observed volatility cubes as well as of the cube on 20 October 2021 in dark red. Displayed are the means of the Gaussian VAE-approximated posteriors $q_\theta(z|x_i)$. Blue dots indicate encodings of market-observed data, while the grey dots show the encodings of the synthetic VAE training data. Furthermore, the paths that were traced by the latent encodings of the sample average imputed cubes after between 1 and 2000 Markov chain steps are shown in red for five different runs of the VAE-imputation algorithm

We note here that in our setup, it is of no importance to make sure whether the imputed volatility cubes after Gibbs sampling exhibit arbitrage or not, since we opt to use the VAE-imputed cube as a mere input for the calibration procedure of a subsequent stochastic volatility model like the SABR model. Maintaining no-arbitrage conditions then depends of the choice of the specific stochastic model employed; see, e.g., Johnson and Nonas (2009).

Convergence of the imputed sample averages (7) can be monitored at each step by standard techniques for computation of confidence intervals from Markov chain Monte Carlo theory, e.g., the Overlapping Batch Means estimator of Flegal and Jones (2011). Note that a 95%-confidence error of ϵ does not imply a 95% chance of the interval $[\hat{x}_{\text{miss}} - \epsilon, \hat{x}_{\text{miss}} + \epsilon]$ to cover the masked value of the observed volatility cube that shall be imputed, but merely a 95% chance of the interval to cover the expected value of x_{miss} given x_{obs} , whatever this value will be. Thus, we cannot expect the imputed cube to "converge" to the true observed cube but merely to an approximation of the cube that we would expect after seeing the observed values and given the underlying distribution of data that was inferred from our synthetic training data.

Convergence can also be monitored in the latent space instead of in the observable space. Figure 5 displays the first two principal components of the ten-dimensional latent encodings of the synthetic and market-observed volatility cubes as well as in dark red the latent encoding of the volatility cube from 20 October 2021 that is used for Fig. 4. Furthermore, the paths that were traced by the latent encodings

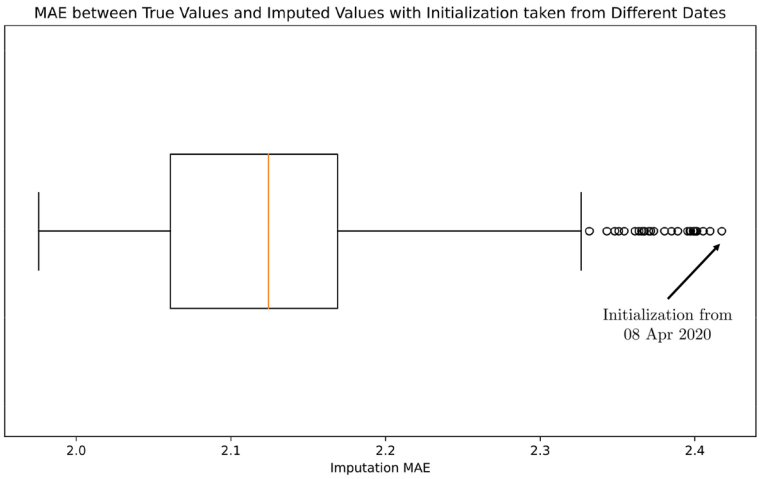

Fig. 6 MAE of imputation of the volatility cube on 21 October 2021 obtained by initializing the Pseudo-Gibbs algorithm with starting values interpolated from cubes of different dates. All outlier points above the top whisker correspond to initializing the algorithm with cubes from between 02 Mar 2020 and 16 Apr 2020 during the outbreak of the Corona pandemic

of the sample average imputed cubes between 1 and 2000 Markov chain steps are shown in red for five different runs of the Pseudo-Gibbs imputation algorithm. It is seen that the latent encodings of the sample averages converge to some encoding near the encoding of the true underlying cube.

Studying the underlying non-euclidean geometry of the latent space of the VAE model that is displayed in Fig. 5 provides an interesting possibility for further research. Lately, it has been noted (see Arvanitidis et al. 2018, and Hauberg 2018) that, empirically, Euclidean latent space distances carry little information about the relationship between data points and that an interpretation of the latent space as a Riemannian manifold appears more promising. Hence, it could be fruitful to investigate whether the paths traced by the imputed cubes via the Pseudo-Gibbs algorithm in Fig. 5 are in concordance with such a geometric structure, i.e., whether these paths represent geodesics on the respective manifolds between the initial latent representation of the missing cube and the latent representation of the true underlying cube.

Finally, we test the robustness of convergence of the Pseudo-Gibbs imputation method with respect to a perturbation of the starting values of the algorithm. We use different starting points for the volatility cube on 20 October 2021 from Fig. 4. Instead of the canonical choice of starting values from 20 October 2021, we use values from the dates between 21 August 2019 and 29 September 2021. For each different date used for initialization, we then measure the imputation MAE across the cube, as we did before. Figure 6 shows a boxplot of the imputation MAEs across the different initialization dates. While the MAE of imputation amounted to 1.9123 basis points for the imputation correctly initialized by interpolating the values of the cube on 20 October 2021 (see Fig. 4), the MAEs of imputation with perturbed

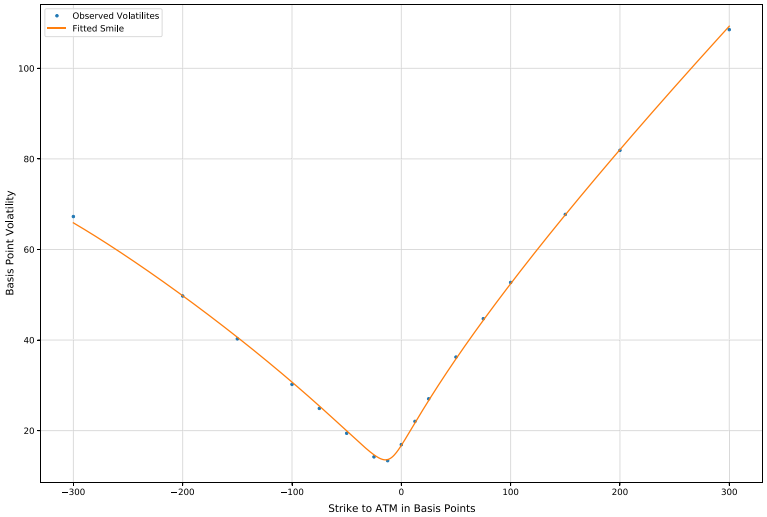

Fig. 7 SABR fitted volatility smile for the 1 year by 1 year swaption on 20 Oct 2021. The calibrated parameters are $\alpha = 0.0086$, $\nu = 1.0732$, and $\rho = 0.6506$ and the mean absolute error of fit is 0.35 basis points

starting values vary only between 2.0 and approximately 2.4, demonstrating the robustness of our approach with respect to the starting values of the algorithm. The worst imputation performance is obtained by initializing the sampling with values from volatility cubes following the market crash during the beginning of the Corona pandemic in March 2020.

After obtaining an imputed reconstruction of a swaption volatility cube containing missing values, these reconstructions can be used to calibrate certain market-standard stochastic volatility models whose calibration would otherwise fail because of a too sparse data domain. We follow this procedure in the following section using the example of the SABR stochastic volatility model.

5.2 Fitting the SABR model to imputed volatility smiles

In the following, for each tenor–maturity combination, we fix β to a value of 0.5 and fit the other three parameters of the SABR stochastic volatility model by a standard least-squares minimization. The shift introduced is $b = 0.04$ or 400 basis points to ensure positivity of all forward swap rates on each date as well as of all strikes. Figure 7 shows the calibrated smile the model produces for the swaption with maturity and tenor of 1 year on 20 October 2021.

Since, for this swaption, all volatility quotes except from the at-the-money point are missing in the volatility cube of between 21 Aug 2019 and 24 Apr 2020, the aim is now to examine how much the fit of the SABR model varies when the true volatility quotes except from the at-the-money point are replaced by ones from the Pseudo-Gibbs imputation and by ones from a simple interpolation. Figure 8 shows

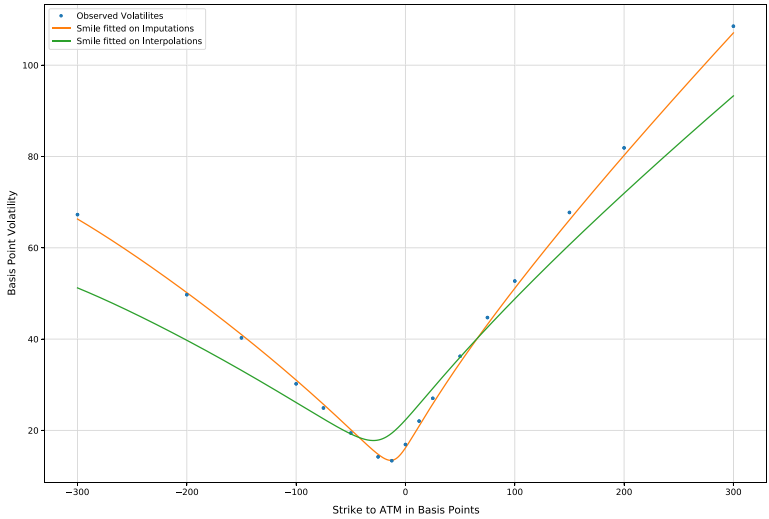

Fig. 8 SABR fitted volatility smile for the 1 year by 1 year swaption on 20 Oct 2021 when all but the at-the-money point are replaced by the imputed values from the Pseudo-Gibbs algorithm and by linearly interpolated values on the cube respectively. The calibrated parameters are $\alpha = 0.0083$, $\nu = 1.0622$, and $\rho = 0.6169$ in the first case and $\alpha = 0.0117$, $\nu = 0.7349$, and $\rho = 0.6561$ in the second case

the calibrated SABR volatility smile for the same swaption on the same date, where except for the at-the-money point all other volatility quotes are replaced by a) the Pseudo-Gibbs imputed volatility values and b) volatility values from a simple linear interpolation (respectively, extrapolation on the boundary of the observed market cube) of the swaption cube at the missing values.

The linearly interpolated volatility cube clearly fails to adequately rebuild the true volatility structure due to the large amount of missing values which results in an insufficient fit of the smile and a mean absolute deviation from the true smile of around 5.84 basis points. This does not happen to the same extent with the Pseudo-Gibbs imputed volatility cube which exhibits only a mean absolute deviation of approximately 1.05 basis points from the true smile.

An analysis of the mean absolute error of the SABR fits calibrated on VAE-imputed volatilities in the same manner on all different swaptions on the (T_0, T_M) -grid can be found in Fig. 9. The worst fits are reached for swaptions with very short maturity of 1 month which seems natural given the reconstruction behavior of the model shown in Fig. 4. This is in particular in congruence with mean absolute errors for SABR fits calibrated on the observed volatility which exhibit largest misfits in the high tenor-low maturity range as well. For swaptions with maturity between 6 months and 2 years and tenor between 1 and 5 years, the best fit is obtained. A similar goodness-of-fit behavior can be found when averaging the mean absolute errors shown here over all samples in the test set with a maximum MAE of 2.76 basis points obtained for the 1 month–10 year contract. To put this into perspective, a difference of 2.76 basis points in Bachelier implied volatility makes up a 7.37% price difference for the at-the-money swaption contract.

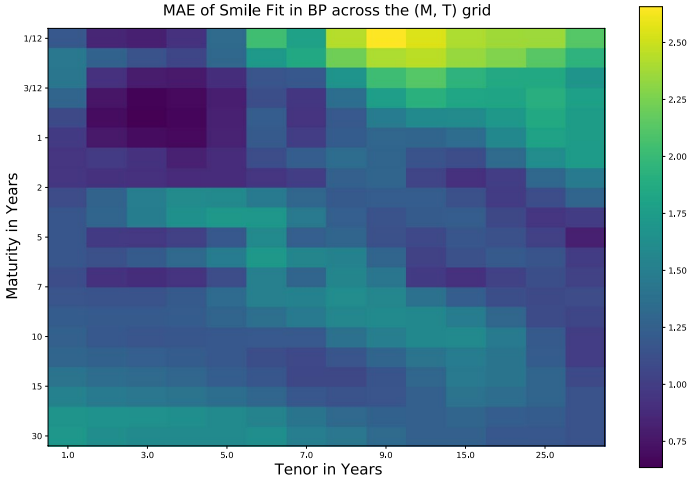

Fig. 9 Mean absolute errors between observed volatilities of different swaptions on 20 Oct 2021 and SABR fitted volatility using the imputed values from the Pseudo-Gibbs algorithm

Next to the direct evaluation of the volatility fits, other evaluation metrics can be applied when judging the quality of a SABR fit that is applied to a Pseudo-Gibbs imputed volatility smile. In particular, the delta hedging performance of the approach is analyzed in the following subsection.

5.3 Delta hedging using VAE-imputed volatility data

In this section, we apply the Pseudo-Gibbs imputation algorithm in the context of delta hedging. Specifically, we will hedge a synthetically simulated European swaption contract maturing 1 year in the future with a tenor of 1 year ($T_0 = 1$, $T_M = T_0 + 1$). The notational prerequisites are summarized briefly in Sect. 2 as well as Appendices A and B . The precise methodology for the hedging study is as follows:

1) On the (T_0, T_M)-grid described in Sect. 4, we fix α_{T_0,T_M}, ν_{T_0,T_M} and ρ_{T_0,T_M} parameters, which were calibrated to the swaption cube on 20 October 2021. β is fixed to a value of 0.5.
2) For each maturity–tenor combination, we generate a 1 year timeseries of forward swap rates F_t by a discretization of the SABR stochastic differential equation. To obtain realistic dependencies between forward swap rates of different maturities and tenors, we use the same normally distributed increments of the Brownian motion processes for trajectory generation on the whole grid. The swaption corresponding to the simulated timeseries for a maturity and a tenor of 1 year is the one we want to hedge.
3) At each point in time, we reconstruct a theoretical swaption volatility cube from the simulated paths by interpolating the forward swap rates on the (T_0, T_M)-grid and by applying (2). Afterwards, we mask 70% of the points on the swaption

volatility cube at each point in time. This is the volatility cube the practitioner is supposed to observe in the market.

4) We reconstruct the full cube using the trained VAE-imputation model at each point in time. Afterwards, the values α_{T_0,T_M}, ν_{T_0,T_M}, and ρ_{T_0,T_M} are calibrated from the reconstructed cube on the (T_0, T_M)-grid. To obtain SABR parameters corresponding to the maturity of the contract we want to hedge, we interpolate the obtained SABR parameters on the (T_0, T_M)-grid once again to the present maturity of the original 1 year swaption.
5) Using the SABR parameters obtained in step 4) at each point in time, we calculate the delta of the swaption that we want to hedge. Using a forward swap with the same maturity and tenor, we obtain a dynamic delta-neutral portfolio as described in Appendix B.

Figure 10 examines the hedging performance of the described approach along a specific simulated path of the 1 year–1 year forward swap rate. We compare the performance of the delta-neutral portfolio obtained by the VAE-imputation algorithm and by interpolation of the cube with missing values to the performance of the theoretical delta-neutral portfolio that could be obtained if the practitioner hedging the swaption had perfect knowledge in advance about the risk-neutral parameters α, β, ν, and ρ generating the observed path. Analogously to the results of Sect. 5.2, this demonstrates the superiority of the VAE imputation approach over interpolation of missing values. The swaption's notional was set to $N = 100\,000$ and payment dates were set to a quarterly tenor after maturity at $T_0 = 1$. For simplicity, we assume 360 trading days per year without weekend effects. Moreover, for the ease of exposition, a deterministic exponential discount rate structure was presumed for the zero-coupon bond values $P(t, T_i)$.

As it is briefly discussed in Appendix B, note that the construction of a delta-neutral portfolio described here does not lead to a perfect replication of a swaption contract by a self-financing portfolio as it is possible in complete stochastic models. Nevertheless, the approach of the current section yields a dynamic assessment of the hedging performance and is comparable with Rebonato et al. (2008).

To track the performance of the described approach along multiple trajectories, we measure the root-mean-squared error (RMSE) between final cumulated predicted price differences and the theoretical final option prices (see the lower part of Fig. 10) along 10 000 simulated paths of the SABR dynamics. The results are shown in Table 1 as percentages of the swaption's notional value of 100 000. One can observe that the differences between the theoretical delta-neutral portfolio and the VAE-imputation portfolio become negligible when using larger rebalancing periods. However, VAE-imputation errors exhibit a much slower decline when decreasing the rebalancing period compared to the theoretical delta-neutral portfolio.

6 Extensions and further research possibilities

In this paper we proposed the usage of variational autoencoders as a novel non-parametric tool for financial data imputation, which has in particular never been applied in the context of sparse financial swaption quotes. The preceding sections

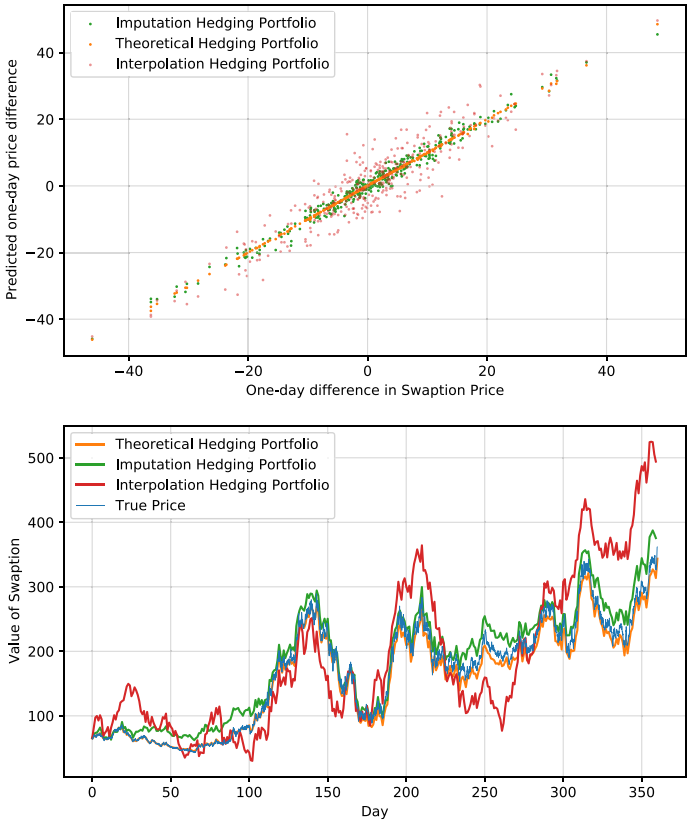

Fig. 10 Upper figure: Actual 1-day changes in swaption price vs. 1-day changes in swaption price predicted by the delta-neutral portfolios using the theoretical hedging portfolio as well as the hedging portfolios obtained from the VAE imputation and interpolation approaches compared in Fig. 8. The R^2 value obtained from a regression of the predicted changes on the actual changes is 0.9778 for the VAE-imputation portfolio, 0.8532 for the interpolation portfolio and 0.9998 for the theoretical hedging portfolio. Lower Figure: The actually realized swaption price was plotted next to the cumulated predicted price differences from the upper figure

Table 1 RMSE between cumulated predicted price differences and theoretical final swaption prices as a percentage of the swaption's notional of 100 000 for rebalancing periods between 1 week and 1 min using the theoretical delta-neutral and the VAE-imputation portfolio

Rebalancing	Theoretical Portfolio	Imputation Portfolio
1 week	$2.059 \cdot 10^{-2}$ %	$3.609 \cdot 10^{-2}$ %
1 day	$1.350 \cdot 10^{-2}$ %	$2.300 \cdot 10^{-2}$ %
1 h	$0.569 \cdot 10^{-2}$ %	$1.482 \cdot 10^{-2}$ %
1 min	$0.155 \cdot 10^{-2}$ %	$1.293 \cdot 10^{-2}$ %

demonstrated how the geometry of an implied swaption volatility cube containing missing values can be inferred by learning stochastic latent representations via VAE in an approximate Gibbs sampling environment, before the imputed estimates of missing quotes can be used to fit the SABR stochastic model on volatilty smiles.

There are plenty of possibilities to extend the basic imputation algorithm presented in the second section. First, Mattei and Frellsen (2018) propose to make use of a Metropolis-within-Gibbs extension of the basic Pseudo-Gibbs imputation of Rezende et al. (2014) to asymptotically sample from the true conditional distribution $p(x_{\text{miss}}|x_{\text{obs}})$ instead of the variational approximation $q(x_{\text{miss}}|x_{\text{obs}})$. Here, instead of sampling $z^{(t+1)}$ from $q_\theta(z|(x_{\text{obs}}, x_{\text{miss}}))$ in step 3) of the algorithm in Sect. 2, we insert a Metropolis–Hastings step in the algorithm which makes sampling from the true posterior asymptotically tractable. This procedure, however, requires very thorough tuning of the Metropolis–Hastings steps due to the very high dimensionality (4998-dimensional volatility data) of the problem. Apart from that, there are a number of variance reduction techniques that can be applied to the basic Gibbs sampling procedure, for example a basic Rao–Blackwellization procedure.

Second, to better capture the highly nonlinear dependence between the components of volatility cubes, other decoder architectures for the pretrained variational autoencoder could prove to be fruitful. For example, convolutional variational autoencoders as well as a different parameterization of the Gaussian covariance matrix as a rank-1 matrix with diagonal correction would be possible. Rezende et al. (2014, Section 4.3) propose the parameterization of the precision $\Sigma^{-1} = D + uu^\top$ where D is diagonal and u is a vector, which allows for arbitrary rotations of the Gaussian distribution along one principle direction with relatively few additional parameters (see Magdon-Ismail and Purnell 2010).

The algorithm presented in Sect. 2 utilizes synthetically generated training data samples, as obtained in Sect. 4. Alternatively, robustness of the Pseudo-Gibbs imputation approach can be studied when the VAE model is trained on samples that already include missing values. Mattei and Frellsen (2019) propose a method based on the importance-weighted autoencoder of Burda et al. (2016) to train deep latent variable models on training data including missing values. Finally, we could also apply a recurrent network structure to improve the predictive power of the variational decoder model by accounting for time dependencies of surfaces or cubes. After all, even if the observed values of today's volatility cube are predicted to come from a cube of a particular shape, the model should predict today's volatility cube to not differ very much of yesterday's cube in shape.

Appendix A: Implied volatility formulas for the SABR model

This appendix summarizes the explicit implied volatility approximations in the SABR model, which allows to compute the current swaption implied volatility from the parameters of the SABR model both in the normal and in the lognormal framework. Let K denote the strike of the swaption, i.e., the fixed interest rate underlying the swap that can be entered by the holder of the swaption at time T_0. Following the asymptotic expansion of Hagan et al. (2002) with slight corrections from

Oblój (2008) and the simplifications in Le Floc'h and Kennedy (2014), we have the approximate formula for $F_t \neq K$ and $\beta \in [0, 1]$

$$\begin{aligned}\sigma_N(t) \approx & \frac{F_t - K}{\hat{x}(\zeta)}\Big[1 + \Big(g + \frac{1}{4}\rho\nu\alpha\beta(F_t + b)^{\frac{\beta-1}{2}}(K + b)^{\frac{\beta-1}{2}} \\ & + \frac{1}{24}(2 - 3\rho^2)\nu^2\Big)(T_0 - t)\Big],\end{aligned} \tag{2}$$

where, for $\beta \in [0, 1]$, the values g, ζ, and $\hat{x}(\zeta)$ are given by

$$\begin{aligned} g &= \frac{1}{24}(\beta^2 - 2\beta)(F_t + b)^{\beta-1}(K + b)^{\beta-1}\alpha^2 \\ \zeta &= \frac{\nu}{\alpha(1 - \beta)}\big((F_t + b)^{1-\beta} - (K + b)^{1-\beta}\big) \qquad (\beta \neq 1) \\ \hat{x}(\zeta) &= \frac{1}{\nu}\log\left(\frac{\sqrt{1 - 2\rho\zeta + \zeta^2} - \rho + \zeta}{1 - \rho}\right),\end{aligned}$$

and where $\zeta = \frac{\nu}{\alpha}\log\left(\frac{F_t+b}{K+b}\right)$ if $\beta = 1$. If $F_t = K$, the normal volatility is given by

$$\sigma_N(t) \approx \alpha(F_t + b)^{\beta}\Big[1 + \Big(g + \frac{1}{4}\rho\nu\alpha\beta(F_t + b)^{\beta-1} + \frac{1}{24}(2 - 3\rho^2)\nu^2\Big)(T_0 - t)\Big]. \tag{3}$$

In the Bachelier model, the forward swap rate under the forward swap measure is modeled by

$$\begin{aligned} \mathrm{d}F_t &= \sigma_N\,\mathrm{d}W_t \\ F_0 &= f, \end{aligned}$$

which possesses the solution $F_t = f + \sigma_N W_t$, i.e., a shifted and scaled Brownian motion. Having obtained the implied normal volatility in the SABR model from (2) or (3) respectively, one can easily obtain the SABR swaption price via the Bachelier-model valuation formulas (see, e.g., Crispoldi et al. (2016)), i.e., for $t \in [0, T_0)$

$$V_t^{\text{Bachelier}} = N \cdot \left[\sum_{i=1}^{m} \delta_i P(t, T_i)\right]\sigma_N\sqrt{T_0 - t}\big(d[\Phi(d) - R] + \varphi(d)\big), \tag{4}$$

where N denotes the notional amount of the swaption, δ_i is a fraction denoting the day-count convention for the period $[T_{i-1}, T_i]$, $P(t, T_i)$ denotes the discount factor for the period $[t, T_i]$ usually measured by the price of an according zero-coupon bond, $d = \frac{F_t - K}{\sigma_N\sqrt{T_0 - t}}$, and Φ and φ denote the cumulative standard normal distribution function and the standard normal density, respectively. The term $\sum_{i=1}^{m} \delta_i P(t, T_i)$ is commonly referred to as the present value of a basis point. The value of R is set to 0 if the swaption is a payer swaption (i.e., the holder of the swaption pays the fixed leg) and set to 1 if the swaption is a receiver swaption (i.e., the holder of the swaption receives the fixed leg). When computing the SABR price of a swaption, (4) is used in conjunction with $\sigma_N(t)$ from (2) or (3).

Appendix B: Delta hedging in the SABR model

The idea of basic dynamic delta hedging a short position in a swaption consists in taking a long position in the forward swap contract corresponding to the swaptions underlying payment structure. The size of this long position at time t will be denoted m_t, while the value $\Delta_t = \frac{\partial V_t^{\text{SABR}}}{\partial F_t}$ will be called the delta of the swaption, where differentiation takes place with respect to the value function in the SABR model. To make the combined portfolio of those two positions independent of fluctuations in the forward swap rate F_t, m_t has to fulfill the condition

$$m_t \frac{\partial V_t^{\text{Swap}}}{\partial F_t} - \Delta_t = 0, \tag{5}$$

where V_t^{Swap} denotes the value of the forward swap contract in the portfolio. To obtain Δ_t in the SABR model, we can decompose

$$\Delta_t = \frac{\partial V_t^{\text{SABR}}}{\partial F_t} = \frac{\partial V_t^{\text{Bachelier}}}{\partial F_t} + \frac{\partial \sigma_N(t)}{\partial F_t} \frac{\partial V_t^{\text{Bachelier}}}{\partial \sigma_N},$$

where $\partial V_t^{\text{Bachelier}} / \partial F_t$ denotes the delta of the swaption in the Bachelier model, i.e., the partial derivative of (4) with respect to F_t and where $\partial V_t^{\text{Bachelier}} / \partial \sigma_N$ denotes the Vega of the swaption in the Bachelier model, i.e., the partial derivative of (4) with respect to σ_N. Plugging in the corresponding partial derivatives calculated from (4) and (2), we obtain for $\beta \in (0, 1)$ and $F_t \neq K$

$$\Delta_t = N \cdot \left[\sum_{i=1}^{m} \delta_i P(t, T_i) \right] \left[\Phi(d) + \sqrt{T_0 - t}\varphi(d)\big(\sigma_N(t) \cdot \kappa + \tau\big) - R \right], \tag{6}$$

where

$$\tau = \frac{F_t - K}{\hat{x}(\zeta)(F_t + b)}(\beta - 1)\left[g + \frac{1}{8}\rho\nu\alpha\beta(F_t + b)^{\frac{\beta-1}{2}}(K + b)^{\frac{\beta-1}{2}} \right](T_0 - t)$$

$$\kappa = (F_t - K)^{-1} - \frac{(F_t + b)^{-\beta}}{\alpha\hat{x}(\zeta)\sqrt{1 - 2\rho\zeta + \zeta^2}}$$

if $F_t \neq K$, whereas for $F_t = K$

$$\tau = \alpha(\beta - 1)(F_t + b)^{\beta-1}\left[g + \frac{1}{8}\rho\nu\alpha\beta(F_t + b)^{\beta-1} \right](T_0 - t)$$

$$\kappa = \frac{1}{2}\left[\beta(F_t + b)^{-1} - \frac{\rho\nu}{\alpha}(F_t + b)^{-\beta} \right].$$

Bartlett (2006) derives an alternative formula for delta in the SABR model that accounts for a forward swap rate change induced jump in instantaneous volatility. For simplicity, we will not focus on Bartlett's delta here. Using the formulas (5) and

(6), the corresponding long position needed in the forward swap for delta neutrality is easily obtained using that V_t^{Swap} is given by

$$V_t^{\text{Swap}} = (1-2R)N \cdot \left[\sum_{i=1}^{m} \delta_i P(t,T_i)\right](F_t - K).$$

Other methods of hedging a swaption include, for example, the use of a portfolio of zero-coupon bonds; see, e.g., Dun et al. (2001). Note that the constructed delta-neutral portfolio approach described here does not replicate the swaption contract perfectly like it is the case, e.g., in the plain-vanilla Black model. Nevertheless, examining the differences between actual swaption price changes and changes in the dynamic portfolio value between two rebalancing dates, as it is done, e.g., by Rebonato et al. (2008), yields an assessment of the hedging performance of a particular stochastic model like it was done in Sect. 5.3.

Appendix C: Details of the Pseudo-Gibbs sampling imputation

In this appendix, we describe the details of the Pseudo-Gibbs sampling imputation method from Sect. 2. Recall that we assume that the remaining data x can be decomposed into an observed and into a missing component by $x = (x_{\text{obs}}, x_{\text{miss}})$ and that the goal lies in sampling from the conditional distribution of x_{miss} given x_{obs}. The Gibbs sampling-inspired approach now samples from the conditional joint distribution of the random vector (x_{miss}, z) given x_{obs}, from which the required conditional distribution of x_{miss} given x_{obs} can be obtained via marginalization of the samples from the distribution of (x_{miss}, z) given x_{obs}. The algorithm for missing data imputation proceeds in the following way:

1) Train a variational autoencoder to learn the encoder distribution $q_\theta(z|x)$ and the decoder distribution $p_\vartheta(x|z)$.
2) Choose starting values $x_{\text{miss}}^{(0)}$ and set $t = 0$.
3) Simulate successively
$$z^{(t+1)} \sim q_\theta(z|(x_{\text{obs}}, x_{\text{miss}}^{(t)})),$$
$$x_{\text{miss}}^{(t+1)} \sim p_\vartheta(x_{\text{miss}}|x_{\text{obs}}, z^{(t+1)}).$$
Here, $p_\vartheta(x_{\text{miss}}|x_{\text{obs}}, z^{(t+1)})$ denotes the conditional distribution of the missing data given the observed data and the current latent code which is obtained by conditioning p_ϑ on x_{obs}. In the following, as it is typical, the decoder p_ϑ is modeled by a multivariate Gaussian with diagonal covariance matrix, and hence, the conditioning on x_{obs} can effectively be omitted.
4) Increment t and return to step 3) until the maximum number T of iterations is reached. Then, go to step 5).
5) Obtain a sequence of marginals $(x_{\text{miss}}^{(t)})_{0\le t\le T}$ by discarding $z^{(t)}$ from the obtained Markov chain of vectors $(x_{\text{miss}}, z^{(t)})_{t\in\mathbb{N}_0}$.

Clearly, the process $(x_{\text{miss}}^{(t)})_{0\le t\le T}$ generated by the above is a Markov chain. Roberts and Smith (1994) give conditions for aperiodicity and irreducibility of the Gibbs chain resulting in convergence to the stationary distribution. Moreover, under rather mild conditions, it can be shown (see Geman and Geman 1984) that the stationary distribution of the above Markov chain comes out as an approximation of the true conditional distribution of x_{miss} given x_{obs} which we call

$$q(x_{\text{miss}}|x_{\text{obs}}) := \int p_\vartheta(x_{\text{miss}}|z, x_{\text{obs}})q_\theta(z|x_{\text{obs}})\,\mathrm{d}z.$$

The above procedure is termed Pseudo-Gibbs sampling. This is due to the fact that in step 3) of the above algorithm, we just use the variational posterior approximation q_θ instead of the true posterior $p(z|x)$. If $q_\theta(z|x)$ and $p(z|x)$ coincide, the above algorithm exactly coincides with the Gibbs sampling framework and we obtain samples from the true conditional distribution

$$p(x_{\text{miss}}|x_{\text{obs}}) = \int p_\vartheta(x_{\text{miss}}|z, x_{\text{obs}})p(z|x_{\text{obs}})\,\mathrm{d}z.$$

Using Birkhoff's ergodic theorem, we can use realizations from the chain to estimate the expectation of any integrable function of x_{miss} conditional on x_{obs} under q.

In practice, the first M values of the chain are discarded as a "burn-in" where M is chosen sufficiently large to ensure convergence of the chain to the stationary

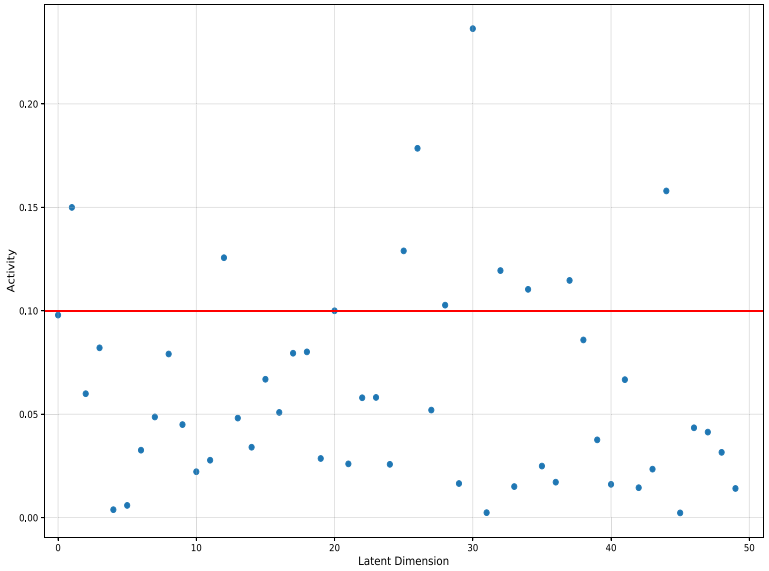

Fig. 11 Latent activity statistics for a VAE model with 50 bottleneck units. Approximately 11 of the 50 units were active in the sense described above, motivating the employed VAE architecture with 10 hidden units

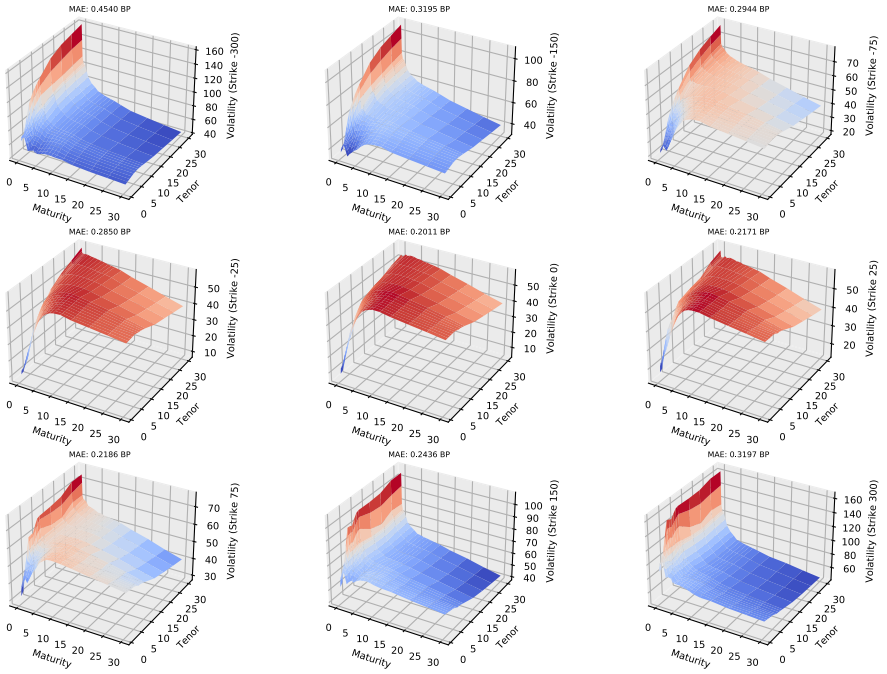

Fig. 12 Approximately 79.6% of the volatility quotes of the cube on 20 October 2021 were masked and treated like missing data. The cube depicted is the obtained cube after missing data imputation with the Pseudo-Gibbs sampling approach described above

distribution. Finally, after obtaining the Markov chain $(x_{\text{miss}}^{(t)})_{M \leq t \leq T}$, we impute the missing values x_{miss} of the data sample by an estimator of the conditional expectation of x_{miss} given x_{obs} under $q(x_{\text{miss}}|x_{\text{obs}})$ which, due to Birkhoff's ergodic theorem, is given by the sample average

$$\hat{x}_{\text{miss}} := \hat{\mathbb{E}}_{q(x_{\text{miss}}|x_{\text{obs}})}(x_{\text{miss}}|x_{\text{obs}}) := \frac{1}{T-M+1} \sum_{t=M}^{T} x_{\text{miss}}^{(t)}. \tag{7}$$

Appendix D: model architecture and considered market cubes

Figures 12 and 13 show the considered market swaption volatility cube observed on 20 October 2021 and its Pseudo-Gibbs reconstruction, the differences of which are shown in Fig. 4.

The trained VAE inference model used throughout the paper was a standard variational autoencoder with a ten-dimensional latent space which was trained for 50,000 epochs on 10,000 synthetically generated swaption cubes by the procedure described in Sect. 4. The choice of dimensionality for the bottleneck layer was based on a latent

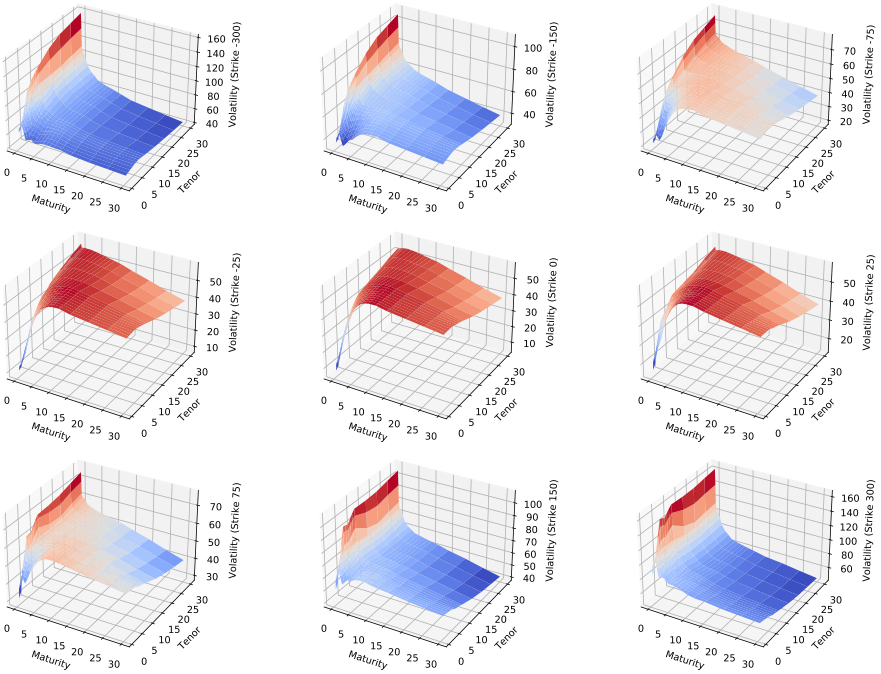

Fig. 13 Market observed volatility cube on 20 October 2021

activity statistic proposed by Burda et al. (2016): The activity of each latent node is measured by

$$A_u := \mathrm{Var}_x\left(\mathbb{E}_{q_\theta(z|x)}(z)\right)$$

and we call a node inactive if $A_u < 0.1$.[2] Figure 11 shows A_u for a trained VAE model with 50 latent units. One can see that of the 50 units approximately 11 units are active which motivates our VAE architecture with 10 hidden units. Using this architecture, all units remain active.

Both the encoder and decoder submodel of the VAE were equipped with four hidden layers with 250, 200, 150, and 100 units, respectively, using ReLU activations. The layer weights were initialized normally distributed with a variance of 1/30. In our experiments, we found that fine-tuning the kernel initializer variance had quite a large impact on the stability of training on synthetic data. The employed optimizer was the Adam algorithm with a learning rate of 10^{-6}.

Acknowledgements We gratefully acknowledge the financial support from the Fraunhofer Institute for Industrial Mathematics ITWM.

[2] In their paper, Burda et al. (2016) chose an activity threshold of 0.01.

Funding Open Access funding enabled and organized by Projekt DEAL. This research was funded by Fraunhofer Institute for Industrial Mathematics ITWM.

Data availability The data that support the findings of this study are available from Refinitiv Eikon, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Andreichenko P (2011) A parsimonious model for the joint evolution of yield vurces and the interest rate smile surface under the objective measure. Msc thesis, University of Oxford

Antonov A, Konikov M, Spector M (2019) Modern SABR analytics: formulas and insights for quants, former physicists and mathematicians. Springer

Arvanitidis G, Hansen LK, Hauberg S (2018) Latent space oddity: on the curvature of deep generative models. In: Proceedings of International Conference on Learning Representations

Bachman P, Precup D (2015) Data generation as sequential decision making. Advances in neural information processing systems. Springer, Cham

Bartlett B (2006) Hedging under the SABR model. Wilmott Mag 04(06):2–4

Brigo D, Mercurio F (2007) Interest rate models: theory and practice. Springer, Berlin

Burda Y, Grosse R, Salakhutdinov R (2016) Importance weighted autoencoders. In: 4th International Conference on Learning Representations (ICLR)

Camino R, Hammerschmidt C, State R (2019) Improving missing data imputation with deep generative models. arXiv:1902.10666

Collier M, Nazabal A, Williams C (2020) VAEs in the presence of missing data. arXiv:2006.05301

Crispoldi C, Wigger G, Larkin P (2016) SABR and SABR LIBOR market models in practice: with examples implemented in Python. Palgrave MacMillan, Hampshire

Dai B, Wang Y, Aston J, Wipf D (2018) Connections with robust PCA and the role of emergent sparsity in variational autoencoder models. J Mach Learn Res 19(1):1–42

Dimitroff G, de Kock J (2011) Calibrating and completing the volatility cube in the SABR model. In: Berichte des Fraunhofer-Instituts for Techno- und Wirtschaftsmathematik (ITWM Report 202)

Du C, Zhu J, Zhang B (2018) Learning deep generative models with doubly stochastic gradient MCMC. IEEE Trans Neural Netw Learn Syst 29(7):3084–3096

Dun T, Schlögl E, Barton G (2001) Simulated swaption delta-hedging in the lognormal forward LIBOR model. Int J Theor Appl Finance 4(4):677–709

Dutilleul P (1999) The MLE algorithm for the matrix normal distribution. J Stat Comput Simul 64(2):105–123

Flegal JM, Jones GL (2011) Implementing MCMC: estimating with confidence. In: Brooks S, Gelman A, Jones GL, Meng X (eds) Handbook of Markov Chain Monte Carlo. CHapman and Hall/CRC, New York, pp 175–197

Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans Pattern Anal Mach Intell 6(6):721–741
Gondara L, Wang K (2018) Mida: multiple imputation using denoising autoencoders. Advances in knowledge discovery and data mining. Springer, Cham, pp 260–272
Goyal V et al (2019) Missing data imputation by principal component analysis (PCA) and fuzzy C means (FCM). Int J Control Autom 12(6):127–134
Hagan P, Kumar D, Lesniewski A, Woodward DE (2002) Managing smile risk. Wilmott Mag 01(02):84–108
Hagan P, Konikov M (2004) Interest rate volatility cube: construction and use. Technical Report, Bloomberg Technical Reports
Hauberg S (2018) Only Bayes should learn a manifold (on estimation of differential geometric structure from data). arXiv:1806.04994
Ipsen NB, Mattei PA, Frellsen J (2021) not-MIWAE: deep generative modelling with missing not at random data. In: ICLR 2021 International Conference on Learning Representations
Ivanov O, Figurnov M, Vetrov D (2019) Variational autoencoder with arbitrary conditioning. In: Proceedings of the 7th International Conference on Learning Representations, pp 1–25
Jäckel P, Rebonato (2000) Linking caplet and swaption volatilities in a GBM/J framework: approximate solutions. Quantitative Research Centre. The Royal Bank of Scotland
Johnson S, Nonas B (2009) Arbitrage-free construction of the swaption cube. Wilmott J 1(3):137–143
Kingma D, Welling M (2013) Auto-encoding variational Bayes. In: Paper presented at 2nd International Conference on Learning Representations, ICLR, Banff, AB, Canada, April 14-16. Technical Report
Kunsági-Máté S, Fáth G, Csabai I (2021) Analyizing the dynamics of the swaption market using neural networks. Eur J Econ 1(2):1–13
Le Floc'h F, Kennedy G (2014) Explicit SABR calibration through simple expansions. SSRN. https://doi.org/10.2139/ssrn.2467231
Lewis S, Matejovicova T, Li Y, Lamb A, Zaykov Y, Allamanis M, Zhang C (2021) Accurate imputation and efficient data acquisition with transformer-based vaes. In: NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications, 2021
Li C, Zhu J, Zhang B (2016) Learning to generate with memory. In: Proceedings of The 33rd International Conference on Machine Learning, pp 1177–1186
Magdon-Ismail M, Purnell JT (2010) Approximating the covariance matrix of GMMs with low-rank perturbations. Int Conf Intell Data Eng Autom Learn 2010:300–307
Ma C, Tschiatschek S, Hernandez-Lobato JM, Turner R, Zhang C (2020) Vaem: a deep generative model for heterogeneous mixed type data. In: 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada
Mattei PA, Frellsen J (2018) Leveraging the exact likelihood of deep latent variable models. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems, pp 3859–3870
Mattei PA, Frellsen J (2019) MIWAR: Deep generative modelling and imputation of incomplete data sets. Proc Int Conf Mach Learn 97:4413–4423
Ma C, Zhang C (2021) Identifiable generative models for missing not at random data imputation. In: 35th Conference on Neural Information Processing Systems (NeurIPS 2021)
Nazabal A, Olmos PM, Ghahramani Z, Valera I (2020) Handling incomplete heterogeneous data using vaes. Pattern Recognit 107:107501
Oblój J (2008) Fine-tune your simle: Correction to Hagan et al. Wilmott Magazine 01/08, pp 102- 104
Qiu YL, Zheng H, Gevaert O (2023) Genomic data imputation with variational auto-encoders. GigaScience 9(8):giaa082
Rebonato R, Pogudin A, White R (2008) Delta and vega hedging in the SABR and LMM-SABR models. Risk Magazine, December 2008
Rezende D, Mohamed S, Wierstra D (2014) Stochastic backpropagation and approximate inference in deep generative models. Proc Int Conf Mach Learn 32(2):1278–1286
Rezende D, Eslami SMA, Mohamed S, Battaglia P, Jaderberg M, Heess N (2016) Unsupervised learning of 3D structure from images. In: Advances in Neural Information Processing Systems, pp 4996–5004
Roberts GO, Smith AFM (1994) Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. Stoch Processes Appl 49(2):207–216
Roskams-Hieter B, Wells J, Wade S (2022) Leveraging variational autoencoders for multiple data imputation. arXiv:2209.15321

Skantzos N, Garston G (2019) The perfect smile. Filling the gaps in the swaption volatility cube. Deloitte Belgium. https://www2.deloitte.com/content/dam/Deloitte/be/Documents/risk/deloitte-be-the-perfect-smile.pdf. Accessed 19 April 2021

Sohl-Dickstein J, Weiss E, Maheswaranathan N, Ganguli S (2015) Deep unsupervised learning using nonequilibrium thermodynamics. In: Proceedings of the 32nd International Conference on Machine Learning 37, pp 2256–2265

Thorin H (2020) Artifical neural networks for SABR model calibration & hedging. Msc thesis, Imperial College London

Vincent P, Larochelle H, Bengio Y, Manzagol PA (2008) Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th International Conference on Machine Learning, pp 1096–1103

Wang H, Chen H, Sudjianto A, Liu R, Shen Q (2018) Deep learning-based BSDE solver for Libor market model with application to Bermudan swaption pricing and hedging. arXiv:1807.06622

West G (2005) Calibration of the SABR model in illiquid markets. Appl Math Finance 12(4):371–385