Autoencoder Market Models for Interest Rates

Alexander Sokol
CompatibL

December 13, 2022

Abstract

We propose a highly optimized latent factor representation of the yield curve obtained by training a variational autoencoder (VAE) to curve data from multiple currencies. A curious byproduct of such training is a "world map of latent space" where neighbors have similar curve shapes, and distant lands have disparate curve shapes. The proposed VAE-based mapping offers a high degree of parsimony, in some cases achieving similar accuracy to classical methods with one more state variable.

In the second part of the paper, we describe four types of autoencoder market models (AEMM) in Q- and P-measure. Each autoencoder-based model starts from a popular classical model and replaces its state variables with autoencoder latent variables. This replacement leads to a greater similarity between the curves generated by the model and historically observed curves, a desirable feature in both Q- and P-measure. By aggressively eliminating invalid curve shapes from its latent space, VAE prevents them from appearing within the model without intricate constraints on the stochastic process used by the classical models for the same purpose. This makes VAE-based models more robust and simplifies their calibration.

Potential applications of the new models and VAE-based latent factor representation they are based on are discussed.

Keywords: Autoencoder Market Model, AEMM, Interest Rates, Nelson-Siegel, Machine Learning, ML, Autoencoder, Variational Autoencoder, VAE

JEL: C01, C14, E43, E47, G12, G17

Contents

1. Introduction

Properties of a term structure interest rate model are to a large extent determined by the choice of its state variables. Having too many negatively affects performance and causes parameter estimation issues. Having too few or choosing them poorly causes the model to miss certain risks. This is why dimension reduction, i.e., decreasing the number of state variables with the least possible loss of accuracy, is of paramount importance.

In one of the early successes of using machine learning for dimension reduction, Kondratyev [1] demonstrated that feedforward neural networks outperform classical regression techniques such as PCA in describing the evolution of interest rate curve shapes in P-measure. Bergeron *et al.* [2] and Buehler *et al.* [3] used VAE to reduce the dimension of the volatility surface. In this paper, we derive a highly optimized VAE-based representation of the yield curve and propose a new category of interest rate models in Q- and P-measure that produce VAE-generated curve shapes to which minimal corrections are applied to keep the model arbitrage-free.

Dimension reduction is a compression algorithm, not unlike those used to compress an image. The maximum possible degree of compression depends on the universe of images the algorithm is designed for. Because JPEG and similar general-purpose image compression algorithms impose no restrictions on what the image can depict, they have a moderate rate of compression (around x10 for JPEG). For these general-purpose algorithms, dimensions of the compressed data (i.e., bits of the compressed file) are local, in the sense that each of them encodes information from a group of nearby pixels.

Variational autoencoders (VAE) are machine learning algorithms that provide a fundamentally different type of compression. The rate of compression they can achieve is multiple orders of magnitude higher than general-purpose compression algorithms. Such a tremendous performance gain can only be achieved by training the algorithm to compress a specific type of image, such as the image of a human face. In the process of aggressively eliminating implausible combinations of pixels in pursuit of better compression, something quite remarkable happens – dimensions of the compressed image acquire meaning.

When using VAE to encode images of a human face, dimensions of the compressed data (the latent variables, named after the Latin word "lateo" which means "hidden") become associated with realistic changes to the image of a human face, such as adding a smile or changing hair color. This happens for the simple reason that the only combinations of pixels not eliminated by training on a large library of human face images are those that correspond to realistic faces. In machine learning, this "feature extraction" effect is frequently a more important objective than compression itself. The latent factors obtained in this manner are global because they can affect pixels that may be far away from each other in the image (e.g., a dimension that encodes hair color).

Let us now contemplate what a similar approach can do for the interest rate term structure models. The first thing to consider is what we should be compressing. In order to build a VAE-based counterpart to a stochastic volatility term structure model such as SABR-LMM [4], we would need to compress both the yield curve and the volatility surface into a single latent space (throughout the paper, we will be using the term "volatility surface" generically to describe volatility surface or

volatility cube). For deterministic volatility term structure models, the volatility surface is a function of the yield curve, and accordingly the yield curve is the only thing we need to compress. This is the model type we will focus on in this paper. VAE-based stochastic volatility models will be described in a separate publication.

Continuing with the image analogy, a smoothing spline fit to the yield curve is similar to JPEG in the sense that its dimensions and the structure it imposes are both local. On the other hand, the Nelson-Siegel [5] basis is similar to VAE in the sense that its dimensions (roughly corresponding to the level, slope, and convexity) and the structure it imposes (for example, not permitting a curve to have both minimum and maximum at the same time) are both global. Having global dimensions leads to a higher degree of compression for the Nelson-Siegel basis compared to the smoothing spline.

Is aggressive dimension reduction necessary for term structure interest rate models, and can machine learning help find a more effective way to achieve it than the Nelson-Siegel basis? We answer both questions in the affirmative. Many of the interest rate models popular with the practitioners, including multi-factor short rate models [6, 7], Cheyette model [8] and others, are Markovian in a small number of state variables, usually between two and four. Considering the aggressive dimension reduction that must occur when an extraordinary variety of historical yield curve shapes is compressed into a small number of state variables, a sophisticated compression algorithm is clearly required. And yet, the majority of classical models use an exogenously specified SDE or factor basis whose selection is driven by criteria unrelated to optimal compression.

In this paper, we describe several models in Q- and P-measure that start from a popular classical model specification and replace the classical model's state variables by autoencoder latent variables. This replacement leads to greater similarity between curves generated by the model and historically observed curves, a desirable feature in both Q- and P-measure. We called the new model category "autoencoder market models", or AEMM.

Aggressively eliminating unfeasible curve shapes by VAE training creates state variables that represent only valid curves. This prevents AEMM from generating unrealistic curve shapes without using intricate constraints on curve dynamics.

The improvement in accuracy of mapping historical curve observations to model state variables brought about by the use of autoencoders can be measured from the first principles, providing a rigorous way to compare the proposed machine learning approach with its classical counterparts. Our results indicate that the use of autoencoders leads to a significant and measurable improvement in the accuracy of representing complex curve shapes compared to classical methods with the same number of state variables. In turn, this makes AEMM perform better compared to the corresponding classical models.

The rest of the paper is organized as follows. We describe machine learning architecture and present the results of using autoencoders to compress LIBOR and OIS swap curves in Chapter 2. After that, we describe how to convert four distinct classical model types to AEMM by switching from classical state variables to VAE latent variables. In each case, the objective of such conversion is to make curve shapes generated by the model closer to the prevailing historical curve shapes for the model's currency than would have been possible with the original classical model. The conversion of Q-measure models is discussed in Chapter 3 and P-measure models

in Chapter 4. The paper concludes with Chapter 5 where we summarize key results.

2. Autoencoders for the Yield Curve

2.1. Curve Representation

There are three options for how historical yield curve observations can be converted to autoencoder input: (a) raw market quotes (e.g., swap rates, bond yields, etc.), (b) continuously compounded zero coupon rates, the parameterization used by the Nelson-Siegel basis [5], and (c) piecewise constant forward rates, called unsmoothed Fama-Bliss rates [9] in economic literature. With option (a), the curve builder uses autoencoder output. With options (b) and (c), the autoencoder uses curve builder output. Option (a) has the lowest dimension of input space and does not involve curve construction before the data is provided to the autoencoder, making the results easier to replicate. However, it requires that times-to-maturity align perfectly across currencies, which happens for some market quotes but not others.

We will focus on encoding LIBOR and OIS swap curves in this paper. Practitioners construct the long end of the curve from swap quotes, and its short end from other liquid instruments. While times-to-maturity of swap rates are constant and match across most currencies, shorter-dated instruments differ by currency and some of them have variable times-to-maturity (e.g., interest rate futures). Including these instruments would be incompatible with option (a) and would also increase the dimension of input space.

It is well known that even with its three latent factors, the Nelson-Siegel [5] basis has difficulty fitting the short end of the curve. This difficulty was specifically cited by Svensson [10] as the primary reason for the introduction of the Nelson-Siegel-Svensson basis with four latent factors. The need for an additional latent factor is plainly visible by looking at the historical curve shapes, where the initial segment built from instruments with maturity of 2 years or less has frequent shape variations that are not strongly correlated with the rest of the curve.

In this paper, our objective will be to develop an alternative to the Nelson-Siegel basis for compressing swap rates with maturities from 2 years to 30 years. An alternative to the Nelson-Siegel-Svensson basis for compressing curves with additional instruments for the short end will be described in [11].

Limiting our compression target to the curve segment from 2y to 30y has two major benefits. Because times-to-maturity of the swap rates are standard and aligned, we can choose option (a) for curve representation. Under this option, autoencoder input is simply the set of $N = 7$ liquid swap rates with maturities from 2y to 30y. This choice of input, made possible by limiting our encoding to swap rates only, contributes to replicability because VAE results can be reproduced without curve building, an intricate process that requires subjective choices and heuristics to stitch data from multiple instrument types into a well-behaved curve.

The second, equally important, benefit of compressing only the swap rates is avoiding interpolation. With the training algorithm described in this paper, VAE does not receive any information whatsoever, directly or indirectly, about what swap maturities are or even the maturity order when generating curve shapes. This makes for a more rigorous test of VAE where the possibility of "fitting the curve builder" rather than "fitting the curve" is completely eliminated.

2.2. Classical Nelson-Siegel Basis

Before building an autoencoder-based curve representation, we will briefly review our classical benchmark. The Nelson-Siegel basis is specified in terms of the continuously compounded yield $R(t, t+\tau)$ of a zero coupon bond observed at time t for time-to-maturity $\tau = T - t$ ("zero rate"). The zero rate is the average of the instantaneous forward rate $f(t, T')$ over time interval $t < T' < t + \tau$:

$$R(t, t+\tau) = \frac{1}{\tau} \int_t^{t+\tau} f(t, T') dT' \tag{1}$$

The canonical form of the Nelson-Siegel basis is given by:

$$R(t, t+\tau) = \beta_1 + \beta_2 \left(\frac{1 - e^{-\lambda\tau}}{\lambda\tau} \right) + \beta_3 \left(\frac{1 - e^{-\lambda\tau}}{\lambda\tau} - e^{-\lambda\tau} \right) \tag{2}$$

where the three latent factors $\beta_{1,2,3}$ have simple interpretation as the parallel shift, slope, and curvature of the instantaneous forward curve.

To gain insight into the origins of the peculiar linear-exponential form of the Nelson-Siegel basis, it is highly illuminating to review the justification for choosing this form presented by its authors in [5]. In their paper, Nelson and Siegel described two simple parametric forms for the instantaneous forward rate $f(t, T)$. One of these forms was swiftly rejected due to having too many parameters, while the other became the canonical Nelson-Siegel formula (2) after being converted from the instantaneous forward rate $f(t, T)$ to the zero rate $R(t, t+\tau)$.

The first, subsequently rejected, form represented the instantaneous forward rate $f(t, T)$ as the sum of a constant and two mean reverting terms, each with its own rate of decay λ_i:

$$f(t, t+\tau) = \beta_1 + \beta_2 e^{-\lambda_2\tau} + \beta_3 e^{-\lambda_3\tau} \tag{3}$$

When a similar functional form with two exponents is encountered in the context of two-factor Hull-White (HW2F) model [6], the typical calibration makes the two rates of decay $\lambda_{2,3}$ very different, one representing slow and the other fast reversion to the mean. Nelson and Siegel however did exactly the opposite and proposed to make the difference between the two parameters $\lambda_{2,3}$ infinitesimally small. Using Taylor expansion of (3) in $d\lambda = \lambda_2 - \lambda_3$ and omitting higher order terms, they arrived at the linear-exponential form of their basis:

$$f(t, t+\tau) = \beta_1 + \beta_2 e^{-\lambda\tau} + \beta_3 \lambda\tau e^{-\lambda\tau} \tag{4}$$

Substituting this form into (1), we obtain the canonical expression (2) for the zero rate.

An important takeaway from this brief review of the origins of the Nelson-Siegel formula is that its distinctive linear-exponential form is somewhat of an accident, resulting from an attempt to reduce the number of parameters after starting from a more conventional exponential form and omitting higher order terms in the Taylor expansion. While the Nelson-Siegel basis achieves higher degree of parsimony in representing curve shapes compared to local representations such as the smoothing spline, it is clear that achieving the maximum degree of compression was not the explicit objective of its selection. As a result, large areas of the three-dimensional

latent space of the Nelson-Siegel basis correspond to unrealistic curves. We will aim to increase the degree of parsimony by using VAE to aggressively eliminate unfeasible curve shapes, creating a more compact latent space where historical data is densely packed.

2.3. Machine Learning Architecture

2.3.1. Variational Autoencoder (VAE)

A conventional autoencoder consists of an encoder followed by a decoder. The encoder performs compression of a high dimensional vector of input variables (input vector) into a low-dimensional vector of latent variables (latent vector). The decoder performs decompression (i.e., approximate reconstruction) of the input vector from the latent vector. The encoder and decoder are trained together to minimize the difference between the original and reconstructed input vector (reconstruction loss) over a large number of data samples as shown in Figure 1(a).

For our purposes, encoding and decoding must be continuous including derivatives (continuity) with no rapid changes in gradient (regularity). Conventional autoencoders have no inherent mechanism of ensuring continuity and regularity and must resort to ad-hoc methods such as adding randomness to the neural network weights. The variational autoencoder (VAE) developed by Kingma and Welling [12] pursues the same objective in a more systematic way by encoding the input vector into the parameters of a distribution in latent space rather than a single point. During training, a random sample is drawn from that distribution in a way that permits backpropagation ("reparameterization trick") and the input is reconstructed from this sample as shown in Figure 1(b). The distribution we will use for this purpose is diagonal normal $N(\mu_k, \sigma_k)$ parameterized by its mean μ_k and standard deviation σ_k for each dimension $k = 1 \dots K$ of the latent space. In this case, the dimension of the encoder output is $2K$ and the dimension of the decoder input is K.

The sampling step in VAE acts as a powerful regularizer because it penalizes reconstruction loss not only for the training samples, but also their latent space vicinity. This makes VAE not only extraordinarily resistant to overfitting, but also less data-hungry than other machine learning algorithms. With the smaller size of training datasets in capital markets compared to other areas where machine learning is used, the latter feature is especially valuable.

The contribution to VAE loss from each observation is the sum of L2 swap rate reconstruction loss D_{L2} scaled by the number of input swap rates N and β-weighted Kullback-Leibler divergence D_{KLD} between encoder output $N(\mu_k, \sigma_k)$ for that observation and K-dimensional diagonal standard normal distribution $N(0, 1)$:

$$
\begin{aligned}
D_{\mathrm{VAE}}(\boldsymbol{S}, \boldsymbol{S'}, \boldsymbol{\mu}, \boldsymbol{\sigma}) &= \frac{1}{N} D_{\mathrm{L2}}(\boldsymbol{S}, \boldsymbol{S'}) + \beta D_{\mathrm{KLD}}(\boldsymbol{\mu}, \boldsymbol{\sigma}) \\
D_{\mathrm{L2}}(\boldsymbol{S}, \boldsymbol{S'}) &= \sum_{n=1}^{N} (S'_n - S_n)^2 \\
D_{\mathrm{KLD}}(\boldsymbol{\mu}, \boldsymbol{\sigma}) &= \frac{1}{2} \sum_{k=1}^{K} \left(\sigma_k^2 + \mu_k^2 - 1 - \ln(\sigma_k^2) \right)
\end{aligned}
\tag{5}
$$

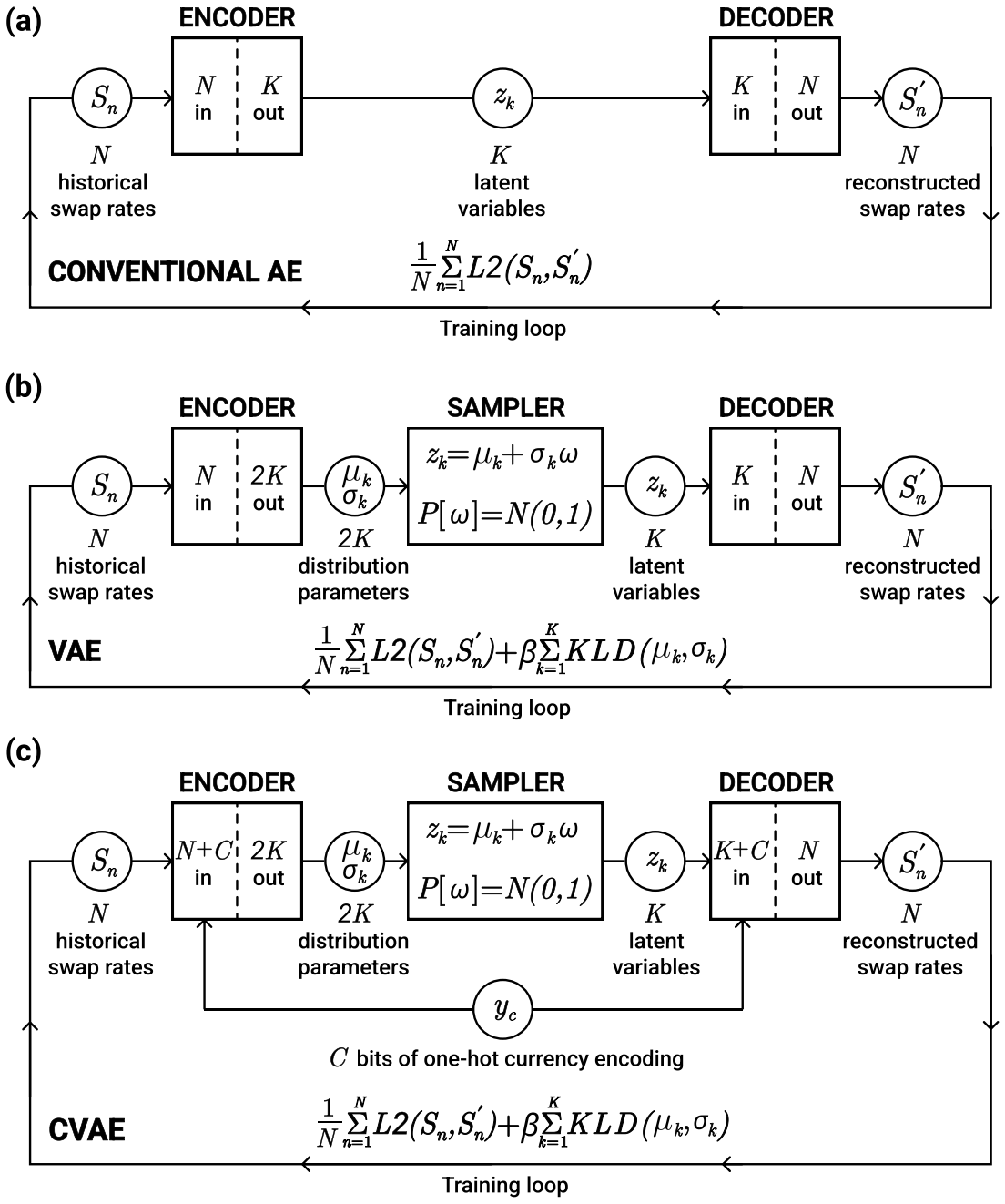

Figure 1: (a) Conventional autoencoder (AE) (b) Single- or multi-currency variational autoencoder (VAE) (c) Multi-currency conditional VAE (CVAE).

where S_n are historical swap rates and S'_n their reconstruction. We will be using boldface to indicate vectors throughout the paper. The two components of D_{VAE} are shown schematically in Figure 2.

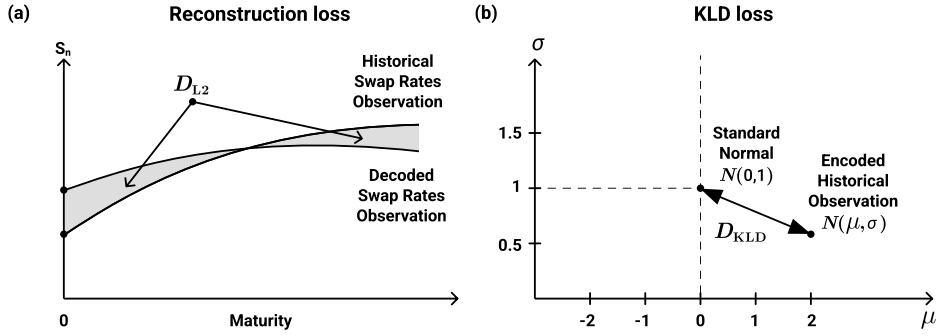

Figure 2: Two components of loss for a single historical observation of swap rates S_n used in VAE training: (a) L2 reconstruction loss between the historical swap rates S_n and decoded swap rates S'_n (b) KLD of the distribution $N(\mu_k, \sigma_k)$ obtained by encoding the historical swap rates S_n and diagonal standard normal $N(0, 1)$ in latent space.

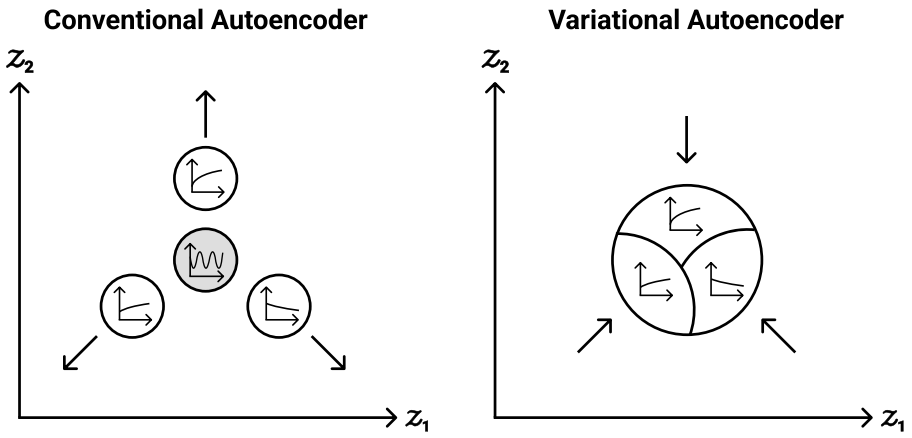

Figure 3: Areas where distinct historical curve shapes (non-shaded circles) are encoded in the latent space of (a) conventional autoencoder and (b) variational autoencoder (VAE). Invalid curve shapes (shaded circle) appear inside gaps in latent space where no historical data is encoded. These gaps are suppressed by the KLD term in VAE.

The reconstruction loss D_{L2} acts as a repelling force that pushes historical curve shapes away from each other in latent space to ensure each can be accurately decoded without mixing. This causes the latent space of most conventional autoencoders to contain large gaps where no historical data is encoded. In the absence of a penalty term that would discourage the appearance of such gaps, they reduce reconstruction loss by creating physical separation between distinct yield curve shapes as shown in

Figure 3(a). When a point inside one of these latent space gaps is decoded, it will likely produce an unrealistic curve shape because there are no historical data points nearby.

In VAE, the KLD loss D_{KLD} pulls (μ_k, σ_k) toward $(0, 1)$ as shown in Figure 3(b). However, if each observation were to be encoded into the same point $(0, 1)$, the decoder would be unable to distinguish between them and L2 reconstruction loss would increase dramatically. With small enough β, the mild attractive force exerted by KLD loss makes areas of latent space where distinct curve shapes are encoded move closer to each other and "touch", but do not overlap. This is the mechanism by which VAE eliminates gaps in latent space and generates a continuous and well-regularized mapping between the input space and the latent space, as shown in Figure 3(b).

After training, the sampling step is eliminated because the randomness it adds is no longer needed. This is usually done by discarding σ_k and using μ_k to obtain a deterministic mapping. Doing so is not always optimal because the center $\boldsymbol{\mu}(\boldsymbol{S})$ of the distribution that minimizes the sum of L2 and KLD loss during training is not necessarily the point in latent space that minimizes L2 loss for encoding a specific input vector S_n, regardless of whether or not such vector comes from an observation in the training dataset.

A post-processing step described in [11] increases the accuracy of VAE mapping by performing gradient descent minimizing L2 loss starting from the center $\boldsymbol{\mu}(\boldsymbol{S})$ of the distribution produced by the encoder. This additional step must be implemented in a way that ensures continuity and regularity of the mapping from the input space to the latent space.

2.3.2. Conditional VAE (CVAE)

When VAE is trained to single-currency data, its only input is the yield curve. When it is trained to multi-currency data, the currency label can be provided as an additional input to both the encoder and decoder as shown in Figure 1(c) in order to tailor the latent representation to the prevailing curve shapes for a specific currency. The architecture that makes data label part of both the encoder and decoder input was proposed by Sohn, Lee, and Yan [13] and is known as "conditional variational autoencoder", or CVAE.

In some publications, the term CVAE is used to describe a different architecture where the data label is an input to the encoder but an output of the decoder, rather than an input to both like in Figure 1(c). This alternative type of CVAE, proposed by Kingma, Rezende, Mohamed and Welling [14], generates both the reconstructed data and a prediction for the data label. For example, it can simultaneously reconstruct the shape of a handwritten digit and determine what digit it is. Doing so only makes sense when the mapping between the data and the label is unique and can be learned. In our case, the mapping is not unique as similar curve shapes occur for different currencies, nor are we interested in teaching VAE to play the guessing game "what currency is this curve for". Accordingly, we will not be using this alternative type of CVAE in this paper.

Encoding the currency label as an ordinal number (e.g., the sequential number of currency in an alphabetical sequence) and making this number an additional autoencoder input as shown in Figure 4(a) would cause the algorithm to erroneously

assume that currencies with ordinal numbers next to each other are similar, causing bias. Such bias can be avoided if each of C currency labels is converted into a binary sequence of length C where the bit that corresponds to the observation's currency is set to 1 and all other bits to 0. This approach, called "one-hot encoding", is the standard way to avoid label ordering bias in classifiers. With one-hot encoding, the total dimension of input space is $N + C$, with N axes for the swap rates S_n and C additional axes for the currency bits y_c that have two possible values of 0 or 1 as shown in Figure 4(b). The dimensions of CVAE encoder input are shown in Figure 5(a) and the dimensions of the decoder input in Figure 5(b).

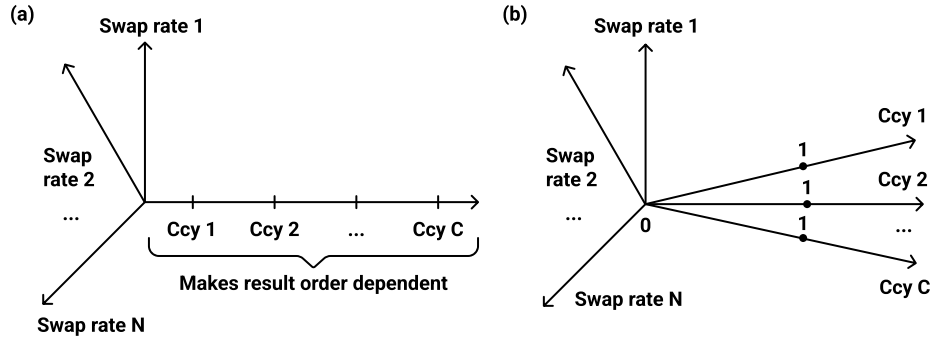

Figure 4: Ordinal (a) vs. one-hot (b) encoding of the currency.

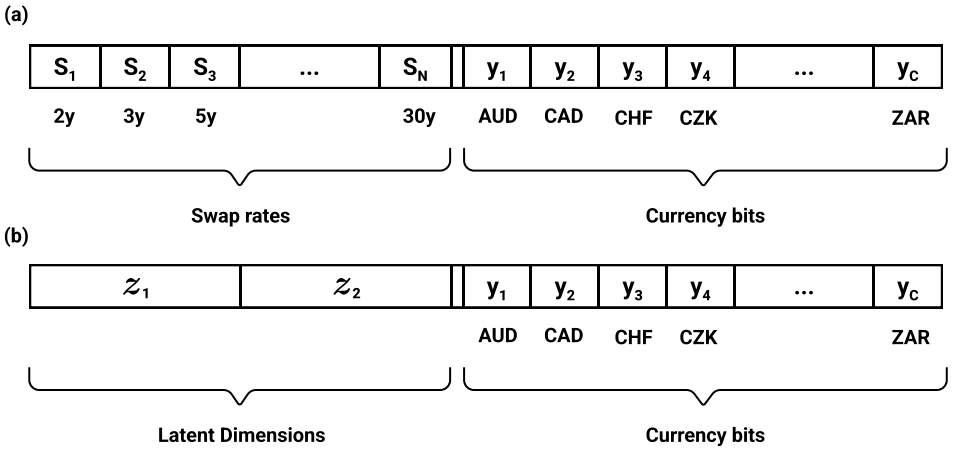

Figure 5: Dimensions of the encoder input (a) and decoder input (b) for CVAE.

CVAE trained to multi-currency data can improve reconstruction accuracy compared to VAE trained to single-currency data by performing interpolation in the C-dimensional subspace of one-hot currency encoding. The extent to which CVAE relies on such interpolation is greater for currencies with limited amount of training data, such as new currencies, currencies where swap rates became liquid only recently, or when prior data cannot be used, e.g., after the removal of an FX rate peg. As the amount of data increases with the passage of time, CVAE output will gradually

become less dependent on interpolation between currencies and in the limit of infinite time series length will closely match the output of VAE trained to single-currency data.

2.3.3. Multi-Currency VAE

CVAE reduces reconstruction loss by making currency-specific adjustments to both the encoder and decoder, in effect creating a separate hyperplane of dimension K for each currency within the $K+C$ dimensional space of K latent dimensions and C one-hot currency dimensions as schematically shown in Figure 6(a). Each of the C

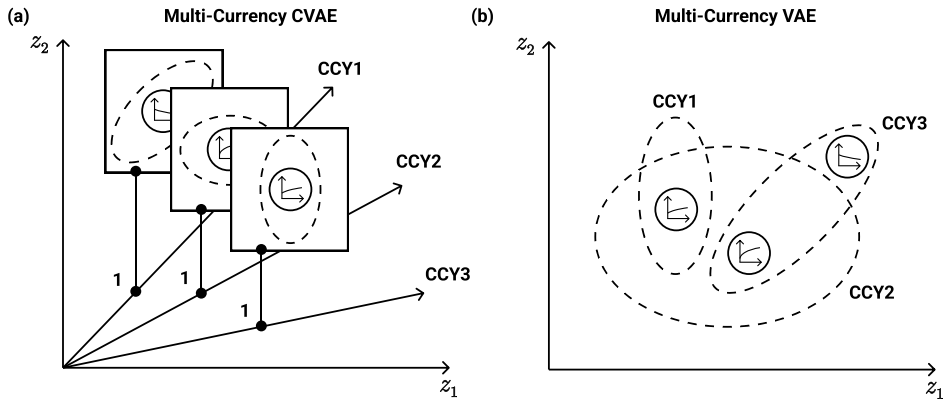

Figure 6: Latent space geometry for (a) multi-currency CVAE and (b) multi-currency VAE. Dashed lines schematically show the historical data envelope for each currency.

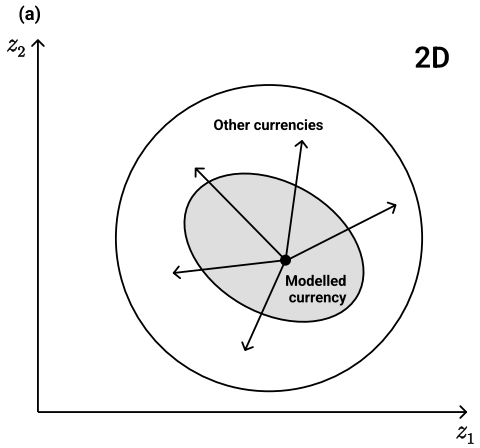

Figure 7: Moving outside the currency's historical data envelope in shared latent space of multi-currency VAE.

currency hyperplanes in CVAE has its own historical data envelope shown by dashed lines in Figure 6(a).

To be used in model construction, any mapping from the yield curve to the model's state variables must be capable of extrapolation outside the historical data envelope for the currency being modelled in order to describe curve shapes that have not occurred for that currency in the past. CVAE performs such extrapolation for each K-dimensional hyperplane separately, guided by data in other hyperplanes.

An alternative and potentially more effective way to perform extrapolation outside the currency's historical data envelope is to encode the swap rates from all currencies into a shared latent space, as shown in Figure 6(b). This can be accomplished by going back to the non-conditional VAE architecture described in Figure 1(b) and training it to multi-currency data. When the stochastic process reaches an area where no data is encoded for the currency being modelled, at first no extrapolation would be required as the mapping will be guided by historical data from adjacent currencies as shown in Figure 7. Extrapolation would only kick in when the stochastic process reaches even more distant areas of latent space where no historical data for any currency is encoded, a rare occurrence in a properly calibrated model.

Layer	Input Dimension	Output Dimension	Activation
Encoder 1	7	4	Tanh
Encoder 2	4	4	None
Sampler	4	2	
Decoder 1	2	4	Tanh
Decoder 2	4	7	Sigmoid

Table 1: Neural network configuration for single-currency VAE.

Layer	Input Dimension	Output Dimension	Activation
Encoder 1	7	7	Tanh
Encoder 2	7	4	None
Sampler	4	2	
Decoder 1	2	4	Tanh
Decoder 2	4	7	Tanh
Decoder 3	7	7	Sigmoid

Table 2: Neural network configuration for multi-currency VAE.

Layer	Input Dimension	Output Dimension	Activation
Encoder 1	7+13	9	Tanh
Encoder 2	9	4	None
Sampler	4	2	
Decoder 1	2+13	7	Tanh
Decoder 2	7	7	Tanh
Decoder 3	7	7	Sigmoid

Table 3: Neural network configuration for multi-currency CVAE. Where two numbers are specified, the second number refers to $C = 13$ additional dimensions used to provide one-hot bits of currency label.

2.4. Results

2.4.1. Data and Configuration

The training dataset consists of daily observations of 2y, 3y, 5y, 10y, 15y, 20y, and 30y LIBOR and OIS swap rates for AUD, CAD, CHF, CZK, DKK, EUR, GBP, JPY, MXN, NOK, SEK, USD, and ZAR. Historical data is visualized in Figure 8. Only those observation dates where each swap maturity was present were included. Because our objective is to build an autoencoder for the curve shapes rather than the distribution of their shocks, the improvement from using more frequent than monthly (e.g., daily) observations would not be significant as prevailing curve shapes are well sampled even at the monthly frequency. As a practical consideration, data vendors provide longer historical time series for monthly observations compared to daily observations. We found that results obtained using daily observations are indeed quite similar given the same time series length.

As Diebold and Li noted in [15], using unequally spaced liquid swap rates to compute reconstruction loss provides a natural way to overweight shorter maturities where the curve has more structure. Keeping this in mind, we assigned equal weight to all swap rates in the L2 reconstruction loss term. We do not estimate the LIBOR-OIS spread, expecting VAE to learn how to represent both types of curves directly from the data.

Neural network configuration for single-currency VAE is described in Table 1, multi-currency VAE in Table 2., and multi-currency CVAE in Table 3. Before passing the input swap rates to the encoder, they are mapped from the interval with the lower bound of $S_n = -5\%$ and higher bound of $S_n = 25\%$ to the $(0, 1)$ interval using linear transformation. The distribution $N(\mu_k, \sigma_k)$ is encoded using mean and logvar. We used the value of $\beta =$ 1e-7 for all three architectures (single- and multi-currency VAE and multi-currency CVAE).

2.4.2. Dimension of Latent Space

With multi-currency VAE architecture, one-dimensional latent space lacks "capacity" to represent the variation of curve shapes between currencies. It would produce an identical sequence of curves for every currency, i.e., the proverbial "model for the average currency" of no practical value. Things are somewhat better for CVAE architecture where latent space is not shared across currencies, and each currency can have its own sequence of curve shapes. However, a model constructed using CVAE with one latent dimension will have many of the shortcomings of classical one-factor models, including 100% correlation between shocks to rates of different maturities.

When shared latent space has more than one dimension, different currencies can be encoded into its different areas, preserving individual characteristics of curves for each currency. The Nelson-Siegel basis accomplishes this task with three latent dimensions. Our analysis shows that VAE can do nearly as well with only two latent dimensions. We will focus on multi-currency VAE with two latent dimensions for the remainder of this paper. Results for other configurations will be described in [11] and elsewhere.

One may argue that a model based on VAE with as few as two latent dimensions will miss certain risks. We believe that any such concerns should be considered

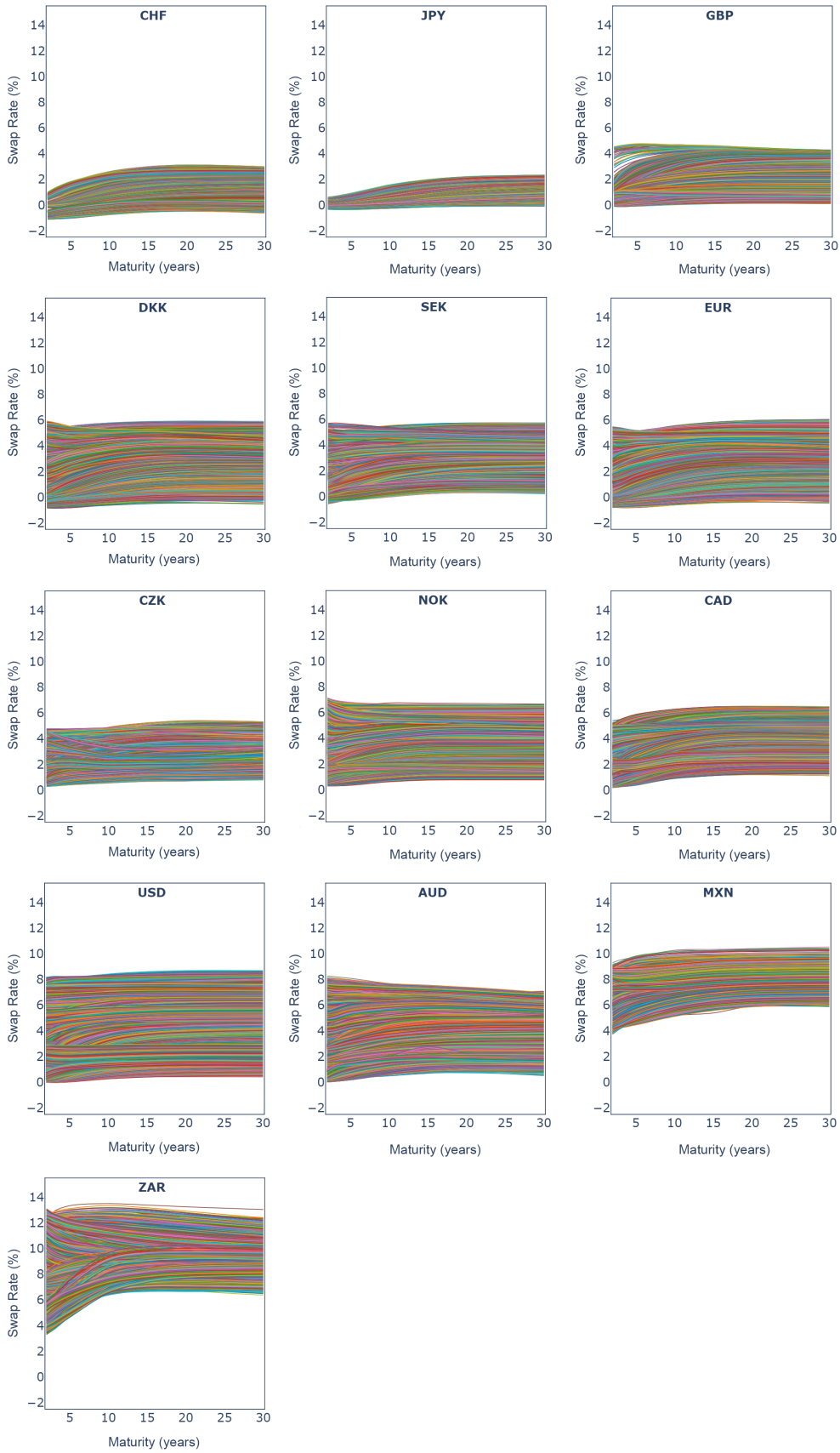

Figure 8: Historical LIBOR and OIS swap rates for all currencies.

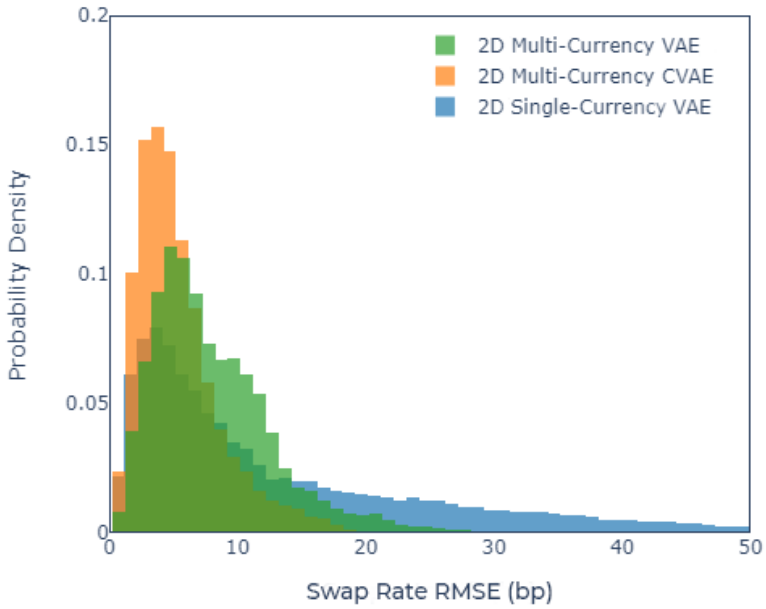

Figure 9: Distribution of in-sample root-mean-square error (RMSE) of swap rate reconstruction for single-currency VAE, multi-currency CVAE, and multi-currency VAE with two latent dimensions across all currencies, maturities, and observation dates. The vertical axis is probability density in arbitrary units and the horizontal axis is swap rate RMSE.

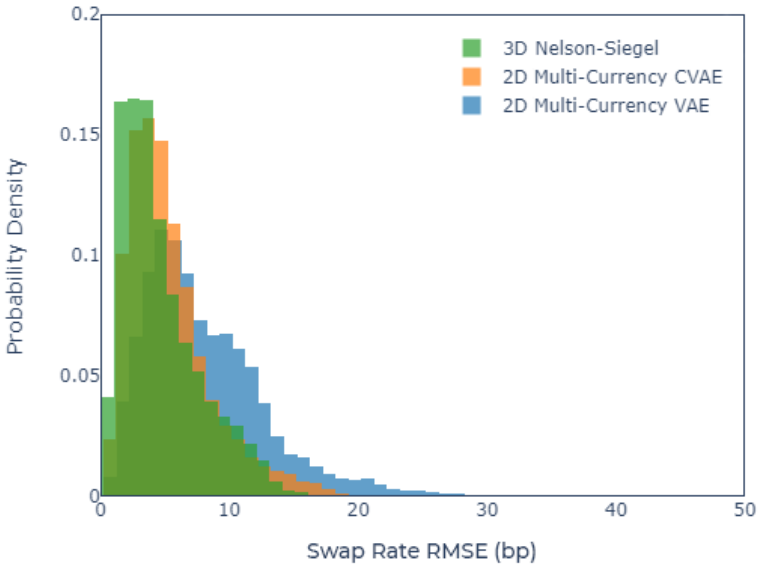

Figure 10: Distribution of in-sample root-mean-square error (RMSE) of swap rate reconstruction for single-currency VAE and multi-currency CVAE, with two latent dimensions vs. the classical Nelson-Siegel basis with three latent dimensions across all currencies, maturities, and observation dates. The vertical axis is probability density in arbitrary units and the horizontal axis is swap rate RMSE.

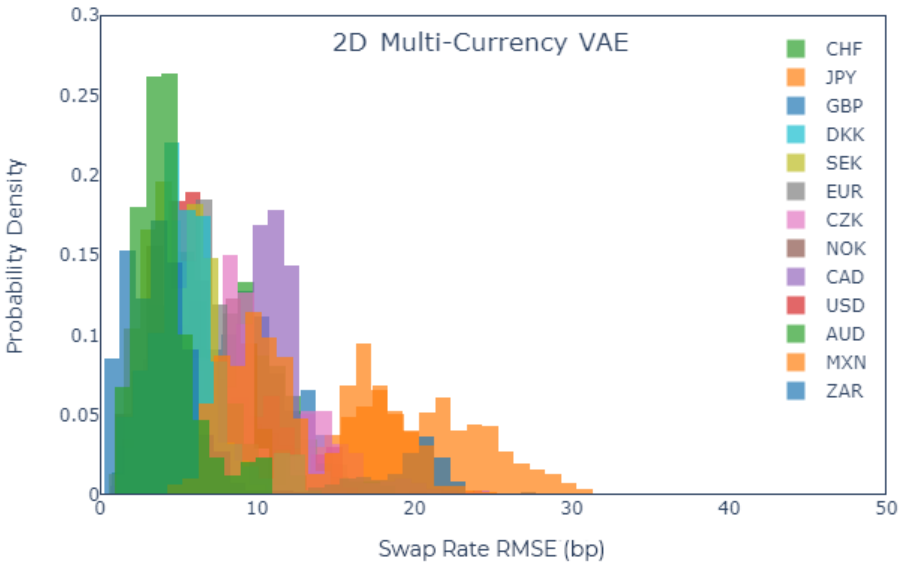

Figure 11: Distribution of in-sample root-mean-square error (RMSE) of swap rate reconstruction by currency across all maturities and observation dates. The vertical axis is probability density in arbitrary units and the horizontal axis is swap rate RMSE.

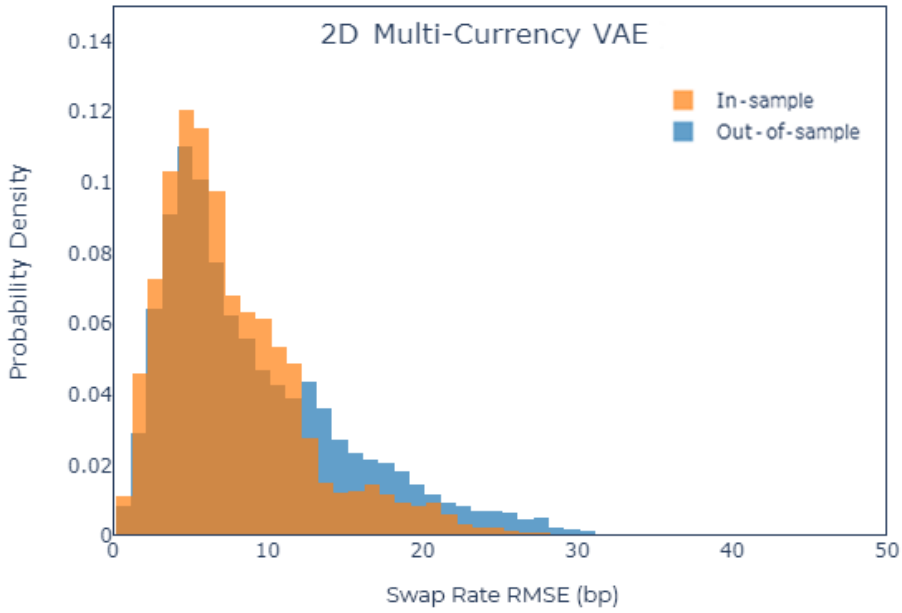

Figure 12: Distribution of in-sample vs. out-of-sample root-mean-square error (RMSE) of swap rate reconstruction across all currencies, maturities, and observation dates. The vertical axis is probability density in arbitrary units and the horizontal axis is swap rate RMSE. There is no indication of overfitting.

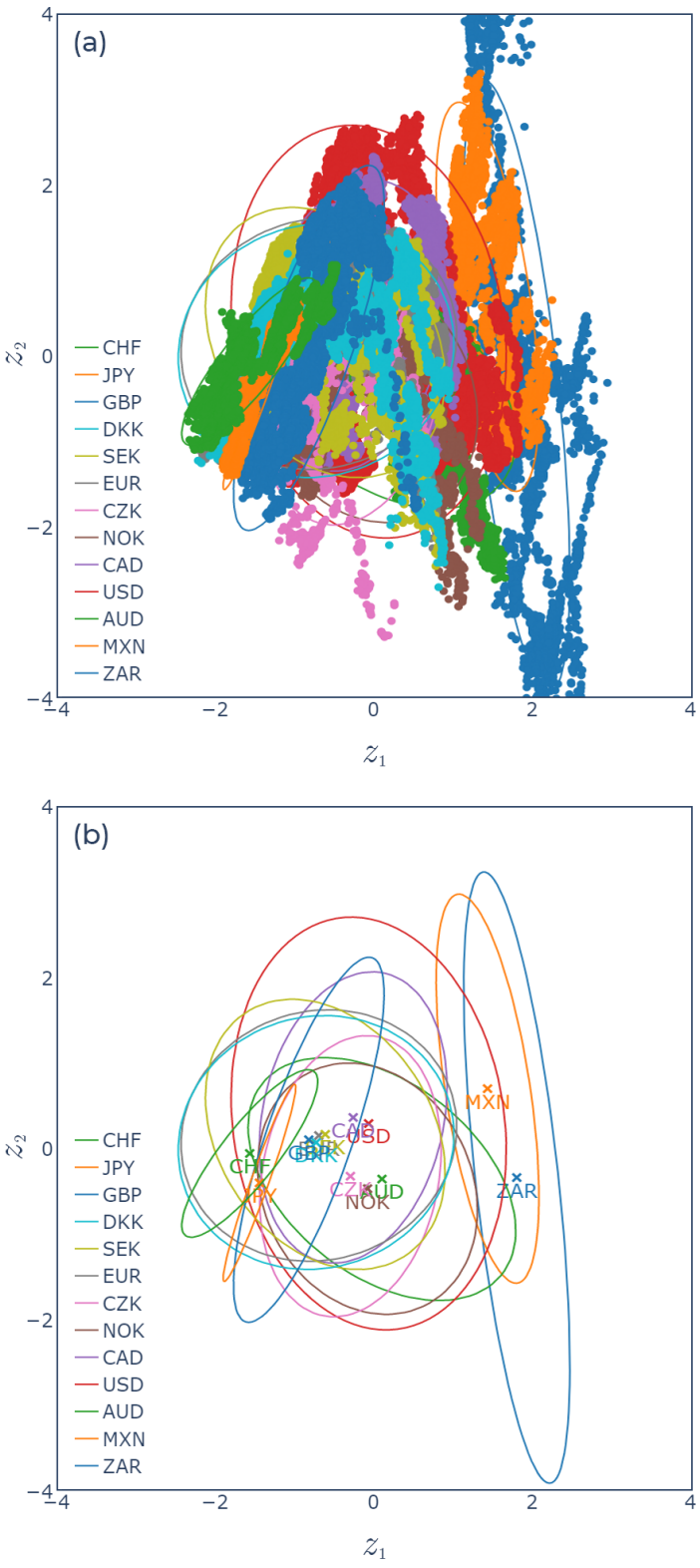

Figure 13: World map of latent space obtained using VAE trained to multi-currency data. Panel (a) shows latent space mapping of daily historical swap rate observations for all currencies. Each ellipse in panel (b) encloses two standard deviations of data along each principal axis for one currency.

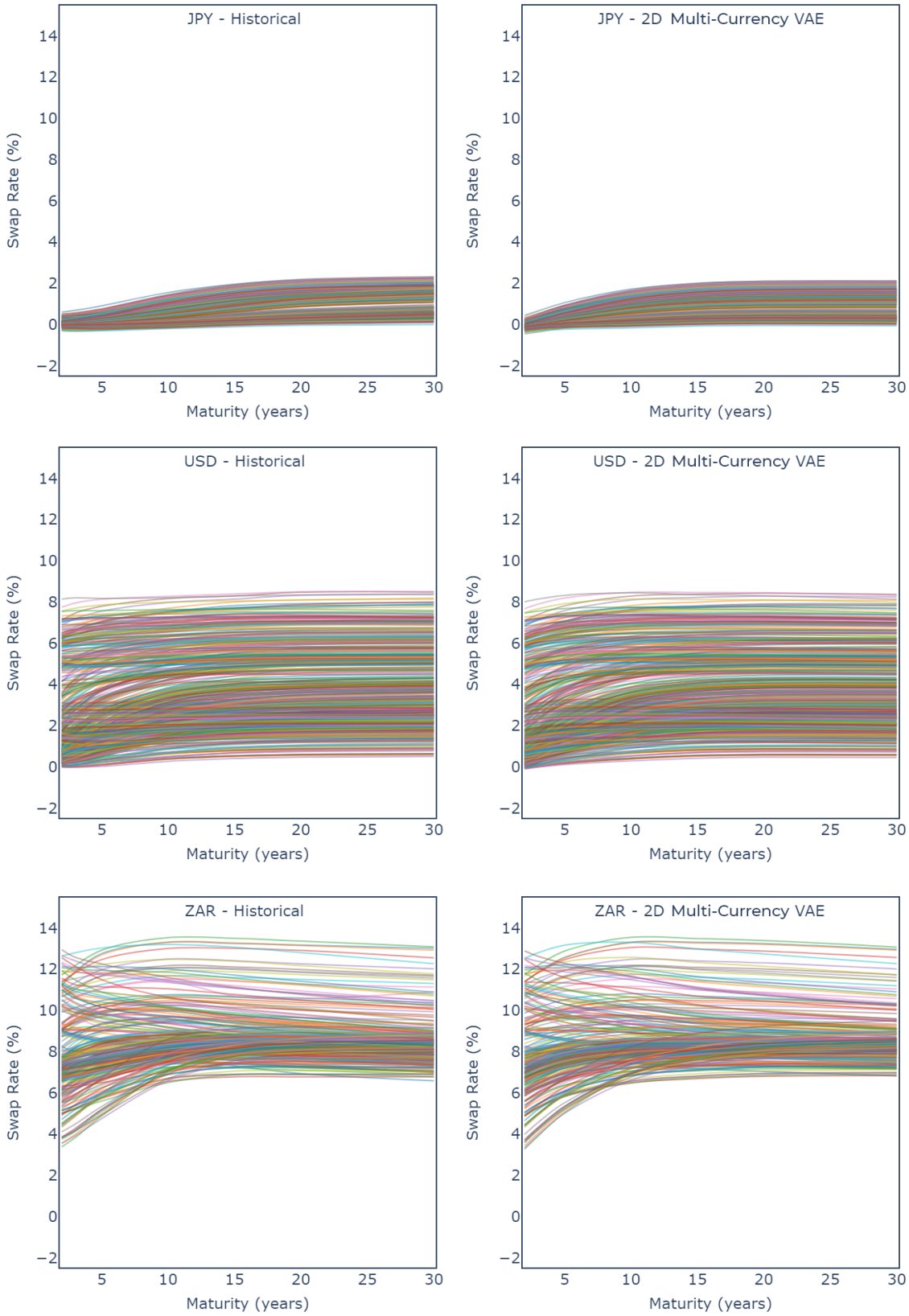

Figure 14: Historical (left) vs. reconstructed (right) swap rates for JPY, USD, and ZAR (from top to bottom) using VAE trained to multi-currency data.

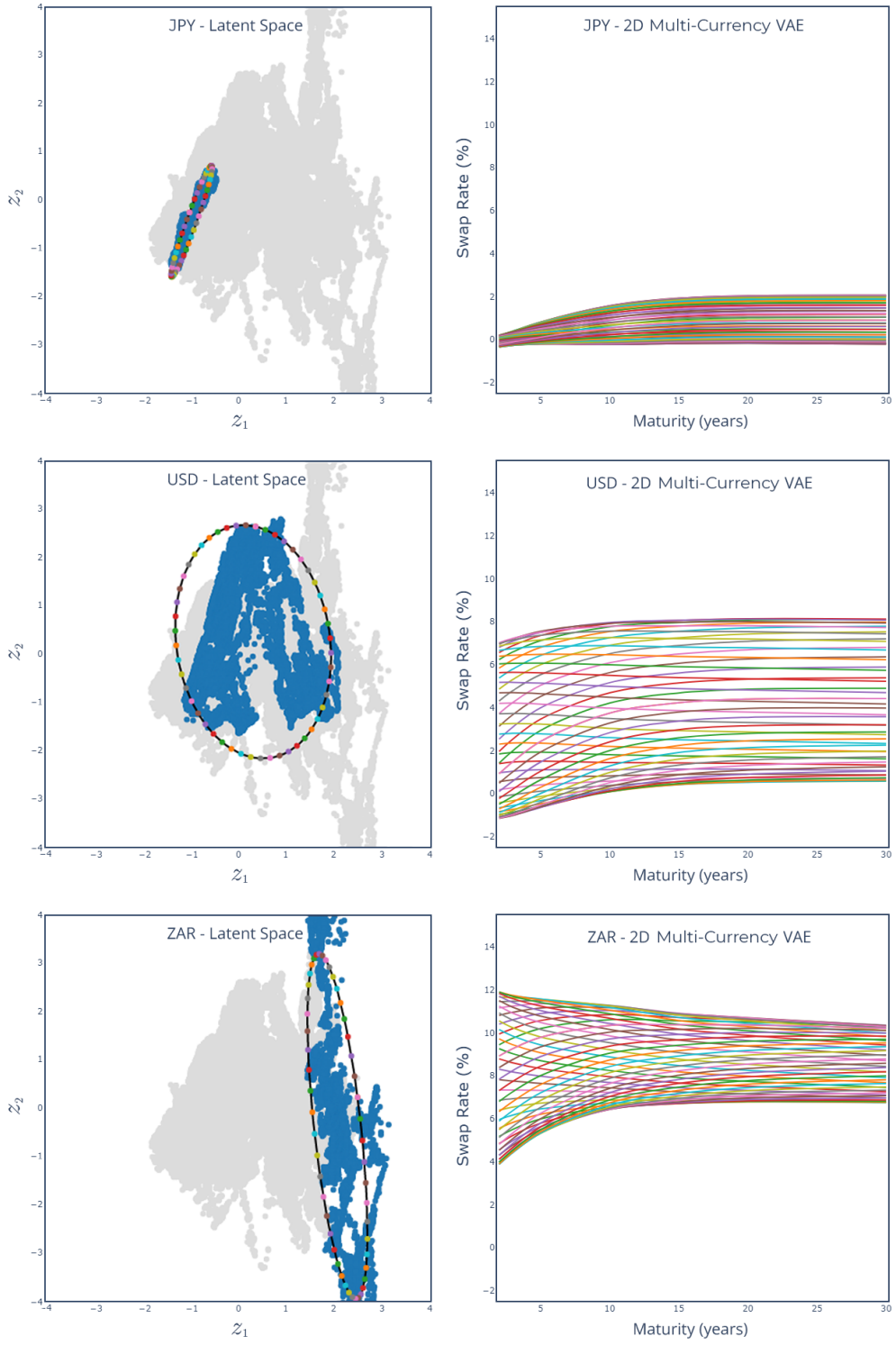

Figure 15: The ellipse in latent space that encloses two standard deviations of data (left) along each principal semi-axis vs. curves obtained by moving around its perimeter (right) for JPY, USD, and ZAR (from top to bottom). Each marker on the ellipse perimeter in the left panel corresponds to a single curve of matching color in the right panel. Blue markers show the latent space mapping of daily historical swap rate observations for the same currency and gray markers for all other currencies.

in light of the purpose for which the model is used, and balanced against greater challenges of parameter estimation for models with more factors. Having a highly parsimonious, densely packed latent space means that residual risks that a VAE-based model does not capture will be smaller compared to a classical model with the same number of state variables.

If having two state variables is still deemed insufficient even when greater parsimony of VAE-based models is taken into account, our approach can be used for higher dimensional VAE with minimal changes. Based on the evidence presented here, we will conjecture that VAE-based models will continue to have more effective compression than classical models even as the number of state variables increases.

2.4.3. Comparison to Nelson-Siegel

We first compare the overall accuracy of single-currency VAE, multi-currency VAE, and multi-currency CVAE. The distribution of in-sample root-mean-square error (RMSE) of swap rate reconstruction for each method across all currencies, maturities, and observation dates is shown in Figure 9. We find that multi-currency methods (VAE and CVAE) are significantly more precise than single-currency VAE and will not consider single-currency training any further.

Of the two multi-currency methods, CVAE is somewhat better because it can make use of the currency label, but the difference between multi-currency CVAE and multi-currency VAE is not dramatic, indicating that even two-dimensional latent space has sufficient capacity to encode swap rates across all currencies without making use of the currency label to aid the encoding. This finding is quite remarkable if we consider how large the differences between prevailing curve shapes are from one currency to the next.

We are now ready to compare the best machine learning methods (multi-currency VAE and CVAE) to the classical Nelson-Siegel basis (Figure 10). This is not an apples-to-apples comparison because the Nelson-Siegel basis has three latent dimensions while both machine learning methods only have two. We found that multi-currency CVAE has similar accuracy to the Nelson-Siegel basis despite having one fewer dimension, and multi-currency VAE has somewhat lower but still comparable accuracy. This result is in line with our expectation that training to the historical data will produce more effective compression than an exogenously specified parametric form. The performance demonstrated by VAE here seems even more impressive if we recall that VAE has no information whatsoever about what swap rate maturities are or even the maturity order, while such added information is central to the Nelson-Siegel representation.

2.4.4. Method Selection

If swap curve reconstruction accuracy were our only consideration, we would choose multi-currency CVAE over multi-currency VAE. However, multi-currency VAE has shared latent space for all currencies while multi-currency CVAE does not. Our analysis shows that the benefits of shared latent space to model construction outweigh the moderate loss of accuracy in multi-currency VAE compared to multi-currency CVAE. Based on this reasoning, we believe multi-currency VAE is the optimal encoding method for AEMM construction. All subsequent examples in this paper will

use multi-currency VAE with two latent dimensions unless otherwise noted.

Figure 11 shows the distribution of in-sample of swap rate reconstruction RMSE by currency across all maturities and observation dates. We found that multi-currency VAE performs well for all currencies in the dataset, with moderate variation in accuracy between currencies. The error is somewhat larger in currencies with shorter time series, but even in those currencies it rarely exceeds 20bp. Figure 12 shows in-sample vs. out-of-sample distribution of root-mean-square error (RMSE) of swap rate reconstruction across all currencies, maturities and observation dates. In-sample results were obtained by using the entire dataset for both training and measurement. Out-of-sample results were obtained by using the data for 2011 and prior years for training, and subsequent years for measurement without overlap. The decrease in accuracy for out-of-sample results is surprisingly minor despite a much shorter training period. There is no indication of overfitting.

2.4.5. Generated Curves and World Map

Figure 14 is the comparison of historical swap rates (left) vs. reconstructed, i.e., encoded and then decoded, swap rates (right) for three representative currencies JPY, USD, and ZAR. Interpolation between swap maturities is provided for visual guidance only. This figure makes it possible to visually assess the agreement between the original and reconstructed curves achieved using only two latent dimensions, with latent space shared by all currencies.

Figure 13 shows the "world map of latent space" obtained using multi-currency VAE. Panel (a) shows latent space mapping of daily historical swap rate observations for all currencies. Each ellipse in panel (b) encloses two standard deviations of data along each principal semi-axis for one currency. The coordinates in latent space were rotated such that the horizontal axis predominantly encodes interest rate levels and the vertical axis curve shapes. Around the middle of panel (b), the lower cluster of three currencies has higher prevalence of inverted curves compared to the rest of the currencies above.

Figure 15 shows the result of decoding latent space points around the ellipse perimeter for the same three representative currencies: JPY, USD, and ZAR.

3. Models in Q-Measure

When interest rate term structure models were first introduced, their primary purpose was pricing derivatives for the front office where performance benefits of an analytical solution were key to model adoption. To admit an analytical solution, the model must assume highly stylized behaviors such as a specific form of volatility, constant speed of mean reversion, and others. Practitioners are well aware that financial markets follow these behaviors at best approximately, but willing to tolerate model error for the performance gain of an analytical solution.

When a model is used for pricing a single derivative instrument with calibration to its natural hedges, the model's role is akin to interpolation, and the impact of approximations made in pursuit of an analytical solution is reduced. A well-known example where interpolation works perfectly is pricing a multi-callable instrument when the model is calibrated to European options with a similar underlying. There

are however many calculations, including XVA among others, where the instrument or portfolio to be priced has little in common with the calibration instruments. For these calculations, the dictum "imply from market prices what you can (really) hedge, and estimate econometrically what you cannot" by Rebonato *et al.* [16] creates motivation for replacing stylized behaviors that admit an analytical solution by those derived from the historical data. Using VAE-based state variables provides a way to do so.

3.1. Forward Rate Models

3.1.1. HJM and LMM Models

We consider a general form of forward rate model in Q-measure that describes the joint evolution of a contiguous sequence of simple forward rates as correlated Brownian motion:

$$dF_n = \mu_n(t, \boldsymbol{F})dt + \sigma_n(t, \boldsymbol{F})dw_n \tag{6}$$

where $\boldsymbol{F} = (F_n)$ is the vector of forward rates for borrowing over (T_n, T_{n+1}) and the stochastic drivers are multivariate standard normal with correlation $\rho_{nn'} = \langle dw_n dw_{n'} \rangle$. The rolling numeraire is obtained by compounding F_n over each period:

$$B_n(T_{n+1}) = \prod_{n'=0}^{n} (1 + \delta_{n'} F_{n'}) \tag{7}$$

where δ_n is the daycount fraction for period n.

Well-known special cases of this model include Gaussian HJM with infinitesimal time intervals and normal volatility [17]:

$$\sigma_n(t, \boldsymbol{F}) = \sigma_n(t) \tag{8}$$

the LIBOR Market Model (LMM) with lognormal volatility [18, 19, 20]:

$$\sigma_n(t, \boldsymbol{F}) = F_n \sigma_n(t), \tag{9}$$

the Andersen-Andreasen model with CEV volatility [21] and the SABR-LMM model with stochastic volatility [4].

To perform discounting to time t of a contingent cashflow paid at time T, its amount is multiplied by the model's stochastic discount factor $SDF(t, T)$, defined as the ratio of model's numeraire B at times t and T:

$$SDF(t, T) = \frac{B(t)}{B(T)} \tag{10}$$

A Q-measure model with K-dimensional vector of state variables $\boldsymbol{X} = (X_k)$ where $k = 1 \dots K$ must satisfy the following no-arbitrage condition for the price of any contingent claim V that pays no cashflows between times t and T:

$$V(t) = E_Q \left[V(T) SDF(t, T) \middle| \boldsymbol{X}(t) \right] \tag{11}$$

where $V(t)$ and $V(T)$ are the contingent claim prices at time t and T respectively, and $E_Q \left[\cdot | \boldsymbol{X}(t) \right]$ is the expectation calculated over the model's Q-measure probability density conditional on the state variable vector $\boldsymbol{X}(t)$.

A model that specifies a perfectly reasonable probability distribution of $\boldsymbol{X}(t)$ but not one specifically selected to satisfy the constraints (11) will exhibit arbitrage. These constraints are universal in nature, and we will not be changing them when converting classical models to AEMM. Our goal will be to convert AEMM into a form where these constraints can be applied the same way as they are applied for the corresponding classical model.

In forward rate models described by (6), the no-arbitrage constraints (11) can only be satisfied if the drift $\mu_n(t, \boldsymbol{F})$ is completely determined by the volatility $\sigma_n(t, \boldsymbol{F})$:

$$\mu_n(t, \boldsymbol{F}) = \sigma_n(t, \boldsymbol{F}) \sum_{n'} \frac{\delta_{n'} \rho_{nn'} \sigma_{n'}(t, \boldsymbol{F})}{1 + \delta_{n'} F_{n'}} \tag{12}$$

This makes correlation $\rho_{nn'}$ the sole mechanism of controlling the movement of forward rates F_n relative to each other.

Forward rate models described by (6) use correlation $\rho_{nn'}$ to produce realistic curve shapes by making rates for nearby time intervals more correlated, and rates for distant time intervals less correlated. This restricts the movement of nearby rates relative to each other in order to maintain the continuity of the curve. As a result, the probability of producing unrealistic curve shapes such as those with multiple local extrema is suppressed, but not completely eliminated. The drawback of this indirect approach is that the control over curve shapes it offers becomes increasingly tenuous over long time horizons, with a higher probability of unrealistic curve samples being generated.

Things improve somewhat if the rank of the correlation matrix is reduced by using a volatility basis (also called volatility kernel in some publications) with a small number of stochastic drivers $K = 2{-}4$ instead of N drivers in the original model:

$$\sigma_n(t, \boldsymbol{F}) = \sum_k \sigma_k(t, T_n, \boldsymbol{F}) dw_k \tag{13}$$

where $\sigma_k(t, T_n, \boldsymbol{F})$ is the shock to $F_n(t)$ at time t from stochastic driver dw_k. However, there is still no guarantee that an exogenously specified parametric basis or one obtained by performing PCA on curve shocks will produce accurate curve shapes over long time horizons and across all interest rate levels. This is what we will aim to improve by converting classical forward rate models to AEMM.

3.1.2. AFNS and FHJM Models

Our review of classical forward rate models would not be complete without describing the family of models that fit the curve using a classical latent factor basis such as Nelson-Siegel and its extensions, and derive forward rate dynamics consistent with that basis up to corrections required to make the model arbitrage-free. This model family includes the Arbitrage-Free Nelson-Siegel (AFNS) model by Christensen, Diebold, and Rudebusch [22, 23] and the Factor HJM (FHJM) model by Lyashenko and Goncharov [24].

Both model families, the former represented by HJM/LMM and the latter by AFNS/FHJM, are specified via an SDE for the forward rates (6). They differ in how they derive the volatility basis. HJM/LMM deal exclusively with curve shocks relative to the initial forward curve. At no point in model construction do they attempt to fit the curve itself.

AFNS/FHJM on the other hand begin by fitting the curve to the Nelson-Siegel parametric form or its extensions, and derive the volatility basis from transformations of that fit. Despite this apparent difference, the two model families are closely related. A general framework developed in [24] clarifies the connection between them and can be used to specify several important types of classical forward rate models, including Cheyette, as well as the original AFNS as one of its special cases.

At present, there is a lack of consistency between state variables of the popular P-measure models, many of which use the Nelson-Siegel basis and its extensions, and state variables of the popular Q-measure models that use a variety of other specifications. Multiple authors including [24, 25, 26, 27] describe the ability to share state variables between Q- and P-measure models as highly beneficial. The ability to share Nelson-Siegel state variables between P- and Q-measure has been cited as motivation in the development of both AFNS and FHJM. We will retain the ability to construct Q- and P-measure models with shared state variables after replacing the Nelson-Siegel latent factors with VAE latent variables in AEMM.

3.1.3. Forward Rate AEMM

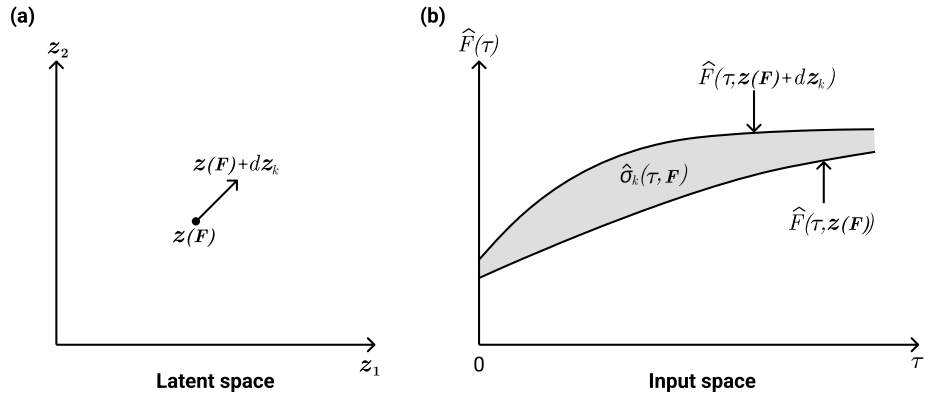

Figure 16: Volatility basis $\hat{\sigma}_k(\tau, \boldsymbol{F})$ in forward rate AEMM is partial derivative of $\hat{F}(\tau, \boldsymbol{z})$ with respect to latent space shift $d\boldsymbol{z}_k$.

In order to make model-generated curve shapes match historical curve shapes better, we propose to derive a dynamic volatility basis $\hat{\sigma}_k(\tau, \boldsymbol{F})$ from the VAE-generated family of curve shapes $\hat{F}(\tau, \boldsymbol{z})$ as shown in Figure 16:

$$\hat{\sigma}_k(\tau, \boldsymbol{F}) = \left. \frac{\partial \hat{F}(\tau, \boldsymbol{z})}{\partial \boldsymbol{z}_k} \right|_{\boldsymbol{z}(\boldsymbol{F})} \tag{14}$$

where $\boldsymbol{z}(\boldsymbol{F})$ is the point in latent space obtained by encoding $\boldsymbol{F}$. By "dynamic" we mean that the basis depends on the forward curve $\boldsymbol{F}$ before the shock, unlike the static basis obtained by PCA or other means. This approach bears certain parallels to AFNS and FHJM, where the volatility basis is also derived from curve fit transformations, except in our case the fit is provided by VAE rather than the Nelson-Siegel basis or its extensions. However, important differences will soon emerge in how the model handles convexity-induced deviations from its curve fit.

The expression (14) specifies the basis up to the maximum swap rate maturity, in our case $\tau = 30y$. Common sense and a rigorous argument in [28] require that any extrapolation we use beyond that horizon eventually converges to zero, so that the impact of a random shock at any finite time t is no longer felt at infinite maturity:

$$\hat{\sigma}_k(\tau, \boldsymbol{F}) \to 0 \text{ for } \tau \to \infty. \tag{15}$$

The proposed basis (14) is compatible with both time-*of*-maturity:

$$\sigma_k(t, T_n, \boldsymbol{F}) = \sigma_k(t)\hat{\sigma}_k(T_n, \boldsymbol{F}) \tag{16}$$

and more popular time-*to*-maturity factorization of volatility:

$$\sigma_k(t, T_n, \boldsymbol{F}) = \sigma_k(t)\hat{\sigma}_k(T_n - t, \boldsymbol{F}) \tag{17}$$

In addition to deriving the maturity profile of volatility basis, expression (14) also produces an estimate for the historical dependence of realized volatility on the interest rate level. Whether or not to take advantage of this is up to the model user. To use an exogenously specified volatility skew, $\hat{\sigma}_k(\tau, \boldsymbol{F})$ is rescaled to obtain the desired parametric form of volatility without affecting the accuracy of matching VAE-generated curve shapes. Skipping this step will cause the model to use historical skew.

Even though AEMM makes shocks to the curve equal to the difference between two adjacent curves in VAE-generated family, the drift (12) would cause deviation of curves within the model from VAE-generated curves to gradually accumulate unless specific steps are taken to correct for this effect. Filipovic [29] demonstrated that such deviation cannot be eliminated completely. The magnitude of this deviation is $O(\sigma^2)$ for time-of-maturity factorization of volatility (16), and $O(1)$ for the more popular time-to-maturity factorization of volatility (17) unless the basis is invariant to time translation, in which case it is also $O(\sigma^2)$. Serendipitously, the Nelson-Siegel basis is already time translation-invariant, but not the Nelson-Siegel-Svensson basis to which Christensen, Diebold, and Rudebusch [22] proposed to add one more term to restore the invariance. Because doing the same is not possible for VAE, we will proceed in a somewhat different way from here.

Imposing even an approximate time translation invariance constraint on VAE to continue with the model construction approach of AFNS/FHJM would require adding more latent variables. This would significantly reduce its parsimony and cause the historical data to fill only a small fraction of the expanded latent space. Instead, we avoid the accumulation of deviation between the curves generated by the model and those generated by VAE by continuously re-encoding the curve after each timestep, a calculation that has similar speed to an analytical formula. Each jump in latent space due to a random shock is followed by a small additional jump to the closest VAE-generated curve after applying the time shift and the drift (12) as shown in Figure 17. This procedure guarantees that at each timestep curve shocks are based on transformations of the nearest VAE-generated curve, as intended.

3.2. Multi-Factor Short Rate Models

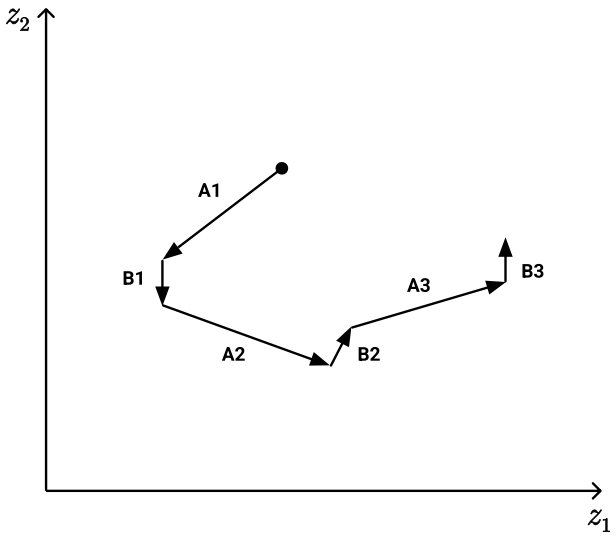

Figure 17: Stochastic evolution of forward rate AEMM in latent space where shocks to the curve from stochastic drivers dw_k denoted by A are alternating with adjustments in latent space due to the drift denoted by B.

3.2.1. Multi-Factor Hull-White Model

The two-factor Hull-White (HW2F) model [6] was originally specified as a normal mean reverting process for the short rate $r(t)$ whose target of mean reversion $u(t)$ in turn follows a normal mean reverting process:

$$\begin{aligned} dr &= -a_r(\theta(t) + u - r)dt + \sigma_r(t)w_r \\ du &= -a_u u dt + \sigma_u(t)dw_u \end{aligned} \tag{18}$$

where $a_{r,u}$ are the rates of mean reversion and $\sigma_{r,u}$ are the normal volatilities of $r(t)$ and $u(t)$ respectively, and stochastic drivers $dw_{r,u}$ are multivariate normal with correlation $\rho_{ru} = \langle dw_r dw_u \rangle$. The dependence of $\theta(t)$ on time is set such that the model matches the initial yield curve for every maturity, making it arbitrage-free. We define the short rate $r(t)$ as the interest rate for borrowing over an infinitesimal time period between t and $t + dt$. While an alternative definition of $r(t)$ using a small but finite investment period is helpful for certain numerical calculations, we will not use it here. The HW2F model is Markovian, a property we will preserve when converting it to short rate AEMM.

The HW2F model can be specified via an alternative, symmetric set of equations described by Brigo and Mercurio [7] and known as G2++. The symmetric specification is especially convenient for adding more than two factors to obtain G(K)++ model, where K is the number of factors. The G(K)++ model represents the short rate as the sum of K correlated mean reverting stochastic variables x_k and a deterministic time-dependent shift $\phi(t)$:

$$\begin{aligned} r(t) &= \phi(t) + \sum_k x_k(t) \\ dx_k &= -a_k x_k dt + \sigma_k(t)dw_k \end{aligned} \tag{19}$$

where a_k are the rates of mean reversion, σ_k are the normal volatilities, and stochastic drivers dw_k are multivariate normal with correlation $\rho_{kk'} = \langle dw_k dw_{k'} \rangle$. The G(K)++ model is made arbitrage-free by calculating the time dependence of deterministic shift $\phi(t)$ that makes the model match the initial yield curve for every maturity. When volatility $\sigma_k(t)$ is normal, the state variables $\boldsymbol{x} = (x_k)$ can be set to zero at $t = 0$ without loss of generality.

The numeraire asset of this model is the money market account accruing continuously compounded interest at the short rate $r(t)$. The stochastic discount factor (SDF) is given by:

$$SDF(t,T) = \exp\left(-\int_t^T r(t')dt'\right) \tag{20}$$

The discount factor $DF(t,T,\boldsymbol{x}(t))$ for maturity T seen at time t is a deterministic function of the models' state variable vector $\boldsymbol{x}(t)$:

$$DF(t,T,\boldsymbol{x}(t)) = E_Q\left[\exp\left(-\int_t^T r(t')dt'\right)\middle|\boldsymbol{x}(t)\right] \tag{21}$$

Let $f(t,T,\boldsymbol{x}(t))$ be the instantaneous forward rate seen at time t for investing over the infinitesimal time interval $(T, T+dT)$. Its limit for $T \to t$ is the short rate:

$$f(t,t,\boldsymbol{x}(t)) = r(t,\boldsymbol{x}(t)) \tag{22}$$

The relationship between the discount factor $DF(t,T,\boldsymbol{x}(t))$ and the instantaneous forward rate $f(t,T,\boldsymbol{x}(t))$, each a deterministic function of of the models' state variable vector $\boldsymbol{x}(t)$, is given by:

$$DF(t,T,\boldsymbol{x}(t)) = \exp\left(-\int_t^T f(t,T',\boldsymbol{x}(t))dT'\right) \tag{23}$$

The instantaneous forward rate $f(t,T)$ in G(K)++ has the following form:

$$f(t,T) = f(0,T) + \sum_k x_k(t)e^{-a_k(T-t)} + f_{\text{conv}}(t,T,\boldsymbol{x}(t)) \tag{24}$$

where $f_{\text{conv}}(t,T,\boldsymbol{x}(t)) \sim O(\sigma^2)$ is convexity adjustment. In what follows we will occasionally omit the last argument $\boldsymbol{x}(t)$ to simplify notation.

The curve shape given by (24) is defined relative to the initial curve shape, and is therefore unable to capture the difference between how the curve moves for low rates vs. high rates. We will be able to reflect this difference in AEMM, at the same time adding currency-specific aspects of curve dynamics.

3.2.2. Multi-Factor Short Rate AEMM

Short rate models generate curve shapes in a different way than forward rate models do. In forward rate models, the change in curve shape as a result of a random shock at time t is determined by the dependence of volatility basis $\sigma_k(t,T,\boldsymbol{F})$ on maturity T. In short rate models, it is determined by the rate at which mean reverting drift between shock time t and maturity time T returns the short rate to its initial "trajectory". The rate of exponential decay in volatility basis of forward rate models

plays the same role as the rate of mean reversion in short rate models. It follows that what we previously accomplished for forward rate models by changing the volatility basis we can accomplish for short rate models by changing the drift.

The multi-factor short rate AEMM is given by:

$$\begin{aligned} r(t) &= \phi(t) + \sum_k x_k(t) \\ dx_k &= \mu_k(t, x_1 \dots x_k)dt + \sigma_k(t)dw_k \end{aligned} \tag{25}$$

Because drift is now non-linear, initial values of x_k cannot be set to zero and are instead selected to minimize the forward rate residual $\phi(t)$. We will calibrate the new drift term $\mu_k(t, x_1 \dots x_k)$ that replaced constant mean reversion $-a_k x_k$ of the classical model to match the VAE-generated family of curve shapes, with minimal corrections to keep the model arbitrage-free.

Notation $\mu_k(t, x_1 \dots x_k)$ for the drift of state variable x_k in (25) means it depends on x_k and all of the preceding state variables $x_{k'}$ where $k' < k$, but not the subsequent state variables where $k' > k$. This constraint helps achieve a hierarchical model definition where lower factors are responsible for a greater share of overall drift and variance than higher factors. This definition of drift brings short rate AEMM closer to the original specification of the HW2F model where one state variable is reverting to the other.

The calibration target for $\mu_k(t, x_1 \dots x_k)$ is $\hat{f}_k(\tau, z_1 \dots z_k)$ obtained by decomposing VAE-generated instantaneous forward rate into a sum of components, where k-th component depends on latent space dimensions $z_1 \dots z_k$:

$$\hat{f}(\tau, \boldsymbol{z}) = \sum_{k=1}^{K} \hat{f}_k(\tau, z_1 \dots z_k) \tag{26}$$

The target curves for calibrating drift of the first state variable are shown in Figure 18(a) and second state variable in Figure 18(b). The first state variable of AEMM captures a significant share of curve variation. The contribution of each added factor is progressively smaller, providing a convenient hierarchical model specification.

There are two drift factorizations we can consider. The first is time-homogeneous factorization $\mu_k(x_1 \dots x_k)$ that matches the VAE-generated family of curves on average for all time horizons. The second is time-*of*-maturity factorization $\mu_k(t, x_1 \dots x_k)$ that matches the VAE-generated family of curves exactly for a single time horizon which we will once again choose to be $t = 0$ like we did for the forward rate models, with $O(\sigma^2)$ error for other time horizons.

The option of time-to-maturity factorization we previously used for forward rate models is not available here because all short rate models, including the classical ones, use the same time t to advance along the time axis and the maturity axis and cannot shift one relative to the other. Its closest analog for short rate models is the time-homogeneous factorization $\mu_k(x_1 \dots x_k)$. This factorization is equivalent to making the overall mean reversion speed of the short rate dependent on the level of interest rates but not explicitly on time. Arguments supporting the existence of such dependence in P-measure have been presented in [30, 31]. Unless miraculously canceled out by the risk premium, the same dependence should exist in Q-measure as well. Fitting time-homogeneous drift to VAE-generated curve shapes provides a new way to estimate its level in Q-measure.

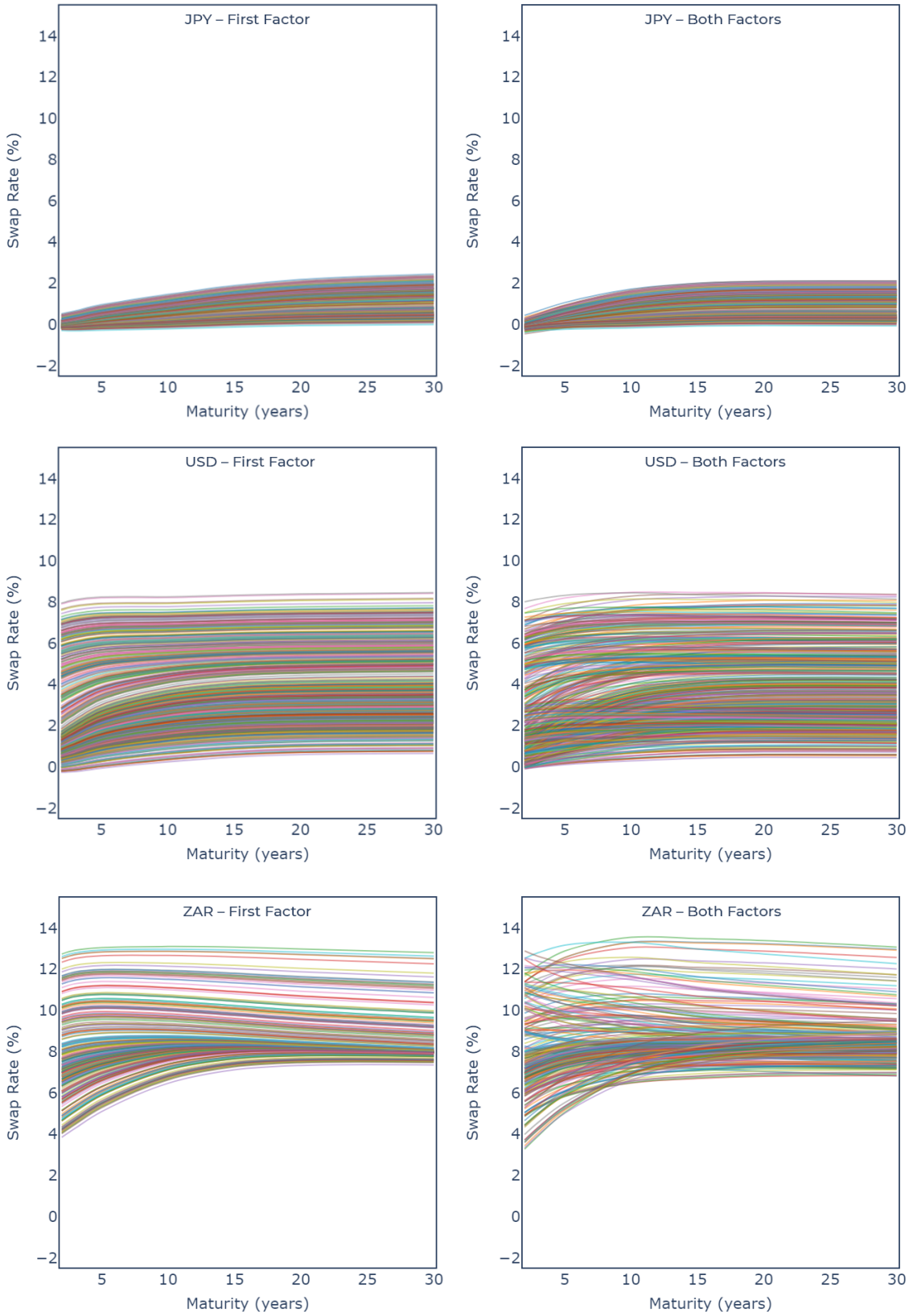

Figure 18: VAE-generated target curves for calibrating drift of (a) the first state variable vs. (b) both state variables for JPY, USD, and ZAR (from top to bottom).

The fitting procedure must ensure that the resulting drift is mean-reverting to a single equilibrium point in state space $\boldsymbol{x}$ in order to comply with the constraint that "long forward rates can never change" [28], which is another way of saying that the impact of a random shock at any finite time t must no longer be felt at infinite maturity $T \to \infty$.

The time-of-maturity factorization requires finding $\mu_k(t, x_1 \dots x_k)$ that reproduces the entire set of VAE-generated forward rates $\hat{f}(\tau, \boldsymbol{z})$ at model origin, or any other single time horizon, exactly for any maturity τ and latent vector $\boldsymbol{z}$. Note that such fit will be exact everywhere in latent space only if the mapping between $\boldsymbol{z}$ and each forward rate is monotonic, which we found to be the case for the multi-currency VAE we developed. A non-monotonic mapping may cause local deviations from the exact fit, but not a global one. The $t \to \infty$ extrapolation of drift must be mean-reverting to a single equilibrium point in state space $\boldsymbol{x}$.

The time-of-maturity factorization has a remarkable property. While it can only be made exact across all maturities T for a single time horizon t which we chose to be $t = 0$, its deviation from the exact fit at other time horizons is caused solely by $O(\sigma^2)$ convexity effects. This means curve shapes produced by the model stay close to VAE-generated curve shapes across all maturities T until relatively long time horizons t. Any attempt to make the agreement exact across all maturities is thwarted by convexity adjustments that keep the model arbitrage-free [29]. However even an approximate agreement with $O(\sigma^2)$ accuracy is still highly beneficial.

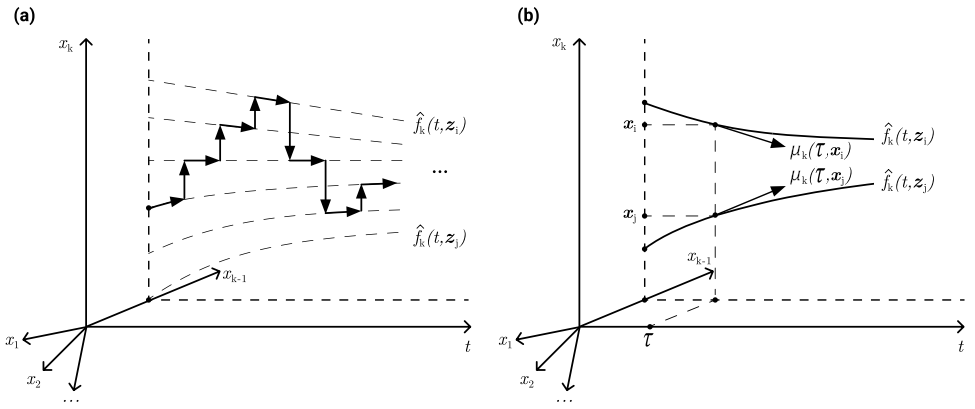

Figure 19: (a) Family of state variable trajectories under time-of-maturity factorization of drift that produces the desired VAE-generated sequence of curve shapes. (b) The leading $O(1)$ term in drift $\mu_k(t, x_1 \dots x_k)$ that produces the desired trajectories. Convexity correction $O(\sigma^2)$ to the drift is included in the fitting procedure but not shown on the drawing.

The choice between time-homogeneous and time-of-maturity factorization of drift, like many other calibration choices, is a matter of preference. One disadvantage of the time-of-maturity factorization is that with the passage of time we will be leaving in the past an increasingly long initial segment of the curve, until only the longest maturities are left. We will refer to this effect as "snipping off" of the head of term structure. This behavior occurs with respect to all time-dependent parameters of short rate models, and should not be considered a surprise. In fact, snipping off the head of the term structure of drift is exactly like snipping off the head of the term

structure of volatility, a universally accepted market practice for this model type. One may even argue that if the head of the initial forward curve must be snipped off to satisfy a martingale property that is not model-specific, the head of any term structure of drift should be snipped off as well. For the reader not convinced by this argument, time-homogeneous calibration, although approximate rather than exact even at $t = 0$, eliminates any such concerns as it has no term structure to snip.

4. Models in P-measure

Interest rate models in P-measure have important practical applications across the financial industry. At short time horizons, they are used for market and liquidity risk. At long time horizons – for economic forecasting, macro investing, insurance reserve requirements, and limit management.

For time horizons measured in days, calculations can be performed by sampling directly from the historical time series of risk factor returns. This model-free approach is the foundation of historical value-at-risk (HVaR) and expected shortfall (ES) methodologies. For longer time horizons, the number of independent returns in the historical time series is insufficient for direct sampling. In this case, a P-measure model is required.

The first P-measure models to be developed used the same type of SDE as Q-measure models, but with time-homogeneous parameters. After enjoying a period of considerable popularity in the 1990s, the SDE-based approach fell out of favor as the practitioners increasingly turned to latent factor models based on the Nelson-Siegel basis and its extensions. This model category will be the focus of our review of P-measure models.

4.1. Autoregressive Models

We will begin with P-measure models for forecasting and risk that are calibrated to the historical time series. The premise of historical calibration is that no observation date is special, so that past observations can be treated as independent samples drawn from the model's P-measure probability distribution. This naturally leads to time-homogeneous model specification.

Of course, time-homogeneity does not mean that the interest rate drift is zero. Because any interest rate is confined to a finite range, its unconditional drift is indeed zero when averaged across an infinitely long observation period. However, its drift conditional on the initial state must be positive for lower-than-average rates and negative for higher-than-average rates to avoid producing "runaway rates".

The standard way to model state-dependent drift in time-homogeneous setting is by estimating parameters of an autoregressive process for the model's state variables. We will refer to models that use this approach as "autoregressive models".

4.1.1. Dynamic Nelson-Siegel Model

Our classical autoregressive P-measure model example is the Dynamic Nelson-Siegel (DNS) model by Diebold and Li [15]. This model is the de-facto standard for yield curve forecasting by central banks. The DNS model represents the yield curve using

the Nelson-Siegel basis (2) or its extension and then estimates the parameters of a separate first-order univariate linear autoregressive AR(1) process [32, 33] for each of its three latent factors $\beta_{1,2,3}(t)$. The DNS model treats the rate of decay λ in the Nelson-Siegel basis as time-independent and specified *a priori*. Diebold and Li set its value to 30 months [15]. As with any other model based on an autoregressive process, the DNS model is time-homogeneous by construction.

Models for the swap curve require the absence of arbitrage at origin ($t = 0$). The DNS model uses a separate stochastic process for the deviations of the curve from the Nelson-Siegel form, setting its initial value such that the curve at origin has no arbitrage. In what follows, we will omit this part of the model for the sake of brevity as AEMM uses an identical approach.

The first-order univariate linear autoregressive AR(1) model represents each of the three Nelson-Siegel latent factors $\beta_k(t + h)$ at the risk horizon $t + h$ as the sum of the deterministic factor forecast and white noise representing factor innovations (i.e., volatility). The forecast for each of the three Nelson-Siegel latent factors $\beta_{1,2,3}$ at time $t+h$ depends only on time-t value of the same latent factor but not the other two:

$$\beta_k(t + h) = (1 - \phi_k)\theta_k + \phi_k\beta_k(t) + \epsilon_k \tag{27}$$

where θ_k is the equilibrium level of the factor (i.e., the target of mean reversion), ϕ_k is the decay multiplier in autoregression, and the white noise ϵ_k for factor β_k, is a serially uncorrelated random variable with time-homogeneous probability density $P[\epsilon_k]$ and the mean of zero:

$$\int_{\epsilon_k} \epsilon P[\epsilon_k] d\epsilon_k = 0 \tag{28}$$

Because ϵ_k has zero mean, P-measure expectation is the deterministic part of (27):

$$E[\beta_k(t + h)|\beta_k(t)] = (1 - \phi_k)\theta_k + \phi_k\beta_k(t) \tag{29}$$

For $h \to 0$, the multiplier $\phi_k \to 1$ and the forecast for $\beta_k(t + h)$ is its preceding value $\beta_k(t)$. For $h \to \infty$, the multiplier $\phi_k \to 0$, the initial state is forgotten, and the forecast is the equilibrium level θ_k. To make the comparison of ϕ_k estimated at different time horizons more intuitive, we can think of the average mean reversion speed a_k over the estimation horizon h defined as:

$$\phi_k = \exp(-a_k h) \tag{30}$$

The DNS model and its extensions are often called "forecasting" models. This name reflects their original use for forecasting yield curve trends, where the objective is to estimate mean future rates (i.e., the forecast), or equivalently the interest rate drift. When forecasting is the objective, probability distribution of ϵ_k is only of interest to the extent it helps estimate forecast accuracy, but not in its own right.

Following their initial development for yield curve forecasting, dynamic latent factor models were applied to the calculation of risk at long time horizons measured in years and decades. This calculation is required for PFE-based limit management and, unlike market risk, cannot be performed by sampling from the time series of returns due to the long risk horizon. When the DNS model is used for calculating PFE, both the forecast and the probability distribution of ϵ_k must be estimated. Unlike for the Ornstein–Uhlenbeck process to which the AR(1) model is frequently compared, the probability distribution of ϵ_k in AR(1) is not necessarily Gaussian. A comprehensive review of DNS estimation methodologies can be found in [34].

4.1.2. Autoregressive AEMM

Constructing autoregressive AEMM is largely similar to constructing DNS, except VAE-based curve representation is used in place of the Nelson-Siegel basis, and VAE latent vector $\boldsymbol{z}$ is used in place of the Nelson-Siegel latent factors $\beta_{1,2,3}$. Minor complications that arise as a result of replacing a linear basis by the non-linear VAE-based representation are easily resolved using readily available multidimensional optimization packages, such as those found in most machine learning libraries.

The encoder component of VAE performs the job of converting the initial curve shape to the initial latent vector. The autoregressive model is then inserted between the encoder and the decoder and calibrated to produce the desired probability distribution in latent space. The decoder converts that distribution to future curve shapes. In a Monte Carlo setting, this is done by simulating paths in latent space first, and then converting each path to Monte Carlo samples of future curves using the decoder. In settings other than Monte Carlo, regression techniques can be used to estimate probability density in latent space and convert it to probability density in input space (i.e., the probability of a given curve shape).

Our results indicate that VAE training to multi-currency data produces highly efficient compression that can represent curve shapes using two latent dimensions with similar accuracy to that of the Nelson-Siegel basis with three latent factors. The resulting reduction in latent space dimension from three to two has the potential to significantly improve the quality of parameter estimation for both forecasting and risk applications. It also reopens the possibility of using a more sophisticated autoregressive model that was ruled out in [15] and subsequent publications due to the high number of model parameters when three or more latent variables are used.

Reduction in the number of latent variables from three to two thanks to the use of VAE in turn reduces the number of parameters, making it possible to consider alternatives to the univariate AR(1) model. Among the classical stochastic model alternatives are the linear vector autoregressive VAR(1) model [35] and Jones model [36, 37]. Each of these models improves one aspect of AR(1).

The AR(1) model used in DNS holds that the forecast for each latent factor depends only on its own prior value, and this dependence is linear. In the case of VAR(1) model, the forecast becomes multivariate but remains linear. In the case of the Jones model, it remains univariate but becomes nonlinear. Other options include Bayesian networks that were previously applied to stochastic factor models in [38] and generative machine learning used for the same purpose in [31].

4.2. Dual-Measure Models

Given the difficulties of estimating historical drift in autoregressive model framework we described in the preceding Section 4.1, harnessing trader's view of the direction where interest rates are headed provides an attractive alternative. Unlike the time-homogeneous forecast of the DNS model, traders may express a highly nuanced view with predictions that depend on the time horizon. For example, they may expect the interest rates to increase for the next two years and then decrease. This view is incorporated into the shape of the yield curve, and can be converted to P-measure forecast if the term structure of the risk premium can also be estimated. Dual-measure models described in this Section provide a framework for doing so.

A dual-measure model consists of two constituent models, one in Q-measure and the other in P-measure, that share the same set of state variables. The two models need not use the same type of SDE, as long as the differences in model construction between Q- and P-measure do not cause their state variables to differ. If both models use the AEMM approach, the state variables are latent variables of VAE on both sides, and this condition is always satisfied.

We will call the two constituent Q- and P-measure models "sides" of the dual measure model. The risk premium incorporated into prices causes additional drift in the Q-measure side relative to P-measure side. During model construction, the Q-measure side is calibrated to market-implied data in the usual way. After that, excess drift due to the risk premium is estimated and subtracted from the drift of the Q-measure side to arrive at the calibration of the P-measure side.

Estimation of P-measure drift is a long-standing problem in interest rates research, which so far eluded a comprehensive and universally accepted solution. The premise of the dual-measure calibration approach is that it is often easier to estimate the risk premium than estimate the drift itself. We will describe the principles of risk premium estimation in the following Subsection 4.2.1. We will then review two classical dual-measure models in Subsection 4.2.2 and introduce dual-measure AEMM in Subsection 4.2.3.

4.2.1. Risk Premium Estimation

It is well known that the yield curve is on average upward sloping. The reason for this effect is the risk premium. Because long-dated bonds have more risk, risk aversion makes their prices lower and their yields higher compared to short-dated bonds. If traders were neutral to risk as a group, the yield curve would be on average flat; with risk-seeking traders, on average downward sloping. With traders risk-averse, as it were, the curve is on average upward sloping.

Because curve slope is determined by the combination of trader's view of the future interest rate trends and their risk aversion rather than risk aversion alone, a downward curve slope (inverted curve) is also possible, but less frequent than the upward slope. For an inverted curve to occur, the downward pull on long-dated yields due to the trader's view that interest rates will fall in the distant future must prevail over the upward pull due to their aversion to riskier long dated instruments. This usually happens when short-term interest rates rise far above their historical average.

Stanton [39] proposed to use the average slope of the yield curve for estimating the risk premium. He observed that the historical average of P-measure forecast for the drift of short-term interest rates must be zero, while the historical average of Q-measure drift is the average forward curve slope. The real world counterpart to the forward rate is the rate at time $t+\tau$ where t is observation time and τ is a fixed offset. Its realized values are the same as realized values of the short rate except for the segments of length τ at each end of the observation interval. It follows that historical average of an unbiased P-measure forecast has the same historical average as spot:

$$\langle E_P[r(t_i+\tau)|X(t_i)]\rangle = \langle r(t_i)\rangle \tag{31}$$

where $r(t)$ is the short rate, $E_P[\cdot|X(t_i)]$ is forecast calculated using the P-measure side of the dual measure model conditional on state variables $X(t_i)$ at time t_i, and the

average is taken over all historical observations t_i in a sufficiently long observation period. The excess drift can be estimated from the difference between historical averages of Q-measure instantaneous forward rate and P-measure forecast of the short rate for the same time offset τ, conditional on the initial state $\boldsymbol{X}$:

$$\Psi(\tau) = \langle f(t_i, t_i + \tau) - E_P[r(t_i + \tau)|\boldsymbol{X}(t_i)]) \rangle \tag{32}$$

where $f(t_i, t_i + \tau)$ is the instantaneous forward rate for investment at time $t_i + \tau$ observed at t_i. While this estimation method becomes exact only in the limit of the infinite observation period length, it can provide a sufficiently accurate estimate for the available historical time series length for most currencies.

4.2.2. HSW and BDL Models

Stanton [39] and Cox and Pedersen [40] used a slightly different methodology to the one described above. In their approach, market price of risk (i.e., the risk premium scaled by volatility) rather than the risk premium was estimated. Subsequently, Ahmad and Wilmott [41] proposed a stochastic model for the market price of risk in the short rate and described its calibration to the historical data.

The average excess drift of the short rate estimated in [39, 40, 41] for short τ measured in months was around 100bp/year. Using the same level of excess drift for rates of all maturities up to 30y rather than the short rate it was derived for, one would arrive at the patently absurd conclusion that the average 30y forward must be 3000bp higher than the average short rate before mean reversion is taken into account. No reasonable amount of mean reversion can reconcile this figure with the historically observed differential between short rates and 30y rates.

Hull, Sokol and White (HSW) [25] proposed to resolve the apparent paradox by assigning a term structure to the market price of risk. In the context of short rate framework used in their publication, they introduced the concept of the "local price of risk" whose value at time t determines the contribution of the yield curve segment from t to $t + dt$ to the market price of risk in bonds with maturity $T > t$. The relationship between the local price of risk at time t in their model and the market price of risk of swaps with maturity T is akin to the relationship between the local volatility in a classical short rate interest rate model and the market implied volatility, in the sense that the latter is a suitably defined average of the former over the time interval between model origin and maturity.

It is well known that a group of 100% correlated assets must have the same market price of risk. Because forward rates of different maturities are not 100% correlated, the market price of risk (i.e., risk premium per unit of volatility) in each rate can be different. Equation (32) provides an estimate of this dependence. The growth of risk premium in the forward rate vs. time-to-maturity τ slows down toward longer maturities. This can be understood by observing that the yield curve is on average less steep for longer maturities compared to shorter maturities.

Barker, Dickinson, and Lipton (BDL) [26] demonstrated that time-inhomogeneity of the local price of risk in [25] is required only for one-factor models, and is not a general limitation of dual-measure models or the short rate framework. They described a 3-factor Hull-White model that can fit the observed term structure of risk premium using time-homogeneous model parameters by distributing the risk premium across multiple model factors, some of which are decaying faster than others.

This approach to risk-premium calibration avoids making the risk premium explicitly depend on model time t while producing accurate estimates of its dependence on time-to-maturity τ.

4.2.3. Dual-Measure AEMM

The P-measure side of the dual-measure model follows the model construction approach we already used for the autoregressive P-measure AEMM in the preceding Section 4.1. To recap, the model uses the encoder to map the initial curve to a vector in latent space, calibrates transition probabilities in latent space to obtain the distribution of future latent space vectors, and uses the decoder to convert these latent vectors to future yield curves.

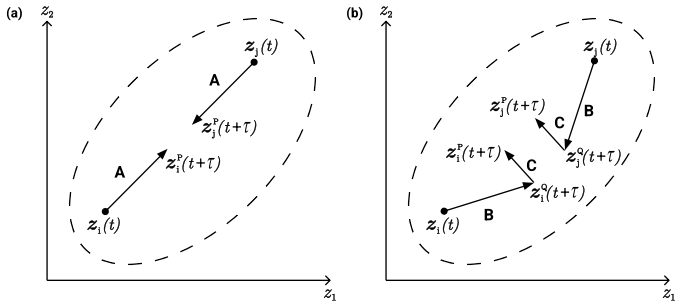

Figure 20: Estimation of P-measure drift in latent space using (a) autoregressive model vs. (b) dual-measure model. Label A shows the estimated P-measure drift, label B estimated Q-measure drift, and label C estimated excess drift due to the risk premium.

Dual-measure AEMM differs from autoregressive P-measure AEMM only in how the estimation of P-measure transition probabilities is performed. In autoregressive AEMM, these probabilities are estimated from the historical data under time-homogeneous assumption. In dual-measure AEMM, their difference from known probabilities on the Q-measure side of dual-measure AEMM is estimated instead. This method of model construction does not require making the risk premium dependent on time as drift adjustment from Q- to P-measure occurs in latent space.

Past studies indicated that in certain markets market-implied volatility may be a better predictor of future volatility than historical volatility. The procedure described above can be used to produce a model with P-measure drift and market-implied volatility. To use the traditional historical calibration of volatility, the above procedure can be used for the estimation of drift and combined with conventional historical estimate of volatility from daily returns.

Figure 21 illustrates the difference between the estimation of drift in autoregressive and dual-measure AEMM. For autoregressive AEMM, P-measure drift (A) conditional on the initial state is estimated directly. For dual-measure AEMM, Q-measure drift (B) is combined with the estimate of excess drift due to the risk premium (C) to obtain P-measure drift.

Procedures for calibrating the term structure of excess drift were described in prior literature [25, 30, 26]. With some important variations, all of them involve

estimating either the average difference $\Psi(\tau)$ between the forward rate with time-to-maturity τ and P-measure forecast for τ-lagged short rate, or the market price of risk driving that difference. By mapping all historical curve shapes to a shared two-dimensional latent space, AEMM creates the opportunity for a new way to perform this estimation.

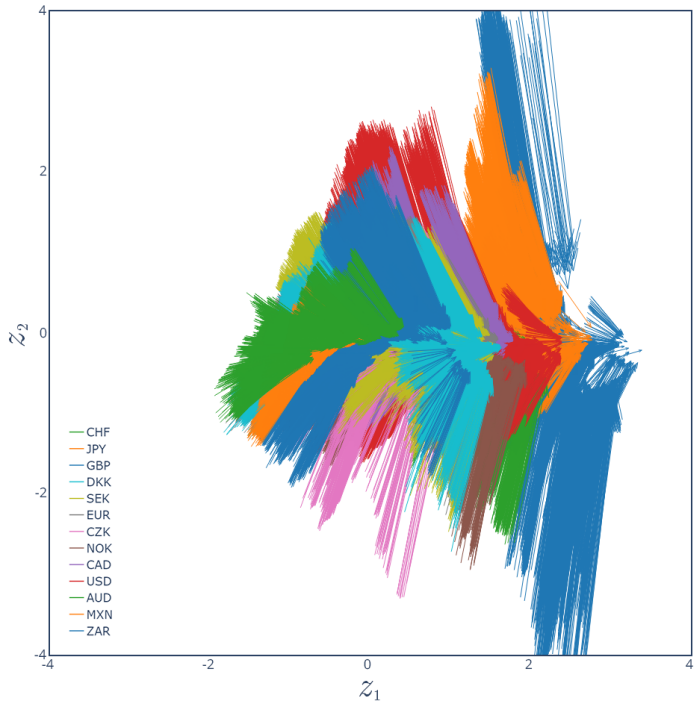

Figure 21: Q-measure "migration map" in latent space for time-to-maturity $\tau = 2y$.

Consider the "migration map" in latent space shown in Figure 21 obtained by "snipping off" the first $\tau = 2y$ of the term structure and re-encoding the curve. This map has two prominent features: the first latent factor z_1 exhibits a large positive drift across the majority of latent space, and the second latent factor z_2 exhibits fast mean reversion toward a line crossing the map horizontally. The former feature is related to the average positive slope of the yield curve, and the second to the flattening of that average slope toward longer maturities. The line toward which z_2 is mean reverting in this chart is where the curve becomes flat on average, given the value of z_1.

In P-measure, the drift is on average zero and the latent space vector is mean reverting toward a fixed position in latent space representing the average curve shape for the currency being modelled. Because this curve shape is on average upward sloping while the target of Q-measure migration in Figure 21 is flat, it follows that excess drift induced by the risk premium has both z_1 and z_2 components. These two components play a similar role to that of the added Hull-White factors in [26] in ensuring that the correct term structure of excess drift can be produced by the model without making the risk premium dependent on model time. Specifically, the drift of z_1 controls the overall difference between short- and long-dated rates, while the drift of z_2 controls how fast the curve flattens for longer maturities.

4.3. Pricing Under P-Measure

A P-measure model will rarely specify every market quote required by its user. For example, a model used for portfolio risk must generate all pricing model inputs for every instrument in the portfolio. Most P-measure models are unable to do so. This problem is particularly acute for P-measure models that simulate the short rate, as these models are unable to calculate risk of anything other than overnight deposits. The models that simulate the entire yield curve but not the volatility surface do better as they are able to price linear instruments, but not interest rate options. To use a P-measure model for risk calculations, practitioners must be able to generate all of the market quotes the model does not specify, but portfolio pricing requires.

Q-measure models sidestep this problem because they can price underlying instruments for the market quotes they do not specify directly. For example, finite-term rates can be obtained by pricing zero coupon bonds, and volatilities by pricing caps and swaptions.

While this approach is not applicable to a standalone P-measure model, it can be used if a P-measure model is paired with a Q-measure model that shares the same state variables. The dual-measure models discussed in the preceding Section 4.2 are already set up this way. This approach can also be used by pairing an autoregressive P-measure model described in Section 4.1 with any Q-measure model that shares the same state variables. The resulting two-model setup is similar to a dual-measure model, except the two models are calibrated independently rather than together.

In this two-model setup, the role of the P-measure model is to calculate the real world probability distribution of the state variables at the risk horizon h. The calculation of market quotes and trade pricing as a function of these state variables becomes the responsibility of Q-measure model.

Figure 22 illustrates how this approach is applied to the Monte Carlo calculation of portfolio risk. The Monte Carlo paths are generated using the P-measure model before the risk horizon and continued using the Q-measure model after the risk horizon. The dependence of trade prices on state variables at the risk horizon is calculated by backward induction from $t > h$ and therefore use only the Q-measure segment of the paths. The probability distribution of these state variables, on the other hand, is calculated for $t < h$ and therefore uses only the P-measure segment of the paths.

Because it is not practical to generate a new set of Q-measure paths for each risk horizon h, the ending state variables of the P-measure segment and the starting state variables of the Q-measure segment will not match exactly. Regression can be used to assign values from one set of state variables to the other. With only two state variables in our VAE, this regression will be more accurate compared to models with more state variables.

One complication that sometimes arises in using the same set of paths for every horizon is that at long horizons, the difference between Q- and P-measure drift causes the two sets of paths to have maximum density in different areas of state space, resulting in a low density of regression inputs in the area of state space where most regression outputs are located. Several ways to address this problem were described in prior literature. Stein [27] described a procedure where switching between the measures is used to generate a single set of paths from which both Q- and P-measure expectations can be computed and Sokol [30] proposed a "path injection" approach

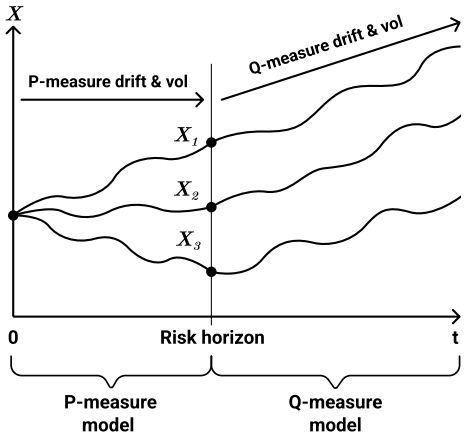

Figure 22: Two-model setup with P-measure model responsible for generating samples of state variable vector $\boldsymbol{X}$ at risk horizon h and Q-measure model responsible for pricing the portfolio for each $\boldsymbol{X}$.

where paths are dynamically added to Q-measure Monte Carlo simulation in order to increase the density of coverage in the area of state space relevant to P-measure calculations.

5. Conclusion

The invention of VAE revolutionized many areas of machine learning, from image processing to natural language recognition. We believe that it holds the same promise for interest rate modelling. In this paper, we describe how four categories of classical interest rate models (two in Q-measure and two in P-measure) can be modified to use VAE for dimension reduction. We propose to call this new model category autoencoder market models (AEMM).

We emphasize that our selection of classical examples for the four model categories discussed in this paper should not be taken as an endorsement of any particular model or category. The purpose of these examples is merely to illustrate how to turn a classical model into its AEMM counterpart. Other classical models can be converted to AEMM by similar means.

Our decision to use autoencoders only for the part of the model that affects curve shapes while leaving other aspects of its classical specification untouched was driven by the desire to maintain continuity with the established market practice in interest rate modelling. In Q-measure, switching to AEMM involves using a special form of volatility basis for forward rate models and a special form of drift for short rate models. In P-measure, model construction involves replacing Nelson-Siegel latent factors by VAE latent variables, and proceeds in the usual way otherwise. It is our hope that conservative use of machine learning to improve a single aspect of the model specification in a transparent and explainable way without making radical changes to the model will facilitate AEMM adoption by the practitioners.

With AEMM, machine learning is only used to create the mapping between yield curve shapes and model state variables. The training process relies on decades of historical data and can be run periodically (e.g., quarterly), with its results carefully

examined before they are used in production. Periodic training is acceptable in this case because machine learning is used in AEMM to replace those aspects of classical models that are usually specified *a priori*, such as the choice of Nelson-Siegel basis to represent the curve.

The accuracy of the resulting mapping can be evaluated and compared to classical methods such as Nelson-Siegel, or previous versions of VAE, using a rigorous process based on measuring curve reconstruction error over the historical dataset. Mapping quality can be examined in detail and even visualized by generating curves from a large number of sample points in latent space. The ability to carefully examine and backtest the quality of machine learning results before using the model in production follows the principles of trustworthy ML. We hope it will be considered during regulatory review of the new model category.

6. Acknowledgments

The principles of training on multi-currency data with one-hot encoding of the currency label have been developed jointly with Oleksiy Kondratyev in a different setting as part of prior research on generative P-measure models. The code used to generate the numerical results was created in collaboration with Svitlana Doroshenko and Andrew Samodurov. The author is grateful for many insights and a detailed review of the early version of this paper by Andrei Lyashenko that led to many improvements, exchange of ideas with Oleksiy Kondratyev during our prior collaboration, and illuminating discussions with Leif Andersen, Marco Bianchetti, Vladimir Chorniy, Igor Halperin, John Hull, Peter Jaeckel, Gordon Lee, Alexander Lipton, Fabio Mercurio, Vladimir Piterbarg, Michael Pykhtin, members of the quant research team at CompatibL, and many others. The author alone is responsible for the errors. No conflicts of interest are reported.

References

[1] A. Kondratyev, "Learning curve dynamics with artificial neural networks," *Risk*, vol. 31, June 2018.

[2] M. Bergeron, N. Fung, J. Hull, Z. Poulos, and A. Veneris, "Variational autoencoders: A hands-off approach to volatility," *The Journal of Financial Data Science*, vol. 4, pp. 125–138, 2022.

[3] H. Buehler, B. Horvath, T. Lyons, I. Perez Arribas, and B. Wood, "A data-driven market simulator for small data environments." Working Paper, SSRN `https://ssrn.com/abstract=3632431`, 2020.

[4] P. Hagan and A. Lesniewski, "LIBOR Market Model with SABR Style Stochastic Volatility." Working Paper, `https://doi.org/10.13140/RG.2.2.22622.89924`, 2006.

[5] C. R. Nelson and A. F. Siegel, "Parsimonious Modeling of Yield Curves," *The Journal of Business*, vol. 60, no. 4, p. 473, 1987.

[6] J. C. Hull and A. D. White, "Numerical Procedures for Implementing Term Structure Models II: Two-Factor Models," *The Journal of Derivatives*, vol. 2, no. 2, pp. 37–48, 1994.

[7] D. Brigo and F. Mercurio, *Interest Rate Models: Theory and Practice - with Smile, Inflation and Credit*. Springer Verlag, 2006.

[8] O. Cheyette, "Markov Representation of the Heath-Jarrow-Morton Model." Working Paper, SSRN `https://ssrn.com/abstract=6073`, 2001.

[9] E. F. Fama and R. R. Bliss, "The information in long-maturity forward rates," *The American Economic Review*, vol. 77, no. 4, pp. 680–692, 1987.

[10] L. E. Svensson, "Estimating Forward Interest Rates with the Extended Nelson & Siegel Method," *Sveriges Riksbank Quarterly Review*, vol. 3, no. 1, pp. 13–26, 1995.

[11] S. Doroshenko, A. Samodurov, and A. Sokol, "Variational Autoencoders for the Yield Curve." To be published, 2022.

[12] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes." Working Paper, arXiv `http://arxiv.org/abs/1312.6114`, 2013.

[13] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *Advances in Neural Information Processing Systems*, vol. 28, Curran Associates, Inc., 2015.

[14] D. P. Kingma, D. J. Rezende, S. Mohamed, and M. Welling, "Semi-Supervised Learning with Deep Generative Models," *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[15] F. X. Diebold and C. Li, "Forecasting the Term Structure of Government Bond Yields," *Journal of Econometrics*, vol. 130, no. 2, pp. 337–364, 2006.

[16] R. Rebonato, K. McKay, and R. White, *The SABR/LIBOR Market Model*. Wiley, 2009.

[17] D. Heath, R. Jarrow, and A. Morton, "Bond Pricing and the Term Structure of Interest Rates: A New Methodology for Contingent Claims Valuation," *Econometrica*, vol. 60, no. 1, pp. 77–105, 1992.

[18] A. Brace, D. Gatarek, and M. Musiela, "The Market Model of Interest Rate Dynamics," *Mathematical Finance*, vol. 7, no. 2, pp. 127–155, 1997.

[19] F. Jamshidian, "LIBOR and swap market models and measures," *Finance and Stochastics*, vol. 1, p. 293–330, 1997.

[20] K. Miltersen, K. Sandmann, and D. Sondermann, "Closed form solutions for term structure derivatives with log-normal interest rates," *The Journal of Finance*, vol. 52, pp. 409–430, 1997.

[21] L. Andersen and J. Andreasen, "Volatility skews and extensions of the libor market model," *Applied Mathematical Finance*, vol. 7, no. 1, pp. 1–32, 2000.

[22] J. H. Christensen, F. X. Diebold, and G. D. Rudebusch, "An Arbitrage-Free Generalized Nelson–Siegel Term Structure Model," *Econometrics Journal*, vol. 12, no. 3, pp. 33–64, 2009.

[23] J. H. Christensen, F. X. Diebold, and G. D. Rudebusch, "The Affine Arbitrage-Free Class of Nelson–Siegel Term Structure Models," *Journal of Econometrics*, vol. 164, no. 1, pp. 4–20, 2011.

[24] A. Lyashenko and Y. Goncharov, "Bridging P-Q Modeling Divide with Factor HJM Modeling Framework." Working Paper, SSRN `https://www.ssrn.com/abstract=3995533`, 2021.

[25] J. Hull, A. Sokol, and A. White, "Short rate joint measure models," *Risk*, no. October, pp. 59–63, 2014.

[26] R. C. Barker, A. S. Dickinson, and A. Lipton, "Simulation in the real world," *Risk Management eJournal*, 2016. `http://www.ssrn.com/abstract=292060`.

[27] H. Stein, "Two measures for the price of one," *Risk Magazine*, 3 2015.

[28] P. H. Dybvig, J. E. Ingersoll, Jr., and S. A. Ross, "Long Forward and Zero-Coupon Rates Can Never Fall," *The Journal of Business*, vol. 69, no. 1, p. 1, 1996.

[29] D. Filipovic, "A Note on the Nelson–Siegel Family," *Mathematical Finance*, vol. 9, no. 4, pp. 349–359, 1999.

[30] A. Sokol, *Long-term Portfolio Simulation: For XVA, Limits, Liquidity and Regulatory Capital*. Risk Books, 2014.

[31] O. Kondratyev and A. Sokol, "Machine Learning for Long Risk Horizons: Market Generator Models." RiskLive conference presentation and to be published, 2000.

[32] G. U. Yule, "On a method of investigating periodicities in disturbed series, with special reference to wolfer's sunspot numbers," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 226, pp. 267–298, 1927.

[33] E. Slutzky, "The summation of random causes as the source of cyclic processes," *Econometrica*, vol. 5, no. 2, pp. 105–146, 1937.

[34] F. X. Diebold and G. D. Rudebusch, *Yield Curve Modeling and Forecasting - The Dynamic Nelson-Siegel Approach*. Princeton University Press, 2013.

[35] C. A. Sims, "Macroeconomics and reality," *Econometrica*, vol. 48, no. 1, pp. 1–48, 1980.

[36] D. A. Jones, "Nonlinear autoregressive processes." Ph.D. Thesis, University of London, `https://spiral.imperial.ac.uk/bitstream/10044/1/22669/2/Jones-DA-1977-PhD-Thesis.pdf`, 1976.

[37] D. A. Jones, "Nonlinear autoregressive processes," *Proceedings of the Royal Society of London, Series A*, vol. 360, no. 1700, pp. 71–95, 1978.

[38] V. Chorniy and A. Greenberg, "Bayesian Networks and Stochastic Factor Models." Working Paper, SSRN `https://ssrn.com/abstract=2688324`, 2015.

[39] R. Stanton, "A Non-Parametric Model of Term Structure Dynamics and the Market Price of Interest Rate Risk," *Journal of Finance*, vol. 52, no. 5, pp. 1973–2002, 1997.

[40] S. H. Cox and H. W. Pedersen, "Nonparametric Estimation of Interest Rate Term Structure and Insurance Applications." Working Paper, Georgia State University, 1999.

[41] R. Ahmad and P. Wilmott, "The Market Price of Interest Rate Risk: Measuring and Modeling Fear and Greed in Fixed Income Markets," *Wilmott*, vol. January, pp. 64–70, 2007.