Factor-Based Neural Network Forecasting of the German Yield Curve

Menno Westenbrink

S4847261

Supervisors:

Alexander Düring
European Central Bank

Artem Tsvetkov
Rijksuniversiteit Groningen

February 25, 2026

Contents

1 Introduction

The yield curve, describing the relationship between interest rates and time to maturity for government bonds, plays a central role in modern finance. It serves as a benchmark for pricing fixed-income securities, reflects market expectations about future economic conditions, and responds dynamically to monetary policy decisions. For central banks such as the European Central Bank, with whom this research is conducted, accurate yield curve forecasts are essential for forward-looking policy decisions. For bond portfolio managers, these forecasts are equally important for managing interest rate risk and identifying relative value opportunities. However, forecasting the yield curve presents significant challenges due to its high dimensionality and complex, time-varying dynamics.

Factor models have become the dominant framework by exploiting the empirical finding that a small number of factors explains the vast majority of yield variation, effectively reducing a high-dimensional problem to a low-dimensional one Nelson and Siegel (1987); Litterman and Scheinkman (1991). Diebold and Li (2006) made the crucial step of combining this dimensionality reduction with time-series forecasting, treating extracted factors as predictable quantities. This thesis follows their framework but evaluates three competing dimensionality reduction approaches on German zero-coupon yield data: Principal Component Analysis on yield levels (Level-PCA), PCA on yield changes (Changes-PCA), and the Nelson-Siegel parametric model. We find that Changes-PCA dominates both alternatives in reconstruction accuracy and is therefore adopted as the basis for forecasting.

A key methodological contribution is the treatment of PCA loadings as a locally stable coordinate system. We demonstrate empirically that rolling loadings L_t change negligibly over a one-month horizon, which allows them to serve as a consistent basis across time: current and all lagged yield changes are projected onto the same loadings L_t, ensuring that inputs and forecast targets are expressed in a common coordinate system. This consistency is essential for a well-posed sequential learning problem and enables a clean three-step forecasting procedure: estimate L_t from a rolling window, project all features into this basis, and forecast the direction of the four principal components one month ahead.

To capture potential nonlinear dynamics in factor evolution, we employ a Long Short-Term Memory (LSTM) network Hochreiter and Schmidhuber (1997), augmented with swaption implied volatility as a forward-looking market variable. The contribution is threefold. First, we systematically evaluate three dimensionality reduction methods and establish Changes-PCA as the superior approach. Second, we formalize and validate the loading stability assumption that underlies the consistent coordinate system. Third, we evaluate whether an LSTM with implied volatility inputs can forecast principal component directions with accuracy above the 50% baseline.

The remainder of this thesis is organized as follows. Section 2 reviews the relevant literature. Section 3 describes the data. Section 4 compares the three dimensionality reduction approaches and validates the stability assumptions. Section 5 presents the LSTM forecasting framework and results. Section 6 concludes.

2 Literature Review

The literature on yield curve modeling and forecasting has evolved considerably over the past decades, progressing from parametric curve-fitting methods to factor-based forecasting frameworks, and more recently incorporating machine learning techniques. This review traces this evolution, beginning with foundational work on dimensionality reduction, proceeding through the development of dynamic factor forecasting models, and concluding with recent applications of neural networks to yield curve analysis. This review is structured to highlight how each contribution addresses specific aspects of the yield curve forecasting problem.

2.1 Parametric Dimensionality Reduction: The Nelson-Siegel Model

The challenge of modeling yield curves across numerous maturities was first addressed systematically by Nelson and Siegel (1987), who proposed a parsimonious functional form capable of generating the diverse shapes observed in practice. The Nelson-Siegel model represents the yield curve at time t for maturity τ as:

$$y_t(\tau) = \beta_{1,t} + \beta_{2,t}\left(\frac{1-e^{-\lambda\tau}}{\lambda\tau}\right) + \beta_{3,t}\left(\frac{1-e^{-\lambda\tau}}{\lambda\tau} - e^{-\lambda\tau}\right) \tag{1}$$

where $\beta_{1,t}$, $\beta_{2,t}$, and $\beta_{3,t}$ are time-varying parameters, and λ is a fixed decay parameter that determines the maturity at which the loading on the curvature factor reaches its maximum. The elegance of this specification lies in its factor interpretation: $\beta_{1,t}$ represents the level (long-term yield), $\beta_{2,t}$ represents the slope (short-term vs. long-term spread), and $\beta_{3,t}$ represents the curvature (medium-term hump). By reducing an N-dimensional yield curve to just three parameters, Nelson and Siegel (1987) demonstrated that complex yield curve shapes could be parsimoniously represented while maintaining economic interpretability.

The key insight of the Nelson-Siegel model is that it achieves dimensionality reduction through an economically motivated functional form rather than through statistical methods. The three factors correspond to parallel shifts (level), tilting (slope), and bowing (curvature) of the yield curve, movements that practitioners had long recognized as the dominant modes of variation. However, the Nelson and Siegel (1987) approach imposes a specific functional structure on how these factors load across maturities, which may be restrictive if the true factor loadings differ from the assumed exponential decay pattern.

2.2 Statistical Dimensionality Reduction: Principal Component Analysis

Litterman and Scheinkman (1991) took a different approach to dimensionality reduction by applying principal component analysis (PCA) to a large dataset of U.S. Treasury yields. Rather than imposing a parametric structure, they let the data reveal the dominant factors driving yield curve variation. Their empirical analysis demonstrated that the first three principal components explain over 95% of the total variance in yield changes across maturities. Moreover, the factor loadings from PCA exhibit patterns strikingly similar to the Nelson and Siegel (1987) factors: the first principal component loads roughly equally across all maturities (level), the second principal component loads with opposite signs on short and long maturities (slope), and the third principal component loads most heavily on intermediate maturities (curvature).

This finding has profound implications for yield curve modeling. It suggests that despite the high dimensionality of observed yields, the effective dimensionality of yield curve movements is remarkably low. The PCA approach offers two advantages over parametric methods like Nelson and Siegel (1987). First, it requires no assumption about the functional form of factor loadings; the loadings are estimated directly from the data. Second, it provides a natural measure of how much variance each factor explains, allowing researchers to determine the appropriate number of factors empirically rather than assuming three factors a priori. The dominance of three factors in explaining yield variation has been confirmed across numerous markets and time periods, establishing factor models as the standard framework for yield curve analysis.

2.3 Dynamic Factor Forecasting: The Diebold-Li Framework

While Nelson and Siegel (1987) and Litterman and Scheinkman (1991) established that yield curves can be effectively represented by three factors, Diebold and Li (2006) made the crucial step of creating a forecasting framework by treating these factors as time series to be predicted. Diebold and Li (2006) combined the Nelson and Siegel (1987) factor structure with vector autoregression (VAR) models to forecast the entire yield curve. Their two-stage approach works as follows. First, at each point in time, they estimate the three Nelson and Siegel (1987) factors $\beta_{1,t}$, $\beta_{2,t}$, and $\beta_{3,t}$ by fitting equation (1) to observed yields. Second, they model the dynamics of these factors using a VAR(1) specification:

$$\begin{bmatrix}\beta_{1,t}\\ \beta_{2,t}\\ \beta_{3,t}\end{bmatrix} = \begin{bmatrix}c_1\\ c_2\\ c_3\end{bmatrix} + \begin{bmatrix}\phi_{11} & \phi_{12} & \phi_{13}\\ \phi_{21} & \phi_{22} & \phi_{23}\\ \phi_{31} & \phi_{32} & \phi_{33}\end{bmatrix} \begin{bmatrix}\beta_{1,t-1}\\ \beta_{2,t-1}\\ \beta_{3,t-1}\end{bmatrix} + \begin{bmatrix}\varepsilon_{1,t}\\ \varepsilon_{2,t}\\ \varepsilon_{3,t}\end{bmatrix} \tag{2}$$

To generate forecasts h periods ahead, they iterate the VAR forward and then reconstruct the yield curve at each horizon using the Nelson and Siegel (1987) formula 1 with the forecasted factor values. This approach transforms a high-dimensional forecasting problem (predicting yields at many maturities) into a low-dimensional one (predicting three factors).

The key empirical finding of Diebold and Li (2006) is that this factor-based VAR approach significantly outperforms a simple random walk benchmark in out-of-sample forecasts, particularly at medium to long horizons. The random walk model, had proven difficult to beat in earlier research, making this result notable. The success of the Diebold and Li (2006) framework stems from exploiting the cross-sectional restrictions implied by the factor structure: rather than forecasting each maturity independently, the model recognizes that yields across maturities move together in predictable ways captured by the three factors. This framework has become the benchmark against which subsequent yield curve forecasting methods are evaluated.

However, the Diebold and Li (2006) approach has an important limitation: the VAR specification assumes that the relationship between past and future factor values is linear. While this assumption simplifies estimation and interpretation, it may be restrictive if yield curve dynamics exhibit regime changes, threshold effects, or other nonlinearities. Moreover, the model relies exclusively on the historical behavior of the factors themselves, ignoring potentially valuable information from other market variables that might help predict future movements.

2.4 Neural Networks for Yield Curve Analysis

The application of neural networks to yield curve modeling has taken several distinct forms, addressing different aspects of the curve analysis problem. We review three key contributions that collectively motivate this approach: one demonstrating the ability of neural networks to capture nonlinear dynamics in interest rate stress testing, one proposing autoencoders as a flexible alternative for dimensionality reduction, and one applying neural networks with market variables to bond return forecasting.

2.4.1 Capturing Nonlinear Dynamics: Kondratyev (2018)

Kondratyev (2018) demonstrated the potential of artificial neural networks to capture complex nonlinear relationships in interest rate markets through an application to yield curve stress testing. While stress testing differs from forecasting, it involves generating plausible adverse scenarios rather than predicting most likely outcomes, Kondratyev (2018)'s work provides important evidence that ANNs can model the nonlinear dynamics of yield curves effectively. Using a feedforward neural network architecture, Kondratyev (2018) showed that ANNs could learn the joint distribution of yield curve changes across maturities, including tail dependencies and non-Gaussian features that are difficult to capture with linear models.

The key insight from this work is that yield curve movements exhibit nonlinear patterns that vary with the market regime, level of rates, and volatility environment. During periods of monetary policy transitions or financial stress, the relationship between factors may differ substantially from normal periods. Neural networks, with their ability to approximate arbitrary nonlinear functions through hidden layers, offer a flexible framework for capturing these regime-dependent dynamics. While Kondratyev (2018)'s application focused on scenario generation rather than forecasting, the demonstrated capability of ANNs to model complex yield curve dynamics motivates their use in this forecasting context.

2.4.2 LSTM-Based Yield Curve Forecasting: Suimon et al. (2020)

Beyond the autoencoder for dimensionality reduction, Suimon et al. (2020) implement a Long Short-Term Memory (LSTM) model to directly forecast the yield curve one month ahead using historical interest rate data. This forms the more relevant methodological comparison for the forecasting component of yield curve modeling.

The LSTM model takes as input the full yield curve across maturities (2-, 5-, 7-, 10-, 15-, and 20-year rates) at four consecutive weekly timesteps. Three weeks prior, two weeks prior, one week prior, and the current date are used. This sequence is mapped to the predicted yield curve one month later:

$$\hat{\mathbf{y}}_{t+h} = \text{LSTM}(\mathbf{y}_t, \mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \mathbf{y}_{t-3}) \tag{3}$$

where $\mathbf{y}_{t-k}$ denotes the vector of observed yields k weeks before the investment date and h represents the one-month forecast horizon. Two LSTM variants are tested: LSTM20 and LSTM50, with 20 and 50 units respectively, both using hyperbolic tangent activation and sigmoid recurrent activation, consistent with the standard LSTM architecture of Hochreiter and Schmidhuber (1997). The model is trained on a rolling window of 5 years of weekly data and retrained annually.

Investment decisions are based on the direction of predicted yield changes: if the model forecasts a yield increase at a given maturity, the corresponding government bond is shorted; if a decrease is forecast, the bond is bought. The investment horizon is one month.

The LSTM outperforms both the autoencoder strategy and the trend-following benchmark in terms of cumulative capital gains, particularly across the 10- and 20-year maturity government bonds. Suimon et al. (2020) attribute this advantage to the LSTM's ability to exploit temporal dependencies in historical interest rate data, which the autoencoder that operates only on the current curve shape cannot capture. The VAR model performs comparably to the LSTM, suggesting that much of the predictive signal is captured already by linear dynamics, with the LSTM providing additional but modest gains through nonlinear pattern recognition.

A limitation noted by Suimon et al. (2020) is the reduced interpretability of the LSTM strategy relative to the autoencoder approach: it is not straightforward to attribute the LSTM's trading signal to either trend-following behaviour or curve-shape mean reversion, as both mechanisms may simultaneously drive the predictions.

2.4.3 Neural Network Forecasting with Market Variables: Verner et al. (2021)

Verner et al. (2021) applied nonlinear autoregressive neural networks (NAR and NARX architectures) to forecast long-term bond prices, representing one of the first attempts to incorporate external market variables into neural network-based fixed-income forecasting. Their approach involved forecasting individual bond prices (rather than yield curves) using two types of models: a pure autoregressive network using only past bond prices, and an augmented network incorporating 50-year EUR interest rate swaps and the VIX volatility index as external inputs.

The NARX (Nonlinear Autoregressive with eXogenous inputs) architecture they employed can be represented as:

$$P_t = f(P_{t-1}, P_{t-2}, ..., P_{t-d}, X_t, X_{t-1}, ..., X_{t-d}) + \varepsilon_t \tag{4}$$

where P_t is the bond price, X_t represents external variables (swap rates and VIX), and $f(\cdot)$ is approximated by a feedforward neural network with multiple hidden layers. Their results showed that both the Levenberg-Marquardt and Scaled Conjugate Gradient training algorithms achieved high in-sample fit ($R^2 > 95\%$), demonstrating the capability of neural networks to model bond price dynamics.

However, their key finding regarding external variables was somewhat negative: incorporating VIX and swap rates provided only marginal improvement over models using price history alone. This result has several potential explanations. First, their focus on individual bond prices rather than the entire yield curve structure means they did not exploit the cross-sectional information that factor models capture. Second, VIX represents equity market volatility and may not be the most relevant measure for fixed-income forecasting; curve-specific measures like swaption implied volatility might be more informative.

3 Data

3.1 Yield Curve Data

The zero-coupon yield curve data for Germany is obtained from IBOXX historical bond data, which have been used to obtain the zero coupon yields. The final dataset spans from 4 January 1999 to 28 January 2026, comprising 6957 observations (weekends excluded) across 15 maturity points ranging from 1 to 30 years.

Table 1 presents summary statistics for the zero-coupon yields across all maturities. The sample period captures multiple interest rate regimes, including the pre-financial crisis era, the European sovereign debt crisis, a prolonged low-interest-rate environment, and the recent monetary policy normalization.

Figure 1 visualizes the evolution of the German zero-coupon yield curve over the sample period.

Table 1: Summary Statistics of German Zero-Coupon Yields

Maturity	Mean	Std. Dev.	Min	Q25	Median	Q75	Max	N
1Y	0.015	0.018	−0.009	−0.002	0.015	0.030	0.051	6,957
2Y	0.016	0.018	−0.010	−0.002	0.017	0.030	0.052	6,957
3Y	0.017	0.018	−0.010	−0.002	0.019	0.030	0.052	6,957
4Y	0.018	0.018	−0.010	−0.001	0.021	0.033	0.052	6,957
5Y	0.019	0.018	−0.010	0.000	0.022	0.035	0.053	6,957
6Y	0.020	0.018	−0.010	0.002	0.023	0.036	0.053	6,957
7Y	0.021	0.018	−0.010	0.003	0.023	0.037	0.055	6,957
8Y	0.022	0.018	−0.009	0.004	0.024	0.039	0.055	6,957
9Y	0.023	0.018	−0.009	0.005	0.025	0.039	0.056	6,957
10Y	0.024	0.018	−0.009	0.006	0.025	0.040	0.056	6,957
12Y	0.026	0.018	−0.008	0.008	0.027	0.041	0.057	6,957
15Y	0.027	0.017	−0.008	0.010	0.028	0.042	0.058	6,957
20Y	0.029	0.018	−0.007	0.011	0.030	0.045	0.059	6,957
25Y	0.031	0.019	−0.006	0.013	0.030	0.047	0.068	6,957
30Y	0.030	0.018	−0.005	0.014	0.030	0.045	0.063	6,957

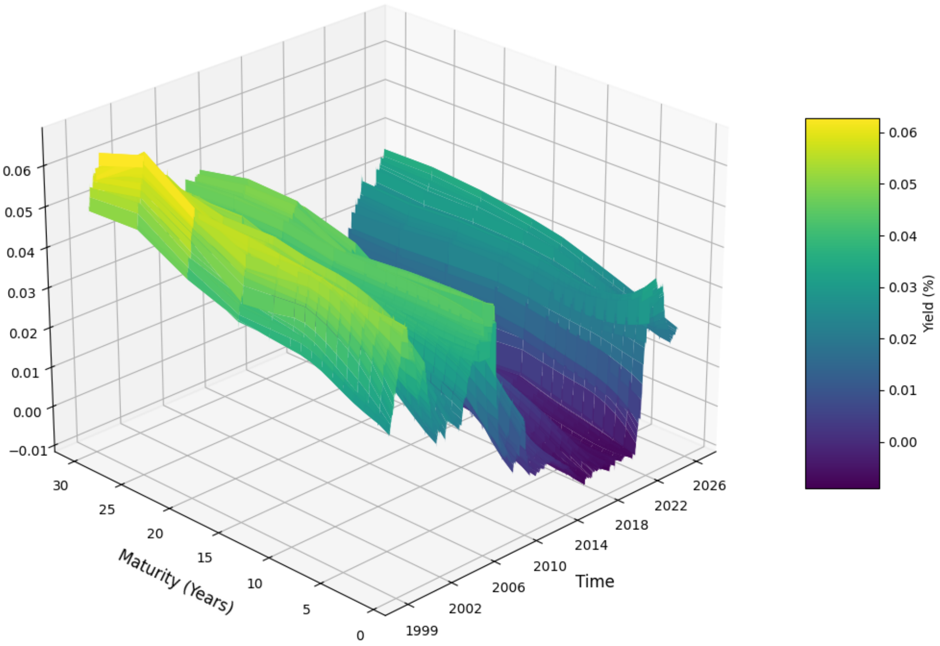


Figure 1: Evolution of the German Zero-Coupon Yield Curve

3.2 Swaption Implied Volatility Data

The implied volatility data is derived from 3-month swaption prices obtained from the statistics database within the ECB. It spans the same sample period as the yield curve data, from January 4, 1999 to January 28, 2026. For dates with missing implied volatility observations (96 dates), values are interpolated using linear interpolation to ensure a complete time series.

Linear interpolation estimates missing values IV_t at time t based on the nearest available observations before and after the missing date:

$$IV_t = IV_{t-k} + \frac{t-(t-k)}{(t+j)-(t-k)}\left(IV_{t+j} - IV_{t-k}\right) \tag{5}$$

where IV_{t-k} and IV_{t+j} are the nearest observed values before and after time t, respectively, with k and j denoting the number of days from t to these observations.

Table 2 presents summary statistics for the 3-month swaption implied volatility.

Figure 2 illustrates the time series evolution of the 3-month swaption implied volatility over the sample period.

Table 2: Summary Statistics of 3-Month Swaption Implied Volatility

Variable	Mean	Std. Dev.	Min	Q25	Median	Q75	Max	N
3M IV	66.632	25.103	26.340	49.580	62.310	79.500	180.310	6,957

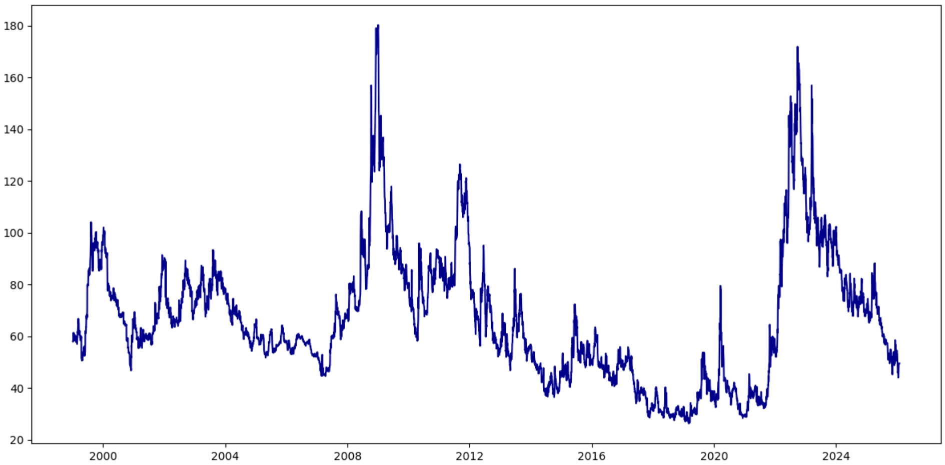


Figure 2: Evolution of 3-Month Swaption Implied Volatility

4 Yield Curve Fitting

4.1 Motivation

Forecasting the entire term structure of interest rates presents a high-dimensional challenge: predicting 15 distinct maturities simultaneously requires modeling complex dependencies. We address this by employing dimensionality reduction techniques that exploit the strong co-movement of yields across maturities. We evaluate three approaches: Principal Component Analysis on yield levels (Level-PCA), Principal Component Analysis on yield changes (Changes-PCA), and the Nelson-Siegel (NS) parametric model. All methods compress the yield curve into a lower-dimensional representation, enabling more efficient forecasting while preserving essential term structure dynamics.

4.2 Principal Component Analysis Approach

4.2.1 Mathematical Framework

Let $\mathbf{y}_t \in \mathbb{R}^{15}$ denote the vector of zero-coupon yields at time t. For Level-PCA, we apply PCA directly to yield levels; for Changes-PCA, we apply PCA to 21-trading-day yield changes:

$$\Delta \mathbf{y}_t = \mathbf{y}_t - \mathbf{y}_{t-h}, \quad h = 21 \tag{6}$$

In both cases, the covariance matrix eigenvalue decomposition yields:

$$\mathbf{\Sigma} = \mathbf{L}\mathbf{\Lambda}\mathbf{L}^T \tag{7}$$

where $\mathbf{L} \in \mathbb{R}^{15 \times k}$ contains the loadings for $k = 4$ factors. Factors are extracted via:

$$\mathbf{f}_t = \mathbf{y}_t \mathbf{L} \quad \text{or} \quad \mathbf{f}_t = \Delta \mathbf{y}_t \mathbf{L} \tag{8}$$

and reconstructed as:

$$\hat{\mathbf{y}}_t = \mathbf{f}_t \mathbf{L}^T \quad \text{or} \quad \widehat{\Delta \mathbf{y}}_t = \mathbf{f}_t \mathbf{L}^T \tag{9}$$

Loadings $\mathbf{L}_t$ are estimated on a rolling window of $\tau = 756$ trading days. For forecasting, we make the key stability assumption:

$$\mathbf{L}_{t+h} \approx \mathbf{L}_t \tag{10}$$

This enables a clean three-step procedure: (1) estimate $\mathbf{L}_t$ from the rolling window ending at t, (2) forecast factors $\hat{\mathbf{f}}_{t+h}$, and (3) reconstruct the future yield curve using the *same* stale loadings $\mathbf{L}_t$:

$$\hat{\mathbf{y}}_{t+h} = \mathbf{y}_t + \hat{\mathbf{f}}_{t+h} \mathbf{L}_t^T \tag{11}$$

The empirical validity of assumption (10) is tested in the following subsections.

4.2.2 Empirical Validation

We test assumption (10) using the German zero-coupon yield data described in Section 2, yielding 6,161 observations. Table 3 reports reconstruction errors when using stale loadings from time t to decompose yields at time $t + 21$, and Table 4 reports the ratio of stale to fresh loading errors, directly measuring loading stability.

Table 3: Reconstruction Errors Using Stale Loadings

Statistic	Level-PCA RMSE	Changes-PCA RMSE
Mean	0.001087	0.000136
Median	0.000645	0.000095
Std Dev	0.000880	0.000148
Min	0.000117	0.000011
Max	0.003933	0.001937

Table 4: Loading Stability Ratios (Stale/Fresh RMSE)

RMSE Ratios	$\frac{\text{Level-PCA Stale}}{\text{Level-PCA Fresh}}$	$\frac{\text{Changes-PCA Stale}}{\text{Changes-PCA Fresh}}$
Mean	1.031	1.062
Median	1.008	1.006

The median stability ratios indicate that using one-month-old loadings degrades reconstruction quality by only 0.8% for Level-PCA and 0.6% for Changes-PCA, providing strong support for assumption (10). Changes-PCA exhibits superior loading stability, suggesting that yield curve dynamics are more stable when measured in first differences.

4.3 Nelson-Siegel Approach

4.3.1 Mathematical Framework

The Nelson-Siegel model (Nelson and Siegel, 1987) parametrically represents the yield curve with three factors (level, slope, curvature) and a decay parameter λ:

$$y(\tau) = \beta_0 + \beta_1 \frac{1 - e^{-\lambda\tau}}{\lambda\tau} + \beta_2 \left(\frac{1 - e^{-\lambda\tau}}{\lambda\tau} - e^{-\lambda\tau} \right) \tag{12}$$

where τ denotes maturity. In matrix form, yields are reconstructed as:

$$\mathbf{y}_t = \mathbf{L}(\lambda_t)\boldsymbol{\beta}_t \tag{13}$$

where $\mathbf{L}(\lambda_t) \in \mathbb{R}^{15\times3}$ contains the maturity-dependent loadings determined by λ_t, and $\boldsymbol{\beta}_t = [\beta_0, \beta_1, \beta_2]^T$ are the factor weights.

For forecasting, we estimate the optimal λ_t by minimizing reconstruction error over the same 3-year rolling window used for PCA. The key assumption is:

$$\lambda_{t+h} \approx \lambda_t \tag{14}$$

Given stable λ, we can refit the betas $\boldsymbol{\beta}_{t+h}$ at the forecast horizon while maintaining consistent yield curve structure.

4.3.2 Empirical Validation

We test assumption (14) using the same data, yielding 6,183 observations. Table 5 shows reconstruction errors when using stale λ from time t to fit yields at time $t + 21$, and Table 6 reports the ratio of stale to fresh errors, directly measuring λ stability.

Table 5: Nelson-Siegel Reconstruction Errors (Stale λ)

Statistic	RMSE
Mean	0.000642
Median	0.000534
Std Dev	0.000374
Min	0.000112
Max	0.002317

Table 6: Nelson-Siegel λ Stability Ratios (Stale/Fresh RMSE)

RMSE Ratio	$\frac{\text{NS Stale}}{\text{NS Fresh}}$
Mean	1.018
Median	1.001

The median stability ratio of 1.001 indicates that using one-month-old λ degrades reconstruction quality by only 0.1%, providing strong support for assumption (14).

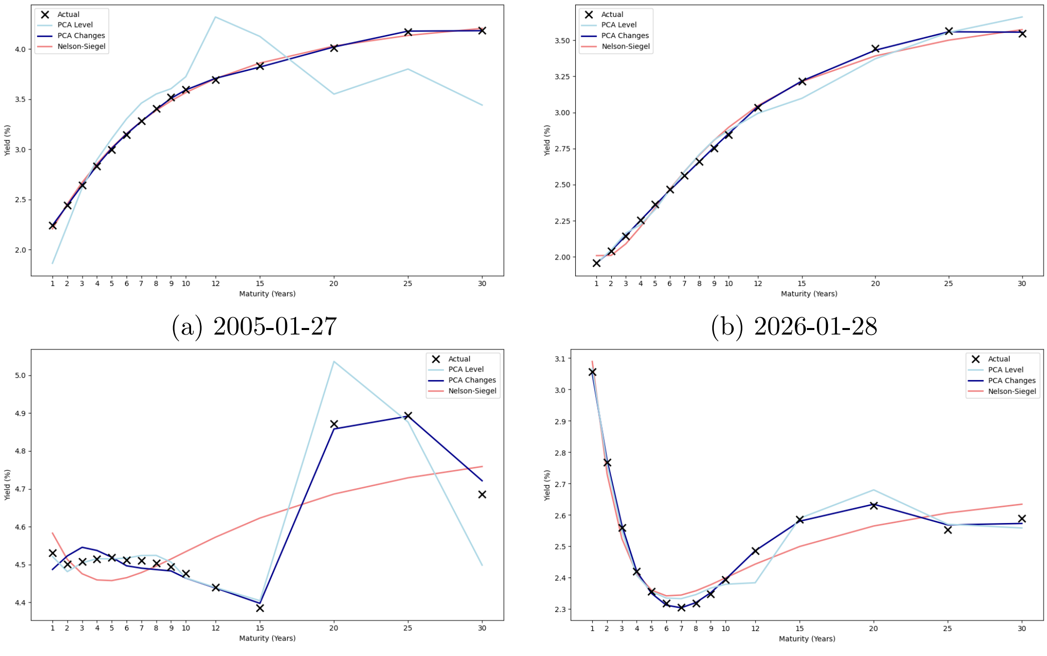
4.4 Comparative Performance

Table 7 presents pairwise RMSE performance comparisons across all three methods. Ratios below 1 indicate the numerator method achieves lower reconstruction errors than the denominator.

Table 7: RMSE Performance Ratios

RMSE Ratios	$\frac{PCA\ Level}{NS}$	$\frac{PCA\ Level}{PCA\ Changes}$	$\frac{PCA\ Changes}{NS}$
Mean	2.300	11.227	0.242
Median	1.258	7.836	0.174
Min	0.080	0.232	0.017
Max	17.534	155.473	3.465

Figure 3 illustrates how this plays out across four yield curve environments. In normal conditions (panels a, b), all three methods fit well except for PCA Level. In irregular environments such as the 2008 financial crisis (panel c) and the 2024 inverted curve (panel d), we see PCA Level breaking down and the inability of Nelson Siegel to capture the shape, while PCA Changes continues to track the actual curve closely.



(a) 2005-01-27

(b) 2026-01-28

(c) 2008-06-24

(d) 2024-06-24

Figure 3: Yield Curve Fits

PCA Changes achieves lower reconstruction errors in 99.1% of observations versus Nelson-Siegel and 98.8% versus PCA Level. PCA Level performs marginally worse than Nelson-Siegel, winning only 41.8% of observations. These results provide strong empirical support for dimension reduction via PCA Changes, forming the foundation for our forecasting framework.

5 Forecasting Principal Components with LSTM

5.1 Motivation and Approach

Having established in Section 4 that Changes-PCA provides superior reconstruction accuracy relative to both Level-PCA and Nelson-Siegel, we adopt this framework as the basis for forecasting. The problem reduces to predicting the future values of four principal components rather than modeling 15 yield maturities simultaneously.

We frame the forecasting problem as a binary classification task: for each PC k at time t, the target variable is the direction of change over a 21-trading-day horizon:

$$y_{k,t} = \nvdash [f_{k,t+21} - f_{k,t} > 0] \tag{15}$$

where $f_{k,t}$ denotes the value of the k-th principal component at time t. This formulation follows Suimon et al. (2020), who similarly cast LSTM-based yield curve forecasting as a directional prediction problem.

5.2 A Consistent Factor Coordinate System

A central methodological consideration is that all inputs to the LSTM must be expressed in a common coordinate system. At each observation t, loadings $\mathbf{L}_t \in \mathbb{R}^{15\times4}$ are estimated on the rolling window $[t-\tau, t)$ with $\tau = 756$ trading days. These loadings define the coordinate system in which all features are constructed.

We forecast the direction of factor changes over a horizon of $h = 21$ trading days. The current state is the h-day cumulative yield curve change ending at t:

$$\Delta\mathbf{y}_t = \mathbf{y}_t - \mathbf{y}_{t-h} \tag{16}$$

projected onto $\mathbf{L}_t$ to obtain the current factor values:

$$\mathbf{f}_t = \Delta\mathbf{y}_t\mathbf{L}_t \tag{17}$$

To capture temporal dynamics, we construct lagged factor observations at $k \in \{5, 10, 15, 20\}$ trading days prior:

$$\Delta\mathbf{y}_{t-k} = \mathbf{y}_{t-k} - \mathbf{y}_{t-k-h} \tag{18}$$

Critically, each lagged observation is projected onto the *same* current loadings $\mathbf{L}_t$:

$$\mathbf{f}_{t-k} = \Delta\mathbf{y}_{t-k}\mathbf{L}_t \quad k \in \{5, 10, 15, 20\} \tag{19}$$

This ensures that $\mathbf{f}_t$ and all $\mathbf{f}_{t-k}$ are expressed in the same basis, making them directly comparable as a time series. Projecting lagged changes onto their own historical loadings $\mathbf{L}_{t-k}$ instead would introduce basis inconsistency, as coordinates would carry different meanings across time steps.

The validity of this approach rests on the loading stability assumption (10) tested in section 4. Since the maximum lag $k = 20$ is less than the forecast horizon $h = 21$, and section 4.2.2 demonstrated that using stale loadings over a 21-day horizon introduces minimal additional reconstruction error, we can safely project all lagged observations onto $\mathbf{L}_t$ without material degradation.

For each observation, we also compute the reconstruction error when representing $\Delta \mathbf{y}_t$ with the four factors:

$$\mathrm{RMSE}_t = \sqrt{\frac{1}{15} \left\| \Delta \mathbf{y}_t - \mathbf{f}_t \mathbf{L}_t^T \right\|^2} \tag{20}$$

This metric quantifies how well the current yield curve change conforms to the estimated factor structure.

In addition to the factor sequences and reconstruction errors, the eigenvalues $\boldsymbol{\lambda}_t = (\lambda_{1,t}, \ldots, \lambda_{4,t})$ from the PCA estimation are included as static features. These capture the time-varying explained variance per component.

The full feature set at time t thus consists of time-varying sequence features and time-invariant static features:

$$\mathcal{X}_t = \{\mathbf{f}_{t-k}, \mathrm{RMSE}_{t-k}, \mathrm{IV}_{t-k}\}_{k \in \{0,5,10,15,20\}} \cup \{\boldsymbol{\lambda}_t\} \tag{21}$$

where IV_{t-k} denotes the 3-month swaption implied volatility at time $t - k$. The sequence features are all expressed within or relative to the coordinate system defined by $\mathbf{L}_t$.

5.3 LSTM Architecture and Training

5.3.1 Model Architecture

The directional forecasting model employs a Long Short-Term Memory (LSTM) network as presented in Hochreiter and Schmidhuber (1997), a type of recurrent neural network designed to capture temporal dependencies in sequential data. Figure 4 illustrates the complete architecture used.

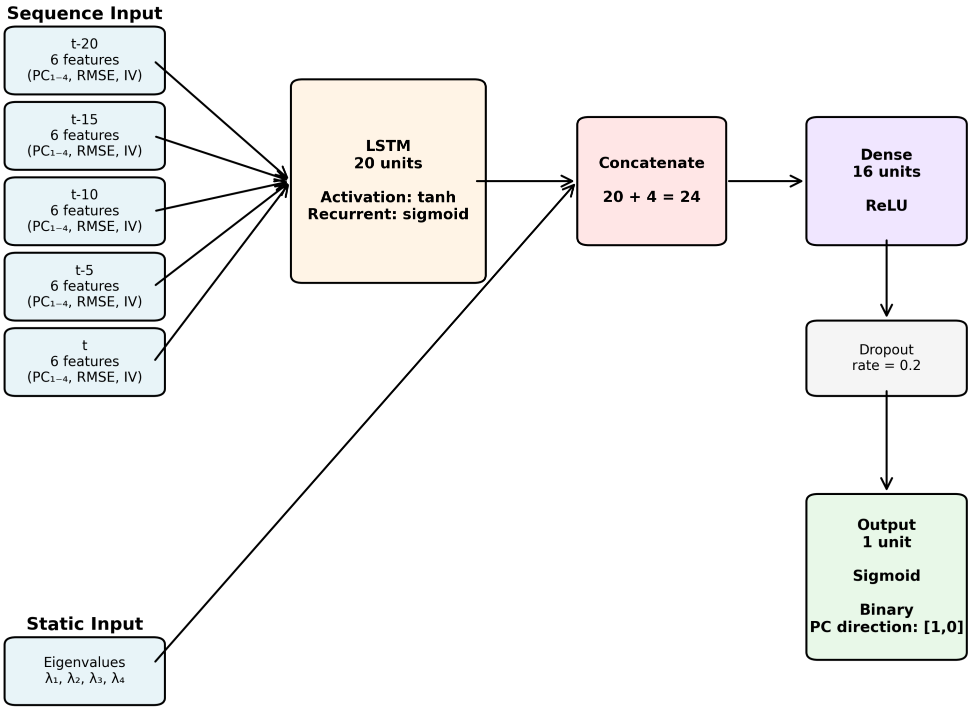


Figure 4: LSTM architecture for principal component direction forecasting.

The LSTM layer processes the sequential input $\{\mathbf{f}_{t-k}, \text{RMSE}_{t-k}, \text{IV}_{t-k}\}_{k\in\{0,5,10,15,20\}}$ through a series of gated computations. At each timestep s, the LSTM cell updates its hidden state $\mathbf{h}_s \in \mathbb{R}^{20}$ and cell state $\mathbf{c}_s \in \mathbb{R}^{20}$ via:

$$
\begin{aligned}
\mathbf{i}_s &= \sigma(\mathbf{W}_i\mathbf{x}_s + \mathbf{U}_i\mathbf{h}_{s-1} + \mathbf{b}_i) \quad \text{(input gate)} && (22)\\
\mathbf{f}_s &= \sigma(\mathbf{W}_f\mathbf{x}_s + \mathbf{U}_f\mathbf{h}_{s-1} + \mathbf{b}_f) \quad \text{(forget gate)} && (23)\\
\mathbf{o}_s &= \sigma(\mathbf{W}_o\mathbf{x}_s + \mathbf{U}_o\mathbf{h}_{s-1} + \mathbf{b}_o) \quad \text{(output gate)} && (24)\\
\tilde{\mathbf{c}}_s &= \tanh(\mathbf{W}_c\mathbf{x}_s + \mathbf{U}_c\mathbf{h}_{s-1} + \mathbf{b}_c) \quad \text{(candidate cell state)} && (25)\\
\mathbf{c}_s &= \mathbf{f}_s \odot \mathbf{c}_{s-1} + \mathbf{i}_s \odot \tilde{\mathbf{c}}_s \quad \text{(cell state update)} && (26)\\
\mathbf{h}_s &= \mathbf{o}_s \odot \tanh(\mathbf{c}_s) \quad \text{(hidden state)} && (27)
\end{aligned}
$$

where $\sigma(\cdot)$ denotes the sigmoid activation, $\tanh(\cdot)$ is the hyperbolic tangent, $\odot$ is element-wise multiplication, and $\mathbf{W}, \mathbf{U}, \mathbf{b}$ are learnable parameters. The final hidden state $\mathbf{h}_5 \in \mathbb{R}^{20}$

encodes the temporal information from all five timesteps.

This output is concatenated with the static eigenvalue features $\boldsymbol{\lambda}_t \in \mathbb{R}^4$ to form a 24-dimensional vector, which is then passed through a feedforward network:

$$\mathbf{z} = \text{ReLU}(\mathbf{W}_1[\mathbf{h}_5; \boldsymbol{\lambda}_t] + \mathbf{b}_1), \quad \mathbf{z} \in \mathbb{R}^{16} \tag{28}$$

$$\mathbf{z}' = \text{Dropout}(\mathbf{z}, p = 0.2) \tag{29}$$

$$\hat{y} = \sigma(\mathbf{w}_2^T \mathbf{z}' + b_2) \in [0, 1] \tag{30}$$

The final output $\hat{y}$ is interpreted as $P(y_{k,t} = 1 \mid \mathcal{X}_t)$, the probability that principal component k will increase over the forecast horizon, where $y_{k,t}$ is the binary direction indicator defined in equation (15). The prediction is classified as $\hat{y}_{k,t} = 1$ if $\hat{y} \geq 0.5$, and $\hat{y}_{k,t} = 0$ otherwise.

5.3.2 Training Methodology

The model is trained using a rolling walk-forward procedure that strictly maintains the temporal ordering of observations. For each prediction at time t:

1. The training window spans $[t - h - \tau, t - h)$ where $\tau = 1000$ trading days and $h = 21$ days is the forecast horizon

2. A fresh model is initialized with random weights (no parameter carry-over between windows)

3. Both sequence and static features are standardized using `StandardScaler` fitted on the training window

4. The model is optimized via binary cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)] \tag{31}$$

   using the Adam optimizer with learning rate $\alpha = 0.001$

5. Training proceeds in mini-batches of size 32 for a maximum of 50 epochs

6. Early stopping halts training if the loss does not improve for 3 consecutive epochs, restoring the best observed weights

7. The trained model predicts the direction at time t using features $\mathcal{X}_t$ scaled by the same training-window statistics

This procedure ensures that each prediction is fully out-of-sample: the model never observes data from time $t-h$ onward during training, and parameters are not influenced by information from future predictions.

5.4 Results

The LSTM model was trained independently for each of the four principal components using 5,080 out-of-sample predictions spanning from 15/03/2006 to 30/12/2025. Table 8 presents the directional forecast accuracy for each component.

Table 8: Overall directional accuracy by principal component

Component	Predictions	Accuracy	Up	Down
PC1	5,080	66.85%	2,525 (49.7%)	2,555 (50.3%)
PC2	5,080	65.93%	2,521 (49.6%)	2,559 (50.4%)
PC3	5,080	64.94%	2,498 (49.2%)	2,582 (50.8%)
PC4	5,080	66.44%	2,493 (49.1%)	2,587 (50.9%)

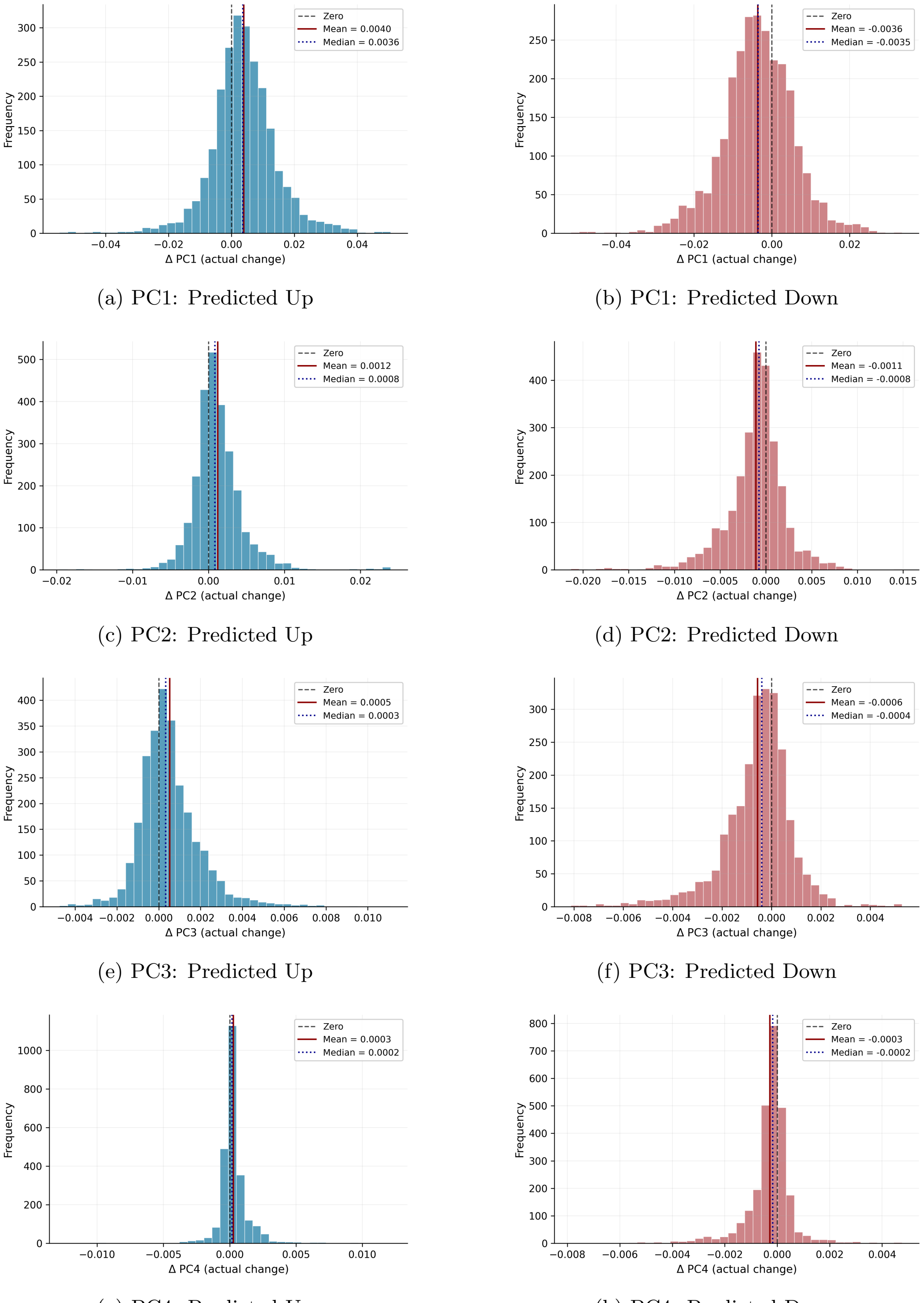
All four components achieve directional accuracy between 65% and 67%, substantially above the 50% baseline for balanced binary classification. The class distributions are approximately balanced across all components, confirming that the yield curve exhibits no systematic directional bias over the sample period.

Table 9 reports the conditional distribution of actual PC changes given the model's directional prediction. The mean values confirm that the model's predictions align with the realized direction: when the model predicts an increase, the average actual change is positive; when it predicts a decrease, the average actual change is negative.

Table 9: Conditional statistics of actual PC changes by predicted direction

PC	Pred	n	Mean	Median	Std	Q25	Q75	Min	Max
PC1	Up	2,395	0.00396	0.00363	0.01047	−0.00172	0.00946	−0.0546	0.0507
PC1	Down	2,685	−0.00360	−0.00353	0.00925	−0.00890	0.00226	−0.0516	0.0335
PC2	Up	2,554	0.00121	0.00083	0.00342	−0.00057	0.00261	−0.0196	0.0240
PC2	Down	2,526	−0.00108	−0.00075	0.00323	−0.00240	0.00057	−0.0213	0.0149
PC3	Up	2,645	0.00053	0.00034	0.00148	−0.00034	0.00119	−0.0047	0.0111
PC3	Down	2,435	−0.00057	−0.00039	0.00143	−0.00119	0.00022	−0.0081	0.0053
PC4	Up	2,446	0.00027	0.00017	0.00147	−0.00014	0.00056	−0.0124	0.0122
PC4	Down	2,634	−0.00028	−0.00018	0.00090	−0.00049	0.00008	−0.0079	0.0048

Figure 5 displays the empirical distributions of actual PC changes conditional on the model's prediction. The separation between the up and down distributions is visible across all components, with PC1 exhibiting the strongest separation and largest conditional means in absolute terms.



(a) PC1: Predicted Up

(b) PC1: Predicted Down

(c) PC2: Predicted Up

(d) PC2: Predicted Down

(e) PC3: Predicted Up

(f) PC3: Predicted Down

(g) PC4: Predicted Up

(h) PC4: Predicted Down

Figure 5: Conditional distributions of actual principal component changes by predicted direction.

6 Conclusion

References

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

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9:1735–1780.

Kondratyev, A. (2018). Learning curve dynamics with artificial neural networks. *SSRN Electronic Journal*, (3041232).

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

Nelson, C. R. and Siegel, A. F. (1987). Parsimonious modeling of yield curves. *The Journal of business (Chicago, Ill.)*, 60(4):473–489.

Suimon, Y., Sakaji, H., Izumi, K., and Matsushima, H. (2020). Autoencoder-based three-factor model for the yield curve of japanese government bonds and a trading strategy. *Journal of Risk and Financial Management*, 13(4).

Verner, R., Tkáč, M., and Tkáč, M. (2021). Improving quality of long-term bond price prediction using artificial neural networks. *Kvalita inovácia prosperita*, 25(1):103–123.