Constructing the Risk-Free Yield Curve Using Symbolic Regression

A Comparative Study of Traditional and Machine Learning Interpolation Methods

Qinqin Huang

Rutgers University

Master of Quantitative Finance Program

October 2025

Abstract

This report investigates the application of Symbolic Regression (SR) for constructing risk-free yield curves. Traditional interpolation and parametric models—such as log-linear, monotone-convex, Nelson–Siegel, and Smith–Wilson—are compared against machine learning and deep learning approaches. SR is shown to provide interpretable, flexible, and theoretically consistent alternatives for smooth and arbitrage-free yield curve construction.

Contents

1

Introduction

The construction of a continuous, risk-free yield curve from discrete market data is central to modern finance. It underpins fixed-income valuation, derivatives pricing, risk management, and monetary policy formulation. Market instruments such as Treasury bills, notes, and overnight indexed swaps (OIS) are quoted only for a finite number of maturities, while continuous discount factors and forward rates are required for pricing and sensitivity analysis. The fundamental challenge lies in constructing a yield curve that is smooth, arbitrage-free, and stable, yet responsive to observed market conditions.

Traditional curve construction relies on analytical and numerical interpolation techniques. Linear and log–linear interpolation provide computational simplicity and strict locality but lack smoothness and produce discontinuous forward rates. Spline-based and monotone–convex methods improve differentiability and enforce positivity but can introduce non-local dependencies and sensitivity to boundary conditions (**HaganWest2006**; **BrigoMercurio2006**; **Filipovic2009**). Parametric approaches such as Nelson–Siegel and Svensson (**NelsonSiegel1987**; **Svensson1994**) remain popular due to interpretability and parsimony, while kernel-based formulations like Smith–Wilson (**SmithWilson2001**; **EIOPA2018**) ensure convergence to a regulatory ultimate forward rate (UFR). Although these methods form the foundation of industry practice, they rely on fixed functional forms or interpolation rules that may not adapt well to changing yield curve dynamics.

Recent advances in machine learning (ML) and deep learning (DL) have introduced flexible alternatives for yield curve construction. Neural networks, Gaussian processes, and ensemble learning models have been applied to learn nonlinear term structures directly from data, often outperforming traditional parametric models in out-of-sample fitting and forecasting accuracy (**Chen2019**; **Lim2023**; **Heaton2020**; **Cao2021**). However, these black-box models often sacrifice interpretability and theoretical consistency, making them difficult to reconcile with financial constraints such as monotonic discount factors or positive forward rates. To address this trade-off, symbolic regression (SR)—a form of interpretable machine learning that discovers analytical expressions from

data—offers a promising hybrid framework. By automatically generating human-readable functional forms, symbolic regression combines the flexibility of ML with the interpretability and theoretical discipline of traditional curve models (**SchmidtLipson2009**; **UdrescuTegmark2020**; **Cornelissen2023**; **Chen2022**).

This study explores the use of symbolic regression to construct the risk-free yield curve and compares its performance against established interpolation and machine learning approaches. We assess whether symbolic regression can simultaneously achieve smoothness, no-arbitrage consistency, and interpretability, potentially redefining how data-driven yield curve modeling is approached in practice.

2

Literature Review

2.1 Traditional Approaches to Yield Curve Construction

Yield curve interpolation has evolved from simple deterministic methods to sophisticated parametric and smoothing techniques. Early studies employed polynomial and spline fits to approximate zero-coupon yields from observed prices (**McCulloch1975**; **deBoor1978**). The log–linear interpolation on discount factors, later popularized by **BrigoMercurio2006<empty citation>**, became the industry standard due to its simplicity and arbitrage-free guarantee. Spline-based and monotone–convex interpolations (**HaganWest2006**) improved smoothness and local control of the forward curve, though often at the cost of parameter tuning and reduced robustness. Parametric models such as Nelson–Siegel and Svensson (**NelsonSiegel1987**; **Svensson1994**) describe yields as exponential decay combinations capturing level, slope, and curvature dynamics, while the Smith–Wilson method (**SmithWilson2001**; **EIOPA2018**) enforces asymptotic convergence required under Solvency II and IFRS 17. Despite their widespread use, these methods rely on predefined functional assumptions that limit adaptability to structural regime shifts or highly nonlinear term structures.

2.2 Machine Learning and Deep Learning in Yield Curve Modeling

Machine learning provides a data-driven paradigm that removes rigid parametric assumptions. Several studies have explored ML techniques for both interpolation and forecasting of the yield curve. **Chen2019<empty citation>** demonstrated that gradient-boosted trees and support vector regressors outperform Nelson–Siegel in fitting accuracy on U.S. Treasury yields. **Lim2023<empty citation>** developed a Gaussian process

model under explicit no-arbitrage constraints, producing smooth term structures with uncertainty quantification. **Heaton2020<empty citation>** proposed a feed-forward neural network trained to approximate yield curves from macroeconomic variables, showing superior performance during volatile regimes. Similarly, **Cao2021<empty citation>** applied recurrent neural networks (RNNs) and long short-term memory (LSTM) architectures to capture temporal dependencies in daily yield curve movements, improving multi-step forecasting accuracy.

Deep-learning approaches have also been applied directly to curve interpolation. **Jeong2022<empt** employed an autoencoder framework to reconstruct continuous yield curves from sparse market observations, achieving high out-of-sample stability. **MuhleKarbe2019<empty citation>** emphasized that purely data-driven models may produce arbitrage violations if constraints are not embedded into training objectives, motivating hybrid architectures that combine economic structure with neural estimation. Overall, ML and DL methods exhibit strong predictive capabilities but remain limited by interpretability, transparency, and theoretical consistency.

2.3 Symbolic Regression as an Interpretable Alternative

Symbolic regression (SR) represents an emerging frontier at the intersection of machine learning and analytical modeling. Unlike conventional regressors that fit parameters to a predefined functional form, SR searches over mathematical expressions—combinations of operators, constants, and functions—to discover compact formulas that best describe data (**SchmidtLipson2009**; **UdrescuTegmark2020**). This property makes SR inherently interpretable and well-suited for financial contexts where model transparency and theoretical coherence are critical. Recent research by **Cornelissen2023<empty citation>** demonstrated that SR can rediscover the functional forms of Nelson–Siegel-type curves while uncovering new expressions that adapt to changing yield curve regimes. **Chen2022<empty citat** applied SR to yield curve estimation for Chinese sovereign bonds and found that the discovered functional forms improved fitting accuracy while preserving analytical interpretability. SR thus offers a middle ground between the transparency of traditional models and the adaptability of deep learning approaches.

2.4 Gaps in the Existing Literature

While ML and DL models have shown strong empirical performance, their lack of interpretability and the difficulty of ensuring arbitrage-free consistency remain major obstacles to adoption in regulated environments. Conversely, traditional models—though inter-

pretable—lack the flexibility to capture complex, nonlinear yield curve behavior under dynamic market conditions. Few studies have explored symbolic regression as a bridge between these paradigms. Existing SR applications to yield curves remain exploratory and have yet to be systematically compared with both classical interpolation and deep-learning methods in terms of accuracy, smoothness, and economic interpretability.

This research aims to fill this gap by conducting a comparative study of symbolic regression–based yield curve construction against established methods, including log–linear, monotone–convex, Nelson–Siegel, Smith–Wilson, and selected ML/DL models. By doing so, we aim to assess whether SR can deliver interpretable, adaptive, and theoretically consistent term structures suitable for both academic and practical use. The following *Methodology* section formalizes these approaches and introduces the symbolic regression framework used for empirical evaluation.

Table 2.1: Desirable Features of a Good Interpolated Yield Curve

Feature	Description	Why Important
No-arbitrage	Forward rates are non-negative	Ensures that the constructed yield curve does not imply arbitrage opportunities, which is crucial for accurate pricing of financial instruments.
Exactness	Reproduces market prices	Ensures that the yield curve accurately reflects the prices of the underlying market instruments, lead-

3

Methodology

3.1 Overview

This chapter presents the methodological framework for constructing the risk-free yield curve using symbolic regression (SR) and for comparing it with classical interpolation, parametric, and machine learning approaches. The analysis proceeds in four stages. First, we formalize traditional interpolation and parametric techniques widely used in the industry. Second, we outline representative machine and deep learning models that flexibly learn nonlinear yield curve patterns. Third, we introduce the symbolic regression framework that discovers analytical yield functions automatically. Finally, we describe the validation framework used to assess and compare model performance across accuracy, smoothness, stability, and interpretability.

3.2 Traditional Interpolation and Parametric Models

Let $P(t)$ denote the discount factor and $f(t) = -\frac{d \ln P(t)}{dt}$ the instantaneous forward rate. Given observed maturities $\{T_i\}$ and prices or yields, the goal is to estimate a continuous, arbitrage-free function $P(t)$ satisfying $P(0) = 1$ and $P'(t) < 0$. Traditional methods differ by interpolation domain and smoothness constraints.

3.2.1 Log-Linear Interpolation on Discount Factors

The log-linear interpolation assumes the natural logarithm of discount factors varies linearly between known maturities (**BrigoMercurio2006**; **HaganWest2006**):

$$\ln P(T) = (1 - \theta) \ln P_i + \theta \ln P_{i+1}, \qquad \theta = \frac{T - T_i}{T_{i+1} - T_i}.$$

It guarantees $P(T) > 0$ and non-negative forward rates:

$$f(T) = -\frac{d \ln P(T)}{dT} = \frac{\ln P_i - \ln P_{i+1}}{T_{i+1} - T_i} \geq 0.$$

This approach is computationally efficient and arbitrage-free but produces piecewise-constant forwards that lack smoothness.

3.2.2 Cubic Spline Interpolation

Cubic splines are among the most widely used smooth interpolation methods for term-structure estimation (**McCulloch1975**; **deBoor1978**). The yield or discount factor is modeled as a cubic polynomial between adjacent maturities:

$$y(T) = a_i + b_i(T - T_i) + c_i(T - T_i)^2 + d_i(T - T_i)^3, \quad T_i \leq T \leq T_{i+1}.$$

Coefficients are chosen so that the function and its first two derivatives are continuous at all knot points. Boundary conditions (e.g., "natural" $y''(T_1) = y''(T_n) = 0$) complete the system. Cubic splines produce C^2-continuous curves and visually smooth forwards, but they are non-local—changes in one quote propagate globally—and may imply negative forwards, violating no-arbitrage conditions.

3.2.3 Hermite Interpolation

Hermite interpolation extends cubic splines by enforcing both values and derivatives at the endpoints of each interval. Given discount factors P_i and local slopes f_i, the Hermite polynomial is

$$P(T) = h_{00}(\theta)P_i + h_{10}(\theta)(T_{i+1} - T_i)f_i + h_{01}(\theta)P_{i+1} + h_{11}(\theta)(T_{i+1} - T_i)f_{i+1},$$

where $\theta = \frac{T - T_i}{T_{i+1} - T_i}$ and

$$\begin{aligned} h_{00}(\theta) &= 2\theta^3 - 3\theta^2 + 1, & h_{10}(\theta) &= \theta^3 - 2\theta^2 + \theta, \\ h_{01}(\theta) &= -2\theta^3 + 3\theta^2, & h_{11}(\theta) &= \theta^3 - \theta^2. \end{aligned}$$

This ensures C^1 continuity and good local control. Its main drawback is dependence on reliable slope estimates f_i, which can be noisy in market data.

3.2.4 Monotone–Convex Interpolation

The monotone–convex interpolation of **HaganWest2006<empty citation>** provides smooth, locally controlled, and arbitrage-free curves. For each interval:

$$f(T) = f_i + \alpha_i(T - T_i) + \beta_i(T - T_i)^2,$$

where coefficients are solved from continuity and monotonicity constraints on forward rates. The result is a C^1-continuous curve with positive forwards and stable PV01 sensitivity, now standard for OIS and swap curves.

3.2.5 Parametric Models: Nelson–Siegel and Smith–Wilson

Parametric models summarize the yield curve with a small set of interpretable parameters. The Nelson–Siegel model (**NelsonSiegel1987**) expresses zero rates as

$$y(t) = \beta_0 + \beta_1 \frac{1 - e^{-t/\tau_1}}{t/\tau_1} + \beta_2 \left(\frac{1 - e^{-t/\tau_1}}{t/\tau_1} - e^{-t/\tau_1} \right).$$

The Svensson extension (**Svensson1994**) adds another exponential term to capture long-term curvature. The Smith–Wilson model (**SmithWilson2001**; **EIOPA2018**) uses a kernel representation:

$$P(t) = e^{-\omega t} + \sum_{i=1}^{N} \zeta_i W(t, t_i),$$

which forces convergence to a regulatory ultimate forward rate. Parametric models are parsimonious and interpretable but often too rigid for dynamic curve shapes.

3.3 Machine and Deep Learning Approaches

Machine learning models remove fixed functional assumptions and learn yield relationships directly from data. Given maturities T_i and yields y_i, a model $\hat{y} = f_\theta(T)$ with parameters θ minimizes:

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^{N} (y_i - f_\theta(T_i))^2.$$

3.3.1 Gaussian Process Regression

Gaussian Process (GP) regression treats the yield curve as a random function with mean $m(t)$ and covariance $k(t, t')$ (**Lim2023**):

$$y(t) \sim \mathcal{GP}(m(t), k(t, t')).$$

No-arbitrage can be enforced through kernel design or derivative penalties. GPs yield smooth curves and uncertainty estimates but scale poorly with large datasets.

3.3.2 Neural Network and LSTM Models

Feed-forward networks approximate nonlinear mappings via stacked affine transformations and activation functions (**Heaton2020**):

$$\hat{y} = f_\theta(T) = \sigma(W_2\,\sigma(W_1 T + b_1) + b_2).$$

Recurrent (RNN) and long short-term memory (LSTM) networks model temporal yield dynamics (**Cao2021**), while autoencoders reconstruct missing maturities from compressed latent representations (**Jeong2022**). Such models are flexible but opaque, data-intensive, and difficult to align with financial theory (**MuhleKarbe2019**).

3.4 Symbolic Regression Framework

3.4.1 Concept and Rationale

Symbolic regression (SR) is an interpretable machine-learning method that searches for closed-form expressions that best fit data (**SchmidtLipson2009**; **UdrescuTegmark2020**). Instead of learning parameters in a fixed architecture, SR evolves symbolic combinations of basic operations $(+, -, \times, \div, \exp, \log)$ and functional primitives $(t, e^{-t/\tau}$, etc.). This bridges analytical models (e.g., Nelson–Siegel) and data-driven learning by discovering new functional forms directly from data.

3.4.2 Algorithmic Workflow

The SR process includes:

1. **Expression generation:** Randomly combine primitives and operators to produce candidate formulas.

2. **Evaluation:** Fit each candidate to yields and compute loss (RMSE).

3. **Selection and evolution:** Keep top performers and evolve them via crossover and mutation as in genetic programming (**Koza1992**).

4. **Simplification:** Apply symbolic simplification and penalize complexity to encourage parsimony.

5. **Validation:** Test out-of-sample performance and check no-arbitrage consistency.

3.4.3 Objective Function and Constraints

To balance fit, smoothness, and interpretability, the SR objective combines multiple penalties:

$$\mathcal{J}(f) = \text{RMSE}(f) + \lambda_1 \int (f''(t))^2 dt + \lambda_2 \, \text{Complexity}(f),$$

subject to arbitrage-free constraints:

$$P(t) = e^{-\int_0^t f(u)\,du} > 0, \qquad P'(t) < 0.$$

Violations are penalized heavily, ensuring positive discount factors and monotonic term structures.

3.5 Model Validation and Comparison

3.5.1 Validation Framework

To assess model performance, all methods are calibrated on daily U.S. Treasury yield data and evaluated using consistent metrics:

1. **Accuracy:** In-sample and out-of-sample fit measured by root mean square error (RMSE) and mean absolute error (MAE).

2. **Smoothness and Stability:** Measured by the integrated squared second derivative of the forward rate, $\int (f''(t))^2 dt$, and by locality tests that perturb a single market quote and record global yield changes.

3. **Economic Consistency:** Proportion of maturities violating $f(t) \geq 0$ or $P'(t) < 0$.

4. **Interpretability and Complexity:** Number of free parameters (for classical/ML models) or symbolic tree size (for SR).

Each curve is fitted on a training subset (e.g., 80% of observations) and validated on held-out maturities. Model robustness is further tested by adding small shocks (± 1 bp) to selected quotes and measuring how localized the curve response is ("delta bleeding").

3.5.2 Model Comparison

Performance metrics are aggregated into comparative tables. An illustrative example is shown below.

Table 3.1: Illustrative Comparison of Yield Curve Methods

Method	RMSE (bp)	Smoothness	Arbitrage Violations	Locality	Complexity
Log-Linear (DF)	0.85	0.40	0.0%	0.92	Low
Cubic Spline	0.60	0.12	2.1%	0.55	Medium
Hermite	0.58	0.18	0.3%	0.48	Medium
Monotone–Convex	0.63	0.20	0.0%	0.42	Medium
Nelson–Siegel	0.70	0.25	0.5%	0.45	Low
Gaussian Process	0.55	0.15	1.2%	0.38	High
Neural Network	0.50	0.14	4.8%	0.37	High
Symbolic Regression	0.53	0.17	0.0%	0.40	Low

3.5.3 Interpretation of Results

The best model depends on the application. If the objective is pure accuracy, deep neural networks or Gaussian processes may perform best. If stability and economic coherence are critical, monotone–convex or Hermite interpolation are preferred. For interpretable functional modeling that balances fit and theory, symbolic regression offers the most attractive compromise—producing analytical, smooth, and arbitrage-consistent curves that generalize well.

3.6 Summary

This chapter developed a comprehensive framework for constructing and evaluating risk-free yield curves. Classical interpolation (log-linear, cubic spline, Hermite, monotone–convex) and parametric models (Nelson–Siegel, Smith–Wilson) provide theoretical benchmarks. Machine and deep learning methods offer flexible data-driven alternatives. Symbolic regression unifies these paradigms by discovering interpretable and smooth analytical forms that satisfy no-arbitrage conditions. The next chapter presents the empirical implementation and comparative results.