Optimization
A Journal of Mathematical Programming and Operations Research

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/gopt20

Revisiting the fitting of the Nelson–Siegel and Svensson models

Dirk Banholzer, Jörg Fliege & Ralf Werner

To cite this article: Dirk Banholzer, Jörg Fliege & Ralf Werner (2024) Revisiting the fitting of the Nelson–Siegel and Svensson models, Optimization, 73:10, 3021-3053, DOI: 10.1080/02331934.2024.2389242

To link to this article: https://doi.org/10.1080/02331934.2024.2389242

Published online: 26 Aug 2024.

OPTIMIZATION
2024, VOL. 73, NO. 10, 3021–3053
https://doi.org/10.1080/02331934.2024.2389242

Taylor & Francis
Taylor & Francis Group

Revisiting the fitting of the Nelson–Siegel and Svensson models

Dirk Banholzer[a], Jörg Fliege[a] and Ralf Werner[b]

[a]Mathematical Sciences, University of Southampton, Southampton, UK; [b]Institut für Mathematik, Universität Augsburg, Augsburg, Germany

ABSTRACT

The Nelson–Siegel and the Svensson models are two widely used models for the term structure of interest rates. These models are quite simple and intuitive, but fitting them to market data is numerically challenging and various difficulties have been reported. In this paper, we provide a novel mathematical analysis of the fitting problem based on parametric optimization. We formulate the fitting problem as a separable nonlinear least-squares problem, in which the linear parameters can be eliminated. We provide a thorough discussion on the conditioning of the inner part of the reformulated problem and show that many of the reported difficulties encountered when solving it are inherent to the problem formulation itself and cannot be tackled by choosing a particular optimization algorithm. Our stability analysis provides novel insights that we use to show that some of the ill-conditioning can be avoided, and that a suitably chosen penalty approach can be used to address the remaining ill-conditioning. Numerical results indicate that this approach has the expected impact while being independent of any choice of a particular optimization algorithm. We further establish smoothness properties of the reduced objective function, putting global optimization methods for the reduced problem on a sound mathematical basis.

ARTICLE HISTORY
Received 1 June 2023
Accepted 17 July 2024

KEYWORDS
Parametric optimization; stability analysis; Nelson–Siegel model; Svensson model; yield curve fitting

2020 MATHEMATICS SUBJECT CLASSIFICATIONS
49-02; 49-04; 91-02; 91-08; 91G30

1. Introduction

Due to their appealing features, the model of Nelson–Siegel [1] and its extension by Svensson [2] have become very popular with practitioners in the financial industry to represent the term structure of interest rates. By means of simple parametric functions that rely on few parameters only, both models are parsimonious and yet able to capture the shapes of most of the observed term structures of interest rates in the market. Their extensive popularity is reflected by the fact that they are widely employed by financial institutions, e.g. by national banks (see, e.g. [3]). In particular, Svensson's extension is used on a daily basis by

CONTACT Ralf Werner ralf.werner@math.uni-augsburg.de

the European Central Bank, see Ref. [4], and the Deutsche Bundesbank, see Ref. [5], to model yield curves constructed from market data. Further applications can be found in the recent paper [6]. In addition, also quite recently, deep learning approaches–specifically autoencoders–have been successfully applied to yield curve data. Since the Nelson–Siegel and Svensson models represent an approach strongly related to such kind of dimension reduction techniques, renewed interest in these models has arisen from a data science perspective, cf. especially [7,8].

To apply the Nelson–Siegel and the Svensson models in practice they need to be made consistent with observed data, i.e. their parameters need to be chosen such that model rates best match given (market) rates. This optimization procedure, frequently called *fitting*, however, bears several numerical difficulties, as reported, for instance, in Refs [9–12]; see also Ref. [6] for further references. Both models are highly nonlinear and non-convex so that the objective function, usually some kind of root- or mean-square error, contains multiple local minima. Moreover, it is well-known that the models suffer from severe multi-collinearity in certain regions of the parameter space.

To avoid non-convexity, a popular yet straightforward technique that has been adopted by some authors is to provide predetermined values to parameters that appear in a nonlinear fashion in the model and to use ordinary linear least-squares methods to obtain the remaining parameters (see Refs e.g. [1,10,12–14]). Some of these approaches only consider one pre-specified value (where the value is based e. g. on economic reasoning or hindsight) which, however, limits the models and reduces some of their flexibility in reproducing different types of curves. Other approaches consider several potential values for the nonlinear parameter(s) and can typically be either classified as a grid search method or as a (heuristic) global optimization method for the reduced global optimization problem. However, so far no analysis has been provided that puts such methods on a sound mathematical basis. In particular, the literature lacks continuity and/or smoothness results for the reduced objective function.

To avoid multi-collinearity, some authors have suggested to not consider regions of the nonlinear parameter space which lead to such multi-collinearity (see, for instance, Ref. [15]), while others have proposed to tackle the fitting problem by a suitable choice of optimization algorithm (e.g. Ref. [6] suggest a genetic algorithm). We will argue that while the first approach is reasonable, it still bears some difficulties. For the second approach, we will prove that the issue of multi-collinearity is an inherent aspect of the problem and cannot be addressed choosing the optimization algorithm in a particular way. Note that although this result was already indicated in Ref. [15], no formal proofs nor a mathematical precise analysis were given.

In essence, although the main difficulties in fitting these models have been recognized in various sources, no fully satisfying analysis nor remedy has been presented in the literature so far.

In view of these findings, we propose a novel mathematically rigorous analysis of fitting Nelson–Siegel and Svensson models. Given that both models are linear combinations of specific nonlinearly parameterized basis functions, it is well-known that the problem of matching model rates to given rates can be formulated as a separable nonlinear least-squares problem. In particular, this allows to express the linear model parameters as an ordinary linear least-squares solution that depends on the nonlinear parameters. On substituting the optimal solution into the original objective function, we arrive at an at most two-dimensional non-convex and potentially non-differentiable optimization problem in the nonlinear parameters only. This reduction is the basis for our analysis and quantification of the ill-conditioning of the problem; an analysis which has not been carried out before in such a way in the literature. Accordingly, by means of such a stability analysis, we can exactly identify the regions of the parameter space that lead to the ill-conditioning and consequently untrustworthy values. After demonstrating that some of this ill-conditioning can be avoided by adding more short and/or long term tenors to the fitting problem, we then argue that the most reliable and efficient way to address the remaining ill-conditioning is by penalizing the reduced objective function, where the parameters of the penalization can be adjusted as to yield sufficient stability in the linear parameters. One of the main findings of our analysis is that if the optimal solution is obtained in a region with a high condition number this can be interpreted as an over-specification of the model for the data at hand. As our subsequent analysis will show, this ill-conditioning is caused in full by high collinearity of the basis functions. For the first time, this also shows in a rigorous way that the model parameters cannot be properly identified (due to offsetting effects) – independent of the optimization algorithm employed.

A further interesting result is that the reduced objective function is smooth, and thus Lipschitz continuous, in a large compact set containing the global optimum, which for the first time puts global optimization methods for the reduced problem on a sound mathematical basis.

The remainder of this paper is structured as follows. In Section 2, we briefly review the modelling framework of the Nelson–Siegel and the Svensson models. In Section 3, we describe the traditional fitting procedure of these models and show how the (partial) linear structure of the models can be exploited, while in Section 4, we provide a thorough analysis of the inherent ill-conditioning of the problem and present an approach to solve the fitting problems by means of penalization. In Section 5, we support our theoretical findings with a brief computational study. Finally, Section 6 provides our conclusions.

2. Model specification

Let us start by mentioning that both the Nelson–Siegel and the Svensson model have their thorough foundations in interest rate theory, see, for instance, Ref. [12],

where more details on the models can be found. For the purpose of this paper, it suffices though to assume that some kind of rates $y(\tau)$ (e.g. zero rates, swap rates, CDS spreads, etc.) are given for selected maturities $\tau \in [0, T]$, with fixed horizon date $T > 0$, which we want to approximate by either the Nelson–Siegel or the Svensson family of functions. This rather pragmatic point of view is also employed, e.g. by Sokol [8].

2.1. Nelson–Siegel model

In Ref. [1], the following model curve is proposed

$$y_{\lambda,\beta}(\tau) = \beta_1 + \beta_2 \left(\frac{1 - e^{-\lambda_1 \tau}}{\lambda_1 \tau} \right) + \beta_3 \left(\frac{1 - e^{-\lambda_1 \tau}}{\lambda_1 \tau} - e^{-\lambda_1 \tau} \right), \tag{1}$$

where $\beta_1, \beta_2, \beta_3 \in \mathbb{R}$ denote the linear coefficients and $\lambda_1 \geq 0$ the shape parameter.

Although Nelson and Siegel's model is quite simple, it can assume a variety of shapes depending on the four parameters which have a clear interpretation: β_1 describes the long rate, the sum $\beta_1 + \beta_2$ accounts for the short rate, and β_3 and λ_1 determine the height and position of the hump of the curve, respectively.

2.2. Svensson model

To allow for an even greater flexibility in the curves and to improve the fit, the author [2] proposes to extend Nelson and Siegel's model by adding a further term. Svensson's extension often provides a better fit to long maturities than the Nelson–Siegel model, see, e.g. [16]. The corresponding model curve is given as

$$\begin{aligned} y_{\lambda,\beta}(\tau) = \beta_1 + \beta_2 \left(\frac{1 - e^{-\lambda_1 \tau}}{\lambda_1 \tau} \right) + \beta_3 \left(\frac{1 - e^{-\lambda_1 \tau}}{\lambda_1 \tau} - e^{-\lambda_1 \tau} \right) \\ + \beta_4 \left(\frac{1 - e^{-\lambda_2 \tau}}{\lambda_2 \tau} - e^{-\lambda_2 \tau} \right). \end{aligned} \tag{2}$$

with $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$.

Unlike other authors, we do not impose any restrictions on the linear parameters $\beta_1, \beta_2, \beta_3$ and β_4 at this point. This is justified by the fact that for example interest rates may well become negative, as developments in financial markets have shown, see, e.g. [17].

3. Fitting of model curves

The aim of a fitting procedure is to determine model parameters such that they best match available data. Fitting can thus be seen as defining an optimization problem, of which several different variants exist, and of choosing and executing an optimization algorithm. Different variants of fitting differ from each

other in the formulation of the objective function as well as in the choice of the optimization algorithm used to solve the problem.

3.1. Traditional approach

3.1.1. General setup

Given the descriptions of the Nelson–Siegel and the Svensson models in (1) and (2), respectively, the model curves[1] can be expressed as

$$y_{\lambda,\beta}(\tau) = \sum_{j=1}^{l} \beta_j \phi_j(\lambda;\tau), \tag{3}$$

where the continuously differentiable basis functions ϕ_j have the form

$$\phi_1(\lambda;\tau) = 1, \qquad \phi_2(\lambda;\tau) = \frac{1-e^{-\lambda_1\tau}}{\lambda_1\tau},$$
$$\phi_3(\lambda;\tau) = \frac{1-e^{-\lambda_1\tau}}{\lambda_1\tau} - e^{-\lambda_1\tau}, \quad \phi_4(\lambda;\tau) = \frac{1-e^{-\lambda_2\tau}}{\lambda_2\tau} - e^{-\lambda_2\tau}.$$

Using (3) and letting $0 < \tau_1 < \ldots < \tau_m \leq T$ denote some set of predefined maturities at which given rates $\hat{y}_1, \ldots, \hat{y}_m \in \mathbb{R}$ are available, the fitting of the Nelson–Siegel and the Svensson models to these rates can then be described[2] in the least-squares sense as solving the problem

$$\min_{\lambda\in\Lambda,\,\beta\in\mathbb{R}^l} \left\{ F(\lambda,\beta) := \left\| \Phi(\lambda)\beta - \hat{y} \right\|_2^2 \right\}, \tag{4}$$

where $\beta \in \mathbb{R}^l$ and $\lambda \in \Lambda$ are unknown parameters, $\Lambda \subset \mathbb{R}_+^k$ is without loss of generality a closed set of positive k-dimensional real numbers (see Section 3.2.2 for the relevant reasoning), $\Phi(\lambda) \in \mathbb{R}^{m\times l}$ denotes the matrix of basis functions with entries $\{\Phi(\lambda)\}_{i,j} = \phi_j(\lambda;\tau_i)$, $i = 1,\ldots,m$, $j = 1,\ldots,l$, and $\hat{y} = (\hat{y}_1,\ldots,\hat{y}_m)^\top \in \mathbb{R}^m$ presents the vector of given rates. We further assume that $m > l + k$ holds, i.e. the number of maturities is greater than the dimension of the problem so that (4) defines an overdetermined least-squares problem with more observations than unknowns.

Note that we use throughout the convention

$$\phi_2(0;\tau) := 1 = \lim_{\lambda_1\to 0^+} \phi_2(\lambda;\tau) \quad \text{and}$$

[1] For $l = 3$, $\lambda \in \mathbb{R}_+$, we obtain the Nelson–Siegel model, whereas for $l = 4$, $\lambda \in \mathbb{R}_+^2$, we obtain the Svensson model. To allow for larger values of l, further basis functions would need to be defined. Our subsequent analysis is specifically tailored for the Nelson–Siegel and Svensson models, but can partially be generalized to larger models as well.

[2] To take care of potential heteroscedasticity in the data, positive weights may additionally be included in the objective function F. The latter may then be formulated as $F(\lambda,\beta) := \|W(\Phi(\lambda)\beta - \hat{y})\|_2^2$, where the elements of the weight matrix $W = \operatorname{diag} w_1,\ldots,w_l$ are typically set equal to the reciprocals of the variances of the residuals, which may be estimated, for instance, from historical data (e.g. Ref. [18]). However, for ease of exposition, we do not use any weights in our analysis and the numerical calculations.

$$\phi_3(0;\tau) := 0 = \lim_{\lambda_1 \to 0^+} \phi_3(\lambda;\tau),$$

where both limits are understood in the supremum norm on $C([0, T])$ (i.e. limits are uniform in τ).

Further, note that other possibilities for modeling the fitting problem exist as well. For instance, one could use other functions than the sum-of-squares to measure the fitting error, such as the ∞-norm, the 1-norm, or any monotone transformation thereof. While the main idea of our analysis still remains valid for these formulations, our analysis exploits the special structure of the given optimization problem. Different variants will require different definitions of a condition number of the inner problem, which might lead to a much more involved analysis. Let us point out here that (4) appears to represent the most popular formulation and is also used in other contexts, e.g. training autoencoders [8].

3.1.2. Review of existing approaches

To deal with the numerical difficulties involved in the fitting, several different approaches have been presented in the literature which will be reviewed in the following.

In order to avoid solving a non-convex least-squares problem, the idea of splitting the problem and employing a grid search on the reduced problem was already proposed in the original paper [1]: consider the shape parameter λ_1 only on a finite grid of different values in a reasonable interval and estimate for each of these values the remaining parameter β by solving a linear least-squares problem. The optimal solution among all sets of parameters was then chosen as the one yielding the highest *coefficient of determination* R^2.

The idea of fixing the shape parameter at pre-specified values in the estimation of the Nelson–Siegel model was adopted by Fabozzi et al. [10], Diebold and Li [13], De Pooter [12], and other authors, albeit with a different strategy for the choice of λ_1. Considering that the shape parameter determines the position of the hump of the zero rate curve, the literature [10,13] fixed the parameter λ_1 in the Nelson–Siegel model in such a way that the maximum of the zero rate curve was attained for different sets of data at a maturity of 5.38 and 2.5 years, respectively. The latter value was also used by De Pooter [12] for his data set. While the reason for these particular choices was motivated by historical observations, the author [14] fixed the nonlinear parameter in hindsight at a value which provided the lowest fitting error over the time horizon considered. By setting the nonlinear parameter to a single pre-specified value, some of the numerical problems can be resolved. However, a significant amount of flexibility of the models is lost this way. In particular, no such strategy guarantees that the fixed parameter is suitable, let alone optimal, for all individual curves. Moreover, this simple strategy has only been applied to fitting the Nelson–Siegel model. In the case of the Svensson model, two nonlinear parameters λ_1 and λ_2 would need to be fixed at adequate values, which is a much more demanding task. We want to emphasize

that the original grid search idea proposed in Ref. [1] is at present merely a heuristic approach, as no mathematical rigorous reasoning has been given as to why such a strategy should be able to approximate the true global optimum.

If separability is ignored and all parameters are estimated simultaneously, the corresponding optimization problem (4) is non-convex and may thus have several local minima. Unsurprisingly, using nonlinear optimization techniques, various authors hence have noted that the success crucially depends on the choice of the initial values, see e.g. Ref. [19] for the Nelson–Siegel model as well as [12,20] for both models. To mitigate the danger of getting stuck in a local optimum, [12] suggested to carefully choose the initial values by applying the above strategy of fixing the shape parameter. In Ref. [20], it is indicated that it would be necessary to run any local optimization algorithm from many different initial values and therefore suggest a multi-start framework in which they run a local optimizer for (4) from a selected subset of randomly generated points.

The difficulties in fitting are further elevated by the potential multi-collinearity in the models, as analysed by De Pooter [12], Gilli et al. [15], and Annaert et al. [21], for instance. It is pointed out in Ref. [12] that the linear parameter estimates β are sensitive to the choice of the shape parameter λ and that the fitting procedure as given via (4) can result in optimal parameter sets that lead to a very good fit but include extreme values, especially for the Svensson model[3]. Since the degree of multi-collinearity seemed to be influenced only by the nonlinear parameter λ (apart from maturities), the most common technique for preventing multi-collinearity is to restrict its parameter space in an appropriate manner, see Refs [12,15]. Given the economical interpretation of the shape parameter, the author [12] constrained the parameter in both models to lie in a small interval that implies that the humps of the resulting zero rate curves are between one and five years of maturity for a data set with up to ten years of maturity. Similarly, to avoid the case in the Svensson model where λ_1 and λ_2 lie too close together, he restricted λ_2 so that the second hump occurs at a maturity which is at least one year shorter than the first hump. In contrast to the above interpretation of the shape parameter, the study [15] discussed multi-collinearity in the Nelson–Siegel and the Svensson models (albeit not in a completely rigorous fashion) and constrained the range of λ to those values that yield factor loadings that are not too highly correlated. If factors become too highly correlated, many different parameter sets typically have very similar objective function values so that the factors can no longer be uniquely identified. Nevertheless, their final restriction of the parameter space excludes regions that may contain a potential global minimum with moderately correlated factor loadings. An approach different from restricting the parameter space was chosen by Annaert et al. [21], who improved the suggested grid search of Nelson–Siegel by a ridge regression

[3] An example was provided by Gimeno and Nave [22] for fitting the Svensson model, albeit with a different objective function. They reported that extreme and often offsetting optimal values for the linear parameters β_3 and β_4 occur whenever the corresponding nonlinear parameters λ_1 and λ_2 are similar to each other.

to stabilize the estimated parameters and hence prevent multi-collinearity. More precisely, for an optimal nonlinear parameter λ_1^* causing high collinearity, they iteratively re-estimated the corresponding linear coefficients until the condition number of the modified linear least-squares problem falls below a given threshold. The approach is extendable to the Svensson model in a straightforward manner. The main disadvantage of this approach is that the changes in the linear parameters might result in a significant deterioration of the model fits.

In what follows, we provide a novel analysis supplementing existing approaches for fitting both the Nelson–Siegel and the Svensson models. As already mentioned, this analysis is based on the observation that the corresponding optimization problem can be reformulated as a separable nonlinear least-squares problem, which allows to avoid collinearity issues substantially and which renders the global optimization problem computationally tractable as its dimension is reduced significantly. Even though the special structure of the objective function was already recognized by Angelini and Herzel [23] and Gauthier and Simonato [20], no theoretical justification in the sense of our Theorem 3.1 was provided, not to mention the subsequent implications on the treatment of the ill-conditioning of the inner problem.

3.2. Dimension reduction in fitting models

3.2.1. Main idea

Since the model rates $y_{\lambda,\beta}$ in both the Nelson–Siegel and the Svensson model are expressed as a linear combination of nonlinear basis functions in which the parameters λ and β occur independently, cf. formula (3), the original minimization problem (4) evidently presents a *separable nonlinear least-squares problem*, see e.g. Ref. [24], Section 9.4.

Hence, for any given $\lambda \in \Lambda$, some optimal linear parameter $\beta^* = \beta^*(\lambda)$ will always exist and can be obtained by solving the standard linear least-squares problem

$$\min_{\beta \in \mathbb{R}^l} F(\lambda, \beta), \tag{5}$$

for fixed $\lambda \in \Lambda$. Its solution is given by

$$\beta^*(\lambda) = \Phi(\lambda)^\dagger \hat{y}, \tag{6}$$

where $\Phi(\lambda)^\dagger$ denotes the Moore-Penrose pseudoinverse of $\Phi(\lambda)$, see, e.g. [24], Sections 1.1.4 and 1.2.5. Note that β^* solves (5) if and only if β^* satisfies the *normal equations* of (5):

$$\Phi(\lambda)^\top \Phi(\lambda) \beta^* = \Phi(\lambda)^\top \hat{y}.$$

Accordingly, if the columns of $\Phi(\lambda)$ are linearly independent, i.e. $\operatorname{rank}(\Phi(\lambda)) = l$, the unique least squares solution is given by $\beta^*(\lambda) = (\Phi(\lambda)^\top \Phi(\lambda))^{-1} \Phi(\lambda)^\top \hat{y}$.

If $\text{rank}(\Phi(\lambda)) < l$, the least squares solution $\beta^*(\lambda)$ is not unique, and any such solution has the same residual $\Phi(\lambda)\beta^*(\lambda) - \hat{y}$. In this case, the Moore-Penrose pseudoinverse assigns the solution with minimum length $\|\beta^*(\lambda)\|_2$, which is uniquely defined.

On substituting the optimal solution into the objective function F, the original problem (4) can be decomposed into an outer and inner optimization problem

$$\min_{\substack{\lambda\in\Lambda\\ \beta\in\mathbb{R}^l}} F(\lambda,\beta) = \min_{\lambda\in\Lambda} \underbrace{\min_{\beta\in\mathbb{R}^l} F(\lambda,\beta)}_{=:H(\lambda)} = \min_{\lambda\in\Lambda} H(\lambda), \tag{7}$$

where the objective function H takes the semi-analytic form

$$H(\lambda) = F\left(\lambda,\beta^*(\lambda)\right) = \left\|\Phi(\lambda)\Phi(\lambda)^\dagger\hat{y} - \hat{y}\right\|_2^2, \tag{8}$$

in which the linear parameter β has been eliminated.

The outer problem (7) is a non-convex optimization problem in the nonlinear parameter $\lambda \in \Lambda$. For each function evaluation of the objective function H in (8), the inner problem (5) needs to be solved which represents an unconstrained low-dimensional linear least-squares problem in the parameter β. Once the optimal nonlinear parameter λ^* has been obtained by solving the outer problem (7), the unique corresponding optimal linear parameter $\beta^*(\lambda^*)$ can be derived via Equation (6).

3.2.2. Theoretical justification

The justification for employing the proposed technique is given by the following Theorem 3.1. For a proof of Theorem 3.1, see Ref. [25], Theorem 2.1. This result establishes a strong relationship between critical points of the original objective function F and the new objective function H, as well as between their global minimizers.

Theorem 3.1: *Assume that in the open set Ω, the matrix of basis functions $\Phi(\lambda)$ has constant rank $0 < q \leq l$.*

(a) *If λ^* is a critical point, resp. a global minimizer, of H in Ω and*

$$\beta^* = \Phi(\lambda^*)^\dagger\hat{y}, \tag{9}$$

then (λ^*,β^*) is a critical point, resp. a global minimizer, of F in $\Omega\times\mathbb{R}^l$ and $F(\lambda^*,\beta^*) = H(\lambda^*)$.

(b) *If (λ^*,β^*) is a global minimizer of F in $\Omega\times\mathbb{R}^l$, then λ^* is a global minimizer of $H(\lambda)$ in Ω and $H(\lambda^*) = F(\lambda^*,\beta^*)$. Furthermore, if there is a unique β^* among the minimizing pairs of F, then β^* must satisfy (9).*

We note that the equivalence between the critical points of both objective functions relies on the assumption that the rank of the matrix $\Phi(\lambda)$ is locally constant

on an open set Ω, while the constant rank condition is obviously not necessary for the equivalence of the global minimizers.

Concerning the corresponding Moore-Penrose pseudoinverse $\Phi(\lambda)^\dagger$, in our setup the constant rank condition further allows to establish continuity and even smoothness of H on Ω, see the subsequent Corollary 3.1 based on the following Theorem 3.2. For this purpose, let $D_\lambda \Phi(\lambda)$ denote the Fréchet derivative of the matrix $\Phi(\lambda)$ with respect to λ. For a proof of Theorem 3.2, let us refer to [25], Theorem 4.3. Note that our Equation (10) equals equation (4.12) in Ref. [25].

Theorem 3.2: *Assume that in the open set Ω, the matrix of basis functions $\Phi(\lambda)$ has constant rank $0 < q \leq l$. Further, let $\Phi(\lambda)$ be Fréchet differentiable with respect to λ in Ω. Then, for any $\lambda \in \Omega$, we have that the following identity holds:*

$$\begin{aligned} D_\lambda \Phi(\lambda)^\dagger = &-\Phi(\lambda)^\dagger D_\lambda \Phi(\lambda) \Phi(\lambda)^\dagger \\ &+ \left(\Phi(\lambda)^\top \Phi(\lambda)\right)^\dagger D_\lambda \Phi(\lambda)^\top \left(I - \Phi(\lambda)\Phi(\lambda)^\dagger\right) \\ &+ \left(I - \Phi(\lambda)^\dagger \Phi(\lambda)\right) D_\lambda \Phi(\lambda)^\top \left(\Phi(\lambda)\Phi(\lambda)^\top\right)^\dagger. \end{aligned} \tag{10}$$

From the differentiability of the Moore-Penrose pseudoinverse on Ω, it immediately follows with (8) and (10) that the objective function H is differentiable on Ω with respect to λ as well, so that formulas for its gradient $\nabla_\lambda H(\lambda) = (D_\lambda H(\lambda))^\top$ can be established, see the following Corollary 3.1. In Corollary 3.1 we cover the Svensson model; the Nelson–Siegel model can be easily recovered by setting $\lambda_2 = 0$, $\beta_4^*(\lambda) = 0$ and neglecting the second column of $D_\lambda H(\lambda)$.

Corollary 3.1: *Assume that in the open set Ω, the matrix of basis functions $\Phi(\lambda)$ has constant rank $0 < q \leq l$ and that $l = 4$. Further, let $\Phi(\lambda)$ be Fréchet differentiable with respect to λ in Ω. Then, for any $\lambda \in \Omega$, H is differentiable and it holds:*

$$D_\lambda H(\lambda) = -2\hat{y}^\top \left(I - \Phi(\lambda)\Phi(\lambda)^\dagger\right) \left[\beta_3^*(\lambda)\,(\tau \circ e^{-\lambda_1 \tau}), \beta_4^*(\lambda)\,(\tau \circ e^{-\lambda_2 \tau})\right], \tag{11}$$

where "$\circ$" denotes the Hadamard product of componentwise vector multiplication.

Proof: Differentiability of H has already been noted above. Further, according to Ref. [25], equation (4.7), the derivative of H with respect to λ can be written as

$$D_\lambda H(\lambda) = -2\hat{y}^\top \left(I - \Phi(\lambda)\Phi(\lambda)^\dagger\right) D_\lambda \Phi(\lambda) \Phi(\lambda)^\dagger \hat{y}. \tag{12}$$

Since $D\Phi(\lambda) \in \mathbb{R}^{k \times (m \times l)}$ is a tensor, its first and second slice with partial derivatives with respect to λ_1 and λ_2 have the matrix forms

$$[D_\lambda \Phi(\lambda)]_1 = \frac{\partial}{\partial \lambda_1} \Phi(\lambda) = \left[\mathbf{0}, D_{\lambda_1} \phi_2(\lambda; \tau), D_{\lambda_1} \phi_2(\lambda; \tau) + \tau \circ e^{-\lambda_1 \tau}, \mathbf{0}\right]$$

and

$$[D\Phi(\lambda)]_2 = \frac{\partial}{\partial \lambda_2}\Phi(\lambda) = \left[\mathbf{0}, \mathbf{0}, \mathbf{0}, D_{\lambda_2}\phi_4(\lambda;\tau)\right],$$

respectively, where

$$D_{\lambda_1}\phi_2(\lambda;\tau) = \frac{\partial}{\partial \lambda_1}\phi_2(\lambda;\tau) = \frac{e^{-\lambda_1\tau}}{\lambda_1} - \frac{1-e^{-\lambda_1\tau}}{\lambda_1^2\tau},$$

$$D_{\lambda_1}\phi_3(\lambda;\tau) = \frac{\partial}{\partial \lambda_1}\phi_3(\lambda;\tau) = D_{\lambda_1}\phi_2(\lambda;\tau) + \tau \circ e^{-\lambda_1\tau}, \quad \text{and}$$

$$D_{\lambda_2}\phi_4(\lambda;\tau) = \frac{\partial}{\partial \lambda_2}\phi_4(\lambda;\tau) = \frac{e^{-\lambda_2\tau}}{\lambda_2} - \frac{1-e^{-\lambda_2\tau}}{\lambda_2^2\tau} + \tau \circ e^{-\lambda_2\tau}$$

denote the derivative of the i-th basis function $\phi_i(\lambda;\tau)$ with respect to λ_k. Since $D_{\lambda_1}\phi_2(\lambda;\tau)$ does not depend on λ_2, we can set

$$h_2(\lambda_1) := D_{\lambda_1}\phi_2((\lambda_1, 0);\tau).$$

As previously noted, for each $\lambda \in \Omega$ the inner problem in β possesses at least one optimal solution $\beta^*(\lambda)$ with $\beta^*(\lambda) = \Phi(\lambda)^\dagger \hat{y}$. We can thus rewrite (12) as

$$\begin{aligned} D_\lambda H(\lambda) = -2\hat{y}^\top \left(I - \Phi(\lambda)\Phi(\lambda)^\dagger\right) \\ \times \left[\left(\beta_2^*(\lambda) + \beta_3^*(\lambda)\right) h_2(\lambda_1) + \beta_3^*(\lambda)\,(\tau \circ e^{-\lambda_1\tau}),\right. \\ \left.\beta_4^*(\lambda) h_2(\lambda_2) + \beta_4^*(\lambda)\,(\tau \circ e^{-\lambda_2\tau})\right]. \end{aligned}$$

Now, $h_2(\lambda_1;\tau) = -\phi_3(\lambda;\tau)/\lambda_1$ and $h_2(\lambda_2;\tau) = -\phi_4(\lambda;\tau)/\lambda_2$, and from the normal equations of the inner problem for fixed λ

$$\Phi(\lambda)^\top \Phi(\lambda)\beta^*(\lambda) = \Phi(\lambda)^\top \hat{y},$$

it follows that any column of $\Phi(\lambda)$ is orthogonal to $(\Phi(\lambda)\Phi(\lambda)^\dagger - I)^\top \hat{y}$. Hence, $D_\lambda H(\lambda)$ can be simplified to

$$D_\lambda H(\lambda) = -2\hat{y}^\top (I - \Phi(\lambda)\Phi(\lambda)^\dagger)[\beta_3^*(\lambda)\,(\tau \circ e^{-\lambda_1\tau}), \beta_4^*(\lambda)\,(\tau \circ e^{-\lambda_2\tau})].$$

■

Note that while Equation (12) holds in the general situation, the specific form of $\nabla_\lambda H(\lambda)$ in (11) only holds for the Nelson–Siegel and the Svensson models. Moreover, since the matrix norms of $\Phi(\lambda)$ and $\Phi(\lambda)^\dagger$ and the Fréchet derivatives $D\Phi(\lambda)$ and $D\Phi(\lambda)^\top$ are bounded on bounded domains for constant rank, the Fréchet derivative in (10) is bounded as well. In particular, the Moore-Penrose pseudoinverse is locally Lipschitz continuous on Ω, as well as the objective function H as a composition of locally Lipschitz continuous functions. Thus, H is globally Lipschitz continuous on any compact subset $\Lambda \subset \Omega$.

It is a natural next step to investigate for which values of $\lambda \in \mathbb{R}_+^k$ the matrix $\Phi(\lambda)$ might fail to have full rank. We will prove that in the Svensson model a rank deficiency may only occur for points which are not global minimizers, while in the Nelson–Siegel model, rank deficiency will only happen for a few isolated points in the general situation; and most importantly, does not happen at all in our specific situation.

Proposition 3.1: *Let $m \geq 3$ and $\tau_1, \tau_2, \tau_3 \in \mathbb{Q}$. Then in the Nelson–Siegel model we have*

$$|\{\lambda \in \mathbb{R}_+ : \mathit{rank}(\Phi(\lambda)) < 3\}| \leq 4.$$

Proof: We first note that the extreme case $\lambda = 0$ corresponds to $\Phi(\lambda)$ having rank 1, as $\Phi(\lambda)$ only contains columns of 0's or columns of 1's. Therefore, let $\lambda > 0$ in the following. To prove the claim it is sufficient to consider the upper 3×3 matrix of $\Phi(\lambda)$, i.e. ignore further maturities, as the rank of the full matrix is always equal to or larger than the rank of this submatrix:

$$A := \begin{pmatrix} \phi_1(\lambda;\tau_1) & \phi_2(\lambda;\tau_1) & \phi_3(\lambda;\tau_1) \\ \phi_1(\lambda;\tau_2) & \phi_2(\lambda;\tau_2) & \phi_3(\lambda;\tau_2) \\ \phi_1(\lambda;\tau_3) & \phi_2(\lambda;\tau_3) & \phi_3(\lambda;\tau_3) \end{pmatrix}.$$

Since $\phi_3(\lambda;\tau) = \phi_2(\lambda;\tau) - e^{-\lambda\tau}$, we can subtract the second column from the third, then multiply the third column by -1, swap second and third column, and obtain the matrix B with $\det(B) = \det(A)$:

$$B := \begin{pmatrix} 1 & e^{-\lambda\tau_1} & \dfrac{1-e^{-\lambda\tau_1}}{\lambda\tau_1} \\ 1 & e^{-\lambda\tau_2} & \dfrac{1-e^{-\lambda\tau_2}}{\lambda\tau_2} \\ 1 & e^{-\lambda\tau_3} & \dfrac{1-e^{-\lambda\tau_3}}{\lambda\tau_3} \end{pmatrix}.$$

We can then multiply the last column of B by λ and substitute $z := e^{-\lambda}$ (note $0 < z < 1$) to obtain the matrix C with $\det(C) = \lambda \det(B)$:

$$C := \begin{pmatrix} 1 & z^{\tau_1} & \dfrac{1-z^{\tau_1}}{\tau_1} \\ 1 & z^{\tau_2} & \dfrac{1-z^{\tau_2}}{\tau_2} \\ 1 & z^{\tau_3} & \dfrac{1-z^{\tau_3}}{\tau_3} \end{pmatrix}.$$

Letting $t \in \mathbb{N}$ be the least common denominator of τ_1, τ_2 and τ_3, hence $\tau_i = \nu_i/t$ for some $\nu_i \in \mathbb{N}$, $i = 1, 2, 3$, we can further substitute $y = z^{1/t}$ and obtain

$$1/t \cdot \det(C) = \left(\frac{1}{\nu_2} - \frac{1}{\nu_3}\right) y^{\nu_3+\nu_2} + \left(\frac{1}{\nu_3} - \frac{1}{\nu_1}\right) y^{\nu_3+\nu_1} + \left(\frac{1}{\nu_1} - \frac{1}{\nu_2}\right) y^{\nu_2+\nu_1}$$

$$+\left(\frac{1}{\nu_1}-\frac{1}{\nu_2}\right)y^{\nu_3}+\left(\frac{1}{\nu_3}-\frac{1}{\nu_1}\right)y^{\nu_2}+\left(\frac{1}{\nu_2}-\frac{1}{\nu_3}\right)y^{\nu_1}.$$

Thus, $1/(t\cdot y^{\nu_1})\cdot\det(C)$ yields a polynomial in y with a constant term and five monomial terms. By Descartes' rule of sign, this polynomial can have at most five distinct positive real roots as it has at most five sign changes in the coefficients. Further, since the product of the first and the last coefficient is positive, the number of roots has to be even, hence there are at most four different positive real roots. Since we can further divide the remaining polynomial by the leading factor, which yields a constant term of 1, we know that there must be at least one positive real root larger than 1 if there is any positive real root smaller than 1 (as the product of all roots equals 1). Thus, there are at most three positive real roots in the open interval $(0,1)$, which proves the claim. ■

Remark 3.1: Since $1/(t\cdot y^{\nu_1})\cdot\det(C)$ has to vanish for $y=z=1$, one might be able to show that the polynomial is, from there on, strictly increasing. This would actually show that there are no roots besides $z=0$ and $z=1$. Unfortunately, we have not been able to prove this yet. Nevertheless, we have checked a variety of choices for the maturities τ and we have never found an instance where $\Phi(\lambda)$ becomes rank deficient in the Nelson–Siegel model.

Fortunately, for the specific choice of maturities which we consider in Section 5, it is quite easy to prove that $\Phi(\lambda)$ always has full rank:

Remark 3.2: For the specific choice of maturities as in Section 5, where $m\geq 3$ and $\tau_1=1$, $\tau_2=2$, and $\tau_3=3$, we have that $t=1$ and, most importantly,

$$\det(C)=1/6\cdot(z-1)^4 z,$$

which shows that $\Phi(\lambda)$ has full rank for all $\lambda>0$. Thus, in the Nelson–Siegel model, we can choose $\boldsymbol{\Omega}=\mathbb{R}_+$ as minimization region, where the case $\lambda=0$ can be covered separately in an easy fashion.

From now on, we make the assumption on the maturities that $\lambda=0$ is the only point of rank deficiency for $\Phi(\lambda)$ in the Nelson–Siegel model, i.e. we require that the maturities $\tau_1,\ldots,\tau_m$ are such that

$$\text{in the Nelson–Siegel model holds: } \{\lambda\in\mathbb{R}_+ :\ \operatorname{rank}(\Phi(\lambda))<3\}=\{0\}. \tag{FRNS}$$

(FRNS) is an assumption, which is satisfied in our numerical setup according to Remark 3.2. Further, (FRNS) can be easily checked (e.g. with symbolic computing toolboxes) for other maturity choices.

Under assumption (FRNS) let us now consider the Svensson model in more detail. We first consider the case that a related full rank assumption for the Svensson model holds:

$$\text{(FRNS) holds and in the Svensson model we have:}$$
$$\forall\lambda_1>0\ \exists\lambda_2>0:\operatorname{rank}(\Phi(\lambda))=4. \tag{FRSv}$$

Remark 3.3: For the specific choice of maturities as in Section 5, where $m \geq 4$ and $\tau_k = k$ $(k = 1, \ldots, 4)$, we can prove by similar considerations as in the proof of Proposition 3.1 that the upper 4×4 matrix of $\Phi(\lambda_1, \lambda_2)$ has full rank for all choices of $\lambda_1 > 0$ when we set $\lambda_2 = \ln(4)$ if $\lambda_1 < \ln(3)$ and $\lambda_2 = \ln(2)$ otherwise. Alternatively, this also follows from Remark 3.4 as an easy consequence. In summary, (FRSv) holds.

Under (FRSv), we can now show that there is at least one global optimizer λ^* of H such that $\Phi(\lambda^*)$ has full rank. This statement is made precise in the following Proposition 3.2. Proposition 3.2 allows us to restrict the global minimization of H to regions where $\Phi(\lambda)$ has full rank.

Proposition 3.2: *Let (FRSv) hold. Then*

$$\min_{\lambda \in \mathbb{R}^2_+} H(\lambda) = \min_{\lambda \in \mathbb{R}^2_+ : \operatorname{rank}(\Phi(\lambda)) = 4} H(\lambda)$$

Proof: Let $\bar{\lambda}$ be a global minimizer of H on $\mathbb{R}^2_+$ with $\operatorname{rank}(\Phi(\bar{\lambda})) < 4$. Then $\operatorname{rank}(\Phi(\bar{\lambda})) = 3$, as the first three columns of $\Phi(\lambda)$ are independent by assumption (FRNS). Thus, there exist $\hat{c}_1, \ldots, \hat{c}_4 \in \mathbb{R}$ with $\hat{c}_4 \neq 0$ such that

$$\sum_{k=1}^{4} \hat{c}_k \phi_k(\bar{\lambda}; \tau) = 0,$$

or, equivalently, for $c_k = -\hat{c}_k / \hat{c}_4$:

$$\phi_4(\bar{\lambda}; \tau) = \sum_{k=1}^{3} c_k \phi_k(\bar{\lambda}; \tau).$$

Using $H(\bar{\lambda}) = \min_{\beta \in \mathbb{R}^4} F(\bar{\lambda}, \beta) = F(\bar{\lambda}, \beta^*(\bar{\lambda}))$ for some optimal $\beta^*(\bar{\lambda})$ yields

$$\begin{aligned} H(\bar{\lambda}) &= \min_{\beta \in \mathbb{R}^4} F(\bar{\lambda}, \beta) = F(\bar{\lambda}, \beta^*(\bar{\lambda})) = F(\bar{\lambda}, \beta_1^*(\bar{\lambda}), \ldots, \beta_4^*(\bar{\lambda})) \\ &= F(\bar{\lambda}, \beta_1^*(\bar{\lambda}) - c_1 \beta_4^*(\bar{\lambda}), \ldots, \beta_3^*(\bar{\lambda}) - c_3 \beta_4^*(\bar{\lambda}), 0). \end{aligned}$$

This means that $\bar{\lambda}_2$ is such that ϕ_4 is already contained in the linear hull of the first three basis functions. Since now $\beta_4 = 0$, we can choose any other value $\hat{\lambda}_2$ instead of $\bar{\lambda}_2$ without changing the value of F. We choose $\hat{\lambda}_2$ such that $\Phi(\bar{\lambda}_1, \hat{\lambda}_2)$ has full rank (i.e. $\operatorname{rank}(\Phi(\bar{\lambda}_1, \hat{\lambda}_2)) = 4$), which is possible due to assumption (FRSv) and obtain:

$$\begin{aligned} H(\bar{\lambda}) &= \ldots = F(\bar{\lambda}, \beta_1^*(\bar{\lambda}) - c_1 \beta_4^*(\bar{\lambda}), \ldots, \beta_3^*(\bar{\lambda}) - c_3 \beta_4^*(\bar{\lambda}), 0) \\ &= F((\bar{\lambda}_1, \hat{\lambda}_2), \beta_1^*(\bar{\lambda}) - c_1 \beta_4^*(\bar{\lambda}), \ldots, \beta_3^*(\bar{\lambda}) - c_3 \beta_4^*(\bar{\lambda}), 0) \\ &= \min_{\beta \in \mathbb{R}^3} F((\bar{\lambda}_1, \hat{\lambda}_2), \beta_1, \ldots, \beta_3, 0) \end{aligned} \tag{X}$$

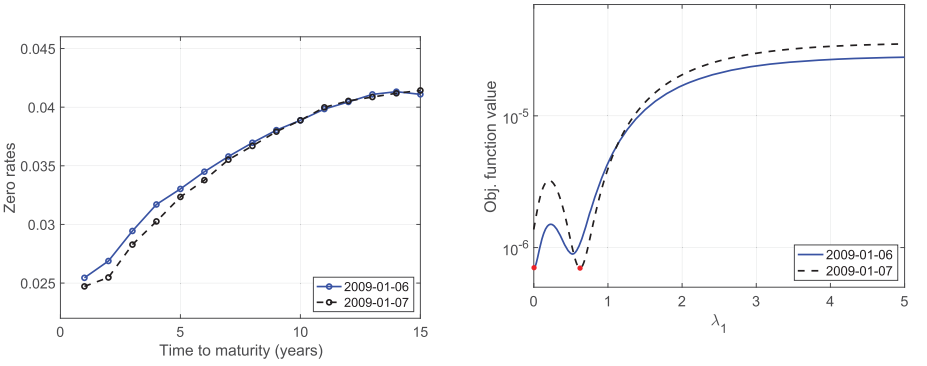

Figure 1. Term structures and objective functions H with corresponding global minima when the Nelson–Siegel model is fitted to the data on 6 January 2009 and 7 January 2009, respectively. (a) Term structures. (b) Objective functions H and global minima (red points).

$$\geq \min_{\beta\in\mathbb{R}^4} F((\bar{\lambda}_1, \hat{\lambda}_2), \beta_1, \ldots, \beta_3, \beta_4)$$
$$= H((\bar{\lambda}_1, \hat{\lambda}_2)).$$

This shows the claim. ■

Since the set $\{\lambda \in \mathbb{R}^2_+ \mid \operatorname{rank}(\Phi(\lambda))) = 4\}$ is open, we choose $\mathbf{\Omega} := \{\lambda \in \mathbb{R}^2_+ \mid \operatorname{rank}(\Phi(\lambda)) = 4\}$ for the Svensson model. For an approximate characterization of $\mathbf{\Omega}$ let us refer to Remark 3.4. Note that Figure 2 indicates that Φ has full rank in the Svensson model as long as $\lambda \in \mathcal{L} := \{(\lambda_1, \lambda_2) \in \mathbb{R}^2_+ \mid \lambda_1 > 0, \lambda_2 > 0, \lambda_1 \neq \lambda_2\}$ with potential exception of the two bent curves visible in Figure 2. This can indeed be rigorously shown in our setup:

Remark 3.4: For the specific choice of maturities as in Section 5, where $m \geq 4$ and $\tau_k = k$ $(k = 1, \ldots, 4)$, we have that with

$$\mathcal{B} := \{(\lambda_1, \lambda_2) \in \mathbb{R}^2_+ \mid \lambda_1 = \ln(v(e^{-\lambda_2})) - \ln(u(e^{-\lambda_2}))\},$$

where $u(s) = 1 - s + 2s\ln(s)$ and $v(s) = s^2 - 4s^2\ln(s) - s + 2s\ln(s)$, that

$$\mathcal{L}\backslash\mathcal{B} \subset \{\lambda \in \mathbb{R}^2_+ \mid \operatorname{rank}(\Phi(\lambda)) = 4\} \subset \mathcal{L}.$$

In analogy to the proof of Proposition 3.1, the first inclusion (the second is obvious) can be proved by showing, for $a := e^{-\lambda_2}$,

$$\det \begin{pmatrix} 1 & z^{\tau_1} & \dfrac{1-z^{\tau_1}}{\tau_1} & \dfrac{1-a^{\tau_1}}{-\tau_1\ln(a)} - a^{\tau_1} \\ 1 & z^{\tau_2} & \dfrac{1-z^{\tau_2}}{\tau_2} & \dfrac{1-a^{\tau_2}}{-\tau_2\ln(a)} - a^{\tau_2} \\ 1 & z^{\tau_3} & \dfrac{1-z^{\tau_3}}{\tau_3} & \dfrac{1-a^{\tau_3}}{-\tau_3\ln(a)} - a^{\tau_3} \\ 1 & z^{\tau_4} & \dfrac{1-z^{\tau_4}}{\tau_4} & \dfrac{1-a^{\tau_4}}{-\tau_4\ln(a)} - a^{\tau_4} \end{pmatrix}$$

$$= -\frac{z}{24}(z-1)^4 \frac{a-1}{\ln(a)}(a-z)\left(u(a)z + v(a)\right).$$

This shows that the determinant of the above matrix can only vanish for given a if $z = a$ or if $z = -v_a/u_a$. Whether Φ has full rank on $\mathcal{B}$ remains open, as we have only considered the upper 4×4 block of Φ.

Under a slightly stronger assumption than (FRSv), we can show a stronger statement than in Proposition 3.2. For this purpose, let us introduce the assumption (RNS) on the maturities τ_i, $i = 1, \ldots, m$:

$$\exists 0 < \lambda_1^{(1)} < \ldots < \lambda_1^{(m)} : \det\left(\phi_3((\lambda_1^{(j)}, 0); \tau_k)_{k,j=1,\ldots,m}\right) \neq 0. \tag{RNS}$$

In other words, we can find m different values for λ_1 such that the m versions of the basis function ϕ_3 form a basis of the $\mathbb{R}^m$.

Remark 3.5: Assume (FRNS); then assumption (FRSv) immediately follows from (RNS): since the m versions of the basis function ϕ_3 (and thus also those of ϕ_4) form a basis of $\mathbb{R}^m$, we can always pick one of these m values for λ_2 to get a fourth column of $\Phi(\lambda)$, which is not contained in the linear hull of the first three columns.

Remark 3.6: For the specific choice of maturities as in Section 5, where $m = 15$ and $\tau_k = k$ $(k = 1, \ldots, 15)$, we can show numerically that

$$\begin{aligned}\{\lambda_1^{(1)}, \ldots, \lambda_1^{(m)}\} &= \{0.01, 0.04, 0.09, 0.17, 0.28, 0.42, 0.58, 0.79, 1.05, 1.36, 1.77, 2.31, 3.10,\\ &\quad 4.43, 8.47\}\end{aligned}$$

yields a set of values which satisfies (RNS).

Similar considerations show that for our numerical tests the basis function ϕ_2 is also rich enough, i.e. ϕ_2 satisfies (RNS), given

$$\begin{aligned}\{\lambda_1^{(1)}, \ldots, \lambda_1^{(m)}\} &= \{0.01, 0.02, 0.06, 0.12, 0.20, 0.32, 0.47, 0.65, 0.89, 1.18, 1.56, 2.06, 2.77,\\ &\quad 3.99, 7.62\}.\end{aligned}$$

We can now strengthen our result above for the Svensson model to the following Theorem 3.3. Note that Theorem 3.3 is not relevant for the Nelson–Siegel model, as for the Nelson–Siegel model the matrix $\Phi(\lambda)$ always has full rank under our given assumptions.

Theorem 3.3: *Let λ^* be the global minimizer of $H(\lambda)$ on $\mathbb{R}^2_+$, let $H(\lambda^*) > 0$ and let (RNS) hold. Then, the matrix $\Phi(\lambda^*)$ has full rank.*

Proof: Let $\bar{\lambda}$ be a global minimizer of H with $H(\bar{\lambda}) > 0$ and assume $\Phi(\bar{\lambda})$ does not have full rank. Then, in complete analogy to the proof of Proposition 3.2 up to the equality marked (X), we have:

$$\begin{aligned} H(\bar{\lambda}) &= \min_{\beta\in\mathbb{R}^3} F((\bar{\lambda}_1,\bar{\lambda}_2),\beta_1,\ldots,\beta_3,0) = F((\bar{\lambda}_1,\bar{\lambda}_2),\bar{\beta}_1,\ldots,\bar{\beta}_3,0) \\ &= F((\bar{\lambda}_1,\hat{\lambda}_2),\bar{\beta}_1,\ldots,\bar{\beta}_3,0). \end{aligned}$$

Since $H(\bar{\lambda}) > 0$, the residual $\hat{y} - \Phi(\bar{\lambda})\bar{\beta}$ (with $\bar{\beta}_4 = 0$) does not equal 0. Further, due to (RNS), we can choose $\hat{\lambda}_2$ not only in such a way that $\Phi((\bar{\lambda}_1,\hat{\lambda}_2))$ has full rank, but we can further choose it such that $\phi_4((\bar{\lambda}_1,\hat{\lambda}_2);\tau)$ is not orthogonal to the residual $\hat{y} - \Phi(\bar{\lambda})\bar{\beta}$ and hence

$$\begin{aligned} H(\bar{\lambda}) &= F((\bar{\lambda}_1,\hat{\lambda}_2),\bar{\beta}_1,\ldots,\bar{\beta}_3,0) > \min_{\beta_4\in\mathbb{R}} F((\bar{\lambda}_1,\hat{\lambda}_2),\bar{\beta}_1,\ldots,\bar{\beta}_3,\beta_4) \\ &\geq \min_{\beta\in\mathbb{R}^4} F((\bar{\lambda}_1,\hat{\lambda}_2),\beta) = H((\bar{\lambda}_1,\hat{\lambda}_2)). \end{aligned}$$

This shows that $\bar{\lambda}$ cannot be a global minimizer, hence the claim follows. ■

If $H(\lambda^*) = 0$, not much can be said about the rank of $\Phi(\lambda^*)$. For example, one might be given data $\hat{y}$ which is already in the linear hull of the first three basis functions. Then the choice of λ_2 does not play a role and one can choose λ_2 in a way such that a rank deficit of Φ occurs. In our numerical tests, $H(\lambda^*) = 0$ has never occurred, indicating that this is indeed a rare event in practise. Further, Proposition 3.2 tells us that there is at least one other global minimizer without a rank deficit.

While Proposition 3.2 already allows us to consider only points where Φ has full rank, Theorem 3.3 additionally yields that points λ with rank deficient $\Phi(\lambda)$ have worse function values than the global minimizer. Thus, these regions can be avoided by the minimization routine, which provides the basis for our penalty approach in the next section.

Let us finally remark that establishing global Lipschitz continuity of H on compact $\Lambda \subset \Omega$ provides the main basis for all global optimization approaches for the reduced problem; a result which so far has been missing in the corresponding literature.

We are now in a situation where we could apply any reasonable global method to the optimization of H. However, instead of more involved strategies, we remain with the most simple grid search approach introduced by Nelson and Siegel [1] for three reasons: first, the dimension of the global optimization problem is reduced to one or two and thus grid search is computationally feasible, second, we are given the natural lower bound of 0 for H and can thus easily judge the

quality of potential solutions, and third, this method is most easily extendable to the ideas presented in the following section. As our numerical investigations in Section 5 show, this approach already yields encouraging results.

4. Stability analysis

One of the main issues that arise in the minimization of the reduced optimization problem (7) is the stability of optimal solutions. To assess the quality of optimal solutions, note that the evaluation of H solely depends on the solution of the inner problem $\beta^*(\lambda) = \Phi(\lambda)^\dagger \hat{y}$. Hence, the stability of the inner optimal solution β^* of the separable least-squares problem can essentially be analysed by applying perturbation theory to linear least-squares problems, see e.g. Ref. [24], Section 1.4. Accordingly, there are two different scenarios in which optimal solutions can become sensitive with respect to perturbations of either the data vector $\hat{y}$ or the matrix $\Phi(\lambda)$. The first scenario concerns the projection of $\hat{y}$ onto the span of $\Phi(\lambda)$ and turns out to be of relevance if both components are nearly orthogonal to each other. In such case, the projected $\hat{y}$ is much smaller than $\hat{y}$ itself so that minor changes in $\hat{y}$ may affect the linear solution $\beta^*(\lambda)$ greatly. However, since both the Nelson–Siegel and the Svensson model are able to fit a variety of different shapes with high accuracy, this scenario never occurs for these models and the sensitivity to perturbations in $\hat{y}$ can be neglected.[4] The second issue pertains to the conditioning of the matrix $\Phi(\lambda)$ and thus is influenced solely by the factor loading structure that is imposed by the models. In this case, optimal solutions of the linear least-squares system respond strongly to perturbations in $\Phi(\lambda)$ if the matrix is ill-conditioned, i.e. if some of the columns of $\Phi(\lambda)$ are almost linearly dependent. Since this is a more subtle issue, in the remaining part of this section we provide a thorough analysis of the potential ill-conditioning of $\Phi(\lambda)$ and how it can be dealt with. In particular, we use the condition number of the matrix $\Phi(\lambda)$ to measure the sensitivity of an optimal solution $(\lambda^*, \beta^*(\lambda^*))^\top$, which also corresponds to the condition of the problem of evaluating $H(\lambda)$. In this way, we are able to quantify – and manage – the ill-conditioning with our enhanced approach, in contrast to previous approaches.

Let us point out that the above considerations are not to be confused with the dependence of the optimal λ^* on the data $\hat{y}$, as this is a different issue. Recall that the main relationship here is that changes in $\hat{y}$ imply changes in λ^*, which in turn imply changes in β^*. While the above considerations contemplate the question, how (small) changes in $\Phi(\lambda^*)$ (due to small changes in λ^*) impact β^*, the general dependence of λ^* with respect to $\hat{y}$ is of different nature. Our penalty approach, which we introduce later, stabilizes the local behaviour of β^* for small changes in λ^* as it takes care of the condition number of $\Phi(\lambda)$. However, no approach whatsoever will be able to prevent large changes in λ^* given small changes in $\hat{y}$

[4] If models are used where perturbations to $\hat{y}$ turn out to be relevant, the following analysis can be extended by adjusting the condition number to include $\hat{y}$, see, e.g. Ref. [24], Subsection 1.4.3.

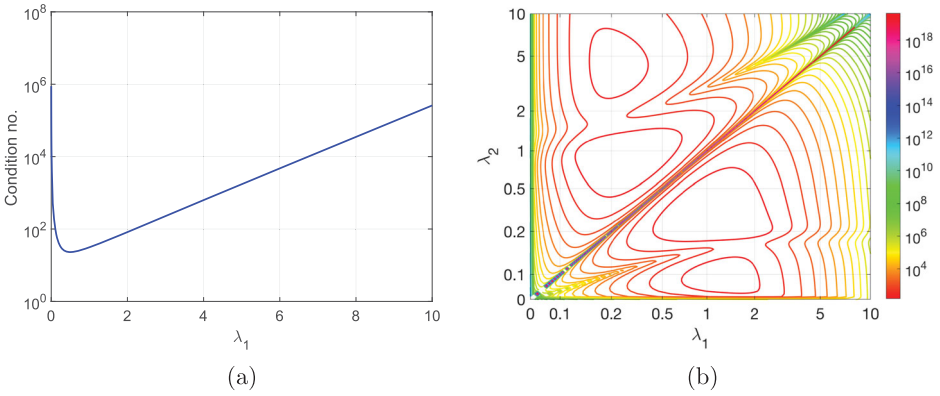

Figure 2. A two-dimensional line plot and a contour plot of the condition number of the matrix $\Phi(\lambda)$ as a function of the nonlinear parameter λ for the Nelson–Siegel and the Svensson models, respectively, with maturity vector $\tau = (1, 2, \ldots, 15)^\top$. (a) Nelson–Siegel model. (b) Svensson model.

due to the inherent non-convex structure of the fitting problem in λ, as the following example shows: Let us take $\hat{y}$ from the subsequent dates 6 January 2009 and 7 January 2009 and look at corresponding objective functions H in the Nelson–Siegel model. As we can see in Figure 1, interest rates do not change much from one day to the other, but the optimal λ^* moves from one local minimum to the other. Similar behaviour can be observed in the Svensson model, where this happens more often due to a larger number of local minima, which all have quite similar objective values (compare Figure 6 for an illustrative instance). Unfortunately, such a behaviour in λ (and thus in β correspondingly) can never be avoided by any fitting method due to the inherent structure of the models, unless one gives up on the global optimality of λ. Moreover, Figure 1 again indicates the unwanted property of *unidentifiability* of the Nelson–Siegel model, i.e. there exists a data vector $\hat{y}$ such that two different λ values yield the same optimal quality of approximation.

4.1. The inherent ill-conditioning of $\Phi(\lambda)$

Recall that the matrix function $\Phi(\lambda)$ in the Nelson–Siegel and the Svensson models can be written as

$$\Phi(\lambda) = [\phi_1(\lambda; \tau), \ldots, \phi_l(\lambda; \tau)], \tag{13}$$

with basis functions ϕ_j, $j = 1, \ldots, l$. This implies that the degree of ill-conditioning of $\Phi(\lambda)$ depends on both the vector of shape parameters λ and the vector of predefined maturities τ, but not on the data $\hat{y}$.[5] To be able to quantify the degree of ill-conditioning of the rectangular matrix $\Phi(\lambda) \in \mathbb{R}^{m \times l}$, we consider its

[5] For the available data described in Section 5.1, we have $m = 15$ and $\tau = (1, 2, \ldots, 15)^\top$.

singular value decomposition according to Ref. [24], Theorem 1.2.1, where

$$\Phi(\lambda) = U(\lambda) \begin{pmatrix} \Sigma(\lambda) & 0 \\ 0 & 0 \end{pmatrix} V(\lambda)^T,$$

for unitary matrices $U(\lambda) \in \mathbb{R}^{m\times m}$ and $V(\lambda) \in \mathbb{R}^{l\times l}$, and the diagonal matrix $\Sigma(\lambda)$ containing the singular values of $\Phi(\lambda)$. Using this decomposition, the condition number of the rectangular matrix $\Phi(\lambda)$ is then defined as follows, cf. [24], Definition 1.4.2.

Definition 4.1: The condition number of $\Phi(\lambda) \in \mathbb{R}^{m\times l}$ is given by

$$\kappa\left(\Phi(\lambda)\right) = \|\Phi(\lambda)\|_2 \left\|\Phi(\lambda)^{\dagger}\right\|_2 = \frac{\sigma_1(\lambda)}{\sigma_q(\lambda)},$$

where $0 < q \leq l$, $\sigma_1(\lambda) \geq \sigma_2(\lambda) \geq \ldots \geq \sigma_q(\lambda) > 0$ are the nonzero singular values of $\Phi(\lambda)$, and $\|\cdot\|_2$ denotes the matrix 2-norm.

The condition number describes how solutions of the linear least-squares problems are affected by small perturbations. If the condition number is 'large', i.e. solutions are affected greatly, the problem is said to be ill-conditioned, see e.g. Ref. [26], Chapter 3. A more accurate interpretation of ill-conditioning is subject to the problem at hand and depends on the application. For our setup, we will give a suitable idea of a large condition number in Section 4.2.

The effect of having obtained an optimal nonlinear solution λ^* with ill-conditioned matrix $\Phi(\lambda^*)$ may become especially apparent in that some of the values of the corresponding linear parameter $\beta^*(\lambda^*)$ turn out to be very large (and offsetting), with values being proportional to the degree of ill-conditioning. This, though, is in contradiction to the intuitive economic interpretation that all model parameters have.

Figure 2 shows a line plot and a contour plot of the condition number of the matrix $\Phi(\lambda)$ as a function of the parameter λ for the Nelson–Siegel model and Svensson's extension, respectively. From the subfigures, we can observe that the main difficulties in the fitting of both models arise when the shape parameter λ is either very small or becomes increasingly large, or, in the case of the Svensson model, when $\lambda_1 \approx \lambda_2$. The severity of the ill-conditioning in the latter case is illustrated by the elevated diagonal in the contour plot of the condition number, see Figure 2(b). Disproportional large condition numbers can also be observed beneath and above the diagonal in form of slightly bent curves for very small and increasingly large values of the parameter λ, respectively. These curves exactly represent the set $\mathcal{B}$ defined in Remark 3.4. Note that our numerical computation indicates that in both cases the linear dependence of the columns is only approximate and thus does not lead to a rank-deficient matrix $\Phi(\lambda)$ (although the upper 4×4 block is rank-deficient according to Remark 3.4).

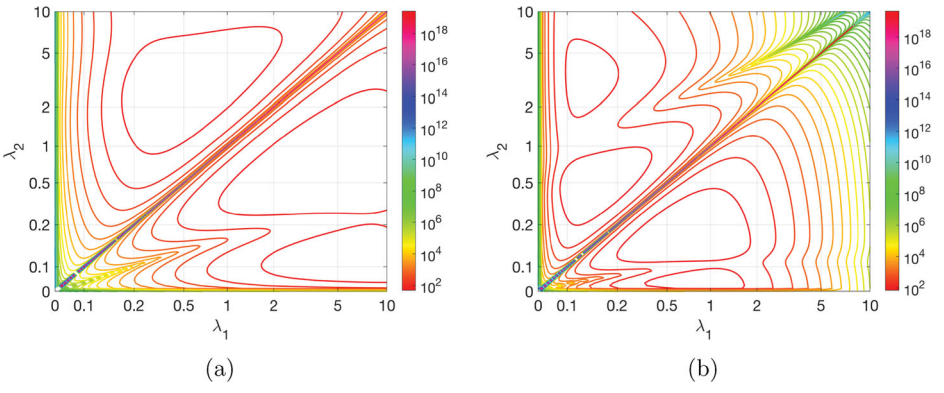

Figure 3. Contour plots of the condition number of the matrix $\Phi(\lambda)$ as a function of the nonlinear parameter λ for the Svensson model with different maturity vectors $\tau^{(1)} = (1/12, 1/4, 1/2, 1, 2, \ldots, 15)^\top$ and $\tau^{(2)} = (1, 2, \ldots, 15, 24, 36, 60)^\top$. (a) Maturity vector $\tau^{(1)}$. (b) Maturity vector $\tau^{(2)}$.

The approximate linear dependence between the columns of the matrix $\Phi(\lambda)$ can be mitigated by considering observations with shorter and/or longer maturities, in addition to the observations already used in the model. As an example, the impact on the condition number of the matrix $\Phi(\lambda)$ when including short and long maturities into the vector of maturities τ is depicted in Figure 3(a ,b), respectively. Accordingly, the inclusion of short maturities can considerably improve the degree of ill-conditioning in the region with increasingly large λ's and hence enlarge the parameter space for which a solution may be acceptably stable. Similarly, the inclusion of long maturities can improve the degree of ill-conditioning in the region where λ is small.

Let us point out again that if the global optimal solution λ^* leads to an ill-conditioned $\Phi(\lambda^*)$, then it can not be guaranteed that the parameters of the model can be identified with high accuracy – independent of the method used. Hence, ill-conditioning is an issue with the Nelson–Siegel and the Svensson models themselves, which can occur for certain types of curves, i.e. certain shape parameters λ. Usually, simply shaped curves (e.g. flat curves, i.e. λ close to 0) lead to ill-conditioned solutions as in these cases the models are over-specified. Hence, the condition number can act as an indicator for the over-specification of the model.

4.2. A penalty approach for avoiding ill-conditioned $\Phi(\lambda)$

The most obvious way of dealing with ill-conditioning in the fitting of the Nelson–Siegel and the Svensson models is to restrict the parameter space according to the condition number of the matrix $\Phi(\lambda)$. However, this approach is rather inconvenient, as it bears several issues. Whereas the simple relation between condition number and nonlinear parameter may still allow for an adequate

derivation of constraints for the Nelson–Siegel model, see Figure 2(a), it is a fairly demanding task to constrain the parameter space for the Svensson model, see Figure 2(b). Due to the irregularly distributed condition numbers over the parameter space, a suitable restriction only seems possible if the parameter space is modified accordingly, either through transformation or decomposition, or both. In any case, though, the derivation of constraints remains prone to inaccuracies as it presently depends on the visual amenability of the condition number in one or two dimensions. It thus also lacks a theoretical foundation.

Finally, the approach is somewhat inflexible since minor changes in the models, or even the use of other models that share the same separable structure, require the constraints to be readjusted. Because of these reasons, we follow a different approach that deals with the ill-conditioning of the matrix $\Phi(\lambda)$ in a more general way, still ensuring the separability of the problem. The approach relies on a penalization of the objective function if the condition number of $\Phi(\lambda)$ exceeds a maximum allowed level and is described hereinafter.

To penalize large condition numbers in the objective function H of the reduced optimization problem, we consider the function

$$H^{\text{pen}}(\lambda) = \left\| \Phi(\lambda)\Phi(\lambda)^{\dagger}\hat{y} - \hat{y} \right\|_2^2 + \eta \, [\kappa\,(\Phi(\lambda)) - \kappa_{\max}]^{+}, \tag{14}$$

where $\eta > 0$ denotes the weight of the penalization, $\kappa_{\max}$ the maximum condition number whose exceedance is penalized, and $[\,x\,]^{+} = \max\{x, 0\}$.

Adding a penalty term to the objective function avoids optimal solutions being situated in regions with relatively high condition numbers. Because of the direct relation between the nonlinear parameter λ and the condition number $\kappa(\Phi(\lambda))$ in the objective function H^{pen}, the impact of the condition number can be controlled more effectively than for any restriction of the parameter space. This is a particular advantage in case there are no easy-to-identify regions of the parameter space in which the condition number is large, such as for the Svensson model. A further benefit of the approach lies in its flexibility, as it only requires by choosing the weight parameter η and the maximum unpenalized condition number $\kappa_{\max}$.

To determine the maximum unpenalized condition number $\kappa_{\max}$ in the penalization of the objective function H^{pen}, we consider the stability of optimal linear solutions under perturbations of the matrix $\Phi(\lambda^*)$, using the results found in Ref. [24], Section 1.4.3. More specifically, assuming that $\left\|\delta\Phi(\lambda^*)\right\|_2 < \sigma_l(\lambda^*)$ holds to ensure a full and constant rank under perturbation, it follows from formula (1.4.18) that the absolute change in $\beta^*(\lambda^*)$ with respect to small perturbations in $\Phi(\lambda^*)$ can be bounded by

$$\left\|\delta\beta^*(\lambda^*)\right\|_2 \leq \frac{\|\delta\Phi(\lambda^*)\|_2}{\|\Phi(\lambda^*)\|_2}\,\kappa\left(\Phi(\lambda^*)\right)\left[\left\|\beta^*(\lambda^*)\right\|_2 + \frac{\|r(\lambda^*)\|_2}{\|\Phi(\lambda^*)\|_2}\kappa\left(\Phi(\lambda^*)\right)\right], \tag{15}$$

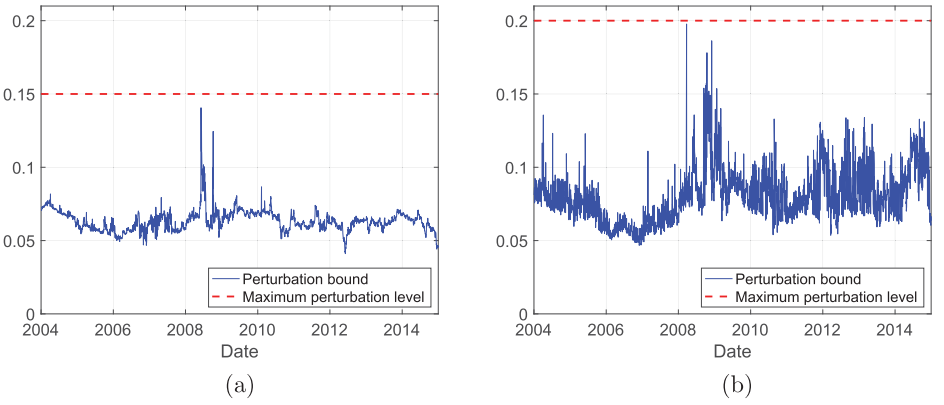

Figure 4. Perturbation bounds according to inequality (15), with $\|\delta\Phi(\lambda^*)\|_2 = \sigma_l(\lambda^*)$ for fitting the Nelson–Siegel and the Svensson models to the dataset described in Section 5.1. The condition number $\kappa_{\max}$ is set to 100 and 180, respectively, corresponding to an acceptable perturbation level of 0.15 and 0.2, respectively. (a) Nelson–Siegel model. (*b*) Svensson model.

where $r(\lambda) = \Phi(\lambda)\Phi^{\dagger}(\lambda)\hat{y} - \hat{y}$ denotes the residual vector[6]. This is a first-order estimate for least squares solutions that can be derived from the normal equations for a perturbed solution by ignoring second-order terms, taking norms, and using the singular value decomposition of $\Phi(\lambda^*)$.

Now, since the perturbation bound on the right-hand side of (15) is mainly influenced by the condition number of $\Phi(\lambda^*)$, it can be controlled to a certain extent by the maximum allowed condition number $\kappa_{\max}$ of H^{pen}. In particular, to avoid situations in which optimal linear solutions are too sensitive to perturbations in $\Phi(\lambda^*)$, the value of $\kappa_{\max}$ should be chosen in such a way that the absolute change in $\beta^*(\lambda^*)$ does not exceed a reasonable level for all fittings in the worst case. For the dataset given in Section 5.1, we therefore fix the maximum acceptable perturbation in the optimal linear solutions at a level of 0.15 and 0.2, respectively. The parameter β is thus not allowed to change by more than 0.15 and 0.20, respectively, to exclude disproportional large movements. This is about five to ten times the average rate level or standard deviation of rates, respectively, cf. Figure 5 for historical zero rate levels and their oscillations. By reverse engineering inequality (15) such that, when solving the sequence of fitting problems, the right-hand side of (15) with the obtained λ^*'s (and $\|\delta\Phi(\lambda^*)\|_2 = \sigma_l(\lambda^*)$) does not exceed 0.15 and 0.20, respectively, these levels then imply that the condition number $\kappa_{\max}$ needs to be set to approximately 100 and 180, respectively, to guarantee reasonably stable and moderate parameters. For the latter values, the time series of perturbation bounds resulting from the fitting of the models to the given data are depicted in Figure 4, along with the corresponding maximum perturbation levels.

[6] Note that we have disregarded perturbations in the data vector $\hat{y}$ and therefore set $\delta\hat{y} = 0$. If deemed relevant, the perturbation bound in (15) can easily be adjusted accordingly.

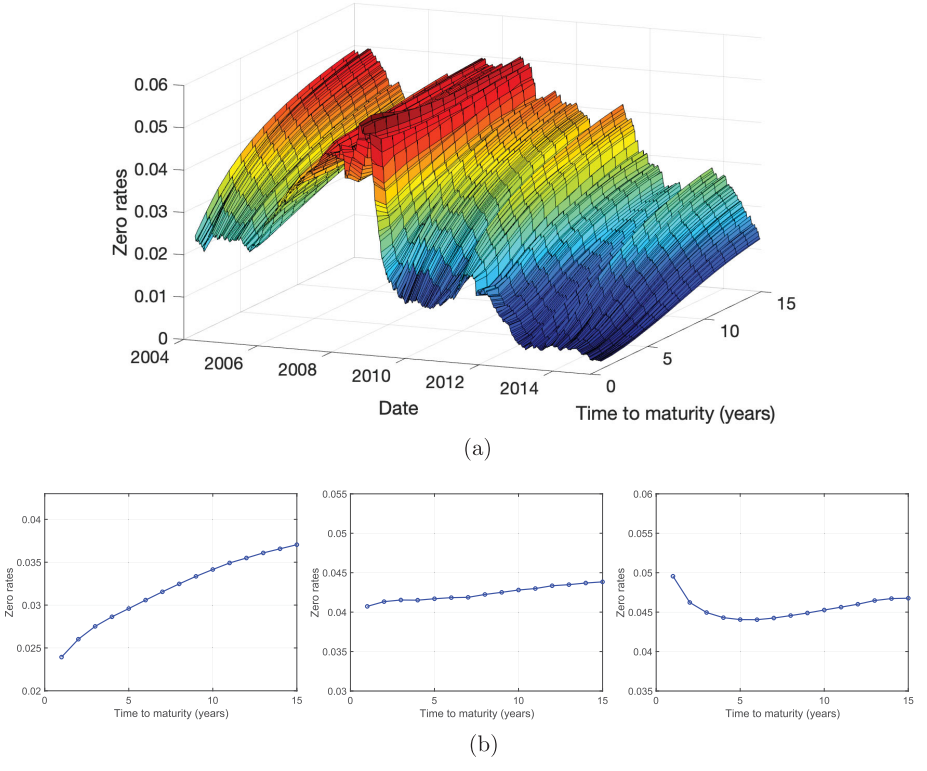

Figure 5. (Selected) market zero rate curves. (a) Market zero rates of every fifth business day from 1 January 2004 to 31 December 2014. (b) From left to right: Examples of upward sloping (as of 13 October 2005), nearly flat (as of 14 February 2007), and inverted market zero rate curves (as of 9 September 2008).

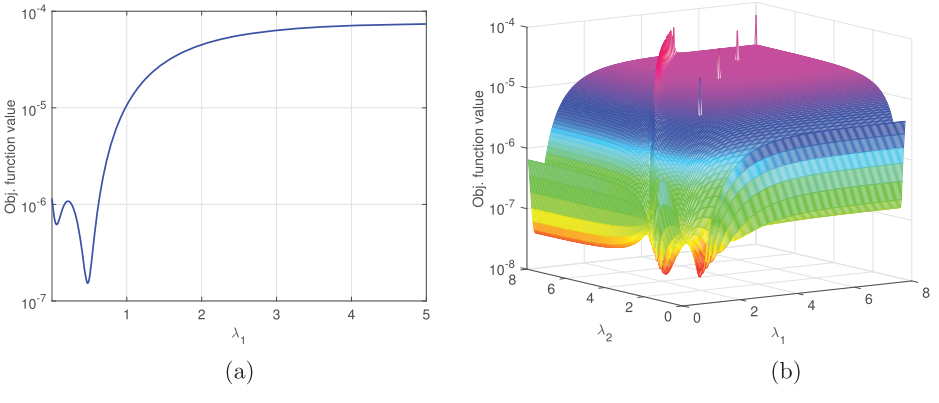

Figure 6. The objective function H for the Nelson–Siegel and the Svensson models for data from 16 March 2004. (a) Nelson–Siegel model. (b) Svensson model.

In contrast to $\kappa_{\max}$, the weight η of the penalization is less relevant for the minimization of the objective function H^{pen}. To keep the resulting function values within a reasonable range in ill-conditioned regions of the parameter space, we have set $\eta = \times 10^{-6}$.

5. Numerical analysis

In this section, we assess the numerical aspects of our method when applied to historical zero rate data. We begin by briefly describing the underlying data that we have used and then illustrate a typical objective function for each of the models. Finally, we conduct a brief computational study in which we show that a) our grid search approach is effective and that b) the penalty approach indeed improves the condition of the problem without a significant deterioration in solution quality.

5.1. Data

To fit the models to some data, we have chosen market zero rates $\hat{y}_1, \ldots, \hat{y}_m$, in order to stay with the original interpretation of the models. However, other kinds of data, like swap rates, etc. can have been used as well, i.e. the models can also be directly fitted to observed swap rate quotations. For our analysis, we have used Euro swap rates which can be converted into the corresponding zero rates by the usual bootstrapping technique. More specifically, we use daily swap par rates with maturities from one to 15 years, which are observed in the time period from 1 January 2004 to 31 December 2014, as obtained from Bloomberg L.P.[7]. The resulting dataset hence consists of 2769 daily zero rate curves with 15 maturities each, to which the models are fitted.

For the given dataset, Figure 5(a) shows the evolution of the zero rate curves over time. It can be observed that the curves vary considerably and assume several different shapes. Apart from the typical upward sloping shapes, nearly flat and inverted zero rate curves can also be found in the dataset, cf. Figure 5(b).

5.2. Illustration of objective function

To provide some qualitative analysis of the objective functions under consideration, we have plotted two representative examples for a selected date in Figure 6, ignoring any penalization term. As can be seen from both subfigures, the objective function H typically has one or two local minima in case the Nelson–Siegel model is fitted, and it usually exhibits between three and five different regions in which the local minima are situated in the case of fitting the Svensson model. Naturally, the exact number of local minima depends on the market data. For

[7] The respective Bloomberg tickers are 'EUSA1 CMPN Curncy', 'EUSA2 CMPN Curncy',..., 'EUSA15 CMPN Curncy', where we use the last quote 'Px_Last' of each day.

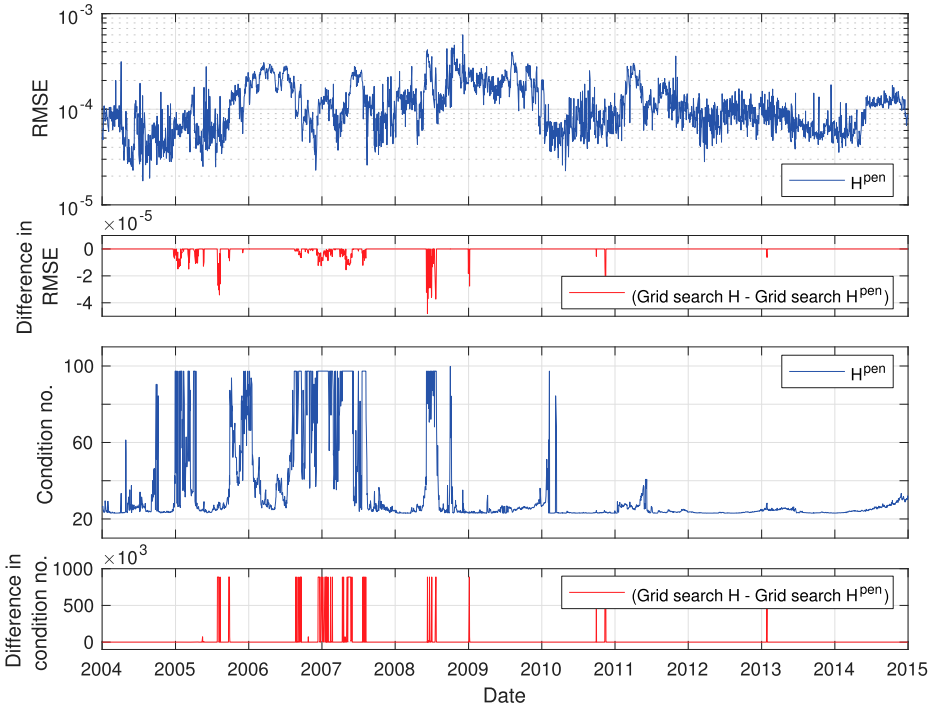

Figure 7. RMSEs and condition numbers obtained by fitting the Nelson–Siegel model to the given data, using a grid search to minimize H and H^{pen}.

certain instances, this number may change, due to adding a penalization term. However, the overall impact on the fitting quality is negligible in these cases, as can be seen in Figures 7 and 8.

From Figure 6, it becomes apparent that for the objective function H, there are different regions in which the objective exhibits different types of behaviour: there are small regions of the parameter space in which H is rather insensitive to any parameter changes and larger regions where changes in the parameter result in considerable differences in the function values. Moreover, local minima are commonly situated in areas of the parameter space that are characterized by narrow and flat valleys. Let us point out again that especially in the Svensson model, local minima often have almost the same objective value which might cause jumps in λ from one region to the other over the course of time.

5.3. Numerical results

In the following, we present numerical results in which we compare how minimizing the objective function H^{pen} compares with minimizing H for a grid search method[8]. The obtained results are then analysed in terms of model fit

[8] We have also tested a multi-start strategy starting a pattern-search method in each grid point of a slightly coarser grid which yields higher accuracy in λ^* at the cost of a higher numerical effort. The main result for the Nelson–Siegel model is that the time series of the optimal λ^* remains quite unchanged. With the exception of a

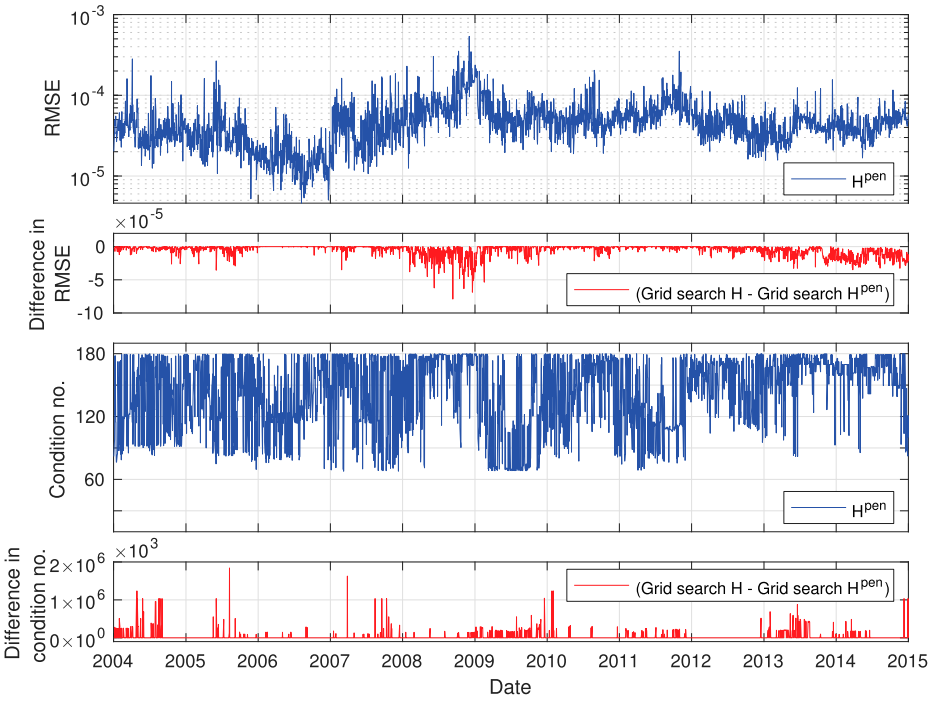

Figure 8. RMSEs and condition numbers obtained by fitting the Svensson model to the given data, using a grid search to minimize H and H^{pen}.

Table 1. Lower (LB) and upper (UB) bounds for fitting the Nelson–Siegel and the Svensson models to the given data using objective functions H and H^{pen}.

(a) Nelson–Siegel model		
	λ_1	
LB	$\times 10^{-3}$	
UB	5	
(b) Svensson model		
	λ_1	λ_2
LB	$\times 10^{-4}$	$\times 10^{-8}$
UB	4	15

and solution quality, where we visualize the time series of fitting errors on a logarithmic scale by using the (monotone) root-mean-square error (RMSE) measure $\sqrt{\frac{1}{m}\widetilde{F}(\lambda^*, \beta^*)}$, with $\widetilde{F}(\lambda^*, \beta^*)$ denoting the minimum objective function value of the approach considered. This can be interpreted as the average error in terms of basis points (bps).

To solve the series of fitting problems with objective functions H and H^{pen} by the considered method, we use the constraints as described in Table 1.

few instances with larger differences, there are only slight changes in most λ^* values. Especially Figure 9 remains mainly unchanged.

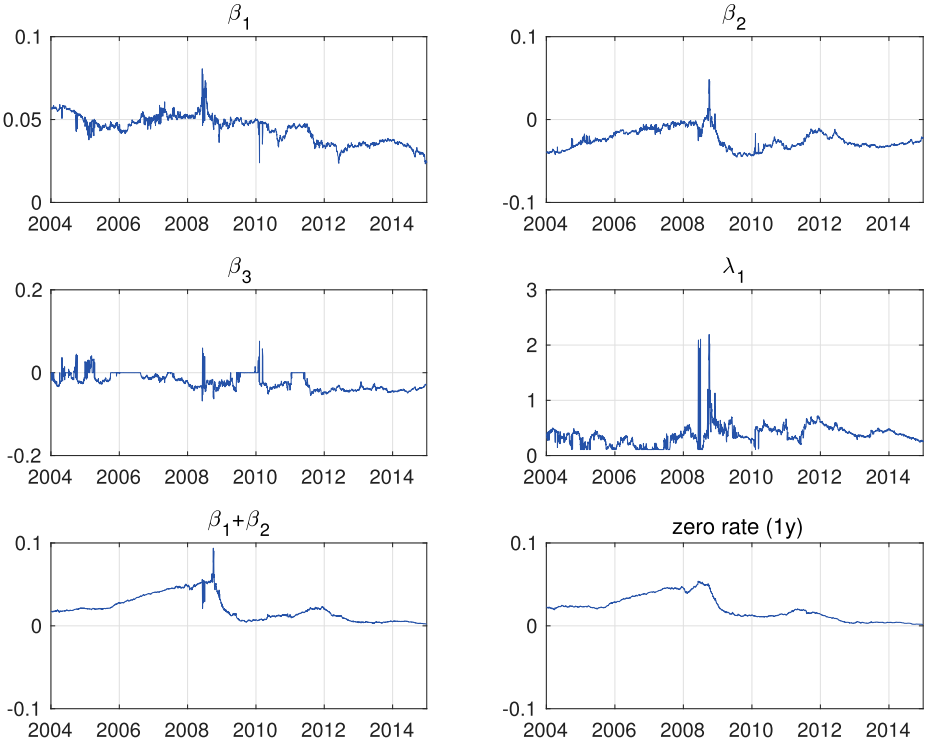

Figure 9. Parameters obtained by fitting the Nelson–Siegel model to the given data, using a grid search to minimize H^{pen}, along with $\beta_1 + \beta_2$ and the estimated 1*y* zero rate.

All numerical computations[9] were carried out in Matlab [27].

We consider a grid search method in which we evaluate each of the objective functions H and H^{pen} at an equidistant grid of points and then take the point with the lowest function value as an optimal solution. The grid size in each dimension is set to 2000 and 250 in the case of the Nelson–Siegel and Svensson model, respectively. Note that it is feasible to use such a fine grid as we have reduced the problem dimension to one or two.

Some interesting straightforward observations can be made based on the results reported in Figures 7 and 8 and Tables 2 and 3:

- The average fit of the Nelson–Siegel model is quite good; on average, the fitting error per tenor is roughly one basis point.
- There are some instances, where the fit is worse than one basis point, and the error exceeds five basis points only in a few instances.
- As expected, the average fit of the Svensson model is even better, with a smaller number of bad instances.

[9] The source code as well as the data and the underlying analysis are available upon request from the first author.

Table 2. Mean and standard deviation of the RSMEs for fitting the Nelson–Siegel and the Svensson models to the given data, using a grid search to minimize H and H^{pen}.

(a) Nelson–Siegel model		
	Mean	Std. Dev.
H^{pen}	1.17×10^{-4}	0.69×10^{-4}
H	1.16×10^{-4}	0.69×10^{-4}
(b) Svensson model		
	Mean	Std. Dev.
H^{pen}	0.50×10^{-4}	0.36×10^{-4}
H	0.45×10^{-4}	0.34×10^{-4}

Table 3. Mean and standard deviation of the run times for fitting the Nelson–Siegel and the Svensson models to the given data, using a grid search to minimize H and H^{pen}.

(a) Nelson–Siegel model		
	Mean	Std. Dev.
H^{pen}	0.1996s	0.0927s
H	0.2119s	0.1134s
(b) Svensson model		
	Mean	Std. Dev.
H^{pen}	7.1114s	1.9349s
H	6.0102s	0.8149s

- Working with H^{pen} instead of H does not impact the error in a significant manner, both for the Nelson–Siegel and the Svensson model. While maintaining the solution quality, a much better condition number can be obtained in a variety of instances.
- We observe that the grid search is fully effective since solutions with very high quality are found. Although it might be possible to refine these solutions further by some local search, the obtained solution quality is already sufficient for practical purposes. Let us emphasize again that jumps between successive λ^* values cannot be avoided due to the non-convex character of the fitting problem in λ.

In accordance with the fitting errors obtained by the grid search, Figures 9 and 10 show the evolution of fitted parameters for the Nelson–Siegel and the Svensson model, respectively. The evolution of fitted parameters of the Nelson–Siegel model looks by and large as expected: small changes in the data from one day to the next typically imply that all fitted parameters do not change much (depending on the conditioning of the problem), as long as the global minimum does not switch from one region to the other (which also happens, especially for the Svensson model). Even though the collections of all fitted parameters of

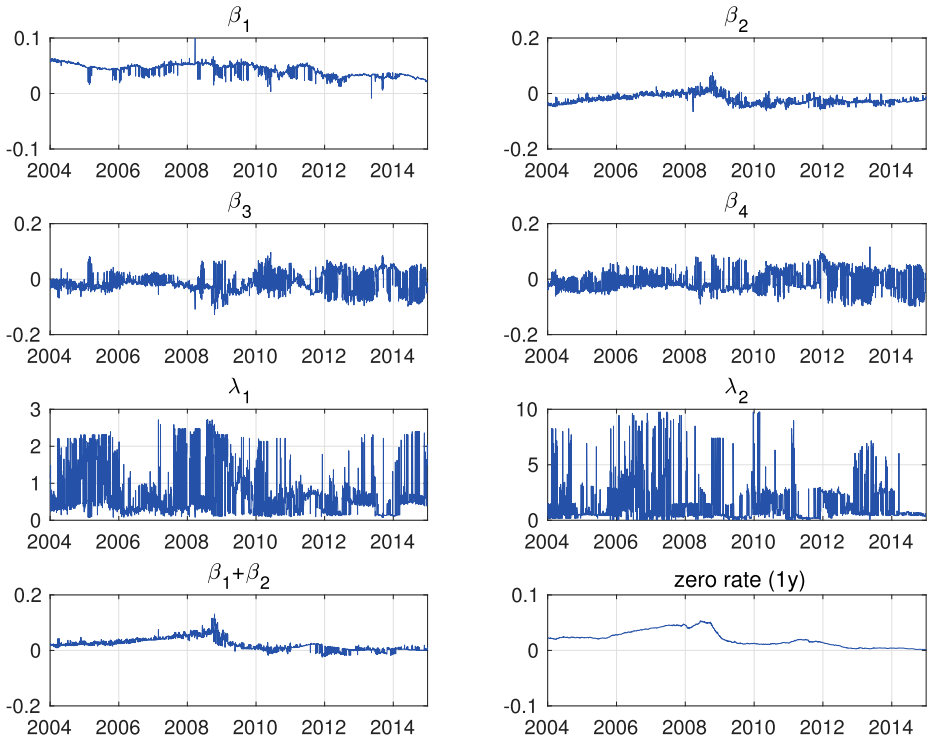

Figure 10. Parameters obtained by fitting the Svensson model to the given data, using a grid search to minimize H^{pen}, along with $\beta_1 + \beta_2$ and the estimated 1*y* zero rate.

the Svensson model in Figure 10 look severely more irregular than in the Nelson–Siegel model at first sight, a closer examination reveals that the parameters actually still behave as just described: Since in the Svensson model there are typically more local minima as in the Nelson–Siegel model and objective values of local minima are more similar to each other in the Svensson model than in the Nelson–Siegel model, more jumps between the regions occur for the Svensson model as for the Nelson–Siegel model.

Whether day-by-day changes of the nonlinear parameter λ (with according to changes in the inner linear parameter β) are acceptable for the application at hand has to be decided on a case-by-case basis by the model user. For example, the models described here might be fully adequate as (compressed) descriptive models of the term structure, but they might not be suitable for econometric analysis or dynamic term structure modelling because of these changes.

6. Conclusion

In this paper, we have presented a mathematical analysis of the fitting of the Nelson–Siegel and the Svensson models to given rates. The analysis is based on the fact that the fitting problem can be formulated as a separable nonlinear

least-squares problem which allows to eliminate the linear model parameters in the objective function and to optimize over the remaining nonlinear parameters. Besides smoothness results for the reduced objective function, our analysis especially shows that the fitting of the Nelson–Siegel and the Svensson models may become ill-conditioned in certain regions of the parameter space. To deal with this issue and to avoid estimated parameters become too sensitive with respect to perturbations, we propose to penalize the least-squares objective function if the condition number of the matrix of basis functions exceeds a pre-specified level. This reformulation leads to an easy and effective handling of ill-conditioning. Numerical results based on market data indicate that our proposed penalization method substantially improves the solution in terms of robustness while maintaining solution quality.

By our analysis, the strengths and weaknesses of both models have been made more transparent, especially with respect to econometric aspects of the fitting.

Notes

1. For $l = 3$, $\lambda \in \mathbb{R}_+$, we obtain the Nelson–Siegel model, whereas for $l = 4$, $\lambda \in \mathbb{R}_+^2$, we obtain the Svensson model. To allow for larger values of l, further basis functions would need to be defined. Our subsequent analysis is specifically tailored for the Nelson–Siegel and Svensson models, but can partially be generalized to larger models as well.
2. To take care of potential heteroscedasticity in the data, positive weights may additionally be included in the objective function F. The latter may then be formulated as $F(\lambda, \beta) := \|W(\Phi(\lambda)\beta - \hat{y})\|_2^2$, where the elements of the weight matrix $W = \operatorname{diag} w_1, \ldots, w_l$ are typically set equal to the reciprocals of the variances of the residuals, which may be estimated, for instance, from historical data (e.g. Ref. [18]). However, for ease of exposition, we do not use any weights in our analysis and the numerical calculations.
3. An example was provided by Gimeno and Nave [22] for fitting the Svensson model, albeit with a different objective function. They reported that extreme and often offsetting optimal values for the linear parameters β_3 and β_4 occur whenever the corresponding nonlinear parameters λ_1 and λ_2 are similar to each other.
4. If models are used where perturbations to $\hat{y}$ turn out to be relevant, the following analysis can be extended by adjusting the condition number to include $\hat{y}$, see, e.g. Ref. [24], Subsection 1.4.3.
5. For the available data described in Section 5.1, we have $m = 15$ and $\tau = (1, 2, \ldots, 15)^\top$.
6. Note that we have disregarded perturbations in the data vector $\hat{y}$ and therefore set $\delta\hat{y} = 0$. If deemed relevant, the perturbation bound in (15) can easily be adjusted accordingly.
7. The respective Bloomberg tickers are 'EUSA1 CMPN Curncy', 'EUSA2 CMPN Curncy',..., 'EUSA15 CMPN Curncy', where we use the last quote 'Px_Last' of each day.
8. We have also tested a multi-start strategy starting a pattern-search method in each grid point of a slightly coarser grid which yields higher accuracy in λ^* at the cost of a higher numerical effort. The main result for the Nelson–Siegel model is that the time series of the optimal λ^* remains quite unchanged. With the exception of a few instances with larger differences, there are only slight changes in most λ^* values. Especially Figure 9 remains mainly unchanged.
9. The source code as well as the data and the underlying analysis are available upon request from the first author.

Acknowledgments

The authors would like to thank all involved referees for the efforts and their valuable feedback.

Disclosure statement

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

Funding

This work was supported by the Engineering and Physical Sciences Research Council [grant no. EP/M50662X/1], DEVnet GmbH, and the IRIDIS High-Performance Computing Facility and associated support services at the University of Southampton.

References

[1] Nelson CR, Siegel AF. Parsimonious modeling of yield curves. J Bus. 1987;60(4):473–489. doi: 10.1086/jb.1987.60.issue-4
[2] Svensson LEO. Estimating forward interest rates with the extended Nelson and Siegel method. Sveriges Riksbank Quart Rev. 1995;3:13-–26.
[3] BIS. Zero-coupon yield curves: technical documentation [BIS Paper No. 25, Bank for International Settlements]. 2005.
[4] ECB. The New Euro Area Yield Curves [Monthly Bulletin, February 2008, European Central Bank]. 2008.
[5] Schich ST. Schätzung der deutschen Zinsstrukturkurve [Discussion Paper Deutsche Bundesbank, Series 1, 04/1997]. 1997.
[6] Lakhany A, Pintar A, Zhang A. Calibrating the Nelson-Siegel-Svensson model by genetic algorithm. Available from: https://arxiv.org/abs/2108.01760. 2021.
[7] Kondratyev A. Learning curve dynamics with artificial neural networks. Available at SSRN: https://ssrn.com/abstract = 3041232. 2018.
[8] Sokol A. Autoencoder market models for interest rates. Available at SSRN: https://ssrn.com/abstract = 4300756. 2022.
[9] Cairns A, Pritchard D. Stability of descriptive models for the term structure of interest rates with application to German market data. Br Actuar J. 2001;7(3):467–507. doi: 10.1017/S1357321700002439
[10] Fabozzi FJ, Martellini L, Priaulet P. Predictability in the shape of the term structure of interest rates. J Fixed Income. 2005;15(1):40–53. doi: 10.3905/jfi.2005.523089
[11] Diebold FX, Rudebusch GD, Aruoba SB. The macroeconomy and the yield curve. J Econom. 2006;131(1):309–338. doi: 10.1016/j.jeconom.2005.01.011
[12] De Pooter M. Examining the Nelson-Siegel class of term structure models: In-sample fit versus out-of-sample forecasting performance [Discussion Paper 2007-043/4, Tinbergen Institute, Erasmus University]. 2007.
[13] Diebold FX, Li C. Forecasting the term structure of government bond yields. J Econom. 2006;130:337-–364. doi: 10.1016/j.jeconom.2005.03.005
[14] De Rezende RB. Giving flexibility to the Nelson-Siegel class of term structure models. Revista Brasileira De Finanças. 2011;9(1):27–49.
[15] Gilli M, Grosse S, Schumann E. Calibrating the Nelson-Siegel-Svensson model [COMISEF Working Paper Series No. 31]. 2010.
[16] Diebold FX, Yield curve modeling and forecasting: the dynamic Nelson-Siegel approach. Princeton, NJ: Princeton University Press; 2013.

[17] Atkins R. Europe Shows Negative interest rates not absurd – and might work. The Financial Times Limited. 2014 Sep 18. Available from: http://www.ft.com/cms/s/0/db1f5da4-3e89-11e4-a620-00144feabdc0.html#axzz41Cvbt4Ux. 2014.
[18] Strutz T. Data fitting and uncertainty. Wiesbaden, Germany: Vieweg+Teubner Verlag; 2011.
[19] Virmani V. On estimability of parsimonious term structure models: an experiment with the Nelson-Siegel specification. Appl Econ Lett. 2012;19:1703–1706. doi: 10.1080/13504851.2012.657343
[20] Gauthier G, Simonato JG. Linearized Nelson-Siegel and svensson models for the estimation of spot interest rates. Eur J Oper Res. 2012;219(2):442–451. doi: 10.1016/j.ejor.2012.01.004
[21] Annaert J, Claes AGP, De Ceuster MJK. A ridge regression approach. Int Rev Econ Finance. 2013;27:482–496. doi: 10.1016/j.iref.2013.01.005
[22] Gimeno R, Nave JM. A genetic algorithm estimation of the term structure of interest rates. Comput Stat Data Anal. 2009;53(6):2236–2250. doi: 10.1016/j.csda.2008.10.030
[23] Angelini F, Herzel S. Consistent initial curves for interest rate models. J Derivat. 2002;9(4):8–17. doi: 10.3905/jod.2002.319182
[24] Björck A. Numerical methods for least squares problems. Philadelphia (PA): Society for Industrial and Applied Mathematics; 1996.
[25] Golub GH, Pereyra V. The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. SIAM J Numer Anal. 1973;10:413-–432. doi: 10.1137/0710036
[26] Demmel JW. Applied numerical linear algebra. Philadelphia (PA): Society for Industrial and Applied Mathematics; 1997.
[27] MATLAB. Version 9.12 (R2022a). Natick (MA): The MathWorks Inc.; 2022.