NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Victor Lapshin, Sofia Sokhatskaya

CHOOSING THE WEIGHTING COEFFICIENTS FOR ESTIMATING THE TERM STRUCTURE FROM SOVEREIGN BONDS

BASIC RESEARCH PROGRAM
WORKING PAPERS

SERIES: FINANCIAL ECONOMICS
WP BRP 73/FE/2018

Victor Lapshin[1], *Sofia Sokhatskaya*[2]

CHOOSING THE WEIGHTING COEFFICIENTS FOR ESTIMATING THE TERM STRUCTURE FROM SOVEREIGN BONDS[3]

Estimates of the term structure of interest rates depend heavily on the quality of the market data from which it is constructed. Estimated rates can be incorrect because of observation errors and omissions in the data. The usual way to deal with the heteroscedasticity of observation errors is by introducing weights in the fitting procedure. There is currently no consensus in the literature about the choice of such weights. We introduce a non-parametric bootstrap-based method of introducing observation errors drawn from the empirical distribution into the model data, which allows us to perform a comparison test of different weighting schemes without implicitly favoring one of the contesting models – a common design flaw in comparison studies. We use government bonds from several countries with examples of both liquid and illiquid bond markets. We show that realistic observation errors can greatly distort the estimated yield curve. Moreover, we show that using different weights or other modifications to account for observation errors in bond price data does not always improve the term structure estimates, and often worsens the situation. Based on our comparison, we advise to either use equal weights or weights proportional to the inverse duration in practical applications.

JEL Classification: E43.

Keywords: term structure of interest rates, zero-coupon yield curve, bond prices, weights, cross-validation.

[1] National Research University Higher School of Economics. E-mail: vlapshin@hse.ru

[2] National Research University Higher School of Economics.

[3] Support from the Basic Research Program of the National Research University Higher School of Economics is gratefully acknowledged

1 Introduction

Estimating the term structure of interest rates is a subject of great importance in financial economics, particularly in asset pricing, risk management, and long-dated liabilities such as pensions and life insurance. Forward rates may serve as indicators of monetary policy.

The resulting estimates of the interest rate curve depend heavily on the quality of the market data used. There is extensive literature introducing various methods of accounting for observation errors in the data. For instance, the most popular method is to weight the data by inverse bond duration assuming that longer maturities (and longer durations), correspond to larger price errors (Ahi et al., 2016; Bliss, 1997; Cruz-Marcelo et al., 2011; Gimeno and Nave, 2009). However, the efficiency of these methods has not been investigated separately. Up until now, there was no comparison of the numerous methods, therefore the choice of the method to account for observation errors was usually ad hoc and without an evidence-based rationale.

The purpose of this research is twofold. Our first objective is to demonstrate how errors in the input data can affect the estimates of the term structure of interest rates – by using data on government bonds from several different bond markets (in the alphabetical order): English, French, German, Greek, Italian, Portuguese, Russian, and Spanish. Our second objective is to check whether conventional methods of accounting for data errors actually improve the estimation results.

For estimating the term structure of interest rates, we use the two most popular methods: the Svensson model (a parametric method) and cubic smoothing splines for the spot rate (a non-parametric method). For both approaches, least squares with different weights are applied. We propose a non-parametric method of introducing observation errors drawn from the empirical distribution into the model data. In addition, we perform a real-data-based comparison based on estimating the posterior predictive power. We estimate it via a leave-out-one cross-validation.

The paper proceeds as follows: Section 2 provides the literature review. Section 3 and 4 give details of the data and the comparison methodology. Section 5 discusses the empirical results. Section 6 concludes.

2 Literature review[4]

Laurini and Ohashi (2015) investigated the reasons for the difference between yields to maturity and estimated forward rates, which was observed in many empirical studies. The authors considered that this discrepancy is due to data errors that led to a shift in the constructed term structure of interest rates. The causes of data errors lie in microstructural market effects.

Ahi et al., (2018) used weighted least squares using the inverse duration, assuming that longer maturities (and longer durations), correspond to larger price errors. That is, they solve the following optimization problem:

$$\sum_{\mathrm{i}=1}^{\mathrm{N}} w_i \left(P_i - \hat{P}_i(\theta)\right)^2 \to \min_{\theta} \quad ,$$

$$w_i = \frac{1/D_i}{\sum_{j=1}^{N}\left(1/D_j\right)} \ ,$$

where D_i is the Macaulay duration of bond i, $\widehat{P}_i(\theta)$ is the model price for bond i and the parameter vector θ, P_i is the observed price for bond i.

Bliss (1997) compared several weighting methods. The best of them were weighting by term to maturity and by duration. However, the author asserted that the use of duration was more reasonable, since more frequent errors in market data were associated with a longer duration rather than with a long maturity period. Similar weights were used by Cruz-Marcelo et al. (2011) based on the assumption that the bond price data with shorter maturity were more reliable.

Gimeno and Nave (2009) stated that small changes in the yields of short-term bonds caused more significant changes in the yield curve than similar changes in the yields of long-term securities. They used weights inversely proportional to the square root of the duration:

$$w_i = \frac{1}{\sqrt{D_i}}.$$

Ioannides (2003) weighted the data using the inverse square of the duration:

$$w_i = \frac{1}{D_i^2}.$$

Bliss (1997) suggested that closing prices are calculated erroneously with respect to bid and ask prices. Therefore, for constructing a yield curve the author used bid and ask prices instead of the closing prices. The least squares functional employed in the paper only penalized prices outside the bid-ask interval:

[4] Note that the objective of the paper is not to propose yet another term structure fitting model. Instead, we study the effect of observation errors and the efficiency of various approaches to dealing with them. Therefore, we just use the most widespread basic term structure estimation techniques – one parametric and one spline-based. Our literature review is limited to the problem we're investigating, i.e. methods used to deal with observation errors.

$$\sum_{i=1}^{N}(w_i e_i)^2,$$

where e is the discrepancy term, which is non-zero only if the fitted price $\hat{P}$ lies outside the bid-ask spread:

$$e = \begin{cases} \hat{P} - P^{Ask}, & \hat{P} > P^{Ask} \\ P^{Bid} - \hat{P}, & \hat{P} < P^{Bid} \\ 0, & \text{otherwise.} \end{cases},$$

Note that this approach can easily be combined with any weighting scheme.

Hladikova and Radova (2012) compared several methods of weighting: by the number of bonds, by the reverse duration, by the inverse bid-ask spread and by the sum of the inverse duration and the inverse spread using data on Czech government bonds. To assess the performance of the weighting schemes, the authors used the MSE for prices for yields to maturity, weighted MSE for the smoothness of the constructed functions, and the stability of the model coefficients when one bond was excluded from the sample. Different indicators gave preference to different weighting methods, so it was not possible for the authors to select the best weighting scheme using their criteria.

Thus, different authors have introduced different methods, which, in their opinion, should help reduce the effect of data errors. However, studies of their performance faced two main obstacles. The first concerns relative performance. One should be very careful about choosing the comparison criteria, since it's very easy to accept a comparison criterion the construction of which favors one of the methods. Since the methods differ by the functionals they are optimizing, choosing a specific comparison criterion might easily introduce a bias towards the method which optimizes this exact functional (e.g. a simple non-weighted sum-of-squares will by construction penalize all weighted schemes and favor a non-weighted scheme, because the latter optimizes the same functional that is used to compare the models). To overcome this, we use two comparison criteria. One measures the discrepancy in the whole estimated yield curve rather than estimated bond prices, thus avoiding potential bias. The other tests the out-of-sample predictive power of the models via a leave-out-one cross-validation. We describe these criteria in detail later.

The second obstacle comes from the fact that many comparison criteria require knowing the 'true' bond prices without noise. This is impossible using real data, so model data must be used with artificially introduced errors. However, choosing a model distribution for observation errors presents a challenge since the true distribution is very complex and the performance of the methods greatly depends on it.

This problem has been given some consideration in the literature. The following is a review of the different approaches to introducing simulated observation errors into model data.

Error modeling

Bliss (1997) analyzed the ex-post errors of fitting spline models and an extension of the Nelson-Siegel model. He found that there is a significant interdependence of the errors in time and between models. For instance, if the error was positive in one model, the error for the same bond for the next day and the error for the same bond in another model will likely be positive too. The author also regressed the errors on various bond features: maturity, age (number of years after the issuance), time, trade volume, bid-ask spread and some calculated indicators. All regressors were significant, however, some coefficients (maturity and volume) had different signs for different models, which, considering the dependence of errors between models, was alarming. An empirical connection between the fitting errors and bond duration was also found by Fleming and Whaley (1994) and Carcano and Nicola (2011).

Several researchers conducted tests for robustness to possible errors in the input data via the scheme in Figure 1. Term structure is estimated from the initial data. Then randomly generated errors are added to the original data. The distribution of errors is chosen according to the assumptions on the real fitting errors. Then new yield curves are constructed based on the data with errors. Estimation errors for all maturities are usually calculated as the difference between the initial (true) yield curve and the curve estimated from noisy data. After repeating the procedure several times, the resulting sets of estimation errors are aggregated – usually their mean and variance is calculated.

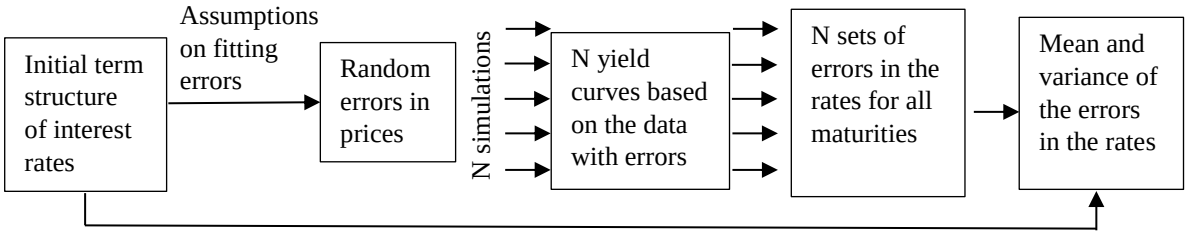

Figure 1. A flowchart for checking the resistance to errors in the input data.

In comparing five methods for estimating the term structure, Jordan and Mansi (2003) investigated the effect of interpolation and observation errors on the estimate. For this purpose, the authors generated model spot rate term structures and added normally distributed errors to the calculated model prices. They assumed a theoretical error term that was homoscedastic in yields, but heteroscedastic in prices:

$$P_i^e = P_i + \varepsilon_i, \qquad \varepsilon_i = \sigma \left(\frac{dP}{dy}\right)_i \eta_i, \qquad \eta_i \sim N(0,1).$$

To generate errors, they used the following procedure.

Step 1: calculate the sensitivity of bond prices with respect to their returns:

$$\left(\frac{dP}{dy}\right)_i = MD_i \cdot P_i,$$

where MD_i is the modified duration of bond i.

Step 2: draw η_i from the standard normal distribution.

Step 3: calculate σ so that the modelled observation error variance $\sum_{i=1}^{n}(\sigma P_i MD_i \eta_i)^2$ is equal to the estimated actual error variance S (exogeneous):

$$\sigma = \frac{S}{\sum_{i=1}^{n}(P_i MD_i \eta_i)^2}.$$

The interpolation error was found to be a significant factor, not dominated by the observation error. Two methods from five, the Nelson and Siegel and the Mansi and Phillips models, stood out in terms of good interpolation properties and robustness with respect to observation errors.

Bertocchi et al. (2000) built a yield curve based on the Black-Derman-Toy model (BDT). They conducted a similar study of the model sensitivity with respect to random changes in the initial term structure. Prices with errors were simulated by adding normally distributed random perturbations to the 'true' prices. It turned out that errors due to small perturbations of the yield curve propagated to large errors in the short-term interest rates. The acceptable magnitude of perturbations for the optimal solution to be stable was found to be very small.

For constructing a Bayesian hierarchical model, Cruz-Marcelo et al. (2011) assumed a normal distribution of bond prices with the mean equal to the sum of the discounted cash flows and the variance directly proportional to the bond duration.

Laurini and Ohashi (2015) and Ubukata and Fukushige (2009) modeled errors with an autoregressive process with normally distributed innovations, the coefficients for which were taken from empirical studies. The authors also assumed a non-zero correlation between the errors for different bonds, which were also calculated from the observed fitting errors. Bolder (2006) also assumed a normal error distribution.

Another kind of the error distribution was used by Ahi et al. (2018) and Bliss et al. (2011). Ahi et al. used the Svensson model to construct a yield curve for four bond markets: American, Brazilian, Mexican and Turkish. The goals were to compare various optimization methods, but they also checked the models' robustness with respect to data errors.

The errors they introduced into bond prices were uniformly distributed within the bid-ask interval:

$$\varepsilon_i = \left(P_{ask,i} - P_{bid,i}\right) * \varphi_i\,,$$
$$\varphi_i \sim U\,[0,1].$$

Thus, the methods of error simulation proposed in the literature are diverse and involve both different types of distributions and different dependencies on parameters such as duration and spread. Unfortunately, real data errors are unobservable, therefore it is impossible to verify these assumptions, however, it is generally considered highly unlikely that observation errors are, in fact, normally distributed. Moreover, introducing normally distributed errors with equal variances implicitly favors the ordinary least squares estimation method (it being the maximum likelihood estimate for this observation error setting). The homoscedastic-in-yields scheme described above implicitly favors weighted least squares with the weights being inversely proportional to the durations.

To overcome this obstacle, we introduce a non-parametric bootstrap-like method of sampling observation errors from the empirical distribution while approximately preserving both correlations and marginal distributions.

3 Data

To ensure accurate results we use market quotations of government bonds of Russia and several countries of the Eurozone from 2016 to the end of 2017 (data from 2017 is used directly in the construction of the yield curve, data from 2016 is needed for auxiliary calculations) obtained from Bloomberg. The sample includes only bonds denominated in the home currency with fixed coupons and without embedded options. We exclude bonds with the maturity less than three months at the settlement date and illiquid bonds (those with less than 10 trading days for 2017). The total sample includes 70 English bonds, 145 French bonds, 102 German bonds, 42 Greek bonds, 105 Italian bonds, 24 Portuguese bonds, 30 Russian bonds, 55 Spanish bonds. While the dataset is homogeneous for each country as there no significant changes in the yield curves during the period considered, we get enough observations for our research. Including several countries in the sample allows us to compare the results for bond markets of different liquidity.

4 Methodology

Basic Models

For each country, we select all trading days from 01 Jan 2017 to 31 Dec 2017 with price data for at least 10 bonds. For each of these days, at market closing prices, zero-coupon yield curves are constructed using two methods.

1) A parametric method (Svensson model) defining the spot rate $r(t,\beta)$ for the term to maturity t and the parameter vector $\theta = (\beta_0, \beta_1, \beta_2, \beta_3, \tau_1, \tau_2)^T$:

$$r(t,\theta) = \beta_0 + \beta_1 \frac{1-e^{(-t/\tau_1)}}{t/\tau_1} + \beta_2 \left(\frac{1-e^{(-t/\tau_1)}}{t/\tau_1} - e^{(-t/\tau_1)} \right) + \beta_3 \left(\frac{1-e^{(-t/\tau_2)}}{t/\tau_2} - e^{(-t/\tau_2)} \right).$$

The model parameters are determined from minimizing the least squares functional:

$$\sum_{i=1}^{N} \left(P_i - \hat{P}_i(\theta) \right)^2 \to \min_{\theta} \quad ,$$

where $\hat{P}_i(\theta)$ is the model price of bond *i* equal to the sum of the discounted cash flows $C_i(t_j)$ promised at the corresponding terms $t_j, j = 1..M_i$ with parameter vector $\theta = (\beta_0, \beta_1, \beta_2, \beta_3, \tau_1, \tau_2)^T$, and P_i is its observed price:

$$\hat{P}_i(\theta) = \sum_{j=1}^{M_i} C_i(t_j) e^{-r(t_j,\theta)\cdot t_j}$$. A spline method. We used

$$r(t,\theta) = \begin{cases} r_1(t) = a_1 + a_2 t + a_3 t^2 + a_4 t^3, & t \in [0, T_1] \\ r_2(t) = b_1 + b_2 t + b_3 t^2 + b_4 t^3, & t \in [T_1, T_2] \\ \vdots & \\ r_M(t) = k_1 + k_2 t + k_3 t^2 + k_4 t^3, & t \in [T_{M-1}, T] \end{cases},$$

where $\theta = (a_1, a_2, a_3, a_4, b_1, \ldots, k_4)^T$ is the parameter vector. The optimization functional also includes a regularization smoothness term:

$$\sum_{i=1}^{N} \left(P_i - \hat{P}_i(\theta) \right)^2 + \alpha \int_0^T \left(r''(\tau,\theta) \right)^2 d\tau \to \min_{\theta} \quad ,$$

where α is the regularization parameter. The second derivative of the spot rate with respect to the term to maturity $r''(t,\theta)$ is used as the criterion for smoothness to avoid overfitting. The usual problem with smoothing splines is choosing the right amount of smoothing α. If α is too low, the

yield curve will fluctuate strongly trying to fit the noisy data but a high smoothing coefficient increases the approximation errors $P_i - \hat{P}_i(\theta)$ and worsens the fit.

We choose the smoothness coefficient so that the sum of the squares of the fitting errors for the spline method are approximately equal to the same quantity for the Svensson model. To deal with outliers, we clip the Svensson fitting error to its 95% quantile from above for this purpose as follows. If the sum of the squares of the fitting errors for a given date is less than its 95% quantile for all dates, then the sum of the squares of the spline fit errors should be approximately equal. If, however, the sum of the squared fitting errors is greater than the 95% quantile, then the spline model is tuned to produce the same sum of error squares as the 95% quantile. This method of selecting the smoothness coefficient is used for all variants of the spline model.

We used splines for the spot rate curve, the forward rate curve and the discount function. The results are almost identical, so we report only one of these settings – the spot rate. Our results do not really depend on the number of knot points for the spline (in the range from 10 to 40 knot points), since most of the regularization is done via the regularization coefficient α. We use 12 knot points in our spline model, from one day to the longest maturity.

Dealing with errors

For both approaches (parametric and non-parametric), several variations of the method of least squares with different weights are used:

$$\sum_{i=1}^{N} (w_i \varepsilon_i(\theta))^2 \to \min_{\theta} \quad \#(1)$$

for the parametric model and

$$\sum_{i=1}^{N} (w_i \varepsilon_i(\theta))^2 + \alpha \int_0^T \left(r''(\tau, \theta)\right)^2 d\tau \to \min_{\theta}$$

for the spline model, where $\varepsilon_i(\theta) = P_i - \hat{P}_i(\theta)$ are the fitting errors, w_i are the weights to be chosen, where θ is the parameter vector of the respective model.

We apply the most commonly used methods of accounting for errors and introduce several new ones, which could improve the results of estimating the yield curve.

1. Standard model – unweighted least squares (‘Standard’):

$$w_i = 1 .$$

This is the basic model with which we compare all the subsequent modifications. Note that due to the particular form of the optimization problem (1) above, the optimal yield curve depends on weights w_i up to a multiplicative constant. Therefore, we do not normalize the weights in any way.

2. Inverse duration weights (‘1/D’):

$$w_i = \frac{1}{D_i}.$$

The most commonly used weighting scheme in studies, based on the assumption that the longer the maturities (and correspondingly higher duration), the larger the observation errors in the data.

3. Duration weights ('D'):

$$w_i = D_i .$$

Reverses the previous scheme.

4. Inverse bid-ask spread weights ('1/S'):

$$w_i = \frac{1}{S_i}.$$

It is logical to assume that observation errors should be greater in the prices of less liquid bonds. The bid-ask spread is often considered as a liquidity indicator (for liquid securities the spread is smaller). Weighting by the inverse spread could possibly improve the results by assigning more weight to more liquid bonds.

5. Penalizing atypical spreads ('1/log(S/MS)').

An improvement of the previous scheme – we recognize that various bonds have different typical spreads and penalize atypical spreads (compared to the average spread for the previous 180 days) – because both unusually large and unusually small bid-ask spreads could indicate a departure from equilibrium state:

$$w_{it} = \frac{1}{1 + \left|\log\left(\frac{S_{it}}{MS_{it}}\right)\right|}.$$

where MS_{it} is the average spread for bond *i* for the previous 180 days:

$$MS_{it} = \frac{1}{180}\sum_{h=t-181}^{t-1} S_{ih} .$$

6. An improvement of the previous scheme. Instead of an arbitrary penalty for unusual spreads, we construct an empirical probability density function and assign weights according to the likelihood of the bid-ask spread data ('HIST'):

   Step 1: We calculate a histogram of $\log\left(\frac{S_{it}}{MS_{it}}\right)$ for the bond in question for the previous six months (this histogram updates daily).

   Step 2: For the current spread value S_{it}, we calculate $\log\left(\frac{S_{it}}{MS_{it}}\right)$ and its corresponding empirical frequency calculated during Step 1. This is taken as the weight for bond *i*.

Note that this approach can assign exactly zero weights if the observed spread is atypical enough.

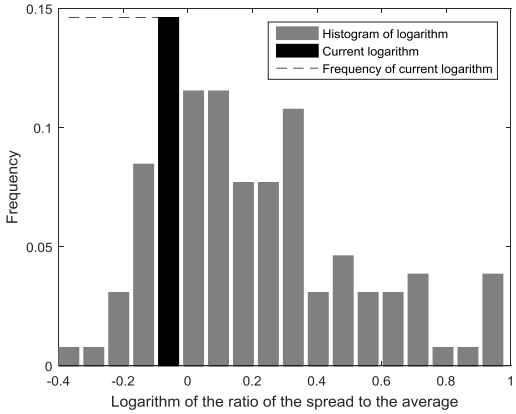

Figure 2. Determination of the probability weight with which the logarithm of the ratio of the spread to the average takes the current value

7. Only penalize model prices outside the observed bid-ask spread ('Bid-Ask'):

$$\sum_{i=1}^{n} e_i^2 \to \min_{\theta} \quad ,$$

where

$$e_i = \begin{cases} \widehat{P}_i(\theta) - P_i^{Ask}, & \widehat{P}_i(\theta) > P_i^{Ask}, \\ P_i^{Bid} - \widehat{P}_i(\theta), & \widehat{P}_i(\theta) < P_i^{Bid}, \\ 0, & \text{otherwise.} \end{cases}$$

This model is used by some authors on the basis that closing prices can possibly carry erroneous information about the form of the yield curve. Consequently, they stipulate that it is better to use bid and ask quotations.

Comparison Methodology

We compare the estimation methods described above using two approaches to measuring the fitting quality. The fitting quality could refer to either the accuracy of the resulting term structure estimate or the accuracy of the predicted bond prices. In theory, they could be very different – with abundant data (many bonds), weighting would not impact the term structure estimate, however, if the model incorrectly gives more weight to the bonds with more variable quotes, then their predicted price distribution would have less variance and might be way off the real data. Therefore, for choosing the weighting scheme, we recommend cross-validation, but we also report the quality of the term structure fitting.

The two approaches are described in detail in what follows. A high-level summary of their differences is presented in Table 1.

Table 1. Comparison of two tests

	Bond Price Fitting	**Term Structure Fitting**
Fitting quality definition	Expected log-likelihood of new (out-of-sample) bond quotes.	The difference between the estimated zero-coupon yield curve and the 'true' one.
Optimal model	Correctly identifies observation error distribution.	Correctly identifies the term structure – regardless of observation errors.
Quality compared using	Leave-out-one cross-validation on the real data.	Model data with model errors generated with a special algorithm.
Useful for	Estimating term structure for further bond-related work: pricing bonds, estimating effects of various factors on bond prices, etc.	Estimating the term structure for uses not directly connected with bonds (discounting, pricing derivatives, monetary policy).

Comparison of Term Structure Fitting

The main problem is designing a simulation scheme so that the random observation errors introduced into bond prices do not favor one estimation method over the others. For example, if we use i.i.d. normal errors, then the best way to estimate the yield curve is to use equal weights. Thus, the best estimation scheme is actually determined by the assumed distribution of observation errors. Once we assume a specific form for the observation error distribution, there is no need for a comparison of various estimation methods. One can easily devise a method tailored specifically for this kind of distribution (e.g. maximum likelihood estimate).

Unfortunately, the true error distribution is unknown. Therefore, we use a specially designed bootstrap-based procedure to make sure that the distribution of the modeled errors is as close to the real distribution as possible. The comparison thus makes sense since we do not assume any kind of error distribution *a priori*.

Step 1: For each estimation method k and for each date t, yield curves $r_{t,k}(\cdot)$ are estimated and the model prices $\hat{P}_{i,t,k}$ are calculated. For each date t, we randomly choose the base ('true') yield curve for that date $r_{t,base}(\cdot)$ at random from all estimated yield curves $r_{t,k}(\cdot)$ for that date.

The sole purpose of this step is to choose a realistic 'true' yield curve for further experiments.

Step 2: For each bond i for each date t and for each model k, a set of errors is formed $\varepsilon_{i,t,k}$:

$$\varepsilon_{i,t,k} = P_{i,t} - \hat{P}_{i,t,k},$$

where $P_{i,t}$ are the observed bond prices. $\varepsilon_{i,t,k}$ are the fitting errors on date t for all bonds traded on that day. We assume that this sample is representative of the true observation errors at least in the sense of the marginal distribution (we filter out outliers for each bond individually).

The above steps are only performed once. The next steps are designed to replicate the correlations between observation errors for different bonds and are performed each time we need to simulate a noisy set of bond prices (i.e. for every date *t*, but nothing stops us from simulating many scenarios for the same date *t*).

Step 3: To obtain noisy bond prices for date *t*, we calculate theoretical prices for all bonds *i* using the base yield curve $\hat{P}_{i,t,base}$ and add random errors $\hat{\varepsilon}_i$ to them. The errors are determined in the next two steps:

Step 4: From all dates in the dataset, we randomly choose the reference date t_t^* (one for all bonds *i*, but different for each simulation t).

Step 5: For each bond *i*:

- a random date $\tau_{i,t}$, separated from the reference date t_t^* by no more than T days is chosen;
- a random fitting model $\kappa_{i,t}$ is chosen, in order not to give preference to one of the models, as noted above;
- the model observation error $\hat{\varepsilon}_i$ for the bond *i* is taken to be equal to real fitting error for the randomly selected date $\tau_{i,t}$ and the randomly selected fitting model $\kappa_{i,t}$:

$$\hat{\varepsilon}_i = \varepsilon_{i,\tau_{i,t},\kappa_{i,t}}.$$

The parameter T regulates the correlation structure of the modeled errors. Small values of T ensure that the model errors for the same day are taken from neighboring days (from the same day if $T = 0$), which ensures the plausibility of this combination of errors. However, the assortment of real fitting errors to sample from, is very small in this case. On the other hand, large values of T mean that real fitting errors are sampled from all the available history almost independently for each bond.

Note that since bond fitting errors have previously been found to be persistent over time, this sampling scheme is actually sensible for $T > 0$.

In our simulation, we use $T = 20$.

Step 6: For all fitting models *k*, yield curves $r_{t,k}^{est}(\cdot)$ are estimated from the noisy model data generated in the previous steps.

Step 7: For each model k, we calculate the yield curve fitting error:

$$\varepsilon_{t,k}^{est}(\tau) = r_{t,k}^{est}(\tau) - r_{t,base}(\tau),$$

where $r_{t,k}^{est}(\tau)$ is the zero-coupon yield for term to maturity τ, estimated from the noisy data for date t using model k (see Step 6), and $r_{t,base}(\tau)$ is the corresponding 'true' model zero-coupon rate (see Step 1).

Note that it would be more logical to use relative errors, rather than absolute ones. However, as the yields in some European countries (Germany, France) reach zero and negative values during the time period considered, the relative errors become inadequately large. Therefore, we use the absolute errors.

Steps 1–7 are repeated for each date *t* in the dataset. For each date *t*, steps 3–7 are repeated $N_{random} = 100$ times with new random noisy prices to obtain a sample $\varepsilon_{t,k,j}^{est}(\tau), j = 1..N_{random}$ from the yield curve fitting error distribution.

The error statistics reported below are calculated from the joint sample of all fitting errors for all simulations for all dates in the dataset.

Comparison of Bond Price Fitting

This comparison technique is based on estimating the posterior predictive power using real data. We estimate it via a leave-out-one cross-validation as described below.

Step 1: One bond i is excluded from the sample. Using the truncated sample, a yield curve $r_{-i,t,k}(\cdot)$ is constructed via fitting model *k*.

Step 2: The model price $\hat{P}_{-i,t,k}$ of the excluded bond i and the model prediction error $\varepsilon_{-i,t,k}$ are calculated from the yield curve $r_{-i,t,k}(\cdot)$.

Step 3: Since all our estimation methods feature independent observation errors, the posterior predictive error distribution for the excluded bond coincides with its prior distribution (normal with standard deviations w_k). Therefore, the expected out-of-sample log-likelihood for model *k*, ELL_k can easily be approximated as:

$$ELL_k = \frac{1}{T}\sum_{t=1}^{T}\left[\frac{1}{n}\sum_{i=1}^{n}\log p_{i,t,k}\left(\varepsilon_{i,t,k}\right)\right],$$

where $p_{i,t,k}(x)$ is the predictive probability density function of the observation error for bond i on date t according to model k:

- For models 1–6 (‘Standard’, ‘D’, ‘1/D’, ‘1/S’, ‘1/log(S/MS)’, ‘HIST’), $p_{i,t,k}(\varepsilon)$ is a normal pdf with the standard deviation equal to the inverse of the corresponding weight $w_{i,t,k}$:

$$p_{i,t,k}(\varepsilon) = \frac{w_{i,t,k}C_{i,t,k}}{\sqrt{2\pi}} \exp\left[-\frac{\varepsilon^2 w_{i,t,k}^2 C_{i,t,k}^2}{2}\right].$$

Note that up until now, the weights could have been specified up to a multiplicative constant C: the change of variables $w'_{i,t,k} = C \cdot w_{i,t,k}$ does not change the optimization problem in (1). ELL_k clearly depends on this constant (these constants), so to compute $p_{i,t,k}(\varepsilon)$, we estimate the respective constant via maximum likelihood. Since $p_{i,t,k}(\varepsilon)$ is the predictive density for the observation error $\varepsilon_{-i,t,k}$, in computing it we cannot use information on bond i. Therefore, we can estimate:

$$C_{i,t,k} = \arg\max \sum_{\substack{j=1 \\ j\neq i}}^{n} \left[\log\frac{w_{j,t,k}C_{i,t,k}}{\sqrt{2\pi}} - \frac{\varepsilon_{j,t,k}^2 w_{j,t,k}^2 C_{i,t,k}^2}{2}\right] = \sqrt{\frac{n-1}{\sum_{\substack{j=1 \\ j\neq i}}^{n} \varepsilon_{j,t,k}^2 w_{j,t,k}^2}}.$$

The values for $C_{i,t,k}$ can be computed after solving the optimization problem (1), but before making the predictions, because the fitting does not depend on $C_{i,t,k}$, only the predictive distribution does.

- For model 7 (‘Bid-Ask’):

$$p_{i,t,k}(\varepsilon) = \frac{w_{j,t,k}C_{i,t,k}}{\sqrt{2\pi} + s_{i,t}w_{j,t,k}C_{i,t,k}} \exp\left[-\frac{w_{i,t,k}^2 C_{i,t,k}^2 \phi^2(\varepsilon, s_{i,t})}{2}\right],$$

where $\phi(\varepsilon, s) = \begin{cases} \varepsilon + \frac{s}{2}, & \varepsilon \leq -\frac{s}{2}; \\ \varepsilon - \frac{s}{2}, & \varepsilon \geq \frac{s}{2}; \\ 0, & -\frac{s}{2} \leq \varepsilon \leq \frac{s}{2}; \end{cases}$

$s_{i,t}$ is the bid-ask spread of bond i at date t, $C_{i,t,k}$ is the multiplicative constant to be estimated via maximum likelihood. As before,

$$C_{i,t,k} = \arg\max \sum_{\substack{j=1 \\ j\neq i}}^{n} \left[\log\frac{w_{j,t,k}C_{i,t,k}}{\sqrt{2\pi} + s_{i,t}w_{j,t,k}C_{i,t,k}} - \frac{\varepsilon_{j,t,k}^2 w_{j,t,k}^2 C_{i,t,k}^2}{2}\right].$$

A closed form solution for $C_{i,t,k}$ is available, but it requires solving a cubic equation, so the formula is too complex to be included here. As before, this constant can be computed before making the prediction for the price of bond i.

These steps are repeated for each date t forming a set of expected log-likelihoods to be compared, but we also record the individual prediction errors $\varepsilon_{-i,t,k}$, as their mean and standard deviation (taken across the random simulations and across the time dimension t) could serve as a measure of robustness of the weighting models considered.

5 Empirical Results

The Effect of Observation Errors on Term Structure Estimations

Figure 3 shows an example of the yield curves estimated by the Svensson model for $N_{dates} = 252$ days. As bonds with a maturity of less than 90 days were excluded from the sample, the constructed curves at short terms are excessively volatile, that is, they have anomalously low or high values. Therefore, the term structure of interest rates is also considered only for terms from 90 days.

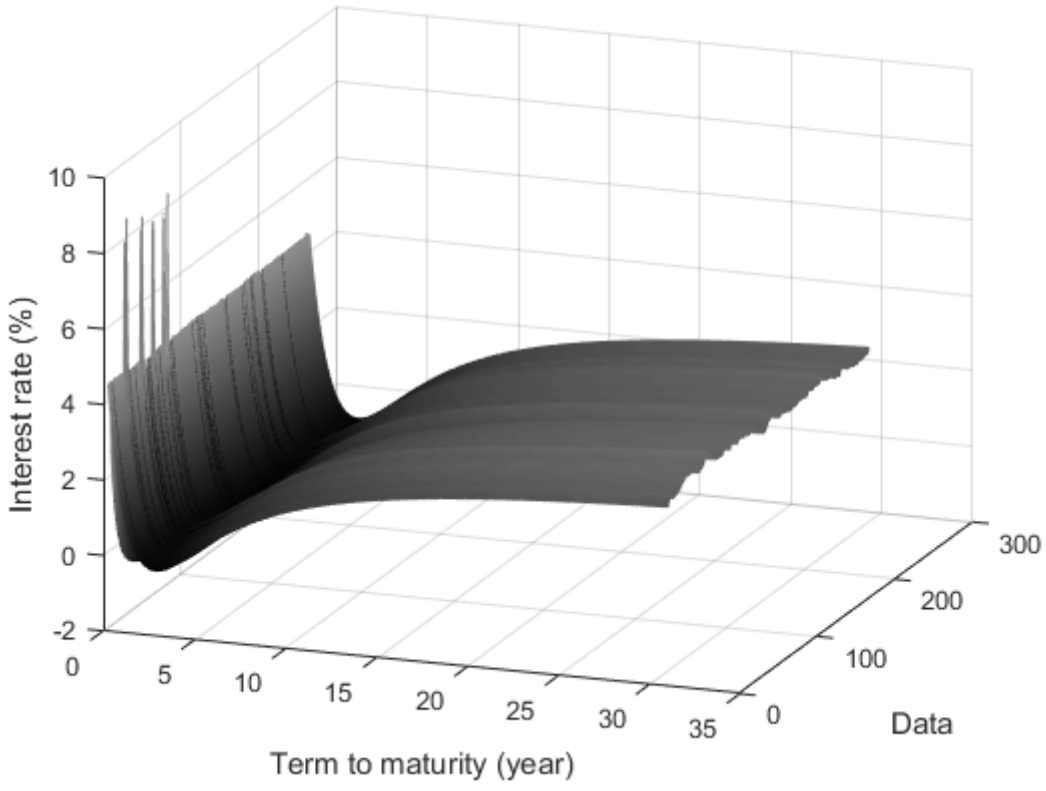

Figure 3. The yield curves for different days for the Svensson model – Spain.

Next, $N_{random} = 100$ sets of random errors are added to the model prices for each bond for each day according to the procedure described above. Then, for each noisy set of model bond prices, zero-coupon yield curves are estimated via every fitting model. For every day, we get $N_{random} \cdot N_{models}$ interest rate term structures constructed from noisy data (Figure 4).

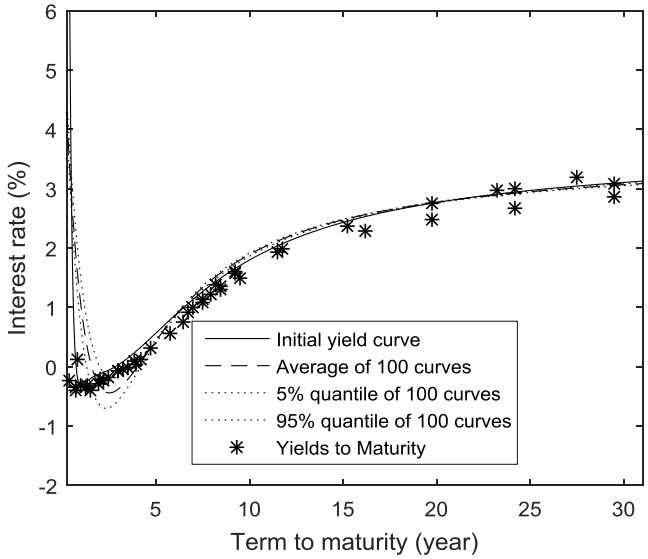

Figure 4. The yield curve for the Svensson model at market prices – Spanish data on 3 Feb 2017, the average for 100 noisy curves, 5% and 95% quantiles for noisy curves.

The estimation error distribution would ideally have a zero mean and a small standard deviation. However, Figure 5 shows that this is not always the case – in practice, noisy estimates are usually biased away from zero for a given date.

We consider four term to maturity intervals for convenience – short-term (from 90 days to 1 year), medium-term (1–8 years), long-term (8–49 years) and all maturities (90 days to 49 years).

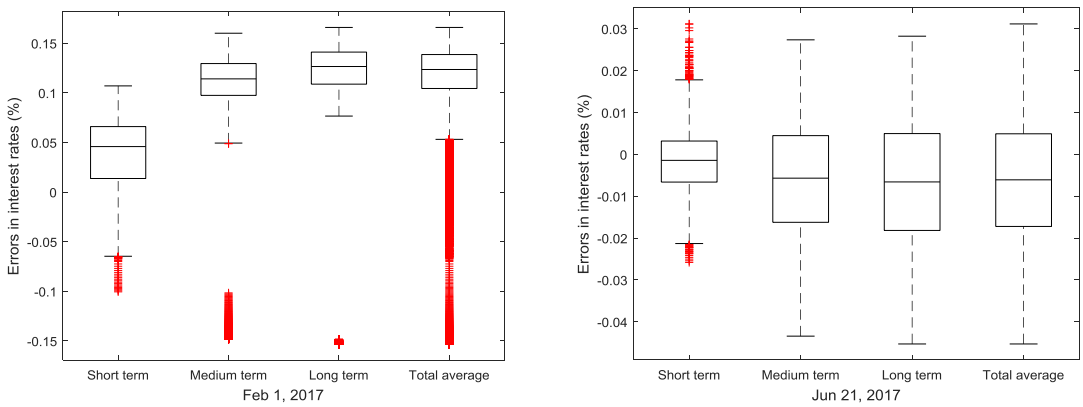

Figure 5. Box plots of interest rate estimation errors for the Svensson model – UK data for 3 Feb 2017 and 21 Jun 2017.

The means of the error distributions are non-zero for a given date. What distribution do these non-zero means follow over different dates? Figure 6 plots estimated means for all dates.

To assess the impact of data errors on the estimation results for all models, we retain the term to maturity dimension τ and compute the mean $M_{t,k}(\tau)$ and the standard deviation $D_{t,k}(\tau)$ of the errors across all random samples within each date.

The mean and the standard deviation of errors in rates behave in a similar way (Figures 6, 7) – they have maximum values for small terms (means from -10% to 5%, standard deviations from 0.1% to 10%), then fall and fluctuate at a certain level for medium and long terms to maturity (means from -0.3% to 0.3%, standard deviations from 0.03% to 0.2%). The average of the mean errors for different days varies around zero while the average standard deviation is about 0.1%.

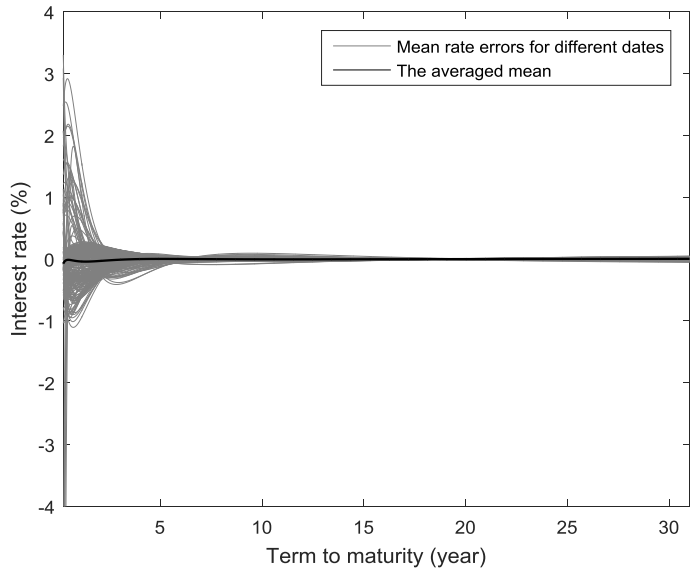

Figure 6. Mean of interest rate estimation errors for the parametric model – UK, various dates.

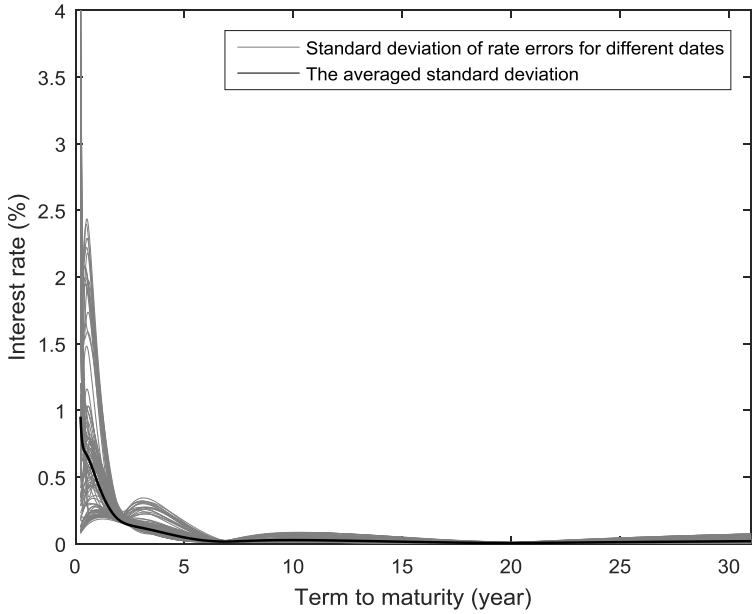

Figure 7. Standard deviation of interest rate estimation errors for the parametric model – UK, various dates.

One can see that realistic observation errors in bond prices can, in general, greatly affect the estimated term structure.

Term Structure Fitting

To compare the term structure estimates, for each fitting method, k, we calculate the median absolute deviation of the fitting errors for each term to maturity τ: $M_k(\tau) = \text{median}_{t,j}\left|\varepsilon_{t,k,j}^{est}(\tau)\right|$.

Table 2 presents an example of the median term structure deviation levels with additional aggregation over the term to maturity τ. We also report significance levels for the null hypothesis of insignificant differences vs. the standard model. Almost all differences turn out to be significant – mainly due to large sample sizes (hundreds of thousands).

The significance levels are produced using the Wilcoxon-Mann-Whitney test. According to Fay and Proschan (2010), it should be preferred to the t-test in our case – many observations which are subject to errors and outliers (due to possible numerical optimization issues). Note that although the Wilcoxon-Mann-Whitney test does not, strictly speaking, test for the difference in medians for non-equal distributions, we report median values to provide additional intuition.

Table 2. Median term structure fitting errors $M_k(\tau)$ for the French data. Less is better. Significance levels via Wilcoxon-Mann-Whitney test comparing the $M_k(\tau)$ levels with the Standard model (*). (+) indicates the best model.

Weighting Scheme	**From 90 months to the year**		**From 1 to 8 years**		**From 8 to 49 years**		**All**	
	Parametric	Non-parametric	Parametric	Non-parametric	Parametric	Non-parametric	Parametric	Non-parametric
Standard	0,129% +++	1,54%	0,05% +++	0,128%	0,007% +++	0,062%	0,007% +++	0,066%
D	0,497% ***	0,56% ***	0,231% ***	0,107% ***	0,024% ***	0,062% *	0,027% ***	0,065% ***
1/D	2,836% ***	0,565% ***	0,216% ***	0,108% ***	0,276% ***	0,056% ***	0,271% ***	0,058% ***
1/S	2,681% ***	0,702% ***	0,181% ***	0,129% **	0,279% ***	0,043% ***,+++	0,279% ***	0,046% ***,+++
1 / log(S/MS)	0,137% ***	1,798% ***	0,056% ***	0,172% ***	0,007% ***	0,082% ***	0,008% ***	0,09% ***
HIST	0,223% ***	0,567% ***	0,084% ***	0,105% ***	0,016% ***	0,059% ***	0,018% ***	0,062% ***
Bid-ask	0,235% ***	0,565% ***	0,056% ***	0,12% ***	0,033% ***	0,051% ***	0,035% ***	0,055% ***

Different from the standard model: '***' 0.001 '**' 0.01 '*' 0.05

Best in column: '+++' 0.001 '++' 0.01 '+' 0.05

We also report the results without aggregation over the term-to-maturity dimension. Figure 8 schematically depicts the results. Each horizontal line corresponds to a weighting scheme as indicated via the legend. For a given weighting scheme, a portion of the line is black for those terms to maturity, for which its $M_k(\tau)$ is the best (several schemes might be marked as best if the difference is insignificant). A portion of the line is dark gray if for these maturities the scheme

performs worse than the best, but better than the standard model. Light gray line segments mean the model differs from the standard model insignificantly. Finally, white regions mean the model performs worse than the standard model.

One can see that for the parametric model any modification is significantly worse than the standard model with equal weights. While this might seem counterintuitive at the first glance (e.g. one would expect better results for the 'D' scheme for longer maturities and for '1/D' scheme for shorter maturities), it can be explained.

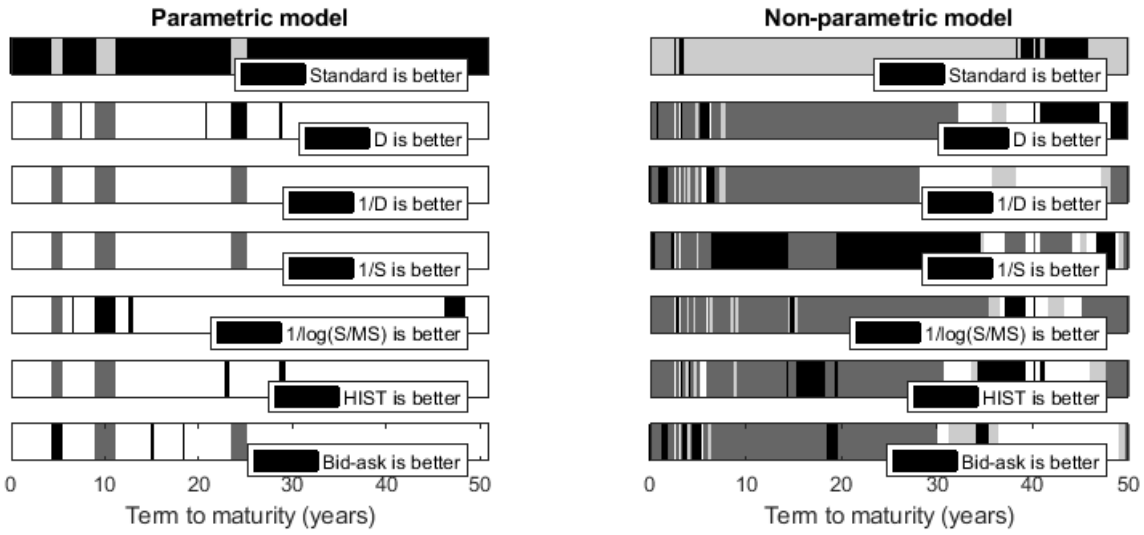

Figure 8. A graphical representation of the comparison results. French data. Black = the best, dark grey = better than the standard, light gray = insignificantly different from the standard, white = worse than standard.

Consider, for example, the '1/D' scheme. It gives much more weight to short-term bonds and less weight to long-term bonds. It is to be expected that the estimate of the long-term rates will be worse for the '1/D' scheme than for the standard approach. However, the estimate of the short-term rates will also be worse, because the Svensson parametric equation is not flexible enough. Some bonds have much more weight than others and fitting the curve to these bonds (very short-term bonds in this case) will inevitably mess with other maturities. Basically, instead of normal fitting for all maturities we have overfitting for some bonds and underfitting for others. The same applies to other approaches – variance in weights decreases the effective amount of data used to fit the curve, which results in overfitting of some bonds and underfitting of others.

The non-parametric model has more interpretable results – 'D' is better for longer maturities, '1/D' – for shorter, etc. The results for other countries do not exhibit a single pattern. They are summarized in Table 4 below (see the Appendix for the detailed results).

Bond Price Fitting

Using cross-validation, we estimate the posterior predictive log-likelihood for all weighting schemes. Since the underlying stochastic model is the same (independent normally distributed observation errors), the log-likelihood values are directly comparable – more is better. Once again, we assess the significance of the differences in log-likelihoods using the Wilcoxon-Mann-Whitney test, which is preferred due to outliers in the data. As before, we report median values for the predictive log-likelihood even though the statistical test we use does not actually test for the difference in medians in a general case.

An example is presented in Table 3. Results for other countries are reported in the Appendix. They are summarized in Table 4 below.

Table 3. Results of cross-validation (the median value of the expected log-likelihood). More is better. French data. Significance levels via Wilcoxon-Mann-Whitney test comparing the ELL levels with the Standard model (*). (+) indicates the best model.

Weighting Scheme	Parametric model	Non-parametric model
Standard	-1,71+++	-3,96
D	-2,85***	-5,17***
1/D	-2,04***	-2,89***
1/S	-3,27***	-3,53***
1/log(S/MS)	-2,06***	-4,21***
HIST	-2,75***	-1,89***, +++
Bid-Ask	-1,99***	-3,97

Different from the standard model: '***' 0.001 '**' 0.01 '*' 0.05

Best in column: '+++' 0.001 '++' 0.01 '+' 0.05

Aggregated results

We report the aggregated statistics for all countries. We indicate the preferred method according to the two approaches (term structure fitting and bond price fitting) and within the two models (parametric and non-parametric).

Table 4. Summary of comparison results for various countries. Reporting the best method (methods) for each setting.

Country	Term Structure Fitting		Bond Price Fitting	
	Parametric	Non-parametric	Parametric	Non-parametric
France	Standard	1/S, other equal	Standard	HIST
Germany	Standard	Various	Standard	Standard and

				1/log(S/MS)
Greece	Standard	Standard	1/D	1/D
Italy	Standard	All equal	Standard	1/D
Portugal	Standard	1/D and Standard	1/D	1/D
Russia	Standard	Standard	Standard	Standard
Spain	Standard	1/D and Standard	1/S	1/S, Standard
United Kingdom	1/D, D and Standard	Standard, D and 1/D	1/D	1/D and HIST

6 Conclusion

Our results are twofold. We propose a non-parametric bootstrap-based method of sampling real-like observation errors in model bond price data. This allows us to estimate the magnitude of errors in estimating the zero-coupon yield curve from bond prices.

We stipulate that this procedure does not introduce an unjust bias towards one weighting scheme and use it to compare different weighting schemes. The weighting schemes for the comparison are taken from the literature along with several other plausible choices.

The results of the comparison are somewhat surprising. If one is interested in the quality of fitting the zero-coupon yield curve, the standard method (equal weights) is preferred for almost all datasets for the Svensson model and mixed results with no discernible pattern for the spline method. However, real-life bond price fitting errors can be assumed to be roughly proportional to the inverse duration, which is one of the most common weighting schemes used in the literature. We think that other schemes lose in this comparison due to the fact that their weights are too volatile themselves (e.g. the bid-ask spread changes too much for a given bond from one day to another, which introduces additional volatility into the model).

We can summarize the practical takeaway as follows:

1. Typical observation errors in bond prices are relatively large, resulting in fitting discrepancies in the zero-coupon yield curve of about 10–100 basis points.
2. In light of such large discrepancies, we do not advise to introduce weights for bonds, as most choices result in overfitting for bonds with large weights and underfitting for bonds with small weights.

3. The variance of the real-life observation errors in bond prices can be considered to be proportional to the inverse duration of the bond, which is one of the most popular choices in the literature.

Therefore, if one feels like introducing varying weights for bonds, they should be proportional to the inverse duration. However, equal weights are equally fine if one is interested in the term structure estimates and not in the bond price distribution.

References

1. Ahi E., Akgiray, V., Sener, E. Robust term structure estimation in developed and emerging markets // Annals of Operations Research. 2018. Vol. 260. Issue 1-2. pp. 23-49.
2. Bertocchi M., Moriggia, V., Dupacova, J. Sensitivity of Bond Portfolio's Behavior with Respect to Random Movements in Yield Curve: A Simulation Study // Annals of Operations Research, Vol. 99. No. 1. 2000. pp. 267-286.
3. Bliss R. Testing term structure estimation methods. Advances in Futures and Options Research. Vol. 9. 1997. pp. 197–231.
4. Bolder D. J. Modelling Term-Structure Dynamics for Risk Management: A Practitioner's Perspective. 2006.
5. Carcano N., Dall'O H. Alternative models for hedging yield curve risk: An empirical comparison // Journal of Banking and Finance. Vol. 35. No. 11. 2011. pp. 2991-3000
6. Cruz-Marcelo Al., Ensora K. B., Rosnera G. L. Estimating the Term Structure With a Semiparametric Bayesian Hierarchical Model: An Application to Corporate Bonds // Journal of the American Statistical Association. Vol. 106. No. 494. 2014. pp. 387-395
7. Fay M.P., Proschan M.A. Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules // Statistical Surveys. 2010. Vol. 4. pp. 1 – 39.
8. Fleming J., Whaley R.E. The value of wildcard options // Journal of Finance. Vol. 49. 1994. pp. 215–236
9. Gimeno R., Nave J.M. A genetic algorithm estimation of the term structure of interest rates // Computational Statistics and Data Analysis. Vol. 53. No. 6. 2009. pp. 2236-2250.
10. Hladikova H., Radova J. Term Structure Modelling by Using Nelson-Siegel Model. European Financial and Accounting Journal. Vol. 7. No. 2. 2012.pp. 36-55.
11. Jordan J.V., Mansi S.A. Term structure estimation from on-the-run Treasuries // Journal of Banking and Finance. Vol. 27 No. 8. 2003. pp. 1487-1509.
12. Ioannides M. A comparison of yield curve estimation techniques using UK data // Journal of Banking and Finance. Vol. 27 No. 1. 2003. pp. 1-26.
13. Laurini M.P., Ohashi A., A noisy principal component analysis for forward rate curves. European Journal of Operational Research. Vol. 246. 2015. pp. 140–153
14. Ubukata M., Fukushige M. Estimation and inference in the yield curve model with an instantaneous error term // Mathematics and Computers in Simulation Vol. 79. 2009. pp. 2938–2946.

Appendix

In what follows, we present the results for all countries in our dataset.

Term Structure Fitting

Germany

Table 5. Median term structure fitting errors $M_k(\tau)$ for the German data. Less is better. Significance levels via Wilcoxon-Mann-Whitney test comparing the $M_k(\tau)$ levels with the Standard model (*). (+) indicates the best model.

Weighting Scheme	From 90 months to the year		From 1 to 8 years		From 8 to 49 years		All	
	Parametric	Non-parametric	Parametric	Non-parametric	Parametric	Non-parametric	Parametric	Non-parametric
Standard	0,245%+++	1,055%	0,023%+++	0,208%+++	0,007%+++	0,999%	0,008%+++	0,895%
D	0,86%***	2,28%***	0,1%***	0,278%***	0,028%***	0,996%**	0,034%***	0,92%***
1/D	2,486%***	0,574% (***,+++)	0,309%***	0,282%***	0,207%***	1,001%***	0,228%***	0,85%
1/S	0,506%***	0,669%***	0,078%***	0,298%***	0,186%***	0,954% (***,+++)	0,17%***	0,778%.
1/log(S/MS)	0,334%***	1,03%***	0,031%***	0,275%***	0,01%***	0,996%***	0,012%***	0,871%***
HIST	2,274%***	0,693%***	0,097%***	0,226%***	0,029%***	1,007%***	0,034%***	0,874%***
Bid-ask	0,307%***	0,974%***	0,033%***	0,223%***	0,012%***	1,001%***	0,014%***	0,885%.

Different from the standard model: '***' 0.001 '**' 0.01 '*' 0.05

Best in column: '+++' 0.001 '++' 0.01 '+' 0.05

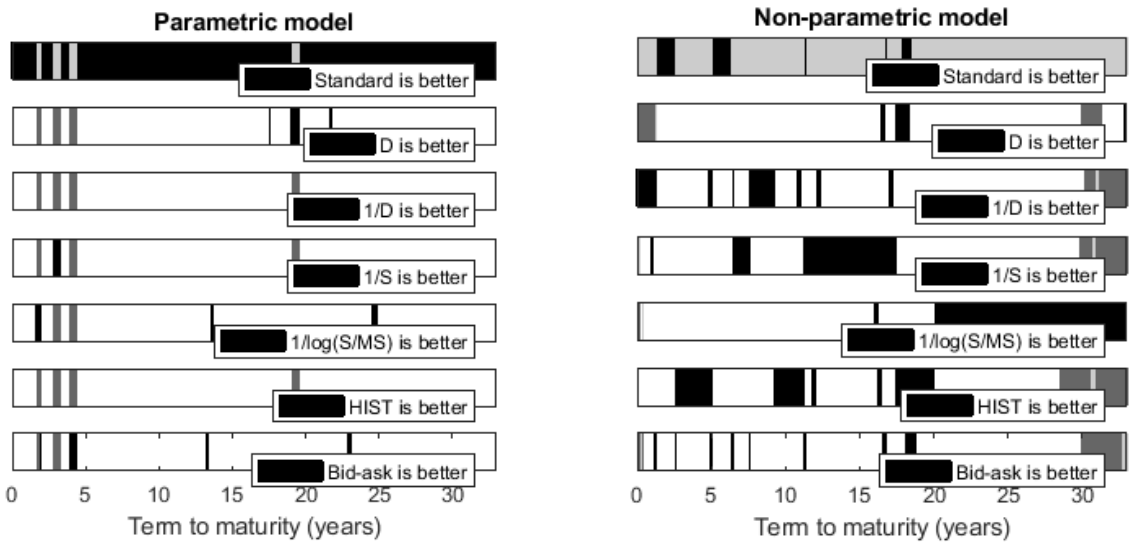

Figure 9. A graphical representation of the comparison results. German data. Black = the best, dark grey = better than the standard, light gray = insignificantly different from the standard, white = worse than standard.

Greece

Table 6. Median term structure fitting errors $M_k(\tau)$ for the Greek data. Less is better. Significance levels via Wilcoxon-Mann-Whitney test comparing the $M_k(\tau)$ levels with the Standard model (*). (+) indicates the best model.

Weighting Scheme	**From 90 months to the year**		**From 1 to 8 years**		**From 8 to 49 years**		**All**	
	Parametric	Non-parametric	Parametric	Non-parametric	Parametric	Non-parametric	Parametric	Non-parametric
Standard	0,457%+++	0,126%+++	0,228%+++	0,15%+++	0,175%+++	0,352%+	0,178%+++	0,268%+++
D	25,404%***	0,284%***	0,815%***	0,339%***	0,296%***	0,842%***	0,398%***	0,541%***
1/D	2,776%***	0,381%***	0,441%***	0,215%***	0,602%***	0,63%***	0,579%***	0,456%***
1/S	1,056%***	0,887%***	0,454%***	0,492%***	0,57%***	2,457%***	0,557%***	1,45%***
1/log(S/MS)	7,518%***	0,804%***	0,502%***	0,756%***	0,344%***	2,744%***	0,369%***	1,694%***
HIST	7,524%***	0,197%***	1,053%***	0,283%***	0,649%***	0,853%***	0,69%***	0,599%***
Bid-ask	4,666%***	0,186%***	0,414%***	0,167%***	0,239%***	0,355%*	0,264%***	0,281%***

Different from the standard model: '***' 0.001 '**' 0.01 '*' 0.05

Best in column: '+++' 0.001 '++' 0.01 '+' 0.05

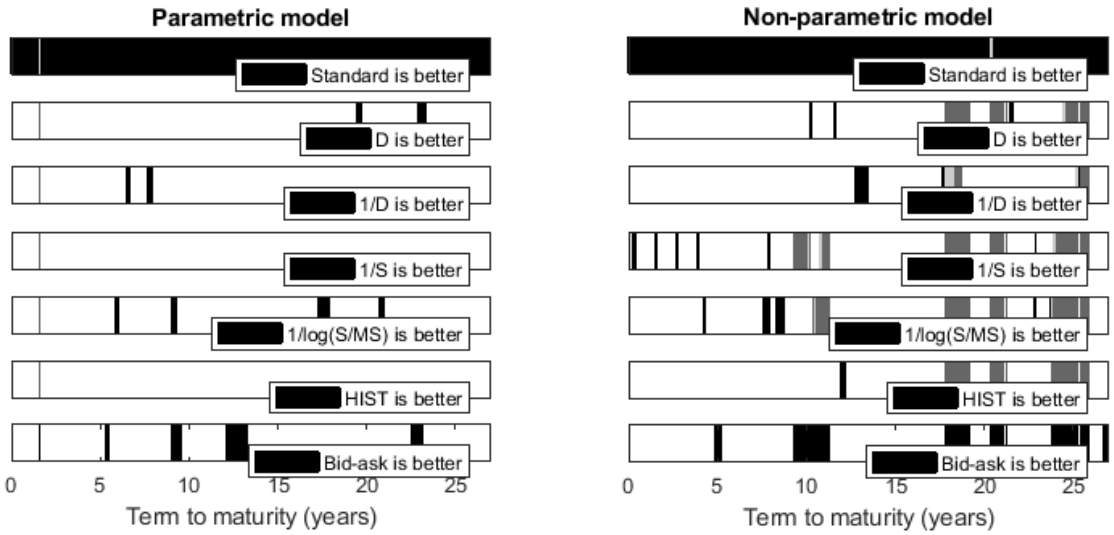

Figure 10. A graphical representation of the comparison results. Greek data. Black = the best, dark grey = better than the standard, light gray = insignificantly different from the standard, white = worse than standard.

Italy

Table 7. Median term structure fitting errors $M_k(\tau)$ for the Italian data. Less is better. Significance levels via Wilcoxon-Mann-Whitney test comparing the $M_k(\tau)$ levels with the Standard model (*). (+) indicates the best model.

Weighting Scheme	From 90 months to the year		From 1 to 8 years		From 8 to 49 years		All	
	Parametric	Non-parametric	Parametric	Non-parametric	Parametric	Non-parametric	Parametric	Non-parametric
Standard	0,139%+++	2,078%	0,031%+++	0,288%	0,018%+++	0,278%	0,018%+++	0,274%
D	8,913%***	12,563%***	0,275%***	0,432%***	0,071%***	0,308%***,+++	0,076%***	0,339%***,+++
1/D	2,518%***	2,309%***	0,444%***	0,262%***,+++	0,395%***	0,339%***,+++	0,405%***	0,32%***,+++
1/S	2,363%***	1,314%***	0,248%***	0,57%***,+++	0,257%***	1,002%***,+++	0,257%***	0,803%***,+++
1/log(S/MS)	0,157%***	0,673%***,+++	0,056%***	0,443%***,+++	0,023%***	0,522%***,+++	0,024%***	0,501%***,+++
HIST	0,492%***	1,29%***	0,149%***	0,319%***,+++	0,066%***	0,452%***,+++	0,069%***	0,411%***,+++
Bid-ask	0,164%***	1,616%	0,039%***	0,309%***,+++	0,019%***	0,436%***,+++	0,019%***	0,396%***,+++

Different from the standard model: '***' 0.001 '**' 0.01 '*' 0.05

Best in column: '+++' 0.001 '++' 0.01 '+' 0.05

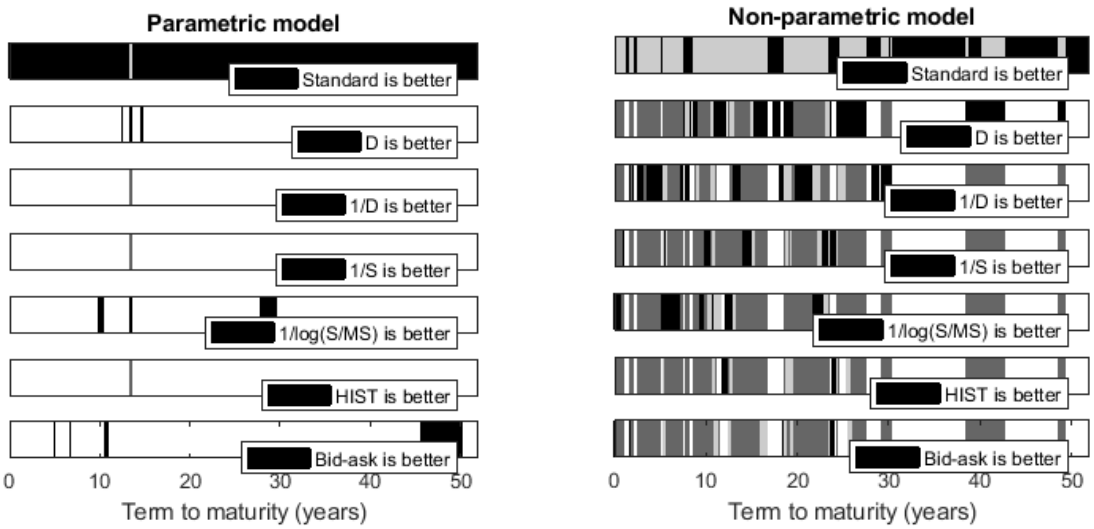

Figure 11. A graphical representation of the comparison results. Italian data. Black = the best, dark grey = better than the standard, light gray = insignificantly different from the standard, white = worse than standard.

Portugal

Table 8. Median term structure fitting errors $M_k(\tau)$ for the Portuguese data. Less is better. Significance levels via Wilcoxon-Mann-Whitney test comparing the $M_k(\tau)$ levels with the Standard model (*). (+) indicates the best model.

Weighting Scheme	From 90 months to the year		From 1 to 8 years		From 8 to 49 years		All	
	Parametric	Non-parametric	Parametric	Non-parametric	Parametric	Non-parametric	Parametric	Non-parametric
Standard	0,307%+++	0,148%	0,061%+++	0,108%	0,024%+++	0,105%+++	0,028%+++	0,106%+++
D	1,826%***	0,266%***	0,107%***	0,163%***	0,03%***	0,118%***	0,036%***	0,126%***
1/D	1,712%***	0,095%***	0,199%***	0,1%***,+++	0,2%***	0,13%***	0,199%***	0,117%***
1/S	1,738%***	0,09%***,+++	0,304%***	0,115%***	0,168%***	0,146%***	0,182%***	0,134%***
1/log(S/MS)	0,567%***	0,19%***	0,076%***	0,162%***	0,027%***	0,17%***	0,032%***	0,167%***
HIST	0,84%***	0,174%***	0,155%***	0,121%***	0,048%***	0,125%***	0,057%***	0,121%***
Bid-ask	0,497%***	0,149%	0,071%***	0,115%***	0,033%***	0,116%***	0,037%***	0,117%***

Different from the standard model: '***' 0.001 '**' 0.01 '*' 0.05

Best in column: '+++' 0.001 '++' 0.01 '+' 0.05

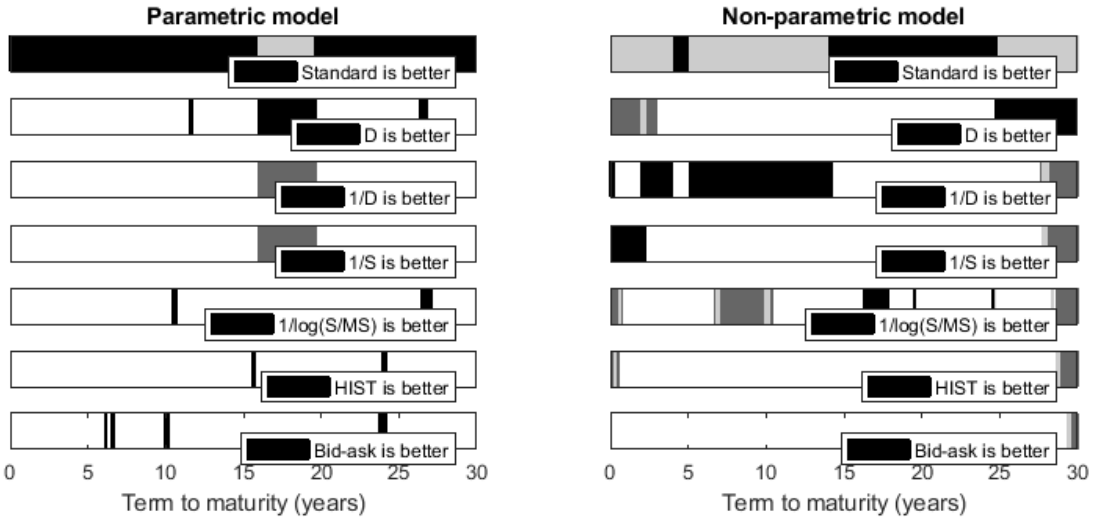

Figure 12. A graphical representation of the comparison results. Portuguese data. Black = the best, dark grey = better than the standard, light gray = insignificantly different from the standard, white = worse than standard.

Russia

Table 9. Median term structure fitting errors $M_k(\tau)$ for the Russian data. Less is better. Significance levels via Wilcoxon-Mann-Whitney test comparing the $M_k(\tau)$ levels with the Standard model (*). (+) indicates the best model.

Weighting Scheme	From 90 months to the year		From 1 to 8 years		From 8 to 49 years		All	
	Parametric	Non-parametric	Parametric	Non-parametric	Parametric	Non-parametric	Parametric	Non-parametric
Standard	0,235%+++	0,084%+++	0,037%+++	0,036%+++	0,042%+++	0,038%+++	0,041%+++	0,036%+++
D	1,263%***	0,09%***	0,059%***	0,038%***	0,044%***	0,04%***	0,049%***	0,038%***
1/D	0,257%***	0,191%***	0,044%***	0,038%***	0,063%***	0,04%***	0,054%***	0,039%***
1/S	0,322%***	0,17%***	0,108%***	0,099%***	0,123%***	0,072%***	0,116%***	0,078%***
1/log(S/MS)	0,594%***	0,363%***	0,1%***	0,108%***	0,254%***	0,151%***	0,176%***	0,136%***
HIST	1,485%***	0,098%***	0,17%***	0,052%***	0,391%***	0,055%***	0,312%***	0,053%***
Bid-ask	0,386%	0,09%***	0,086%***	0,039%***	0,12%***	0,042%***	0,105%***	0,04%***

Different from the standard model: '***' 0.001 '**' 0.01 '*' 0.05

Best in column: '+++' 0.001 '++' 0.01 '+' 0.05

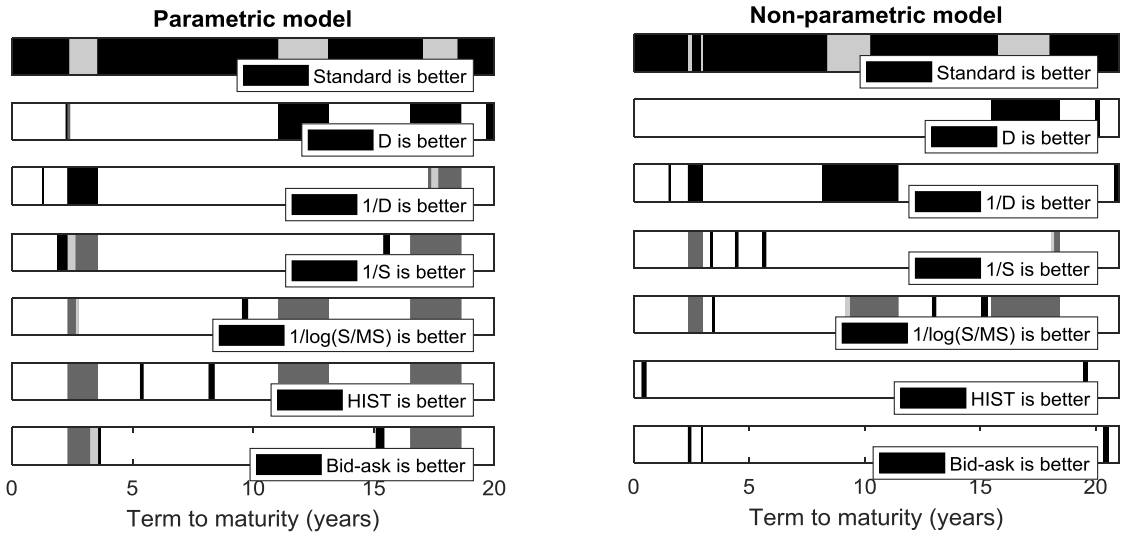

Figure 13. A graphical representation of the comparison results. Russian data. Black = the best, dark grey = better than the standard, light gray = insignificantly different from the standard, white = worse than standard.

Spain

Table 10. Median term structure fitting errors $M_k(\tau)$ for the Spanish data. Less is better. Significance levels via Wilcoxon-Mann-Whitney test comparing the $M_k(\tau)$ levels with the Standard model (*). (+) indicates the best model.

Weighting Scheme	From 90 months to the year		From 1 to 8 years		From 8 to 49 years		All	
	Parametric	Non-parametric	Parametric	Non-parametric	Parametric	Non-parametric	Parametric	Non-parametric
Standard	0,197%+++	0,14%	0,05%+	0,067%	0,007%+++	0,029%	0,01%+++	0,034%
D	0,785%***	0,849%***	0,186%***	0,356%***	0,042%***	0,043%***	0,05%***	0,058%***
1/D	1,96%***	0,088%***,+++	0,307%***	0,061%***,+++	0,296%***	0,038%***	0,305%***	0,041%***
1/S	0,705%***	0,105%***	0,111%***	0,076%***	0,103%***	0,046%***	0,108%***	0,051%***
1/log(S/MS)	0,32%***	0,721%***	0,074%***	0,22%***	0,011%***	0,153%***	0,014%***	0,158%***
HIST	0,292%***	0,29%***	0,073%***	0,128%***	0,025%***	0,041%***	0,029%***	0,049%***
Bid-ask	0,207%***	0,141%**	0,051%*	0,078%***	0,009%***	0,041%***	0,012%***	0,045%***

Different from the standard model: '***' 0.001 '**' 0.01 '*' 0.05

Best in column: '+++' 0.001 '++' 0.01 '+' 0.05

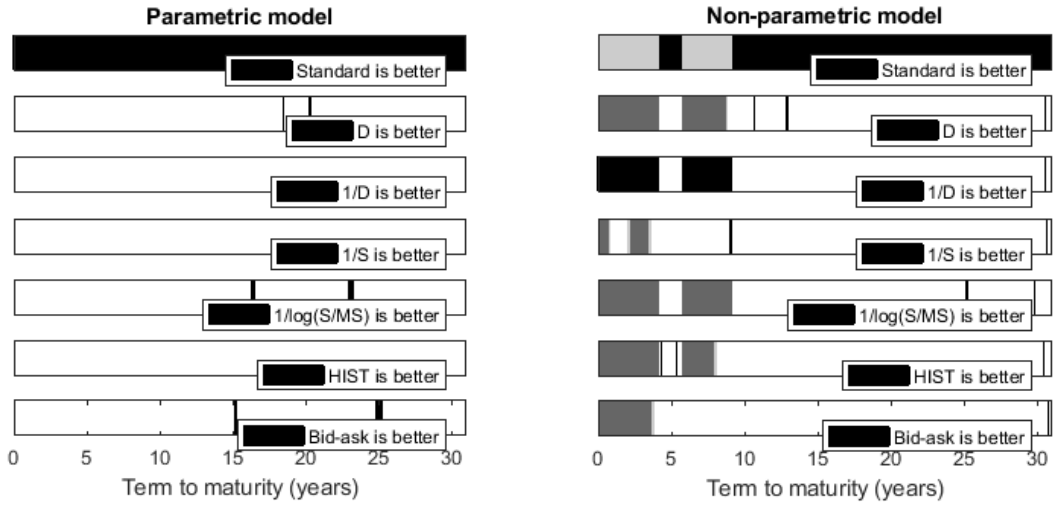

Figure 14. A graphical representation of the comparison results. Spanish data. Black = the best, dark grey = better than the standard, light gray = insignificantly different from the standard, white = worse than standard.

United Kingdom

Table 11. Median term structure fitting errors $M_k(\tau)$ for the United Kingdom data. Less is better. Significance levels via Wilcoxon-Mann-Whitney test comparing the $M_k(\tau)$ levels with the Standard model (*). (+) indicates the best model.

Weighting Scheme	From 90 months to the year		From 1 to 8 years		From 8 to 49 years		All	
	Parametric	Non-parametric	Parametric	Non-parametric	Parametric	Non-parametric	Parametric	Non-parametric
Standard	3,727%	0,095%+++	0,306%	0,059%+++	0,086%+++	0,058%+++	0,103%+++	0,057%+++
D	4,701%***	2,346%***	0,8%***	0,191%***	0,115%***	0,079%***	0,133%***	0,091%***
1/D	0,393%***,+++	0,175%***	0,115%***,+++	0,119%***	0,193%***	0,053%***,+++	0,18%***	0,059%***
1/S	1,052%***	0,2%***	0,16%***	0,093%***	0,16%***	0,124%***	0,171%***	0,118%***
1/log(S/MS)	3,776%***	0,267%***	0,287%***	0,137%***	0,093%***	0,083%***	0,113%***	0,091%***
HIST	3,771%***	0,184%***	0,294%***	0,112%***	0,115%***	0,076%***	0,134%***	0,08%***
Bid-ask	3,799%***	0,13%***	0,331%***	0,077%***	0,091%***	0,075%***	0,108%***	0,075%***

Different from the standard model: '***' 0.001 '**' 0.01 '*' 0.05

Best in column: '+++' 0.001 '++' 0.01 '+' 0.05

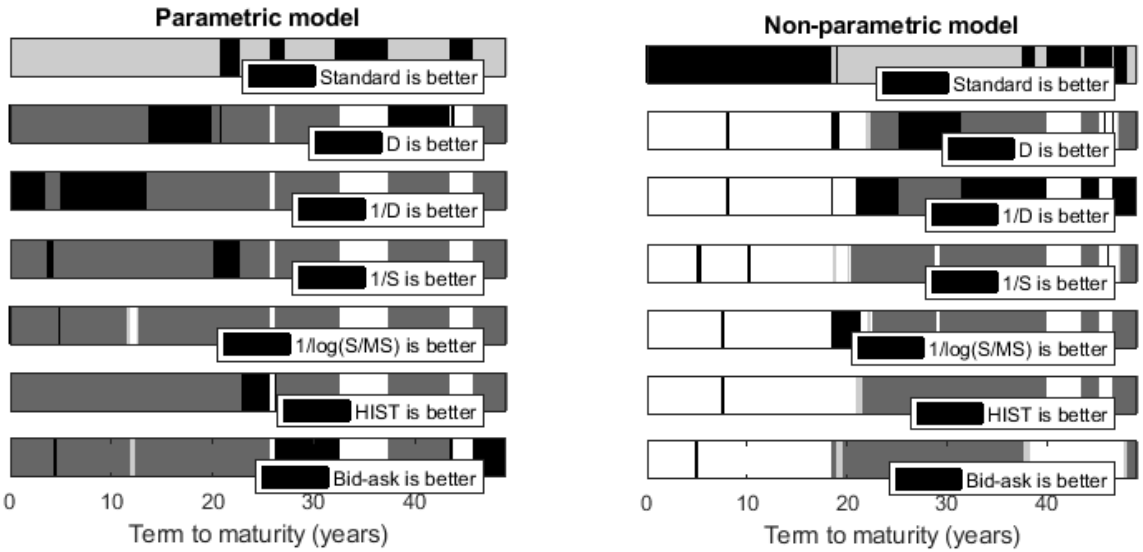

Figure 15. A graphical representation of the comparison results. United Kingdom data. Black = the best, dark grey = better than the standard, light gray = insignificantly different from the standard, white = worse than standard.

Bond Price Fitting

Germany

Table 12. Results of cross-validation (the median value of the expected log-likelihood). More is better. German data. Significance levels via Wilcoxon-Mann-Whitney test comparing the ELL levels with the Standard model (*). (+) indicates the best model.

Weighting Scheme	Parametric model	Non-parametric model
Standard	-0,66+++	-1,38
D	-1,95***	-1,98***
1/D	-0,81***	-2,16***
1/S	-0,94***	-2,58***
1/log(S/MS)	-1,15***	-1,36*,+
HIST	-1,99***	-2,72***
Bid-ask	-2,46***	-1,52***

Different from the standard model: '***' 0.001 '**' 0.01 '*' 0.05

Best in column: '+++' 0.001 '++' 0.01 '+' 0.05

Greece

Table 13. Results of cross-validation (the median value of the expected log-likelihood). More is better. Greek data. Significance levels via Wilcoxon-Mann-Whitney test comparing the ELL levels with the Standard model (*). (+) indicates the best model.

Weighting Scheme	Parametric model	Non-parametric model
Standard	-5,26	-4,8
D	-10,97***	-11,95***
1/D	-3,85***,+++	-3,8***,+++
1/S	-9,34***	-14,52***
1/log(S/MS)	-5,41***	-5,29***
HIST	-5,25***	-5,99***
Bid-ask	-4,63***	-4,21***

Different from the standard model: '***' 0.001 '**' 0.01 '*' 0.05

Best in column: '+++' 0.001 '++' 0.01 '+' 0.05

Italy

Table 14. Results of cross-validation (the median value of the expected log-likelihood). More is better. Italian data. Significance levels via Wilcoxon-Mann-Whitney test comparing the ELL levels with the Standard model (*). (+) indicates the best model.

Weighting Scheme	Parametric model	Non-parametric model
Standard	-2,55	-2,65
D	-3,38***	-3,39***
1/D	-2,75***	-2,43 ***,+++
1/S	-3,72***	-3,07***
1/log(S/MS)	-2,77***	-2,92***
HIST	-3,02***	-3,21***
Bid-ask	-2,63***	-2,69***

Different from the standard model: '***' 0.001 '**' 0.01 '*' 0.05

Best in column: '+++' 0.001 '++' 0.01 '+' 0.05

Portugal

Table 15. Results of cross-validation (the median value of the expected log-likelihood). More is better. Portuguese data. Significance levels via Wilcoxon-Mann-Whitney test comparing the ELL levels with the Standard model (*). (+) indicates the best model.

Weighting Scheme	Parametric model	Non-parametric model
Standard	-1,52	-2,44
D	-3,48***	-4,66***
1/D	-0,96***,++	-1,39 ***,+++
1/S	-1,16***	-1,91***
1/log(S/MS)	-1,62*	-2,7***
HIST	-4***	-4,93***
Bid-ask	-1,6.	-2,46

Different from the standard model: '***' 0.001 '**' 0.01 '*' 0.05

Best in column: '+++' 0.001 '++' 0.01 '+' 0.05

Russia

Table 16. Results of cross-validation (the median value of the expected log-likelihood). More is better. Russian data. Significance levels via Wilcoxon-Mann-Whitney test comparing the ELL levels with the Standard model (*). (+) indicates the best model.

Weighting Scheme	Parametric model	Non-parametric model
Standard	-1,46++	-1,37
D	-1,78***	-1,66***
1/D	-1,54***	-1,5***
1/S	-2,29***	-2,36***
1/log(S/MS)	-2,17***	-1,46***
HIST	-2,56***	-1,64***
Bid-ask	-1,51**	-1,42***

Different from the standard model: '***' 0.001 '**' 0.01 '*' 0.05

Best in column: '+++' 0.001 '++' 0.01 '+' 0.05

Spain

Table 17. Results of cross-validation (the median value of the expected log-likelihood). More is better. Spanish data. Significance levels via Wilcoxon-Mann-Whitney test comparing the ELL levels with the Standard model (*). (+) indicates the best model.

Weighting Scheme	Parametric model	Non-parametric model
Standard	-1,17	-1,29
D	-2,06***	-2,21***
1/D	-1,04***	-1,17***
1/S	-0,75***,+++	-1,32
1/log(S/MS)	-1,27***	-1,48***
HIST	-2,06***	-1,9***
Bid-ask	-2,09***	-1,54***

Different from the standard model: '***' 0.001 '**' 0.01 '*' 0.05

Best in column: '+++' 0.001 '++' 0.01 '+' 0.05

United Kingdom

Table 18. Results of cross-validation (the median value of the expected log-likelihood). More is better. United Kingdom data. Significance levels via Wilcoxon-Mann-Whitney test comparing the ELL levels with the Standard model (*). (+) indicates the best model.

Weighting Scheme	Parametric model	Non-parametric model
Standard	-19,93	-11,89+++
D	-10,78***	-14,46***
1/D	-3,22***,+++	-3,92***
1/S	-32,32***	-30,58***
1/log(S/MS)	-7,24***	-8,74***
HIST	-7,31***	-4,01***
Bid-ask	-12,91***	-7,95***

Different from the standard model: '***' 0.001 '**' 0.01 '*' 0.05

Best in column: '+++' 0.001 '++' 0.01 '+' 0.05

Victor Lapshin
National Research University Higher School of Economics (Moscow, Russia). School of Finance.
E-mail: vlapshin@hse.ru

Sofia Sokhatskaya
National Research University Higher School of Economics (Moscow, Russia). School of Finance.