Functional data analysis for Brazilian term structure of interest rate

Lucélia Viviane Vaz[†]
Rodrigo Jardim Raad[‡]

Abstract This paper analyzes Brazilian nominal yield curves based on a functional data analysis framework. Specifically, we use functional principal component analysis to describe sources of variability in yield curves and their related level, slope, and curvature. We also present a functional linear regression model to investigate macroeconomic determinants of the yield curves. We conclude that level shocks strongly explain variability in interest rate curves. Slope changes are the second-largest source of variability. The slope of the yield curve is negatively affected by the nominal exchange rate and Selic reference rate, and positively affected by Brazil's risk and industrial capacity utilization. We also infer that the following explanatory variables: expected inflation, Selic reference rate, Brazil risk, and industrial capacity utilization, all have positive effects on the level of the yield curves. The variables Selic, Brazil risk, and the nominal exchange rate positively impact curvature. The yield curve is negatively impacted by industrial capacity utilization and expected inflation.

Keywords: Functional data analysis; Functional principal component analysis; Term structure of interest rate.
JEL Code: G1, G12.

1. Introduction

The term structure of interest rates (TSIR) is a crucial tool to guide the decision-making process of investors, regulators, risk managers, and others. The database analyzed in this paper can be viewed as observations of a single random function, since the term structure of interest rates defines a relation between the yield of a bond and its maturity. The term structure of interest rates and its sources of variability provide essential information about monetary policy, interest rate risk factors, and fixed-income trading decisions. It is also important to understand the dynamics of bond portfolio management, derivatives pricing, and risk management, among other objectives. Recently, modeling of TSIR evolution has been an active area of research. Many authors seek components, typically additive, answerable for well-defined characteristics of the interest rate curves (Cox et al., 1985). A seminal paper on

Submitted on April 27, 2020. Revised on April 29 and October 26, 2021. Accepted on November 29, 2021. Published online in March 2022. Editor in charge: Marcelo Fernandes.

[†] Federal Center for Technological Education of Minas Gerais, Brazil: lucelia@cefetmg.br
[‡] Universidade Federal de Minas Gerais, Brazil: rjraad@gmail.com

Brazilian Review of Finance (Online), Rio de Janeiro, Vol. 20, No. 1, March 2022, pp. 82–104 ISSN 1679-0731, ISSN online 1984-5146
©2022 Sociedade Brasileira de Finanças, under a Creative Commons Attribution 3.0 license

this topic is that of Litterman and Scheinkman (1991). Using principal component analysis, the authors identify three components that explain around 98% of the variability of US bond prices. These components affect movements in the interest rate curves' level, slope, and curvature.

Term structure models adopted by major central banks can be classified as parametric and spline-based models, according to the Bank for International Settlements (BIS, 2005). Almeida and Faria (2014) find that parametric models fit the yield curve in a parsimonious way. Such methods are typically used in macroeconomic studies, in which smoothness and the ability to capture common movements are as important as model accuracy. This class of models includes the three-factor exponential model of Nelson and Siegel (1987), its four-factor extension proposed by Svensson (1994), and their corresponding dynamic extensions proposed by Diebold and Li (2006) and Pooter (2007).

Spline-based models are made up of several low-order polynomials, which are smoothly linked over the range of maturities. Therefore, splines might estimate a larger number of parameters, with the correspondent fitting curves being less smooth than standard models. Important benchmarks in this class include McCulloch (1975) and Vasicek and Fong (1982). More recent extensions include the penalized spline models of Waggoner (1997) and Chava and Jarrow (2004).

The analysis in this paper is similar to that of Litterman and Scheinkman (1991), and we also estimate a functional linear regression model to investigate the macroeconomic determinants of the yield curves for Brazil. One innovation on Litterman and Scheinkman (1991) is to consider the set of observations as points of a smooth function. The real-valued covariates for the functional regression model include (a) industrial capacity utilization, (b) expected inflation (Broad National Consumer Price Index - IPCA), (c) Selic overnight interest rate of reference, (d) variation of the logarithm of the nominal exchange rate(BRL/USD) (nominal exchange rate, hereafter), and (e) Brazil risk (EMBI+). Our approach is to use functional principal component analysis (PCA) to identify the yield curves' level, slope, and curvature. We then compute the scores of the yield curves on each principal component, and use these scores to assess the relation of level, slope, and curvature to macroeconomic variables.

The main difference between Litterman and Scheinkman (1991) and this paper is that noise present in the data is corrected by imposing smoothness restrictions in the estimation. From an economic perspective, Inoue and Rossi (2019) point out that applying functional data models to term structure and its associated shift provides a more general way to study the impact of monetary policy shocks. Actually, the scalar shocks considered (exogenous movement in the short-term interest rate, forward guidance, and others) can lead to an exogenous shift in the entire yield curve associated with unexpected monetary policy decisions.

Our proposed framework follows a typical analysis of functional data. It begins by using smoothing techniques to represent each observation as a functional object. This first procedure can correct potential problems induced by measurement errors and other types of local disturbances. This is not the case when observations of the yield curves are treated as a multivariate data set (models for repeated measures, longitudinal data of mixed effects, and structural equations). As observed by Levitin et al. (2007), we then set aside the original data, and use the estimated curves for the functional regression model. More specifically, we use cubic spline interpolation to obtain monthly yield curves. After that, the curves form the set of dependent variables in a functional linear regression model with the covariates mentioned above.

The set of dependent variables is composed of curves from January 4, 2010 to December 20, 2018, obtained from Bloomberg. The basic elements of our database are the interest rates of interbank deposits. Price quotations are expressed as a percentage rate per annum, compounded daily, based on a 252-day year. The contracts are those expiring for $t = 1,2,\ldots,39$ months ahead.

We find that a large source of the variability in interest rate curves is due to level shocks. Moreover, it is not affected by the nominal exchange rate. All other variables: expected inflation, Selic reference rate, Brazil risk, and industrial capacity utilization, are positively related to the yield curves' level. Similar results for Brazilian yield curves are presented by Fernandes et al. (2020).

Slope changes are the yield curves' second-largest source of variability. The slope is not associated with expected inflation. The relation between scores on the second principal component and macroeconomic variables is negative for the nominal exchange rate and expected Selic. This means that higher values of the latter variables are related to lower scores, that is, lower-sloped curves. On the other hand, the relationship is positive for Brazil's risk and industrial capacity utilization.

The Selic overnight reference rate, Brazil risk,[1] and the nominal exchange

[1] Brazil risk expresses the credit risk that foreign investors are subject to when investing in the country. The EMBI+ (Emerging Markets Bond Index Plus), calculated by J.P. Morgan Chase Bank, is a weighted index composed of external debt instruments actively traded and denominated in dollars from governments of emerging countries. Its calculation is a weighted average

rate positively affect the yield curvature. Meanwhile, industrial capacity utilization and expected inflation negatively affect curvature. In other words, higher values of the last two variables are associated with lower scores of the third principal component (curves with lower second derivative).

This paper is organized as follows. The second section deals with the methodology of functional models, while also detailing underlying theory. In the third section, we provide descriptive statistics of the databases and the analysis of principal components for the yield curves. The fourth section is devoted to the results of the functional linear regression model of the yield curves against macroeconomic covariates and the traditional linear regression of the scores on the principal functional components against macroeconomic covariates. Our conclusions are presented in the fifth section.

2. Empirical model

To deal with functional data, we must create a suitable representation of each functional object. Here, we use a cubic spline. In the mathematical appendix, we present important considerations. Each vector in $\mathbb{R}^n$ is a column vector.

2.1 Functional linear model

Let us consider $\boldsymbol{y} := \{y_t\}_{t\in\mathbb{Z}}$ wherein[2] each y_t is a real-valued random function[3] with common domain $(0,N]$. More specifically, y_t is the yield curve for month t and domain $(0,39]$. Set uci_t the industrial capacity utilization, $ipca_t$ the average expected inflation, $selic_t$ the Selic interest rate, exr_t the nominal exchange rate, and br_t the Brazil risk. Define[4] $\mathscr{F}_t$ as the information generated by $\{uci_t, ipca_t, selic_t, exr_t, br_t\}$. We assume that

$$\begin{aligned} E(y_t(n)|\mathscr{F}_t) = \mu(n) + \beta_1(n)uci_t + \beta_2(n)ipca_t \\ + \beta_3(n)selic_t + \beta_4(n)exr_t + \beta_5(n)br_t \end{aligned} \tag{1}$$

where $\mu(n)$ is a function that plays the same role as the constant in traditional regression models and the functions $\beta_1(n)$, $\beta_2(n)$, $\beta_3(n)$, $\beta_4(n)$, and $\beta_5(n)$ are coefficients related to each variable.

of the daily returns paid by these securities applied to the previous day's index.

[2] When there is no ambiguity, we omit the variable ω from y_t for the sake of simplicity.

[3] That is, interest rates are given as a Carathéodory function $y : \Omega \times (0,N] \to \mathbb{R}$ where Ω is an underlying probability space. See appendix for more details.

[4] $\mathscr{F}_t$ is also known as the σ-algebra generated by uci_t, $ipca_t$, $selic_t$, exr_t, and br_t.

2.2 Functional principal component analysis

Functional principal component analysis is a key technique to explore features characterizing functions, mainly when the variance-covariance and correlation functions can be challenging to interpret. For $\mathbb{R}^K$-valued data, the principal component analysis is based on the spectral decomposition of the underlying covariance matrix. This is also the case for functional data. More specifically, we determine a set of functions that capture, in decreasing order, the sources of variability of the data (Hall, 2011).

Given the functional decomposition (see (4) in the appendix for more details)

$$\tilde{y}(\omega,n) = \sum_{i=1}^{I} \phi_i(\omega) f_i(n) \text{ for all } n \in (0,N]$$

and an observation of T realizations of $\tilde{y}$, say $\{\tilde{y}(\omega_n,n)\}_{t\leq T}$, let Φ be the $T \times I$-matrix with column vectors $\left(\phi_i(\omega_t) - \sum_{t\leq T} \phi_i(\omega_t)/T\right)_{t\leq T}$ for each $i \in \{1,\ldots,I\}$. Then a direction $\bar{\rho} \in \mathbb{R}^I$ which maximizes variability satisfies

$$\bar{\rho} = \operatorname{argmax}\{\rho^\top \Phi^\top \Phi \rho : \rho \in \mathbb{R}^I \text{ and } \rho^\top \rho = 1\}.$$

Consider $(\lambda_i)_{i\leq I}$ the eigenvalues of $\Phi^\top\Phi$ with $(\lambda_i)_{i\leq I}$ satisfying $\lambda_k \geq \lambda_\kappa$ for $k \leq \kappa \leq I$ with $\{v_i\}_{i\leq I}$ their respective orthonormal basis of eigenvectors. Write V as the $I \times I$ matrix with columns $\{v_i\}_{i\leq I}$. Then each $\rho \in \mathbb{R}^n$ with $\rho^\top\rho = 1$ can be written as[5] $\rho = V\dot{\rho}$ where $\dot{\rho}^\top\dot{\rho} = 1$. Write D_λ as the $I \times I$-diagonal matrix with $(\lambda_i)_{i\leq I}$ into its diagonal. Thus, given $\rho \in \mathbb{R}^I$ with $\rho^\top\rho = 1$, we get

$$\rho^\top\Phi^\top\Phi\rho = \dot{\rho}^\top V^\top \Phi^\top \Phi V \dot{\rho} = \dot{\rho}^\top V^\top V D_\lambda \dot{\rho} = \dot{\rho}^\top D_\lambda \dot{\rho} = \sum_{i\leq I} \dot{\rho}_k^2 \lambda_i \leq \lambda_1.$$

Therefore $\dot{\rho} = (1,0,\ldots,0)$ and $\rho = V\dot{\rho} = v_1$ is the direction which maximizes variability. Finally, define $\hat{y}_1(n) = \sum_{k\in I} v_{1k} f_k(n)$ as the estimated functional principal component. Furthermore, considering now the orthogonal space spanned by $(v_i)_{i>\iota}$ for $\iota > 1$, then we obtain an analogous ι-maximal variability with correspondent functional component $\hat{y}_\iota(n) = \sum_{k\geq\iota} v_{\iota k} f_k(n)$.

2.3 Permutation test

In this paper, we deal with a regression setting where the variations of a functional response are explained by a group of real-valued covariates. Cardot et al. (2004) propose a procedure to check if a real-valued group of covariates

[5] Note that $V^{-1} = V^\top$.

has an effect on a functional response $y_t(n)$. Let $\hat{y}_t(n) = E(y_t(n)|\mathscr{F}_t)$ and $\bar{y}(n) = \frac{1}{T}\sum_{t=1}^{T} y_t(n)$, according to Ramsay and Silverman (2005), the functional version of the univariate F-statistic is given by

$$F(n) = \frac{\sum_{t=1}^{T}(\hat{y}_t(n) - \bar{y}(n))^2/(K-1)}{\sum(y_t(n) - \hat{y}(n))^2/(T-K)} \tag{2}$$

where K is the number of real-valued covariates and T is the total number of observed dependent curves.

3. Database

Table 1 presents descriptive statistics of the yield curves. The average curve is typically upward-sloping, the average rate for 1 month ahead is 10.29%, and for 39 months ahead is 11.48%. This reflects the increasing term premia. The yield curve volatility decreases with maturity. The standard deviation is 2.54% for 1 month ahead and 1.76% for 39 months ahead.

In Figure 1, we observe the term structure of interest rate curves obtained from cubic spline interpolation for December 2009, June 2010, June 2012, December 2005, and June 2016. It shows that the shape and position of the yield curves vary substantially over time. The visual overhaul of the curves allows us to affirm that they move along the vertical axis (level displacement) and present a particular concavity and slope change. For instance, the June 2012 curve is less sloped than December 2019. Furthermore, the December 2019 curve is concave, but this is not true for June 2010.

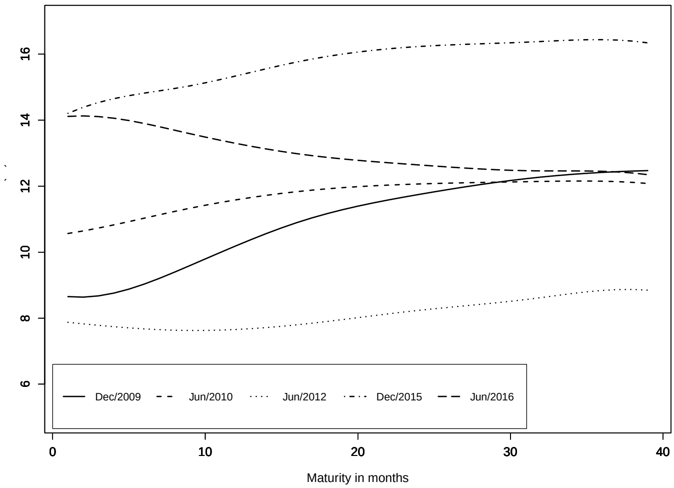
Figures 2 to 4 show the functional mean curve of the interest rate plus/minus a suitable multiple of the first three principal components. The trick of adding/subtracting the principal components to the average was initially used by Ramsay and Silverman (2005), and it is an essential aid to interpreting principal components.[6] It allows us to understand the effect of each principal component on the average. They are responsible for 95%, 4%, and 0,4% of variability, respectively. The first principal component shows that much of the variability of the interest rate curves' average is roughly due to vertical displacements. Such displacements do not affect the shape of the curve. Months with large scores on the second principal component show their first maturities with interest rates lower than the mean, and the opposite for long-term maturities (interest rates greater than the mean). This behavior corresponds to a change in the slope of the interest rate curve. Months with high coordinates

[6] More details on functional principal component analysis are given by Ramsay and Silverman (2005). We use the R package's command pca.fd to estimate the functional principal components.

Table 1
Descriptive statistics for nominal yield curves observed from January 2010 to December 2018 (% per annum)

maturity	mean	max	min	s.d.
1	10.29	14.26	6.34	2.54
2	10.29	14.34	6.33	2.55
3	10.30	14.50	6.27	2.56
4	10.32	14.67	6.24	2.57
5	10.32	14.83	6.22	2.56
6	10.34	15.02	6.22	2.57
7	10.35	15.17	6.21	2.55
8	10.38	15.35	6.21	2.54
9	10.41	15.50	6.23	2.55
10	10.43	15.62	6.27	2.51
11	10.46	15.72	6.29	2.50
12	10.50	15.79	6.32	2.50
13	10.56	15.85	6.30	2.47
14	10.60	15.94	6.73	2.43
15	10.62	16.00	6.49	2.48
16	10.71	16.13	6.48	2.44
17	10.72	15.91	6.96	2.35
18	10.74	15.66	6.72	2.34
19	10.86	16.32	6.75	2.35
20	10.83	16.06	7.27	2.18
21	10.89	15.84	7.00	2.19
22	11.01	16.47	7.25	2.23
23	10.96	16.14	7.50	2.07
24	11.01	15.96	7.70	2.09
25	11.14	16.53	7.71	2.11
26	11.09	16.18	7.83	1.97
27	11.14	16.09	7.82	2.00
28	11.28	16.63	7.84	2.01
29	11.20	16.20	7.98	1.89
30	11.24	16.20	7.93	1.92
31	11.37	16.69	7.97	1.93
32	11.28	16.22	8.11	1.81
33	11.34	16.27	8.01	1.84
34	11.46	16.70	8.08	1.86
35	11.37	16.23	8.23	1.76
36	11.40	16.33	8.11	1.80
37	11.53	16.71	8.19	1.80
38	11.44	16.24	8.38	1.71
39	11.48	16.37	8.21	1.76

Figure 1
Yield curves obtained through spline interpolation



in this component present more sloped curves. The third principal component shows a change in the curvature of the mean interest rate curve.

Table 2 has descriptive statistics for the covariates considered here. Average inflation expectation (Broad National Consumer Price Index - IPCA) accumulated rate for the next 12 months (% a.a.). The Selic overnight interest rate of reference (% a.m). The variation of the logarithm of the nominal exchange rate (BRL/USD). The measure of Brazil's risk is the EMBI+ (Emerging Markets Bond Index Plus); the unit of measure for this index is the base point, with ten base points equal to 0.1%. Industrial capacity utilization (%) is a proxy for Brazil's real economic activity.

Table 2
Descriptive statistics for covariates

covariate	mean	max	min	s.d.
expected inflation (% per annum)	5.40	7.29	3.40	0.90
Selic interest rate (Over/Selic - (% per month)	0.82	1.22	0.47	0.20
nominal exchange rate	–0.24	0.11	–25.06	2.41
Brazil risk (EMBI+)	146.27	531.29	253.33	77.66
industrial capacity utilization (%)	80.42	85.00	74.90	2.73

Figure 2
The mean yield curve added(+) to and subtracted(−) from the first principal component

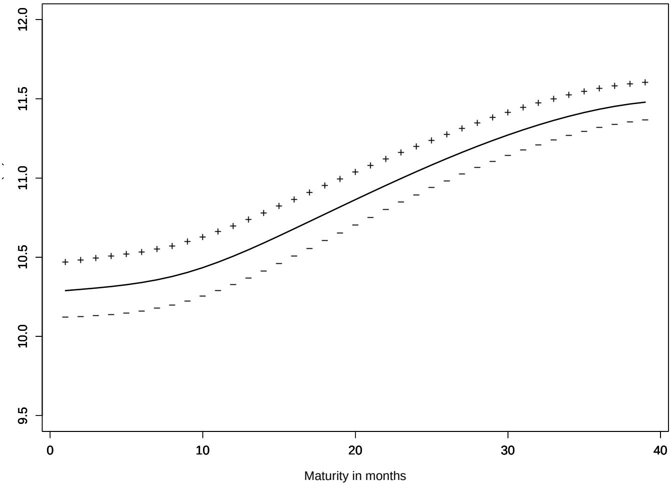


Figure 3
The mean yield curve added(+) to and subtracted(−) from the second principal component

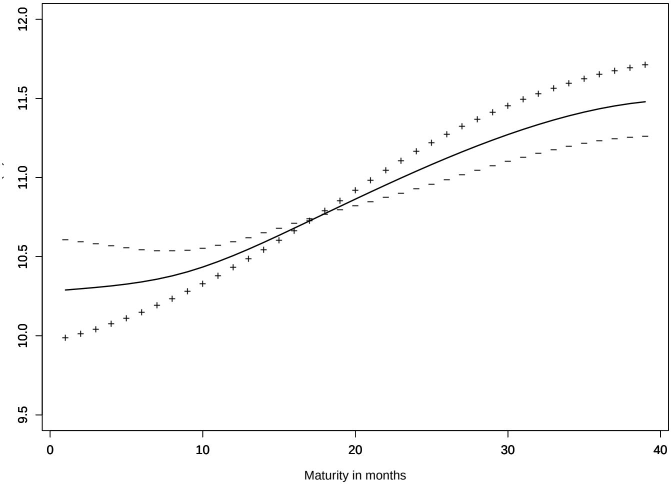
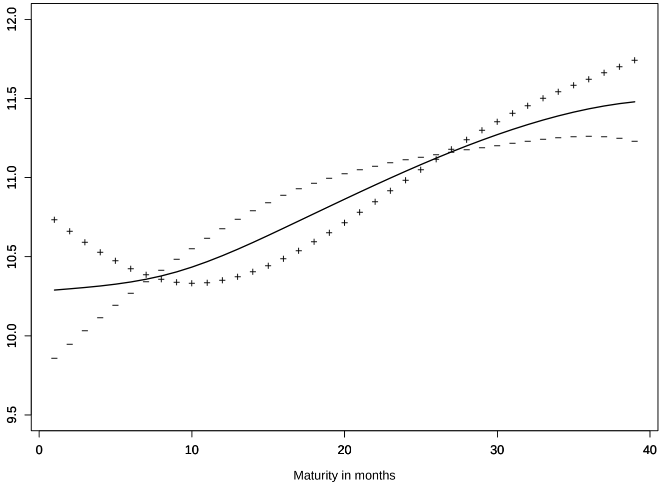


Figure 4
The mean yield curve added(+) to and subtracted(−) from the third principal component



4. Regression results

Figures 5 to 9 show estimated coefficients for each explanatory variable. Since our coefficients are functions with the same domain as the yield curves, they can capture changes in the dependent variable's level, slope, and curvature. Nevertheless, in this approach, it is impossible to specify which variable most affects each component of the yield curve. Apart from the nominal exchange rate coefficient, the estimated coefficients are positive. For the Selic specifically, the result aligns with the expectation hypothesis. The coefficient is decreasing through the maturities, indicating that long-term maturities are weakly related to current values of Selic, compared with short-term maturities.

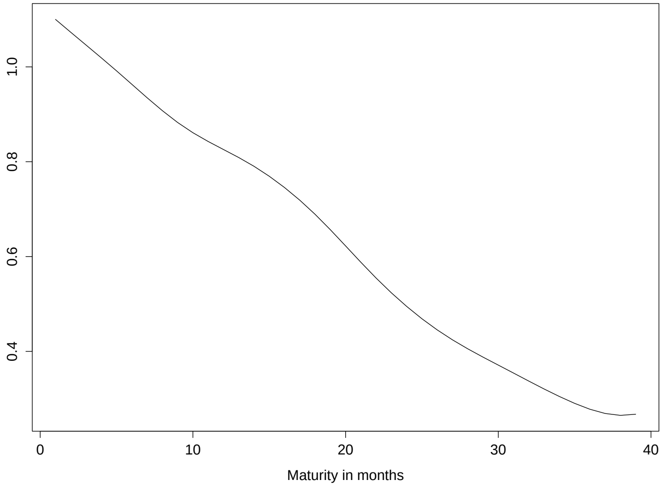
Our results allow us to claim that higher values of the variation in the current nominal exchange rate reduce the slope of the yield curves, that is, the term premia. The coefficient associated with the nominal exchange rate is negative for most yield curve domains, and more intense for long-term maturities. In the next section, we show the relationship between macroeconomic variables and the components of the term structure of the interest rate. Our results show that the nominal exchange rate is related only to the yield curve's slope for a 5% level of significance.

The positive relation between the Brazil risk and yield curve reflects the fact that in the face of a higher level of risk, investors demand a higher level of compensation (Fernandes et al., 2020). The macroeconomic literature points out the existence of inertia in economic activity (Franses and Paap, 2004). Thus, current values of industrial production should be more strongly related to long-term-maturity interest rate, as we find here.

We expect an inverse relationship between expected inflation (IPCA) and interest rates, since lower interest rate levels should increase real activity, implying higher prices. However, the Brazilian monetary authority has used interest rates to keep inflation on target since implementing an inflation targeting regime on June 1, 1999. This use of the interest rates explains the positive coefficient of expected inflation.

Figure 10 shows the pointwise values of the observed F-statistic and the pointwise 0.95 quantiles of the distribution induced by the null hypothesis. The model is pointwise statistically significant (at a 5% level) to explain the variability of the term structure. The observed F-statistic to access the overall[7] significance of the model is 12.23, and the 0.95 quantile of the null distribution is 0.13.

Figure 5
The estimated coefficient for the Selic reference rate



[7] More details about the F-type test for overall significance of the model can be found in Ramsay and Silverman (2005) and Zhang (2014)

Figure 6
The estimated coefficient for industrial capacity utilization

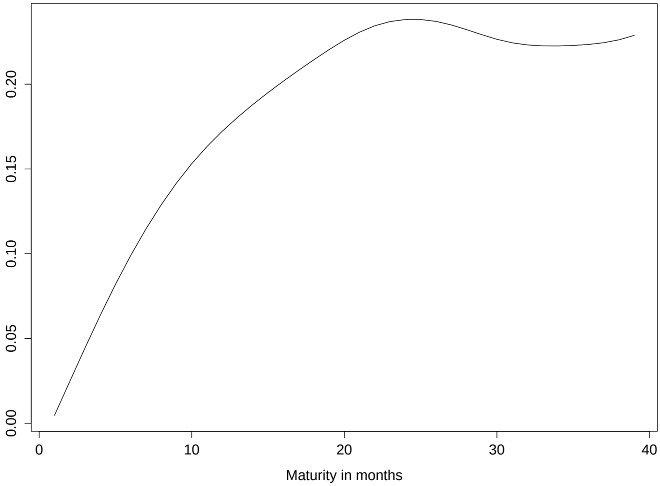


Figure 7
The estimated coefficient for expected inflation (IPCA)

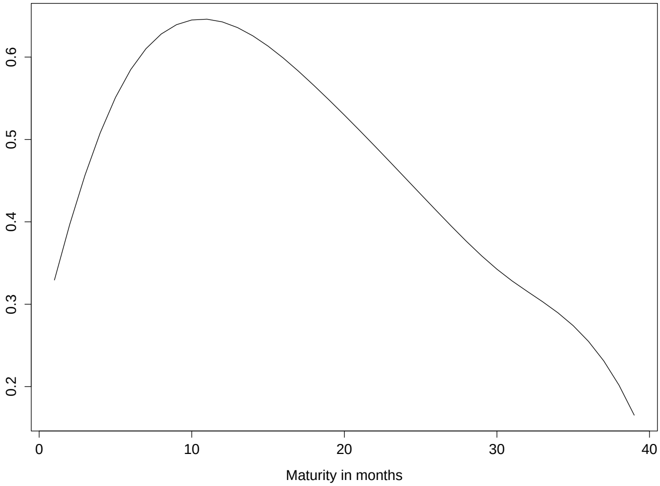


Figure 8
The estimated coefficient for the nominal exchange rate

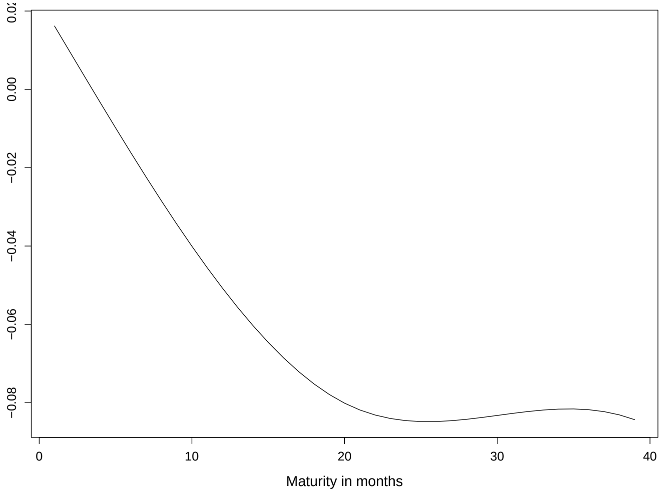


Figure 9
The estimated coefficient for the Brazil risk

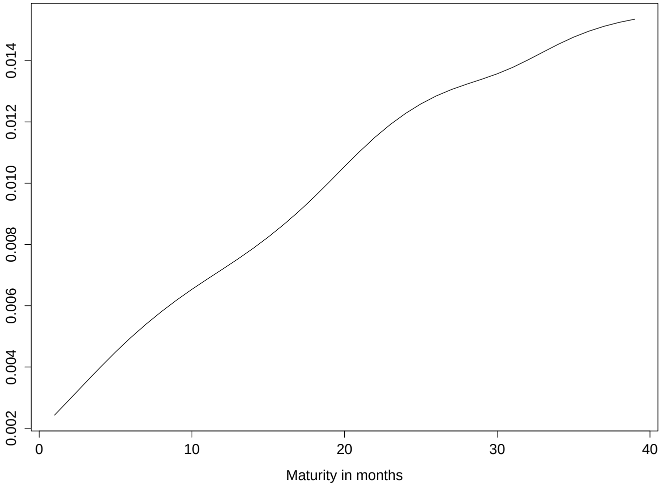
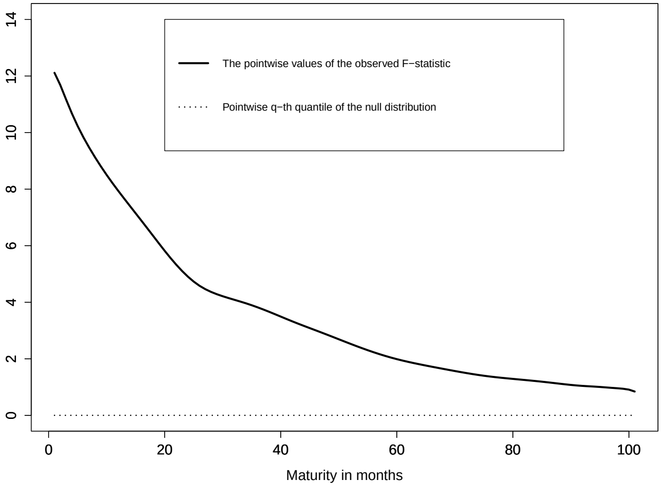


Figure 10
Permutation test for the coefficients of the functional linear regression model



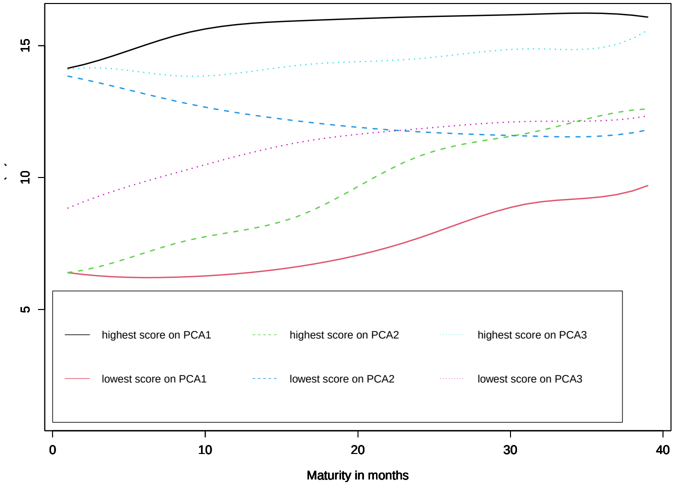
4.1 Regression results for term structure of interest rate components

Figure 11 shows curves selected according to their scores on the first, second, and third principal components. The maximum(minimum) on the first principal component corresponds to a high(low)-level curve. Higher(lower) scores are associated with high(low) sloped curves for the second principal component. Concave functions present lower scores in the third principal component. That is, the second derivative of the curve is increasing with the score. To understand the relation between level, slope, and curvature and the macroeconomic covariates, we conduct a traditional linear regression of each set of scores on the set of macroeconomic covariates.

The nominal exchange rate does not affect the level of the yield curves. All other variables, such as expected inflation, Selic interest rate, Brazil risk, and industrial capacity utilization, are positively related to the term structure of the interest rate.

The yield curve slope correlates significantly with the risk premium, so the results presented here can be extended to the latter. Concerning the Selic reference rate and nominal exchange rate, the relation is negative (Fernandes et al., 2020). Larger values of these variables are related to smaller scores corresponding to lower sloped curves. In contrast, the slope increases with Brazil's risk and industrial capacity utilization. Notably, the negative sign of

Figure 11
Curves selected according to their scores on the first, second, and third principal components



the Selic reference rate coefficient means that agents demand higher risk premiums in the face of low Selic reference rate values. Nevertheless, we could also expect the opposite, since a low Selic reference rate would be related to a favorable macroeconomic scenario. It positively affects the agent's expectations, so they would not require a higher risk premium. The Brazilian Central Bank has used the Selic reference rate since 2008/2009 to keep inflation on target, so the current low values of the Selic reference rate can be related to high inflation levels and recession scenarios in the future. It justifies the negative sign of the Selic reference rate. The Brazil risk coefficient signal is in line with expectations, since a higher Brazil risk should be related to a higher risk premium. Additionally, higher levels of industrial capacity utilization increase inflation expectations and pressure to increase the risk premium.

Traditionally, the curvature of the yield curve is related to an anticipation of economic contraction. Expected inflation and industrial capacity utilization negatively affect the curvature of the yield curve. This means that larger values of these variables are related to lower scores on the third principal component, curves with lower second-order derivatives (concave curves). On the other hand, the Selic reference rate, the nominal exchange rate, and Brazil risk positively affect the curvature of the yield curve. Higher values of the latter variables induce comparatively flat curves that correspond to the anticipation of an economic contraction.

Table 3
Linear regression of the principal components against covariates

covariate	estimate	s.d.	*t*-stat.	*p*-value
		PCA 1 - level		
intercept	–24.8520	4.8310	–5.1443	0.0000
expected inflation	0.4634	0.1476	3.1389	0.0022
Selic interest rate	6.6021	0.6435	10.2600	0.0000
nominal exchange rate	–0.3541	0.2472	–1.4323	0.1551
Brazil risk	0.0578	0.0138	4.1769	0.0001
industrial capacity utilization	0.1808	00.0594	3.0457	0.0030
		PCA 2 - slope		
intercept	–6.3030	1.8075	–3.4872	0.0007
expected inflation	–0.0434	0.0552	–0.7861	0.4336
Selic interest rate	–1.7234	0.2408	–7.1584	0.0000
nominal exchange rate	–0.2216	0.0908	–2.4415	0.0163
Brazil risk	0.0317	0.0051	6.2467	0.0000
industrial capacity utilization	0.0826	0.0222	3.7194	0.0003
		PCA 3 - curvature		
intercept	1.2066	0.4504	2.6790	0.0086
expected inflation	–0.0659	0.0138	–4.7868	0.0000
Selic interest rate	0.0875	0.0600	1.4580	0.1479
nominal exchange rate	0.0386	0.0230	1.6796	0.0961
Brazil risk	0.0053	0.0013	4.0895	0.0001
industrial capacity utilization	–0.0141	0.0055	–2.5557	0.0121

5. Concluding remarks

In this paper, through functional principal component analysis, we identify that the significant source of variability (around 95%) of yield curves is due to level displacements. Using a functional linear regression model, we present the coefficients for the macroeconomic variables: (a) industrial capacity utilization, (b) expected inflation (Broad National Consumer Price Index - IPCA), (c) Selic Over (d) the nominal exchange rate(BRL/USD), and (e) Brazil risk (EMBI+). In the functional linear regression model approach, it is not possible to specify which variable most affects each component of the yield curve. To understand the relation between level, slope, and curvature and the macroeconomic covariates, we conduct a traditional linear regression model of each set of scores on principal components over the set of macroeconomic covariates.

References

Akram, T. and Uddin, S. A.-H. (2021). An empirical analysis of long-term Brazilian interest rates, *PloS one* **16**(9): e0257313.

Almeida, C. and Faria, A. (2014). Forecasting the Brazilian term structure using macroeconomic factors, *Brazilian Review of Econometrics* **34**(1): 45–77.

Billingsley, P. (2008). *Probability and measure*, John Wiley & Sons.

BIS (2005). *Zero-coupon yield curves: Technical documentation*, Bank for International Settlements.

Caldeira, J. F., Gupta, R., Suleman, M. T. and Torrent, H. S. (2020). Forecasting the term structure of interest rates of the BRICS: Evidence from a nonparametric functional data analysis, *Emerging Markets Finance and Trade* pp. 1–18.

Caldeira, J. and Torrent, H. (2017). Forecasting the US term structure of interest rates using nonparametric functional data analysis, *Journal of Forecasting* **36**(1): 56–73.

Cardot, H., Goia, A. and Sarda, P. (2004). Testing for no effect in functional linear regression models, some computational approaches, *Communications in Statistics-Simulation and Computation* **33**(1): 179–199.

Charalambos, D. and Aliprantis, B. (2013). *Infinite Dimensional Analysis: A Hitchhiker's Guide*, Springer-Verlag Berlin and Heidelberg GmbH & Company KG.

Chava, S. and Jarrow, R. A. (2004). Bankruptcy prediction with industry effects, *Review of Finance* **8**(4): 537–569.

Cox, J. C., Ingersoll, J. E. and Ross, S. A. (1985). A theory of the term structure of interest rates, *Econometrica* **53**(2): 385–407.
URL: http://www.jstor.org/stable/1911242

da Silveira, G. B. and Bessada, O. (2003). Análise de componentes principais de dados funcionais–uma aplicação às estruturas a termo de taxas de juros, *Working Papers Series 73*, Central Bank of Brazil, Research Department.
URL: https://ideas.repec.org/p/bcb/wpaper/73.html

De Boor, C. (1978). *A practical guide to splines*, Vol. 27, Springer-Verlag New York.

Diebold, F. X. and Li, C. (2006). Forecasting the term structure of government bond yields, *Journal of Econometrics* **130**(2): 337 – 364.
URL: https://doi.org/10.1016/j.jeconom.2005.03.005

Duffee, G. R. (2002). Term premia and interest rate forecasts in affine models, *Journal of Finance* **57**(1): 405–443.

Fernandes, M., Nunes, C. and Reis, Y. (2020). What drives the nominal yield curve in Brazil?, *Brazilian Review of Econometrics* **40**(2): 267–284.

Ferraty, F., Mas, A. and Vieu, P. (2007). Nonparametric regression on functional data: Inference and practical aspects, *Australian & New Zealand Journal of Statistics* **49**(3): 267–286.

Franklin JR., S. L., Duarte, T. B., Neves, C. R. and Melo, E. F. L. (2012). A estrutura a termo de taxas de juros no Brasil: Modelos, estimação e testes, *Economia Aplicada* **16**(2): 255–90.
URL: http://dx.doi.org/10.1590/S1413-80502012000200003

Franses, P. H. and Paap, R. (2004). *Periodic time series models*, OUP Oxford.

Ganem, M. and Baidya, T. K. N. (2011). Assimetria e prêmio de risco na estrutura a termo de juros brasileira, *Revista Brasileira de Finanças* **9**(2): 277–301.

Green, P. J. and Silverman, B. W. (1993). *Nonparametric regression and generalized linear models: A roughness penalty approach*, CRC Press.

Hall, P. (2011). Principal component analysis for functional data: methodology, theory, and discussion, *The Oxford handbook of functional data analysis*, Oxford: Oxford University Press, pp. 210–234.

Heath, D., Jarrow, R. and Morton, A. (1992). Bond pricing and the term structure of interest rates: A new methodology for contingent claims valuation, *Econometrica* **60**(1): 77–105.

Hicks, J. R. (1946). *Value and Capital*, 2 edn, Clarendon Press, Oxford.

Hoffman, K. and Kunze, R. (1971). *Linear Algebra*, New Jersey: Englewood Cliffs.

Hull, J. and White, A. (1990). Pricing interest-rate-derivative securities, *Review of Financial Studies* **3**(4): 573–592.

Inoue, A. and Rossi, B. (2019). The effects of conventional and unconventional monetary policy on exchange rates, *Journal of International Economics* **118**: 419–447.

Levitin, D. J., Nuzzo, R. L., Vines, B. W. and Ramsay, J. (2007). Introduction to functional data analysis., *Canadian Psychology* **48**(3): 135.

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

McCulloch, J. H. (1975). An estimate of the liquidity premium, *Journal of Political Economy* **83**(1): 95–119.

McLeod, A. I. (1978). On the distribution of residual autocorrelations in Box-Jenkins models, *Journal of the Royal Statistical Society. Series B (Methodological)* **40**(3): 296–302.
URL: http://www.jstor.org/stable/2984693

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

Neto, A. A. (2003). *Finanças corporativas e valor*, Atlas.

Pooter, M. D. (2007). Examining the Nelson-Siegel class of term structure models, *Tinbergen Institute Discussion Papers 07-043/4*, Tinbergen Institute.

Ramsay, J. O. and Dalzell, C. J. (1991). Some tools for functional data analysis, *Journal of the Royal Statistical Society. Series B (Methodological)* **53**(3): 539–572.
URL: http://www.jstor.org/stable/2345586

Ramsay, J. O. and Silverman, B. W. (2005). *Functional data analysis*, 2 edn, Springer, New York.

Reinsch, C. H. (1967). Smoothing by spline functions, *Numerische mathematik* **10**(3): 177–183.

Royden, H. L. and Fitzpatrick, P. (1988). *Real analysis*, Vol. 32, Macmillan New York.

Senturk, D. and Muller, H.-G. (2010). Functional varying coefficient models for longitudinal data, *Journal of the American Statistical Association* **105**(491): 1256–1264.
URL: https://doi.org/10.1198/jasa.2010.tm09228

Stona, F. and Caldeira, J. F. (2019). Do US factors impact the Brazilian yield curve? Evidence from a dynamic factor model, *North American Journal of Economics and Finance* **48**: 76–89.

Svensson, L. E. (1994). Estimating and interpreting forward interest rates: Sweden 1992-1994, *NBER Working Papers 4871*, National Bureau of Economic Research.
URL: https://EconPapers.repec.org/RePEc:nbr:nberwo:4871

Vasicek, O. A. and Fong, H. G. (1982). Term structure modeling using exponential splines, *Journal of Finance* **37**(2): 339–348.

Vasicek, O. A., Fong, H. G. and Vasicek, O. A. (2015). Term structure modeling using exponential splines, *Finance, Economics, and Mathematics* p. 49.

Vaz, L. V. and da Silveira Filho, G. B. (2017). Functional autoregressive models: An application to Brazilian hourly electricity load, *Brazilian Review of Econometrics* **37**(2): 297–325.

Waggoner, D. F. (1997). Spline methods for extracting interest rate curves from coupon bond prices, *FRB Atlanta Working Paper 97-10*, Federal Reserve Bank of Atlanta.
URL: https://EconPapers.repec.org/RePEc:fip:fedawp:97-10

Wahba, G. (1990). *Spline models for observational data*, SIAM.

Zhang, J.-T. (2014). *Analysis of Variance for Functional Data*, Monographs on Statistics and Applied Probability, Chapman and Hall/CRC.

A. Mathematical appendix

Suppose that time is defined in the interval $(0,N]$ endowed with the standard Lebesgue measure λ characterizing smoothness and that uncertainty is represented by a probability space $(\Omega,\rho,\mathscr{S})$ where ρ is the objective probability governing uncertainty and $\mathscr{S}$ its information sigma-algebra. Consider $C^2((0,N],\mathbb{R})$ as the set of all twice differentiable functions with domain $(0,N]$ and codomain $\mathbb{R}$. Clearly, $C^2((0,N],\mathbb{R})$ is a vector space.[A1] A smooth functional data could be summarized by a measurable map $y:\Omega\times(0,N]\to\mathbb{R}$ such that $y(\omega,\cdot)\in C^2((0,N],\mathbb{R})$ for each $\omega\in\Omega$. Write S_{F} as the set of all such maps. Given an observation of $\omega\in\Omega$, we could also observe the function $\tilde{y}_\omega:(0,N]\to\mathbb{R}$ defined by $\tilde{y}_\omega=y(\omega,\cdot)$. This is the precise definition of a random function, and we will denote it shortly by $\tilde{y}$ or $\tilde{y}(n)$ instead of $\tilde{y}_\omega$ or $\tilde{y}_\omega(n)$ when there is no ambiguity.

The model embodies two types of data. The first is a functional data denoted by $y\in S_{\mathrm{F}}$ representing the interest rates and maturities of the contracts on each day. This variable is observed over a partition $0=n_0<n_1<\cdots<n_I=N$ of $(0,N]$. The second is denoted by measurable real-valued functions $x=(x_1,\ldots,x_{\mathrm{K}})$ defined on $(\Omega,\rho,\mathscr{S})$ and representing the control variables. As we observe discrete values of y, we build the estimation in two stages. In the first stage, we use a smoothing metric to estimate y, and in the second we perform an OLS estimation between the random variables $y(n)$ and $x(n)$ for each n in the usual way. The tradeoff between smoothness and bias on an efficient fit may be addressed through the following problem of minimization. Given a realization $\omega\in\Omega$ then we observe the values $\dot{y}_i=y(\omega,n_i)$ for all $i\leq I$ and solve

$$\tilde{y}(\omega,n) \;=\; \underset{f\in C^2\left((0,N],\mathbb{R}\right)}{\operatorname{argmax}} \left\{\sum_{i\in I}(\dot{y}_i-f(n_i))^2+\sigma\int_{(0,N]}(f''(n))^2\lambda(dn)\right\}. \qquad (3)$$

where σ is the smoothness parameter. The first term in (3) is a measure of goodness of data fit to the function f, and the second is related to the smoothness penalty of the estimated function, weighted by the curvature f''. The positive constant σ is the smoothing parameter. Large values of σ produce smoother curves. It can be shown (Wahba, 1990) that the unique function $\tilde{y}$ which minimizes (3), for a fixed σ, is a natural cubic spline with knots at the n_i for $i\leq I$. Although each solution $\tilde{y}$ of (3) depends randomly on each realization $(\dot{y}_i)_{i\leq I}$, there is a deterministic spline basis $\{f_i\}_{i\leq I}\subset\mathbb{S}_{\mathrm{B}}(M)$ and

[A1] See Charalambos and Aliprantis (2013) or Royden and Fitzpatrick (1988) for detail about vector spaces.

time independent random coefficients $\{\phi_i\}_{i\leq I}$ such that $\tilde{y}$ can be decomposed in the following way:

$$\tilde{y}(\omega,n) = \sum_{i=1}^{I} \phi_i(\omega) f_i(n) \text{ for all } n \in (0,N]. \tag{4}$$

Therefore, the estimation can be thoroughly defined in terms of realizations of the random coefficients $\{\phi_i\}_{i\leq I}$. The functions f_k obtained from a minimization of the induced estimation metric is a cubic spline (Reinsch, 1967). Thus we can work on the subset of cubic splines on $C^2(0,N)$ as detailed below.

The decomposition of random functions in terms of splines[A2] can be summarized as the following for variables and parameters of the functional model. A *spline* is a function $f : (0,N] \to \mathbb{R}$, piecewise polynomial. Each spline f is associated to the fixed partition $0 = n_0 < n_1 < \cdots < n_I < n_I = N$ in such a way that the restriction of f to each $[n_i,n_{i+1}]$ and $i \in \{0,1,\ldots,I\}$ is a polynomial f_i. The *degree* of the spline f, denoted by m, is the greatest degree of the $f_i's$. We also impose that f and its $M-1$ first derivatives are continuous. The points $n_0,n_1,\ldots,n_I$ are called *knots* of the spline f. A spline is called *natural* when both f_0 and f_n are polynomials of degree one. We call the vector space of all splines with degree M by $\mathbb{S}(M)$ and $\mathbb{S}(3)$ the set of cubic splines. Clearly, $\mathbb{S}(M)$ is a vector subspace of $C^2((0,N],\mathbb{R})$ and has finite dimension[A3]. Indeed, consider $Ind_i : (0,N] \to \{0,1\}$ the indicator function on the interval $[n_i,n_{i+1}]$ for all $i \leq I$. Then $\mathbb{S}(M)$ is generated by a finite spline basis $\mathbb{S}_{\mathrm{B}}(M) \subset C^2((0,N],\mathbb{R})$ of some fixed linear combinations of functions[A4] $f : (0,N] \to \mathbb{R}$ that can be written for some $m \in \{0,1,2,\ldots,M\}$ and $i \in \{0,1,\ldots,I\}$ as $f(n) = n^m Ind_i(n)$ for all $n \in (0,N]$.

For the second stage of estimation, define $E[z] = \int_\Omega z(\omega)\rho(d\omega)$ and consider the space of all $z : \Omega \to \mathbb{R}$ with $E[z] = 0$ and $E[z^2] < \infty$. Let us call this space $L^2(\Omega,\rho,\mathscr{S})$. It is easy to see that $L^2(\Omega,\rho,\mathscr{S})$ is a Hilbert space under the inner product $\langle z,\dot{z}\rangle = E(z\dot{z})$. The norm is defined in the standard way $||z|| = \langle z,z\rangle$.

Given a linear independent random vector $x : \Omega \to \mathbb{R}^{\mathrm{K}}$ with $x = (x_1,\ldots,x_{\mathrm{K}})$ there exists an orthonormal basis $\hat{x} = (\hat{x}_1,\ldots,\hat{x}_{\mathrm{K}})$, spanning the vector sub-

[A2] For further details see, for example, Green and Silverman (1993) and De Boor (1978).

[A3] More precisely, $\mathbb{S}(M)$ has dimension at most $M(I+1)$. A vector subspace of a finite-dimensional vector space has finite dimension. For example, if $\mathbb{S}(M)$ does not have finite dimension, then one can obtain a linear independent subset of $\mathbb{S}(M)$ with $m > M(I+1)$ elements that is a contradiction.

[A4] The space $\mathbb{S}'(M)$ generated by the set of such functions $f : (0,N] \to \mathbb{R}$ is finite dimensional and $\mathbb{S}(M)$ is a vector subspace of $\mathbb{S}'(M)$.

space generated by x and the conditional expectation[A5] of $y(n)$ over x is given by (Billingsley, 2008)

$$E[y(n)|\hat{x}] = \langle y(n), \hat{x}_1 \rangle \hat{x}_1 + \cdots + \langle y(n), \hat{x}_{\mathrm{K}} \rangle \hat{x}_{\mathrm{K}}.$$

Therefore, since there exists a $k \times k$ real-valued matrix mapping $\hat{x}$ on x then there exist constants $(\alpha_1(n), \ldots, \alpha_{\mathrm{K}}(n))$ such that

$$E[y(n)|x](\omega) = \alpha_1(n)x_1(\omega) + \cdots + \alpha_{\mathrm{K}}(n)x_{\mathrm{K}}(\omega). \tag{5}$$

Define $\varepsilon : \Omega \to \mathbb{R}$ by $\varepsilon(\omega,n) = y(\omega,n) - E[y(n)|x](\omega)$. Then $E(\varepsilon) = 0$, $\langle \varepsilon, x_k \rangle = 0$ for all $k \leq \mathrm{K}$ and the true random variables satisfy

$$y(\omega,n) = \alpha_1(n)x_1(\omega) + \cdots + \alpha_{\mathrm{K}}(n)x_{\mathrm{K}}(\omega) + \varepsilon(\omega,n) \text{ for all } (\omega,n) \in \Omega \times (0,N]. \tag{6}$$

Even if y and x are not centralized variables we can rewrite the conditional expectation (5) as

$$\begin{aligned} y(\omega,n) = {} & \mu(n) + \beta_1(n)x_1(\omega) \\ & + \cdots + \beta_{\mathrm{K}}(t)x_{\mathrm{K}}(\omega) + \varepsilon(\omega,n) \text{ for all } (\omega,n) \in \Omega \times (0,N]. \end{aligned}$$

[A5] We write $y(\omega,n)$ shortened to $y(n)$ to simplify.