# Contents

# 1 Forecasting discount curves with Kernel Regularized Least Squares

## 1.1 Introduction

In this chapter, we apply Kernel Regularized Least Squares (KRLS) learning methods to Yield Curve forecasting. By *Yield Curve*, we actually mean *discount curves*. That is, we consider that the curves used in the examples do not include any counterparty credit risk, and focus on the forecasting problem. Two types of formulations of the spot rates' forecasting problem are tested here. One relying on the popular Dynamic Nelson-Siegel (DNS) framework from [DL06], and another one, in which we apply the KRLS directly to the discount curves' observation dates and time to maturities, to model the spot rates.

In the DNS framework [DL06], each cross-section of yields observed over time is fitted by using the Nelson-Siegel [NS87] model. The fitting of each cross-section observed over time, produces three time series of parameters (more details in the next section) representing the evolution of the level, slope, and curvature of the Yield Curve. A KRLS model is applied to forecasting the time series of parameters, using a technique which is similar to the one described in [**exterkate2016nonlinear**]. And to finish, the forecast obtained for the trivariate time series are plugged into the Nelson-Siegel model formula, to deduce forecast for the cross-sections of yields.

The second approach based on KRLS is a machine learning/data-driven one, in which we put no specific constraint on the model to reproduce the specific Yield Curve stylized facts. The regularization parameters inherent to the KRLS models will act as implicit constraints, that cause the model to converge as close as possible to reproducing these stylized facts. In this latter approach, we are mostly interested in the model with the *best* out-of-sample error. As a consequence, the technique as is, is probably less adapted than the former framework based on DNS (in its arbitrage-free version) to no-arbitrage pricing (if no-arbitrage pricing is required).

To introduce KRLS, we start by describing the ridge regression [**hoerl1970ridge**] and the *kernel trick* applied to ridge regression. Then, we make a link between the ridge regression and KRLS.

In a ridge regression setting, we want to explain an observed variable $y \in \mathbb{R}^n$, as a linear function of $p$ predictors stored in a matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$. For $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, p\}$ we have:

$$X_{ij} =: \mathbf{x}_i^{(j)}$$

We will denote the $i^{th}$ row of $\mathbf{X}$ as $\mathbf{x}_i$, and its $j^{th}$ column as $\mathbf{x}^{(j)}$. Hence, we are searching for the parameters $\beta = (\beta_1, \dots, \beta_p)^T$ verifying:

$$ArgMin_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \left( y_i - \mathbf{x}_i^T \beta \right)^2$$

under the constraint

$$||\beta||_2^2 \leq s$$

The solution to this problem is given directly by the formula:

$$\hat{\beta} = \left( \mathbf{X}^T \mathbf{X} + \lambda I_{p \times p} \right)^{-1} \mathbf{X}^T y$$

where $\lambda$ is a Lagrange multiplier having a unique correspondance with $s$, and a regularization parameter preventing the model from overfitting the observed data contained in $y$. In the case where we want to explain $y$ as a function $\Phi$ of the predictors, we have a similar expression:

$$\hat{\beta} = \left( \Phi(\mathbf{X})^T \Phi(\mathbf{X}) + \lambda I_{p \times p} \right)^{-1} \Phi(\mathbf{X})^T y$$

where:

$$\Phi(\mathbf{X})_{ij} = \Phi(\mathbf{x}_i^{(j)})$$

Now, by using the Woodbury identity (cite Gene H. Golub and Charles F. van Loan. Matrix Computations and cite Max Welling The Kalman filter, Lecture Note) for $\mathbf{P}$ and $\mathbf{R}$ positive definite

$$\left( \mathbf{P}^{-1} + \mathbf{B}^T \mathbf{R}^{-1} \mathbf{B} \right)^{-1} \mathbf{B}^T \mathbf{R}^{-1} = \mathbf{P} \mathbf{B}^T \left( \mathbf{B} \mathbf{P} \mathbf{B}^T + \mathbf{R} \right)^{-1}$$

The solution to the ridge regression problem can be re-written as:

$$\hat{\beta} = \Phi(\mathbf{X})^T \left( \Phi(\mathbf{X}) \Phi(\mathbf{X})^T + \lambda I_{n \times n} \right)^{-1} y$$

This relationship can be useful in the case where $n << p$. That is, when there is a high number of predictors compared to the number of observations (cite Exterkate (2016)). Indeed, with this new relationship, we are no longer calculating/inverting a $p \times p$ matrix, but a $n \times n$ matrix. That's the *kernel trick*. And if some new observations arrive, and are stored in $\mathbf{X}^*$, the new values predicted by the model will be given by:

$$y^* = \Phi(\mathbf{X}^*)\hat{\beta} = \Phi(\mathbf{X}^*)\Phi(\mathbf{X})^T \left(\Phi(\mathbf{X})\Phi(\mathbf{X})^T + \lambda I_{n\times n}\right)^{-1} y$$

Which we re-write as:

$$y^* = \mathbf{K}^* \left(\mathbf{K} + \lambda I_{n\times n}\right)^{-1} y$$

$\mathbf{K}$ is a *kernel*; the empirical covariance matrix of $\Phi(\mathbf{X})^T$ (modulo a $1/p$ factor), in the case where the rows of $\Phi(\mathbf{X})$ are centered. Now, in the case of KRLS, the problem we are trying to solve is:

$$ArgMin_{c\in\mathbb{R}^n} \sum_{i=1}^{n} \left(y_i - K_i^T c\right)^2$$

where $K_i$ is the $i^{th}$ row of $\mathbf{K}$, with:

$$\mathbf{K}_{ij} =: K(\mathbf{x}_i, \mathbf{x}_j) = f(||\mathbf{x}_i - \mathbf{x}_j||_1) \ \ or \ \ f(||\mathbf{x}_i - \mathbf{x}_j||_2)$$

The cost of computing the whole kernel $\mathbf{K}$, for any $i$ and $j$ is a quadratic function of the number of observations, $n$. Meaning that, the most interesting cases for using the KRLS method would be those in which $n$ is not too high.

As described in (cite Halmueller & Hazlett (2013)), two approaches can be used to interpret/motivate the KRLS learning method: a similarity-based view and the superposition of Gaussians view. Here, we refer only to the first one, which is the most intuitive to us, and is also the one described in (cite Ferwerda, Hainmueller et Hazlett (2017)). Indeed here, the $i^{th}$ observation of the response, $y_i$ is explained as a linear combination of functions measuring the similarity/dissimilarity between its

characteristics gathered in $\mathbf{x}_i$, and the other observations from the training set, $\mathbf{x}_j$, $j \neq i$:

$$y_i = \sum_{j=1}^{n} c_j K(\mathbf{x}_i, \mathbf{x}_j) \tag{1.1}$$

But again, to prevent the model from overfitting the observed data and not being able to generalize well, we need to constrain the parameters $c$ under a certain norm defined by the matrix $\mathbf{K}$ (cite Hofmann, Schoelkopf, Smola (2008)):

$$||c||_K^2 = c^T \mathbf{K} c \leq s$$

So that the solution to this new (constrained) problem is:

$$\hat{c} = (\mathbf{K} + \lambda I_{n \times n})^{-1} \mathbf{y}$$

$\lambda$ is a Lagrange multiplier having a unique correspondance with $s$, and a regularization parameter. And for new observations arriving for the model, we have a solution which is identical to the one that we had for kernel ridge regression:

$$y^* = \mathbf{K}^* (\mathbf{K} + \lambda I_{n \times n})^{-1} y$$

Many other types of kernels could be envisaged for $\mathbf{K}$, allowing to take into account nonlinearities and the various complexities of the covariance structure. One of the most popular kernels is the Gaussian kernel, also called *squared exponential* kernel, defined for $i < j$ by:

$$K_{Gauss}(\mathbf{x}_i, \mathbf{x}_j) = \sigma^2 exp\left(-\frac{||\mathbf{x}_i - \mathbf{x}_j||_2^2}{2l^2}\right)$$

where $l$ is a characteristic length-scale controlling the distance between peaks of the covariance function and $\sigma^2$ is the marginal variance, obtained when $\mathbf{x}_i = \mathbf{x}_j$. Both $l$, $\sigma^2$ are used as the learning model's hyperparameters, along with the regularization parameter $\lambda$.

This kernel is however often judged as being too smooth for most typical optimization problems (cite Rasmussen et al.). Some other kernels that could be interesting for machine learning (cite Rasmussen et al.) belong to the Matérn class of covariance

functions. If we define $r := ||\mathbf{x}_i - \mathbf{x}_j||_2$, the most used for machine learning problems (cite Rasmussen et al.) are:

$$\mathbf{K}_{ij} = K_{3/2}(r) = \sigma^2 \left(1 + \frac{\sqrt{3}r}{l}\right) exp\left(-\frac{\sqrt{3}r}{l}\right)$$

and

$$\mathbf{K}_{ij} = K_{5/2}(r) = \sigma^2 \left(1 + \frac{\sqrt{5}r}{l} + \frac{5r^2}{3l^2}\right) exp\left(-\frac{\sqrt{5}r}{l}\right)$$

Figure 1.1 below, provides more insights on the kernels we have just defined; $K_{Gauss}$, $K_{Matérn3/2}$ and $K_{Matérn5/2}$, for $\sigma^2 = 1$ and $l = 1$.

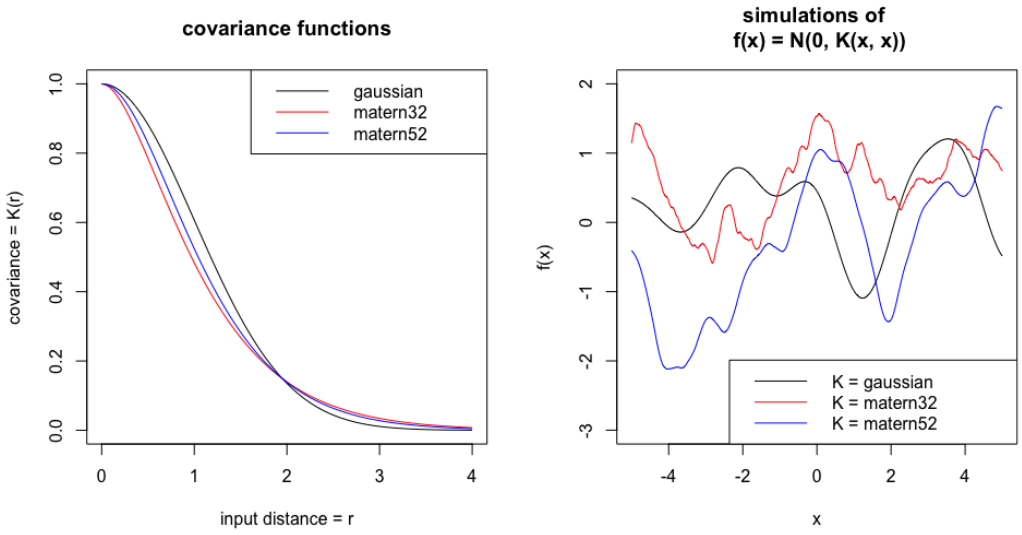


**Fig. 1.1:** **Left:** covariance functions; **Right:** random simulations from a multivariate Gaussian distribution $\mathcal{N}(\mathbf{0}, K)$ , with $\sigma^2 = 1$ and $l = 1$. The sample functions on the right were obtained using a discretization of the x-axis of 1000 equally-spaced points

We observe in figure 1.1 (**left**) that: the smoother the kernel, the higher the covariance associated to observations that are close to each other (that are similar). This relationship is inverted as the distance between the observations grows, but with a lower magnitude. The Gaussian kernel is the more flexible of the three kernels. Then, comes the kernel Matérn 5/2, and to finish, the kernel Matérn 3/2 (cf. figure 1.1, **right**).

An interesting feature of KRLS learning, is the possibility to derive estimators for the marginal effects of the covariates $\mathbf{x}^{(j)}$ on the response. For example, (and like in

Ferwerda, Hainmueller et Hazlett (2017)), since we have the relationship (1.1), we can write for a fixed $j_0 \in \{1, \dots, p\}$ and for any $k \in \{1, \dots, n\}$, :

$$\frac{\partial y_i}{\partial \mathbf{x}_k^{(j_0)}} = \sum_{j=1}^{n} c_j \frac{dK(\mathbf{x}_i, \mathbf{x}_j)}{d\mathbf{x}_k^{(j_0)}} = c_k \frac{\partial K(\mathbf{x}_i, \mathbf{x}_k)}{\partial \mathbf{x}_k^{(j_0)}} \tag{1.2}$$

That's an approximation of how much of an increase (at the first order) we obtain in $y_i$, for a slight change in $\mathbf{x}_k^{(j_0)}$. An average marginal effect of the $j_0^{th}$ covariate on the $i^{th}$ observation of the response $y$ can thus be obtained as:

$$\frac{1}{n} \sum_{k=1}^{n} \frac{\partial y_i}{\partial x_k^{(j_0)}} = \frac{1}{n} \sum_{k=1}^{n} c_k \frac{\partial K(\mathbf{x}_i, \mathbf{x}_k)}{\partial \mathbf{x}_k^{(j_0)}} \tag{1.3}$$

Two types of KRLS' model formulations will be described in sections 1.2 and 1.3. One relying on the famous Dynamic Nelson-Siegel (DNS) framework, and another one explaining the spot rates as a function of the time to maturity and date of observation.

More specifically in sections **??** and 1.3.2, we derive formulas for the sensitivities of the response variable $y_i$ to a change in the covariates $\mathbf{x}^{(j)}$, in each one of the two frameworks. We do this for Gaussian, Matérn 3/2 and Matérn 5/2 kernels. These sensitivities can then be plugged into formula 1.3, in order to obtain average marginal effects of the covariates on the response.

## 1.2 Description of the DNS-KRLS model and model's sensitivity

### 1.2.1 The DNS-KRLS model

In the DNS framework ([DL06]), the spot interest rates observed at time $t$, for time to maturity $\tau$ are modeled as:

$$R_t(\tau) = \alpha_{1,t} + \alpha_{2,t} \left( \frac{1 - e^{-\tau/\eta}}{e^{-\tau/\eta}} \right) + \alpha_{3,t} \left( \frac{1 - e^{-\tau/\eta}}{e^{-\tau/\eta}} - e^{-\tau/\eta} \right) \tag{1.4}$$

If these spot interest rates $R_t(\tau)$ are observed at increasing dates $t = t_1 < \ldots < t_n$, for increasing time to maturities $\tau = \tau_1 < \ldots < \tau_p$, the factor loadings in the DNS framework are the vectors (of length $p$):

$$(1, \ldots, 1)^T$$

and

$$\left( \frac{1 - e^{-\tau_1/\eta}}{e^{-\tau_1/\eta}}, \ldots, \frac{1 - e^{-\tau_p/\eta}}{e^{-\tau_p/\eta}} \right)^T$$

and

$$\left( \frac{1 - e^{-\tau_1/\eta}}{e^{-\tau_1/\eta}} - e^{-\tau_1/\eta}, \ldots, \frac{1 - e^{-\tau_p/\eta}}{e^{-\tau_p/\eta}} - e^{-\tau_p/\eta} \right)^T$$

These vectors are used to represent respectively the level, slope, and curvature of the Yield Curve. Estimations of $\alpha_{i,t}, i = 1, \ldots, 3$ are obtained for each cross-section of yields (that is, for each fixed date $t$) by taking a fixed $\eta$, and doing a least squares regression of the spot rates observed at time $t$ on these factor loadings.

The three time series $(\alpha_{i,t})_t, i = 1, \ldots, 3$ associated to the loadings for each cross-section of yields, are those that we wish to forecast simultaneously, by using KRLS learning. For doing this, we store the most contemporaneous values of the three time series $(\alpha_{i,t})_t, i = 1, \ldots, 3$ in a response matrix $\mathbf{Y}$, and their lags in a matrix of predictors $\mathbf{X}$.

Considering the $p \in \mathbb{N}^*$ time series $(\alpha_t^{(j)})_{t \geq 0}, j = 1, \ldots, p$ (with $p = 3$), observed at $n \in \mathbb{N}^*$ discrete dates. We are interested in obtaining simultaneous forecasts of the $p$ time series at time $n + h$, $h \in \mathbb{N}^*$, by allowing each of the $p$ variables to be influenced by the others (in the spirit of VAR models, see [**lutkepohl2005new**]). We use $k < n$ lags of each of the observed $p$ time series. Hence, the output variables (columns of $\mathbf{Y}$) to be explained are:

$$Y^{(j)} = \left( \alpha_n^{(j)}, \ldots, \alpha_{k+1}^{(j)} \right)^T \tag{1.5}$$

for $j \in \{1, \ldots, p\}$. Where $\alpha_n^{(j)}$ is the most contemporaneously observed value of the $j^{th}$ time series, and $\alpha_{k+1}^{(j)}$ was observed $k$ dates earlier in time for $(\alpha_t^{(j)})_{t \geq 0}$. These output variables are stored in:

$$\mathbf{Y} \in \mathbb{R}^{(n-k) \times p}$$

and the predictors are stored in:

$$\mathbf{X} \in \mathbb{R}^{(n-k)\times(k\times p)}$$

where $\mathbf{X}$ consists in $p$ blocks of $k$ lags, for each one of the observed $p$ time series. For example, the $j_0^{th}$ block of $\mathbf{X}$, for $j_0 \in \{1, \ldots, p\}$ contains in columns:

$$\left(\alpha_{n-i}^{(j_0)}, \ldots, \alpha_{k+1-i}^{(j_0)}\right)^T \tag{1.6}$$

with $i \in \{1, \ldots, k\}$. If we consider the $p = 3$ time series $(\alpha_{t_1}^{(1)}, \ldots, \alpha_{t_5}^{(1)})$, $(\alpha_{t_1}^{(2)}, \ldots, \alpha_{t_5}^{(2)})$ and $(\alpha_{t_1}^{(3)}, \ldots, \alpha_{t_5}^{(3)})$ observed at $n = 5$ dates $t_1 < \ldots < t_5$, with $k = 2$ lags, the response variables are stored in:

$$\mathbf{Y} = \begin{pmatrix} \alpha_{t_5}^{(1)} & \alpha_{t_5}^{(2)} & \alpha_{t_5}^{(3)} \\ \alpha_{t_4}^{(1)} & \alpha_{t_4}^{(2)} & \alpha_{t_4}^{(3)} \\ \alpha_{t_3}^{(1)} & \alpha_{t_3}^{(2)} & \alpha_{t_3}^{(3)} \end{pmatrix}$$

The predictors are stored in:

$$\mathbf{X} = \begin{pmatrix} \alpha_{t_4}^{(1)} & \alpha_{t_3}^{(1)} & \alpha_{t_4}^{(2)} & \alpha_{t_3}^{(2)} & \alpha_{t_4}^{(3)} & \alpha_{t_3}^{(3)} \\ \alpha_{t_3}^{(1)} & \alpha_{t_2}^{(1)} & \alpha_{t_3}^{(2)} & \alpha_{t_2}^{(2)} & \alpha_{t_3}^{(3)} & \alpha_{t_2}^{(3)} \\ \alpha_{t_2}^{(1)} & \alpha_{t_1}^{(1)} & \alpha_{t_2}^{(2)} & \alpha_{t_1}^{(2)} & \alpha_{t_2}^{(1)} & \alpha_{t_1}^{(1)} \end{pmatrix}$$

It is also possible to add other regressors to $\mathbf{X}$, such as dummy variables, or indicators of special events. In this situation, and as discussed in [**exterkate2016nonlinear**], we can avoid the constraining of these dummy variables, in a Kernel ridge regression with unpenalized terms. Here, we consider only the inclusion of the observed time series' lags in the model.

### 1.2.2 Sensitivity of the response to a change in the covariates

Here, the response is the matrix $\mathbf{Y}$ described in the previous section. Thus, we are deriving the sensitivity of level, slope, and curvature, to the changes in their associated lags.

We let $r^2$ be:

$$r^2 := ||\mathbf{x}_i - \mathbf{x}_k||_2^2 = (\mathbf{x}_i - \mathbf{x}_k)^T(\mathbf{x}_i - \mathbf{x}_k)$$

Where $\mathbf{x}_i$ is the $i^{-th}$ line of matrix $\mathbf{X}$. Hence:

$$\frac{\partial r^2}{\partial \mathbf{x}_k^{(j_0)}} = -2\left(\mathbf{x}_i^{(j_0)} - \mathbf{x}_k^{(j_0)}\right)$$

And:

$$\frac{\partial r}{\partial \mathbf{x}_k^{(j_0)}} = \frac{1}{2r}\frac{\partial r^2}{\partial \mathbf{x}_k^{(j_0)}} = -\frac{1}{r}\left(\mathbf{x}_i^{(j_0)} - \mathbf{x}_k^{(j_0)}\right)$$

As a consequence:

- For the **Gaussian kernel**

$$(\mathbf{x}_i, \mathbf{x}_k) \mapsto \sigma^2 exp\left(-\frac{r^2}{2l^2}\right)$$

We have:

$$\begin{aligned}
\frac{\partial K_{Gauss}(\mathbf{x}_i, \mathbf{x}_k)}{\partial \mathbf{x}_k^{(j_0)}} &= -\frac{\sigma^2}{2l^2} exp\left(-\frac{r^2}{2l^2}\right)\frac{\partial r^2}{\partial \mathbf{x}_k^{(j_0)}} \\
&= -\frac{\sigma^2}{2l^2} exp\left(-\frac{r^2}{2l^2}\right)\left[-2\left(\mathbf{x}_i^{(j_0)} - \mathbf{x}_k^{(j_0)}\right)\right] \\
&= \frac{\sigma^2}{l^2} exp\left(-\frac{r^2}{2l^2}\right)\left(\mathbf{x}_i^{(j_0)} - \mathbf{x}_k^{(j_0)}\right) \\
&= \frac{\left(\mathbf{x}_i^{(j_0)} - \mathbf{x}_k^{(j_0)}\right)}{l^2} K_{Gauss}(\mathbf{x}_i, \mathbf{x}_k)
\end{aligned}$$

- For the **Matérn 3/2 kernel**

$$(\mathbf{x}_i, \mathbf{x}_k) \mapsto \sigma^2\left(1 + \frac{\sqrt{3}r}{l}\right) exp\left(-\frac{\sqrt{3}r}{l}\right)$$

We have:

$$\begin{aligned}
\frac{\partial K_{3/2}(\mathbf{x}_i, \mathbf{x}_k)}{\partial \mathbf{x}_k^{(j_0)}} &= \sigma^2 \frac{\sqrt{3}}{l} exp\left(-\frac{\sqrt{3}}{l} r\right) \left[\frac{\partial r}{\partial \mathbf{x}_k^{(j_0)}} - \frac{\partial r}{\partial \mathbf{x}_k^{(j_0)}} \left(1 + \frac{\sqrt{3}}{l} r\right)\right] \\
&= \sigma^2 \frac{\sqrt{3}}{l} \frac{\partial r}{\partial \mathbf{x}_k^{(j_0)}} exp\left(-\frac{\sqrt{3}}{l} r\right) \left[1 - \left(1 + \frac{\sqrt{3}}{l} r\right)\right] \\
&= \sigma^2 \frac{\sqrt{3}}{l} exp\left(-\frac{\sqrt{3}}{l} r\right) \frac{\sqrt{3}\left(\mathbf{x}_i^{(j_0)} - \mathbf{x}_k^{(j_0)}\right)}{l}
\end{aligned}$$

- For the **Matérn 5/2 kernel**

$$(\mathbf{x}_i, \mathbf{x}_k) \mapsto \sigma^2 \left(1 + \frac{\sqrt{5}r}{l} + \frac{5r^2}{3l^2}\right) exp\left(-\frac{\sqrt{5}r}{l}\right)$$

  We have:

$$\begin{aligned}
\frac{\partial K_{5/2}(\mathbf{x}_i, \mathbf{x}_k)}{\partial \mathbf{x}_k^{(j_0)}} &= \sigma^2 exp\left(-\frac{\sqrt{5}}{l} r\right) \left[\left(\frac{\sqrt{5}}{l} \frac{\partial r}{\partial \mathbf{x}_k^{(j_0)}} + \frac{5}{3l^2} \frac{\partial r^2}{\partial \mathbf{x}_k^{(j_0)}}\right) - \frac{\sqrt{5}}{l} \frac{\partial r}{\partial \mathbf{x}_k^{(j_0)}} \left(1 + \frac{\sqrt{5}r}{l} + \frac{5r^2}{3l^2}\right)\right] \\
&= \sigma^2 exp\left(-\frac{\sqrt{5}}{l} r\right) \left[\frac{5}{3l^2} \frac{\partial r^2}{\partial \mathbf{x}_k^{(j_0)}} - \frac{\sqrt{5}}{l} \frac{\partial r}{\partial \mathbf{x}_k^{(j_0)}} \left(\frac{\sqrt{5}r}{l} + \frac{5r^2}{3l^2}\right)\right] \\
&= \sigma^2 exp\left(-\frac{\sqrt{5}}{l} r\right) \left(\mathbf{x}_i^{(j_0)} - \mathbf{x}_k^{(j_0)}\right) \left[-2\frac{5}{3l^2} + \frac{1}{r}\frac{\sqrt{5}}{l} \left(\frac{\sqrt{5}r}{l} + \frac{5r^2}{3l^2}\right)\right] \\
&= \sigma^2 exp\left(-\frac{\sqrt{5}}{l} r\right) \frac{5}{3l^2} \left(1 + \frac{\sqrt{5}r}{l}\right) \left(\mathbf{x}_i^{(j_0)} - \mathbf{x}_k^{(j_0)}\right)
\end{aligned}$$

From these expressions of the sensitivities, we can derive the average marginal effect of a covariate on the response, as demonstrated in equation 1.3. These formulas will be valid in the DNS framework described in section 1.2, where only one length-scale $l$ parameter is required. Similar types of formulas could be derived in the KRLS framework from section , but they would include two length-scale parameters.

## 1.3 Description of the KRLS model applied to observed dates and time to maturities

### 1.3.1 Description of the model

In this other framework, we consider that the response variable is the spot interest rate observed at time $t$, for time to maturity $\tau$, $R(t,\tau)$. The predictors are the observation date, and the time to maturity.

In this setting, we use the following weighted distance between the vectors $(t,\tau)$ in the Gaussian, Matérn 3/2 and Matérn 5/2 kernels:

$$r = \sqrt{\frac{(t_i - t_j)^2}{l_1^2} + \frac{(\tau_i - \tau_j)^2}{l_2^2}}$$

So that here, the spot rates values are explained as linear combination of distances between vectors of time to maturities and observation dates $(t_i,\tau_i)$ and $(t_j,\tau_j)$. In this setting, if we consider $10$ spot rates observed at $2$ dates $t_1 < t_2$ and $5$ time to maturities $\tau_1,\ldots,\tau_5$, the response variable is:

$$\mathbf{Y} = (R(t_1,\tau_1),\ldots,R(t_1,\tau_5),R(t_2,\tau_1),\ldots,R(t_2,\tau_5))^T$$

and the predictors are

$$\mathbf{X} = \begin{pmatrix} \tau_1 & t_1 \\ \vdots & \vdots \\ \tau_5 & t_1 \\ \tau_1 & t_2 \\ \vdots & \vdots \\ \tau_5 & t_2 \end{pmatrix}$$

If some new observations arrive at time $t_3$ in the model, these new observations will be stored in:

$$\mathbf{X}^* = \begin{pmatrix} \tau_1 & t_3 \\ \vdots & \vdots \\ \tau_5 & t_3 \end{pmatrix}$$

In this other setting, it is also possible to add other regressors such as dummy variables, or indicators of special events. Again, and as suggested in the previous section and in [**exterkate2016nonlinear**], we can avoid the constraining of these dummy variables, in a Kernel ridge regression with unpenalized terms.

For example, if we wanted to add another indicator $(I_t)_t$ observed at times $t_1 < t_2$, we would have to consider the following matrix of predictors:

$$\mathbf{X} = \begin{pmatrix} \tau_1 & t_1 & I_{t_1} \\ \vdots & \vdots & \vdots \\ \tau_5 & t_1 & I_{t_1} \\ \tau_1 & t_2 & I_{t_2} \\ \vdots & \vdots & \vdots \\ \tau_5 & t_2 & I_{t_2} \end{pmatrix}$$

And another weighted distance in the Gaussian, Matérn 3/2 and Matérn 5/2 kernels, taking into account the new indicator $(I_t)_t$:

$$r = \sqrt{\frac{(t_i - t_j)^2}{l_1^2} + \frac{(\tau_i - \tau_j)^2}{l_2^2} + \frac{(I_{t_i} - I_{t_j})^2}{l_3^2}}$$

### 1.3.2 Sensitivity of the spot rates to a change in observation date and time to maturity

In this framework, and as mentioned in the previous section, we consider the following measure of similarity/dissimilarity between $\mathbf{x}_i = (t_i, \tau_i)$ and $\mathbf{x}_k = (t_k, \tau_k)$:

$$r^2 = \frac{(t_i - t_k)^2}{l_1^2} + \frac{(\tau_i - \tau_k)^2}{l_2^2}$$

The associated kernels are the following ones:

$$(\mathbf{x}_i, \mathbf{x}_k) \mapsto K_{Gauss}(r) = \sigma^2 exp\left(-\frac{r^2}{2}\right)$$

$$(\mathbf{x}_i, \mathbf{x}_k) \mapsto K_{Matérn3/2}(r) = \sigma^2 \left(1 + \sqrt{3}r\right) exp\left(-\sqrt{3}r\right)$$

and

$$(\mathbf{x}_i, \mathbf{x}_k) \mapsto K_{Matérn5/2}(r) = \sigma^2 \left(1 + \sqrt{5}r + \frac{5}{3}r^2\right) exp\left(-\sqrt{5}r\right)$$

It is possible to obtain the sensitivity of the spot rates to a change in the observation date, $\frac{\partial R(t_k, \tau)}{\partial t_k}$ for any $\tau$. Even if it is actually very difficult to predict in which direction

the spot rates will move, this type of indicators could still serve as decision-assistance tools for risk management. We have:

$$\frac{\partial r^2}{\partial t_k} = -\frac{2(t_i - t_k)}{l_1^2}$$

and

$$\frac{\partial r}{\partial t_k} = -\frac{1}{r}\frac{(t_i - t_k)}{l_1^2}$$

Thus, we have:

$$\begin{aligned}
\frac{\partial K_{Gauss}(r)}{\partial t_k} &= -\frac{\sigma^2}{2} exp\left(-\frac{r^2}{2}\right)\frac{\partial r^2}{\partial t_k} \\
&= \sigma^2 exp\left(-\frac{r^2}{2}\right)\frac{(t_i - t_k)}{l_1^2}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial Matérn3/2}{\partial t_k} &= \sigma^2\sqrt{3}\frac{\partial r}{\partial t_k} exp\left(-\sqrt{3}r\right)\left[1 - (1 + \sqrt{3}r)\right] \\
&= -3\sigma^2 r exp\left(-\sqrt{3}r\right)\frac{\partial r}{\partial t_k} \\
&= 3\sigma^2 exp\left(-\sqrt{3}r\right)\frac{(t_i - t_k)}{l_1^2}
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial Matérn5/2}{\partial t_k} &= \sigma^2 exp\left(-\sqrt{5}r\right)\left[\left(\sqrt{5}\frac{\partial r}{\partial t_k} + \frac{5}{3}\frac{\partial r^2}{\partial t_k}\right) - \sqrt{5}\frac{\partial r}{\partial t_k}\left(1 + \sqrt{5}r + \frac{5}{3}r^2\right)\right] \\
&= \sigma^2 exp\left(-\sqrt{5}r\right)\left[\frac{5}{3}\frac{\partial r^2}{\partial t_k} - \sqrt{5}\frac{\partial r}{\partial t_k}\left(\sqrt{5}r + \frac{5}{3}r^2\right)\right] \\
&= \sigma^2 exp\left(-\sqrt{5}r\right)\left[-\frac{10}{3}\frac{(t_i - t_k)}{l_1^2} + \frac{\sqrt{5}}{r}\frac{(t_i - t_k)}{l_1^2}\left(\sqrt{5}r + \frac{5}{3}r^2\right)\right] \\
&= \sigma^2 exp\left(-\sqrt{5}r\right)\frac{5}{3}\left[1 + \sqrt{5}r\right]\frac{(t_i - t_k)}{l_1^2}
\end{aligned}$$

In the next section, 1.4, we present the results obtained by the DNS-KRLS model from section 1.2 and the KRLS model from section 1.3 on a training/testing dataset[1].

## 1.4 Numerical examples

In this section, we present the results obtained by the DNS-KRLS model from section 1.2 and the KRLS model from section 1.3. The examples are not exhaustive benchmarks, but aim at illustrating the forecasting capabilities of the models.

We use calibrated discount rates data from Deutsche Bundesbank website, observed on a monthly basis, from the beginning of 2002 to the end 2015. There are 167 curves, observed at 50 maturities in the dataset. Only $15$ time to maturities are used in these examples (in years): $1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 15, 20, 25, 30$.

In figure 1.2, we present the data that we use, and table 1.1 contains a summary of these data; the minimum, maximum, median, first and third quartiles of the discount rates observed at given maturities. There are alternate cycles of increases and decreases of the discount rates, with generally a decreasing trend. Some of the discount rates, at the most recent dates, and lower maturities, are negative.

**Tab. 1.1:** Summary of observed discount rates from Deutsche Bundesbank website, from 2002 to the end 2015

| Maturity | Min | 1st Qrt | Median | 3rd Qrt | Max |
|---|---|---|---|---|---|
| 1 | -0.116 | 0.858 | 2.045 | 3.072 | 5.356 |
| 5 | 0.170 | 1.327 | 2.863 | 3.807 | 5.146 |
| 15 | 0.711 | 2.616 | 3.954 | 4.702 | 5.758 |
| 30 | 0.805 | 2.594 | 3.962 | 4.814 | 5.784 |

A rolling forecasting methodology (see [**bergmeir2015note**]) is implemented in order to obtain the benchmarks between the models. It is described in figure 1.3. A fixed 12 months/36 months-length window for training the model, and the following 12 months/36 months for testing, the origin of the training set is then advanced of 1 month, and the training/testing procedure is repeated. The measure of forecasting performance is the Root Mean Squared Error ($RMSE$).

We use similar grids for all the models, in order to ease the comparability of the results, and avoid too much manual tweaking of the hyperparameters and overtraining the available data.

[1] for a more complete treatment, a validation set would be added, in order to verify that the models are not 'overtrained' on this training/testing set

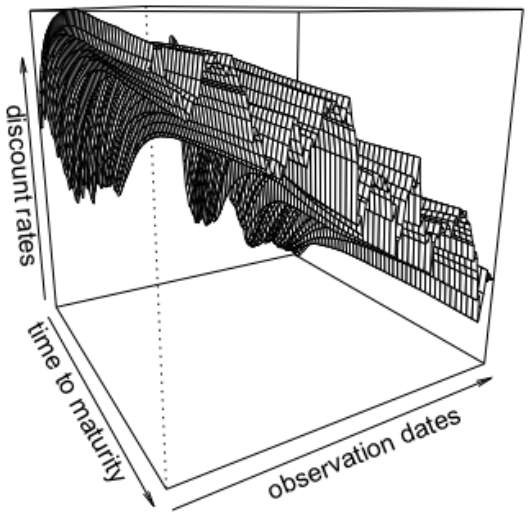


**Fig. 1.2:** Observed discount rates from Deutsche Bundesbank website, from 2002 to the end 2015

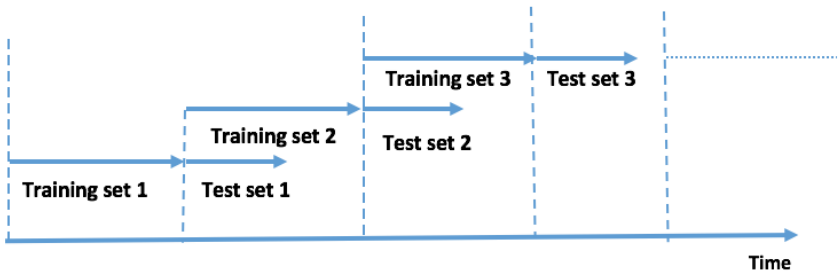


**Fig. 1.3:** rolling forecasting cross-validation sets

Hence for both models, DNS-KRLS (from section 1.2) and KRLS (from section 1.3), we consider $5$ values of $\sigma$ (variance parameter), $l$, $l_1$, $l_2$ (length-scale parameters for Gaussian, Matérn 3/2 and Matérn 5/2 kernels) and $\lambda$ (the regularization parameter for all the kernels) regularly spaced between $[10^{-2}, 10^{2}]$: $0.01, 0.1, 1, 10, 100$.

For the additional parameter $\eta$ in the DNS-KRLS model, we use 5 values comprised (regularly spaced) between the minimum of the observed time to maturities and the maximum of the observed time to maturities (on $[1, 30]$): $1, 8.25, 15.5, 22.75, 30$.

### 1.4.1 Cross-validation results

The results obtained after the cross-validation procedure are reported in table 1.2 and 1.3. The results obtained by considering an automatic ARIMA modeling (cite Hyndman et al.) of the three time series are also indicated, to serve as a benchmark.

**Tab. 1.2:** Average out-of-sample RMSE for training set length = 12 months and test set length = 12 months

| Model | $\sigma$ | $l$ | $l_1$ | $l_2$ | $\lambda$ | $\eta$ | **RMSE** |
|---|---|---|---|---|---|---|---|
| Gaussian | 10 | - | 0.01 | 10 | 10 | - | 0.5839150 |
| Matérn 3/2 | 100 | - | 1 | 100 | 0.01 | - | **0.5136373** |
| Matérn 5/2 | 1 | - | 0.01 | 10 | 0.1 | - | 0.5781184 |
| DNS-Gaussian | 0.1 | 10 | - | - | 0.01 | 15.5 | 0.6041652 |
| DNS-Matérn 3/2 | 100 | 1 | - | - | 0.01 | 22.75 | **0.6038667** |
| DNS-Matérn 5/2 | 100 | 1 | - | - | 0.01 | 15.5 | 0.6041580 |
| DNS-ARIMA | - | - | - | - | - | 1 | 0.6751660 |

**Tab. 1.3:** Average out-of-sample RMSE for training set length = 36 months and test set length = 36 months

| Model | $\sigma$ | $l$ | $l_1$ | $l_2$ | $\lambda$ | $\eta$ | **RMSE** |
|---|---|---|---|---|---|---|---|
| Gaussian | 10 | - | 0.1 | 10 | 100 | - | 1.170690 |
| Matérn 3/2 | 100 | - | 1 | 10 | 0.01 | - | **1.003246** |
| Matérn 5/2 | 10 | - | 10 | 100 | 10 | - | 1.134833 |
| DNS-Gaussian | 0.01 | 0.01 | - | - | 0.01 | 30 | 1.264533 |
| DNS-Matérn 3/2 | 0.01 | 0.01 | - | - | 0.01 | 30 | **1.264533** |
| DNS-Matérn 5/2 | 0.01 | 0.01 | - | - | 0.01 | 30 | 1.264533 |
| DNS-ARIMA | - | - | - | - | - | 1 | 1.281937 |

For the DNS model and all the types of kernels (Gaussian, Matérn 3/2 and Matérn 5/2), the *optimal* number of lags is respectively equal to $5$ and $6$ for the $12$-months and $36$-months horizons. The last 12 months and 36 months of the data are respectively considered as training sets for obtaining these graphs.

We observe that no matter the length of the training/testing window (either 12 months or 36 months), or the method employed (either DNS-KRLS or KRLS), the Matérn 3/2 kernel performs better than the other models. It is even performing better on this specific problem with a KRLS model, considering the similarities between vectors of time to maturities and observation dates. The other kernels are probably too flexible for the purpose, so that they are both overfitting the data a bit.

In the next sections 1.4.2 and 1.4.3, we examine these results further, by looking at the out-of-sample root mean squared error (RMSE) obtained over time, and the projected discount/discount factors'/discrete forward curves obtained with the optimal parameters.

### 1.4.2 Out-of-sample RMSE over time

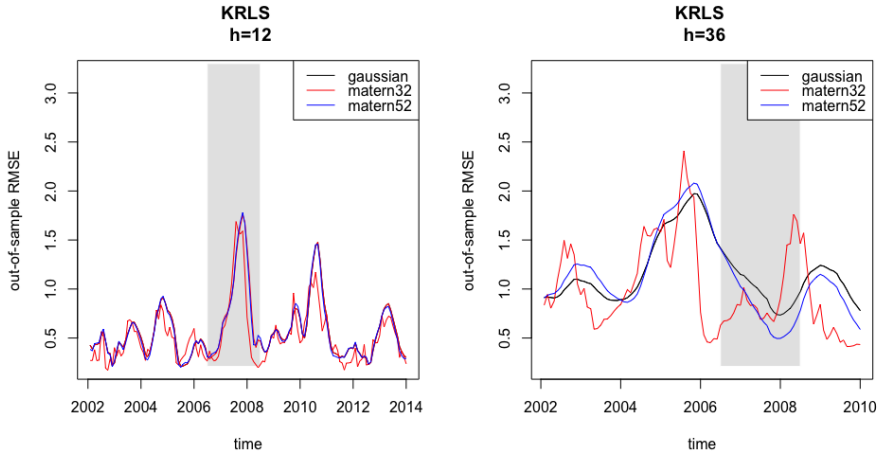


**Fig. 1.4:** Out-of-sample RMSE over time for KRLS models, with horizon = 12 and horizon = 36

More details on these boxplots are given in appendix 1.6.2.

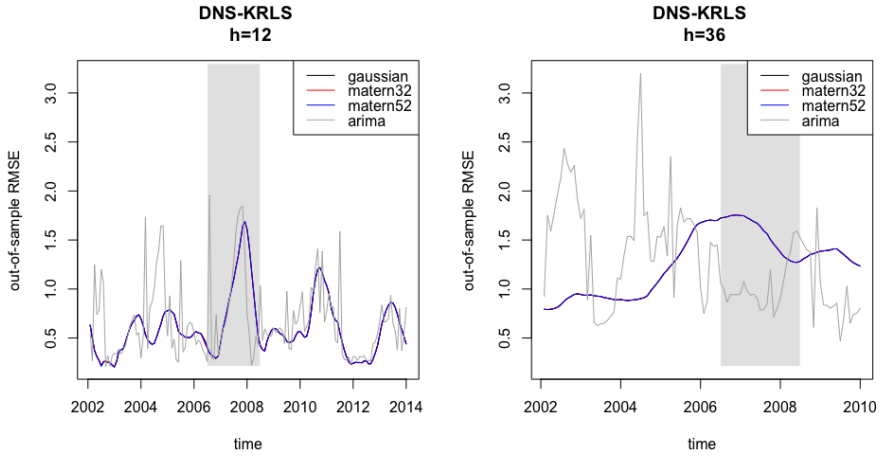


**Fig. 1.5:** Out-of-sample RMSE over time for DNS-KRLS, with horizon = 12 and horizon = 36

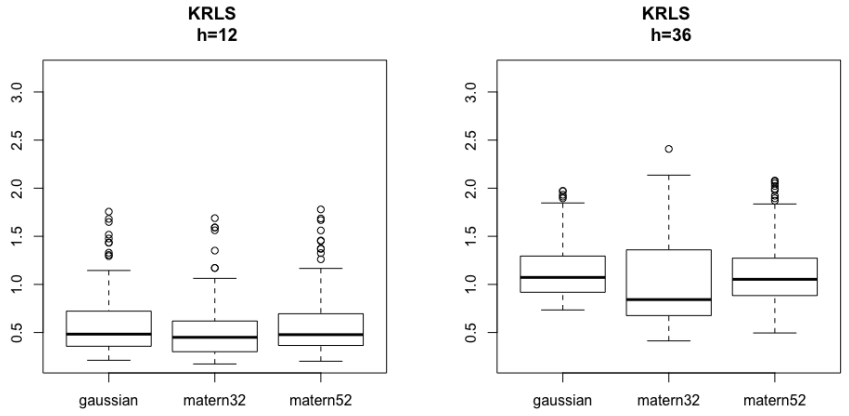


**Fig. 1.6:** Boxplots of Out-of-sample RMSE over time, for KRLS models

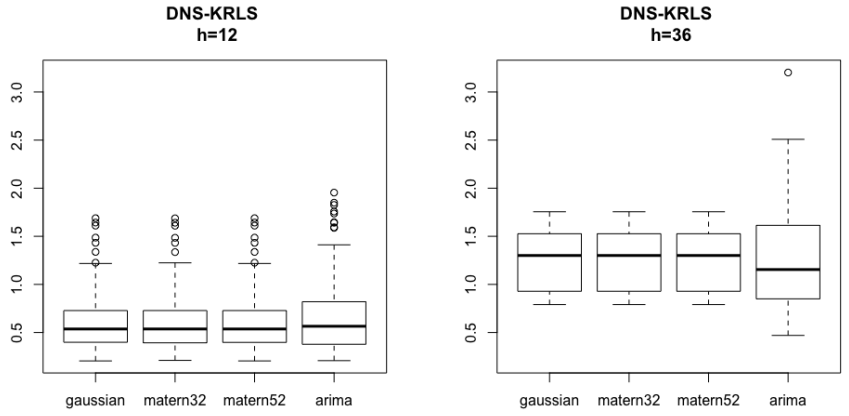


**Fig. 1.7:** Boxplots of Out-of-sample RMSE over time, for DNS-KRLS models

### 1.4.3 Implied forecast term-structure of discrete forward rates

The following figure, 1.8, presents the 12-months ahead and 36-months ahead ($h = 12$ and $h = 36$) forecasts obtained in the KRLS framework, by considering a Matérn 3/2 kernel. Similarly, figure 1.9, presents the 12-months ahead and 36-months ahead ($h = 12$ and $h = 36$) forecasts obtained in the DNS-KRLS framework, by considering a Matérn 3/2 kernel for forecasting the factors.

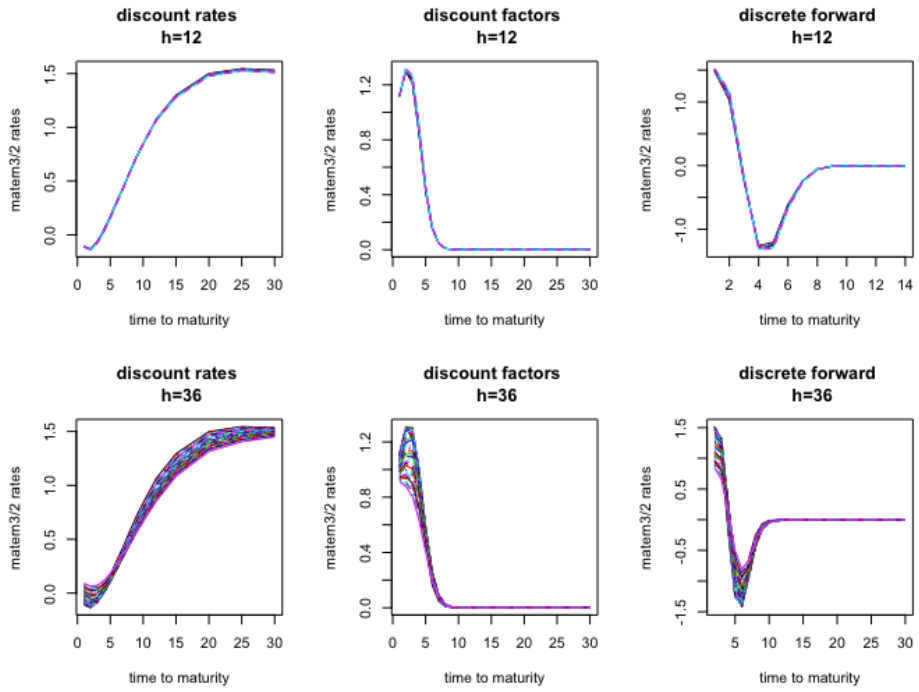


**Fig. 1.8:** Forecasts of discount rates, discount factors and discrete forward rates for Matérn 3/2 for horizon = 12 and horizon = 36

These (figures 1.8 and 1.9) can be compared to a more familiar model, the one in which the level, slope, and curvature are modelled separately with an ARIMA model in a DNS framework.

We observe that the discount (rates and factors) and forward curves obtained by each model do, indeed, exhibit the same patterns as actual market discount and discrete forward curves. In particular, for the projected negative rates, we observe projected discount factors that are greater than 1.

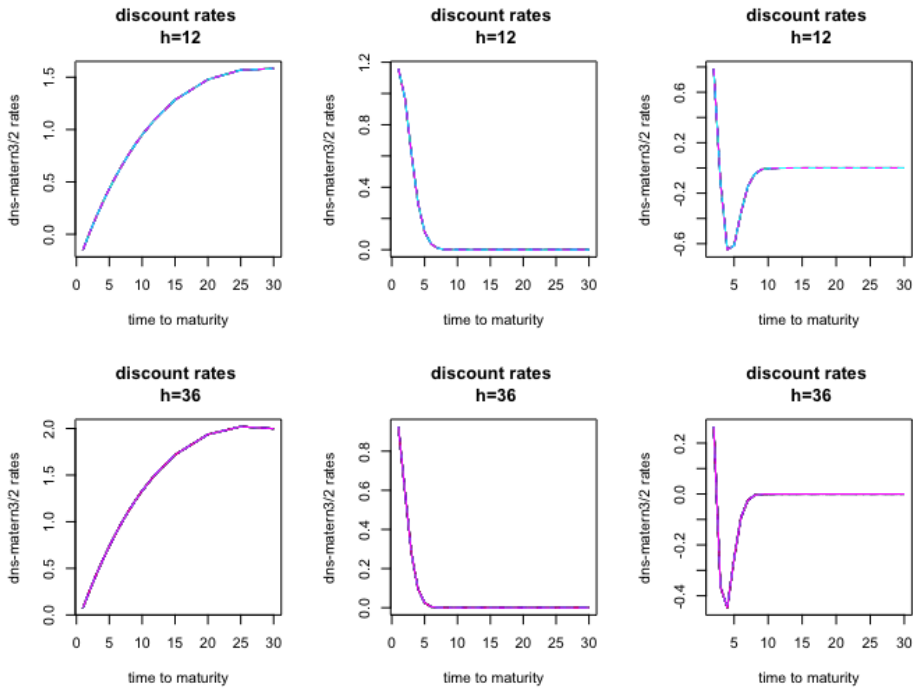


**Fig. 1.9:** Forecasts of discount rates, discount factors and discrete forward rates for DNS-Matérn 3/2 for horizon = 12 and horizon = 36

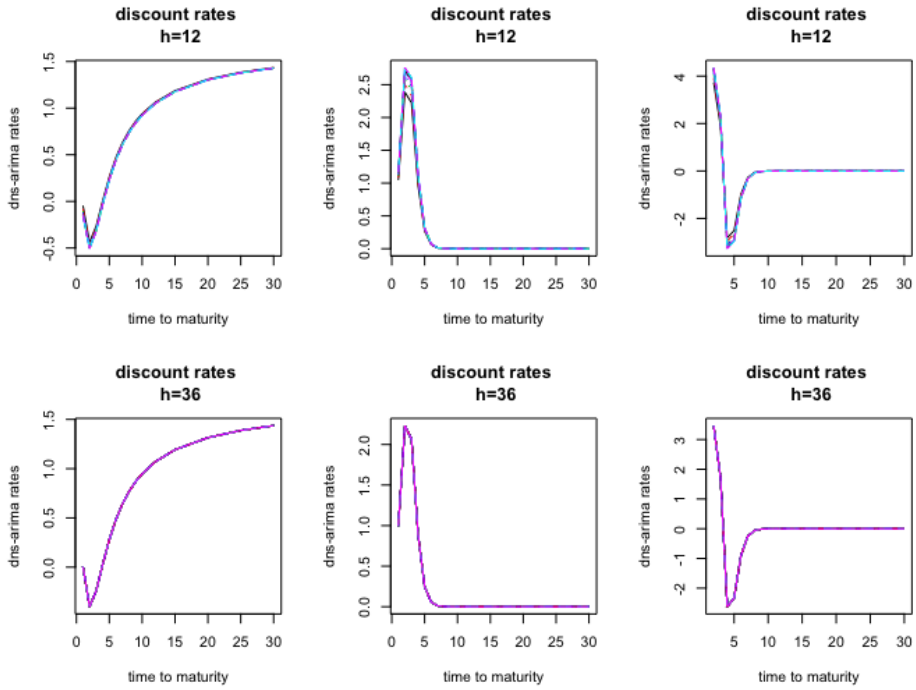


**Fig. 1.10:** discount rates, discount factors and discrete forward rates for DNS-ARIMA

## 1.5 Conclusion

In this chapter, we discussed the forecasting of discount curves, by using Kernel Regularized Least Squares (KRLS). The KRLS techniques are capable of learning nonlinear response variables, by taking into account various and complex types of covariance structures between the predictors. The response must have a relatively low number of examples, though, because of the quadratic cost of computing the kernels.

The model is highly interpretable, as it explains the responses as linear combinations of similarities/dissimilarities between the examples. Some sentivity indicators of the response as a function of the predictors are derived, and they could consitute useful decision-assistance tools.

Two types of KRLS models are considered here, specifically for the discount curves. One relying on the famous Dynamic Nelson-Siegel (DNS) framework, and another one explaining the spot rates as a function of the time to maturity and date of observation. Both types of KRLS models deliver some robust forecasts of the discount curves, as the kernels hyperparameters implicitly constrain the model to reproduce the Yield Curve's stylized facts.

## 1.6 Appendix

### 1.6.1 Summary of out-of-sample errors for all the models (in %)

**KRLS, h = 12** (cf. figure 1.6)

```
gaussian          matern32          matern52
Min.   :0.2114   Min.   :0.1726   Min.   :0.2012
1st Qu.:0.3582   1st Qu.:0.3009   1st Qu.:0.3669
Median :0.4828   Median :0.4498   Median :0.4777
Mean   :0.5839   Mean   :0.5136   Mean   :0.5781
3rd Qu.:0.7216   3rd Qu.:0.6158   3rd Qu.:0.6914
Max.   :1.7562   Max.   :1.6886   Max.   :1.7795
```

**KRLS, h = 36** (cf. figure 1.6)

```
  gaussian          matern32          matern52
Min.   :0.7343   Min.   :0.4130   Min.   :0.4953
1st Qu.:0.9193   1st Qu.:0.6775   1st Qu.:0.8860
Median :1.0727   Median :0.8422   Median :1.0529
Mean   :1.1707   Mean   :1.0032   Mean   :1.1348
3rd Qu.:1.2824   3rd Qu.:1.3548   3rd Qu.:1.2639
Max.   :1.9723   Max.   :2.4080   Max.   :2.0797
```

**DNS-KRLS, h = 12** (cf. figure 1.7)

```
gaussian          matern32          matern52           arima
Min.   :0.2048   Min.   :0.2103   Min.   :0.2048   Min.   :0.2077
1st Qu.:0.4016   1st Qu.:0.3934   1st Qu.:0.3997   1st Qu.:0.3779
Median :0.5366   Median :0.5368   Median :0.5366   Median :0.5649
Mean   :0.6042   Mean   :0.6039   Mean   :0.6042   Mean   :0.6752
3rd Qu.:0.7281   3rd Qu.:0.7277   3rd Qu.:0.7281   3rd Qu.:0.8190
Max.   :1.6884   Max.   :1.6870   Max.   :1.6884   Max.   :1.9544
```

**DNS-KRLS, h = 36** (cf. figure 1.7)

```
gaussian          matern32          matern52           arima
```

```
 Min.   :0.7900   Min.   :0.7900   Min.   :0.7900   Min.   :0.4698
 1st Qu.:0.9294   1st Qu.:0.9294   1st Qu.:0.9294   1st Qu.:0.8527
 Median :1.3005   Median :1.3005   Median :1.3005   Median :1.1545
 Mean   :1.2645   Mean   :1.2645   Mean   :1.2645   Mean   :1.2819
 3rd Qu.:1.5158   3rd Qu.:1.5158   3rd Qu.:1.5158   3rd Qu.:1.6033
 Max.   :1.7549   Max.   :1.7549   Max.   :1.7549   Max.   :3.2024
```

### 1.6.2 Summary of KRLS Matérn 3/2 and DNS-ARIMA forecasts (in %) for horizon = 12 and horizon = 36

**KRLS Matérn 3/2**

**horizon = 12** (cf. figure 1.8)

```
1y                 5y                10y               30y
 Min.   :-0.1135   Min.   :0.1538   Min.   :0.8424   Min.   :1.509
 1st Qu.:-0.1105   1st Qu.:0.1588   1st Qu.:0.8441   1st Qu.:1.514
 Median :-0.1091   Median :0.1636   Median :0.8456   Median :1.520
 Mean   :-0.1097   Mean   :0.1632   Mean   :0.8453   Mean   :1.520
 3rd Qu.:-0.1084   3rd Qu.:0.1680   3rd Qu.:0.8468   3rd Qu.:1.526
 Max.   :-0.1078   Max.   :0.1708   Max.   :0.8471   Max.   :1.533
```

**horizon = 36** (cf. figure 1.8)

```
1y                  5y                10y               30y
 Min.   :-0.11114   Min.   :0.1066   Min.   :0.6661   Min.   :1.449
 1st Qu.:-0.08006   1st Qu.:0.1120   1st Qu.:0.6823   1st Qu.:1.472
 Median :-0.02946   Median :0.1273   Median :0.7121   Median :1.495
 Mean   :-0.02263   Mean   :0.1324   Mean   :0.7262   Mean   :1.494
 3rd Qu.: 0.03020   3rd Qu.:0.1490   3rd Qu.:0.7619   3rd Qu.:1.517
 Max.   : 0.09165   Max.   :0.1804   Max.   :0.8374   Max.   :1.535
```

**DNS-ARIMA**

**horizon = 12** (cf. figure 1.10)

```
1y                  5y                10y               30y
 Min.   :-0.15142   Min.   :0.2295   Min.   :0.9237   Min.   :1.429
 1st Qu.:-0.14819   1st Qu.:0.2305   1st Qu.:0.9242   1st Qu.:1.429
```

```
Median :-0.13914   Median :0.2333   Median :0.9256   Median :1.429
Mean   :-0.12468   Mean   :0.2379   Mean   :0.9279   Mean   :1.430
3rd Qu.:-0.11405   3rd Qu.:0.2412   3rd Qu.:0.9296   3rd Qu.:1.431
Max.   :-0.04527   Max.   :0.2628   Max.   :0.9405   Max.   :1.434
```

**horizon = 36** (cf. figure 1.10)

```
1y                 5y                10y               30y
Min.   :0.003647   Min.   :0.2782   Min.   :0.9482   Min.   :1.437
1st Qu.:0.003647   1st Qu.:0.2782   1st Qu.:0.9482   1st Qu.:1.437
Median :0.003647   Median :0.2782   Median :0.9482   Median :1.437
Mean   :0.003647   Mean   :0.2782   Mean   :0.9482   Mean   :1.437
3rd Qu.:0.003647   3rd Qu.:0.2782   3rd Qu.:0.9482   3rd Qu.:1.437
Max.   :0.003647   Max.   :0.2782   Max.   :0.9482   Max.   :1.437
```

# Bibliography

[DL06] Francis X Diebold and Canlin Li. „Forecasting the term structure of government bond yields“. In: *Journal of econometrics* 130.2 (2006), pp. 337–364 (cit. on pp. 1, 6).

[NS87] Charles R Nelson and Andrew F Siegel. „Parsimonious modeling of yield curves“. In: *Journal of business* (1987), pp. 473–489 (cit. on p. 1).

# List of Figures

# List of Tables

# Colophon

This thesis was typeset with LaTeX $2_\varepsilon$. It uses the *Clean Thesis* style developed by Ricardo Langner. The design of the *Clean Thesis* style is inspired by user guide documents from Apple Inc.

Download the *Clean Thesis* style at `http://cleanthesis.der-ric.de/`.

# Declaration

You can put your declaration here, to declare that you have completed your work solely and only with the help of the references you mentioned.

*City, August 26, 2015*

Ricardo Langner