Using extracted forward rate term structure information to forecast foreign exchange rates

Fearghal Kearney[*] Mark Cummins[†] Finbarr Murphy[‡]

Abstract

The difficulty of beating the random walk in forecasting spot foreign exchange rates is well documented. In this paper, we propose a functional principal component-based scalar response model which we benchmark versus leading VECM frameworks. Our approach leads to near systematic outperformance in terms of a comparison of performance measures, and to multiple instances of statistically significant improvements in forecast accuracy. Overall, our results provide evidence that the forward rate term structure contains substantial information about the evolution of the spot exchange rate. Finally, a stylised trading strategy is employed to demonstrate the potential economic benefits of our approach.

Keywords: foreign exchange; forward rate term structure modelling; functional data analysis; multiple hypothesis testing.

JEL classification: C4; F31; G12; G17.

forthcoming in the Journal of Empirical Finance

[*]Queen's Management School, Queen's University Belfast, BT9 5EE, UK. E-mail: F.Kearney@qub.ac.uk.
[†]Dublin City University Business School, Dublin City University, Dublin 9, Ireland. E-mail: Mark.Cummins@dcu.ie.
[‡]Kemmy Business School, University of Limerick, Limerick, Ireland. E-mail: Finbarr.Murphy@ul.ie.

1 Introduction

Foreign exchange is the largest asset class in the world with the Bank for International Settlements reporting that trading levels in foreign exchange markets average $5.1 trillion per day.[1] Many stakeholders are exposed to foreign exchange risk including banks, speculators, traders, multinational firms, importers, and exporters. Modelling foreign currency cash flows, making investment decisions, and designing hedging strategies, are all greatly dependent on expectations of future foreign exchange rate movements. For this reason it is very important to forecast exchange rates in a practical out-of-sample setting. One approach is to try to glean information from forward foreign exchange contracts. However, despite research to date concluding that the forward rate is not the optimal predictor of future spot rates, evidence has been mixed as to whether or not information is embedded in the term structure of these forward foreign exchange rates.[2] We propose tackling this problem using a new functional data approach to exploit all underlying information contained in the term structure. More specifically, we ask three main questions that contribute to our understanding of the informational content of forward exchange rates: i) Does the forward rate term structure contain information about the evolution of spot exchange rates?; ii) How does a functional principal component-based forecasting model perform at extracting this information?; and iii) Can our approach inform economically profitable trading strategies?

Meese and Rogoff (1983a,b) ascertain that standard exchange rate models do not have the ability to beat forecasts implied by the random walk in the short run. In an attempt to explain this, Engel and West (2005) and Engel et al. (2008) demonstrate that such models imply a near random walk process for the exchange rate, so their power to beat the random walk in out-of-sample forecasts is low. Furthermore, as mentioned above, it has been demonstrated that the forward rate is not the optimal predictor of future spot rates (see for example, Sarno 2005). Despite this, the question as to whether or not there is information embedded in forward foreign exchange rates persists. Clarida and Taylor (1997) seek to answer this by moving beyond single-equation methods and conclude that forward premia information is in fact considerable. Their restricted vector error correction model (VECM) constitutes the leading challenger to the seminal work of Meese and Rogoff (1983a,b). The approach is applied in an out-of-sample framework resulting in 40% lower forecasting errors than those implied by the naive random walk benchmark. These results are confirmed by Clarida et al. (2003) and Sager and Taylor (2014), who establish statistically significant outperformance using more formal Diebold-Mariano (1995) tests. Furthermore, Clarida et al. (2003) propose an extension to

[1] https://www.bis.org/publ/rpfx16fx.pdf

[2] Previous literature that establishes that the forward exchange rate is not the optimal predictor of future spot rates includes Hansen and Hodrick (1980), Frankel (1980), Bilson (1981), Frankel and Rose (1995) and Taylor (1995)

incorporate a nonlinear VECM, citing that it more closely aligns with the nonlinearities empirically identified in foreign exchange data. They also empirically demonstrate its additional forecasting merit over the linear approach.

In this paper, we add to the existing literature by seeking to extract the informational content of forward foreign exchange rates through the novel proposal of a functional data analysis-based forecasting model. In parallel, our framework further tests if exchange rates are in fact predictable and if the simple risk neutral efficient market hypothesis holds. In relation to extracting the informational content of the foreign exchange forward curve, we contribute by moving the problem to a functional domain to improve on the forecasting performance achieved by the leading benchmark models. To this aim, we adopt the scalar response model proposed by Horvath and Kokoszka (2012). For comparative purposes with previous studies, we initially present a direct comparison of forecasting performance measures. However, we also apply formal tests to identify instances of statistically significant outperformance for the scalar response model over linear and nonlinear VECM formulations, as well as random walk benchmarks. In doing so, we uniquely apply the Romano and Wolf (2010) resampling based procedure to directly address the multiple comparison problem that arises from simultaneously testing a number of hypotheses concurrently. Our use of this multiple hypothesis testing technique also ensures we set a very high statistical hurdle towards adjudicating on the effectiveness of the novel functional model.

This paper contributes by showcasing the merits of functional data analysis in financial economics, a discipline that has not seen widespread use of such techniques. In our exchange rate setting, we use functional data analysis to establish the complex dependency relation between the forward rate term structure and future spot exchange rates. The flexible functional data approach we propose exhibits the capacity to accurately capture the forward rate term structure process, whilst mitigating the need to impose restrictive data structure assumptions on the exchange rate system. . This leads to statistically significant forecasting improvements over the leading VECM benchmarks. We also find that utilising the functional model to inform a stylised trading strategy results in economically profitable exchange rate portfolios. Furthermore, our results provide robust evidence of improved forecasting performance relative to various random walk specifications, indicating that the forward rate term structure contains statistically significant information about the evolution of the spot exchange rate, above what is embedded in the historic spot rate series. Finally, our results provide additional evidence supporting the rejection of the simple risk neutral efficient market hypothesis.

The remainder of the paper is organised as follows. Section 2 introduces the theoretical basis for our work, and a description of prior approaches from Clarida and Taylor (1997) and Clarida et al. (2003). Section 3 presents the functional scalar response model we propose. Section 4 introduces the forecast evaluation framework used in the analysis, in addition to the multiple hypothesis testing framework of Romano and Wolf (2010). Section 5 presents and discusses the empirical results, with Section 6 concluding the paper and drawing implications for future studies.

2 Theoretical background

This section provides a description of the theoretical context for our work and presents the technical details of the VECM benchmarks use in our comparative analysis.

2.1 Risk neutral efficient market hypothesis

A major strength of both our proposed functional forecasting model and the Clarida and Taylor (1997) framework, is that they work in spite of the failure of the simple risk neutral efficient market hypothesis (RNEMH) and are agnostic to the precise cause of rejection. RNEMH is predicated on both risk-neutrality and rational expectations, and postulates that the k-period forward rate at time t, f_t^k, is equal to the expectation of the spot rate at time $t+k$, s_{t+k}. This is conditional on information available at time t, Ω_t:

$$0 \equiv f_t^k - E\left(s_{t+k} \mid \Omega_t\right).$$

In other words, it hypothesizes that the forward rate is the optimal predictor of the future spot rate. The RNEMH is derived from the combination of two hypotheses, namely, covered and uncovered interest parity (CIP and UIP, respectively). CIP states that the k-period eurodeposit interest rate differential between the domestic, denoted r^k, and foreign, denoted $r^{k'}$, countries is equal to the spot-forward premium, $f_t^k - s_t$:

$$0 \equiv r_t^k - r_t^{k'} - \left(f_t^k - s_t\right).$$

UIP is a related no-arbitrage condition that is satisfied without the use of a forward contract. It deems that the interest rate differential is equal to the expected forward rate:

$$0 \equiv r_t^k - r_t^{k'} - E\left(s_{t+k} - s_t \mid \Omega_t\right).$$

Empirically, it has been shown that CIP holds (Taylor, 1987, 1989) whereas Chaboud and Wright (2005) show that UIP is rejected at horizons above a few hours, yet Chinn and Merdith (2004) and Lothian (2016) find that UIP cannot be rejected at horizons above five years. Therefore, given the average investor's time horizon, it can be taken that UIP does not hold empirically.[3] It follows that the simple RNEMH has been decisively rejected (Sarno 2005 and Engel 2014). Various phenomena have been proposed to explain the rejection, including the presence of risk premia (Alvarez et al. 2009 and Lustig et al. 2011), consumption externalities (Moore and Roche 2010), institutional investor flows (Froot and Ramadorai 2005), monetary volatility (Moore and Roche 2012), microstructure effects (Dunne et al. 2010), rational bubbles (Lewis 1989), and the well documented peso problem (Rogoff 1979, Evans and Lewis 1995, and Burnside et al. 2011). The success of our proposed functional model is contingent on the existence of empirical departures from the RNEMH, therefore it serves as an indirect test for its failure.

2.2 Clarida and Taylor (1997) VECM

To date, the restricted vector error correction model (VECM) of Clarida and Taylor (1997) is the leading challenger to the seminal work of Meese and Rogoff (1983a,b). For this reason, we adopt the VECM as a comparative benchmark model, alongside the traditionally used random walk. Clarida and Taylor (1997) move beyond single-equation methods and conclude that the information contained in the forward premiums is in fact considerable. The approach is applied in a dynamic recursive out-of-sample forecasting framework that results in root mean squared error and mean absolute error metrics that are up to 50% lower than those implied by the random walk. The framework is also adopted by Clarida et al. (2003) and Sager and Taylor (2014) who confirm the results and demonstrate statistically significant outperformance when applying the model to different data sets.[4] We now outline the theoretical basis for the Clarida and Taylor (1997) approach.

The framework of Clarida and Taylor (1997) shows that, given stationary departures from the RNEMH, γ_t, both spot and forward rate series inherit a common stochastic drift. Based on Beveridge and Nelson (1981) and Stock and Watson (1988), Clarida and Taylor (1997) express the spot exchange rate, s_t, as the sum of two processes:

$$s_t = z_t + q_t, \tag{1}$$

with z_t representing a random walk with drift and q_t being a zero mean stationary process with finite variance. Clarida and Taylor (1997) then make the assumption that γ_t is $I(0)$, leading to:

[3] Baba and Parker (2009) provide a discussion of dislocations of covered interest rate parity during the financial crisis.

[4] A heteroskedastic VECM is also proposed by Clarida et al. (2003), however the introduction of nonlinearities shows only marginal benefits over the standard homoskedastic VECM.

$$f_t^k = \gamma_t + k\theta + E_t\left(q_{t+k} \mid \Omega_t\right) + z_t, \tag{2}$$

where θ is a constant, representing the drift component of the random walk process, z_t. Comparing (1) and (2), we see that both the spot, s_t, and the forward series, f_t^k, share a common stochastic trend, z_t. As defined above, θ, γ_t and $E_t\left(q_{t+k} - q_t \mid \Omega_t\right)$ all constitute $I(0)$ series. It follows, therefore, that the forward premium, $f_t^k - s_t$, is also stationary, and that the forward and spot rates are cointegrated according to the vector $[1, -1]$:

$$f_t^k - s_t = \gamma_t + k\theta + E_t\left(q_{t+k} - q_t \mid \Omega_t\right). \tag{3}$$

Given that this is true for any forecasting horizon, k, the cointegrating relationship can be generalised to an $(N+1)$-dimensional system, comprised of the spot and N forward rates, $\left\{s_{t,} f_t^{k_1}, f_t^{k_2}, ..., , f_t^{k_N}\right\}$. In this case, an N-sized vector encompassing the system's forward premia represent the system's cointegrating equilibria. The strength of the approach is that it identifies both the components and coefficient parameters defining the system's cointegrating space. Consistent with Engle and Granger (1987), a system of spot and N forward rates can be well represented by a vector error correction model (VECM). Therefore, following Clarida and Taylor (1997), we estimate a restricted linear VECM using the maximum likelihood method of Johansen (1991), to obtain 4-, 13-, 26-, and 52-week ahead forecasts of the foreign exchange spot rate.

2.3 Clarida et al. (2003) MSVECM

Clarida et al. (2003) extend the linear VECM of Clarida and Taylor (1997) to a nonlinear multivariate Markov-switching framework in recognition of the nonlinearities evidenced in exchange rate movements (see for example, Engel and Hamilton 1990). However, the literature on nonlinear modelling of exchange rates still generally fail to beat random walk models in out-of-sample forecasting. In contrast, Clarida et al. (2003) demonstrate that nonlinear Markov-switching VECM forecasts are significantly superior to the random walk and, to some extent, the linear VECM forecasts. This is achieved in a dynamic recursive out-of-sample forecasting framework. They define their model, the Markov-Switching-Intercept-Heteroskedastic-VECM (MSIH-VECM) as:

$$\mathit{\Delta} y_t = v(z_t) + \sum_{i=1}^{p-1} \mathit{\Gamma}_i \mathit{\Delta} y_{t-i} + \mathit{\Pi} y_{t-1} + u_t, \tag{4}$$

where y_t is a K-dimensional observed time series vector; $\mathit{\Gamma}_i = -\sum_{j=i+1}^{p} \mathit{\Pi}_j$ are matrices of parameters;

$\Pi = \alpha\beta^{'}$, with β representing the matrix of cointegrating vectors; the error terms $u_t = NIID(0, \Sigma(z_t))$; and $z_t \in \{1, \ldots, M\}$ being a time dependent regime state variable, with M representing the total number of regimes assumed. This model specification allows for regime shifts in the intercept and the error term variation, which are respectively characterised by the regime state dependent vector $v(z_t)$ and the regime state dependent variance-covariance matrix $\Sigma(z_t)$.More specifically, Clarida et al. (2003) find the term structure of forward exchange rates to be well modelled using a multivariate three-regime Markov-switching VECM. The nonlinear VECM forecasts are strongly superior to the random walk model and outperform the linear VECM, in particular at longer forecasting horizons. To assess the efficacy of our functional scalar response model, we also employ an MSIH(3)-VECM(1) model, defining three regimes and one lag in the VECM system.

3 Functional scalar response model

Functional data analysis (FDA) provides a functional representation of the process underlying a data set; the process is defined over a continuum, where continuum values are most commonly represented in terms of time or space. In this paper, the functions we consider are defined over the space domain spanned by the tenors of the forward contracts, k. The function serves to characterise the forward foreign exchange rate dynamics in the spirit of the VECM approach of Clarida and Taylor (1997). While the VECM describes the dynamic relationship between the spot rate and the discretely observed forward exchange rates, the FDA approach describes this relationship over a continuum of forward exchange rates. FDA has many advantages over current modelling approaches: it accurately captures the forward rate term structure dynamics through its functional representation of the discretely observed forward curve; there is no imposed parametric structure, unlike that assumed with the VECM; it is computationally efficient, and so appealing for large scale empirical work; and it results in a process that can be evaluated on an arbitrarily fine grid, allowing foreign exchange contracts of any maturity tenor to be modelled. These and other advantages of FDA are outlined in Ramsay and Silverman (2005).

To begin, we assume an inherent link or *smoothness* between weekly observed spot and forward foreign exchange rates. Let $x_t\left(k_q\right) \equiv \left\{ f_{t,}^{k_0} f_t^{k_1}, f_t^{k_2}, ..., , f_t^{k_N} \right\}$ be the forward exchange rate curve for N available maturities. For notational convenience the representation, $f_t^{k_0}$, is utilised for the spot rate at time t. From the discretely observed $x_t\left(k_q\right)$, we uncover a continuous functional or curve representation, which we denote $\tilde{x}_t(k)$. This curve describes the forward rate term structure dynamics. When constructing the curve, a vector of n bases, denoted $\phi_1\left(k\right), ..., \phi_n\left(k\right)$, must first be specified. The decision of which basis system to choose is driven by the known characteristics of the underlying data. For instance, when modelling periodic data, a Fourier basis expansion, comprised of successive sine/cosine terms, is most commonly applied. However, the forward curve in foreign exchange markets does not exhibit strong cyclical variation, so we choose the popular and flexible B-spline for the basis function system - essentially a number of polynomials joined together smoothly at fixed points called knots.[5] The number and positioning of the knots are derived from knowledge of the complexity of the underlying process over particular ranges. In line with practice, we place knots in the range spanned by the observable discrete forward rate tenors $k_q : k_0 \leq ... \leq k_N$, with polynomials describing the tenor interval between the knots.

The functional structure or curve is then subsequently approximated as a weighted combination of these bases:

$$\tilde{x}_t(k) = c_1\phi_1(k) + c_2\phi_2(k) + ... + c_n\phi_n(k),$$

where $c_1, ..., c_n$ represent the parameters of the expansion coefficients. As in Ramsay and Silverman (2005), the coefficients c_j are chosen by minimising the sum of squared errors.

Constructing $\tilde{x}_t(k)$ relies on the assumption that there is a inherent link between consecutive observations along the forward rate tenor curve at a given point in time, t. This is a reasonable assumption that does not in itself constitute a failure of the RNEMH. However, we now proceed to use $\tilde{x}_t(k)$ for forecasting, with the view that the market mechanism imparts significant information to the term structure of the forward rates, an exercise dependent on departures from the RNEMH.

Second, we specify the dynamic relationship between spot and forward exchange rates. A functional regression model is used for this purpose, which provides a means to predict the evolution of the spot rate. The classical regression model seeks to describe the dependency between a scalar response variable and a specified set of scalar explanatory variables. However, in functional regression at least one of the observed variables is a curve. Given that the explanatory variable adopted in our study is the forward rate term structure, and we wish to predict future scalar spot rates, we employ the use of the scalar response/functional explanatory ("scalar response" henceforth) regression model of Horvath and Kokoszka (2012).

The functional forecasting model is given by

$$s_{t+\kappa} = \alpha' + \sum_{j=1}^{3} \rho_{t,j}\beta_j + \varepsilon_t. \tag{5}$$

where the $s_{t+\kappa}$ is the κ-week ahead spot rate, $\rho_{t,j}$ is the principal component score for each observation t

[5] In this paper, a second order polynomial or polygonal, is specified, significantly aiding computational efficiency. Another key strength of B-spline representation is that at any one point along the curve it simplifies to a polynomial that can be easily evaluated (De Boor, 2001).

on each principal component j and β_j is the functional sensitivity of the future spot rate to the dynamics of the lagged functional forward exchange rate curve. Three functional latent factors provides a good fit for the forward rate term structure. Please see Appendix Section A.1 for further mathematical details.

This approach offers dual benefits. First of these benefits is that it provides, for a future spot rate to be predicted, a sensitivity coefficient to the lagged forward rate term structure that is itself a function of the forward tenor k; in contrast, classical regression only permits scalar coefficients. This facilitates a dynamic link between spot and lagged forward rates and therefore allows the full forward rate term structure to inform predictions of the spot rate. Second of these benefits is that it models nonlinear relationships present in the forward exchange rate system through (i) the nonlinearity inherent in the forward exchange rate functional curve used as the explanatory variable and (ii) the nonlinearity of the functional sensitivity coefficient. In contrast, classical regression demands linearity in the coefficients. Such a feature of the FDA approach we propose is important given the evidence for nonlinearities in exchange rate markets (see for example, Baillie and Cho 2014).

4 Forecast evaluation

In this section, we define the loss functions used for the forecasting exercise and address the issue of multiple comparisons bias in our hypothesis testing setting.

4.1 Loss functions

The out-of-sample forecasts for a given horizon κ are obtained using a recursive scheme. Each week an additional observation is added to an expanding training window and the models are re-estimated. We choose this testing framework in line with Clarida and Taylor (1997), Clarida et al. (2003) and Sager and Taylor (2014). It ensures that forecasting is conditional only on information available at the time of the forecast, while the weekly expansion and re-estimating procedure serves to incorporate all available up-to-date information into the prediction. The accuracy of the forecasts are evaluated using the following measures:

1. Mean absolute error (MAE) is a measure of the average absolute difference between the forecast, $\hat{s}_{t+\kappa}$, and the corresponding realised observation, $s_{t+\kappa}$. It measures the average error magnitude in the forecasts, regardless of direction and serves to aggregate the errors into a single measure of predictive

power. Formally,

$$MAE \equiv \frac{1}{T-\kappa} \sum_{i=1}^{T-\kappa} |s_{t+\kappa} - \hat{s}_{t+\kappa}| .$$

2. Root mean squared error (RMSE) is a measure of the average squared difference between the values predicted by a model and the values realised. The RMSE is defined as the square root of the mean squared error, and again serves to aggregate the errors into a single measure of predictive power. Formally,

$$RMSE \equiv \sqrt{\frac{\sum_{i=1}^{T-\kappa} (s_{t+\kappa} - \hat{s}_{t+\kappa})^2}{T-\kappa}},$$

We present three different levels of forecast evaluation of our proposed functional scalar model. Firstly, we assess performance across the models through a direct comparison of forecasting measures. This is in line with the approach of Clarida and Taylor (1997). Secondly, we apply the Diebold-Mariano (1995) test to formally establish statistically significant outperformance; a testing approach consistent with Clarida et al. (2003) and Sager and Taylor (2014). Thirdly, in an important extension of the existing literature, we employ the use of a stepwise resampling based technique to correct for the multiple comparisons bias inherent in our multiple hypothesis test setting. This forecasting evaluation framework offers robust cross-model comparison, allowing us to ascertain scalar response outperformance relative to benchmark models that mitigates concerns over false discoveries. The next section details this multiple hypothesis testing technique. While the multiple comparisons bias correction addresses an important statistical issue in our forecasting exercise, our use of multiple hypothesis testing techniques also ensures that we do not overstate the effectiveness of our novel functional scalar response model.

4.2 Multiple hypothesis testing

The robust testing framework we employ adjusts for the likelihood that *seemingly* significant outperformance of one model relative to another at conventional statistical significance levels might in fact be a random artefact. The multiple comparisons problem states that given multiple simultaneous hypothesis tests, statistically significant results may be found by pure chance alone and so represent false discoveries rather sound findings. As we are simultaneously testing 96 hypotheses - given that we consider two performance measures, four comparative benchmark models, four forecasting horizons, and three currencies - such concerns over multiple comparisons bias cannot be ignored and must be addressed. Furthermore, given the novelty of the FDA application proposed in our empirical study, the approach we take is conservative and ensures that we set a very high statistical hurdle towards adjudicating on the effectiveness of the technique.

To control for multiple comparisons bias in our testing, we employ the operative balanced stepdown procedure of Romano and Wolf (2010). This recursive resampling based generalised multiple hypothesis testing procedure offers a more powerful and flexible approach to controlling for the multiple comparisons problem than previous multiple hypothesis testing techniques. It improves upon formerly proposed single step procedures, by allowing for subsequent iterative steps to identify additional hypothesis rejections, and it offers balance by construction in the sense that each hypothesis is treated equally in terms of power. It works by controlling the probability that some defined $k \geq 1$ or more false discoveries occur among a family of S tests, and in so doing offers greater power in identifying statistical significance than overly conservative techniques such as the reality check of White (2000) and the superior predictive ability technique of Hansen (2005) that control only for one or more false discoveries.[6] Consistent with the notation of Romano and Wolf (2010), the following definition is made for the *generalised familywise error rate*:

$$k\text{-}FWER_\theta = P_\theta \left\{ \text{reject at least } k \text{ null hypothesis } H_{0,s} : s \in \mathcal{I}(\theta) \right\}.$$

$\mathcal{I}(\theta)$ is defined as the set of true null hypotheses and k is user-defined. The generalised familywise error rate defines the probability of making k or more false discoveries in a multiple testing setting. A significance level α is then chosen such that the k-FWER$\leq \alpha$. See Appendices A.2 and A.3 for further implementation details.

With this definition in place, we build towards a formal testing framework to identify outperformance in the foreign exchange forecasting models. The following hypotheses are therefore considered:

$$H_0 : \theta \equiv \theta_{benchmark} - \theta_{SR} \leq 0$$

$$H_1 : \theta \equiv \theta_{benchmark} - \theta_{SR} > 0$$

[6]In an attempt to stay consistent with the notation of Romano and Wolf (2010) we reuse the letters k and s here. In this context, k represents the lower bound number of false discoveries controlled for in the Romano and Wolf (2010) framework and not the forward tenors as defined in previous sections, and s represents an index for the simultaneous hypothesis tests undertaken and not the spot rate.

where θ_{SR} is a given forecast evaluation measure for the functional scalar response model, and $\theta_{benchmark}$ is the corresponding measure for the comparative benchmark model: VECM of Clarida and Taylor (1997), MSIH-VECM of Clarida et al. (2003), or random walks with or without drift. Aligning closely with the forecast evaluation performance measures set out in Section 4.1, we consider MAE and MSE as our two performance measures for our hypothesis testing. Under the null hypothesis, where the difference in performance measures is negative, the scalar response model fails to outperform the stated benchmark model.

5 Data and empirical results

This section first describes the data used in our forecasting experiment and then presents the informal comparison of forecasting measures and the results of formally testing forecasting errors. The section concludes by considering whether the forecasting ability of the functional scalar response model translates into investment performance under a stylised trading strategy exercise.

5.1 Data

Our data set comprises observations of spot, and 4-, 13-, 26-, and 52-week forward rates for Euro, Japanese Yen and British Sterling, all versus the U.S. Dollar.[7] Weekly exchange rates are obtained over the period of the 26th week of 1990 (02-Jul-1990) to the 26th week of 2014 (30-Jun-2014), a total of 1,253 observations for each exchange rate series. Following Sager and Taylor (2014) and Della Corte et al. (2009), our Euro series is proxied by the German Deutschemark over the July 1990 to January 1999 period.[8] As in Clarida et al. (2003), we designate all but the final three years of the data set as the in-sample period. The data set is sourced from Thomson Reuters Datastream. The strong theoretical priors outlined in Section 2.2 dictate that forward premia, $f_t^k - s_t$, for each currency span the cointegration space according to the vector $[1, -1]$.[9] Therefore, we proceed by restricting the basis of the cointegration space through imposing the following condition on the VECM:

[7]We choose the same three currency pairs as Sager and Taylor (2014), who cite that they are the most actively traded pairs according to the Bank for International Settlements (2010).

[8]The use of a weekly data frequency is in line with Clarida and Taylor (1997), Clarida et al. (2003) and Sager and Taylor (2014).

[9]As in Clarida et al. (2003) and Sager and Taylor (2014), we proceed with the restrictions, $[1, -1]$, despite the likelihood ratio test indicating that the null hypothesis of four linearly independent forward premiums comprising the basis for the cointegration space is rejected. Clarida et al. (2003) conclude that although the departures from the precise overidentifying restrictions are statistically significant, they are very small in magnitude.

Table 1: Results of forecasting exercises: Dollar-Euro

κ (weeks)	SR (level)	VECM (ratio)	MSIH-VECM (ratio)	RW (ratio)	RWD (ratio)
Root mean square error (RMSE)					
4	0.0247	0.971	0.832	0.977	0.972
13	0.0368	0.936	0.856	0.947	0.935
26	0.0440	0.949	0.925	0.970	0.954
52	0.0621	1.090	1.113	1.139	1.132
Mean absolute error (MAE)					
4	0.0201	0.973	0.836	0.979	0.975
13	0.0319	0.952	0.937	0.972	0.966
26	0.0361	0.936	0.910	0.949	0.933
52	0.0556	1.153	1.187	1.213	1.236

The performance measure for the functional specification for each forecasting horizon is given in the second column of the table with the third to sixth columns containing the ratio of the scalar response performance measure to the corresponding VECM and random walk performance measures, respectively. Therefore, superior relative performance by the scalar response model is indicated by a ratio of less than 1. The forecast period is from July 2011 to July 2014. "SR" corresponds to "scalar response", "VECM" corresponds to the vector error correction model framework of Clarida and Taylor (1997), "MSIH-VECM" corresponds to the Markov switching intercept heteroskedastic VECM of Clarida et al. (2003), "RW" corresponds to a no-change random walk without drift, and "RWD" corresponds to a random walk with drift.

$$\beta' x_t = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 \\ -1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} s_t \\ f_t^4 \\ f_t^{13} \\ f_t^{26} \\ f_t^{52} \end{bmatrix}.$$

The VECM is dynamically estimated through the maximum likelihood method of Johansen (1991) to obtain 4-, 13-, 26-, and 52-week ahead forecasts.[10] The sample expands recursively with the optimised VECM being re-estimated at each time step (weekly).. The out-of-sample forecasting performance of all models are outlined in the next section.

5.2 Informal comparison

The goal of the paper is to assess the usefulness of the functional model set out in Section 3 to predict spot exchange rates using full forward curve information. To this end, RMSE and MAE measures are adopted to examine out-of-sample forecasting performance. The results presented in Tables 1, 2 and 3, compare the forecasting accuracy of our proposed scalar response model against those of VECM and random walk based

[10]For further technical VECM estimation details, the reader is directed to Johansen (1991) and Clarida and Taylor (1997). A first-order lag is chosen in line with Clarida and Taylor (1997) who cite algorithmic instability using higher-order lag specifications.

Table 2: Results of forecasting exercises: Dollar-Sterling

κ (weeks)	SR (level)	VECM (ratio)	MSIH-VECM (ratio)	RW (ratio)	RWD (ratio)
Root mean square error (RMSE)					
4	0.0213	0.993	0.885	0.998	0.998
13	0.0299	0.992	0.899	1.001	0.999
26	0.0365	0.960	0.903	0.969	0.952
52	0.0430	0.825	0.822	0.842	0.795
Mean absolute error (MAE)					
4	0.0166	0.979	0.845	0.982	0.978
13	0.0239	0.972	0.868	0.983	0.980
26	0.0291	0.932	0.857	0.941	0.921
52	0.0360	0.896	0.885	0.912	0.872

The performance measure for the functional specification for each forecasting horizon is given in the second column of the table with the third to sixth columns containing the ratio of the scalar response performance measure to the corresponding VECM and random walk performance measures, respectively. Therefore, superior relative performance by the scalar response model is indicated by a ratio of less than 1. The forecast period is from July 2011 to July 2014. “SR” corresponds to “scalar response”, “VECM” corresponds to the vector error correction model framework of Clarida and Taylor (1997), “MSIH-VECM” corresponds to the Markov switching intercept heteroskedastic VECM of Clarida et al. (2003), “RW” corresponds to a no-change random walk without drift, and “RWD” corresponds to a random walk with drift.

Table 3: Results of forecasting exercises: Dollar-Yen

κ (weeks)	SR (level)	VECM (ratio)	MSIH-VECM (ratio)	RW (ratio)	RWD (ratio)
Root mean square error (RMSE)					
4	0.0277	0.989	0.826	0.994	0.967
13	0.0575	0.927	0.872	0.936	0.881
26	0.0922	0.894	0.875	0.905	0.827
52	0.1444	0.919	0.948	0.935	0.809
Mean absolute error (MAE)					
4	0.0214	1.020	0.815	1.023	1.004
13	0.0439	0.956	0.848	0.964	0.902
26	0.0667	0.865	0.870	0.878	0.779
52	0.1261	0.948	0.972	0.970	0.803

The performance measure for the functional specification for each forecasting horizon is given in the second column of the table with the third to sixth columns containing the ratio of the scalar response performance measure to the corresponding VECM and random walk performance measures, respectively. Therefore, superior relative performance by the scalar response model is indicated by a ratio of less than 1. The forecast period is from July 2011 to July 2014. “SR” corresponds to “scalar response”, “VECM” corresponds to the vector error correction model framework of Clarida and Taylor (1997), “MSIH-VECM” corresponds to the Markov switching intercept heteroskedastic VECM of Clarida et al. (2003), “RW” corresponds to a no-change random walk without drift, and “RWD” corresponds to a random walk with drift.

alternatives. The performance measure for the functional specification is given in the second column of the table with the third to sixth columns containing the ratio of the scalar response performance measure to the corresponding VECM and random walk performance measures, respectively. As in Clarida and Taylor (1997), superior relative performance of the proposed model is indicated by a ratio of less than one with ratios calculated for each of the 4-, 13-, 26-, and 52-week forecasting horizons.

A direct comparison of the performance measures indicates that the scalar response model generally outperforms both the linear and nonlinear VECMs as well as the random walk benchmarks with and without drift. This result is broadly similar across all currencies, with the exception of some measure specific underperformance exhibited at the 4-week forecasting horizon for the Japanese Yen. The pockets of underperformance exhibited at the 52-week forecasting horizon for the Euro could be attributed to instability in the extrema values the constructed function.[11] Overall, these out-of-sample results are promising in that they show almost systematic outperformance of the scalar response model over the VECM and random walk approaches. This provides an initial indication of the ability of our functional model to extract useful forecasting information from the term structure beyond what has been achieved using approaches proposed to date. However, to test whether this outperformance holds statistically, we proceed by implementing our formal testing framework.

5.3 Formal testing

The literature is split on how best to evaluate forecasting performance. While Meese and Rogoff (1983a,b) and Clarida and Taylor (1997) infer model superiority using a direct comparison of performance measure differences, such as those presented in the previous section, both Clarida et al. (2003) and Sager and Taylor (2014) formally test for statistically significant outperformance. Aligning ourselves with these latter studies, we formally test the hypothesis of outperformance. We firstly perform Diebold-Mariano (1995) tests for outperformance and then apply the multiple hypothesis testing technique of Section 4.2 to mitigate concerns over multiple comparisons bias.

Recall that the following hypotheses are considered:

$$H_0 : \theta \equiv \theta_{benchmark} - \theta_{SR} \leq 0$$

[11] We note, as outlined in Section 2, that the function characterising the forward rate term structure is defined over the tenor range of 0 weeks to 52 weeks, with the constructed function subsequently being used as an explanatory covariate in the scalar response specification. In constructing the 52-week tenor, there is only one single weekly forward rate data point available to define the coefficients in this range. Such a set-up has been noted to lead to instability in the estimations of the range (see for example, Ramsay and Silverman 2005), and in turn, as may be the case here, can lead to more volatile forecast predictions in the range. A possible remedy is to use neighbouring forward rates to enrich the data, the 9 month and 2 year tenors for instance, however the use of additional forward rates would be inconsistent with the approach taken in previous studies, against which we aim to provide a relative comparison.

Table 4: Significant outperformance: Dollar-Euro

κ (weeks)	SR Vs VECM	SR Vs MSIH-VECM	SR Vs RW	SR Vs RWD
Difference in MSEs				
4	$0.00004^{\dagger}$*	$0.00027^{\dagger}$*	$0.00003^{\dagger}$*	$0.00004^{\dagger}$*
13	$0.00019^{\dagger}$*	$0.00050^{\dagger}$*	0.00016*	0.00019*
26	0.00021*	0.00033*	0.00012	0.00019
52	−0.00061	-0.00074	−0.00088	-0.00084
Difference in MAEs				
4	$0.00057^{\dagger}$*	$0.00397^{\dagger}$*	0.00042*	0.00051*
13	0.00161*	0.00215	0.00090	0.00113
26	0.00246*	0.00356*	0.00196	0.00258
52	−0.00740	-0.00875	−0.00975	-0.01062

"SR Vs VECM", "SR Vs MSIH-VECM", "SR Vs RW" and "SR Vs RWD" respectively denote, for a given performance measure, the comparison of the scalar response model with (i) the VECM framework of Clarida and Taylor (1997), (ii) the Markov-switching intercept-heterosekdastic VECM framework of Clarida et al. (2003) , (iii) a driftless random walk, and (iv) a random walk with drift. The mean difference in the square error (MSE) and absolute error (MAE) performance measures are reported, with positive numbers representing outperformance of the scalar response model over the stated benchmark model. The forecast period is July 2011 to July 2014. The symbol † is used to represent an instance of statistically significant outperformance after applying the resampling based balanced operative stepdown framework of Romano and Wolf (2010). The symbol * is used to represent an instance of statistically significant outperformance after applying the Diebold-Mariano (1995) test at a 5% significance level. A first order loss function is specified for the Differences in MAEs group and a second order loss function is specified for the Differences in RMSEs group.

$$H_1 : \theta \equiv \theta_{benchmark} - \theta_{SR} > 0$$

where θ_{SR} is a given forecast evaluation measure (MSE or MAE) for the functional scalar response model, and $\theta_{benchmark}$ is the corresponding measure for the given VECM or random walk benchmark. The results are shown in Tables 4, 5 and 6. The difference in MSE and MAE performance measures are reported, with positive numbers representing outperformance of the scalar response model over the stated benchmark model. The symbols * and † are used to represent an instance of statistically significant outperformance after applying the Diebold-Marino (1995) and Romano and Wolf (2010) approaches, respectively.

Firstly, as can be seen from the implementation of the Diebold-Mariano (1995) tests, significant outperformance can be observed in the majority of considered cases at a 5% level. This provides strong support for the superiority of our functional scalar response model relative to the standard benchmarks but in particular the leading VECM and MSIH-VECM models. However, as expected, implementing the multiple hypothesis testing framework of Romano and Wolf (2010) is much more conservative than the common practice of reporting results at the conventional statistical significance levels. Despite this high statistical hurdle, we still find many instances of significant outperformance in both the Euro and Japanese Yen. More specifically, we find eight instances of statistically significant functional outperformance for the Euro, evidenced at the shorter forecasting horizons. In the case of the Japanese Yen, results are even more promising, with nineteen

Table 5: Significant outperformance: Dollar-Sterling

κ (weeks)	SR Vs VECM	SR Vs MSIH-VECM	SR Vs RW	SR Vs RWD
Difference in MSEs				
4	0.00001	0.00013†	0.00000	0.00000
13	0.00001	0.00021*	0.00000	0.00000
26	0.00011	0.00030*	0.00009	0.00014
52	0.00087*	0.00089*	0.00076*	0.00107*
Difference in MAEs				
4	0.00035	0.00304†*	0.00031	0.00037
13	0.00070	0.00364*	0.00041	0.00049
26	0.00213	0.00487*	0.00181	0.00250
52	0.00416	0.00468	0.00345	0.00527

"SR Vs VECM", "SR Vs MSIH-VECM", "SR Vs RW" and "SR Vs RWD" respectively denote, for a given performance measure, the comparison of the scalar response model with (i) the VECM framework of Clarida and Taylor (1997), (ii) the Markov-switching intercept-heterosekdastic VECM framework of Clarida et al. (2003), (iii) a driftless random walk, and (iv) a random walk with drift. The mean difference in the square error (MSE) and absolute error (MAE) performance measures are reported, with positive numbers representing outperformance of the scalar response model over the stated benchmark model. The forecast period is July 2011 to July 2014. The symbol † is used to represent an instance of statistically significant outperformance after applying the resampling based balanced operative stepdown framework of Romano and Wolf (2010). The symbol * is used to represent an instance of statistically significant outperformance after applying the Diebold-Mariano (1995) test at a 5% significance level. A first order loss function is specified for the Differences in MAEs group and a second order loss function is specified for the Differences in RMSEs group.

Table 6: Significant outperformance: Dollar-Yen

κ (weeks)	SR Vs VECM	SR Vs MSIH-VECM	SR Vs RW	SR Vs RWD
Difference in MSEs				
4	0.00002	0.00036†*	0.00001	0.00005*
13	0.00054†*	0.00104†*	0.00047†*	0.00095†*
26	0.00213†*	0.00261†*	0.00188†*	0.00392†*
52	0.00386†*	0.00237*	0.00298†*	0.01103†*
Difference in MAEs				
4	-0.00041	0.00485†*	-0.00048	-0.00009
13	0.00200	0.00785†*	0.00164	0.00476*
26	0.01041†*	0.00996†*	0.00925†*	0.01895†*
52	0.00689*	0.00366	0.00394	0.03088†*

"SR Vs VECM", "SR Vs MSIH-VECM", "SR Vs RW" and "SR Vs RWD" respectively denote, for a given performance measure, the comparison of the scalar response model with (i) the VECM framework of Clarida and Taylor (1997), (ii) the Markov-switching intercept-heterosekdastic VECM framework of Clarida et al. (2003), (iii) a driftless random walk, and (iv) a random walk with drift. The mean difference in the square error (MSE) and absolute error (MAE) performance measures are reported, with positive numbers representing outperformance of the scalar response model over the stated benchmark model. The forecast period is July 2011 to July 2014. The symbol † is used to represent an instance of statistically significant outperformance after applying the resampling based balanced operative stepdown framework of Romano and Wolf (2010). The symbol * is used to represent an instance of statistically significant outperformance after applying the Diebold-Mariano (1995) test at a 5% significance level. A first order loss function is specified for the Differences in MAEs group and a second order loss function is specified for the Differences in RMSEs group.

identified instances of statistically significant outperformance of the functional scalar response model across all of the forecasting horizons. However, in contrast, there are only two identified hypothesis rejections for the British Pound under the Romano and Wolf (2010) procedure.[12]

While the results do not show systematic outperformance in this stringent statistical setting, they are particularly encouraging, in that the functional model demonstrates multiple instances of outperformance against the benchmark alternatives of the widely lauded VECM and MSIH-VECM frameworks as well as the notoriously hard-to-beat random walk benchmarks. The predictive results reported here underscore two of the benefits the functional methods offer: the ability to capture and extend to an infinite-dimension the information embedded in the forward foreign exchange rate curve on a given date and to relate this to future spot rate observations; and the ability to capture nonlinearities effectively through both the functional explanatory variables and the functional sensitivity coefficients. The findings presented here should motive empirical researchers to conduct further investigation into functional methods for forecasting purposes.

5.4 Trading Strategy

While the forecasting performance of the functional scalar response model is notable, such an ability to forecast foreign exchange rates may not necessarily translate into an ability to exploit the identified inefficiencies from an investment perspective. We therefore seek to shed some light on the economic significance of the forecasting performance of the functional scalar response model through implementing a stylised trading strategy implementation in the spirit of Sager and Taylor (2014).

Following the basic trading rule of Sager and Taylor (2014), we use the forecasts emanating from the functional scalar response model as trading signals to generate measures of investment performance. Specifically, we take as a buy signal the case where the four-week forecast of the spot rate is above the four-week forward rate, while we take as a sell signal the opposing case where the four-week forecast of the spot rate is below the four-week forward rate. There is no neutral position, trades are rebalanced weekly and transaction costs are not considered. The motivation for the focus on the four-week investment horizon is premised on its popularity with asset managers and hedge funds (Sager and Taylor, 2014). We align the evaluation of the trading strategy performance with our out-of-sample period. Results are reported in Table 7 for each of the three currencies considered (against the dollar) and an equally weighted portfolio of all three currencies.

The results suggest an ability on the part of the functional scalar response model to generate positive

[12] These two instances of outperformance are both identified when comparing SR to the nonlinear MSIH-VECM approach at the 4-week ahead forecasting horizon, mirroring what Clarida et al. (2003) and Sager and Taylor (2014) find in that the nonlinear VECM extension is a relative underperformer at 4-weeks ahead in comparison to longer forecasting horizons.

Table 7: Trading Strategy Results

	EUR		GBP		JPY		Equal-Weight Portfolio	
	SR	VECM	SR	VECM	SR	VECM	SR	VECM
Cumulative return	0.079	-0.001	-0.048	-0.052	0.111	-0.224	0.047	-0.092
Mean return	0.001	-0.000	0.000	-0.001	0.001	-0.002	0.000	-0.001
Std. dev. return	0.013	0.013	0.010	0.010	0.015	0.014	0.007	0.007
t-ratio	0.612	-0.005	-0.500	-0.538	0.760	-1.549	0.657	-1.354
p-value	0.542	0.996	0.618	0.592	0.449	0.124	0.512	0.179
Information ratio	0.441	-0.004	-0.361	-0.388	0.548	-1.117	0.474	-0.976
Returns skewness	0.097	0.202	-0.053	0.007	-0.753	-0.387	-0.910	-0.441
Returns kurtosis	2.490	2.518	2.173	2.185	6.270	5.912	5.589	4.576

Trading results are generated following the basic trading rule of Sager and Taylor (2014), whereby a buy signal is taken when the four-week forecast of the spot rate is above the four-week forward rate, while a sell signal is taken when the four-week forecast of the spot rate is below the four-week forward rate. There is no neutral position, trades are rebalanced weekly and transaction costs are not considered. The evaluation period is July 2011 to July 2014. "SR" corresponds to the functional scalar response model and "VECM" corresponds to the vector error correction model framework of Clarida and Taylor (1997).

trading returns, as evidenced in the case of the euro and yen currencies, although a negative return is reported for sterling. By comparison, following the trading signals from the VECM, as proposed by Sager and Taylor (2014), leads to negative cumulative returns across all currencies during this period. Furthermore, the positive information ratios calculated for the functional scalar response model suggest an ability to outperform the benchmark, although a caveat is that the reported investment returns are found to be statistically insignificant. While a full scale assessment of investment performance is out of scope of the current study, our findings should motivate future work in this direction.

6 Conclusion

It has been proven that the forward rate is not the optimal predictor of the future spot rate (Hansen and Hodrick 1980, Frankel 1980, Bilson 1981, Frankel and Rose 1995, Taylor 1995). However, the market mechanism may still impart a significant degree of information to the forward rates. The informational content of the forward rate term structure has been most successfully exploited by Clarida and Taylor (1997) and Clarida et al. (2003) with their dynamic VECM approaches predicting spot exchange rates out-of-sample with high precision. Building on this work we offer a novel functional data analysis alternative to exploit the informational content of the forward rates.

While the functional model does not conclusively beat the VECM approaches across all forecasting horizons, it shows great promise as a forecasting tool. The scalar response model leads to near systematic outperformance in terms of a direct comparison of performance measures, coupled with multiple instances of statistically significant outperformance as identified under our formal testing framework. These favourable functional results are cast in the context of remarkable VECM performance documented in numerous studies to date. The use of the flexible functional framework we propose has the advantage of removing the need to impose prescriptive assumptions on the system of foreign exchange rates. Clarida and Taylor (1997) outline the advantages of moving beyond single-equation methods, whereas this study achieves its forecasting performance by exploiting an infinite-dimensional space representation. The successful use of functional data analysis in our analysis should motivate additional empirical research to explore these techniques further and to build on the emerging finance and economics literature in this space. Our empirical analysis also serves to highlight the importance of controlling for multiple comparisons bias in multiple testing settings, such as that constructed here, as the absence of such controls lead to falsely overstating levels of outperformance by over 40%. Additionally, given the novelty of the functional model proposed, we use the multiple hypothesis testing framework to set a very high statistical hurdle to clear before making conclusive judgments on its forecasting performance.

The improvement in forecasting performance relative to the random walk indicates that the forward rate term structure does indeed contain important information about the evolution of the spot exchange rate, above what is embedded in the historic spot rate series itself. Further to this, the results reinforce the rejection of the risk neutral efficient market hypothesis. Elliott and Ito (1999) and Dunis and Miao (2007) highlight how even small pockets of predictability can be exploited profitably. Our study provides additional informal evidence supporting the view that exchange rates are indeed predictable to some degree with an assessment of profitability of the scalar response framework leading to positive cumulative returns for two of the three currency pairs considered. Therefore, we further vindicate the use of forward bias currency strategies.

A Appendix

A.1 Functional scalar response model

Formally, we utilise the scalar response framework to find the dependency between the current day, t, forward rate term structure curve, $\tilde{x}_t(k)$, and the κ-week ahead spot rate, $s_{t+\kappa}$, as follows:

$$s_{t+\kappa} = \alpha + {}_{\Omega_k} \beta\,(k)\,\tilde{x}_t(k)dk + \varepsilon_t, \tag{6}$$

where Ω_k is the defined forward exchange rate tenor range. As outlined above, $\beta\,(k)$ describes the functional sensitivity of the future spot rate to the dynamics of the lagged functional forward exchange rate curve, which as a function allows for nonlinearities. An important implementation issue similar to that encountered in classical linear regression arises however, as there must be fewer explanatory variables than observations.[13] Using a curve as an explanatory variable naturally gives an infinite-dimensional predictor of a finite set of responses. This means that an exact fit, leading to $\varepsilon = 0$, is always possible. It also means that an infinite number of possible $\beta(k)$ coefficients produce the same predictions. In a similar manner to dimension reduction achieved using standard discrete principal components analysis, a functional principal component based representation of the explanatory variable can be used to solve this underdetermination issue. To this end, we move on to our third and final key step and outline the procedure for obtaining functional principal components as proposed and detailed by Ramsay and Silverman (2005); the interested reader is directed to this reference for more technical discussion of the below concepts.

Functional principal component analysis (FPCA) proceeds in a manner similar in principle to standard principal component analysis (PCA) but involves the search for functional probes or principal components, $\xi\,(k)$, that correspond to probe scores or principal component scores, $\rho\,(\tilde{x}_t\,(k))$, and that capture the highest possible levels of variation in the data. The probe scores seek to identify the most important types of variation and are formally defined as:

$$\rho\,(\tilde{x}_t\,(k)) \equiv \;\xi\,(k)\,\tilde{x}_t\,(k)\,dk.$$

These probe scores are continuous analogs to the principal component scores derived via discrete summation under standard PCA. As mean is a common mode of variation across functional observations, it is removed,

[13] Classical definitions of weak and strict stationarity do not directly extend to this seperable Hilbertian space (Horvath and Kokoszka, 2012).

with the residuals, $\tilde{x}_t(k) - \bar{\tilde{x}}(k)$, being probed. The probe score variance,

$$Var\left[\xi(k)\left(\tilde{x}_t(k) - \bar{\tilde{x}}(k)\right)^2 dk\right],$$

corresponding to probe $\xi(k)$, is then calculated by means of the following optimisation of total variance:

$$\mu \equiv \max_{\xi}\left\{\sum_{t=1}^{T}\rho^2\left(\tilde{x}_t(k)\right)\right\},$$

subject to the natural size restriction of $\xi^2(k)\,dk = 1$. With parallels to standard PCA language, μ is referred to as the eigenvalue and $\xi(k)$ is referred to as the eigen*function* of the variance-covariance function $v\left(k, k^{'}\right) \equiv (T-1)^{-1}\sum_{t=1}^{T}\left[\tilde{x}_t(k) - \bar{\tilde{x}}(k)\right]\left[\tilde{x}_t\left(k^{'}\right) - \bar{\tilde{x}}\left(k^{'}\right)\right]$. To ensure that each principal component function captures a distinct mode of variation, each new component is required to be orthogonal to those computed previously, i.e. $\xi_h(k)\,\xi_l(k)\,dk = 0$ for all $h = 1, ..., l-1$. It follows that for each functional observation, $\tilde{x}_t(k) = \bar{\tilde{x}}(k) + \sum_{j\geq 0}\rho_{t,j}\xi_j(k)$, where $\rho_{t,j}$ is the principal component score for each observation t on each principal component j.

For our empirical analysis, we find that specifying three functional latent factors provides a good fit for the forward rate term structure, capturing a large proportion of total variance. To evaluate the functional forecasting model, we regress the scalar response variable, i.e. the κ-week ahead spot price $s_{t+\kappa}$, on the functional principal components of the constructed forward rate term structure curve, $\tilde{x}_t(k)$. We can now define our final functional forecasting model as:

$$s_{t+\kappa} = \alpha^{'} + \sum_{j=1}^{3}\rho_{t,j}\beta_j + \varepsilon_t. \tag{7}$$

The link between Eq.(6) and Eq.(7) can be readily observed by noting that with $\rho_{t,j} = \xi_j(k)\left(\tilde{x}_t(k) - \bar{\tilde{x}}(k)\right)dk$ then $\beta(k) \approx \sum_{j=1}^{3}\rho_{t,j}\xi_j(k)$.

A.2 Balanced stepdown procedure

The balanced stepdown procedure of Romano and Wolf (2010) improves upon formerly proposed single step procedures, by allowing for subsequent iterative steps to identify additional hypothesis rejections. It also offers balance by construction in the sense that each hypothesis is treated equally in terms of power. The stepdown procedure is constructed such that at each stage, information on the rejected hypotheses to date is used in re-testing for significance on the remaining hypotheses. Assume a set of test statistics $T_{n,s} = \hat{\theta}_{n,s} \equiv \hat{\theta}_{benchmark}^{n,s} - \hat{\theta}_{SR}^{n,s}$ associated with the hypothesis tests, where n is the sample size of the data used for estimation. Introducing some notation, let $H_{n,s}(\cdot, P_\theta)$ denote the distribution function of

$\left(\hat{\theta}_{n,s} - \theta_s\right)$ and let $c_{n,s}(\gamma)$ denote the γ-quantile of this distribution. The confidence interval

$$\left\{\theta_s : \hat{\theta}_{n,s} - \theta_s \leq c_{n,s}(\gamma)\right\}$$

then has coverage probability γ. Balance is the property that the marginal confidence intervals for a population of S simultaneous hypothesis tests have the same probability coverage. Within the context of controlling the generalised k-FWER, the overall objective is to ensure that the simultaneous confidence interval covers all parameters $\theta_s, s = 1, ..., S$, except for at most $(k-1)$ of them, for a given limiting probability $(1-\alpha)$, while at the same time ensuring balance (at least asymptotically). So, what is sought is that

$$P_\theta \left\{\hat{\theta}_{n,s} - \theta_s \leq c_{n,s}(\gamma) \quad \text{for all but at most } (k-1) \text{of the hypotheses}\right\}$$

$$\equiv P_\theta \left\{H_{n,s}\left(\hat{\theta}_{n,s} - \theta_s, P_\theta\right) \leq \gamma \quad \text{for all but at most } (k-1) \text{of the hypotheses}\right\}$$

$$\equiv P_\theta \left\{k\text{-}\max\left(H_{n,s}\left(\hat{\theta}_{n,s} - \theta_s, P_\theta\right)\right) \leq \gamma\right\} = 1 - \alpha.$$

Letting $L_{n,\{1,...,S\}}(k, P_\theta)$ denote the distribution of $k\text{-}\max\left(H_{n,s}\left(\hat{\theta}_{n,s} - \theta_s, P_\theta\right)\right)$, the appropriate choice of the coverage probability γ is then $L^{-1}_{n,\{1,...,S\}}(1-\alpha, k, P_\theta)$.

Given that P_θ is unknown, it is necessary to use appropriate bootstrapping techniques to generate an estimate of the coverage probability $L^{-1}_{n,\{1,...,S\}}\left(1-\alpha, k, \hat{P}_\theta\right)$, under $\hat{P}_\theta$. Therefore, from this development it is possible to define the simultaneous confidence interval

$$\left\{\theta_s : \hat{\theta}_{n,s} - \theta_s \leq H^{-1}_{n,s}\left(L^{-1}_{n,\{1,...,S\}}\left(1-\alpha, k, \hat{P}_\theta\right), \hat{P}_\theta\right)\right\}.$$

The right-hand side of the above inequality will form the basis of the critical value definitions used within the stepdown procedure. See Romano and Wolf (2010) for further technical details. Note that although the above development was made assuming the full set of hypothesis tests, it equally applies to any subset $K \subseteq \{1, \ldots, S\}$. Hence, the balanced stepwise algorithm may now be described as follows.

- **Step 1:** Let A_1 denote the full set of hypothesis indices, i.e. $A_1 \equiv \{1, \ldots, S\}$. If for each hypothesis test, the associated test statistic $T_{n,s}$ is less than or equal to the corresponding critical value estimate, $\hat{c}_{n,A_1,s}(1-\alpha, k) \equiv H^{-1}_{n,s}\left(L^{-1}_{n,A_1}\left(1-\alpha, k, \hat{P}_\theta\right), \hat{P}_\theta\right)$, then fail to reject all null hypotheses and stop the algorithm. Otherwise, proceed to reject all null hypotheses $H_{0,s}$ for which the associated test statistics exceeds the critical value level, i.e., where $T_{n,s} > \hat{c}_{n,A_1,s}(1-\alpha,k)$.

- **Step 2:** Let R_2 denote the set of indices for the hypotheses rejected in Step 1 and let A_2 denote the indices for those hypotheses not rejected. If the number of elements in R_2 is less than k, i.e., $|R_2| < k$, then stop the algorithm, as the probability of k or more false discoveries is zero in this case. Otherwise, the appropriate critical value to be applied for each hypothesis test s at this stage is calculated as follows:

$$\hat{d}_{n,A_2,s}(1-\alpha,k) = \max_{I \subseteq R_2, |I|=k-1} \{\hat{c}_{n,K,s}(1-\alpha,k) : K \equiv A_2 \cup I\}.$$

  Hence, additional hypotheses from A_2 are rejected if $T_{n,s} > \hat{d}_{n,A_2,s}(1-\alpha,k), s \in A_2$. If no further rejections are made then stop the algorithm.

  ⋮

- **Step j:** Let R_j denote the set of indices for the hypotheses rejected up to Step $(j-1)$ and let A_j denote the indices for those hypotheses not rejected. The appropriate critical value to be applied for each hypothesis test s at this stage is calculated as follows:

$$\hat{d}_{n,A_j,s}(1-\alpha,k) = \max_{I \subseteq R_j, |I|=k-1} \{\hat{c}_{n,K,s}(1-\alpha,k) : K \equiv A_j \cup I\}.$$

  Hence, additional hypotheses from A_j are rejected if $T_{n,s} > \hat{d}_{n,A_j,s}(1-\alpha,k), s \in A_j$. If no further rejections are made then stop the algorithm.

At each step j in the stepwise procedure, the hypotheses that are not rejected thus far are re-tested over a smaller population of hypothesis tests than previously. The size of this smaller population is given $(|A_j| + k - 1)$, which includes all the hypotheses within A_j, in addition to $(k-1)$ hypotheses drawn from those hypotheses already rejected, i.e., drawn from R_j. Given that control of the generalised k-FWER is the premise of the procedure, it is expected that there are at most $(k-1)$ false discoveries amongst the set of hypotheses rejected R_j. However, it is not known which of the rejected hypotheses may represent false discoveries. Hence, it is necessary to circulate through all combinations of R_j, of size $(k-1)$, in order to obtain the appropriate critical values. A maximum critical value $\hat{d}_{n,A_j,s}(1-\alpha,k)$ must be determined for each hypothesis test s. This adds an additional layer of computational burden on the algorithm.

A.3 Operative Method

In requiring to circulate through all subsets of R_j, of size $(k-1)$, in order to obtain the maximum critical value to apply at each stage of the stepdown procedure, the algorithm can become highly, if not excessively, computationally burdensome. Depending on the $|R_j|$ and the value of k, the number of combinations ${}^{|R_j|}C_{k-1}$ can become very large. Romano and Wolf (2010) therefore suggest an operative method that reduces this computational burden, while at the same time maintaining much of the attractive properties of the algorithm.[14]

It is first necessary to be able to order the hypothesis tests rejected up to step $(j-1)$ in terms of significance. To this end, it is noted that marginal p-values can be obtained as follows:

$$\hat{p}_{n,s} \equiv 1 - H_{n,s}\left(\hat{\theta}_{n,s}, \hat{P}_{\theta}\right).$$

This gives the following ascending order for the significance of the hypothesis tests:

$$\hat{p}_{n,r_1} \leq \hat{p}_{n,r_2} \leq \cdots \leq \hat{p}_{n,r_{|R_j|}},$$

where $\left\{r_1, r_2, \ldots, r_{|R_j|}\right\}$ is the appropriate permutation of associated hypothesis test indices that gives this ordering. As before, a maximum number of combinations, N_{max}, at each step of the algorithm is defined. Then an integer value M is chosen such that ${}^{M}C_{k-1} \leq N_{max}$, leading to the calculation of the critical values as follows:

$$\hat{d}_{n,A_j,s}(1-\alpha,k) = \max_{I \subseteq \left\{r_{\max\left(1,\left|R_j\right|-M+1\right)},\ldots,r_{\left|R_j\right|}\right\}, |I|=k-1} \left\{\hat{c}_{n,K,s}(1-\alpha,k) : K \equiv A_j \cup I\right\}.$$

What this serves to do is to replace circulating through all the hypothesis tests rejected to date with that of circulating through only the M least significant hypothesis tests rejected. Of course, in the case where $M \geq |R_j|$, then this amounts to circulating through all the hypotheses rejected. Although this approach is premised on the assumption that the (up to $k-1$) false discoveries lie within the least significant hypotheses rejected so far, it does offer significant computational efficiencies for the algorithm. It is this operative method that is used for the empirical analysis in subsequent sections.

[14] Attractive properties include conservativeness, which allows for finite sample control of the k-FWER under P_{θ}, and provides asymptotic control in the case of contiguous alternatives.