

# A Dynamic Model of Expected Bond Returns: a Functional Gradient Descent Approach

Francesco Audrino\* Giovanni Barone-Adesi

*Institute of Finance, University of Lugano, Via Buffi 13, 6900 Lugano,  
Switzerland*

---

## Abstract

A multivariate methodology based on Functional Gradient Descent to estimate and forecast time-varying expected bond returns is presented and discussed. Backtesting this procedure on US monthly data, empirical evidence of its strong forecasting potential in terms of the accuracy of the predictions is collected. The proposed methodology clearly outperforms the classical univariate analysis used in the literature.

*Key words:* Term structure, Bond returns, Functional gradient descent, Semi-parametric VAR-GARCH models.

---

## 1 Introduction

The joint dynamics of bond returns at different maturities have been object of numerous theoretical models, but relatively little empirical study. The high dimensionality of the necessary models has forced earlier studies to rely on univariate methods as in Fama (1984). However, the disadvantage of this approach is that it suffers because of its lack of power, that prevents reaching firm conclusions on bond expected returns. Other studies overcome the curse of dimensionality problem by using variance reduction or similar techniques. In particular, there is a growing body of literature on the behavior of the bond return dynamics that is based on factor models, starting with Brennan and Schwartz (1982) and Stambaugh (1988), to the more recent studies by Ilmanen (1995), Ang and Piazzesi (2003) and Cochrane and Piazzesi (2005). In his work, Ilmanen (1995) ran regressions of monthly excess returns on bond

---

\* Corresponding author: Tel: +41586664789; fax: +41586664647.  
E-mail address: francesco.audrino@lu.unisi.ch.

in different countries on a term spread, the real short rate, stock returns, and bond return betas. In particular, he found that a model with a small set of global instruments can forecast 4 to 12 percent of monthly variation in excess bond returns. In a similar way, Cochrane and Piazzesi (2005) studied the time variation in expected bond returns by running regressions of one-year excess bond returns on initial forward rates. They found that a single tent-shaped factor predicts well excess returns on 1-5 year maturity bonds. Nevertheless, factor models restrict a priori expected returns and the shape of the term structure in ways that may conflict with empirical observations. Moreover, the information loss due to the use of a factor structure can have significant drawbacks in terms of the accuracy of the expected bond return estimates and predictions.

Recently Audrino and Bühlmann (2003) proposed an adaptation of the functional gradient descent technique (FGD), developed in Friedman et al. (2000) and Friedman (2001) mainly to solve the classification problem in the machine learning context, to cope with high dimensional models in economics. The main contribution of the FGD technique is that it allows to overcome the curse of dimensionality without resorting to variance reduction techniques. As a consequence, the whole correlation structure among the bond returns at different maturities is taken into account in the estimation and prediction. Audrino et al. (2005) applied the FGD approach to study the stability of the term structure of interest rates. In particular, the authors filtered the data with FGD and focussed on an in-sample analysis of the loading stability from a three factor decomposition of the yield curve. They did not investigate further the estimating and forecasting ability of the FGD technique. We extend their approach to study in-sample and out-of-sample FGD forecasts of the time-varying dynamics of realized returns on notional zero coupon US bonds with maturities ranging from one month to ten years. These forecasts also differ from yields to maturity previously studied by Audrino et al. (2005) because of the different investment horizon and the presence of time-varying liquidity premia.

Our initial return data are obtained applying Babcock's formula (1984) to spot rates from US monthly bond prices over the period January 1965 to December 1999. Most of our analysis concerns bonds with maturities up to 1 year because of data limitations. The main result of our analysis is that the accuracy of expected bond return forecasts is significantly improved by the inclusion of past lagged returns at different maturities in their prediction. This can be of primary importance when estimating and forecasting term premia.

The plan of the paper is as follows. Section 2 presents our model and the corresponding estimation procedure. The empirical results for zero coupon US bond returns at twelve different maturities are presented in Section 3.

Section 4 summarizes and concludes.

## 2 The model

This section describes the multivariate dynamic model we use to estimate and forecast expected bond returns at different maturities. In addition, we briefly present the estimation procedure based on Functional Gradient Descent (FGD) introduced in Audrino et al. (2005) that can be applied to it.

### 2.1 The general dynamic setting

For the purposes of our study, the multivariate data of interest are index bonds at various maturities. To construct such data from yields we rely on the formula developed by Babcock (1984). Consider the multivariate time series of yields to maturity on day  $t$

$$\{y_{t,T_i}; t = 1, \dots, n, i = 1, \dots, d\}, \quad (1)$$

where  $T_i, i = 1, \dots, d$ , are the different maturities. Suppose that the investment horizon is 1 month long. To construct an index bond representing a 1-year treasury issue we use the series of rates of return produced by a bond that is a 1-year bond at the beginning of each month and becomes a 11-month bond at the end of each month. In the first month of the sample, the yield of the 1 year bond equals  $y_{1,T_i}$ , with  $T_i = 1$  year. At the end of the month (or at the beginning of the next one), the yield of the 11-month bond equals  $y_{2,T_j}$ , with  $T_j = 11$  months. The following equation developed by Babcock (1984) provides for an extremely good approximation for the rate of return  $r_{2,T_i}$  on the 1-year bond for the month

$$r_{2,T_i} = y_{1,T_i} + \left(1 - \frac{D_{1,T_i}}{s}\right)(y_{2,T_j} - y_{1,T_i}). \quad (2)$$

The symbol  $s$  in (2) represents the length of the time interval over which the rate of return is computed, in our case one-twelfth of a year. The multivariate time series  $\{D_{t,T_i}\}$  refer to the  $T_i$ -bond's duration at the beginning of each month  $t$  of the sample. In particular, in our empirical investigation of Section 3, we consider US zero-coupon bonds, such that for each  $T_i, i = 1, \dots, d$ , the time series of bond durations  $\{D_{t,T_i}\}$  are constant and equal the time to maturity  $T_i$ . We assume stationarity of the multivariate time series of rates of return for the month

$$\{r_{t,T_i}; t = 2, \dots, n, i = 1, \dots, d\} \quad (3)$$

constructed using (2).

Our goal is to find time-varying estimates and forecasts for the conditional first moments of the series of rates of returns given in (3), i.e. time-varying estimates and forecasts for the expected bond returns. To this purpose, we consider a general multivariate dynamic model, in the class of the semi-parametric VAR-GARCH models. In particular, we assume that the dynamics of the multivariate conditional mean  $\mu_t = \mathbb{E}[\mathbf{r}_t \mid \mathcal{F}_{t-1}]$  of  $\mathbf{r}_t = (r_{t,T_1}, \dots, r_{t,T_d})^T$ , where  $\mathcal{F}_{t-1}$  denotes the information available up to time  $t-1$ , i.e.  $\mathcal{F}_{t-1}$  the  $\sigma$ -algebra generated by  $\{\mathbf{r}_v; v \leq t-1\}$ , are specified by

$$\begin{aligned} \mathbf{r}_t &= \mu_t + \xi_t, \quad \mu_t = (\mu_{t,1}, \dots, \mu_{t,d})^T \quad \text{and} \\ \mu_{t,i} &= G_i(\{r_{t-j,T_k}; j = 1, 2, \dots, k = 1, \dots, d\}), \quad i = 1, \dots, d, \end{aligned} \quad (4)$$

where we assume that  $\{\xi_t\}_t$  is a sequence of i.i.d. multivariate Gaussian innovations having mean zero and time-varying covariance matrix  $\text{Cov}(\xi_t) = V_t$ . In particular, we assume that the dynamics of the covariance matrix  $V_t$  of the innovations follow a generalization of the multivariate constant conditional correlation (CCC)-GARCH model introduced by Bollerslev (1990), where individual volatilities are allowed to be also of a general semi-parametric form exactly as in Audrino et al. (2005). The functional form introduced in (4) for the conditional means implies dependence on past multivariate observations. This assumption allows for the exploitation of the cross-information included (in particular) in bond returns at nearer maturities, with a significant improvement in the accuracy of the forecasts. This is one of the nice features of such a type of model.

## 2.2 The FGD estimation procedure

As Audrino et al. (2005) have already shown, a slightly modified version of the classical procedure based on the Functional Gradient Descent (FGD) technique introduced firstly in the financial context by Audrino and Bühlmann (2003) is a powerful strategy to construct accurate estimates and predictions for the expected bond returns  $\mu_t$ , exploiting all the available information included in the multivariate bond return process. Moreover, this estimation procedure is computationally feasible independently of the dimension of the multivariate bond return time series, overcoming the curse of dimensionality.

The algorithm used in this study is the same introduced in Audrino et al. (2005), Appendix A. Therefore, we briefly review here the main ideas of such an estimation procedure, referring to the above cited work for a detailed description and for a discussion of its reliability.

First of all, we restrict the conditional mean functions  $G_i(\cdot) : \mathbb{R}^{pd} \rightarrow \mathbb{R}$  in

(4) to depend on the first  $p$  lagged multivariate observations, similarly to the standard VAR( $p$ ) model. The main idea of FGD is to find the estimates for the conditional mean functions  $G_i(\cdot)$  that minimize a suitable empirical criterion, under the constraint requiring that the optimal solutions  $\hat{G}_i(\cdot)$  must be of an additive form. In every iteration step of the algorithm, a new term is added in the expansion of the  $G_i(\cdot)$  functions. This new term is estimated using a statistical procedure  $\mathcal{S}$  via a (constrained or penalized) least square fitting. The statistical procedure  $\mathcal{S}$  can be any reliable procedure for forecasting; typical examples for  $\mathcal{S}$  are regression trees, projection pursuit, splines or neural nets. The only requirement on  $\mathcal{S}$  is that it must be of a simple form (i.e. must involve the estimation of a small number of parameters) to avoid overfitting.

In our study, a suitable empirical criterion is given by the negative log-likelihood function in (4)

$$- \sum_{t=p+1}^n \log \left( (2\pi)^{-d/2} \det(V_t)^{-1/2} \exp(-\xi_t^T V_t^{-1} \xi_t / 2) \right), \quad (5)$$

with  $\xi_t = \mathbf{r}_t - \mu_t = \mathbf{r}_t - \mathbf{G}(\mathbf{r}_{t-p}^{t-1})$ . We choose  $\mathcal{S}$  to be a regression tree with three end-nodes. We find that this statistical procedure represents a good trade-off between flexibility and simplicity, since it involves only the estimation of 5 parameters (i.e. 2 threshold and 3 location parameters). The complexity of  $\mathcal{S}$  can be further reduced via shrinkage toward zero. This means the introduction of a parameter  $\nu$ ,  $0 < \nu \leq 1$ , that multiplies every term in the additive expansion of the functions  $G_i(\cdot)$ . Obviously, this reduces the variance of the statistical procedure  $\mathcal{S}$  by the factor  $\nu^2$ . In particular, in our empirical investigations in Section 3 we use a shrinkage factor of 0.5. For all details, see Audrino et al. (2005).

To stop the algorithm, we use the following cross-validation scheme: we split the (in-sample) estimation period into two sets, the first of size  $0.7n$  used as training set and the second of size  $0.3n$  used as test set. The optimal stopping value  $\hat{M}$  (and consequently the optimal number of terms in the additive expansion of the functions  $G_i(\cdot)$ ) is then chosen to optimize the cross-validated negative log-likelihood. Note that during the estimation procedure, we have also to find the optimal number of past lags  $p$  in the multivariate series of bond returns to be used as predictors in constructing the conditional mean functions. The optimal number  $\hat{p}$  is the one that optimizes the cross-validated negative log-likelihood.

Summarizing, the functional form that the vector of expected bond returns  $\mu_t$  can take is

$$\mu_t = \mathbf{G}(\mathbf{r}_{t-j}, j = 1, 2, \dots) = \mathbf{G}_0(\mathbf{r}_{t-j}, j = 1, 2, \dots) + \sum_{m=1}^M \sum_{k=1}^3 \gamma_k^{(m)} I_{\left[ \mathbf{r}_{t-p}^{t-1} \in \mathcal{R}_k^{(m)} \right]}, \quad (6)$$

where  $\mathbf{G}_0(\mathbf{r}_{t-j}, j = 1, 2, \dots)$  are some initial estimates,  $I$  is the indicator function, and  $\gamma_k^{(m)}$  and  $\mathcal{R}_k^{(m)}, k = 1, \dots, 3$ , are respectively the vectors of location parameters and the cells of the predictor space that are constructed when fitting by least squares the regression tree with three end-nodes. In particular, in our empirical investigations of Section 3 we use the estimates from a standard diagonal VARMA(1,1) model as starting functions  $\mathbf{G}_0(\cdot)$ .

### 3 An empirical investigation of the US term structure

This section presents the results of our estimations for different zero-coupon bond (rate of) returns for the month constructed on the US term structure.

#### 3.1 Data

In our empirical investigation, we use monthly U.S. zero-coupon bond yields at  $d = 12$  different maturities from 1 to 12 months. The data comes from the Center for Research in Security Prices (CRSP) database. In particular, we downloaded the data from the Fama Treasury Bill 12-month Term Structure Files. The time period considered is between January 1965 and December 1999, for a total of 420 monthly observations. In the backtest analysis of the following sections, we split the whole data sample in an in-sample period until December 1994 (360 observations) and in a subsequent out-of-sample period from January 1995 to December 1999 (60 observations). We estimate the optimal functional forms for the expected bond returns, given in equation (6), using the in-sample data. Then, we backtest the accuracy of the forecasts over the out-of-sample period, keeping the estimated optimal parameters fixed.

Note that in addition to the data described above and reported in this study, we also performed a similar analysis on the whole US term structure, with maturities going from 1 month to 10 years. We downloaded the data from *Datastream International*, for the time period between April 1997 to September 2005, for a total of 102 observations. However, since the data at monthly maturities longer than 1 year are available for this time period only, we were able only to perform an in-sample analysis. In-sample results for maturities longer than 1 year are qualitatively the same as those reported in the next sections. Results are available upon request.

In Figure 1, we provide a three-dimensional plot of our term structure data. As expected, the average yield curve is upward sloping and long rates are less volatile than short rates.

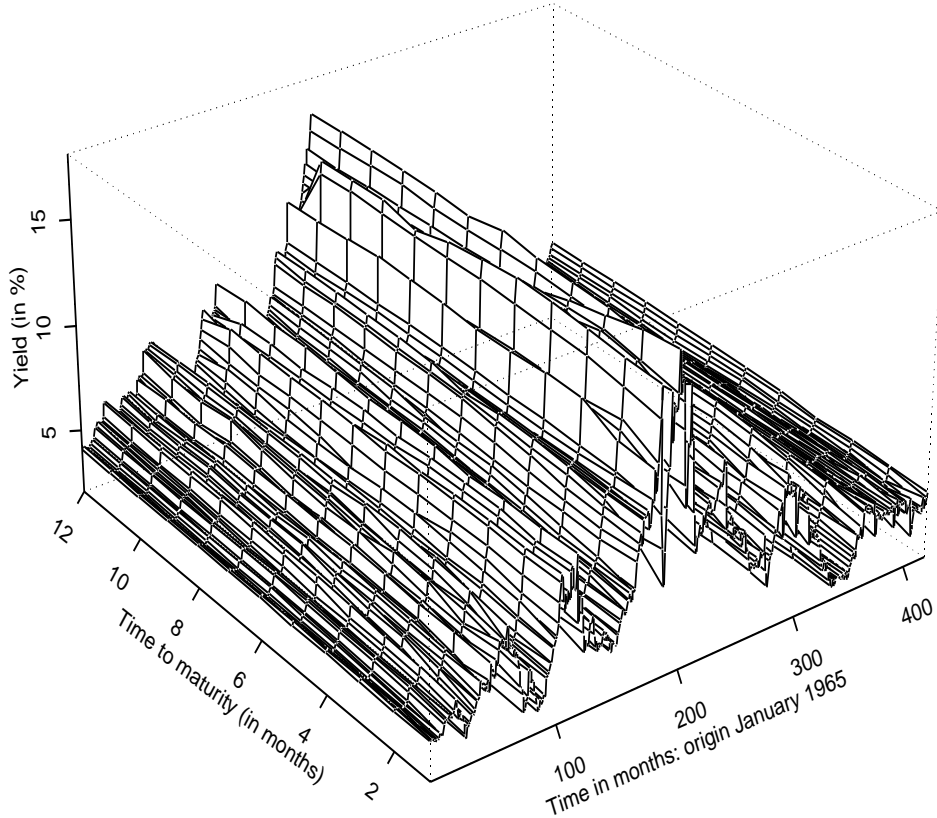


Fig. 1. Three-dimensional plot of the term structure data: the sample consists of 420 monthly US zero-coupon bond yields downloaded from the CRSP for the time period between January 1965 and December 1999, at all monthly maturities from 1 month up to 1 year.

More interesting for our purposes is the analysis of the bond returns computed using the formula given in equation (2). Summary statistics of the bond return time series are given in Table 1.

The mean of the bond returns tends to increase from the 1 month to the 6 months maturity and remains fairly constant for longer maturities. In contrast, the standard deviation of the bond returns increases significantly from short to long maturities, implying higher variability of the bond returns at longer maturities. At all maturities there is (strong) evidence of persistent autocorrelation in the bond return series. We compute classical Ljung-Box statistics testing for autocorrelations in the individual bond return series up to the 10th order. The test statistics are for all maturities significant at the 5% level or better. Moreover, we also compute Ljung-Box statistics testing for cross-correlations in the multivariate series of bond returns. Once again, most of the test statistics are rejected at the 5% level. All our findings yield empirical support for the use of model (4) for the analysis of the (time-varying)

Maturity	Mean	St. Dev.	$\hat{\rho}(1)$	$\hat{\rho}(5)$	$\hat{\rho}(10)$
1 month	6.189	2.525	0.955	0.850	0.755
2 months	6.554	2.759	0.875	0.778	0.700
3 months	6.837	3.073	0.771	0.662	0.620
4 months	6.851	3.337	0.652	0.511	0.532
5 months	7.024	3.716	0.534	0.438	0.464
6 months	7.076	4.131	0.466	0.289	0.366
7 months	7.039	4.504	0.430	0.202	0.308
8 months	7.244	4.953	0.390	0.155	0.287
9 months	7.316	5.442	0.355	0.128	0.246
10 months	7.166	5.988	0.304	0.080	0.208
11 months	7.222	6.498	0.281	0.045	0.174
12 months	7.245	6.637	0.416	0.031	0.147

Table 1

Summary statistics for the bond returns computed using formula (2) at all monthly maturities from 1 month to 1 year. The sample consists of 420 monthly data for the time period between January 1965 and December 1999. St. Dev. and  $\hat{\rho}(x)$  mean sample standard deviation and sample autocorrelation at lag  $x$ , respectively.

dynamics of the expected bond returns.

### 3.2 Estimation results

This section presents the results of our estimation of expected bond returns on the US data presented in Section 3.1 for the in-sample time period between January 1965 and December 1994. Results for the choice of the number of past lags  $p$  that are included in the additive expansion of the conditional mean functions  $G_i(\cdot)$  in (6) as well as those for the choice of the optimal stopping parameter  $M$  in the FGD algorithm are presented in Table 2.

Table 2 shows that the optimal number of past lags  $\hat{p}$  is 2, with a consequent  $\hat{M} = 49$  optimal number of FGD iterations. Note that this finding means that the expected bond return functions  $G_i(\cdot)$  depends on the whole history of its return time series  $\{r_{s,T_i}, s < t\}$ , incorporated in the starting functions  $G_{0,i}$ , as well as on the first two lagged multivariate observations  $\mathbf{r}_{t-1}$  and  $\mathbf{r}_{t-2}$ . Therefore, cross-information included in the bond return series is relevant for prediction only when considering at most the previous two months. Including



$p$	$M$	Cross-validated neg. log-likelihood
1	32	5147.49
2	49	5053.88
3	32	5115.81
5	21	5309.31

Table 2

In-sample estimation results from the FGD procedure introduced in Section 2.2. The in-sample time period starts in January 1965 and ends in December 1994, for a total of 360 monthly observations.  $p$  and  $M$  denotes the number of past lags in the multivariate bond return series that are considered as predictors in the estimation of the conditional mean functions and the number of iterations of the FGD algorithm, respectively.

more past multivariate information in the estimation procedure increases the accuracy in-sample, but leads to overfitting.

When analyzing more in details the structure of the additive terms in (6) we can make the following comments. First, most of the time (about 80%) the relevant predictors chosen by the FGD procedure in the regression tree construction are past lags of bond returns at a different maturity than the one for which we are estimating the conditional mean. This implies that cross-information is fundamental in increasing the accuracy of the starting estimates. This result is not surprising, since the starting functions are already constructed by considering the whole information included in the individual bond return series. Second, in general the estimated optimal location parameters  $\gamma_k^{(m)}$  in (6) are small (in magnitude), once again to protect against overfitting.

To end this section, we plot in Figure 2 the estimated dynamics for the expected bond returns at two representative maturities (6 months and 1 year). For a comparison, these optimal dynamics are superimposed on those coming from a classical univariate analysis of each bond return series. Note that the latter are used as starting functions in the FGD procedure.

Figure 2 clearly shows that differences in the estimated expected bond returns are very small. Nevertheless, as we will see in the next sections these little differences between the final FGD estimates and the starting functions become statistically significant when considering performance measures for the expected bond returns. Figure 2 also shows that the dynamics of the estimated conditional means are time-varying. Therefore, a model that estimates and forecasts constant expected bond returns may yield very poor and misleading results.

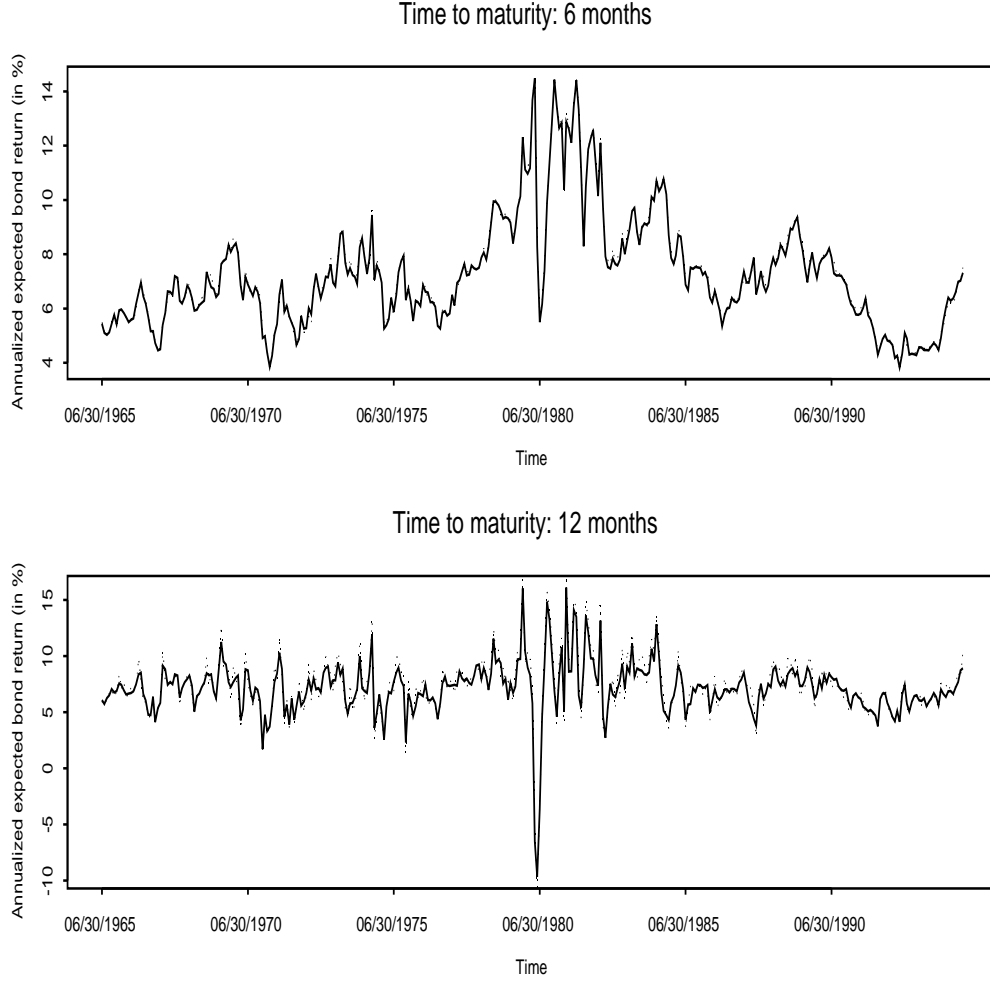


Fig. 2. Estimated expected bond return dynamics at two representative maturities of 6 months (top) and 12 months (bottom). The in-sample period goes from January 1965 to December 1994, for a total of 360 monthly observations. The time-varying dynamics are estimated using the FGD procedure (solid lines) and a classical univariate analysis (dotted lines).

### 3.3 Out-of-sample performance results

We evaluate in our investigation the out-of-sample performance of the FGD based multivariate procedure, also in comparison to a standard univariate ARMA(1,1) (-GARCH(1,1)) analysis of each individual bond return series. Note that the conditional mean estimates from the latter model are used as starting functions in the FGD algorithm. The out-of-sample period goes from January 1995 to December 1999, for a total of  $n_{\text{out}} = 60$  monthly observations.

We quantify the out-of-sample performance of the expected bond return pre-

dictions obtained using the competitive approaches by means of individual (i.e. for each maturity) and global mean absolute error (MAE) and mean square errors (MSE). More in details, individual out-of-sample MAE and MSE of the predictions are measured by

$$\text{MAE}_i = \frac{1}{n_{\text{out}}} \sum_{t=1}^{n_{\text{out}}} |\xi_{t,T_i}| = \frac{1}{n_{\text{out}}} \sum_{t=1}^{n_{\text{out}}} |r_{t,T_i} - \hat{\mu}_{t,T_i}| \quad \text{and} \quad (1)$$

$$\text{MSE}_i = \frac{1}{n_{\text{out}}} \sum_{t=1}^{n_{\text{out}}} |\xi_{t,T_i}|^2 = \frac{1}{n_{\text{out}}} \sum_{t=1}^{n_{\text{out}}} |r_{t,T_i} - \hat{\mu}_{t,T_i}|^2, \quad i = 1, \dots, d, \quad (2)$$

and the global out-of-sample MAE and MSE by

$$\text{gMAE} = \frac{1}{d} \sum_{i=1}^d \text{MAE}_i = \frac{1}{d} \sum_{i=1}^d \frac{1}{n_{\text{out}}} \sum_{t=1}^{n_{\text{out}}} |r_{t,T_i} - \hat{\mu}_{t,T_i}| \quad \text{and} \quad (3)$$

$$\text{gMSE} = \frac{1}{d} \sum_{i=1}^d \text{MSE}_i = \frac{1}{d} \sum_{i=1}^d \frac{1}{n_{\text{out}}} \sum_{t=1}^{n_{\text{out}}} |r_{t,T_i} - \hat{\mu}_{t,T_i}|^2. \quad (4)$$

In the formula above,  $\hat{\mu}$  are the expected bond returns estimated from the in-sample data and evaluated on the out-of-sample data. In addition to these performance measures, we also report the value of the out-of-sample negative log-likelihood (5).

Table 3 summarizes the out-of-sample performance measures obtained using the FGD procedure and a standard univariate ARMA(1,1) analysis of each bond return series.

Table 3 shows that differences of performance between the expected bond predictions are very small, confirming our visual inspection of Section 3.2. In particular, we find that the average gain obtained when using the FGD procedure instead of a classical univariate analysis is about 0.5% for both the gMAE and the gMSE measures. However, the individual gains range from about -4% to 15%, suggesting that they are relevant for some maturities. The gain in the negative log-likelihood obtained using FGD is more evident: about 7-8%.

It can be difficult to judge whether the FGD procedure yields significant improvements over a classical analysis in the accuracy of the expected bond returns on the basis of the results in Table 3 (i.e. in terms of percentages). It is easier to determine whether one procedure is significantly better than the other by performing some statistical tests on differences of performance. This is done in the next section.

Maturity	MAE		MSE		neg. log-lik.	
	FGD	univariate	FGD	univariate	FGD	univariate
1 month	0.3013	0.3013	0.2106	0.2106		
2 months	0.4108	0.4435	0.3301	0.3798		
3 months	0.3984	0.4107	0.3018	0.3119		
4 months	0.4355	0.4420	0.3550	0.3589		
5 months	0.6534	0.6611	0.7052	0.7164		
6 months	0.8815	0.8896	1.1908	1.2359		
7 months	1.0235	1.0192	1.6640	1.6667		
8 months	1.2521	1.2737	2.4231	2.4659		
9 months	1.4687	1.4987	3.4667	3.5598		
10 months	1.7042	1.6717	4.3430	4.1835		
11 months	1.8756	1.8767	5.1496	5.2362		
12 months	2.0114	1.9926	6.4479	6.3701		
global	1.0347	1.0401	2.2156	2.2247	450.604	487.387

Table 3

Out-of-sample goodness of fit measures for the expected bond returns obtained using the FGD procedure introduced in Section 2 and a classical univariate analysis of the bond return series (univariate). MAE, MSE and neg. log-lik. denote the out-of-sample mean absolute and mean square error of the conditional mean predictions, and the out-of-sample negative log-likelihood of the estimation, respectively. The out-of-sample period goes from January 1995 to December 1999, for a total of 60 monthly observations.

### 3.4 Two tests on differences of performance

We test formally whether differences in the out-of-sample model performances, as highlighted in the last section, are statistically significant. To this purpose, we make use of the  $t$ -type and sign-type performance tests, as proposed by Audrino and Bühlmann (2004) extending previous work done by Diebold and Mariano (1995). In particular, we test for significance of the difference in the (g)MAE and (g)MSE performance measures (1)-(4) of the FGD procedure against standard univariate ARMA(1,1) models.

The tests are defined as follows. Let  $\tilde{U}_t$  be the realized out-of-sample loss associated at time  $t$  with a given model and based on a given loss function  $\tilde{U}$ .

By applying a suitable functional form for  $\tilde{U}$  we have, for instance, that

$$\sum_{t=1}^{n_{\text{out}}} \tilde{U}_{t;\text{model}} = \text{chosen performance measure.} \quad (5)$$

Note that in our case, the “chosen performance measure” in (5) may be one of  $\text{MAE}_i$ ,  $\text{MSE}_i$ , for  $i = 1, \dots, d$ ,  $\text{gMAE}$  or  $\text{gMSE}$ . The realized loss difference at time  $t$  between  $\text{Model}_1$  and  $\text{Model}_2$  is

$$\widehat{D}_t = \tilde{U}_{t;\text{model}_1} - \tilde{U}_{t;\text{model}_2}, \quad t = 1, \dots, n_{\text{out}}. \quad (6)$$

We test the null hypothesis that differences  $\widehat{D}_t$  have mean zero against the one-side alternative of mean less (bigger) than zero, i.e. the estimates from  $\text{model}_1$  ( $\text{model}_2$ ) are better than the ones from  $\text{model}_2$  ( $\text{model}_1$ ). Moreover, we also test the null hypothesis that the frequency of negative differences has mean  $1/2$  against the one-side alternative that their mean is more (less) than  $1/2$ . This allows us to investigate whether there is a systematic difference between the estimates from the two models. Note also that this second type of test is more robust against outliers. For this purpose, we use versions of classical  $t$ - and sign-tests, adapted to the case of dependent observations. The exact definition of the tests is presented in Appendix A.

Results of the tests for the real data example under investigation are summarized in Table 4. Note that results are not reported for the bond return series at 1 month since the predictions are exactly the same using the two competitive approaches. Negative values of  $t$ -type statistics mean that the FGD procedure has to be preferred to the individual  $\text{ARMA}(1,1)$  estimation. The contrary is valid for the sign-type test statistics.

In half of the cases (six out of twelve individual tests), there is at least one statistically significant result (at the 5% confidence level or better) in favor of the FGD procedure, either based on the  $\text{MAE}_i$  or  $\text{MSE}_i$  ( $i = 1, \dots, 12$ )  $t$ -type and sign-type tests. The individual  $\text{ARMA}(1,1)$  predictions are clearly beaten by the FGD ones in the cases of time to maturities 2, 3, 5 and 8 months. By contrast, for no time to maturity the individual  $\text{ARMA}(1,1)$  expected bond return predictions are significantly better (at the 10% confidence level) than the FGD ones. These results are confirmed (although at a lower confidence level) when performing the tests on the whole twelve-dimensional time series. Sign-type test statistics for both the  $\text{gMAE}$  and  $\text{gMSE}$  performance measures are significant at the 10% confidence level, preferring the FGD based estimation to the classical one.

Maturity	Performance measure			
	MAE		MSE	
	t-type	sign-type	t-type	sign-type
1 month	—	—	—	—
2 months	−2.1159	2.1646	−1.9509	2.1646
	(0.0172)	(0.0152)	(0.0255)	(0.0152)
3 months	−2.3506	1.4582	−1.5678	1.4582
	(0.0094)	(0.0724)	(0.0585)	(0.0724)
4 months	−1.6998	−0.6039	−0.8255	−0.6039
	(0.0446)	(0.2729)	(0.2045)	(0.2729)
5 months	−1.6190	1.9789	−0.9795	1.9789
	(0.0527)	(0.0239)	(0.1637)	(0.0239)
6 months	−0.3105	0.4409	−0.8326	0.4409
	(0.3781)	(0.3296)	(0.2025)	(0.3296)
7 months	0.4783	0.7595	−0.1717	0.7595
	(0.3162)	(0.2238)	(0.4318)	(0.2238)
8 months	−1.5838	1.8726	−0.8748	1.8726
	(0.0566)	(0.0306)	(0.1908)	(0.0306)
9 months	−1.0186	0.5969	−0.7216	0.5969
	(0.1542)	(0.2753)	(0.2353)	(0.2753)
10 months	1.2276	−1.1155	0.9588	−1.1155
	(0.1098)	(0.1323)	(0.1688)	(0.1323)
11 months	−1.2296	1.9367	−0.8356	1.9367
	(0.1094)	(0.0264)	(0.2017)	(0.0264)
12 months	0.2532	−0.8973	0.3126	−0.8973
	(0.4001)	(0.1848)	(0.3773)	(0.1848)
Global	−0.9473	1.4213	−0.5096	1.3489
	(0.1717)	(0.0776)	(0.3052)	(0.0887)

Table 4

Tests for a difference in the out-of-sample performance of expected bond return predictions obtained using the FGD procedure (Model<sub>1</sub>) against those from a classical univariate ARMA(1,1) analysis of each bond return series (Model<sub>2</sub>). The table gives the values of the relevant test statistics and the corresponding  $p$ -values (below, in parentheses). Columns 2 and 4 present generalized  $t$ -type tests on the series of differences of performance losses  $\tilde{U}_t$ . Columns 3 and 5 present generalized sign-type tests on the series of differences of performance losses  $\tilde{U}_t$ . Negative values of  $t$ -type statistics mean that Model<sub>1</sub> has to be preferred to Model<sub>2</sub>. The contrary is valid for the sign-type test statistics.

## 4 Conclusions

We proposed a procedure based on functional gradient descent (FGD) to estimate the time-varying dynamics of the multivariate time series of bond returns at different maturities. In contrast to the classical univariate analysis of each individual bond return time series, the FGD approach allows for the incorporation of cross-effects, by modelling the dynamics of the expected bond return vector as a function of past multivariate observations. Thus, the information included in the past realizations of bond returns at nearer maturities (when statistically relevant) can be exploited to improve the accuracy of expected bond return estimates and forecasts.

In our real data investigation on artificial zero coupon US bonds with maturities ranging from one month to one year, we found strong empirical evidence of the higher predictive potential of the FGD procedure. In particular we showed using two different type of statistical tests that expected bond return predictions from the FGD approach are at least in half of the cases significantly more accurate than those from a classical univariate analysis.

The analysis of our study can be extended to long-term bond with time to maturity longer than 1 year. Based on some preliminary findings, we believe that results for long maturities are qualitatively the same as those reported in this study. The FGD methodology is very general and allows for increases of the predictor space in a very simple way. Therefore, some future research on bond return dynamics can be devoted to forecasting expected bond returns by exploiting the additional exogenous information included, for example, in some macroeconomic variables such as indicators for inflation and real activity.

## 5 Acknowledgements

Funded by a grant from the Foundation of Research and Development of the University of Lugano and by a grant from the National Center of Competence in Research “Financial Valuation and Risk Management” (NCCR FINRISK).

## A Appendix: t-type and sign-type tests

Consider differences  $\widehat{D}_t$ ,  $t = 1, \dots, n_{\text{out}}$ , of performance terms. The t-type test statistic in the case of dependent observations introduced in Section 3.4 is

$$\sqrt{n_{\text{out}}} \frac{\overline{D}}{\widehat{\sigma}_{D;\infty}}, \quad \text{where } \overline{D} = \frac{1}{n_{\text{out}}} \sum_{t=1}^{n_{\text{out}}} \widehat{D}_t. \quad (\text{A.1})$$

In (A.1),  $\widehat{\sigma}_{D;\infty}^2 = (2\pi)\widehat{f}_D(0)$ , where  $\widehat{f}_D(0)$  is a smoothed periodogram estimate at frequency zero, based on  $\widehat{D}_1, \dots, \widehat{D}_{n_{\text{out}}}$ ; see, for example, Brockwell and Davis (1991). The motivation for this estimate is based on the assumption that  $\{\widehat{D}_t\}_t$  is stationary (conditional on the training data) and satisfies suitable dependence conditions, e.g. mixing. Then, conditional on the training data,

$$\begin{aligned} \sqrt{n_{\text{out}}}(\overline{D} - E[\widehat{D}_t]) &\implies \mathcal{N}(0, \sigma_{D;\infty}^2) \quad (n_{\text{out}} \rightarrow \infty), \\ \sigma_{D;\infty}^2 &= \sum_{k=-\infty}^{+\infty} \text{Cov}[\widehat{D}_0, \widehat{D}_k] = (2\pi)\widehat{f}_D(0), \end{aligned} \quad (\text{A.2})$$

where  $\widehat{f}_D(0)$  is the spectral density at zero of  $\{\widehat{D}_t\}_t$ .

Thus, using (A.2) for the test statistic in (A.1), and conditional on the training data,

$$\sqrt{n_{\text{out}}} \frac{\overline{D}}{\widehat{\sigma}_{D;\infty}} \implies \mathcal{N}(0, 1) \quad (n_{\text{out}} \rightarrow \infty) \quad (\text{A.3})$$

under the null hypothesis.

Analogously, the version of the sign test in the case of dependent observations introduced in Section 3.4 is based on the frequency of negative differences

$$\widehat{W}_t = I_{\{\widehat{D}_t \leq 0\}}, \quad t = 1, \dots, n_{\text{out}},$$

for the null hypothesis that the frequency of negative differences  $\widehat{W}_t$  has mean  $\frac{1}{2}$  against the alternative of mean greater than  $\frac{1}{2}$ . The test statistic is given by

$$\sqrt{n_{\text{out}}} \frac{\overline{W} - \frac{1}{2}}{\widehat{\sigma}_{W;\infty}}, \quad \text{where } \overline{W} = \frac{1}{n_{\text{out}}} \sum_{t=1}^{n_{\text{out}}} \widehat{W}_t \quad (\text{A.4})$$

and  $\widehat{\sigma}_{W;\infty}^2$  as in (A.1) but based on  $\widehat{W}_1, \dots, \widehat{W}_T$ . As in the derivation of the t-type test above, we have, conditional on the training data,

$$\sqrt{n_{\text{out}}} \frac{\overline{W} - \frac{1}{2}}{\widehat{\sigma}_{W;\infty}} \implies \mathcal{N}(0, 1) \quad (n_{\text{out}} \rightarrow \infty) \quad (\text{A.5})$$

under the null hypothesis.



## References

- Audrino, F., Barone-Adesi, G. and Mira, A., 2005. The Stability of Factor Models of Interest Rates. *Journal of Financial Econometrics* 3, No. 3, 422–441.
- Audrino, F. and Bühlmann, P., 2003. Volatility Estimation with Functional Gradient Descent for Very High-Dimensional Financial Time Series. *Journal of Computational Finance* 6, No. 3, 1–26.
- Audrino, F. and Bühlmann, P., 2004. Synchronizing multivariate financial time series. *Journal of Risk* 6, No. 2, 81–106.
- Ang, A. and Piazzesi, M., 2003. A no-arbitrage vector autoregression of term structure dynamics with macroeconomic and latent variables. *Journal of Monetary Economics* 50, 745–787.
- Babcock, G., 1984. Duration as a Link Between Yield and Value. *Journal of Portfolio Management*, Summer and Fall.
- Bollerslev, T., 1990. Modelling the coherence in short-run nominal exchange rates: a multivariate generalized ARCH model. *The Review of Economics and Statistics* 72, 498–505.
- Brennan, M.J. and Schwartz, E.S., 1982. An equilibrium model of bond pricing and a test of market efficiency. *Journal of Financial Quantitative Analysis*, September, 301–329.
- Brockwell, P.J. and Davis, R.A., 1991. *Time Series: Theory and Methods*. Springer, New York.
- Cochrane, J.H. and Piazzesi, M., 2005. Bond risk premia. *American Economic Review* 95, 138–160.
- Diebold, F.X. and Mariano, R.S., 1995. Comparing predictive accuracy. *Journal of Business and Economic Statistics* 13, 253–263.
- Fama, E.F., 1984. Term premiums in bond returns. *Journal of Financial Economics*, December, 529–546.
- Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. *Annals of Statistics* 29, 1189–1232.
- Friedman, J.H., Hastie, T. and Tibshirani, R., 2000. Additive logistic regression: a statistical view of boosting. *Annals of Statistics* 28, 337–407 (with discussion).
- Ilmanen, A., 1995. Time varying expected bond returns in international bond markets. *Journal of Finance* 50, 481–506.
- Stambaugh, R.F., 1988. The information in forward rates: implications for models of the term structure. *Journal of Financial Economics* 22, 3–25.