# Stock Return Prediction based on a Functional Capital Asset Pricing Model

Ufuk Beyaztas
Department of Statistics
Marmara University

Kaiying Ji
Discipline of Accounting, Governance and Regulation
University of Sydney

Han Lin Shang *
Department of Actuarial Studies and Business Analytics
Macquarie University

Eliza Wu
Discipline of Finance
University of Sydney

**Abstract**

The capital asset pricing model (CAPM) is frequently used to capture a linear relationship between the daily returns of an asset and a market index. We extend this model to an intraday high-frequency setting by proposing a functional CAPM. The functional CAPM is a stylized example of a function-on-function linear regression with a bivariate functional regression coefficient. The two-dimensional regression coefficient measures the cross-covariance between cumulative intraday asset returns and market returns. We apply it to the Standard and Poor's 500 index and its constituent stocks to demonstrate its practicality. We investigate the functional CAPM's in-sample goodness-of-fit and out-of-sample prediction for an asset's cumulative intraday return. The findings suggest that the proposed functional CAPM methods have both superior model goodness-of-fit and forecast accuracy in comparison to the traditional CAPM. In particular, the functional methods produce better model goodness-of-fit and prediction accuracy for those stocks that are traditionally considered less price-efficient or information-opaque.

*Keywords: Cumulative intraday returns; Function-on-function linear regression; Regression coefficient surface; S&P 500 index*
*JEL code: G12; G17*

*Correspondence: Department of Actuarial Studies and Business Analytics, Macquarie University, NSW 2109, Australia; Telephone: +61(2) 9850 4689; Email: hanlin.shang@mq.edu.au

# 1 Introduction

The capital asset pricing model (CAPM) (Sharpe 1964, Linter 1965) has been the cornerstone of theoretical and empirical finance. It is a benchmark model in asset pricing research due to its simplicity and efficiency for measuring the cost of equity. CAPM assumes that securities markets are highly competitive and efficient so that relevant information about companies is quickly and universally distributed and absorbed. It also assumes that these markets are dominated by rational, risk-averse investors, who seek to maximize returns on their investments, that is, they demand higher returns for greater risks. Therefore, in an efficient market without arbitrage opportunities, asset prices take into account all available information and the expected return of a stock is influenced by its correlation with the market index, which is measured by beta.

Since the introduction of CAPM, much attention has been placed on improving its model specifications (see, e.g., Casabona and Vora 1982, Jagannathan and Meier 2002, Bernardo et al. 2007). For example, considerable evidence suggests that elements in the CAPM model, that is, market risk premium, risk-free interest rates, and betas, are time-varying (see, e.g., Bollerslev et al. 1988, Fama and French 1989, D'Souza et al. 1989, Jagannathan and Wang 1996, Ghysels 1998, Wang 2003, Ang and Chen 2003, Zhou and Paseka 2017), and conditional modeling that explicitly allows for temporal variation in the factor loading results in statistically significant and economically meaningful improvements (see, e.g., Ball and Kothari 1989, Bera et al. 1988, Wiggins 1992, Braun et al. 1995, Ellis 1996, Fletcher 2002, Andersen et al. 2003, Bali et al. 2009, 2017, Zhou and Paseka 2017). Bali et al. (2017) argue that generalized autoregressive conditional heteroskedasticity-based time-varying conditional betas help explain cross-sectional variation in expected stock returns.

Building on the conditional CAPM, Bollerslev et al. (2016) use high-frequency-based estimates and propose a pricing framework that uses a continuous beta to reflect smooth intraday co-movements in the market and two rough betas associated with intraday price discontinuities, or jumps, during the active part of the trading day and overnight close-to-open return, respectively. Furthermore, Bollerslev et al. (2022) propose "granular betas" that provide a much more refined look at the inherent dependencies between an asset and a given set of factors.

The literature on time-varying beta presents many models that consider the time points at which close prices are observed but these models largely overlook changes in the intraday price curves, that is, how a financial instrument shifts from time to time. High-frequency financial data, that is, observations on financial instruments taken at a granular scale, such as time-stamped transaction-by-transaction or tick-by-tick data, may better address this deficiency. With the advent of new exchanges, every online transaction is recorded and compiled into databases, such as Island electronic communication networks or data for individual bids to buy and sell.

Recent advances in computer storage and data collection have enabled researchers in finance to record and analyze high-frequency data to understand market microstructure related to price discovery and market efficiency (see, e.g., Campbell et al. 1997, Engle 2000, Andersen 2000, Andersen et al. 2001, 2003, Goodhart and O'Hara 1997, Ghysels 2000, Wood 2000, Gençay et al. 2001, Gouriéroux and Jasiak 2001, Lyons 2001, Andreou and Ghysels 2002, Gençay et al. 2001, Tsay 2010, Papavassiliou 2013, Goldstein et al. 2014, Bollerslev et al. 2022). This stream of research demonstrates that high-frequency-based sampling allows more accurate factor representations and improved asset pricing predictions compared to conventional lower-frequency estimates, resulting in more efficient ex-post mean-variance portfolios (Bollerslev and Zhang 2003, Cenesizoglu et al. 2016, Hollstein et al. 2020). Andersen et al. (2004, p.13) provide further evidence suggesting that high-frequency beta measures are capable of "more clearly highlighting the dynamic evolution of individual security betas" compared to the results obtained from lower frequency daily data. Aue et al. (2012) introduces a modified functional CAPM and sequential monitoring procedures and suggests that the functional data-analytic approach performs better in detecting a structural break, in other words, time variability in the betas.

High-frequency intraday financial data are examples of dense functional data in statistics, represented in graphical form as curves (see, e.g., Andersen et al. 2024). As an integral part of functional data analysis (Ramsay and Silverman 2005, Ferraty and Vieu 2006) and time series analysis (Kokoszka and Reimherr 2017, Peña and Rsay 2021), functional time series consist of random functions observed at a time interval. Functional time series can be grouped into two categories, regardless of whether the continuum is also a time variable. On the one hand, functional time series can arise from measurements obtained by separating an almost continuous time record into consecutive intervals, for example, days, weeks, or years (see, e.g., Hörmann and Kokoszka 2012). We refer to such data structures as *sliced* functional time series, examples of which include the intraday price or volatility curves of a financial stock (see, e.g., Shang 2017, Shang et al. 2019, Andersen et al. 2021). On the other hand, the functional variable can be other continuous variables, such as maturity-specific yield curves (see, e.g., Hays et al. 2012) or near-infrared spectroscopy wavelength (see, e.g., Shang et al. 2022). In either case, the object of interest is a discrete-time series of functions with a continuum (see, e.g., Wang et al. 2008, Kokoszka and Zhang 2012, Kokoszka et al. 2017). In Section 2, we describe our motivating data sets – 5-minute cumulative intraday returns (CIDRs) of the Standard and Poor's 500 (S&P 500) indexes and their constituents. These data sets belong to the first type of functional time series.

If prices are modeled as a univariate time series of discrete observations, the underlying process that generates these observations cannot be fully discovered. The advantages of functional time series include: 1) Thanks to continuity, we can study the temporal correlation between two intraday functional objects. 2) The beta function estimate is a two-dimensional image capturing (cross-) correlation between an asset and its stock index, so we can study the cross-sectional dependence between two random points on a functional object. 3) We handle missing values via interpolation or smoothing techniques. 4) By converting a univariate time series to a time series of functions, we implicitly overcome the "curse of dimensionality" (Donoho 2000), where nonparametric and semiparametric techniques can be implemented (see, e.g., Ferraty and Vieu 2006, Aneiros-Pérez and Vieu 2006).

We propose an extension of a CAPM tailored for high-frequency financial data, termed a functional CAPM. The functional CAPM has recently been considered in a working paper of Pedersen (2022), but we present a novel way of estimating the regression coefficient and investigating the difference between the CAPM and its functional version from an aspect of firm characteristics. It is designed to explain how much variability (i.e., information) in the market cumulative intraday return can explain the variability in an asset. Our functional CAPM can be cast as a function-on-function linear regression, where our objective is to estimate the bivariate functional regression coefficient. The bivariate functional regression coefficient measures a linear relationship between a functional response (i.e., CIDRs of an asset) and a functional predictor (i.e., CIDRs of a market index). Through the functional CAPM, we can predict the in-sample conditional expectation of the CIDRs of an asset with the estimated bivariate regression coefficient function and evaluate the model's goodness of fit via a functional $R^2$. The findings in this study suggest that the proposed functional CAPM methods present superior performance in model goodness-of-fit for those less price-efficient stocks, and better prediction accuracy for information opaque stocks.

Our article is structured as follows. In Section 2, we describe a financial stock market index and its respective constituent stocks. In Section 3, we introduce the functional CAPM to estimate the bivariate regression coefficient function, which captures the linear relationship between the S&P 500 index and its constituent stock. We apply the functional CAPM to model CIDRs by selecting assets representing various asset classes. In Section 4, we display the estimated regression coefficient functions obtained from the functional principal component regression (FPCR), functional partial least squares regression (FPLSR), and penalized function-on-function regression (PFLM). Intraday $R^2$, root mean squared error (RMSE), and root mean squared prediction error (RMSPE) are presented to summarize the goodness-of-fit and out-of-sample prediction for the intraday financial data. The integrated values of these errors can be used to evaluate and compare the overall goodness-of-fit among different constituent stocks in Section 5.1. In Section 5.3, we relate the differences in forecast accuracy between the classical CAPM and functional CAPM with firm characteristics and corporate governance. Section 6 concludes, and offers some ideas on how the methodology presented can be further extended. Details of the FPCR, FPLSR, and PFLM techniques are presented in Appendixes A-C.

# 2 Empirical data analysis

## 2.1 Data and sample selection

The S&P500 (SPX) is a financial stock market index that tracks the performance of around 500 large companies listed on stock exchanges in the United States. The intraday tick history for the S&P500 (SPX) index and its constituent stocks were obtained from the Refinitiv Datascope (`https://select.datascope.refinitiv.com/DataScope/`). We consider daily cross-sectional returns from January 4, 2021, to December 31, 2021. In 2021, there were 252 trading days. Following early work by Bollerslev and Zhang (2003), we downloaded transaction prices for

SPX and each constituent stock at each 5-min interval from 09:30 to 16:00. For each stock, the sampling process yielded 78 data points per day and approximately 20,000 data points in the sampling period for analysis. This study utilized 10 million intraday data points for the purposes of demonstration. The sample size could be expanded without computing power constraints.

We obtained data on the financial balance sheet and the firm Global Industry Classification Standard (GICS) sector from Compustat. Daily data on returns, prices, market capitalizations, and volumes were obtained from the Center for Research in Security Prices (CRSP). Institutional holdings were downloaded from 13F filings. Analyst following and earnings forecast accuracy data were downloaded from the $I/B/E/S$ summary file, and board structure information was obtained from the BoardEx database (https://wrds-www.wharton.upenn.edu/pages/about/data-vendors/boardex/). Winsorization was not performed since dealing with extreme observations is a part of the functional modeling.

**Table 1:** *Summary statistics of the S&P 500 index and firm characteristics.*

| Variables | Min | 1st Quantile | Median | Mean | 3rd Quantile | Max |
|---|---|---|---|---|---|---|
| SPX intraday price | 3667 | 4074 | 4298 | 4274 | 4488 | 4808 |
| $\ln MC_{i,t}$ | 7.889 | 9.843 | 10.434 | 10.555 | 11.139 | 14.659 |
| $\ln P_{i,t}$ | 1.778 | 4.164 | 4.781 | 4.799 | 5.436 | 8.112 |
| $LEV_{i,t}$ | 0.102 | 0.505 | 0.653 | 0.640 | 0.774 | 0.990 |
| $VOL_{i,t}$ (million) | 2.191 | 20.075 | 40.540 | 91.834 | 89.521 | 1900.558 |
| $ILLIQ_{i,t}$ ($10^{-9}$) | 0.001 | 0.031 | 0.063 | 0.084 | 0.108 | 1.872 |
| $BidAskSpread_{i,t}$ | -8.452 | -7.991 | -7.806 | -7.758 | -7.622 | -4.361 |
| $Coverage_{i,t}$ | 0.2624 | 2.5871 | 2.8502 | 2.8059 | 3.1061 | 3.9269 |
| $Accuracy_{i,t}$ | 0.0000 | 0.0314 | 0.0957 | 0.3114 | 0.2962 | 6.0809 |
| $InstoHold_{i,t}$ | 0.002 | 0.728 | 0.829 | 0.846 | 0.907 | 3.763 |
| $Independent_{i,t}$ | 0.5556 | 0.8333 | 0.9000 | 0.8723 | 0.9167 | 1.0000 |
| *$Duality_{i,t}$ | 0.0000 | 1.0000 | 1.0000 | 0.8187 | 1.0000 | 1.0000 |

* Binary variable, there are 411 ones and 91 zeros.

Table 1 presents summary statistics for the S&P 500 mid-point 5-minute intraday prices observed over 252 days. We also show the summary statistics of the firm characteristics; detailed variable definitions are presented in Section 5.3.

## 2.2 Cumulative intraday returns of S&P 500

We considered five-minute resolution data, 78 data points, covering the period from 9:30 to 16:00 Eastern standard time. For each asset, intraday 5-minute close price data, $P_t(u_i)$, are available on each trading day, which we used to construct a sequence of CIDRs (see also Rice et al. 2020)

$$\mathcal{X}_t^j(u_i) = 100 \times [\ln P_t^j(u_i) - \ln P_t^j(u_1)], \qquad i = 2, 3, \dots, 78 \tag{1}$$

where $j$ denotes an asset in the S&P 500 index, $i$ denotes the $i^{\text{th}}$ intraday period, $t$ denotes a given trading day and $\ln(\cdot)$ denotes a natural logarithm. From (1), we take the inverse transformation to obtain the 5-minute intraday price

$$P_t^j(u_i) = \exp^{\frac{\mathcal{X}_t(u_i)}{100}} \times P_t^j(u_1),$$

where $P_t^j(u_1)$ denotes the beginning close price on day $t$.

In a given day $t$, we observe CIDRs of the S&P 500 index and its constituent stocks over the intraday period. Let $\mathcal{X}_t(u)$ be the functional predictor, consisting of the CIDR of a market index. Let $\mathcal{Y}_t^j(v)$ be the functional response, consisting of the CIDR of an index's constituent stock. Let $\boldsymbol{\mathcal{Y}}^j(v) = [\mathcal{Y}_1^j(v), \mathcal{Y}_2^j(v), \dots, \mathcal{Y}_n^j(v)]^\top$ and $\boldsymbol{\mathcal{X}}(u) = [\mathcal{X}_1(u), \mathcal{X}_2(u), \dots, \mathcal{X}_n(u)]^\top$ be two functional time series of response and predictor, respectively. Let $\boldsymbol{\mathcal{Y}}^{j,c}(v) = \boldsymbol{\mathcal{Y}}^j(v) - \boldsymbol{R}^f$ and $\boldsymbol{\mathcal{X}}^c(u) = \boldsymbol{\mathcal{X}}(u) - \boldsymbol{R}^f$ denote excess intraday returns of an asset and a stock, where $\boldsymbol{R}^f = [R_1^f, R_2^f, \dots, R_n^f]^\top$. $R_t^f$ can be computed by dividing the daily treasury par yield curve rate at one-year maturity by $(251 \times 78)$ intraday trading time over 251 trading days.

# 3 Functional Capital Asset Pricing Model (CAPM)

Before presenting the proposed functional CAPM, we provide an intuitive understanding of how the curves are obtained from discrete data.

## 3.1 From discrete data points to curves

With high-frequency financial data, the data points are observed discretely. For our intraday financial data, they are observed densely at an equally-spaced grid, that is five-minute time interval. In practice, a basis function expansion can be utilized to convert the discrete data points into a continuous function. Because of this, an advantage of our functional data-analytic approach is that it can address non-synchronicities, that is, time intervals of irregular lengths, in asset returns when measured at high frequencies (Dimson 1979, Lewellen and Nagel 2006, Gilbert et al. 2014, Boguth et al. 2016).

Among all possible basis functions, those widely used are polynomial basis functions (which are constructed from the monomials $\phi_k(u) = u^{k-1}$), Bernstein polynomial basis functions (which are constructed from $1, 1-u, u, (1-u)^2, 2u(1-u), u^2, \dots$), Fourier basis functions (which are constructed from $1, \sin(\omega u), \cos(\omega u), \sin(2\omega u), \cos(2\omega u), \dots$), radial basis functions, wavelet basis functions, spline basis functions, and orthogonal basis functions. Our functional predictor and response are approximated by 20 $B$-spline basis functions:

$$\mathcal{Y}_t^{j,c}(v) = \sum_{k=1}^{20} \widehat{z}_{t,k}^j \widehat{\pi}_k^j(v) = \widehat{\boldsymbol{Z}}_t^j \widehat{\boldsymbol{\Pi}}^j(v), \tag{2}$$

$$\mathcal{X}_t^c(u) = \sum_{m=1}^{20} \widehat{a}_m \widehat{\gamma}_m(u) = \widehat{\boldsymbol{A}}_t \widehat{\boldsymbol{\Gamma}}(u), \tag{3}$$

where $\widehat{\boldsymbol{\Pi}}^j(v)$ and $\widehat{\boldsymbol{\Gamma}}(u)$ are the bases, and $\widehat{\boldsymbol{Z}}$ and $\widehat{\boldsymbol{A}}$ are the corresponding coefficient matrices.

This *pre-smoothing* step allows us to mitigate the curse of dimensionality by choosing to work in a square-integrable function space. From these discrete data points, a continuous curve can be constructed from $[\mathcal{X}_t(u_2), \dots, \mathcal{X}_t(u_{78})]$ representing six and half hours of trading at the New York Stock Exchange on a trading day.

A univariate time series of $19,327$ discrete returns was converted into $n = 251$ days of CIDR curves. In Figure 1, we present CIDRs for the S&P 500 and BlackRock Inc. Using a functional KPSS test of Horváth et al. (2014), both CIDR series are stationary with $p$-values of 0.669 and 0.546, respectively.

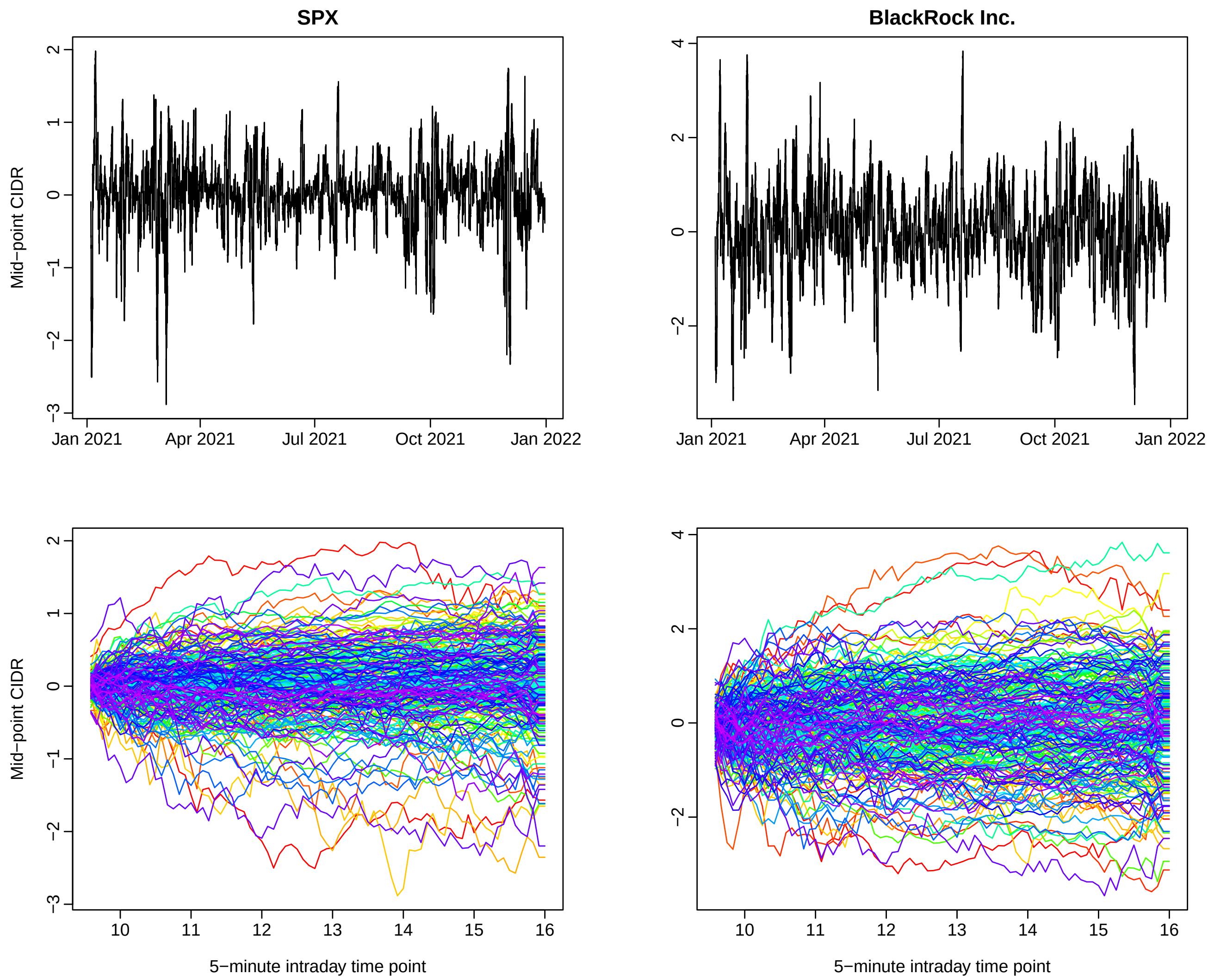


**Figure 1:** *Univariate and functional time series plots of the CIDRs of the SPX index and BlackRock Inc.*

## 3.2 Functional CAPM

The classic CAPM can be expressed as:

$$R_i = R_f + \beta_i(R_m - R_f) + \vartheta_i,$$

where $R_i$ and $R_m$ denote daily asset and market returns, $R_f$ denotes a risk-free rate of interest, $\beta_i$ is a real-valued slope parameter associated with the asset, and $\vartheta_i$ denotes an error term with a mean of zero and finite variance.

The functional CAPM captures a linear relationship between a centered functional predictor and a centered functional response through an unknown bivariate regression coefficient function, known as beta surface. The functional CAPM can be expressed as

$$\mathcal{Y}_t^j(v) = R_t^f + \int_{\mathcal{I}} \beta^j(u,v)[\mathcal{X}_t(u) - R_t^f]du + \varepsilon_t^j(v), \tag{4}$$

where $R_t^f$ denotes the intraday risk-free rate of interest, $\mathcal{X}_t^c(u) = [\mathcal{X}_t(u) - R_t^f]$ denotes the intraday market risk premium, $\beta^j(u,v)$ is the bivariate regression coefficient function associated with the $j^{\text{th}}$ constituent stock, $\varepsilon_t^j(v)$ denotes an independent and identically distributed (i.i.d.) random error function, and $u, v \in \mathcal{I}$ denote a function support range of $\mathcal{I}$ (i.e., intraday trading period between 9:30 and 16:00 Eastern Time).

The functional CAPM is a special case of the concurrent function-on-function linear regression (see, e.g., Ramsay and Dalzell 1991). The direct estimation of the regression coefficient in the functional CAPM is an ill-posed problem due to the singularity and curse of dimensionality. Since the functional predictor and response belong to infinite-dimensional function space, we consider projecting the functional predictor and response onto orthonormal and $B$-spline bases.

### 3.3 Estimation of regression coefficient function

Within the framework of functional CAPM in (4), it is crucial to accurately estimate the bivariate regression coefficient function $\beta^j(u,v)$ from a finite sample. Toward this end, we explore three distinct methodologies: FPCR, FPLSR, and PFLM. The PFLM relies on general basis expansion techniques such as $B$-spline basis functions. The FPCR and FPLSR adopt a data-driven dimension reduction paradigm, and they entail projecting infinite-dimensional curves onto finite-dimensional spaces of orthonormal bases. In contrast, PFLM may necessitate a larger number of basis functions to approximate the functional regression coefficient, which can potentially lead to model overfitting and reduced prediction accuracy.

In FPCR, the components used for approximating the bivariate regression coefficient function $\beta^j(u,v)$ are derived from the covariance among functional predictors alone. A few of the leading principal components generally comprise most of the variance between the functional predictors. These latent components may not necessarily be important to prediction accuracy (see, for example Delaigle and Hall 2012). FPLSR addresses this issue by leveraging both response and predictors during the extraction of latent components, thereby capturing more information with comparably fewer terms. Additionally, studies have demonstrated that FPLSR offers more accurate parameter function estimation compared to FPCR, albeit with greater dimension reduction (see, e.g., Aguilera et al. 2010, Beyaztas and Shang 2022, Saricam et al. 2022).

In both FPCR and FPLSR, the data-driven orthogonal bases may lack of smoothness to the functional parameter. However, this can lead to significant under-smoothing if the functional parameter exhibits considerably more smoothness than the higher-order FPLSR and FPCR scores (see, e.g., Ivanescu et al. 2015, Beyaztas et al. 2024). Consequently, including one or two additional latent components may alter the shape and interpretation of the functional parameter (see also Crainiceanu et al. 2009). Contrarily, in PFLM, the penalty term applied during the estimation phase imposes a specific level of smoothness on the parameter estimate, thereby preventing overfitting.

The comprehensive details regarding the methodologies of FPCR, FPLSR, and PFLM for estimating the bivariate regression coefficient function $\beta^j(u,v)$ in (4) have been deferred to Appendixes A-C.

# 4 Illustration of the functional CAPM

We apply the FPCR and FPLSR to estimate the bivariate regression coefficient function in the functional CAPM. We consider the CIDRs of the S&P 500 index as the functional predictor and of BlackRock Inc stock as the functional response for the purposes of demonstration. In Figure 2, we display the estimated regression coefficient functions obtained from the two function-on-function regression models. The regression coefficient function, estimated by the FPCR, shows intense activity between 10:00 and 11:30 and between 13:30 and 14:30. In contrast, the regression coefficient function estimated by the FPLSR demonstrates intense activity between 10:00 and 11:30 and between 15:00 and 16:00.

The difference between the estimated beta surfaces is because of the characteristics of the basis functions. In the FPCR, the basis functions obtained from the functional predictors are orthonormal, and there are no off-diagonal elements. Thus, its estimated beta surface is comparably smooth. In contrast, the FPLSR basis functions obtained from the functional predictors and responses are not orthogonal. The inverse square root of the inner product

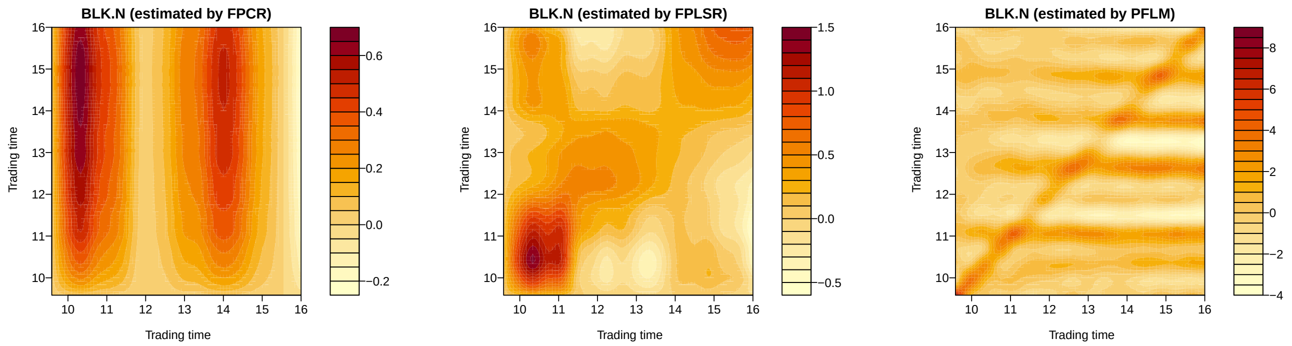


**Figure 2:** *Plots of the estimated bivariate regression coefficient functions $\widehat{\beta}^j(u,v)$ when the CIDR curves of a market index are the realizations of the functional predictor and the CIDRs of an asset are the functional response realizations. The bivariate regression coefficient function is estimated via the FPCR, FPLSR, or PFLM on the left, bottom, and right panels, respectively.*

matrices plays an essential role in computing the regression coefficient surface of the FPLSR. In the FPLSR basis functions, the off-diagonal elements of the inner product matrices capture the linear dependence between the functional predictor and response at different intraday periods. As a result, the FPLSR can show more local features than the FPCR.

# 5 Estimation accuracy of the response

Using BlackRock Inc.'s CIDRs, we evaluate and compare the model performance to traditional CAPM, FPCR, and FPLSR. We firstly measure the in-sample goodness-of-fit and estimation accuracy of FPCR and FPLSR by computing $R^2$ and root-mean-square error (RMSE) and then compute root-mean-square percentage error (RMSPE) to measure one-step-ahead out-of-sample prediction accuracy. Further, we investigate whether these performance measures are impacted by firm characteristics, such as industry sector, firm size, leverage, liquidity, and valuation uncertainty.

## 5.1 In-sample goodness-of-fit

While the estimated regression coefficient functions differ, we compute an intraday version of $R^2$ and RMSE as two in-sample goodness-of-fit criteria. The intraday $R^2$ and RMSE criteria extend from conventional linear models, defined as

$$R^2(v) = 1 - \frac{\sum_{t=1}^{n}[\mathcal{Y}_t^j(v) - \widehat{\mathcal{Y}}_t^j(v)]^2}{\sum_{t=1}^{n}[\mathcal{Y}_t^j(v) - \overline{\mathcal{Y}}^j(v)]^2}, \qquad v \in \mathcal{I},$$

$$\text{RMSE}(v) = \sqrt{\frac{1}{n}\sum_{t=1}^{n}[\mathcal{Y}_t^j(v) - \widehat{\mathcal{Y}}_t^j(v)]^2}$$

$$= \sqrt{\frac{1}{n}[1 - R^2(v)]\sum_{t=1}^{n}[\mathcal{Y}_t^j(v) - \overline{\mathcal{Y}}^j(v)]^2},$$

where $\widehat{\mathcal{Y}}_t^j(v)$ represents the fitted values obtained from the functional CAPM, using the estimated regression coefficient function.

In Figure 3, we compute the intraday $R^2$ and RMSE between the observed and fitted CIDRs for Goldman Sachs. The FPLSR consistently produces larger $R^2$ and smaller RMSE values than those obtained from the FPCR. Thus, the off-diagonal elements of the inner product matrices help estimate the response. It is evident that $R^2$ for FPLSR peaks around 11am, on average, and remains relatively still over the trading day. However, on the other hand, $R^2$ for FPCR peaks much later, around 1pm, and drops in the last two trading hours. It indicates that FPLSR generally offers better in-sample model goodness-of-fit for longer trading hours. A possible reason is that FPLSR extracts latent components by maximizing covariance between functional predictor and response variables.

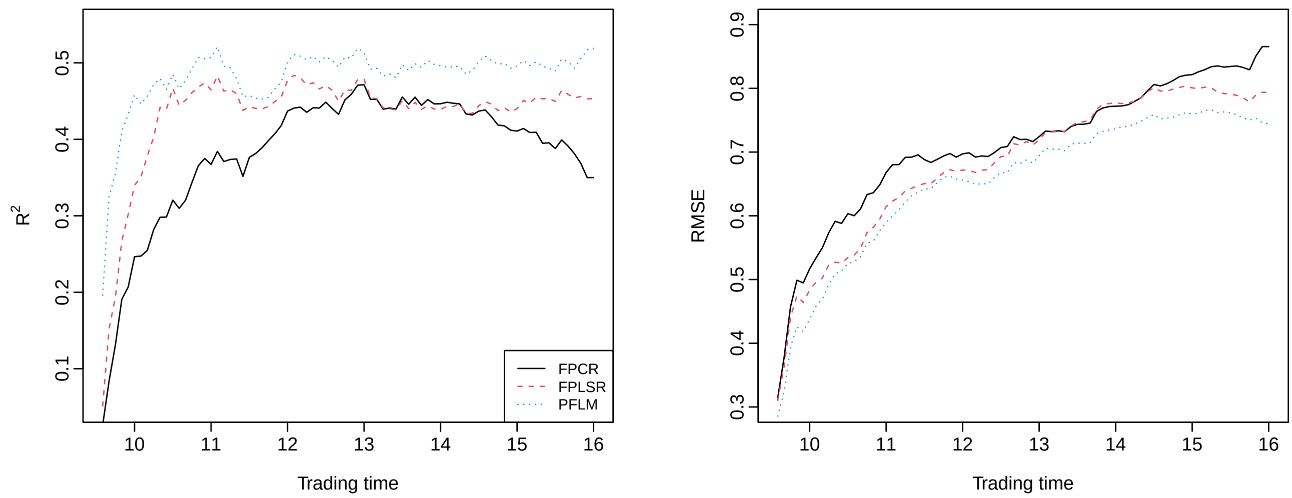


**Figure 3:** *Intraday $R^2$ and in-sample RMSE values between the observed and fitted CIDRs for BlackRock Inc.*

Averaged $R^2$ and RMSE are useful if one requires single numerical measures of fit. They can be expressed as

$$
R^2 = \int_{\mathcal{I}} R^2(v)dv = \frac{1}{78}\sum_{i=1}^{78} R^2(v_i)
$$
$$
\text{RMSE} = \int_{\mathcal{I}} \text{RMSE}(v)dv.
$$

The larger the $R^2$ value, the better the functional CAPM can capture the overall linear relationship between predictor and response. The smaller RMSE often reflects the larger $R^2$ value.

We employ a linearity test between the decentered functional response and decentered functional predictor as proposed by Garcia-Portugués et al. (2021) to examine the linear relationship between the intraday returns of an asset, denoted as $\boldsymbol{\mathcal{Y}}^{j,c}(v)$, and the stock $\boldsymbol{\mathcal{X}}^{c}(u)$. The null hypothesis is formulated as follows:

$$
H_0 : m_\beta \in \mathcal{L} = \left\{ m_\beta(\boldsymbol{\mathcal{X}}^c)(v) = \int_{\mathcal{I}} \beta^j(u,v)\boldsymbol{\mathcal{X}}^c(u)du : \quad \beta^j \in \mathcal{L}^2[u \times v] \right\},
$$

where $m_\beta$ is a Hilbert-Schmidt operator between $\mathcal{L}^2$ spaces that can be represented integrally using a bivariate kernel $\beta$, that is $\boldsymbol{\mathcal{Y}}^{j,c} = m_\beta(\boldsymbol{\mathcal{X}}^c) + \epsilon^j$. This test defines $H_0$ through an integral regression operator obtained by projecting the functional covariate and response into finite-dimensional functional directions. The Cramér–von Mises statistic, which integrates these directions, measures the deviation of the empirical process from its expected zero mean. The statistic is calibrated using an efficient wild bootstrap on the residuals.

The results of the linearity test, conducted at a 0.05 significance level, reveal a clear linear relationship for 449 firms between $\boldsymbol{\mathcal{Y}}^{j,c}(v)$ and $\boldsymbol{\mathcal{X}}^{c}(u)$. This indicates that the intraday returns of these firms' assets can be adequately explained by a linear functional CAPM. Conversely, for the remaining 58 firms, we observe $p$-values close to zero (i.e., $p$-value $< 0.05$), suggesting a potential nonlinear relationship between these firms' intraday returns and the stock. This implies that a more complex model may be required to accurately capture the relationship for this subset of firms.

## 5.2 Out-of-sample prediction accuracy

Apart from in-sample estimation accuracy, we compare one-step-ahead out-of-sample prediction accuracy among the methods by computing the RMSPE. The total RMSPE is defined as follows:

$$\text{RMSPE} = \int_{\mathcal{I}} \text{RMSPE}(v)dv,$$

where

$$\text{RMSPE}(v) = \sqrt{\frac{1}{n_{\text{test}}}\sum_{\zeta=1}^{n_{\text{test}}}[\mathcal{Y}_{\zeta}^{j}(v) - \widehat{\mathcal{Y}}_{\zeta}^{j}(v)]^2},$$

where $\widehat{\mathcal{Y}}_{\zeta}^{j}(v)$ is the predicted response function.

We consider an expanding-window approach to compare the out-of-sample prediction performance of the methods. For both datasets, we divide $n = 251$ into two parts; a training sample consisting of first $n_{\text{train}} = 200$ curves and a test sample consisting of the remaining $n_{\text{test}} = 51$ curves. Using the entire observations in the training sets, we obtain the one-step-ahead forecast for the 201$^{\text{st}}$ curve. We increase the training sample by one day and obtain the forecast for the 202$^{\text{st}}$ curve. This procedure is repeated until the training samples cover the entire dataset.

## 5.3 Firm characteristics and model performance

We compare the model performance measures among the four methods and investigate whether firm characteristics impact these measures. For the traditional CAPM, the input is the daily closing price and beta is a real value. In contrast, the parameters in the FPCR, FPLSR and PFLM methods are estimated using the intraday data and beta is a two-dimensional surface. Intuitively, the outcomes of the four methods are not directly comparable due to the nature of the data. Hence, we have performed an integration for the functional methods so that the mean and median of the four performance measures for the traditional CAPM, FPCR, FPLSR and PFLM are comparable.

Table 2 displays the in-sample estimations using the total $R^2$ and RMSE values between the holdout and estimated responses. For the stocks in the S&P 500 index, we summarize the results by their GICS sectors. The results presented in Table 2 indicate that the model goodness-of-fit (measured by $R^2$) varies greatly amongst different industries. However, the rankings of the model fitness produced by the four methods are consistent. The traditional CAPM produces higher $R^2$ values than the FPCR and the FPLSR methods but under-performs the PFLM method. We find that Information Technology, Financials, and Industrials are the sectors with the highest $R^2$ values, while Utilities and Consumer Staples are the sectors with the lowest goodness of fit. This is consistent with existing literature that finds that stocks in some industries have less noisy prices and reflect relatively more firm-specific private information. Hence, this is reflected in the better goodness of fit for firms in these particular industries (see, e.g., Brogaard et al. 2022).

When investigating in-sample fitting (measured by RMSE), the PFLM method again outperforms the other comparable methods with the smallest mean and median estimation errors. The sectors with better in-sample goodness of fit are Consumer Staples and Utilities, while the Energy and Consumer Discretionary sectors have relatively larger in-sample prediction errors. Overall, the PFLM method presents superiority with in-sample estimation accuracy.

**Table 2:** *Computed total $R^2$ and total RMSE values between the traditional daily CAPM and functional CAPM sorted by the 11 sectors. N = 488.*

| | Total $R^2$ | | | | Total RMSE | | | |
|---|---|---|---|---|---|---|---|---|
| Sector | CAPM | FPCR | FPLSR | PFLM | CAPM | FPCR | FPLSR | PFLM |
| Energy (21) | 0.160 | 0.093 | 0.118 | 0.172 | 2.400 | 1.591 | 1.570 | 1.525 |
| Materials (27) | 0.207 | 0.139 | 0.180 | 0.222 | 1.813 | 1.117 | 1.092 | 1.066 |
| Industrials (69) | 0.234 | 0.154 | 0.205 | 0.246 | 1.897 | 0.993 | 0.963 | 0.939 |
| Consumer Discretionary (59) | 0.202 | 0.136 | 0.180 | 0.220 | 2.028 | 1.292 | 1.258 | 1.228 |
| Consumer Staples (31) | 0.105 | 0.060 | 0.111 | 0.145 | 1.223 | 0.843 | 0.820 | 0.804 |
| Health Care (61) | 0.149 | 0.111 | 0.148 | 0.189 | 1.678 | 1.042 | 1.020 | 0.997 |
| Financials (66) | 0.247 | 0.186 | 0.231 | 0.268 | 2.013 | 1.001 | 0.973 | 0.952 |
| Information Technology (76) | 0.254 | 0.192 | 0.253 | 0.300 | 2.454 | 1.090 | 1.047 | 1.015 |
| Communication Services (20) | 0.141 | 0.115 | 0.157 | 0.202 | 2.159 | 1.140 | 1.109 | 1.082 |
| Utilities (28) | 0.079 | 0.044 | 0.087 | 0.131 | 1.211 | 0.852 | 0.832 | 0.812 |
| Real Estate (30) | 0.160 | 0.070 | 0.121 | 0.151 | 1.630 | 0.987 | 0.960 | 0.944 |
| Mean | 0.176 | 0.118 | 0.163 | 0.204 | 1.864 | 1.086 | 1.059 | 1.033 |
| Median | 0.160 | 0.115 | 0.157 | 0.202 | 1.897 | 1.042 | 1.020 | 0.997 |

Table 3 displays the model comparison on out-of-sample forecast accuracy using the total RMSPE values. It suggests that FPCR is the best performer with the smallest RMSPE and PFLM is ranked the second. The traditional CAPM is the worst performer with the highest forecast errors. This observation is consistent regardless of the stock's GICS sector. The results further confirm the wider applicability of the FPCR for out-of-sample forecasting (see, e.g., Wang and

Cao 2023, for a comparison between the FPCR and FPLSR).

In summary, the PFLM method performs best in model goodness-of-fit and in-sample estimation accuracy, while FPCR performs best in out-of-sample forecast. The above mentioned functional methods presents superior performance than the traditional CAPM. This is because the traditional CAPM uses only one observation for a trading day (closing price recorded at 16:00), that is, it is highly aggregated and, consequently, ignores the rich information embedded in the intraday fluctuation. On the other hand, the functional methods utilize intraday data (78 data points) for a single trading day. The functional methods utilize $B$-spline basis functions to reconstruct smooth curves from discrete data. With an adequate amount of smoothing, it mitigates the data measurement error that is often ignored by the traditional CAPM.

**Table 3:** *Computed total RMSPE values between the traditional daily CAPM and functional CAPM sorted by the 11 sectors. N = 488.*

| | Total RMSPE | | | |
|---|---|---|---|---|
| Sector | CAPM | FPCR | FPLSR | PFLM |
| Energy (21) | 1.966 | 1.228 | 1.466 | 1.265 |
| Materials (27) | 1.536 | 0.917 | 1.130 | 0.938 |
| Industrials (69) | 2.102 | 0.867 | 1.230 | 0.878 |
| Consumer Discretionary (59) | 2.332 | 1.134 | 1.572 | 1.139 |
| Consumer Staples (31) | 1.287 | 0.732 | 1.186 | 0.753 |
| Health Care (61) | 1.855 | 0.918 | 1.225 | 0.931 |
| Financials (66) | 2.768 | 0.832 | 1.097 | 0.837 |
| Information Technology (76) | 3.485 | 0.982 | 1.466 | 0.985 |
| Communication Services (20) | 2.911 | 1.026 | 1.459 | 1.034 |
| Utilities (28) | 1.135 | 0.669 | 1.063 | 0.669 |
| Real Estate (30) | 2.064 | 0.855 | 1.333 | 0.875 |
| Mean | 2.131 | 0.924 | 1.293 | 0.937 |
| Median | 2.064 | 0.917 | 1.230 | 0.931 |

We next investigate whether model performance varies with firm characteristics. We sort each of the S&P 500 firms by firm characteristics, then adopt the two-sample $t$ tests to investigate whether model performance (measured by $R^2$, RMSE, and RMSPE) is significantly different between the top and bottom decile sub-sample groups. We follow Brogaard et al. (2022) and investigate the following firm characteristics: log market capitalization ($\ln MC_{i,t}$), log stock price ($\ln P_{i,t}$), and leverage ($LEV_{i,t}$). We follow Kumar (2009) and adopt volume turnover ($VOL_{i,t}$) to proxy valuation uncertainty, which is measured as the ratio of the number of shares traded in a month and the number of shares outstanding. Next, we include two measures of illiquidity and trading costs to capture limits to arbitrage. The first is the illiquidity measure ($ILLIQ_{i,t}$) in Amihud (2002) and Bali et al. (2009), which is defined as the average ratio of the daily absolute return to the (dollar) trading volume on that day. The second is the stock's average effective bid-ask spread ($BidAskSpread_{i,t}$), which is measured by the natural logarithm of the average daily effective spread.

Table 4 presents the results of mean differences for model performance when comparing firms with different characteristics. For model goodness-of-fit, as measured by $R^2$, all four methods suggest significantly superior performance for larger ($\ln MC_{i,t}$) and higher-priced ($\ln P_{i,t}$) stocks. This is consistent with prior studies in the literature showing that larger firms are more transparent and liquid, and are therefore more efficiently priced (Brogaard et al. 2022). The three functional methods yield significantly better model goodness-of-fit for stocks with higher illiquidity ($ILLIQ_{i,t}$), and FPCR and PFLM perform better for firms with higher bid-ask spread ($BidAskSpread_{i,t}$). It indicates that for less liquid stocks with greater trading costs, the functional methods provide superior model goodness-of-fit when compared to the traditional CAPM.

When measuring in-sample fitting, all four methods produce significantly smaller RMSE for stocks with greater market capitalization ($\ln MC_{i,t}$), higher illiquidity ($ILLIQ_{i,t}$), and greater trading costs ($BidAskSpread_{i,t}$). It is also evident that the three functional methods perform significantly better for higher-priced stocks ($\ln P_{i,t}$) and stocks with lower monthly volume turnover ($VOL_{i,t}$), while the performance difference for the traditional CAPM was not observed for such stocks. The test results for the out-of-sample forecast accuracy are similar. The RMSPE is significantly smaller for for stocks with greater market capitalization ($\ln MC_{i,t}$) and greater trading costs ($BidAskSpread_{i,t}$). Similarly, the three functional methods performs significantly better for firms with higher illiquidity ($ILLIQ_{i,t}$), while the performance difference for the traditional CAPM was not observed.

**Table 4:** *The impact of firm characteristics on model performance.*

| Firm characteristic | | $R^2$ CAPM | FPCR | FPLSR | PFLM | RMSE CAPM | FPCR | FPLSR | PFLM | RMSPE CAPM | FPCR | FPLSR | PFLM |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\ln MC_{i,t}$ | High | 0.227 | 0.167 | 0.212 | 0.260 | 1.494 | 0.869 | 0.845 | 0.819 | 1.554 | 0.800 | 1.051 | 0.798 |

Continued on next page

| Firm characteristic | | $R^2$ CAPM | FPCR | FPLSR | PFLM | RMSE CAPM | FPCR | FPLSR | PFLM | RMSPE CAPM | FPCR | FPLSR | PFLM |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Low | 0.152 | 0.108 | 0.152 | 0.191 | 2.324 | 1.324 | 1.289 | 1.261 | 2.881 | 1.118 | 1.644 | 1.137 |
| | t.statistic | 3.208 | 3.676 | 3.323 | 3.901 | -3.059 | -6.246 | -6.248 | -6.383 | -2.699 | -4.555 | -5.239 | -4.911 |
| | p.value | 0.002 | 0.001 | 0.001 | 0.000 | 0.003 | 0.000 | 0.000 | 0.000 | 0.010 | 0.000 | 0.000 | 0.000 |
| | Sig. | *** | *** | *** | *** | ** | *** | *** | *** | *** | *** | *** | *** |
| $\ln P_{i,t}$ | High | 0.235 | 0.173 | 0.220 | 0.268 | 2.173 | 1.064 | 1.032 | 1.002 | 3.347 | 0.962 | 1.371 | 0.960 |
| | Low | 0.160 | 0.108 | 0.148 | 0.188 | 2.094 | 1.344 | 1.314 | 1.284 | 2.145 | 1.087 | 1.530 | 1.110 |
| | t.statistic | 3.361 | 4.576 | 4.470 | 5.066 | 0.269 | -3.903 | -3.990 | -4.085 | 1.762 | -1.949 | -1.511 | -2.343 |
| | p.value | 0.001 | 0.000 | 0.000 | 0.000 | 0.789 | 0.000 | 0.000 | 0.000 | 0.083 | 0.054 | 0.134 | 0.021 |
| | Sig. | *** | *** | *** | *** | | *** | *** | *** | | | | * |
| $LEV_{i,t}$ | High | 0.192 | 0.135 | 0.182 | 0.220 | 2.091 | 1.081 | 1.052 | 1.028 | 2.939 | 0.958 | 1.339 | 0.970 |
| | Low | 0.221 | 0.160 | 0.214 | 0.258 | 2.591 | 1.129 | 1.093 | 1.061 | 3.518 | 0.993 | 1.411 | 0.996 |
| | t.statistic | -0.331 | -1.093 | -0.722 | -0.943 | 0.665 | 0.616 | 0.588 | 0.601 | 0.347 | -0.146 | 1.429 | 0.037 |
| | p.value | 0.742 | 0.278 | 0.472 | 0.348 | 0.508 | 0.540 | 0.558 | 0.550 | 0.730 | 0.884 | 0.157 | 0.971 |
| | Sig. | | | | | | | | | | | | |
| $VOL_{i,t}$ | High | 0.164 | 0.110 | 0.159 | 0.201 | 1.726 | 1.048 | 1.017 | 0.992 | 1.998 | 0.890 | 1.302 | 0.901 |
| | Low | 0.187 | 0.136 | 0.182 | 0.226 | 2.352 | 1.035 | 1.006 | 0.979 | 3.288 | 0.907 | 1.294 | 0.920 |
| | t.statistic | -0.932 | -0.799 | -1.153 | -0.880 | -0.107 | 3.236 | 3.273 | 3.254 | -2.423 | 2.849 | 1.227 | 2.943 |
| | p.value | 0.354 | 0.426 | 0.252 | 0.381 | 0.915 | 0.002 | 0.002 | 0.002 | 0.019 | 0.006 | 0.224 | 0.005 |
| | Sig. | | | | | | *** | *** | *** | * | ** | | ** |
| $ILLIQ_{i,t}$ | High | 0.185 | 0.126 | 0.169 | 0.212 | 1.848 | 1.047 | 1.021 | 0.995 | 1.879 | 0.868 | 1.208 | 0.881 |
| | Low | 0.152 | 0.113 | 0.161 | 0.202 | 2.286 | 1.140 | 1.107 | 1.081 | 3.222 | 0.999 | 1.417 | 1.010 |
| | t.statistic | -1.654 | -2.558 | -2.355 | -3.005 | 2.151 | 4.807 | 4.786 | 5.011 | 1.940 | 2.402 | 3.238 | 2.795 |
| | p.value | 0.102 | 0.012 | 0.021 | 0.004 | 0.035 | 0.000 | 0.000 | 0.000 | 0.058 | 0.018 | 0.002 | 0.006 |
| | Sig. | | * | * | ** | ** | *** | *** | *** | | * | ** | ** |
| $BidAskSpread_{i,t}$ | High | 0.216 | 0.135 | 0.187 | 0.226 | 1.539 | 0.983 | 0.952 | 0.930 | 1.796 | 0.850 | 1.259 | 0.864 |
| | Low | 0.193 | 0.133 | 0.177 | 0.220 | 1.719 | 1.046 | 1.018 | 0.992 | 2.119 | 0.895 | 1.262 | 0.901 |
| | t.statistic | 0.938 | 2.400 | 1.679 | 2.040 | 4.499 | 6.562 | 6.483 | 6.471 | 3.170 | 6.317 | 5.782 | 6.140 |
| | p.value | 0.351 | 0.018 | 0.096 | 0.044 | 0.000 | 0.000 | 0.000 | 0.000 | 0.003 | 0.000 | 0.000 | 0.000 |
| | Sig. | | * | | * | *** | *** | *** | *** | ** | *** | *** | *** |

*** significance at 0.001, ** significance at 0.01, * significance at 0.05

Definitions of firm characteristics are defined: $\ln MC_{i,t}$ is the log market capitalization for firm $i$ at time $t$ and $\ln P_{i,t}$ is the log price. $LEV_{i,t}$ is leverage measured by the ratio of total liabilities to total assets. $VOL_{i,t}$ monthly volume turnover is a valuation uncertainty proxy, which is measured as the ratio of the number of shares traded in a month and the number of shares outstanding.$ILLIQ_{i,t}$ is the Amihud (2002) measure of illiquidity, which is measured by the average ratio of the daily absolute return to the (dollar) trading volume on that day. $BidAskSpread_{i,t}$ is the average effective bid-ask spread, which is measured by the natural logarithm of the average daily effective spread.

Next, we investigate model performance as a result of external monitoring and corporate governance factors. We adopt analyst following variables as well as institutional holdings to measure the external monitoring effect. Financial analysts gather information from diverse sources, assess current firm performance, forecast future prospects, and provide buy, hold, or sell recommendations to investors. Prior literature indicates analysts mitigate information asymmetry across various dimensions, acting as external monitors for firm managers. Consequently, they impact on firms' investment, financing decisions, stock prices, liquidity, and valuation (He and Tian 2013, Hong et al. 2000, Derrien and Kecskes 2013). To investigate the effect of analysts following on model performance, we adopt the widely used analyst coverage ($Coverage_{i,t}$) (He and Tian 2013) and analyst forecast accuracy ($Accuracy_{i,t}$) (Payne 2008) measures. The $Coverage_{i,t}$ is measured as the natural logarithm of one plus the average of the 12 monthly numbers of earnings forecasts, and the numbers of forecasts are obtained from the $I/B/E/S$ summary file. Following Payne (2008), we measure analyst consensus forecast as the mean of the forecast made by each analyst for firm $i$ in year $t$ prior to the client's earnings announcement date. We then deflate the absolute differences between the earnings consensus forecast and actual earnings by the latest available stock price. A smaller forecast error in the measurement represents a greater forecast accuracy ($Accuracy_{i,t}$). Institutional holding ($InstoHold_{i,t}$) is measured as the percentage of outstanding shares held by institutional investors. It is commonly acknowledged that institutional investors are informed traders in markets responsible for impounding information in prices and contribute to market efficiency through their trading (Boehmer and Kelley 2009). It is expected that firms under greater external monitoring effect provide higher quality firm specific information to enable better model goodness-of-fit and smaller forecast errors. We adopt two additional board structure variables to measure effective corporate governance. Following Linck et al. (2008), we use board independence measured by the proportion of outside independent directors ($Independent_{i,t}$), and board leadership by a dummy variable of one if the CEO is the chair of the board ($Duality_{i,t}$).

Table 5 presents the results of mean differences for model performance when comparing firms with different levels of external monitoring and corporate governance. For model goodness-of-fit, all three methods perform significantly better for firms with higher analyst coverage ($Coverage_{i,t}$) and greater analyst forecast accuracy ($Accuracy_{i,t}$). This is consistent with studies that find that firms with greater analyst following produce more transparent firm-specific information, which improves price efficiency. In addition, the three functional methods provide greater $R^2$ for firms with a lower proportion of independent directors, while no difference is observed for the traditional CAPM method. For in-sample estimation, all four methods provide significantly smaller estimation errors for stocks with higher analyst forecast accuracy, and the three functional methods perform better for firms with higher institutional holding ($InstoHold_{i,t}$). For out-of-sample prediction, when using the functional methods, RMSPE is significantly smaller for firms with better analyst forecast accuracy ($Accuracy_{i,t}$) and higher institutional holding ($InstoHold_{i,t}$). Consistent with prior findings, the three functional methods produce much smaller RMSE than the traditional CAPM method for in-sample estimation and out-of-sample prediction.

**Table 5:** *The impact of analyst following and corporate governance on model performance.*

| Analyst following & corporate governance | | $R^2$ | | | | RMSE | | | | RMSPE | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | CAPM | FPCR | FPLSR | PFLM | CAPM | FPCR | FPLSR | PFLM | CAPM | FPCR | FPLSR | PFLM |
| $Coverage_{i,t}$ | High | 0.259 | 0.176 | 0.225 | 0.270 | 1.816 | 1.052 | 1.021 | 0.991 | 1.806 | 0.964 | 1.328 | 0.954 |
| | Low | 0.170 | 0.106 | 0.151 | 0.188 | 1.938 | 1.064 | 1.037 | 1.015 | 2.638 | 0.928 | 1.362 | 0.943 |
| | t.statistic | 4.185 | 5.366 | 5.056 | 5.393 | -0.351 | -0.200 | -0.275 | -0.413 | -1.227 | 0.622 | -0.388 | 0.197 |
| | p.value | 0.000 | 0.000 | 0.000 | 0.000 | 0.727 | 0.842 | 0.784 | 0.680 | 0.225 | 0.536 | 0.699 | 0.844 |
| | Sig. | *** | *** | *** | *** | | | | | | | | |
| $Accuracy_{i,t}$ | High | 0.185 | 0.125 | 0.171 | 0.211 | 2.195 | 1.160 | 1.131 | 1.104 | 2.059 | 0.976 | 1.409 | 0.998 |
| | Low | 0.210 | 0.142 | 0.190 | 0.233 | 2.100 | 1.105 | 1.074 | 1.047 | 2.859 | 0.966 | 1.333 | 0.975 |
| | t.statistic | -2.690 | -3.665 | -3.154 | -3.604 | 2.298 | 5.695 | 5.649 | 5.716 | 2.068 | 4.484 | 4.284 | 4.712 |
| | p.value | 0.009 | 0.000 | 0.002 | 0.001 | 0.024 | 0.000 | 0.000 | 0.000 | 0.041 | 0.000 | 0.000 | 0.000 |
| | Sig. | ** | *** | ** | *** | * | *** | *** | *** | * | *** | *** | *** |
| $InstoHold_{i,t}$ | High | 0.188 | 0.132 | 0.187 | 0.223 | 1.995 | 1.065 | 1.030 | 1.009 | 2.694 | 0.908 | 1.339 | 0.920 |
| | Low | 0.167 | 0.117 | 0.160 | 0.201 | 1.929 | 1.209 | 1.182 | 1.153 | 2.162 | 0.998 | 1.322 | 1.013 |
| | t.statistic | 0.502 | 1.885 | 0.681 | 0.844 | 4.316 | 6.615 | 6.471 | 6.453 | 2.260 | 5.939 | 2.450 | 5.573 |
| | p.value | 0.660 | 0.162 | 0.559 | 0.475 | 0.000 | 0.000 | 0.000 | 0.000 | 0.029 | 0.000 | 0.075 | 0.000 |
| | Sig. | | | | | *** | *** | *** | *** | * | *** | | *** |
| $Independent_{i,t}$ | High | 0.196 | 0.116 | 0.158 | 0.202 | 1.816 | 1.074 | 1.049 | 1.023 | 1.816 | 0.890 | 1.245 | 0.910 |
| | Low | 0.210 | 0.147 | 0.192 | 0.229 | 1.839 | 1.186 | 1.157 | 1.131 | 2.055 | 0.990 | 1.391 | 1.004 |
| | t.statistic | -1.067 | -2.556 | -2.546 | -2.806 | -1.070 | -0.851 | -0.755 | -0.715 | -1.039 | -1.702 | -1.362 | -1.232 |

Continued on next page

| Analyst following & | | $R^2$ | | | | RMSE | | | | RMSPE | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| corporate governance | | CAPM | FPCR | FPLSR | PFLM | CAPM | FPCR | FPLSR | PFLM | CAPM | FPCR | FPLSR | PFLM |
| | p.value | 0.289 | 0.013 | 0.013 | 0.006 | 0.288 | 0.398 | 0.453 | 0.477 | 0.302 | 0.093 | 0.177 | 0.222 |
| | Sig. | | * | * | ** | | | | | | | | |
| $Duality_{i,t}$ | High | 0.196 | 0.134 | 0.181 | 0.222 | 1.946 | 1.080 | 1.050 | 1.024 | 2.306 | 0.924 | 1.315 | 0.935 |
| | Low | 0.188 | 0.127 | 0.173 | 0.212 | 1.825 | 1.102 | 1.073 | 1.049 | 2.178 | 0.941 | 1.317 | 0.958 |
| | t.statistic | 0.558 | 0.769 | 0.837 | 1.082 | 0.819 | -0.515 | -0.543 | -0.602 | 0.397 | -0.478 | -0.027 | -0.613 |
| | p.value | 0.578 | 0.443 | 0.404 | 0.281 | 0.414 | 0.607 | 0.588 | 0.549 | 0.692 | 0.634 | 0.978 | 0.541 |
| | Sig. | | | | | | | | | | | | |

*** significance at 0.001, ** significance at 0.01, * significance at 0.05

Definitions of firm characteristics: $Coverage_{i,t}$ is the natural logarithm of one plus the average of 12 monthly numbers of earnings forecasts. $Accuracy_{i,t}$ is measured by the absolute analyst earnings forecast error deflated by the latest available stock price, where the forecast error is equal to analyst consensus forecast minus actual earnings. $InstoHold_{i,t}$ is the percentage of outstanding shares held by institutional investors. $Independent_{i,t}$ is the proportion of outside independent directors.$Duality_{i,t}$ is a dummy variable that equals one if the CEO is the chair of the board.

In summary, this study finds that the PFLM proposed here provide superior performance in model goodness-of-fit and in-sample fitting than the traditional CAPM, while the FPCR method produces the best out-of-sample prediction amongst the comparable models. The results suggest that the functional models have significantly better goodness-of-fit for firms that are less liquid and with higher bid-ask spread. Similar findings are evident in predictions that functional methods perform better for firms with lower monthly volume turnover, greater illiquidity, and greater bid-ask spread. Such firms are usually considered as more information opaque (Amihud 2002, Brogaard et al. 2022), so that is where we see the use of higher-frequency data and functional methods as much superior for asset pricing.

# 6 Conclusion

Given an ever-increasing amount of high-frequency financial data, we propose an extension of the CAPM in which the predictor and response are function-valued variables. We consider FPCR, FPLSR and PFLM to estimate the bivariate regression coefficient function in this concurrent function-on-function linear regression. The estimated regression coefficient function measures a linear relationship between the functional predictor and response. We can obtain fitted responses with the estimated regression coefficient function and compare their values with the holdout ones. Via the intraday and total $R^2$ and RMSE, we evaluate and compare the goodness-of-fit between a market index and its constituents using the functional CAPM. Via the RMSPE, we also study its out-of-sample forecast accuracy, which can be divided into various GICS sectors. The findings in this study suggest that the PFLM, along with high-frequency data, presents superior in-sample goodness-of-fit, while the FPCR method performs best in terms of out-of-sample predictions, followed closely by the PFLM. Compared with the traditional CAPM, the functional CAPM provides better model goodness-of-fit and prediction accuracy for less price-efficient or information-opaque stocks.

There are at least three ways in which the methodology presented can be extended. First, we used one-year intraday data, but the analysis can be conducted for a longer period. Second, the functional CAPM is an example of the concurrent function-on-function linear model. Following Corsi (2009) and Hollstein et al. (2020), one extension is to add lagged variables of the response variable $\mathcal{Y}_{t-1}^{j}(v)$, $\mathcal{Y}_{t-5}^{j}(v)$ and $\mathcal{Y}_{t-22}^{j}(v)$ representing the past daily, weekly, and monthly CIDRs of the $j^{\text{th}}$ stock. Following Qi and Luo (2019), the other extension is to consider a *nonlinear* function-on-function regression, where the beta surface can be estimated non-parametrically. These represent opportunities for further investigation.

## Acknowledgment

The authors thank participants at the International Conference on Trends and Perspectives in Linear Statistical Inference for their insightful comments and suggestions. This work was supported by an Australian Research Council Discovery Project (grant no: DP230102250) and the Scientific and Technological Research Council of Turkey (TUBITAK) (grant no: 120F270).

## Conflict of interest

The author has no competing interests to declare that are relevant to the content of this article.

# Appendix A Functional principal component regression

Functional principal component analysis extracts latent components based on the largest variance explained in each predictor and response variable (see, e.g., Shang 2014, Wang et al. 2016). Let us denote

$$C_{\mathcal{Y}^{j,c}}(v_1, v_2) = \text{cov}[\mathcal{Y}^{j,c}(v_1), \mathcal{Y}^{j,c}(v_2)]$$
$$\mathcal{C}_{\mathcal{X}^c}(u_1, u_2) = \text{cov}[\mathcal{X}^c(u_1), \mathcal{X}^c(u_2)]$$

as the empirical covariance functions of $\mathcal{Y}^{j,c}(v)$ and $\mathcal{X}^c(u)$, respectively. By Mercer's theorem (Mercer 1909), we have the following representations for the covariance functions:

$$\mathcal{C}_{\mathcal{Y}^{j,c}}(v_1, v_2) = \sum_{k\geq 1} \widehat{\lambda}_k^j \widehat{\phi}_k^j(v_1)\widehat{\phi}_k^j(v_2),$$
$$\mathcal{C}_{\mathcal{X}^c}(u_1, u_2) = \sum_{m\geq 1} \widehat{\delta}_m \widehat{\psi}_m(u_1)\widehat{\psi}_m(u_2),$$

where $\{\widehat{\phi}_k^j(v) : k = 1, 2, \ldots\}$ and $\{\widehat{\psi}_m(u) : m = 1, 2, \ldots\}$ are the empirical orthonormal eigenfunctions corresponding to the estimated eigenvalues $\{\widehat{\lambda}_1^j \geq \widehat{\lambda}_2^j \geq \ldots\}$ and $\{\widehat{\delta}_1 \geq \widehat{\delta}_2 \geq \ldots\}$. In practice, most of the variability in a functional variable can be captured by the first few eigenfunctions.

There are at least five approaches for selecting the number of retained principal components: (1) the scree plot or the fraction of variance explained by the first several functional principal components (Chiou 2012); (2) the pseudo-versions of the Akaike information criterion and Bayesian information criterion (Yao et al. 2005); (3) the cross-validation with one-curve-leave-out (Rice and Silverman 1991); (4) the bootstrap technique (Hall and Vial 2006); and (5) the eigenvalue ratio criterion (Li et al. 2020). In this study, we chose the first $K_j$ and $M$ eigenfunctions, which explain at least 95% of the total variation in the data, to project the functional response and functional predictor onto orthonormal basis expansions (Beyaztas and Shang 2023).

By Karhunen-Loève expansion (Karhunen 1947, Loève 1978), the realizations of the functional variables can be approximated by:

$$\mathcal{Y}_t^{j,c}(v) \approx \sum_{k=1}^{K_j} \widehat{c}_{t,k}^j \widehat{\phi}_k^j(v) = \widehat{\boldsymbol{C}}_t^j \widehat{\boldsymbol{\Phi}}^j(v), \tag{5}$$

$$\mathcal{X}_t^c(u) \approx \sum_{m=1}^{M} \widehat{d}_{t,m} \widehat{\psi}_m(u) = \widehat{\boldsymbol{D}}_t \widehat{\boldsymbol{\Psi}}(u), \tag{6}$$

where $\widehat{c}_{t,k}^j = \int_{\mathcal{I}} \mathcal{Y}_t^{j,c}(v)\widehat{\phi}_k^j(v)dv$ and $\widehat{d}_{t,m} = \int_{\mathcal{I}} \mathcal{X}_t^c(u)\widehat{\psi}_m(u)du$ are the projections of $\mathcal{Y}_t^{j,c}(v)$ and $\mathcal{X}_t^c(u)$ onto orthonormal basis functions, respectively. Let $\widehat{\boldsymbol{C}}_t^j = (\widehat{c}_{t,1}^j, \widehat{c}_{t,2}^j, \dots, \widehat{c}_{t,K_j}^j)$ and $\widehat{\boldsymbol{D}}_t = (\widehat{d}_{t,1}, \widehat{d}_{t,2}, \dots, \widehat{d}_{t,M})$ be two matrices of estimated principal component scores, and let $\widehat{\boldsymbol{\Phi}}^j(v) = [\widehat{\phi}_1^j(v), \widehat{\phi}_2^j(v), \dots, \widehat{\phi}_{K_j}^j(v)]^\top$ and $\widehat{\boldsymbol{\Psi}}(u) = [\widehat{\psi}_1(u), \widehat{\psi}_2(u), \dots, \widehat{\psi}_M(u)]^\top$ be two matrices of empirical orthonormal basis functions (e.g., functional principal components). In addition, the error function $\varepsilon_t^j(v)$ admits the expansion with the same basis function in $\mathcal{Y}_t^{j,c}(v)$ as follows:

$$\varepsilon_t^j(v) \approx \sum_{k=1}^{K_j} \widehat{e}_{t,k} \widehat{\phi}_k^j(v) = \widehat{\boldsymbol{e}}_t \widehat{\boldsymbol{\Phi}}^j(v),$$

where the random error function $\widehat{e}_{t,k} = \int_{\mathcal{I}} \widehat{\varepsilon}_t^j(v)\widehat{\phi}_k^j(v)dv$ is i.i.d., and denote $\widehat{\boldsymbol{e}}_t = (\widehat{e}_{t,1}, \widehat{e}_{t,2}, \dots, \widehat{e}_{t,K_j})$ and $\widehat{\boldsymbol{\Phi}}^j(v) = [\widehat{\phi}_1^j(v), \widehat{\phi}_2^j(v), \dots, \widehat{\phi}_{K_j}^j(v)]^\top$.

The bivariate coefficient function can be represented by:

$$\begin{aligned}\widehat{\beta}^j(u,v) &= \sum_{m=1}^{M} \sum_{k=1}^{K_j} \widehat{\beta}_{m,k} \widehat{\psi}_m(u) \widehat{\phi}_k^j(v) \\ &= \widehat{\boldsymbol{\Psi}}^\top(u) \widehat{\boldsymbol{\beta}}^j \widehat{\boldsymbol{\Phi}}^j(v),\end{aligned} \tag{7}$$

where $\widehat{\beta}_{m,k} = \int_{\mathcal{I}}\int_{\mathcal{I}} \widehat{\beta}^j(u,v)\widehat{\psi}_m(u)\widehat{\phi}_k^j(v)dudv$ and $\widehat{\boldsymbol{\beta}}^j$ is a $(M \times K_j)$ real-valued matrix.

By substituting the basis expansion forms of the functional variables in (5), (6) and (7) into the functional CAPM, we have the following representation:

$$\begin{aligned}\widehat{\boldsymbol{C}}_t^j \widehat{\boldsymbol{\Phi}}^j(v) &= \int_{\mathcal{I}} \widehat{\boldsymbol{D}}_t \widehat{\boldsymbol{\Psi}}(u) \widehat{\boldsymbol{\Psi}}^\top(u) \widehat{\boldsymbol{\beta}}^j \widehat{\boldsymbol{\Phi}}^j(v) du + \widehat{\boldsymbol{e}}_t \widehat{\boldsymbol{\Phi}}^j(v), \\ &= \widehat{\boldsymbol{D}}_t \widehat{\boldsymbol{\beta}}^j \widehat{\boldsymbol{\Phi}}^j(v) \int_{\mathcal{I}} \widehat{\boldsymbol{\Psi}}(u) \widehat{\boldsymbol{\Psi}}^\top(u) du + \widehat{\boldsymbol{e}}_t \widehat{\boldsymbol{\Phi}}^j(v).\end{aligned} \tag{8}$$

Due to the orthonormality property of bases $\widehat{\boldsymbol{\Phi}}^j(v)$ and $\widehat{\boldsymbol{\Psi}}(u)$, we have $= \int_{\mathcal{I}} \widehat{\boldsymbol{\Psi}}(u)\widehat{\boldsymbol{\Psi}}^\top(u)du = \mathbf{1}$.

By dividing $\widehat{\mathbf{\Phi}}^j(v)$ from both sides, (8) reduces to

$$\widehat{\boldsymbol{C}}_t^j = \widehat{\boldsymbol{D}}_t \widehat{\boldsymbol{\beta}}^j + \widehat{\boldsymbol{e}}_t, \qquad t = 1, 2, \dots, n.$$

Let $\widehat{\boldsymbol{D}} = (\widehat{\boldsymbol{D}}_1, \widehat{\boldsymbol{D}}_2, \dots, \widehat{\boldsymbol{D}}_n)$ and $\widehat{\boldsymbol{C}}^j = (\widehat{\boldsymbol{C}}_1^j, \widehat{\boldsymbol{C}}_2^j, \dots, \widehat{\boldsymbol{C}}_n^j)$, $\widehat{\boldsymbol{\beta}}^j$ can be estimated via ordinary least squares

$$\widehat{\boldsymbol{\beta}}^j = (\widehat{\boldsymbol{D}}^\top \widehat{\boldsymbol{D}})^{-1} \widehat{\boldsymbol{D}}^\top \widehat{\boldsymbol{C}}^j.$$

The estimate of the bivariate coefficient function is obtained as

$$\widehat{\beta}^j(u, v) = \widehat{\mathbf{\Psi}}^\top(u) \widehat{\boldsymbol{\beta}}^j \widehat{\mathbf{\Phi}}^j(v). \tag{9}$$

# Appendix B Functional partial least squares regression

From (2) and (3), the functional CAPM can also be expressed as:

$$\widehat{\boldsymbol{Z}}_t^j \widehat{\mathbf{\Pi}}^j(v) = \int_{\mathcal{I}} \widehat{\boldsymbol{A}}_t \widehat{\mathbf{\Gamma}}(u) \widehat{\beta}^j(u, v) du + e_t^j(v) \tag{10}$$

We multiple (10) by $\widehat{\mathbf{\Pi}}^j(v)$ on both sides, and integrating with respect to $v$, we obtain

$$\int_{\mathcal{I}} \widehat{\boldsymbol{Z}}_t^j \widehat{\mathbf{\Pi}}^j(v) \widehat{\mathbf{\Pi}}^j(v) dv = \int_{\mathcal{I}} \int_{\mathcal{I}} \widehat{\boldsymbol{A}}_t \widehat{\mathbf{\Gamma}}(u) \widehat{\beta}^j(u, v) \widehat{\mathbf{\Pi}}^j(v) du dv + \int_{\mathcal{I}} e_t^j(v) \widehat{\mathbf{\Pi}}^j(v) dv$$

Denote by $\widehat{\mathbf{\Pi}}^j = \int_{\mathcal{I}} \widehat{\mathbf{\Pi}}^j(v) [\widehat{\mathbf{\Pi}}^j(v)]^\top dv$ and $\widehat{\mathbf{\Gamma}} = \int_{\mathcal{I}} \widehat{\mathbf{\Gamma}}(u) \widehat{\mathbf{\Gamma}}^\top(u) du$ the symmetric matrices of the inner products of the $B$-spline basis functions. Let $\epsilon_t^j = \int_{\mathcal{I}} e_t^j(v) \widehat{\mathbf{\Pi}}^j(v) dv$. From (9), we observe $\widehat{\beta}^j(u, v) = \widehat{\mathbf{\Gamma}}^\top(u) \widehat{\boldsymbol{\beta}}^j \widehat{\mathbf{\Pi}}^j(v)$, then

$$\begin{aligned} \widehat{\boldsymbol{Z}}_t^j \widehat{\mathbf{\Pi}}^j &= \widehat{\boldsymbol{A}}_t \widehat{\boldsymbol{\beta}}^j \widehat{\mathbf{\Pi}}^j \widehat{\mathbf{\Gamma}} + \epsilon_t^j \\ \widehat{\boldsymbol{Z}}_t^j (\widehat{\mathbf{\Pi}}^{\frac{1}{2}})^j [(\widehat{\mathbf{\Pi}}^{\frac{1}{2}})^j]^\top &= \widehat{\boldsymbol{A}}_t \widehat{\boldsymbol{\beta}}^j (\widehat{\mathbf{\Pi}}^{\frac{1}{2}})^j [(\widehat{\mathbf{\Pi}}^{\frac{1}{2}})^j]^\top (\widehat{\mathbf{\Gamma}}^{\frac{1}{2}}) [(\widehat{\mathbf{\Gamma}}^{\frac{1}{2}})]^\top + \epsilon_t^j. \end{aligned} \tag{11}$$

By multiplying both sides of (11) by $[(\mathbf{\Pi}^{-\frac{1}{2}})^j]^\top$, we obtain

$$\begin{aligned}\widehat{\boldsymbol{Z}}_t^j(\widehat{\mathbf{\Pi}}^{\frac{1}{2}})^j &= \widehat{\boldsymbol{A}}_t\widehat{\boldsymbol{\beta}}^j(\widehat{\mathbf{\Pi}}^{\frac{1}{2}})^j(\widehat{\mathbf{\Gamma}}^{\frac{1}{2}})(\widehat{\mathbf{\Gamma}}^{\frac{1}{2}})^\top + \epsilon_t^j \times [(\mathbf{\Pi}^{-\frac{1}{2}})^j]^\top \\ &= \widehat{\boldsymbol{A}}_t\widehat{\mathbf{\Gamma}}^{\frac{1}{2}}\widehat{\beta}^j(\widehat{\mathbf{\Pi}}^{\frac{1}{2}})^j(\widehat{\mathbf{\Gamma}}^{\frac{1}{2}})^\top + \tilde{\epsilon}_t^j.\end{aligned}$$

Denote by $\widehat{\boldsymbol{Z}}^j = (\widehat{Z}_1^j, \widehat{Z}_2^j, \ldots, \widehat{Z}_n^j)$, $\widehat{\boldsymbol{A}} = (\widehat{A}_1, \widehat{A}_2, \ldots, \widehat{A}_n)$ and $\tilde{\boldsymbol{\epsilon}}^j = (\tilde{\epsilon}_1^j, \tilde{\epsilon}_2^j, \ldots, \tilde{\epsilon}_n^j)$. Using a multivariate partial least squares, we obtain

$$(\widehat{\boldsymbol{Z}}^j)^\top(\widehat{\mathbf{\Pi}}^{\frac{1}{2}})^j = \widehat{\boldsymbol{A}}^\top\widehat{\mathbf{\Gamma}}^{\frac{1}{2}}\mathbf{\Omega}^j + \tilde{\boldsymbol{\epsilon}}^{\boldsymbol{j}}, \tag{12}$$

where

$$\mathbf{\Omega}^j = \widehat{\boldsymbol{\beta}}^j(\widehat{\mathbf{\Pi}}^{\frac{1}{2}})^j(\widehat{\mathbf{\Gamma}}^{\frac{1}{2}})^\top, \tag{13}$$

and $\tilde{\boldsymbol{\epsilon}}^{\boldsymbol{j}}$ represent the coefficient and residual matrices, respectively. From (12), we apply the multivariate partial least-squares method of Beyaztas and Shang (2020) to estimate the coefficient matrix $\mathbf{\Omega}^j$. From (13), we obtain the estimated $\widehat{\boldsymbol{\beta}}^j$,

$$\widehat{\boldsymbol{\beta}}^j = \widehat{\mathbf{\Gamma}}^{-\frac{1}{2}}\mathbf{\Omega}^j(\widehat{\mathbf{\Pi}}^{-\frac{1}{2}})^j. \tag{14}$$

The bivariate regression coefficient function $\beta(u, v)$ in the functional CAPM can be estimated from (9) and (14) as follows:

$$\widehat{\beta}^j(u, v) = \widehat{\mathbf{\Gamma}}^\top(u)\left[\widehat{\mathbf{\Gamma}}^{-\frac{1}{2}}\mathbf{\Omega}^j(\widehat{\mathbf{\Pi}}^{-\frac{1}{2}})^j\right]\widehat{\mathbf{\Pi}}^j(v).$$

# Appendix C Penalized function-on-function regression

A penalized function-on-function (PFLM) regression methodology is adopted to estimate the bivariate regression coefficient function $\beta^j(u, v)$. In this approach, the regularized estimate of $\beta^j(u, v)$ can be obtained by minimizing the objective function:

$$\underset{\beta^j(u,b)}{\arg\min} \sum_j \left[\mathcal{Y}_t^{j,c}(v) - \int_{\mathcal{I}} \beta^j(u, v)\mathcal{X}_t^c(u)du\right]^2 + \frac{\kappa}{2}\mathcal{J}(\beta),$$

where $\mathcal{J}$ represents a roughness penalty on $\beta^j(u,v)$, and $\kappa$ serves as the smoothing parameter controlling the degree of shrinkage in $\beta^j(u,v)$.

As outlined in Section 3.1, we make the assumption that $\mathcal{Y}_t^{j,c}(v)$ and $\mathcal{X}_t^c(u)$ are densely observed, such that $\mathcal{Y}_t^{j,c}(v_i) = \mathcal{Y}_t^{j,c}(v_j)$ and $\mathcal{X}_t^c(u_i) = \mathcal{X}_t^c(u_j)$ for $i = 1,\dots,78$. The regression coefficient function is assumed to be represented by tensor product B-spline basis functions $\{\widehat{\pi}_k^j(v), k = 1,\dots,20\}$ and $\{\widehat{\gamma}_m(u), m = 1,\dots,20\}$, given by:

$$\beta^j(u,v) = \sum_{k=1}^{20}\sum_{m=1}^{20} \widehat{b}_{km}\widehat{\pi}_k^j(v)\widehat{\gamma}_m(u),$$

where $\widehat{b}_{km}$ represents the B-spline basis expansion coefficients. Let $\Delta_r$ denote the length of the $r^{\text{th}}$ interval in $\mathcal{I}$, such that $\Delta_r = i_{r+1} - i_r$. Following a similar approach to Ivanescu et al. (2015), we employ numerical integration to approximate $\int_{\mathcal{I}} \beta^j(u,s)\mathcal{X}_t^c(u)du$ as follows:

$$\begin{aligned}\int_{\mathcal{I}} \beta^j(u,v)\mathcal{X}_t^c(u)du &\approx \sum_{i=1}^{19} \Delta_r \beta^j(u_i,v)\mathcal{X}_t^c(u_i) \\ &= \sum_{r=1}^{19} \Delta_r \sum_{k=1}^{20}\sum_{m=1}^{20} \widehat{b}_{km}\widehat{\pi}_k^j(v)\widehat{\gamma}_m(u_i)X_t^c(u_i) \\ &= \sum_{k=1}^{20}\sum_{m=1}^{20} \widehat{b}_{km}\widehat{\pi}_k^j(v)\widetilde{\widehat{\gamma}}_m, \end{aligned} \tag{15}$$

where $\widetilde{\widehat{\gamma}}_m = \sum_{r=1}^{19} \Delta_r\widehat{\gamma}_m(u_i)X_t^c(u_i)$. Then, the functional CAPM model is approximated as follows:

$$\mathcal{Y}_t^{j,c}(v) = \sum_{k=1}^{20}\sum_{m=1}^{20} \widehat{b}_{km}\widehat{\pi}_k^j(v)\widetilde{\widehat{\gamma}}_m + \varepsilon_t^{j,c}(v).$$

Consider a $20 \times 20$ dimensional matrix of basis expansion coefficients denoted by $\widehat{\boldsymbol{b}} = (\widehat{b}_{km})_{km}$. The penalty functional $\mathcal{J}(\beta)$ is approximated in the following manner:

$$\begin{aligned}\widetilde{\mathcal{J}}(\beta) &= \int_{\mathcal{I}}\int_{\mathcal{I}} \left[\frac{\partial^2}{\partial v^2}\beta(u,v)\right]^2 dudv + \int_{\mathcal{I}}\int_{\mathcal{I}} \left[\frac{\partial^2}{\partial u^2}\beta(u,v)\right]^2 dudv \\ &= \widehat{\boldsymbol{b}}^\top(\widehat{\boldsymbol{\Gamma}} \otimes \boldsymbol{P}_y + \boldsymbol{P}_x \otimes \widehat{\boldsymbol{\Pi}}^j)\widehat{\boldsymbol{b}}, \end{aligned} \tag{16}$$

where $\widehat{\boldsymbol{\Pi}}^j = \int_{\mathcal{I}} \widehat{\boldsymbol{\Pi}}^j(v)(\widehat{\boldsymbol{\Pi}}^j)^\top(v)dv$, $\widehat{\boldsymbol{\Gamma}} = \int_{\mathcal{I}} \widehat{\boldsymbol{\Gamma}}(u)(\widehat{\boldsymbol{\Gamma}})^\top(u)du$, and $\boldsymbol{P}_y$ and $\boldsymbol{P}_x$ are the penalty matrices, with $(kk')^{\text{th}}$ and $(mm')^{\text{th}}$ entries; $P_{y,kk'} = \int_{\mathcal{I}} (\widehat{\pi}_k^j)^{(2)}(v)(\widehat{\pi}_{k'}^j)^{(2)}(v)dv$ and $P_{x,mm'} = \int_{\mathcal{I}} \widehat{\gamma}_m^{(2)}(u)\widehat{\gamma}_{m'}^{(2)}(u)du$ for $k,k',m,m' = 1,\dots,20$.

By employing the estimated integral in (15) and the approximated penalty functional described in equation (16), the estimation of $\widehat{\boldsymbol{b}}$ can be achieved by minimizing:

$$\underset{\widehat{\boldsymbol{b}}}{\arg\min} \sum_j \sum_{i=1}^{78} \left[ \mathcal{Y}_t^{j,c}(v_i) - (\widetilde{\widetilde{\gamma}}^\top \otimes (\widehat{\boldsymbol{\Pi}}^j)^\top (v_i))\widehat{\boldsymbol{b}} \right]^2 + \frac{\kappa}{2} \widetilde{\mathcal{J}}(\beta), \tag{17}$$

where $\widetilde{\widetilde{\gamma}} = [\widetilde{\widetilde{\gamma}}_1, \ldots, \widetilde{\widetilde{\gamma}}_{20}]^T$. Accordingly, the regularized estimate of $\beta^j(u,v)$ is obtained as follows:

$$\widehat{\beta}^j(u,v) = (\widehat{\boldsymbol{\Gamma}}(u) \otimes (\widehat{\boldsymbol{\Pi}}^j)^\top (v))\widehat{\boldsymbol{b}}^*,$$

where $\widehat{\boldsymbol{b}}^*$ is the estimates of $\widehat{\boldsymbol{b}}$ obtained by minimizing (17).

The estimation of $\widehat{\boldsymbol{b}}$ is achieved through a penalized least squares approach. In this method, determining the optimal value of $\kappa$ is crucial for efficient estimation results. Various information criteria, including the Bayesian information criterion (BIC), generalized cross-validation, and modified Akaike information criterion, are available for this purpose. We propose using BIC to identify the optimal smoothing parameter due to its simplicity and computational efficiency. The BIC for determining the optimal smoothing parameter can be computed as follows:

$$\text{BIC}(\kappa) = N \times \ln \left\| \sum_j \left[ \mathcal{Y}_t^{j,c}(v) - \widehat{\mathcal{Y}}_{t,\kappa}^{j,c}(v) \right] \right\|_2^2 + \ln(N),$$

where $\widehat{\mathcal{Y}}_{t,\kappa}^{j,c}(v)$ represents the estimate of $\mathcal{Y}_t^{j,c}(v)$ with the smoothing parameter $\kappa$. It's important to note that the optimal value of the penalty parameter $\kappa$ is determined using a standard grid-search approach with a predefined set of candidate values for $\kappa$.

# References


Aguilera, A. M., Escabias, M., Preda, C. and Saporta, G. (2010), 'Using basis expansions for estimating functional PLS regression applications with chemometric data', *Chemometrics and Intelligent Laboratory Systems* **104**, 289–305.

Amihud, Y. (2002), 'Illiquidity and stock returns: Cross-section and time-series effects', *Journal of Financial Markets* **5**(1), 31–56.

Andersen, T. G. (2000), 'Some reflections on analysis of high-frequency data', *Journal of Business and Economic Statistics* **18**(2), 146–153.

Andersen, T. G., Bollerslev, T., Diebold, F. X. and Labys, P. (2001), 'The distribution of realized exchange rate volatility', *Journal of the American Statistical Association: Applications and Case Studies* **96**, 42–55.

Andersen, T. G., Bollerslev, T., Diebold, F. X. and Labys, P. (2003), 'Modeling and forecasting realized volatility', *Econometrica* **71**, 579–625.

Andersen, T. G., Bollerslev, T., Diebold, F. X. and Wu, J. (2004), Realized beta: Persistence and predictability, *in* T. B. Fomby and D. Terrell, eds, 'Econometric Analysis of Financial and Economic Time Series', Emerald Group Publishing Limited, Bingley, pp. 1–39.

Andersen, T. G., Su, T., Todorov, V. and Zhang, Z. (2024), 'Intraday periodic volatility curves', *Journal of the American Statistical Association: Theory and Methods* **119**(546), 1181–1191.

Andersen, T. G., Thyrsgaard, M. and Todorov, V. (2021), 'Recalcitrant betas: Intraday variation in the cross-sectional dispersion of systematic risk', *Quantitative Economics* **12**, 647–682.

Andreou, E. and Ghysels, E. (2002), 'Rolling sample volatility estimators: Some new theoretical, simulation and empirical results', *Journal of Business and Economic Statistics* **20**, 363–376.

Aneiros-Pérez, G. and Vieu, P. (2006), 'Semi-functional partial linear regression', *Statistics and Probability Letters* **76**, 1102–1110.

Ang, A. and Chen, J. C. (2003), 'CAPM over the long run: 1926–2001', *Journal of Empirical Finance* **14**, 1–40.

Aue, A., Hörmann, S., Horváth, L., Hušková, M. and Steinebach, J. G. (2012), 'Sequential testing for the stability of high-frequency portfolio betas', *Econometric Theory* **28**, 804–837.

Bali, T. G., Cakici, N. and Tang, Y. (2009), 'The conditional beta and the cross-section of expected returns', *Financial Management* **38**(1), 103–137.

Bali, T. G., Engle, R. F. and Tang, Y. (2017), 'Dynamic conditional beta is alive and well in the cross section of daily stock returns', *Management Science* **63**, 3531–3997.

Ball, R. and Kothari, S. P. (1989), 'Nonstationary expected returns: Implications for tests of market efficiency and serial correlation in returns', *Journal of Financial Economics* **25**, 51–74.

Bera, A., Bubnys, E. and Park, H. (1988), 'Conditional heteroscedasticity in the market model and efficient estimates of betas', *The Financial Review* **23**(2), 201–214.

Bernardo, A. E., Chowdhry, B. and Goyal, A. (2007), 'Growth options, beta, and the cost of capital', *Financial Management* **36**(2), 5–17.

Beyaztas, U. and Shang, H. L. (2020), 'On function-on-function regression: Partial least squares approach', *Environmental and Ecological Statistics* **27**, 95–114.

Beyaztas, U. and Shang, H. L. (2022), 'A robust functional partial least squares for scalar-on-multiple-function regression', *Journal of Chemometrics* **36**(4), e3394.

Beyaztas, U. and Shang, H. L. (2023), 'Robust functional linear regression models', *The R Journal* **15**(1), 212–233.

Beyaztas, U., Shang, H. L. and Saricam, S. (2024), 'Penalized function-on-function linear quantile regression', *Computational Statistics* **in-press**.

Boehmer, E. and Kelley, E. (2009), 'Institutional investors and the information efficiency of prices', *Review of Financial Studies* **22**(9), 3563–3594.

Boguth, O., Carlson, M., Fisher, A. and Simutin, M. (2016), 'Horizon effects in average returns: The role of slow information diffusion', *Review of Financial Studies* **29**(8), 2241–2281.

Bollerslev, T., Engle, R. F. and Wooldridge, J. M. (1988), 'A capital asset pricing model with time-varying covariances', *Journal of Political Economy* **96**, 116–131.

Bollerslev, T., Li, L. Z. and Todorov, V. (2016), 'Roughing up beta: Continuous versus discontinuous betas and the cross section of expected stock returns', *Journal of Financial Economics* **120**, 464–490.

Bollerslev, T., Patton, A. J. and Quaedvlieg, R. (2022), Granular betas and risk premium functions, Technical report, Duke University.
**URL:** https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4258489

Bollerslev, T. and Zhang, B. Y. B. (2003), 'Measuring and modelling systematic risk in factor pricing models using high-frequency data', *Journal of Empirical Finance* **10**, 533–558.

Braun, P. A., Nelson, D. B. and Sunier, A. M. (1995), 'Good news, bad news, volatility, and betas', *Journal of Finance* **50**, 1575–1603.

Brogaard, J., Nguyen, T. H., Putnins, T. J. and Wu, E. (2022), 'What moves stock prices? The roles of news, noise, and information', *The Review of Financial Studies* **35**(9), 4341–4386.

Campbell, J. Y., Lo, A. W. and MacKinlay, A. C. (1997), *The Econometrics of Financial Markets*, Princeton University Press, New Jersey.

Casabona, P. A. and Vora, A. (1982), 'The bias of conventional risk premiums in empirical tests of the capital asset pricing model', *Financial Management* **11**(2), 90–96.

Cenesizoglu, T., Liu, Q., Reeves, J. J. and Wu, H. (2016), 'Monthly beta forecasting with low-, medium- and high-frequency stock returns', *Journal of Forecasting* **35**(6), 528–541.

Chiou, J.-M. (2012), 'Dynamical functional prediction and classification, with application to traffic flow prediction', *The Annals of Applied Statistics* **6**(4), 1588–1614.

Corsi, F. (2009), 'A simple approximate long-memory model of realized volatility', *Journal of Financial Econometrics* **7**(2), 174–196.

Crainiceanu, C. M., Staicu, A.-M. and Di, C. (2009), 'Generalized multilevel functional regression', *Journal of the American Statistical Association: Theory and Methods* **104**(488), 177–194.

Delaigle, A. and Hall, P. (2012), 'Methodology and theory for partial least squares applied to functional data', *The Annals of Statistics* **40**(1), 322–352.

Derrien, F. and Kecskes, A. (2013), 'The real effects of financial shocks: Evidence from exogenous changes in analyst coverage', *Journal of Finance* **68**(4), 1407–1440.

Dimson, E. (1979), 'Risk measurement when shares are subject to infrequent trading', **7**, 197–226.

Donoho, D. L. (2000), High-dimensional data analysis: The curses and blessings of dimensionality, *in* 'AMS Conference on Math Challenges of the 21st Century', Los Angeles, CA.
**URL:** https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.329.3392

D'Souza, R. E., Brooks, L. D. and Oberhelman, H. D. (1989), 'A general stationary stochastic regression model for estimating and predicting beta', *The Financial Review* **24**(2), 299–317.

Ellis, D. M. (1996), 'A test of the conditional CAPM with simultaneous estimation of the first and second conditional moments', *The Financial Review* **31**(3), 475–498.

Engle, R. F. (2000), 'The econometrics of ultra-high-frequency data', *Econometrica* **68**(1), 1–22.

Fama, E. F. and French, K. R. (1989), 'Business conditions and expected returns on stocks and bonds', *Journal of Financial Economics* **25**, 23–49.

Ferraty, F. and Vieu, P. (2006), *Nonparametric Functional Data Analysis*, Springer, New York.

Fletcher, J. (2002), 'Examination of conditional asset pricing in UK stock returns', *The Financial Review* **37**(3), 447–468.

Garcia-Portugués, E., Alvarez-Liébana, J. and G. Alvarez-Pérez, W. G.-M. (2021), 'A goodness-of-fit test for the functional linear model with functional response', *Scandinavian Journal of Statistics* **48**(2), 502–528.

Gençay, R., Dacorogna, M. M., Muller, U. A., Pictet, O. and Olsen, R. (2001), *An Introduction to High-Frequency Finance*, Academic Press, San Diego, CA.

Ghysels, E. (1998), 'On stable factor structure in the pricing of risk: Do time-varying betas help or hurt?', *Journal of Finance* **53**, 549–573.

Ghysels, E. (2000), 'Some econometrics recipes for high-frequency data cooking', *Journal of Business and Economic Statistics* **18**(2), 154–163.

Gilbert, T., Hrdlicka, C., Kalodimos, J. and Siegel, S. (2014), 'Daily data is bad for beta: Opacity and frequency-dependent betas', *Review of Asset Pricing Studies* **4**, 78–117.

Goldstein, M. A., Kumar, P. and Graves, F. C. (2014), 'Computerized and high-frequency trading', *The Financial Review* **49**(2), 177–202.

Goodhart, C. A. E. and O'Hara, M. (1997), 'High frequency data in financial markets: Issues and applications', *Journal of Empirical Finance* **4**(2-3), 73–114.

Gouriéroux, C. and Jasiak, J. (2001), *Financial Econometrics: Problems, Models and Methods*, Princeton University Press, New Jersey.

Hall, P. and Vial, C. (2006), 'Assessing the finite dimensionality of functional data', *Journal of the Royal Statistical Society: Series B* **68**(4), 689–705.

Hays, S., Shen, H. and Huang, J. Z. (2012), 'Functional dynamic factor models with application to yield curve forecasting', *The Annals of Applied Statistics* **6**(3), 870–894.

He, J. and Tian, X. (2013), 'The dark side of analyst coverage: The case of innovation', *Journal of Financial Economics* **109**(3), 856–878.

Hollstein, F., Prokopczuk, M. and Simen, C. W. (2020), 'The conditional capital asset pricing model revisited: Evidence from high-frequency betas', *Management Science* **66**(6), 2474–2494.

Hong, H., Lim, T. and Stein, J. (2000), 'Bad news travels slowly: Size, analyst coverage, and the profitability of momentum strategies', *Journal of Finance* **55**(1), 265–295.

Hörmann, S. and Kokoszka, P. (2012), Functional time series, *in* T. S. Rao, S. S. Rao and C. R. Rao, eds, 'Handbook of Statistics', Vol. 30, North Holland, Amsterdam, pp. 157–186.

Horváth, L., Kokoszka, P. and Rice, G. (2014), 'Testing stationarity of functional time series', *Journal of Econometrics* **179**(1), 66–82.

Ivanescu, A. E., Staicu, A. M., Scheipl, F. and Greven, S. (2015), 'Penalized function-on-function regression', *Computational Statistics* **30**(2), 539–568.

Jagannathan, R. and Meier, I. (2002), 'Do we need capm for capital budgeting', *Financial Management* **31**(4), 55–77.

Jagannathan, R. and Wang, Z. (1996), 'The conditional CAPM and the cross-section of expected returns', *Journal of Finance* **51**, 3–53.

Karhunen, K. (1947), 'Über lineare methoden in der wahrscheinlichkeitsrechnung', *Annales Academiae Scientiarum Fennicae* **37**, 1–79.

Kokoszka, P. and Reimherr, M. (2017), *Introduction to Functional Data Analysis*, Chapman and Hall/CRC, Boca Raton.

Kokoszka, P., Rice, G. and Shang, H. L. (2017), 'Inference for the autocovariance of a functional time series under conditional heteroscedasticity', *Journal of Multivariate Analysis* **162**, 32–50.

Kokoszka, P. and Zhang, X. (2012), 'Functional prediction of intraday cumulative returns', *Statistical Modelling* **12**(4), 377–398.

Kumar, A. (2009), 'Hard-to-value stocks, behavioral biases, and informed trading', *Journal of Financial and Quantitative Analysis* **44**(6), 1375–1401.

Lewellen, J. and Nagel, S. (2006), 'The conditional capm does not explain asset-pricing anomalies', *Journal of Financial Economics* **82**, 289–314.

Li, D., Robinson, P. M. and Shang, H. L. (2020), 'Long-range dependent curve time series', *Journal of the American Statistical Association: Theory and Methods* **115**(530), 957–971.

Linck, J., Netter, J. and Yang, T. (2008), 'The determinants of board structure', *Journal of Financial Economics* **87**(2), 308–328.

Linter, J. (1965), 'The valuation of risk assets and the selection of risky investments in stock portfolio and capital budgets', *Review of Economics and Statistics* **47**, 13–37.

Loève, M. (1978), *Probability Theory*, Vol. II of *Graduate Texts in Mathematics*, 4th edn, Springer-Verlag, New York.

Lyons, R. (2001), *The Microstructure Approach to Exchange Rates*, MIT Press, Cambridge, Massachusetts.

Mercer, J. (1909), 'Functions of positive and negative type, and their connection with the theory of integral equations', *Philosophical Transactions of the Royal Society A* **209**(441-458), 415–446.

Papavassiliou, V. G. (2013), 'A new method for estimating liquidity risk: Insights from a liquidity-adjusted CAPM framework', *Journal of International Financial Markets, Institutions & Money* **24**, 184–197.

Payne, J. (2008), 'The influence of audit firm specialization on analysts' forecast errors', *Auditing: A Journal of Practice & Theory* **27**(2), 109–136.

Peña, D. and Rsay, R. S. (2021), *Statistical Learning for Big Dependent Data*, Wiley, Hoboken, New Jersey.

Pedersen, J. (2022), Functional data analysis: Estimating beta in the capital asset pricing model for high-frequency U.S. stock data, Technical report, University of Copenhagen.
**URL:** https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4310423

Qi, X. and Luo, R. (2019), 'Nonlinear function-on-function additive model with multiple predictor curves', *Statistica Sinica* **29**(2), 719–739.

Ramsay, J. O. and Dalzell, C. J. (1991), 'Some tools for functional data analysis', *Journal of the Royal Statistical Society, Series B* **53**(3), 539–572.

Ramsay, J. O. and Silverman, B. W. (2005), *Functional Data Analysis*, 2nd edn, Springer, New York.

Rice, G., Wirjanto, T. and Zhao, Y. (2020), 'Tests for conditional heteroscedasticity of functional data', *Journal of Time Series Analysis* **41**, 733–758.

Rice, J. A. and Silverman, B. W. (1991), 'Estimating the mean and covariance structure nonparametrically when the data are curves', *Journal of the Royal Statistical Society: Series B* **53**(1), 233–243.

Saricam, S., Beyaztas, U., Asikgil, B. and Shang, H. L. (2022), 'On partial least-squares estimation in scalar-on-function regression models', *Journal of Chemometrics* **36**(12), e3452.

Shang, H. L. (2014), 'A survey of functional principal component analysis', *AStA Advances in Statistical Analysis* **98**, 121–142.

Shang, H. L. (2017), 'Forecasting intraday S&P 500 index returns: A functional time series approach', *Journal of Forecasting* **36**(7), 741–755.

Shang, H. L., Cao, J. and Sang, P. (2022), 'Stopping time detection of wood panel compression: A functional time-series approach', *Journal of the Royal Statistical Society: Series C* **71**(5), 1205–1224.

Shang, H. L., Yang, Y. and Kearney, F. (2019), 'Intraday forecasts of a volatility index: Functional time series methods with dynamic updating', *Annals of Operations Research* **282**(1-2), 331–354.

Sharpe, W. F. (1964), 'Capital asset prices: A theory of market equilibrium under conditions of risk', *Journal of Finance* **19**, 425–442.

Tsay, R. S. (2010), *Analysis of Financial Time Series*, 3rd edn, John Wiley & Sons, Inc, Hoboken, New Jersey.

Wang, H. and Cao, J. (2023), 'Nonlinear prediction of functional time series', *Environmetrics* **34**(5), e2792.

Wang, J.-L., Chiou, J.-M. and Müller, H.-G. (2016), 'Functional data analysis', *Annual Review of Statistics and Its Application* **3**, 257–295.

Wang, K. Q. (2003), 'Asset pricing with conditioning information: A new test', *Journal of Finance* **58**, 161–196.

Wang, S., Jank, W. and Shmueli, G. (2008), 'Explaining and forecasting online auction prices and their dynamics using functional data analysis', *Journal of Business and Economic Statistics* **26**(2), 144–160.

Wiggins, J. B. (1992), 'Betas in up and down markets', *The Financial Review* **27**(1), 107–123.

Wood, R. A. (2000), 'Market microstructure research databases: History and projections', *Journal of Business and Economic Statistics* **18**(2), 140–145.

Yao, F., Müller, H.-G. and Wang, J.-L. (2005), 'Functional data analysis for sparse longitudinal data', *Journal of the American Statistical Association: Theory & Methods* **100**(470), 577–590.

Zhou, J. and Paseka, A. (2017), 'Unconditional tests of linear asset pricing models with time-varying betas', *The Financial Review* **52**(3), 373–404.